Cartography M.Sc.

Master thesis

Developing a system for suggesting alternative ways of visualizing data

Iaroslav Boretskii

Technical University of Munich

Technische Universität Wien, Vienna University of Te

Technische Universität Dresden

ITC University of Twente.

2022

Statement of Authorship

Herewith I declare that I am the sole author of the submitted Master's thesis entitled:

"Developing a system for suggesting alternative ways of visualizing data"

I have fully referenced the ideas and work of others, whether published or unpublished. Literal or analogous citations are clearly marked as such.

Munich, 09.09.2022 Iaroslav Boretskii

UNIVERSITY OF TWENTE.

Master Thesis

Developing a system for suggesting alternative ways of visualizing data

Iaroslav Boretskii
M.Sc. Thesis

March-September 2022

Supervisors:
Prof. Dr. M.-J. Kraak (Chair)

Thesis Assessment Board:
Prof. Dr. M.-J. Kraak (Chair)
Dr.-Ing. H. Kumke (Reviewer, TUM)
Dr. B. Köbben (M.Sc.Thesis Coordinator, Chair)

University of Twente's Faculty
of Geo-Information Science
and Earth Observation Department
of Geo-Information Processing

Acknowledgements

As long as I can remember, from my early years, only three things were truthfully exciting to me: watching motorsports, making public speaking, and designing maps and diagrams. The first in this work is completely irrelevant, second — maybe partially, as I will defend this work in front of the scientific audience soon; the third thing, finally, has properly found a practical embodiment in this work, which is written myself. This became possible through a wonderful programme such as **Cartography M.Sc** and all the involved in it professors, coordinators, students, and partners. I am proud to graduate from this programme, having the astonishing privilege of its accomplishment with dignity. This experience has changed my life forever, completely refactoring how things can be.

I am tremendously grateful for the organization of the entire study process during these difficult two years; I say thanks for the impeccable support and presence of those who were involved. I would like to highlight **Menno-Jan Kraak, Juliane Cron, Edyta Bogucka, Barend Kobben** and **Paolo Raposo** as they had been aspiring to me and stayed close from the beginning till the end of this Master's Thesis.

Regarding the contents, I must confess that I did not have enough time and resources to complete this work the way I originally foresaw it. Nevertheless, it is now done and I am very happy about it. Writing this work has taught me a lot from a project management perspective, data visualization' philosophy, and conceptual design processes. I'm pleased with how it all came together in the end. My extra gratitude and beloved words of respect to the whole masters' experience go to:

- Darya Lapo for caring, being around with support, patience, and mutual support throughout these two years;

- Liam, Sam, Stevie, Vivien, Lydia, Hassam, Syed, Majk, Hector, Zulfa, Camilo and **Leo** for a good friendship and fun times;

- My dad Andrey and **mom Zhanna, my sister Kate** and **her husband Ilya,** for raising me and preparing me for the programme; for being always in touch when the support was heavily needed;

- To all my other friends back in Russia and wherever they are now, to all acquaintances and colleagues, to all people whom I have met during the last two years; thank you for your warm memories and support.

Abstract

The UN Agenda for the Sustainable Development Goals (SDGs) is an actual, engaged framework for a resilient progress in ecological, social and political aspects. Most of the targets and indicators for the 17 Goals already have an established methodology and can be expressed in a measure of value, and thus can potentially be visualized.

The goals can be used for analysis in various ways: comparisons between countries and territories, for assessing dynamics, identifying qualitative and quantitative transformations, classification, for possible forecasting. The selection on data visualization options for such topics represents a strong challenge: the type of data or presentation format may be not well understood, the design or visualization type itself (especially maps) may be inappropriate, the audience can be misplaced; the external support to see potential successful diagram variations, including the inclusion of some data transformation option, is often needed. Therefore, research on mapping and visualization within the SDGs is still needed; as a contribution to this ongoing overall research, this work attempts to develop an interactive system of guidelines and hints for proper visualization of SDG data, realized as a recommender system.

The current thesis attempts to find an approach to close the gap between the original data and its visualization outputs through an elaboration on suggestion system for the SDG indicators. The peculiar focus is to implement the alternative options for the indicators by considering the data transformation.

Contents

List of acronyms

SDG	Sustainable Development Goal
SDGs	Sustainable Development Goals
MDGs	Millennium Development Goals
DIKW	Data, Information, Knowledge, Wisdom
UN	United Nations
ICA	International Cartographic Association
LoM	Level of Measurement
ATL	Attribute, Location and Time model
KPI	Key Performance Indicators
DVS	Data Visualization Society
RSOs	Research Sub-Objectives
RQs	Research Questions
ML	Machine Learning

Chapter 1

Introduction

Data visualization holds a significant place in science, media, history, technology and modern daily life. We see multiple data-driven methodologies alongside with wide availability of computational technologies and miscellaneous devices. In this situation, an understanding and presentation of data through visual form stands out as an important aspect to develop and involve. Data visualization is one of the approaches to describe and explore big numeric datasets about miscellaneous phenomenons, converting them into summarized pictures and figures (Tufte, 2001). The information perception' quickness alongside with precision of understanding is connected to the final true goal of data visualization: express the important and valuable gist of different concepts and phenomenons to the readers (Few, 2004). The visualization's objective is the insight, not pictures, so it does imply the amplification of cognition (Card, Mackinlay, & Shneiderman, 1999). Thus, computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively (Munzner, 2014).

Data visualization can be powerful with regards to cognitive advantages and effective communications. But it does not always works straightforward: inappropriate visualization types and styles are not rare occurrences still; *graphical literacy* is a presently topical issue. Thus, there is a space to come up with support that could recommend appropriate types of visualizations or improve some aspects of it, depending on the input and output parameters. It implies certain disintegration of the graphics and visualizations with attempt to build a framework that could describe each type of visual artefacts in their core characteristics. A theoretical legacy in data visualization' classification can be synthetically used to come up with an algorithm that could be used to provide more instant options to look upon and be inspired from.

Sustainable Development Goals (SDGs), as a particular case, can benefit from the usage of data visualization. The existing indicators that are describing SDGs are characterized by potentially visualized and mapped values. Graphics, therefore, are able to effectively communicate SDG challenges and successes; they lead to a better global awareness (Kraak, Roth, Ricker, Kagawa, & Le Sourd, 2020). Combination of data visualization' taxonomy and recommender' computer-based algorithm might bring a benefit for the visualization and analysis of SDG data.

1.1 Research Identification

1.1.1 The Main Research Objective

The Main Research Objective of this study is **to implement a system for suggesting multiple ways of data visualization particularly for SDGs.** The system will be elaborated in an algorithm and realized in a prototype. The system can be regarded as *a recommender* or *suggestion* system that can assist in design process decisions. A theoretical analysis of the available research on relevant topic is the basis that will justify the choice of the main features of this prototype.

1.1.2 Research Sub-Objectives and Research Questions

The Main Research Objective is subdivided into Research Sub-Objectives (RSOs) with corresponding Research Questions (RQs) and described as follows:

RSO A: Investigate the conceptual basis behind the data visualization and SDGs

• RQ A1 What are the main features of data visualization as a concept?

• RQ A2 What methodologies classify and allocate different data visualization types?

• RQ A3 How do data transformation and representation affect the data graphics output?

• RQ A4 What concept stands behind Sustainable Development Goals?

• RQ A5 What are the challenges for the user in visualizing SDG indicators??

RSO B: Investigate the methodologies of data visualization recommender systems

• RQ B1 What is the reasoning behind the recommender systems for data visualization?

• RQ B2 Which methodologies are used to implement the recommendation systems for data visualization types?

RSO C: Conceptualize the recommender system with respect to the RSO A and RSO B findings

• RQ C1 Which data visualization conceptual aspects (from RSO A) are in the focus of the recommender system?

• RQ C2 What characteristics does the recommender system have?

RSO D: Define software and interface requirements for a system and design the prototype

• RQ D1 What requirements are applied to the prototype?

• RQ D2 Which technology or software is used to design the prototype?

• RQ D3 What are the interface-related (UX/UI) demands that should be considered to design the prototype?

1.1.3 Mapping for Sustainable Development Goals

One of the main motivations for the thesis elaborations is tied to the expansion of the research related to the SDGs' presentation in the context of cartography and visualization.

The key publication, on which the recommendation system will be studied and made, is the Mapping For a Sustainable World (Kraak et al., 2020).

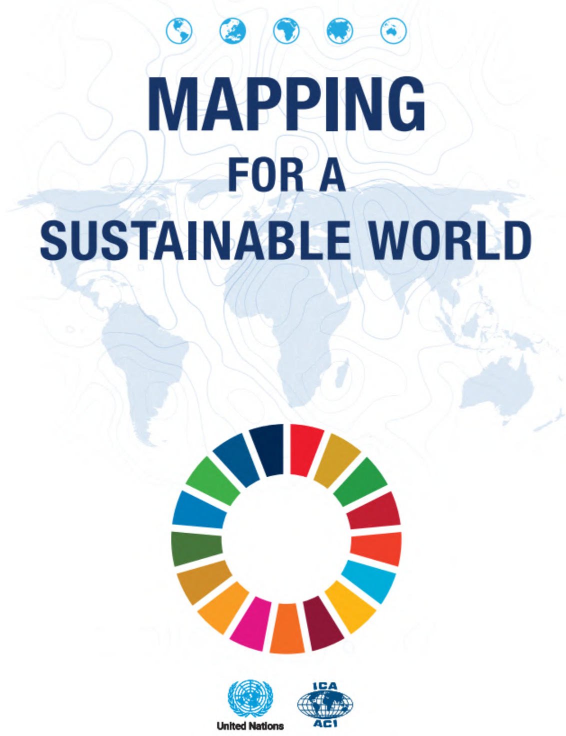

Figure 1.1: Mapping for a Sustainable World book cover [Kraak et al. (2020)]

The book reveals the corresponding aspects of SDGs's relation to geospatial data, describing SDG indicators, foundational design decisions in the cartographic workflow, the common map types and diagrams for representing the SDG indicators considerations for map use environments, etc.

Thus, this thesis can be treated as an attempt to expand the already existing research on the subject of sustainable development and coherent use of geospatial data and information presentation.

1.2 Thesis outline

Literature review and background investigation is the starting point to understand the context and obtain background knowledge in order to answer the research questions from RSO A. The data visualization conceptions and SDGs overview are the first chapters to be presented.

Chapter of Methodology combines the comparison of the automated visualization together with author's final decisions on the conceptual system output. The description and comparison of different data visualization systematization' methodologies and frameworks are main empirical methods to use firstly. Methodology' review for RSO B is implemented through comparative analysis and further synthesis to understand the key features of data visualization recommender systems. RSO C implies the synthesis of the investigated concepts from RSO A and RSO B into a recommender system' conceptualization.

Prototype and Results. RSO D is related to the pre-realization of the recommender system; it is mainly about the definitions and restrictions of the software capabilities according to the recommender system features and specifications. The practical realization of the recommender system as a prototype is included here.

Chapter 2

Data visualization background and related works

2.1 Data vs. Information

There are many perspectives on the definition of *data visualization*. It is important to understand both words and how they support each other and turn themselves into a proper definition. Nowadays we see that *data* is used equivocally with the terms of *information* or *knowledge*, which is obviously misleading, since a logical consensus implies that *data* is not *information* and *information* is not *knowledge* (as well as *information visualization* is not *data visualization*).

Data, information, and knowledge can be viewed in a hierarchical relationship, which is known as Data, Information, Knowledge, Wisdom (DIKW) pyramid (Figure 2.1). It is a popular model that classifies the human's understanding in the perceptual and cognitive space, and which is often referred to in publications on data visualization studies as well.

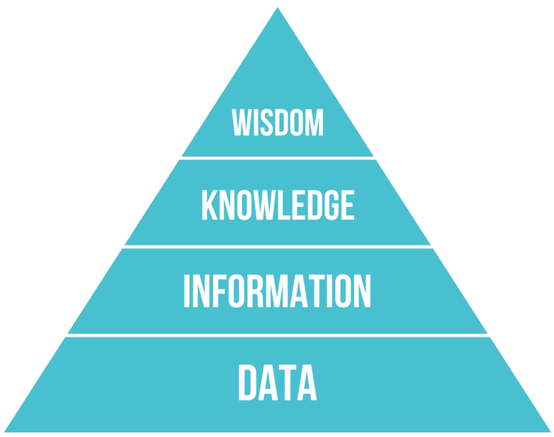

Figure 2.1: DIKW Pyramid

Data described here as **facts, or signals, in some potentially usable, relevant form.** The data in irrelevant and unusable forms have no meaning for further conversion. Information stands out as data' derivative; it adds a *cognitive layer*; shortly, by adding semantics of

'who', 'what', 'where', and 'when' to data, it becomes information. Knowledge appears as a derivative of information and implies **a direct application to answer questions related to causality, structure, or genesis.** Knowledge further involves theories, concepts, persuasions, opinions, and definitions. Wisdom in this structure appears as a kind of the final point, which is described as **guidelines for action that directly corrects and influence one's behavior.**

With the recent technological advancements, data and information as concepts obtained *a special status.* More data-driven approaches are introducing the social, economics, and life sciences, which turned them having more operationalization-based methodologies, multiple measurable quantitative indicators, and further qualitative descriptions. In this regard, data and information can be treated as entities in computational space (Chen et al., 2009). Chen, Laramee Ebert are describing data in this extent as **computerized representations of models and attributes of real or simulated entities,** while information represents **the data as the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings**. A primitive example of DIKW hierarchy in a real situation is presented in Table 2.1, showing clearly how all entities are getting semantically different from each other. Thus, although data and information are interchangeable, it is *data* that stands first, as it is *yet to become useful*, while information contains a valuable meaning as such.

Table 2.1: DIKW Explanatory Example

DIKW Entity	Possible example description
Data	For the last 50 years there are 37 200 hourly-divided observations of air temperature in ℃ during the month of August in the point of 30° N 31° E. Weather conditions are qualitatively described too as "sunny", "overcast" or "rainy".
Information	The average daily temperature in August for the last 50 years at this point is 34.7 ℃. 95 percent of the observations were described as "sunny". of This point is the city of Cairo, Egypt's capital.
Knowledge	Cairo can be classified to have a hot desert climate (e.g. by Koppen system) due to its location and temperature indicators in August. In sum, it is very hot during the day in Cairo.
Wisdom	In August during the day in Cairo it is better to walk in the shadows and wear a light shirt and a cap on a head and stayed hydrated to avoid overheating or sunstroke.

2.2 Data visualization synonyms and their definitions

There are many definitions of data visualization too. Unlike the term, for example, *cartography,* which was once defined and accepted by the International Cartographic Association (ICA) as an art, science, and technology of making and using maps, with data visualization we see multiple and scattered terms and definitions' variations across publications and works. Moreover, it is often used alongside (or instead) with synonym terms of "information design", "data graphics", "infographics", "visual communication", "visualization", "graphic representation of data", or even narrower names of "charts", "diagrams", "graphs", "plots", "dashboards". Indeed, all of those terms to a certain degree are interchangeable or at least tightly connected in a certain order. Such a disparity takes its roots from a relative immaturity of data visualization as a discipline, science, or technology. It is a recent invention that was and still is requiring a vast variety of skills in mathematics, statistics, data processing, empirical methods, art, psychology, perception, and graphics (Tufte, 2001). To embark, we will review several definitions of data visualization and its synonyms.

"The academic pioneer of data visualization" E. Tufte mentions *data graphics*, which stands as **"a display of measured quantities by means of the combined use of points, lines, coordinates systems, colors, symbols, words and shadings"** (Tufte, 2001). Such wording reflects the structural, mathematical aspect of data visualizations itself, mainly about quantitative information. The definition' connotation is quite similar to Bertin's graphics classification theory (Bertin, 1967), which will be reviewed in chapter 2.8. Another, more general, but still worth mentioning definition' example is given by Engelhardt in his "The language of graphics"; he speaks there broadly about *a graphical representation*, which is described as **"a visible artifact on a more or less flat surface that was intentionally created to express an information"** (Engelhardt, 2002). Engelhardt genuinely mentions the aspect' duet that shapes graphical representation' term — two-dimensionality (which is referred as to "escaping flatland" notions (Tufte, 1990)) and purposefulness. A definition of graphics by Levi can be valuable here; he outlines it as **"translations of numbers in the form of a drawing, design or plan to explain or illustrate something"** (Lewi, 2008). The verbs *explain* and *illustrate* express a direct purposeful action that graphics executes. Same connotations are followed by an influential data-journalist and designer Cairo, when he speaks about *a visualization*, described as **"any kind of visual representation of information designed to enable communication, analysis, discovery, exploration, etc."** (Cairo, 2016). Cairo precisely notions the visualization in a social dimension, including communication. In this matter, a concept of *infographics* often refers to communication; e.g., Laptev defines *infographics* as **a field in communicative design which based on a graphical representation of information, connections, numeric datasets and knowledge** (Laptev, 2012). Information graphics are considered here as a subsidiary field inside of another discipline, and also refers to the graphical representation by means of key data visualization elements.

Finally, a *data visualization* definition is given by Chiasson and Gregory, claiming it as **the process by which data are visualized, or presented, after the data cleaning process, and involves making choices about which data will be visualized, how data will be visualized, and what message will be shared with the target audience of the visualization** (Chiasson & Gregory, 2014). We see quite a bulky, but the comprehensive definition that attempts to cover all adjacent aspects of data visualization. Being technological, mathematical, artful, user-oriented, and complex within its essence, data visualization needs to be treated multi-dimensionally.

Back to the DIKW model, it gives a certain priority to information over data, so the same status should probably stay in the visualization-related terms, which is not exactly true. Since the 2010s were characterized by a broad popularisation of *big-data* concept, not only in science but rather in social and economical aspects, the term data visualization became a common popular expression for anything that deals with either information or knowledge or datasets that should be visually expressed. This also can be seen in the last decade through a significant increase in data visualization' mentioned among the academic publications when it is compared to, e.g., information visualization (Figure 2.2).

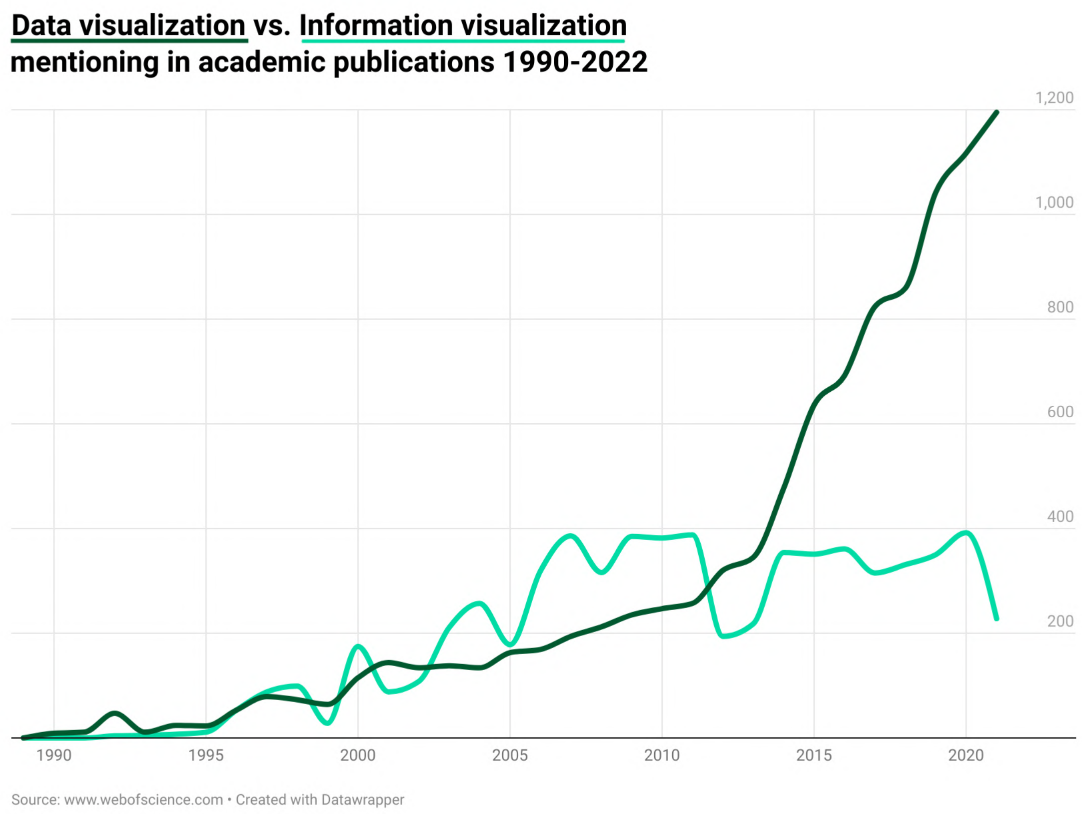

Figure 2.2: Dynamics of absolute number of publications mentioning data visualization and information visualization [Web of Science]

To summarise the presented definitions and take into consideration all aforementioned disparities in their wordings, it is possible to generally outline the specific areas that data visualization encompasses. Data visualization deals with **a translation of qualitative, quantitative, and spatio-temporal data into a correspondingly appropriate visual representation that might be relevant to its users.** From this we can identify these common aspects of data visualization:

1. Translation relates to the mediums and methods by which the conversion into the graphics is made.

2. Data, as it stands as a mathematical, statistical, geographical, and methodological unified background alongside computational algorithms and data collection.

3. Correspondence implies that certain data types can be potentially converted into certain visual representations only, as well as certain visualization should be decoded into respective data types or values.

4. Appropriate visual representation, as it relates to the graphical excellence, typography, device environment, visual rules, semiotics, and sociocultural context that data visualization is following or should follow.

5. Relevance for the users is expressing the social implication, e.g. how the data visualization helps, what insights are given, and what the perceptual basis stands behind it. This relevance can be expressed in the question of "why does the data visualization matter?".

We may also identify two logical outer-inner characteristics of data visualization as a concept:

6. Data visualization as a whole, having all aforementioned aspects been united as a solitary process or system, having contributors involved in decision-making and feedback.

7. And **"A look from above"** side, when data visualization itself is seen in the historical, social, and philosophical space of other terms and concepts, it is possible to analyze its global and local influences.

These aspects' divisions can serve as a background for a proper formulation and understanding of data visualization terms in the future, as well as it can serve as an appropriate framework for the literature review in the following chapters.

2.3 Data representation

A huge role in data visualization is played directly by the types, forms, views, or concisely, *a data representation.* Identical terms and definitions disparity can be seen in this field respectively. Although simply said, we can focus on the fact that **how data is presented, and structured affects the way how it can be visualized.** For example, the possible arrangement of some elements along the horizontal axis is formally obvious, if something is happening over time, or mapping can be a good solution if things are happening somewhere.

Several examples of data representation' taxonomy considerations can be reviewed (Munzner, 2014), (Chiasson & Gregory, 2014), (Tory & Möller, 2002), (Ware, 2019).

According to the aforementioned works on the Semiology of Graphics in chapter 2.2 (Bertin, 1967), two fundamental data forms of data values and data structures might be distinguished. Such a definition has refactored into the proposition of entities and relations with coherent attributes, which rather had been applied in the database theory first and then got adopted for data visualization. Ware considers this taxonomy, claiming an entity as a general object of interest. Something that connects the entities, and relates them is specified as a relationship (Ware, 2019). Both interest and relationships inherently contain attributes, and properties (e.g., if a pear is considered as an entity, then the color is its attribute; the same way a pear tree has related to pear through an attribute of parenting, etc.). Further, Ware defines next the dimensional complexity of the data, which is expressing how complex the entity is concerning its given attributes. The dimensional multiplicity evokes the need for another categorization, which is expressed through types of numbers: category data, integer data, and real-number data.

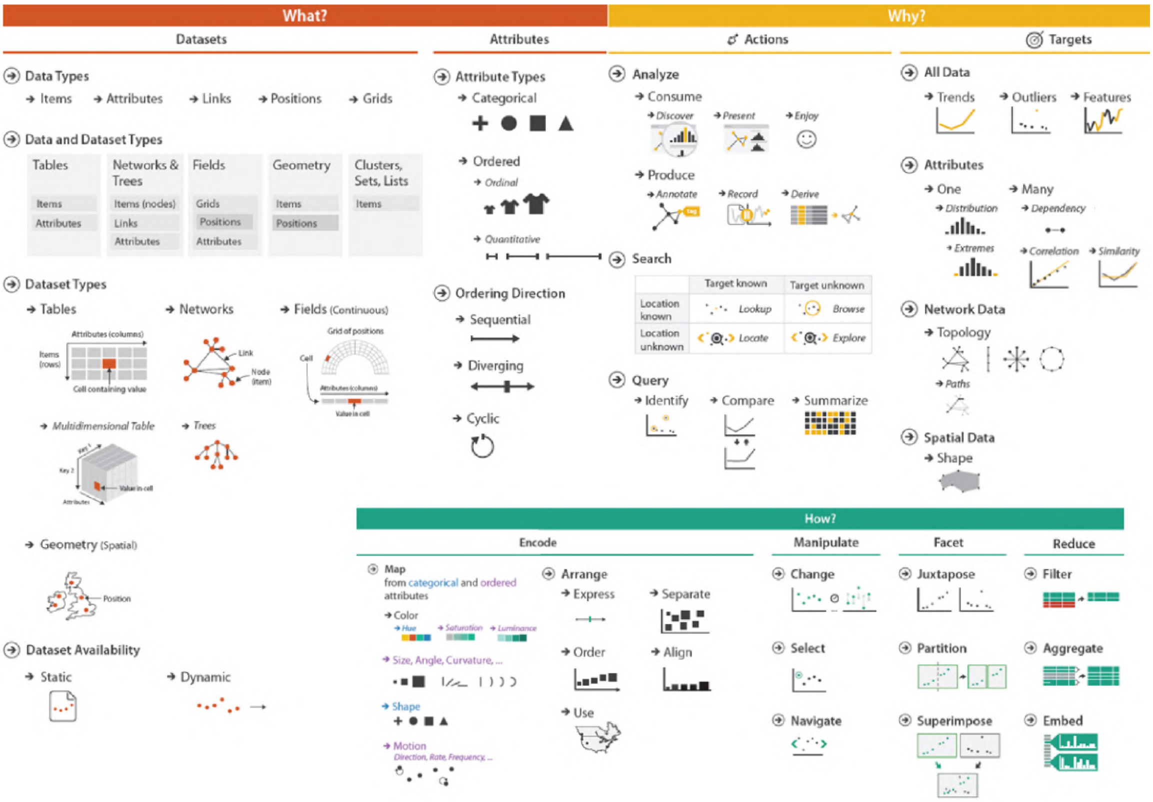

Figure 2.3: "Why, what and how can be visualized" [Munzner (2014)]

Additionally, Tory and Möller with respect to other research (Ware, 2019) identify the data characteristics through several dependent and independent variables and their distinct type for each: firstly, whether it is scalar, vector, tensor, or another, and whether it is discrete or continuous (Tory & Möller, 2002). Such an approach is familiar to the users of Tableau Software, which is mentioned in the following chapter.

