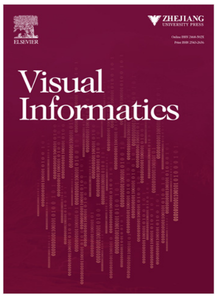


# A theoretical model for pattern discovery in visual analytics

Natalia Andrienko, Gennady Andrienko, Silvia Miksch,
Heidrun Schumann, Stefan Wrobel

PII: S2468-502X(20)30093-0
DOI: https://doi.org/10.1016/j.visinf.2020.12.002
Reference: VISINF 90

To appear in: *Visual Informatics*

Received date : 11 November 2020
Revised date : 21 December 2020
Accepted date : 21 December 2020

Please cite this article as: N. Andrienko, G. Andrienko, S. Miksch et al., A theoretical model for pattern discovery in visual analytics. *Visual Informatics* (2020), doi: https://doi.org/10.1016/j.visinf.2020.12.002.

# A theoretical model for pattern discovery in visual analytics

Natalia Andrienko[a,b,*], Gennady Andrienko[a,b], Silvia Miksch[c], Heidrun Schumann[d], Stefan Wrobel[a,e]

[a] Fraunhofer Institute IAIS, Sankt-Augustin, Germany
[b] City, University of London, London, UK
[c] TU Wien, Vienna, Austria
[d] University of Rostock, Rostock, Germany
[e] University of Bonn, Bonn, Germany

## Abstract

The word 'pattern' frequently appears in the visualisation and visual analytics literature, but what do we mean when we talk about patterns? We propose a practicable definition of the concept of a pattern in a data distribution as a combination of multiple interrelated elements of two or more data components that can be represented and treated as a unified whole. Our theoretical model describes how patterns are made by relationships existing between data elements. Knowing the types of these relationships, it is possible to predict what kinds of patterns may exist. We demonstrate how our model underpins and refines the established fundamental principles of visualisation. The model also suggests a range of interactive analytical operations that can support visual analytics workflows where patterns, once discovered, are explicitly involved in further data analysis.

*Keywords:* Visual analytics, data distribution, pattern, abstraction, data organisation, data arrangement, data variation, pattern discovery

[*] Corresponding author
*Email address:* natalia.andrienko@iais.fraunhofer.de (Natalia Andrienko)

## 1. Introduction

### 1.1. Motivation

We began to feel a need in a conceptual and theoretical model for visual analytics when we started teaching visual analytics to students of a data science course [1]. Every year it is necessary to explain the students what visual analytics is, why and for what purposes they will need to use visual analytics in their job, how to utilise visual analytics techniques in practice, what principles are important to obey, and why these principles exist. It turned out to be not easy to explain these things clearly and convincingly to practice-oriented and computation-minded people. In particular, when we tell the students that visualisation is required for observing distributions and detecting patterns, we need to explain them the meaning of the terms “distribution” and “pattern”. We want the students to understand that the meaning of “distribution” is not limited to statistical distribution of values of a variable, and this requires us to give a general definition which would cover the concepts of statistical, spatial, temporal, and, desirably, also other principally possible distributions. We need to teach the students how to find patterns in distributions, and this requires defining what a pattern is and what kinds of patterns, and why, can exist in different types of distributions.

Although this work has been originally motivated by pedagogical needs, we believe that having a clear conceptual and theoretical background can also be beneficial for visual analytics science as well as engineering. Explicitly defined rather than intuitively understood concepts can potentially enable systematic approaches to conducting research work and to developing new methods and procedures. Solid theoretical foundations of visual analytics could be especially helpful when entering new application domains or dealing with new types of data.

### 1.2. Goals and purposes

With this work, we pursue the following **goals**:

- Introduce an explicit working definition of the concept of *pattern in data* (Section 4.1).
- Describe how properties of data determine the *types of possible patterns* that can exist in the data (Section 4.2).
- Draw implications for the possible visual analytics approaches to discovering patterns existing in data (Section 7).
- Use the explicit definition of a pattern to explain some of the existing principles of visualisation design (Section 7.3, 7.4).
- Describe how patterns that have been discovered can be utilised in further data analysis (Section 6).

We expect that the proposed theoretical model will be useful for the following **purposes**:

- For data analysis *practitioners*: provide a ground for informed and reasoned anticipation of the possibly existing types of patterns in given data and selection of techniques for finding these patterns.
- For *developers* of visual analytics methods and procedures: provide foundations for
  - methodical design of approaches and analytical workflows involving discovery and exploitation of different types of patterns;
  - development of approaches to guiding users and supporting the externalisation of the knowledge gained by them in the process of data analysis.
- For visual analytics *researchers*: underpin systematic development of principles and general approaches to analysing different kinds of data.
- For *students* of visual analytics and/or data science: enable better understanding of patterns, and how they can be used in visual data analysis.

### 1.3. Main ideas

The essence of our model can be summarised in the following statements:

- A *pattern* consists of relationships between multiple elements of at least two data components.
- A pattern is such a combination of relationships that allows multiple elements to be perceived and/or represented holistically *as a single object*, as, for example, a cluster, a trend, or a correlation.
- The *types of relationships* existing between elements of data (such as ordering and distance relationships) determine the possible *types of patterns* that can be made by these elements.
- *Pattern discovery*, which involves abstraction, is a principal way to understand synoptic relationships between data components.
- To discover patterns, analysts investigate *distributions* of elements of one (or more) components with respect to elements of another component and relationships between these elements.
- Analysts can use discovered patterns in next steps of data analysis by applying *analytical operations*, such as aggregating, grouping, comparing, and others.

In the following, after reviewing the related work, we shall explain, justify, and elaborate these statements.

## 2. Related work

Here we discuss how the concept of pattern is treated in different scientific disciplines.

### *2.1. Patterns in Mathematics*

Modern mathematicians tend to see mathematics as a science of patterns [2]. It is argued that the primary subject of study in mathematics is not the individual mathematical objects but rather the structures (patterns) in which they are arranged [3]. Here, the term ‘pattern’ is used as a synonym for ‘structure’.

A pattern consists of one or more objects, called *positions*, which stand in various *relationships*. Positions as such have no distinguishing features. Only within a pattern positions may be identified or distinguished, since the pattern containing them provides a *context* for so doing. Thus, in a triangle ABC, the points A, B, C can be differentiated when considered as triangle vertices, but taken in isolation they are indistinguishable from each other and from other points.

A context provided by a pattern can be viewed as a *representation system* [4], and the use of different systems of representation results in seeing different aspects. Thus, the same thing can be seen as a table, as a composition of table-parts, as a collection of molecules, etc., and all these views are correct. Oliveri [4] emphasises that the aspect we perceive is not a property of an object itself but a *relation* between it and other objects.

To summarise, mathematicians define patterns as arrangements of objects in which only *relationships* between the objects are important but not properties of the objects themselves. Patterns have properties that are based on the relationships between the objects and do not apply to the objects taken separately. Giving different representations to the same objects allows perceiving different patterns, which can complement each other.

Mathematicians define pattern types according to the branches of mathematics [2]: arithmetic deals with patterns of numbers, mathematical logics with patterns of reasoning, calculus with patterns of change and motion, geometry with shapes and symmetry, and topology with patterns of connectivity and reachability.

In our work, we deal with patterns existing in data, i.e., made by elements of data. Like mathematicians, we acknowledge the key role of relationships in forming patterns. Pattern types can be defined based on the types of relationships existing between data elements.

### *2.2. Patterns in Statistics*

There is no explicit definition of the concept of data pattern in statistics; nevertheless, the expressions “data patterns” or “patterns in data” are extensively used in statistical literature [5, 6]. Patterns in data distributions are commonly described in terms of centre, spread, shape (or form), and presence of particular features, such as gaps and outliers. Several types of patterns are specifically defined for time series data [7], namely, trend, seasonal, cyclic, and irregular (random) patterns. Trend patterns are further differentiated into linear, exponential, and other subtypes.

The concept of distribution, in turn, is defined as a function that associates each value of a variable with its probability [8]. Statistics considers various forms of distributions [9], such as normal, uniform, bimodal, long tail, etc.

While the definition of distribution in statistics is limited to probability distribution, we give a more general definition covering also spatial distributions as well as other imaginable kinds of distributions. Another extension is consideration of relationships between data elements and the role of these relationships in forming data patterns. Thus, the types of patterns that can be found in time series data are made by specific relationships (namely, ordering and distances) between time steps and between corresponding values of a variable.

### *2.3. Patterns in Geography-related Sciences*

All sciences studying phenomena that occur on the Earth, including natural, social, and economic phenomena, are concerned with analysing spatial distributions and spatial patterns. A pattern in a spatial distribution is defined in terms of the *arrangement* of individual entities in space and the geographic *relationships* among them [10, 11]. Geographic analysis usually involves observing and describing spatial patterns, testing whether the observed pattern differs from a null model, such as complete randomness, and fitting empirical data to theoretical models for the purposes of prediction [12]. Spatial patterns are characterised by specific metrics of concentration or dispersion, eccentricity, randomness, clustering, etc. [11]. An important characteristic is spatial autocorrelation indicating how an object or feature located in space is influenced by similar objects or features in the neighbourhood [10].

It is acknowledged that patterns that can be observed in spatial distributions are dependent on the spatial scale of analysis [13, 14]. Thus, the kinds of patterns that can exist in the global distribution of a biological species are very different from the possible kinds of patterns in a local distribution of individuals belonging this species.

Our definitions of the concepts of distribution and pattern cover, in particular, the concepts of spatial distribution and spatial pattern. Our model can explain the role of spatial relationships in forming spatial patterns.

### *2.4. Patterns in Information Theory*

In information theory [15, 16], the term 'pattern' may refer to any distinct arrangement of symbols or to a combination of pixels in an image, regardless of whether it is meaningful or interesting. In the context of an application, all possible data patterns collectively define a so-called *alphabet*, where each pattern is a *letter*. In data compression, the resources used to encode different patterns are optimised according to the probability of the patterns in the data space. In image processing and computer vision, patterns are broadly divided into groups, which are mathematically specified. Various algorithms were developed to differentiate patterns in one group from others. They make use of different information-theoretic metrics for pattern recognition, matching, segmentation, registration, etc. [17, 18].

Ideas and techniques from information theory have been used for characterising and studying pattern recognition by humans. Chen et al. [19] noticed that humans' ability to identify interesting patterns when they are overlapped with other patterns and to connect interesting patterns when they are distributed away from each other bears some resemblance to the family of techniques called multiplexing in tele- and data communication. The researchers used information theory to explain this phenomenon of visual multiplexing in visualisation. In a survey of a large collection of empirical studies concerning visualisation [20], the studies were categorised according to the main independent variables: contexts (e.g., tasks, applications), patterns (e.g., clusters and changes), and values (e.g., data values and statistics). It was noticed that patterns were in the focus of about 50% of the studies. Kijmongkolchai et al. [20] also conducted an empirical study to detect and measure human's knowledge used in reasoning about time series patterns. They found that the human's prior knowledge on pattern identification brought more benefit than that on context awareness and statistical estimation. The benefits were measured using the information-theoretic metric for cost-benefit analysis [21].

Importantly, the process of pattern perception and recognition by humans involves abstraction. Since the information-theoretical view of a pattern does not accommodate the notion of abstraction, it cannot support the description of the phenomenon of pattern discovery by means of visual analytics.

### *2.5. Patterns in Data Mining*

Data mining is defined as an automatic or semi-automatic process of discovering *useful* patterns in data [22]. A pattern is defined as "an expression $E$ in some language $L$ describing facts in a subset $F_E$ of a set of facts $F$ so that $E$ is simpler than the enumeration of all facts in $F_E$" [23, p.7]; in other words, a pattern is defined as a synoptic *representation* of multiple data items.

Han [24] states that types of patterns can be defined according to data mining functionalities, which include: characterisation and discrimination; mining of frequent patterns, associations, and correlations; classification and regression; cluster analysis; outlier analysis. In practice, what is usually called 'pattern types' in data mining literature rather refers to the existing forms of outputs of data mining methods, such as decision trees, classification rules, clusters, frequent item sets, frequent sub-sequences, etc. [25, 26, 22]. There is no underlying scheme for a more systematic definition of possible pattern types.

An important difference of our conceptual model is acknowledging that patterns objectively exist in data regardless of any representation or someone's awareness of their existence. By defining a pattern as a structure formed by relationships between data elements, we provide a basis for anticipating what kinds of patterns can exist in given data.

### *2.6. Patterns in Visualisation and Visual Analytics*

Similarly to statistics, visualisation literature often uses the expressions "pattern(s) in data" or "data pattern(s)", although there is no commonly adopted explicit definition of what this term means. Thus, Munzner treats the term 'pattern' as a synonym to 'trend' [27], whereas others use this term as self-evident without explaining what they mean by it. There was an attempt to adapt the data mining definition: a pattern was defined as a parsimonious representation of essential features of a behaviour in the form of a description in some language (natural, formal, or graphical) or a mental image of the behaviour [28, p. 85].

