# An Interdisciplinary Model for Graphical Representation

**Accepted to CIFMA workshop**

G. Antonio Pierro[1,2], Alexandre Bergel[3], Roberto Tonelli[2], and Stéphane Ducasse[1]

[1] Université de Lille, Inria, CNRS, Centrale Lille, UMR 9189 – CRIStAL, France
[2] Università degli Studi di Cagliari, Italy
[3] DCC Universidad de Chile, Chile

**Abstract.** The paper questions whether data-driven and problem-driven models are sufficient for a software to automatically represent a meaningful graphical representation of scientific findings. The paper presents descriptive and prescriptive case studies to understand the benefits and the shortcomings of existing models that aim to provide graphical representations of data-sets. First, the paper considers data-sets coming from the field of software metrics and shows that existing models can provide the expected outcomes for descriptive scientific studies. Second, the paper presents data-sets coming from the field of human mobility and sustainable development, and shows that a more comprehensive model is needed in the case of prescriptive scientific fields requiring interdisciplinary research. Finally, an interdisciplinary problem-driven model is proposed to guide the software users, and specifically scientists, to produce meaningful graphical representation of research findings. The proposal is indeed based not only on a data-driven and/or problem-driven model but also on the different knowledge domains and scientific aims of the experts, who can provide the information needed for a higher-order structure of the data, supporting the graphical representation output.

***Keywords—*** Data visualization, interdisciplinary model, data-driven model, problem-driven model.

## 1 Introduction

Graphical representations of data are fundamental for the understanding of scientific knowledge, as readers often rely on what the experts visually represent in their publications to understand the underlying data-set and interpret their potential scientific meaning [1]. Figures and diagrams not only show the relevant data that support key research findings, but also provide visual information on the interactions among different operations required in scientific reasoning [2, 3]. Being able to adequately and precisely visualize data is also a pillar on which decisions can be made, as proposed by different dashboards in the market.

Data visualization has various purposes, such as to make abstract thinking on data series or sets more concrete and (mentally) manipulable, to help readers identify and evaluate some features of the data, to let users see the possible underlying trends, patterns, processes, mechanisms, etc. of the phenomena considered and studied [4]. The way data are visualized can therefore have important epistemic implications for scientific knowledge, as data visualization is not an

“interpretation-free” practice, i.e. a neutral process of data presentation in terms of scientific understanding. There are indeed several ways to transform data into a visual format, each of them entailing different possibilities for data interpretation.

Nowadays data visualization plays a significant role in the large adoption of data-driven and machine learning approaches and techniques. In this frame, the definition of what a visualization is can be object of debate. A visualization could be defined as a reusable component, which is achieved through a dedicated software library. For instance, some software for data visualization are MATLAB and Mathematica. Despite the large amount of tools offered by these software, surprisingly, it is left to the practitioner to actually manipulate the data to achieve a ready-to-be-used graphical representation. Previous research proposed data-driven models that exploit existing software libraries or adopt a framework-agnostic approach (D. A. Keim, 2002 [5]) based on data types to be visualized.

The paper aims at designing a framework for a software, named Miró, which instead allows the users to produce meaningful graphical representation in an automatic way without the need to manually transform the data. First of all, we aim to verify the benefits and the shortcomings of existing data-driven and problem-driven models, by presenting some case studies. The case studies focus on the problem of visually representing specific data-sets collected in different scientific domains for different (descriptive vs. prescriptive) scientific aims. The case studies suggest that data-driven models can actually provide a visualization that fits the domain knowledge and scientific aims of the experts in the case of descriptive sciences, but present some limitations in the case of prescriptive sciences. Finally, the paper draws some conclusion from the case studies, presenting an alternative interdisciplinary perspective for data visualization. A comprehensive model for graphical representation is then presented, which integrates a data-driven approach with an approach that guides the experts on a specific domain field to achieve the intended visualization, based on their aims, knowledge and hypotheses. Miró adopts this interdisciplinary perspective and is based on a visualization engine developed in Pharo and named Roassal [6].

