Insight Beyond Numbers: The Impact of Qualitative Factors on Visual Data Analysis

Benjamin Karer, Hans Hagen, and Dirk J. Lehmann

Abstract—As of today, data analysis focuses primarily on the findings to be made inside the data and concentrates less on how those findings relate to the domain of investigation. Contemporary visualization as a field of research shows a strong tendency to adopt this data-centrism. Despite their decisive influence on the analysis result, qualitative aspects of the analysis process such as the structure, soundness, and complexity of the applied reasoning strategy are rarely discussed explicitly. We argue that if the purpose of visualization is the provision of domain insight rather than the depiction of data analysis results, a holistic perspective requires a qualitative component to to be added to the discussion of quantitative and human factors. To support this point, we demonstrate how considerations of qualitative factors in visual analysis can be applied to obtain explanations and possible solutions for a number of practical limitations inherent to the data-centric perspective on analysis. Based on this discussion of what we call qualitative visual analysis, we develop an inside-outside principle of nested levels of context that can serve as a conceptual basis for the development of visualization systems that optimally support the emergence of insight during analysis.

Index Terms—Visualization, Reasoning, Qualitative Aspects

1 Introduction

In the social sciences, a number of open questions are vividly discussed, such as whether raw and unbiased data actually exists and whether neutral and objective analysis is actually possible and how it has to be defined (e.g. [19]). Asking this kind of questions is important as the answer has a considerable impact on how a field understands itself. Yet, there typically is no absolute answer. Instead, there is a plethora of positions, each with its own implications on how research within the scope of this position has to be conducted. For example, there is an interesting collection of literature discussing issues with purely data-centric analysis from a socio-technological perspective (e.g. [5]). For visualization as a discipline, such an existential question might be to define what actually constitutes insight. Although it is a common argument that the mission of visualization is to support its users in obtaining insight, there is no commonly accepted definition about what insight actually is [9] and the sensemaking processes applied to obtain insight are still not well understood [42]. As visualization as a field has reached the maturity to ask this kind of self-defining questions, it also has advanced enough to critically reflect and question its approaches to accomplish its mission.

In this paper, we critically reflect visualization's current perspective on insight. Our reflection is based on the premise that all data analysis is always directed towards obtaining insights into a domain – either to achieve a structural understanding of states of affairs observable within the domain or to enable an informed decision within the context of this domain. The emphasis here is clearly on insight into <u>the domain</u>. Yet, we find that the notion of insight that is commonly applied in visualization is <u>not</u> referring to insight into the domain but rather restricts itself to insights into how to read and interpret the visualization correctly (e.g. by legends, explanations, guidance etc.) or to insights about properties of the data (e.g. reading off visualized results of quantitative analysis).

Today's data analysis is usually conducted in a context-free manner. Hence, although the computationally obtained results are valid, their interpretation is up to a human analyst. Visualization supports this analysis by providing graphical representations of the data and of analysis results. Yet, this presentation typically does not involve additional domain information that is not encoded in the data or the computed results. Regarding the search for insight into the domain, we identify two major issues with this approach: First, without any context information, there is no way to be certain that the analysis conducted is actually relevant to the problem or that a computationally valid and sound solution actually captures a relevant feature of the data. Second, as long as there is no reflection of the obtained knowledge about the data back into the domain, one can never know whether this knowledge is applicable, relevant, or even valid in the whole domain.

For better intuition, let us provide an example:
Consider an analysis of the gothic-style architecture of medieval cathedrals in western Europe based on a collection of photographs or construction plans. To allow for numerous large openings in the walls that could be filled with glass windows, architects constructed a complex network of arches in order to concentrate the weight of the ceilings and walls onto a small number of support elements. As a result, a typical pattern in gothic-style medieval walls is the spandrel – a triangular area of which one side is part of an arch. Identifying the spandrel to be a pattern in the architecture data therefore is a sound and perfectly valid result of data analysis. However, spandrels are not defining or even characteristic elements of gothic-style architechture but only a side-effect of the usage of arches as support structures to allow open spaces for more windows. Hence, gothic architecture is not about spandrels, it is about arches. Yet, this conclusion is not evident from the data alone – it can only be reached by considering the patterns found in the data within the context of the domain of architecture and construction.

Therefore, to offer the analyst the holistic perspective necessary to reflect insights into the data with respect to the domain and context, visualization needs to go beyond computing and displaying numerical results. In particular, visualization has to

1. offer means to qualitatively evaluate findings resulting from the analysis process in their appropriate context
2. offer ways to map obtained insights into the data back to the domain in order to determine their implications for the domain

Only findings that can be validated against the domain – either empirically or by formal proof – can become insights into the domain.

This paper is divided into two parts. Each of them discusses one of the above postulates in more detail. In the first part, we review related work in order to understand what aspects contribute to an insight to be found in visualization. This review motivates what we call qualitative visual analysis – an explicit consideration of context information in the visualization and in the analysis process for the qualitative evaluation of analysis findings. Towards a better understanding how such

• Benjamin Karer is with the Federal Criminal Police Office of Germany. E-mail: benjamin.karer@bka.bund.de.
• Hans Hagen is with TU Kaiserslautern. E-mail: hagen@cs.uni-kl.de.
• Dirk J. Lehmann is with the University of Magdeburg and with IAV GmbH. E-mail: dirk@isg.uni-magdeburg.de.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

considerations can support the analysis process, we discuss a number of limitations of purely data-centric analysis and how they can be alleviated by additional context information. In the second part, we deepen the discussion on how reasoning crosses the boundaries between different levels of context. To this end, we develop an inside-outside-principle of qualitative visual analysis that allows us to propose possible approaches for the integration of context information into a visualization.

2 Related Work

Discussing the value of visualization for practitioners, Fekete et al. emphasize visualization's capability to support identifying models as a starting point for analysis [17]. They argue that automatic analysis and (exploratory) visual analysis answer different analysis questions and thus should attempt to support each other. The combination of visual and automatic data analysis is also the central idea of visual analytics [31]. Sacha et al. propose a high-level model of knowledge generation with visual analytics formalizing the obtainment of knowledge as a cyclic process of repeated automatic data processing and human reasoning [46]. Andrienko et al. interpret visual analytics as a goal-driven process whose result is the construction of a mental model of the data [3]. More detailed discussion in this direction requires mappings between the graphical display, a person's understanding of this display, the available models for reasoning, and the phenomenon represented by the visualization. Vickers et al. apply category theory and semiotics to propose a theoretical framework formalizing this set of connections [61].

We argue that models like these can serve as guidance for the discussion of analysis and reasoning processes for visualization applications in order to establish a certain degree of standardization and comparability across different papers. In order to apply those models correctly, it is important to understand the basic concept they all share: the idea that the analysis process leads to the obtainment of insight. In the following discussion, we therefore review how insights are seen within visualization as a field, what kinds of insight can be obtained during the process of visual data analysis[FR], and what aspects contribute to the obtainment of insight.

