Visualizing Uncertainty in Sets

Christian Tominski, *University of Rostock, DE*
Michael Behrisch, *Utrecht University, NL*
Susanne Bleisch, *FHNW University of Applied Sciences and Arts Northwestern Switzerland, CH*
Sara Irina Fabrikant, *University of Zurich, CH*
Eva Mayr, *University for Continuing Education Krems, AT*
Silvia Miksch, *TU Wien, AT*
Helen Purchase, *Monash University, AU*

Abstract—Set visualization facilitates the exploration and analysis of set-type data. However, how sets should be visualized when the data is uncertain is still an open research challenge. To address the problem of depicting uncertainty in set visualization, we ask (i) which aspects of set type data can be affected by uncertainty and (ii) which characteristics of uncertainty influence the visualization design. We answer these research questions by first describing a conceptual framework that brings together (i) the information that is primarily relevant in sets (i.e., set membership, set attributes, and element attributes) and (ii) different plausible categories of (un)certainty (i.e., certainty, undefined uncertainty as a binary fact, and defined uncertainty as quantifiable measure). Following the structure of our framework, we systematically discuss basic visualization examples of integrating uncertainty in set visualizations. We draw on existing knowledge about general uncertainty visualization and previous evidence of its effectiveness.

Analysing set data encompasses consideration of the sets themselves, the elements within the sets, and attributes of both the sets and the elements. Take for example academic courses at a university (e.g., Biology, Mathematics) as sets, and the students enrolling in those courses as the set elements. Set visualizations aim to express such set-type data visually to support analysis and better understanding. Relevant set analytical questions involve set memberships (i.e., who is enrolled in which course), set cardinality (i.e., how many students are in a course), or set intersections (i.e., which course combinations are favored by students). Such kinds of set analytical questions can be answered, for example, with the help of Euler diagrams, Venn diagrams, and bipartite node-link representations. While set visualizations themselves are an active research frontier [1], there are far fewer research activities focusing on the implications of uncertainty for set visualization [2]. For our courses-and-students example, we might not know exactly how many students are enrolled in a course or how old they are.

In fact, it is challenging to design visual representations of sets where uncertainty is involved (see **Table 1** A.). This is because both the set data themselves and also the information about their uncertainty need to be communicated to a reader. The interpretation of this uncertainty has a major impact on decisions that are made based on the data, not only for simple applications such as course planning, but also for more complex scenarios like comparison of ensemble forecasting models or gene-to-phenotype mapping.

So far, the literature offers little insight into the implications of uncertainty for set visualization [2]. In particular, a distinction of classes of uncertainty in the context of set-type data is missing. Only if we know, however, what types of uncertainty are relevant for set-type data can we design expressive visual representations. Therefore, the main objective of this paper is to systematize uncertainty considerations for

XXXX-XXX © 2023 IEEE
Digital Object Identifier 10.1109/XXX.0000.0000000

This article has been accepted for publication in IEEE Computer Graphics and Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCG.2023.3300441

set visualization. To this end, we devise a framework that brings together (a) different facets of set data that might be affected by uncertainty, and (b) different types of uncertainty that might influence the visualization design. As the primarily relevant data facets in sets, our framework lists: set membership, set attributes, and element attributes. In terms of different types of (un)certainty, we distinguish: certainty, undefined uncertainty as a binary fact, and defined uncertainty as quantifiable measure.

From our framework, we derive interesting combinations of data facets and types of uncertainty that would benefit from future dedicated visualization strategies. For each combination, we sketch initial thoughts on possibly useful visualization designs. While some cases are rather straightforward, others seem to be more intricate to deal with. In any case, we make use of previous empirical evidence on uncertainty visualization (see Table 1 B.) to inform our example designs.

Before developing our framework, we will next introduce basic set and uncertainty terminology.

Sets & Uncertainty

Set theory has been investigated in mathematical logic in the nineteenth century by Cantor [3] to describe collections of objects, called sets, and their elements. Sets do not impose any ordering on their elements. Sets may overlap, making well-defined relations between sets possible, including containment, exclusion, and intersection. Moreover, both sets and elements may have various attributes associated with them. Accordingly, the primarily relevant data characteristics (D) for set-type data are: (i) set membership, (ii) set attributes, and (iii) element attributes.

Uncertainty (U) relates to information that is unknown, vague, or of varying accuracy. So, a good starting point is to think about what is known and what is unknown. In a perfect world, we know the data and assume they are accurate. There is no uncertainty, which we denote as $U = 0$. For set-type data this means that we know for certain the sets and the elements, their memberships, and their attributes.

However, in the real world, uncertainty is commonly encountered in everyday life [4]. It is inherent to any piece of information and thus also present in any dataset, data model, or visualization and has been studied in many scientific disciplines and academic fields [5]. Given the universal relevance of uncertainty, it is not surprising to find various notations and categorizations in different fields. Terms like aleatoric uncertainty or stochastic uncertainty are used in mechanical engineering. But also more common terms such as incertitude, probability, or ignorance appear in the context of uncertainty (see Table 1 A.).

The common theme behind the heterogeneous landscape of terminology is that uncertainty is present in all parts of the data-driven scientific research process [6], starting with measurement and data capture, data transformations and processing, data modeling and visualization, and finally human inference and decision making with visual data displays [7]. Uncertainty in data sources can be of locational, temporal, and/or semantic nature [6]. Uncertainty issues in data capturing can stem from, e.g., data provenance, acquisition methods, and measurement inaccuracies. Uncertainty also arises and gets further propagated in data transformations and processing, including data modeling. Uncertainty can also occur in data portrayal methods, and lead to perceptual issues of the viewer of uncertainty-depicting visualizations. Finally, uncertainty might arise in human interpretations and decision making. Hence, we argue that uncertainty should always be also considered in sets and their visualization.

An important question to ask is how much do we actually know about the uncertainty in our data? Here, we distinguish two scenarios. One scenario is that we know that there is uncertainty, but we cannot tell accurately where it is, what it is, or how much of it exists. In other words, we know for a fact that uncertainty is present, but no further details. We denote this as $U > 0$. In the second scenario, we also know that uncertainty exists, and we know with certainty where, what, and how much of it is in our data. For the sake of simplicity, we denote this as $U = p$. The letter p is a strong simplification of what could be known about the uncertainty and p can take different forms. When set membership is certain, one can say either $a \in X$ or $a \notin X$. Under uncertainty, p might denote a probability of a being a member of X, $P(a, X) = p$, which is a notation known from fuzzy sets. We could also say that p denotes a more complex probability distribution, e.g., $p = \mathcal{N}(\mu, \sigma^2)$, based on which set membership is decided. In relation to the data attributes of elements or sets, we may understand p as the probability value or probability distribution of an attribute taking a particular data value. Additionally, it is common for uncertain attribute values to specify them via a range of possible values, in which case $p = [l, u]$ is some interval with a lower and upper bound of l and u.

