Accepted Manuscript

Geometrical and topological approaches to Big Data

Václav Snášel, Jana Nowaková, Fatos Xhafa, Leonard Barolli

PII: S0167-739X(16)30185-6
DOI: http://dx.doi.org/10.1016/j.future.2016.06.005
Reference: FUTURE 3071

To appear in: *Future Generation Computer Systems*

Received date: 6 March 2016
Revised date: 25 May 2016
Accepted date: 6 June 2016

Please cite this article as: V. Snášel, J. Nowaková, F. Xhafa, L. Barolli, Geometrical and topological approaches to Big Data, *Future Generation Computer Systems* (2016), http://dx.doi.org/10.1016/j.future.2016.06.005

Highlights

- An overview of state-of-the-art in Geometrical and Topological Approach to Big Data
- Trends in Geometrical and Topological Approach to Big Data
- Big Data Visualization
- Discussion of Current Techniques and Future Trends to Address the Applications

Geometrical and Topological Approaches to Big Data☆

Václav Snášel[a], Jana Nowaková[a], Fatos Xhafa[b], Leonard Barolli[c]

[a]*Department of Computer Science, Faculty of Electrical Engineering and Computer Science, VŠB - Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava - Poruba, Czech Republic*
[b]*Department of Computer Science, Technical University of Catalonia, C/Nord, Omega Bld, C/Jordi Girona 1-3, 08034 Barcelona, Spain*
[c]*Department of Information and Communication Engineering, Faculty of Information Engineering, Fukuoka Institute of Technology (FIT), 3-30-1 Wajiro-higashi, Higashi-ku, Fukuoka 811-0295, Japan*

Abstract

Modern data science uses topological methods to find the structural features of data sets before further supervised or unsupervised analysis. Geometry and topology are very natural tools for analysing massive amounts of data since geometry can be regarded as the study of distance functions. Mathematical formalism, which has been developed for incorporating geometric and topological techniques, deals with point cloud data sets, i.e. finite sets of points. It then adapts tools from the various branches of geometry and topology for the study of point cloud data sets. The point clouds are finite samples taken from a geometric object, perhaps with noise. Topology provides a formal language for qualitative mathematics, whereas geometry is mainly quantitative. Thus, in topology, we study the relationships of proximity or nearness, without using distances. A map between topological spaces is called continuous if it preserves the nearness structures. Geometrical and topological methods are tools allowing us to analyse highly complex data. These methods create a summary or compressed representation of all of the data features to help rapidly uncover particular patterns and relationships in data. The idea of constructing summaries of entire domains of attributes involves understanding the relationship between topological and geometric objects constructed from data using various features.

A common thread in various approaches for noise removal, model reduction, feasibility reconstruction, and blind source separation, is to replace the original data with a lower dimensional approximate representation obtained via a matrix or multi-directional array factorization or decomposition. Besides those transformations, a significant challenge of feature summarization or subset selection methods for Big Data will be considered by focusing on scalable feature selection. Lower dimensional approximate representation is used for Big Data visualization.

The cross-field between topology and Big Data will bring huge opportunities, as well as challenges, to Big Data communities. This survey aims at bringing together state-of-the-art research results on geometrical and topological methods for Big Data.

Keywords:
Big data, Industry 4.0, Topological data analysis, Persistent homology, Dimensionality reduction, Big Data visualization

1. Introduction

Big Data is everywhere as high volumes of varieties of valuable precise and uncertain data can be easily collected or generated at high velocity in various real-life applications. The explosive growth in web-based storage, management, processing, and accessibility of social, medical, scientific and engineering data has been driven by our need for fundamental understanding of the processes which produce this data. It is predicted that volume of the produced data could reach 44 zettabytes in 2020 [9]. The enormous volume and complexity of this data propel technological advancements realized as exponential increases in storage capability, processing power, bandwidth capacity and transfer velocity. This is, partly, because of new experimental methods, and in part because of the increase in the availability of high-powered computing technology. Massive amounts of data (Big Data) are too complex to be managed by tra-

Email addresses: vaclav.snasel@vsb.cz (Václav Snášel), jana.nowakova@vsb.cz (Jana Nowaková), fatos@cs.upc.edu (Fatos Xhafa), barolli@fit.ac.jp (Leonard Barolli)

ditional processing applications. Nowadays, it includes the huge, complex, and abundant structured and unstructured data that is generated and gathered from several fields and resources. The challenges of managing massive amounts of data include extracting, analysing, visualizing, sharing, storing, transferring and searching such data. Currently, traditional data processing tools and their applications are not capable of managing Big Data. Therefore, there is a critical need to develop effective and efficient Big Data processing techniques. Big Data has five characteristics: volume, velocity, variety, veracity and value [15]. Volume refers to the size of the data for processing and analysis. Velocity relates to the rate of data growth and usage. Variety means the different types and formats of the data used for processing and analysis. Veracity concerns the accuracy of results and analysis of the data. Value is the added value and contribution offered by data processing and analysis.

Modern data science uses so-called topological methods to find the structural features of data sets before further supervised or unsupervised analysis. Geometry and topology are very natural tools for analysing massive amounts of data since geometry can be regarded as the study of distance functions. Besides the heterogeneity of distance functions, another issue is related to distance functions on large finite sets of data. The mathematical formalism which has been developed for incorporating geometric and topological techniques deals with point clouds, i.e. finite sets of points equipped with proximity or nearness or distance functions [10, 11]. It then adapts tools from the various branches of geometry and topology for the study of point clouds [14]. The point clouds are finite samples taken from a geometric object, perhaps with noise.

Geometrical and Topological methods are tools for analysing highly complex data [10]. These methods create a summary or a compressed representation of all of the data features to help rapidly uncover patterns and relationships in data. The idea of constructing summaries of entire domains of parameter values involves understanding the relationship between geometric objects constructed from data using various parameter values e.g. [103].

One problem with Big Data analysis, which is very actual, is that the currently used methods based on model creation, simulation of the created model and then assessment, whether the original data corresponds to data obtained using the created model - model verification cannot be applied. The described process is useful and appropriate for solving classic problems such as physical problems, because the theoretical background of these problems has been researched and understood enough, so it could be reconstructed to fit the model. For Big Data processing, the first problem is that we are not able to define the concrete hypothesis of the data feature which could be tested. Due to this, for the Big Data problem, the same approach as with the classic physical problem cannot be used. Therefore, the main aim of the research is not to define a model, but to be able to mine accurately and automatically interesting features of Big Data sets. In many cases, the data to be examined is often based on shapes that are not easy to capture using traditional methods [68].

A common thread in various approaches for noise removal, model reduction, feasibility reconstruction, and blind source separation, is to replace the original data with a lower dimensional approximate representation obtained via a matrix or multi-directional array factorization or decomposition. Besides those transformations, a significant challenge of feature summarization or subset selection methods for Big Data will be considered by focusing on scalable feature selection. Lower dimensional approximate representation is used for Big Data visualization to be able to visualize data in understandable form. This approach – dimensionality reduction can be also understood as a method for feature compression, see Fig. 6.

The whole paper is organized as follows: in Section 2, a brief introduction to Big Data technologies is given. In the next Section 3, a brief motivational example is presented. A short mathematical background is introduced in Section 4. This part contains a brief review of topology, metric space, homology and persistent homology theory, manifolds and Morse theory. In the following Section 5, a brief review of homology and persistent homology theory is introduced. Various applications of geometrical and topological methods are presented in Section 6. Big Data visualization is discussed in Section 7. In this section, we discuss methods to create a summary or compressed representation of all of the data features to help visualize hidden relationships in data. This is followed by the section described and introduced new, perspective Big Data challenges. This paper ends with conclusions in Section 9.