A necessary prerequisite for applying such models is to specify what exact purpose a given visualization is made for. As Card, Mackinlay, and Shneiderman phrase it, "the purpose of visualization is insight, not pictures" [6] (p. 6). Although insight might be the most common answer to the question what visualization is for, it is typically not very specific as it is generally hard to specify what this insight would actually be. Chen and Edwards identified a collection of schools of thought to help structuring the different opinions on what is the purpose of visualization [10]. According to them, other directions than the insight-oriented approach tend to answer this question in terms of supporting thought and general cognitive processes (e.g. [6]), communicating information (e.g. [14, 23]), or optimizing the visualization, e.g., with regard to the amount of data to be analyzed or the efficiency of the reasoning with the visualization in terms of the time saved [11, 12]. According to Chen and Edwards, most authors, however, follow a more pragmatic approach and motivate the purpose of a visualization by its utility for the specific application. We take an alternative viewpoint, identifying the purpose of a visualization not with its specific task within an analysis or other kind of setup but rather with the aims associated with the setup itself and how well the visualization fits this setup by supporting these aims. As the further discussion indicates, this "quality of fit"[FR] can be assessed by characterizing the different kinds of insight a visualization application allows its users to obtain.[S7]

2.1 Insight

Unfortunately, there is no clear and commonly agreed upon definition what actually constitutes an insight in visualization. Chang et al. propose two different characterizations [9]. The first is the notion of a sudden "Aha"-effect[FR], which is one of several approaches to treat insight in cognitive psychology. The second one is a gradual increase of knowledge and appears to be the one more commonly used implicitly and informally within the visualization community. Saraiya et al. characterize insight as individual findings in the displayed information [49]. Other authors extend this definition by the requirement to involve reasoning. In their knowledge generation framework for visual analytics, Sacha et al. specify insight as the generation of additional knowledge by interpreting a finding in combination with domain knowledge [46]. Yi et al. conduct a literature review on the processes how people obtain insight. Their findings can be summarized as getting an overview, then adjusting the view to identify patterns and matching these patterns with existing knowledge in order to draw conclusions [65]. North extends these notions of insight further, characterizing insight as being complex, deep, relevant, unexpected, and qualitative [40]. Ishack et al. study the effect of users' domain knowledge on the insights reported in visual data analysis [29]. Relating their findings to the notion of distributed cognition, they conclude that insights are neither generated in the user's head nor are they hidden in the data but instead are constructed during the user's interaction with the data visualization. Although they differ in the details, all the discussed definitions agree in the following:[S8]

Definition: (Visualization) Insight

An insight marks a step forward in the interpretation and analysis in the form of a change of the user's knowledge or understanding.

Thereby, an insight necessarily involves the individual viewer. Rather than being an object or entity to be mined from the data, an insight is a step of progress of certain significance, an emergent property of the visual data interpretation and analysis process. It can therefore be understood as a function on the viewer's knowledge and understanding. While its execution is clearly a cognitive action of the viewer, it is not as clear how this function is generated. In visualization, there are currently two major directions of thought in this regard which Chen and Edwards refer to as weak and strong insightism [10]. Weak insightism requires the viewer to actively incorporate additional knowledge into the interpretation of the visualized information in order to *obtain* insight. Strong insightism suggests that insight can be directly *provided* by the visualization.[S7] Towards a better understanding of how insight emerges in visual data analysis, we need a more detailed discussion about the kinds of insight that can be obtained.

2.2 Insight and Reasoning

In their 1996 article on External Cognition, Scaife and Rogers discuss reasoning with data visualization regarding the analysis of data with respect to a mental model [50]. Cleveland and McGill define graphical perception as the act of decoding the quantitative and qualitative information encoded in visual data representations [13]. Quantitative information can be obtained from within the data context – they constitute of numerical or otherwise deterministic values and properties that can be derived from this information. Qualitative observations also involve additional levels of context.

User studies reveal that global tasks involving inference from a number of observations are harder to perform then local tasks where information can be read directly from the depiction, especially if the tasks involve knowledge from outside the data context [22, 44]. Casner applies a similar argument to motivate the transformation of cognitive tasks into perceptive tasks that provably yield the same results but are substantially easier to perform [7, 8]. Trafton et al. describe three different kinds of insights to be obtained during the process of analyzing visualized data[FR], requiring increasingly complex reasoning [58, 59]. Because their observations are based on findings from several empirical studies, we adopt this distinction and identify the types of insight to be obtained during the process of visual data analysis[FR] as follows:[S8]

Definition: Insight into the Visualization

Insights that determine how the viewer interprets the visualization.

This is the kind of insight involved with understanding the displayed information either by the viewer's own visual literacy or by mapping information provided by legends or manuals to the display.[S8]

Definition: Insight into the Data

Insights affecting the viewer's knowledge about statistical and other *structural information about the data.*

Insight into the data reveals information such as relationships, patterns, or trends. This is the kind of information that can be obtained from purely quantitative data-centric analysis without further interpretation of the obtained results. The analysis yielding this kind of insight can thus in principle be done completely automatically. Once viewers understand the structural aspects of the data, they can attempt to draw more complex conclusions from the observations.[S8]

Definition: Insight into the Domain
Insights affecting the viewer's understanding of the domain.

These insights are obtained by reflecting and interpreting the findings observed in the visualization with respect to the viewer's knowledge about the domain under investigation. As a consequence, this type of insight directly affects the viewer's understanding of the domain.[S8]

In the above formulation, the three categories of insight match the three-level model proposed by Ware [63]. Regarding the mapping of high-level tasks to these categories, Nazemi and Kohlhammer associate the first level with search tasks, the second with exploration, and the third with the analysis process [39]. Where the first and second category are concerned, Smuc et al. report results from a user study revealing that while insight about the data can only be obtained after insight about the visualization, viewers only need to understand those parts of the visualization that are relevant for the insights they intend to obtain [54]. This observation emphasizes Kosslyn's remark on the importance of pragmatics in visualization: visualization performance depends on how well it supports its purpose [35].

Remarkably, all of these models have in common that reasoning about the data with respect to domain information is regarded as the most complex part of the analysis process. The understanding of insight in visualization as discussed above commonly refers to insight about the data (e.g. North [40]) or into the domain (e.g. Sacha et al. [46]). Yet, the discussions of what kind of insight can be found in specific visualization applications rarely apply those theoretical frameworks and tend to stick with the presentation on what kinds of insight into the visualization are required to obtain certain insights into the data. Of course, it is hard to assess how users map the insights about the data to their domain knowledge. Still, we have to remark that if it cannot be explained exactly how a given visualization is meant provide to insight into the domain, this is probably because it simply does not do so. Of course, the visualization may still support its user in obtaining those insights by deriving them from the insights into the data offered by the visualization. We only suggest to be careful when discussing what kinds of insight a visualization actually provides or only supports in order to prevent promising too much. Towards a better understanding of how insight into the visualization and into the data can be related to the domain, we now turn to a discussion of levels of context involved with obtaining insights into the domain.

2.3 Context

The interpretation of data depends on the context. Data without context may well also be free of meaning. To support reasoning, a visualization should thus also reflect available context information [5]. In his Views on Visualization, Van Wijk proposes different high-level models for the visualization process [60]. One of his key findings is that the visualization itself – not only the interpretation by its viewer – is subjective: It does not only depend on the data but also on the algorithms and data structure used as well as other factors determined by the programmer. Remarkably, this finding generalizes to automatic analysis. We therefore define the information that can be learned from data processing as the data context:[R2]

Definition: Data Context
The data context contains all information that can be obtained from the raw or processed data by read-off without further interpretation.

The conclusions drawn from analytical reasoning do not only depend on the data context. For example, the ability to understand the information encoded into the graphical display correctly depends on the viewers' individual skill and experience. These factors contribute to an individual user context.[R2] Pohl and Doppler conducted an extensive literature review of different theoretical approaches to the explanation of sensemaking processes and an empirical user study investigating whether these techniques can be applied to visualization [42]. Among other interesting findings, they observe that the reasoning strategies applied to draw conclusions from the displayed information are highly dependent on user-specific aspects such as individual skill, experience, and background knowledge. Similar findings are reported by other studies. Ishack et al. recognized that viewers with more domain knowledge tend to provide more sophisticated reports applying their domain knowledge to evaluate their findings in the data [29]. Sheidin et al. report that character traits influence performance in working with time series visualization [52]. These results match with earlier findings by Petre and Green who recognized that obtaining information from visualizations also depends on the users' experience [41]. Other empirical evidence shows that not only the users' experience but also their general background knowledge can benefit the analysis process, especially knowledge about the patterns to search for in a specific visualization [32]. Regarding the design and creation of visualizations, it has been found that integrating users' prior knowledge into a visualization and thus allowing them to directly reflect their prior knowledge with respect to the data benefits recall and comprehension, even if the users only possess little prior knowledge on the data set [33].[S7]

Definition: User Context
The user context determines the influence of the user's background on the interpretation of information in the data context. This includes but is not limited to factors such as visualization literary, domain knowledge, experience, and personal reasoning strategies.