## 2 Data-driven and Problem-driven Models

In the field of data visualization computing, researchers proposed different approaches to a comprehensive data-model, i.e. a model able to provide a meaningful graphical representation of a data-set for some scientific aims. Some authors advocated graphical representation techniques or visualization frameworks [7] based on data-driven models. The data-driven model approach is based on the idea that a comprehensive data-model is based on a prior data classification that can guide the automatic creation of a meaningful graphical representation. In general, the data-driven model describes the data characteristics of the data-set, such as the size (the number of rows), the data type (string, number, boolean) and the dimension (the number of the variables to represent), to categorize the data. Keim [5] proposed a data-driven visualization model based on the data types to be visualized, the visualization technique and the technique of visual interaction with data, ranging from standard and projection to distortion and "link&brush".

Other authors, especially in the context of big data visualization, proposed graphical representation techniques based on a problem-driven model [8]. The problem-driven model provides the researchers with the possibility to perform specific tasks on specific variables of the data-set, such as visualizing a variable distribution, performing a linear regression between two variables to see an eventual relationship via a scatter plot, comparing their composition via a pie chart, etc.

On the one hand, adopting a problem-driven model does not necessarily mean abandoning data-driven models. The problem-driven model may be tightly linked to the data-driven model, because the data-driven model imposes constraints on the graphical representation of data which might conditioning how the problem can be solved. For instance, in the case of time series,

there are graphs that are less appropriate than others or that are simply wrong depending on the data classification: the time data-type is indeed a constraint given or inferred from the data-driven model. On the other hand, a graphical representation that is guided only by a data-driven model would not allow the users to further act on data to have their final intended graphical representation. In the software where a problem-driven model is also envisaged, the user can interfere with the final graphical representation of the data. The user can indeed act on and guide the graphical representation to be produced.

The main disadvantage of the problem-driven model is that it might be negatively influenced by the users' previous hypotheses or scientific aims. On the contrary, a data-driven model is neutral under this respect: of course it is based on a prior classification, but the users might not know it. Without the users' interference, the final graphical output of a data-driven model might indeed have the advantage of questioning the researchers' prior goals and solicit a belief revision. Especially when a graphical output is unexpected and not corresponding to previous scientific goals, it might bring about further research or action.

Both the models assume that the data-set contains the information useful to produce a meaningful graphic representation. This may not always be the case. Scientific studies based on data-sets make use of graphical representations to better interpret their results. Among these studies, it is possible to find descriptive as well as prescriptive studies. The former aim to describe phenomena as they are, observing, recording, classifying, and comparing them [9]. The latter aim to provide the conditions for how phenomena should be, thus supporting inferences for data interpretation and decision and/or action to perform on data. Of course, a scientific study could be both descriptive and prescriptive, also depending on the scientific goals. The development of new decision-aiding technology should be tailored for both [10], also in the case of graphical representation [11]. The paper is therefore driven by the question on how a model should be to provide a meaningful graphical representation of a data-set to support the inferences and/or the decision a researcher wants to draw, in both the case of descriptive and prescriptive scientific studies.

In the paper we propose a general distinction between a model for descriptive studies and a model for prescriptive studies. Within these two models, it is possible to specify sub-models, specific for scientific domain and particular data types involved in the study [12]. Both the models can be used whenever a study has both descriptive and prescriptive scientific aims, as it is often the case.

## 3 Research Questions and Hypotheses

The paper aims to discuss the strengths and limitations of existing models for data visualization, by considering and discussing some case studies coming from publications of different scientific domains and having different scientific aims.

The research addresses the following questions: Q1) Are data-driven models sufficient for a software to help the researchers to automatically create the intended visual form for a data-set? Q2) In the case the data-driven models are not sufficient, what could be the best way to overcome their limitations? Q3) Can the existing libraries or programs fit a data-driven model perspective and at the same time overcome their shortcomings?

To answer the research questions, we advanced the following hypotheses: H1) The data-driven models might support the creation of meaningful graphical representation only for some specific scientific aims, such as the researchers' aims to provide a descriptive data analysis. H2) For scientific aims going beyond descriptive analysis, the existing data-driven models might not be sufficient. The data-driven models might need to be integrated into a more comprehensive and interdisciplinary data-model to overcome their eventual limitations. H3) Existing software libraries are data-driven and might not be sufficient to help researchers to find the intended visual

form for prescriptive scientific aims. They might need further implementation to allow the users to perform different manipulation on data, such as transformation, accommodation and integration with complementary data, to achieve the intended graphical output.

## 4 Case Studies Evaluation

We analyzed data-sets which are representative of two different scientific approaches: 1) descriptive and 2) prescriptive studies. In particular we provide a detailed analysis of some case studies, coming from 1) the domain of software metrics, in the wider field of AI, and 2) the field of human mobility and sustainable development. The analysis can be extended to further case studies in different scientific domains.