Quite a comprehensive view on the matter of data representation is provided by Munzner, as she considers the data representation taxonomy differently, providing the big picture with involved data and dataset types, availability, attribute type, and ordering direction (Munzner, 2014). She demonstrates illustrative pictures, answering on *what* can be visualized, *why* and *how* (Figure 2.3). Munzner is emphasizing that the data type is representing its structural or mathematical interpretation; regarding the datasets, it is rather *the approach* that combines these data types into bigger compositions. Moreover, actions and targets can stand as reasoning for working with data, bringing the encoding, manipulations, facets and reduction as the arranging methods.

Tables, networks (trees), fields, and geometries are the four fundamental dataset type categories; clusters, sets, and lists are rather derivative dataset types. Usually, all of them are made up of different arrangements of the following five types of data: elements, attributes, links, positions, and grids. Datasets might be available whether as a static file or dynamic source which is processed and updated instantly. Attribute types (referred to as

Table 2.2: Levels of Measurement Stevens (1946)

Property	Level of Measurement	Measure Property	Math Operator	Central Tendency	Variability
	Nominal	Classification, Membership	= $\neq$	Mode	Qualitative Variation
Qualitative	Ordinal	Comparison, Level	$>$ $<$	Median	Interquartile Range
	Interval	Difference, Affinity	+ -	Mean	Deviation
Quantitative	Ratio	Magnitude, Amout	$\times$ $\div$	Geometric Mean	Coefficient of variation

types of numbers by Ware) and narrowed to categorical and ordered types, where the final one can be either quantitative or ordinal. The ordering direction may be circular, divergent, or sequential.

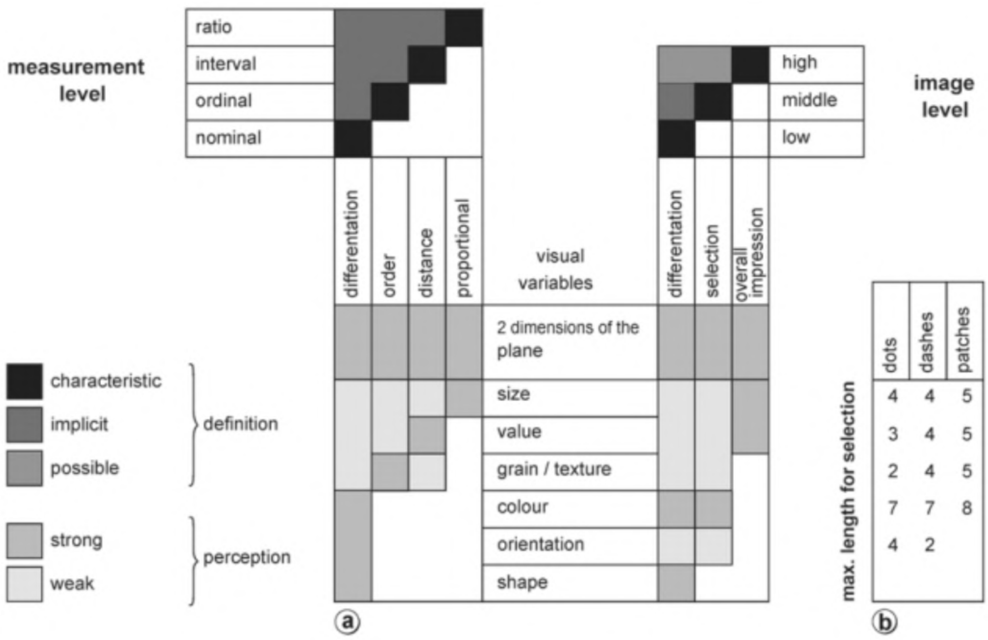

Figure 2.4: Cartographic information analysis [Kraak and Ormeling (2020)]

The reviewed aspects of "basic data types" (Chiasson & Gregory, 2014) or "attribute types" are rather known as levels of measurement, **which describe the nature of the information within the values assigned to variables** (Kirch, 2008). This data's feature is a special focus of a current study, as this allows us to understand the *who* and *what* of the data, as well as to properly define the following corresponding visualization. The foremost adopted method was introduced by Stevens, as he proposed four distinctive incremental groups, presented and described through their characteristics at Table 2.2 (Stevens, 1946). Multiple social sciences are still seriously depending on Stevens' methodology, although it was respectively criticized, as well as edited by Stevens himself (Chrisman, 1998). Nonetheless, such a classification of attributes' semantics is still in use by data visualization authors and cartographers 2.4.

The scale of measurement itself represents a subject that changes, therefore allowing the data to be transformed. The elemental task of data transformation can be expressed by emphasizing, generalizing, and revealing (Zhou & Feiner, 1998). This particular aspect of data representation will remain in focus for the visualization recommender system in chapter 4.

2.4 Tools, software and mediums

The translation part of the data visualization as the concept can be reviewed through the available and popular mediums by means of which the data visualization is implemented. Since the contemporary focus of data visualization is highly characterized by its computer-based nature, the key graphic realization of the data is mainly made through the "computational methods, transforming the symbolic into geometric, enabling researchers to observe their simulations and computations" (DeFanti, Brown, & McCormick, 1989). The variety of the phenomenons and methodologies provide the competitive basis for the advancing of the visualization tools as products.

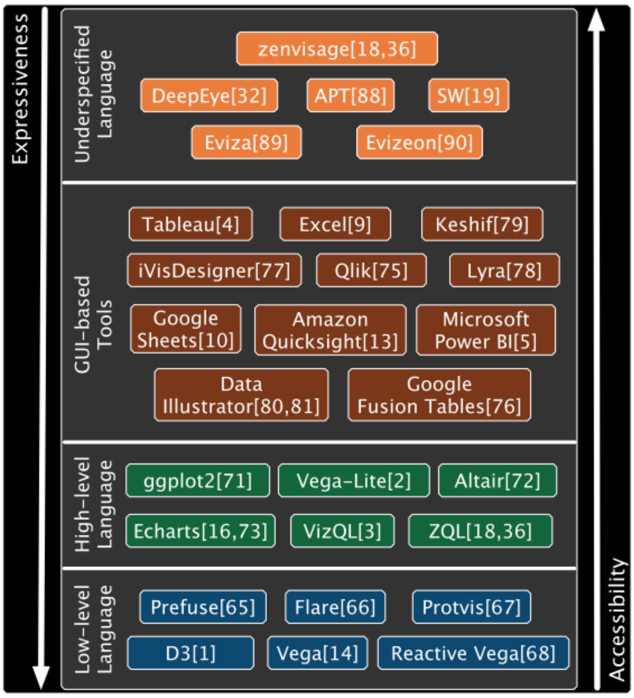

Figure 2.5: Several data visualization tools with regards to their usability specifications [Qin et al. (2020)]

The distribution of the dataviz tools' popularity is uneven due to the multiple affections, but general market competition trends can be recognized. Seeing the Figure 2.7, such tools as *Pen Paper* with *Other physical materials* also remind us that data visualization can be created by means not only the software but even more extravagant mediums. Technically, we are dealing with a definite variety of programs or instruments that data visualization specialist is picking to realize the fullness of their potential according to the initial needs.

Annual Data Visualization Society (DVS) surveys results and their visualizations are demonstrating the feasible range of influential factors (DVS, 2019,2020):

1. Associated areas and task types are representing the professional basis where the data visualization is needed. Depending on the specialist's work field, the medium option can be picked depending on different cases, especially the key role (Figure 2.6).

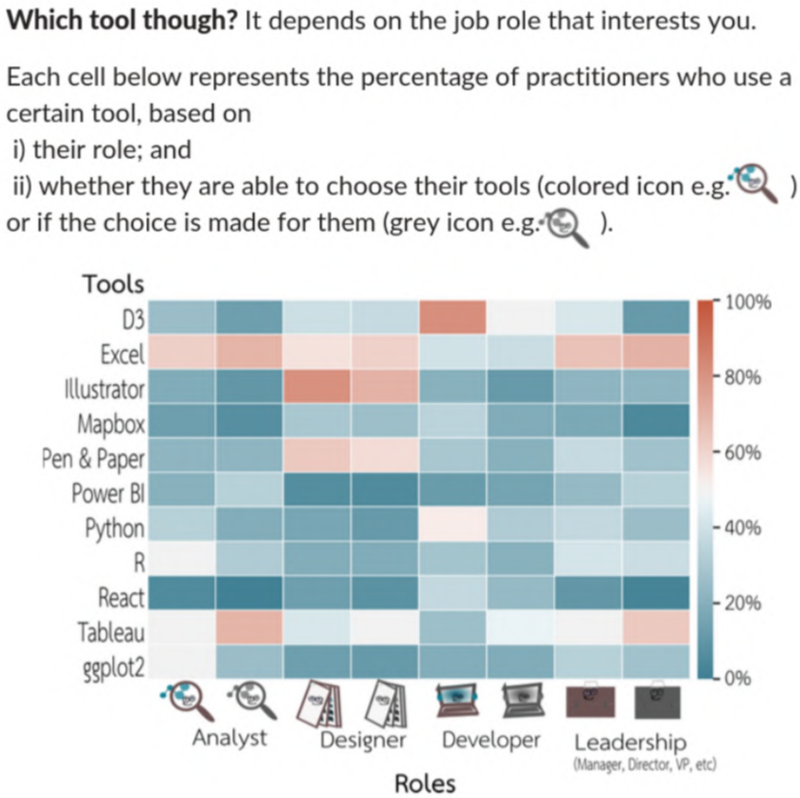

Figure 2.6: Decisions on the data visualization tool with respect to the role of the specialist Khoo (2019)

E.g., the analysts tend to seek analytical software first (Excel, Tableau) to conduct the calculations and explore the data, while graphic designers would pick art-related tools (Canva, Adobe Illustrator) to present visually appealing infographics or reports. Likewise, web developers are likely to choose the web tools (D3, VegaLite) to insert certain interactivity into the graphics. Finally, cartographers or GIS specialists surely concentrate on the geovisualization packages (MapBox, ArcGIS).

2. Functionality is smoothly translating from the first factor; although it was explained on the matter of associated areas, the program packages can fit the certain roles of specialists, but might not respond to the functional needs (e.g., seamless compatibility with certain file types, web embedding option, etc.). In Figure 2.8 this can be seen that certain instruments are still reported as popular among multiple associated areas.

3. Learning curve and usage complexity represents the skill-related issue that can refactor the final choice of the software. E.g., R Studio, being a powerful, open-source instrument for data analysis, cleaning, and visualization might not be in favor for some due

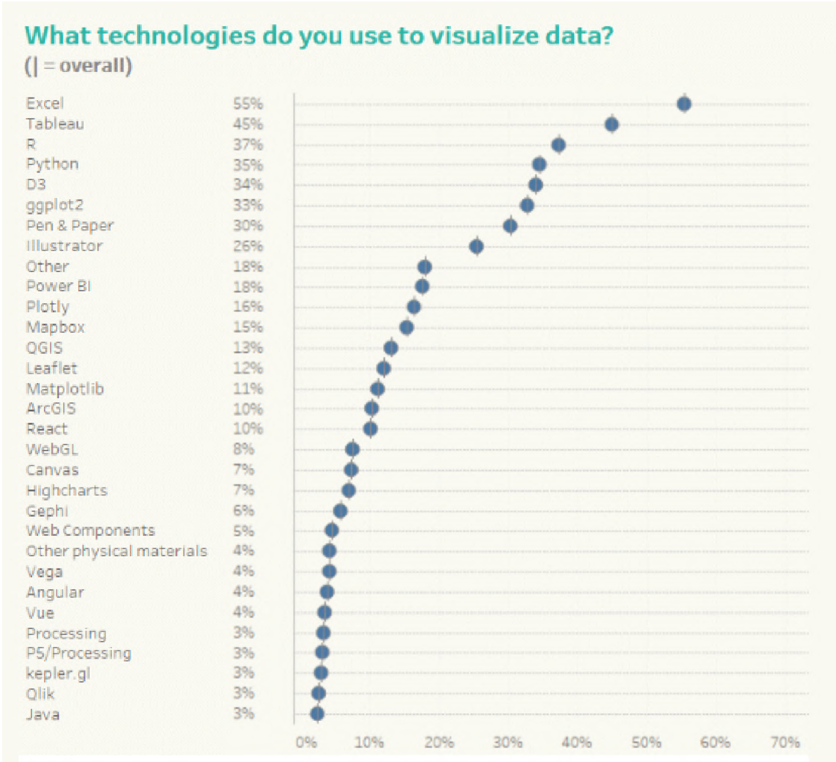

Figure 2.7: The popularity of several data visualization tools [Wexler (2019)]

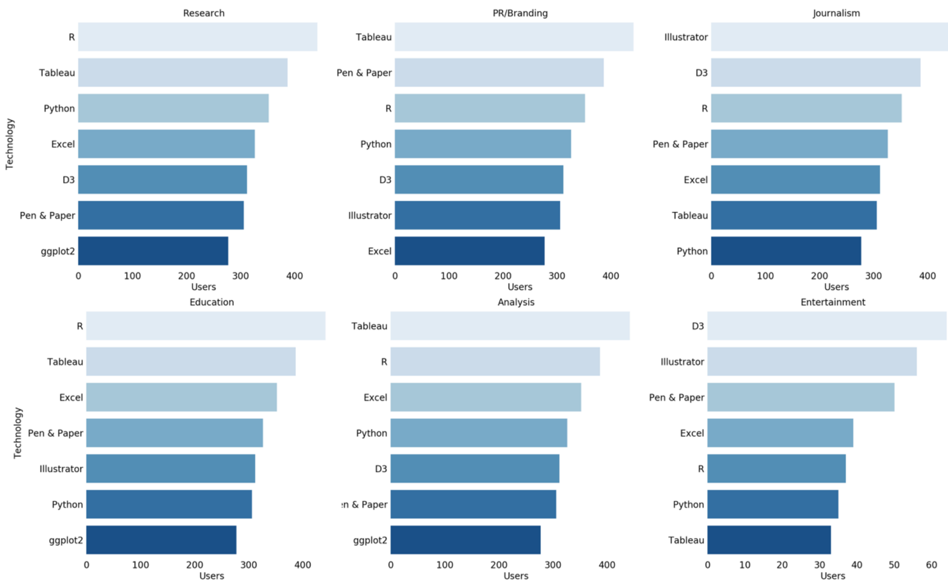

Figure 2.8: The dataviz tools popularity, aggregated by the associated area Garrod (2020)

to the code-based working process. Likewise, Tableau might be ignored by some dataviz-related specialists due to its specific data interpretation and visualization implementation.

4. Finally, **an availability** is strongly determined by the market positioning of the tool, expressed in its licensing. An illustrative example is a direct comparison between ArcGIS and QGIS for geospatial visualization needs. Both are cutting-edge geospatial packages, sharing the dissimilarity in this aspect. While QGIS is open-source, free-to-use software, ESRI is targeting the commercial realization of its GIS products, which makes it not easily available for smaller teams or individuals.

Thus, the proper visualization tool for the data visualization specialists should fit their field of work, have an essential functionality for the task specifications, respond to the skill level (or to the estimated educability) and be financially suitable. According to the reviewed survey results, the instruments such as Tableau, Microsoft Excel, RAWGraphs, Google Studio, Datawrapper, R Studio, Flourish, QGIS, D3.js, and Python-based data visualization libraries are seemingly in high demand. Table 2.3 contains the highlights of some data visualization software.

Table 2.3: Several DataViz Software with Functionality and Pricing

DataViz Software	Major Functions	Costs
Tableau	Creating interactive charts and maps with data	Tableau Public is Free. Tableau Desktop is also free through education license for students and instructors.
Microsoft Excel	Preparing and cleaning dataset	Part of Microsoft Office
Google Studio	Cleaning dataset and creating charts	Free.
Datawrapper	Creating interactive charts and maps with data	Free version is good for educational purposes.
R Studio	Creating interactive charts and maps with data (coding skill is required)	Desktop version is free and good for educational use.
Flourish	Creating interactive charts and maps	Free version is good for educational use.
iNZight	Data cleaning and exploration	Free and open-source.
BatchGeo Pro	Creating maps	99$ per month.
ArcGIS	Creating maps	100$ per year.
QGIS	Exploring data; creating, editing and managing data; publishing maps	Free and Open-Source.
Infogram	Creating interactive charts, maps and infographics	Free for students and educators under education license.
D3.js	Creating interactive charts and maps (JavaScript/coding skill is required)	Free and open-source.
Storyline.js	Creating annotated and interactive line charts	Free and open-source.
Canva	Creating infographics	Free version is limited, pro version is 12.95$ per month for 5 users

2.5 Correspondence: data visualization classification

To handle the diagrams or types of maps systematically, it is useful to analyze them on the matter of some specific features that could make them appropriate; starting from the mid-twentieth century, multiple academic authors were proposing *deconstructive* theories on the graphics. The pioneer ideas in this matter were described in Semiology of Graphics (Bertin, 1977). He proposed the set of graphic variables that can represent a particular encoded value or category; Bertin discerns 6 possible differences to consider: size, lightness or value, grain or texture, hue, orientation, and shape. Such variables exist not only as a sole entity but rather imply multiple recommendations on the use of each. As it was demonstrated back in Figure 2.4, Bertin's visual variables are corresponding to the concrete levels of measurement, e.g. encoding nominal (qualitative) values with color, shape, and orientation. Bertin's approach in the deconstruction of the graphic was supported by multiple followed publications afterwards (Twyman, 1979), (Tversky, 1995), (Richards, 1984), (Card & Mackinlay, 1997).

Twyman's *Schema for the Study of Graphic Language* was focused on a layout configuration of the textual and graphical elements. The matrix (Figure 2.9) was intended to illustrate the wide range of approaches open in graphic language and the effects on reading and viewing strategies as well as cognitive processes. Such an approach, although, does not consider the habitual data visualization types as such, but rather focuses on the grouping and positioning parameters.

Method of configuration

Mode of symbolization	Pure linear	Linear interrupted	List	Linear branching	Matrix	Non-linear directed viewing	Non-linear most options open
Verbal/ numerical	1	2	3	4	5	6	7
Pictorial & verbal/ numerical	8	9	10	11	12	13	14
Pictorial	15	16	17	18	19	20	21
Schematic	22	23	24	25	26	27	28

Figure 2.9: Twyman's matrix of the visualization configuration method [Twyman (1979)]

Interesting vision was proposed by Card and MacKinlay, as they define the individual image elements as variables that can be described from the point of controlled, automatic and interactive parameters. According to the framework, all kinds of images is possibly to deconstruct in this way to highlight the common features of the data visualization. Figure 2.10 represents the possible example on how the graphics can be described within the proposed conceptual rules.

Conceptual framework of The 'DNA of visualization' can be reviewed as one of the recent advancements on the matter of visual grammar and syntax (Engelhardt & Richards, 2018). The framework is intended to help designers to generate visualization options; The definition of the fundamental building blocks, or syntactic constituents opens the door for various combinatorial possibilities. A visual representation of data in a framework is considered as the expressing meaning by way of graphic relationships between graphic components. Graphic components may vary (e.g., shapes, lines, symbols, glyphs) and may be organized into graphic relationships by certain principles. There are types of information that are strictly assigned to definite visual encodings, which then are expressed through visual components and supportive means of layout principles, directions and reference elements (Figure 2.11.

Data				*Controlled*		*Automatic*					*Interaction*	
Variable	D	F	D’	CP	M	R	X	Y	Z	T	V	W

Symbol *Meaning*

Variable Name of case or variable dimension

D Data Type ::= N (Nominal),
O (Ordinal),
Q (Quantitative).
Q_x (Intrinsically spatial),
Q_{lon} (Geographical)
NxN (Set mapped to itself - graphs)

F Function for recoding data ::= f (unspecified)
$>$ (filter)
s (sorting)
mds (multidimensional scaling)
↑ (interactive input to a function)

D’ Recoded Data Type (see D)

CP Control Processing tx (text)

M Mark types ::= P (Point), L (Line),
S (Surface), A (Area), V (Volume)

R Retinal properties ::= C (Color), S (Size),
— (Connection), [] (Enclosure)

XYZT Position in space time ::= N, O, Q,
* (non-semantic use of space-time)

V View transformation ::=hb(hyperbolic mapping)

W Widget ::= sl(slider) rb(radio buttons)

Example: Profit Landscape

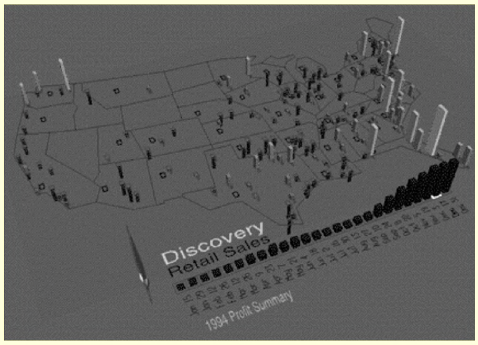

Image Description

Variable	D	F	D’	CP	M	R	X	Y	Z	T	V	W
Offices					L							
Lon.	Q_{lon}						Q					
Lat.	Q_{lat}							Q				
Profit	Q					Sz			Q			
		f	N			C						

Figure 2.10: Card and MacKinlay framework example [Card and Mackinlay (1997)]

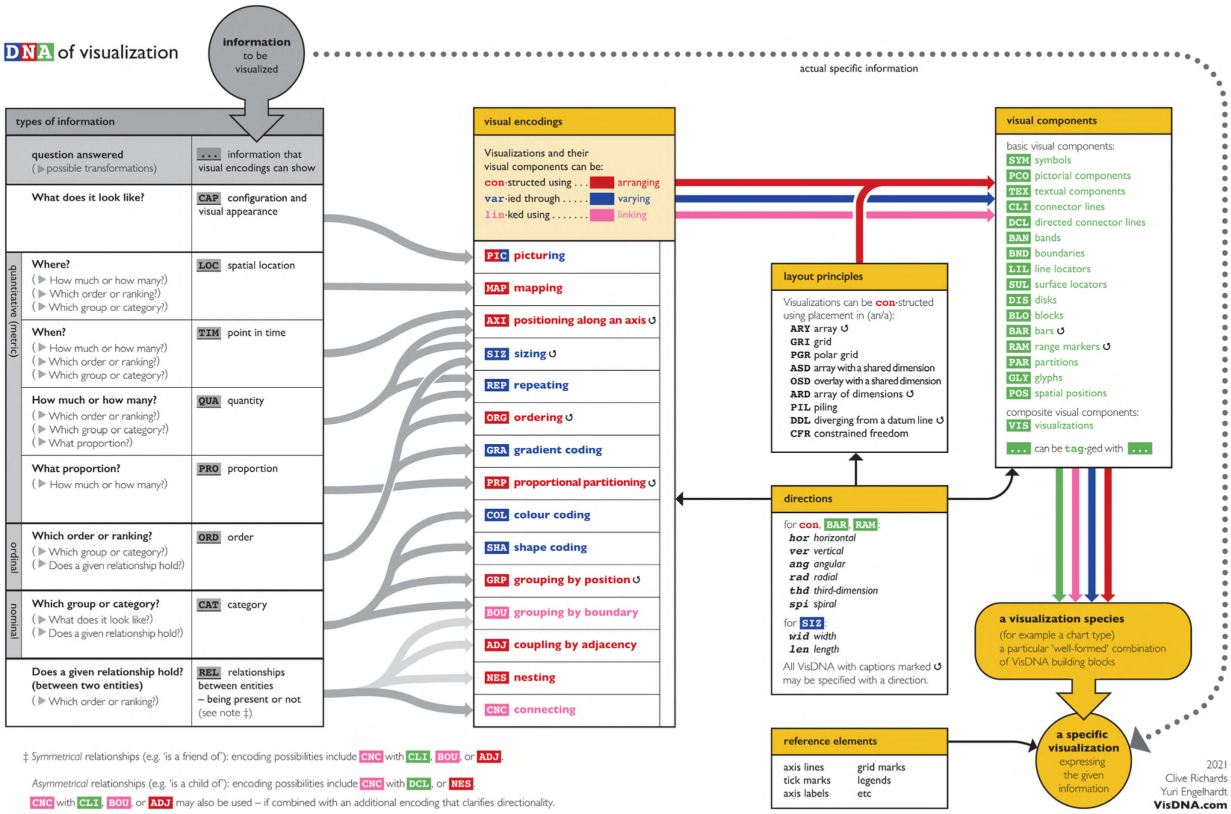

Figure 2.11: The DNA of visualization framework [Engelhardt and Richards (2021)]

Figure 2.12 contains the vivid example of the framework application, describing a pie chart as a distinctive specimen. Firstly it is identifying the visual encodings that stands out as the main method and then deconstruct the graphics further on the matter of used visual components. We see that the pie chart is described here pretty clear and straightforward, which means that by doing so with the vast variety of multiple specimens, it would be possible to distinguish the visualization methods more systematically.