2. Big Data Technologies during Time

The manner in which data is stored, transmitted, analysed and visualized has varied over time; the rise of all fields of human activities is always connected with an increase in technological possibilities, as with the political situation, development of the socio-economical arrangement and industry. In 1936 Franklin D. Roosevelt's administration in the USA, after Social Secu-

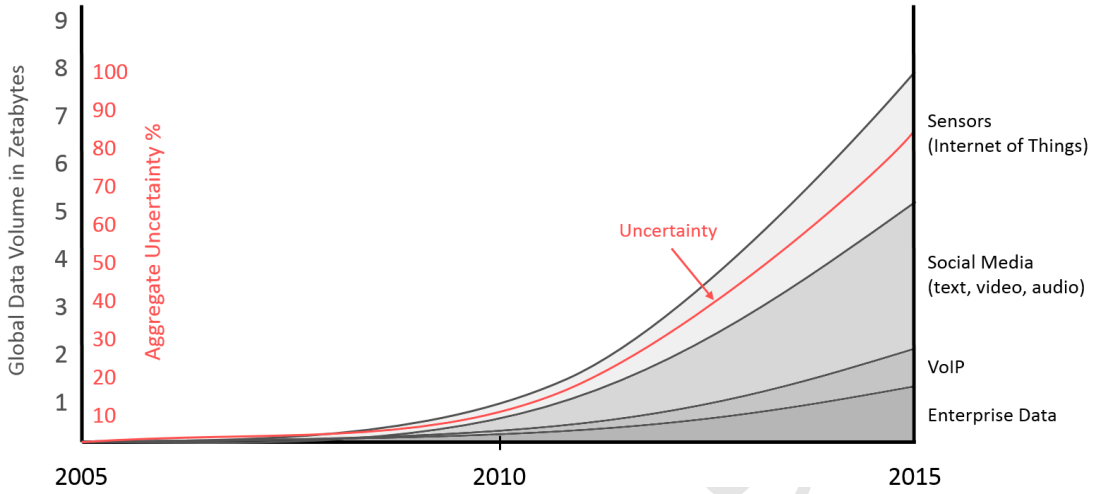

Figure 1: Big Data source with depicted data uncertainty [121, 48].

rity became law, ordered from IBM the development of the punch card-reading machine to be able to collect data from all Americans and employers. This biggest accounting operation of all time, as it was called at that time, can be considered as the first major data project [124, 125, 126, 127]. As already mentioned, the political situation always has a big influence on the rise of technology and the main mover of its development has always been war and money. During World War II, the British invented, in 1943, a machine *Colossus* to decipher German codes. The device, which searched for patterns in encrypted messages at a rate of 5000 characters per second, is known as the first data-processing machine [18, 124]. Big Data as a term has been one of the biggest trends in recent years, leading to an increase in research, as well as industry and government applications [62, 110, 113]. The continued improvements in high-performance computing and high resolution sensing capabilities have resulted in data of unprecedented size and complexity. Data is deemed a powerful raw material that can impact multidisciplinary research.

2.1. Data Storage

We face a wave of data; the amount of data is so big that a lot of information is never looked at by anybody [60]. The next problematic aspect of data is that a big part of it is redundant, e.g. one video due to many existing video formats, its resolution and subtitles in many languages [53] takes up lots of space, which is necessary from an informational point of view but generally it does not bring anything new. The manner in which data is stored has changed: what was sufficient in 1965, when the US Government decided to found the first data centre to store 175 million sets of fingerprints and 742 million tax returns and store data onto magnetic computer tape [123], is nowadays unusable. Traditionally, persistent data is still stored using hard disk drives (HDD) [54] with all the disadvantages which they have, such as boundaries on their access times, a lifetime limited by mechanical (moving) parts, and DRAM (volatile memory) with faster access. The trend is to replace HDDs with solid-state drives (SSD) as a type of non-volatile memory (NVM) [63, 84]. Other types of NVM, which are now also on the rise, are phase-change memory (PCM) and memristors. These will be integrated as byte/addressable memory on a memory bus or stacked directly on a chip (3D-stacking) [87]. All existing storage architectures, such as storage area networks (SAN), network-attached storage (NAS) and direct-attached storage (DAS), were ordinarily used before large-scale distributed systems were required and the aforementioned architectures met their limitations [63, 84].

2.2. Data Transmission

Cloud computing and cloud data storage are, nowadays, very popular. Users do not have time and do not want to maintain data storage and computing hardware, so the easiest way is to send data to the cloud [41]. However, this modern technology also has its limits - the volume of communication capacity and security [66, 84]. Cloud computing is still considered a hot trend.

2.3. Data Processing / Analysis

The next question is not where to store data, but how to store it and what platform to use to analyse it. The

classical approach to managing structured data is divided into two parts: the first is to store the data set, and the second is a related database for retrieval of stored data. Large-scale structured data set management is often based on data *warehouse* and *data mart*, which are both Standard Query Language (SQL) based. SQL is more reliable, and straightforward and analytic platforms such as Cloudera Impala and SQLstream run on it [84]. Moreover, recently, the Not Only SQL (NoSQL) database approach is often used in order to avoid using the Relation Database Management System (RDBMS) [49]. The most popular management systems using NoSQL databases are Hbase, Apache Cassandra, SimpleDB, Google BigTable, Apache Hadoop, MapReduce, MemchaceDB and Voldemort [84].

The analytic methods of Big Data are still under investigation. To help deal with Big Data, cloud computing, and then granular computing, biological computing systems, and quantum computing, are under consideration [84].

3. Motivation Examples

The general problem with statistical physics is the following: given a large collection of atoms or molecules, given the interaction laws among the constituents of this collection of particles, and given the laws of dynamic evolution, how can we predict the macroscopic physical properties of matter composed of these atoms or molecules?

The typical feature-based model [42, 61] looks for the most extreme examples of a phenomenon and represents the data by these examples, but to describe a large system, this model is not appropriate. A solution to this problem in statistical physics is based on feature summarization or subset selection methods.

This consists of the fact that, for a very large n, the volume of an n-dimensional figure is concentrated near its surface [55, 70].

It is not hard to see that the volume of an n-dimensional ball of diameter d should be expressed by the formula $V_n d^n$, where V_n is constant and does not depend on d. For example, the volume of a spherical ring between spheres of radius 1 and $1 - \epsilon$ equals

$$V_n \left(1 - (1 - e)^n\right), \tag{1}$$

which, for a fixed and arbitrarily small ϵ, but increasing n, it approaches b_n. A 20-dimensional watermelon with a radius of 20 *cm* and skin with a thickness of 1 *cm* is nearly two-thirds skin

$$\left(1 - \left(1 - \frac{1}{\epsilon}\right)^n\right) = 1 - e^{-1}. \tag{2}$$

This circumstance plays a significant role in statistical mechanics. Consider, for example, the simplest model of gas in a reservoir consisting of n atoms, which we shall assume are material points with mass 2 (in an appropriate system of units). We represent the instantaneous state of the gas by n three-dimensional vectors $(v_1, \ldots, v_n)$ of the velocities of all molecules in the physical Euclidean space; that is, by a point in the three n-dimensional coordinate space R^{3n}. The square of the lengths of the vectors in R^{3n} has a direct physical interpretation as the energy of the system (the sum of the kinetic energies of the atoms)

$$E = \sum_{i=1}^{n} |v_i|^2 . \tag{3}$$

For a macroscopic volume of gas under normal conditions, n is of the order of 10^{23} (Avogadro's number), so that the state of the gas can be described only on a sphere of an enormous dimension, whose radius is the square root of its energy.

We may conclude that a model of a large system (Big Data) must be based on feature summarization or compression or subset selection methods.

The increasing amount of VoIP, social media, and sensors data [121, 48] emphasizes the need for methods to deal with the uncertainty inherent in these data sources. **Currently about 80% of data is uncertain** see Fig. 1. We can face the problem of uncertainty via application of topological methods. The number of components or holes is not something that changes with small changes. This is vital to an application in cases where data is very uncertain.

4. Mathematical Background

In this section, we summarize the theoretical concepts which are necessary for Big Data processing as presented in the rest of the paper.

4.1. Topology

A topological space [52, 78, 92] is a set of points along with a topology; that is, a collection of subsets that are referred to as open sets. Intuitively, a set U is open if, starting from any point in U and going in any direction, it is possible to move a little and still stay inside the set. It turns out that the notion of an open set provides a fundamental way of how to speak about the

nearness of points, although without explicitly having a concept of distance defined in the considered topological space. Thus, once a topology has been defined, we are allowed to introduce properties such as continuity, connectedness, and closeness, which are all based on some notion of nearness.

A topological space is a set X and a set τ of subsets of X satisfying the following axioms:

- $\emptyset$ and X are in τ,
- if $U_1, U_2, \ldots, U_n$ are in τ, then so is $\bigcap_{i=1}^{n} U_i$,
- if U_i, $i \in I$ are in τ, then so is $\bigcup_{i \in I} U_i$.

A map f between topological spaces is said to be continuous if the inverse image of every open set is an open set. A homeomorphism is a continuous bijection whose inverse is also continuous. Two topological spaces (X, τ_X), (Y, τ_Y) are said to be homeomorphic if there exists a homeomorphism $f : X \to Y$. From the viewpoint of topology, homeomorphic spaces are essentially identical. Properties of topological space which are preserved up to homeomorphisms are said to be topological invariants.

The notion of metric is a straightforward generalization of Euclidean distance through its three properties listed there. Given a nonempty set X, we say that a mapping $d : X \times X \to \mathbb{R}$ is a metric if it satisfies the following properties:

- for all points x and y, $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$,
- for all points x and y, $d(x, y) = d(y, x)$,
- for all points x, y and z $d(x, y) + d(y, z) \leq d(x, z)$.

The pair (X, d) is called a metric space. If the metric d is understood from the context we will often refer to X as being a metric space. A systematic description of metric has been given by Deza [21, 71].