### 4.1 Descriptive Case Studies

As to descriptive scientific studies, we considered first of all the case of a study on the performance evaluation of different frameworks in AI [13]. The case study proposes a set of meaningful visual representations of a benchmark data-set for the performance evaluation of different Deep Learning (DL) models and frameworks. The Authors calculated the accuracy and the throughput of five classification problems for the DL models and frameworks. The output data-set was made of a series of two categorical data (the name of the framework and the DL model) and two physical data.

We selected this study for three reasons: 1) The work aims to provide a significant graphical representation of the performance metrics of different frameworks; 2) The work also aims to extend the graphical representation to other frameworks, to be applied to other works and thus be generalized. 3) The study's data-set presents a number of variables and categories, which are not trivial to represent as a whole to obtain a meaningful graphical representation [14].

When analyzing the study case, we found that there is a data-driven model, specifically Keim's data-model, that provides us with a significant representation of the data-set, without any accommodation and/or transformation of the data and, more importantly, without any addition of further information by the user. Indeed, by applying Keim's data-model, the data-set is well within multi-dimensional category and so the meaningful graphical representation technique should be a "heat-map graph", where the colour is represented by the categorical data and the two physical data (accuracy and throughput) are represented in a 2D coordinate system. Therefore, as to what concerns RQ1, "Do data-driven models support the creation of meaningful graphical representation", the answer is positive. As the Keim's data-model is sufficient to have a proper graphical representation, we do not need to cope with RQ2 on how to improve it for this specific case study. As to what concern RQ3, the existing libraries for producing data visualizations alone cannot give that expected output, even though based on a data-driven model. However, throughout a data-driven model such as the Keim's model and some accommodation of the data, the existing libraries could provide the expected automatic visual representation, starting from the raw data-set.

Other descriptive case studies concern, for instance, static programming analysis and focus on the correlation between numerical variables, such as the number of lines of code, cohesion, coupling or cyclomatic complexity [15] and categorical variables, such as the name of the package included in the analyzed software. This type of studies' authors often choose to represent their data-sets via a bar graph where the bar length represents the numerical value and the categorical variable is represented by the different color of the bar or by a label. Also in these cases, the graphical output can thus be provided by a data-driven model such as Keim's model. The analysis

can be extended to other descriptive case studies in different disciplines (e.g. biology [16], and sociology [17]), where Keim's data-model is sufficient to provide the categorization for descriptive scientific aims.

### 4.2 Prescriptive Case Studies

In the case of prescriptive scientific studies we first considered an interdisciplinary study on human mobility [18]. The Authors collected the data using smartphones and smartwatches worn by several participants over 2 weeks. Through these devices, they collected three kinds of data: 1) motion sensor data, 2) physiological data, 3) environmental data. For the purposes of this case study, we are interested in the second data-set collecting information about electrocardiographic (ECG) data, such as heart beat and blood pressure. The data-set has the following characteristics: 1) data are multidimensional, as each row of the data set contains both spatial coordinates (longitude and latitude) and physiological data (heart rate, in beats per minute), provided by the optical heart rate sensor of the smartwatch; 2) the row data series consists of over 1 millions of data.

One of the purposes of the research paper was to use physiological data to infer the user's stress and emotion level to identify places within a University campus area that are perceived as dangerous by the majority of participants. We selected this research for the following reasons:

- The research covers different domains: mobile computing, sensing systems, human mobility profiling and cardiology.
- As in the previous case study, the data-set presents a number of variables and categories, which are not trivial to represent in an overall meaningful graphic representation.

If we apply the Keim's model to the data-set, the graphic representation output is a "heatmap chart", where the position is represented in a 2D-coordinate system and the heart rate beat is represented by color hue. This type of representation may not be enough meaningful for the aims of the study, when based only on the data-set collected by the devices. Indeed, the data-set is not per se sufficient to have a meaningful representation: the danger zones' classification needs other, additional data, such as the normal resting heart rate range and the dangerous heart rate range, to be properly represented.

Figure 1 shows the graphical representation produced considering the additional data, the normal and dangerous heart rate ranges. These additional data are used to represent the different zones on the map with colors having different opacity (color with opacity 1 for the dangerous zones and transparent color for the zones considered safe).

