1 2 9 0

UNIVERSIDADE D
COIMBRA

Barbara do Nascimento Ramos

DATA GLYPHS

A CASE STUDY

Dissertation in the context of the Master in Design and Multimedia, advised by Professor Catarina Maçãs and Evgheni Polisciuc and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

September of 2022

Faculty of Sciences and Technology
Department of Informatics Engineering

DATA GLYPHS:
a case study

Barbara do Nascimento Ramos

Dissertation in the context of the Master in Design and Multimedia,
advised by Professor Catarina Maçãs and Evgheni Polisciuc and presented to the
Department of Informatics Engineering of the Faculty of Sciences
and Technology of the University of Coimbra.

September of 2022

1 2 9 0

UNIVERSIDADE Đ
COIMBRA

Abstract

We live in a society characterized by better decisions made with the help of data. Data is everywhere, our society shapes itself through it and, with the years passing by, it is becoming greater in dimension. A lot of this data now available is multi-variate in nature. The bigger the volume and complexity are, the harder tasks to detect, classify and measure characteristics and relations within data. When there is no certainty of the characteristics of a dataset due to its size, or there are no clear targets, visualization can be used to simplify the information.

Regarding multivariate data, glyph-based visualization is one of the possible techniques commonly adopted in data science. This dissertation aims to study data glyphs, analyzing the definition, the different classifications, and the design process around them. In the end, and applying the previously studied concepts, three visual explorations are produced as glyph design alternatives to represent a dataset related to audiological tests carried out in the population of Portugal. These glyphs were evaluated through semi-structured tests with users and then an analysis of the performance result of each of the glyphs is presented, evaluating them in terms of learning and memorization.

Keywords

Information visualization, data glyphs, glyph-based visualization, Multivariate data visualization;

Acknowledgment

Primeiramente, faço um agradecimento especial aos meus orientadores Catarina Maçãs e Evgheni Polisciuc pela paciência e atenção, me auxiliando e guiando até o último minuto, literalmente.

A minha mãe, que mesmo não compreendendo em alguns momentos, me apoiou incondicionalmente. Obrigada por, mesmo com o Atlântico entre nós, estar sempre presente, foi minha primeira inspiração na vida e sem você nada disso seria possível!

A minha amiga Bea, que foi uma amizade fiel durante esses dois anos. Dentro de dias solitários em um país com vários rostos desconhecidos, sua amizade foi um carinho acolhedor que me motivou a chegar até aqui. Agradeço também às minhas amigas Irina, Maria e Diva que tornaram os dias mais leves mesmo nas fases mais turbulentas. Um agradecimento especial também vai a Valéria, que com seu jeito inclusivo foi minha primeira amizade dentro do Curso, obrigada por estar aberta a me conhecer e deixar meus dias mais felizes.

Um agradecimento exclusivo deve ser feito ao meu amigo Zé, foi meu porto seguro nessa experiência intensa. Obrigada por estar sempre disponível, seja para os momentos de risadas e comemorações ou os momentos de tristeza e saudade.

Por fim, agradeço imensamente ao Francesco, que chegou no meio dessa experiência e deixou-a muito mais agradável. Obrigada por estar ao meu lado todos os dias, me apoiando até nos momentos mais difíceis.

This work was funded by project A4A: Audiology for All (CENTRO-01-0247-FEDER-047083) financed by the Operational Program for Competitiveness and Internationalisation of PORTUGAL 2020 through the European Regional Development Fund and by the FCT - Foundation for Science and Technology, I.P./MCTES through national funds (PIDDAC), within the scope of CISUC R&D Unit - UIDB/00326/2020 or project code UIDP/00326/2020.

Contents

1. Introduction

> Having all the world's information at our fingertips doesn't make communication easier: it makes it harder.
>
> (Laszlo Bock in Knaflic, 2015, p.IX)

We live in an era characterized by the promise of better decisions made through data (Munzner, 2014, p. 2). Data is everywhere and all sectors receive more and more pressure to work with it. When talking about data-driven decision-making in a context that has well-defined questions, it is possible to use purely computational techniques from fields such as statistics and machine learning. With such techniques, jobs that were previously done by humans can now be completely automated with a computer-based solution. However, it is not always possible to have well-defined questions for data and problems in hand. There are thousands of possible questions and approaches one can take when working with data, and often people don't know which questions or approaches are the right ones.

According to Ward (2002), with the increasing volume and complexity of data to which we have immediate access, the ability to effectively analyze the data has become increasingly strained. Much of the data available today are multivariate in nature, whether it be from spreadsheets, multispectral sensors, or complex computations. Each row or data entry may have anywhere from two to several thousand entries, and the larger the number and size of the entries, the harder it is to detect, classify, and measure features and relations of interest. In such cases, a fully computerized process may not be the best. Different factors can come into play. The computational models may not be prepared to cope with uncertainty in data or the heterogeneous nature of data may even make it impossible to create such models.

In scenarios of uncertainty and when dealing with experimental and exploratory data analysis, it might be better to add people in the process. This is because the human perceptual system is incredibly powerful in finding patterns and discrepancies that statistical techniques may not notice (Manguire, 2015). In these cases, the use of visualizations is powerful, expanding human visual perception capabilities and serving as a tool for safer and more effective data analysis. A good visualization can assist users in several tasks, including observing trends and outliers; creating connections between data and experience, knowledge and intuition; visual generation and evaluation of hypotheses; monitoring the correction and performance of computer models, and effective communication with others. However, for the aforementioned goals to be achieved, the visualization must be effective, that is, it has to show the data correctly and clearly. Today, with a multitude of digital tools, everyone can make visualization and, for this reason, Knaflic (2015) states that bad graphics and visualizations are everywhere. The author adds to the issue that historically visualizations were reserved for scientists or non-scientists with highly technical positions, while today, with the democratization of technology and the possibility of using digital tools to create graphics, without a clear path, the best intentions and efforts, can lead to bad results.

Knaflic (2015) states that effective data visualization can make the difference between success and failure when it comes to communicating study findings, raising money for nonprofit associations, presenting to the board, or simply conveying a point of view to the audience. According to the author, visualizations need to be solid, have a story to tell, and show data clearly and objectively. Additionally, the visualizations must also have a precisely made design to effectively support the user with an array of tasks, which, as mentioned earlier, is one of the principle objectives of visualizations.

When it comes to visualization techniques, there are enumerable approaches that depend not only on the volume and type of data being represented but also on the final objective of the visualization. Multivariate data visualization is a class of visualization techniques dealing with the presentation of many data attributes at once. In the context of visualizing multivariate data, one of the possible techniques is data glyphs.

Data glyphs are composite graphical objects that use their visual and geometric attributes to encode multidimensional data (Borgo et al., 2013). For example, arrows, mostly used in visualizing vector fields, are glyphs whose visual variables can be used to encode attributes, other than directions. Another more complex, yet efficient, example is the Star Glyph, composed of some lines equally spaced and radially arranged, whose lengths encode the magnitude of a given data. These are just a few examples of countless variations of designs for data glyphs. There are multiple types of glyphs with different designs and concepts, exploring the entire spectrum, from the most pictorial representations to the most abstract ones.

This project intends to explore different techniques to design glyphs, study the impact of the produced glyph and its components in terms of visual perception and understanding of the data and evaluate its efficiency through user tests. This involves a series of steps such as analysis and exploration of the data, including, for example, the use of classification techniques; design of the glyphs, exploring different visual alternatives, and; evaluation of glyphs through tests with users carried out in person and with the aid of a crowdsourcing model. Finally, the performance analysis of the different developed glyphs and improvement requirements is delivered.

1.1. Scope and Objectives

The project proposed in this document is carried out in the context of the National project with the title A4A: Audiology for All, financed by the Fundo Europeu de Desenvolvimento Regional (FEDER), through the Programa Operacional Regional do Centro (Centro 2020). The purpose of the A4A project consists of the development of innovative services and features allowing the optimization of Audiology services and associated health care, provided by Evollu and its partners, services, and mobile applications to support hearing and innovative algorithms that improve speech perception. /speech and make them available in service format.

This dissertation has as its main objective the application of information visualization techniques to transform a set of multivariable data related to Portuguese hearing health into data glyphs, that allow obtaining a general visual aspect of the problem and thus characterizing the Portuguese population. This project also aims to achieve the following specific targets:

- Develop a theoretical basis and state-of-the-art research regarding glyphs and visualizations;
- Develop various forms of data glyphs (e.g., exact, using existing variables, non-exact, using symbols such as Chernoff faces or icons, or figurative images or pictograms);
- Study the visual ranking qualities of elements applied to a glyph;
- Study the efficiency of the elements used from a series of parameters;
- Develop a method of validating glyphs with users and apply it.

1.2. Dissertation Summary

The first chapter of this document, which is the current one, presents an introduction to the project.

The following chapter presents instead a summary and the analysis of major studies carried out on information visualization and glyphs. This chapter is divided into three parts: the first deals with theoretical studies on visualization and the main references in the area; the second focuses on glyphs and all their characteristics, including design; the third and final part presents the concept of empirical research and the different possible approaches for validating visualizations.

Chapter three presents the methodology that will be used in the present work, describing each of the steps according to the process developed by Maguire (2015) based on Munzner's (2014) model.

The fourth chapter presents the entire practical project report. Divided between the steps of the methodological process, this chapter begins with a presentation on the domain related to the data and ends with a description of the glyph validation process.

Finally, the project results are discussed in the concluding chapter, as well as some improvement requirements presented from the validation results presentations, thus presenting possibilities for future work.

2. Literature Review

This section is intended to make a complete review of all topics relevant to the development of the project, forming a solid theoretical basis, capable of guiding decision-making throughout the dissertation.

This chapter is divided into three parts. First, a brief introduction is made that aims to present a conceptualization of what information visualization is and why to use it.

In the second part, the main theorists regarding visualization principles are presented. The objective is to present a series of principles at the design level that will guide the production of glyphs. In this section, Bertin's concepts of visual channels (1968), design principles, and Tufte's ratio-ink concept will be presented, in addition to presenting some considerations of the principles of Gestalt psychology.

Finally, the third part is dedicated to discussing about glyphs and their productions. The objective is to present the main research regarding glyphs and to present a glyph production methodology pointing out its different moments. The definition of what glyphs are at the production level, data abstraction, translation comment for visual channels, rendering, and arrangement of glyphs is covered. In addition, some considerations are made about the validation of the view.

2.1. Information Visualization

Tamara Munzer makes the following cut in terms of information visualization.

> Computer-based visualization systems provide visual representations of datasets designed to help people to perform tasks more effectively. Visualization is appropriate when there is a need to augment human capabilities rather than replacing people with computational methods of decision-making. The design space of possible visualization languages is huge, and includes visual considerations. Visualization is full of trade-offs, and most possibilities in the design space are ineffective for a particular task, so validating the effectiveness of a design is both necessary and difficult. Visualization designers must take into account three very different types of resource limitations: computer, human, and display limitations. Visualization usage can be analyzed in terms of why the user needs it.
>
> (Munzner, 2014. p. 1)

In this definition, some important points can be highlighted: the presence of people, the presence of the computer, the need to perform tasks, and the limitations.

Munzner (2014) states that human perception is powerful and completely necessary within contexts of uncertainty. When there is no developed computational model, or when there are still no clear objectives and methods for the development of the project, a fully computerized system does not present effective results. On the other hand, human perception, supported by good visualizations, has the power to find patterns and trends even when there are still no well-defined questions.

Computing enables the construction of visualizations that allow users to explore and present a large set of data, something that would be completely infeasible to do by hand (Munzner, 2014, p.4). It is possible to manually create visual representations of data, either completely hand-made on paper with drawing materials or using vector illustration softwares where each item is individually organized and colored. Such visualizations often use small datasets with hundreds of items and can take hours or days to accomplish. However, most databases that need visualizations are much larger, ranging from thousands to millions of data. Also, most datasets change dynamically over time. For these reasons, implementing a visualization with the aid of computer systems, beyond saving a lot of time and effort for the designer, allows real-time visualization of the data.

A visualization serves to increase human capabilities. A human, in contrast to how a machine works, has the perception to visualize patterns where models are not yet available. Examples of tasks they can perform are observing trends and outliers, creating connections between data and experience, intuition, etc (Manguire, 2015) That is, visualizations always aim to perform one or more tasks. Munzner (2014) states that a visualization that serves one task may be ill-suited for another, using the same dataset. Hence, the tasks should be defined not only according to the dataset (e.g., the types of data), but also to the user's needs. Visualizations can be used in the most varied scenarios and reasons. Munzner (2014), lists some reasons why people do visualization:

- **To discover something new.** Visualization can be used within an exploratory analysis, when we are not sure what to expect from the data and want to have a greater view of it. A visualization can be very illuminating in this context, as patterns, or deviations from them, may not be noticeable when data is presented only in numbers. An example of this is a dataset designed by Francis Anscombe, in 1973, to illustrate how another dataset with identical descriptive statistics can have very different structures: these structures are obvious when the dataset is shown graphically, as seen in figure 1.

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

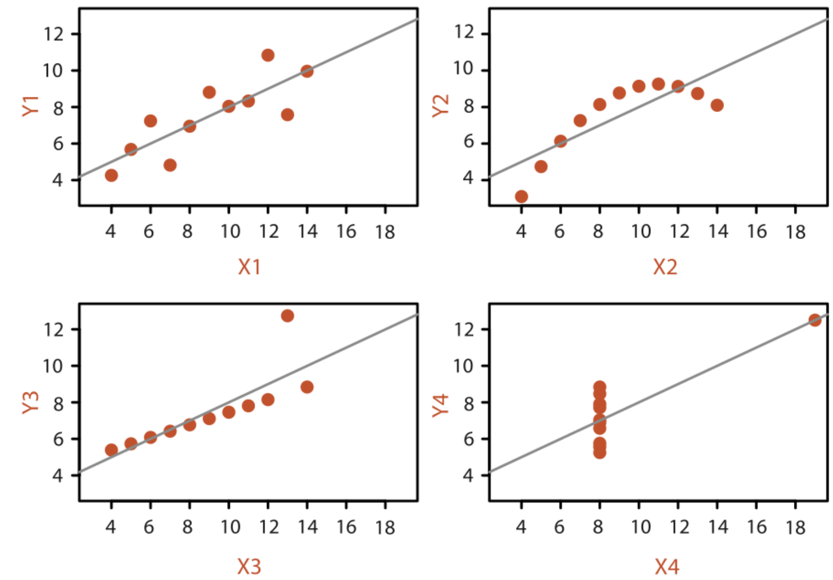


Figure 1: Anscombe's Quartet.
Source: Munzner, 2014, p.8.

- **To synthesize something known.** Visualizations can be used to present something that is already known in order to better communicate the information to others. Munzner (2014, p.3)states that this type of visualization is used every day by the most diverse newspapers and magazines (i.e., The Guardian and Publico).

- **To automate something.** For example, when an automated computational model has not yet been developed, at least three visualization opportunities can be mapped (Munzner, 2014, pp.3-4): (i) one before starting any development, to have a clear understanding of the analysis requirements; (ii) another one that is not aimed at end users, but at designers and developers, where the objective is to monitor the system; and (iii) an end-user-focused visualization can be made to legitimize the computer-based system, thereby increasing user confidence.

In any of the situations presented before, visualization helps to compare data faster and with less effort. As shown in Anscombe's Quartet of figure 1, even with a small amount of numbers it is difficult to do a complete analysis, not because it is not possible to read the numbers, but because it requires more effort to read each value, memorize, and then compare them. The next subchapters will discuss how to develop a proper visualization so that an effective communication of the data occurs and thus helps the comparison and analysis of the same.

2.2. Foundations of the Visualization

In information visualization, signs are basic geometric elements depicting items or links. Typically, a signal represents more than its graphical aspects: in fact, it represents anything but the mark itself. The process of turning some information into a mark is called visual coding. The reading of signs is directly linked to the viewer's interpretation and therefore generates different readings depending on the context and culture presented — semiology is the science that structures these signs and their different impacts. Studies indicate that the effectiveness of a coding channel, that is, how a certain data will be represented, depends on its type. For example, the channels and visual variations that convey magnitude information are a good match for ordered data type and those that convey differentiating information are a good match for categorical data (Munzer, 2014, p.95).

According to Munzer (2014) and Bertin (1967), the world of design space for graphic construction and visualizations can be described by two aspects: graphic elements called marks, and visual channels that control their appearance. Even complex visual encodings can be broken down into components that can be analyzed in terms of their branding and channel structure.

Learning to reason about marks and channels is important to develop effective and efficient visualization. This chapter presents studies that suggest graphic principles and recommendations based on the analysis of signs and the cultures that use them, not only in terms of information visualization, but in any spectrum that contains graphic marks.

For organization purposes, this chapter is divided between the principles and studies of the main authors on the subject in a chronological way. Starting with Bertin's graphic semiology, where the concept of visual channels is approached, proceeding with Tufte's studies about the principles of graphic integrity and graphic excellence, and concluding with the gestalt principles in visual perception.

2.2.1. Semiology of Graphics

Jacques Bertin is a French cartographer and theorist. In 1967, he wrote the book "Graphique Semiology", which is considered the first and most comprehensive attempt to provide a theoretical basis for the use of graphic elements with regard to developments in graphics and maps.

Bentin, through examples in more than 1000 maps and diagrams, synthesized the principles of graphic communication using standard logical rules applied to writing and topography. The author classified a series of elements that he called retinal variables, six fundamental visual variables, which are at the base of the construction of any graphic sign: size, color, texture, orientation, volume, and shape. Bertin also presented a set of rules for the development of a graphic system (Borgo et al, 2013) where he classified which types of retinal variables are most suitable for each type of data: numeric, ordinal, and categorical.

In this subchapter, some of the concepts developed by Bertin will be presented, as well as each type of retinal variable, describing in more detail their possibilities, challenges of application, and limitations. At the end of this study, it is intended to have a comprehensive theory on the best ways to apply visual channels during glyph development.

Analysis of Information

> As a "language" for the eye, graphics benefit from the ubiquitous properties of visual perception. As a monosemic system, it forms the rational part of the world of images
>
> (Bertin, 1967, p.6)

According to Bertin (1967, p.2), eyes and ears have distinct systems of perception where each of the information systems can be classified according to the possibility of interpreting its meaning. There are three systems: monosemic, polysemic, or pansemic.

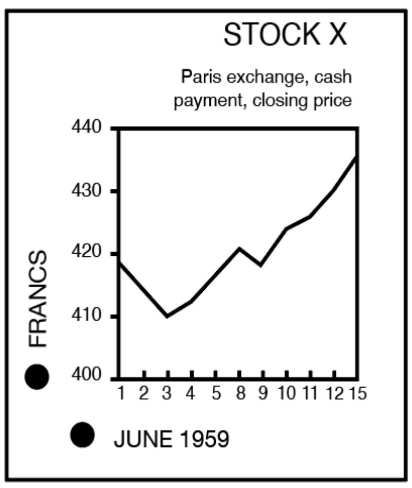
Monosemic perception systems, both visual and auditive, are those that do not allow for different interpretations—they are objective. Graphics are considered monosemic systems in visual perception, since each sign, that is, each visual element presented in the graphic, precedes the observation of the collection of signs. This means that graphics by nature do not give rise to various interpretations: the legends serve as an anchor that presents to the spectator the function of each presented element, thus avoiding any subjectivity or communication failure.

With regard to visual perception systems, when meaning becomes subjective, it is classified as a polysemic system. Bertin (1967, p.2) presents figurative images as an example, which despite clearly representing an object, leaves room for interpretations. Based on this, abstract paintings are the extreme form of polysemy, becoming pansemic.

Furthermore, when categorizing the sensory variables of the different perception systems (eyes and ears), Bertin states that the ears have only two variables at their disposal: sound and time. So the perception of sound is very linear, while the eyes have three variables: two related to the variations in the plane (e.g., the spatial arrangement of elements in dimension x and dimension y) and a third that represents the variations of the mark, that is, the visual characteristics that a certain element has in addition to its position. As a spatial system, graphs instantly communicate the relationship between these three variables.

Graphic representations are the transcription of "information" known through the intermediation of any system of signs to a graphic system. Bertin (1967, p.6) states that for any transcription, a separation of content and form is necessary, emphasizing that the content are the elements regardless of the sign system — the information or data to be translated — and the recipient is represented by the means available in a given system and the laws that govern its use — the representational form. Bertin also makes a categorization of the content elements, where he states that the "information" is a set of correspondences of variables with an invariant, graphically traducing the content of a thought.

Information can be divided into two concepts, the invariant and the component. The invariant is the central notion, common to all pertinent correspondences, and the component is the variational concept. In addition, there are elements that consist of different and identifiable parts of the components. To clarify these concepts, Bertin presents an example. In Figure 2 it is possible to Observe the following information: on June 3, 1959, share X on the Paris stock exchange was quoted at 410 francs; on 8 June it was quoted at 420 francs. In this example, "stock X" is the invariant data, and the information "time" and "amount of francs" are elements that vary. By placing this information on a graph, as shown in figure 2, it is possible to observe the relationship of the invariant data with the title, and of the variables with the spatial components, x and y-axis. This relationship between the variants and the arrangement of the element is named by Bertin as the planar dimension.



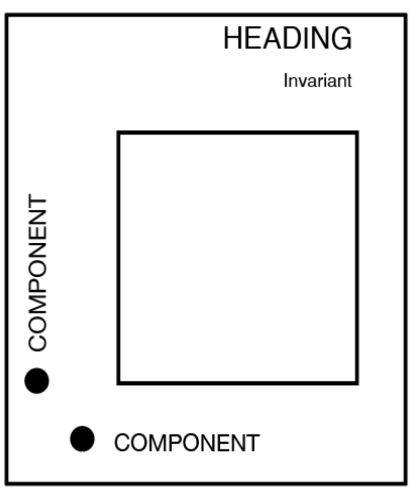


Figure 2 and 3: Invariant, Component and element relationship. **Source:** Bertin, 1967, p. 16 and 24

Based on the structure of the graphs, Berlin describes three steps to analyze the information. The first step is to determine the number of components. When analyzing the example presented in figure 2, there is an invariant, the stock X, and two components, the time and the number of francs. In the second stage, it is necessary to determine the length of the components, which means, describing the number of elements that can be identified in a given component. The third and last stage is the identification of the relationship between the components and the elements. This can be defined in three levels of organization:

- **Qualitative:** they are a class channel that can facilitate the grouping of all elements of a variable despite differing values. Can be divided into associated and differentiated groups, which means, it always involves two perceptual approaches: (i) this is similar to that (similarity/association); or (ii) this is different from that (differentiation).

- **Ordered:** components consist of elements with a natural sequence that can be ranked. (i.e., this is more than that or less than that).

- **Quantitative:** when the elements are countable units, with constant numerical ratios between one another.

According to Bertin, the designer needs to analyze the quantitative, ordered, or differential nature of the data to be transmitted and adopt the corresponding visual variable, with the choice of signs being conditioned by the limited properties of the visual variables.

There are numerous perceptual variables based on the human senses, the eye is the intermediary for most of them. Mapping all the possibilities of each visual variation in a broad sense is an impossible activity, so Bertin's research is limited to a scenario with a flat sheet or white paper of standard size and under normal lighting. Within these limits, the graphic system is considered to have eight variables. A sign can vary in relation to the two dimensions of the plane, x and y. It can further vary in size, value, texture, color, orientation, and shape. Within the plane, this mark can be represented by three types, named by Bertin as Implementation or representation class, a point (a position without an area), a line (a linear position without an area), or an area.

The Plane

The plane is the pillar of graphic representation, as mentioned before. The plane is homogeneous and has two dimensions, x and y, offering a series of visual possibilities that must be explored. A tag is a primitive form of representation, a visual blot used to represent the elements of a graph. Bertin states that this mark can represent a point, a line, or an area in the plane—elementary figures of the geometric plane that are termed "implementations":

- **The point:** points are a type of mark that has no dimension and only represents a position in the space or on the plane without an associated length or area. The meaning of the point depends only on the position associated with the plane, where its size or other rendering characteristics should not interfere with its meaning. A point can vary in position but it will never represent a line or an area.

- **The line:** has only one dimension, representing what has length as in geometry the connection between two points. Bertin points out that essentially a line is the edge of an area. Thus, a line means a phenomenon with a measurable length but associated area. The meaning of a line is independent of its width or other visual characteristics applied to its rendering.

- **The area:** it has two dimensions, that is, it is the only shape that can have a measurable fill size. The space of this mark has the meaning of its size, so the total space occupied by the mark is converted by the symbol.

In the organization of the plane using the two dimensions, named by Bertin as an imposition, there are sublevels of organization that the author classified into four groups: (i) diagram, where the correspondence on the plane is among all the elements of one component and all the elements of another component; (ii) network, the correspondence is between all the elements of the same component; (iii) graphic map, the correspondence is among all the elements of the same geographic component, inscribed on the plane according to the observed distribution; and (iv) symbolism, the correspondence on the plane can be established between a single element and the reader (road signs, various codes based on shape, industrial color codes, etc.). The correspondence is exterior to the graphic image.

