The VLDB Journal
https://doi.org/10.1007/s00778-019-00588-3

SPECIAL ISSUE PAPER

Making data visualization more efficient and effective: a survey

Xuedi Qin[1] · Yuyu Luo[1] · Nan Tang[2] · Guoliang Li[1]

Received: 31 December 2018 / Revised: 16 October 2019 / Accepted: 21 October 2019

Abstract
Data visualization is crucial in today's data-driven business world, which has been widely used for helping decision making that is closely related to major revenues of many industrial companies. However, due to the high demand of data processing w.r.t. the volume, velocity, and veracity of data, there is an emerging need for database experts to help for efficient and effective data visualization. In response to this demand, this article surveys techniques that make data visualization more efficient and effective. (1) *Visualization specifications* define how the users can specify their requirements for generating visualizations. (2) *Efficient approaches for data visualization* process the data and a given visualization specification, which then produce visualizations with the primary target to be efficient and scalable at an interactive speed. (3) *Data visualization recommendation* is to auto-complete an incomplete specification, or to discover more interesting visualizations based on a reference visualization.

Keywords Data visualization · Visualization languages · Efficient data visualization · Data visualization recommendation

1 Introduction

Data visualization, which transforms abstract data into physical visions (for example, length, position, shape, color, and so on), is a powerful means to present compelling stories of data to humans who are more visually oriented. Nowadays, all organizations have more data than ever at their disposal. Consequently, more and more organizations use data and advanced analytics to inform strategic and operational decisions. Data visualization is a natural fit for both giving a good overview of massive data, and making it easier to interpret the results of data analytics to data scientists.

The Blossom of Data Visualization Undoubtedly, data visualization has made great strides in many fields, contributed by multiple communities.

The *computer graphics* community has significantly advanced the technology of rendering beautiful yet self-interpretable visualizations using e.g., D3 [1].

The *visualization community* makes it easy for users to specify and interact with visualizations, such as D3 [1], Vega-Lite [2], VizQL [3], Tableau [4], and Microsoft Power BI [5].

The *database community* has significantly improved the user experience of seeing and interacting with data visualization in real time, even for big data (e.g., for millions or billions of records). For example, Hyper DB [6–8] is the back-end engine to power up Tableau [4], and the Falcon project (available at GitHub https://github.com/uwdata/falcon) makes D3 [1] highly scalable supported by Apache Spark.

In addition, data visualization has also been extensively used in many database-related applications, such as Excel [9], Google Sheets [10], Oracle Data Visualization Desktop [11], IBM DB2 [12], Amazon Quicksight [13], Microsoft Power BI [5], and many others.

✉ Guoliang Li
liguoliang@tsinghua.edu.cn

Xuedi Qin
qxd17@mails.tsinghua.edu.cn

Yuyu Luo
luoyy18@mails.tsinghua.edu.cn

Nan Tang
ntang@hbku.edu.qa

[1] Department of Computer Science and Technology, Tsinghua University, Beijing, China

[2] Qatar Computing Research Institute, HBKU, Doha, Qatar

Published online: 19 November 2019

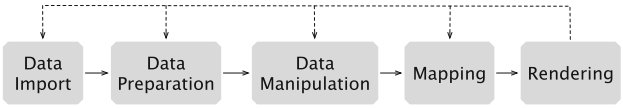


Fig. 1 The data visualization pipeline

The Pipeline of Data Visualization A typical *iterative data visualization pipeline*[1] is shown in Fig. 1.

1. *Data import* is to retrieve the required data from a desired data source.
2. *Data preparation* is to prepare the imported data for visualization, by e.g., normalizing values, correcting erroneous entries, and interpolating missing values.
3. *Data manipulation* is to select the data to be visualized (*a.k.a.* filtering from the visualization community) and possibly with other common operations such as joining and grouping.
4. *Mapping* is to map the data obtained from the above process to geometric primitives (e.g., points and lines), together with their attributes (e.g., color, position, and size).
5. *Rendering* is to transform the above geometric data into a visual representation.

Based on the pipeline, we have identified three directions that make data visualization more efficient and effective, yet relevant to database researchers.

(1) ***Visualization Specifications*** Visualization specifications provide various ways that users can specify what they want. There have been a great many studies from both visualization [1,2,14–16] and database community [3,17–19] on visualization specifications. We include it in this survey for two reasons:

– *Self-completeness:* It is important for readers to know how to generate data visualizations.
– *Language design perspective:* It mainly serves the "Mapping" component of the pipeline (Fig. 1), by specifying how to map different information to visual elements. However, it has some overlap with the "Data Manipulation" component, e.g., grouping and ordering operations can be specified in either step, which triggers a design choice problem between database languages (such as SQL) and visualization languages (see Sect. 2 for more details).

(2) ***Efficient Approaches for Data Visualization***

In order to effectively involve users in the iterative pipeline, the process of creating data visualizations must be efficient and scalable, especially for the two components, "Data Manipulation" and "Mapping". Many researchers have tried both interfacing with powerful and mature data processing engines (such as translating visualization queries to SQL queries to be evaluated over RDBMSs [17,20–23]), and customizing existing systems for data visualization tasks (such as HyperDB [6–8] for Tableau). There are also approximate solutions [24,25] and progressive solutions [26–28] to cope with big data, in order to provide real-time response. Both visualization [1,27,29–31] and database communities [22,24,32–34] have signification contributions on efficient visualization.

(3) ***Data Visualization Recommendation*** Precisely specifying a visualization is hard, even for experts, simply because the understanding of *what data to visualize*, *which story to tell*, and *how to visualize* is a trial and error exercise [17,22,35,36]. Hence, it is important that the visualization system can smartly guide users by providing recommendations. Several systems [18,32,36,37] allow users to provide an ambiguous specification, and the system will either automatically complete the visualizations, or provide recommendations. The works [20,38–41] from visualization community and [17,18,22,42] from database community tackle the problem of visualization recommendation from various angles.

Related Surveys Most existing surveys on visualization focus on a specific topic, such as graph visualization [43–45], linked data visualization [46–48], ontology visualization [49], high-dimensional data visualization [50], temporal data visualization [51]. We survey techniques from a different perspective.

For *visualization specifications*, Mei et al. [52] give a survey about classification, data source, presentation medium, etc., of visualization languages. We survey visualization languages from the stack perspective and emphasize how these languages are used from a practical perspective. There have also been some surveys [53,54] on exploratory data analysis tools, which are complementary to our *interactive data visualization*—we have added a discussion in the corresponding section.