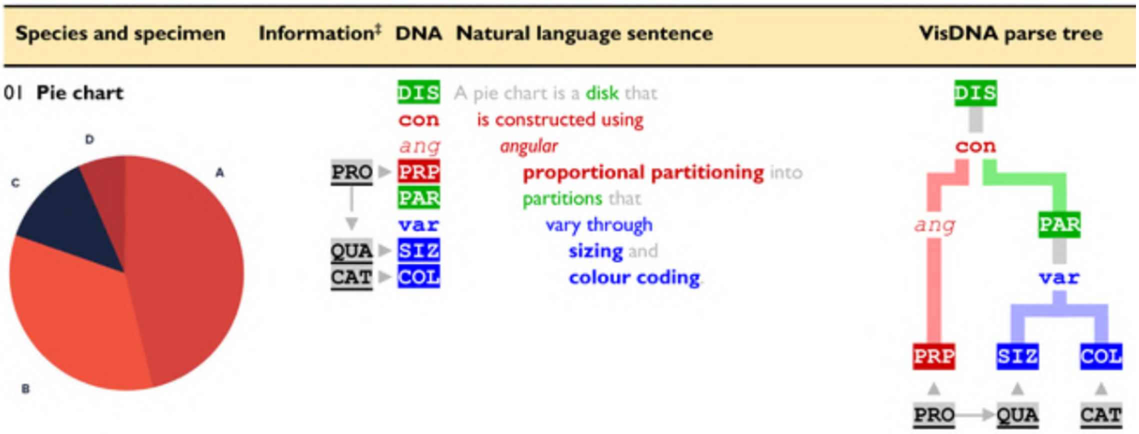

Figure 2.12: The example of pie chart been described by means of DNA of visualization syntax [Engelhardt and Richards (2021)]

The reviewed examples of so-called *graphic taxonomies, syntax* or *visual grammars* can be regarded as academic-based theories or conceptual methodologies that can substantiate the selection of certain visualization types. They are reasonably well-described and analyzed, although might not be well understood by most of the users straightaway. In this regard, it is interesting to review some arbitrary *enthusiastic* frameworks for the proper data visualization types picking. Most of them have a vivid, graphic nature, bringing the possible options at once depending on certain filters or users' needs.

Very enthusiastic approach is made in *A periodic table of data visualization methods* (Figure 2.13). Although the main inspiration for the arrangement of the graphic examples is taken from Mendeleev's table, this representation does not follow the inner periodic laws; it is used to present the huge variety of visualization methods clearly, proposing the associative metaphor with the periodic table. Such a view allows for to investigate of multiple methods on a low, generic level of classification.

The other example is discerning the visualization types by the answer on a *what you would like to show* question. It implies that certain visualization methods fit better for a particular data presentation task, e.g. changes over time or inner distribution of the dataset. The simplified set of the visualization methods was classified by the Financial Times Graphic Team (Financial Times, 2021). The highlighted categories of deviation, correlation, ranking, distribution, change over time, magnitude, part-to-whole, spatial and flow contain the drawn icons with the corresponding specimens and a short description (Figure 2.14).

One of the most comprehensive classifications and collections of the data visualization methods is provided by Ferdio with their DataVizProject, containing over 100 different types

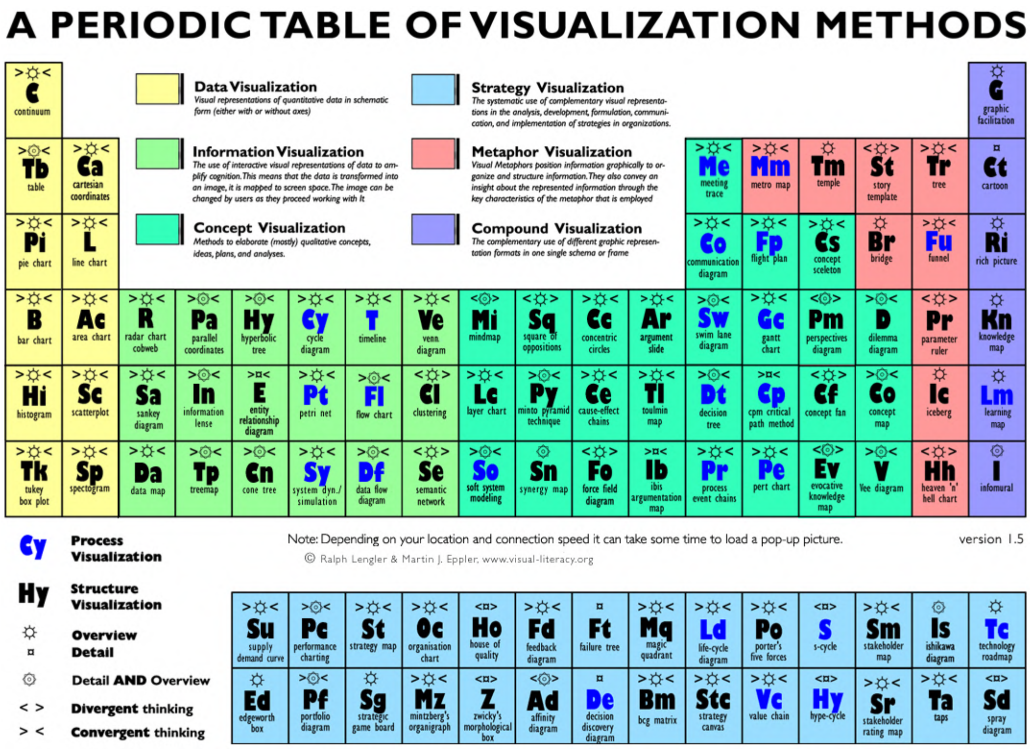

Figure 2.13: A periodic table of visualization methods [Visual-literacy.org (2007)]

Figure 2.14: A snippet from FT visual vocabulary guide [Financial Times (2021)]

and corresponding examples for each one. Family, input, function, and shape are the main collection item categories; The enabled filters allow us to find the needful visualization types more properly on a deeper level (Figure 2.15).

Therefore, the provided references on the different taxonomies and data visualization collections demonstrate that decent work has been made to systematize the data visualization types as such. We can highlight that the conceptual classification of the visualization types allows us to pick the right graph or diagram faster and more effectively, giving the user instant examples of appropriate specimen options.

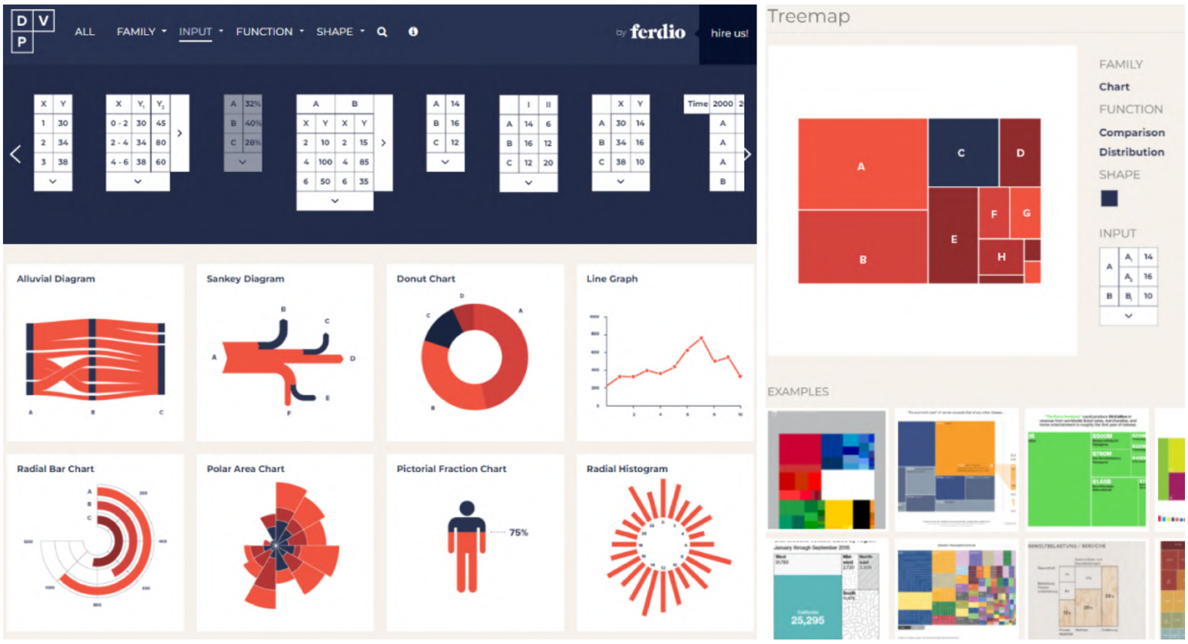

Figure 2.15: The DataVizProject website with Treemap example in detail [Data Viz Project by Ferdio (2022)]

2.6 Aesthetic, graphical and art principles

Probably, one of the most *obvious* influential aspects of data visualization is narrowed to its visual figuration and shape. The corresponding term of *graphic design* can be involved relevantly here. Although, still graphic design inherently keeps multiple interpretations. E.g., it is considered as **an activity that enables and organizes visual communication in a society** (Frascara, 1988). Indeed, such an approach considers graphic design as a problem-solving tool, where the accumulated visual expressions' experiences and research is used for decision-making. They can be shortly traced to **layout principles, typography** and **colour systematization;** in cartography these aspects are usually at foremost important in the considerations on map design (Tyner, 2014).

Multiple works are specifying and identify the character of the layout for certain design decisions (Muehlenhaus, 2013), (Djonov & Leeuwen, 2013). There is a definite focus on the fact that the design layout likely should follow the functional purpose of the communication that is intended to be implemented. Leborg systematically describes the possible design considerations of the layout principles through *Abstract, Concrete, Activities, and Relations* classes, dividing them in detail deeper (Leborg, 2006). E.g., in the case with Relations, he defines plural distinctive variations on how the abstract elements might visually interact on a 2D plane 2.16.

The prominent concepts of clarity, order, contrast, balance, unity, and harmony are the ones that are considered by some as the goals of a good data visualization or map design (Tyner, 2014). Mainly, such layout concepts are related to better and faster reading alongside the perceptual aesthetics enjoyment from *graphical excellence.* An importance of the design

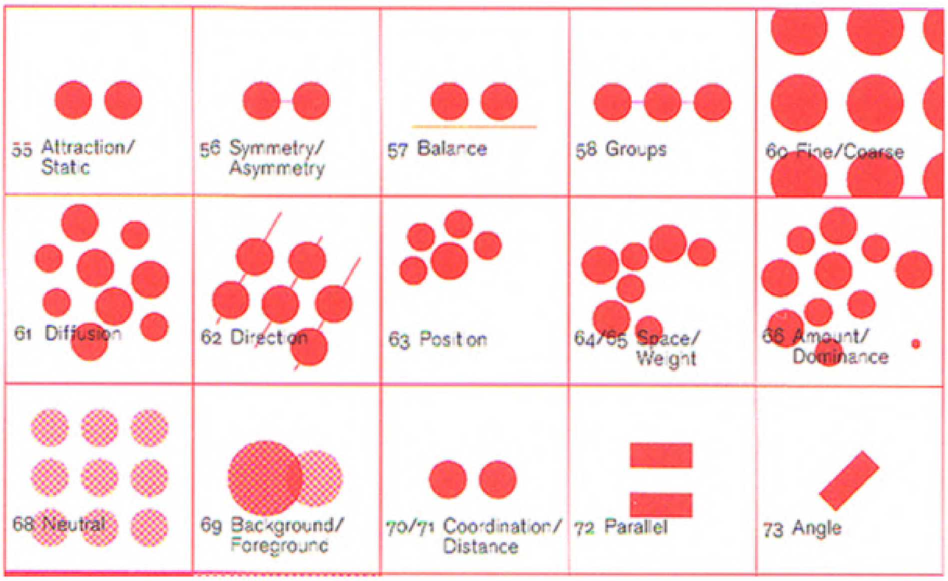

Figure 2.16: Several relations examples for the visual components [Leborg, 2006]

decision on the layout of the element stays actual due to multiple use environments and enriched capabilities of data integration for the reports and, e.g., business intelligence.

The other component of graphical principles for data visualization is embodied in the textual side, as symbols and words can label, explain, direct, engage and involve in a hierarchy of concepts (Tyner, 2014). This is not only about the wording contents of the data visualization, but rather it is the ”instrumental” character that shapes the graphical expression. The terms of typefaces, fonts, lettering, spacing, kerning and punctuation are valuable and important parameters to the text (Korolkova, 2007); respective importance is expressed for the data visualizations. Labeling and annotation are relevant to accent notable patterns, trends, and anomalies, as a clear graphic association becomes the key goal of the textual components (Kraak et al., 2020).

Since human eyes are specifically sensitive to the matter of light and are capable of distinguishing around a million grades, the challenge of color is something that can be thought of as beneficial for data visualization (Tufte, 1990). Certain scientific work with respect to the colour systematization was done and still in process (Itten, 1970), (Smith, 1978), (Fairchild, 2013) (Muth, 2022). The color itself, being a derivative of light, represents itself as a complex phenomenon with multiple factors to consider when it is used for further purposes, especially concerning the data visualization design. The factors of the involved use environment (paper, screen, matrix type), selected color model (HSV, HSL, RGB), color schema (diverging, sequential), cultural context, and users' readability due to the color blindness and disruptions makes it extremely complex to properly use this aspect for the data visualization (Kraak et al., 2020).

Concerning data visualizations, there are a large number of views on how to involve all aforementioned graphic elements or design in the work. Surely, the historical perspective connects the cultural and epoch features with its prominent visual examples.

Back to Playfair or Minard times, this was a period of great development of classical statistics and mathematics and their accompanying graphical methods. The visualizations are embodied by hand on paper, color schemes depended largely on the tools that are available rulers, pens, pencils, or paints. The typographic features are largely adopted from the available and understandable typefaces (in particular, neat calligraphic handwritings, serifs, and ligatures). Finally, the general approach may also have been dictated by the philosophical views of the authors (for example, Minard's focus on the French' Army losses precisely through the thickness of the lines, highlighting especially the dark color of the troops who left defeated (Kraak, 2014)).

Our days, dictated by the data-technology revolution, are generating a different demand for the visual design of information. Contemporary approaches for data visualization design is focusing on web development, interactive operators, and user-centered design, but still involve the aforementioned layout, colors, and typography principles. Nevertheless, as it was mentioned at the beginning, graphic embodiment, graphic design is largely aimed at solving some task or problem, no matter in what historical period we live. In some cases, the task description can be not as straightforward as it may seem to be. Possible visual distortions and experiments may lead to extravagant graphic decisions. In this respect, one may recall Edward Tufte versus Nigel Holmes overlooks the visualizations. While Tufte's approach is proposing to eliminate the so-called chartjunk and all irrelevant graphical excesses on plots, maps, and graphs, Holmes' extravagant visualizations for the British press in the 1980s and 1990s were containing contextual pictures, exactly contrasting to the ET's approaches 2.17.

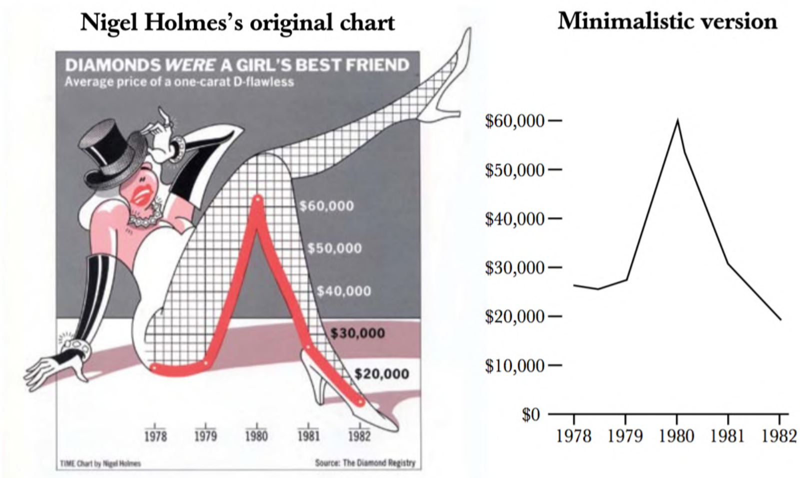

Figure 2.17: Two diffrent approaches on a chart' visual performance [Beautiful Design]

The emotional, eye-catching nature of visualizations can also, to a certain extent, be its goal and essence. In this respect, graphic designers in their educational studying and work should see and develop themselves as psychologists rather than artists, because the quality in graphic design is measured by the changes it produces in the audience (Frascara, 1988). In this regard, it is interesting to look further at the data visualization's place in the socially relevant context.

2.7 Relevance, context and audience

Since we shed a little light on the communication aspect of graphic design, it is important to address the contextual features of data visualization which might make them more relevant for particularly involved parties of communication. The systematic, well-established scope of the usage context is representing a decent methodological challenge.

Basically, next aspects can be highlighted with respect to multiple references (Göker & Myrhaug, 2002), (MacEachren, 2004), (MacEachren & Kraak, 2001):

1. The deep neurological aspect represents itself the inner brain and mental specificities about the processing and understanding of visual information and data. In this regard we ask questions like ”what parts of the brain are getting activated when the reader is involved” or ”how the reader perceives the space and relations on a plane”, etc. (Reuter, Tukey, Maloney, Pani, & Smith, 1990). Multiple sample user studies can bring certain conclusions with approximate confidence, so then they can be applied to the whole population.

2. The cognitive accessibility aspect is involving higher-levels of perceptive features, such as specific skills of literacy, numeracy, visual and spatial thinking, technology, and what is, more importantly, the motivation to investigate. Depending on this broad set of skills, the following decisions on the graphical and functional output can be made.

3. Professional context considers the readers and users from their specialization with the visualization. Cartography-cubed model, which expresses the map use tasks and map interactivity (can be reapplied to the matter of data visualization respectively) concerning the kinds of an audience is in good support here (Figure 2.18). Specialist audiences are small and trained to explore or confirm patterns in mapped datasets. These results are then synthesized into an explanation for visual presentation to general audiences that are wide and diverse (Kraak et al., 2020).

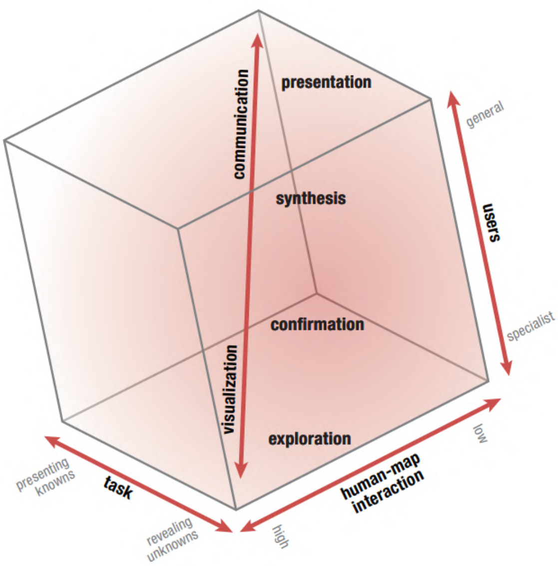

Figure 2.18: Cartography-cubed model organizes the broad map use [MacEachren et al. (1994)]

4. Sociological context considers the data visualization in relation to the bigger communal levels, comparing the relevance of wide population groups that are sharing same qualities. The differences in perception among different groups of age, sex, nation, race, material status, etc. can be involved in the decision-making for the visualization purpose and view. The known inclination of the color blindness for male readers rather than female, e.g., might be relevant example in this extent (National Eye Institute, 2015).

5. Cultural and geographic context connects the possible visual trends with the cultural once in a particular state of society around the globe. Obvious structural and culture differences among the countries might impact the data visualizations significantly. In this extent it is early Soviet graphics that can be commented and considered with respect to the cultural peculiarities. 1920-1930s Soviet infographics' were characterized by a huge influence of Vienna Development Method and eccentric socialistic thrust (Laptev, 2018). The immutable attribute of such graphics was an ideological red color, allowed to grasp intense focus on the communistic advances with comparison to, e.g., capitalistic states (Figure 2.19).

Figure 2.19: "The Red Army's defence power is ahead of the capitalistic countries' armies" — Soviet infographics demonstrates the importance of *cultural context* for the information visualizations [Kurganov, ISOSTAT, 1939]

6. Generic environment and spatio-temporal context itself is including the physical surrounding of the reader, involving the categories of time, location, direction, speed, temperature, etc. into the factors of relevance for the readers. Indeed, these parameters are highly important for the visualization interaction. The outdoor poster allocation of data visualization would have a different contextual complexity compare to the phone's screen artworks. The road signs, considered as the information graphics in a certain sense, are developed with the consideration of the potential read time with the allowed moving speed.

Thus, multiple contextual types are involving important parameters, forming the potential relevance for particular population groups and individuals. Accounting of all involved factors is rather presents a stochastic model with dedicated level of confidence. The feedback process, therefore, may allow to improve the contextual understanding behind the data visualization usage.

2.8 Data visualization as a process and system

As we can see, data visualization involves multiple important components. The technical, scientific, artistic, social, and other aspects must be organized to produce data visualizations. The involved components' connection and workflow conceptions are various here, but certain examples can be reviewed. Figure 2.20 contains five distinctive examples of what can be referred as *data visualization process, cartographic process* or *data visualization pipeline.* These examples do not necessarily mean the same steps and workflow; they are demonstrated to highlight the procedural, systematic side of the data visualization so that it should be treated as a complex concept. We may identify how such a process can be constructed.

Under **(a)** the example by Tyner is brought here. It is specified with the stages of planning, data analysis, presentation, and editing concerning the cartographic production (Tyner, 2014). This short, generic high-to-low schema of the mapping (or visualization) process is more focused on the creative side of the issue; almost the same approach is proposed in the aforementioned Mapping For a Sustainable World (Kraak et al., 2020), bringing the cartographic workflow example **(e).** Here we see already a cyclical character of the process, as the evaluation and editing part may influence the project goals and decisions. It implies that the users' feedback, an important aspect of user-centered design, is permanently taken into consideration when the project is planned. The schema's complexity like **(a)** and **(b)** may vary as the specific project plan will be employed based on the given mapping context. These examples, although primarily applied to the maps, are showing a little, common complexity of the data visualization creation process itself.

Tory and Möller **(c)** mention *the visualization process* as a loop, when the data gets rendered by the means of graphical software and interface, then gets perceived in the visual form by the user; from the user's side, there is a direct interaction with the visualization and the actual retrieval of the data (Tory & Moller, 2004). This schema is mostly about the user-vis interaction, it avoids the design process' reference, as it was applied for **(a)** and **(b)** examples. Similar process is presented on the example **(d)** from *The value of visualization*; here ellipses denote processes that transform inputs to outputs and boxes denote *containers* (Van Wijk, 2005). Van Wijk's data visualization process proposes the gained knowledge as a model that depends on the actual image output, the current knowledge of the user, and the particular properties of the perception and cognition of the user. Moreover, the user may decide to adapt the specification of the visualization, based on his current knowledge, to explore the data further if it has inherent interactivity.

Such a view of the data visualization process does not necessarily focus on the way how the visualization looks and what kind of data is involved, but it allows us to see the user's place in the visualization process more concretely. An example **(e)** shares the same approach but contains a little bit more details, e.g. parallel knowledge-based system (Chen et al., 2009). The objectives of knowledge-assisted visualization pipelines include sharing domain knowledge among different users and reducing the burden upon users to acquire knowledge about complex visualization techniques. This particular aspect will be discussed

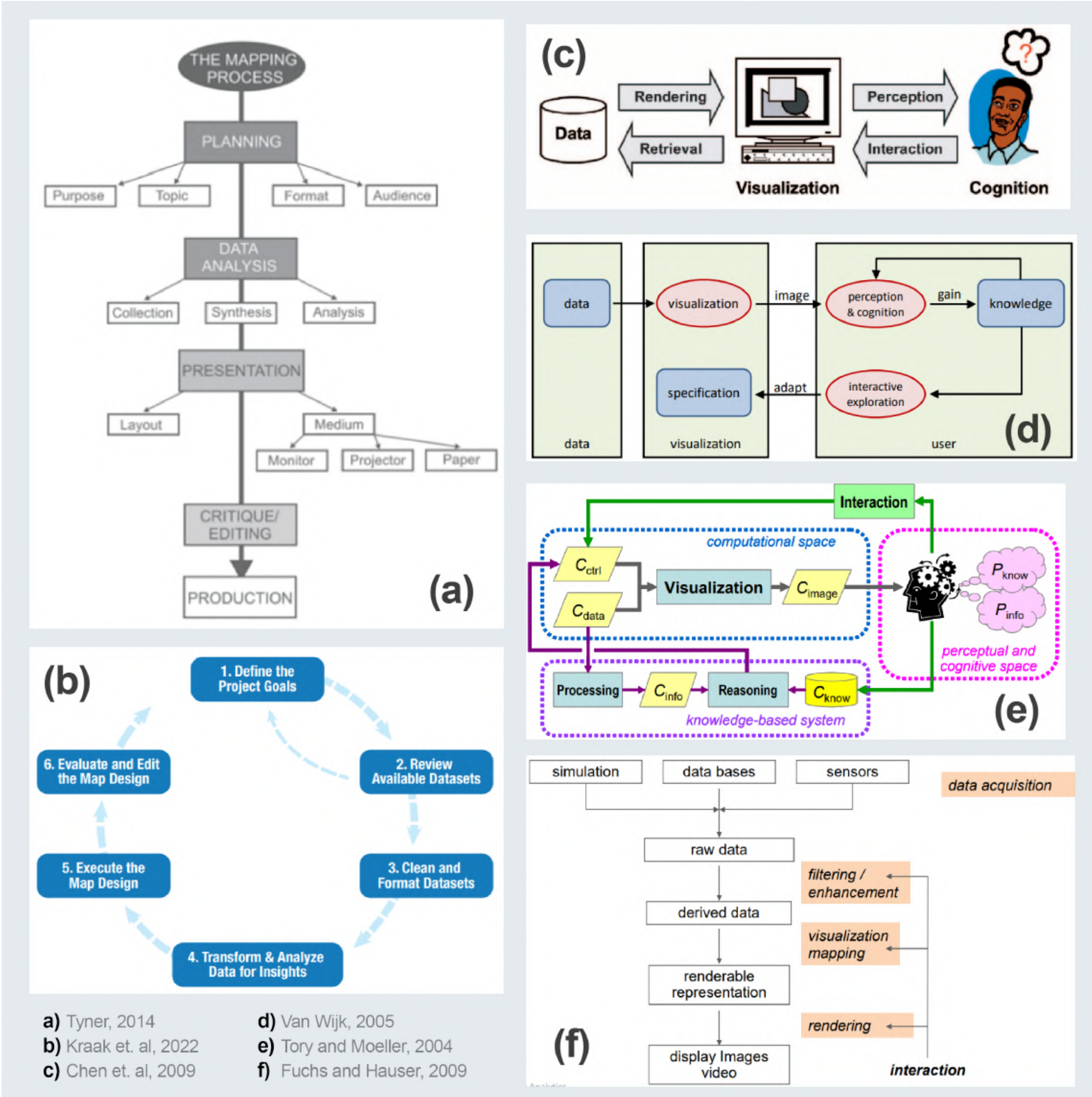

Figure 2.20: Several examples of data visualization (or mapping) process

in the next chapter.