Visual analytics can be seen as a model building activity [29] in which an analyst creates a model, in particular, a mental model, of the analysis subject. A model needs to be general, i.e., refer to multiple observations taken together rather than represent each observation separately. Collins et al. [30] argue that, in order to generalise, analysts should be able to perceive multiple data items together and conceptualise them jointly as a meaningful whole. Such a whole is called a *pattern*. Collins et al. propose the following definition of a pattern: "a representation of a collection of items of any kind as an integrated whole with specific properties that are not mere compositions of properties of the constituent items". This is similar to the definition given in data mining; a pattern is also defined as a representation rather than an objectively existing structure.

According to Bertin, understanding of data means "discovering combinational elements which are less numerous than the initial elements yet capable of describing all the information in a simpler form" [31, p. 166]. In fact, what is called "combinational elements" here corresponds to what is usually meant by a pattern in data: it is a structure formed by multiple elements, and it can be described holistically without enumerating these elements.

Perception of patterns from visual representations of data is extensively discussed in the Colin Ware's book [32]. Pattern perception involves seeing multiple visual elements (a.k.a. "marks", in Bertin's terms) as an integrated whole. The first attempt to understand this process was undertaken by the Gestalt School of psychology [33, 34]. Ware discusses the Gestalt "laws" of pattern perception and shows how they translate into principles of visualisation design. The Gestalt laws refer to certain *relationships* between visual marks, such as proximity (in the display space), similarity, smooth continuity, symmetry, and relative size. Visible patterns can emerge due to these relationships. Acknowledging that data patterns are formed by relationships between data elements leads to an obvious implication that visual representations can effectively and correctly reveal patterns existing in data when the relationships between the marks representing data elements correspond to the relationships between the data elements.

In our theoretical model, we strive to give definitions that can underpin the main principles of visualisation. We attach high importance to relationships between data elements as pattern-forming forces and to the phenomenon of abstraction, which is involved in perception and representation of multiple related data elements as a unified whole.

Our use of the term 'theoretical' corresponds to the definition of a theory as "a set of interrelated constructs (concepts), definitions, and propositions that present a systematic view of phenomena by specifying relations among variables, with the purpose of explaining and predicting the phenomena" [35, p.11].

We do not pretend that our model can describe everything in visual analytics. Visual analytics is concerned not only with finding patterns in data but also with other analytical activities, such as search for specific information (e.g., clues to identify a criminal) or inspection of the performance of a computer model. Our theoretical model refers only to the process of finding patterns in data. It is an important type of analytical activity addressed in a large part of the visual analytics research. We believe that this research will benefit from the clarification of the concept of pattern in data.

## 3. Distribution

We begin presenting our theoretical model with defining and explaining the concept of *distribution*. We describe *relationships* within data components and establish a *formal notation* of the introduced concepts. This provides us with the necessary background to define and discuss *patterns*.

### *3.1. Definition of data distribution*

Among multiple existing definitions of the term "distribution", the following ones express the meaning relevant to our model: "the position, arrangement, or frequency of occurrence (as of the members of a group) over an area or throughout a space or unit of time" [36] and "the way that something is shared or exists over a particular area or among a particular group of people" [37]. An important part of these definitions is that something is positioned or spread over or throughout or among something else; the latter may be, in particular, space, time, or a group of people.

We shall build on these definitions to generate a more specific definition of *distribution of data*, or *data distribution*. A data distribution involves at least two components of data. For example, in the VAST Challenge 2011 dataset [38], the data records describing the microblog messages include the following components: microblog users (denoted by identifiers), times when the messages were posted, locations from where they were posted, and message texts. Besides, the data provided for the challenge include a map of the territory and daily weather records specifying, in particular, the wind speed and direction. Furthermore, since message texts consist of words, the set of the words is also a component of the data. To solve the challenge, analysts need to consider the distributions of the messages and of the words over the time and space, and the distribution of the wind parameters over the time.

This example demonstrates that data components are usually sets consisting of certain elements: people, messages, words, spatial locations, time moments, particular values of wind parameters, etc. Data describe *connections between elements of different components*. Thus, each message text is connected to a particular person, time moment, spatial location, and words that are used in the text. Each word is connected to the messages in which it is used. Each time moment is connected to the messages that were posted at that moment, each spatial location is connected to the messages posted from it, and so on.

A *data distribution* consists of connections between elements of two or more structural components of data. A data component is a set of items of the same kind, e.g., a set of entities, or attribute values, or category labels, or references to places or times. Data components are typically represented by fields of database records or by table columns. Data components involved in a distribution are not treated semantically equally. Each time when we talk about a distribution we say that one component (or a group of components) is *distributed over* another component. It means that the second component is treated as a kind of *base* for the first component. Generally, the base of a distribution must not necessarily be space, time, or a group of people, as stated in the definitions from the dictionaries, but it can consist of elements of any nature (these may also be compound elements consisting of several simpler ones). For example, we can consider the distribution of the words over the messages, in which the base is the set of messages. When we consider the distribution of the messages over space and time, the base consists of compound elements comprising spatial locations and time moments.

The concept of distribution assumes that the elements of the base are regarded as a kind of *positions* that can be occupied by elements of another component, or as *holders* of elements of another component. Thus, space and time provide positions for messages, messages can be seen as positions for words, or as holders of words, people can be seen as holders (i.e., owners) of the messages they have produced, time units can be seen as holders of particular values of wind parameters, etc.

We shall use the term *overlay of the distribution* to refer to a set of elements that are connected to positions or holders in the base: in a metaphorical sense, this set is *laid over the base*. Like the base, the overlay may consist of any kind of elements, including compound elements. The elements of the overlay are *instances (occurrences)* of elements of some data component that is distinct from the base. This data component can be called the *domain* of the overlay.

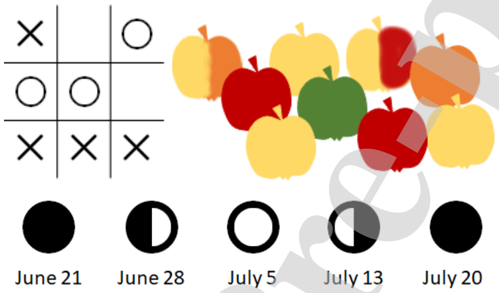


Figure 1: Simple examples of distributions. Top left: a distribution of symbols over a grid in a tic-tac-toe game. Top right: a distribution of colours over a set of apples. Bottom: a distribution of the shapes of the moon over time.

Let us illustrate the concepts of distribution, base, and overlay by simple examples shown in Fig. 1. In the tic-tac-toe game, players create distributions of crosses and noughts (X and O symbols) over a $3 \times 3$ grid (top left). Here, the base is the grid; the cells are the elements of the grid, which can serve as positions to the symbols. The overlay is the set of instances of the symbols X and O placed in particular positions in the grid. The domain of the overlay is the set of symbols {X, O}.

The upper right part of Fig. 1 demonstrates a distribution of colours over a set of apples. Here, the set of apples is the base of the distribution. The apples are holders of different colour instances, which make the overlay of the distribution. The domain of the overlay is the set of colours {yellow, orange, red, green}.

The lower part of Fig. 1 shows a distribution of the moon shapes over time.

Here, the time is the base of the distribution, and the overlay consists of different shapes of the moon arranged in a particular way. The illustration in Fig. 1 does not show the full distribution. The base of the full distribution includes all intermediate dates between those specified in the picture and also extends to the past and to the future beyond the period shown. The overlay of the full distribution includes instances of all intermediate shapes between the shapes shown in the picture. The domain of the overlay is the set of all possible unique shapes the moon can have.

Data distributions are analysed in order to understand relationships between components of data, for example, the relationship between the moon shape and the course of time. It is mostly a matter of common sense or convenience which of the data components should be viewed as the base and which as the overlay. Thus, it is more natural to see the time as the base for the moon shapes than the set of possible moon shapes as the base for different dates and times. It is also more natural to consider the grid in the tic-tac-toe game as the base for the X and O symbols than the other way around. As apples can be easily treated as holders of colours, it is natural to see the set of apples as the base for the colours and less natural to consider the set of colours as the base for the apples.

The examples in Fig. 1 illustrate an important property of the base of a distribution: it consists of *unique elements*. This means that, when a data component is chosen as a distribution base, the base is composed of a single occurrence, or instance, of each element of this component. The overlay is formed by the elements from another data component that are connected to each element of the base. It may happen that the same element of the other component is connected to more than one element of the base. Hence, the overlay will contain multiple instances of the same element: multiple cross and nought symbols, multiple instances of the same colour, re-occurrences of the same moon shape, etc. It may also happen that two or more overlay elements are connected to the same element of the base, as two colours can be connected to the same apple. The tic-tac-toe example demonstrates that some elements of the base may have no connected elements of the overlay.

Let us summarise our discussion in the following definition of a data distribution:

**Definition 1**: Let $S^B$ and $S^\Omega$ be two sets, and let the elements of $S^B$ be treated as as positions or holders of elements from $S^\Omega$. The **distribution** of $S^\Omega$ over $S^B$ is the set of all connections between elements of $S^B$ and elements of $S^\Omega$ that are specified in data, i.e., $D(S^\Omega/S^B) = \{(e^B, e^\Omega)|e^B \in S^B, e^\Omega \in S^\Omega\}$. The set $S^B$ is called the **base** of the distribution. The set of all **instances of elements** from $S^\Omega$ that occur in $D(S^\Omega/S^B)$ in connection with elements of $S^B$ is called the **overlay** of the distribution, and the set $S^\Omega$ is the **domain of the overlay**. The elements of $S^\Omega$ are called **prototypes** with respect to their instances occurring in the overlay.

We shall use the symbol $B$ or $B(D)$ to denote the base of a distribution and the symbol $\Omega$ or $\Omega(D)$ to denote the overlay. According to Definition 1, $B = S^B$, whereas $\Omega$ is not the same as $S^\Omega$. $\Omega$ may contain multiple instances of the same element of $S^\Omega$, while some other elements of $S^\Omega$ may be absent in $\Omega$. Since each overlay element has its prototype in the overlay domain, it can be said that overlay elements are linked to their prototypes by instantiation relationships. We shall call the set of these instantiation relationships the *composition of the overlay*.

**Definition 2**: The **composition** of the overlay of a data distribution is the set of instantiation relationships between the elements of the overlay and their prototypes in the domain of the overlay.

Overlay composition can be described in terms of the number of instances of each element of the overlay domain. Thus, the overlay composition in the tic-tac-toe game (top left of Fig. 1) consists of four instances of the symbol X and three instances of the symbol O. The overlay composition in the set of coloured apples (top right of Fig. 1) includes five instances of the yellow colour, three instances of red, two instances of orange, and one instance of green. In the distribution of the moon shapes (Fig. 1, bottom), the overlay composition includes two instances of the "new moon" shape (i.e., dark disc) and one instance of each other shape.

### 3.2. Within-component relationships

Within any data component, elements may be linked by relationships. There are two major groups of relationships: qualitative and metric. Qualitative relationships can be represented by logical statements (predicates) saying if a relationship exists or not. Examples are relationships of equivalence, ordering, adjacency, or kinship. Metric relationships can be represented by numeric values. Examples are relationships of distance, similarity, or intensity of communication.

Some of existing relationships may be intrinsic, belonging to the very nature of a data component. For example, there are intrinsic relationships of ordering and distance between elements of a temporal component, i.e., between time units, and intrinsic relationships of distance between elements of a spatial component, i.e.., spatial locations. Intrinsic relationships are usually not represented in data explicitly, but, when needed, explicit representations can be obtained in well-known ways. Non-intrinsic qualitative relationships, such as kinship, need to be represented explicitly in data. Non-intrinsic metric relationships, such as similarity, need to be computed by appropriate methods.

**Definition 3**: The set of all relationships existing between elements of a data component is called the **organisation** of this data component.

For example, the organisation of the set of grid cells in the tic-tac-toe game includes qualitative relationships of adjacency, horizontal ordering, and vertical ordering. The set of symbols {X, O} has no relationships except identity: X=X, O=O, X≠O. The set of apples and the set of colours on the top right of Fig. 1 also have only identity relationships between their elements. The organisation of the set of dates in Fig. 1, bottom, includes qualitative relationships of linear ordering and metric relationships of distance (i.e., time difference) between the elements. The organisation of the set of moon shapes includes relationships of ordering and distance between the sizes of the visible parts of the moon and same-different relationships between the sides (right or left) of the visible parts in the moon disc.

Definition 3 relates to the Bertin's concept of the *level of organisation*, which may be qualitative (nominal), ordered, or quantitative. Each level implies particular types of relationships between data elements: the qualitative level has no ordering and no metric relationships, the ordered level has ordering relationships but no metric relationships, and the quantitative level implies the existence of both ordering and metric relationships. However, there may be components that have metric relationships but no ordering (e.g., 2D or 3D space), and there may be components with partial ordering relationships (e.g., ancestor-descendant relationships among people). Therefore, we introduce a more general definition of organisation as a set of all existing relationships between elements. Since all possible combinations of different types of relationships cannot be arranged into a single sequence of levels, we use the term "organisation" rather than "level of organisation". Our definition also corresponds to the term "mathematical structure" (of data) used by Kindlmann and Scheidegger [39].