For *efficient approaches for data visualization*, Keim et al. [55] consider how to integrate databases, data visualization, and data analysis so a user can easily work in one system, but without a discussion for efficiency. Idreos et al. [56] surveyed the techniques which aim to improve efficiency in the data exploration cycles, but we focus on techniques about how to construct visualizations efficiently. Bikakis [57] gives an overview of current systems and techniques for big data visualization, but with a less detailed discussion.

[1] Note that, our pipeline and terminologies used in this paper are slightly different than those used in the visualization community. Please refer to https://infovis-wiki.net/wiki/Visualization_Pipeline for more details.

For *data visualization recommendation*, although there have been many works [58–61] about recommendation systems and works about recommendation for different tasks, e.g., QOS-aware web services [62], social software [63], E-commerce [64], and there is no survey about data visualization recommendation, where we survey how different systems recommend insightful visualizations for users automatically.

2 Visualization specifications

2.1 The specification of data visualizations

Generally speaking, data visualization languages consist of three parts: data, marks (or visual cues), and the mapping between them.

- **Data**
  - *Records*: the data that need to be visualized.
  - *Transformation*: the operations—such as group, bin, filter, and sort—are used to transform the specified data records.
- **Marks** (or visual cues)
  - *Type:* the visual representation for data records, such as bar, line or point.
  - *Size:* the width, height of the visualization.
  - *Legend:* the legend information.
  - *Miscellaneous:* other properties, such as the width and color of a bar.
- **Mapping**: maps data to corresponding marks.

GUI-based visual operations are typically translated into data visualization languages.

2.2 A categorization of data visualization languages

A commonly used strategy to categorize data visualization languages is based on their expressiveness, as shown in the left side of Fig. 2. Apparently, the lower level of a language, the more expressive it is. Higher level languages encapsulate some low-level details by providing sensible defaults and adding more constraints (e.g., Excel [9] provides templates for supported visualizations). Another dimension to understand different levels of visualization specification languages is through their *accessibility* (or easy-to-use): the higher level the language, the easier to use, as also shown in Fig. 2.

Low-level Languages We refer to low-level languages as those that the users need to specify all mapping elements [1, 14,65–68].

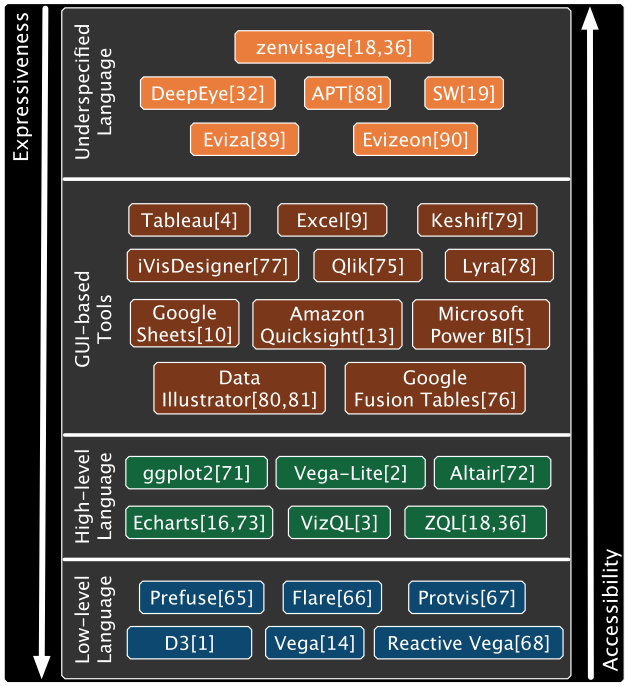


Fig. 2 An overview of data visualization specifications. Data visualization specifications are classified to four types: low-level language, high-level language, GUI-based tools and underspecified Language. The higher level the data visualization specification is, the easier it is to use and the less expressive it is.

Prefuse [65] and Flare [66] are Java-based visualization libraries; they encapsulate visual items as a Java class, which have many visual attributes, and the languages map data to these visual attributes by setting predefined functions. Protovis [67] is a declarative JavaScript-based graphical toolkit; it uses simple graphical marks (bar, area, line, etc.) with specified visual attributes. D3 [1] is a development of Protovis and is more effective in dealing with users' interaction (e.g., brushing and linking [69]). Vega [14] and Reactive Vega [68] are similar to Protovis and D3, but they provide declarative composable interaction grammars.

High-level Languages High-level languages [2,3,16,18,36, 70–73] encapsulate the details of visualization construction, such as the mapping function, as well as some properties for marks such as canvas size, legend, and other properties.

ggplot2 [71] is built on top of Wilkinson's work in 2005 "*The Grammar of Graphics* [70]"; it is a layered grammar of graphics embedded in R language. Vega-Lite [2] is a higher development of Vega and Reactive Vega; it also supports composable interaction design but provides concise grammars. Recently, Altair [72] made Vega-Lite available to the Python community. Echarts [16,73] is a latest development in declarative visualization languages designed to support quick visualization creation for non-programmers. VizQL [3] develops from the Polaris system [20] and is the visualization