Overall, the characteristics of set data D and the types of uncertainty U form the basis for a conceptual framework of uncertainty in set visualization, which will be described next.

TABLE 1. Uncertainty visualization literature overview.

A. General Uncertainty Visualization

N. Gershon, “Visualization of an imperfect world,” *IEEE Computer Graphics and Applications*, vol. 18, no. 4, 1998. DOI: 10.1109/38.689662

H. Griethe and H. Schumann, “The visualization of uncertain data: Methods and problems,” in *Proceedings of Simulation and Visualization (SimVis)*, SCS Publishing House, 2006

K. Brodlie *et al.*, “A review of uncertainty in data visualization,” in *Expanding the Frontiers of Visual Analytics and Visualization*, Springer, 2012. DOI: 10.1007/978-1-4471-2804-5_6

G. Bonneau *et al.*, “Overview and state-of-the-art of uncertainty visualization,” in *Scientific Visualization*, Springer, 2014. DOI: 10.1007/978-1-4471-6497-5_1

D. Sacha *et al.*, “The role of uncertainty, awareness, and trust in visual analytics,” *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, 2016. DOI: 10.1109/TVCG.2015.2467591

J. S. Mason *et al.*, “Special issue introduction: Approaching spatial uncertainty visualization to support reasoning and decision making,” *Spatial Cognition & Computation*, vol. 16, no. 2, 2016. DOI: 10.1080/13875868.2016.1138117

S. Dübel *et al.*, “Visualizing 3D Terrain, Geo-Spatial Data, and Uncertainty,” *Informatics*, vol. 4, no. 1, 2017. DOI: 10.3390/informatics4010006

A. Jena *et al.*, “Uncertainty visualisation: An interactive visual survey,” in *Pacific Visualization Symposium*, IEEE, 2020. DOI: 10.1109/PacificVis48177.2020.1014

B. Empirical Studies on Uncertainty Visualization

A. M. MacEachren *et al.*, “Visualizing geospatial information uncertainty: What we know and what we need to know,” *Cartography and Geographic Information Science*, vol. 32, no. 3, 2005. DOI: 10.1559/1523040054738936

A. M. MacEachren *et al.*, “Visual semiotics & uncertainty visualization: An empirical study,” *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, 2012. DOI: 10.1109/TVCG.2012.279

C. Kinkeldey *et al.*, “How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies,” *Carto. J.*, vol. 51, no. 4, 2014. DOI: 10.1179/1743277414Y.0000000099

H. Guo *et al.*, “Representing uncertainty in graph edges: An evaluation of paired visual variables,” *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 10, 2015. DOI: 10.1109/TVCG.2015.2424872

J. Hullman *et al.*, “Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering,” *PloS one*, vol. 10, no. 11, 2015. DOI: 10.1371/journal.pone.0142444

G. McKenzie *et al.*, “Assessing the effectiveness of different visualizations for judgments of positional uncertainty,” *Intl. J. Geograph. Inform. Sci.*, vol. 30, no. 2, 2016. DOI: 10.1080/13658816.2015.1082566

T. Gschwandtner *et al.*, “Visual encodings of temporal uncertainty: A comparative user study,” *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, 2016. DOI: 10.1109/TVCG.2015.2467752

M. Kay *et al.*, “When(Ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems,” in *Proceedings of the CHI*, ACM, 2016. DOI: 10.1145/2858036.2858558

M. Korporaal *et al.*, “Effects of uncertainty visualization on map-based decision making under time pressure,” *Frontiers in Computer Science*, vol. 2, 2020. DOI: 10.3389/fcomp.2020.00032

A. C. Robinson, “Representing the presence of absence in cartography,” *Annals of the American Association of Geographers*, vol. 109, no. 1, 2019. DOI: 10.1080/24694452.2018.1473754

C. Bors *et al.*, “Exploring Time Series Segmentations Using Uncertainty and Focus+Context Techniques,” in *EuroVis Short Paper Proceedings*, Eurographics Association, 2020. DOI: 10.2312/evs.20201040

I. Kübler *et al.*, “Against all odds: Multicriteria decision making with hazard prediction maps depicting uncertainty,” *Annals of the American Association of Geographers*, vol. 110, no. 3, 2020. DOI: 10.1080/24694452.2019.1644992

C. Uncertainty in Set Visualization

C. Vehlow *et al.*, “Visualizing fuzzy overlapping communities in networks,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, 2013. DOI: 10.1109/TVCG.2013.232

L. Zhu *et al.*, “Visualizing fuzzy sets using opacity-varying freeform diagrams,” *Inf. Vis.*, vol. 17, no. 2, 2018. DOI: 10.1177/1473871617698517

J. Görtler *et al.*, “Bubble treemaps for uncertainty visualization,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, 2018. DOI: 10.1109/TVCG.2017.2743959

M. Sondag *et al.*, “Uncertainty treemaps,” in *Pacific Visualization Symposium*, IEEE, 2020. DOI: 10.1109/PacificVis48177.2020.7614

TABLE 2. Framework of uncertainty in set visualization with relevant set characteristics and categories of (un)certainty.

		Data characteristics (D)		
		Set membership	**Set attributes**	**Element attributes**
Types of uncertainty (U)	**Certainty** No uncertainty in the data U = 0	**II. Set visualization**		**I. Multivariate visualization**
	Undefined uncertainty Uncertainty in the data, but it is undefined U > 0	Focus of this work: **Uncertainty in set visualization**		**III. Uncertainty visualization**
	Defined uncertainty Uncertainty in the data, and it is well-defined U = p			

A Framework for Uncertainty in Set Visualization

In terms of data characteristics D, the framework distinguishes: **set membership**, **set attributes**, and **element attributes**. Related to uncertainty U, we use the different plausible types of (un)certainty: **certainty** ($U = 0$), **undefined uncertainty** as a binary fact ($U > 0$), and **defined uncertainty** as a quantifiable measure ($U = p$). The framework is depicted in **Table 2**, whose columns and rows respectively represent D and U. The cells of the table correspond to different combinations of data characteristic and type of uncertainty for which adequate visualization methods are needed.

The most interesting cells in Table 2 are marked in orange, and will be described later. For the green cells in the table, established visualization methods already exist. Our framework identifies three relevant areas (I.-III.) in this context. First, when the data are certain ($U = 0$), multivariate visualization methods can be used to depict element attributes. Second, for certain set memberships and set attributes, one can use existing set visualization methods. Third, when the attributes of individual data elements are uncertain ($U > 0$ or $U = p$), uncertainty visualization gets involved.

For multivariate visualization (I.), we refer to the existing visualization literature [8], [9]. For the green cells set visualization (II.) and uncertainty visualization (III.), we provide further details below because they directly inform the design of uncertain set visualizations.

Set Visualization

Alsallakh et al. [1] provide a comprehensive overview of set visualization methods. They identified six categories of techniques: Euler-based diagrams, overlays, node-link diagrams, matrix-based techniques, aggregation-based techniques, and scatter plots (see **Figure 1** for examples). Specific techniques in these categories such as, for example, BubbleSets, KelpFusion, OnSet, and Parallel Sets have different strengths and weaknesses and are suited for different set-analytic tasks (e.g., find, count, or filter elements; determine set cardinality, intersections, or unions; understand value distribution in the intersection of groups of sets). For a detailed discussion of individual techniques and their effectiveness for certain tasks, we refer to the original survey article [1].

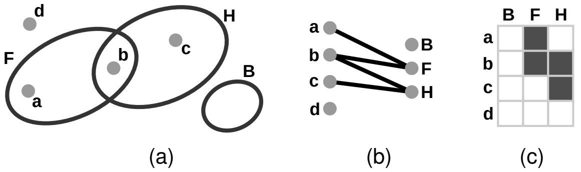

FIGURE 1. Examples of common set visualizations: (a) Euler/Venn diagrams, (b) bipartite node-link diagrams, and (c) matrices, all representing the same data.

Uncertainty Visualization

Designing visual representations of uncertain data is challenging, mainly due to the fact that not only the data D themselves need to be encoded visually, but also the information about their uncertainty U needs to be communicated. Above all, visualization users must be able to extract all the encoded information (both the data and their uncertainty) from the visualization, which can be formulated abstractly as a pipeline, inspired by algebraic visualization design [10]:

$$(D, U) \xrightarrow{m} V \xrightarrow{i} (D', U').$$

The visualization designer defines a mapping m of data D and uncertainty U to create a visual representation V. Through an interpretation i of the visual representation V, human observers extract their own versions of data D' and uncertainty information U'. The scientific challenge is to understand the cognitive process of i and to devise mappings m so that ideally $D = D'$ and $U = U'$ for all human observers. The congruence of D and D', as well as U and U', can serve as a guiding principle for the visualization of uncertain data.

There are many ways to depict data uncertainty (see Table 1 A.). Much empirically-grounded research on the representation of uncertainty exists, primarily in geospatial visualization, but also for graph visualization, statistical visualization, and temporal visualization (see Table 1 B.).

The empirically validated uncertainty visualization framework proposed by MacEachren and colleagues [11] is an attractive candidate to directly transfer more broadly to the depiction of uncertainty in general and in sets specifically. MacEachren et al. first empirically assessed the intuitiveness of a visual variable (e.g., location of symbol, size, color value, color hue, color saturation, texture, orientation) to judge the suitability of abstract or iconic point symbols for depicting data variation in a given category of uncertainty. Second, they also measured the relative performance of the most intuitive point symbol depiction of uncertainty with a focus on symbol effectiveness for a typical use case: assessing and comparing the aggregate uncertainty in two regions of a graphic display. Their stimuli are generic enough so that findings can be transferred to many data visualization types and use cases, including the visualization of sets.

Based on their studies, MacEachren et al. derived generalizable design guidelines. For example, the visual variables fuzziness and relative location and distance (from a known location in the center of a crosshairs) work particularly well for the depiction of uncertainty in point symbols. Color value and arrangement are also rated highly. Both size and transparency are potentially usable. On the other hand, color saturation of a point symbol, often cited as intuitively related to uncertainty, was ranked quite low. Later we apply the knowledge and guidelines from MacEachren et al.'s studies to inform the design of set visualizations including uncertainty.

Now that we have dealt with the ’easier’ green boxes from Table 2, we will move on to discuss the ’tougher’ orange box, the visualization of uncertain set characteristics.

Uncertainty in Set Visualization

In comparison to general uncertainty visualization, the representation of uncertain set data has received less attention (see Table 1 C.). A few attempts exist for the visualization of fuzzy sets by Vehlow et al. and Zhu et al., where set membership has a defined uncertainty. For set attributes, we are aware of only two prior works, a treemap representation for set size in set hierarchies with defined uncertainty by Sondag et al. and a circle packing visualization by Görtler et al. Visual representations of undefined uncertainty within sets have not been developed until now.

Given the scarcity of visualization methods for uncertain information in sets, we next present examples of visualization designs for each relevant orange cell from our framework from Table 2.

Design Examples for Uncertainty in Set Visualization

It is sensible to begin the process of depicting uncertainty by constructing a visualization of the data that is ’certain’, and then subsequently adapting or augmenting it as necessary to depict the uncertainty. Gershon [12] calls this *intrinsic* representation of uncertainty as opposed to *extrinsic* representations where uncertainty information is shown in separate auxiliary displays, like a supplementary diagram or text. The decision on whether to use intrinsic or extrinsic representations may depend on the complexity of both the data and the uncertainty.

Uncertain Set Membership

Communicating set membership is essential for set visualization [1]. In the following, we use the dataset with students and courses from **Figure 2** for illustration. Following Cantor’s [3] notation, elements are denoted by small letters, whereas sets are denoted with capital letters. For elements a (Alex), b (Ben), c (Chris), and d (Dana) membership is certain, and we also know that set B (Biology) is empty. We are uncertain, however, about the membership of elements e (Eva) and f (Frank) as well as of set M (Math).

Visualizing certain set membership In general, certain set membership ($U = 0$) can be represented in two different ways: implicitly or explicitly. Implicit representations do not use a dedicated graphical mark to visualize set membership, but rather some relation between existing marks. A common implicit example was already shown in Figure 1 (a) where sets are visualized as ellipses and elements of sets are visualized as

Elements

ID	*Name*
a	Alex
b	Ben
c	Chris
d	Dana
e	Eva
f	Frank

Sets

ID	*Course*
B	Biology
F	French
H	History
M	Math

Membership

Element	*Set*	*Uncertainty*
null	B	U=0
a	F	U=0
b	F	U=0
b	H	U=0
c	H	U=0
d	null	U=0
e	undef	U>0 / U=p
f	undef	U>0 / U=p
undef	M	U>0 / U=p

FIGURE 2. Example dataset with certain and uncertain set memberships. 🅭🅯

dots within the ellipses. In this case, set membership is implicitly encoded through *inclusion* of the dots in ellipses. In addition to inclusion, also adjacency and overlap are possible implicit representations [13].

In contrast to implicit representations, explicit representations have dedicated graphical marks to represent set memberships (in addition to marks representing sets and elements). Examples include bipartite node-link diagrams and matrices as shown in Figures 1 (b) and (c). In node-link diagrams, sets and their elements are both depicted as dots, and their set membership is visualized by explicitly drawing links between set dots and their element dots. For matrices, sets and their elements are assigned to matrix columns and matrix rows, respectively. Set membership is explicitly represented by the matrix cells, whose content indicates which and where elements are members of a set, for example, by a certain fill color or symbol.

Visualizing uncertain set membership When uncertainty needs to be considered, it is necessary to vary the representation of set membership in order to communicate either the fact that undefined uncertainty is present ($U > 0$) or the exact information we may have about the defined uncertainty ($U = p$). Varying an implicit representation of set membership (i.e., inclusion, adjacency, or overlap of graphical elements) is difficult. Where in Figure 1 (a) should we place the dots for the uncertain elements *e* and *f* of our data and how should we draw the ellipse for set *M*? The problem is that graphical marks may or may not include, be adjacent, or overlap other marks, but there are no other states that could be used to indicate uncertainty. So, for implicit representations, we would first need to add further graphical marks before uncertainty could be encoded. In contrast, explicit representations already have dedicated marks for set membership, which offer several options for perceivable variation to visualize the uncertainty of set memberships following the guidelines offered by MacEachren et al. [11].

Next, we sketch two example designs for the case of explicit representations: (i) bipartite node-link diagrams and (ii) matrices. The uncertain set memberships (elements *e* and *f* and set *M*) of the data from Figure 2 will be used for illustration.

i. Bipartite node-link diagrams In bipartite node-link diagrams, we may vary the visual properties of links to communicate uncertainty. **Figure 3** (a-c) shows certain set memberships as bold dark links. The figure further shows three different variants of encoding uncertainty. The fact that uncertainty is present ($U > 0$) can be visualized by varying line width and color value for uncertain memberships as in Figure 3(a). This makes certain memberships (bold dark lines) easily distinguishable from uncertain memberships (thin gray lines). Note, however, that elements with uncertain membership must be linked to all possible sets, and vice versa for uncertain sets. This may lead to visual clutter when many membership links are uncertain.

A design goal could thus be to reduce link clutter. Therefore, the variant in Figure 3 (b) replaces the full-length links for uncertain memberships with small link fans, which are graphically less demanding. This way, clutter can be reduced, but readers need to mentally connect the elements to all possible sets.

Finally, for the variant in Figure 3 (c), we assume that we know exact probability values for possible set memberships ($U = p$). This allows us to maintain the explicit connection of elements and sets, and also to encode the different probability values per membership by varying lightness and width of lines. Thinner and lighter lines indicate lower probability values.

ii. Matrices For matrices (see Figures 3 (d-f)), one can follow a similar strategy of varying the explicit representation of set membership. While we changed the graphical properties of 1D lines in the case of bipartite node-link diagrams, we now adapt the 2D cells of a matrix. If only the presence of uncertainty is known ($U > 0$), then we can differentiate certain and uncertain set memberships by varying the fill color of matrix cells as in Figure 3 (d). However, this solution might again draw too much attention to the uncertain information, simply because many cells need to be marked. To better balance certain and uncertain information, one can reduce the size of the cell marks as indicated in Figure 3 (e).

Continuing on this line of thought, exact quantitative information about the uncertainty ($U = p$) can be encoded by varying size and color of matrix cells together as indicated in Figure 3 (f). These and similar encodings in matrix cells have already proven effec-

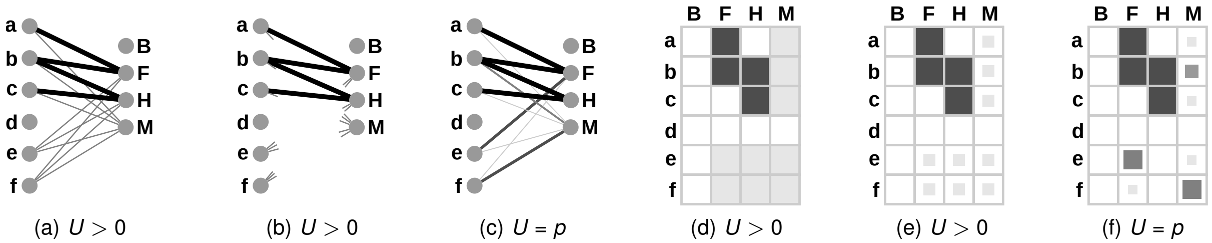

FIGURE 3. Variants of visualizing uncertain set membership in bipartite node-link diagrams and matrix representations. 🅭🅯

tive for visualizing multivariate graphs, in particular for showing the weight of edges [14].

Overall, our examples show that explicit representations of set membership can be successfully adapted to communicate different types of (un)certainty. However, the visual saliency of the uncertain information and the certain information need to be cleverly balanced depending on the communicative goal of the visualization. While set membership is just a single piece of information that may be uncertain, the design of visualizations becomes more complicated when multiple uncertain set attributes are involved.

Uncertain Set Attributes

We now assume that the relationship between the sets and the elements is given, and we are interested in visualizing an overall aggregated property of the sets. We are not interested in representing the elements themselves (indeed, there may be too many to represent explicitly).

While set attributes are a separate column in our framework in Table 2, they are related to the other two columns of set membership and element attributes. Set membership directly determines the set attribute of set size, and the values of element attributes may be the basis for derived set attributes. For the case that set size is certain ($U = 0$), there are known visualizations [1], and uncertain set size ($U > 0$, $U = p$) relates to uncertainty in set membership, for which designs were discussed in the previous section. This section, therefore, focuses on the more general case of visualizing set attributes derived from elements.

We use the following scenario for illustration: There are students (elements) who can enroll in various courses (sets), and among other information we know the students' residential status (domestic or international) and their age. The two derived set attributes that we are interested in are (i) the proportion of international students in a course, the 'international residential ratio', called IRR, and (ii) the average age of the students, called AA. The IRR and AA can be derived for each course individually and also for possible subsets (i.e., unions, intersections, differences). In the following, we present example visualizations for IRR and AA for the three types of (un)certainty identified in our framework. For each type, we first indicate the raw data with attribute values for individual elements, and then the actual visual representation of the derived set attribute values. Again, our examples are informed by published guidelines for uncertainty visualization [11].

Visualizing certain set attributes We start with the case $U = 0$. The residential status and the ages of the students are the element attributes, and their enrollment in different courses is set membership. In **Figure 4** (top, left), the first three circles indicate the raw data as filled (domestic) and empty (international) dots. The three circles below are colored in shades of blue to represent the aggregated IRR values for the sets and their intersections. Following the same pattern, the individual ages of students are given in Figure 4 (top, right) and the corresponding aggregated AA values are visualized using shades of green.

It is important to note that even representing set attributes without uncertainty ($U = 0$) using a simple visual variable like color value has its challenges, since we cannot easily represent the aggregated value of the whole sets as well as that of the subsets created by the relationships between them. Hence, our examples represent aggregated information for the subsets; additional visual variables or supplementary visualizations would be required if the aggregates of the entire sets need to be communicated as well.

Visualizing set attributes with undefined uncertainty In the case of undefined uncertainty ($U > 0$), although we have a residential status and an age associated with each student, we know that some of this information is incorrect. Because we do not know for which students this might be the case, we have to assume there is uncertainty throughout. Figure 4 (middle row) shows how such general uncertainty can be added to

This article has been accepted for publication in IEEE Computer Graphics and Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCG.2023.3300441

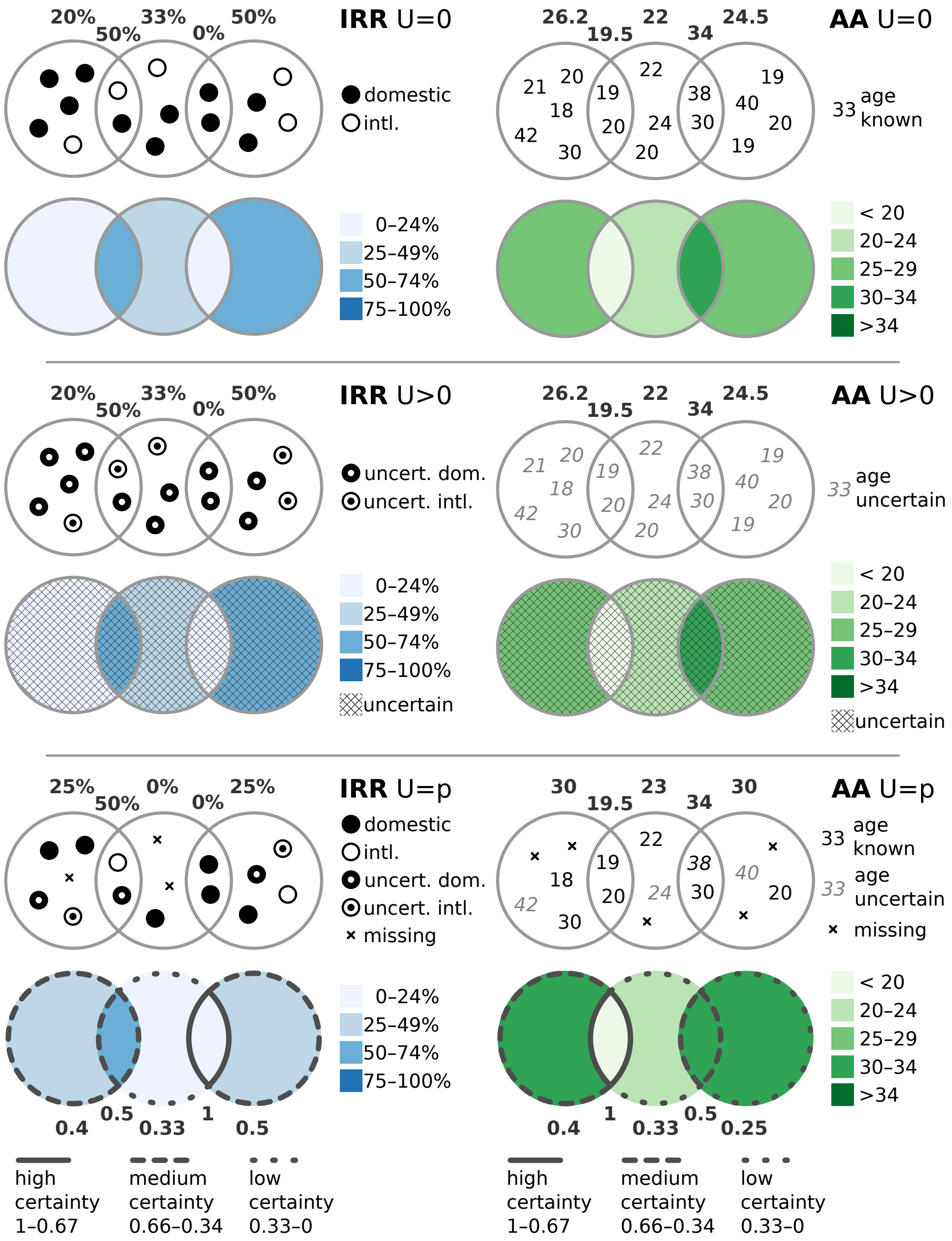

FIGURE 4. Visualizing set attributes IRR (left) and AA (right) when $U = 0$ (top), $U > 0$ (middle), and $U = p$ (bottom). 🅭🅯

the visualization using texture, more precisely, a hash pattern overlaid on the colored set representations. However, a global overlay like this may interfere with the perception of colors. As an alternative, such an overall uncertainty could be noted as a general disclaimer statement in the caption or the text (rather than being explicitly added to the visualization by the use of additional visual variables).

Visualizing set attributes with defined uncertainty In the case of defined uncertainty ($U = p$), we know which courses each student is enrolled in, and we know the residential status of some students. The uncertainty lies in the fact that there are some students for whom we do not know their residential status or age (missing values), and/or there are some students for which we know that the information given may be incorrect (uncertain values).

In this case, the visualization designer has to make choices relating to how both the aggregated value and its (un)certainty are calculated. The calculation of the aggregated value can:

1) ignore the elements with missing values as well as those with uncertain value, or
2) ignore the elements with missing values, and use the given values for the uncertain elements.

The certainty can be calculated as:

3) the proportion of elements for which the value is certainly known, in relation to the total number of elements in the (sub)set; or
4) the proportion of elements for which the value is certainly known, in relation to the total number of elements for which values have been given (i.e. ignoring the elements with missing values).

In the examples in Figure 4 (bottom), we visualize IRR (left) and AA (right) using options 2 and 3. That is, missing values are ignored when computing the aggregated IRR or AA value, but are taken into account when computing the certainty of the aggregated value. Note that the visualization in the bottom row uses the outline dash pattern to indicate certainty (not uncertainty).

Possible use of other visual variables In the examples above, we chose to use variations in color value, line dashes, and texture to represent aggregated data values and uncertainty. While other visual variables could be used instead (for example, color hue, line weight), we argue that using size variation to represent uncertainty is not helpful, even though it is commonly used in other data visualizations. Despite evidence that simple error bars in a bar chart are not as easy to interpret as they may appear [15], uncertainty is often depicted with error bars or grayed out areas which indicate proportional uncertainty corresponding to the size differences as in **Figure 5** (a). It is not recommended though to apply such 'size-aware' principles to common set visualizations as in Figures 5 (b) and (c), which focus on depicting set membership. The size of the graphical objects simply has no meaning, and hence, inappropriate inferences could be made about the extent of the uncertainty.

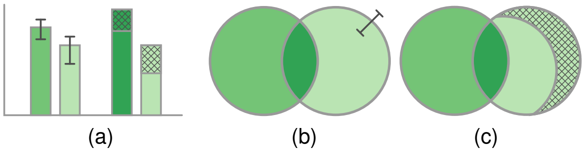

FIGURE 5. Representing uncertainty with markers indicating size variation as in (a) is **not recommended** for set visualizations where the area of the set does not relate to the set attribute value as in (b) and (c). 🅭🅯

Adding supplementary information to depict uncertainty The examples above extend traditional Venn diagrams for uncertainty visualization by varying well-established visual variables and existing graphical elements. An alternative idea is to add supplementary graphical elements to the original diagram. For example, in a bipartite node-link diagram, sets can be represented as small pie charts indicating the proportion of international students and uncertainty (see **Figure 6**, top). In this way, any set aggregate attribute (and its uncertainty) can be added to the (certain) set representation. Still, the association between the supplementary information and the set (or subset) it refers to must be made clear, using, for example, visual cues, like proximity or links. However, this visualization does not explicitly show (un)certainties of the intersections. Should this be required, one may use a matrix representation as in Figure 6 (bottom), which shows the statistics for each intersection and the overall sets.

Uncertain Element Attributes

Finally, we present exemplary design suggestions for visualizing uncertain element attributes (third column in Table 2). Continuing our scenario, we now wish to depict the age distribution (element attribute) of the students (elements) enrolled in our courses (sets). For the defined uncertainty case ($U = p$), we know the value of an element attribute, and we also know the nature and value of the uncertainty. For example, for a given course with twenty enrollments, fifteen students

This article has been accepted for publication in IEEE Computer Graphics and Applications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCG.2023.3300441

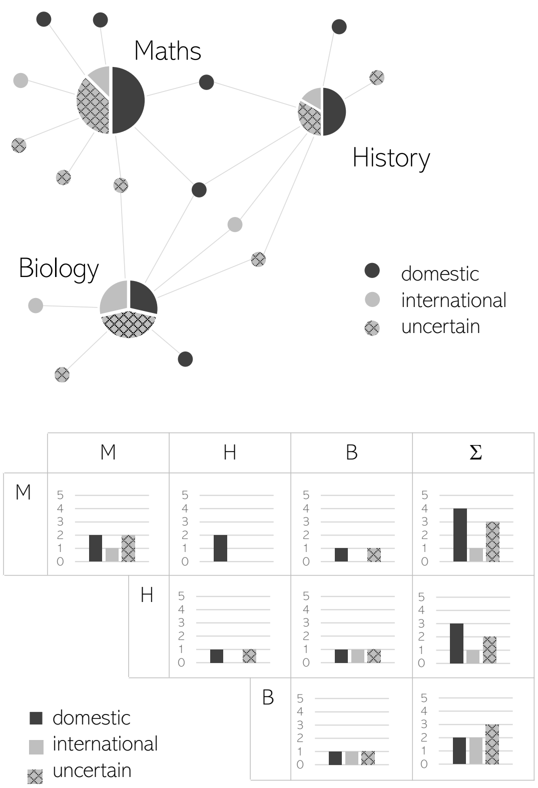

FIGURE 6. Set attributes with defined uncertainty visualized in a bipartite node-link diagram (top) and as a matrix of bar charts (bottom).

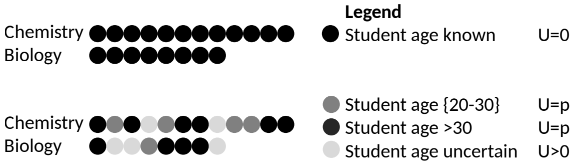

FIGURE 7. Top row ($U = 0$): Two courses (sets) have twenty enrolled students (elements) where all individual ages (element attribute) are known (i.e., black point symbols). Bottom row ($U > 0$ and $U = p$): Two courses (sets) have twenty enrolled students where the degree of uncertainty in student ages varies from (i) completely unknown, that is, point symbol denoted with the lightest shade of gray, to (ii) mostly unknown, (i.e., age above 30 yrs.) shown with medium gray point symbols, and to (iii) somewhat unknown (i.e., within a given age range 20-30 yrs.), assigned dark gray point symbols.

have already enrolled and we have been given their birth year from the university registrations office. For five students who are also interested in taking the course, we only know their age range at this point, as they ticked the age range box, i.e., 20-30 yrs. in the course enrollment questionnaire. The uncertainty lies in the fact that there are some students for whom we do not know their actual age with certainty, but we do have information on their age within a given range and above a certain threshold.

Following MacEachren et al. [11], we again employ the intuitively understood visual variable of color value (i.e., varying shades of gray) to denote variation of uncertainty in set element attributes (see **Figure 7**). According to MacEachren et al.'s studies, other visual variables such as opacity, fuzziness, texture, and arrangement could have been used as well. The same visual variables to show uncertainty in set elements may also be used to denote uncertainty of set membership or set attributes, as discussed before.

In the undefined uncertainty case ($U > 0$), we know the value of an element attribute, but we do not know the type and value of the uncertainty. For our example, we may again say that for a given course with twenty seats, fifteen students have already enrolled, and we know their birth year. But for the five students who are also interested in taking the course, we do not have any age information. In this case, following the logic in Figure 7 (bottom), we denote the uncertainty on some of the students' age with the lightest shade of gray.

For the easy case of certainty ($U = 0$), we know the value of an element attribute with certainty. For our data example with twenty enrolled students we know their age with certainty. In this case, we can depict the age attribute of our elements with any of the commonly known multivariate visualization methods [8], [9] and principles that are appropriate for ratio level data, i.e., dots, bar charts, box plots with commonly used visual variables.

Discussion

Our investigations of uncertainty in set visualization has resulted in a conceptual framework that links types of uncertainties (i.e., $U = 0$, $U > 0$, and $U = p$) and data characteristics in sets (i.e., set memberships, set attributes, and element attributes), as summarized in Table 2. The framework also names the different subfields of visualization that can inform successful visual communication of set uncertainties, and together with the literature provided in Table 1, can be a useful starting point for further research. We have addressed the relevant cells in our framework and outlined several design alternatives for set visualization drawing upon existing visualization research.

A further outcome of our work is the identification of challenges that warrant further detailed investigations.

Evaluation. Our framework implicitly outlines a road map to empirically study the appropriateness and usefulness of the various ways of depicting uncertainty in sets. This must happen considering the target audience's background and training (e.g., graphicacy, domain expertise, disciplinary context), which influence the interpretation of uncertainty visualizations. For example, depicting the uncertainty of a set element attribute in a Venn diagram with a lighter gray value of the element mark could be interpreted as uncertainty of set membership of that element. The evaluation of uncertainty visualizations, as a whole, is still in its infancy, with a range of important questions to be tackled.

Perception of uncertainty depictions. In general, the depiction of uncertainty U needs to be balanced with respect to the depiction of the actual set data D. Carelessly adding an uncertainty depiction to a data visualization can lead to clutter or overemphasis. For example, in our bipartite node-link diagrams from Figure 3 (a), uncertain set memberships require links between all uncertain set elements and all possible sets, which easily leads to visual clutter and gives much greater saliency to the uncertain information rather than to the certain information. In Figure 3 (b), we employed small link fans to reduce visual clutter, but other, more general alternatives should be explored and evaluated for their effectiveness.

Task- and context-specific challenges. The visualization community is well aware that user tasks and the application context are essential ingredients in designing expressive, effective, and efficient visualization solutions [9], [16], [17]. While specific visualization tasks were considered in previous work on general set visualization [1], our work exemplified solutions prototypically. We leave a comprehensive task- and context-specific exploration of the design space for future research. Ideally, evaluations with real-world application scenarios including tasks of varying complexity beyond our small data example would be useful. For example, ensemble forecasting models or gene-to-phenotype mapping would serve as exciting test scenarios.

Uncertainty propagation and missing data. Uncertainty is not a static concept and interdependency might occur due to data processing chains. For example, the uncertainty of a set attribute may be directly dependent on set membership and set element attribute uncertainties. Consequently, uncertainty propagation should be specifically addressed in a future version of the framework; the communication of uncertainty propagation in a set visualization will pose interesting depiction challenges. For simplicity, we also treated missing and uncertain data equally in our framework. Future work should extend our proposal to disentangle these two concepts further.

Temporal and spatial uncertainty. While we considered uncertainty in set visualization, we mostly ignored the spatial and temporal context of sets, which poses additional challenges for their visualization [18]. Similarly to what can be said about the dynamics of uncertainty propagation, temporal uncertainty itself also relates to time-varying changes to uncertainty, including uncertainty states with respect to points in time, perdurance, and the evolution of uncertainty in unfolding events [19]. Likewise, uncertainty in a spatial frame of reference requires special consideration. When several domains with data uncertainty need to be understood in context, more scalable designs are needed to balance the visualization according to the needs of users [20]. When visualizing sets and set elements that represent spatial and temporal data, consideration will need to be given to the particular nature of spatial and temporal uncertainty.

Overall, there is an extensive design space to be explored. With this contribution, we call to action to further extend the outlined framework, to systematically evaluate the already offered design solutions, and to revise the framework with empirical evidence where necessary.

Conclusion

We set out to devise a conceptual framework to describe how uncertainty in set data could be visualized by first finding answers to still open research questions: (i) Which aspects of set type data could be affected by uncertainty, and (ii) Which characteristics of uncertainty could influence the visualization design. Based on this framework, we then systematically discussed set visualization examples with integrated uncertainty information. We also provided a set of open challenges in the hope that these may inspire future research on uncertainty in set visualization.

Recommendations. We emphasize two high-level recommendations in this call to action that we identified early on during our work:

- *Data first, uncertainty second*: It is practical to start with the visual encoding of the certain data, followed by the encoding of the uncertain aspects.
- *Be aware of visual misinterpretations by the users*: Test your designs with users, as interpretation and understanding of uncertainty are likely

challenging for many users; visual solutions might be misread by the target audience. Ample labeling and adding legends and explanations accompanying the uncertainty visualization will help to guide the users. In some cases, we even found, it may be more effective to communicate uncertain information by non-visual means.

ACKNOWLEDGMENT

This work was initiated by Dagstuhl Seminar 22462 on Set Visualization and Uncertainty.

References

[1] B. Alsallakh *et al.*, "The state-of-the-art of set visualization," *Comput. Graph. Forum*, vol. 35, no. 1, 2016. DOI: 10.1111/cgf.12722.

[2] S. Bleisch *et al.*, "Set Visualization and Uncertainty (Dagstuhl Seminar 22462)," *Dagstuhl Reports*, vol. 12, no. 11, pp. 66–95, 2023. DOI: 10.4230/DagRep.12.11.66.

[3] G. Cantor, "Beiträge zur Begründung der transfiniten Mengenlehre," *Mathematische Annalen*, vol. 46, no. 4, 1895. DOI: 10.1007/BF02124929.

[4] D. V. Lindley, *Understanding Uncertainty*, revised edition. Wiley, 2013. DOI: 10.1002/9781118650158.

[5] A. Jena *et al.*, "Uncertainty visualisation: An interactive visual survey," in *Pacific Visualization Symposium*, IEEE, 2020. DOI: 10.1109/PacificVis48177.2020.1014.

[6] UC Santa Barbara: National Center for Geographic Information and Analysis, *Uncertainty in Geospatial Information Representation, Analysis, and Decision Support*, https://escholarship.org/uc/item/14v6587w, 1998.

[7] D. Sacha *et al.*, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, 2016. DOI: 10.1109/TVCG.2015.2467591.

[8] M. O. Ward *et al.*, *Interactive Data Visualization: Foundations, Techniques, and Applications*, 2nd ed. A K Peters/CRC Press, 2015, ISBN: 9781482257373. DOI: 10.1201/b18379.

[9] C. Tominski and H. Schumann, *Interactive Visual Data Analysis* (AK Peters Visualization Series). CRC Press, 2020, ISBN: 9781498753982. DOI: 10.1201/9781315152707.

[10] G. L. Kindlmann and C. E. Scheidegger, "An algebraic process for visualization design," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, 2014. DOI: 10.1109/TVCG.2014.2346325.

[11] A. M. MacEachren *et al.*, "Visual semiotics & uncertainty visualization: An empirical study," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, 2012. DOI: 10.1109/TVCG.2012.279.

[12] N. Gershon, "Visualization of an imperfect world," *IEEE Computer Graphics and Applications*, vol. 18, no. 4, 1998. DOI: 10.1109/38.689662.

[13] H. Schulz *et al.*, "The design space of implicit hierarchy visualization: A survey," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 4, 2011. DOI: 10.1109/TVCG.2010.79.

[14] B. Alper *et al.*, "Weighted graph comparison techniques for brain connectivity analysis," in *2013 ACM SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Paris, France, April 27 - May 2, 2013*, W. E. Mackay *et al.*, Eds., ACM, 2013. DOI: 10.1145/2470654.2470724.

[15] M. Correll and M. Gleicher, "Error bars considered harmful: Exploring alternate encodings for mean and error," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, 2014. DOI: 10.1109/TVCG.2014.2346298.

[16] H.-J. Schulz *et al.*, "A Design Space of Visualization Tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, 2013. DOI: 10.1109/TVCG.2013.120.

[17] S. Miksch and W. Aigner, "A Matter of Time: Applying a Data-Users-Tasks Design Triangle to Visual Analytics of Time-Oriented Data," *Computers & Graphics: Special Section on Visual Analytics*, vol. 38, 2014. DOI: 10.1016/j.cag.2013.11.002.

[18] S. I. Fabrikant *et al.*, "Visual analytics for sets over time and space (dagstuhl seminar 19192)," *Dagstuhl Reports*, vol. 9, no. 5, 2019. DOI: 10.4230/DagRep.9.5.31.

[19] T. Gschwandtner *et al.*, "Visual encodings of temporal uncertainty: A comparative user study," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, 2016. DOI: 10.1109/TVCG.2015.2467752.

[20] S. Dübel *et al.*, "Visualizing 3D Terrain, Geo-Spatial Data, and Uncertainty," *Informatics*, vol. 4, no. 1, 2017. DOI: 10.3390/informatics4010006.

Christian Tominski, is a scientist at the Institute for Visual & Analytic Computing at the University of Rostock, Germany. His main research interests are in visualization of and interaction with data. He is particularly interested in effective and efficient techniques for interactively exploring and editing complex data. Christian has published numerous papers on new visualiza-

tion and interaction techniques. He co-authored three books, including a book on the visualization of time-oriented data in 2011, a book focusing on interaction for visualization in 2015, and a more general book about interactive visual data analysis in 2020.

Michael Behrisch, is an Assistant Professor for Visual Analytics at the Utrecht University, Netherlands. In his research, he focuses on novel visual interactive techniques, algorithmic approaches, and integrated visual analytics systems to support users in navigating and exploring large complex datasets. One central research objective is to automatically assess the interestingness of visualizations and show only potentially important views from a large exploration space to reduce the users' cognitive load.

Susanne Bleisch, is a Professor of Geovisualization and Visual Analytics at the Institute of Geomatics, FHNW University of Applied Sciences and Arts Northwestern Switzerland. Her main research interests are in geoinformation visualization, currently focusing on bridging the gap between exploratory and communicative geovisualizations, as well as suitable visual analytics processes in interdisciplinary research collaborations such as studying group processes or climate change factors and implications. Susanne holds a PhD in Geographic Information Science from City University London, UK. Email her at susanne.bleisch@fhnw.ch.

Sara Irina Fabrikant, is a Professor of Geography, leading the Geographic Information Visualization and Analysis (GIVA) group at the GIScience Center of the Geography Department at the University of Zurich, Switzerland. Her research and teaching interests lie in geographic information visualization and geovisual analytics, GIScience and cognition, graphical user interface design and evaluation, including dynamic cartography. She is currently an elected member of the Swiss Science Council (SCC) and a Co-Initiator and past Co-Director of the UZH Digital Society Initiative (DSI). Other academic service included Vice-President of the International Cartographic Association (ICA) 2015-2019, and program committee co-chairing of the international conferences AGILE 2008, GIScience 2010, and COSIT 2015 & 2022.

Eva Mayr, is a postdoctoral researcher at the University for Continuing Education in Krems, Austria. Her main research interests are the cognitive processes during interaction with information visualizations, in particular in "casual", informal learning settings and in the and digital humanities. Eva holds a PhD in applied cognitive and media psychology from the University of Tübingen, Germany. Contact her at eva.mayr@donau-uni.ac.at.

Silvia Miksch, is University Professor and head of the Visual Analytics research unit (CVAST) at the Institute of Visual Computing and Human-Centered Technology, TU Wien. She served as paper co-chair of several conferences including IEEE VAST 2010, 2011, 2020, and VIS Overall Papers Chair (IEEE VIS 2021) as well as EuroVis 2012 and on the editorial board of several journals including IEEE TVCG and Computer Graphics Forum. She acts/acted in various strategic committees, such as the chair of the EuroVis steering committee and the VIS Executive Committee. In 2020 she was inducted into The IEEE Visualization Academy. Her main research interests are Visualization/Visual Analytics (particularly Focus+Context and Interaction), Space, and Time.

Helen Purchase, is currently Professor in the Faculty of Information Technology at Monash University, Australia. Her early research career focused on investigating the principles underlying graph layout algorithms from a human comprehension perspective. From this she extended into wider empirical information visualization research, including image complexity, graph animation, multivariate data, causality and social networks. Her book “Experimental human-computer interaction: a practical guide with visual examples” (CUP, 2012), gathers together her experiences of conducting a range of visual experiments. She has also authored several publications in the area of educational technology.