### *3.3. Aspects of a data distribution*

In a data distribution, the elements of the overlay get *arranged* according to the organisation of the base. For example, the instances of the X and O symbols in the tic-tac-toe game get arranged according to the relationships of adjacency, horizontal ordering, and vertical ordering between the grid cells in which they have been put. The colour instances in Fig. 1, top right, are arranged by the identity relationships between the apples holding them: two colour instances have either a common holder or distinct holders. The instances of the moon shapes in the lower part of Fig. 1 are arranged in a row by the ordering and distance relationships between their temporal positions. Let us introduce a formal definition of the overlay arrangement:

**Definition 4**: **Arrangement relationships** between elements of the overlay of a data distribution are the relationships between the corresponding elements of the base. The **arrangement** of the overlay of a data distribution is the set of the arrangement relationships between the overlay elements.

In addition to arrangement relationships, overlay elements are linked by the relationships pertaining to the organisation of the overlay domain $S^\Omega$, i.e., by the relationships that exist between the prototypes of the overlay elements. Let $b_1$ and $b_2$ be two elements of the distribution base $B$, and let $o_1$ and $o_2$ be the elements of the overlay $\Omega$ connected to $b_1$ and $b_2$, respectively. The domain-pertinent relationships between $o_1$ and $o_2$ can be treated as the way in which the overlay *varies* between position or holder $b_1$ and position or holder $b_2$. It is important to note that the domain-pertinent relationships between $o_1$ and $o_2$ are considered in connection to their positions in the base $b_1$ and $b_2$ and the relationships existing between these positions, or, in other words, in connection to the arrangement relationships between $o_1$ and $o_2$ (Definition 4).

**Definition 5**: The **variation** of the overlay of a distribution with respect to the base consists of the domain-pertinent relationships between the overlay elements (i.e., relationships belonging to the organisation of the overlay domain) considered in connection to the arrangement relationships between the overlay elements.

Generally, base elements may contain or hold multiple elements from the overlay domain $S^\Omega$ or no such elements. The variation of the overlay with respect to the base includes relationships between any two instances of the elements of the overlay domain, either having distinct holders or the same holder. The relationships of having distinct or same holders are a part of the overlay arrangement; hence, Definition 5 is applicable. To deal with cases when base elements have no connected overlay elements, we shall assume that the overlay domain includes a special *null* element denoting the absence of any other element. The null element has no relationships to the other elements except of being not identical to any other element.

The composition, arrangement, and variation of the overlay of a data distribution will be called the *aspects* of the distribution. The concept of a data

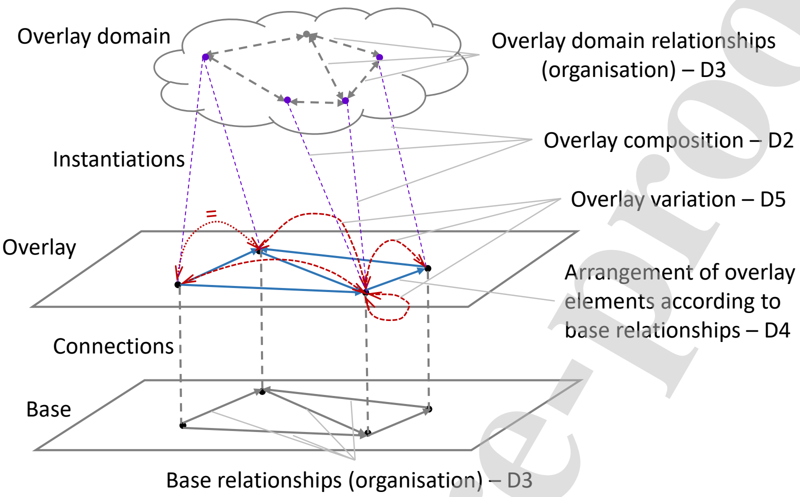

Overlay domain
Overlay domain relationships (organisation) – D3
Instantiations
Overlay composition – D2
Overlay variation – D5
Overlay
=
Arrangement of overlay elements according to base relationships – D4
Connections
Base
Base relationships (organisation) – D3

Figure 2: A schematic illustration of the definitions of a data distribution and its aspects. The colours distinguish the composition (purple), arrangement (blue), and variation (red) of the overlay. The labels D2 to D5 refer to the definitions from 2 to 5.

distribution and its aspects are schematically illustrated in Fig. 2.

Let us introduce a formal notation for the aspects of a distribution, which will help us to clarify what they are made of and how they are related to each other. We already use the symbol $B$ to denote the distribution base and $\Omega$ for the overlay. The notation $C^{\Omega}$ will refer to the *composition* of the overlay (Definition 2). The overlay composition is determined by the existing connections of overlay domain elements to base elements (Definition 1). To reflect this dependency, we shall use the expression $C^{\Omega}(B)$.

The symbol $Or$ will denote the *organisation* of a set (Definition 3). The expression $Or^B$ refers to the organisation of the base and $Or^{\Omega}$ to the organisation of the overlay, which is the same as the organisation of the overlay domain. The symbol $Ar^{\Omega}$ will refer to the *arrangement* of the overlay elements according to the organisation of the base (Definition 4). $Ar^{\Omega}$ is imposed by $Or^B$, i.e. it is a function of $Or^B$. To emphasise this dependency, we shall denote the arrangement of the overlay as $Ar^{\Omega}(Or^{B})$. In the tic-tac-toe example, $Ar^{\Omega}(Or^{B})$ consists of the particular placements of the cross and nought symbols in the grid cells. For the apples, $Ar^{\Omega}(Or^{B})$ consists of the particular colouring of each apple, including both the unicolour and bicolour variants. For the moon shape, $Ar^{\Omega}(Or^{B})$ is the particular sequence of the instances of the moon shapes corresponding to the sequence of the days.

The *variation* of the overlay, i.e., the relationships between the overlay elements within the arrangement (Definition 5) can be represented by the notation $Var^{\Omega}(Ar^{\Omega}(Or^{B}), Or^{\Omega})$. This means that the variation exists within a specific arrangement $Ar^{\Omega}(Or^{B})$ and involves relationships from the overlay organisation $Or^{\Omega}$. In the tic-tac-toe example, the variation is the manner in which the cell content changes as the grid is traversed. In the example with the apples, the variation consists of the similarities and differences between the apples in terms of their colouring. In the example with the moon shapes, the variation is the manner in which the moon shape changes from day to day along the time, i.e., how each shape in the succession relates to the previous one.

The formal notation reflects the asymmetric roles of the base and overlay of a distribution: while the base is considered as an independent component, the overlay is composed and arranged according to the base. The composition $C^{\Omega}(B)$ is the instantiation relationships between instances *connected to the base elements* and their prototypes. The arrangement $Ar^{\Omega}(Or^{B})$ is the structure made of the *base-specific relationships* between the positions or holders of these instances. In turn, the variation of the overlay $Var^{\Omega}(Ar^{\Omega}(Or^{B}), Or^{\Omega})$ depends on the arrangement and, through the arrangement, on the *organisation of the base*.

Using these concepts, we can formulate the ***general task of analysing a distribution*** as follows: given a data distribution $D(S^{\Omega}/S^{B})$, characterise the composition, arrangement, and variation of the overlay, i.e., $C^{\Omega}(B)$, $Ar^{\Omega}(Or^{B})$, and $Var^{\Omega}(Ar^{\Omega}(Or^{B}), Or^{\Omega})$.

## 4. Patterns

### *4.1. Patterns in a distribution*

Usually, the purpose of analysing a distribution is to understand how two or more data components are related *in general*, i.e., as wholes. For example, the distribution of the symbols over the tic-tac-toe grid in Fig. 1, top left, would be examined to see whether there is a linear arrangement of three instances of the same symbol, irrespective of the specific positions of the symbols. The distribution of the colours over a set of apples (Fig. 1, top right) could be analysed for estimating the probabilities of finding apples of different colours rather than for ascertaining the colour of each particular apple. The temporal distribution of the moon shapes (Fig. 1, bottom) would be studied to understand how the moon shape changes over time in general, regardless of particular dates.

Data specify connections between individual elements of components. We shall thus call these connections *elementary*. In contrast, relationships between components as wholes will be called *synoptic*. Synoptic relationships are not mere compositions of elementary connections but have a higher level of generality. Understanding synoptic relationships based on elementary connections requires *abstraction*, which means that multiple elementary connections are united and considered all together.

How can elementary connections be unified? What is the force that can glue them together? It is the relationships between the elements within the data components, i.e., the relationships that belong to the internal organisation of the components. Let us illustrate this statement using the simple examples from Fig. 1.

On the top left, the organisation of the tic-tac-toe grid (i.e., the set of the spatial relationships between the cells) allows us to unite individual cells into horizontal, vertical, and diagonal lines. Simultaneously, the equivalence relationships between the symbol instances allow us to unite multiple instances of the same symbol in a group. The combination of the relationships between the cells and between the symbol instances allows us to consider groups of cells with equivalent symbol instances as certain shapes. In the set of the apples, there are only identity relationships between the apples, i.e., each apple is distinct from all others. This does not give an opportunity for unification. However, the colour instances can be grouped according to the equivalence relationships, and the groups can be characterised in terms of their sizes (i.e., colour frequencies) and intersections. In our example, the group of the instances of the yellow colour intersects with the groups of the instances of the red and orange colours.

At the bottom of Fig. 1, the ordering relationships between the time steps unite all time steps into a single time line and, simultaneously, arrange the different moon shapes into a succession. Then, the relationships between two neighbouring shapes in the succession can be seen as the change from the earlier to the later shape. If similar changes occur successively, they can be unified and considered all together as a *trend*. Thus, if we characterise the moon shape in terms of the visible fraction of the whole moon disc, we can unite the shapes of the first two weeks into the trend of increase of the visible fraction from 0 to 100% and the shapes of the following two weeks into the trend of decrease of the visible fraction from 100 to 0%. It is also possible to consider the succession of the changes in more detail, e.g., by taking into account on which side (right or left) of the moon disc the changes happen.

In all these examples, we used relationships between elements of data components for unifying multiple elements and multiple elementary connections. We described the objects resulting from the unification as shapes, groups, or trends, without referring to the elementary connections anymore; hence, we performed the operation of *abstraction*. According to the common understanding, the objects that we have obtained are examples of different kinds of patterns existing in data. Hence, a pattern in data is, generally, a combination of multiple connections and relationships between elements of data components such that there exists an operation of abstraction allowing to treat all these connections and relationships together as a single object. Given the principal possibility of considering one data component as the base of a data distribution and the other(s) as the overlay domain(s), we can use the previously introduced concepts to formulate the definition of a pattern in a data distribution:

**Definition 6**: A **pattern in a data distribution** is a subset of the *relationships* involved in the composition, arrangement, or variation of the overlay over the base such that there exists an operation of *abstraction* allowing to treat this subset as a *unified whole.*

We emphasise that a pattern consists of relationships, not of elements. Therefore, the same pattern (i.e., the same combination of relationships) may occur in data multiple times so that each occurrence connects different elements. For example, the pattern "three equal symbols next to one another" may occur several times in one tic-tac-toe game, and it may connect either crosses or noughts. Moreover, one and the same pattern may occur in different datasets and even in data of different nature. Thus, the pattern "three equal symbols next to one another" may also occur in a text or in musical notation. The groups of elements from the base and from the overlay that are connected by a pattern will be called the base and the overlay of the pattern, respectively.

**Definition 7**: The **base of a pattern** is the subset of the elements from the base of the overall distribution whose relationships and connections contribute to the pattern. The **overlay of a pattern** is the subset of the elements from the overlay of the distribution that are connected to the base of the pattern.

We shall use the notation $\beta$ and $\omega$ to refer, respectively, to the base and overlay of a pattern; $\beta \subseteq B$, $\omega \subseteq \Omega$. Accordingly, the expressions $C^{\omega}(\beta)$, $Ar^{\omega}(Or^{\beta})$, and $Var^{\omega}(Ar^{\omega}(Or^{\beta}), Or^{\omega})$ denote the composition, arrangement, and variation of the overlay of the pattern, which are different aspects of a pattern.

Definition 6 implies that patterns *objectively exist* in data. A pattern is not a product of observation or computation, it is a combination of relationships and connections that actually exist in data. Observation or computation can involve an abstraction operation that brings these connections and relationships together and represents as a unified whole. Hence, the product of observation or computation is a *representation of a pattern* rather than the pattern itself. There may be different forms of representation: verbal, symbolic (e.g., a formula), schematic, etc. The pattern itself does not depend on the representation form and on the way in which this representation has been obtained.

We shall use the term *abstracted pattern* to refer to a holistic representation of an objective pattern in any form and medium:

**Definition 8**: An **abstracted data pattern** is a representation of an objective pattern as a unified whole regardless of the form, language, and medium of the representation. An abstracted data pattern may represent the composition, arrangement, and/or variation of the pattern overlay with respect to the base.

The concept of an abstracted pattern corresponds to the definition of a pattern in data mining cited in Section 2.5. The set of facts in our case consists of all connections and relationships between elements involved in an objective pattern. However, our definition of an abstracted pattern refers not only to explicit expressions in some languages but also to internal representations constructed in the mind of a human observing objective patterns. The definition of an abstracted pattern is also consistent with the definition given by Collins et al. [30]. Unlike these previous definitions, Definition 8 emphasises the existence of an objective data pattern represented by an abstracted pattern.