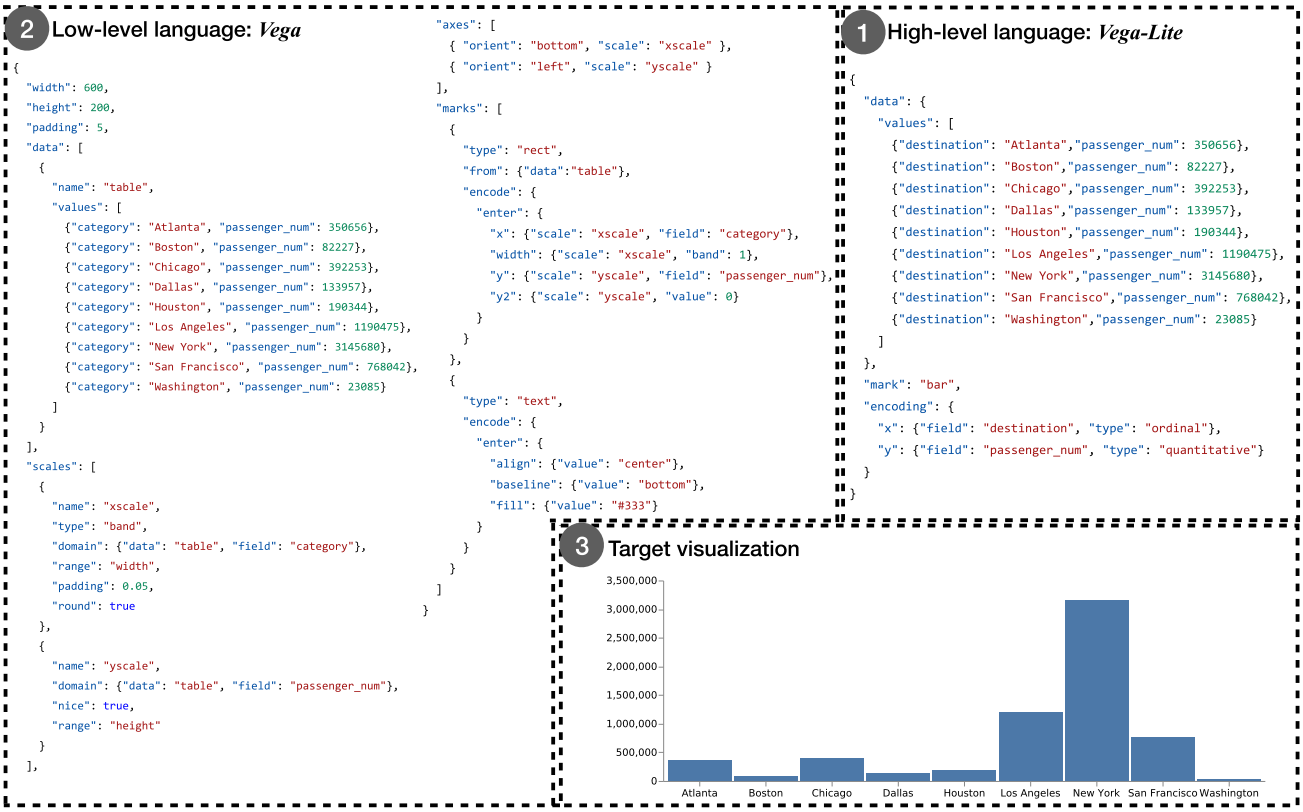


Fig. 3 Example of low- and high-level visualization languages. The target visualization (③) is a bar chart showing the passenger_num of different destinations. And we can use both low- (②) and high-level (①) visualization language to specify ③

specification language of Tableau. ZQL [18,36] of Zenvisage [18,36] employs a tabular structure language—each row in the table is a visualization specification.

Now, let us show the difference between different levels of visualization languages by an example.

Example 1 Table 1 is an excerpt of flight delay statistics. And Fig. 3 shows high-level (Fig. 3-①) and low-level specifications (Fig. 3-②) of a bar chart (Fig. 3-③) about *passenger_num* with *destination* in Table 1. Users can specify Fig. 3-③ by Vega-Lite in Fig. 3-①, and then Vega-Lite is compiled to Vega (Fig. 3-②), finally users will get the target visualization (Fig. 3-③).

Note that, in low-level languages, users have to specify the mapping function. For example, the "scales" in the Vega specification specifies the mapping function of the target visualization. The "xscale" denotes placing the categorical elements (*Atlanta*, *Boston*, *Chicago*, etc.) to the pixel range ([0, 600], specified by "range":"width") of X-axis averagely. And the "yscale" denotes mapping the data range ([0, 3500000], range of *passenger_num*) to the pixel range ([0, 200], specified by "range":"height") of Y-axis linearly. But in high-level language, users only need to specify the mark type, e.g., bar and do not need to specify the mapping function between data and mark. □

Note that, most of the low- and high- level languages listed in the survey are declarative languages (where the users only need to specify "what" they want) except Prefuse and Flare. Prefuse and Flare are procedural languages, because they are Java-based visualization libraries, and users should initialize panels, add visual elements, etc.

2.3 GUI-based visual operations

Compared with using declarative visualization languages to specify visualizations as discussed in Sect. 2.2, a more user-friendly way of providing a specification is to follow the "direct manipulation principle" [74], a widely used concept in the human-computer interaction aspect.

We have listed state-of-the-art GUI-based tools (Tableau [4], Qlik [75], Excel [9], Google Sheets [10], Amazon Quicksight [13], Microsoft Power BI [5], Google Fusion Tables [76], iVisDesigner [77], Lyra [78], Keshif [79], Data Illustrator [80,81]) in Fig. 2. Figure 4 shows an example of visual specification in Tableau using the flight delay data.

Remarks Our main purpose of discussing GUI-based visual operations is to show different ways that users can specify visualizations. Regardless of using declarative languages or visual operations to specify visualizations, the common prob-

Table 1 An excerpt of flight delay statistics of Chicago O'Hare International (Jan–Dec, 2015), where *scheduled* is the scheduled time to take off, *carrier*, *destination*, *departure delay (min)*, *arrival delay (min)*, *passengers* are the carrier, destination, departure delay, arrival delay, passenger number of the flight, respectively

A. scheduled	B. carrier	C. destination	D. departure delay (min)	E. arrival delay (min)	F. passengers
01-Jan 00:04	AA	New York	−5	2	173
01-Jan 06:43	MQ	Atlanta	9	2	132
01-Jan 09:30	EV	Chicago	13	17	127
01-Jan 00:04	AA	Boston	22	10	141
01-Jan 00:04	MQ	New York	19	13	232
01-Jan 00:04	UA	Los Angeles	0	−2	119