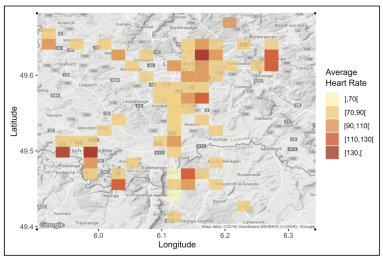


**Fig. 1.** Places that are perceived as dangerous by the majority of users through the use of colors with different shades.

Therefore, as to what concerns RQ1, the answer is that the data-driven model is not sufficient to give the intended graphic representation. Indeed the authors considered complementary data that are not merely added to the existing categories considered by the data-driven model, but rather organize in a higher-order structure and provide the cues to interpret the data-set to have a meaningful representation of the zones considered dangerous. The complementary data do shape the authors' interpretation of the data-set as they provide some intervals (the heartbeat rates intervals), as conditions to classify dangerous vs. safety zones. Indeed, the graphical representation 1 can be prescriptively used by experts in urban development for strategic planning to improve safety in public places.

As to RQ2, the solution to overcome the limitations of the data-driven model could be the possibility of inserting further data types into the data-set, relating the average heartbeat rates stored in the original data-set with the heartbeat rates intervals considered normal and dangerous. Furthermore, the data must be re-sampled taking into account the new knowledge, the normal resting heart rate range, coming from a different domain, the cardiology. However, this solution requires specific knowledge from the cardiology domain which may be different from the researchers' knowledge performing the data analysis.

Finally, regarding RQ3, data visualization libraries alone cannot help to obtain the expected output. Indeed, different tasks should be foreseen to achieve the intended outcome through a software, including the data visualization libraries:

- the program should make use of a data-driven model, such as the Keim's model.
- the program should give the user the possibility to add other data type. In the prescriptive case study, the data-type are intervals (conditioning the interpretation of the other data), also coming from a different scientific domain, i.e. cardiology.
- the program should give the researchers the possibility to further categorize the data-set via the additional knowledge. The program must provide the data-set with an higher-order structure to achieve the graphic representation meaningfully corresponding to the authors' scientific aims.
- Once adopting this workflow, the program might use the data visualization library to generate the intended graphic representation.

Another example of prescriptive studies concern the correlation between air pollution and respiratory illnesses [19]. The research findings come from data belonging to different domains such as 1) prescriptive data conditions in health information systems, 2) the air quality index (AQI) data provided by the World Health Organization (WHO), and 3) the descriptive data coming from particular air pollution electrical sensors. The descriptive data alone, in particular the concentration of microscopic particles with a diameter of 2.5 $\mu$m or less, are not sufficient to produce a graphical representation apt to meet the prescriptive aims of the study (see Figure 2), i.e the sustainable development program in urban and rural areas affected by air pollution.

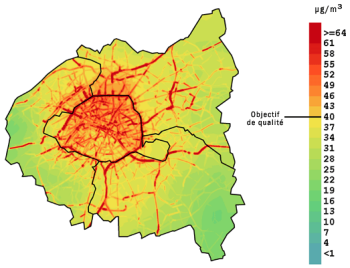


**Fig. 2.** Areas affected by air pollution.

## 5 An Interdisciplinary Model

In the field of graphical representation, interdisciplinary models have been proposed to cope with the limitations of both previous data-driven and problem-driven models. For instance, Hall et al. [20] proposed a trans-disciplinary model which allow the experts in a particular domain to be supported by visualization experts. Their work is very interesting as the interaction between experts with skills in different domains could greatly influence the production of meaningful graphical representations to display cues for scientific findings.

However, the prescriptive case study examined in this paper cannot be solved through this trans-disciplinary approach. Of course a competence in visualization is welcome, but cannot per se highlight the conditions of meaningfulness, which come from another scientific domain in the prescriptive case studies. Therefore an interdisciplinary model is needed which integrates knowledge and practice coming from different scientific domains in the process of visualization. Figure 3 proposes the main elements of the interdisciplinary model.

- The source domain/s is/are the domain/s from which the data are collected.
- The complementary domain/s is/are the domain/s from where to collect the data required to interpret the source domain/s data.
- The blended domain [21] is given by the intersection between the source domain/s and the complementary domain/s, where some new insight could emerge.
- The data model is the model driving the software in the process of data categorization and visualization.