The same objective pattern can be described very roughly in a short and simple expression or in a more refined and accurate manner using a longer and more complex expression. The possible expressions differ in their *degree of abstraction*: the more details are included, the lower the abstraction. For example, the expression “increasing trend in the morning” has a higher degree of abstraction than “increase by factor 1.6 in the interval from 8:00 till 10:00”.

A synoptic relationship between two or more data components can be understood and characterised by finding objectively existing data patterns and representing them by abstracted data patterns. This process is called *pattern discovery.*

### *4.2. Pattern types*

Patterns can be categorised first of all according to the data distribution aspects whose relationships are involved in the patterns, i.e., composition, arrangement, and variation. Based on this principle, we distinguish *composition patterns*, *arrangement patterns*, and *variation patterns*.

Composition patterns can be abstracted into frequency distributions or probability distributions (in the statistical sense) of the elements of the overlay domain. Composition patterns involve instantiation relationships between overlay elements and their prototypes (Definition 2) and do not involve any relationships from the organisation of the base of the data distribution. Relationship from the organisation of the overlay domain $Or^{\Omega}$ may be utilised in the abstraction operation applied to a composition pattern. For example, when the overlay domain consists of numeric values, the ordering and distance relationships between the values are usually involved in the construction of the frequency or probability distribution. On this basis, composition patterns can be further categorised as normal, exponential, left- or right-skewed, long-tailed, fat-tailed, etc.

Arrangement patterns are formed by relationships between base elements as signified by the expression $Ar^{\omega}(Or^{\beta})$. Types of arrangement patterns can be distinguished according to the types of relationships between the base elements. Thus, the pattern type commonly known as “spatial cluster” refers to an arrangement of overlay elements by relationships of spatial distance between positions in a spatial base. A well-known example is the cholera outbreak in London in 1854, when John Snow discovered that the deaths from cholera were arranged into a spatial cluster around the Broad Street. Arrangement patterns involving relationships of ordering between base elements (such as temporal ordering) may refer to the density of overlay elements along the order (high or low number of overlay elements corresponding to sub-sequences of consecutive base elements) and existence of gaps (positions in the order without corresponding overlay elements). When the base is time, the density of the overlay elements is usually referred to as temporal frequency.

Variation patterns involve relationships both from the organisation of the base (incorporated in the overlay arrangement) and from the organisation of the overlay, as signified by the expression $Var^{\omega}(Ar^{\omega}(Or^{\beta}), Or^{\omega})$. Consequently, possible types of variation patterns can be defined according to the types of the relationships between the base elements and between the overlay elements. For example, the pattern type known as "trend" involves relationships of linear ordering in the base, which arrange the overlay elements in a sequence, and metric relationships between the elements of the overlay such that the relationships along the sequence can be treated as changes and linked into series of similar changes.

While there are specific terms denoting particular types of patterns, such as trend, peak, plateau, fluctuation, cluster, alignment, etc., the vocabulary of the existing terms does not fully cover the variety of possible types of patterns. It may be unfeasible (and not very useful) to enumerate and label all possible types of patterns. It appears more reasonable to understand the roles of different kinds of relationships existing in the base and overlay of a distribution in forming patterns. This will allow one to anticipate the types of patterns that can potentially exist in a given data distribution without the need to know the terms denoting these pattern types. Our conceptual model introduced in Section 3 creates prerequisites for gaining such an understanding. Let us briefly describe the effects of different relationships.

Arrangements of overlay elements are formed by relationships existing in the base of a distribution $(Or^{B})$. The types of such relations include (but are not limited to) the following:

- *Identity*: overlay elements may be arranged in terms of having distinct or same (identical) holders.
- *Ordering*:
  - *Linear*: arrange overlay elements into a sequence.
  - *Cyclic* (e.g., temporal): arrange overlay elements into a succession of subsequences corresponding to consecutive cycles.

- *Distances*: create an arrangement of overlay elements where one element can be close to or far from another. The arrangement can be characterised in terms of the density of the overlay elements: uniform or variable, high or low, existence of clusters and empty regions, etc.

- *Neighbourhood* (*adjacency*): arrange the overlay elements into contiguous regions.

- *Direction* (e.g., spatial): arrange subsets of overlay elements into sequences similarly to linear ordering relationships.

Relationships existing in the domain of the overlay of a distribution ($Or^{\Omega}$) are involved in the variation of the overlay over the base. The expression $Var^{\Omega}(Ar^{\Omega}(Or^{B}), Or^{\Omega})$ signifies that the variation also involves arrangement relationships between the overlay elements which, in turn, are determined by the organisation of the base. Hence, the effects of the domain-pertinent relationships in the overlay need to be considered together with the possible arrangements of the overlay elements according to relationships in the base, as it is done in the following list:

- *Identity* or *equivalence*: create groups of identical or equivalent overlay elements, which can be characterised in terms of arrangement with respect to the base, e.g., contiguous, split into parts, or dispersed. Identical overlay elements may re-occur in a linear or cyclic arrangement, be aligned along some direction, have close positions in the base, etc.

- *Ordering*: may (or may not) be related to arrangement with respect to the base: increase or decrease of the element order along a sequence, regions with lower- or higher- order elements, etc.

- *Distances*: realise themselves as amounts of difference or change between positions in the base and thus create patterns of change: gradual, abrupt, moderate, etc.

- *Neighbourhood*: may or may not be preserved in an arrangement with respect to the base, i.e., neighbouring overlay elements may be close or distant in the arrangement.
- *Direction*: may or may not be same or similar along a sequence or within a region, may consistently change along a sequence, etc.

Data distributions where the base or the overlay domain have distance relationships between the elements may contain *outliers*. Distance relationships existing in the base may be responsible for outliers in the overlay arrangement. An *outlier in an arrangement* is an overlay element whose position in the base (i.e., the base element it is connected to) has a large distance to the positions of all other overlay elements. For example, a spatial outlier is an overlay element located in a spatial base far away from the bulk of the overlay elements. Distance relationships existing in the overlay may be responsible for outliers in the overlay variation. An *outlier in a variation* is an overlay element whose prototype in the overlay domain has a large distance to the prototypes of all other overlay elements. For example, an outlier in a distribution of values of a numeric attribute over any kind of base is an instance of a value that is much higher or much lower than all other values instantiated in the overlay.

A question arises: should an outlier be treated as a pattern type? In terms of our conceptual model, a pattern consists of relationships, not of elements. Accordingly, a particular outlying element of an overlay is not a pattern. However, its large distances to other elements, considered together with much smaller distances between those other elements, is a pattern. This type of pattern can be called *outlierness*, leaving the term "outlier" for applying to elements.

### *4.3. Patterns in common types of data*

The most common, frequently encountered types of data components include

- discrete entities, as the cross and nought symbols on the top left of Fig. 1 and apples on the top right;

- time, as the sequence of the days in Fig. 1, bottom, or sequentially ordered abstract steps, as the positions of words in a text or genes in a DNA molecule;

- space: a continuous or discrete set of locations, as the grid cells in Fig. 1, top left;

- attributes, as the colours on the upper right of Fig. 1 and the moon shape characteristics (width and disc side) in the lower part of Fig. 1.

These types have different organisations, i.e., different relationships between data elements. Sets of discrete entities, by default, have only identity relationships, i.e., same or distinct, between the elements. Time has ordering relationships and may also have metric distance relationships between the elements, i.e., how far in time one element is from another. Space has distance and/or neighbourhood (adjacency) relationships between the elements. Two- and three-dimensional spaces (and, more generally, multidimensional spaces) may also have direction relationships between elements. Attributes may have different organisations of the value sets, usually called scales of measurement: nominal, ordinal, interval, and ratio. These organisations differ in the presence or absence of ordering relationships and metric relationships of distance and ratio.

When a set of entities is a base of a distribution, it can create an arrangement of the overlay elements in terms of being connected to same or distinct entities. Such an arrangement may contain patterns of co-occurrence (i.e., some elements from the overlay base may repeatedly co-occur in connection with same entities) or exclusion (e.g., some elements never co-occur, or some element never occurs together with any other element).

When time is a base of the distribution, relationships of linear and cyclic ordering and temporal distance between time units create linear and cyclic arrangements of entities or attribute values corresponding (i.e., connected) to these time units. In a time-based arrangement of entities, there may be such patterns as high or low temporal frequency, temporal gaps (absence of entities for a time period), temporal clusters (groups of temporally close entities), and regular appearance of entities (i.e., with equal time intervals in between). For attribute values distributed over a temporal base, patterns of value variation with respect to the temporal arrangement are usually of interest. Patterns of temporal variation, such as trend, periodicity, peak, or plateau, result from the combination of the time-based arrangement of the attribute values and relationships of ordering and distance between the attribute values themselves. For attributes with qualitative (categorical) values, such as labels denoting types of entities, there may be such patterns as re-occurrence of particular value sequences or regular re-appearance of some values.

Space as a base creates arrangements of entities and attribute values according to relationships of spatial distance, neighbourhood, and direction between locations. In a space-based arrangement of entities, there may be such patterns as spatial clusters or regions of high and low density. Space-based arrangement of attribute values together with ordering and distance relationships between the values themselves can form such spatial patterns as "hot" and "cold" spots, i.e., regions of high and low attribute values, respectively. Spatial trend patterns involve relationships of spatial direction between spatial locations and relationships of ordering and, possibly, distance, between attribute values. Examples are increase or decrease from north to south or from centre to periphery.

What has not been discussed so far is the types of components in network, or graph data. In terms of our conceptual model, such data include two components: the set of all possible pairs of *nodes* and the set of *links* connected to some node pairs. The organisation of the set of node pairs in an undirected graph consists of the adjacency relationships: two pairs are adjacent if they have a common node. In a directed graph, there are relationships of adjacency and partial ordering: pairs (a,b) and (b,c) are adjacent, and the former precedes the latter in the order. The links can be considered as discrete entities or as values of a binary attribute specifying whether a pair is linked or not. In a weighted graph, the links with their weights can be treated as values of a numeric attribute; the value 0 can signify the absence of a link. If we consider the distribution of the links over the set of the node pairs, the latter will be the distribution base, and the links will make the overlay. The composition of the overlay is the set of actually existing links. The adjacency and ordering relationships in the base arrange the links into various structures, such as clusters, cliques, paths, stars, trees, and cycles. Such structures are usually considered as possible types of patterns in a graph. The concept of variation is relevant when there are some relationships between the links as such. In particular, when the links are weighted, there are metric relationships of the weight difference.

A graph as a whole can be considered as a base for other data components whose elements are connected to the nodes or links of the graph. In this case, the organisation of the base consists of the relationships represented by the links, and the elements of the other components are arranged by these relationships.

### *4.4. Patterns in selected examples of visual data analysis*

#### *4.4.1. Cluster and calendar view*

This example is based on a well-known paper by van Wijk and van Selow [40]. The data under analysis consist of two components: time, consisting of hourly time steps with the total length of one year, and numeric values of the power demand of a facility for each time step. The analyst wants to understand the variation of the power demand over time; hence, the time is the base of the distribution, and the overlay consists of the instances of the values of the attribute "power demand" recorded in each hour. The organisation of the base includes several kinds of ordering relationships: linear ordering and orderings in the daily, weekly, and seasonal cycles. These relationships create a complex arrangement of the attribute values, which is hardly possible to represent adequately in a single visual display. Therefore, the variation of the attribute values with respect to this arrangement is hard to grasp comprehensively using purely visual means.