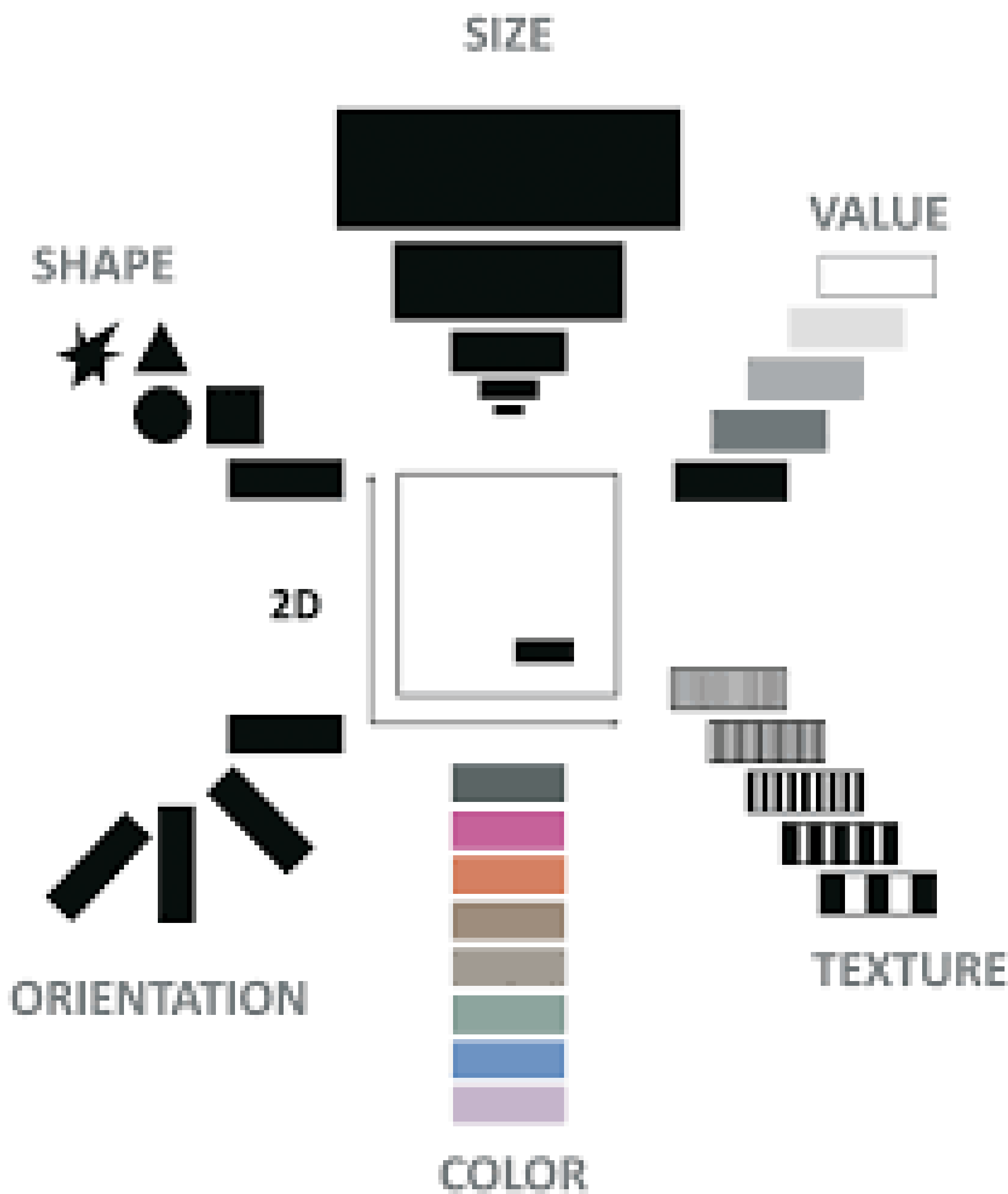
Regarding graphic composition, Bertin (1967) refers that a graph has invariant and variant components. As far as variant elements are concerned, they are organized into three general categories, variations in plane and brand. Also, a mark can be a point, a line, or it can have an area. However, when it comes to the brand, the author has classified a series of variations that can be used for encoding information, named retinal variations. Below, each of these variations and the recommendations for use described by Bertin (1967) will be presented.

Retinal Variables

As discussed earlier, above the plane variables there are several visual variables. Experimental psychology defines deep perception as the result of a series of factors, some of which are: the decrease in size of a known object; a decrease in values of a known contrast; a reduction in the known texture of an object; a decrease in the color saturation of known objects; orientation and form determinations (Bertin, 1967, p.60).

The author states that all these variations, called retinal variables, can be used by the graphic designer when he or she intends to add a component to the representation, thus varying not only in its position about the plane but also in its visual characteristics. Next, each of the retinal variables classified by Bertin (1967) will be presented.

Figure 4: The visual variables.
Source: Bertin, 1967, p.43.



Size

It refers to the variation of height and width of a visual element present in the image, that is, it is the variation of the dimension of the graphic mark. Thus, in the implementation, this property can only be applied in the meaning type “point” or “line”, in a two-dimension representation the area cannot vary in size. Bertin points out that in the implementation, for a qualitative and orderly perception, the number is almost unlimited, being limited just to the display and the eyes. However, in selective perception, size variation is very limited.

Value

According to Bertin, the variation in value is a continual progression of how the eye perceives grays coming from black to white, referring to tonal variation—the total amounts of black and white in one. The given surface will be called the value. It is important to emphasize that this variation, between black and white, happens independently of the represented color - blue, red, or green - the higher the value, the stronger the black (or white, as in computers). Concerning the level of organization of the implementation, a value range is ordered.

Texture

To Bertin (1967), a texture variation is a series of photographic reductions of a pattern of marks without causing the value of the area being distributed to vary. At a given value, the texture is the number of marks contained in a certain area, starting with the null texture value, where the marks are numerous but being small they are imperceptible. As the value increases, a “zoom” is applied to the texture and the brands get bigger and consequently less numerous. In the implementation, textures can be applied to areas along their entire length. In points and lines it can be applied but in a very small number of steps.

Color

It is observed that there is a difference between the color (matrix) and the variation of value, which was mentioned earlier. The author claims that color is the noticeable difference that can be achieved in areas that have the same value. To build a series of values it is necessary to add either a little white to have lighter values or more black to darker values.

The value that has no addition of black or white is called "pure hue" or "saturated hue". It is important to understand that the pure hue value is not in the same tonal range for all colors, in other words, it is not a constant value but varies with the color

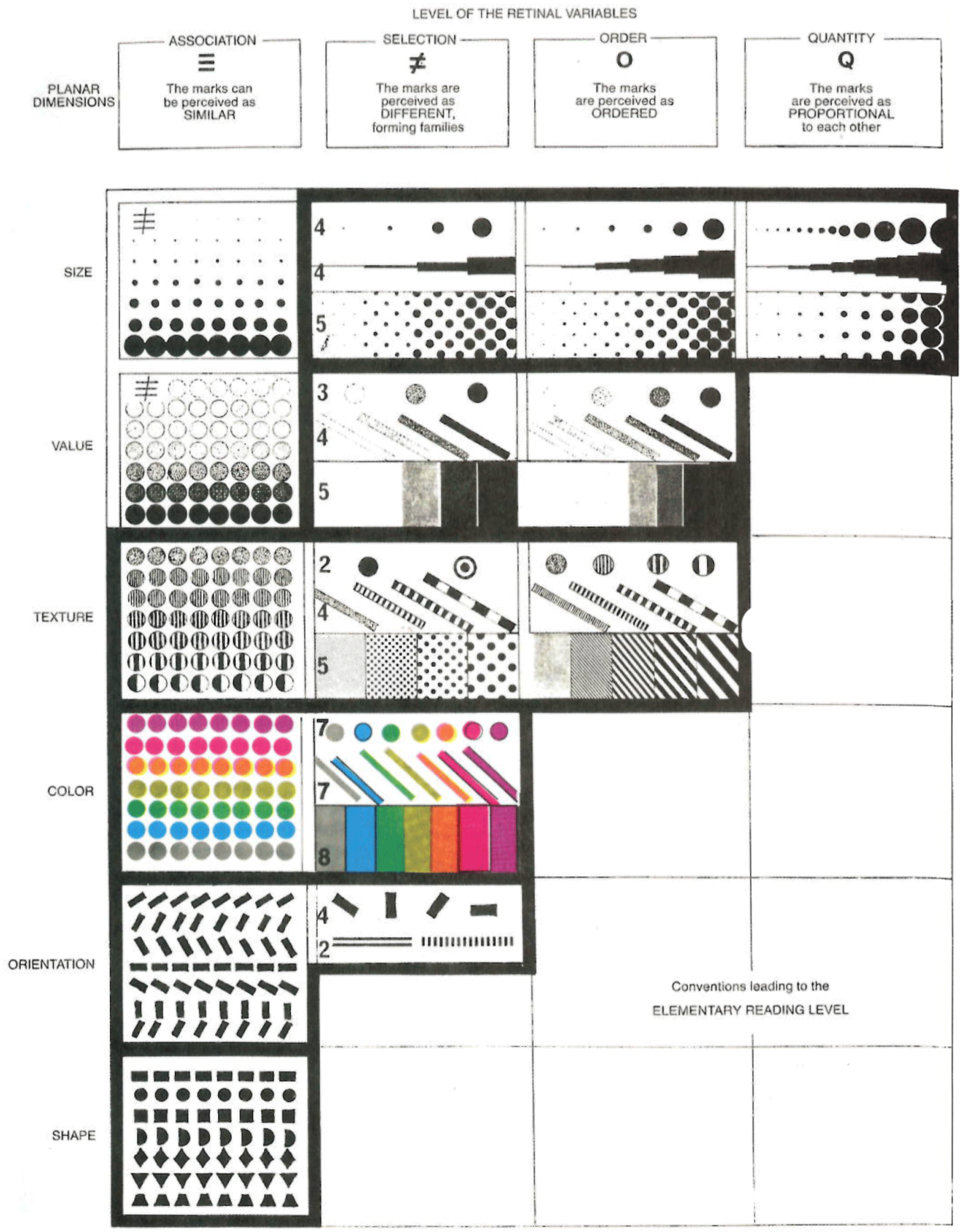
Orientation

Orientation is the differential angle between the fields created by several parallel signals that constitute a perceptible stimulus for orientation variation. A point or mark can vary in a large number of interpretations without changing its position, simply rotating around its axis. However, it is important to emphasize that the eye is only sensitive to changes in the orientation of linear marks. In the implementation, Bertin states that in point representation, orientation is the only available variation which can differentiate signs of equal visibility (Bertin, 1967, p.93). Linear variation is limited to two orientations and the area representation is easy to construct, but it can only be used for combination or selective variables.

Shape

A mark that has a constant area can have numerous shape possibilities. The shape perception variable represents the form that a mark takes. Bertin (1967) states that shape variation is associative and can be used when the image has a density of signs, however it is not selective — it is not possible to ask the user to identify a certain shape. In the implementation, shapes representing points are easier to identify if they are geometric, while amorphous shapes are difficult to distinguish from each other. The shape variation in the line representation can translate different concepts according to their angularity. Finally, the representation of areas can contain a large number of shapes.

Figure 5: Table of properties of retinal variables. **Source**: Bertin, 1967, p96.



Looking at the table of properties of retinal variables made by Bertin (1967), it is possible to notice that the size, value, and texture have good properties to transmit the perception of the ordering; the size is the only one that can transmit the proportional perception and variables like shape, color or orientation can effectively transcribe the notion of diversity, that is, they are good visual properties to use in categorical data.

2.2.2. Principles of Graphic Excellence

> Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.
>
> (Tufte, 2001, p.51)

Although visualizations are a great way to convey information and, as presented in the studies by Bertin (1967), have a series of visual channels that generate an immense possibility of graphic representations, some reservations must be taken for the production of efficient visualizations.

Edward Tufte is an American statistician recognized for his studies in the production of graphs and visualizations, where he presents a series of principles for graphic excellence. Through a series of examples of multivariate graphs, such as data maps, time series, spatiotemporal narrative drawings, and relational graphs, the author talks about how to narrate, investigate, and summarize data clearly, accurately, and efficiently using graphic design. Tufte (2001, originally published in 1983), argues that graphs can be more revealing than conventional statistical calculations. However, the author adds that due to a series of factors, bad graphics are easily found, in terms of data distortion, excesses in graphic stains, or even contextualization. With this in mind, Tufte presents the following principles to guide the production of a graphic of excellence:

Show the Data and Focus on the Substance

Every graphical display should show the data. In a visualization focused on the data that is the main thing to think about, the spectator needs to think about the substance of the information, not about how beautiful it is, nor regarding the methodology or the technology used for its production. The spectator cannot think about anything other than the information communicated there.

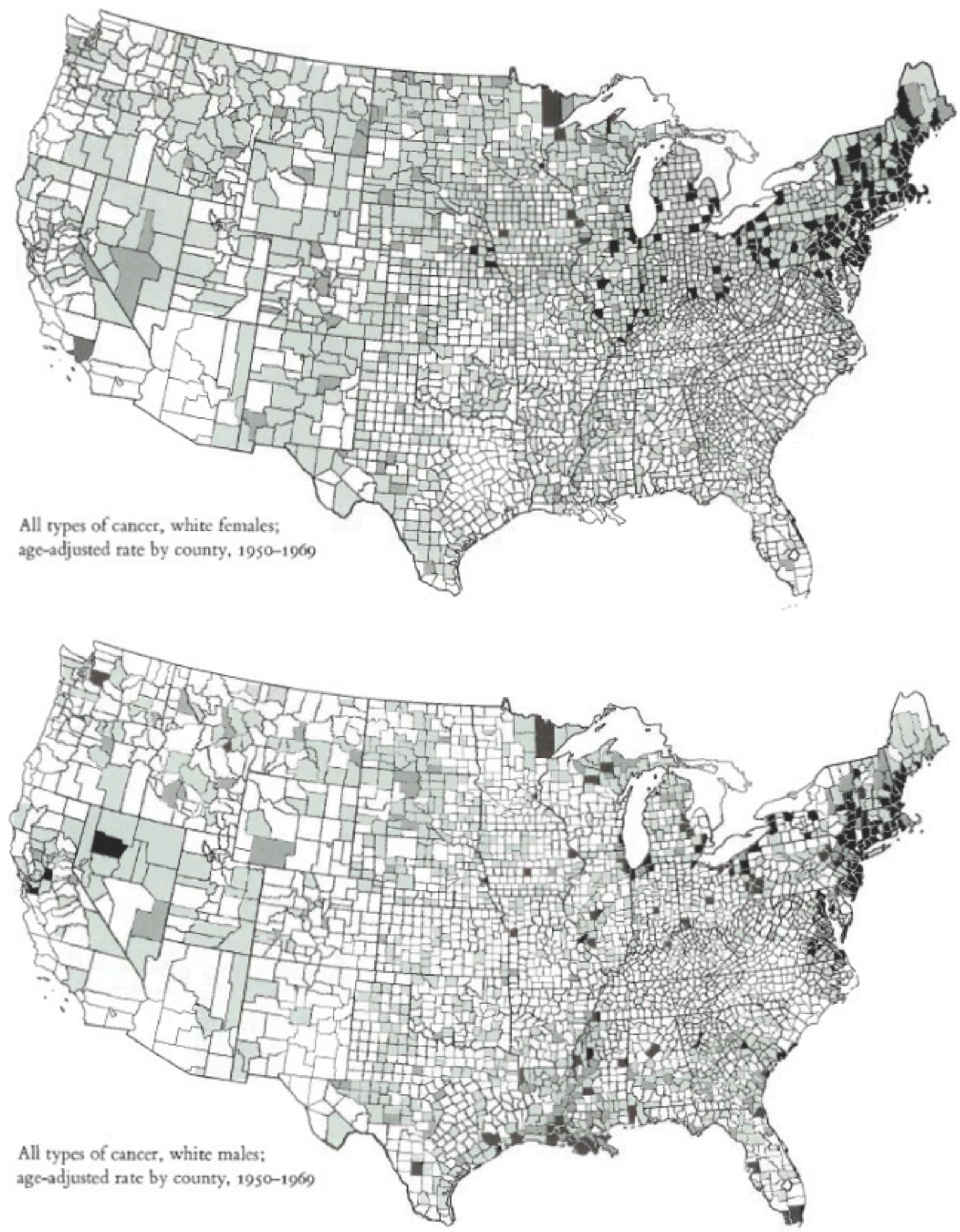


Figure 6: All types of cancer, white females and males. **Source**: Tufte, 2001, p.17

Tufte (2001) says that maps are powerful because they can quickly and naturally direct the attention, thus allowing the exploration of the substantial content of the data. When looking at a map, a viewer will hardly think about its graphic quality or the technology used, but focuses directly on the content. For example, the maps below are part of a series of 6 age-adjusted maps that report the death rate from various types of cancer for the 3,056 counties in the United States. Each of these maps carries around 21,000 numbers. Only images can carry such a small amount of data in such a small space. Furthermore, all this data, thanks to the graph, can be thought of in many different ways at many different levels of analysis, from contemplating general patterns to detecting very precise count details by municipality (Tufte, 2001).

Avoid Distorting Data

Krok (2021) states that although a recipient may be impacted only by changing font size and style within a text, graphics and tables are the most susceptible to distortion. For there to be no distortion in a graph, its visual representation must be consistent with their numerical representation (Tufte, 2001). An example of a bad graph with regard to data distortion and inconsistency in visual representation is represented in Figure 7. The visualization is from a newspaper article in which it represents a series of fuel economy standards to be met by car manufacturers, starting at 18 miles per gallon in 1978, moving in steps of 27.5 until 1985, which would mean an increase of 53%. However, when analyzing the graph it is noticed that the use of perspective within the representation of the road makes the magnitude of the rows not match the growth represented in the data. Making a calculation based on the proportions of the graphic representation, the increase would be 783% and not 53%.

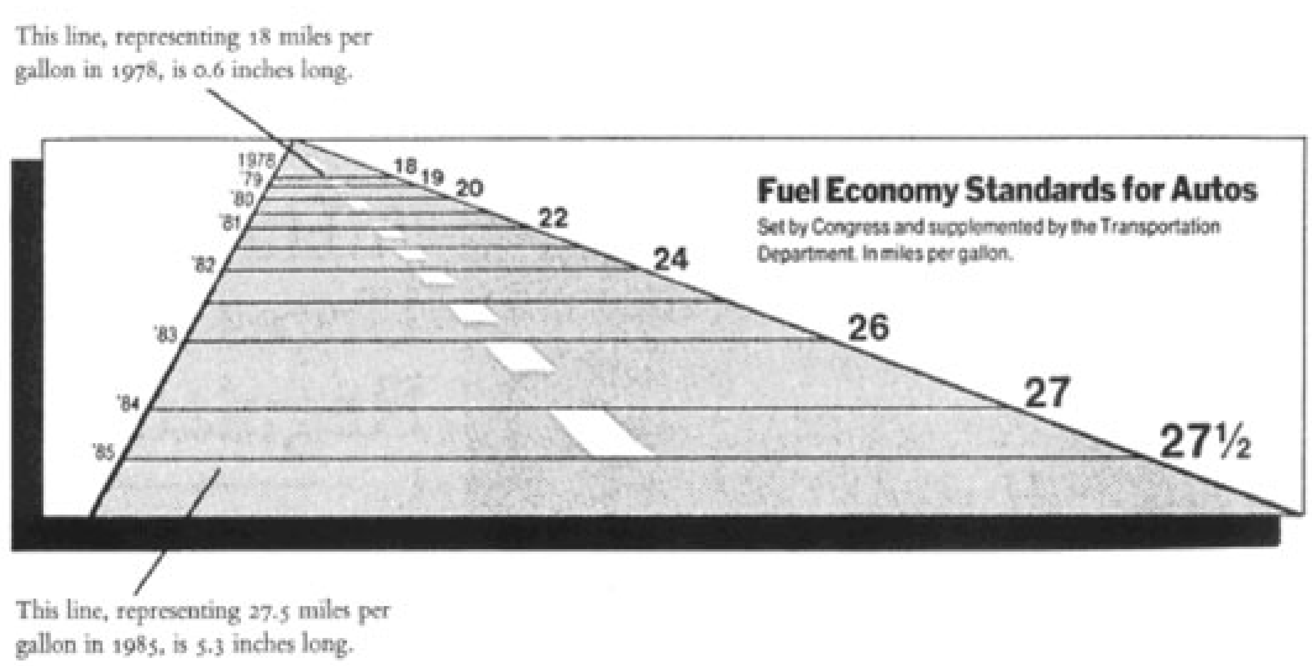


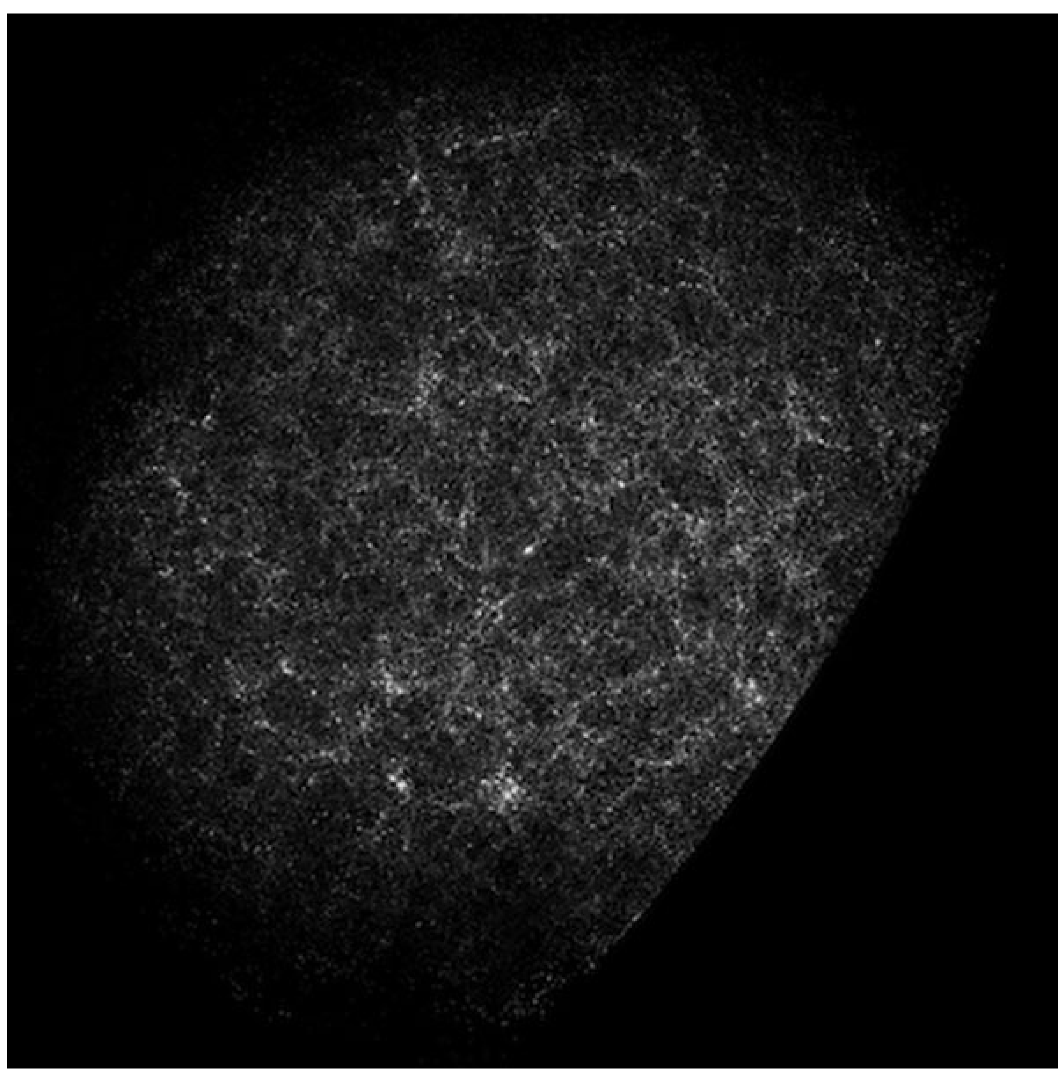
Figure 7: Fuel economy Standards for autos. **Source**: Tufte, 2001, p.57

Distortions such as these must be avoided because when it comes to graphic reading, the recipient of the message unconsciously does everything to limit their cognitive effort (Krok, 2021) and may have an erroneous interpretation of the data represented there. Thus, it is important to ensure their maximum readability when creating the message and avoid distorting the presented data.

Many Numbers in a Small Space

According to Tufte (2001), a good graph presents the largest number of data within the smallest space. The author praises the map produced by Seldner et al. (1977) which counts the galaxies through a grid of rectangles. The number of galaxies per rectangle is represented by the intensity of the white, the lighter areas show a higher density of galaxies. Tufte (2001) also states that maps are a powerful method when it comes to plotting a lot of data in small spaces.

Figure 8: New reduction of the Lick catalog of galaxies.
Source: Seldner, 1977.



Make Large Data Sets Coherent

In addition to showing the least amount of data in the smallest space, attention must be paid to the coherence of the representation. Graphs, according to Tufte (2001), should make large volumes of data coherent and easy to understand, and allow different levels of reading. In figure 9, it is possible to see a graph that presents the temperatures for the entire year of 2003, a total of 3,322 numbers are presented. Daily maximum and minimum temperatures are plotted against the long-term average, showing highs and lows during the year. This chart presents a large collection of numbers successfully comparing data and telling a story.