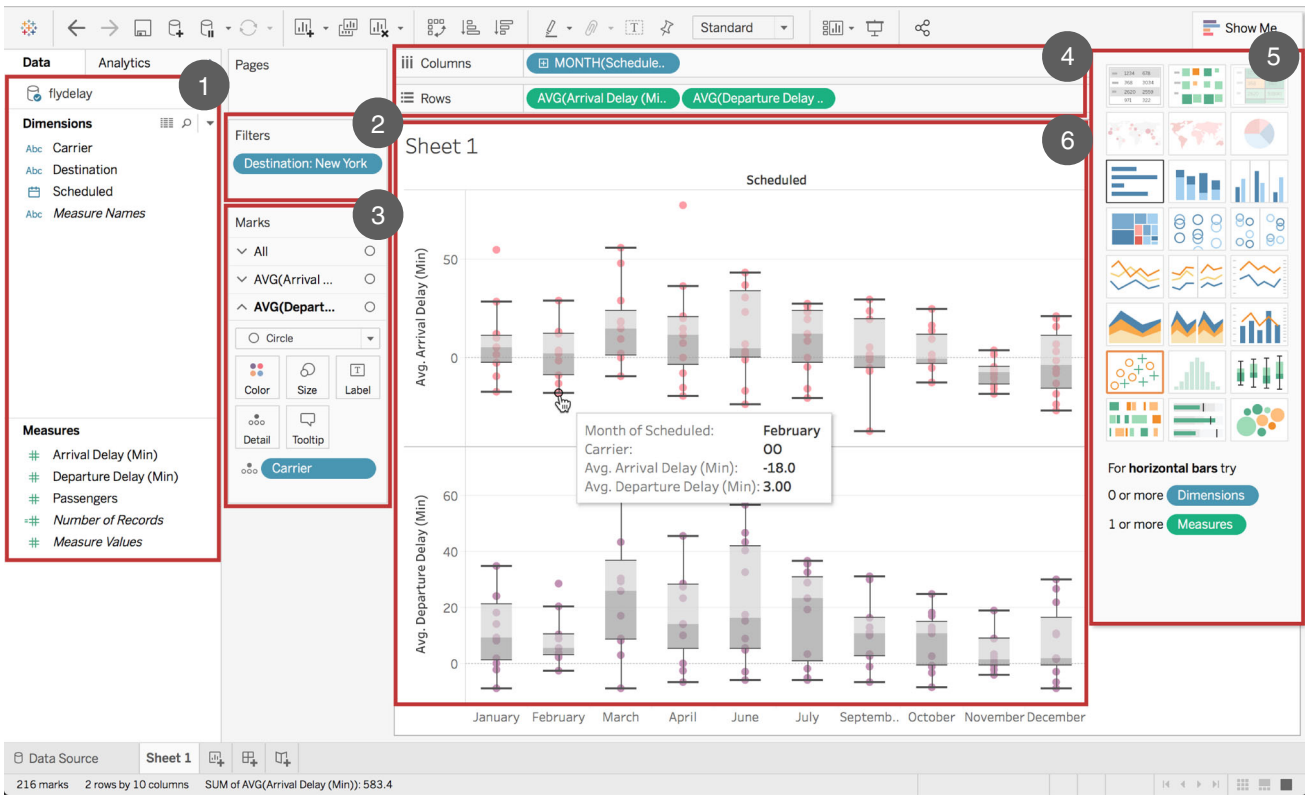


Fig. 4 An example of visual specification in Tableau using the flight delay data. ① displays the attributes of the loaded data, and users can drag attributes here to ④. ④ specifies the column attributes, row attributes, aggregation functions, and so on, for the specified visualization. The visualization of Tableau is in tabular structure. And the *Columns* and *Rows* in ④ denote the column attributes (i.e., *MONTH(scheduled)*) and row attributes (i.e., *AVG(departure delay (min))* and *AVG(departure delay (min))*) of the table. Users can choose the filter condition and visual mapping of marks in ② and ③, respectively. Also, users can click in ⑤ to specify the chart type. ⑥ displays the final specified visualizations to users, which is a box-and-whisker plot of average departure delay and average arrival delay with each month and carrier in 2017

lems of making the process efficient and smart are the same. Hence, classifying the applications of different tools is out of the scope of this article, for which please see the slides[2] of Jeff Heer for an introduction of these tools.

Interactive Data Visualization The rationality behind interactive data visualization is that in many cases, data visualization is a process of exploration, where the users need to keep refining the specification (e.g., add/remove/change attributes, change chart type) of current explored visualization until getting their desired visualizations in the exploration process.

We show two categories of interactive data visualization, Polaris and Tableau, using step-by-step query refinement to

[2] https://courses.cs.washington.edu/courses/cse442/17au/lectures/CSE442-Tools.pdf.

Fig. 5 Faceted navigation in DeepEye: visualization ① is the root for exploration, and the suggested facets for visualization ① are *bin size* and *category*; Once the user chooses the facet *bin size*, she gets visualization ②. Visualization ② is different from ① only in the bin size (① and ② are binned by weekday and date, respectively), and the other visualization elements (e.g., X-axis, Y-axis, chart type) of ① and ② are the same. Then, DeepEye suggests 3 facets for ②: *chart type*, *category*, *bin size*

create multidimensional visualizations. Moreover, DeepEye and Voyager enable facet exploration and help users easily navigate the visualization.

(1) Stepwise Query Refinement Polaris [20] and Tableau [4] provide chart templates to show multidimensional visualizations. Multidimensional visualizations are shown in two-dimensional plane organized in a tabular structure (Fig. 4). Using tabular structure (e.g., 2 Rows × 1 Column in Fig. 4) to display visualizations of different attributes (e.g., *AVG(departure delay (min))* and *AVG(departure delay (min))* in Fig. 4) or different values of the same attribute is called "small multiples" [82], which is convenient to compare and analyze different attributes (different values of the same attribute). The "small multiples" are widely used in data visualization systems, such as Voyager [83], VizDeck [84], Show Me [39], Profiler [85], [86], [87], etc. Users can gradually drag multiple attributes to the rows, columns, layers of the tabular visualization, pick the appropriate visualization type, mapping of data to visual properties, etc., to build desired visualizations step-by-step.

(2) Faceted Navigation DeepEye [37] supports faceted navigation to help users explore the visualization design space. Users can type in keywords, then DeepEye recommends relevant visualizations to users. Once a user chooses one interested visualization V, she can do a further navigation by different facets. The facets include *chart type*, *X-axis*, *Y-axis*, *category*, *bin size*, *group column*, and DeepEye will recommend visualizations which have the corresponding different facets with V while maintaining the other visualization elements unchangeable once users select one facet to explore. Also, users can choose the *similar trend* or *different trend* facets, and then DeepEye will recommend visualizations which have similar or different trend with V. Figure 5 shows a faceted navigation example on Table 1. Similar to DeepEye, Voyager [41] allows users to explore the visualization space by recommending visualizations which have the same or one more other attribute with current explored visualization.

Remark Although GUI-based interactive tools provide simple interfaces to quickly construct common visualizations, which is of great importance for non-technical people, there may be limited chart types in the templates, and it is also not flexible to change details of visualizations, such as bar width, and color mapping, etc. Hence, in practice, similar to high-level visualization languages, GUI-based interactive tools are typically used for quickly prototyping, or for finding useful visualizations. Afterward, low-level languages (e.g., D3) will be used for fine tuning or reimplementing the desired visualizations.

2.4 Underspecified specifications

Visualizations are meaningless if they cannot give insights of the data. However, in many cases, the users do not really know all aspects of the data at hand, because the data might be large and the data can be frequently updated. Hence, it poses a requirement of supporting underspecified specifications.

Generally speaking, for underspecified specifications, users only provide some "hint", and it is the task of the visualization systems to interpret the underspecified input, in (possibly) different ways.

The first type of hint is "reference-based", where the users provide a reference visualization as a seed and the system suggests visualizations based on the reference. zenvisage [18,36] returns similar or dissimilar visualizations (e.g., similar trends in line charts) with a user provided reference visualization.