The analyst tackles the problem by decomposing it. First, the analyst focuses on the segments of the overall arrangement corresponding to the daily cycles

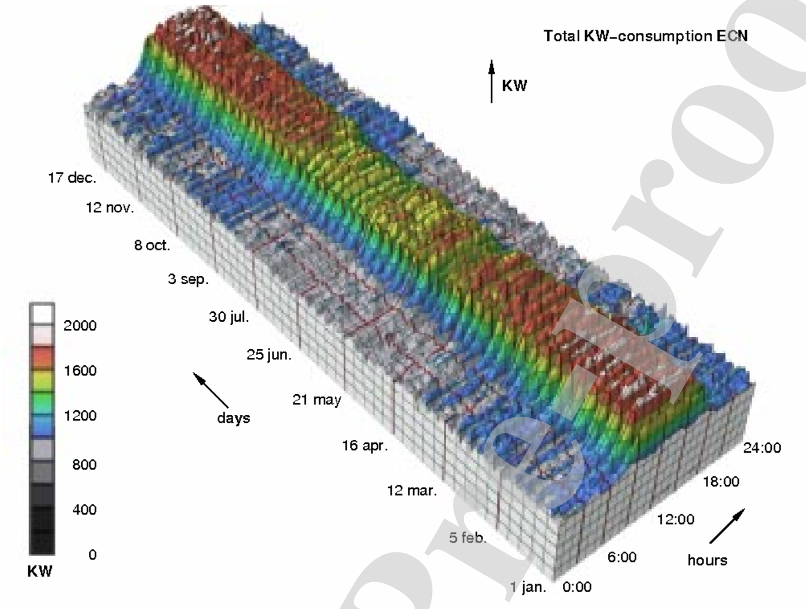


Figure 3: Variation of attribute values over a temporal base with linear and cyclic ordering relationships is hard to understand from a purely visual representation (source: [40]).

and the respective variation of the attribute values within each day. It can be expected that similar patterns of daily variation exist in different days, and this expectation is supported by the visual display in Fig. 3. The analyst uses a clustering technique to capture these similarities. The clustering puts days with similar sequences of hourly attribute values into groups. To see the common pattern of the daily variation in each group, the analyst aggregates the individual value sequences in each group into sequences of the hourly mean values. The resulting sequences are represented in a line chart, as shown in the right part of Fig. 4. The horizontal axis represents the linear ordering of the hourly intervals in a day. The attribute values are represented by vertical positions of points, and consecutive points are connected by lines. This technique represents the variation patterns by the shapes of the lines. It is possible to observe similarities and differences between the daily patterns corresponding to the clusters. All

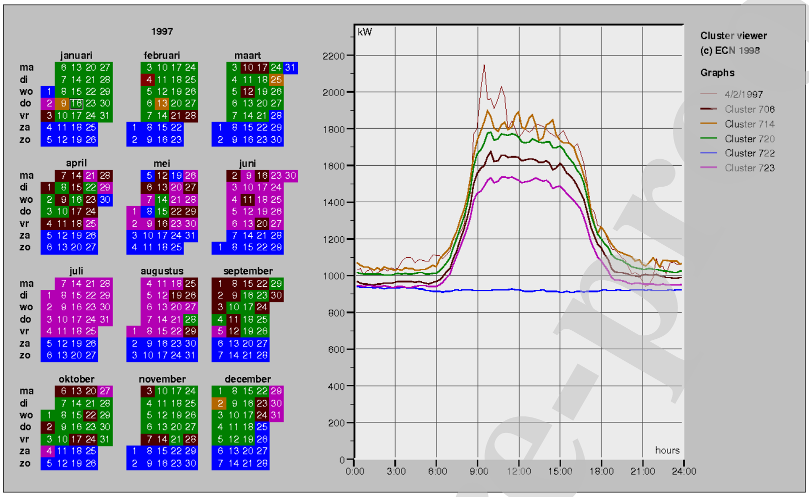


Figure 4: Right: patterns of daily variation are captured by means of clustering. Left: a calendar view shows the distribution of the daily patterns with respect to the weekly and seasonal cycles (source: [40]).

but one pattern can be holistically described, e.g., as low values in the night, steep increase in the morning, plateau during the day, and gradual decrease in the evening. The patterns mostly differ in the level of the daytime plateau. The remaining pattern can be characterised as constantly low values over the whole day. According to our theoretical model, these patterns are made by the temporal ordering and distance relationships between the elements (hours) of the base (time of the day) together with the quantitative difference relationships between the elements of the overlay (attribute values). The attribute values are arranged into sequences by the temporal relationships. The relationships between the values create the variation along the sequences.

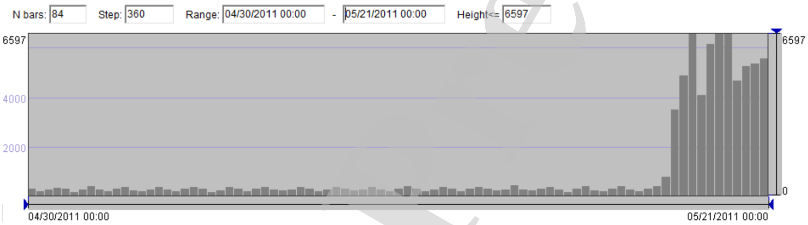
In the next step, the analyst treats the set of extracted daily patterns as a new component of the data. Each pattern is treated as a single entity. The analyst studies the distribution of the occurrences of these entities over the time. Now, the daily cycle relationships, which are are incorporated in the daily patterns, do not participate in arranging the overlay elements, i.e., the daily patterns. To consider the arrangement and variation of the daily patterns by the weekly and seasonal cyclic relationships, the analyst creates a calendar display shown on the left of Fig. 4. The elements of the base, i. e., the days of the year in this case, are visually represented by square marks organised according to the weeks and months. The pattern occurrences are represented by different colours. The display effectively enables perceptual unification of closely located marks painted in the same colour, and also unification of the marks arranged in the rows, the columns, and the monthly blocks. The analyst observes repeated weekly patterns, in which the Saturdays and Sundays are painted in blue and the other days have a different colour. A seasonal pattern is also noticeable, with the green colour occurring in colder months of the year and magenta in the summer. There are multiple disruptions of the seasonal green-magenta pattern by intrusions of other colours, mostly dark brown and orange. These disruptions correspond to the differences in the midday value levels between the patterns.

In this example, visual discovery of variation patterns is supported by displays in which the relationships between the base elements ($Or^B$) and the corresponding arrangements of the overlay elements ($Ar^{\Omega}(Or^B)$) are represented using one or two dimensions of the display space. The variations of the overlay elements ($Var^{\omega}(Ar^{\omega}(Or^{\beta}), Or^{\omega})$) are represented using either the remaining display dimension (as in the daily time series display on the right of Fig. 4) or, in Bertin's terms, a retinal visual variable, namely, the colours in the calendar display.

### *4.4.2. VAST Challenge 2011 (Mini Challenge 1)*

The challenge requires an investigation of the circumstances of an epidemic outbreak in a fictive city Vastopolis [38]. The data are geographically referenced microblog messages (further called tweets), some of which include keywords indicating disease symptoms, such as fever, chills, aches and pains, etc. The time span of the data is three weeks from April 30 to May 20, 2011. An analyst needs to find out when and where the outbreak started and how it developed.

Here, the data include the following components: set of tweets, set of people, set of keywords, time, and space. The analyst first selects a subset of the relevant tweets, i.e., those whose texts contain occurrences of keywords indicating the disease symptoms. To find out when the outbreak started, the analyst investigates the distribution of the relevant tweets over time, specifically, how the tweets are arranged along the time period under study. For this purpose, the analyst uses a time histogram (Fig. 5), in which the heights of chronologically ordered bars represent the numbers of the tweets that were posted in each day. As expected, the outbreak start is signified by a pattern of sharp increase of the tweet numbers. This happened in the last three days of the studied period.



N bars: 84 Step: 360 Range: 04/30/2011 00:00 - 05/21/2011 00:00 Height<= 6597

Figure 5: The histogram of the tweet posting times reveals a sharp increase of the frequency of tweets mentioning disease symptoms in the last three days of the time span of the data.

The analyst focuses on these three days, i.e., selects the data from this time interval for the further analysis. To analyse only the new disease cases, the analyst discards the secondary tweets posted by the same individuals after posting their first messages mentioning the disease symptoms. The analyst uses map displays to investigate the spatial distribution of the selected tweets in the three days of the outbreak (Fig. 6). The maps show the arrangements of the tweets according to the spatial distance and direction relationships between the locations in the spatial base. On the first day of the outbreak, a dense spatial cluster of tweets appeared in the city centre, with some extension in the eastern direction. On the second day, the cluster in the centre remains but does not extend to the east anymore. Additionally, two dense clusters, or a single cluster divided in two parts by a river, appeared in the south-western part of the city. On the third day, the south-western cluster almost vanished, as the spatial density of the tweets notably decreased, whereas the central cluster preserved.

The different behaviours of the clusters over time suggest that they may differ by other characteristics. To reveal the differences of the message contents between the clusters, the analyst studies the respective compositions of the keywords using word cloud displays, in which word frequencies are represented by font sizes (Fig. 7). Observing the differences between the frequency distribution patterns, the analyst concludes that there were two different kinds of illness: a flu-like disease in the centre and stomach disorders on the southwest. The latter appeared one day later than the former. However, the shapes and relative spatial arrangement of the clusters suggest that the two diseases might have a common origin somewhere at a motorway bridge crossing the river.

The analyst extracts the subset of messages posted in the vicinity of the bridge on the day before the outbreak start, examines the keyword composition, and finds indications of a truck crash, fire, and spilling of the truck cargo in the river. The analyst also looks at additional data concerning the weather and the river flow direction. The analyst concludes that the smoke from the fire contaminated the air, which was transmitted by the wind eastwards and caused the flu-like symptoms, whereas the spilled substance contaminated the water in the river and caused the stomach disorders downstream along the river.

The analyst compares the spatial distributions of the primary and secondary disease-related tweets on the third day (Fig. 8) and observes multiple compact dense clusters of secondary tweets. Using the background map, the analyst finds out that most of the clusters are located at hospitals. The analyst examines the keyword composition of the tweets in these clusters and finds out that the most frequent keywords from the tweets posted at the hospitals correspond to the most frequent keywords that occurred in the primary tweets posted in central-eastern area. The analyst selects the subset of people who came to hospitals, studies the trajectories made of their previous tweet locations, and determines that at least 95% of them had visited the central-eastern area after the truck accident and before coming to hospitals. This indicates that people with serious

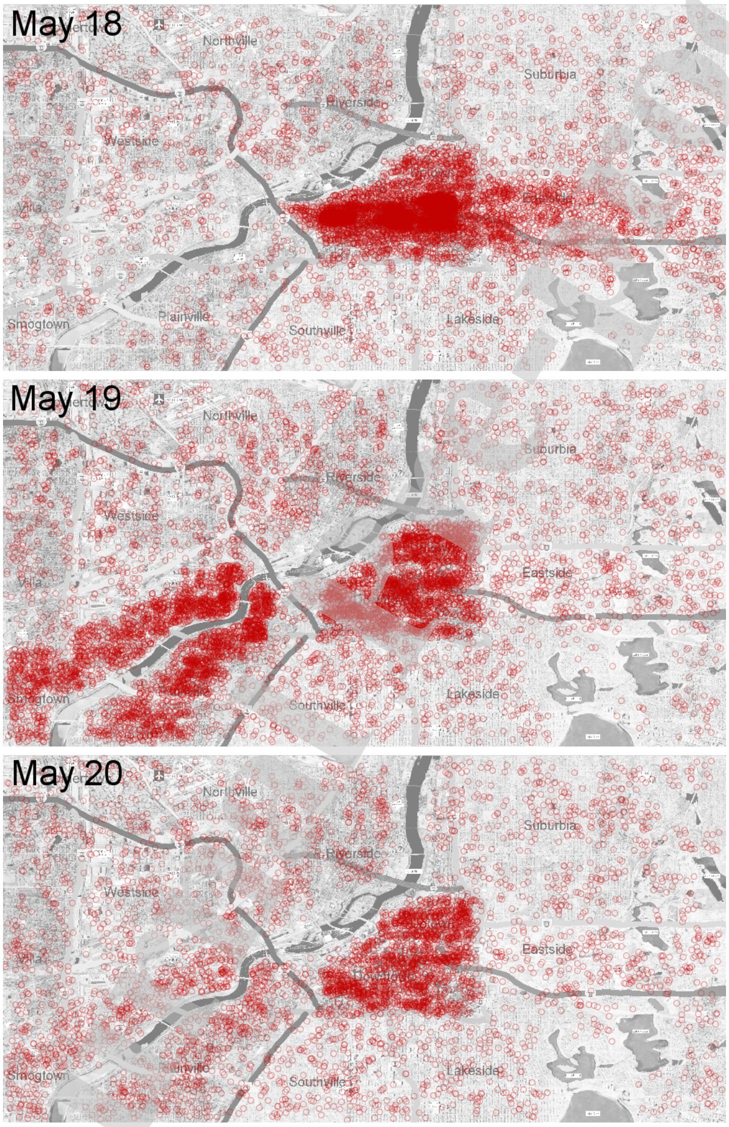
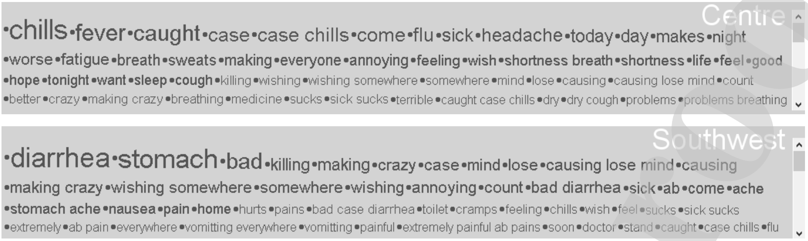


Figure 6: The spatial distribution of the primary outbreak-related tweets posted on three consecutive days.



Centre

•chills•fever•caught•case•case chills•come•flu•sick•headache•today•day•makes•night •worse•fatigue•breath•sweats•making•everyone•annoying•feeling•wish•shortness breath•shortness•life•feel•good •hope•tonight•want•sleep•cough•killing•wishing•wishing somewhere•somewhere•mind•lose•causing•causing lose mind•count •better•crazy•making crazy•breathing•medicine•sucks•sick sucks•terrible•caught case chills•dry•dry cough•problems•problems breathing

Southwest

•diarrhea•stomach•bad•killing•making•crazy•case•mind•lose•causing lose mind•causing •making crazy•wishing somewhere•somewhere•wishing•annoying•count•bad diarrhea•sick•ab•come•ache •stomach ache•nausea•pain•home•hurts•pains•bad case diarrhea•toilet•cramps•feeling•chills•wish•feel•sucks•sick sucks •extremely•ab pain•everywhere•vomitting everywhere•vomitting•painful•extremely painful ab pains•soon•doctor•stand•caught•case chills•flu

Figure 7: The text cloud displays represent the keyword compositions for the central cluster of tweets (top) and for the cluster on the southwest (bottom).