Fig. 2 shows how we can transform cloud points to nearness structure (topology space) and distance structure (metrics space).

4.2. Manifolds

The natural, higher-dimensional analogue of a surface is an n-dimensional manifold, which is a topological space with the same local properties as Euclidean n-space. Because they frequently occur and have applications in many other branches of mathematics, manifolds are certainly one of the most important classes of topological spaces.

A topological manifold is a space M locally homeomorphic to $\mathbb{R}^n$. That is, there is a cover $\mathcal{A} = \{U_\alpha\}$ of M by open sets along with maps $\phi_\alpha : U_\alpha \to \mathbb{R}^n$ that ϕ_α are homeomorphisms. The cover $\mathcal{A} = \{U_\alpha\}$ is called an atlas. This tuple (U_α, ϕ_α) is called a chart. Such local homomorphism is called a coordinate system on U_α and enables the identification of any point $u \in U_\alpha$ with an n-tuple of $\mathbb{R}^n$. M is an n-dimensional manifold with a boundary if every point has a neighbourhood homeomorphic to an open set of either $\mathbb{R}^n$ or the half-space $\{u = (u_1, \ldots, u_n) \in \mathbb{R} \,|\, u_n \geq 0\}$.

Suppose that (U_α, ϕ_α) and (U_β, ϕ_β) are two charts for a manifold M such that $U_\alpha \cap U_\beta$ is non-empty.

The transition map

$$\tau_{\alpha,\beta} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta), \tag{4}$$

is the map defined by

$$\tau_{\alpha,\beta} = \phi_\beta \circ \phi_\alpha^{-1}. \tag{5}$$

Note that since ϕ_α and ϕ_β are both homeomorphisms, the transition map $\tau_{\alpha,\beta}$ is also a homeomorphism, see Fig. 3. Depending on the type of the transition functions (e.g., smooth, analytic, piecewise smooth, Lipschitz), the manifold is consequently named (e.g. smooth manifold, analytic manifold, etc.). A compact manifold is a manifold that is compact as a topological space. A closed manifold is a compact manifold without a boundary. An important property of a manifold concerns orientability. Then, a manifold M is called orientable if there exists an atlas $\mathcal{A} = \{(U_i, \phi_i)\}$ on it such that the Jacobian of all transition functions $\phi_{i,j}$ from one chart to another is positive for all intersecting pairs of regions. Manifolds that do not satisfy this property are called non-orientable. We prefer here to skip the technicalities needed to formally define such a notion, referring the reader to [59, 26] for further details.

4.3. Algebraic Topology

The approach adopted by algebraic topology is the translation of topological problems into an algebraic language, to solve them more easily. There are classics resources of algebraic topology [45, 47, 73, 99]. These resources are written without high-level formalism.

In persistent homology, we ultimately want to compare topological spaces based on the characteristic holes that they encompass. Because we usually operate with finite point clouds in data analysis, we first need to discretize the space to add the notion of connectivity. That is done through the creation of simplicial complexes. A p-simplex σ is the convex hull of $p+1$ linearly independent points $x_0; x_1, \ldots, x_p \in R^d$ [116]. More intuitively,

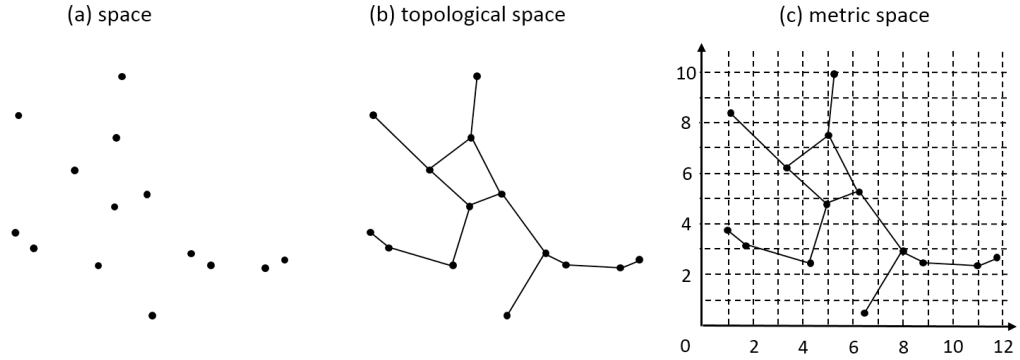

Figure 2: Spaces [118] (depicted on Hercules constellation).

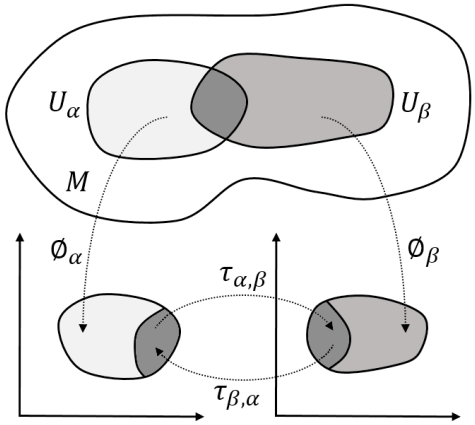

Figure 3: Charts on a manifold.

a 0-simplex is a vertex, an 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so forth. A simplicial complex K is a finite set of simplices such that, for $\sigma \in K$, all of its faces are also in K.

The core idea of persistent homology is to analyse how holes appear and disappear, as simplicial complexes are created. To do this, a filtration is constructed. An increasing sequence of ϵ values, i.e., distance values, produces a filtration, such that a simplex enters the sequence no earlier than all its faces.

The Vietoris-Rips complex, Fig. 4, is one of the most popular complexes in persistent homology. For a non-negative real number ϵ, the Vietoris–Rips complex $V(K, \epsilon)$ at scale ϵ is defined as follows:

$$V(K, \epsilon) = \{\sigma \subset K \mid d(x, y) \leq \epsilon\} \text{ for all } x, y \in \sigma. \quad (6)$$

For $\epsilon \leq \epsilon'$, we have $V(K, \epsilon) \subseteq V(K, \epsilon')$, so considering the different values of the scale ϵ yields a filtered simplicial complex. The dimension of the Vietoris-Rips complex is bounded only by the size of K, therefore, in practice, it is necessary to put a limit on the dimension of the simplices that one allows in the construction of the Vietoris-Rips complex.

The Čech complex, Fig. 4, is defined as set of simplices such that $\epsilon/2$-ball neighborhoods have a point of common intersection.

A typical case is the construction of algebraic structures to describe topological properties, which is the core of homology theory, one of the main tools of algebraic topology. In [112], persistent homology is presented as a new approach to the topological simplification of Big Data via measuring the lifetime of internal topological features during a filtration process. This approach was assessed as being exploitable in many scientific and engineering applications. In [83], a broad view is given of the theory of persistence, including also its topological and algorithmic aspects, and an elaboration on its context to quiver theory on the one hand, to data analysis on the other. This book also contains many open problems in topological data analysis.

Another concept of the persistence is the Survival Signature [16]. This concept has became a popular tool for analysis and assessment of system reliability. Samaniego introduced this topic in [91]. However, signatures are applicable only in systems with a single type of component, as all its components have to be characterized as exchangeable random quantities. This work is from the theoretical part of research, so its practical usage for real systems is limited, because real systems tend to have components of multiple types. Signatures can not be used for analysing the reliability of networks, it is caused by the existence, at least, of two different kinds of components - links and nodes.

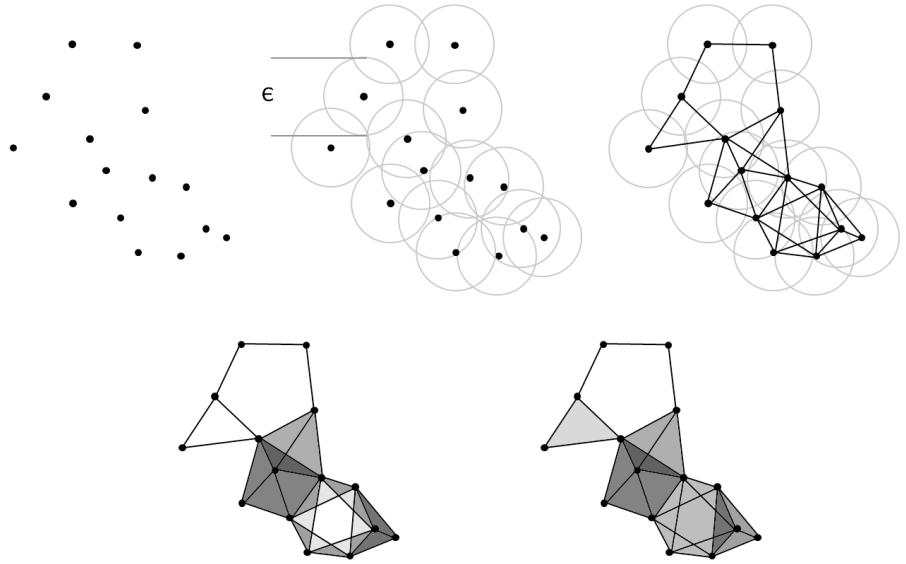

Figure 4: Čech (lower left) and Rips (lower right) complex built on the fixed set of points (upper left) with the depiction of the formation of the both complexes.