The second type of hint is "keyword-based", in a Google style. APT [88] accepts user's data viewing goals of desired columns, for example, "*present the departure delay and scheduled relations*". In other words, APT specifies the *columns* to be visualized and then recommends visualizations which satisfy the goals. DeepEye [32] is a recent system that accepts keyword inputs as data viewing goals and provides recommended visualizations. For example, the user may input "*show me line charts about electricity*", and DeepEye

will recommend line charts which also contain the column "electricity" to users. The demo of DeepEye can be found at http://deepeye.tech. A similar tool[3] that supports keyword inputs, called "Ask Data", was recently released by Tableau, which allows user to get answers without the need to know the structure of the data, such as "*what is the average price by variety*". SW [19] accepts users' window-based constraints (e.g., "*identify all windows in which the average departure delay* $>$ 50") about desired visualization windows (a window is a rectangular region in a visualization).

The third type of hint is "natural language-based", which considers the context of user inputs and system states in the data exploration cycle instead of one-shot in "keyword-based" hint. Eviza [89] and Evizeon [90] are two recent visualization systems which provide natural language interfaces for visual analysis cycles. For example, in Evizeon [90], the user first types "show me the spike of measles in the UK", and Evizeon will show the user the spike in the line of measles outbreaks in the UK. Then, the user types "mumps over there", and Evizeon will show the user the mumps outbreaks in the zone of the spike of measles.

Discussion (1) We categorize visualization languages organized as a stack (see Fig. 2), which is different from the survey [52] that categorizes visualization languages based on *graphic library*, *declarative*, *chart typology*, *data source*, *presentation medium*, and so on. (2) The survey [54] that evaluates different exploratory data analysis (EDA) tools for different applications is complementary to our survey, because we focus on how interactive data visualization tools construct visualizations through iterative interaction (i.e., *stepwise query refinement*, *faceted navigation*) with users.

3 Efficient approaches for data visualization

In this section, we will discuss efficient approaches for data visualization; it is important because the data visualization life-cycle is always iterative (see Fig. 1), with human-in-the-loop.

In the following, we will first describe *exact data visualization* that computes precise visualization as fast as possible (Sect. 3.1). Sometimes, however, providing exact visualizations may not always be doable because of the large size of data and high complexity of queries, *approximate data visualization* that provides fast, but approximate visualizations are ideal for this case (Sect. 3.2). Furthermore, instead of only producing one-shot approximate visualizations, *progressive data visualization* gradually refines the intermediate results (Sect. 3.3).

[3] https://www.tableau.com/about/blog/2018/11/ask-data-simplifying-analytics-natural-language-98655.

Table 2 gives a summary of the techniques to be discussed in this section.

3.1 Exact data visualization

Many data visualization systems [17,21–23] read data from databases. They may also manipulate data by SQL statements and then use visualization tools to render the visualizations.

Query Translation A natural way to reuse many mature (DBMS) systems is to translate the visualization queries to the queries those systems accept. For example, DeepEye [17, 21], Polaris [20], SeeDB [22,23] get data by issuing SQL queries to the databases. By creating a mapping between the primitives of visualization language and SQL language, we can convert the target visualization language to a SQL query.

Example 2 The visualization f_1 in Table 3 specified by a ZQL query can be translated to a SQL query Q_1 as shown below.

```
Q1 : SELECT carrier, SUM(passengers)
     FROM flight delay
     GROUP BY carrier
     WHERE destination="New York";
```

The attributes of X- and Y-axes, i.e., *carrier* and *passengers*, can be mapped to the projection clause followed the keyword `SELECT`. The *Constraints* can be mapped to the filter condition following the `WHERE` clause. The *Viz*, i.e., *y = sum(passengers)* means that the SQL query should `GROUP BY` *carrier* and apply *sum* to *passengers*. □

Integrating Visualization Systems with DBMSs Although using query translation is natural, there are some disadvantages. One main issue is that many functionalities are repeated, resulting in non-unified optimization techniques with different assumptions and performance in server (i.e., the database side) and client (i.e., the visualization side), leaving the developers confused to choose the suitable optimization techniques. For example, the database engine and visualization tool may both support the *filter* operation; consequently, one can filter data records by either issuing a SQL query to database or by the function *filter* of JavaScript in the front end—choosing to *filter* at database or visualization tool (e.g., the front end) is difficult. Another main issue is that decoupled methods are hard to maintain, extend and optimize [92] for interactive visualizations, which requires continuously issuing queries to modify visualizations.

Intuitively, a promising way to solve the above problems is to tightly couple (or integrate) data retrieval and rendering together to speedup the process of visualization creation. Ermac [91], a Data Visualization Management System (DVMS), is a research attempt on this direction. It supports two relations: `data` and `scales`, where relation

Table 2 A summary of efficient data visualization, where we summarize the widely studied problems in efficient data visualization in column *Problem*, the corresponding techniques and references for each problem in column *Technique* and the target for solving the problem in column *Target*

Problem	Technique	Target
Exact Data Visualization	Query Translation [17,20–23]	Accelerate Visualization Exploration Process
	Integrating Visualization Systems with DBMS [91,92]	
	Column Stores [22,23,29,85]	
	Indexes [93–95]	
	Parallel Computation [22,23,31,96]	
	Prediction and Prefetching [19,31,34,97,98]	
Approximate Data Visualization	AQP [24,25]	Enable Quick Visualization Creation
	Incremental Sampling [26–28]	
	Human Perception [26,99]	
Progressive Data Visualization	Hierarchical Aggregation [31,93,100,101]	Enable Progressive Visualization Creation

Table 3 An ZQL query which returns a bar chart about the SUM(passengers) of different carriers to "New York", where *Name* denotes the visualization name specified by the ZQL query, *X* and *Y* denote attributes of the *X*- and *Y*-axes, *Constraints* specifies the constraints which the data used to generate the visualization should satisfied, and *Viz* specifies the visualization type and aggregation function on attribute of *Y*-axis

Name	X	Y	Constraints	Viz
f_1	carrier	passengers	destination = "New York"	bar. (y = sum (passengers))

`data` include the data records to be visualized and references to the rendered visual elements; relation `scales` denote the mapping from data ranges to visual encoding ranges. A visualization in Ermac is represented as a Logical Visualization Plan (LVP), and LVP is compiled into a SQL-like query. The SQL-like query deals with the `data` and `scales` relations, and the query constitutes a Physical Visualization Plan (PVP), and then PVP can be optimized by the traditional database optimization techniques. During query execution, Ermac uses rendering placement and psychophysical approximation techniques to reduce latency. It also uses visualization features to support automatic lineage-based interaction, visualization estimation, recommendation, and so forth.