The final example **(f)** on the data visualization pipeline can be also demonstrated as one of the generic views of what a data visualization pipeline can look like. Fuchs and Hauser propose such a pipeline, specifying several aspects (Fuchs & Hauser, 2009). Data acquisition is presented by the simulation, databases, and sensors; working with raw and derived data through filtering and enhancement implies the solutions to obtain useful data; finally, the renderable representation is made using *graphical primitives* (points, lines, etc.) and *visual channels* (color, texture, etc.). Although the users' concept is not directly involved in this example, such a data visualization pipeline involves the aforementioned in the previous sections aspects of data representation, correspondence, and graphical principles straightforwardly.

Therefore, we see that the data visualization process involves multiple entities and reasonable connections in between. Moreover, the process itself can be measured and treated on the matter of its effectiveness, considering the time and resources spent to create the visualization or how much knowledge was obtained by the user. One of the ways to implement the effectiveness into the data visualization pipeline is to incorporate the automatic elements that would save time and narrow down the possible variations in decision-making moments. The review of this issue is presented in the next section.

2.9 Automatic visualization and visualization recommendation systems

According to the recent comprehensive study on automatic visualization by Zhu, the recommender systems for visualization types can be developed by three methodologies (Zhu, Sun, Jiang, Zha, & Liang, 2020):

1. Data-driven model is involving the computational methods of Machine Learning (ML) and deep neural networks in the field of data visualization types. The visual elements of various charts or maps can get deconstructed, then quantified and qualified, so the ML-based model can be trained to predict visualizations, and visual elements are reconstructed to create recommendations ((Hu, Bakker, Li, Kraska, & Hidalgo, 2019)). The reviewed in chapter 2.8 approaches are useful here, as they may contain the elementary grammar structure that compiles the visualizations. The drawbacks of such an approach are plainly in the "black-boxed" specification of the ML algorithms.

One of the examples of data-driven model is presented with *Text-to-Viz* prototype that can create multiple data visualization outputs based on the wording inputs (Cui et al., 2019). Figure 2.21 arbitrary demonstrates the possible processed prompt' results.

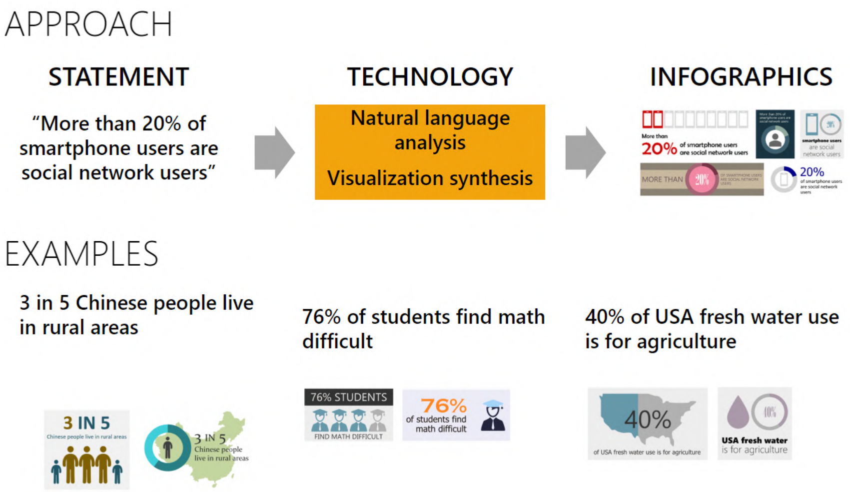

Figure 2.21: Text-to-Viz automatic visualization schema explanation [Cui et al. (2019)]

2. Expert knowledge-based model is representing a more simple approach with defined rules and logical restrictions, forming itself *a knowledge representation* of the graphs, plots, maps, and charts with corresponding characteristics. Control System Theory and Artificial Intelligence introduce four types of knowledge representation (Rubanov & Filatov, 2010):

- Productional Rules (IF...THEN),
- Predicative Logic Language (A $\rightarrow$ B),
- Frame (Connected set of large structural units),
- Semantic Networks (Graphs),

With regards to the data visualization recommendation, this can be explained when the proposition of the final decision on the data visualization method is made (Figure 2.22).

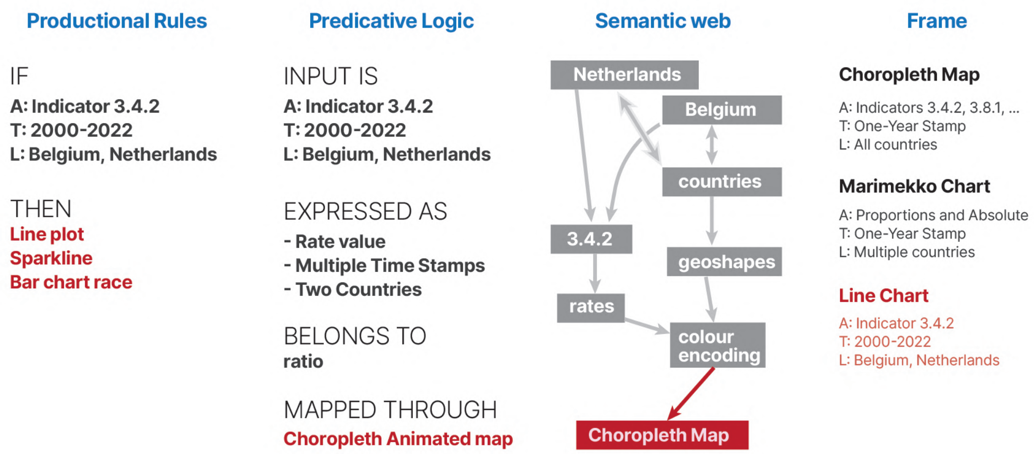

Figure 2.22: Knowledge representations with examples of visualization types, considering their ATL model

Since the first two knowledge representations are rather generic, simplified, and partial examples and can be represented through lists of expressions, the semantic web and frame methodologies are looking more promising to design the scalable recommendation system.

Frames are data structures used to divide knowledge into substructures by representing ”stereotyped situations” (Minsky, 1974). The frame-based approach of knowledge representation applies to several reviewed recommendation systems from the data visualization enthusiasts (such as the aforementioned in chapter 2.5 Financial Times Visual Vocabulary example (Figure 2.14)).

A semantic network can be expressed as a directed or undirected graph consisting of vertices (concepts) and edges (semantic relations) between concepts mapping or connecting semantic fields (Sowa, 1987). Semantic webs’ most prominent and actual application is represented through the World Wide Web and Search Engine Systems (e.g., Google, Yahoo, Bing).

The advanced programmed realization of expert-based methodology can be found in a well-noted feature of Tableau Software called *ShowMe*; it represents an illustrative example of a knowledge-based recommendation system (Mackinlay, Hanrahan, & Stolte, 2007); depending on the given input type of the columns, the interactive window makes it available to select general visualization types, e.g. vertical bar charts, maps (Figure 2.23)

The obvious drawback of this approach lies in the inherent methodological constraints of possible output scenarios, such as the accepted scale of data representations, the set of visualization types, etc.

3. Hybrid model, concatenates the aforementioned approaches into a new one. Zhu

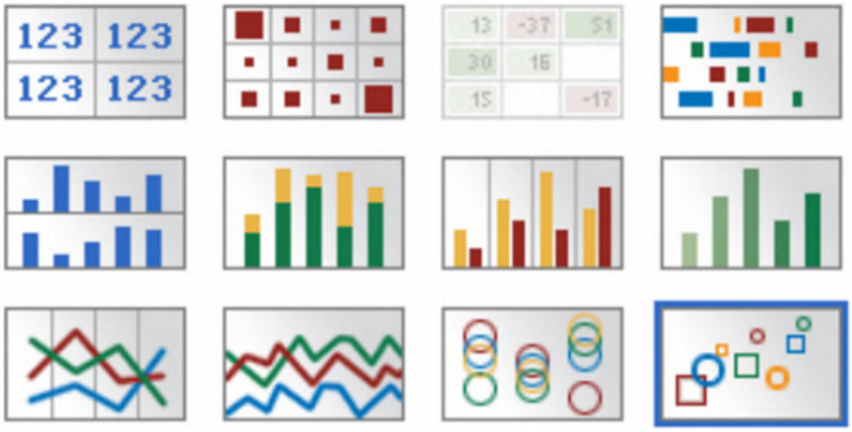

Figure 2.23: Original view of ShowMe feature in Tableau Software; it is representing the knowledge-based automatic visualization recommender system method [Mackinlay et al. (2007)]

claims it as an ”ideal method for building models of automatically generating visualizations”. In this extent the peculiar example of *DeepEye* project can be demonstrated. The model of suggesting the visualization types based on the two parts of *offline* part with two trained machine learning models and *online* part where all possible visualizations are generated to qualify and rank the recommendations (Figure 2.24).

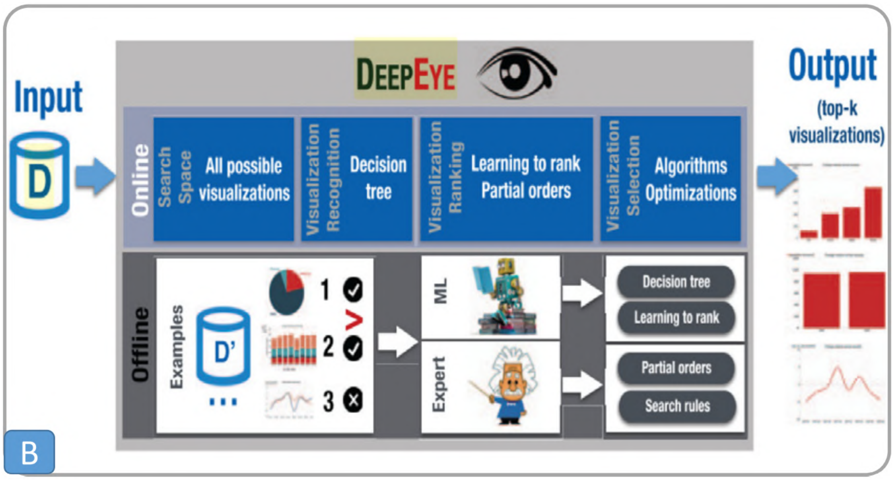

Figure 2.24: DeepEye methodological pipeline is representing the advanced methodology of data-driven and knowledge-based approaches [Luo et al. (2018)]

In general, the design challenges of such recommendation systems are mainly related to the visualization type collection with corresponding identification of visual elements (the applied framework visual grammar) and the defined rules (or algorithms) of automatic recommendations and generation of visualization.

Chapter 3

Sustainable Development Goals

In September 2015, the United Nations in New York proposed a set of global SDGs, which consisted of 17 goals that relate to the most important global problems. Since the United Nations (UN) is the largest international influential organization that deals with worldwide social and humanitarian issues, the current sustainable development framework is actively used by the authorities and various organizations. For the proper analysis and better cognition of such concepts, data visualizations and maps come with support, which offers insights into geographical patterns at multiple scales and displays trends over time (Dykes, MacEachren, & Kraak, 2005).

3.1 Brief overview

The Sustainable Development Goal (SDG) are closely associated with the concepts of operationalization, indexes, and frameworks. It is considered that the global economics, demographic and social processes may be subjected to some valuable estimation, which could bring a basis for further qualitative assessments and comparisons.

Figure 3.1: Millenium Development Goals (MDGs) are the direct conceptual predecessors of the SDGs [The United Nations]

The forerunner of the SDGs was expressed through the Millennium Development Goals (MDGs), first established in September 2000 by the UN. Such a framework was an early-stage attempt to summarize global environmental and social problems based on data collections since the 1970s through properly defined goals with respective indicators (Hák, Janousková, & Moldan, 2016). Having a fairly established methodology, it was supposed to achieve the 8 decent goals which had multiple targets and following indicators up to 2015 (Figure 3.1).

Although the overall progress of MDGs achievement was uneven among different countries (UN Economic and Social Council, 2012), it is considered that the MDGs represent a quite successful framework for further usage as a balanced blueprint for economical and social development. As a result, since 2015 the UN is promoting a new goals-based initiative, which is now known as the 2030 Agenda for Sustainable Development, or Sustainable Development Goals (Figure 3.2).

Figure 3.2: The 17 Sustainable Goals by the United Nations [The United Nations]

The sustainable development goals are independent of each other, although the understanding of such a serious framework works better through the perspective of ecology, economy, and society (Figure 3.3). Goals 13 (Climate Action), 14 (Life Below Water), 15 (Life on Land), and 6 (Clean Water and Sanitation) are mainly focused on the ecological balance, mainly through the implications of impact reduction of human activities. The economy is covered by the following goals: 8 (Decent Work and Economic Growth), 9 (Industry, Innovation, and Infrastructure), 10 (Reduced Inequalities), and 12 (Responsible Consumption and Production). These goals are aimed to present inclusive, equal, progressive economic growth with the design of sustainable urban and rural infrastructures alongside the developing of rational consumption and production models. The social aspect is the most diverse and foremost, having eight goals, such as 1 (No Poverty), 16 (Peace, Justice and Strong Institutions), 3 (Good Health and Well-Being), etc.

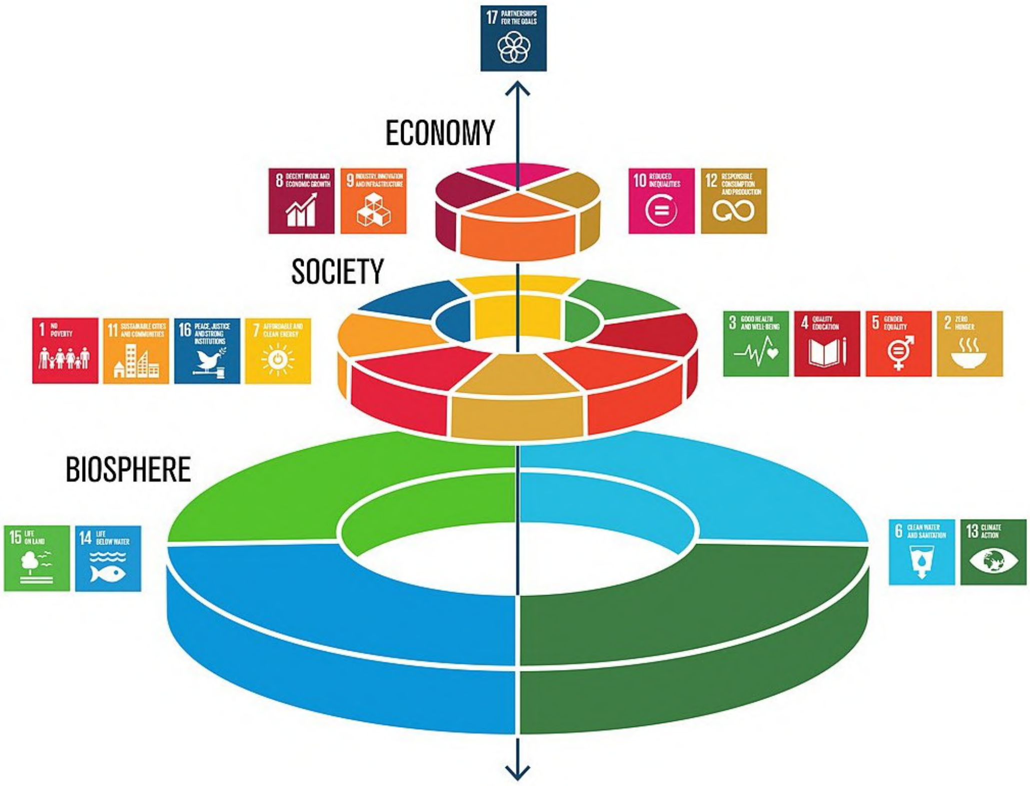

Figure 3.3: The conceptual structure of the SDGs with regards to the aspects [By Azote for Stockholm Resilience Centre, Stockholm University]

These goals are mainly directed to bring the inequalities among the countries and communities to the possible zero, providing more decent social opportunities for all. The final goal 17 (Partnership for the Goals) is proposing the policymakers, authorities, and all involved parties unite and collaborate for the achievement of the goals, as well as to contribute to the achievement of the goals worldwide.

Each of the **17 goals** of this agenda are divided into **specific targets,** which contain well-developed methodology, expressed by **231 indicators.** These indicators are divided into Tiers, which assign them to established, conceptually clear methodologies with rather constant data collection (*Tier I*) or not regular data collection (*Tier II*), or put them into the *Tiering Pending* category when methodology and/or data are not available yet. As of the 4th of February 2022, the tier classification contains 136 Tier I indicators, 91 Tier II indicators and 4 indicators that have multiple tiers (different components of the indicator are classified into different tiers) (United Nations Statistics Division, 2022a).

To understand the better possible strategies on how to support the designers, cartographers, or reporters with suitable data visualization solutions, it is foremost important to investigate the SDG data on the matter of its meta characteristics.

3.2 Indicators tiers and data characteristics

The obvious complexity of the problems and aspects that are covered by the SDGs is appropriately expressed in the structural and logical characterization of its targets and indicators, including their original methodological background and original measurement approach. This, therefore, makes it difficult to conduct the explorative analysis and decide on the possible visualization representations. Tiers (I / II / Pending) classification makes it easier to see where the possible analysis and study can be conducted with available data. Additionally, the indicators themselves represent a subject to analyze through their publicly available metadata (United Nations Statistics Division, 2022b).

Table 3.1: SDG Indicator Metadata Sections and Subsections

Metadata section	**Metadata subsection**
0. Indicator information	0.a. Goal
	0.b. Target
	0.c. Indicator
	0.d. Series
	0.e. Metadata update
	0.f. Related indicators
	0.g. International organisations(s) responsible for global monitoring
1. Data reporter	1.a. Organisation
2. Definition, concepts, and classifications	2.a. Definition and concepts
	2.b. Unit of measure
	2.c. Classifications
3. Data source type and data collection method	3.a. Data sources
	3.b. Data collection method
	3.c. Data collection calendar
	3.d. Data release calendar
	3.e. Data providers
	3.f. Data compilers
	3.g. Institutional mandate
4. Other methodological considerations	4.a. Rationale
	4.b. Comment and limitations
	4.c. Method of computation
	4.d. Validation
	4.e. Adjustments
	4.f. Treatment of missing values (i) at country level and (ii) at regional level
	4.g. Regional aggregations
	4.h. Methods and guidance available to countries for the compilation of the data at the national level
	4.i. Quality management
	4.j Quality assurance
	4.k Quality assessment
5. Data availability and disaggregation	—
6. Comparability / deviation from international standards	—
7. References and Documentation	—

Table 3.1 represents the metadata attributes that are describing each SDG indicator in separate documents. By looking at the metadata complications, it becomes obvious that the systematization of 231 indicators for the possible assistance in graphical representation requires a strong methodology in distinguishing patterns and features that these indicators share.

Firstly, the verbal expression of the indicators immediately represents its measuring essence. For Tier I indicators it is possible to apply the classification by the level of measurement in accordance with Stevens' methodology (United Nations, 2017). Moreover, for each Level of Measurement (LoM), the indicators are expressed through the *mappable / visualizable* values, for which certain visualization types can be picked or suggested (Kraak et al., 2020). Figure 3.4 represents the indicators' Tiers with their respective mappable values (as of April 2020).

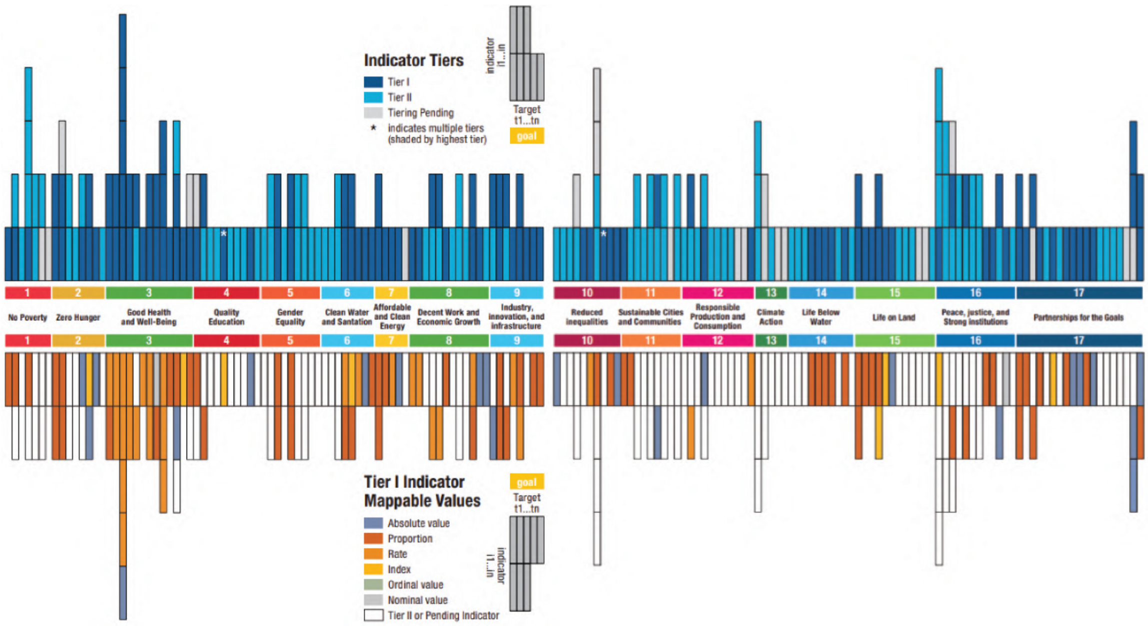

Figure 3.4: Indicator tiers and their mappable values [Kraak et al. (2020)]

Most of the enumerated SDG indicators are expressed through the proportions or rates related to the population. However, these enumerated attributes are rarely evenly distributed within the enumeration units. Thus, there is a certain place for the proper normalization and transformations to ensure proper comparability of the countries with regard to their data. Figure 3.5 explains how this exact transformation can be conducted. Chapters 4.2.1 and 4.2.2 are concentrated on this algorithm represents its impact on the visual output.

There are annually collected indicators data for countries. Statistical mapping distinguishes the triad model, where the Attributes (what?), Location (where?) and Time (when?) are forming the factual basis for the exploration and explanation of multiple phenomenons that are unfolding in the chrono-, geo- and attributive- dimensions (Kraak & Ormeling, 2020). SDG datasets can be treated within this ATL model to emphasize and specify the attributes of the dataset.

Figure 3.6 demonstrates the example data where the columns are treated concerning

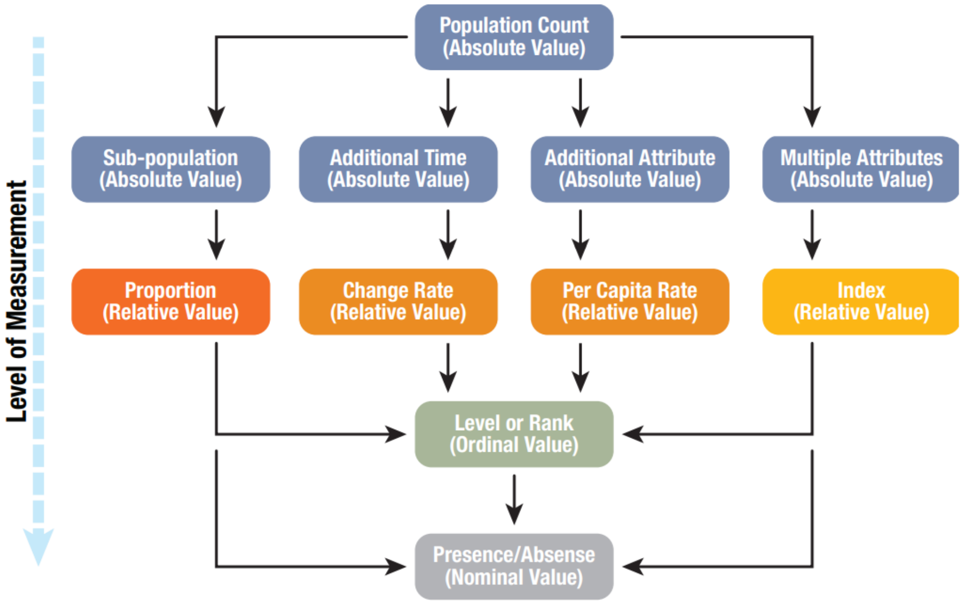

Figure 3.5: Data transformation options for enumerated, population-based attributes [Kraak et al. (2020)]

such a division. A country in the indicator datasets represents the *nucleus*, a basic entity that exists in time with some qualities and quantities, expressing different levels of dynamics over the available years. Although countries may collect the data on the local levels and track the SDGs progress on an even more detailed scale, the initial methodology implies that all UN member states are the main geographical representations to look at and analyze.

The UN also applies global regional aggregations, forming groupings, which are used to monitor broad SDG progress, as they are defined under the M49 standard (Department Of Economic and Social Affairs Statistics Division, 1999). One of the examples of global regional aggregations is demonstrated through the specific SDG regional groupings (Figure 3.7). Such groupings allow us to compare different regions with each other, as well as to analyze the countries within their neighboring geographical areas. More territorial aggregations are possible based on certain methodologies, parameters, or arbitrary selections.