While the ability to obtain information from the data or its visualization depends mainly on the data and user context, the question what the user will attempt to conclude from the observations depends heavily on the analytical task to be solved and the domain of interest in which the analysis is conducted.[R2] This is also a result of a study conducted by Streeb et al. who investigated the arguments of a range of visualization practitioners from different disciplines on why they did (or did not) apply visualization [56]. They found that most arguments were influenced not only by the data and the user, but also by the task.[S7]

Being directly concerned with the solution of specific low-level or high-level analysis questions, task-based design primarily captures the analysis context. In this regard, a crowdsourcing study reports that the same visualization performs differently in the contexts of different tasks [48]. From the opposite perspective, this is reflected by different interaction patterns users of visual analytics applications show while performing different tasks [21] and an observed large impact of domain expertise on interaction patterns and report detail [29, 41]. These findings suggest to dynamically adapt the visualization to the analysis and the user context. In this direction, Golemati et al. proposed a context-adaptive visualization environment that extends the typical focus of automatic visualization on the domain and data by an explicit consideration of user profiles and preferences [20]. In the intelligence sector, user models for adaptive visualization have been reported to support distinguishing relevant from non-relevant information which in turn allows optimizing relevance-based visualizations [1].

Definition: Analysis Context
The analysis context determines the influence of the analysis aim and task on the interpretation of information. In a multi-stage analysis process or in a process involving different tasks to perform, this context may change frequently.

The interpretation of information being found in the data or in a visualization depends heavily on the domain under investigation.[R2] It is therefore common good practice to tailor visualizations to the domain [38, 45, 51]. As a result, there is a plethora of techniques trying to incorporate the domain's perspective into the design. Tory and Möller point out expert interviews to be useful especially in the early phase of design but also remark that they should be complemented by user studies evaluating the resulting application against the design goals [57]. Understanding of the domain context can be obtained for example from involving users in participatory design, applying structured interviews [51, 53] or by field studies [15, 24, 26]. Domain-analysis is a specific technique useful when working in environments with existing software solutions [16]. Task-oriented design aims at the characterization of analysis tasks and their decomposition into interaction workflows [2, 64].

Definition: Domain Context

The domain context specifies the domain's influence on the interpretation of information in the data context.

A holistic perspective on data analysis benefits from synergies between data-centric and user- or task-centric considerations [43]. We feel that, despite the significant added value of such a perspective, the qualitative aspects of data analysis are underrepresented in the discussions of visualization systems. Frameworks attempting to explain the effectiveness of visualization and visual analytics do exist but are rarely explicitly applied to explain a visualization application or even to motivate design choices.

3 Qualitative Visual Analysis

Automatic analysis approaches relying on data processing can necessarily only yield results that are intrinsic to the data. In visual data analysis, results can instead also be inferred by interpreting the data considering additional knowledge from outside the data context. This interpretation enriches the available quantitative information by a context-specific valuation and thereby translates it into a *qualitative* representation of the information that is applied for reasoning about the data. It is the interpretation that makes the viewer consider a value high or low, a point cloud dense or sparse, or a developing trend positive or negative. Interpreting the displayed information with respect to knowledge from outside the data context in order to translate the available quantitative information into a qualitative representation[S8] therefore is an essential part of data analysis and is commonly performed by an analyst reasoning about the displayed information. However, it can also be predetermined by the assignment of fixed interpretations to certain configurations and values of parameters and variables.

In the above discussion, we have seen that a collection of different levels of context contribute to the interpretation of the information depicted in visualizations and that there are different types of insight to be obtained. Relating the types of insight with the levels of context they involve, we can explain the difference in complexity of obtaining insights about the visualization, the data, and the domain as the result of the involvement of additional levels of context. Following this approach in the opposite direction describes what kinds of knowledge from outside the data needs to be incorporated into the analysis process in order to obtain insights of a specific type.

The kind of reasoning we describe here results in a descriptive model of the observations reflecting the behavior of the data. It is based on principles, rules, and predictions rather than only on observable facts. The insights obtained in such a model define an adequate and accurate description of the analyzed data and its interpretation with respect to the inherent data context, the analysis context determined by the task, the analyst's individual expertise, and the general domain background. It is this kind of reflective and interpretative reasoning that forms the transition of insights into the data into the conclusions that constitute insight into the domain. To better distinguish this kind of reasoning from those parts of the analysis process that yield insights into the visualization and into the data, we propose the following definition:

Definition: Qualitative Visual Analysis

Qualitative visual analysis answers the questions for observable principles independent of quantitative factors as well as for the interpretation of observations with respect to the relevant context.

Qualitative visual analysis is therefore primarily concerned with the process and the results of reasoning about what is depicted and perceived within a visualization. Compared to today's dominant focus on insight into the data, it constitutes an additional perspective, broadening the scope of the discussion by considerations how exactly a visualization is meant to support insights into the domain and how those insights are to be obtained by a viewer. Current literature on visual data analysis tends to discuss these aspects only implicitly – if they are covered at all. This is especially true for those applications that consider visualization only as a tool for the communication of the results of computational analysis rather than as a means to support the actual analysis. Our impression is that this view is especially prominent in the wide spread applied areas of Data Science, Business Intelligence, and Industrial Analytics. Visualization as a field of research instead is aiming to enable obtaining insights into the domain. We argue that, being a central aspect of visual data analysis, the path to domain insight should be discussed much more explicitly in visualization literature.

Especially since the questions asked by qualitative visual analysis are quite complex, such a discussion should be grounded in a solid mathematical foundation. In order to frame the scope of such a discussion, we propose an orientation along four principles that we name *comprehension*, *validity*, *precision*, and *interpretation*:

Comprehension

A discussion of qualitative visual analysis should point out exactly what observations there are, how they are related to the data, and how they are interpreted and evaluated. Being primarily concerned with inferred rather than computed or measured properties of data does not automatically mean to introduce ambiguity or arbitrariness. Of course, there remains a certain degree of subjectivity in the process of interpretative analysis, no matter whether it is introduced by the designer of an automatic analysis procedure or by the analyst evaluating and interpreting the data's representation. Yet, despite this subjectivity, the individual analytical reasoning relies on logical considerations and is contextually framed by the data, the user's background, the analysis question, and the domain. Therefore, the provenance of insights being obtained during[FR] qualitative visual analysis can and should be assessed in terms of a comprehensible inference structure. Other than explaining how insights are obtained, the comprehensibility provided by such an inference structure will also make it much easier for other analysts to reproduce or falsify analysis results.

Validity

Drawing conclusions requires the application of logic. Hence, the reasoning strategy being applied can be described exactly by a proper logical calculus or similar kind of inference system. Therefore, it is possible to precisely determine whether the conclusions drawn are correct with respect to this strategy. Qualitative visual analysis should therefore critically reflect the provenance of insights being obtained and analyze their correctness within the applied reasoning structures as well as the soundness of the reasoning structures themselves.

Precision

There is a common misconception regarding qualitative descriptions as imprecise or otherwise not exact. As a consequence, imprecise statements of a form like "*A* behaves somewhat like..." or "there seems to be some regularity in *B*" are often excused as being "qualitative descriptions"[FR]. However, they are not – at least not in the sense of qualitative analysis. The idea of qualitative visual analysis is not to prefer vagueness over exactness. What can be measured, should be measured – with the condition that the measure applied must be semantically meaningful with clear implications for the analysis result. Statements in qualitative visual analysis should be formalized by predicates or functions in formal logic. As a rule of thumb, a valid observation in qualitative visual analysis is completely describable in terms of a clear provenance structure based on valid conclusions made in a logically sound inference structure. Consequently, valid observations in qualitative visual analysis are exact and precise logical statements.

Interpretation

Qualitative visual analysis is not only concerned with reasoning about data but also with interpreting the conclusions drawn with respect to the analysis question and the domain context. To this end, it explicitly allows a valuation of observations and conclusions. Such a valuation can, for example, be established in terms of suitable logical predicates. It is, however, essential to properly ground the valuation in the contexts of the user's knowledge and the analysis task. The user's background primarily determines how the displayed information is being interpreted in terms of visual literacy and visualization understanding, whereas the analysis context drives the decision on what is considered relevant. Insights into the domain cannot be obtained without interpreting the observations with respect to their meaning for the domain. In any case, every interpretation is a result of reflecting observations and conclusions in a proper context. Discussions of qualitative visual analysis therefore should investigate these interpretations in detail and reconstruct which aspects of which level of context contribute to a certain interpretation. As a result, interpretations become explainable and therefore can be presented with the same formal rigor applied to conclusions.