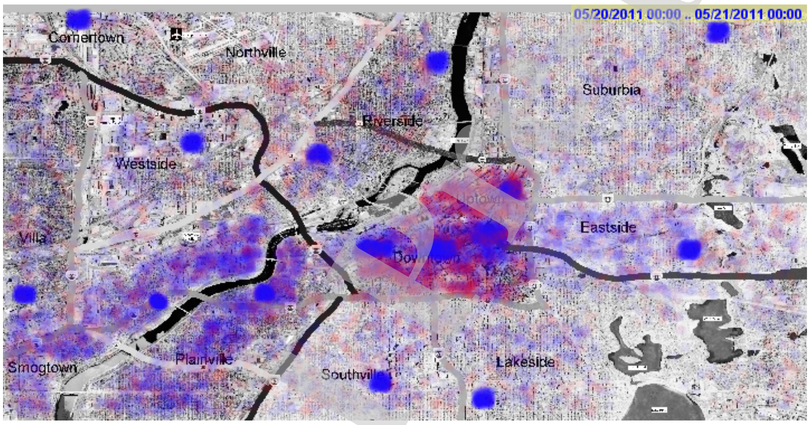


Figure 8: The spatial distributions of the primary and secondary outbreak-related tweets (red and blue dots, respectively) on the third day of the outbreak.

medical conditions, most probably, had been infected while being in the central-eastern area rather than anywhere else.

This example focuses on analysis of data distributions regarding overlay composition $C^{\Omega}(B)$ or arrangement $Ar^{\Omega}(Or^{B})$. The arrangement with respect 970 to time was visualised in a time histogram, where one display dimension was utilised to represent the temporal ordering and distance relationships. The arrangements with respect to space were visualised in maps, so that the spatial relationships existing in the physical space were represented by the spatial relationships between positions in the display space. The arrangement with respect to the composition of the space and time (i.e., the spatio-temporal distribution of the tweets in three days of the outbreak) was examined by decomposing the complex spatio-temporal base into slices along the temporal dimension, i.e., consecutive days, and considering the spatial arrangement in each slice.

### *4.5. Handling distributions with more than two data components*

The previous discussions mostly referred to distributions involving two data components. Relationships between more than two components can be handled in two complementary ways: *decomposition*, i.e., considering the relationships for each pair of components, and *integration*, i.e., considering the Cartesian product of two or more components as a single component. The latter approach is recommendable when the components have similar organisation, i.e. same types of relationships between the elements. Thus, it is quite common to integrate multiple numeric attributes into an abstract multidimensional space, which can be considered as a single component. It is also common to treat the physical space and time as an integrated domain called space-time continuum. The rationale for the recommendation is that the elements of the integrated set will be linked by relationships derived from the relationships existing in the original sets. Thus, distances between elements of a Cartesian product can be derived from distances within the original sets, and directions in the integrated set may be defined as compositions of ordering relationships from the original sets. If the organisation of the original sets are incompatible, i.e., have no relationship types in common, it is harder to derive meaningful relationships in the integrated set.

Section 4.4 contains examples of decomposition applied when the base of a distribution is composed of several components (namely, space and time in the epidemic outbreak example) or has a complex organisation, such as the nested cyclic orderings in the power demand example. In these cases, the distribution base was decomposed into slices, or segments. In the space-time case, the spatio-temporal base was decomposed into spatial slices corresponding to time steps (days). In the case with the temporal cycles, the base was divided into one-day segments. The analysts considered the parts of the distribution corresponding to the slices or segments and discovered patterns in these distributions. In the epidemic outbreak case, the analyst constructed a mental representation of the whole distribution from the discovered patterns by determining their relationships. In the power demand analysis, there were too many segments and respective patterns to deal with; therefore, the analyst used a computational method (clustering) to group the patterns by similarity and represent them by the average patterns of the groups.

## 5. Relationships between patterns

A data distribution may contain more than one pattern. Relationships between two or more patterns existing in the same distribution can be described in terms of the relationships between their bases and between the overlays. Obviously, the compositions of the bases and of the overlays, i.e., the sets of the elements they consist of, can be linked by general between-set relationships, such as inclusion and intersection. Besides, when there are specific organisation relationships, such as ordering and distances, between the elements within the base and/or within the overlay, these relationships can also link patterns. Thus, regarding the bases, patterns may be arranged in a particular order within the base, or lie at a certain distance from each other, or be adjacent, etc. For example, in Fig. 1, bottom, the trend pattern of the increase of the visible fraction of the moon is followed by the decreasing trend. In Fig. 6, there are spatial distance and direction relationships between the bases of the spatial clusters.

Regarding the overlays, patterns may be similar or different in terms of the composition, arrangement and/or variation of the overlay elements. Thus, in Fig. 7, we see different patterns of the word composition in two subsets of texts. As mentioned in Section 4.4.2, the word composition pattern in the central cluster of tweets was similar to the pattern in the secondary tweets concentrated at the hospitals. In Fig. 4, right, the patterns of daily variation consist of very similar changes and differ only in the highest attribute values.

Patterns involving sequential arrangement of the overlay element may also be opposite in terms of their variation, as, for example the increasing trend pattern in the morning and decreasing trend in the evening.

Relationships that exist between patterns unite simpler patterns into more complex patterns. Thus, the trend patterns of the changes of the moon shape make together a sequential pattern consisting of the increasing trend following by the decreasing trend. On a longer time period, there exists a pattern of periodic repetition of the same sequential pattern, i.e., it consists of multiple similar sequential patterns following one another. Similarly, the daily variation patterns in Fig. 4, right, are composed of shorter patterns of the night low values, morning sharp increase, daytime plateau, and evening gradual decrease. The calendar display on the left of Fig. 4 demonstrates how the daily patterns are organised into weekly and seasonal patterns.

These examples are consistent with the statement of Resnik [3] who said that the premier relationships among patterns are structural *similarity* and structural *containment*. Let us discuss these and other possible relationships in terms of our definitions.

### *5.1. Similarity*

**Definition 9**: Two or more objective patterns are **similar** if they can be represented by the same abstracted pattern.

For example, each curve in the plot in Fig. 4, right, is an abstracted pattern representing multiple objective patterns of daily variation of the power demand.

The concept of pattern similarity does not imply that patterns need to be in the same distribution, or in distributions with a common base or a common overlay, or involve occurrences of the same elements. It is only essential that patterns *involve the same relationships*. For example, the daily patterns of the variation of the power demand in Fig. 4 may be similar to the daily patterns of the variation of the amount of traffic in a city, or the number of employees present at their working places, etc.

In Fig. 8, there are multiple patterns of spatial arrangement that can be represented by the same abstracted pattern “dense compact spatial cluster”; hence, all these patterns are similar.

As we discussed, abstracted patterns representing the same objective pattern may differ in the degree of abstraction. This means that the similarity between objective patterns may be dependent on the degree of abstraction in representing them. Thus, all curves in Fig. 4, right, except the blue one, can be considered similar because they can be represented by a common abstracted pattern “low values in the night followed by sharp increase in the morning, then plateau with small fluctuations during the daytime, followed by a gradual decrease to the night low values”. However, if the rates of the morning increase and evening decrease are taken into account, the curves are different.

### *5.2. Containment*

**Definition 10**: An objective pattern $X$ **includes**, or **contains**, an objective pattern $Y$, denoted $Y \subset X$, when the base of $X$ includes the base of $Y$: $\beta(Y) \subset \beta(X)$. The pattern $Y$ is called a *sub-pattern* of $X$, and $X$ is a *super-pattern* of $Y$.

We have had already many examples of containment relationships. Thus, the pattern of the variation of the moon shape contains the trend patterns of the increase and decrease of the visible area of the moon. The daily variation patterns in Fig. 4, right, contain sub-patterns of uniformly low values, rapid growth, high plateau, and decrease. The weekly and seasonal patterns in Fig. 4, left, contain daily sub-patterns. The south-western spatial cluster in Fig. 6, centre, contains two sub-clusters separated by the river.

### *5.3. Repetition*

In a distribution, there may be two or more similar patterns with non-intersecting bases. In such a case, it can be said that some pattern repeatedly occurs in the distribution, and this repetition itself is a pattern. More specifically:

**Definition 11**: A **repetition pattern** is a super-pattern containing two or more *similar* sub-patterns with non-overlapping bases together with relationships existing between the pattern bases.

For example, a daily variation pattern in Fig. 4, right, contains two instances of a pattern of uniformly low values, one at the beginning and one at the end of the day. On the left of Fig. 4, we see patterns of multiple repetitions of several daily patterns, as well as multiple repetitions of weekly patterns composed of five consecutive repetitions of one daily pattern followed by two repetitions of another daily pattern. Figure 8 exhibits a spatial repetition pattern containing multiple dense clusters of outbreak-related tweets located around hospitals.

A repetition pattern can be called *regular* if the organisation of the distribution base includes distance relationships between the elements, and the distances between the bases of similar patterns are (approximately) equal. If the base organisation also includes ordering relationships, a pattern of regular repetition can be called *periodic*. Thus, the distribution in Fig. 4, left, contains a pattern of periodic repetition of the weekend pattern of daily variation, which is represented by blue colour. Besides, there are sub-patterns with periodic repetition of particular weekly patterns, sometimes with small disruptions. The variation of the moon shapes considered on a longer time period than shown in Fig. 1 is also periodic.

### *5.4. Cross-overlay relationships*

Not only patterns that exist in the same distribution can be linked by relationships but also patterns existing in two or more distributions with a common base. More specifically, relationships can exist between the bases of the patterns, and such relationships may be quite important. They may hint at correlations or even causal relationships between phenomena or events. Potentially related patterns may have the same or overlapping bases, or there may be a particular relationship (such as a temporal lag) between the pattern bases.

**Definition 12**: **Cross-overlay relationships** between patterns existing in distributions of distinct overlays over a common base consist of relationships between the bases of the patterns.

For example, the spatial pattern of the repeated dense clusters of outbreak-related tweets visible in Fig. 8 is related to the pattern of the spatial distribution of the hospitals in Vastopolis, as the base of each cluster includes the position of one hospital. This relationship indicates that many infected people came to hospitals. Other examples of cross-overlay relationships are those between the south-western dense cluster of tweets and the river position (the spatial base of the cluster overlaps with the spatial base of the river), between the central dense cluster and the wind direction at the time of cluster emergence (the temporal base of the cluster coincides with the temporal base of the pattern of the western wind), and between the spatial position and time of the track crash event and the spatial positions and times of both clusters. The latter example demonstrates spatial and temporal shifts between the pattern bases.

## 6. Use of patterns in further data analysis

One of the benefits of having a clear definition of a pattern is the possibility to define in a systematic way various operations that can be applied to patterns in the course of data analysis. Consequently, designers of systems for data analysis can implement system functions and interaction techniques supporting these operations.

In accord with the model building view [29], discovered patterns are integrated in an overall model of the analysis subject, and this model is used for description, prediction, and/or decision making. Here we do not consider these final uses of analysis outcomes but discuss how discovered patterns can be utilised in the further data analysis. We begin with considering specific examples and then use our conceptual model to define in a systematic way the possible actions that can be applied to patterns or their constituents in the process of analysis.

### *6.1. Specific examples of pattern use*

In section 4.4.1 and Section 4.4.2 we described two examples of visually analysing data. Let us look how analysts in these examples used the patterns they had discovered.

**E1**: In Section 4.4.1, several repeating patterns of daily variation of the power demand discovered by means of cluster analysis (Fig. 4, right) were considered as elements of an overlay set distributed over a base consisting of the days of the year. The further analysis was applied to the distribution of these patterns over this base (Fig. 4, left). This is the most obvious example of application of further analysis steps to discovered patterns.

Section 4.4.2 provides the following examples. **E2**: After seeing the pattern of high increase of the number of disease-related tweets in the last three days, the analyst focused the further analysis on these three days (Fig. 5). **E3**: The analyst selected the tweets that formed the increase pattern in time and considered their distribution in space (Fig. 6). **E4**: After detecting two dense spatial clusters of posted tweets (Fig. 6, centre), the analyst considered and compared the keyword compositions of the respective messages (Fig. 7). **E5**: Observing particular spatial relationships between the two clusters, the analyst came to the hypothesis of a common reason and origin of both and inferred the likely place and time of the event that might cause the appearance of these clusters. **E6**: The analyst compared the shapes of the clusters and their positions in space and time (Fig. 6, top and centre) with the spatial position and flow direction of the river and with the wind direction at the time of the cluster emergence and drew conclusions concerning the disease transmitting mechanisms. **E7**: Analysing the spatial distribution of the secondary outbreak-related tweets posted on the last day, the analyst noticed a pattern of repetition of compact dense clusters and found out that these clusters were located around hospitals (Fig. 8). **E8**: Comparing the keyword compositions of the tweets posted at the hospitals, in the central-eastern area, and in the south-western area, the analyst observed similar frequency patterns in the two former compositions. **E9**: The analyst selected the subset of people who came to the hospitals, studied the spatial relationships of their previous tweets to the central-eastern and south-western areas, and thereby ascertained that most people had previously visited the central-eastern area and, most probably, had been infected while being there.