Attributes					Location				Time					
Goal	Target	Indicator	SeriesCode	SeriesDescription	GeoArea	GeoAreaName	Location	Units	2000	2001	2002	2003	…	2017
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	112	Belarus	ALLAREA	PERCE	80.62	80.81	80.99	81.17	…	94.52
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	100	Bulgaria	ALLAREA	PERCE	96.84	96.84	96.84	96.81	…	96.95
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	203	Czechia	ALLAREA	PERCE	96.32	96.32	96.44	96.55	…	97.88
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	348	Hungary	ALLAREA	PERCE	50.51	50.51	50.51	50.51	…	89.57
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	616	Poland	ALLAREA	PERCE					…	99.16
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	498	Republic of Mold	ALLAREA	PERCE	40.42	42.32	44.27	46.26	…	72.88
6	6.1	6.1.1	SH_H2O_SAFE	Proportion of popu	642	Romania	ALLAREA	PERCE	81.65	81.61	81.59	81.63	…	81.92

Figure 3.6: Generic SDG dataset representation with regards to ATL aspects [Kraak et al. (2020)]

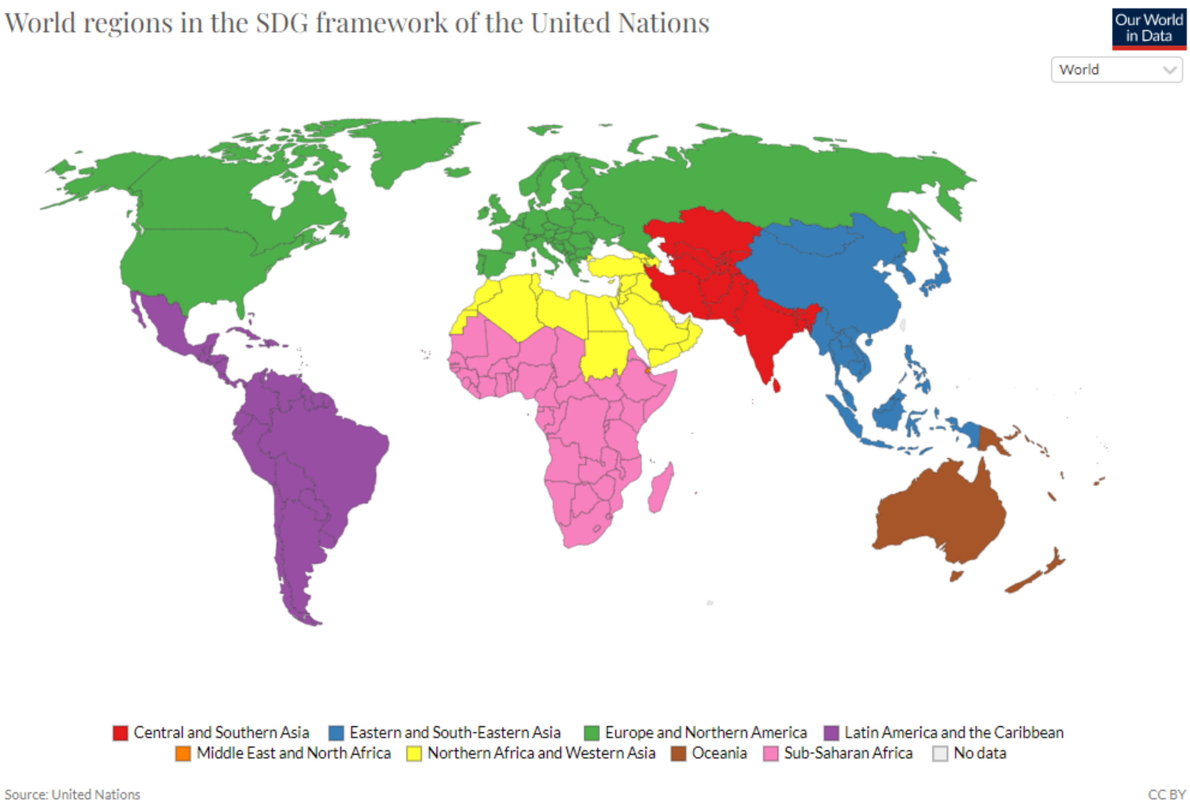

Figure 3.7: Generic SDG dataset representation with regards to ATL aspects [Our World in Data, United Nations]

The annual dynamics of indicators provide the opportunity to track existential, numerical, and qualitative changes in countries. Most of the indicators are available starting from the year 2000 annually, although the exclusions are not rare. Several indicators are gathered within 3-year or 5-year intervals (3.4.2. Suicide mortality rate), while some may be collected monthly (2.c.1 Indicator of food price anomalies). It is important to look at the indicators distinctively. Representations of changes in space over time are also not excluded, e.g. indicator 15.1.1 (Forest area as a proportion of total land area) is possible to consider through the disappearance, emergence, or landscape transformation of forest cover. Although the SDG data model itself does not contain spatial specification, it is possible to address certain issues about indicators through more detailed analysis and representations.

Thus, we can conclude that the common data structure of the indicators has been worked out by the UN in a quite clear way, having for most of them well-described metadata, collection methodology, and the values themselves over the available time series.

3.3 Challenges and factors to impact SDG visualization

As we have outlined the principles of data visualization and reviewed the metadata and characteristics of the SDG indicators, we are now able to summarize the possible factors that shape the final data visualization output, after which we may be able to formulate the conceptual basis for the recommendation system.

The proposed in Chapter 2 different aspects of data visualization for the literature review can help us to specify the possible challenges to the SDG indicators visualization.

1. Translation of the data into the visual representations with regards to SDG can be expressed initially through the available mediums, tools, and software in use. The analysis from the 2.4 demonstrates that the available instruments for the analysis and the presentation are not homogeneous in terms of availability, as well as the possible functionality. Whether it is open source or software for a profit, the possibilities are certainly different — so does the visualization. Although, because the SDGs are considered a part of the open data movement and it is openly offered and shared for everyone to use (Kraak, Ricker, & Engelhardt, 2018), the software and visualization methodologies are following the same direction, providing tutorials for visualizations in the open source programming frameworks (E.g., QGIS) (Houtman & Roth, 2021).

2. The data aspect is represented through several challenges. We have considered the data characteristics in the previous chapter and may outline the definite inferences.

Firstly, although, the majority of the indicators are assigned to Tier I up to now, it is first and foremost important to actually have something to visualize — if **the data availability** is scattered, this should be addressed respectively in the reports and visualizations, or correct interpolations should be made to correctly represent the data.

The factor of potential **qualitative disaggregation** plays a big role in how to divide and present. While some indicators (3.3.3) may not have any inner qualitative disaggregation, most of them do (e.g., by gender for 1.3.1 or by the level of education for 4.a.1). Adding more complexity into the attributive part of Attribute, Location and Time model (ATL) model, qualitative disaggregation should be kept in mind.

The final factor for data-related issues, **the pivot calculations, statistical transformations aggregations** is the step that may refactor the numeric content to be presented. While, e.g., the world map of countries with the proportional indicator may depict the spatial, global context, the visualization of the share of the countries that are below or above a certain percentage, which was calculated through the pivot tables, may express a completely different message.

3. Correspondence in this context is rather related to the configuration of ATL in focus. Depending on what is the what-where-when of users' interest, this significantly determines the possible variations on a possible list of options for the visual representation.

Attribute itself may not be left solitary within the indicator, but rather could be transformed through its level of measurement. The upcoming chapters 4.2.1 and 4.2.2 shed the light on how the transformation of the indicator's level of measurement may impact its visual output. Moreover, the attributes themselves may be visualized in a mix, so the variate combination of multiple indicators with corresponding levels of measurement adds more challenge.

Additionally, second aspect relates to **location.** Different spatial scales of analysis are possible, so it involves the indicator into more issues for the visualization representation. Moreover, when the mapping method is selected to present the data, an array of map design considerations gets involved. Map projection, spatial scale, color schemes, typography —

data, by getting more ”spatial” in its geographic aspect, bring more issues to consider and to be aware of for the final visualization decision. Additionally, non-spatial visualization is also representing the challenging aspect.

In the end, selected **temporal resolution** determines if the visualization implies the dynamic factor, where the aforementioned compiled configuration of attribute and time gets unfolded over the years. Whether the data is looked through the one year, before and after consequences, and within the linear time series, this will limit the options for visualization even more.

4. Visual appropriateness, expressed through the aesthetics, visual rules and other aspects are formed from several challenges.

Of course, it is foremost the actual **visualization type,** which is selected accordingly to the ATL configuration. Among the possible diagram options and maps, it is important to finally pick one that will reflect the data. The main focus of the current Master's thesis is related to this actual factor. considering the reasoning of correspondence parameters.

Since the map or diagram components that are encoding the data itself are not really isolated and universal, it is **visual components, encodings layout principles** that shape the selected visualization type into the properly designed graphic work. Such a factor is interdependent with the usage and readers' context, available understanding, and planned goal of the visualization. By selecting pictorial elements, sizes, colors, and layout it is possible to estimate the possible success of the visualization in the explanatory dimension.

Usage environment is figured here as the visual appropriateness reasoning. Whether it is the printed report or interactive dashboard, the graphic design decision is made with the following considerations of what the environmental possibilities are.

5. Relevance for the users implies several points to work with.

The readers, **the audience** itself should be investigated on the matter of mutual expectations. Briefly, the key differences between the users are expressed in their *accessibility* (e.g. ability to obtain or extract useful information from the visualization), *expertise* (whether the user is familiar with the subject area or diagram/map reading), *skills* (such as reading, literacy, spatial and numeric understanding) and *motivation* (whether the topic or visual forms induces the user). Cultural, ethnic, religious, demographic, and political considerations are representing an additional, multidimensional set of issues that influence the relevance for the users.

Immediately emerging aspect of **users feedback,** being the chain's element of data visualization pipeline, is able to skew the intended outputs and assist in important reviewing and editing (e.g., mistakes, typos, suspicious representations). Moreover, the lack of feedback itself may lead to incorrect interpretations of the message's impact.

Finally, it is the **actuality and topic** of goals and indicators that are presented. Although it was highlighted in the motivational aspect of the users' differences, the presence of the topic in the social field plays a significant role in how this visualization should or might be presented like. A good example of an emerging climate change topic (respectively related to Goal 13 from the SDGs) vividly represents a topicality factor, when the alarming focus on a problem encourages to address and communicate visually with a stunning eagerness

(Hawkins, 2020). Hawkins' warming stripes (Figure 3.8) in this case represent a great example of when such a complex concept and "hot topic" as climate change can be expressed through simple, data-driven minimalistic forms.

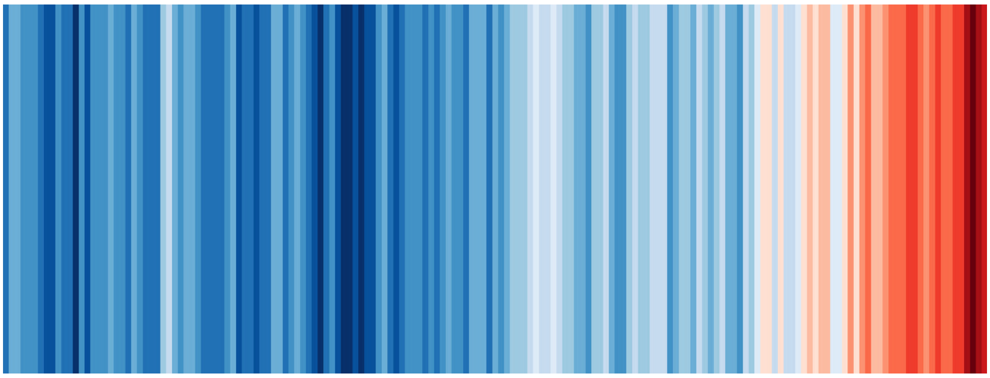

Figure 3.8: "Warming Stripes" visualization is demonstrably representing the annual temperature differences over the last years through the diverging color scheme [Ed Hawkins]

6. In the end, the **data visualization as a whole process** of the indicators implies two more relevant factors. The aforementioned detailed aspects behind the data visualization, when they get combined into the systematic, step-by-step process, now obtain system-related features.

Visualization' objective plays an already built-in role in how the graphics might look. The overlook of the graphics as a means of communication implies a feasible set of questions that visualization answers or attempts to answer. The task of the visualization as a whole should include clear wording and measurable Key Performance Indicators (KPI).

It would not be out of place to mention **the available resources and expertise** of data graphics creators. This is applicable for the categories such as time deadlines, computational abilities, the background and level of designer's skills, and human resources. The differentiation among these elements generates an array of possible outputs. The only cartographer with a generic laptop is less likely will manage to present the maps for the outlined visualization goal, but a decent tech team of data analysts, managers, and designers preserve a bigger potential for high-quality work in collaboration.

Therefore, the SDG data visualization inherently contains a plural list of interrelated challenges. The key premise of this study is to attempt designing the conceptual basis for the factors that are in our focus of interest, mainly related to the factors of correspondence. Moreover, a potential visualization recommender system is influencing the data visualization process as a whole, as it allows to save time within the possible consideration of the diagram types or maps.

The accessible instances of SDG data visualization collection can be reviewed on the subject of the challenges that may affect the data visualization types. Miscellaneous reports' examples might be of the interest (Pirlea & Whitby, 2020), (Sachs, Lafortune, Kroll, Fuller, & Woelm, 2022), (ICA, 2016), (Lynch & Sachs, 2021), (SDG Center for Africa and Sustainable Development Solutions Network, 2019).

Among the variety of factors and challenges associated with the visualization of SDG indicators, it was decided to focus on the issues of attributes, location, and time, and most precisely on the option of a possible transformation of the attribute' scale of measurement. The reasons for the chosen focus are as follows.

First, in general, the idea of presenting the same data and the same story by different visual methods is of particular interest to the author. The confrontation between form and content rests more in the field of art, however, experimenting with the expression of the same information in different forms allows us to find new solutions in visual communication. Using the attribute transformation for such a task looks like an option in this case, which is technically easy to implement in form of a prototype (see chapter 5).

Secondly, in the reviewed SDG-related materials on recommendations for mapping, data transformation with the measurement scale was systematically taken into account (Kraak et al., 2018), (Kraak et al., 2020). The peculiar wishes of the thesis supervisor in this regard were also taken into consideration by the author.

It is also worth noticing that since the research objectives on the system development were the main ones, it was relevant to find simple, minimal elements that could be used as a basis for the choice of visual forms. The involvement of relevance and audience parameters, visual appropriateness (graphic design), as well as issues of software selection or optimization of pipeline processes, seemed to the author to be more complex aspects, the systematization of which would require much more time and effort than the time allotted for writing the paper.

Chapter 4

SDG visualization recommender system: methodology

Based on the completed research, the preliminary outlined features of the system can be narrowed down to these two broad points:

1. It should **consider the LoM transformation as something that defines** the possible *alternative* variations of the **data visualization methods** for SDG indicators;

2. The system should **consider the variations with regards to the Attribute-Location-Time parameters** of the input data, thus preventing the possible visualization recommendations based on this configuration;

The system's requirements can be described as follows:

1. Contain the SDG indicators' description with the original level of measurement and its possible transformations;

2. Identify the original ATL complexity of the dataset of interest;

3. Based on the described meta-input, bring the possible data visualization solutions to the user together with default and alternative solutions;

Within an analyzed framework from chapter 2.9 of possible variations in the field of automatic visualizations and recommender systems, it is chosen to focus on **designing the expert knowledge-based methodology** due to multiple important implications.

Firstly, the SDGs themselves represent a closed, finite set of concepts and relations alongside with inherent metadata, applications, and data characteristics. As we are focusing on the SDGs data visualization only, there is probably no need to implement the broad data-driven methodology, involving the ML computations.

Moreover, the presented framework of ATL data representation with regards to the indicator datasets introduces an additional level of restriction, therefore simplifying and narrowing the regarded SDG indicator data representation understanding.

Finally, the dedicated period for the thesis' completion is likely to be deficient alongside other tasks to study the ML methods in detail. So, it was decided to leave the possible deployment of ML algorithms' elements for future research, focusing on an elementary, expert concept of a system for proposing data visualization outputs.

Thus, the conceptual foundation for the visualization recommendation system will be embodied through the expert knowledge-based approach.

4.1 Conceptual presentation: ATL matrix

Since we have outlined the key peculiarities of the knowledge-based approach for visualization systems, the actual conceptualization of our recommendation algorithm can be implemented.

Regarding our interest in the selection of visualization methods by ATL and the scale of measurements, it seems practical to apply the *frame* approach (see Figure 2.22). We can describe a particular visualization type from the point of this ATL model — what spatial entities of Location are involved, how does the Time stretch, and what type of Attribute is in the dataset. Location can be classified in a way that both maps and diagrams are simultaneously offered to the users. The aspect of time already appears systematic, since the annual dynamics of data collection for SDGs implies possible absolute (one year), relational (only two points, before and after), and linear outcomes of the temporal parameter. Concerning the attributes, such a formalization is not as straightforward, but the key interpretation of an attribute in our case can be thought of as the measurement scale itself. As we have already determined, it has a significant impact on the graphical expression and content of the communicated message. Locations decided to treat simply with four possible complexity levels: one country (or region), two countries, more countries, and all countries (Figure 4.1).

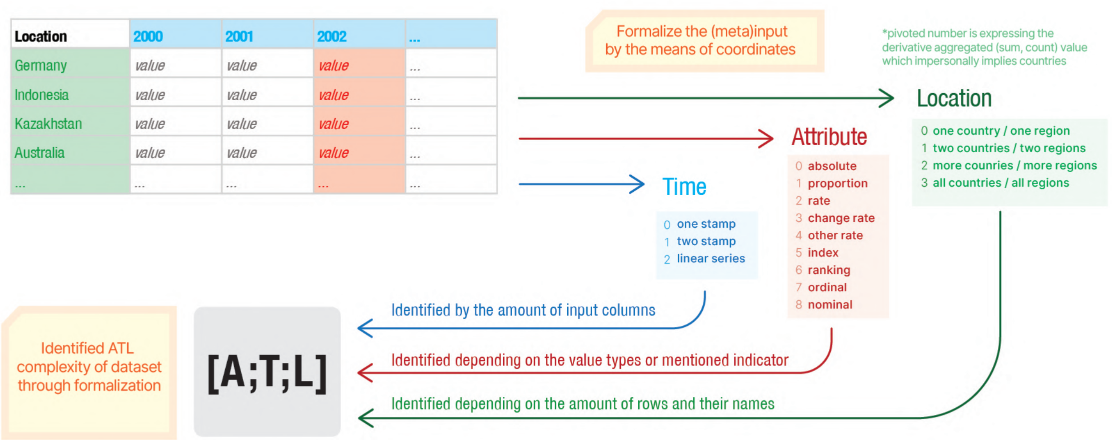

Figure 4.1: ATL matrix contains the formalized complexity levels of corresponding aspects of attribute, time and location

It is possible to represent such a relationship of descriptive parameters in the form of a three-dimensional matrix with the corresponding coordinates (Figure 4.2).

All possible graphical representations (diagrams, maps, plots, charts) can be located in this matrix. It means that they should be described utilizing coordinates on the matter of their appropriateness concerning the ATL-complexity (Figure 4.3).

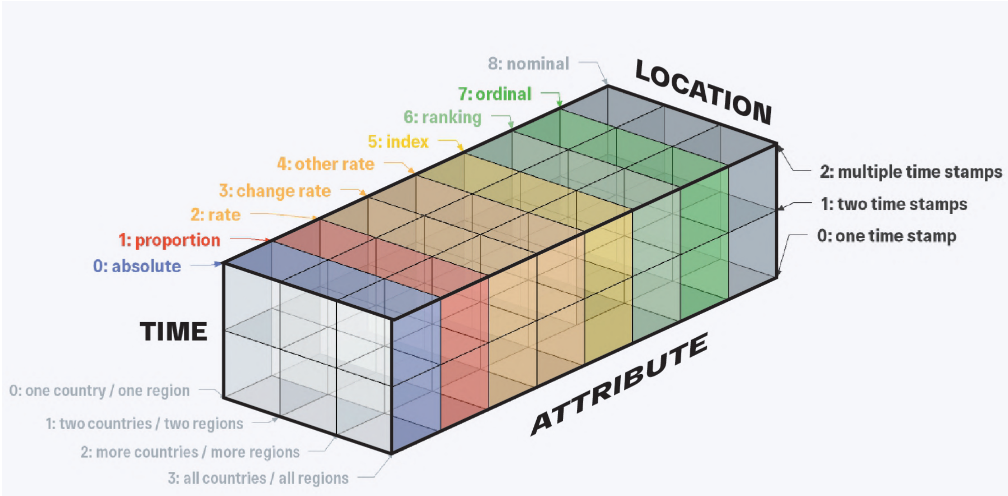

Figure 4.2: ATL matrix contains the formalized complexity levels of corresponding aspects of attribute, time and location

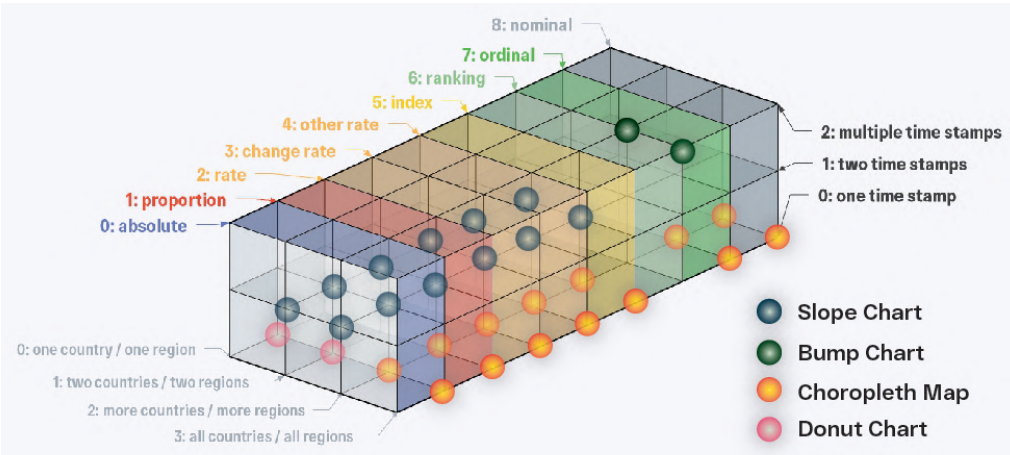

Figure 4.3: The location of some examples of different visualizations in the matrix

The coordinates of the resulting ATL description bring users to the allocated set of possible visualization that can correspond to the initial input. Even more broadly, it is possible to see certain *planes,* which contain the proper visualization types on it (Figure 4.4).

Back to Figure 3.5, we also implement the possibility for the data transformation with regards to the levels of measurement or, in our case, attributes. The matrix's usability is mainly defined by its included variety of data visualization types. Because the description and assignment of the ATL aspects to the whole diversity of data visualization is a demanding task within the obvious time shortage for the thesis completion, it is just enough, for now, to highlight that the more alternative user is having, the more effective and helpful the matrix' performance is. The next subsection is rewinding the design decisions on the low-level ATL complexity, but this time already involves the aforementioned matrix. Figure 4.5 demonstrates the work of data transformation in the example of proportion, going down to the ordinal, ranking, and nominal attributes.

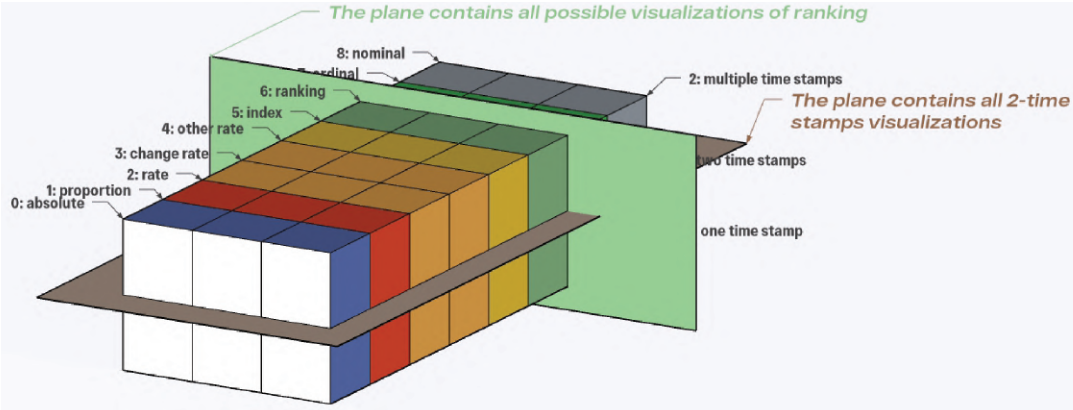

Figure 4.4: Matrix contains corresponding planes that contain data visualization types

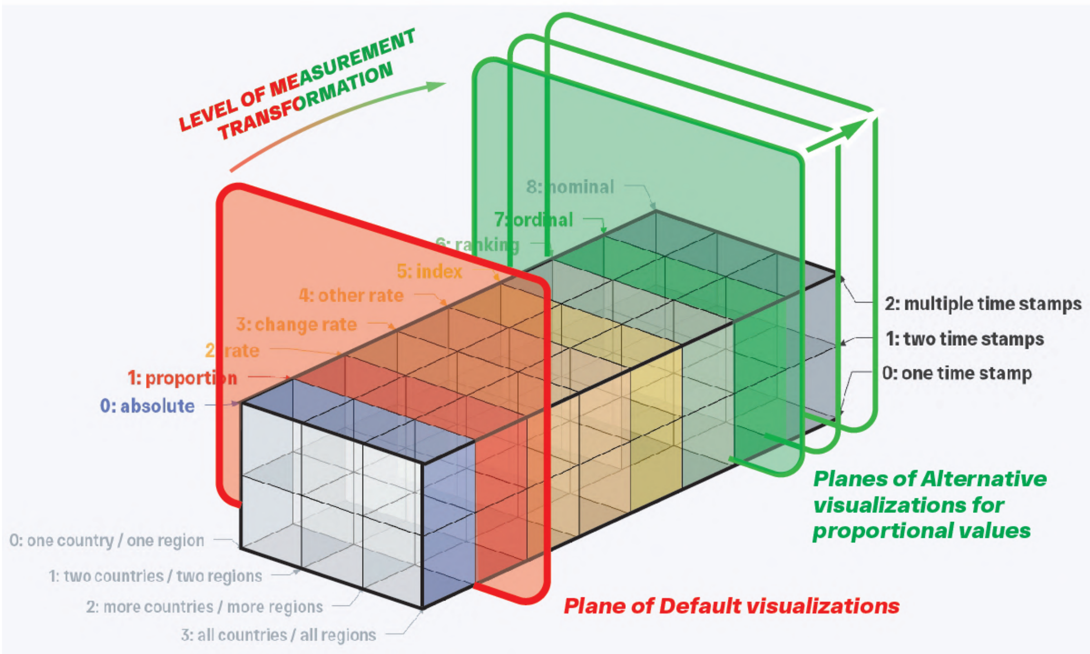

Figure 4.5: Matrix contains corresponding planes that contain data visualization types

4.2 Recommender system validation: two examples

The validation of the system as well as the demonstration of the data transformation' effect on the visual output can be reviewed here.

4.2.1 Visualizing low level of ATL complexity: Germany vs. Poland

The demonstration of this aspect can be firstly presented with a low level of ATL complexity. It was arbitrarily selected to consider two neighboring countries, e.g. of Germany and Poland in one indicator (5.5.2. Proportion of women in senior and middle management positions) at the absolute moment of time, e.g., in 2018. The dataset is available in the UNStats repository and starts from the year 2000 (UN Department of Economical and Social Affairs, 2022). The filtered selected data with the values are presented in Table 4.1.