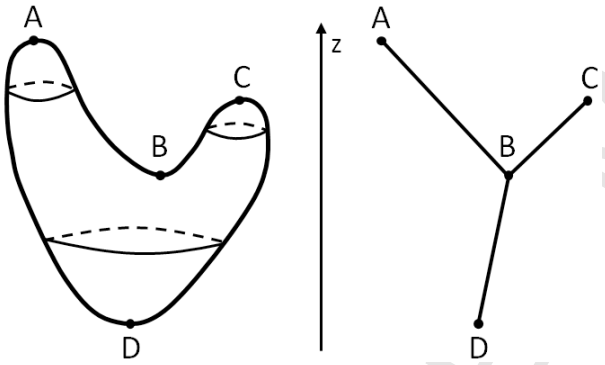

Figure 5: Morse function and Reeb graph on the 2-sphere.

4.4. Morse Theory

In this Section, we report some classical results in Morse theory [76], which constitutes the essential mathematical root for Reeb graphs.

Morse theory can be seen as the investigation of the relation between functions defined on a manifold and the shape of the manifold itself. The key feature of Morse theory is that information on the topology of the manifold is derived from the information about the critical points of real functions defined on the manifold.

Morse theory is a means of relating the global features of (in the classical setting) a Riemannian manifold M with the local features of critical points of smooth $\mathbb{R}$-valued functions on M. Recall that $h : M \to \mathbb{R}$ is Morse if all critical points of h are non-degenerate, in the sense of having a non-degenerate Hessian matrix of second partial derivatives. Denote by $Cr(h)$ the set of critical points of h. For each $p \in Cr(h)$, the Morse index of p, $\mu(p)$, is defined as the number of negative eigenvalues of the Hessian at p. The theory identifies at points which level sets of the function undergo topological changes, and it relates these points via a complex. In particular, Morse theory provides the mathematical background underlying several descriptors, such as Reeb graphs, size functions, persistence diagrams and Morse shape descriptors. For a detailed overview of Morse theory, see [74, 81].

Let $h : M \to \mathbb{R}$ be a continuous function defined on a domain M. For each scalar value $a \in \mathbb{R}$, the level set $h^{-1}(a) = \{x \in M \,|\, h(x) = a\}$ may have multiple connected components. The Reeb graph Fig. 5 of h, denoted by $Rb_h(M)$, is obtained by continuously identifying every connected component in a level set to a single point. In other words, $R_h(M)$ is the image of a continuous surjective map $\Phi : X \to Rb_h(X)$, where $\Phi(x) = \Phi(y)$ if, and only if, x and y come from the same connected component of a level set of h. For a detailed overview of a Reeb graph, see [26].

5. Topological Data Analysis

Geometry is understood and used mainly as quantitative mathematics, while topology, on the other hand, provides a formal language for a qualitative approach. In topology, it is studied relationships of nearness or

proximity, but without using distances. A map between topological spaces is called continuous if the nearness structures are retained. Nowadays, in algebra, we study maps that preserve product structures; for example, group homomorphisms between groups, and one of the largest areas of growth in pure mathematics this century has been the solution of topological problems by casting them into a simpler form using groups. This theory is called algebraic topology and, like analytical geometry and differential geometry before it, there is considerable interplay with some of the most fundamental ideas in computer science.

Topological data analysis aims to provide additional tools for analysing data sets that appear in engineering and science. The goal is not to replace current techniques because these techniques still supply an additional and powerful approach for mining intuitive features (as well as not-so-intuitive) in data collections. Proposed approaches focus on the data shape, and can be implemented to data sets of high dimensions.

As computational topology has undergone progress, now we are able to deduce topological invariants from data. The input of these procedures is often in the form of a point cloud, regarded as possibly noisy observations from an unknown lower-dimensional set whose topological features, which could have information potential, were lost during a sampling procedure. Sampling data is a way to get sublinear algorithms. Sublinear algorithms are a recent development in theoretical computer science, statistics, and discrete mathematics, which address the mathematical problem of understanding global features of a data set using limited resources. Often enough, to determine important input features, one does not need to actually look at the entire input. The field of sublinear algorithms [110] makes precise the circumstances when this is possible and combines discrete mathematics and algorithmic techniques with a comprehensive set of statistical tools to quantify errors and give trade-offs with sample sizes. The output is a collection of data summaries that are used to estimate the topological features of data collection. There are software packages for computing topological invariants from data [72, 80, 101, 102, 122].

By using homology, the features of a topological space such as an annulus, sphere, torus, complicated surface or manifold can be measured. Homology is so helpful that, thanks to it, it is possible to differentiate spaces from one another using the quantified connected components, trapped volumes, topological circles, etc. On a finite set of data points, a (noisy) sampling from an underlying topological space can be seen. The homology of the data can be measured using the connections' proximate data points; changing the scale of which these connections are made and finding the features of the data is persistent regardless of changing the scale. This approach is called persistent homology, and it is considered to be the most useful and helpful method for finding the topological structure of a discrete data set. Persistent homology has found its place in various application areas for its ability to discover the topological structure of data.

Persistent homology for data analysis has been studied by many researchers in mathematics and computer science, e.g. Carlsson [10], Edelsbrunner and Harer [26], Ghrist [36], Oudot [83] and Zomorodian [117, 118].

To drive the reader through the bunch of approaches and frameworks revised here, we must first introduce the basic notions of mathematical concepts such as topological space, manifold, map, metric and transformation. We also provide a brief overview of algebraic topology.

How do we find the topological structure of the data sets? A technology called persistent homology analysis was proposed to solve this problem [10, 14, 36]. The topological structure of the data sets is now one of the major areas where mathematicians and computer scientists have focused considerable attention. The geometric structure of massive amounts of data, or Big Data, will be critical in data analysis. We predict there will be many more new findings in theory and practice.

Historically, geometrical and topological techniques have been deployed as independent alternatives in the analysis of a variety of data types. However, the continuing increases in size, dimensionality, number of variables, and uncertainty create new challenges that traditional approaches cannot address. New methods based on geometrical and topological techniques are needed to support the management, analysis and visualization of Big Data [3, 10, 26, 36, 117, 118].

An essential part of Big Data processing is the need for different types of users to apply visualizations [3, 30, 85] to understand a result of Big Data processing. Recently, it became apparent that a large number of the most interesting structures and world phenomena could be described by networks. Developing a theory for very large networks is a significant challenge in Big Data research [64]. Big Data is one of the main science and technology challenges of today.

In mathematical science, homology is a general procedure to associate a sequence of abelian groups or modules to a given topological space and/or manifold [26, 43]. The idea of homology dates back to Euler and Riemann, although the homology class was first rigor-

ously defined by Henri Poincaré, who built the foundation of modern algebraic topology. The topological structure of a given manifold can be studied by defining the different dimensional homology groups on the manifold such that the bases of the homology groups are isomorphic to the bases of the corresponding topological spaces. In a computational point of view, we can approximate the given manifold using a triangulated simplicial complex, on which homology groups can be further defined. There exist some methods, such as Delaunay triangulation, which can be used for the triangulation of a manifold or topological spaces. And, there are many triangulation software packages, such as TetGenand CGAL. The Cartesian representation is one of the most important approaches to scientific computing. Due to this homology analysis being based on a cubical complex, it has been a popular field for researchers in recent years. Kaczynski et al. described homology analysis in the cubical complex very systematically in [51].

6. Application of Computational Geometry and Topology

Persistent homology creates a multiscale representation of topological structures via a scale parameter relevant to topological events [27, 33, 90, 119]. In the past decade, persistent homology has been developed as an efficient computational tool for the characterization and analysis of topological features in large data sets [27, 119, 120]. Persistent homology can be maintained continuously, despite the filtration process, over a range of spatial scales in persistent homology analysis. Persistent homology, by its nature, when compared to traditional computational topology [13, 56, 105] and/or computational homology (which results in truly metric-free or coordinate-free representations), exhibits one additional dimension - the filtration parameter. This additional parameter finds its use in building some crucial geometry or quantitative information into the topological invariants, so that the birth and death of isolated components, cavities, circles, loops, pockets, rings or voids at all geometric scales can be defined by topological measurements. For the visualization of topological persistence [115], a Barcode representation has been proposed, in which various horizontal line segments or bars are used to interpret the persistence of the topological features.