A further development [92] of Ermac is proposed to provide a SQL-like language, DeVIL, to represent both static and interactive visualizations. In DeVIL, *Marks* and *Pixels* are two visual relations to express visualizations which are expressed in SQL-like queries. DeVIL models the user inputs as event streams and database relations and enables the interactive visualizations by executing SQL-like queries in joined visualizations and event relations iteratively to update the visualizations and response to user's inputs. By modeling the static and interactive visualizations as declarative database relations, visualization designers are released from event-driven programming, making programming process more standardized and code more scalable. The work of [92] also proposes many optimization techniques (e.g., concurrency control and streaming framework) for interactive visualizations in DVMS.

Column Stores In data management, a key performance factor is the data layout, e.g., row-based and column-based layouts, which may have a huge performance difference for OLAP applications. In terms of data visualization, the users are typically interested in only a few columns. Naturally, column-stores may achieve better performance, compared with row-stores, which have been adopted in SeeDB [22,23], Profiler [85], and TDE [29].

Indexes Indexes are widely used to improve search performance by essentially cutting down the number of records/rows in a table that need to be examined. Naturally, they play an important role in improving data visualization performance. FlashView [94] builds a hierarchical tree-based index to support users' selections with continuous filtering conditions. The work of [95] builds a tree-based index for the data which is to be queried instead of the whole dataset and gradually refines the index when more data are queried. imMens [31] and Nanocubes [93] build datacubes which precompute aggregation results for different data slices to reduce query execution time by accessing the precomputed aggregation

results instead of the raw data. Hashedcubes [102] also builds datacubes for real-time big data visualization. Hashedcubes uses pivot arrays to construct datacubes, while Nanocubes is tree-based. And Hashedcubes achieves lower memory usage and lower query time compared with Nanocubes. Gaussian Cubes [30] is a development of Nanocubes which supports more visualization analysis task types. For example, Gaussian Cubes precomputes sufficient statistics information in the datacubes to support model fitting.

Falcon [103] uses indexing techniques to reduce interaction time for brushing and linking in visualization. The visualization that the user is interacting (i.e., brushing) with is *active view*, and the other visualizations are *passive*. For current active visualization, Falcon builds index for each passive visualization. The index stores the data which should be highlighted in the passive visualization, and the data are in the form of array, where each entry of it stores cumulative counts. Thus, Falcon can calculate the data to be highlighted in the passive visualizations in constant time given the start and end position in the active visualization. Since Falcon only maintains index for active visualization, it has much smaller index than imMens [31], Nanocubes [93], etc.

Parallel Computation Parallel computation has also been widely used for query processing in data visualization systems [22,23,31,96]. The aggregation queries on data tiles in imMens [31] are parallelized using the dense index representation of a data tile. SeeDB [22,23] executes multiple SQL queries of visualization candidates in parallel during visualization ranking. Harald et al. [96] provide a multi-threading architecture for interactive visualization exploration. The architecture maintains a main application thread to capture users' interaction requests and multiple visualization threads for each visualization to process the visualization of this thread. Furthermore, whether the main thread and visualization threads are asynchronous or synchronous depends on the types of the users' interaction requests.

Prediction and Prefetching One important step of data visualization is data exploration—users continuously browse their interested visualizations to get a sense of what to visualize. Oftentimes, the current explored visualization is usually inspired from the previous one. In other words, users may get the next visualization by changing parameters of current visualization or zooming in/out to get detailed/overall information, etc. Evidently, predicting the following data that users may be interested, and then prefetching/caching data which may be used in the next step during current exploration can speedup the exploration process, and these techniques have been used in many visualization systems [19,31,34,97,98,100,104].

We categorize the prefetch and prediction technologies to two types, based on:

1. Currently explored visualizations [19,31,34,100], or
2. Historical data [34,97,98,105,106].

(1) Currently Explored Visualizations. XmdvTool [34] clusters tuples in different granularity to support users' hierarchical navigations. It enables users to continuously explore data in the structured-based brush [107]. Hence, it needs to predict the next direction of the user and then prefetches and caches the data in that direction during the idle time. The caching system uses the least recently used (LRU) as the replacement policy and the current explored visualization-based prefetching strategy is to randomly pick a direction from the position of the current explored data.

Following the above hierarchical navigations, instead of prefetching only one piece of data on the tree hierarchies, it is also natural to prefetch different levels' representation of the present data, as used in imMens [31] and [100].

Another angle for a good prefetching is based on the size of prefetched data (e.g., SW [19]), instead of the direction on the hierarchy that the user will explore. More specifically, SW finds all windows that satisfy users' constraints inputs (e.g., "*identify all windows in which the average departure delay* $>$ 50"). SW iteratively explores all possible visualization windows to find "good" (i.e., satisfy users' constraints inputs) windows. When a window is being explored, SW prefetches the neighbor windows in all directions, but the size of the data to be prefetched in each direction should be decided by the algorithm. SW first computes whether the prefetched window satisfies the constraints by sampling data of the window. If the result is true, the prefetching in this direction is stopped, and the prefetched window is to be explored further (the exploring process is same as the current window). Otherwise, SW continues to prefetch in that direction by increasing the sampling rate until the data in this direction are all prefetched or the constraints are satisfied. Note that SW wants to find all "good" windows in the search space, and thus, it must explore all possible windows, and by exponentially increasing the sampling size, SW can terminate the exploration quickly with less prefetching times, thus getting all "good" windows with less time.

(2) Historical Data. Next, we will discuss techniques that leverage historical trajectories [34,97,98,105] for prefetching.

When historical data are available, naturally, systems can do more complicated yet meaningful inference than randomly picking a direction as discussed above. More specifically, XmdvTool [34] proposes three strategies to prefetch the data based on historical data:

- the *direction*: select the most likely direction based on the users' previous trajectory tracking,
- the *focus*: select the direction with hot regions, and

– the *vector*: select the direction based on the vectors of the movement trajectories of the users, in the form of <*start position*, *width*, *level*>, where *start position* is the start location and orientation of the movement, *width* is the moving distance of the movement, *level* is the aggregation hierarchy of the data explored in the movement. It uses the mean or exponential weighted average of previous trajectory vectors to select the directions.

Recently, machine learning-based approaches have also been studied. ForeCache [98] partitions data to blocks or data tiles in different levels and predicts data tiles to users. There are two stages of data prediction:

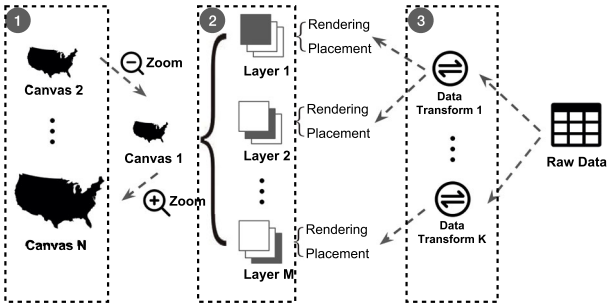
– *Predicting analysis phases:* it predicts the users' exploration phase by a Support Vector Machine (SVM) model, and the features include position, panning, and zooming information of users' exploration traces.
– *Predicting data tiles:* it uses the corresponding strategies to recommend prefetched data: ① action-based strategy using Markov chain which accepts sequences of users' movement (e.g., {*left, left, left*}) as a state and a move from state to state as a transition (e.g., "*right*"); ② signature-based strategy which recommends similar data tiles with users' previous explored data.