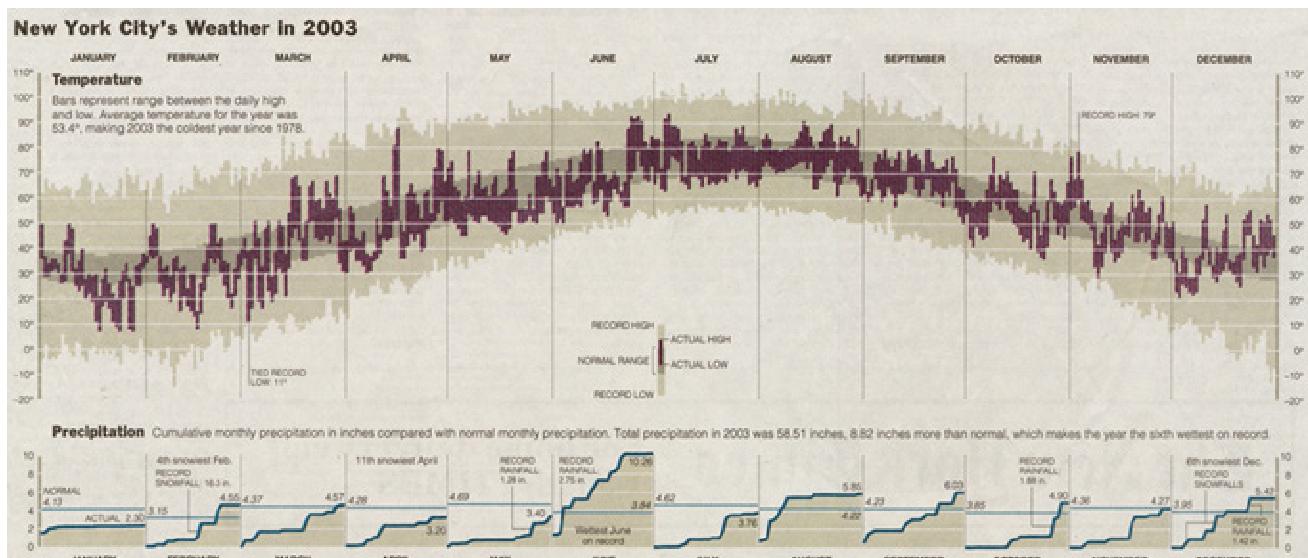


Figure 9: New York City" Weather in 2003. **Source**: https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00014g.

Encourage Comparison

A well-designed visualization encourages the eye to compare different pieces of data. For example the small-multiple below, which shows the levels of three air pollutants—nitrogen oxides, carbon monoxide, and hydrocarbons—located on a two-dimensional surface four times a day. A small multiple is a type of graph in which the same graphic structure is repeated several times, in this case 12. Tufte (2001) says that this type of representation is economical because, once the viewers understand the design that a slice has, they have access to the data in all other slices, allowing them to focus on changing the data without spending time and effort on understanding the graphs. The author also states that multiple time-series, as in the case of the graph below, encourage not only the comparison with each series over time, as any time-series does, but also between the different series presented, here are the nitrogen series, carbon monoxide and hydrocarbons.

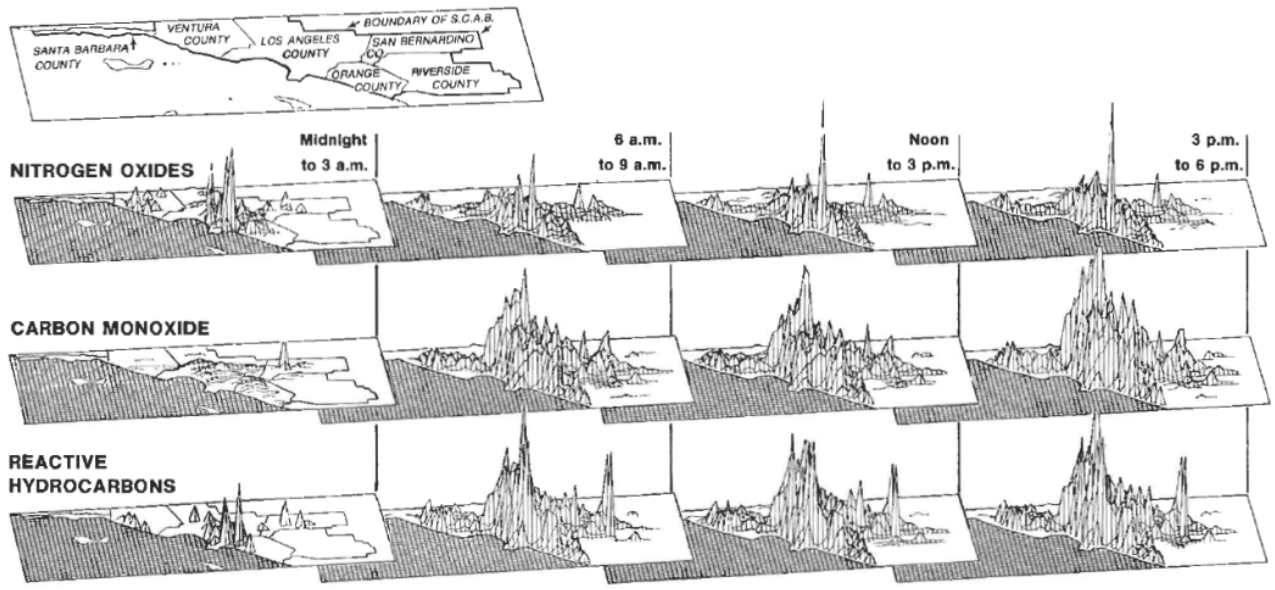


Figure 10: The air pollution display, example of small multiple. **Source**: Tufte, 2001, p.42

Have a Clear Purpose

Graphic displays must serve a reasonably clear purpose: description, explorations, tabulation, or decoration (Tufte, 2021). A good example is a graph by Marey about the schedule of trains departing from Paris to Lyon in 1880. The arrivals and departures of a station are located along the horizontal line, while stations, located on the vertical line, are separated in proportion to their actual distance. The slope of the line reflects the speed of the train: the more vertical the line, the faster the train. The intersection of two lines locates the time and place where trains going in opposite directions pass each other. It can be seen that, due to its clear objective, it does not have unnecessary information or graphic stains that hinder the reading of the chart; the entire project is easy to read and makes it possible to compare the different times and duration of transport.

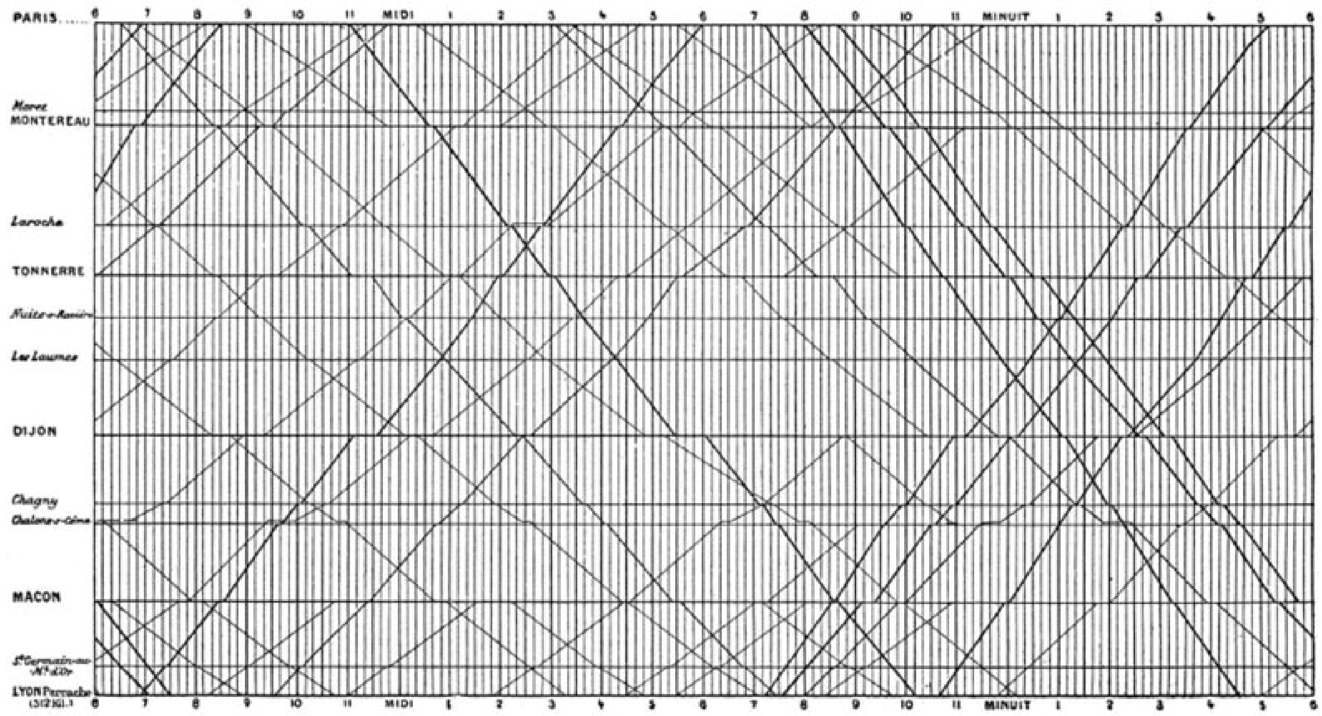


Figure 11: Marey's graphical.
Source: Tufte, 2001, p.31.

Reveal the Data and Be Integrated

It is important to reveal data at various levels of detail, from a broad overview to a fine structure and a graph needs to be closely integrated with the statistical and verbal description of a data set. A great example is a classic by Charles Joseph Minard (1781-1870). Drawn in 1861, the graph combines data maps and time series to depict the losses suffered by Napoleon's army in the Russian campaign in 1812, starting from the left on the Polish-Russian border, the thickness of the brown line shows the size of the army when it invaded Russia in 1812, as time goes by, the expressive decrease of the army can be seen. It is noticed how the graph is very well integrated into its different layers of information and legends, so that the viewer will translate the information in an immersive way without losing focus with anything or needing to read to look for information in legends away from the map. According to Tufte (2001), this may be the best statistical graph ever created.

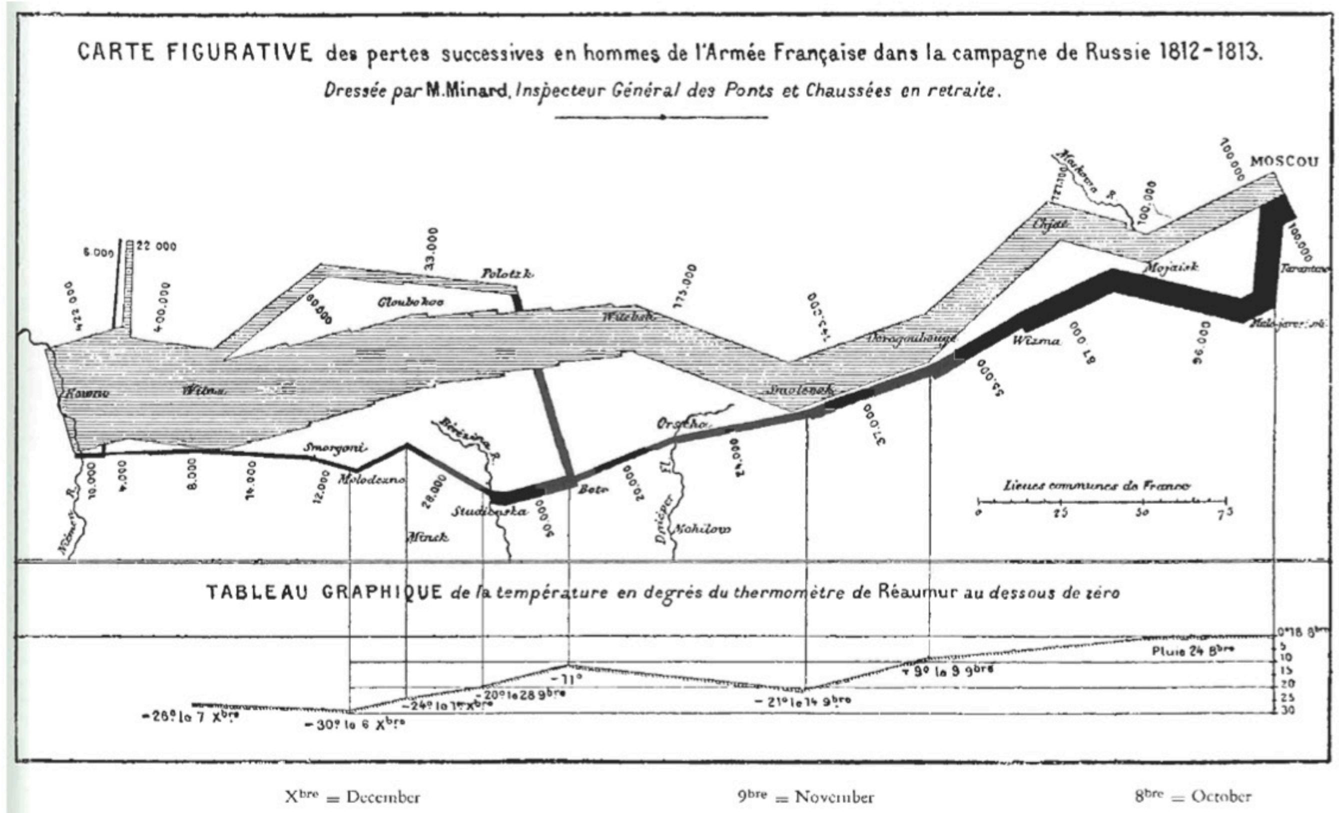


Figure 12: Graphic oh Charles Joseph Minard. **Source:** Tufte, 2001, p.41.

Tufte summarizes all the above observations in 5 points that lead to graphic excellence:

- Graphical excellence is the well-designed presentation of interesting data - a matter of substance, statistics, and design
- Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency
- Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
- Graphical excellence is nearly always multivariate
- Graphical excellence requires telling the truth about the data

With the studies of Tufte (2001), it is possible to perceive that it is not enough to have knowledge about the visual channels, but one must also be aware of the different implications of visual perception. The composition of the graph as a whole must be cautious and always be focused on showing the data in the best possible way.

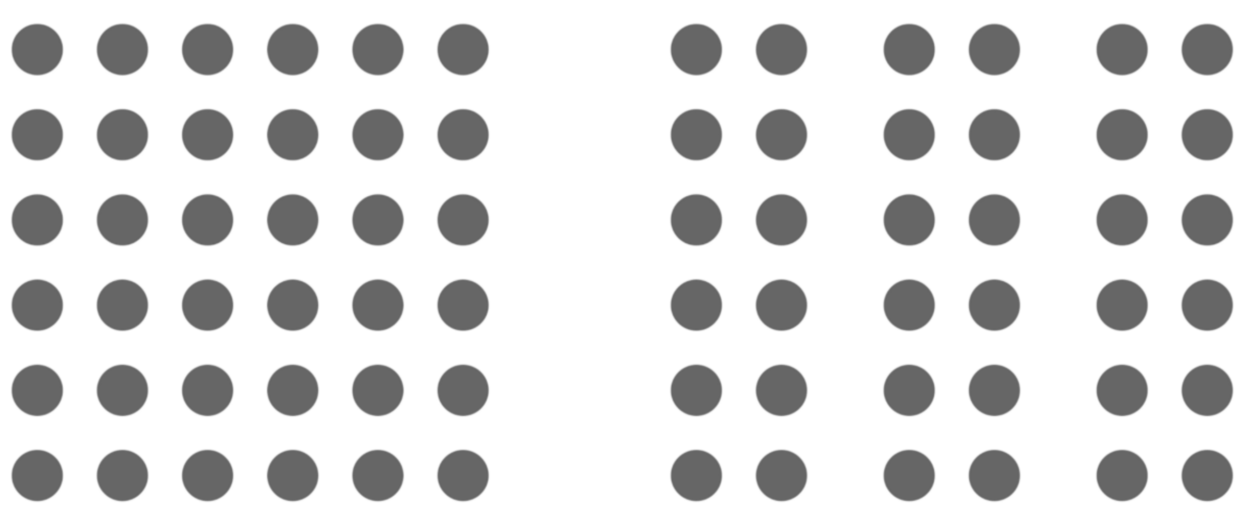
2.2.3.Gestalt Principles

In all visualizations, a critical aspect related to users is the abilities and limitations of the visual system. If the purpose of the visualization is to accurately convey information with images, perceptual skills must be considered (Ward et al., 2010, p.35). Gestalt is a school of psychology, started in 1912, that proposed a set of laws and principles describing how human visual perception perceives patterns. Gestalt laws can be used as design principles for effective ways to improve pattern detection and perceptual inferences (Meirelles, 2013, p.22). Next, each of the laws proposed by the gestalt will be presented.

Proximity

Proximity describes the tendency of close visual objects/elements to form a visual unit, that is, when joining the elements within the spatial variation of the display, they pass the perception of belonging to the same visual group. Analyzing the example below, two different groups can be seen. This is due to the law of proximity, since the elements do not have other visual variations such as color, size, or shape. Proximity always refers to local characteristics, that is, as the elements are spatially organized, the use of proximity is a good strategy to facilitate the detection and search of associated data.

Figure 13: law of proximity.
Source: Meirelles, 2013, p.19.



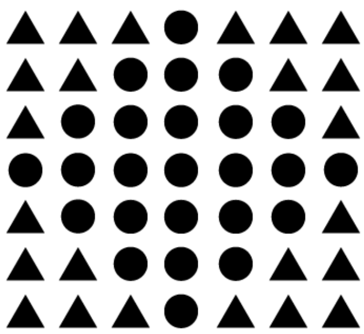


Figure 14: law of similarity.
Source: Meirelles, 2013, p.51.

Similarity

It is the law that describes the tendency to form groups into similar visual elements. This applies to non-local features, that is, they are not related to their spatial position, but features such as color, shape, texture, etc. The use of the law of similarity is recommended for categorical associations or to apply the notion of hierarchy between elements.

In conclusion, when visual elements are spread out in space but have some visual characteristics in common, as in figure 14, we tend to unite them in the same category or group.

Figure 15: Common fate.
Source: Meirelles, 2013, p.73.

Common Fate

It is the tendency to associate elements that are moving in the same direction. The parallel lines represent a graphic element that makes it easy to understand this grouping. Notice, for example, in figure 15: in the upper part (A) there is a single unit due to the proximity, but in the lower part (B) it is possible to divide it into two distinct groups due to the change of orientation of half of the elements.

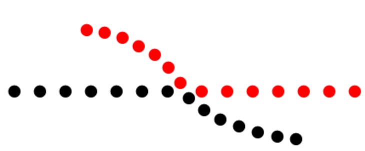


Figure 16: Law of continuity.
Source: Meirelles, 2013, p.58.

Continuity

The law of continuity talks about our brain's tendency to create flows and sequences from a series of elements that have some spatial or symbolic connection, such as alignments or markers. It also represents the tendency for objects to accompany others in the sense of achieving a form. Experiences of good continuity are commonly found on maps, where the contours of states, roads, or rivers can be easily distinguished (Meirelles, 2013, p.58). The law of continuity is also widely applied in the architecture of large buildings and text layouts, for it seeks to establish the best possible perception for the eyes.

Closure

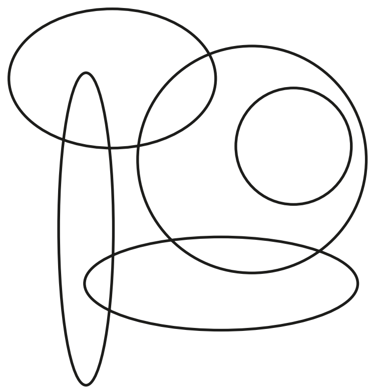


Figure 17: Law of closure
Source: Meirelles, 2013, p.33.

The principle of closure talks about the performance of closing contours and uniting limited visual elements in a single way. Even when the elements are superimposed, as in figure 17, there is a tendency, which is influenced by the principle of good continuity, to separate the units and apply a union of shapes, thus visualizing a large group of different closed shapes. It is like the mind "enclosing" the missing parts and "closing" the visual element, so when an object is fragmented but the small fragments have similarities in shape or alignment, we tend to bring them together (Meirelles, 2013).

> When we perceive data representation such as Venn diagrams, for example, we make use of the closure principle to extract information. The closure principle plays a significant role in distinguishing the sectors and the levels of hierarchy in the SmartMoney map of the market. Each sector, or container, has a clear boundary represented by thicker and lighter lines
>
> (Meirelles, 2013, p. 33)

Segregation Between Figure and Ground

The segregation between figure and background is the law of the gestalt that talks about the tendency of the brain to unite forms, judging them as main. Normally, the brain classifies images according to what surrounds the objects, thus establishing depth relationships. In other words, if an object is surrounded by another, with these having any kind of visual variation - color, texture - the eye tends to focus on one of them, which is called a figure and all other more distant elements are considered as background. When the lines and the contrast between the figure and background are very unequal, this separation becomes a little ambiguous. This is the case of the famous image of Rubin's vase shown in Figure 18, where it is possible to see two profiles on a black background or a vase on a white background.

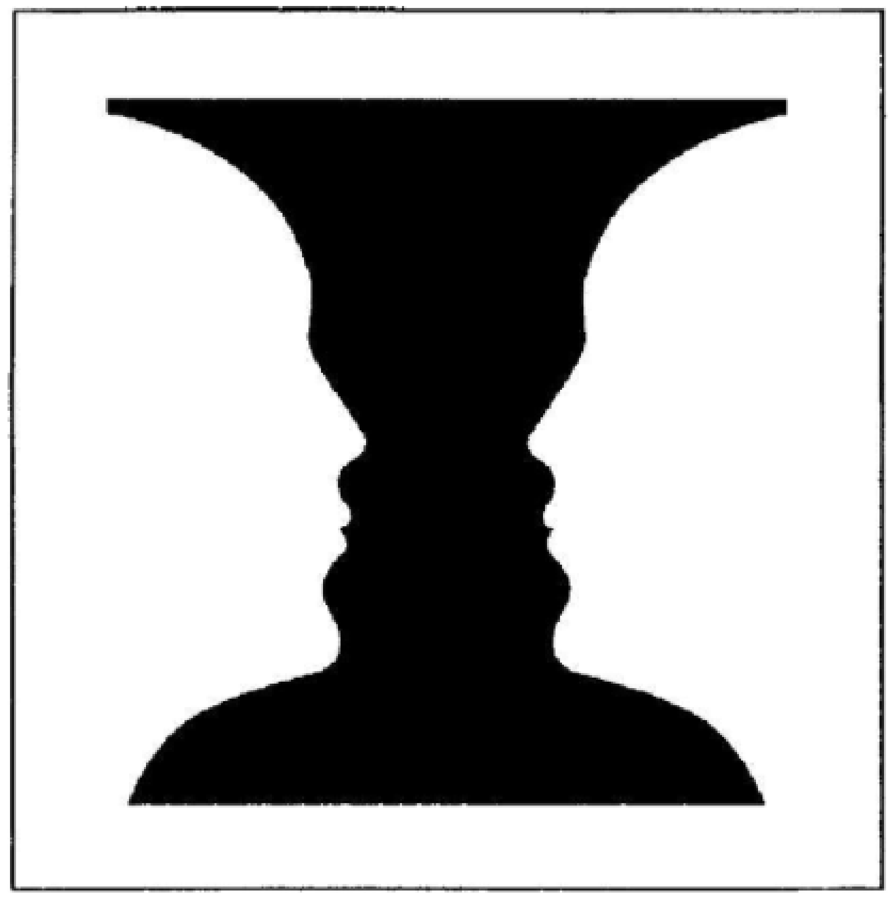


Figure 18: Rubin's vase.
Source: Meirelles, 2013, p.126.

These are the main laws when it comes to visual construction within Gestalt psychology. In addition to those mentioned above, Ware (2019) also presents the law of simplicity and the law of familiarity. The law of simplicity is related to the tendency of the human brain to reduce visual elements to the simplest and easiest possible way. This law is important in that, because, when a layout has a series of simple shaped elements, there is a greater probability of the user to mark them, in the sense of remembering and understanding them in an integrated way. Those are very important concepts when it comes to data representation: after all, the intention is that users understand the data contained therein in the visualization and retain the information presented.

As we saw in each of the topics described, the human brain is always trying to form patterns in order to understand all the visual impulses received. The law of familiarity summarizes this tendency to standardize and categorize new elements within a repertoire of already known elements.

2.2.4. Conclusion

In this chapter, we looked at some concepts behind visualizations and what principles should be considered when producing them. With the theories of graphic semiology (Bertin, 1969) it was possible to understand the different parts that make up a visualization, the mark and the plan, and how it can be manipulated in the most different ways through visual channels. It has also been observed that some visual channels are more appropriate for some types of data than others, making visualizations that use inadequate visual channels have unsatisfactory results with regard to readability.

The following shows a little of why there are bad visualizations, from the studies of Tufte (2001) issues such as data distortion, lack of clarity and objective in the production of visualization, excessive use of graphic stains in addition to the existence of visualizations were addressed. Weak graphics that do not have enough data or are of poor quality data. Everything was approached through what Tufte (2001) defines as principles for graphic excellence. Finally, the principles of Gestalt psychology were presented, which explain a little about how visual perception works and how certain graphic compositions convey certain interpretations and perceptions such as groupings, sensation of movement, and highlighting effects. These studies are important to realize strategies that are useful when working with visualizations to produce desired effects concerning visual perception.

In the next chapter, all the concepts discussed here will be discussed again within specific contexts. After realizing what makes up visualizations, through Bertin (1969), what makes visualizations good or bad and the principles of graphic excellence and a better understanding of visual perception through Gestalt principles, in the next chapter it will be presented how to apply these concepts within data glyphs.

2.3.Data Glyphs

Data glyphs are a design technique used to represent multivariate data. Multivariate data consists of some number of points, *m*, where each point is defined by an n-vector of values. This data can be viewed as an *mxn* matrix, where each row represents a data point and each column represents a variable. A variable can be scalar or vector, nominal or ordinal, and it may or may not have a distance metric, ordering relationship, or absolute zero (Ward, 2008).