Table 4.1: The initial filtered data for Germany and Poland of the indicator 5.5.2 for 2018

Country	5.5.2_2018, %
Germany	28.58
Poland	39.52

Thus, there is an immediate understanding that the data is representing a one-time stamp comparison of two countries with proportional values. We can describe this through our matrix as coordinates (1;0;1). We see that the identified coordinates from the input are bringing us to the point in the matrix that already contains multiple data visualization methods. The highlighted lines inside the matrix are identifying the point with all the corresponding collections (Figure 4.6).

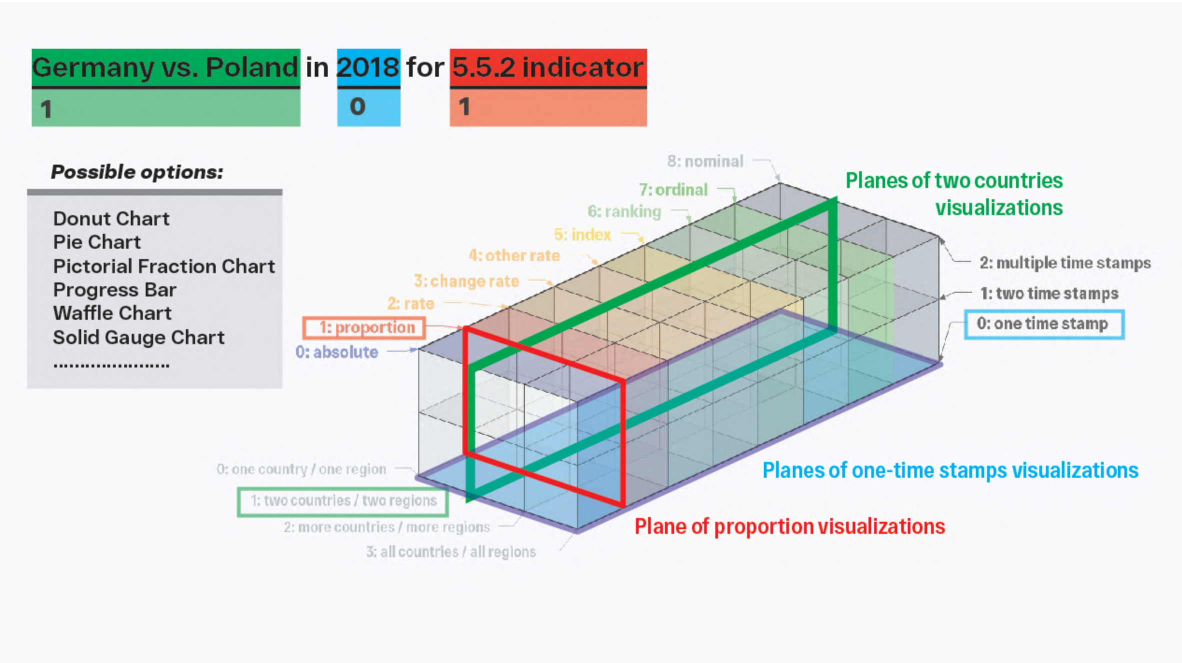

Figure 4.6: Work of matrix with the input of Germany vs. Poland example

Spontaneous decisions on the visualization might be obvious: the conventional proportional diagram might better represent the part-to-whole structure, according to the aforementioned in chapter 2.5 recommendation systems. The basic visual components, although, can be selected differently. We could choose from multiple solutions, such as donut charts, progress bars, waffle charts, pie charts, etc. The reasoning behind a proper picking of the right one among the listed options is not straightforward, but we can agree that through the iterative process a better solution for the data visualization can be found. The usage of donut charts seems promising, as is depicted in Figure 4.7. But the contextual factor of topics and audience might lead to something that is presented in Figure 4.8.

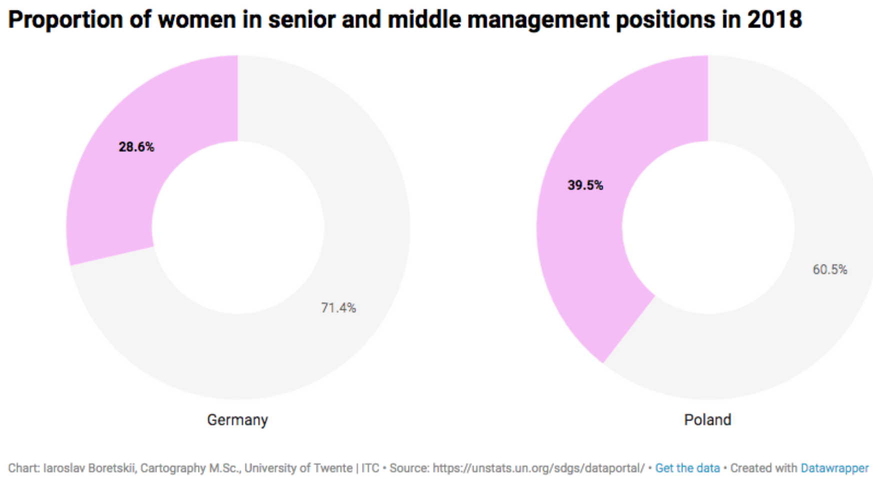

Figure 4.7: The representation of proportional values for two geographical entities can be made through the comparative diagrams, such as Donut Chart

The possibilities for the visualization of percentages, although, are limited to the possible visual encoding of proportional values. It is possible to initiate a transformation of the level of measurement to demonstrate these tiny visual changes. According to the schema from Figure 3.5, we can move down towards the ordinal and nominal representations. As our spatial focus is relatively small to make global inferences, we need to extend our location scope by adding more countries to assign the qualitative values to Germany and Poland.

When we move down the matrix in direction of the attribute coordinates, we can see new options to pick from concerning the data transformation been made (Figure 4.9).

How we distinguish the thresholds for the ordinal and nominal transformations can be narrowed down to three possibilities. The first, most representative and scientific, is based on a statistical-based application, thanks to which we apply such concepts as mode, mean, median, quartiles, etc. In this case, we are dealing directly only with the statistics of the available dataset, splitting the distribution into more representative bins. In the second case, we can rely on a methodological approach that, in addition to statistical and quantitative indicators, can introduce some omissions of arbitrary nature or neglect the nature of the data themselves, referring to customary or established classification, e.g. Human Development Index with 4 ordinal categories for countries to be assigned (United Nations, 2021). Third, the final approach implies arbitrary decisions on the way how classification is made. This approach can be largely caused by the specifics of the task and the personal preferences of the authors of the project; despite the possibility of a particular manipulation, this method sometimes can be used. However, in any of the three approaches, it is preferable to specify the applicable restrictions and methods, so the readers will be able to grasp the original, unbiased picture.

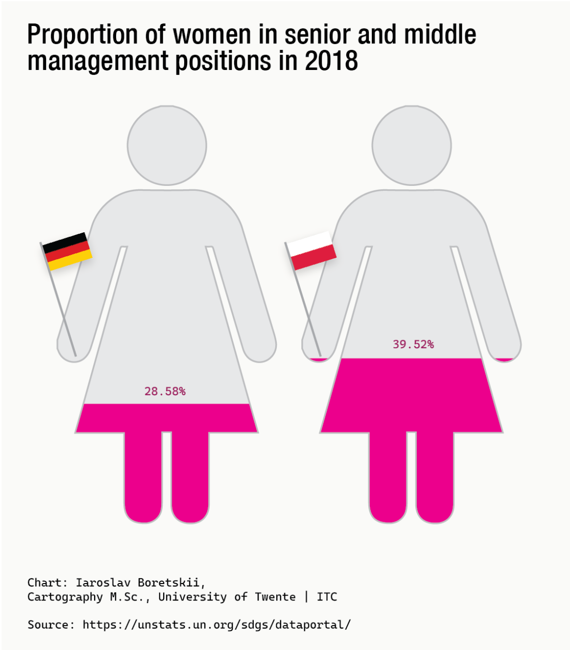

Figure 4.8: By adding more contextuality through the topic or audience, the same level of measurement of the data may be presented more exotically

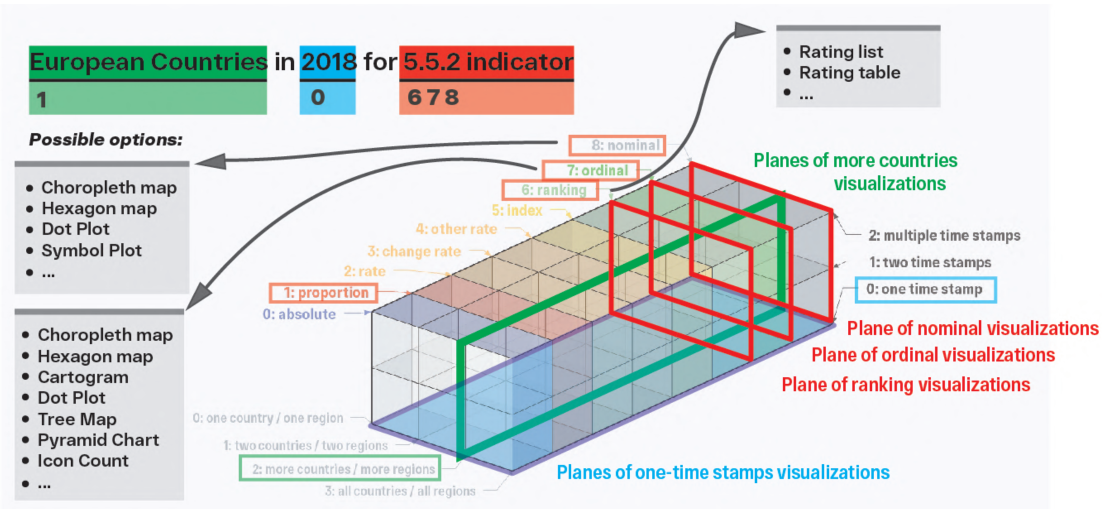

Figure 4.9: Available options with data transformation been applied

To present the possible variation and disruption in the visual presentation of the same data, it was decided to proceed with a statistical-based application. Table 4.2 describes the inner characteristics of the world countries.

Our countries of interest in the distribution of the values among the other states with

Table 4.2: The central tendency analysis for the indicator 5.5.2 in the 2018 dataset

COUNT_NonNull	MEDIAN	AVERAGE	QUARTILE1	QUARTILE2	QUARTILE3
64	32.67	32.24328	27.76	32.67	37.9375

available data allow us to identify the breaks of 27.76 — 32.16 — 37.93 percent that create a respective very high-high-medium-low sequence of classes. The possible representation of such a distribution with regards to inner classes can be visualized, e.g., through the dot plots, containing highlighted annotations of Germany and Poland (Figure 4.10).

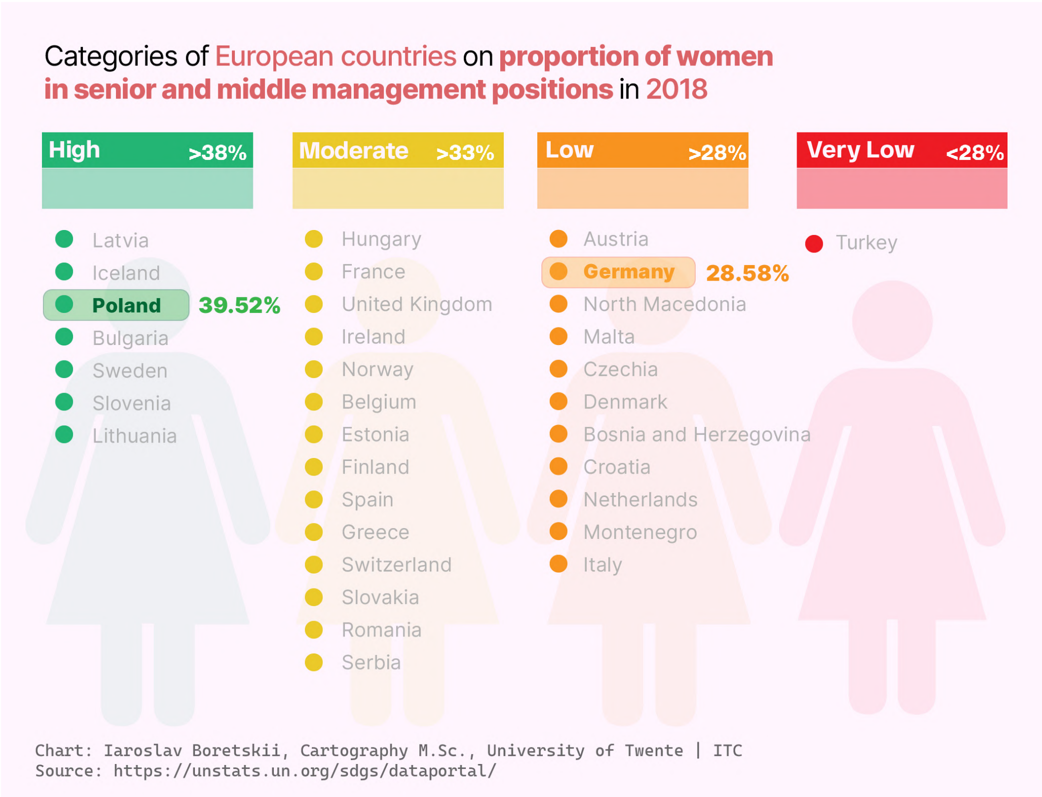

Figure 4.10: The dot plot solution for ordinal classes visualization

Thinking spatially, as the geographical aspect of the data got complicated, the attempt to represent the ordinarily classified data through maps seems not that meaningless. This time we may preserve the same classification, though shrink our area of interest to the European countries. For the experimental extravagance, the hexagonal cartogram is selected for the representation, expressing the qualitative values through color differentiation respectively (Figure 4.11).

The nominal representation of such an indicator can be also derived from the ordinal classification that has been proposed. E.g., to divide the countries in a more simplified manner with regards to the indicators' performance, it is relevant to describe them from the point of required actions to reach the expected threshold. So, with the same ordinal datasets, we can convert "moderate" and "bad" countries to the category of "an effort should be made", while the good / very good states might be left with the neutral label of nominal "achievement" (Figure 4.12).

Figure 4.11: The message is simplified, though the Germany-Poland focus is preserved. The regional context demonstrates the qualitative differences in a more detailed way, although at first, the two proportional values were in focus

There is an obvious contextual, informative deformation of the message that the SDG indicator visualization can express if an LoM-based data transformation comes on stage. It is not yet manageable to recognize the possible typification' of what this transformation would be able to bring to the users and readers, but this should be addressed that such an aspect of data transformation makes a significant impact on the graphical output. But as this experiment was made on a relatively low level of complexity, the options within a more sophisticated ATL structure should be presented.

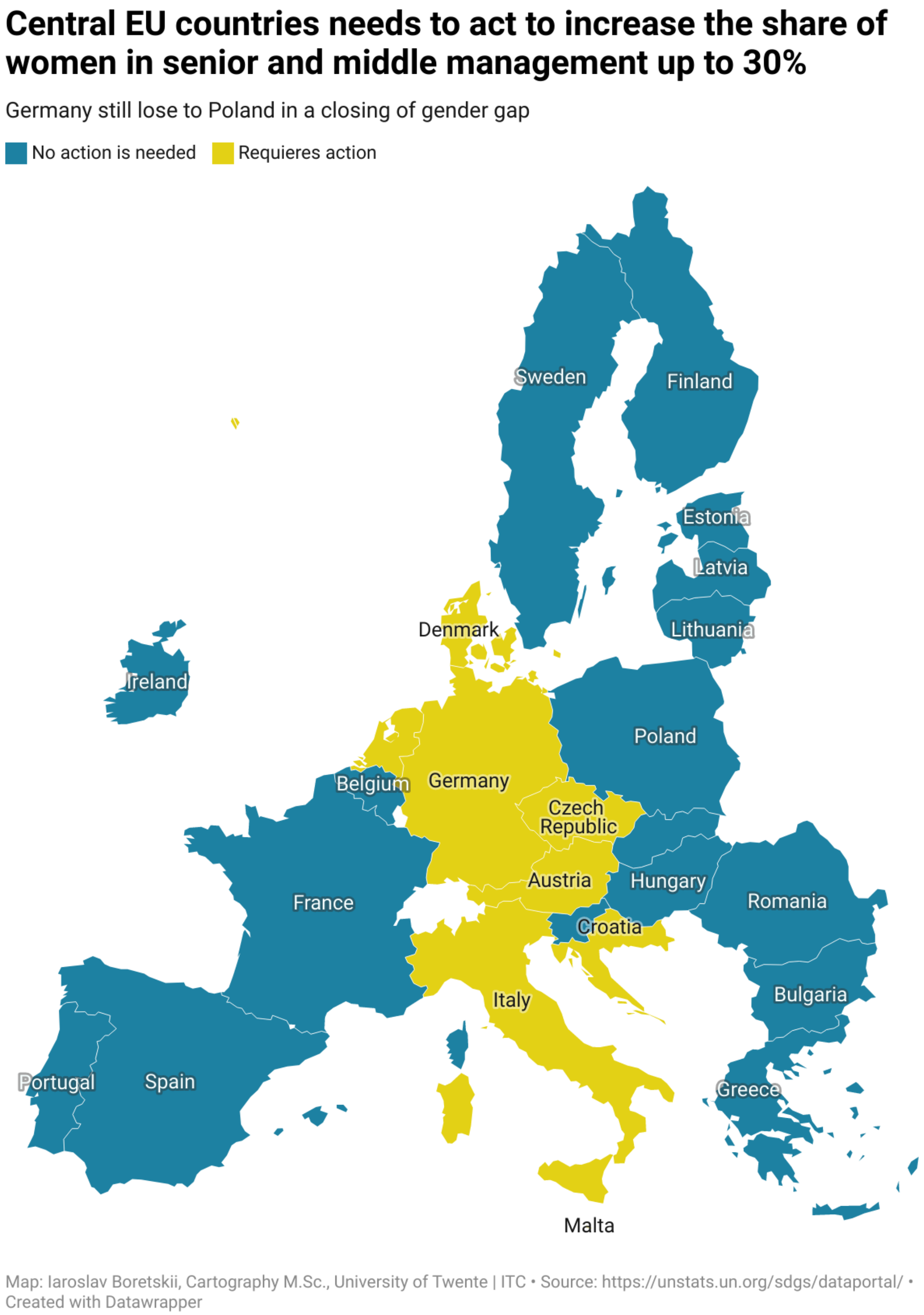

Figure 4.12: Focusing on the EU countries with the nominal transformation it becomes more feasible to see the spatial patterns that Poland and Germany are nominally following

4.2.2 Visualizing higher ATL complexity: Goal 3 in Asia over time

The second scenario involves more countries, extended time, and the original indicators based on the index and rate. We have an attempt to build an instant visualization with indicators 3.4.2 Suicide mortality rate and 3.8.1 Coverage of essential health services for Central and Southern Asia countries in the time frame of 2005-2015 (Table 4.3). Data is obtained from the likewise origin of the UN SDG database repository (UN Department of Economical and Social Affairs, 2022).

Table 4.3: The initial filtered data for Central and Southern Asia countries for two indicators of 3.8.1 and 3.4.2 in a two-time stamps range of 2005-2015

Country	3.4.2-2005, per 100.000 inhabitants	3.4.2-2015, per 100.000 inhabitants	3.8.1-2005, index, 0 to 140	3.8.1-2015, index, 0 to 140
Afghanistan	10.8	8.6	52	68
Bangladesh	13.8	10	60	92
Bhutan	11.8	12.4	78	118
India	17.1	13.7	35	52
Iran	15	18.6	108	140
Kazakhstan	68.2	39.3	60	74
Kyrgyzstan	21.7	16.1	53	68
Nepal	27	30.8	58	102
Pakistan	26	26.4	60	84
Sri Lanka	77.4	48.4	100	126
Tajikistan	5	5.4	46	67
Turkmenistan	25	9.4	59	69
Uzbekistan	13.9	12.7	56	72

Initially, looking at the complex structure of data representation, we see a more complex spatial structure (more countries) with two separate time stamps. Moreover, the indicator value types themselves are not represented by proportions as in the previous example, but by index and rates. The matrix can offer some of the original options (Figure 4.13). The cartographic theory offers a choice of a static map with specific graphical variables, a series of static maps, and animated maps (Kraak & Ormeling, 2020). Exploratory visualizations with original numbers that are used in the preliminary analysis could have been the final implementation if no transformations were applied, so the series of static maps also could have been selected as the visualization decision (Figure 4.14).

Although, there is a space for a certain LoM-based transformation, and in our favor is to come up with an instant, concise visualization instead. Picking the thresholds on the ordinal and nominal measurement levels transformation is the task to complete here and as was discussed earlier, multiple methods of data transformation from numeric to qualitative can be made (Chapter 2.3). The mix of the statistical, methodological, and arbitrary classifications in our case are involved respectively. We may outline the further steps of data transformation to design the final visualization decision:

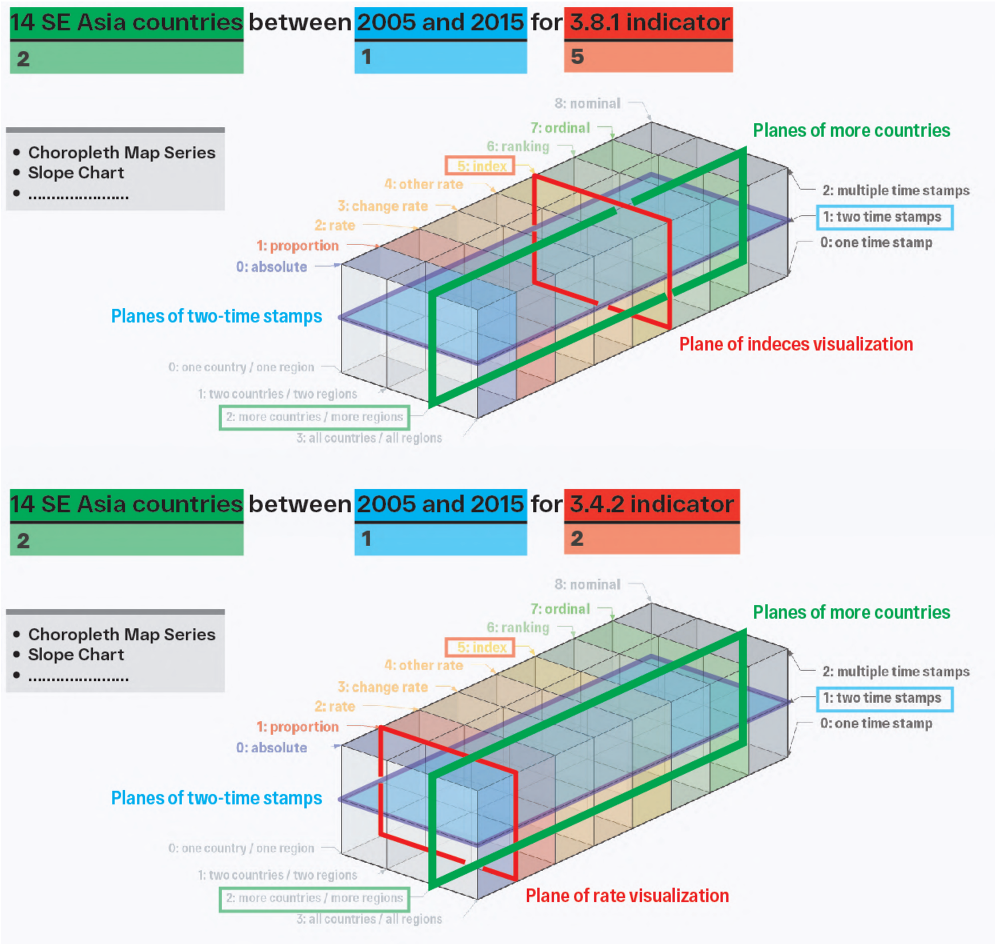

Figure 4.13: Available options for the second example

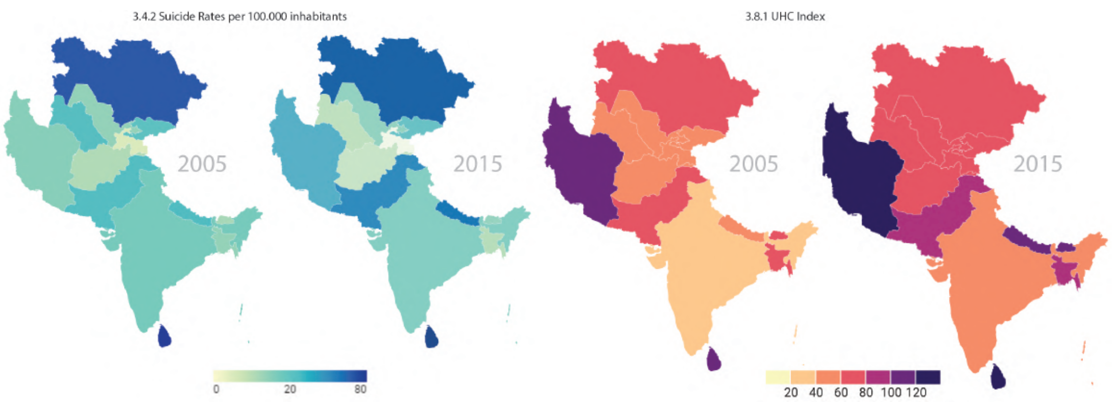

Figure 4.14: Generic, exploratory choropleth maps series of the aforementioned indicators with arbitrary color scheme

1) Quantitatively compare all countries in the region between these two-time stamps for each indicator

2) For 2005 and 2015 time stamps assign an ordinal estimation to all countries based on the indicator 3.4.2 by WHO ordinal classification (World Health Organization, 2021).

3) For 2005 and 2015 time stamps assign an ordinal estimation to all countries based on the indicator 3.8.1 by Sustainable Development Report classification (The Sustainable Development Report, 2021).

The temporal aspects are involved as follows:

4) Assign an ordinal estimation of the ordinal change rate for 3.4.2 with regards to global progress

5) Assign an ordinal estimation of the ordinal change rate for 3.8.1 with regards to global progress

Since it is quite a common approach to compare the countries' indexes by ranking them, we may apply a ranking (which is likewise ordinal) transformation to the 3.8.1 indicator:

6) Assign global ranks in accordance with the 3.8.1 indicator for 2005 and 2015

7) Assign an ordinal estimation of the ranking change rate for 3.8.1 regional-wise

Instead of having classified (or continuous) values for each time stamp of one attribute, ordinal-(ranking)-based facts on whether the progress for countries was done between the two-time stamps or not are likely to work better in a general context. Once again, the LoM-based data transformation simplifies the message details, reveals valuable, factual information, and focuses attention on the main things in a verbal way; in the end, with ordinal and qualitative data the mass audience must not involve deeper in the meanings of the values themselves behind the indicators. The transformed data output is presented in Table 4.4.