Experiments have shown that ForeCache achieves 25% higher prediction accuracy than the prediction strategies in XmdvTool [34].

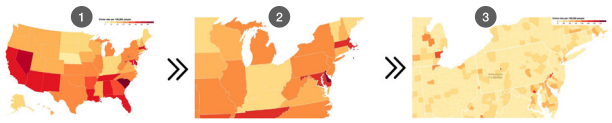
Case Studies using Kyrix and Tableau In the following, we will discuss two case studies using Kyrix, an interactive scalable data visualization system, and Tableau, one of the most successful visualization tools.

Kyrix [108] is an interactive scalable data visualization system. Kyrix provides declarative visualization specification interface in front end and effective scalable visualization processing in back-end, where the user zooms in to see detailed information and zooms out to overview in scalable visualization.

1. *Visualization Specifications in Front end.* There are two abstractions in the visualization specification language of Kyrix: *canvas* and *jump*. A *canvas* contains a static visualization, where the data of the visualization are specified by a SQL query and the transformation and rendering function can be specified by existing visualization libraries (e.g., D3 [1], Vega [14]). A *jump* specifies the source and destination canvas and the transition type when panning or zooming.
2. *Efficient Approaches for Data Visualization in Back-end.* There are two important improvements in Kyrix: *fetching granularity* and *indexing*. For *fetching granularity*, Kyrix splits raw data to static data tiles of fixed size.



(a) Declarative model of Kyrix [121]. A canvas (①) contains multiple layers (②), and a layer can be got by specifying a rendering and placement function for the transformed data (③).



(b) Zoomable crime rate map of US by Kyrix [121]. ① is the original visualization, and ②,③ are more detailed visualizations by zooming in.

Fig. 6 Declarative model and zooming example of Kyrix [108]

The tiles of current visualization together with a dynamic box which encompasses these tiles are sent to the front end, and the box is recalculated when the tiles of current visualization are out of it. Compared to fetching large or small tiles, Kyrix can adjust the size of dynamic box by different algorithms, providing a way to neutralize the network time and prefetching size. For *indexing*, Kyrix builds Btree [109] or hash indexes on the tile id of a tuple to support quick fetching.

Figure 6a shows the declarative model of Kyrix. A *canvas* (Fig. 6a-①) in Kyrix is a level of detail of data, and users can zoom in/out to see more canvas of different levels of details. A *canvas* may contain more than one *layer* (Fig. 6a-②) (e.g., background layer, line layer, etc.), and there should be a *rendering function* and *placement function* for the transformed data (Fig. 6a-③) of each *layer*, where the *rendering function* defines how to map data to visual objects, and the *placement function* gives the location of the visualized data by the *fetching granularity* and *indexing* strategies. Figure 6b is a zoomable crime rate map of USA by Kyrix.

TDE [29] is a data engine customized for visualization in Tableau 6.0. TDE optimizes the data engine mainly in the following perspectives.

1. *Column-oriented Storage and Compression.* Due to the high I/O cost of Tableau's former database Firebird and data of visualizations usually stored in different columns, column-oriented storage and compression techniques have been designed to solve this problem in TDE. TDE mainly uses dictionary compression strategy, and there

are two compression mechanisms for dictionary compression: *heap compression* for variable width types and *array compression* for fixed width types. Then, the compressed columns can be represented by the *dictionary tokens* (i.e., the dictionary keys) which reference the dictionary values during the query execution.

2. *Operator Reordering.* Selection operators and operators with single compressed columns are pushed down in the SQL query plan tree.
3. *Cardinality Reduction.* For columns with high cardinality columns, TDE automatically transforms these columns to higher hierarchies, e.g., transforms column *Time* with 2500 distinct values to column *Year* with 7 distinct values, then pushes the operators with *Time* in SQL query plan tree down and replaces *Time* with *Year*.
4. *Other Visualization Support.* TDE provides domain information (e.g., the cardinality, maximum and minimum values of the domain) of columns. This domain information can be used to choose the level of detail of a visualization for users. TDE also supports progressive reporting and termination control (i.e., terminate long running visualization queries) when executing visualization queries.

The recent effort of Tableau 10's server data engine is to customize a highly efficient main-memory system Hyper [6–8]. Hyper is used as the data engine to power all versions of Tableau, such as Tableau Server, Tableau Desktop, Tableau Online, and Tableau Public. In particular, Hyper is used to support efficient creation, refresh, query extraction, and cross-database joins.

3.2 Approximate data visualization

When the data volumes grow exponentially, traditional data processing modules cannot provide fast interactive processing results. To bridge the gap between data volumes and interactivity, many works [24–28,99] speedup data processing phase by leveraging approximate query processing (AQP) that provides approximate visualization results.

We discuss approximate data visualization from three perspectives: *AQP*-based approaches that leverage techniques from AQP; *incremental sampling*-based approaches that link incremental query processing to visualization; and *human perception*-based approaches that capture the cognitive limitations of human perception. A summary of the contents to be discussed is shown in Table 4 and Fig. 7.

AQP-based A straightforward way for generating approximate visualizations in interactive time is leveraging the techniques of AQP. Using the representative subset of the data can provide users with approximate visualizations for online interaction by sacrificing the quality. We will review

Table 4 Summary of approximate data visualization systems, where we summarize the papers and algorithms for approximate data visualization systems in column *Paper* and *Algorithm*, respectively, and the supported visualization and query types of each algorithm in column *Visualization Types* and *Query Types*

Paper	Algorithm	Visualization Types				Query Types					
		Bar	Pie	Line	Heatmap	GROUP BY	WHERE	ORDER BY	COUNT	SUM	AVG
Sample+Seek [24]	Uniform Sampling & Measure-biased Sampling	✓	✓	×	×	✓	✓	×	✓	✓	✓
Pangloss [25]	Optimistic Visualization Based on AQP	✓	×	×	✓	×	×	×	×	×	×
SampleAction [27]	Sampling-based Incremental Visualization	✓	×	✓	×	✓	✓	✓	✓	✓	✓
INCVISAGE [28]	Sampling-based Incremental Visualization	×	×	✓	✓	✓	✓	✓	✓	✓	✓
IFOCUS [26]	IFOCUS Algorithm	✓	×	✓	✓	✓	✓	×	✓	✓	✓
PFunk-H [99]	Human Perceptual Model with Sampling-based AQP	✓	×	×	×	×	×	×	✓	✓	✓