Glyphs can be considered as small visual elements in which unique data points are individually encoded, assigning their dimensions to one or more tags and their visual variables (Fuchs et al, 2016). According to Borgo et. al (2013), a glyph-based visualization is a common form of visual design in which a dataset is represented by a collection of visual objects, the glyphs. In a broad interpretation, Borgo (2013) states that:

- A glyph is a small visual object that can be used independently and constructively to represent attributes of a data record or the composition of a set of data records;

- Each glyph can be placed independently of the others, while in some cases, the glyphs can be spatially connected to convey the topological relationships between data records or geometric continuity of the underlying data space; and

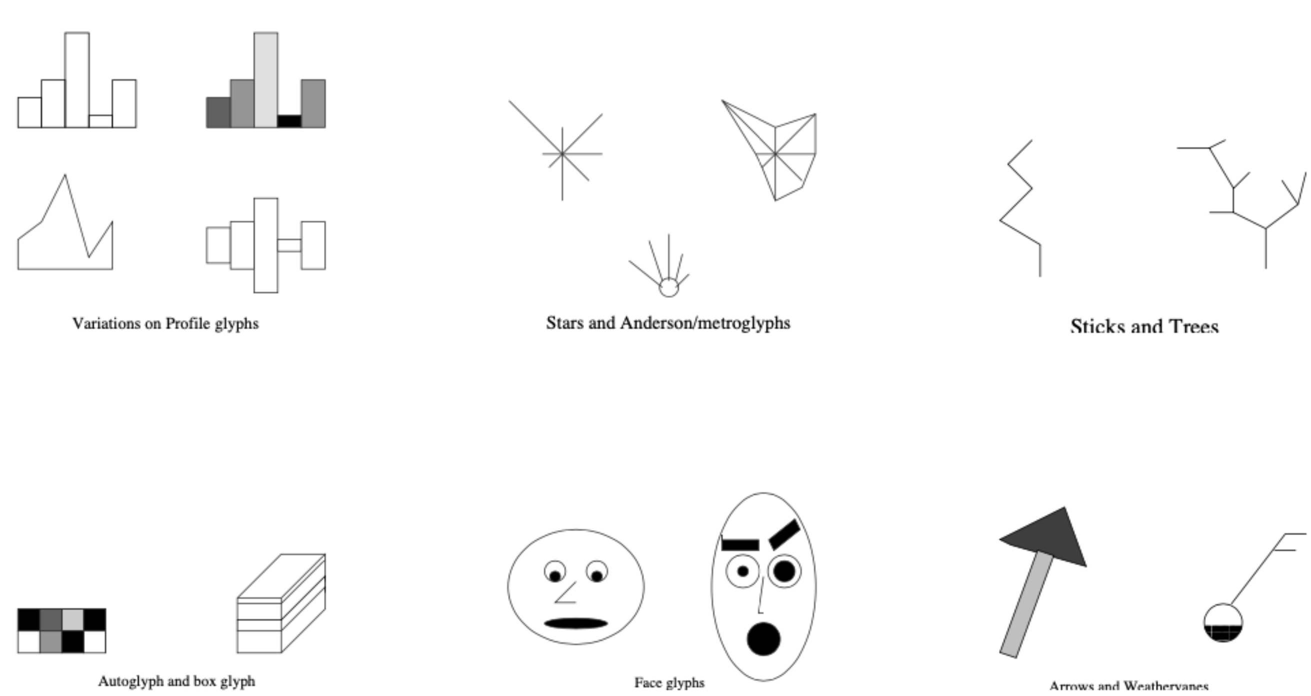
- Glyphs are a type of visual sign that can make use of visuals from other types of signs, such as icons, indices, and symbols.

Data glyphs have a long history; one of the first glyphs designed within the 1950s were the metroglyphs, which use line length to encode data (Fuchs et al, 2016). An example much discussed in the literature is the Chernoff face, a glyph that encodes values through the characteristics of a cartoon face, such as the length of the nose or the orientation of the eyebrows (Chernoff, 1973). Another well-known type of glyph is the Star Glyph which uses the length of evenly spaced rays emanating from the center (Ward, 2002). Over the years, many different glyph variations have been introduced to better fit certain types of data or to solve specific tasks more effectively (Fuchs et al, 2016).

According to Chung et al. (2013), in the era of the data deluge, there is a great cost-benefit ratio in the use of glyph-based visualization, as glyphs have a differential of not using spatial attributes for their development, that is, in addition to glyphs, it is possible to use the plane variables, x and y, to display additional information. Moreover, glyphs have great flexibility in creation regarding the use of visual channels for the representation of attributes of multidimensional data, being able to even use semantic relations for the encoding of the data when choosing the appropriate visual channels. This flexibility allows designers to create new and innovative glyph representations for specific data, tasks, or contexts (Fuchs et al, 2016).

Over time, glyphs of the most diverse types have been recorded. Some are only recommended for specific applications such as fluid flow visualization or time visualization, others however are more general and can be applied in many different ways. Ward (2002) cataloged the most diverse types of existing glyphs—in Figure 19, we present some of the most common.

Figure 19: Types of glyphs.
Source: Ward, 2008.

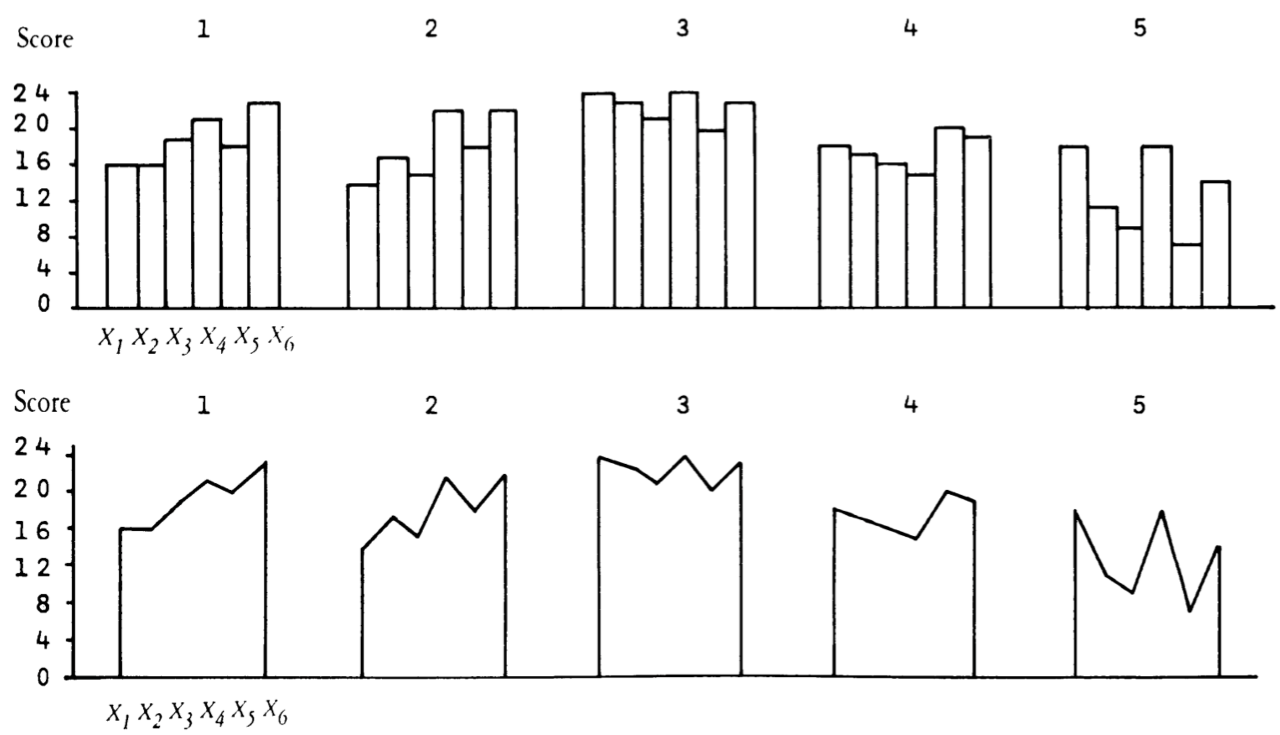


Profiles

It is a type of glyph that uses height and color to represent bars (Ward, 2002). According to DuToit et al. (2012), this type of representation is the most natural and simple way to represent a p-dimensional observation, using p-columns or vertical bars. The bars are placed side by side with the heights respectively proportional to the p-observations.

How the bar diagram is constructed for each observation vector. Alternatively, the midpoints of the bars can be joined together to give the profiles a more continuous appearance. The overall appearance of the profiles is naturally very dependent on the ordering of the variables.

Figure 20: Two alternatives for profile representation. **Search:** DuToit et al., 2012, p. 56.



Star Glyph

The star glyph is one of the most widely used glyphs (Fanea et al., 2005) for multivariate data. The glyph represents the attributes as equal angled axes with an outer line connecting the data value points on each axis. The length of each axis is proportional to the magnitude of the variable with respect to the maximum value of all variables. Each star represents a separate data record and each axis of the star represents different attributes of datasets. Similar records share particular characteristics that will share the same corresponding branch of the star. (Kamsani et al., 2015).

An example is shown in Figure 21, where the size of the star can be considered as a measure of the total performance in the Ability Tests. It is noticed that students 3 and 5, in general, did, respectively, the best and the worst. Within Star Glyphs, when values are negative, private variable numbers can be converted to positive values via a zero-point transformation (adding a constant to each variable's value).

Chan (2006) states that the viewer's experience when reading a star glyph-based view can be improved by manipulating the dimensions of the attributes. For this, an interactive visualization can be used, allowing users to visualize the data at different angles, axis, and manipulation of attributes (Kamsani et al., 2015). The application of colors can also enhance the experience, bringing instant recognition of similarities or differences of large data items and the relationship of expressed attributes (Keim, 2002).

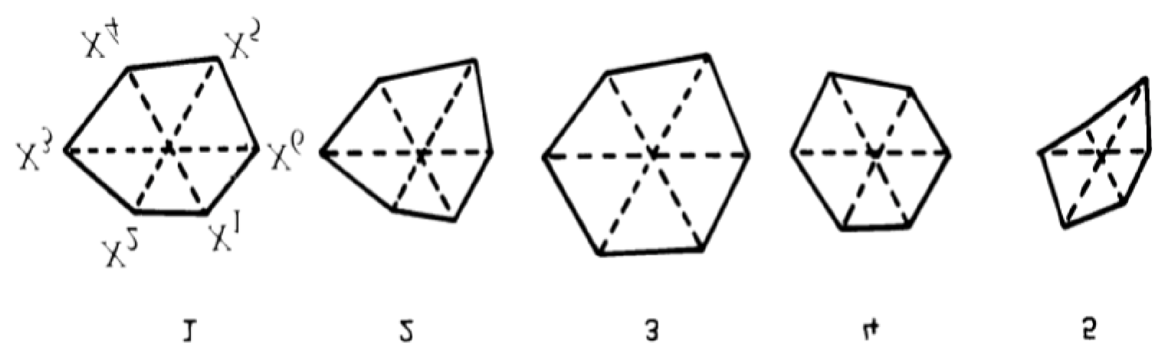


Figure 21: Star representations of five pupils' scores in six ability tests. **Search**: DuToit et al., 2012, p. 66.

Anderson / Metroglyph

Anderson (1960) proposed a representation where the measures of each of the variables in a specific case are represented as radii of a circle. This type of glyph differs from the ones presented previously because, firstly, all the rays are drawn from the top of the circle outwards and, secondly, the end points of the rays are not connected, as in the case with the stars.

Figure 22 shows the glyphs of the five students' scores on the six skill tests also represented in Figure 21 with the Star Glyph. According to Anderson (1960), all the information in a glyph can be grouped through an index that is calculated as the sum of the lengths of all the radii. If a ray can have a length of only 0, 1, or 2, the index will range between 0 and 2p. If there is a large number of important variables (more than seven), the four or five having the highest linear relationship can be turned into a single index as described above. The indices along with the values of the other variables can then be used in glyph representations. The index value of each of the five glyphs in the X appears below each glyph (DuToit et al., 2012).

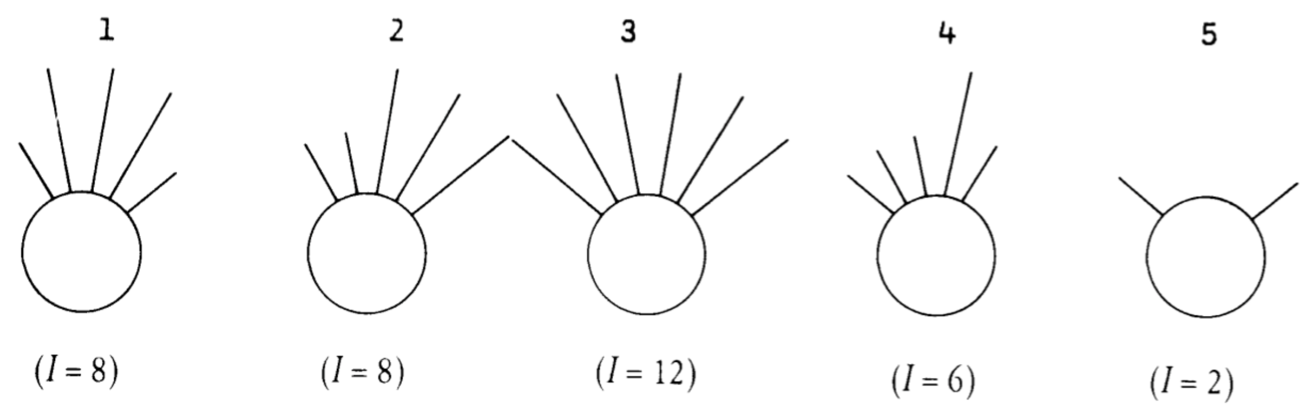


Figure 22: Representations of five pupils' scores in six ability tests with metroglyph. **Search:** DuToit et al., 2012, p. 68.

Stick Figures

According to Keim (2002), the Stick figure is an iconic display technique, which allows the visualization of larger amounts of data and, therefore, is more suitable for data mining. As indicated by the name, the icon is some kind of stick figure. Two dimensions are mapped to the display dimensions and the remaining dimensions are mapped to the angles and/or lengths of the puppet icon's limbs. Figure 23 shows a visualization using stick figures for multispectral images from the NOAA-7 AVHRR meteorological satellite; it is possible to see how various textures are created in the visualization, and this is done through the data densities. If the data items are fairly dense in relation to the screen dimensions, the resulting visualization has texture patterns that vary with the characteristics of the data and are therefore detectable by preattentive perception (Pickett and Grinstein, 1988).

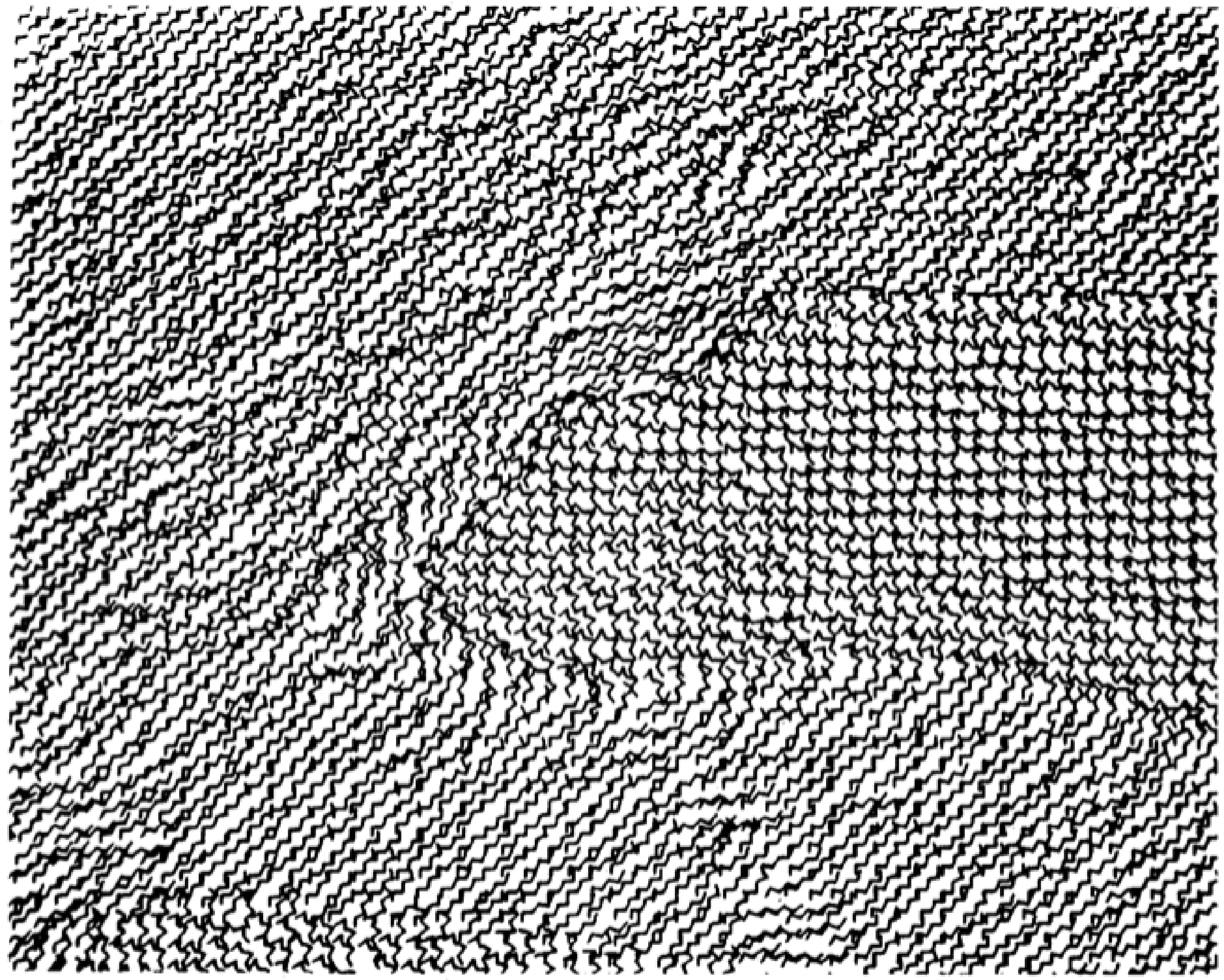


Figure 23: Multispectral images from the NOAA-7 AVHRR weather satellite. **Search**: Pickett and Grinstein, 1988.

Faces and Chernoff Faces

One of the most common type of glyphs (Prohaska, 2007) is based on representing data variables by different characteristics of a cartoon face, such as face shape, eye size and position, nose length, and mouth curvature. This category of glyphs became known through the studies of Hermann Chernoff (1973), in which he introduced what we begin today as Chernoff faces (Figure 19): the author generated faces having up to eighteen distinct facial parameters. The fundamental reason for using Chernoff faces is that humans can easily and quickly recognize distinct faces and notice changes in even small and subtle facial features (Morris et al., 2000). However, Ebert et al. (1999) states that research on facial recognition is significantly far from conclusive, and many questions still need to be answered. The way in which humans recognize faces does not contain a unique process that is not used in recognizing other objects. Therefore, other objects that are recognized with the same process would be just as effective for viewing as faces.

One of the biggest problems with trying to define how people recognize faces is that various facial features have very different meanings depending on whom you ask. Within ethnicities, races, sexes, religions, and even social classes, faces and their features have varied connotations. When using Chernoff's faces, it must be taken into account that a human face can be interpreted very differently among various observers (Prohaska, 2007). The author also states that personal preferences about facial features may prevent an objective perception of available information. However, the belief still exists that certain faces or facial features produce remarkably uniform impressions on observers (Ebert et al., 1999).

Regarding the advantages of using face glyphs, Prohaska (2007) states that they are a very good method to discover the maximums and minimums of a database. Exceptional parameter values can be identified quite easily because the corresponding facial feature must look very attractive.

It was possible to see that glyphs can vary greatly in form and purpose of application. This statement is even more evident when within the studies of Fuchs et al. (2016), which makes a systematic review of a series of projects involving glyph-based visualizations and there are presented some glyph models that were not cataloged by Ward (2008), such as flowers and cars (Figure 24). In addition, visual alternatives were found to use faces other than the already known Chernoff model.

It is important to emphasize that, like any other visualization method, the use of glyphs is not exempt from limitations in its communication (Ward, 2008). As seen above, the characteristics of a glyph can vary greatly, presenting different limitations and flaws for different situations, ranging from limitations related to the visual channels used to represent the data or even the media in which the visualizations are presented. According to Ward (2008) most mappings, if not all of them, introduce biases in the process of interpreting the relationships between dimensions.

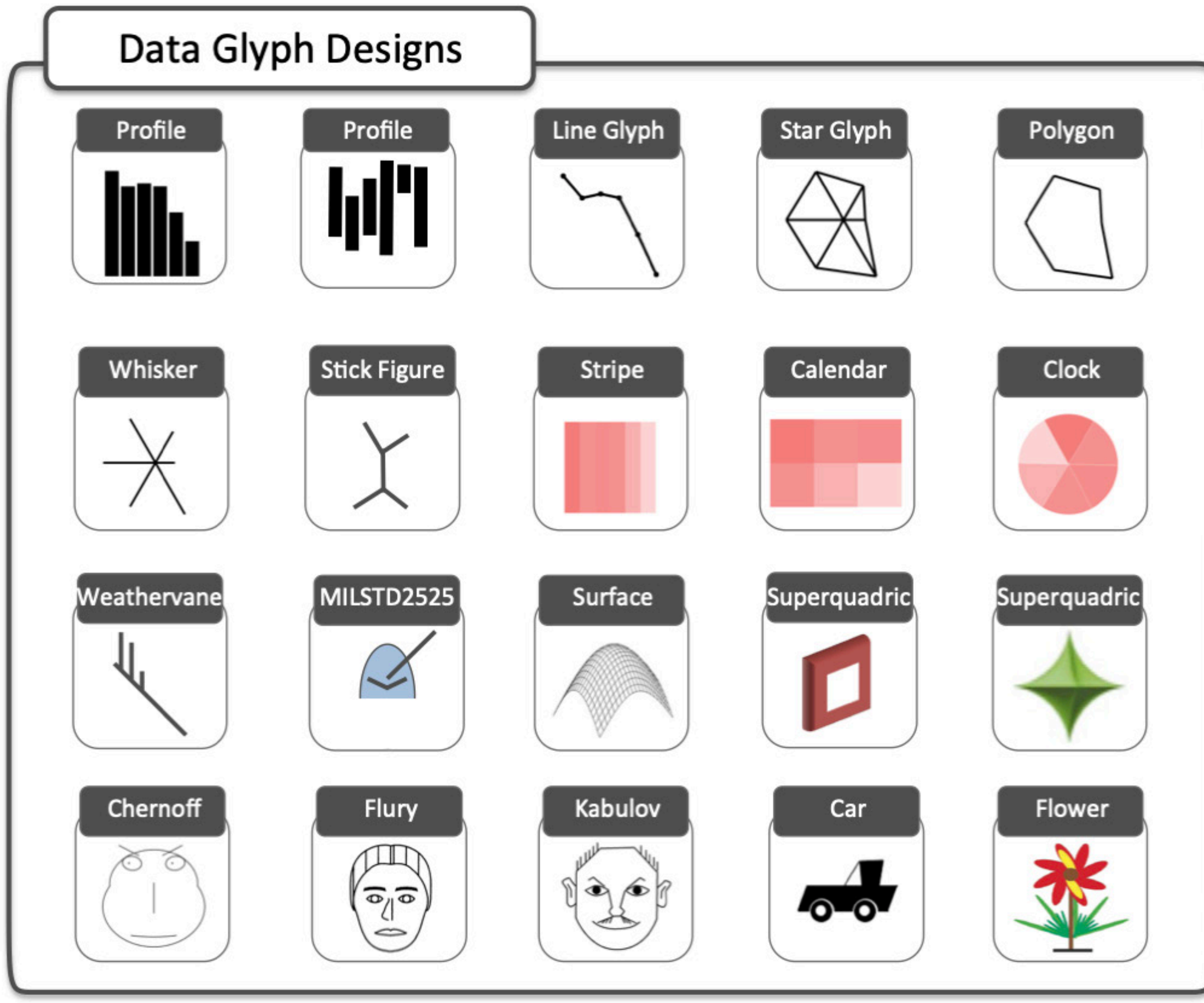


Figure 24: A selection of the different glyph designs found in the study by Fuchs et al(2016). **Search**: Fuchs et al., 2016.

2.3.1. Limitations and Biases

Like any other visualization method, glyphs have their flaws and limitations. Some of them, as seen earlier, are more present in some glyph categories than others. One of the limitations pointed out by Ward (2002) is that some relationships between data are easier to perceive than others, such as data dimensions mapped to adjacent components. Furthermore, the perception of visual channels is very variable at the level of perception, for example, our ability to accurately measure length is superior to attributes such as orientation and color (Ward, 2002). The perception of content can also vary greatly between different observers, or the same observer in different contexts; studies show that color is extremely sensitive to context (Levkowitz, 1997). Finally, there are also limitations based on the medium being used to communicate the information. Screen space and resolution are limited, and displaying too many glyphs at once can lead to overlaps or very small glyphs.

In addition to the limitations mentioned above, Ward (2008) states that one of the most common criticisms of glyphs is that there is an implicit bias in most mappings, which means that some attributes or relationships between attributes are easier to perceive than others. Within this panorama, Ward (2008) categorizes the possible biases glyph-based visualization is subjected to, whether they are perceptual bias, proximity, or clustering.

- **Perception:** it is easier to compare some graphical attributes than others. Cleveland (1987) conducted a study showing that users measure elements of length along a common axis more accurately than measuring angles. Hence, according to this study, the differences between the values contained in a bar graph are better perceived than the ones contained in the rays of a pie chart.

- **Proximity:** it is easier to compare and measure adjacent elements than those that are farther away. One type of glyph that is always subject to this bias is the star glyph: it is much easier to compare the difference of a ray to adjacent rays than to compare it to a ray on the opposite side.

- **Grouping:** graphic attributes can be grouped semantically or perceptually, resulting in a biased sense. For example, if we map two variables to the size of a face's ears, the relationship between those variables can be much easier to discern than mapping one to the shape of the eye and the other to the size of the mouth.