Table 4.4: The initial filtered data for Central and Southern Asia countries for two indicators of 3.8.1 and 3.4.2 in a two-time stamps range of 2005-2015

Country	**3.4.2-2005, ordinal estimation**	**3.4.2-2015, ordinal estimation**	**3.4.2, ordinal progress**	**3.8.1-2005, ranking global**	**3.8.1-2015, ranking global**	**3.8.1, delta ranking**	**3.8.1, delta ranking ordinal**
Afghanistan	Moderate	Normal	Positive	110	99	11	Good ascent
Bangladesh	Moderate	Moderate	Positive	69	10	59	Great ascent
Bhutan	Moderate	Moderate	No Change	10	4	6	Good ascent
India	Moderate	Moderate	Positive	142	142	0	No change
Iran	Moderate	Moderate	Negative	1	1	0	No change
Kazakhstan	Extremely serious	Very serious	Very Positive	69	65	4	Good ascent
Kyrgyzstan	Serious	Moderate	Positive	107	99	8	Good ascent
Nepal	Serious	Very serious	Negative	80	8	72	Great ascent
Pakistan	Serious	Serious	No Change	69	23	46	Great ascent
Sri Lanka	Extremely serious	Very serious	Very Positive	2	3	-1	No change
Tajikistan	Normal	Normal	No Change	128	103	25	Good ascent
Turkmenistan	Serious	Normal	Very Positive	74	94	-20	Big drop
Uzbekistan	Moderate	Moderate	Positive	88	78	10	Good ascent

Looking at the resulting picture, the changes for many countries in terms of the suicide rate and the UHS index are evident in many respects. We see that according to the classification of the number of suicides, the qualities of many countries leaned in a more positive direction. It can be determined that if the country has changed by one quality (degree), then these changes over these years can be defined as *positive*, and if more, then even as *very positive.*

In the case of the healthcare accessibility index, we chose the ranking of countries and can assess the dynamics from the position of overtaking countries in the ranking of this index. We also see an overall positive trend for most countries. Moreover, according to the data, five countries from these two regions are standing in the top ten of the global UHC index ranking. It is possible to indicate serious overtaking in the ranking by some, a slight or good ascent, and no (weak, imperceptible) change. The choice of ranking method, in this case, throws out some significant hidden factors (e.g., problems in other countries may weaken them in the world rankings, while the rest without much change may turn out to be higher by keeping the same value with no overall progress behind). However, these elaborations on data transformation and corresponding visualization do not encourage manipulation of the data and its visual encoding, but rather demonstrate the potential for a possible ordinal and nominal transformation.

Figure 4.15 represents the final static map with specific graphical variables, expressing the temporal character of the indicators 3.4.1 and 3.8.2. Finally, the visual representation of this complexity consists of encoded ordinal grades of change rate, applied specifically for Central and Southern Asia countries.

The color, graphical and semantic aspects of this map are tied to some personal feelings of the author on how the data can be visualized. In general, we see the combination of two attributes that are *tonally* different and thus delimit themselves. The size of the territory of countries within the regions made it possible to easily place ordinal arrow icons as one attribute in the centroids of countries' polygons, which are eventually encoded utilizing a color to represent different attribute changes over time (choropleth map).

Thus, we see that the SDGs contain within themselves rather complex features of formation through indicators and targets. Moreover, these indicators themselves are quite detailed and mostly unambiguously described concepts from the standpoint of metadata and content, which, under the conditions of the Attribute-Location-Time model, can be transformed and visualized in a large variety of ways. However, the challenges behind visualizing this data are very acute and require special attention to overcome them.

Therefore, the provided pictures and examples have vividly demonstrated the gist of the selected concept of the recommender system for the SDG indicators. The prototype implementation of the system is representing another step for the system validation and the completion of the thesis; the next chapter is specifically focusing on this particular part of the work.

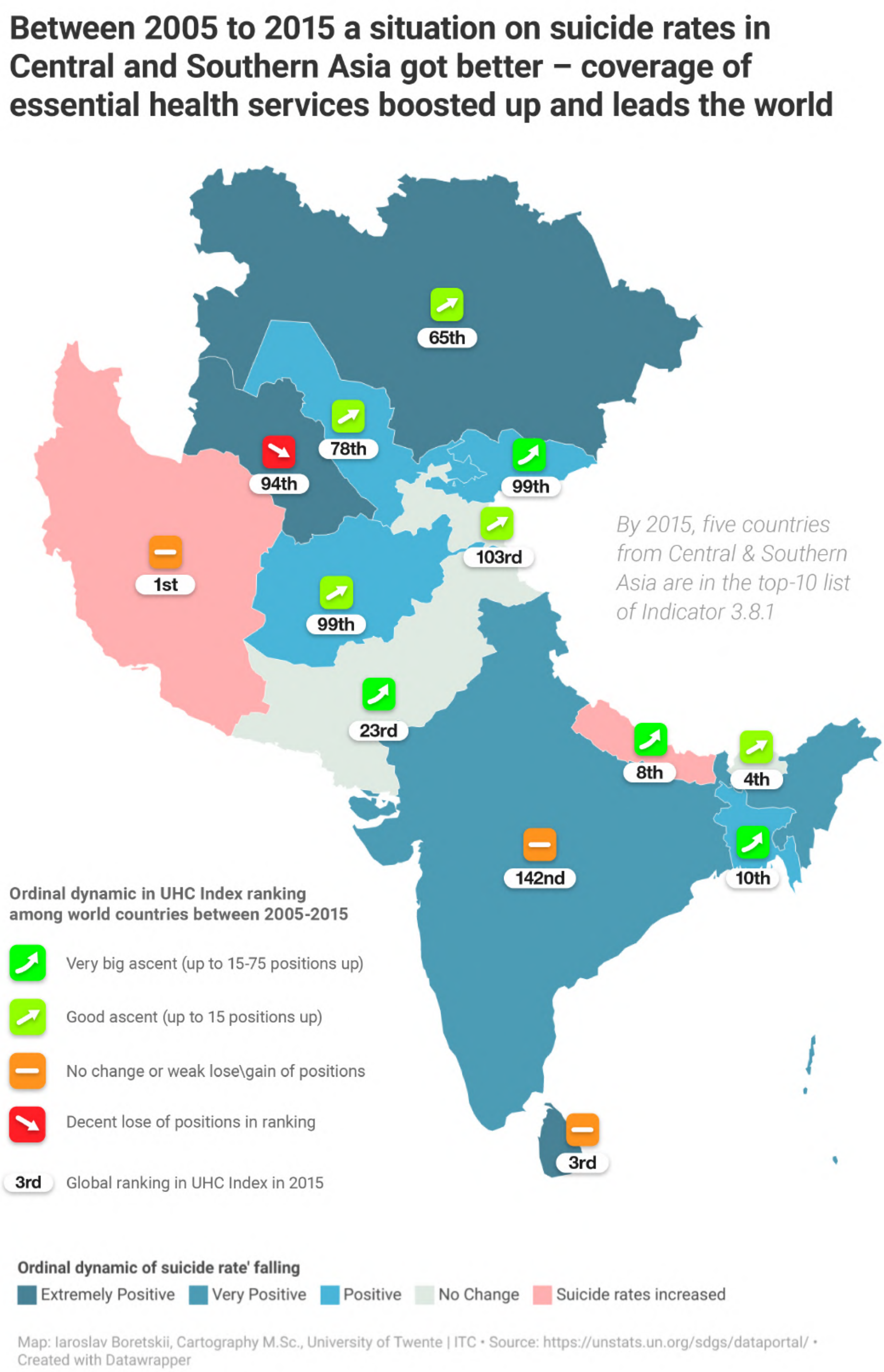

Figure 4.15: Central and Southern Asia countries concerning SDG 3.4.1 and 3.8.2 indicators between 2005 and 2015

Chapter 5

Prototype implementation

The practical part of the current research, as it was outlined in the introduction, is the design and development of an interactive prototype. The conceptual practices from the previous chapter are taken into account. We are focused on several aspects during the prototype development. To implement the prototype for the conceptual model and literature review, the following requirements were followed:

Functionality and Accessibility are represented through the

options to determine the dataset configuration (coordinates) through time, location, and indicator' level of measurement and its possible transformations employing interactive elements (dropdown menus or sliders); additionally, the presence of an enriched, demonstrative set of visualization types must be given to the user as recommendations, based on the selection. The key feature of the prototype is that it includes the inherent rules of transformations, bringing the alternative options that are based on the data transformation schema. Regarding accessibility, the prototype should be designed in a way that it can be available via web environment through desktop devices for all users with an internet connection.

Understandability and Learnability of prototype relates to design (UX/UI) and implies clear elements and layout, considering the variety of inner users' specifications. The inclusion of the informational, exploratory blocks, allowing the users to seamlessly understand the prototype' and system' gist is also relevant for this task. This can be achieved by the usage of popular instruments that are currently in use for working with data and visualizations (see section 2.4).

5.1 Functionality and Accessibility

To implement the desired functionality, it was important to design the inherent data model and knowledge base representation. The proposed matrix model implies the presence of its size through coordinates. This can be digitally implemented by utilizing short tables, which are connected to one database (Figure 5.1).

Three tables *attribute, location* and *time* are the matrix' axes. The table *vis types* represents the collection of distinctive visualization types (e.g., Bar Chart, Slope Chart, Cartogram), most of which were taken from Mapping For a Sustainable World book' sections

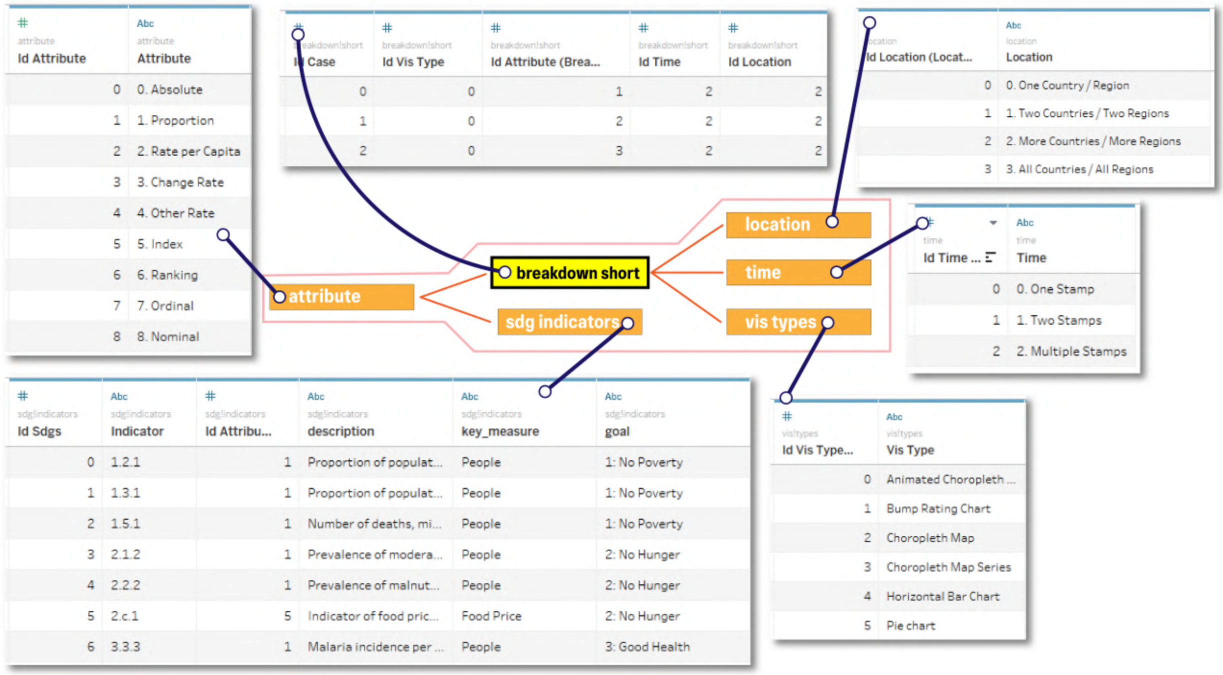

Figure 5.1: The back-end side of the prototype with all involved data

on diagrams and map types. The *sdg indicators* table contains the SDG indicators numbers with corresponding goals and inherent attributes (level of measurement), connected to the aforementioned table. The resulting table of *breakdown short* involves all the tables in one, allowing us to use as the main data source of the prototype main interface. There are five *id columns* that were joined together in a resulting table where multiple case scenarios with certain data visualization types and corresponding ATL-coordinates are generated. The central dependent table remains scalable as more different *vis types* can be added and described through matrix parameters.

Such a data model with multiple tables was designed to satisfy the data handling in the program framework of Tableau Software, which was decided to use as a developing environment for a prototype. Because the prototype's goal is rather a demonstration and validation of conceptual ideas, Tableau Software fits well for the design of dashboards, so the prototype itself can be considered as such. Additionally, the available licensing allows for seamless publish the dashboard on the web through Tableau Public, so the accessibility issue is resolved here.

5.2 Understandability and Learnability. UX/UI

The understandability and learnability of the prototype can be achieved by focusing on making it easy, comfortable *usability,* which is critical to the overall user experience (Kraak et al., 2020). The prototype visual outcome was a matter of multiple iterations, as some of the features and interaction scenarios were often revised. Figure 5.2 represents the first mockup of the MetaDataViz interface.

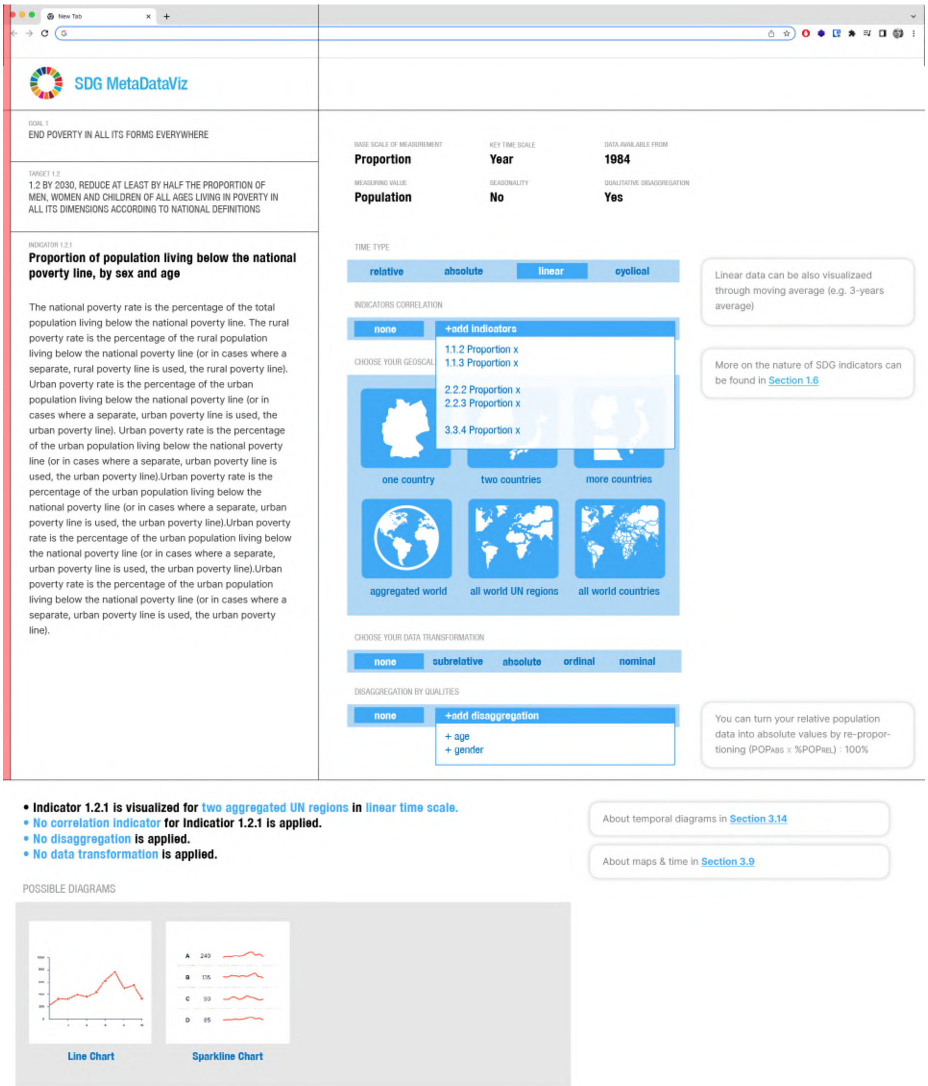

Figure 5.2: SDG MetaDataViz first mock-up tries to comprehend the possible web-based implementation of the conceptual system

The initial idea of the prototype is intended to present multiple interactive filters alongside corresponding indicator information at the top so that certain data visualization options can be selected. Further down the description of the filtering request alongside the available diagrams and maps options are presented. Additionally, it was implied to add the hyperlinked footnotes to the aforementioned Mapping For a Sustainable World book over the prototype sections. The early stage prototype design has founded the desirable UX/UI structure of the prototype.

After the conceptual system was finally refined and described, the decision on the Tableau Software usage has been made as it was implying the non-code solutions for a prototype. Although Tableau itself is quite a complex tool for learning all of its features, the public dashboards that can be designed there can be relatively easy to use. The realization of the prototype is made via such a dashboard, which is designed through a connection of multiple worksheets with interdependent filters in one layout. Figure 5.4 shows the first designed version of Tableau-based dashboard. It includes the filtering options at the top with inherent options beneath. Alternative options imply the available data transformation attributes. Here it is important to mention the restrictions by Tableau, as the solution for the automatic alternative options bringing is not as straightforward as it should be; the user should specify the available alternative options according to the transformation schema from the Figure (3.5). Moreover, it was decided to provide users with the *Mappable Values* table alongside the data transformation schema. In the final version of the prototype, this feature as well as the transformation schema was eliminated due to the vast graphically occupied space in the prototype; it was decided to keep more visual clarity and simplicity.

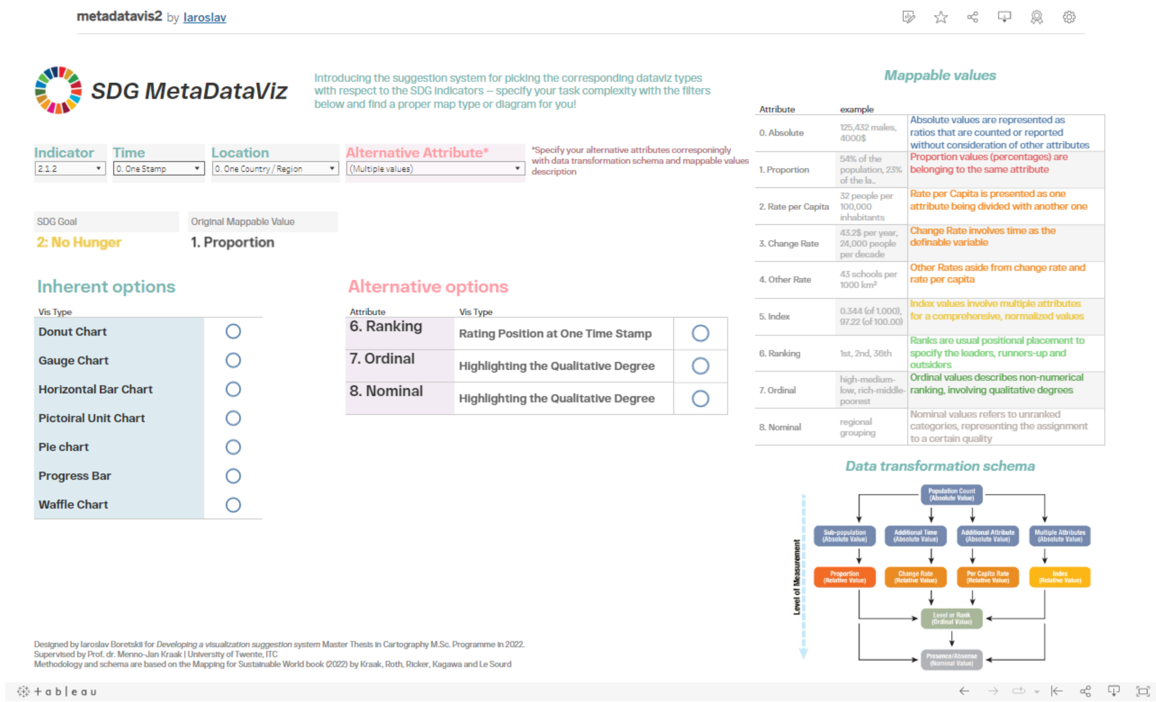

Figure 5.3: SDG MetaDataViz prototype in Tableau Public, first version

Multiple iterations on the design process of the prototype were done; the final iterations allowed to outline of the distinctive interface sections:

1) Filtering responsive area;
2) Metadata and information block;
3) Originally advised options;
4) Alternative options, brought concerning the data transformation abilities.

The combination of these blocks was designed in the final version iteration of the prototype (Figure 5.4).

By interacting with the upper-left part of the responsive filters, the advised and alternative options sections are updated; they are presented as instant icons with corresponding names. The majority of the used pictures in the prototype are taken from the aforementioned in section 2.5 data visualization collection (Data Viz Project by Ferdio, 2022) concerning the Creative Commons Licensing, implying non-commercial use as well as the direct referencing to the authors. The advised options are dependent on the ATL filters only, while the

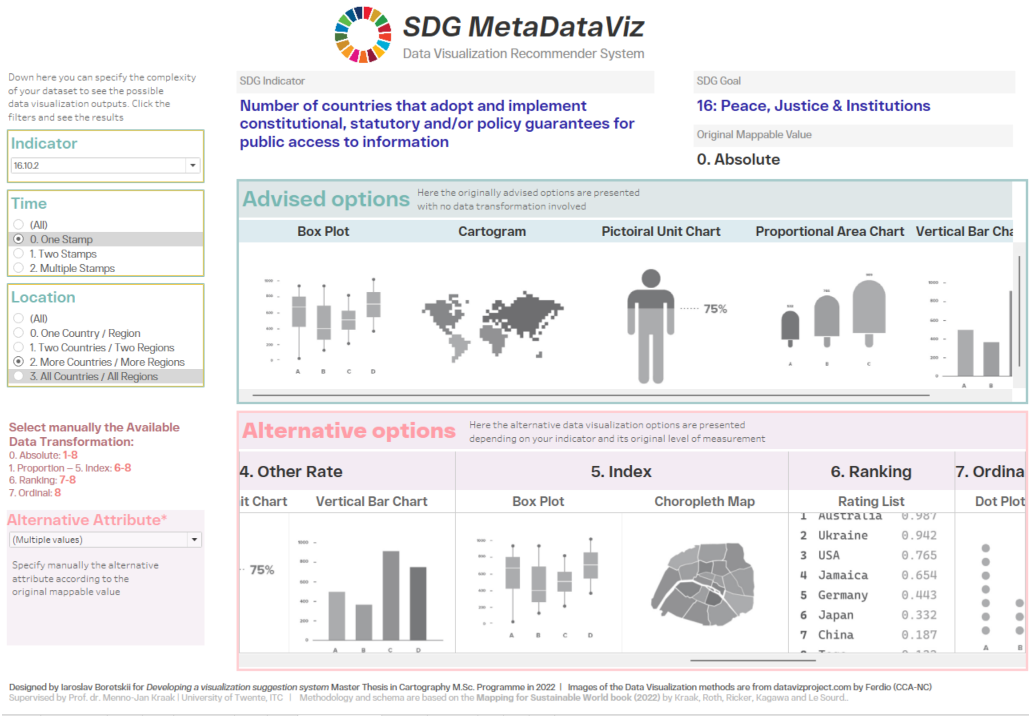

Figure 5.4: SDG MetaDataViz prototype in Tableau Public, final version

alternative options should be refactored by alternative attribute filtering manually. Nonetheless, the alternative options area is also getting updated with all the applied filters. Finally, the informative top blocks of SDG indicator, Goal, and Original Mappable Value are updated concerning the filters, bringing the use support. The learnability issue for the prototype was decided to solve through small text footnotes that explain the purpose of the corresponding area and give instructions on what should be done to update the options output. Figure 5.5 demonstrates the altering of the options' output depending on the filters. E.g., the top example is involving 2.2.2 example with Two-Stamps time complexity and More countries location complexity. The system output allows seeing the instant recommendations for both original and non-transformed levels of measurement. With such a configuration it is possible to use, e.g., *Rating List with Position Gain or Lost Addition* to see the two-stamp change of the indicator which has been undergoing the data transformation. The bottom example demonstrates the 4.b.1 indicator being visualized for One-Stamp for Only country. We see that the advised options on absolute number visualization for a country are not that enriched, while the alternative options are various and can be used in the design process. It is important to mention that the order in which the recommendations (both advised and alternative) specimens are given is not following a defined methodology or rules. It is only implied that all of the recommended options can fit the defined ATL parameters and their possible level of measurement transformation, but the final decision on which visualization type to pick stays behind the user. The subject of data visualization recommendations' order remains the methodological challenge.

The conclusions of the prototype evaluation and limitations are reviewed in the following chapters.

Figure 5.5: SDG MetaDataViz prototype state examples depending on the filter configuration

Chapter 6

Results

This chapter describes the achieved results of the research, focusing on the outcomes and their pros and cons. The literature review allowed us to see multiple aspects of data visualization and focus on some of them. SDG indicators and visualization challenges have been also reviewed to see where the visualization recommender system for the indicators might be helpful. Finally, the conceptual model of the recommender system was elaborated, focusing on the suggestion of the visualization types according to the attribute (more precisely, its original level of measurement and available transformation options), time, and location. Such a realization can be presented through a matrix with coordinates of each parameter's complexity. The practical realization of the system employing the Tableau-based prototype has allowed us to demonstrate the feasibility of the developed matrix.

6.1 Recommender system constraints

The foremost constraints of the system are related to the given time and location configuration. By introducing a classification of location on a scale of one-two-more-all, I do not take into account the regional, real geographical features of the dataset. The attributes and their chosen levels of measurement with corresponding mappable values are also represented by a limited range of examples; obviously, there could be many more added. The system does not take into account the possibility of creating pivot tables by country. For example, for the task of visualizing the number of countries assigned to some categories, several approaches can be applied. Their number can be considered both as an absolute number or as a percentage, which can lead to the proposal of different types of visualization (e.g., 34 countries against 17 percent of the whole number of countries can be differently visualized).