These examples demonstrate the main purpose of pattern discovery in visual analytics: *patterns are involved in analytical reasoning; analysts use them to make hypotheses and draw conclusions.* This main use of discovered patterns is supported by interactively performed analytical operations, such as selection, extraction of connected elements from other data components, aggregation, and unified representation. Our theoretical model allows us to define the set of possible analysis operations on patterns in a systematic manner.

### *6.2. General analysis operations on patterns*

According to our model, a pattern has its base $\beta$ and overlay $\omega$, which are subsets of the base $B$ and overlay $\Omega$ of the overall data distribution. The internal contents of a pattern is connections and relationships between the elements of its base and its overlay (Definition 6). The base and overlay elements may also have other connections and relationships, external with respect to the pattern. Analytical operations can be applied to the internal pattern contents or exploit the external connections or relationships.

Operations on **internal contents of individual patterns**:

- *Characterise* pattern contents: derive (in particular, computationally) synoptic characteristics of a pattern from the elements of its base and overlay and their relationships, e.g., the number of elements in a spatial cluster and their spatial density.
- *Aggregate* a pattern: represent a pattern as a single element of data (as in E1).
- *Refine* a pattern: divide $\beta$ or $\omega$ into subsets (e.g., the outbreak-related tweets into primary and secondary), investigate and characterise the parts of the overall data distribution including the elements of each subset and the connected elements of the other component.

Operations on **comparing contents of several patterns**:

- *Compare* patterns in terms of relationships they include; e.g., compare daily variation patterns in E1, compare word frequency patterns in E4 and E8.
- *Group* patterns by similarity of their contents; e.g., create clusters of similar daily patterns in E1.
- Represent similar patterns by a *common abstracted pattern* and treat each pattern as an instance of this abstracted pattern (E1).

Operations using **relationships of $\beta$ and $\omega$ to external elements of $B$ and** $\Omega$:

- *Determine relationships of a pattern to the rest of the distribution*, e.g., determine the relative time of the tweet number increase pattern in E2 and the amount of the increase with respect to the average number.
- *Determine relationships between patterns* in the same distribution, e.g., between spatial clusters of tweets (E5).
- *Unite patterns into compound patterns* (super-patterns), e.g., unite the compact dense clusters of tweets posted around the hospitals (Fig. 8) and use them all together, as in E8 and E9.

Operations using **connections of base or overlay elements to elements of other components**:

- *Extract elements of other components* connected to the elements of the pattern base or overlay, e.g., extract spatial locations of the tweets in E3, words of the messages in E4 and E8, people who posted the tweets in E9.
- *Characterise a pattern using elements of other components*, e.g., characterise spatio-temporal clusters of tweets in terms of keyword occurrences (E4, E8).

- *Determine cross-overlay relationships* of patterns in a distribution of a component $\Omega_1$ over a base $B$ to patterns or elements of the distribution of another component $\Omega_2$ over the same base $B$ (E6, E7).

This section emphasises that discovery of distribution patterns is a part of an analytical workflow, in which the patterns that have been discovered are used in various ways in reasoning and further analysis. This emphasis is specific to visual analytics, whereas data mining, statistics, and other disciplines developing techniques for data analysis are primarily concerned with pattern discovery and, possibly, interpretation but not with the further use.

## 7. Discussion of model implications

### *7.1. Summary of the model*

The definitions and statements we have formulated earlier can be briefly summarised as follows:

- A *data pattern* is a combination of relationships between connected elements of two or more data components. Elements of one of the components make the pattern *base*, the remaining elements make the *overlay*. The relationships between the overlay elements are considered in connection to the base and to relationships existing between the base elements. A pattern does not include its base or overlay elements; it only includes the system of relationships between the elements.

- An *objectively existing data pattern* can be represented and treated as a single object. Any such representation is called an *abstracted pattern*. *Similarity* of objective patterns means the possibility to represent them by the same abstracted pattern.

- Patterns may be linked by containment or intersection relationships between the sets of their base and/or overlay elements as well as by relationships made from elementary relationships between the base or overlay elements. Patterns linked by relationships form *composite patterns*.

- Once discovered, patterns can be utilised in the further data analysis through applying interactive analytical operations to their internal contents and external relationships and connections.

Let us now discuss the meaning of this model for the visual analytics science and practice.

### *7.2. Need for pattern discovery*

Understanding relationships among data components is one of major general tasks for which visual analytics techniques are applied. A visual analytics process often aims at building a model (particularly, a mental model in the analyst's mind) of some subject of analysis, and the model needs to represent relationships between components (aspects) of the subject in a generalised way [29]. The requirement of the generality means that multiple connections between individual elements of data need to be unified.

As we have explained in this paper, unification of multiple elementary connections is possible owing to relationships that exist between elements within data components. These relationships unite multiple elements and elementary connections into structures that can be considered and represented holistically. Such structures are usually called *patterns*. Hence, general relationships between components of data and/or analysis subject can be understood and modelled by discovering patterns in data distributions. Therefore, pattern discovery can be regarded as a fundamental operation in visual analytics processes.

There are two approaches to pattern discovery: computational and visual. Computational pattern discovery is done by specially designed algorithms. This requires precise specification of patterns to seek, i.e., what relationships must exist between elements. Besides, parameter tuning is often needed, such as setting the minimal number of elements in a pattern, maximal distance or difference between elements, minimal frequency, etc. An algorithm will find patterns matching the given specification and nothing else. Hence, in the context of the general task of gaining an overall understanding of the relationships between data components, pattern discovery algorithms do not do the full job, as they will not find potentially relevant patterns beyond the specifications received. Still, when particular types of patterns are expected to exist in the data, it makes sense to employ algorithms designed to detect patterns of these types. The possible pattern types can be predicted based on the types of the relationships existing within the data components, as discussed in Section 4.2.

Visual pattern discovery relies on the human capability to see patterns in visual representations of information. The use of this capability does not require an exact specification of what to look for, and a human observer can detect patterns of various types. However, the visual representation must fulfil the following **requirements**:

- Since patterns are formed by relationships between data elements, the visualisation must faithfully show the existing relationships.
- The visualisation must not provoke seeing non-existent relationships, to preclude generation of false patterns.
- Since patterns need to be considered and represented holistically, the visualisation should facilitate perceptual unification of multiple elements.

These requirements logically follow from the conceptual model introduced in the previous section. At the same time, they are consistent with the established principles of the visualisation introduced by Bertin and further developed by other researchers.

### *7.3. Principle of correspondence*

The first two requirements to visual representation can be seen as two sides of a single principle of correspondence: relationships that can be perceived by a human observer from a visual display must correspond to relationships actually existing in data. This statement is consistent with the Bertin's formulation of the principle of the correspondence between the organisation level of a data component and the perceptual properties of the visual variable that should be used for representing this component [31]. Mackinlay [41] referred to this principle using the term "expressiveness" (of a visual variable). Based on our conceptual model, the principle of correspondence can be explained by the necessity to make relationships involved in objectively existing patterns perceivable by a human so that the relationships between values of a visual variable can be intuitively translated into relationships between the data elements represented.

Talking about organisation of a data component, Bertin considered only ordering and metric (quantitative) relationships. We have discussed in Section 4.2 how other types of relationships, such as equivalence, spatial direction, neighbourhood, cyclic ordering, can also be important. Besides, there may be application-specific relationships, e.g., hierarchical relationships or links in a network. Hence, the Bertin's concept of organisation level is insufficient for describing the variety of possible organisations. The existing assortment of visual variables is also insufficient for representing all types of relationships that may exist within data components. Thus, there is no visual variable that could represent a cyclic or a hierarchical organisation. Such organisations are usually represented using other means, such as particular layouts of visual marks. For example, a cyclic organisation can be represented by a radial, spiral, or matrix layout. In node-link diagrams, relationships are represented by special linear marks connecting nodes. The treemap technique [42] uses a nested layout for representing hierarchical relationships.

Bertin considered various layouts (called "impositions"), including networks and maps, in separation from the concept of set organisation, whereas a layout is no less a means to represent relationships within a set than a visual variable. We propose to treat these and other possible means of representing relationships between elements equally and thus to state the **fundamental principle of visualisation** in the following way:

*Analysis-relevant relationships between data elements need to be represented by appropriate means of visual expression, including visual variables, layout of visual marks, special marks, spacing, etc. These means of visual expression must support the perception of existing relationships and preclude the perception* *of non-existing relationships.*

Spacing between display components, such as bars in a bar chart, is often used when the visual variable 'position', which is perceived as continuous, represents a discrete set. This is an example of an approach to precluding perception of non-existent relationships.

The principle we have formulated can be called the *principle of correspondence of visualisation means to relationships existing in data*. This is not a new principle; although it was not stated in this way until recently (see [43]), visualisation designers have been always following it by using empirically established conventions of choosing particular visual means for representing different kinds of data. The proposed explicit formulation clearly states: what visualisation designers need to care about primarily is the *relationships* existing in data. By matching the possible types of relationships, including those discussed in Section 4.2, to the visual means capable to convey them, it is possible to transform the tacit conventions into explicit rules of visualisation design.

There are other theoretical models that consider the requirement of correspondence between data and visualisation from different perspectives. Kindlmann and Scheidegger [39] care about the correspondence between the so-called "mathematical structure of the underlying data" (i.e., data types and organisation) and the "mathematical structure in the perception of visualizations". They formulate their three principles stating that the visualisation must be invariant to the internal representation of the data and that changes in the data must result in noticeable, meaningful, and unambiguous changes of the display. One of the principles is called "The Principle of Correspondence", but, unlike ours, it refers to data changes rather than relationships within the data. Demiralp et al. [44] propose a model that treats visualisation as a data embedding that must preserve structures existing in the data. The model focuses on relationships between data items that can be represented as distances. The idea is that distances perceived from the visualisation must correspond to the actual distances between data items. Wattenberg and Fisher [45] focus on the kinds of relationships that organise data into groups and hierarchies. They propose a formal model that can describe the organisation of an arbitrary grey-scale image as, supposedly, would be perceived by an observer. A visualisation designer can compare the structure reconstructed by the model with the actual data structure and thus check if the image conveys the data structure correctly. Unlike those works, our model explicitly acknowledges the pattern-forming role of different kinds of relationships between data items and explains the fundamental principle of correspondence between data and visualisation by the need to correctly convey data patterns to human observers and analysts.

### *7.4. Principle of unification*

According to our model, pattern discovery involves unification of multiple elements and abstraction, that is, integrated representation of these elements as a single object. Consequently, visual displays of data should not only correctly represent objectively existing data patterns but also support perceptual abstraction from multiple elements and elementary relationships to holistic representations. This corresponds to the Bertin's concepts of the overall and intermediate reading levels as opposite to the elementary level involving perception of individual elements and relationships [31]. Bertin also introduced the concept of image as "the meaningful visual form, perceptible in the minimum instant of vision" [31, p. 11]. A single image providing answers to questions of all three levels allows us to perceive patterns as units. Visualisations with more images require integration across images, which may hinder holistic perception.

Hence, in designing visual representations for data analysis, it is essential to support integrated perception of multiple relationship instances. For example, in a line chart, multiple points are connected by line segments; as a result, a large number of ordering and distance relationships are integrated into a single line that is perceived as a unit. In plots or maps where elements are represented by dots, multiple neighbouring dots can be perceptually integrated into shapes according to the Gestalt law of proximity [33]. This capacity of the human's perception is also utilised in projection displays where distances in the projection space represent degrees of similarity, semantic relatedness, or other kind of relationships whose strength can be expressed numerically [46, 47, 48].

Abstractive perception can be promoted by smoothing, e.g., using kernel density estimation techniques [49], which, however, hide the original elements and relationships. Tufte advocated creation of displays supporting both micro- and macro-readings [50], such that multiple small visual marks can be perceived all together. Bae and Watson [51] study the use of five cues stimulating visual grouping, namely, proximity, colour similarity, common region, connectivity, and alignment, separately and in combinations. They assess the strengths of the different cues and find that complex structures can be more effectively communicated by combining two or more grouping cues.

The development of visual analytics science and technology would benefit from a systematic survey of the existing approaches suitable for supporting abstractive perception, and it would also be appropriate to evaluate these techniques empirically.

### *7.5. Directions for empirical research*

As we have mentioned, the need to support abstractive perception of patterns calls for empirical research on how different techniques can promote such kind of perception.

Since relationships play the key role in forming data patterns, the existing means of visual representation require empirical evaluation of their capabilities to enable perception of various types of relationships. The empirical studies that were conducted so far mostly referred to the ability of display users to perceive values rather than relationships. Hence, there is a need in further studies focusing primarily on relationships.

Knowing the types of relationships involved in the organisation of data components, it is possible to predict what kinds of patterns may exist, irrespective of the existence of specific terms denoting these kinds of patterns. This possibility can be used for testing the capabilities of a particular visualisation to convey correctly and effectively the kinds of patterns that can exist in data with components of given types. For this purpose, one can construct an artificial dataset with these data types that includes this or that kind of pattern as the "ground truth", and check if users can efficiently spot the incorporated patterns.

### *7.6. Practical utilisation of the theoretical model*

An analyst who wants to discover patterns in a distribution can use the model to

- understand which aspects of a distribution are relevant to analysis goals: composition, arrangement, or variation (Section 3);
- understand what kinds of relationships between elements are involved in these relevant aspects and need to be taken into account (Section 4.2);
- find appropriate means for representing these relationships;
- decompose a distribution over a complex base with several kinds of relationships into a combination of distributions with simpler bases;
- understand what relationships can exist between patterns and determine these relationships (Section 5);
- build an analytical workflow involving appropriate operations on patterns (Section 6).