Efficient computational algorithms such as the pairing algorithm [25, 35], Smith normal form [26, 119] and Morse reduction [27, 38, 39], have been proposed to track topological variations during the filtration process [7, 105]. Some of these persistent homology algorithms have been implemented in many software packages, namely Perseus [80], JavaPlex [102] and Dionysus [122]. In [82], guidelines are provided for the computation of persistent homology with a good introduction on how to make our implementations.

In the past few years, persistent homology has been applied to image analysis [2], image retrieval [32], chaotic dynamics verification [51], sensor networks [97], complex networks [46], data analysis [10, 67, 88, 108], computer vision [6, 12], shape recognition [23] and computational biology [114].

Advances in medicine, particularly in genetic engineering, have increased the amount of genome-wide gene expression data, but the number of pattern recognition methods, which could be useful in this area, is still not huge enough. To be able to find interesting and adequate enough fact patterns in such huge amounts of data connected with some level of noise is still a big challenge. In [20], a new approach to Pattern detection in gene expression data is presented.

Recovering or inferring a hidden structure from discrete samples is a basic problem in data analysis, omnipresent in a wide range of applications. Data often shows a considerable high-dimension; for the understanding and finding of interesting information, which are hidden in data, it is necessary to approximate it with a low-dimensional or even with one-dimensional space, because many important aspects of data are often internally low-dimensional. Morse theory and Reeb graphs are a simple but significant scenario, where the hidden space has a graph-like geometric structure, such as the branching filamentary structures formed by blood vessels.

In [19], a straightforward and efficient algorithm is presented to approximate the Reeb graph $Rb_f(M)$ of a map $f : M \to \mathbb{R}$ from point data sampled from a smooth and compact manifold M.

In [4], an overview of the mathematical properties of Reeb graphs is given. In [34], the authors introduced a framework to extract, as well as to simplify, a one-dimensional skeleton from unorganized data using the Reeb graph. They apply a proposed algorithm for molecular simulation. The input is molecular simulation data using the replica-exchange molecular dynamics method [79]. It contains $250K$ protein conformations, generated by 20 simulation runs, each of which produces a trajectory in the protein conformational space. Simulations at low energy should provide a good sampling of the protein conformational space around the native structure of this protein.

7. Big Data Visualization

The emergence of Big Data has brought about a paradigm shift through computer science, such as the fields of computer vision, machine learning, and multimedia analysis. Visual Big Data, which is specifically about visual information such as images and videos, accounts for a large and important part of Big Data. Many theories and algorithms have been developed for visual Big Data in recent years, among which the dimensionality reduction technique [28, 96, 100] plays an increasingly important role in the analysis of visual Big Data. Unfortunately, conventional statistical and computational tools are often severely inadequate for processing and analysing large-scale, multi-source and high-dimensional visual Big Data. The combination dimensionality reduction and visual Big Data will bring about huge opportunities as well as challenges to these communities. In recent years, this area has gained much attention, thanks to the development of nonlinear spectral dimensionality reduction methods, often referred to as manifold learning algorithms, see [69].

The authors of [98] presented a tool for extracting a feature of data using selected dimension reduction techniques. From the verified methods were chosen non-negative matrix factorization, singular value decomposition, semi-discrete decomposition, a novel neural network-based algorithm for Boolean factor analysis, and two cluster analysis methods as well. As the benchmark, the so called bars problem was applied. The authors proposed generating sets of artificial signals as a Boolean sum of the given number of bars and then analysing it using selected dimension reduction techniques. From the results, it was deduced that Boolean factor analysis is the most suitable method for this kind of data.

Data or information visualization is used to synthesize information and knowledge from massive, dynamic, ambiguous, uncertain, noisy and often conflicting data. Information visualization is a broad research area that aims to aid users in exploring, understanding, and analysing data through progressive, iterative visual exploration [95]. The rise of the field of Big Data caused the need for development of areas closely connected with it, such as machine learning, computer vision and multimedia analysis. With the boom in Big Data and deep data analytics, visualization is being widely used in a variety of data analysis applications [17, 24]. Big Data visualization is one of the most needed fields for Big Data processing. Many theories and algorithms were developed, and are still being developed, to help visualize Big Data, because the well known and used already-developed tools and statistical methods are very often appropriate for the nature of large-scale, multi-source and high-dimensional visual Big Data. To understand the knowledge and relationships, which are "hidden in pure data", it is necessary to be able to understand the relationship between geometric objects constructed from data using various parameter values. To be able to visualize data in a more understandable form based on the same principle, the aggregation of original attributes is used, using various techniques, among which the dimensionality reduction technique [28, 96] plays an increasingly important role. The connection between dimensionality reduction techniques and visual Big Data will introduce huge opportunities as well as challenges to the community interested in this area [113].

Dimensionality reduction techniques are based on assignment of high-dimensionality space to lower and, ideally, to low-dimensional space (3D, 2D) to be better able to visualize data (Fig. 6) and to solve a fundamental problem in a variety of data analysis tasks - to find an appropriate representation for the given data, see [104]. Dimensionality reduction methods can be divided into two groups, linear and non-linear methods. The linear methods transform original variables into a new variable using the linear combination of the original variables. From linear dimensionality reduction techniques this can be named Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Multi-Dimensional Scaling (MDS), Linear Discriminant Analysis (LDA), Canonical Correlations Analysis (CCA), Maximum Autocorrelation Factors (MAF), Slow Feature Analysis (SFA), Sufficient Dimensionality Reduction (SDR), Locality Preserving Projection (LPP), Under complete Independent Component Analysis (UICA), Probabilistic PCA (PPCA), Factor Analysis (FA), Linear Regression (LR), or Distance Metric Learning (DML) [5, 8, 29, 75].

Non-linear methods include e.g. Non-linear Manifold Learning Methods (Laplacian Eigenmaps (LE), Locally Linear Embedding (LLE), Isomap, Hessian Eigenmap, Semi-Definite Programming (SDE), Manifold Based Charting, Local Tangent Space Alignment (LTSA), Diffusion maps, Parallel vector Field Embedding (PFE), Geodesic Distance Function Learning (GDL) and Parallel Field Alignment for cross media Retrieval (PFAR)), Discriminative locality alignment (DLA), or Generalized Eigenvectors for Multiclass (GEM) [1, 5, 8, 29, 58].

The latest research in the area of decomposition methods has introduced some new approaches, which enable the restrictions and limits of conventional methods to be dealt with. Wang et al. [111] proposed the

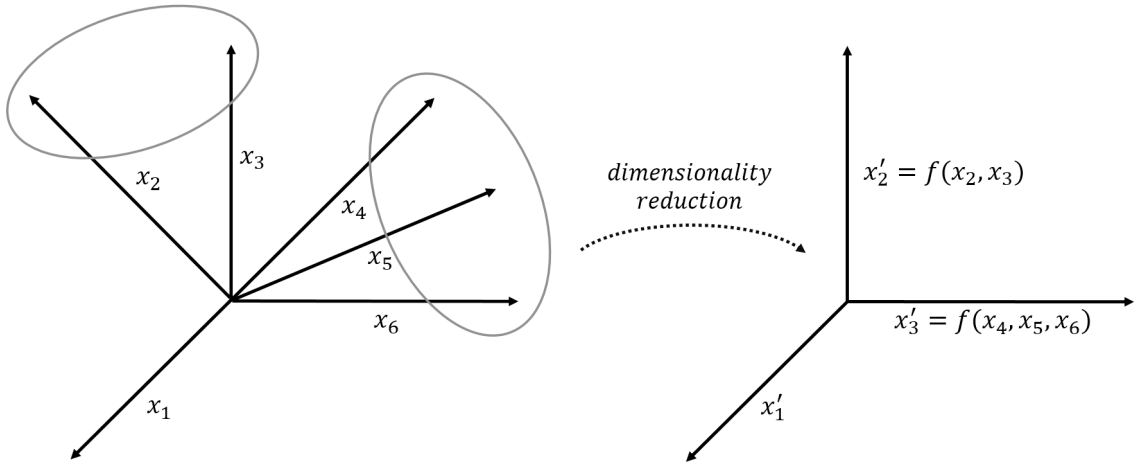

Figure 6: Principle of dimensionality reduction.

generalized Discriminative Generalized Eigendecomposition (DGE) method based on the idea that better separation of a multi-dimensional feature could be helpful in finding better discriminant vectors. DGE can deal with Gaussian and non-Gaussian distribution. In [50], the combination of LDA and LPP as the RElevant Local Discriminant Analysis RELDA algorithm is presented, which has an analytical form of the globally optimal solution, and it is based on eigendecomposition, too. Some new interesting variants of LPP are introduced in [94].

8. Big Data Challenges

One of the biggest challenges for Big Data research, we face today, is digitization. Digitization is the main part of cyber-physical systems (CPS) which introduces the fourth stage of industrialization, commonly known as Industry 4.0. A strategic initiative called Industrie 4.0 (Industry 4.0) has been proposed and adopted by the German government as a part of the High-Tech Strategy 2020 Action Plan [128]. Industry 4.0 or the fourth industrial revolution, is a collective term embracing some contemporary automation, data exchange, and manufacturing technologies. Similar strategies have also been proposed by other main industrial countries, e.g., Industrial Internet [129] by the USA and Internet+ [65] by China. Industry 4.0 is also referred as Smart Factory, Cyber-Physical Production Systems or Advanced Manufacturing, but the meaning is mostly the same. It is defined as a collective term for technologies and concepts of value chain organizations which draw together CPS, the Internet of Things and the Internet of Services. Smart factory modelling based on virtual design and simulation has emerged as a part of the mainstream activities geared towards reducing product design cycle. The smart factory is characterized by a self-organized multi-agent system [109] supported with Big Data based feedback and coordination.

The Industry 4.0 describes a CPS oriented production system [77, 89, 106, 107] that integrates production facilities, warehousing systems, logistics, and even social requirements to establish the global value creation networks [31]. Big data and cloud computing for Industry 4.0 are viewed as data services that utilize the data generated in Industry 4.0 implementations but are not independent as Industry 4.0 components [37, 44]. For Industry 4.0 and Smart Manufacturing processes dealing with large data storage, sharing data, processing and analysing have become key challenges to computer science research. Some examples of these include efficient data management, additional complexity arising from analysis of semi-structured or unstructured data and quick time critical processing requirements. To resolve these issues, understanding of this massive amount of data, advanced visualization and data exploration techniques are critical [93].

Analytics based on Big Data has emerged only recently in the manufacturing world, where it optimizes production quality, saves energy, and improves equipment service [22, 40]. In an Industry 4.0 context, the collection and complete evaluation of data from many heterogeneous sources will become standard to support real-time decision-making.

Example of Big Data collection is, in research area, well known **D**ata**B**ase systems and **L**ogic **P**rogramming

(DBLP) Computer Science Bibliography, which provides bibliographic information on major computer science journals and proceedings. Nowadays, more than 3.35 million records, which contain titles of articles, their authors, years of publication is indexed in DBLP and since 2011, more than 300 thousands records have been added every year. DBLP is database with open access and due to its content and its size, it is very interesting resource for evolution analysis of co-author networks and could be considered as one of the example of Big Data dataset in nonindustrial world [86],[57].

9. Conclusion

The last few years have seen a great increase in the amount of data available to scientists, engineers, and researchers from many disciplines. Modern data science uses topological methods to find the structural features of data sets before further supervised or unsupervised analysis. The size of data at present is huge and continues to increase every day. Datasets with millions of objects and hundreds, if not thousands, of measurements, are now commonplace in areas such as image analysis, computational finance, bio-informatics, and astrophysics. The variety of data being generated is also expanding. The velocity of data generation and its growth is increasing because of the proliferation of IoT, sensors connected to the Internet. This data provides opportunities that allow businesses across all industries to gain real-time business insights. We present motivational examples to show that, for large amounts of data, we need a new model for data processing. This model must be based on feature summarization instead of classical methods based on feature selection. We also face, with uncertainty, Big Data. The geometrical and topological method helps us to solve this problem.

In this study, we presented a review of the rise of geometrical and topological methods which can be used for Big Data processing. We proposed a geometrical and topological view of the Big Data model.

Acknowledgement

This work was supported by the Czech Science Foundation under the grant no. GACR GJ16-25694Y and in part by Grant of SGS No. SP2016/68 and SP2016/97, VŠB - Technical University of Ostrava, Czech Republic.

References

[1] M. Belkin, N. Partha, Laplacian eigenmaps for dimensionality reduction and data representation, Neural computation 15.6 (2003): 1373-1396.

[2] P. Bendich, H. Edelsbrunner, M. Kerber, Computing robustness and persistence for images, IEEE Trans. Vis. Comput. Graph. 16 (2010) 1251–1260.

[3] J. Bennett, F. Vivodtzev, V. Pascucci, Topological and Statistical Methods for Complex Data, Springer 2015.

[4] S. Biasotti, D. Giorgi, M. Spagnuolo, B. Falcidieno, Reeb graphs for shape analysis and applications, Theoretical Computer Science 392 (2008) 5–22.

[5] E. Bertini, A. Tatu, D. Keim, Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization, IEEE Transactions on Visualization and Computer Graphics 17(12) (2011) 2203–2212.

[6] S. Biasotti, A. Cerri, A. Bronstein, M. Bronstein, Recent Trends, Applications, and Perspectives in 3D Shape Similarity Assessment, Computer Graphic Forum, DOI: 10.1111/cgf.12734, 2015.

[7] P. Bubenik, P. T. Kim, A statistical approach to persistent homology, Homol. Homotopy Appl. 19 (2007) 337–362.

[8] C. J. C. Burges, Dimension reduction: A guided tour, Now Publishers Inc, 2010.

[9] D. Butler, A world where everyone has a robot: why 2040 could blow your mind, Nature 530(7591) (2016).

[10] G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46, 255–308 (2009).

[11] G. Carlsson, Topological pattern recognition for point cloud data Acta Numerica 23, 289–368, 2014.

[12] G. Carlsson, T. Ishkhanov, V. De Silva, A. Zomorodian, On the local behavior of spaces of natural images. Int. J. Comput. Vis. 76(1) (2008) 1–12.

[13] H. W. Chang, S. Bacallado, V. S. Pande, G. E. Carlsson, Persistent topology and metastable state in conformational dynamics, PLoS ONE 8(4) (2013) e58699.

[14] L. M. Chen, Digital and Discrete Geometry: Theory and Algorithms, Springer, 2014.

[15] H. Chen, R. H. L. Chiang, V. C. Storey, Business intelligence and analytics: From big data to big impact, MIS Quarterly 36(4) (2012) 1165–1188.

[16] F. P. A. Coolen, T. Coolen-Maturi, Generalizing the signature to systems with multiple types of components, Complex Systems and Dependability. Springer Berlin Heidelberg, 7th International Conference on Dependability and Complex Systems (DepCoS-RELCOMEX), (2013) 115-130.

[17] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, X. Tong, Textflow: towards better understanding of evolving topics in text, IEEE Trans. Vis. Comput. Graph. 17(12) (2011) 2412–2421.

[18] B. J. Copeland, Colossus, Its origins and originators, IEEE Annals of the History of Computing (2004) 38–45. DOI: 10.1109/MAHC.2004.26.

[19] T. K. Dey, Y. Wang, Reeb Graphs: Approximation and Persistence, Discrete & Computational Geometry 49(1) (2013) 46–73.

[20] M. L. Dequeant, S. Ahnert, H. Edelsbrunner, T. M. A. Fink, E. F. Glynn, et al. Comparison of Pattern Detection Methods in Microarray Time Series of the Segmentation Clock. PLoS ONE 3(8): e2856 (2008). doi:10.1371/journal.pone.0002856.

[21] M. M. Deza, E. Deza, Encyclopedia of Distances, Springer, 2014.

[22] P. ODonovan, K. Leahy, K. Bruton, D. T. J. OSullivan, An industrial big data pipeline for datadriven analytics maintenance