In addition, another important limitation is the impossibility to configure visualization types at some syntactic level, taking into account the SDGs. For example, the visualization of temporal dynamics of a certain indicator can be realized by arranging visual artifacts along both the vertical and horizontal as well as the radial axis. Instead of taking into account some grammatical feature of visualization of time (largely inspired by the example of VisDNA), in this conceptual system, it is necessary to take into account parameters ATL only, thus, assigning the distinctive data visualization specimens manually and only within the ATL framework.

Finally, the main constraint of the recommender system is its limited focus on the ATL parameters only. This system does not consider in any way the possible peculiarities of visualizations in terms of the choice of graphic design, suitable visualization software, audience, etc. In this respect, it is poorly scalable and is locked only on the marked features of the data representation. However, it can serve as an excellent filter for refining the system already into a hybrid-based model (see section 2.9).

6.2 Prototype characteristics

It is important to stress that no user study and third-party usability testing were done. It is possible to evaluate the prototype only on the prescribed requirements that were broadly determined at the beginning of chapter 5. It is better to think about the prototype in a demonstrative context, as the initial purpose of the prototype is to validate the conceptual system in a computer-based environment.

The prototype is publicly available through Tableau Public as a dashboard and the highlighted link below:

MetaDataVizBeta by Iaroslav Boretskii — Tableau Public

All the important elements of the prototype are responsive and clickable, been able to be used in a desktop web environment smoothly. There are only three interactive actions that can be done with the prototype:

- Three basic responsive filters;
- One alternative recommendations filter;
- Horizontal scrolling of the presented data visualization specimens for alternative and advised options.

These interactive elements and their corresponding interface elements are supported by inherent footnotes that allow the user to understand the prototype's functionality in an easy learning manner. In sum, the prototype follows the generic demonstrative requirements and should be further developed with consideration of actual user study to find the future drivers of its improvement.

Chapter 7

Conclusions and discussion

The chapter concludes the research and answers the research questions. Limitations and future research suggestions are provided

7.1 Conclusions

A general conclusion can be provided at the end of this research. I was able to see that users, especially cartographers and data visualization designers are facing various challenges in SDG indicators data visualization possibilities; the selected matrix methodology implies that some of the aspects were taken into account in this research to partially assist in the design process. Still, there is an obvious need for user study to evaluate the final output of the prototype and conceptual model that has been developed.

7.2 Answers to research questions

• **RQ A1 What are the main features of data visualization as a concept?**

Data visualization deals with the translation of qualitative, quantitative, and spatio-temporal data into a correspondingly appropriate visual representation that might be relevant to its users. This involves data visualization in multiple technical, artful, and social aspects that are very important to consider when the data visualization is getting created.

• **RQ A2 What methodologies classify and allocate different data visualization types?**

There are systematical and arbitrary methodologies that were reviewed in chapter 2. Firstly reviewed examples are using a grammar-based approach where the graphics break down into syntactic components. The second is rather authorial where distinctive data visualization specimens are assigned to particular categories (e.g., function, input type, etc.) with provided examples.

• **RQ A3 How do data transformation and representation affect the data graphics output?**

The level of measurement data transformation allows us to present the data in a more generalized manner; by converting the numbers into qualitative descriptions, it is qualities and degrees that can be discerned in graphics, not the numbers.

- **RQ A4 What concept stands behind Sustainable Development Goals?**

SDGs are aimed to achieve resilient progress by bringing down the economical, ecological and social global inequalities. Data-driven methodology of the goals allows us to track progress feasibly by comparing global indicators among countries over time.

- **RQ A5 What are the challenges for the user in visualizing SDG indicators?**

The challenges are respectively migrating from the corresponding data visualization aspects. The use of data visualization technology, metadata- and data-related issues, involving data transformation and decision-making in the design process can be highlighted as one the significant ones.

- **RQ B1 What is the reasoning behind the recommender systems for data visualization?**

The recommender systems bring multiple options in possible data visualization variants. Moreover, our example includes transformation options concerning the level of measurement, which allows us to see even more options for visualizing the same datasets.

- **RQ B2 Which methodologies are used to implement the recommender systems for data visualization types?**

There are data-driven, expert knowledge-based, and hybrid approaches to developing a visualization recommender system. The first is based on the computational methods of machine learning and deep neural networks and suggests the visualization types by analyzing the given prompt. The second methodology involves certain rules and logical restrictions, forming a knowledge representation of the visual artifacts with corresponding characteristics. The third approach combines the aforementioned methodologies into one.

In my recommender system, there is an expert knowledge-based approach was used.

- **RQ C1 Which data visualization conceptual aspects are in the focus of the recommender system?**

Data representation and data transformation are the main aspects of the recommender system. The conceptual part of the system is focusing on the initial dataset representation; the ability to go down with the level of measurement provides more alternative options for data visualization outputs.

- **RQ C2 What characteristics does the recommender system have?**

The system advises the options for SDG indicators visualization depending on the inner configuration of the dataset, which is described through attribute (its level of measurement), time, and location. The 3-parameters dependence is expressed as a matrix where various data visualization types can be allocated.

The system is focusing on the SDG indicators data representations. Defined matrix parameters are the only variables that are describing distinctive data visualization types. There are defined degrees of time (one stamp, two stamps, and multiple stamps) and location (one country, two countries, more countries, all countries) alongside the attributes that are connected to SDG mappable values (absolute value, proportion, rate, change rate, another rate, index, ranking, ordinal, nominal). Moreover, the possible data transformation options are included, so the visualization options are presented depending on the initial level of measurement of the dataset.

Together these parameters shape the system's characteristics and the following prototype.

• **RQ D1 What requirements are applied to the prototype?**

The prototype should have a corresponding functionality and be able to model the conceptual matrix into a computer-based one. Additionally, the prototype must have accessibility and be published on the web to be available for everyone with an internet connection through a web browser. Moreover, the prototype should be understandable, have a clear interface, and have the logic of interaction between the programming elements. Finally, to provide a better understanding of the prototype usage, some navigational learnability elements should be implemented, such as an interface guide or use tips.

• **RQ D2 Which technology or software is used to design the prototype?**

The prototype is using the programming infrastructure of Tableau Software, published through the free-to-use Tableau Public service. The prototype is designed as an instant dashboard with interactive filters that further bring the data visualization options to the users. The inner data model is designed concerning the Tableau specifications as well as the conceptual model configuration.

• **RQ D3 What are the interface-related (UX/UI) demands that should be considered to design the prototype?**

The prototype's interface should user-centered, be simple, and minimalist in a sense of possible interactions and visual characteristics. It should be responsive, keeping easy usability that leads to a successful and satisfying outcome.

7.3 Research challenges

During the span of the research, some obstacles were correcting the course of the study:

Firstly, the initially proposed topic description was rather vague, having no sign of *a sound methodology*. It has forced me to narrow down the possible research scope, taking SDGs as the key framework for a data visualization recommender system development. It was proved in the literature review chapter that data visualization is a complex concept, so certain compromises and limitations must have been applied to the conceptual system. It became helpful to visualize the aforementioned examples of different ATL parameters; I was able to see what obstacles I meet during the design process. Nonetheless, I was still struggling to find a middle ground between the data transformation schema and the time and location parameters preserved. When I finally managed to grope the conceptual basis in form of a matrix, I felt pretty relieved as I was able to rapidly develop the prototype and validate the system.

It is also important to highlight the affecting circumstances, such as writing the thesis during the COVID-19 pandemic and the military conflict with my home country been involved. It has led to multiple constraints, mainly related to the outcomes of my mental health state.

There were merely five months to develop a recommender system from scratch and formalize all the aspects into one, solid master's work. I could not concentrate steadily during such a short time span to come up with all the cool features being implemented and included.

Finally, I believe that the part of the work that should have been done is the user study to validate the outlined system. Mainly the decisions on how the prototype and conceptual systems should be made are based on the literature review and software examples, but not on the empirical study that could specify the users' needs, especially the ones who are directly involved in the SDG indicators visualization work.

7.4 Future Research and Suggestions

In my opinion, at least three aspects should be considered for future development.

Firstly, the approach to the visualization recommendation system should be developed with the hybrid model, involving the data-driven algorithms in the outlined expert knowledge-based representation. The ML-based examples are demonstrating stunning results, so it would be interesting to see it concerning the SDG specifications.

The second comment is dedicated to the expansion of prototype (or final instrument) technical capabilities. For example, it would be more efficient to add a feature with SDG indicator dataset configuration being automatically determined through the processing of the inserted dataset in *csv* or any other format in the system. This can be done by linking the UN-database API and the usage of data visualization libraries (e.g., VegaLite, D3) in web-based prototypes.

Such a feature can be developed through the processing code that would *read* the dataset and categorize the time, attribute and location respectively. This would make the prototype more advanced and will allow the user to work directly with the dataset specification itself. It would be also nice to alter the prototype programming environment to open-source libraries and instruments to be more independent from other software requirements.

Finally, a proper user study should be made to understand the exact demand character of the prototype, since most of the decisions on prototype design in the thesis were rather hypothetical. The prototype and a conceptual system testing and evaluation were based on the initially proposed requirements and it is important to validate the purposefulness of the recommender system.

List of Figures

References

Bertin, J. (1967). *Semiology of graphics: Diagrams, networks, maps*. Madison: Gauthier-Villars.

Bertin, J. (1977). *). (original: Le graphique et le traitement graphique de l'information)*. Berlin: Walter de Gruyter Co.

Cairo, A. (2016). *The truthful art: Data, charts, and maps for communication*. New Riders Publishing.

Card, S., & Mackinlay, J. (1997). The structure of the information visualization design space. In *Proceedings of viz'97: Visualization conference, information visualization symposium and parallel rendering symposium* (pp. 92–99).

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in information visualization: Using vision to think*.

Chen, M., Ebert, D., Laramee, R., van Liere, R., Ma, K.-L., Ribarsky, W., ... Silver, D. (2009, 03). Data, information, and knowledge in visualization. *Computer Graphics and Applications, IEEE*, *29*, 12 - 19. doi: 10.1109/MCG.2009.6

Chiasson, T., & Gregory, D. (2014). *Data + design: A simple introduction to preparing and visualizing information*. Licensed under a Creative Commons BY-NC-SA 4.0 license. Retrieved from `https://github.com/infoactive/data-design/`

Chrisman, N. (1998, 10). Rethinking levels of measurement for cartography. *Cartography and Geographic Information Science - CARTOGR GEOGR INF SCI*, *25*, 231-242. doi: 10.1559/152304098782383043

Cui, W., Zhang, X., Wang, Y., Huang, R., Chen, B., Fang, L., ... Zhang, D. (2019, November). Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE Transactions on Visualization and Computer Graphics*. Retrieved from `https://www.microsoft.com/en-us/`

research/publication/text-to-viz-automatic-generation-of-infographics-from-proportion-related-natural-language-statements/

Data Viz Project by Ferdio. (2022). Retrieved 2022-08-17, from https://datavizproject.com/

DeFanti, T. A., Brown, M. D., & McCormick, B. H. (1989). Visualization: Expanding scientific and engineering research opportunities. *Computer*, *22*, 12-16. doi: 10.1109/2.35195

Department Of Economic and Social Affairs Statistics Division. (1999). *Standard country or area codes for statistical use.* Retrieved from https://unstats.un.org/unsd/publication/SeriesM/Series_M49_Rev4(1999)_en.pdf

Djonov, E., & Leeuwen, T. V. (2013). Between the grid and composition: Layout in powerpoint's design and use. *Semiotica*, *2013*(197), 1–34. Retrieved 2022-08-23, from https://doi.org/10.1515/sem-2013-0078 doi: doi:10.1515/sem-2013-0078

DVS. (2019,2020). Retrieved from https://www.datavisualizationsociety.org/survey

Dykes, J., MacEachren, A., & Kraak, M. (2005). Exploring geovisualization. *Exploring Geovisualization*, 3.

Engelhardt, Y. (2002). *The language of graphics* (Unpublished doctoral dissertation).

Engelhardt, Y., & Richards, C. (2018). A framework for analyzing and designing diagrams and graphics. In P. Chapman, G. Stapleton, A. Moktefi, S. Perez-Kriz, & F. Bellucci (Eds.), *Diagrammatic representation and inference* (pp. 201–209). Cham: Springer International Publishing.

Engelhardt, Y., & Richards, C. (2021). Retrieved 2022-08-31, from https://visdna.com/

Fairchild, M. D. (2013). *Color appearance models*. John Wiley & Sons.

Few, S. (2004). Show me the numbers. *Analytics Pres*.

Financial Times. (2021, Mar). Charts that work: Ft visual vocabulary guide 2021. Retrieved 2022-08-29, from https://www.ft.com/content/c7bb24c9-964d-479f-ba24-03a2b2df6e85

Frascara, J. (1988). Graphic design: Fine art or social science? *Design Issues*, *5*(1), 18–29. Retrieved 2022-08-17, from http://www.jstor.org/stable/1511556

Fuchs, R., & Hauser, H. (2009). Visualization of multi-variate scientific data. In *Computer graphics forum* (Vol. 28, pp. 1670–1690).

Garrod, M. (2020). *Finding the right tool for the job.* Retrieved from https://github.com/MGarrod1/DVSSC_19

Göker, A., & Myrhaug, H. I. (2002). User context and personalisation. In *Eccbr workshops* (Vol. 2002, pp. 1–7).

Hawkins, E. (2020). *The most important thing to do about climate change is to talk about it.* EGU General Assembly, 2020. Retrieved from https://doi.org/10.5194/egusphere-egu2020-22665

Houtman, L., & Roth, R. E. (2021). Mapping sdgs technical supplement: Version 1.0. Retrieved from http://dx.doi.org/10.5281/ZENODO.5585647 doi: 10.5281/ZENODO.5585647

Hu, K., Bakker, M. A., Li, S., Kraska, T., & Hidalgo, C. (2019). Vizml: A machine learning approach to visualization recommendation. In *Proceedings of chi conference on human* *factors in computing systems* (p. 1–12).

Hák, T., Janousková, S., & Moldan, B. (2016). Sustainable development goals: A need for relevant indicators. *Ecological Indicators*, *60*, 565-573. Retrieved from https://www.sciencedirect.com/science/article/pii/S1470160X15004240 doi: https://doi.org/10.1016/j.ecolind.2015.08.003

ICA. (2016). Retrieved 2022-07-27, from https://icaci.org/maps-and-sustainable-development-goals/

Itten, J. (1970). *The elements of color* (Vol. 4). John Wiley & Sons.

Khoo, A. (2019, Nov). *Tools of the trade.* Retrieved from https://www.datavisualizationsociety.org/annual-survey-challenge/2019/10/22/challenge-title-aakh3

Kirch, W. (2008). Level of measurementlevel of measurement. In *Encyclopedia of public health* (pp. 851–852). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-1-4020-5614-7_1971 doi: 10.1007/978-1-4020-5614-7_1971

Korolkova, A. (2007). Zhivaya tipographika.

Kraak, M. J. (2014). *Mapping time: illustrated by minard's map of napoleon's russian campaign of 1812 / menno-jan kraak*.

Kraak, M.-J., & Ormeling, F. (2020). *Cartography: Visualization of geospatial data*. Fourth edition — Boca Raton; London: CRC Press, 2020.: CRC Press.

Kraak, M. J., Ricker, B., & Engelhardt, Y. (2018). Challenges of mapping sustainable development goals indicators data. *ISPRS International Journal of Geo-Information*, *7*(12). Retrieved from https://www.mdpi.com/2220-9964/7/12/482 doi: 10.3390/ijgi7120482

Kraak, M.-J., Roth, R., Ricker, B., Kagawa, A., & Le Sourd, G. (2020). *Mapping for a sustainable world*.

Laptev, V. (2012). *Izobrazitelnaya statistica. vvedenie v infografiku*. Izdatelstvo Eidos, Saint-Petersburg, Russia. Retrieved from https://www.goodreads.com/book/show/23486950

Laptev, V. (2018). *Russkaya infografika*. Izdatelstvo Politechnicheskogo Universiteta, Saint-Petersburg, Russia. Retrieved from https://www.researchgate.net/profile/Vladimir-Laptev-3/publication/337444662_Russian_Infography/links/5dd7e1c6299bf10c5a275101/Russian-Infography.pdf

Leborg, C. (2006). *Visual grammar: A design handbook (visual design book for designers, book on visual communication)*. Princeton Architectural Press.

Lewi, P. J. (2008). Speaking of graphics. *An Essay*.

Luo, Y., Qin, X., Tang, N., & Li, G. (2018). Deepeye: towards automatic data visualization. In *Proceedings of ieee international conference on data engineering (icde)* (p. 101–112).

Lynch, A., & Sachs, J. (2021). *The united states sustainable development report 2021*. New York: SDSN.

MacEachren, A. M. (2004). *How maps work: representation, visualization, and design*. Guilford Press.

MacEachren, A. M., & Kraak, M.-J. (2001). Research challenges in geovisualization. *Car-*

tography and Geographic Information Science, *28*(1), 3-12. Retrieved from https://doi.org/10.1559/152304001782173970 doi: 10.1559/152304001782173970

MacEachren, A. M., et al. (1994). Visualization in modern cartography: setting the agenda. *Visualization in modern cartography*, *28*(1), 1–12.

Mackinlay, J., Hanrahan, P., & Stolte, C. (2007, 11). Show me: Automatic presentation for visual analysis. *IEEE transactions on visualization and computer graphics*, *13*, 1137-44. doi: 10.1109/TVCG.2007.70594

Minsky, M. (1974). *A framework for representing knowledge.* MIT, Cambridge.

Muehlenhaus, I. (2013). The design and composition of persuasive maps. *Cartography and Geographic Information Science*, *40*(5), 401-414. Retrieved from https://doi.org/10.1080/15230406.2013.783450 doi: 10.1080/15230406.2013.783450

Munzner, T. (2014). *Visualization analysis and design*. CRC Press. Retrieved from https://books.google.de/books?id=dznSBQAAQBAJ

Muth, L. C. (2022, Mar). *A detailed guide to colors in data vis style guides.* Retrieved from https://blog.datawrapper.de/colors-for-data-vis-style-guides/

National Eye Institute. (2015). *Facts about color blindness.* Retrieved 2022-08-17, from https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness

Pirlea, U. S. D. W. M. W., A. F., & Whitby, A. (2020). *Atlas of the sustainable development goals 2020: From world development indicators.* Retrieved from http://datatopics.worldbank.org/sdgatlas/

Qin, X., Luo, Y., Tang, N., & Li, G. (2020, 1). Making data visualization more efficient and effective: a survey. In (Vol. 29, p. 93-117). Springer. doi: 10.1007/s00778-019-00588-3

Reuter, L. H., Tukey, P., Maloney, L. T., Pani, J. R., & Smith, S. (1990). Human perception and visualization. In *Proceedings of the 1st conference on visualization'90* (pp. 401–406).

Richards, C. J. (1984). *Diagrammatics: an investigation aimed at providing a theoretical framework for studying diagrams and for establishing a taxonomy of their fundamental modes of graphic organization.* (Unpublished doctoral dissertation). Royal College of Art.

Rubanov, V., & Filatov, A. (2010). *Intellektual'nye sistemy avtomaticheskogo upravleniya. nechetkoe upravlenie v tekhnicheskih sistemah*. Belgorod: BSTU im. V. G. Shukhova. Retrieved from http://nrsu.bstu.ru/

Sachs, J., Lafortune, G., Kroll, C., Fuller, G., & Woelm, F. (2022). *From crisis to sustainable development: the sdgs as roadmap to 2030 and beyond. sustainable development report 2022*. Cambridge: Cambridge University Press.

SDG Center for Africa and Sustainable Development Solutions Network. (2019). *Sdg center for africa and sustainable development solutions network (2019): Africa sdg index and dashboards report 2019.* Kigali and New York.

Smith, A. R. (1978, aug). Color gamut transform pairs. *SIGGRAPH Comput. Graph.*, *12*(3), 12–19. Retrieved from https://doi.org/10.1145/965139.807361 doi: 10.1145/965139.807361

Sowa, J. F. (1987). *Semantic networks.*

Stevens, S. S. (1946). On the theory of scales of measurement. *Science (New York, N.Y.)*, *103*(2684), 677–680. Retrieved from http://dx.doi.org/10.1126/science.103.2684.677 doi: 10.1126/science.103.2684.677

The Sustainable Development Report. (2021). Retrieved from https://dashboards.sdgindex.org/map/indicators/universal-health-coverage-uhc-index-of-service-coverage

Tory, M., & Möller, T. (2002). A model-based visualization taxonomy. *School of Computing Science, Simon Fraser University*, 39.

Tory, M., & Moller, T. (2004). Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, *10*(1), 72-84. doi: 10.1109/TVCG.2004.1260759

Tufte, E. (1990). *Envisioning information* (Vol. 126). Graphics press Cheshire, CT.

Tufte, E. (2001). *The visual display of quantitative information*. Cheshire: Graphic Press.–2001.–213 p.

Tversky, B. (1995). Cognitive origins of graphic productions. In *Understanding images: Finding meaning in digital imagery* (pp. 29–53).

Twyman, M. (1979). A schema for the study of graphic language (tutorial paper). In *Processing of visible language* (pp. 117–150). Springer.

Tyner, J. A. (2014). *Principles of map design*. Guilford Publications.

UN Department of Economical and Social Affairs. (2022). *Sdg global database gives you access to data on more than 210 sdg indicators for countries across the globe.* Retrieved 2022-07-29, from https://unstats.un.org/sdgs/dataportal/

UN Economic and Social Council. (2012). *Millennium development goals and post-2015 development agenda.* Retrieved 2022-07-27, from https://www.un.org/en/ecosoc/about/mdg.shtml

United Nations. (2017). *General assembly, work of the statistical commission pertaining to the 2030 agenda for sustainable development: report of the secretary-general.* Retrieved from https://documents-dds-ny.un.org/doc/UNDOC/GEN/N17/207/63/PDF/N1720763.pdf?OpenElement

United Nations. (2021). *Human development index, "composite indices — hdi and beyond.* Retrieved 2022-08-17, from https://hdr.undp.org/data-center/documentation-and-downloads

United Nations Statistics Division. (2022a). *Iaeg-sdgs: Tier classification for global sdg indicators.* Retrieved 2022-07-27, from https://unstats.un.org/sdgs/iaeg-sdgs/tier-classification/

United Nations Statistics Division. (2022b). *Sdg indicators: Metadata repository.* Retrieved 2022-07-27, from https://unstats.un.org/sdgs/metadata/

Van Wijk, J. J. (2005). The value of visualization. In *Vis 05. ieee visualization, 2005.* (pp. 79–86).

Visual-literacy.org. (2007). *A periodic table of visualization methods.* Retrieved 2022-08-31, from https://www.visual-literacy.org/periodic_table/periodic_table.html

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

Wexler, S. (2019, Sep). *10 fascinating findings from the data visualization society 2019 community survey.* Retrieved from https://www.datarevelations.com/dvs-survey/

World Health Organization. (2021). Suicide rate estimates, age-standardized, estimates by country. Retrieved from https://apps.who.int/gho/data/node.main.MHSUICIDEASDR?lang=en%7Caccess-date=12

Zhou, M., & Feiner, S. (1998). Visual task characterization for automated visual discourse synthesis. In *Proceedings of the acm human factors in computing systems conference (chi'98* (p. 392–399).

Zhu, S., Sun, G., Jiang, Q., Zha, M., & Liang, R. (2020). A survey on automatic infographics and visualization recommendations. *Visual Informatics*, *4*(3), 24-40. Retrieved from https://www.sciencedirect.com/science/article/pii/S2468502X20300292 doi: https://doi.org/10.1016/j.visinf.2020.07.002

Appendix A

Appendix

A.1 Table of used data visualization types in the system

Visualization Type	Class	Description
Animated Choropleth Map	Map	Uses digital system time to update the map display. Works with multiple time stamps for multiple geoentities.
Bump Rating Chart	Chart	Uses form of a line plot designed for exploring changes in ranking over time.
Choropleth Map	Map	Uses a thematic map in which administrative areas are colored or shaded according to the quantitative range.
Choropleth Map Series	Map	Series of maps allow to see the spatial distribution in detail over time.
Horizontal Bar Chart	Chart	A chart in the form of rectangular horizontal bars. Works good with two and multiple countries at one time stamp.
Pie chart	Diagram	A circular statistical graphic, which is divided into slices to illustrate numerical proportion. One time stamp for one country is the better in usage.
Pie chart series	Diagram	Series of the circular static graphics
Rating List	Table	The table with sorted by ranking locations.
Slope Chart	Diagram	The line chart's with two-time stamps.
Vertical Bar Chart	Diagram	A chart in the form of rectangular vertical bars. Works good with two and multiple countries at one time stamp.
Cartogram	Map	A cartogram is a map in which the geometry of regions is distorted in order to convey the information of an alternate variable.
Donut Chart	Diagram	A donut-like statistical graphic, which is divided into slices to illustrate numerical proportion. One time stamp for one country is the better in usage.
Dot Plot	Chart	Used for a distribution among categories.
Sorted Streamgraph	Diagram	Stacked area graph which is displaced around a central axis, resulting in a flowing. Fits for multiple time stamps of absolute values
Proportional Area Chart	Diagram	A Proportional Area Chart (Icon) is used for comparing proportions
Line Plot	Diagram	A line chart displays data along axis through points connected by straight line segments. Fits for ratio values in dynamics
Dumbell Chart	Diagram	Dot plots with two or more series of data. Fits for multiple- and two-time stamps
Gauge Chart	Chart	Very useful for indication of indeces or proportion with occupied space.
Progress Bar	Diagram	A graphical control element used to visualize the progression (or proportion)
Icon Count	Chart	Works good for visualizing the absolute number of categories entities
Waffle Chart	Diagram	Shows progress towards a target or a completion percentage
Highlighting the Qualitative Degree	Annotation	Used for specifying one country category
Alluvial Diagram	Diagram	Depicts the qualitative changes over time
Rating Position Gain Over Time	Annotation	Highlights the ranking change over time
Nominal \Qualitative Choropleth	Map	Same as choropleth, but implies the ordinal and nominal categories to color
Box Plot	Diagram	Depicts multiple entities through quartiles of their values
Rating Position at One Time Stamp	Annotation	Highlights the ranking at one stamp