Although the above considerations strive quantitative aspects like precision or uncertainty and human factors such as perception and comprehension, these factors are not of central concern for qualitative visual analysis. Rather than on what can be seen and perceived, the focus of qualitative visual analysis is more on how this information is being interpreted and what conclusions can be drawn from it. Hence, we need to consider an additional perspective, complementing the discussion of quantitative and human factors:

Definition: Qualitative Factors in Visualization

The investigation of qualitative factors in visualization is concerned with the reasoning processes and mechanisms governing the emergence of insight from visual data analysis independent of quantitative or human factors.

As an example for the importance of such a discussion, consider a flawed analysis process. Systematic errors in the analysis process or applied reasoning models have a strong impact on the analysis results. In the extreme case, such flaws render the analysis result useless. For this reason, we argue that the consideration of qualitative factors in visualization should dedicate considerate effort to assessing the logical soundness of analysis workflows and the applied reasoning strategies. Other directions for the discussion of qualitative factors include considerations on reasoning complexity, or possible interpretations of displayed content. In the following, we demonstrate how considerations on qualitative factors in visualization can help to address a number of problems that are inherent to today's predominant data-centric approach to visualization and visual data analysis.

4 Applying Qualitative Visual Analysis to Alleviate Limitations of Purely Quantitative Analysis

In this section, we review a collection of limitations inherent to today's predominant quantitative and data-centric approach to analysis and present directions to alleviate these issues inspired by qualitative visual analysis. Although well known, the issues we discuss here are rarely discussed in conjunction or with respect to their general implications. Discussions on these issues typically focus on specific problems they pose for visualization's several sub-disciplines. We collect the several limitations in a concise summary and also point out possible solutions based on qualitative visual analysis. To this end, we discuss each limitation following a fixed structure: First, we describe the problem. Where necessary, we provide an example for better intuition. We then identify reasons for the respective problem related to the four levels of context identified above. In the final step, we leverage this analysis to provide a possible solution. We discuss one of these solutions with respect to the four principles of qualitative visual analysis to demonstrate how the concept can be applied determine the expected utility of design approaches.[S1]

4.1 Missing Essential Domain Characteristics

Qualitative aspects and properties often can only partially be reflected by quantitative measures. Sometimes, this is even impossible. Although there might be indicators allowing an indirect description, the actual object of discourse cannot be reflected completely by quantitative data. Especially if those aspects determine the domain's essential characteristics, purely data-centric analysis might reveal insight about the data but not about the domain. In practical applications working with real-world data, the interpretation of data is often being introduced into the analysis process from outside the data context. In this case, the meaning of observations cannot be concluded from the data.

As an example, consider the following thought experiment: Take a sample of boiled potatoes and estimate their quality by gathering some information on serving temperature, boiling temperature, and the time spent boiling in the pot. Now, let there be strong statistical variations in two categories with clusters in low boiling time and high temperature and high boiling time with low boiling temperature. As a result, the (dependent!) serving temperature only varies slightly and is almost perfect for every dish. Still, there is a wide variety in the costumers' ratings. Why? Because despite having the same serving temperature, the clusters actually indicate that the one group of potatoes is still raw while the other group is overcooked. Although it is essential for the domain (food quality), this interpretation is not part of the data but introduced from the outside during analysis. This kind of interpretation often only becomes relevant during the data analysis and thus after data generation. It therefore cannot be part of any raw data being measured or sampled *without further processing*. Yet, it is an essential aspect of the domain's characteristics.

Reason: In such a setup, the domain is characterized by qualitative properties. These properties rely on interpretations from the user, analysis, and domain context. Although they can be explained in terms of certain equivalence classes in the data, these classes are not part of the data context but are introduced as data interpretations.

Solution: One possible solution attempt is to carefully study the domain context in order to identify relevant equivalence classes either before or during the analysis. Adding these equivalence classes to the data as metadata allows for a semantically meaningful classification of data items that reflects the presence of the identified characteristic domain properties in the data.

4.2 Missing Expression of Meaning

Distance measures are frequently used to find and describe structures like clusters in the data. These structures are defined solely from within the data context and typically rely on a notion of density or sparsity. Hurley and Rickard formalize intuitive properties of sparsity and analyze how well these properties are captured by sparsity measures [28]. They recognize that only few sparsity measures actually capture all identified properties. It therefore can be doubted that the interpretation of sparsity is as intuitive as it might appear. We emphasize that this is indeed a concern as it directly influences the expressiveness and validity of sparsity-based findings for the domain.

As an example, consider a bunch of differently shaped and sized ellipses (i.e., local distributions, patterns) and let them be clustered by either the Euclidean distance of their centers or the cosine similarity of their principle axes. The question now is, which distance measure is more expressive. To answer this question, we need to understand what the respective distance measure tells us about the domain. This question cannot be answered from inside the data because the interpretation requires domain knowledge. This problem has a significant implication for analysis practice: Without determining the domain interpretation of the distance measures we apply, we can impossibly say whether the structure they imply for the data is indeed characteristic for the domain.

Beside the fact that it remains unclear, which of those measures detect semantically relevant structures for the domain, Kleinberg describes in his impossibility theorem [34] that sometimes a semantically relevant measure might not exist at all. Thus, just by looking at the data, it remains unclear whether *any* chosen measure actually captures relevant domain information. As a direct consequence, obtaining an analysis result does not automatically imply an insight into the domain. Yet, the semantic interpretation of distance measures is rarely discussed explicitly in visualization applications.

Reason: For most users, closeness implies similarity (user context) but distances are semantically ambiguous because they only encode the information *that* some data items might be closer to each other than others but neither *why* they are considered closer or more similar nor *what* this means for their interpretation in the domain context.

Solution: This problem can be addressed by inferring a similarity definition from the domain context. Sometimes, such measures can be directly postulated. Where this is not possible, one can sample another set of data from the original domain and have domain experts estimate the similarity between the items in this data subset. Such measures can also be multi-dimensional in order to emphasize different characteristics for different portions of the data. This shift of emphasis can, for example, be achieved by a convex combination of dynamic weights for the measures' components.

4.3 Missing Differentiation

It is a common problem that distance and similarity measures combine multiple dimensions without regarding the units they represent. For incompatible units, the summary of features implied by the distance or similarity measure introduces ambiguity to the interpretation.

Reason: Again, closeness implies similarity for most users. The actual problem here is that distance measures tend to summarize information without a clear reference to its origin. Even if the problem of different scales and variances is solved, the data alone can still not answer the question, whether the summarized information has a meaningful interpretation in the domain context. This is to some extend alleviated by combinations of multiple measures if each measure has a clear interpretation. Yet, this is an important problem for practical applications. If distance or similarity measures that have not been deduced from the domain context are being applied, for example, to cluster the data or to find patterns in the data, one cannot be certain that these clusters or patterns are meaningful for the domain.

Solution: Practical visualization and visual analytics applications to be applied to real-world data should never rely on fixed distance or similarity definitions. Proper representations of similarities and differences determining the comparison between different entities and objects in the data need to be designed with careful consideration of the analysis question and domain context rather than only by statistical properties of the data features.

4.4 Missing Contrast

Another well-known issue with high-dimensional data is the curse of dimensionality. Distance and similarity measures that are very intuitive and descriptive in low dimensions, quickly lose their discriminating power as the data attributes become more sparsely distributed in higher dimensions [4, 25]. Thereby, the search space is increased while the contrast is being decreased which means that attributes become less distinguishable. This effectively hides all non-redundant information-carrying structure, rendering purely global analysis meaningless.

Problem: It is a common problem in real-world applications that too many data dimensions are available where only a subset of these dimensions is actually relevant for the analysis. Since this is a question of the analysis context, the subset of relevant dimensions is not only often unknown at the time the visualization is designed but can also change frequently during the analysis process.