2.3.2. Design Recommendations

A glyph, as seen earlier, is composed of a set of visual channels, each channel encoding a variable from a multivariate data record. According to Borgo et al. (2013), the first criterion is that the visual channel should ideally be able to encode many valid values of the data variable, or collectively, different visual channels of the glyph could encode many data records with different combinations of data values. However, this is not the only criterion and, in many cases, it may not even be the most important criterion (Chung et al., 2013). The following are some criteria and recommendations proposed by several authors in order to produce an appropriate glyph design.

Data Mapping and Ordering

Within the development of the glyphs, each dimension of a dataset will be mapped to a specific graphic attribute. From the different visual attributes available (e.g., position, size, shape, orientation, texture, color) there is a wide range of mappings for possible data glyphs to be made, depending on the type of glyph being produced, type of data, and also the hierarchical order of the data according to the tasks that must be performed. Ward (2008) classified the mappings as follows:

- **One-to-one mappings**, where each data attribute is mapped to a distinct graphical attribute;
- **One-to-many mapping**, where redundant mappings are used to improve the accuracy and ease with which a user can interpret data values. This type of mapping can be used for data that is more complex and difficult to encode visually or data that has greater importance in the task to be performed;
- **Many-to-one mappings**, where several or all data attributes are mapped to a common type of graphic attribute, separated by space, orientation, or other transformation.

In addition to the data mapping format, it is important in the use of glyphs to determine which ordering will be most favorable to the task at hand (Ward, 2008). It should be noted that this importance is more explicit only in some types of glyphs, where the order is more noticeable, such as slash, star, and profile glyphs. Each dimension of a dataset will be mapped to a specific graphic attribute. By modifying the order of dimensions preserving the type of mapping, alternative views of the data are generated, depending on the number of variables presented within the glyph, the sorting possibilities are numerous, based on that Ward (2008) presents four possible strategies for choosing the best ordering of the data.

- **Correlation Oriented Sorting:** This is one of the most common sorting strategies in the literature, it is based on the use of correlations or other similarity measures to sort the data.

- **Symmetry-oriented sorting:** As presented in the laws of Gestalt, the human mind more easily assimilates simple and symmetrical shapes, so one strategy is to organize the data in this way. A user study using monotonicity, that is, the constant change of direction, is more difficult to perceive than the symmetry ordering, where rays with similar values are placed in opposite places. You could say that it's easier to notice changes in patterns when we see simple shapes that don't have many concavities.

- **Data-based orientation:** Another way to sort variables is by selecting one of the attributes present in the database, sorting it in ascending or descending order, and using this organization to represent all the data. This classification form is most commonly used for data that have some temporal relationship.

- **User-based orientation:** Finally, an ordering can always have the user as a reference, representing glyphs that have a specialized target audience, that is, the user group has a very good understanding of the subject matter. In this case, the user can be used to understand which attributes must be correlated more closely and directly, based on this information, the ordering is performed.

Glyph Sorting Criteria

Chung et al. (2015) described a set of design principles for mapping attributes to visually sortable glyphs. According to the author, this significantly increases the usability of glyph-based visualizations for comparative analysis of multivariate data and for supporting the visual search. The author also states that Chernoff Faces and Star Glyphs are some examples of multivariate glyphs where identifying glyphs with similar characteristics is effective but cognitively challenging when determining the ordering of the glyphs. That's because these glyphs are not visually sortable in any obvious way. This becomes more of a challenge when the glyphs are messy. Figure 25 shows the principles of Chung et al. (2015) applied and how the ordering of glyphs in a given spatial configuration is more informative to reveal multivariate trends.

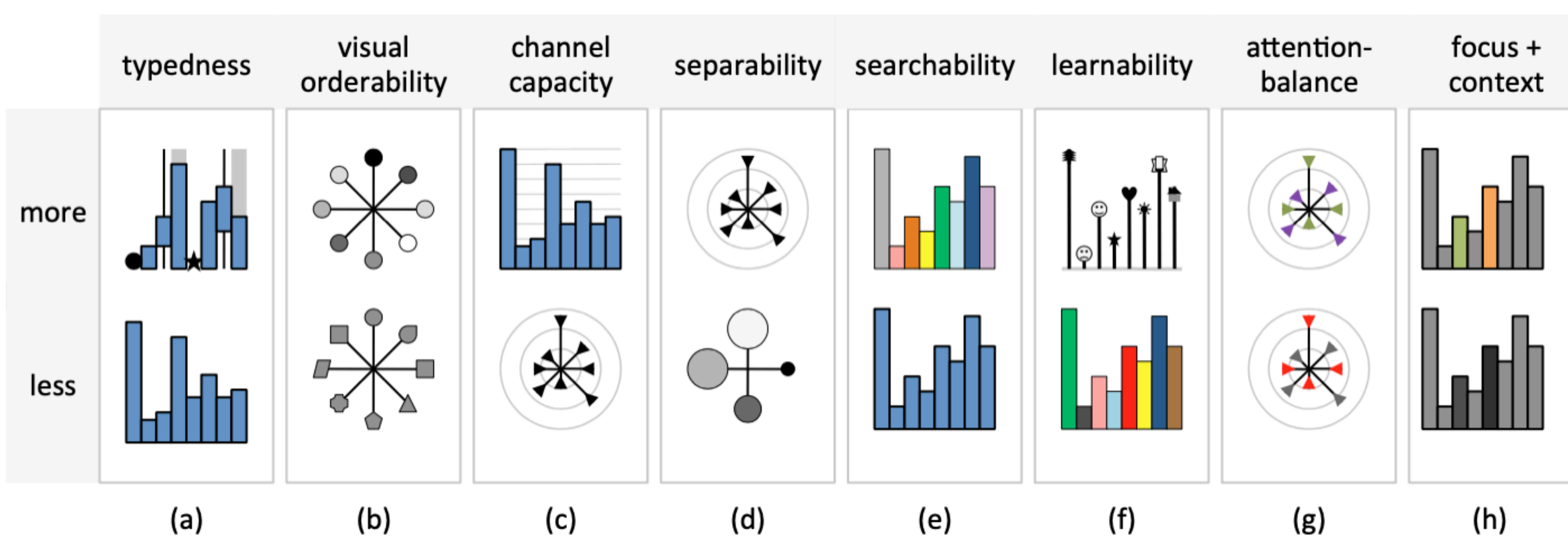


Figure 25: Variations of glyph design in accordance with the design principles of sortable glyph. **Search:** Chung et al., 2015.

- **Typedness.** Each variable in a multivariate dataset can be of a different data type. Typically, they are classified using the scale theory proposed by Stevens (1946): nominal, ordinal, interval, and ratio. As seen earlier, some visual channels are more appropriate for certain types of data. Typedness refers to the appropriate choice of visual channels according to the type of data to be represented (Chung et al., 2015). In Figure 25(a) it is clearer to determine the underlying data types for each variable in the top row glyph than in the bottom row.

- **Visual Orderability**. It is related to the ordering of data through the perception of visual channels, some visual features are naturally more related to quantitative natures, such as size and color intensity, while others, such as shape and texture, would be more difficult approaches to decode. Figure 25(b) shows two examples of glyphs representing 8 variables of the same data type. It is easier to visually order the 8 variables in the upper glyph than in the lower glyph.

- **Channel Capacity.** It refers to the number of values that can be encoded by a visual channel (Borgo et al., 2013). Chung et al. (2015) states that the capacity of a visual channel does not have absolute values, as the number depends on the size of a glyph, as well as many other perceptual factors, such as barely perceptible difference or interference from visual images close to objects. In Figure 25(c) the top glyph has a larger channel capacity, as each bar can visually encode more values than the radial lines in the glyph below. It will always be desirable to use a visual channel with greater capacity.

- **Separability.** When two or more visual channels are integrated into a composite channel, such as combining intensity, hue, and saturation in one color channel, interference between different primitive channels must be minimized. For example, in Figure 25(d), the glyph encodes 8 variables using 4 built-in channels. Each of the circles encodes two variables using the grayscale size and color intensity. That is, the perception of an individual channel is affected by another in an integrated coding, in addition, its ordering may require more cognitive load for a viewer to separate the perception of color intensity with the size

- **Searchability.** As glyphs encode high-dimensional multivariate data, it is necessary to help viewers to quickly search for a specific variable among many other variables (Chung et al., 2015). The search capability is related to using resources that facilitate the quick mapping of certain data attributes according to the need. This can be achieved through the use of resources such as color, as shown in Figure 25(e), or through strategies such as ordering the data. Briefly, searchability refers to the levels of ease when one needs to identify a specific visual channel within a glyph for a specific variable (Borgo et al., 2013).

- **Learning.** One of the weaknesses in glyph-based visualization is its need for learning (Fuchs et al, 2016), so this is an important criterion in many applications. As a type of iconographic representation, reading glyphs requires training, so direct interpretation is not possible, therefore, replacing the need for careful glyph designs to help viewers learn and memorize the association between variables and visual channels without constantly referring to subtitles. Ideally, a glyph design should be easy to learn and easy to remember. Ward (2008) and Maguire (2014) state that the use of metaphors for visual coding can favor the glyph in terms of learning, as it offers semantic associations. Many factors can affect a visual design in the learning context, for example, whether there are well-defined constructive rules, whether there are memorable metaphors, whether it is easy to guess, and so on. Figure 25(f) shows two different levels of learning, when, for example, it is necessary to code the number of greeting cards in different categories. The glyph in the top row is semantically rich and much easier to learn than the one in the bottom row.

- **Attention Balancing.** When a glyph is designed, even though there is a hierarchy of information and a key variable in focus, care must be taken to avoid unbalanced attention between different channels. A glyph should provide information that encourages comparison, in addition to the possibility of changing the focus glyph through interactive visualizations (Ward, 2008). Figure 25(g), for example, shows bright red indicators for some variables. When browsing different glyphs in the visualization, these red triangles are dominant, which can cause an unwanted pop-out effect. Ideally, the levels of attention should correspond to the levels of importance of the variables (Borgo et al. 2013).

- **Focus and Context.** Chung et. al. (2015) states that in multivariate visualization, it is often difficult, often undesirable, to predetermine the focus and context variables. A sort key is considered one of the focuses, the viewer may want to consider another variable as the focus. A focus plus context view refers to the ability to identify and highlight an individual visual channel under a given interactive operation. For example, when a user selects a certain variable as a classification key, it is desirable to highlight the corresponding visual channel so that it differs from other channels, allowing for a better comparative analysis; this is directly linked to the searchability principle.

In addition to these criteria, Chung et al. (2015) also comments on the importance of labels and captions. For example, axis labeling is an essential requirement for any sorting setup, as it allows the viewer to understand the context without referring to the view itself. Bertin (1969) refers to this as external identification. Captions convey the relationships between variables and visual channels and their representation for a given discrete or continuous value. This is often known as internal identification (Bertin, 1969).

The principles presented here are general guidelines to consider when designing glyphs to be sorted interactively in the visualization. However, the author states that they should not be treated as absolute laws. Some cases can lead to conflicting requirements when following some of these principles, or competing for the limited capacity of visual channels for smaller projects. In addition, Borgo et al. (2013) state that some other criteria also play an important role, such as aesthetic appearance.

Glyph Mapping and Appearance

Borg et. al. (2013) in their studies presented a series of design guidelines for the development of glyphs ranging from data mapping and choice of visual variables, which Borgo et al. (2013) calls glyph trimming, to rendering and arranging the glyphs on the screen. Within the guidelines proposed by the author, the following are considered:

- **Use of redundancy.** As presented earlier, it is possible to do a one-to-many mapping within the relationships between data and visual variables. This type of strategy is interesting to reinforce the information to be represented and should be used for data that are at the top of the hierarchy, according to the task to be performed from the visualization Ropinski et al. (2007).

- **Importance-based mapping.** This principle states that visual channels should be chosen based on the importance of the data to be represented, that is, using visual variables and representation strategies to highlight the most important information regarding the task to be performed. In addition, the mapping should guide the user's focus of attention, for example, using more prominent visual stimuli such as color, size, or opacity to encode relevance. An example is the study by Maguire et al (2012) who propose an algorithmic approach to importance-based mapping. The algorithm builds a taxonomy from a list of qualitative terms grouped into classification schemes. The higher a classification scheme is in the taxonomy, the stronger the visual channel for representing that scheme.

- **Simplicity and Symmetry:** Simple and symmetrical shapes facilitate the perception of visual patterns. In addition, simple glyph shapes improve the detection of small shape changes as well as outliers Ware (2004).

- **Orthogonality and Normalization:** When designing glyphs, it is especially important to consider how the different properties of the glyph interact with each other and thus possibly skew the interpretation of the data from the composition as a whole. A challenge in this context is the orthogonality of the different components of the glyph, which means that it must be possible to perceive each visual cue independently. In addition, the distortions introduced by the different properties of glyphs must be taken into account. When using, for example, glyph shape to represent a data variable, this also affects the area (size) of the glyph. Thus, such effects must be normalized to each other.

In addition to the guidelines presented above, Borgo et al. (2013) also mention intuitive mapping based on semantics, also known as natural mapping, stating that the semantics of the data must be incorporated into the glyph mapping.

Natural Mapping

Maguire et al. suggested the importance of establishing a metaphorical association between a visual channel and the concept or concepts to be codified. Because it is a type of iconographic representation, glyph-based visualizations rely on learning. Metaphorical visual representations allow domain-specific coding using "natural mapping". This natural mapping can make it easier for users to infer the meaning of the glyph and require less effort to learn and remember. According to Groh et al. (2018), the use of metaphors as a means of designing interactive systems is based on the idea that abstract structures in information technology do not have a "natural" shape, because data structures consist of relationships, dependencies, or proportions. The author also states that it is only through an image that data structures become visible, navigable, memorable and manipulable. Lakoff and Johnson (2008) refer that metaphors structure the common conceptual system of our culture, so we are able to take those concepts we know and apply them to new problems (Groh et al. 2018). This concept applied in data visualization can make navigation and analysis more intuitive.

The use of metaphors for glyph mapping and projection is widely found in the literature, for example, Groh et al. (2018) makes a study of glyph-based visualization using landscape metaphors to represent large datasets. Fuchs et al. (2015) used the metaphor of leaves to develop glyphs representing forest fires. This project stands out for being able to encode a large number of variations, where the design of the leaf glyphs is controlled by 12 categorical and 14 numerical parameters, adding up to a total of 26 variables (Fuchs et al., 2015) using attributes such as leaf shape, texture, orientation, and color. In addition, the glyphs are positioned in the plane, thus using the axes to represent additional information, in this case, they are temperature and humidity for the x and y axes respectively. Still, on the topic of plants, Stefaner (2014) uses glyphs in the form of flowers to visualize multidimensional data on the characteristics of the country. Each country is represented by a flower. The petals encode the different economic branches with varying sizes and lengths to the corresponding values. Flowers are arranged according to their weighted classification in all dimensions. People can change the layout by changing dimension weights or simply focusing on just one dimension.

2.3.3. Applications

Visualization based on glyphs, as it is a very versatile technique and, as seen, very flexible in terms of aesthetic representation, is used in the most diverse areas. For example, Kovacevic et al. (2020) uses a visualization project based on affective state demonstration. Ropinski et al. (2011) present an in-depth research on the use of glyph-based visualization for spatial multivariate medical data, showing that this type of visualization is already widely used within medicine. Legg et al. (2012) conducted a design study to show the effective use of glyph-based visualization in sports performance analysis. In addition to the existing types and categories of glyphs, it is also possible to classify the glyphs by their application area, below you can see some of the application areas for glyph-based visualization (Ward, 2008):

Medical Visualization

Recent research by Ropinski and Preim (2011) provides an overview of existing glyph-based visualization techniques used in the medical domain and proposes guidelines for developing more valuable glyph representations. A taxonomy of glyphs based on how information is processed when interpreting glyph views is used to classify such techniques. Within the semiotic theory, this consists of a two-phase information process: 1) pre-attentive processing, which is primarily driven by glyph attributes such as size, color, and shape, along with glyph positioning, texture mapping, and filtering of the glyph; 2) processes of attentive stimuli that are based on paradigms of the interaction of glyphs. Examples include a color legend that users can use to formulate more quantitative glyphs and reposition glyphs where the glyph's properties adapt depending on location. Based on this classification, the authors describe eight usage guidelines that compare against modern diffusion tensor imaging and cardiac visualization.

Flow Visualization

A big area for applying glyph-based visualizations is in the flow visualization community. An example is the work of Leeuw and Van Wijk (1993) who, using a probe, placed interactively by users, use glyphs to show the flow of a small region. From the use of arrows, the focus is to show velocity (represented by length and direction), curvature, shear, acceleration, torsion, and convergence.

The study by Martin et al. (2008) proposes to validate the effectiveness of traditional 2D hurricane visualizations by looking at users' ability to mentally integrate the magnitude and direction of flow into a vector field. In particular, the authors focus on evaluating 2D glyphs (or wind barb) - a technology commonly used to represent wind magnitude and direction in weather visualizations. For both magnitude and direction, users had to estimate the value at a given point and estimate the average value in a rectangular region. The authors used an actual hurricane simulation dataset in their study.

Event Visualization

Viewing events and activities is a rapidly growing research topic. A more thorough investigation of embedding visual semantics in glyph designs is explored by Legg et al. (2012). They describe MatchPad: an interactive glyph-based visualization for mapping events and actions in sports rating analysis. Sporting event analysis provides an example where a large number of event types need to be represented in order to facilitate rapid retrieval of information. A comprehensive review of mapping these data is discussed using different levels of abstraction. This includes evaluating abstract icons and colors for encoding each event type. While the approach may be suitable for data attributes with a small number of enumerative values, the range of categorical attributes in sports results in many different shapes or colors, making learning, remembering, or guessing cognitively challenging. Instead, the authors explore the use of metaphorical pictograms that are commonly used in many domain-specific visualizations (e.g. electronic circuit diagrams) and visualization for the masses (e.g. road signs). Metaphorical glyphs can take different forms, from abstract representation to photographic icons, where using appropriate visual channels can provide semantic clues that are easy to learn, remember, or guess. MatchPad adopts a scale-adaptive layout to interactively position glyphs along a timeline based on the viewpoint's zoom factor. This minimizes glyph occlusion which they successfully demonstrate using a case study in Rugby.

Multi-Field Visualization

Due to their multivariate characteristics, geometric shapes are commonly used to represent multiple data attributes. Superquadrics and Barr Angles Preservation Transformations present such an approach by introducing geometric shapes (superquadrics) used to create and simulate three-dimensional scenes. The author defines a mathematical framework used to explicitly define a family of geometric primitives from which their position, size, and surface curvature can be changed by modifying a family of different parameters. Examples of glyphs include a torus, star-shaped, ellipsoid, hyperboloid, and toroid. Additionally, the author describes angle-preserving shape transformations that can be applied to primitives to create geometric effects such as bending or twisting.

Another interesting study is that of Taylor (2002), which provides an overview of successful and unsuccessful techniques for visualizing various scalar fields in a 2D array. The author first hypothesizes that the greatest number of datasets can be displayed by mapping each field to the following: a unique surface feature, applying a different visualization technique to each scalar field, or using textures/glyphs whose features depend on the data sets. Such a framework revealed limitations of up to four scalar fields. This led to the research of two new techniques that proved effectiveness for visualizing multiple scalar fields, (1) data-driven points (DDS) - using different points of various intensities and heights to visualize each dataset, and (2) oriented strips - using strip-like texture glyphs of different orientations to visualize various scalar fields in which luminance is mapped to relative scalar values.

Geo-Spatial Visualization

We often find that geospatial visualization can incorporate cross-disciplinary techniques from other domains and therefore can be classified into more than one category. MacEachren et al. (1992) is an example where the authors present a new approach to visualize reliability in mortality maps using a bivariate mapping. Given a basic geographic map (United States), the technique involves using color-filled regions to represent the data and overlaying textures to represent reliability. Another example is the work of Sanyal et al. (2010) which introduces the use of glyphs, ribbons, and spaghetti charts to interactively visualize ensemble uncertainty in numerical meteorological models. They demonstrate their work on the 1933 Superstorm simulation, where visual mappings illustrate statistical errors (e.g., mean, standard deviation, interquartile range, and 95% confidence intervals) in the data.

Tensor Visualization

The work by Laidlaw et al.(1998) introduces two new methods for viewing Diffusion Tensor (DTI) Images. The first method uses normalized ellipsoids, where the principal axes and radii are mapped to the tensor eigenvectors and eigenvalues respectively. Glyph normalization reduces visual clutter and allows for a complete representation of the dataset. The second method incorporates oil painting concepts to represent seven attributes of tensor data as multiple layers of varying brushstrokes that are composited into a single view. The authors demonstrate their technique on spinal cord DTIs from healthy and diseased mice. The study by Kriz et al. (2005) provides a review of visualization techniques on second-order tensors including: Lame stress ellipsoids, Haber glyphs, Reynolds tensor glyphs, and hyperflow tubes. In addition, the authors introduce a Principal, Normal, and Shear (PNS) glyph to visualize stress tensors and their gradients. The method extends the stress ellipsoids by mapping the shear stress component to the surface color of the ellipsoid.

Kindlmann extends his previous work (2006) to glyph-packing, a new glyph placement strategy. This work aims to improve the discrete nature of glyph-based visualization through the use of regular grid sampling, to a more continuous character, such as texture-based methods, and packaging of the glyphs in the field. A potential energy based on tensors is defined to derive the placement of a system of particles whose final positions will be used to place glyphs. More recently, Schultz and Kindlmann (2010) have introduced superquadric glyphs that can be used to visualize general symmetric second-order tensors that can be non-positive defined. The work extends previous glyph-based methods dealing with tensors with strictly positive eigenvalues, such as diffusion tensors, to the general case, mapping the shape of the glyph to show differences in eigenvalue sign.

Uncertainty Visualization

In addition to the possible applications presented, glyph-based visualizations can be used to quantify and visualize uncertainties. Reasoning with uncertainty is one-sidedly difficult, but researchers are revealing how some types of visualizations can improve decision-making in a variety of diverse contexts, from hazard prediction to health communication, to everyday decisions about traffic. In particular, glyphs are well suited to illustrating uncertainty as they usually clearly show trends and patterns (Borgo et al., 2013).

An example of application is the study by MacEachren et al. (2012) who carried out empirical research evaluating the effects of visualizing different categories of uncertainty using discrete symbols. Using Bertin's Graphic Semiology (1969), they provide information on the effects of using abstract symbols that vary only a single visual variable compared to iconic symbols that are of a more pictorial form. Both sets of symbols underwent two separate experiments that focus on assessing their intuitiveness to represent different categories of uncertainty and effectiveness for a typical map-using task: assessing and comparing aggregate uncertainty across two map regions.

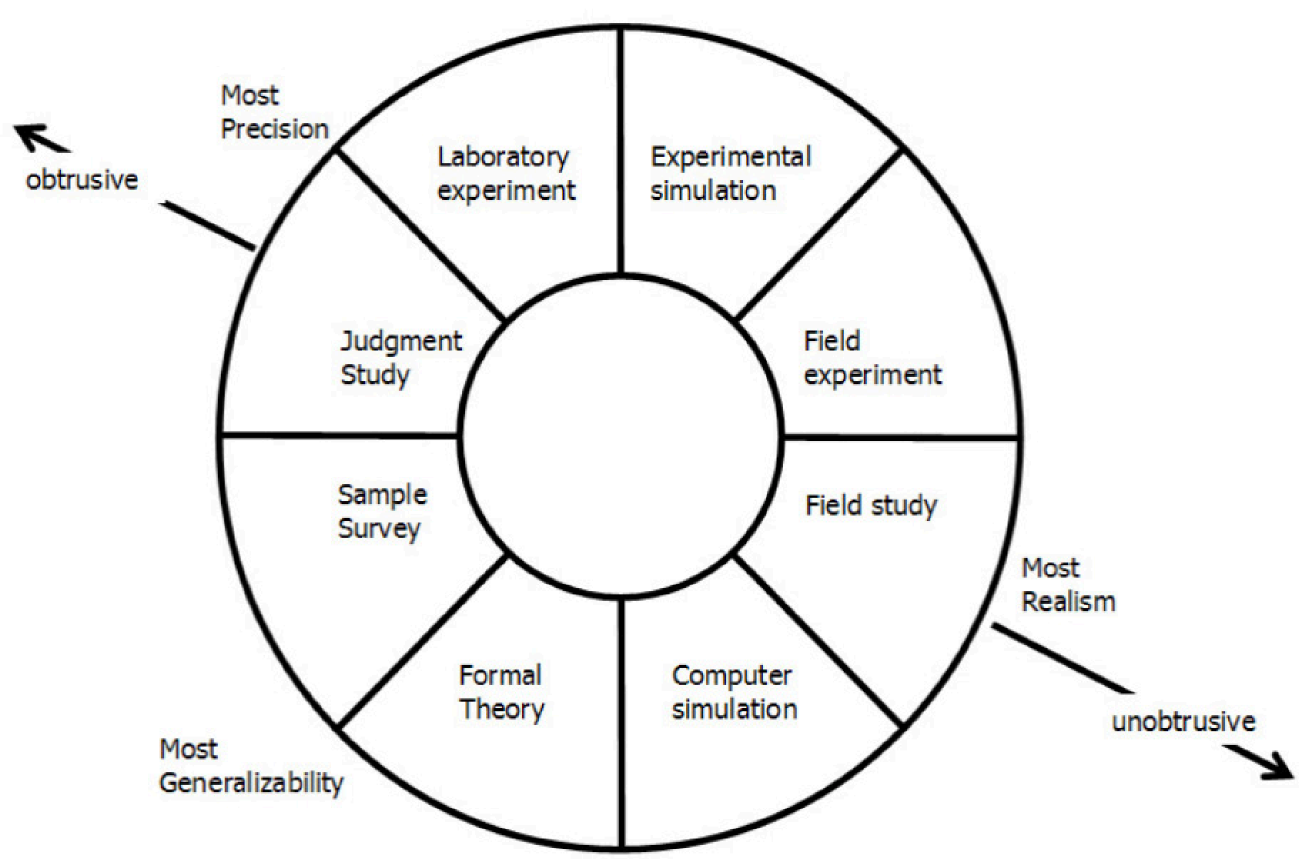
2.3.4. Conclusions

Glyphs are a powerful alternative to visualizations, being able to assume the most diverse forms and be applied in the most diverse areas, offering versatility and countless possibilities of creation. In this chapter, a more in-depth conceptualization of glyphs was presented, as well as some of the most present glyphs in the literature, although there is a catalog of existing types of glyphs (Ward, 2008), due to their iconographic characteristic, new versions appear more and more every day. Due to its great flexibility of creation, some care must be taken when designing a glyph. In this chapter criteria and guidelines for a glyph, and projection were shown, ranging from data mapping and choice of visual variables to rendering. Hence, in addition to other mapping strategies, the use of semantic associations causing the so-called "natural mapping" can significantly favor visualizations regarding the difficulty of learning and memory of the glyph. Finally, some of the different areas that use glyph-based visualizations were presented.