applications in largescale smart manufacturing facilities, Journal of Big Data (2015) 2:25, 1–26.
[23] B. Di Fabio, C. Landi, A MayerVietoris formula for persistent homology with an application to shape recognition in the presence of occlusions, Found. Comput. Math. 11 (2011) 499–527.
[24] J. Dill, R. Earnshaw, D. Kasik, J. Vince, P. Wong, Expanding the Frontiers of Visual Analytics and Visualization, Springer-Verlag (2012).
[25] T. K. Dey, K. Y. Li, J. Sun, C. S. David, Computing geometry aware handle and tunnel loops in 3d models, ACM Trans. Graph. 27 (2008).
[26] H. Edelsbrunner, J. L. Harer, Computational Topology, AMS, 2010.
[27] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, Discrete Comput. Geom. 28 (2002) 511–533.
[28] L. Eldén, Matrix Methods in Data Mining and Pattern Recognition, SIAM, 2007.
[29] I. K. Fodor, A survery of dimension reduction techniques. Tech. Rep. UCRL-ID-148, US Department of Energy, DOI: 10.2172/15002155, (2002)
[30] P. Fox, J. Hendler, Changing the Equation on Scientific Data Visualization, Science 331(6018) (2011) 705–708.
[31] E.M. Frazzon, J. Hartmann, T. Makuschewitz, B. Scholz-Reiter, Towards socio-cyber-physical systems in production networks, Procedia CIRP (7) (2013) 49–54.
[32] P. Frosini, C. Landi, Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval, Pattern Recognit. Lett. 34 (2013) 863–872.
[33] P. Frosini, C. Landi, Size theory as a topological tool for computer vision, Pattern Recognit. Image Anal. 9(4) (1999) 596–603.
[34] X. Ge, I. I. Safa, M. Belkin, Y. Wang, Skeletonization via Reeb Graphs Advances in Neural Information Processing Systems 24 (2011) 837–845.
[35] R. Ghrist, Barcodes: the persistent topology of data, Bull. Am. Math. Soc. 45 (2008) 61–75.
[36] R. Ghrist, Elementary Applied Topology, ISBN 978-1502880857, 2014.
[37] R. Gideon, The Fourth Industrial Revolution: A Davos Reader, Council on Foreign Relations (1860), 2016.
[38] S. Harker, K. Mischaikow, M. Mrozek, V. Nanda, Discrete Morse theoretic algorithms for computing homology of complexes and maps, Found. Comput. Math. (2013), http://dx.doi.org/10.1007/s10208-013-9145-0.
[39] S. Harker, K. Mischaikow, M. Mrozek, V. N, H. Wagner, M. Juda, The efficiency of a homology algorithm based on discrete Morse theory and coreductions, in: Proceeding of the 3rd International Workshop on Computational Topology in Image Context, ImageA (2010) 41–47.
[40] H. Hashem, D. Ranc, An integrative modeling of BigData processing, International Journal of Computer Science and Applications, Volume 12, Issue 1, (2015), 1–15.
[41] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, The rise of "big data" on cloud computing: Review and open research issues, Information Systems 47 (2015) 98–115.
[42] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, Springer, New York, 2011.
[43] A. Hatcher, Algebraic Topology, Cambridge University Press, 2001.
[44] M. Hermann, T. Pentek, B. Otto, Design Principles for Industrie 4.0 Scenarios: A Literature Review. Working Paper, No. 01–2015, Technical University of Dortmund, 2015.
[45] P. J. Hilton and S. Wylie, Homology Theory, Cambridge University Press, 1960.
[46] D. Horak, S. Maletic, M. Rajkovic, Persistent homology of complex networks, J.Stat. Mech. Theory Exp. 2009 (03) (2009) P03034.
[47] S. T. Hu, Homotopy Theory, Academic Press, 1959.
[48] IDC, Worldwide Big Data Technology and Services 2012–2015 Forecast, 1 (2012) IDC #233485.
[49] H. Jing, E. Haihong, L. Guan, D. Jian, Survey on NoSQL database, 6th International Conference on Pervasive Computing and Applications (ICPCA) (2011) 363–366.
[50] P. Jing, Z. Ji, Y. Yu, Z. Zhang, Visual search reranking with RElevant Local Discriminant Analysis, Neurocomputing 173 (2016) 172–180.
[51] T. Kaczynski, K. Mischaikow, M. Mrozek, Computational Homology, Springer-Verlag, 2004.
[52] S. Kalajdzievski, An Illustrated Introduction to Topology and Homotopy. CRC Press, 2015.
[53] K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in big data analytics, J. Parallel Distrib. Comput. 74 (2014) 2561–2573.
[54] V. Kasavajhala, Solid state drive vs. hard disk drive price and performance study. Proc. Dell Tech. White Paper (2011) 8–9.
[55] A. I. Kostrikin, Yu. I. Manin, Linear Algebra and Geometry, Gordon and Breach Science Publishers, 1997.
[56] B. Krishnamoorthy, S. Provan, A. Tropsha, A topological characterization of protein structure, Data Mining in Biomedicine, Springer Optimization and Its Applications (2007) 431–455.
[57] M. Kudelka, M. Radvansky, Z. Horak, P. Kromer, Soft identification of experts in DBLP using FCA and fuzzy rules, Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, SMC (2012), 1942–1947.
[58] J. A. Lee, M. Verleysen, Nonlinear Dimensionality Reduction. Springer (2007)
[59] J. Lee, Introduction to Smooth Manifolds Graduate Texts in Mathematics 218 (2012).
[60] M. Lesk, How Much Information is there in the World?, 1997 - http://www.lesk.com/mlesk/ksg97/ksg.html, (visited on February 2016).
[61] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of Massive Datasets. Cambridge University Press, second edition, 2014.
[62] K.-C. Li, H. Jiang, L. T. Yang, A. Cuzzocrea, Big Data: Algorithms, Analytics, and Applications, CRC Press, 2015.
[63] D. J. Lingenfelter, A. Khurshudov, D. M. Vlassarev., Efficient disk drive performance model for realistic workloads. Magnetics, IEEE Transactions on 50.5 (2014) 1-9.
[64] L. Lovász, Large Networks and Graph Limits, AMS Colloquium Publications, 2012.
[65] K.Q. Li, Premier of the State Council of China,Report on the work of the government, delivered at the third session of the 12th National Peoples Congress, March 5, 2015.
[66] J. Li, X. Chen, Q. Huang, D. S. Wong, Digital provenance: enabling secure data forensics in cloud computing, Future Gener. Comput. Syst. 37 (2014) 259–266.
[67] X. Liu, Z. Xie, D. Yi, A fast algorithm for constructing topological structure in large data, Homol. Homotopy Appl. 14 (2012) 221–238.
[68] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Vejdemo-Johansson, M. Alagappan, J. Carlsson, G. Carlsson, Extracting insights from the shape of complex data using topology. Sci. Rep. 3 (2013). doi:10.1038/srep01236
[69] Y. Ma, Y. Fu, Manifold Learning Theory and Applications. CRC Press (2012)
[70] Yu. I. Manin, Mathematics and Physics, Birkhauser Boston, 1981.
[71] S. Marchand-Maillet, Y. M. Sharaiha, Binary Digital Image Pro-