Solution: A direct approach to this problem is to attempt a semantic dimension reduction based on the current analysis context. This can be achieved, for example, by labelling the dimensions according to their influence on certain concepts. Dimensions can then be filtered by specifying the concepts that are relevant for the current analysis task.

4.5 Missing Artifact Awareness and Structure Heritage

A well-known problem in visualization are aliasing effects – phantom patterns that are not present in the data but occur, for example, when attempting to visualize a continuous field with insufficient sampling density. Although this can be costly, the only way to "cleanly"[FR] eliminate these artifacts is to increase the sampling rate. Similar applies to the data acquisition process. The higher the precision, the more properties and patterns may be embedded in the data, and the fewer patterns in the data occur that are acquisition artifacts (Shannon-Nyquist Theorem). However, based solely on the data, it cannot be determined whether the sampling density is sufficient and whether a certain structure is inherent to the domain or introduced by the acquisition process, or both.

Reason: Data is always just a sample of the actual domain. The data context can thus never completely reflect the domain context.

Solution: The aim should be to ascertain that the data sample is representative for the investigated part of the domain. If this is not possible, the only thing that can be done is to inform the user that findings made in the data may result from flawed sampling. Because data cannot be verified against itself, it is impossible to resolve this problem from within the data context. Therefore, before a pattern or structure observed in the data can be considered for reasoning about the domain, one must attempt to falsify its existence by testing whether it is also observable in the domain.

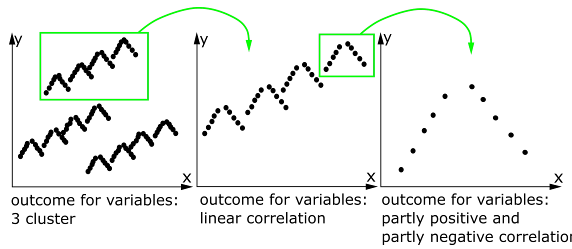

Fig. 1. Scale vs. Granularity: (left to right) a varying graininess of the data may lead to completely different outcomes for the domain.

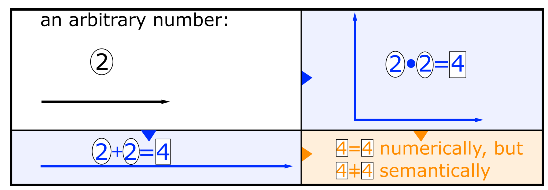

Fig. 2. What does the 4 say? Considering any units, the 4 could be interpreted as a length or an area. Just having the 4 as plain data, the true nature of it remains ambiguous.

4.6 Missing Scale and Unit Invariance

The choice of units and scales is a degree of freedom [18, 36]. The respective design decisions are usually made with respect to the general context, including aspects such as cultural heritage, best practice, or personal experience. Scaling can cause apparent correlations to appear in the data where actually there are none. Figures 1 and 2 illustrate this problem. The effect can actually be achieved quite easily by proper data manipulations [27]. Especially for complex high-dimensional systems, there is a risk of implying apparent but actually non-present correlations in the visualization [62]. Perhaps even worse, the application of measures without proper respect to the units they operate on might imply information-carrying structure in the visualization where there actually is none and thus yield misinterpretations.

Reason: The correct choice of units and scales is not encoded in the data but instead depends on the analysis and domain contexts.

Solution: The analysis context and the domain context determine whether the user needs to study local or global patterns in the data. This information should be leveraged to try to assess proper scaling and units prior to the analysis. The analysis context can also provide proper guidance for different paths for semantic zooming into the data.

4.7 Missing Precision

A common source for uncertainty in the data is the limit of precision of data measurements. By the nature of this kind of uncertainty, the values of variables and parameters are only known to be somewhere within a certain interval but are not known exactly. As a consequence, effects like the surpassing of thresholds may occur slightly sooner or later than suggested by the data. Yet, no matter where exactly within the uncertainty interval an event occurs, its interpretation is the same.

As long as all values within the intervals' boundaries show the same qualitative behavior, a single or few representatives are sufficient to completely describe the observation, capturing all relevant qualitative information. Different interpretations thus specify equivalence classes into which the different possible outcomes of the uncertain data analysis can be grouped. From a qualitative perspective, measurement error and tolerance only become relevant if they indicate multiple alternative interpretations. Note that this also applies to the kind of uncertainty observed in ensemble visualization. Although the details about the parameter values might vary, the ground truth – the observable outcomes of the experiment – remains unchanged.

As an example, consider the worst, average, and best case of the path of an asteroid crossing the course of Earth. Such data is typically generated from an ensemble of simulations with different parameters based on measurements that are imprecise to a certain degree. The three trajectories can be depicted as simple lines as long as no additional uncertainty is involved. Qualitatively, the line will cross the Earth's course in all cases. The uncertainty only needs to be made explicit if the intervals indicate that a collision cannot be ruled out. Even then, depicting the uncertainty would only be necessary close to where a collision would occur. For the rest of the asteroid's path, the uncertainty does not add information relevant for the potential collision event.

Reason: The data context does allow to compute uncertainties but does not contain interpretations of possible outcomes. For practical applications, the relevant point for understanding a problem or making a decision is usually not the uncertainty itself but the different consequences this uncertainty might imply.

Solutions: To address this problem, the relevant possible outcomes (domain context) can be assessed prior to the analysis. Domain experts can be interviewed to obtain (fuzzy) criteria under which the respective outcomes (might) occur. The visualization design can subsequently be adapted to determine whether these criteria are met. In all cases, the view should concentrate on the details of the possible outcomes rather than on the details of the uncertainty.

4.8 Missing Reliability

Incomplete or erroneous input data can be considered to introduce uncertainty in the missing or wrong entries taking the form of local discontinuities or holes in the data. Especially if this problem only affects a single dimension in a vector, ignoring the whole vector one might lose important information. On the other hand, filling the hole or replacing an obviously erroneous value introduces an estimation error. Many algorithms are only stable if there are no missing data entries. The uncertainty here results from the problem that, if the exact values are not known, an estimation can be arbitrarily wrong. While the assumptions being made to generate the estimation typically yield sound results, these results cannot be trusted with perfect confidence, especially where sampling is sparse. At the same time, it is hard to estimate the potential error if there is no additional information available to verify the estimation. As a consequence of this uncertainty, the results obtained from analysis are unreliable if conclusions are drawn from within these regions.

Reason: Flawed data neither contains the information that it is flawed, nor why it is flawed or how to properly compensate for it. Yet, in practice, data almost always has to be cleaned prior to analysis.

Solution: Sometimes, the corrections are quite obvious (e.g. outliers that provide their height in meters instead of centimeters) but proper choice of a correction often requires domain knowledge. Simply discarding a flawed data entry might also not always be a good choice. For example, analysis algorithms might require a constant sampling rate. Flaws in the data also are not always simple to solve. For example, there are measurement processes that tend to be biased due to the temperature of a probe which, however, necessarily increases during the measurement process. Therefore, the measurement is conducted in intervals with different probes which either disrupts the time-line, or results in a distortion pattern based on the systematic error of unknown quantity. For example, where one would expect a linear graph, the probe data might deliver a piece-wise exponential plot. It is necessary to discuss such effects with domain experts as their experience (user context) and knowledge (domain context) determine how they will deal with such an observation during analysis.

4.9 Applying the Principles

Let us apply the four principles introduced above to the last example to show how they can support the assessment how well a visualization does or does not support its users in obtaining specific insights.[S1]

The problem of missing reliability is not directly concerned with displaying information but rather with the effect of data sampling artifacts on what is to be seen in the display. The resulting visualization artifacts are never valid domain observations and should therefore be highlighted as such. In the above examples, a notification of the systematic deviations as part of the visualization's legend would be a comprehensible and sufficiently precise solution to establish this awareness. Discussing the actual expected shape of the displayed information with domain experts also allows to inform the viewer about how the distortion is to be interpreted correctly.[S1]

The same kind of discussion can inform the designer what should *not* be done: It is in general a bad idea to attempt to "visually correct"[FR] systematic measurement errors. First, displaying only the "correction"[FR] without the original data will confuse viewers who are aware of the systematic error and therefore expect to see the corresponding pattern in the visualization. This renders the "corrected"[FR] display even less comprehensible than the original one. Moreover, the "corrections"[FR] cannot be validated against the data due to the exact same problems that motivated their computation in the first place. Hence, there is no way to guarantee their validity. As a consequence, there is also no gain in semantic precision. Indeed, even if it does reduce the numerical error, a viewer used to working with the original pattern might even find it more difficult to interpret the "correction"[FR]. Therefore, visually correcting the patterns of systematic measurement error without referencing the original data also decreases the clarity of the display's interpretation.[S1]

5 Bridging Levels of Context

The above discussion shows that purely quantitative evaluation and reasoning provides us with insight about the visualization or about the data but is unlikely to reveal insight about the domain without further interpretation with respect to the different levels of context. Insights into the data cannot be verified against the data itself. Note that this is not specific to qualitative visual analysis but actually a problem inherent to data analysis in general. The fundamental truth of any model obtained from data analysis can never be reflected from within the model. Before observations can become domain insights, they have to be mapped back into the domain and need to be reflected within the domain context. This reflection can either be explicit or abstract and formal but is essential for obtaining insight into the domain.

The solutions we propose for the different limitations of purely quantitative analysis show that distinguishing the several levels of context and their respective contribution to the interpretation can be leveraged to improve visualizations to better suit the requirements of the specific application. We therefore promote this to a general principle of qualitative visual analysis:

The Inside-Outside Principle of Qualitative Visual Analysis
For the extraction of meaningful information from data, it is necessary to consider both the information to be found or computed inside the data context and the influence of the user context, analysis context, and domain context on how this information is being interpreted.

The interrelations between these different levels of context are illustrated in Figure 3. The traditional approach to the analysis of observations (purple arrows) analyzes them either by quantitative means or by qualitative reasoning towards a conceptual model explaining the observation. This conceptual model combines knowledge from the domain context, the analysis context, and the user context. Quantitative data analysis (green arrows) yields descriptive models for the data rather than for the observations from which the data have been sampled. To be considered for analytical reasoning, these models therefore have to be interpreted and mapped to the conceptual model. Users can obtain findings from analytical reasoning about the conceptual model (red arrows). However, these findings cannot be validated against the model they have been derived from and thus have to be confirmed by validation against external domain knowledge (blue arrows). Only upon verification can the findings be considered confirmed and only upon confirmation can they become insights. Qualitative visual analysis supports this process by establishing a connection between the conceptual model outside and the descriptive model inside the data context.[S6]

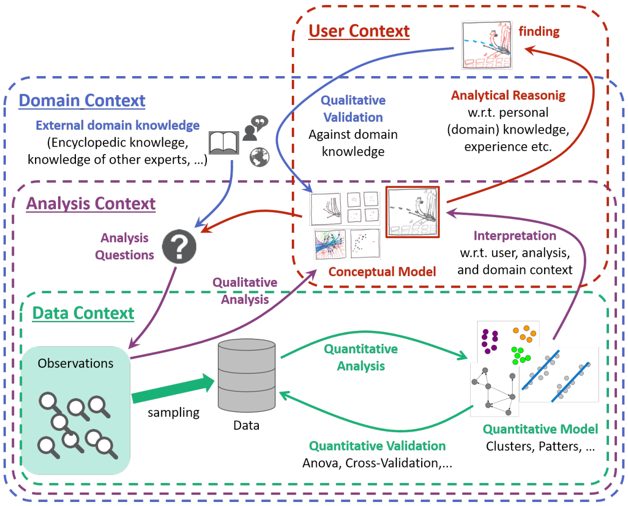

Fig. 3. Data analysis and context. Purely data-centric analysis (green) can only reveal insight into the visualization or into the data. Obtaining more sophisticated insights requires mapping the available information to a conceptual model which a user applies for reasoning. This model combines knowledge from the analysis context (purple), the user context (red), and the domain context (blue). Conclusions drawn from this reasoning can only be considered insights into the domain upon confirmation by validation against the domain.

Studying the qualitative factors in visualization and visual data analysis also has strong implications for fundamental analysis questions. In the following, we apply the inside-outside principle to discuss the different levels of context involved with fundamental questions of visual analysis such as the questions of analytical focus, subjectivity, or the emergence of insight. While the questions of analytical focus and subjectivity are commonly known in the visualization community, we include them as a proof of concept demonstrating how applying the inside-outside principle yields feasible and sound solutions of which some are common practice. Yet, the focus of our discussion is on the emergence of insight.

5.1 Analytical Focus

Visualization commonly tends to concentrate on showing data features regardless of the domain and analysis context. The decision whether to concentrate on objects to be studied or on their features should depend primarily on the analysis question. Otherwise, the visualization renders it hard to draw conclusions about the actual objects of discourse. Notice that this focus might change during the analysis process and the visualization should adapt accordingly.

Reason: The question which objects to study and which of their features to combine for better understanding is a matter of the analysis context, not of the data context.

Solution: One can attempt to model the analysis context beforehand in order to identify the representation that best supports the task. Casner followed this approach to develop an automatic visualization design system [8]. The analytical focus tends to change along with the progress of the analysis. To compensate for these changes, it has been proposed to reintegrate analytical findings into the data as meta information to be leveraged to dynamically adapt the visualization [30].[R2]

5.2 Context-Sensitivity and Subjectivity of Interpretations

The interpretation of data is not absolute but bound to different levels of context. Data analysis as it is performed today places the analyst into one specific reality among multiple possible realities. Different contexts create different realities. Consequently, the result of a specific analysis is also not the only possible outcome.

As a simple example, consider a bar chart showing the calories of certain dishes from which a diet is to be designed for some patient. If the task is to evaluate which is the best dish for the patient, the choice will be different for a patient suffering from obesity than for a patient suffering from starvation, even if though the data on the dishes is exactly the same. In this example, the same data (calories) is to be analyzed within the same domain (diets) for different subjects (patients). The respective patient's conditions shape the analysis context.

In the next example, we keep the data and the analysis task unchanged but only add details about different possible domain contexts after the analysis setup has been specified. Consider a data set of performance values measured for a number of entities capable of performing some task. In this example, a certain percentage of entities performs significantly less than the others. The available options are a relatively low cost solution removing the weak performing entities and replacing them by the other type or to apply costly measures to attempt to increase the performance of the weaker group. Given only this information without further context, the solution might appear to be quite straight forward. As a first domain, let us assume a number of machines in a production complex. For this case, replacement might actually be a good solution. Let us now change the context and let the exact same percentages be measured for the workers rather than the machines. The decision now is likely to be entirely different even though the data is exactly the same.

Another major problem here is the inherent subjectivity of data analysis. Empirical user studies show that domain knowledge has a strong impact on how analysts interact with visualizations and on the results they report. It has been found that more experienced analysts tend to provide more in-depth reports in which they include their own knowledge and conclusions while the reports of novice users would tend to concentrate on what can be read off directly from the data [29]. This is not a result of an ambiguity in the graphical representation but rather a consequence of differing interests, professions, and experiences between the individual analysts and thus different user contexts.

Reason: Human analysis relies heavily on the user context which can of course not be reflected in the data context. Furthermore, all kinds of analysis, whether automatic or human, naturally also depends on the analysis context.

Solution: Although it is admittedly hard to assess the user context, it should at least be tried. Especially where knowledge is concerned, this can be done e.g. by asking an appropriate (analysis context!) set of test questions or simply by asking users to estimate their own competence. Regarding the analysis context, designers should be sure to assess the exact task beforehand and to design the visualization towards the analysis task rather than only towards what seems to reflect the data well. Fortunately, designing visualization towards analysis tasks is established good practice in visualization and is a typical step of modern visualization application or design studies.

5.3 Contingency, Coherence and Insight Emergence

Contingency of information as a qualitative attribute is not necessarily reflected in the data context. Figure 4 shows an example of such a case. Although the images have been transformed until the data alone does not allow recognizing the faces as such, the lack of contingency is compensated by human perception's closure capability.

The example illustrates that insight is not an object or other entity in the data that can be mined or otherwise extracted. Instead, rather than being a compound of the displayed data, insight emerges from the viewer's understanding of the perceived objects and the interpretation of their appearance and arrangement. Emergent properties cannot be found in quantitative properties of the data but instead require a correlation of the displayed information with a coherent structure in the viewer's conceptual model of the data, analysis, and domain context. Reasoning about the data in order to infer complex information with respect to the additional levels of context goes beyond what is explicitly displayed or measured. Hence, redundancy, information entropy or other statistical quantitative estimates can only capture actual contingency.

Coherent structures in the data indicate the contingency of implications or dependencies. Patterns of this kind are typically not highlighted explicitly in the visualization, especially if they are not known or even expected before the analysis starts. Encoding such structures thus re-

Fig. 4. Example data from Mooney's visual closure experiment [37]. The pictures have been manipulated until a viewer without knowledge about characteristics of human facial expressions cannot identify the shapes clearly as faces. Mooney's experiment shows that viewers are capable of mapping their mental image of faces to the shapes, enabling them to correctly determine estimates of age, sex, and even expressed emotion. Contingency of information in visual data analysis is thus not limited by the actual display but rather by the viewer's ability to map the outside knowledge about possible interpretations to the perceived structures.

quires qualitative considerations, connecting the data context and the applied scaling to the available domain knowledge in order to better support the interpretation of observed patterns.

This becomes even harder if the observed data is contingent in the fact that something is absent rather than present. Consider the numbers 14, 142, 407, 43, 24. What do they have in common? Most readers will easily recognize the common feature to be that every number in this sequence contains the digit 4. Now, was it the defining commonality of the sequence 123, 35, 5, 33, 72, 92, 085, 123, 4, 378, 3? For most readers, this question will be much harder to solve. The answer is that none of these numbers contains the digit 6, while all other digits are present. This phenomenon is called the feature positive effect [47]. It states that recognizing absence as a contingent property is deliberately harder than finding commonalities between observations based on the features that are present and thus directly recognizable, especially if one is not trained in identifying absence as a property. Interestingly, the property causing the effect again is an emergent property as it can only be found by matching the fact that something is not there to the corresponding structure in the analyst's mind. Due to the positivist nature of automatic pattern recognition and similar techniques, the identification of absence of a feature as defining property for a class of elements will in most cases not be achieved by automatic analysis relying purely on quantitative considerations.

Reason: Even if an automatic method is able to recognize a pattern in the data based on statistical or other quantitative measures, the interpretation of this pattern's relevance remains with the user. This means that the contingency of information in a data visualization relies on the user's interpretation and therefore is not self-evident. Hence, contingency of interpreted information cannot be proven from within the data context but requires at least the consideration of the user context regarding the ability to detect and interpret the information correctly. As a direct consequence, the same holds for the identification of coherent structures in the display although this additionally involves the user's domain knowledge. Insights that are available via read-off are immediate by nature. These insights typically shape a user's model of the visualization (insight into the visualization), or a user's idea about the data (insight into the data). An analyst attempting to reason about the domain based on the data will gradually accumulate findings and verify hypotheses. During this emergence phase, the actual insight is not yet stable – it is being shaped during the process. This is also accounted for by the experiments of Ishack et al. who remark that, according to their findings, insight is neither generated in the visualization nor in the analyst's mind but is being shaped through the interplay of the analyst's reasoning and the visualization's depiction [29].

Solution: Supporting the reasoning process requires an in-depth analysis of all levels of context and their interrelation in the application at hand. User studies have shown that visualization users tend to apply certain standard strategies, even if the visualization does not optimally support them [42]. Knowledge about these strategies can support the design of more efficient analysis workflows. Although this involves a detailed analysis of all levels of context and might therefore become a costly endeavour, it is worth the effort: For example, it has been found that providing users with widgets containing annotation information for relevant data features can support them in the analysis process [55]. The ultimate goal in this direction would probably be the development of a tool that allows analysts to (partially) externalize their mental models and to leverage those externalizations to interactively evaluate the data and to support reshaping the externalized model.

6 Discussion and Implications for Future Work

In the above discussion, we critically review the dominant data-centric perspective to visual data analysis with respect to its capability to provide insight into the investigated domain. We show how a number of limitations inherent to this data-centric perspective can be alleviated by conclusions drawn from discussing the visual data analysis process from a broader perspective. Towards a conclusion, we summarize our findings and discuss their implications for visualization applications.

From the discussion in Section 4, we learn that it cannot be taken for granted that findings observed in the data or in a visualization are automatically relevant for the domain. Consequently, in order to allow to confirm their validity and relevance, visualization should offer means to validate findings against the domain. Our discussion also suggests that the analysis process can benefit from attempts to strengthen the connection between the domain and the data by integrating domain knowledge into the display. For insight about the visualization, the source of such knowledge is the designers' intention how the visualization is to be used. Insight about the data relies on information from analytical data processing. Insight into the domain can be supported by incorporating domain knowledge that is usually obtained from a liaison with domain experts. From these observations, we conclude that insight that can be obtained during the process of visual data analysis[FR] is already present in the combination of the different levels of context involved. Visualization can offer an interface to access the additional information to be gathered from the different levels of context. Hence, it should not only be designed to optimally represent the data but also to optimally support its users in reasoning about the data and the domain.[S2]

To achieve this, the discussions about visualization applications have to clearly point out how exactly the users are meant to obtain insights. In order to verify these mechanisms, evaluation techniques have to be developed that allow to capture and investigate analytical provenance and to compare this provenance between different users. Research in visualization design should attempt to identify novel techniques and approaches for supporting the users in validating findings they observe against the domain. Theoretical research needs to study the principles underlying the formation of insights in order to support reasoning about how a visualization can best support this process. For example, research in this direction could study the interpretations translating the graphical representation into structures the viewer can reason about.[S1]

We introduce the concept of qualitative visual analysis as a conceptual framework to structure the discussion of the qualitative considerations contributing to a holistic treatment of the analysis process. In Section 4, we apply this concept to explain how a viewer is expected to obtain insights from the solutions we propose for the different problems discussed there. Our discussion shows that our four principles in their current form are applicable to analyse approaches for visualization applications in order to discuss whether they support an analysis from a qualitative perspective. We therefore expect that qualitative visual analysis can serve as a common conceptual framework [1] to structure and align the results obtained from the different directions of further research we point out above.[S4]

7 Conclusion

In this paper, we critically review the data-centric perspective on data analysis currently predominant in visualization as a field of research. We identify a number of limitations common in practical applications of visualization and show how they can be alleviated leveraging qualitative considerations like user experience or domain knowledge. To structure this discussion, we introduce the concept of qualitative visual analysis

[1]metaphorically speaking: "why we should not ignore what is not in the numbers"[FR]

which we apply to discuss the influence of different levels of context on the analysis process. For visualization researchers, our discussion contributes the understanding of the emergent nature of insight to be obtained during the process of visual data analysis[FR]. For practitioners, our examples demonstrate how qualitative considerations can establish a closer connection between the visualization and the domain.[S4]

References

[1] J.-W. Ahn and P. Brusilovsky. Adaptive visualization of search results: Bringing user models to visual analytics. *Information Visualization*, 8(3):167–179, 2009.

[2] R. Amar and J. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 143–150. IEEE.

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

[4] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pp. 217–235. Springer-Verlag, London, UK, 1999.

[5] D. Boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication Society*, 15:662–679, 01 2012.

[6] S. Card, J. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision To Think*. 01 1999.

[7] S. Casner and J. H. Larkin. Cognitive efficiency considerations for good graphic design. p. 12, 08 1989.

[8] S. M. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Trans. Graph.*, 10(2):111–151, 1991. doi: 10.1145/108360.108361

[9] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining insight for visual analytics. *IEEE Computer Graphics and Applications*, 29(2):14–17, March 2009. doi: 10.1109/MCG.2009.22

[10] M. Chen and D. J. Edwards. "isms" in visualization. In M. Chen, H. Hauser, P. Rheingans, and G. Scheuermann, eds., *Foundations of Data Visualization*. Springer, To appear – possibly 2020.

[11] M. Chen, L. Floridi, and R. Borgo. *What Is Visualization Really For?*, pp. 75–93. Springer International Publishing, Cham, 2014.

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

[13] W. S. Cleveland and R. McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985.

[14] B. U. Dur. Data visualization and infographics in visual communication design education at the age of information. *Journal of Arts and Humanities*, 3(5), 2014.

[15] M. Elizabeth Raven and A. Flanders. Using contextual inquiry to learn about your audiences. 20:1–13, 02 1996.

[16] O. J. Espinosa, C. Hendrickson, and J. Garrett. Domain analysis: a technique to design a user-centered visualization framework. In *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*, pp. 44–52. IEEE, 1999.

[17] J.-D. Fekete, J. J. Van Wijk, J. T. Stasko, and C. North. The value of information visualization. In *Information visualization*, pp. 1–18. Springer, 2008.

[18] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff. Selecting the aspect ratio of a scatter plot based on its delaunay triangulation. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2326–2335, 2013.

[19] L. Gitelman. *Raw data is an oxymoron*. MIT Press, Mar. 2013.

[20] M. Golemati, C. Halatsis, C. Vassilakis, A. Katifori, and U. o. Peloponnese. A context-based adaptive visualization environment. In *Tenth International Conference on Information Visualisation (IV'06)*, pp. 62–67, July 2006.

[21] D. Gotz and M. X. Zhou. An empirical study of user interaction behavior during visual analysis. *Technical Report RC24525*, 2008.

[22] J. T. Guthrie, S. Weber, and N. Kimmerly. Searching documents: Cognitive processes and deficits in understanding graphs, tables, and illustrations. *Contemporary Educational Psychology*, 18(2):186 – 221, 1993.

[23] P. H.-G. Visualization within environments supporting human communication. *Future Generation Computer Systems*, 15(1):109 – 117, 1999.

[24] R. Haber and D. McNabb. Visualization idioms: A conceptual model for scientific visualization systems. In L. R. G.M. Nielson, B. Shriver, ed., *Visualization in Scientific Computing*. IEEE Computer Society, 1990.

[25] A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 506–515. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[26] K. Holtzblatt and S. Jones. Contextual inquiry: A participatory technique for system design. 01 1993.

[27] D. Huff and I. Geis. *How to Lie with Statistics*. W. W. Norton, 2010.

[28] N. Hurley and S. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

[29] S. Ishack, C. Baber, and A. Duncan. The effect of domain knowledge on interaction with visual analytics. In *Contemporary Ergonomics and Human Factors 2015: Proceedings of the International Conference on Ergonomics & Human Factors 2015, Daventry, Northamptonshire, UK, 13-16 April 2015*, p. 86. CRC Press, 2018.

[30] B. Karer, I. Scheler, and H. Hagen. Panning for insight: amplifying insight through tight integration of machine learning, data mining, and visualization. In *Proceedings of the Conference on Machine Learning Methods in Visualisation for Big Data*, pp. 1–5. Eurographics Association, 2018.

[31] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Information visualization. chap. Visual Analytics: Definition, Process, and Challenges, pp. 154–175. Springer-Verlag, Berlin, Heidelberg, 2008.

[32] N. Kijmongkolchai, A. Abdul-Rahman, and M. Chen. Empirically measuring soft knowledge in visualization. *Computer Graphics Forum*, 36(3):73–85, 2017. doi: 10.1111/cgf.13169

[33] Y.-S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 1375–1386. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3025453.3025592

[34] J. Kleinberg. An impossibility theorem for clustering. pp. 446–453. MIT Press, 2002.

[35] S. M. Kosslyn. Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3):185–225, 1989.

[36] D. J. Lehmann and H. Theisel. The lloydrelaxer: An approach to minimize scaling effects for multivariate projections. *IEEE Transactions on Visualization and Computer Graphics*, to appear, 2017.

[37] C. M. Mooney. Age in the development of closure ability in children. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 11(4):219, 1957.

[38] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6), 2009.

[39] K. Nazemi and J. Kohlhammer. Visual variables in adaptive visualizations. In *UMAP Workshops*, 2013.

[40] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, 2006.

[41] M. Petre and T. Green. Learning to read graphics: Some evidence that 'seeing' an information display is an acquired skill. *Journal of Visual Languages & Computing*, 4(1):55 – 70, 1993.

[42] M. Pohl and J. Doppler Haider. Sense-making strategies for the interpretation of visualizations—bridging the gap between theory and empirical research. *Multimodal Technologies and Interaction*, 1(3):16, 2017.

[43] A. J. Pretorius and J. J. Van Wijk. What does the user want to see? what do the data want to be? *Information Visualization*, 8(3):153–166, 2009.

[44] R. M. Ratwani, J. G. Trafton, and D. A. Boehm-Davis. Thinking graphically: Extracting local and global information. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 25, 2003.

[45] J. C. Roberts, C. Headleand, and P. D. Ritsos. Sketching designs using the five design-sheet methodology. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):419–428, Jan 2016.

[46] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, Dec 2014. doi: 10.1109/TVCG.2014.2346481

[47] R. Sainsbury. The "feature positive effect" and simultaneous discrimination learning. *Journal of Experimental Child Psychology*, 11(3):347 – 356, 1971. doi: 10.1016/0022-0965(71)90039-7

[48] B. Saket, A. Endert, and Ç. Demiralp. Data and task based effectiveness of basic visualizations. *CoRR*, abs/1709.08546, 2017.

[49] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization*

and Computer Graphics, 11(4):443–456, July 2005. doi: 10.1109/TVCG.2005.53

[50] M. Scaife and Y. Rogers. External cognition: how do graphical representations work? *International journal of human-computer studies*, 45(2):185–213, 1996.

[51] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, Dec 2012.

[52] J. Sheidin, J. Lanir, C. Conati, D. Toker, and T. Kuflik. The effect of user characteristics in time series visualizations. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, p. 380–389. Association for Computing Machinery, New York, NY, USA, 2020.

[53] S. Simon, S. Mittelstädt, D. A. Keim, and M. Sedlmair. Bridging the Gap of Domain and Visualization Experts with a Liaison. In E. Bertini, J. Kennedy, and E. Puppo, eds., *Eurographics Conference on Visualization (EuroVis) - Short Papers*. The Eurographics Association, 2015.

[54] M. Smuc, E. Mayr, T. Lammarsch, A. Bertone, W. Aigner, H. Risku, and S. Miksch. Visualizations at first sight: Do insights require training? In A. Holzinger, ed., *HCI and Usability for Education and Work*, pp. 261–280. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[55] A. Srinivasan, S. Drucker, A. Endert, and J. Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 08 2018. doi: 10.1109/TVCG.2018.2865145

[56] D. Streeb, M. El-Assady, D. Keim, and M. Chen. Why visualize? untangling a large network of arguments. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 09 2019. doi: 10.1109/TVCG.2019.2940026

[57] M. Tory and T. Möller. Evaluating visualizations: do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5):8–11, Sept 2005.

[58] J. G. Trafton, S. Marshall, F. Mintz, and S. B. Trickett. Extracting explicit and implict information from complex visualizations. In *International Conference on Theory and Application of Diagrams*, pp. 206–220. Springer, 2002.

[59] J. G. Trafton and S. B. Trickett. A new model of graph and visualization usage. Technical report, NAVAL RESEARCH LAB WASHINGTON DC, 2001.

[60] J. J. Van Wijk. Views on visualization. *IEEE transactions on visualization and computer graphics*, 12(4):421–432, 2006.

[61] P. Vickers, J. Faith, and B. N. Rossiter. Understanding visualization: A formal approach using category theory and semiotics. *CoRR*, 2013.

[62] T. Vigen. *Spurious Correlations*, vol. ISBN-10: 0316339431. 2015.

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

[64] C. Ware, N. J. Pioch, and E. K. Jones. Visual thinking algorithms for visualization of social media memes, topics, and communities. 2013.

[65] J. Yi, Y. Kang, J. Stasko, and J. Jacko. Understanding and characterizing insights: How do people gain insights using information visualization? p. 4, 01 2008. doi: 10.1145/1377966.1377971