2.4. Evaluation

One of the biggest challenges within information visualization is its validation. This is because in addition to the general assessment challenges of choosing assessment questions, methods, and executing them correctly, within visualizations we are dealing with empirical research. Carpendale (2008) states that in empirical research it is difficult to choose the right focus and questions, making the analysis process and results, such as specific insights and increasing global understanding, difficult to capture and quantify (Lam et al., 2011). In information visualization, there are a wide variety of cognitive reasoning tasks that vary according to the type and characteristics of the data represented. These tasks can be divided into two types: Low-level tasks, such as comparing, contrasting, associating, distinguishing, and classifying data and; High-Level tasks which include activities such as understanding data trends, uncertainties, chance relationships, or predictions. For this reason, it is important to assess the adequacy of representational coding and the readability of visual resources (Carpendale, 2008).

Figure 26 : Types of empirical research methodologies.
Search: Carpendale, 2008.



McGrath (1995) states that within evaluations there are three desirable factors: (i) generalization, which refers to the result that can be applied to people who are not directly linked to the study, which can be extended to other situations; (ii) precision, a result is precise insofar as it can be definitive about the measures that were taken and about the control of the factors that it was not intended to study; and (iii) realism, a result is considered realistic when the context in which it was studied is similar to the context in which it will be used. Ideally, the three factors should occur simultaneously, but existing methodologies are limited and do not support all, favoring only one or supporting up to two simultaneously (Carpendale, 2008).

When it comes to evaluating the effectiveness of glyphs or visualizations in general, there is a variety of evaluation methods available, all of which have their advantages and disadvantages. The appropriate evaluation method should be chosen according to the purpose of the project and evaluation, the time, place, and available participants. Carpendale (2008) mapped different types of empirical research that can be done when the objective is to evaluate and validate visualizations:

- **Field study:** Very common in anthropological research of ethnographic work, this type of study takes place within a real situation and the observer must be as discreet as possible not interfering in the process. According to Carpendale (2008), this type of study has a high level of realism but is not particularly accurate or generalizable.

- **Field experiment:** A field experiment is usually also carried out in a realistic setting; however, an experimenter trades some degree of discretion to obtain more precision in observations. For example, the experimenter may ask participants to perform a specific task while the experimenter is present.

- **Laboratory Experiment:** Opposite of what occurs in the field study. Within this format, the entire scenario is projected. In a controlled environment, it is established how the study will be conducted, what tasks the participants will do, and thus plan the entire study procedure. Then the experimenter makes people participate as fully as possible, following the rules of procedure within the defined situation. Despite lacking realism, this type of study can achieve high levels of precision.

- **Experimental Simulation:** With an experimental simulation, the experimenter tries to maintain as much precision as possible while introducing some realism via simulation. This type of approach is very common in tests that present some kind of danger if applied in a real environment, such as studying driving using a cell phone or under the influence of some substance using a driving simulator. The use of simulation can avoid risky or unethical situations.

- **Judgmental Study:** In a judgmental study, the goal is to piece together a person's response to a set of stimuli in a situation where the setting becomes irrelevant. Much attention is paid to creating "neutral conditions". Ideally, the environment would not affect the outcome. Perceptual studies often use this approach.

- **Sample Survey:** In a sample survey, the experimenter is interested in discovering relationships between a set of variables in a given population. Examples of these types of questions include: of people who discover information visualization tools on the web, how many return frequently, and are their activities social or work? Of the people who have information visualization software available at work, how often do they use it? Considering the increase in examples of information visualization results and software on the web, is there increasing awareness and/or use of information visualization by the general population? In these types of studies, adequate sampling of the population can lead to considerable generalization. However, while careful population selection is extremely important, it is often difficult to control (Carpendale, 2008). For example, in a web-based survey, all the responses returned are those types of people who are willing to take the time, to fill out the questionnaire, etc.

- **Formal theory:** Formal theory is not a separate experimental methodology, but an important aspect of all empirical research that can easily be ignored. As such, it does not involve gathering new empirical evidence and, as a result, is low in accuracy and realism. Here, existing empirical evidence is examined to consider theoretical implications. For example, the results of multiple studies can be considered as a whole to provide a higher-level understanding or goal, or the results can be considered in light of existing theories to expand, adjust, or refute them.

- **Computer Simulation:** It is also possible to make a fully realized study through a computer simulation that has been designed as logically complete. This method is used in battle simulation, search and rescue simulation, etc. This type of strategy can be used to evaluate some views.

Lam et al. (2011) states that although there are several approaches for empirical studies, the determination of an approach that can improve existing practices within information visualization remains undefined. The author also states that for experimenters, part of the problem is a large amount of evaluation methodologies in use. Information visualization draws from diverse disciplines such as psychophysics, social sciences, statistics, and computer science, using methodologies as diverse as laboratory-based factor design studies, field evaluations, statistical data analysis, and automatic image evaluation. To mitigate the problem, Lam et al (2011) presents a study in which seven assessment scenarios that are most commonly encountered are systematically identified and suggestions for appropriate assessment methods for each scenario are presented: assessment of environments and work practices; evaluating visual data analysis and reasoning; evaluating communication through visualization; evaluating collaborative data analysis; evaluating user performance; evaluating the user experience; and automated evaluation of views.

User performance evaluations, according to studies by Lam et al. (2011), are predominantly measured in terms of objectively measurable metrics, such as time and error rate. However, it is also possible to measure subjective performance, such as the quality of the service. The most commonly used metrics are task completion time and task accuracy. The outputs are usually numerical values analyzed using descriptive statistics (such as mean, median, standard deviations, and confidence intervals). Lam et al (2011) states that this type of evaluation needs high precision methods, so that they can result in quantitative and statistically significant data. For this, the most effective approach is the laboratory experiment, within this type of evaluation s methodologies most commonly used involves an experimental design with only a small number of variables changed between experiment conditions, so the full impact of such variables can be measured. A controlled experiment usually requires abstracting real-life tasks to simple tasks that can be performed by large numbers of participants repeatedly in each study session. Due to the need for a relatively large number of participants, researchers often need to recruit non-specialists.

When it comes to the validation of glyph-based visualizations, Maguire (2015) states that, depending on the domain in which the visualization will be applied, it is difficult to get enough attention from a domain expert to carry out a meaningful assessment. Therefore, it is difficult to fully assess the memorability of glyphs, distinguishability, interpretability, and usefulness of a set of glyphs. One of the main challenges is finding ways other than user interaction to measure typical evaluation metrics.

An example of evaluation within glyph-based visualizations using laboratory experiments was the study by Fuchs et al. (2013), where they proposed a controlled experiment to investigate the performance of four different glyph designs (i.e., Line Glyph, Stripe Glyph, Clock Glyph, and the Star Glyph) that use different combinations of visual variables to encode two properties of temporal data in a small multiple scenario. The experiment used a small group of participants (24) within a controlled environment, a silent room with only the participant and the experience. During the evaluations, small groups of glyphs were presented and the participant was asked to perform tasks such as, peak detection, temporal location, and trend detection. The results showed that glyphs perform differently depending on the type of task. Line glyphs turned out to be a good choice for peak and trend detection tasks, but radial encodings are more effective for reading values at specific temporal locations.

3. Methodology

About methods and processes for developing visualizations, Munzner (2014) proposed a waterfall structure, which allows the designer to analyze and validate the visualization problem at four different levels. The first level is classified as the problem identification level, where the target audience and objectives of the visualization are defined. Then, at the level of abstraction, user problems and tasks are mapped to available data in domain-independent forms. The third level refers to visual coding and interaction, where the fundamentals of visualization are applied to represent the data. Finally, the algorithm level is related to the visualization implementation.

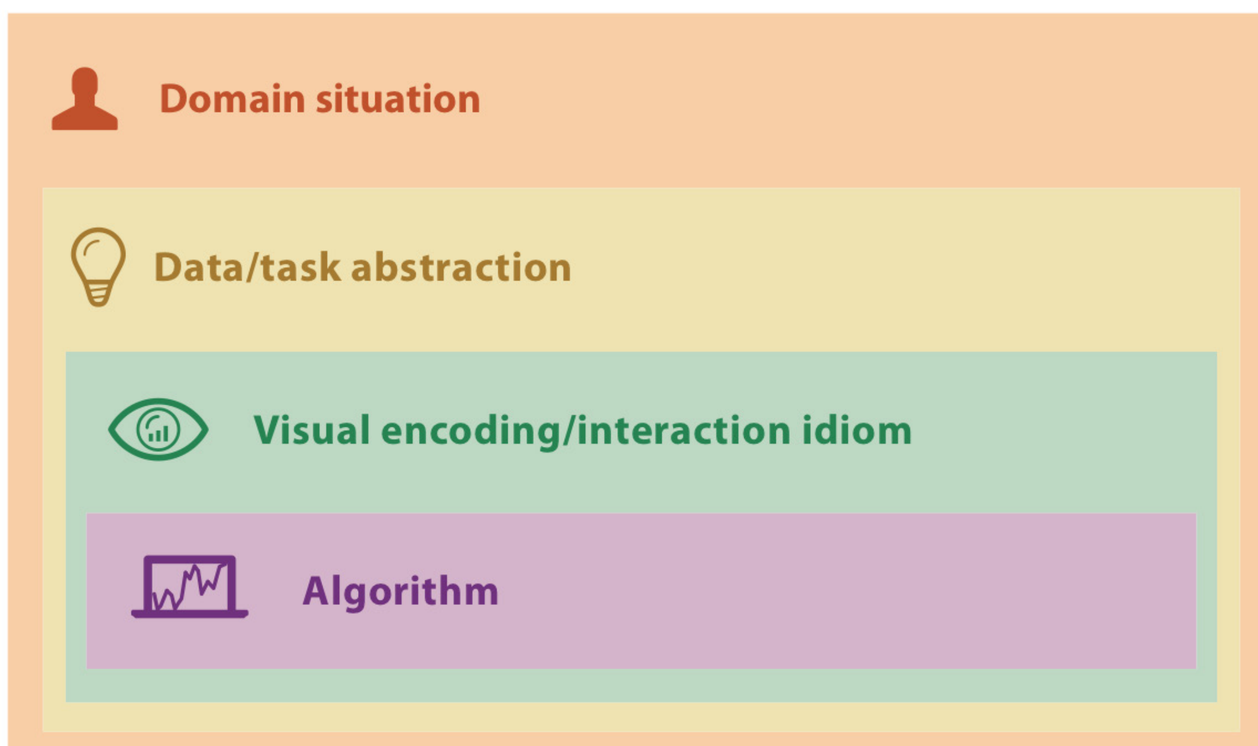


Figure 27: The four nested levels of vis design. **Search:** Munzner, 2014, p.68

With regard to glyph design, Maguire (2015) states that until now it was a largely ad-hoc design process, fueled by the tasks a user wants to perform and performed by a designer. They are based almost entirely on a sample of the raw data, some field data, and intuition. The author also states that this type of approach, despite working in simple cases, can present problems such as: (i) data mapped through incorrect visual channels; (ii) data values mapped to too many channels, or excessive use of a visual channel; (iii) important information being hidden due to details not visible in a job resolution; and (iv) difficulty in interpreting the glyphs by the target audience. To mitigate the problem and provide a model for the systematic creation of glyphs, Maguire (2015) proposes a process for glyph design based on Munzner's nested model (2014). The process consists of seven steps, ranging from familiarization with the content covered in the visualization to the evaluation of the final glyphs.

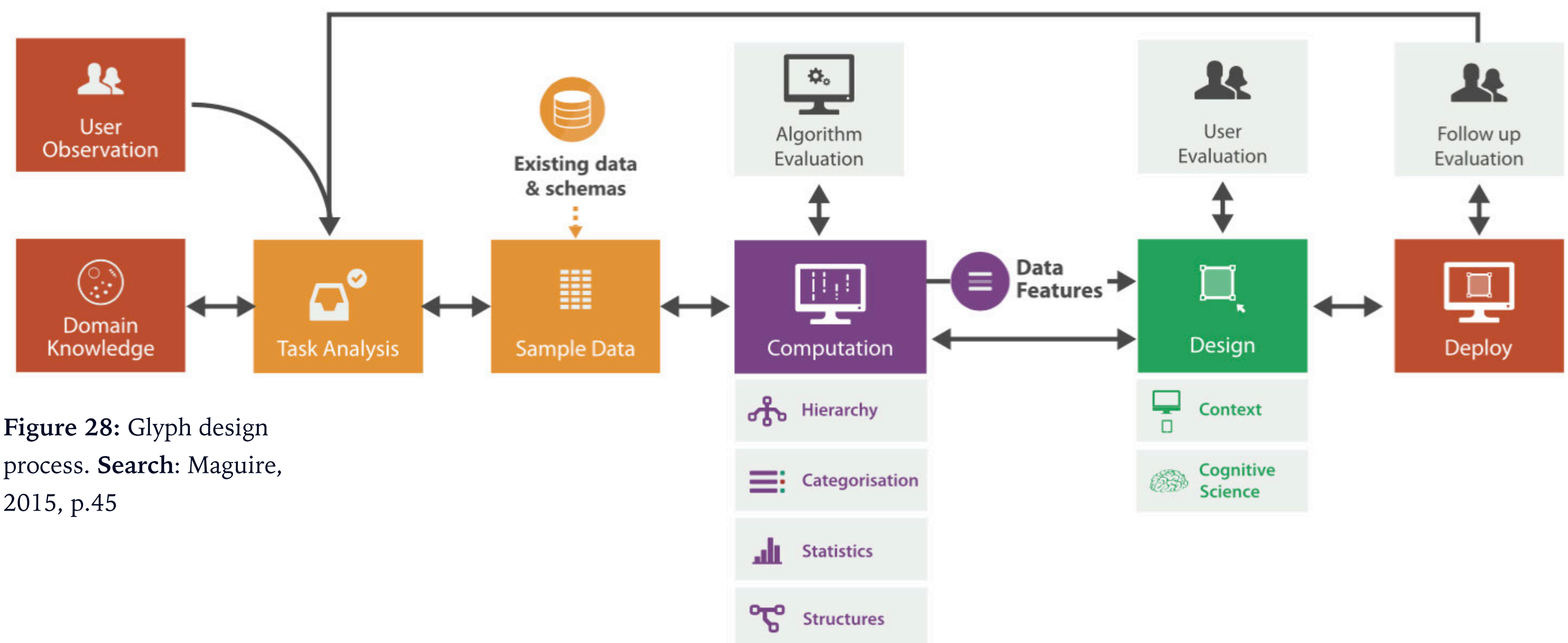


Figure 28: Glyph design process. **Search**: Maguire, 2015, p.45

User Observation

The first step of the glyph design process, user observation, is aimed at a better understanding of visualization needs, user monitoring should be done in their normal working environment. Maguire (2015) states that the knowledge acquired in this type of study can feed: (i) a definition of tasks through interviews and day-to-day observations of current work practices; (ii) designing processes through information about what a user is willing to learn and to remember; and (iii) evaluation to determine whether or not the solution meets the user's expectations and solves the problem for which the glyph is intended. Due to limitations in the project, a user observation step is not presented, but project requirements received through meetings with Evollu are presented.

Domain Knowledge

This refers to the familiarization of the visualization domain, visualizations can be related to the most diverse areas and represent data from the most diverse fields. The designer must have some level of knowledge on the subject so that the choices in terms of visual representation are effective. Maguire (2015) states that a good understanding of the domain will often result in better visualizations that actually solve the problems faced by users. Furthermore, the author states that domain knowledge has additional importance in glyph-based visualization due to the common need for metaphors to support glyph learning and memorization ability.

Task Analysis

Task Analysis, also called data/task abstraction in Munzer's (2014) model, maps the types of visual queries that a user would like to do. These tasks form the basis of data prioritization because it is from the main tasks performed by users that we identify which data should be first in the visualization hierarchy, that is, which should stand out visually. For example, if an important task performed by a user is to find users of a certain age, it is important to make age information more visually available;

Sample Data

Sample data provides an indication of the data values (and ranges) to be encoded by the glyphs. A data representation schema can provide additional information about the format of the sample data and the types of values (eg, categorical or quantitative) that each field contains (Maguire, 2015). This, in turn, can be used in the design phase, where the design principles discussed below can be applied to ensure that the appropriate visual channels are used for specific data values;

Computation

Computing can be deployed to provide additional information about the data (derived data). Maguire (2015) identified four techniques that can be applied individually or combinatorially to aid in the glyph design process:

- **Structure identification** refers to discovering patterns in relational data, for example, connectivity patterns. This information can be used to reduce complexity in network views by removing common or uninteresting patterns, leaving the most important signals behind;

- **Categorizations** provide a way to split data based on multiple categorizations. These schemes can be computationally determined using algorithms such as k-means or hierarchical clustering. Another alternative is to use a more supervised approach to clustering with the help of computation, a large representative dataset, and domain experts in the loop to ensure that categorizations are meaningful and not too granular;

- **Statistics**, the use of statistical analysis is useful as they provide information about the distribution of data, this can feed into the categorization process, for example. Statistical processes can also help identify common/rare data features to encode using glyphs to draw attention to these features in a visualization;

- **Hierarchical Organization** refers to the creation of a taxonomy, a tree representation that hierarchically defines the properties of the items in the tree. This hierarchical organization can be used to structure how a glyph is organized. The imposition of an order on the design should facilitate the best learning and memorization ability of the resulting glyphs since the formation of each glyph must follow a rule.

Design

Stage of the process that involves the selection of visual channels to represent each data variable to be encoded and then the composition of the glyph as a whole, that is, the arrangement of these visual channels within the space. Maguire (2015) reaffirms principles that must be considered when designing glyphs:

- Visual Channel Suitability, related to the typedness presented above, is the correct mapping of the visual channels according to the type of data being worked on.
- Natural mappings suggest the use of semantic relationships to associate the represented data with the visual channel. Maguire (2015) states that natural mapping should occur whenever possible. Using too many colors or abstract shapes, for example, would put too much cognitive load on a user to remember all the mappings. Make those colors or shapes metaphorical, however, the process of decoding glyphs will be much easier.
- Visual hierarchy applies an organization to the visual channels within a glyph, given the importance of data for some tasks. This ensures that the most important data classifiers are available even at the overview level of visualization due to their visual prominence (visual hierarchy and global/local processing and pre-attentive processing);
- Visual channel composition suggests avoiding visual channels that are perceived holistically as opposed to separately. The completeness or separability of a pair of visual channels is not a discrete decision. While it is widely accepted that width and height are the most integral dimensions, while position and color are the most separable, many other dimensions are on a continuum between fully integral and fully separable;
- Visual channel perceptual limitations provide guidance to avoid overusing a visual channel, such as color, whose perception degrades as the number of samples of each hue increases.

- Redundant encoding suggests using multiple encodings for a data item to minimize the chance of error when reading glyphs from a display. The use of redundancy is interesting for emphatic data that is at the top of the hierarchy according to the tasks, or data that is more complex to be visually encoded. Redundant coding can also speed up the visual search.

Evaluation

The assessment provides some level of validation that the glyph design is a success. As seen before, there are several ways to apply validation processes and, as this is empirical research, this step can be complex and difficult. Glyphs can be validated by reviewing previously defined tasks, and even simulating the execution of these tasks. In addition, glyphs can be assessed for distinction, memorization, and learning. That is, it will be evaluated whether or not users can detect differences between the glyphs; if the user can easily learn how the relationships between the glyphs work, it is important to emphasize that, as it is a type of iconographic representation, it requires learning and must be evaluated; and whether the visual attributes used are easy to remember. Maguire (2015) states that if a glyph design is not executed effectively, there will be more iterations in the design to improve performance. Once a project is approved, the glyphs can be deployed to the domain experts' daily work environment. Follow-up studies can be performed over time, and more iterations may be required.

4. Glyphs Development

In summary, the stages of glyph development start with the identification of the problem where a context is presented about the data that will be used for the development of the glyphs. The second stage is the identification of tasks together with the step of familiarization with the dataset, the data cut is presented, that is, what information will be used for the prototype and its function. In the Computation stage, the result of an exploratory analysis performed on the data is presented together with a presentation of the result of an algorithm application for data classification. In the design phase, some developed alternatives and a final sketch of the glyphs are presented. Finally, the process used to evaluate the developed glyphs is presented.

4.1. Evollu and Hearing Loss

Evollu is a technology startup that operates in the field of audiology. Its products aim to facilitate the diagnosis of hearing levels. The company offers solutions both for the general population and for specialists in the area, including pharmacies and specialized offices. Hearing loss is a serious public health problem that can have physical, mental, and social consequences for those affected, as well as for their closest family members and/or people they relate to (Mendes, 2020). When related to age, hearing loss becomes a global public health problem. The World Health Organization reports that approximately 5% of the world's population is hearing impaired, making it the third most common chronic condition (World Health Organization [WHO], 2021). Hypoacusis, whether gradual or sudden, mild or profound, congenital or acquired in adulthood or old age, can have significant effects on communication skills and quality of life (Pratt, 2010).

Hearing Levels

When it comes to hearing levels, a normal hearing human being is able to hear sounds of frequencies between 20 and 20,000 Hz, with the ear being particularly sensitive and responding to a range of sound intensity (Mendes, 2020). The loudest sound that can be heard without being confused with pain is at a level of about 120 dB above the faintest sound it detects. Sounds above 75 dB are considered potentially dangerous. Examples of noise levels considered harmful by experts include lawn mowers, rock concerts, firearms, stereos with insertion sources, motorcycles, tractors, household items (garbage shredders, blenders, food processors/ cutters, etc.), and some noisy toys. All can emit sound above 90 dB and some can reach up to 140 dB, in table 1 you can see the different types of noise and risk levels for hearing health (World Health Organization [WHO], 2021). Any injury that occurs in one or several parts of the auditory system can compromise its function in providing sound information to the brain and the individual can present a partial or total hearing deficit (Raminhos, 2910).

Table 1: Intensity in dB of different types of sound. **Search**: World Health Organization [WHO].

Level	Intensity	Type of sound	Permissible Exposure Time
Painful impulse noise	150 dBP	fireworks at 3 feet, firecracker, shotgun	Not safe
	140 dBP	firearms	Not safe
Painful steady noise	130 dBA	jackhammer	Not safe
	120 dBA	jet plane takeoff, siren, pneumatic drill	Not safe
Extremely loud	112 dBA	rock concert, chainsaw	1 minute
	106 dBA	gas leaf blower, snow blower	3.75 minutes
	100 dBA	tractor, listening with earphones	15 minutes
	94 dBA	hair dryer, kitchen blender, food processor	1 hour
Very loud	91 dBA	subway, passing motorcycle, gas mower	2 hours
Moderate	70 dBA	group conversation, vacuum cleaner	Safe
	60 dBA	typical conversation, dishwasher	Safe
	50 dBA	moderate rainfall	Safe
	40 dBA	quiet room	Safe
Faint	30 dBA	whisper, quiet library	Safe

Regarding hearing loss, according to the World Health Organization (2021), above 35 dB is already considered a hearing loss, and losses caused by noise exposure or aging are irreversible. According to the Bureau International d'Audio Phonologie (1996), the determination of the degree of hearing loss is carried out through the average of the hearing thresholds of the frequencies, obtaining the following degrees.

Table 2: Categorization of hearing loss levels. **Search**: BIAP, 1996.

Levels of hearing loss	Hearing loss range (dB HL)
Normal	≤ 20 dB
Mild	21 to 40 dB.
Moderate	- 1st degree: 41 to 55 dB. - 2nd degree: 56 to 70 dB.
Severe	- 1st degree: 71 to 80 dB. - 2nd degree: 81 to 90 dB
Very Severe	- 1st degree: 91 to 100 dB. - 2nd degree: 101 to 110 dB. - 3rd degree: 111 to 119 dB.
Cophosis	≥ 120 dB

Clinical Evaluation

To carry out the hearing assessment, a medical consultation is required, where additional diagnostic tests are requested, such as a tympanogram, and a tonal and vocal audiogram. Any examination is performed by a Senior Diagnostic and Therapeutic Technician, or an audiologist. According to Mendes (2020), the assessment begins with anamnesis, to know the difficulties perceived by the client, their motivation, clinical and otological, hereditary, and noise exposure history. Subsequently, otoscopy is performed to observe the external auditory canal and the tympanic membrane. This observation is of the utmost importance to verify the presence of cerumen, because if there is a large excess, it jeopardizes the reliability of the audiological assessment, the presence of foreign bodies or exostosis is also analyzed. Then, the psychoacoustic tests are performed: pure tone audiogram, vocal audiogram, and discomfort threshold survey.