Apart from the possible use by data analysts, the model can provide an appropriate basis for practice-oriented teaching of visual analytics. It can also be utilised in designing visual analytics systems providing guidance to users [30, 52, 53]. Intelligent guidance that is not limited to instructing users about system functions may help users in pattern discovery [30, Section 5.5.1], e.g., by informing users about pattern types that can exist in their data and about visual or computational methods that can be used for finding patterns of these types [30, Fig.1]. An intelligent guide can also help users externalise patterns they have discovered, i.e., transform mental images of these patterns into explicit representations. To provide these kinds of user support, the guiding system needs to have a knowledge base enabling prediction of possible pattern types depending on the structure and properties of user's data [30, Section 6]. Our theoretical model can serve as a foundation for such a knowledge base.

## 8. Conclusion

In developing our model, we have built on ideas from systems science [54] and general mathematics [3, 4], and we also generalised and systematised our vast practical experiences from developing visual analytics solutions for various kinds of domains, data, and problems. The model does not include a taxonomy of pattern types, which could hardly be exhaustively itemised. It also does not explicitly refer to data types, which can be defined in multiple ways (e.g., in databases, programming languages, etc.), but refers instead to fundamental properties of data components, particularly, types of relationships between elements.

The model gives a working definition of a pattern in a data distribution, which has been so far a rather vague and not practically utilisable notion in visualisation and visual analytics. To make this definition, we have introduced a system of supporting definitions. By drawing implications from the system of definitions, we have theoretically explained the rationale of some of the existing empirically established principles of visualisation, which may be helpful in teaching these principles. We have outlined how the proposed theoretical model can be used in data analysis practices, but, of course, its practical utility requires extensive testing.

The model can enlighten designers of visual analytics methods and systems concerning possible approaches to supporting pattern discovery. The main idea is to respect and make use of relationships existing in data domains and to find either computational methods extracting combinations of relationships or visual methods allowing human analysts to observe such combinations and perceive them holistically. In this respect, the model suggests a need in empirical studies on perception of visual displays that would specifically focus on perception of relationships rather than judgement of absolute values. Such studies need to evaluate the following: (1) how easy is for a user to see particular relationships between overlay elements and between their positions in the base; (2) whether or not the user may see non-existing relationships; (3) how well items linked by the relationships "stick together" in the user's eyes [33].

Another merit of the given definitions is that they enabled us to describe systematically the analytical operations that can be applied to discovered patterns in the processes of data analysis and analytical reasoning. The task of supporting such processes has primary importance for visual analytics research and design of visual analytics systems. We have defined the range of possible analytical actions that can be applied to patterns or involve patterns. This can inform researchers focusing on supporting analytical processes and designers of systems intended to support such processes.

As a direction for further theoretical research, we see a need in considering in more detail complex bases composed of heterogeneous components, such as *space* × *time*, *entities* × *time*, *entities* × *space* × *time*, etc. The organisations of such bases are very complicated. It would be appropriate to consider what kinds of objective patterns are possible for overlays with different properties.

This discussion gives us a ground to believe that our work makes a valuable contribution to the visual analytics research and can inform and motivate further theoretical researches.

## 9. Acknowledgment

The authors are grateful to Min Chen, Daniel Keim and Jarke van Wijk for intensive discussions and constructive comments on earlier versions of this work.

This research was supported by Fraunhofer Center for Machine Learning within the Fraunhofer Cluster for Cognitive Internet Technologies, by DFG within Priority Programme 1894 (SPP VGI), by EU in project SoBigData++, by SESAR in projects TAPAS and SIMBAD, and by Austrian Science Fund (FWF) project KnowVA (grant P31419-N31).

## References

[1] N. Andrienko, G. Andrienko, G. Fuchs, A. Slingsby, C. Turkay, S. Wrobel, Visual Analytics for Data Scientists, Springer, 2020.

[2] K. Devlin, Mathematics: The Science of Patterns: The Search for Order in Life, Mind and the Universe, Henry Holt and Company, 1996.

[3] M. D. Resnik, Mathematics as a science of patterns, Clarendon Press; Oxford University Press Oxford : Oxford ; New York, 1997.

[4] G. Oliveri, Mathematics. a science of patterns?, Synthese 112 (3) (1997) 379–402.

[5] P. Bruce, A. Bruce, Practical Statistics for Data Scientists: 50 Essential Concepts, O'Reilly Media, 2017.

[6] C. Heumann, M. Schomaker, S. Shalabh, Introduction to Statistics and Data Analysis With Exercises, Solutions and Applications in R, 2017.

[7] W. K. Härdle, S. Klinke, B. Rönz, Introduction to Statistics, 2015. `doi:10.1007/978-3-319-17704-5`.

[8] C. Forbes, M. Evans, N. Hastings, B. J. Peacock, Statistical distributions; 4th ed., John Wiley & Sons, Ltd, New York, NY, 2010.

[9] K. Krishnamoorthy, Handbook of Statistical Distributions with Applications, Statistics: A Series of Textbooks and Monographs, CRC Press, 2006. URL `https://books.google.de/books?id=FEE8D1tRl30C`

[10] Y. H. Chou, Spatial pattern and spatial autocorrelation, in: A. U. Frank, W. Kuhn (Eds.), Spatial Information Theory A Theoretical Basis for GIS, Springer Berlin Heidelberg, Berlin, Heidelberg, 1995, pp. 365–376.

[11] A. Getis, J. H. P. Paelinck, An analytical description of spatial patterns, L Éspace géographique 33 (1) (2004) 61–69. `doi:10.3917/eg.033.0061`.

[12] M. Rosenberg, C. Anderson, Spatial Pattern Analysis, Oxford University Press, New York, NY, 2016. doi:10.1093/OBO/97801998300600144.

[13] M. Souris, Spatial Distribution Analysis, John Wiley & Sons, Ltd, 2019, Ch. 5, pp. 109–175. doi:10.1002/9781119528203.ch5.

[14] M. Borregaard, D. Hendrichsen, G. Nachman, Spatial distribution patterns, Oxford: Elsevier, 2009, pp. 3304–3310.

[15] C. E. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (1948) 379–423.

[16] T. M. Cover, J. A. Thomas, Elements of Information Theory, 2nd Edition, John Wiley & Sons, 2006.

[17] F. E. Ruiz, P. S. Pérez, B. Bonev, Information Theory in Computer Vision and Pattern Recognition, Springer, 2009.

[18] M. Feixas, A. Bardera, J. Rigau, Q. Xu, M. Sbert, Information theory tools for image processing, in: Synthesis Lectures on Computer Graphics and Animation, Vol. 6, 2014, pp. 1–164.

[19] M. Chen, S. Walton, K. Berger, J. Thiyagalingam, B. Duffy, H. Fang, C. Holloway, A. E. Trefethen, Visual multiplexing, Computer Graphics Forum 33 (3) (2014) 241–250.

[20] N. Kijmongkolchai, A. Abdul-Rahman, M. Chen, Empirically measuring soft knowledge in visualization, Computer Graphics Forum 36 (3) (2017) 73–85.

[21] M. Chen, A. Golan, What may visualization processes optimize?, IEEE Transactions on Visualization and Computer Graphics 22 (12) (2016) 2619–2632.

[22] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[23] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[24] J. Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[25] C. C. Aggarwal, Data Mining: The Textbook, Springer Publishing Company, Incorporated, 2015.

[26] W. Klösgen, J. M. Zytkow (Eds.), Handbook of Data Mining and Knowledge Discovery, Oxford University Press, Inc., New York, NY, USA, 2002.

[27] T. Munzner, Visualization Analysis and Design, A.K. Peters visualization series, A K Peters, 2014.
URL `http://www.cs.ubc.ca/%7Etmm/vadbook/`

[28] N. Andrienko, G. Andrienko, Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. `doi:10.1007/3-540-31190-4`.

[29] N. Andrienko, T. Lammarsch, G. Andrienko, G. Fuchs, D. Keim, S. Miksch, A. Rind, Viewing visual analytics as model building, Computer Graphics Forum 37 (6) (2018) 275–299.

[30] C. Collins, N. Andrienko, T. Schreck, J. Yang, J. Choo, U. Engelke, A. Jena, T. Dwyer, Guidance in the human—machine analytics process, Visual Informatics 2 (3) (2018) 166 – 180. `doi:https://doi.org/10.1016/j.visinf.2018.09.003`.

[31] J. Bertin, Semiology of Graphics, University of Wisconsin Press, 1983.

[32] C. Ware, Information Visualization: Perception for Design, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[33] W. Metzger, Laws of Seeing, MIT Press, Cambridge, MA, USA, 2006.

[34] J. Wagemans, J. H. Elder, M. Kubovy, S. E. Palmer, M. A. Peterson, M. Singh, R. von der Heydt, A century of gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization., Psychological bulletin 138 6 (2012) 1172–217.

[35] F. Kerlinger, H. Lee, Foundations of Behavioral Research, PSY 200 (300) Quantitative Methods in Psychology Series, Harcourt College Publishers, 2000.

[36] Merriam-Webster Online, Merriam-Webster Online Dictionary (2009).
URL `http://www.merriam-webster.com`

[37] A. S. Hornby, Oxford Advanced Learner's Dictionary of Current English, 6th Edition, Oxford University Press, 2000.

[38] G. Grinstein, M. Whiting, K. Liggett, D. Nebesh, IEEE VAST Challenge, http://hcil.cs.umd.edu/localphp/hcil/vast11/ (2011).

[39] G. Kindlmann, C. Scheidegger, An algebraic process for visualization design, IEEE Transactions on Visualization and Computer Graphics 20 (12) (2014) 2181–2190. `doi:10.1109/TVCG.2014.2346325`.

[40] J. J. van Wijk, E. R. van Selow, Cluster and calendar based visualization of time series data, in: Proc. IEEE Symposium on Information Visualization (InfoVis), 1999, pp. 4–9.

[41] J. Mackinlay, Automating the design of graphical presentations of relational information, ACM Transactions on Graphics 5 (2) (1986) 110–141. `doi:10.1145/22949.22950`.

[42] B. Johnson, B. Shneiderman, Tree-maps: A space-filling approach to the visualization of hierarchical information structures, in: Proceedings of the 2Nd Conference on Visualization '91, VIS '91, IEEE Computer Society Press, 1991, pp. 284–291.

[43] B. Karer, I. Scheler, H. Hagen, H. Leitte, Conceptgraph: A formal model for interpretation and reasoning during visual analysis, Computer Graphics Forum 39 (6) (2020) 5–18. doi:https://doi.org/10.1111/cgf.13899.

[44] C. Demiralp, C. Scheidegger, G. Kindlmann, D. Laidlaw, J. Heer, Visual embedding: A model for visualization, IEEE Computer Graphics and Applications 34 (1) (2014) 10–15. doi:10.1109/MCG.2014.18.

[45] M. Wattenberg, D. Fisher, Analyzing perceptual organization in information graphics, Information Visualization 3 (2) (2004) 123–133. doi:10.1057/palgrave.ivs.9500070.

[46] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow, Visualizing the non-visual: Spatial analysis and interaction with information from text documents, in: Proceedings of the 1995 IEEE Symposium on Information Visualization, INFOVIS '95, IEEE Computer Society, Washington, DC, USA, 1995, pp. 51–58.

[47] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, P. Dragicevic, Time curves: Folding time to visualize patterns of temporal evolution in data, IEEE Transactions on Visualization and Computer Graphics 22 (1) (2016) 559–568. doi:10.1109/TVCG.2015.2467851.

[48] S. van den Elzen, D. Holten, J. Blaas, J. J. van Wijk, Reducing snapshots to points: A visual analytics approach to dynamic network exploration, IEEE Transactions on Visualization and Computer Graphics 22 (1) (2016) 1–10. doi:10.1109/TVCG.2015.2468078.

[49] N. Willems, H. Van De Wetering, J. J. Van Wijk, Visualization of vessel movements, Computer Graphics Forum 28 (3) (2009) 959–966.

[50] E. Tufte, Envisioning Information, Graphics Press, Cheshire, CT, USA, 1990.

[51] J. Bae, B. Watson, Reinforcing visual grouping cues to communicate complex informational structure, IEEE Transactions on Visualization and Computer Graphics 20 (2014) 1973–1982.

[52] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, C. Tominski, Characterizing guidance in visual analytics, IEEE Transactions on Visualization and Computer Graphics 23 (1) (2017) 111–120.

[53] D. Ceneda, N. Andrienko, G. Andrienko, T. Gschwandtner, S. Miksch, N. Piccolotto, T. Schreck, M. Streit, J. Suschnigg, C. Tominski, Guide me in analysis: A framework for guidance designers, Computer Graphics Forum 39 (6) (2020) 269–288. `doi:https://doi.org/10.1111/cgf.14017`.

[54] G. J. Klir, D. Elias, Architecture of systems problem solving, Da Capo Press, Incorporated, 2002.

Author Contributions Section

All authors have equally contributed to the paper.

declaration

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: