

3 DESIGN FACTORS FOR SUMMARY VISUALIZATION

Visual analytics systems help users navigate large and complex datasets. These datasets often have too much data or too many dimensions to display in one view, requiring designers to engineer systems to first *summarize* available data and then visualize the results. The resulting *summary visualizations* help orient analysts by describing high-level information about the dataset, guiding analyses of particular features, and providing a means for navigating to important subsets of data. Designers and researchers have numerous techniques and design choices for constructing these summary visualizations, but little systematic guidance for reasoning about the trade-offs of different design decisions and their impact on the resulting analyses.

In this chapter¹, we survey summarization in visual analytics, evaluating the relationships between use, analytic affordances, and data summarization methods. We aim to understand, recognize, and characterize limitations in current design choices for effectively summarizing data for visual exploration, focusing on factors such as analysis tasks, data types, and data characteristics. We provide an abstraction of existing summarization methods to support different analyses, and use this survey to propose an initial design space for data summarization for visual analytics. Understanding common links between tasks, data, and techniques used to summarize data for analysis will help guide new tools and opportunities for innovation in visual analytics systems, and for summary visualizations in general.

Summarization in visual analytics serves two primary purposes: to compress the dataset to fit in the available screen-space and to reduce visual complexity to make visualizations easier to interpret. Examples of summary visualizations include histograms that aggregate data across a selected dimension, dimensionally-reduced scatterplots that project high-dimensional data into a lower-dimensional space (see §??), and actor-network diagrams that summarize relationships between entities captured in a text corpora. Summarization is an essential component in most visual analytics tools—we found that more than half of

¹This chapter is part of Alper Sarikaya’s PhD thesis, available at <http://cs.wisc.edu/~sarikaya/research/thesis/sarikaya-thesis.pdf>.

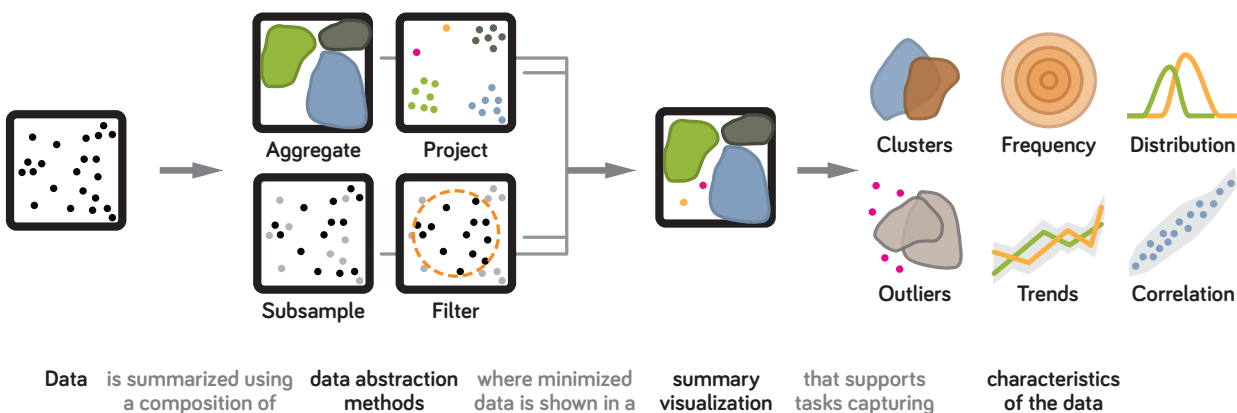


Figure 3.1: A schematic of a generalized process for visual analytics with data summarization. A dataset (left) is reduced using data summarization techniques (center), comprised of four basic methods (aggregate, project, subsample, filter), and is presented visually to support judgments of high-level data characteristics (right). Both the summarization and visual presentation are factors that influence the efficacy of summary visualization to enable viewers to make high-level judgments.

papers surveyed contained a summary visualization. The ubiquity of summarization and its impact on data analytics tools means that designers need a better understanding of the factors that lead to effective summarization to guide the design of effective visualization tools. We construct a preliminary design space of summarization for visual analysis (Figure 3.1) that allows designers to reason how factors are involved in affecting the resulting utility of a summary visualization.

We systematically survey the visualization literature to observe relationships between factors in this design space using quantitative content analysis (QCA) [see Riffe et al., 1998] to analyze summarization in IEEE VIS and EuroVis papers from 2009 to 2015. Our approach quantifies the relationship between analysis tasks, properties of the data, and summarization choices to identify design themes in summary visualizations. Our goal herein is to confirm that the list of summarization methods is sufficient to capture all methods of summarization and also to understand how the method of summarization affects the affordances of a resulting visualization. We use our design space analysis to identify common *themes* observed in summary visualization design. These themes indicate common patterns in how designers *use* summarization to guide analysis, a preference for task *specificity*, the existence of common *design patterns*, and a bias towards certain

design choices for specific types of *data*. These themes highlight key considerations for data summarization, identifying *challenges* in current practices and *opportunities* for new and innovative thinking around summarization in visual analytics.

Contributions: We conduct a systematic survey using QCA to characterize summarization in visual analytics as a function of purpose, data summarization method, analysis task, and data type. In doing so, we make the following contributions:

- A taxonomy of data summarization techniques used in summary visualization (§3.1–3.2),
- A formal survey and analysis of summary design in exploratory tools (§3.3), and
- A description of challenges in summary design (Table 3.2) and potential opportunities for innovation (§3.3–3.4) grounded in existing practice.

This chapter provides a foundation for systematically reasoning about summarization in current and future tools and identifies gaps in our general understanding of summarization in visualization design.

3.1 Background

Visual analytics systems often provide summaries that analysts use to navigate, sift, and winnow through data to create a concise and focused representation of the underlying dataset [Shneiderman, 1996]. A summary within a visual analytics tool communicates properties of a dataset by explicitly using fewer marks than there are datapoints. For example, a scatterplot with points aggregated using KDE could constitute a summary, whereas ‘zoomed-out’ representations of a dataset, such as a standard scatterplot or parallel coordinates plot with many thousands of elements but no data minimization would not qualify as a summary. While individual points may be difficult to distinguish in such zoomed-out representations due to factors such as overdraw (see Fekete and Plaisant [2002] for technical issues, Cui et al. [2006] for understandability issues), such visualizations do not procedurally summarize data. Instead, we focus on methods that explicitly summarize data

(a strategy most overview solutions employ) and analyze how the ways data is summarized and presented affect high-level judgments of the visualized data.

In this work, we consider a **summary visualization** the result of an explicit set of summarization decisions made by the designer, together with the reduced data and the visual representation. Throughout this dissertation, the role of summaries is to convey a “gist” about global and high-level properties of a dataset, as discussed in ???. Design of these summaries should ideally support the needs of the analyst or audience, but the data type, the method of reducing data, and the anticipated use can all affect the resulting design. We draw on these prior characterizations of overview and our own observations to organize the design space of how summaries are used in visualization to generate a grounded codebook for QCA (§3.2.2). From this literature, we propose a design space characterized by *purpose*, *summarization method*, and *task* that guides the design of effective summarization, and discuss these organizations in detail. By basing codes on previous work, we can use these organizations in our effort to observe relationships between these factors for effective summary design.

3.1.1 Data

The type of *data* affects how data is summarized and what global, high-level features an analyst can extract from a summary. Taking note of previous approaches, we can observe how different techniques can affect a resulting visualization based on the hierarchical organization for data, data with multiple dimensions, and dealing with large amounts of data. Hierarchical data can be summarized by visualizing data at different levels within that hierarchy. Elmqvist and Fekete [2010] survey how aggregation techniques, in particular, can reduce the amount and complexity of visualized data. They demonstrate how hierarchical aggregation can be applied to conventional visualization types, even for non-hierarchical data types, and also provide guidelines for effective navigation within hierarchically aggregated visualizations. Elmqvist and Fekete provide a guideline of *visual summary* (G2) that “visual aggregates should convey information about the underlying data.” We consider the effect that aggregation can have on a resulting summary, and also consider how a summary is affected by a broader set of transformation and organizations

of non-hierarchical data.

While other types of data have unique challenges for summarization, a common theme in summarization is dealing with high-dimensional and spatial data. Both Keim and Kriegel [1996] and Keim [2002] have explored visualization techniques for exploring databases, where visual interfaces summarize datapoints and their attributes using overview first, enabling the viewer to explore large amounts of data. Kehrer and Hauser [2013] have surveyed the high-level design and intents of visual analytics overview approaches, exclusively for multifaceted scientific data. They identify many techniques in their survey that lead to summaries for particular types of data, but do not directly draw conclusions about the affordances of different techniques and the cross-applicability of summary designs for different data domains. Leung and Apperley [1993] provide a framework for evaluating visualizations where there is too much data to display each datapoint clearly. This framework helps designers evaluate visual and computational representations of summaries based on their *effectiveness*, *expressiveness*, and *efficiency*; however, it provides no guidance as to how representations might be designed with these qualities in mind.

3.1.2 Purpose

The *purpose* of a visualization describes its intended use. We anticipate that the intended purpose of a summary directly informs effective design. Bertin [1983] presents purpose as a dichotomy: the visualization either communicates previously understood information (presentation-oriented visualization) or supports information processing to address new questions (exploratory visualization). Schulz et al. [2013] refines this division to consider the goals of an analysis: *exploratory*, *confirmatory*, and *presentation*.

We hypothesize that summaries for presentation emphasize specific data characteristics more often than exploratory summaries, and that the intended purpose of a summary (exploratory, confirmatory, and presentation) can inform effective data summarization. This division aligns with recent design guidelines proposed for presentation-oriented visualizations Kosara [2016], advising specificity and compactness over generalizability.

3.1.3 Data Summarization Methods

Methods of data summarization can reduce the scale and complexity of a dataset for display in a summary. We specifically consider methods that summarize data while simultaneously providing a faithful representation of the underlying dataset. Prior work suggests methods of re-organizing data for visualizations—for example, Card and Mackinlay [1997] argue that a small set of functions can be used to process data for visualization: filtering, sorting, multidimensional scaling, and selection by slider. Ellis and Dix [2007] taxonomize clutter reduction techniques for visualizations. Three of these techniques (sampling, filtering, and clustering) explicitly reduce data—however, their work considers summaries only as a means of reducing visual clutter in data space rather than emphasizing particular characteristics of the data.

We derive and propose four methods of data summarization from our observation and reconciliation of the literature: **aggregation**, **subsampling**, **filtering**, and **projection** (see §3.2.2 and §3.3.2 for details). These four categories capture the variety of methods that reduce data for display, and we anticipate that the method used will influence the types of judgments that viewers can make from the data visualized in the resulting visual summary (e.g., exploration of subsampling by Bertini and Santucci [2006]). Understanding the relevant tasks and judgments viewers will perform will help to connect these methods to their support in summary visualizations.

3.1.4 Tasks

The summarization methods used to summarize a dataset directly influence the analysis tasks supported by a derived summary. As an example of this relationship, using kernel density estimation to spatially aggregate values in a scatterplot helps viewers find dense clusters, but obscures local outliers. Our goal in this work is to collect a representative set of overview-level tasks, which capture the high-level information of a dataset. To do so, we look at the multitude of task taxonomies to generate a representative set of analysis tasks.

Task taxonomies have looked at how viewers obtain information from displays (see Andrienko and Andrienko [2006] and Shneiderman [1996] for canonical examples). Amar

et al. [2005] identify a series of low-level tasks used to answer specific queries about a dataset. Ji Soo Yi et al. [2007] outline tasks that analysts perform to guide data interaction and exploration. Zhou and Feiner [1998] explore high-level presentation intents and visual discourse tasks, including “summarize” tasks such as *associate*, *compare*, *distinguish*, and *rank*.

More recent work considers how tasks can drive visualization design (see Rind et al. [2016] for a synthesis of this space). For example, Brehmer and Munzner [2013] looks at how tasks can be abstracted and expressed to support design across different application domains. Pike et al. [2009] look the mutual relationship between user tasks and interaction design. Schulz et al. [2013] describe how designers can reason about tasks using “5 W’s” (and one “H”): why is a task pursued (a task’s *goal*), how is a task carried out (a task’s *means*), what does a task seek (the *target* and *cardinality* of objects), when is a task performed, and who carries out the task? Schulz et al.’s hierarchical synthesis of high-level tasks provides a representative analytic organization that we utilize in designing our codes for the survey. We additionally consider how these questions manifest in existing summaries to identify how tasks might guide effective summary design.

3.2 Methodology

We survey summary visualizations in the research literature to discover patterns in the use, design, and analytic affordances of visualizations using summarization. We are particularly interested in how the methods of data summarization are related to the use and the information communicated by the summary visualizations. To discover these patterns, we use four research questions to ground our exploration of this space. These questions concern the validity of our organization, how different types of summarization affect the resulting affordances communicated by the visualization, and how data and use affects how summarization methods are utilized. In detail, our questions are as follows:

Q1 Do the four proposed summarization methods cover the range of summarization performed for summary visualization design?

- Q2** Does the method of data minimization affect how a resulting visualization can be used?
- Q3** How does the use of summary visualization affect decisions of summary?
- Q4** Does the type of data affect what types of minimization and affordances are appropriate?

To gather the necessary data to address these research questions, we use quantitative content analysis (QCA) [Riffe et al., 1998] that helps to quantify attributes about visual artifacts. This methodology has the advantages of quantitative evaluations (using statistical methods), and can break summary visualizations down into digestible factors to later identify trends between the factors. This in contrast to grounded theory, which could build up concepts from qualitative exploration, but would likely be heavily biased by the sample of summary visualizations chosen. Instead, QCA depends on a static codebook to quantify attributes, evaluated by the coder. To promote ecological validity, we derive the codebook chiefly from existing visualization taxonomies (see §3.2.2), and use the results of the QCA process to validate our organization of summarization methods (**S1**). This methodology confirms the organizations proposed and the data generated through its use highlights trends in summary visualization.

Two data visualization researchers served as the coders for this survey. After a preliminary coding of ten papers, the two coders iterated on codebook definitions to clarify lingering ambiguities and to address emerging concerns regarding measure validity. Of the 180 evaluated manuscripts, 54 randomly-selected papers (30%) were redundantly coded for validation—the Cohen’s kappa measurement for intercoder reliability found substantial agreement between coders ($\kappa = 0.71$, 86% overall agreement). Section 3.3 presents the result of this process, and identifies themes arising from our analysis.

3.2.1 Corpus Construction

To construct a corpus of summary visualizations, we use the data visualization research literature as a collection of peer-reviewed and validated collection of visual analytic systems. This corpus is especially attractive due to the discussion of analysis scenarios in prose in

close relation to the visual presentation of the summary visualization. To compose the corpus, we collected papers from the EuroVis, InfoVis, SciVis/Vis, and VAST conferences from 2009 to 2015 (1,158 papers). As coding each paper through a comprehensive census is untractable, we constructed a representative corpus of papers through simple random sampling from this set, as commonly used in traditional content analysis of large corpora (*cf.* Brubaker et al. [2012]). This resulted in a set of 180 coded papers (48 EuroVis, 53 InfoVis, 48 SciVis, and 31 VAST papers). Each paper was first coded for whether or not they included an implementation or technique that included a summary. Papers containing summaries were considered the artifact (figures and prose), and coded according to the presented protocol. We excluded theory, taxonomy, survey, toolkit, and evaluation papers as the focus of these papers was not on a single, specific visualization design, making application of the codes too subjective as we had no explicit evidence of the designers' intents.

Using examples from the visualization research community allows us to focus on designs whose quality, effectiveness, and utility have been reviewed by external experts in the field and that are tailored for a wide variety of applications. Although visualization designs are also found in conferences outside of the immediate visualization community (e.g., NIPS, VLDB, KDD), specific visualization contributions in these fields are relatively rare and unlikely to appear in a random sample. Further, visualization research papers emphasize novel contributions and techniques that represent the state-of-the-art in visualization specifically, and these papers represent a vetted corpus of summary visualizations that contain explicit rationales for their design discussed within the article, increasing the validity of our coding practices. However, the choice of this corpus biases the results of this study toward exploratory visualizations that are used by researchers or domain experts (not the general public), which we discuss in Section 3.4.

3.2.2 Coding Protocol

Each example in our 180 paper sample was labeled using a predetermined coding protocol designed to characterize four factors of summary visualization design: the visualization *purpose*, the *data* being used, the *data summarization* methods employed, and the *tasks* sup-

ported by the resulting summary visualization. We constructed our codebook by collecting and abstracting categories across 15 existing typologies describing data, purpose, task, and summarization in visualization. Table 3.1 summarizes the coding scheme used in the survey. As mentioned above, a set of 10 papers were used in iteration to revise our codebook for clarity. The final codes are as follows:

Purpose: We capture the purpose, or *goal*, of each visualization by considering whether it supports exploratory (undirected search), confirmatory (directed search) or presentation-oriented (exhibiting analysis results) analyses [Bertin, 1983, Schulz et al., 2013]. These codes describe the high-level intent of the summarization and are treated as three binary (present/absent) codes.

Data: We coded for data type using Shneiderman’s data type taxonomy [1996], with one-dimensional and temporal data collectively coded as *sequence data*, encompassing one-dimensional data on a common axis (e.g., temporal, genomic, or ranked data). While we considered data size as a potential code due to the utility of summarization for large, multidimensional datasets, most systems did not provide specific information about the number of datapoints and dimensions tested and designed their methods for use with more than one dataset. For these reasons, coding for data size would have required significant assumption and extrapolation on the part of the coders, resulting in limited validity. We therefore did not consider data size in our analysis, but it is an important consideration for future work.

Data Summarization: We used our own observations coupled with categories from Schulz et al.’s *reorganization* task [2013], the visualization design space [Card and Mackinlay, 1997], methods for clutter reduction [Ellis and Dix, 2007], and methods of hierarchical abstraction [Elmqvist and Fekete, 2010] to propose a set of four data summarization methods used to reduce data for display:

Aggregation Computationally combining multiple elements (e.g., hierarchical aggregation [Elmqvist and Fekete, 2010]),

Subsampling Subsetting elements based on a stochastic selection of the data (e.g., random subsampling [Bertini and Santucci, 2006]),

Category	Subcategory	Code
Purpose		Exploratory
		Confirmatory
		Presentation
Data Summarization		Aggregation
		Subsampling
		Filtering
		Projection
Task	Means: Navigation	Browsing
		Searching
		Elaborating
		Summarizing
	Means: Relation	Comparison
		Variations
		Relation-seeking
	Characteristics: High-level	Trends
		Outliers
		Clusters
Frequency		
Distribution		
		Correlation
Data		Data type
		Specific data
Other		Additional observations ↔

Table 3.1: Two coders labeled 180 examples from the visualization literature. The coders first identified whether a visual summary was present and then coded each summary according to 22 attributes describing the summary’s purpose, data summarization, and supported tasks (§3.2.2).

- Filtering** Subsetting elements based on properties of the data (e.g., selecting a representative set [Ellis and Dix, 2007]), and
- Projection** Mapping data elements to a set of reduced or derived dimensions (e.g., principal component analysis [Jolliffe, 2002]).

We hypothesize that the data summarization methods used to construct a summary visualization heavily affect the affordances that the summary supports (**S2**). The method of data summarization was coded using a combination of four binary codes (present or absent for each of the above categories).

Task: Our task codes were drawn from the *means* and *characteristics* of Schulz et al.'s taxonomy [2013]. We chose this taxonomy as a general guide over other taxonomies (e.g., Amar et al. [2005], Brehmer and Munzner [2013], Casner [1991], Klein et al. [2006], Roth and Mattis [1990], Springmeyer et al. [1992], Ji Soo Yi et al. [2007], Zhou and Feiner [1998]) as it comprehensively reflected most categories presented in other taxonomies.

Means of navigation describe how summary visualizations support further analysis. These tasks coincide with Springmeyer et al.'s concepts of maneuvering [1992] and Casner's perceptual search operators [1991], and later used in Amar et al. [2005] and Ji Soo Yi et al. [2007] as intent in interaction, Zhou and Feiner [1998] in their modes of "enabling", and Heer and Shneiderman's "interactive dynamics" [2012].

Means of object-object relations describe information foraging tasks, including *comparison* (seeking similarities; see Gleicher et al. [2011], Ji Soo Yi et al. [2007]), *variation* (seeking dissimilarities; see Roth and Mattis [1990], Zhou and Feiner [1998]), *discrepancies* (seeking outliers Roth and Mattis [1990], Zhou and Feiner [1998]), and *relation-seeking* (seeking one of the aforementioned relations for individual objects; see Casner [1991], Heer and Shneiderman [2012]). While we additionally coded for *discrepancy*, this code was removed from our analysis due to poor agreement between coders.

High-level characteristics from Schulz et al. [2013] are used to code specific judgments of high-level data characteristics afforded by summary visualizations. While the source taxonomy does not explicitly define these characteristics, we used the following definitions that were agreed upon by the coders after iteration:

Trends	Estimate high-level changes across a dependent dimension,
Outliers	Identify items that do not match the modal distribution,
Clusters	Identify groups of similar items,
Frequency	Determine how often items appear,
Distribution	Characterize the extent and frequency of items, and
Correlation	Identify patterns between dimensions.

These analysis tasks provide a representative proxy for understanding the informational utility of a summary visualization. Each of these three categories (summarized in Table 3.1) is measured as a combination of binary codes (task supported/unsupported).

Other: We recognize that a codebook constructed *a priori* may not capture all elements of designs and tasks of summarization. To capture traits of summary visualizations not captured by this initial set of codes, we allowed coders to note additional observations about each summary for further exploration.

3.3 Survey Results

Of the 180 papers coded, 104 (58%) contained summary visualizations. Of these, 64 (36% of those 180 total) provided enough detail within the paper to support valid coding across all four categories (described as *fully-coded summaries*). The remaining 40 consist of primarily scientific visualization systems focused on rendering, which provided little to no description of the target purpose or analytic tasks supported by the approach. To avoid over-extrapolation, we only coded those systems for data summarization.

By situating the 104 coded summaries within our design space, we identify factors leading to different design decisions, explore common design themes, and also understand aspects of this space that currently unexplored in visualization. In this section, we use our research questions to highlight significant findings from our analysis, and generate key *themes* (Table 3.2) that describe observations from the survey process. Taken together,

these themes highlight core challenges for and opportunities for innovation in designing summary visualizations. These challenges concern how designers might exploit task specificity (addressing **Q2**), how systems leverage common design patterns (**Q3**), and how data affects design considerations (**Q4**). The full analysis results are available online at http://graphics.cs.wisc.edu/Vis/vis_summaries/.

Challenge	Axis — contributing factors	Theme — observations about the survey data
Use (C1)	Purpose	T1 Summaries serve as a starting point for analysis.
	Purpose	T3 Confirmatory summaries support exploration.
	Purpose × Data Summarization	T5 Designs for communicating specific, known information use aggregation.
	Purpose × Task	T6 Summaries using subsampling emphasize exploration.
	Task	T15 Summaries act as roadmaps to guide detailed exploration by interaction.
Specificity (C2)	Purpose × Task	T2 Exploratory summaries encode a broad set of data characteristics.
	Purpose × Task	T4 Presentation summaries emphasize a small set of specific characteristics.
	Purpose × Data Summarization	T5 Designs for communicating specific, known information use aggregation.
	Data Summarization	T7 Most summaries use more than one data summarization method.
	Data Summarization × Task	T9 Summaries using aggregation support tasks characterizing the entire dataset.
	Data Summarization × Task	T12 Projection and filtering emphasize similar data characteristics.
	Data Summarization × Task	T14 Subsampling supports tasks that are statistically robust to random sampling.
Task	T16 Summaries emphasize patterns that characterize all data and dimensions.	
Design Patterns (C3)	Purpose	T1 Summaries serve as a starting point for analysis.
	Purpose × Data Summarization	T5 Designs for communicating specific, known information use aggregation.
	Data Summarization	T7 Most summaries use more than one data summarization method.
	Data Summarization	T8 Most summaries use aggregation to summarize data.
Data (C4)	Data Summarization	T10 Aggregation is common across all data types.
	Data Summarization	T11 Filtering can be used across all data types.
	Data Summarization	T13 Summaries using subsampling are most common for scientific visualization.

Table 3.2: Our analysis revealed sixteen common design themes in examples of summary visualization. Taken collectively as observations, these themes highlight the challenges in the design of summaries. We use these challenges to reason about the trade-offs in existing designs and to identify underexplored areas of the design space to inform new summary designs.

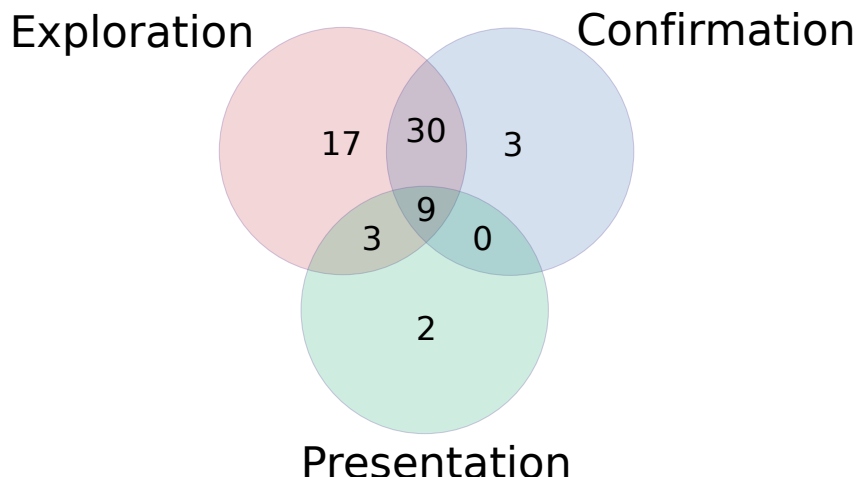


Figure 3.2: The distribution of summaries designed for each purpose over 64 fully-coded summaries (information visualization & visual analytics).

3.3.1 Purpose

Q3 addresses the question of how use of a summary visualization affects the design decisions of summary. To understand how use affects these decisions, we look at the data for statistical trends in the coded purpose of the visualizations. Purpose codified the intended use of summary visualizations for exploration, confirmation, or presentation (Figure 3.2). Most fully-coded summaries supported exploration (92%, 59 of 64), allowing viewers to analyze large collections of data without any *a priori* goals. 66% (42) of summaries were designed for directed analysis (confirmation), while only 22% (14) were explicitly designed to communicate known results (presentation). The dominance of exploration characterizes our first design theme: **summaries most frequently serve as a starting point for detailed analysis (T1)**. 95% (56 of 59) of these exploratory summary designs supported some sort of navigation task and 58% (34) allowed viewers to directly manipulate the granularity of the data encoded in the summary.

Additionally, **exploratory summaries support a broader set of data characteristic tasks (T2)**, such as identifying trends, outliers, clusters, frequency, distribution, and correlation. 70% (41) of exploratory summaries enabled viewers to explore more than three of the six high-level task characteristics (compared to 43% [6 of 14] for presentation) and 12% of summaries (7) supported all six. For example, Chen et al. [2016] uses a set of sum-

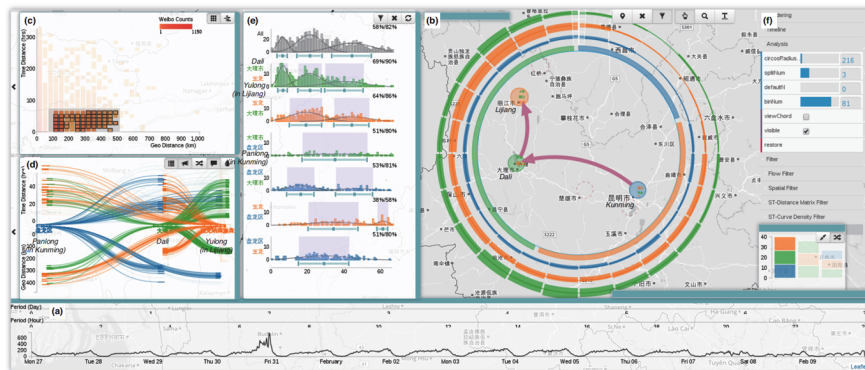


Figure 3.3: A visual summary in the system built by Chen et al. [2016] uses both aggregation and filtering in order to support a wide range of high-level analysis tasks.

marization methods to visualize different patterns across geo-tagged social media data. The resulting system (Figure 3.3) allows analysts to explore aggregate movement trends from social media data, and leverages interaction to enable analysis of the data distribution, frequency, and geospatial-based clusters.

Confirmatory summaries were often also exploratory: 61% of summaries (39 of 64) supported both exploration and confirmation while none were designed for confirmation or presentation alone. Like exploratory designs, confirmatory designs support a broader array of data characteristics than presentation-oriented summarization (68% supported more than three characteristics). This correlation suggests that **summaries designed for confirmation also support exploration (T3)**: confirmatory tools generally allow analysts to not only confirm specific hypotheses about data, but also to further refine and develop additional hypotheses about the data.

In contrast, **presentation summaries often emphasize a small set of data characteristics (T4)**. 57% (8 of 14) of presentation summaries communicated three or fewer coded data characteristics, and only one design communicated all six (Domino [Gratzl et al., 2014], which also supports exploration). All coded presentation summaries used aggregation to summarize data. Of these, 50% (7) used aggregation alone and 35% (5) used aggregation plus filtering. This suggests that **designs communicating specific, known information heavily rely on aggregation (T5)**. Aggregation can summarize data into a small number of precise features to emphasize known findings, encouraging effective presentation [Kosara, 2016]. This theme combined with T2 highlights potential challenges in use and specificity

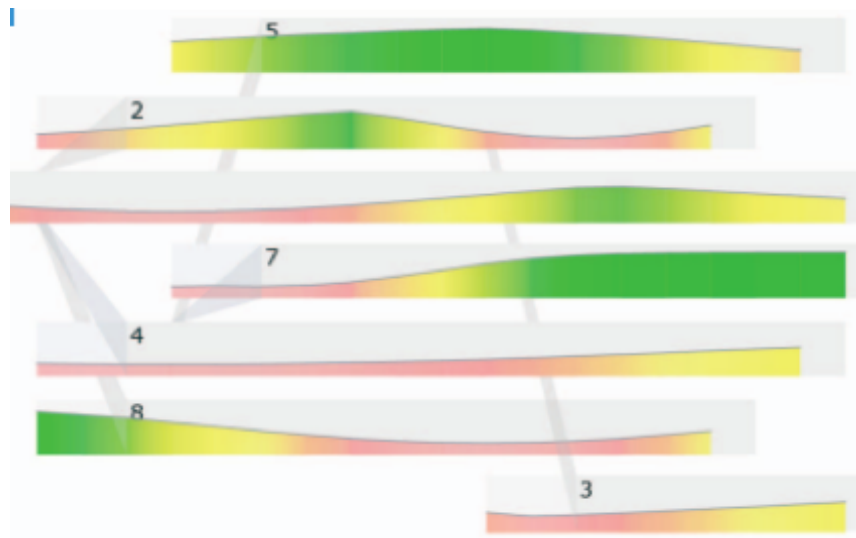


Figure 3.4: World Lines [Waser et al., 2010] aggregates spatial data across different simulation runs to allow viewers to directly search for the simulation with the best outcome.

(§3.4): focusing on specific properties of the data limits viewers' abilities to engage with data to better understand and evaluate presented findings whereas supporting many properties can overwhelm analysts or unnecessarily clutter a summary visualization.

Only five summaries were not explicitly designed for exploration. All five were confirmatory visualizations using aggregation, and none used subsampling. This bias indicates a trade-off between purpose and subsampling. **Subsampling methods favor exploration (T6)** as directed search may be inhibited by stochastically reducing data. Alternatively, aggregation helps guide analysts by presenting precise summarized values for well-defined tasks. For example, World Lines [Waser et al., 2010] uses aggregation to summarize parallel simulations of temporal events enabling comparison across known metrics for disaster planning (Figure 3.4).

3.3.2 Data Summarization Methods

All of the coded summaries employed at least one summary method (Figure 3.5), validating **Q1** that the organization is sufficient to cover the range of summarization operations. Unlike purpose and tasks, data summarization methods were coded for all 104 coded summary visualizations. We found that **most summaries used more than one data sum-**

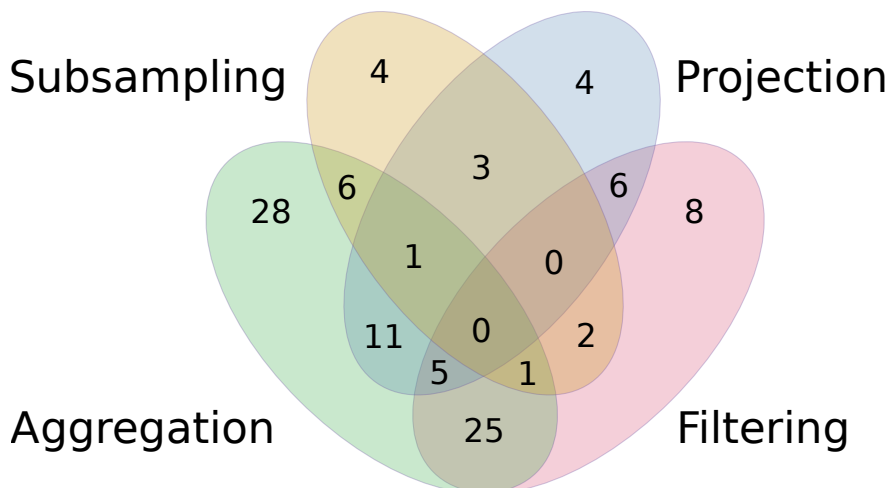


Figure 3.5: The distribution of summary designs using each data summarization method across 104 coded visual summaries.

marization method (T7) (63 summaries of 104, 61%), with 53 (84%) using exactly two. Each summarization method tended to favor a particular set of tasks. Combining summarization methods allows summary designs to leverage the strengths of individual techniques. However, there appear to be limits in how many summarization methods could effectively be composed: none of the coded summaries in our survey used all four summarization methods together. Rather, by understanding how each method is used in conjunction with other factors, we can analyze common design, use, and specificity patterns driven by these techniques. These observations are driven by **Q2**: how does the method of data minimization affect the resulting summary visualization?

Aggregation

Aggregation summarizes data by collecting and collating like-objects together through spatial, organizational, or attribute similarity. **Most surveyed visualizations (74%) use aggregation to reduce data (T8)**, with 27% exclusively using aggregation. **Visualizations frequently used aggregation to support tasks characterizing the entire dataset (T9)**. Of the 64 fully-coded examples, aggregation frequently supported both distribution (42 of 54, 78%) and clusters (43 of 54, 80%).

Visualizations often used these methods of data reduction to take advantage of trade-offs between aggregation and filtering: while aggregation emphasizes characteristics describing

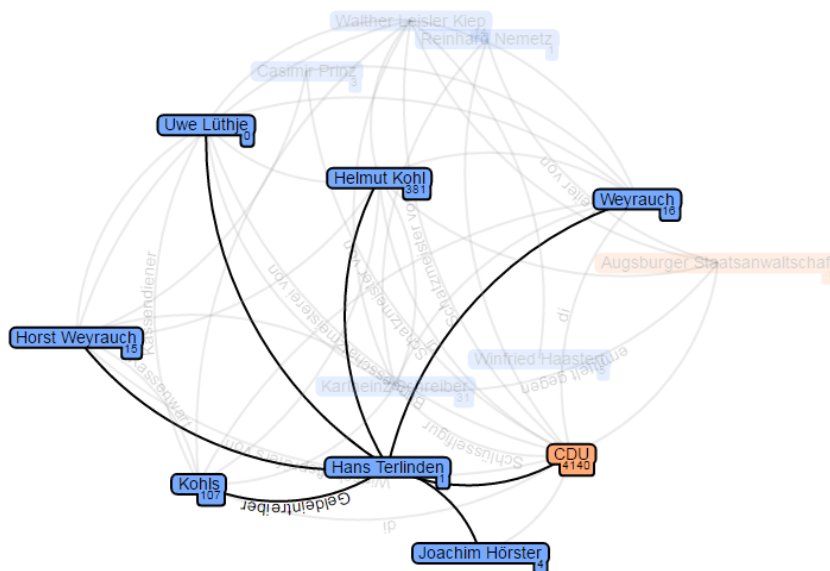


Figure 3.6: Most network summaries, such as Networks of Names [Kochtchi et al., 2014], combine aggregation and filtering to summarize data. The system aggregates different relations across pairs of entities and filters these patterns according to their frequencies to encode the relationships that best characterize the dataset.

multiple datapoints, filtering can help tailor these characteristics towards interesting or relevant collections. For example, the Network of Names [Kochtchi et al., 2014] first aggregates recurring relations in social networks and then filters out uncommon relations to emphasize dominant patterns in large actor networks (Figure 3.6). Filtering can also be used to reintroduce important data values obscured through aggregation, such as outliers in a scatterplot aggregated by density [Mayorga and Gleicher, 2013] (Figure 3.7). We found that summarization without aggregation targeted these kinds of individual value judgments, such as identifying outliers which was supported by 70% of non-aggregate visualizations.

In terms of Q4 (effect of data type on summarization), **aggregation was commonly used across all data types (T10)**, occurring in more than half of the surveyed papers across all data types. The dominance of aggregation across all data types indicates that it is a “default” used in visualization systems. Although aggregation is a powerful technique, it communicates specific properties of a dataset at the cost of the underlying data values. To use aggregation effectively, designers must know what properties of the data are important to the user and how to compute and encode these properties to faithfully represent the underlying data. Summaries using aggregation exchange flexibility for specificity, and crit-

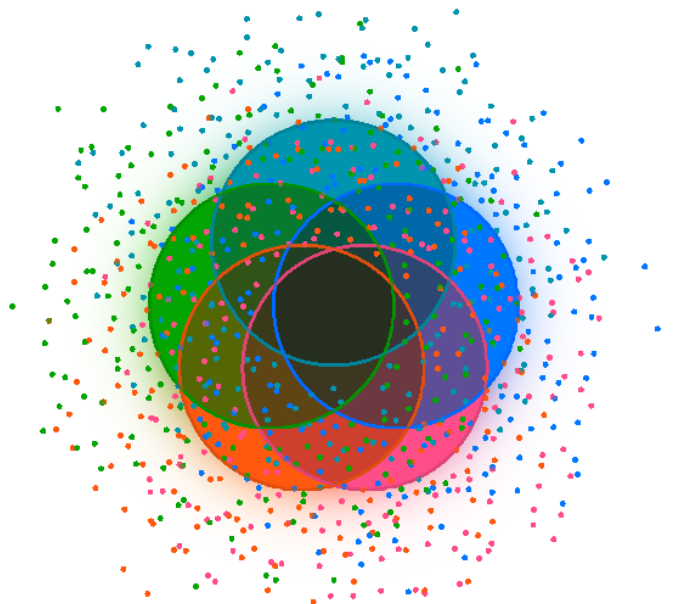


Figure 3.7: Splatterplots [Mayorga and Gleicher, 2013] represent two-dimensional points by combining a kernel density estimation with filtering and subsampling of representative outlier points. Combining aggregation and filtering takes advantage of the trade-offs between these methods to support a broader variety of tasks.

ically examining this trade-off may offer new opportunities for visualizations—discussed in detail in §3.4.

Filtering

Filtering is commonly used to allow analysts to specify meaningful properties of the data or compute representative subsets. 47 visualizations (44%) used filtering; however, filtering was seldom used in isolation (17% of all filtering summaries, supporting T7). Filtering in visualizations allowed analysts to identify clusters (23 of 47 filtering visualizations, 82%), characterize distributions (22 of 47, 79%), and evaluate correlation (17 of 47, 61%). Filtering tended to reduce extraneous data to support and highlight these types of high-level judgments, reflecting the visual information seeking mantra [Shneiderman, 1996]: “overview first, zoom and **filter**, then details on-demand.” Filtering in these cases helps analysts find interesting subsets of the data to explore.

Like aggregation, **filtering supported summary designs for all data types (T11)**. However, visualizations leveraging filtering provided analysts with little information about how

filtering for these properties might bias potential interpretations, again raising challenges for designers around summarization specificity.

Projection

As a method of summarization, projection is used to re-organize data as part of several summarization operations. 30 examples (28%) used projection to summarize data, with most projections summarizing large collections of documents (7 of 30, 23%), 3D fields (9 of 30, 30%) and multi-dimensional datasets (10 of 30, 33%)—highlighting the utility of projection for high-dimensional data (Q4). Similar to filtering, projection was seldom used in isolation (T7), and was commonly paired with either aggregation or filtering (24 summaries, 80%). For example, text documents can use topic modeling to project document vectors into a lower dimensional space and aggregate documents according to these topics (e.g., Cui et al. [2014]).

Projection-based summarization emphasizes similar data characteristics as filtering (T12): locating clusters (17 of 19 summaries, 89%), characterizing distributions (16, 84%), and evaluating correlation (14, 74%). However, projection frequently also enabled outlier analysis (15, 79%). Visualizations can combine filtering and projection to help highlight critical patterns in complex data. For example, Progressive Insights [Stolper et al., 2014] projects patterns onto statistical axes and filters the strongest patterns along each axis to highlight the strongest patterns over each new dimension.

Regarding Q3 in the use of summarization methods, we found that projection was seldom used for presentation (2 of 20, 10%), but instead supported in-depth explorations, as in Progressive Insights. We hypothesize that this is because the mathematical complexity of many methods make it difficult to clearly communicate meaningful narratives about the data, leaving designers to reason closely about use and specificity when using projection techniques (§3.4). However, we acknowledge that this may be biased by our choice of corpus, as we discuss in Section 3.4.

Subsampling

The act of subsampling reduces data for display by stochastically and indiscriminately removing objects from the dataset. While relatively few visualizations used subsampling to reduce data (16% of the 104 sampled), subsampling is also commonly paired with another summarization method (aggregation: 8 visualizations, 47% of subsampled examples; filtering: 3, 18%; and projection: 4, 24%). Similar to projection, subsampling is commonly used as a conjunctive operation to reduce data to manage the complexity of the resulting visualization.

Subsampling was predominantly used for spatial visualization (T13) (11 of 17 examples, 65% of subsampling use), where it reduced the visual complexity of aggregated structural data. Only six fully-coded visualizations used subsampling. These visualizations primarily support trend analysis (5 summaries, 83%) and characterizing distributions (5, 83%). These high-level characteristics indicate that **subsampling can support summarization where the analysis tasks are statistically robust to random sampling (T14)**. While few summaries use subsampling in practice, it is the only data summarization method that does not bias the resulting summary towards any specific attribute of the data. This implies subsampling may be a powerful tool for summaries for novel exploratory visualizations, especially when the target tasks or properties of interest are unknown *a priori*.

3.3.3 Tasks

While several of our prior design themes address relationships between methods of data minimization (Q1, Q2), and data types (Q4), we also explore how the utility of summary visualizations affect design decisions (Q3). The trends here help to inform how summarization affects analytic trade-offs in visualizations (Figure 3.8). From the 64 fully-coded visualizations, we found themes around how designs allow viewers to navigate the dataset, how summarizing different data types prioritize different analyses, and characteristics of the data that summarization universally preserves.

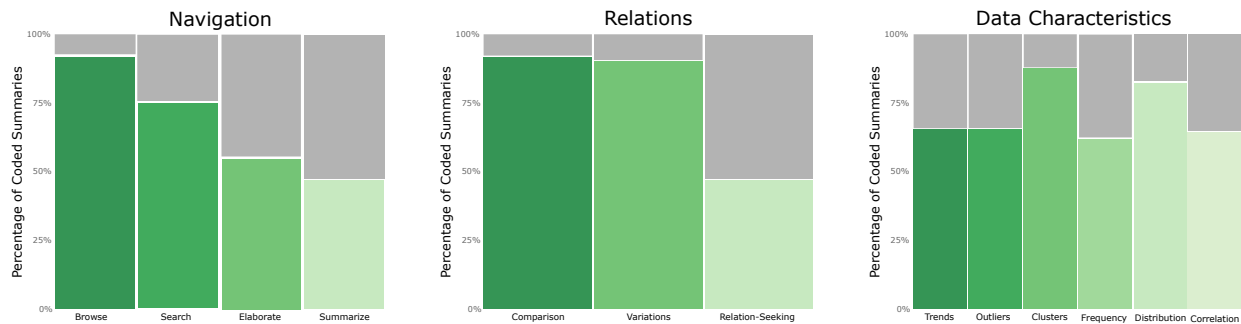


Figure 3.8: The distribution of summary designs supporting different kinds of analysis tasks across 64 fully-coded summaries.

Means of Navigation

Most visualizations presented summarized data to allow browsing for unknown patterns in data (58 of 64, 91%) while a smaller number supported directed search for known patterns (48 of 64, 75%). Among those, 13 visualizations (20%) supported browsing but not searching. These summary methods tended to emphasize relationships across collections of datapoints: all but one emphasized both clusters and outliers, and all but two communicated value distributions. For example, in Brehmer et al.'s juxtaposed matrix and faceted box plots [2016], the aggregate matrix obscures local patterns to prioritize aggregate temporal clusters while box plots encode distribution and outliers (Figure 3.9). This aggregation prevents directed search for individual motifs; however, the interaction between box plots and matrix cells allows viewers to browse for interesting local patterns. This exemplifies how **effective summaries can act as roadmaps to guide user interactions with the data (T15)**. This raises an important challenge for visualization designers to consider when summarizing data: what properties of the data might make for an effective starting point?

Our survey revealed that most designs start with the most abstract available data representation, then allow analysts to drill down into the data to uncover specific details. Many summaries did not allow viewers to change the level of detail without changing the visual representation (28 visualizations, 44%). All of these visualizations used additional supplemental designs to support detailed exploration, supporting **T15**. For example, glyph SPLOMs [Yates et al., 2014] summarize distributions within SPLOMs so viewers can identify scatterplots to explore in detail (Figure 3.10).

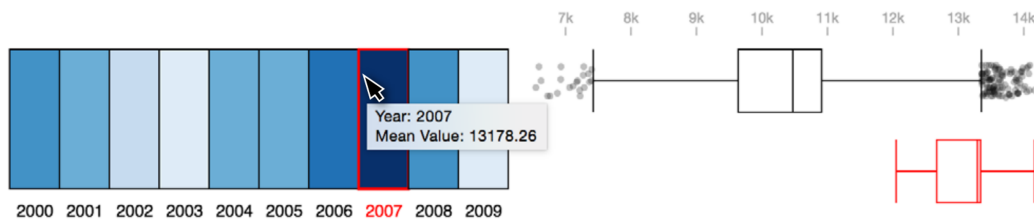


Figure 3.9: Summaries supporting browsing, but not directed search, tended to emphasize properties of collections of datapoints, such as distributions and clustering. For example, Brehmer et al. [2016] use aggregation allow viewers to identify high-level temporal clusters, outliers, and distributions and use interaction to browse for interesting underlying distributions; however, this aggregation obscures smaller scale motifs, preventing viewers from localizing specific patterns.

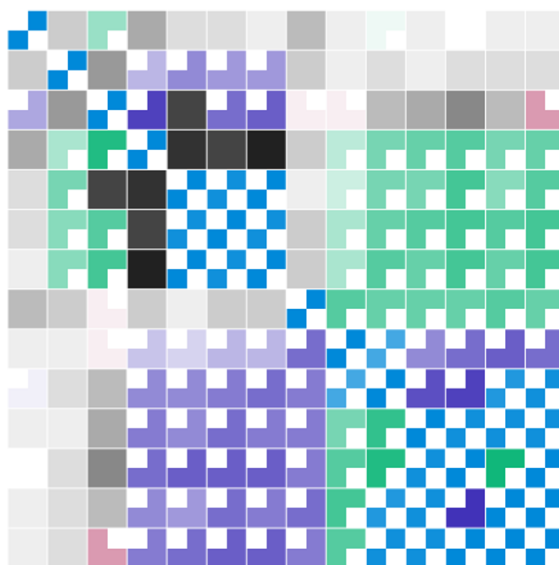


Figure 3.10: Summaries act as roadmaps for exploration, starting with a high-level of abstraction and often requiring viewers to use alternative representations to explore details. For example, glyph SPLOMs [Yates et al., 2014] summarize the quadrants where datapoints are clustered in each scatterplot of a SPLOM. Viewers can then look at specific scatterplots to explore interesting data in more detail.

Means of Relation

Most visualizations used summarization methods that enables viewers to identify similarities (89%) and differences (88%) between collections of datapoints. However, significantly fewer support relation-seeking between individual items (45%), with most of these being network visualizations. Network visualization for relation-seeking use some combination

of aggregation or filtering (2 aggregation, 2 filtering, and 4 aggregation+filtering). The correlation between network data and relation-seeking implies that summarizing network data often requires emphasizing relationships between key portions of the network. Aggregation (by collapsing important collections of nodes or edges) and filtering (by preserving meaningful or common relations) allow designers to meaningfully summarize networks. For example, Networks of Names [Kochtchi et al., 2014] highlights relationships between large collections of entities by first aggregating all entity relations and then filtering on these aggregate frequencies to visualize the most common relations in the dataset (Figure 3.6). The only coded network visualization that did not rely exclusively on aggregation and filtering, SAVE [Shi et al., 2011], did not emphasize relation-seeking and instead focused on multidimensional measures associated with each node. Despite the structural similarities between network and hierarchical data, the latter tended not to support relation-seeking though more visualizations of hierarchical data should be explored before drawing conclusive insights.

Data Characteristics

Summarization most frequently preserved characteristics that describe the entire collection of data and dimensions: clusters (80%) and distributions (75%). Trends (59%), outliers (59%), frequency (56%), and correlation (58%) were roughly equally supported across all visualizations. The bias towards clusters and distributions suggests that **summarization often emphasizes descriptive aggregate patterns across all of the data and dimensions (T16)**, rather than patterns in individual values or relationships between specific dimensions. 11% of coded summary visualizations support all tasks (7 of 64).

We found a bias towards particular task affordances and summarization methods across data type (addressing Q4). For one-dimensional data, many visualizations support discovering clusters (9 of 10, 90%) through aggregation (9, 90%). For 2D data, many visualizations support discovering trends (7 of 8, 88%) and frequency (6, 75%). In comparison, 3D data summarization tends not to support trends or frequency judgments (3 and 1 of 7, 43% and 14%, respectively), but instead preserves distributions (5, 71%). Neither multidimensional nor network data used subsampling (5 summaries of 24, 21%; 0 of 10, 0%; respectively). We

anticipate this bias arises from stochastically removing information that could potentially remove critical structures in the data, such as relations between different levels of hierarchy or across different data dimensions.

3.4 Discussion

Our four research questions lead to observations identified through the QCA process, resulting in 16 design themes (Table 3.2) of summary visualizations in visual analytics. Through the survey process, we confirm that the four methods of summarization are sufficient to encapsulate data re-organization for display in a visual summary (Q1). Here we describe these challenges and opportunities in how viewers *use* summaries, and in how designers consider *specificity* in data summarization, leverage common *design patterns*, and tailor summaries to specific *data*.

3.4.1 Use (C1)

We address the questions of how the use of summary visualizations affects design choices (Q2 and Q3) through the following observations. Summary visualizations frequently serve as a starting point for analysis (T1), providing a roadmap for detailed exploration using alternative views or interactions (T15). To help guide analysis, designers often choose an summarization method and target characteristics based on a visualization's intended *use*: how does the summary guide subsequent interaction and interpretation? To tell an immediate and focused story (e.g., presentation), summarization emphasizes specific patterns (T4) while open-ended analyses are better supported by summarizations encoding a broad set of characteristics (T2).

Challenges: Addressing the use of exploratory summary visualizations (Q3), these visualizations generally present many data characteristics at once, which offers analysis flexibility but might also overwhelm viewers: they may not know which questions to ask first. Exploratory summaries might instead choose to depict subsets of important characteristics to guide viewers through a more targeted analysis sequence. This targeting could

be especially beneficial for domains with established analysis workflows or for guiding novice analysts who could become overwhelmed when faced with too much information.

Opportunities: Summary visualizations that violate assumptions around use may offer interesting trade-offs. For example in response to our research question on the use of summary visualizations, we found that summarization for presentation generally targets a smaller set of data characteristics, whereas exploration supports a larger set. On the surface, this pattern makes sense: presentation tells a story, while exploration searches for unknown patterns in data. However, inverting this pattern may be advantageous. While aggregation can communicate specific information, explicitly visualizing statistics about the data may cause viewers to misinterpret secondary characteristics [Correll and Gleicher, 2015]—for example, trend lines can cause viewers to too liberally label outliers. In response, designs using filtering or subsampling may alleviate potential biases and better familiarize viewers with the data. Further, allowing access to more data properties can allow viewers to construct their own interpretation of the dataset in the context of the arguments made through the visualization.

3.4.2 Specificity (C2)

One of our core questions is how the method of data summarization affects the types of information communicated by a resulting summary visualization (Q2). Existing summaries heavily emphasize data characteristics that describe datasets in aggregate rather than specific data points or dimensions (T16), using aggregation methods to compute and visualize specific patterns in data (T5). However, aggregation explicitly tailors summarized data to specific statistical tasks, visualizing a computed representation rather than the actual raw data. In contrast, subsampling might remove datapoints that are important to a particular story, but also reduces clutter and potentially denoises data while providing immediate access to the underlying data (T6). This trade-off characterizes summarization *specificity*: aggregation can target specific high-level data characteristics but obscures specific values, whereas subsampling and filtering encode individual data values but rely on viewers to estimate characteristics.

Challenges: We found that existing systems favor specificity over data fidelity. Even if important data characteristics are not known *a priori*, aggregation was often used to express generic properties of the entire dataset (T9), such as distributions and clusters (T7). Filtering, subsampling, and projection are seldom used without aggregation; however, designs using these methods frequently preserve the underlying characteristics of the data (T12 and T14).

Our results identify a need to more carefully consider how the broad use of aggregation may bias analysts. Aggregation generally focuses on precisely encoding a specific set of characteristics at the expense of allowing viewers to synthesize their own perspectives from available information. When favoring specificity, designers must carefully consider how their summarizations influence the interpretation of the data, especially as summary visualizations are frequently the first thing that analysts encounter when exploring their data (T1).

Opportunities: Favoring breadth over specificity supports serendipitous exploration of summarized data. Designing for serendipity can foster new discoveries or generate unexplored hypotheses [Thudt et al., 2012] by broadly supporting a plethora of tasks. Subsampling, the least common summarization method in our survey, may be especially helpful in designing for serendipity: subsampling summarized data are statistically unbiased against properties of the dataset. It provides designs with low specificity, but generally preserves aggregate characteristics of the data. Further, stochastic sampling can create summaries that are not subject to the same confirmation biases as targeted filtering or aggregation.

Summary designs should also consider how designs let viewers combine information through visual aggregation. This understanding and explicit use of visual aggregation is just emerging in the visualization literature (see Szafir et al. [2016] for a survey), and our random sample did not identify any summaries explicitly designed for visual aggregation (e.g., Sequence Surveyor [Albers et al., 2011]). However, visual aggregation may allow designers to tailor summaries to specific tasks while using summarization methods. A better understanding of how visual aggregation factors into the specificity of designs is important future work.

3.4.3 Design Patterns (C3)

Several design decisions were reflected in the majority of the coded summary visualizations (Fig. 3.8), helping to address the research question of how summary use affects design decisions (Q3). Understanding these seemingly “default” decisions can guide novel design thinking for summary visualizations, as well as proposing good, starting design foundations. For example, almost all surveyed systems used more than one data summarization method (T7). Compositing summarization methods can emphasize particular data characteristics and increase the number of tasks supported, but has the potential to increase the distance between the representation and the semantics (structure) of the original data. Aggregation, for example, is most commonly paired with other methods (T8), but aggregation techniques are often data-dependent and require viewers to interpret computationally transformed data. In these designs, using multiple summarization techniques to increase task support comes at the expense of usability: the viewer must perform more mental processing to translate visual patterns back to the underlying dataset.

Challenges: The use of design patterns in summarization encourages reproducibility and reduces the analyst’s overhead in learning new systems. However, designers must consider whether a particular design pattern is appropriate given the data type and analysis goals. To date, no concrete guidance exists for understanding design pattern effectiveness. Our results indicate a need to collect and standardize design patterns and evaluate their potential utility. Our design space provides a preliminary scaffold to build this knowledge.

Opportunities: A common design pattern was the use of summary visualizations as a starting point for exploration (T1). While this pattern aligns with conventional visualization guidelines [Shneiderman, 1996], designers might also consider how an analysis might craft a summarization to serve as ending point for an analysis. Insights from exploratory visualizations are often constructed longitudinally, building up as viewers learn more about their data [Saraiya et al., 2006]. Summarizations might arise as descriptors of the insights constructed during an analysis. While no surveyed summaries enabled this inductive summarization, a few visualization systems incorporated annotation within a summary component in order to iteratively refine overviews from insights (e.g., TenniVis [Polk et al.,

2014] and Overview [Brehmer et al., 2014]). For example, Overview lets analysts label and manipulate datapoints to construct understanding across documents. Considering how designs might support summaries as generative artifacts of an analysis, capturing features like provenance, model refinement, and insight development requires moving away from default design patterns to inspire new summarization capabilities and applications.

3.4.4 Data (C4)

Q4 addresses the question of how the data type affects the affordances of a summary visualization. Several themes highlight patterns between specific data types and summarization choices. In some cases, these patterns help guide particular designs. For example, summary visualizations use aggregation and filtering for any data type (**T10** and **T11**), whereas subsampling is generally used for spatial datasets (**T13**). Designers may be able to use common patterns across data types to better reason about how summarization methods might support heterogeneous data, as well as how to adapt summarization techniques across domains.

Challenges: The semantic and statistical properties of the underlying dataset and analysis goals can limit candidate summarization methods. For example, continuous 2D data can be meaningfully summarized using kernel-density estimation (KDE), whereas a kernel does not easily map to hierarchical data. We identified some voids in the factors for particular data types, including lack of frequency support for summaries of three-dimensional data, and a lack of subsampling examples for network and multi-dimensional data. These voids identify places where innovative methods are needed for intuitive summarization.

Opportunities: Specific data types tended to favor specific summarization methods. For example, summaries of document collections and scientific data rely heavily on projection (**T13**). Designers can use this correlation to derive design inspiration in other domains: how might projection effectively summarize datasets that are structurally similar to documents, such as collections of event sequences? An important aspect of understanding and applying our design space in practice will be understanding how different summary approaches might generalize across data types and domain scenarios.

3.4.5 Limitations & Open Questions

This work begins to answer our research questions, taking preliminary steps towards a broader discussion of data summarization in visual analytics. However, our data-driven approach is inherently limited by sampling. Although we anticipate that the collected systems and themes characterize summarization more broadly, we cannot make absolute claims about the generalizability of our results. Instead, this work allows us to identify challenges and opportunities for visualization design that will help extend and enhance a more principled use of data summarization. For example, our observations identified several patterns in summaries designed for particular data types, but our sampling across different data types is limited. Future work could provide deeper coverage across different data types through stratified random sampling to identify biases across different designs. This could inspire both generative guidance for summarizing data across domains and novel design techniques for guiding innovative summarization techniques.

Our dataset is also biased towards exploratory visualizations, which is likely a function of an underlying bias in the visualization research literature [Kosara, 2016]. While we elected to use this literature to ground our coding in the design intents of the authors (§3.2), summaries from other sources, such as data journalism, could help create guidelines that inform summaries for a larger variety of practitioners and uses.

Our observations from this survey begin to answer the four proposed research questions. The four methods of re-organizing and summarizing data (**Q1**) are confirmed as being sufficient—every example of a summary visualization was matched into one or more methods. We identified how methods of minimization affects the resulting affordances and use of visualization (**Q2**), including observations such as how subsampling tends to support tasks dealing with data characteristics that themselves are resistant to missing data (**T14**). We observed that the use of summary visualizations affected design decisions (**Q3**)—as an example, aggregation is used to focus the viewer for presentation tasks (**T5**). Lastly, we identified overrepresentation of summary methods for particular data types (**Q4**), such as how subsampling was most used for spatial datasets (**T13**). The observations from the survey helps to create a clearer picture for the design of summary visualizations.

3.4.6 Conclusion

As datasets grow in size and complexity, effectively leveraging summarization becomes increasingly critical for visual analytics systems. We crafted a design space for summarization and used this design space to evaluate 180 papers from the visualization literature using QCA. Our analysis identified the importance of summarization for visualization (employed in 59% of surveyed manuscripts) and 16 design themes relating visualization purpose, data summarization methods, data types, and analysis tasks. We found trade-offs in the use of different summarization methods and biases in their applications in existing designs. These themes highlight patterns in the design for summarization that can guide viewers using visualization systems.

This work is a critical step in characterizing a design space of summarization and creating a set of design patterns for summary visualization. Our four research questions help to validate the proposed organization of summarization methods (Q1), and identify over- and under-represented trends between the factors of purpose, summarization, affordances, and data type (Q2–4). As a result of this process, these observations comprise a foundation based in realized, visualization design. This foundation provides a base set of guidelines in designing bespoke summary visualizations, and also suggests some potential design defaults for interactive, viewer-centric visualization systems.

As a result of this work, we identify four methods of reducing and re-organizing data for summarization. We have shown that these four methods tend to select the types of high-level information that can be obtained from the resulting visualization. Through a systematic random sample of the visualization literature, we can obtain trends in summary visualization design and use, and highlight correlations that appear. In the following chapter, we explore how different design methods for summarization manifest themselves in a scatterplot design paradigm, and create a framework for understanding what factors (data characteristics and tasks) make some summarization designs more appropriate than others.

BIBLIOGRAPHY

- D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2392 – 2401, Dec 2011. DOI: 10.1109/TVCG.2011.232 [Cited on page 53].
- R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE, 2005. DOI: 10.1109/INFVIS.2005.1532136 [Cited on pages 30 and 36].
- N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, Heidelberg, 2006 [Cited on page 30].
- J. Bertin. *Seminology of Graphics: Diagrams, Networks, Maps*. The University of Wisconsin Press, Madison, Wisconsin, 1983. Translated to English by Albert Biderman and Howard Wainer [Cited on pages 29 and 34].
- E. Bertini and G. Santucci. Give chance a chance- modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006. DOI: 10.1057/palgrave.ivs.9500122 [Cited on pages 30 and 34].
- M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, Dec 2013. DOI: 10.1109/TVCG.2013.124 [Cited on pages 31 and 36].
- M. Brehmer et al. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2271–2280, 2014. DOI: 10.1109/TVCG.2014.2346431 [Cited on page 55].
- M. Brehmer et al. Matches, mismatches, and methods: Multiple-view workflows for energy portfolio analysis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):449–458, 2016. DOI: 10.1109/TVCG.2015.2466971 [Cited on pages 48 and 49].
- J. R. Brubaker, F. Kivran-Swaine, L. Taber, and G. R. Hayes. Grief-stricken in a crowd: The language of bereavement and distress in social media. In *Sixth International AAAI*

- Conference on Weblogs and Social Media*, pages 42–49, Dublin, Ireland, 2012. The AAAI Press [Cited on page 33].
- S. Card and J. Mackinlay. The structure of the information visualization design space. In *Proceedings of the Information Visualization Conference*, pages 92–99, 1997. DOI: 10.1109/INFVIS.1997.636792 [Cited on pages 30 and 34].
- S. M. Casner. Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics*, 10(2):111–151, Apr 1991. DOI: 10.1145/108360.108361 [Cited on page 36].
- S. Chen et al. Interactive visual discovering of movement patterns from sparsely sampled geo-tagged social media data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):270–279, 2016. DOI: 10.1109/TVCG.2015.2467619 [Cited on pages vi, 40, and 41].
- M. Correll and M. Gleicher. Implicit Uncertainty Visualization: Aligning Perception and Statistics. *Proceedings of the 2015 Workshop on Visualization for Decision Making Under Uncertainty*, 2015 [Cited on page 52].
- Q. Cui, M. O. Ward, E. A. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):709–716, 2006. DOI: 10.1109/TVCG.2006.161 [Cited on page 27].
- W. Cui et al. How hierarchical topics evolve in large text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2281–2290, 2014. DOI: 10.1109/TVCG.2014.2346433 [Cited on page 46].
- G. Ellis and A. Dix. A Taxonomy of Clutter Reduction for Information Visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1216–1223, Nov 2007. DOI: 10.1109/TVCG.2007.70535 [Cited on pages 30, 34, and 36].
- N. Elmqvist and J. D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454, 2010. DOI: 10.1109/TVCG.2009.84 [Cited on pages 28 and 34].

- J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *IEEE Symposium on Information Visualization*, volume 2002, pages 117–124, 2002. DOI: 10.1109/INFVIS.2002.1173156 [Cited on page 27].
- M. Gleicher et al. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, Sept 2011. DOI: 10.1177/1473871611416549 [Cited on page 36].
- S. Gratzl et al. Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2023–2032, 2014. DOI: 10.1109/TVCG.2014.2346260 [Cited on page 41].
- J. Heer and B. Shneiderman. Interactive Dynamics for Visual Analysis. *Queue*, 10:30, 2012. DOI: 10.1145/2133416.2146416 [Cited on page 36].
- Ji Soo Yi, Youn ah Kang, J. Stasko, and J. Jacko. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, Nov 2007. DOI: 10.1109/TVCG.2007.70515 [Cited on pages 31 and 36].
- I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, second edition, 2002. DOI: 10.1002/9781118445112.stat06472 [Cited on page 36].
- J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):495–513, 2013. DOI: 10.1109/TVCG.2012.110 [Cited on page 29].
- D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002. DOI: 10.1109/2945.981847 [Cited on page 29].
- D. A. Keim and H. Kriegel. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):923–938, 1996. DOI: 10.1109/69.553159 [Cited on page 29].

- G. Klein, B. Moon, and R. Hoffman. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems*, 21(5):88–92, 2006. DOI: 10.1109/MIS.2006.100 [Cited on page 36].
- A. Kochtchi et al. Networks of Names: Visual exploration and semi-automatic tagging of social networks from newspaper articles. *Computer Graphics Forum*, 33(3):211–220, 2014. DOI: 10.1111/cgf.12377 [Cited on pages 44 and 50].
- R. Kosara. Presentation-Oriented Visualization Techniques. *IEEE Computer Graphics and Applications*, 36(1):80–85, Jan 2016. DOI: 10.1109/MCG.2016.2 [Cited on pages 29, 41, and 56].
- Y. K. Leung and M. D. Apperley. E3: Towards the metrication of graphical presentation techniques for large data sets. In L. J. Bass, J. Gornostaev, and C. Unger, editors, *Proceedings of EWHCI '93*, pages 125–140. Springer, 1993. DOI: 10.1007/3-540-57433-6_44 [Cited on page 29].
- A. Mayorga and M. Gleicher. Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics*, 19(9):1526–1538, 2013. DOI: 10.1109/TVCG.2013.65 [Cited on pages 44 and 45].
- W. A. Pike, J. Stasko, R. Chang, and T. A. O’Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009. DOI: 10.1057/ivs.2009.22 [Cited on page 31].
- T. Polk et al. Tennivis: Visualization for tennis match analysis. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2339–2348, 2014. DOI: 10.1109/TVCG.2014.2346445 [Cited on page 54].
- D. Riffe et al. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 1998 [Cited on pages 26 and 32].
- A. Rind, W. Aigner, M. Wagner, S. Miksch, and T. Lammarsch. Task Cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation. *Information Visualization*, 15(4):288–300, Oct 2016. DOI: 10.1177/1473871615621602 [Cited on page 31].

- S. F. Roth and J. Mattis. Data characterization for intelligent graphics presentation. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 193–200, 1990. DOI: 10.1145/97243.97273 [Cited on page 36].
- P. Saraiya, C. North, and K. Duca. An Insight-Based Longitudinal Study of Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1511–1522, 2006. DOI: 10.1109/TVCG.2006.85 [Cited on page 54].
- H. J. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, 2013. DOI: 10.1109/TVCG.2013.120 [Cited on pages 29, 31, 34, and 36].
- L. Shi et al. SAVE: Sensor anomaly visualization engine. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 201–210. IEEE, 2011. DOI: 10.1109/VAST.2011.6102458 [Cited on page 50].
- B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings IEEE Symposium on Visual Languages*, pages 336–343. IEEE Comput. Soc. Press, 1996. DOI: 10.1109/VL.1996.545307 [Cited on pages 27, 30, 34, 45, and 54].
- R. R. Springmeyer, M. M. Blattner, and N. L. Max. A characterization of the scientific data analysis process. In *Proceedings of the IEEE Conference on Visualization*, pages 235–242, 1992 [Cited on page 36].
- C. D. Stolper et al. Progressive visual analytics: User-driven visual exploration of in-progress analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, 2014. DOI: 10.1109/TVCG.2014.2346574 [Cited on page 46].
- D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11, Mar 2016. DOI: 10.1167/16.5.11 [Cited on page 53].
- A. Thudt et al. The Bohemian Bookshelf: Supporting serendipitous book discoveries through information visualization. In *Proceedings of the SIGCHI Conference on Human*

Factors in Computing Systems, pages 1461–1470. ACM, 2012. DOI: 10.1145/2207676.2208607 [Cited on page 53].

J. Waser et al. World lines. *IEEE Transactions on Visualization and Computer Graphics*, 16(6): 1458–1467, 2010. DOI: 10.1109/TVCG.2010.223 [Cited on page 42].

A. Yates, A. Webb, M. Sharpnack, H. Chamberlin, K. Huang, and R. Machiraju. Visualizing multidimensional data with glyph SPLOMs. *Computer Graphics Forum*, 33(3):301–310, 2014. DOI: 10.1111/cgf.12386 [Cited on pages 48 and 49].

M. Zhou and S. Feiner. Visual task characterization for automated visual discourse synthesis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 392–399, 1998. DOI: 10.1145/274644.274698 [Cited on pages 31 and 36].