A pure tone audiometry exam shows a loss in dB, compared to the normal hearing level, with reference to ISO standards. An average pitch loss is calculated, taking the loss in dB at various frequencies as a starting point. Frequencies are measured in cycles per second (Hertz-Hz) and define the pitch of the sound being heard, bass sounds range from 50 to 60 Hz and high sounds can reach 10,000 Hz or more. Audiometry tests usually evaluate frequencies of 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. A frequency that is not perceived is considered a loss of 120 dB. The total amount is calculated, divided by 4, and rounded to the nearest unit. In the case of an asymmetric hearing loss greater than 15 dB, the average loss level, expressed in dB, is multiplied by 7 for the "good" ear and by 3 for the "bad" ear. The total is then divided by 10.

The Audiologist, when evaluating the audiological tests, must take into account whether the presented hearing loss can be corrected clinically or surgically. If there is no possible surgical correction, the auditory rehabilitation process is started. The data offered by Evollu for the development of the project portray a series of audiological tests carried out, through partners, in the population of Portugal. The data include information such as the average in decibels about the hearing loss of each user, as well as the value in dB for each of the frequencies evaluated, in addition, some demographic data such as the age of the user and the region in which the test was performed are available.

4.2. Data Abstraction

When analyzing the database, it can be seen that it is composed of 36 attributes that generally collect all the characteristics of the ear tests performed by Evollu. It is possible to identify at least five sections in relation to the information portrayed. The first is data related to the profile of the user who is performing the test, the second is the results of the test itself, a third group talks about the characteristics of the devices used to collect the test, and the fourth is identification data of the partner who is undergoing the test, and a fifth talks about the post-test, that is, when the person is contacted to purchase the device. From meetings and goal alignments made together with Evollu, it was identified that the main task within the visualization is to analyze the different types of users (through information such as age, region, and test results) and what is the current status. of the user within the service process, this information is shown through the status attribute in the dataset. Given the project requirements, the design of glyphs focuses on the representation of the data present in the first, second, and fifth groups, that is, information about the user profile, the test result, and the post-test state.

Status

Status is an attribute available in the dataset that includes all the contact between the company and the user after the test has been carried out, divided into a large number of numerical codes that represent the source of the lead, if the user has already been contacted and what the result of the contact was until the final moment of the purchase. For a better visualization of the statuses, and also to facilitate the moment of visual coding in the stage of designing the glyphs, the status codes were grouped into 8. They are: contacted successfully, contact failed, appointment scheduled, missed appointment, service performed, purchase made, purchase not made, and finally, no test with interest in the experience. You can check the different grouped statuses and the original codes in table 3. Other statuses that were not listed before the first contact were ignored.

A perceived problem within the attribute is that its conditions are exclusive: for example, for a user who has the code referring to the purchase made, it is not possible to know how the process of the first contact was, or if the consultation was at home or not. There is no status history, it is only possible to analyze the present state. Although, because there is no history of the status, it is not possible to analyze what leads the user to make the purchase, the status attribute is one of the most important within the information hierarchy, as it presents the user's situation at the moment assisting the user in the decision making of the next steps.

Table 3: Status attribute grouped

Status 1	Status 2	Status 3	Status 4
3 - Do not want to be contacted **10** - Unsuccessful contact **13** - Contacted successfully without interest in service	**12** - Contact postponed **11** - Successful contact	**100** - Appointment scheduled	**105** - Missed the appointment
Status 5	**Status 6**	**Status 7**	**Status 8**
99 - Without test, with interest in experience	**101** - Service performed, with normal hearing **109** - Service performed **103** - Service performed, with hearing loss will decide	**102** - Service performed, with hearing loss, not interested	**110** - Sale Made

Age

The age information of the users who took the test is obtained from the date of birth. In the dataset, the values range from 1989 to 2021, that is, it contains data from people over 90 years old to children under one year old. In addition to the removal of outliers, the age attribute was grouped in intervals of 10 in 10 years. This will make a better analysis of the data and visualization of patterns and trends. However, the largest number of tests are carried out among people between 80 and 60 years old, so in one of the versions of the glyph, which will be presented later, the group of users over 40 years old was used.

City, Lat and Long

In the dataset provided by Evollu, a city attribute is presented, where information about the user's place of address would be, but in most of the data this attribute was filled in by the name of Evolu's partner, mostly names of pharmacies, so the data was used latitude and longitude for the location where the test was performed. The same is grouped in the regions of Portugal: North, Center, Lisbon, Alentejo, and Algarve.

It is important to note that as the latitude and longitude data collected from the test device were used, it includes the location of Evollu's partner, and not necessarily the user's place of residence.

Hearing Loss

Within the data, the arithmetic means of each side are presented, which has a variation in value, from 0 to 108, it was decided to classify this varies according to the hearing loss scale proposed by the Bureau International d'Áudio Phonologie (1996), so the data were grouped into 5 levels of hearing loss: Normal, Mild, Moderate, Severe and Very Severe. The arithmetic mean values, referring to hearing loss, are related to each ear, that is, one side may have a different classification from the other.

Frequencies

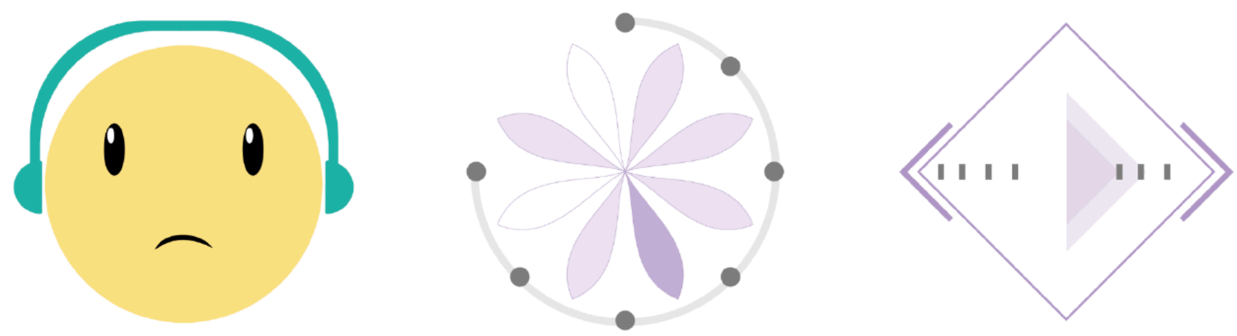
In the data, we have the attribute "result" which is a set of 8 float values, meaning the frequencies 1000 dB, 2000 dB, 4000 dB, and 6000 dB for the right and left ears. As in the hearing loss averages, the value of this attribute varies between 0 and 108. Following the pattern of the data, this attribute was also grouped according to the Bureau International d'Áudio Phonologie (1996), presenting 5 different levels of loss for each ear.

Although in terms of importance, it is below the data such as hearing average, status, age, and region, the information contained within the losses of each frequency can be useful because it allows mapping the type of hearing loss of the user.

4.3. Design

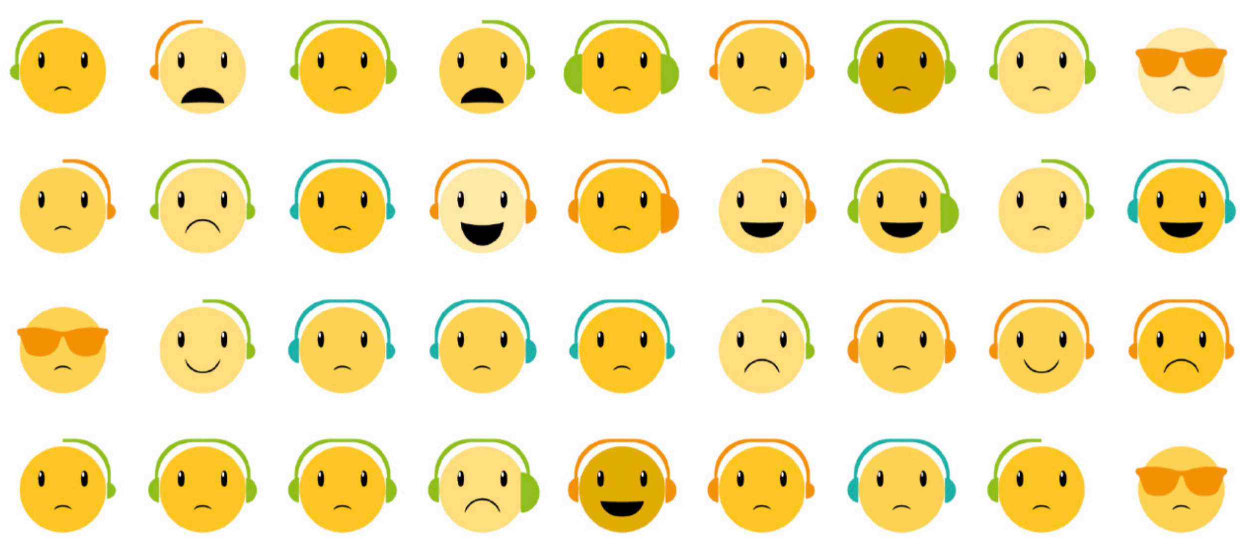
For the development of the glyphs, we chose to make three versions of the design, going from a more figurative representation, using semantic resources for a natural mapping, to a more abstract version that uses only geometric shapes in its composition (Figure 28). A version considered neutral was also developed, so that the evaluation tests can be carried out.

Figure 29: Types of projected glyphs.



Glyph 1 - Emojis

Figure 30: Different types of emoji glyphs



The first glyph uses the faces approach for its visual representation. Faces in emoji format were designed to use semantic relationships for the representation of hearing loss. In addition, the use of faces was also interesting for the representation of different statuses. First, the emojis have headphones that represent the hearing loss. The size of the headphones represents the size of the loss in the respective ear. If the user does not have hearing loss, then the emoji is represented with a pair of sunglasses. If the user has a partial hearing loss, only on one side, then a headset is represented on the respective side, as can be seen in figure 30. In addition to the hearing loss data, the headphones and glasses represent the region in which the test was performed through color (Figure 32). Colors are also used to represent age. In this case, 4 shades of yellow? are presented for the following variations: less than 50 years old, from 51 to 65 years old, from 66 to 80 years old, and over 80 years old. years old.

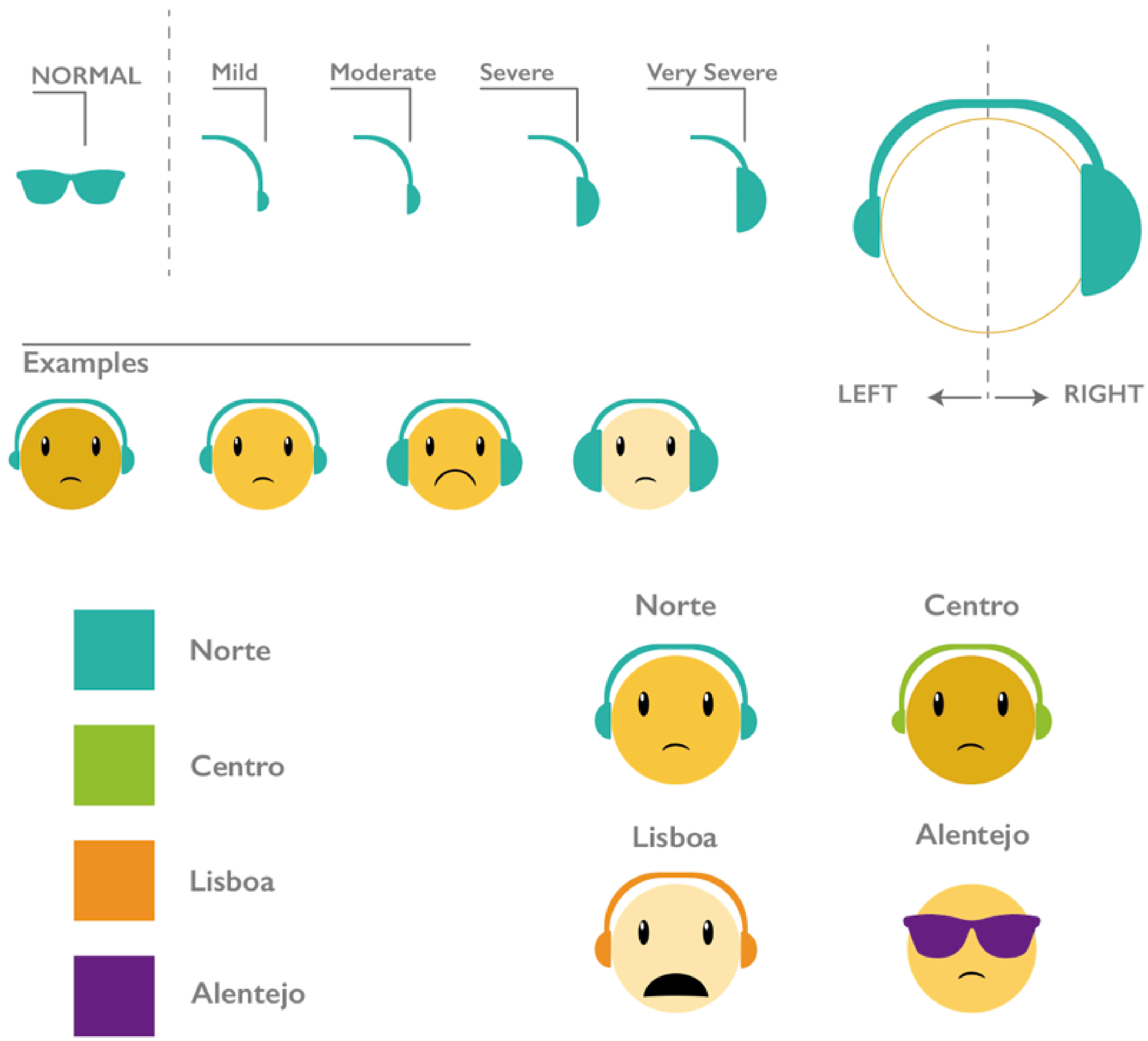
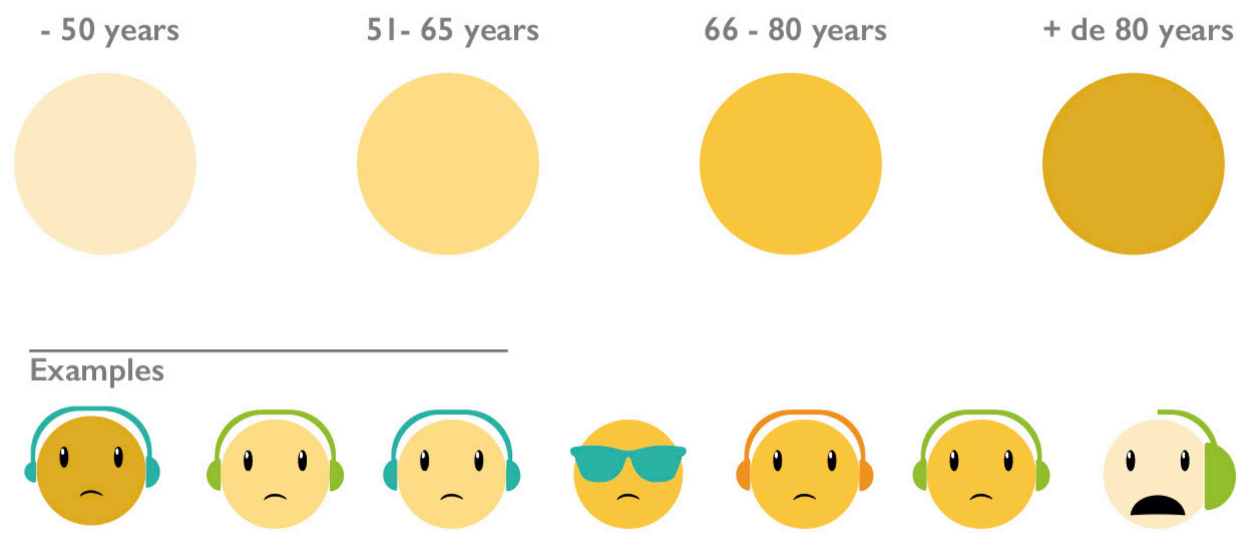


Figure 31 and 32: Representation of "hearing loss" and "region" attributes in the emoji glyph

Finally, the representation of status is done through expression, projected through the drawing of the mouth. The glyph represents the 8 statuses by categorizing "good" statuses, represented by smiling, and "bad" statuses represented by sad mouths. The more intense the expression, the further along the contact process was. For example, a user who has only had the first successful contact will have a small smile, and a user who has already carried out all the stages of the service and in the end made the purchase will have a bigger and more intense smile.



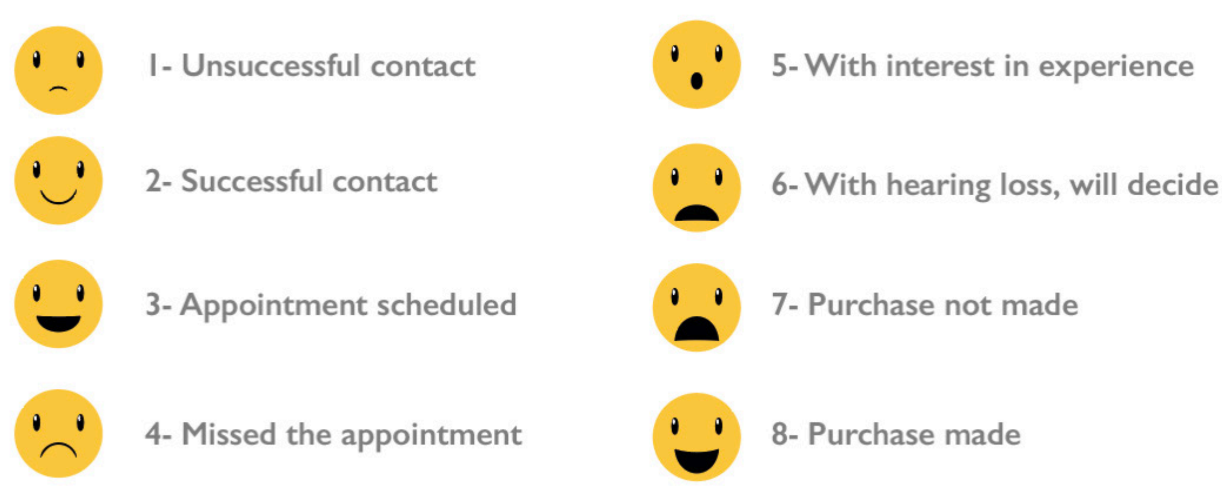


Figure 33 and 34: Representation of "age" and "status" attributes in the emoji glyph

Glyph 2 - Flowers

Figure 35: Different types of flowers glyphs

The second glyph uses the metaphorical approach of flowers in its visual representation. The shape of the petals represents the region in which the test was carried out, with the shape being narrower and with double ends towards the North and more rounded towards the Alentejo (Figure 36). Each petal represents a frequency in the right and left ears respectively, with the highest petals referring to the lowest frequencies and increasing as the petals get lower.

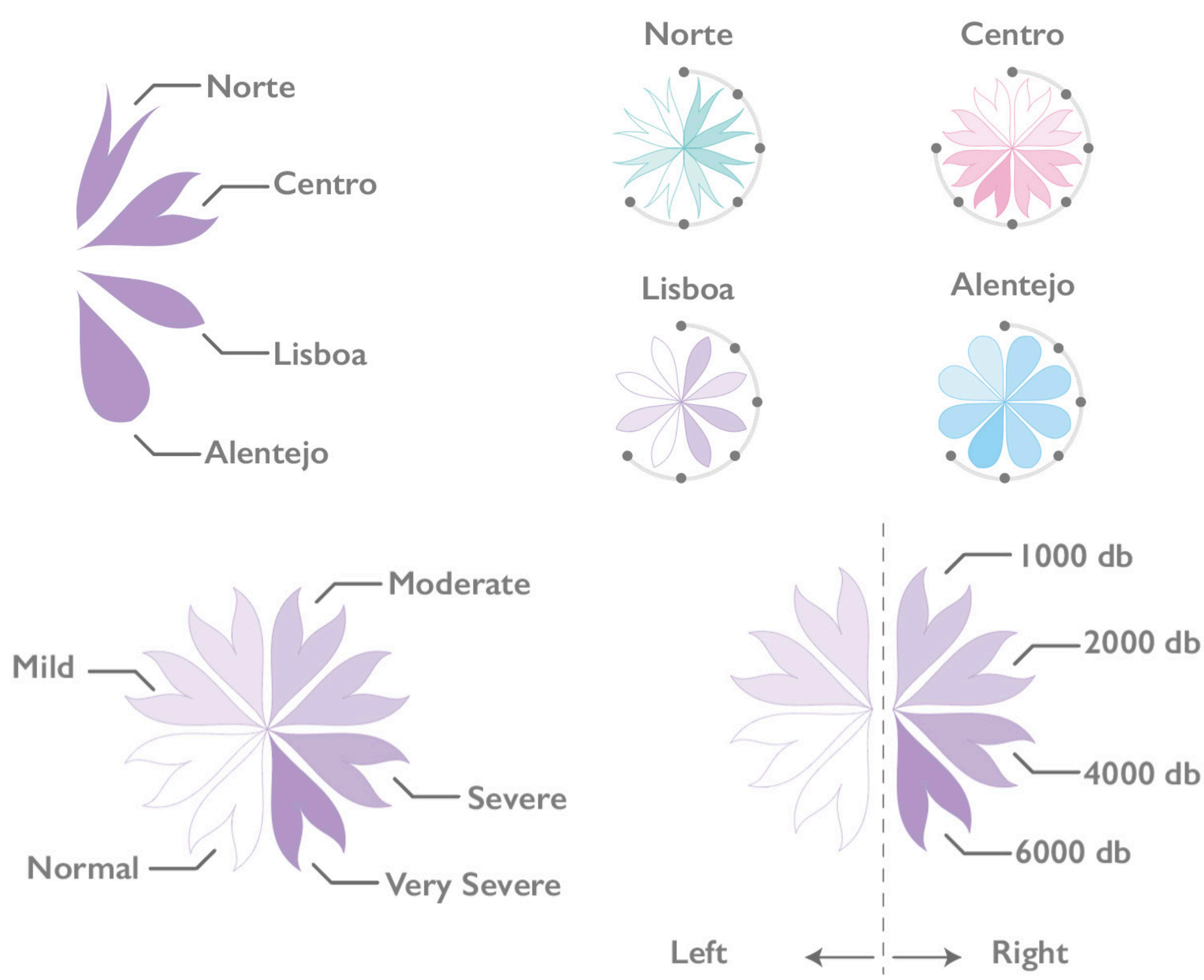


Figure 36 and 37: Representation of “region” and “hearing loss” attributes in the flowers glyph

When any of the frequencies have no hearing loss, it is represented without a fill color. When any of the frequencies have hearing loss values, represented by the opacity of the petal filling color, a very severe loss has an opacity of 100%.

Finally, the representation of the status is done through the color of the petals. As in glyph 1, 8 statuses are represented through the categorization of "good" and "bad" statuses, all "good" statuses, that is, those that represent a successful contact, appointment scheduled or a sale made, are represented with shades of green and blue. The so-called "bad" statuses are represented through warmer tones: reds, pinks, and violets.