cessing, Academic Press, 2000.
[72] C. Maria, GUDHI, simplicial complexes and persistent homology packages, https://project.inria.fr/gudhi/software/, (visited on February 2016).
[73] W. S. Massey, Algebraic Topology: An Introduction, Harcourt Brace and World, 1967.
[74] J. W. Milnor, Morse Theory, Annals of Math Studies, Princeton University Press, 1963.
[75] M. Mizuta, Dimension reduction methods, Handbook of computational statistics, Springer Berlin Heidelberg, (2012) 619–644.
[76] M. Morse, Relations between the critical points of a real function of n independent variables, Transactions of the American Mathematical Society 27 (1925) 345–396.
[77] P. J. Mosterman, J. Zander, Industry 4.0 as a Cyber-Physical System study, Softw Syst Model (2016) 15:17–29.
[78] J. Munkres, Topology: a first course, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.
[79] I.-H. Park, C. Li, Dynamic ligand-induced-fit simulation via enhanced conformational samplings and ensemble dockings: a survivin example. J. Phys. Chem. B. 114 (2010) 5144–5153.
[80] V. Nanda, The Perseus software project for rapid computation of persistent homology, http://www.sas.upenn.edu/ vnanda/perseus/index.html, (visited on February 2016).
[81] L. Nicolaescu, An Invitation to Morse Theory, Springer, 2007.
[82] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington, A roadmap for the computation of persistent homology, (2015) eprint arXiv:1506.08903.
[83] S. Y. Oudot, Persistence Theory: From Quiver Representations to Data Analysis, AMS Mathematical Surveys and Monographs, 2015.
[84] C. L. Philip Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences 275 (2014) 314–347.
[85] J. Pokorny, V. Snasel, Big Graph Storage, Processing and Visualization, in Graph-Based Social Media Analysis, Chapman & Hall/CRC, (2016) 403–430.
[86] M. Radvansky, Z. Horak, M. Kudelka, V. Snasel, Evolution of Author's Profiles Based on Analysis of DBLP Data, New Trends in Databases and Information Systems, Workshop Proceedings of the 16th East European Conference, ADBIS (2012) 317–326.
[87] P. Ranganathan, From microprocessors to nanostores: rethinking data-centric systems, IEEE Comput. 44(1) (2011) 39–48.
[88] B. Rieck, H. Mara, H. Leitte, Multivariate data analysis using persistence-based filtering and topological signatures, IEEE Trans. Vis. Comput. Graph. 18 (2012) 2382–2391.
[89] M. Riedl, H. Zipper, M. Meier, C. Diedrich, Cyber-physical systems alter automation architectures, Annual Reviews in Control 38 (1) (2014) 123–133.
[90] V. Robins, Towards computing homology from finite approximations, in: Topology Proceedings 24 (1999) 503–532.
[91] F. J. Samaniego, System signatures and their applications in engineering reliability, Springer Science & Business Media, (2007).
[92] P. Saveliev, Topology Illustrated, (2016), ISBN-13: 978-1495188756.
[93] K. Schwab, The Fourth Industrial Revolution, World Economic Forum. 2016.
[94] G. Shikkenawis, S. K. Mitra, On some variants of locality preserving projection, Neurocomputing 173 (2016) 196–211.
[95] H. Shiravi, A. Shiravi, A. A. Ghorbani, A survey of visualization systems for network security, IEEE Trans. Vis. Comput. Graph. 18(8), (2012) 1313–1329.
[96] D. Skillicorn, Understanding Complex Datasets Data Mining with Matrix Decompositions, Chapman & Hall/CRC, (2007).
[97] V. D. Silva, R. Ghrist, Blind swarms for coverage in 2-D, Proceedings of Robotics: Science and Systems, Cambridge, USA (2005).
[98] V. Snasel, P. Moravec, D. Husek, A. A Frolov, H. Rezankova, P. Polyakov, Pattern discovery for high-dimensional binary datasets, Neural Information Processing (ICONIP), Part I, LNCS, Volume: 4984 Pages: 861–872, 2008
[99] E. H. Spanier, Algebraic Topology, McGraw-Hill 1966.
[100] H. Strange, R. Zwiggelaar, Open Problems in Spectral Dimensionality Reduction, (SpringerBriefs in Computer Science) . Springer (2014).
[101] A. Tausz, Persistent Homology in R; Available at CRAN http://cran.r-project.org, (visited on February 2016).
[102] A. Tausz, M. Vejdemo-Johansson, H. Adams, Javaplex: A research software package for persistent (co)homology, Software available at http://appliedtopology.github.io/javaplex/. (visited on February 2016).
[103] Ch. M. Topaz, L. Ziegelmeier, T. Halverson, Topological Data Analysis of Biological Aggregation, Models Plos One (2015) doi:10.1371/journal.pone.0126383.
[104] M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, N. Koudas, Non-linear dimensionality reduction techniques for classification and visualization. In: Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 645–651 (2002)
[105] C. Wagner, C. Chen, E. Vucini, Efficient computation of persistent homology for cubical data, in: Topological Methods in Data Analysis and Visualization II, Springer, Heidelberg, Dordrecht, London, New York, 2012.
[106] J. Wan, H. Yan, Q. Liu, K. Zhou, R. Lu, D. Li, Enabling cyber-physical systems with machine-to-machine technologies, International Journal of Ad Hoc and Ubiquitous Computing 13 (3/4) (2013) 187–196.
[107] J. Wan, D. Zhang, S. Zhao, L. Yang, J. Lloret, Context-aware vehicular cyber-physical systems with cloud support: architecture, challenges, and solutions, IEEE Communications Magazine 52 (8) (2014) 106–113.
[108] B. Wang, B. Summa, V. Pascucci, M. Vejdemo-Johansson, Branching and circular features in high dimensional data, IEEE Trans. Vis. Comput. Graph. 17 (2011) 1902–1911.
[109] S. Wang, J. Wan, D. Zhang, D. Li, Ch. Zhang, Towards smart factory for industry 4.0: a self-organized multi-agent system with big data base d fee dback and coordination, Computer Networks 101 (2016) 158–168.
[110] D. Wang, Z. Han, Sublinear Algorithms for Big Data Applications, Springer 2015.
[111] X. Wang, W. Liu, J. Li, X. Gao, A novel dimensionality reduction method with discriminative generalized eigendecomposition, Neurocomputing 173 (2016) 163–171.
[112] B. Wanga, G.-W. Weib, Object-oriented persistent homology, Journal of Computational Physics 305 (2016) 276–299.
[113] F. Xhafa, L. Barolli: Semantics, Intelligent processing and services for big data, Future Generation Computer Systems 37 (2014) 201–202.
[114] K. L. Xia, G. W. Wei, Persistent homology analysis of protein structure, flexibility and folding, Int. J. Numer. Methods Biomed. Eng. 30 (2014) 814–844.
[115] Y. Yao, J. Sun, X. Huang, G. R. Bowman, G. Singh, M. Lesnick, L. J. Guibas, V. S. Pande, G. Carlsson, Topological methods for exploring low-density states in biomolecular folding pathways, J. Chem. Phys. 130(14) (2009) 144115. doi: 10.1063/1.3103496.
[116] X. Zhu, Persistent homology: An introduction and a new text representation for natural language processing. In Proceedings of the 23rd IJCAI, IJCAI13, AAAI Press (2013) 1953–1959.

[117] A. Zomorodian, Advances in Applied and Computational Topology, AMS 2012.
[118] A. Zomorodian, Topology for computing, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2009.
[119] A. Zomorodian, G. Carlsson, Computing persistent homology, Discrete Comput. Geom. 33 (2005) 249–274.
[120] A. Zomorodian, G. Carlsson, Localized homology, Comput. Geom. Theory Appl. 41(3) (2008) 126–148.
[121] IBM Corporation Customer Analytics: Now its personal!, http://www.slideshare.net/IBMExpOne/customer-analytics-now-its-personal, (visited on February 2016).
[122] http://www.mrzv.org/software/dionysus/, (visited on February 2016).
[123] https://news.google.com/newspapers?id=ZGogAAAAIBAJ&sjid=3GYFAAAAIBAJ&pg=933,5465131&dq=data-center&hl=en, (visited on February 2016).
[124] M. van Rijmenam, A Short History Of Big Data, https://datafloq.com/read/big-data-history/239, (visited on February 2016).
[125] http://www-03.ibm.com/ibm/history/, (visited on February 2016).
[126] http://www-03.ibm.com/ibm/history/history/year_1936.html, (visited on February 2016).
[127] https://www.ssa.gov/history/court.html, (visited on February 2016).
[128] Recommendations for implementing the strategic initiative INDUSTRIE 4.0. Final report of the Industrie 4.0 Working Group, http://www.acatech.de/de/publikationen/stellungnahmen/kooperationen/detail/artikel/recommendations-for-implementing-the-strategic-initiative-industrie-40-final-report-of-the-industr.html, (visited on February 2016).
[129] The Industrial Internet Consortium: A Global Nonprofit Partnership Of Industry, Government And Academia, March 2014, http://www.iiconsortium.org/about-us.htm, (visited on February 2016).

*Biographies (Text)

Václav Snášel studied numerical mathematics at Palacký University in Olomouc, PhD degree obtained at Masaryk University in Brno, he teaches as professor at VSB – Technical University of Ostrava. 2001-2009 worked as researcher at The Institute of Computer Science of Academy of Sciences of the Czech Republic. Since 2009 he works as head of research programme at IT4Innovation National Supercomputing Center, currently as dean of the Faculty of Electrical Engineering and Computer Science. He works in a multi-disciplinary environment involving artificial intelligence, bioinformatics, Big Data, knowledge management, machine intelligence, neural network nature and biologically inspired computing, data mining, and applied to various real world problems.

Jana Nowaková received her M.Sc. in Measurement and Control from VŠB–Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science in 2012. Nowadays she continues her studies in Technical Cybernetics. She is interested in addition to fuzzy modelling, data processing, knowledge management, bio-inspired computing, also in statistical data processing in cooperation with University Hospital Ostrava. She works as a researcher in Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava and in IT4Innovation National Supercomputing Center.

Fatos Xhafa received his PhD in Computer Science in 1998 from the Department of Computer Science of the Technical University of Catalonia (UPC), Spain. Currently, he holds a permanent position of Professor Titular (Hab. Full Professor) at UPC. He was a Visiting Professor at University of London, UK, 2009-2010 and Research Associate at Drexel University, USA, 2004/2005. Prof. Xhafa has published in international journals, conferences/workshops, chapters, books and proceedings. He is editor in Chief of IJGUC and IJSSC, Inderscience and of the Elsevier Book Series "Intelligent Data-Centric Systems". His research interests include parallel and distributed algorithms, massive data processing and collective intelligence, optimization, networking, P2P, Cloud computing, security and trustworthy computing, among others.

Leonard Barolli received his BE and PhD from Tirana University, Albania and Yamagata University, Japan in 1989 and 1997, respectively. He has been working as a JSPS Post Doctor Fellow Researcher and Research Associate at Yamagata University, Assistant Professor at Saitama Institute of Technology (SIT) and Associate Professor at Fukuoka Institute of Technology (FIT), Japan. He is currently a Full Professor at Department of Information and Communication Engineering, FIT.
He has published more than 600 papers in refereed journals and international conference proceedings. He is the Steering Committee Co-Chair of IEEE AINA, BWCCA, 3PGCIC, NBiS, INCoS, CISIS, and IMIS international conferences. His research interests include network traffic control, network protocols fuzzy control, genetic algorithms, ad-hoc and sensor networks, IoT, big data, web-based applications and P2P systems. He is a member of SOFT, IPSJ, and IEEE.

Jana Nowakova

Click here to download high resolution image

Fatos Xhafa

Click here to download high resolution image

Leonard Barolli

Click here to download high resolution image