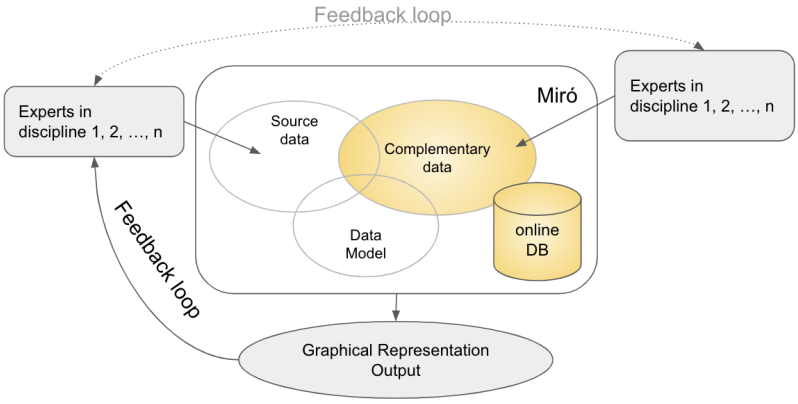


**Fig. 3.** Interdisciplinary Model

As a solution to the prescriptive case studies examined, we propose an interdisciplinary problem-driven approach for the visualization of data coming from different domains. For the aims of descriptive studies, the source domain and the data-driven model are usually sufficient to have meaningful graphical representations. The prescriptive case studies instead show the limits of both data-driven and problem-driven model, as there are scientific aims for which it is not sufficient having both the data models and the data coming from a scientific domain to obtain meaningful graphic reports for the research findings.

In prescriptive studies, two further processes - not envisaged in previous data-driven and problem-driven models - are needed to have meaningful graphical representations of the source data:

- A selection process: when the data collected by the researchers in the source domain are not sufficient, other specific data selected from a different scientific domains might be needed to interpret the source data. These data might indeed be the condition of meaningfulness for data interpretation, and thus for the visual output of the software.
- A transformation process: specific tasks might be needed for the re-interpretation of the data in light of the selected complementary data and the scientific aims of the study. For instance, the source data might need to be re-sampled considering the complementary knowledge.

The scientist's insight needs, therefore, to be entered as complementary data in any software's visual framework, which in turn should make it possible to enter them, interacting with the scientist. In the prescriptive case studies, the interdisciplinary approach is driven by the interaction among experts in different domains (mobile computing and cardiology) and guides the

production of graphical representations, meaningfully representing the areas perceived as dangerous (see Figure 1). The insertion of the relevant complementary data might come not only from experts of another domain, but also from online interactions among experts in different domains and/or online web-based crowd-sourcing selected by the expert users themselves.

This interdisciplinary model might then overcome the limitations of both the data-driven and the problem-drive models, especially when it automatically proposes the complementary data based on the scientific aims of the expert and the relative missing expertise, which could come from an expert in another domain. This approach is the framework for Miró, a software intended to be a guide to build meaningful graphical representations for both descriptive and prescriptive studies, based on a data-set coming from the source domain/s and on a data model eventually able to provide complementary online data. Differently from softwares based on previous data-driven and/or problem-driven models, the Miró's interdisciplinary model allows the user to insert data or select data coming from complementary domain/s, and transform the source data-set to have the intended graphical representation.

In the case study requiring data from both human mobility and cardiology, when the participants to the experiment are considered as a group, their information provides other meaningful cues to identify critical geographical or temporal points. For example, the two figures coming from the prescriptive case studies represent respectively 1) the places that are implicitly perceived as dangerous or risky by most users and 2) the most polluted areas of a city. By analyzing the data-sets and their graphical representations, it emerges that there are data (fields) that make sense only within one or more interval/s [a, b]. Often, the interval information is neither provided within the data-set nor within the single scientific discipline and thus the interval must be set by the scientist and/or by another expert. This needs to be contemplated by the dashboard developer. For instance, in study 1), the heart-rate belongs to the health domain and make the place dangerousness meaningful only if the average value is above a certain threshold. The threshold needs to be provided by a scientist (also following the scientific practices of his/her scientific field), it is not provided by the data-set per sé, especially in interdisciplinary prescriptive studies like 1).

Some data actually come from the data-sets, some other data come from the scientist's interpretation of the data in light of the scientific hypotheses in her/his study. The latter should be provided by the scientist and a dashboard should make it possible to enter them. Prescriptive scientific studies are more likely to need interval information as a condition of meaningfulness to make sense of the data-sets when compared to descriptive scientific studies, which can instead provide meaningful graphical representations based on traditional models. Of course, scientific studies can be both descriptive and prescriptive: Miró can provide a meaningful graphical representation also for these studies as it does not abandon traditional models, but it instead proposes further functionalities.