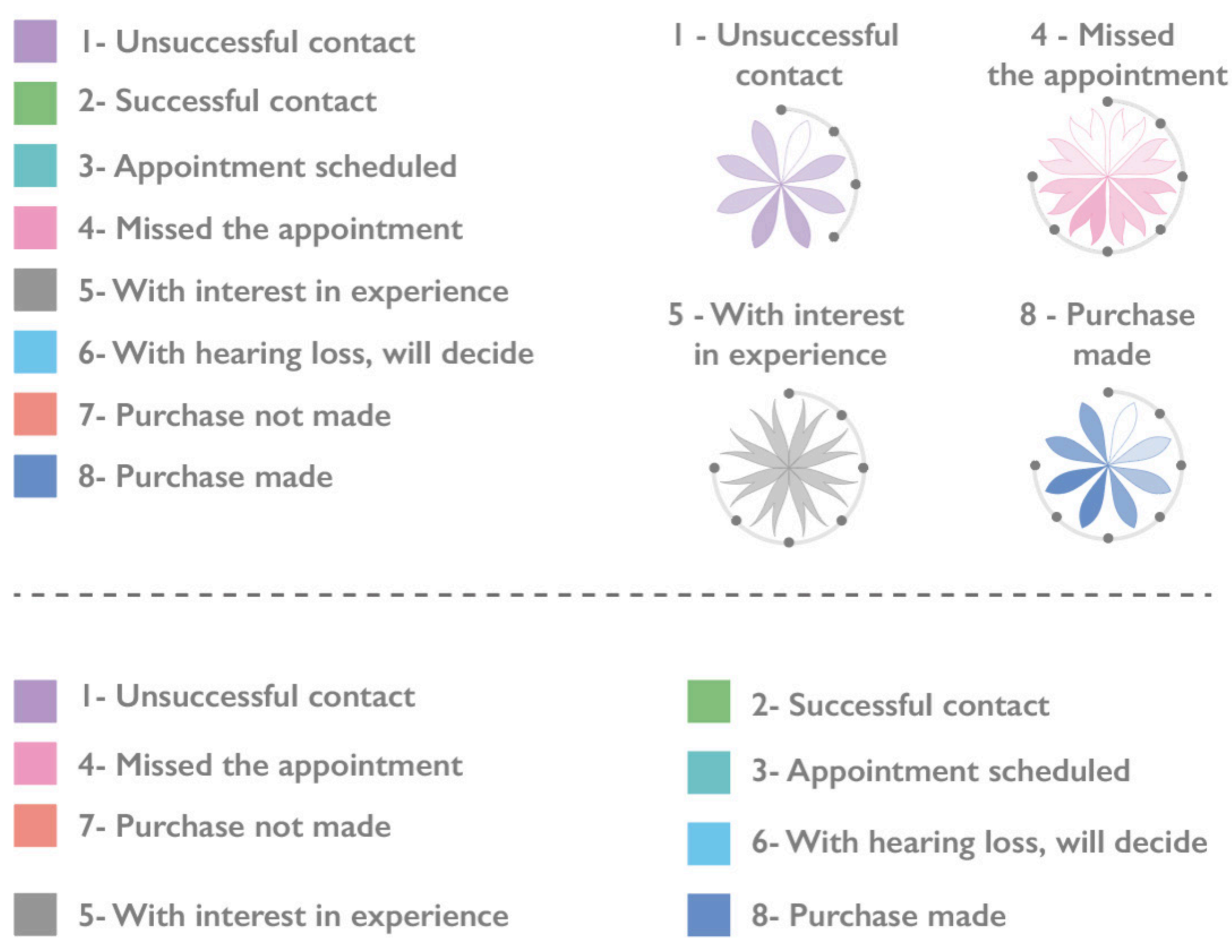
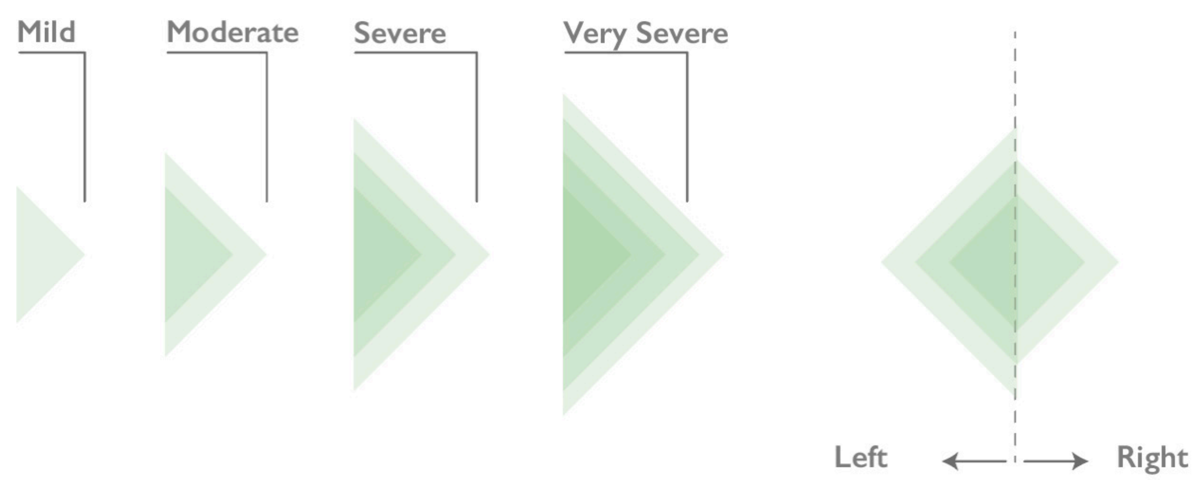


Figure 38: Representation of "status" attribute in the flowers glyph

Glyph 3 - Geometric

The third glyph developed is based on geometric shapes. In the center of the shape, diamonds are superimposed according to the average hearing loss in the right and left ears. Hence, the degree of hearing loss is perceptible both by the size of the area and by the intensity of the color in the diamonds. At the center of the glyph are placed 8 small vertical lines, representing the age of the user. The lines are divided into two sides, that is, 4 lines on the left side (representing users up to 40 years old) and 4 lines on the right side (representing the entire population over 40 years old).

Figure 39: Representation of "hearing loss" attribute in the geometric glyph



The region is represented through external strokes in the diamond area. This representation is made with the aim of making a semantic relationship with the geographical organization of the regions of Portugal, where the first stroke positioned at the highest point of the glyph represents the north. , so the trace is located at lower and lower points as the regions are further south.

Figure 40: Representation of "region" attribute in the geometric glyph

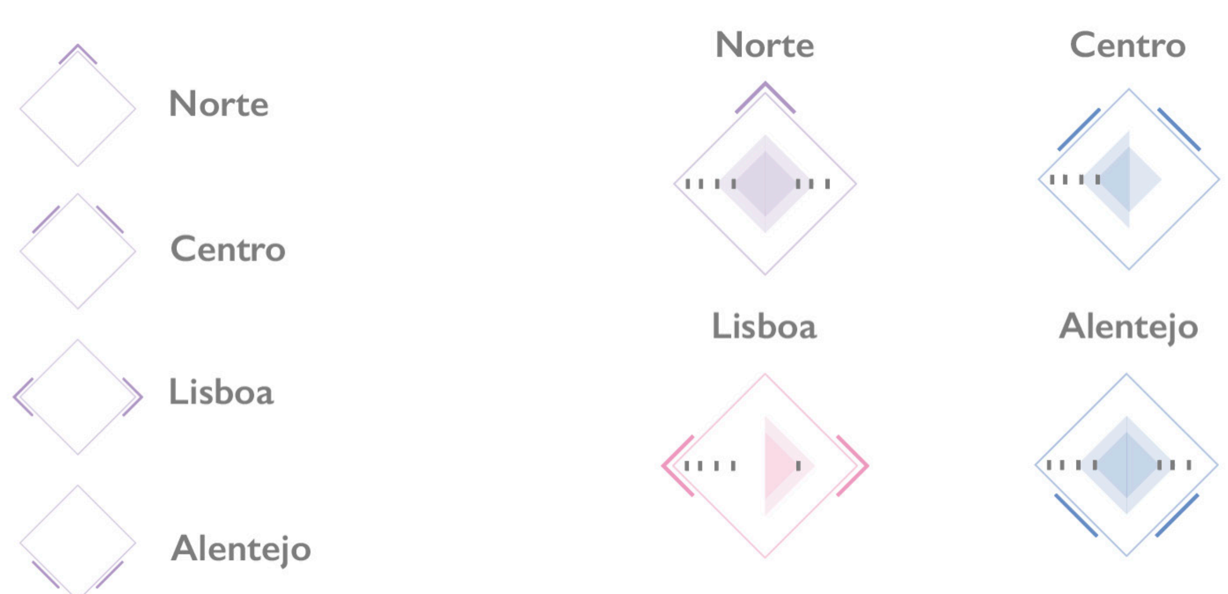
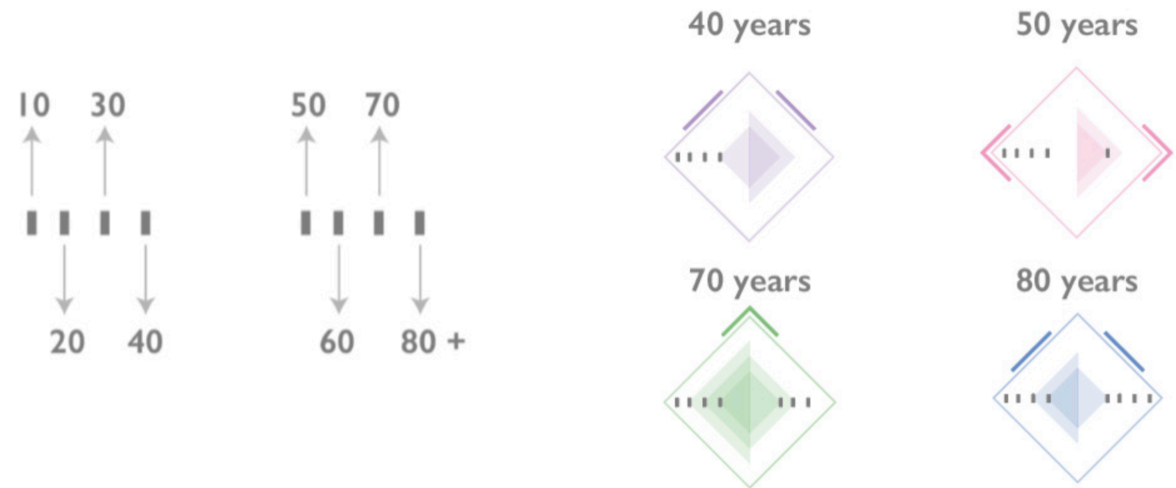
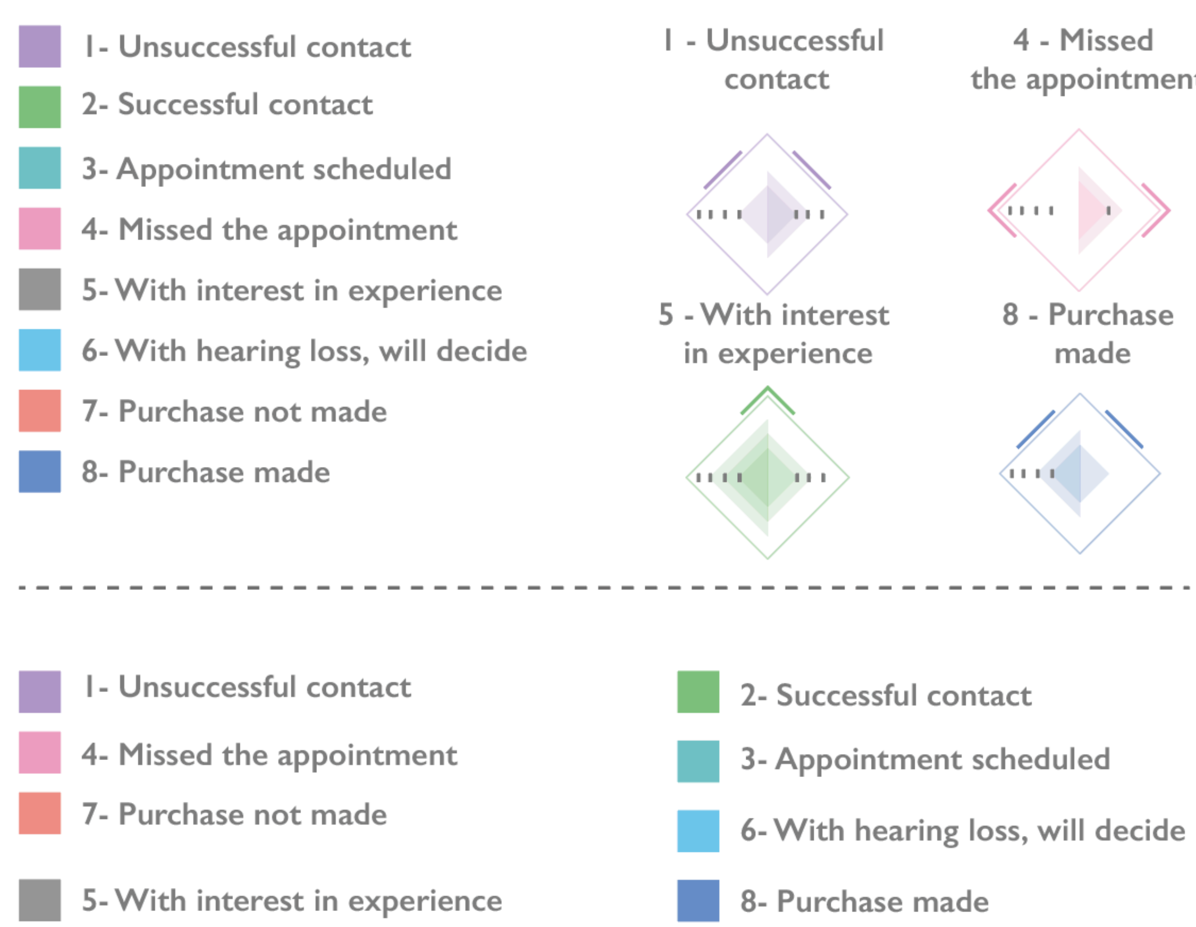


Figure 41: Representation of "age" attribute in the geometric glyph



Finally, the status representation is done through the color of the glyph. As in glyph 2, 8 colors are represented that are divided between blue and green to represent "good" status and warmer colors to represent "bad" status.

Figure 42: Representation of "status" attribute in the geometric glyph



4.4. Evaluation

For the evaluation of the glyphs, forms with pre-selected glyph designs and tests with multiple choice questions were used. right answer.

A total of 48 people participated, most of them from the local student population. All participants had a normal or corrected-to-normal vision and reported no color blindness. Each participant performed the tests referring to two glyphs, that is, each glyph was evaluated 24 times. In addition, each participant received a different combination of the glyphs, so each possible combination was tested twice. Aiming that each user, when performing the test on the second glyph, would be easier to have had previous contact with the type of data to be represented, each glyph was tested 12 times as the first position in the test and 12 times as the second. The task question was always visible on the screen and the time for each answer was counted.

Procedure

The experiment took place in a quiet, closed room at our university. Aside from the participant, the experimenter was the only person present. The test started by asking if the participant was familiar with information visualization and if he knew what data glyphs were. If the participant was not aware, a brief explanation of what it was about was given. However, we sought to do as few oral explanations as possible, basing the entire evaluation on the forms that contained the presentation of the glyph to be evaluated, thus ensuring the maximum possible uniformity throughout the procedure.

- **Section 1:** Before giving a presentation of the data and how it is represented in the glyph, all participants are asked a question about the hearing loss value. The aim was to understand if the visual coding of the glyph was done intuitively, that is, without any presentation, the participants perceived the hearing loss values within the glyphs.

- **Section 2:** After the presentation of the glyph structure, a question section is made to recognize the attributes in the glyph, there are a total of 4 questions. Each question had only one sample glyph and questioned one of the attributes represented there, always with multiple choice answers.

- **Section 3:** Composed of 3 questions, this section tested the ability to remember the glyph. Before receiving the questions, the participant was presented with a glyph for 7 seconds. The participant should try to record as much information as possible and then go to the next page to answer the questions.

- **Section 4:** Section composed of 3 comparison questions. In each question, 3 glyphs are presented and the participant must choose which one best fits the answer. The purpose of this section is, in addition to analyzing the attributes themselves, to see if the differences between the glyphs are easily noticeable.

- **Section 5:** The last section is an opinion, the objective is to collect qualitative information regarding the participant's perception. There are three statements: glyphs are easy to notice, glyphs are visually appealing, and whether glyphs are easy to remember. The participant can give a score from 1 to 5, where 1 completely disagrees and 5 completely agree.

Results

We report significant results, where all glyphs presented an assertiveness rate above 60% in the quantitative analysis, in addition, the rate of all glyphs exceeded glyph 4 which was used as a parameter for the analysis. Within the assessments of assertiveness, the glyph that stood out was 3, which uses geometric shapes for its representation, but glyph 1 obtained very similar results, with a difference of only 0.40% between the two. Regarding the response time, glyph 1 obtained better results, with a difference of almost 1s. Glyphs 2 and 4 had a lower result, despite glyph 2 having an assertiveness above 70%.

Table 4: Assertive rate and response time in glyph validation

	Glyph 1	Glyph 2	Glyph 3	Glyph 4
Assertiveness	76,40%	72,31%	76,80%	67,20%
Time	8,55 s	10,15 s	9,5 s	12,5 s

When analyzing the test results isolating the questions referring to each graphic attribute represented, it is possible to notice that some glyphs present better results in the representation of certain attributes than others. For example, glyph 1, with the use of facial expressions and mouth shape, showed a high rate of assertiveness regarding the representation of status. Another example is glyph 2, which despite having a lower overall result than the other glyphs, has a significantly higher assertiveness rate in representing age. In addition, in the representation of hearing losses, glyph 3 was the one that presented the best results in the perception of hearing loss levels.

Table 5: Assertiveness rate of each glyph in the different attributes.

	Age	Status	Region	Hearing level
Glyph 1	73,33%	96,00%	68,00%	72,00%
Glyph 2	91,03%	63,40%	57,60%	69,20%
Glyph 3	76%	58%	78%	81,33%
Glyph 4	68%	70%	70%	62,67%

Finally, about the qualitative questions asked whether: (i) the glyphs were easy to understand, (ii) the glyphs were visually appealing, and (iii) the glyphs were easy to remember. In all questions, glyph 1 had the highest rating, but all questions had a variation between 3.5 and 4.5 in the answers.

Table 6: Average of results in qualitative questions

	Question 1	Question 2	Question 3
Glyph 1	4,28	4,04	4,55
Glyph 2	3,69	3,64	4,38
Glyph 3	4,12	3.91	4.40
Glyph 4	3,84	3,84	4,2

5. Conclusion

When analyzing the different projected glyphs, it was noticed that a device that became extremely important in the perception of information was the use of semantic relationships. All the attributes that tried to provoke some type of natural mapping had superior performance to the other data represented. This can be seen in the results of glyph 1 that through the use of smiles to represent the status obtained a 96% assertive response while the other glyphs had values below 70% for the same attribute. Another attribute that obtained superior results was glyph 3 with regard to the representation of regions. The same did a semantic representation relating the position of the graphic mark in the glyph with the geographical distribution of the different regions in Portugal, the approach proved to be effective when.

In addition to the use of metaphor, another graphic device that proved to be fundamental for the glyph's effectiveness during the evaluation period was the use of visual resources that are easily measurable. For example, in glyph 2, despite having a lower performance than glyph 1, about age information, participants perceived the information much better, thus having a significantly higher assertiveness rate (91%) than glyph 1 (73.3%). This is because in the glyph two spherical shapes are used to represent age, and this is easily countable, whereas the emoji glyph used the shade of yellow, information that is more difficult to perceive and measure.

Despite having a good performance compared to others, there are points of failure in the representations of glyph 1, firstly, as stated earlier, the representation of age is still very ambiguous and difficult to perceive. This is because the shade of yellow is not so easily measurable, so the value was only well understood when compared to other glyphs, when the participant had to analyze the glyph in isolation, the information ended up being confusing. Another attribute in which glyph 1 had a lower performance was in the representation of regions, also coded through colors. Most participants could not remember the meaning of each of the colors, in addition to the colors that represent north and center were very similar which caused indecision. Glyph 3, which performed the best in assertiveness during the validation process, ended up performing relatively poorly in the status representation (58%), also color-coded. This demonstrates that although colors are an attribute that stands out easily, thus being one of the most easily perceived attributes (Borgo et al., 2013), it is difficult to decode, demanding a lot from the user regarding the memorization of meanings. , especially when it has many variables as was the case with glyph 2.

Glyph 4, despite having a lower performance than all other glyphs, has the second best performance in relation to the representation of status, it is clear that this occurred because, in addition to the use of colors, such as glyphs 2 and 3, they are also used the different rings within the glyph to count, that is, this glyph uses redundancy as a graphic device and managed to achieve a better performance than those who did not follow the same strategy. The use of redundancy is shown as a possible strategy to solve the perceived problems with regard to colors, and can even improve the performance of glyph 3, which already has a good result. Briefly, the following improvement requirements were noticed in the different types of glyphs for future work:

- **Glyph 1:** Improve color coding for regions, look for colors that are more distinguishable from each other. If possible, also use visual redundancies to reinforce the information. Changing the age coding, the use of different shades of yellow for age proved to be difficult to perceive.

- **Glyph 2:** Using another visual resource to represent the region, the shape of the petals proved to be a difficult graphic device to decode, obtaining the worst performance within the validation. Reviewing the use of colors to represent the status, it has many variables and proved difficult to decode.

- **Glyph 3:** As with glyph 2, in reviewing the use of colors for the representation of status, one strategy is to use redundancy with two other visual resources, aiming that it is one of the most important pieces of information within the hierarchy.

Finally, with the following work, it was possible to study the different types of glyphs and visualize in practice how the different visual channels behave. In addition, it was realized how important a validation step is in any information visualization project, in which it was possible to identify the visual impact of each of the glyphs and their flaws. For future work, concerning the project, the need for a redesign stage is highlighted, applying the improvements pointed out here in the requirements generated in the validation stage, after the redesign a new validation stage must be carried out. In addition, the possibility of implementing the glyphs is left open, working with the different arrangement possibilities and rendering techniques. At the theoretical level, a deeper understanding of the use of natural mapping and its impact on visual perception is open for future work.

References

Anderson, E. (1960). A semigraphical method for the analysis of complex problems. Technometrics, 2(3), 387-391.

Anscombe, F. J. (1973). Graphs in statistical analysis. The american statistician, 27(1), 17-21.

Bertin J (1967) Sémiologie graphique. Les diagrammes, les réseaux, les cartes. With Marc Barbut [et al.]. Paris: Gauthier-Villars.(Translation 1983. Semiology of Graphics by William J. Berg). Mapping Ecosystem Services, 70.

Borgo, R., Kehrer, J., Chung, D. H., Maguire, E., Laramee, R. S., Hauser, H., ... & Chen, M. (2013, May). Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In Eurographics (State of the Art Reports)(pp. 39-63).

Carpendale, S. (2008). Evaluating information visualizations. In Information visualization (pp. 19-45). Springer, Berlin, Heidelberg.

Chan, W. W. Y. (2006). A survey on multivariate data visualization. Department of Computer Science and Engineering. Hong Kong University of Science and Technology, 8(6), 1-29.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. Journal of the American statistical Association, 68(342), 361-368.

Chung, D. H., Legg, P. A., Parry, M. L., Bown, R., Griffiths, I. W., Laramee, R. S., & Chen, M. (2015). Glyph sorting: Interactive visualization for multi-dimensional data. Information Visualization, 14(1), 76-90.

Cleveland, W. S., & McGill, R. (1987). Graphical perception: The visual decoding of quantitative information on graphical displays of data. Journal of the Royal Statistical Society: Series A (General), 150(3), 192-210.

De Leeuw, W. C., & van Wijk, J. J. (1993, October). A probe for local flow field visualization. In Proceedings Visualization'93(pp. 39-45). IEEE.

DuToit, S. H., Steyn, A. G. W., & Stumpf, R. H. (2012). Graphical exploratory data analysis. Springer Science & Business Media.

Ebert, D., Morris, C., & Rheingans, P. (1999). An Experimental Analysis of the Effectiveness of Features in Chernoff Faces. University of Maryland Baltimore County.

Fanea, E., Carpendale, S., & Isenberg, T. (2005, October). An interactive 3D integration of parallel coordinates and star glyphs. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005. (pp. 149-156). IEEE.

Fuchs, J., Isenberg, P., Bezerianos, A., & Keim, D. (2016). A systematic review of experimental studies on data glyphs. IEEE transactions on visualization and computer graphics, 23(7), 1863-1879.

Kamsani, I. I., Ariff, N. M., & Abd Khalid, N. E. (2015). Survey of Star Glyph-Based Visualization Technique for Multivariate Data. Australian Journal of Basic and Applied Sciences, 9(26), 61-66.

Keim, D. A., & Kriegel, H. P. (1996). Visualization techniques for mining large databases: A comparison. IEEE Transactions on knowledge and data engineering, 8(6), 923-938.

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.

Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.

Krok, E. (2021). Visualization on charts–manipulations and distortions. Procedia Computer Science, 192, 3932-3944.

Levkowitz, H. (Ed.). (1997). Color theory and modeling for computer graphics, visualization, and multimedia applications. Boston, MA: Springer US.

Maceachren A. M., Roth R. E., O'Brien J., LI B., Swingley D., Gahegan M.: Visualizing semiotics & un- certainty visualization. IEEE Trans. Visualization and Computer Graphics 18, 12 (2012), 2496–2505. 20

Maguire, E. (2015). Systematizing glyph design for visualization. DPhil Thesis.

McGrath, J. E. (1995). Methodology matters: Doing research in the behavioral and social sciences. In Readings in Human–Computer Interaction (pp. 152-169). Morgan Kaufmann.

Meirelles, I. (2013). Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations. Rockport publishers.

Mendes, A. M. A. (2020). Dispositivos médicos de reabilitação auditiva: caracterização do mercado em Portugal (Doctoral dissertation, Instituto Politécnico de Lisboa, Escola Superior de Tecnologia da Saúde de Lisboa).

Morris, C. J., Ebert, D. S., & Rheingans, P. L. (2000, May). Experimental analysis of the effectiveness of features in Chernoff faces. In 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making (Vol. 3905, pp. 12-17). SPIE.

Munzner, T. (2014). Visualization analysis and design. CRC press.

Pickett, R. M., & Grinstein, G. G. (1988, August). Iconographic displays for visualizing multidimensional data. In Proceedings of the 1988 IEEE Conference on Systems, Man, and Cybernetics (Vol. 514, p. 519).

Pratt S.R. Barriers to Hearing Health Care: Current Status and a Glimpse at the Future. American Journal of Audiology. 2010

Prohaska, G., Aigner, W., & Miksch, S. (2007). Glyphs and Visualization of Multivariate Data. Vienna University of Technology.

Raminhos, M. F. L. (2019). Qualidade de vida em indivíduos com perda auditiva: revisão sistemática da literatura (Doctoral dissertation, Instituto Politécnico de Lisboa, Escola Superior de Tecnologia da Saúde de Lisboa).

Ropinski, T., Oeltze, S., & Preim, B. (2011). Survey of glyph-based visualization techniques for spatial multivariate medical data. Computers & Graphics, 35(2), 392-401.

Seldner, M., Siebers, B., Groth, E. J., & Peebles, P. J. E. (1977). New reduction of the Lick catalog of galaxies. The Astronomical Journal, 82, 249-256.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677-680.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103(2684), 677-680.

Tufte, E. (2001). The visual display of quantitative information. Originally published in 1983.

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. Information Visualization, 1(3-4), 194-210.

Ward, M. O. (2008). Multivariate data glyphs: Principles and practice. In Handbook of data visualization (pp. 179-198). Springer, Berlin, Heidelberg.

Ward, M. O., Grinstein, G., & Keim, D. (2010). Interactive data visualization: foundations, techniques, and applications. CRC press.

Ware, C. (2019). Information visualization: perception for design. Morgan Kaufmann.