## 6 Conclusion and Future Works

The paper shows how important might be an interdisciplinary data model, especially in prescriptive studies, to have a software able to provide meaningful graphical representations of data. In the case of descriptive studies, existing models - data-driven models and/or problem-driven models - might be sufficient to produce meaningful graphical representations when providing the data coming from the source domain/s. In the case of prescriptive studies, the existing models might fail to produce meaningful graphical representations when just the collected data coming from the source domain/s are provided. The paper proposed an interdisciplinary approach to overcome the limitations of the existing models via a software-expert interaction. In this framework, the software allows the users to reinterpret and transform the collected source data in the light of the scientific knowledge coming from (online) interaction with other experts or data-sets coming

from complementary domain/s. The graphical representation is made meaningful in the blended domain, thus providing a visual support for new findings.

## References


1. A. Mahling et al. Beyond visualization: Knowing and understanding. In *Lecture Notes in Computer Science*, pages 16–26. Springer Berlin Heidelberg, 1990.
2. William Bechtel and Adele Abrahamsen. Explanation: a mechanist alternative. *Studies in history and philosophy of biological and biomedical sciences*, 36(2):421—441, June 2005.
3. P. Eklund and O. Haemmerlé, editors. *Conceptual Structures: Knowledge Visualization and Reasoning*. Springer Berlin Heidelberg, 2008.
4. Jeff Zacks and Barbara Tversky. Bars and lines: A study of graphic communication. *Memory and Cognition*, 27(6):1073–1079, 1999.
5. D. A. Keim. Information visualization and visual data mining. *IEEE Trans. Vis. Comput. Graph*, 8(1):1–8, 2002.
6. A. Bergel et al. A domain-specific language for visualizing software dependencies as a graph. In *2014 Second IEEE Working Conference on Software Visualization*, pages 45–49, 2014.
7. J. Zhu et al. A data-driven approach to interactive visualization of power systems. *IEEE Transactions on Power Systems*, 26(4):2539–2546, 2011.
8. G. E. Marai. Activity-centered domain characterization for problem-driven scientific visualization. *IEEE Trans. Vis. Comput. Graph*, 24(1):913–922, 2018.
9. D. A. Grimaldi and M. S. Engel. Why Descriptive Science Still Matters. *BioScience*, 57(8):646–647, 09 2007.
10. R.V. Brown and A. Vári. Towards a research agenda for prescriptive decision science: The normative tempered by the descriptive. *Acta Psychologica*, 1-3:33–48, 1992.
11. I. Kim et al. Visualization of neutral model of ship pipe system using x3d. In *Lecture Notes in Computer Science*, pages 218–228. Springer Berlin Heidelberg, 2010.
12. A. Kerren et al. *Information Visualization: Human-Centered Issues and Perspectives*. Lecture notes in computer science. Springer, 2008.
13. D. Velasco-Montero et al. Optimum selection of dnn model and framework for edge inference. *IEEE Access*, 6:51680–51692, 2018.
14. P. Godfrey et al. Interactive visualization of large data sets. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2142–2157, 2016.
15. S. R. Chidamber and C. F. Kemerer. A metrics suite for object oriented design. *IEEE Transactions on Software Engineering*, 20(6):476–493, 1994.
16. O. Roux and J. Bourdon, editors. *Computational Methods in Systems Biology*. Springer International Publishing, 2015.
17. John Salerno, Shanchieh Jay Yang, Dana Nau, and Sun-Ki Chai, editors. *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer Berlin Heidelberg, 2011.
18. S. Faye et al. Characterizing user mobility using mobile sensing systems. *International Journal of Distributed Sensor Networks*, 13(8):155014771772631, 2017.
19. A. Forkan et al. Aqvision: A tool for air quality data visualisation and pollution-free route tracking for smart city. In *2019 23rd InfoVis*, pages 47–51, 2019.
20. K. W. Hall et al. Design by immersion: A transdisciplinary approach to problem-driven visualizations. *IEEE Trans. Vis. Comput. Graph*, 26(1):109–118, 2020.
21. M. Turner and G. Fauconnier. A mechanism of creativity. *Poetics Today*, 20, 09 1999.