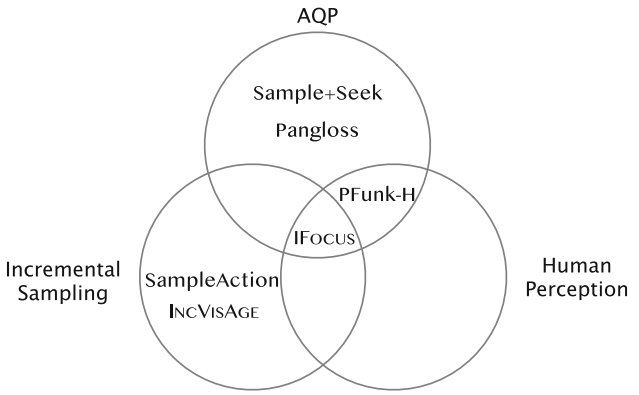


Fig. 7 A classification of approximate data visualization, where the surveyed works are classified as AQP, incremental sampling, and human perception methods

two works [24,25] that mainly focus on the sampling-based AQP techniques.

Sample+Seek [24] is an AQP system for answering visualizations generated from aggregation queries in an interactive speed, and the visualization results are within an error bound specified by users. It first presents the concept of *distribution precision* (e.g., distance between the approximate and exact visualizations) that can represent the precision of total distribution across aggregate groups. Thus, users can specify a *distribution precision* as an error bound. When sampling, for those queries with large data volumes, it uses the *uniform sampling* to answer the `COUNT` aggregation queries and proposes a *measure-biased sampling* technique for approximately answering `SUM` aggregation queries with less predicates, and the key feature of *measure-biased sampling* is to select the rows with probability proportional to its value on the aggregation attribute. Sample+Seek proposes two indexing techniques to speedup sampling: *measure-augmented inverted index* for indexing the categorical dimension to answer the aggregation queries; and *low-frequency group index* for supporting those queries with a conjunction of one or more equi-constraints.

Although there exists a significant difference between approximate and accurate visualizations with a small possibility, users may get frustrated with the visualization tools once a big difference happens. Thus, Pangloss [25], a web-based system powered by Sample+Seek [24], is designed to provide users with approximate visualizations together with exact visualizations. Pangloss provides users with approximate visualizations quickly based on the technique of AQP and then the system still computes the exact results in the background if users click the "remember" button for this visualization. In Pangloss, users can get initial insights from the approximate results and later verify their observations on the precise results.

Incremental Sampling-based Some works [26–28] try to link incremental data query techniques to data visualization. The key idea of approximate visualization with incremental sampling is that the system generates an approximate visualization based on representative samples of dataset rapidly. Then, the system increases the sample size over time to continuously improve the quality of visualizations. The user usually can get some initial insights from the approximate visualizations and decide to terminate if the quality of the visualization is enough to verify these insights.

SampleAction [27] is a tool for visualizing aggregation queries on very large datasets. Given a query, SampleAction rapidly responds to users with partial aggregation results for each group with error bounds (i.e., a bar chart with confidence bounds) based on fixed samples. As the users are waiting, it will narrow its error bound and incrementally improve the visualizations by increasing sampling size in every second.

There may exist significant fluctuations between the adjacent incremental approximate visualizations due to the random sampling in SampleAction [27], which may mislead users during the incremental approximation process. Thus, IncVisAge [28] is designed to solve this problem. IncVisAge [28] is a web-based system, which provides incremental approximate visualizations, typically supporting trendline and heatmap. And there are no significant fluctuations compared with SampleAction [27] during the process of visualization refinement due to the design of the *ISplit* algorithm, thus providing meaningful intermediate visualizations for users. In IncVisAge, a trendline displays aggregation results for all groups, and a segment denotes multiple successive groups of the trendline together with the same approximate aggregation value. A trendline is first initialized as a segment with all groups, and during the iterative process, the *ISplit* algorithm chooses one segment and splits the segment into two segments until there is no segment to split (i.e., all segments have only one group). During each iteration, the *ISplit* algorithm calculates the sampling number for given δ (failure probability) and ϵ (guarantee of error) and then chooses the best segment to be split based on the sampled data.

Figure 8 illustrates what approximate visualizations might be generated by SampleAction and IncVisAge, where significant fluctuation exists during two adjacent visualizations generated by SampleAction, while IncVisAge keeps stable updating. For example, in Fig. 8-①, the visualizations generated by SampleAction at t_1 and t_2 have very different trends. But, IncVisAge only splits one segment to update visualizations and thus keeps stable updating.

Human Perception-based At times, increasing the sample size does not always improve the quality of the visualization. The external reason is that the number of pixels of the screen is finite, and the internal reason is that the cognitive limitations of human perception in identifying small details. Therefore, it is possible for approximate visualization sys-

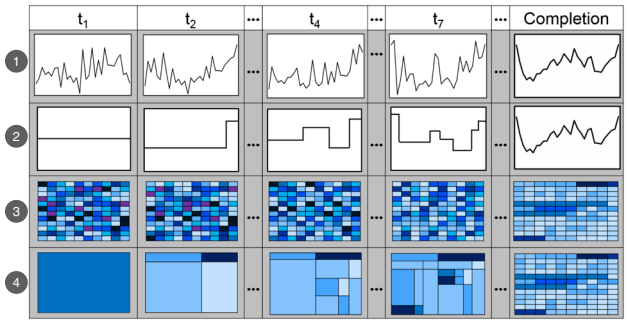


Fig. 8 A schematic diagram (not generated by real data) showing what approximate visualizations might generate by SampleAction and IncVisAge as time progresses (i.e., more samples are sampled) [28]. ① and ② are approximate lines generated by SampleAction and IncVisAge, respectively, using the same data, and ③ and ④ are approximate heatmaps generated by SampleAction and IncVisAge, respectively, using the same data

tems to generate approximate results based on representative samples but with minimal impact on the quality of visualization. Human perception-based approaches stop sampling when there is no obvious difference on human perception between the current approximate visualization and the visualization which is to get by further sampling.

IFocus [26], an online sampling algorithm, can generate an approximate bar chart rapidly and guarantee the pairwise ordering of each bar in a bar chart, because the pairwise ordering in bar charts is important human perceptual focuses. IFocus iteratively draws a sample for each *active* group (all groups are *active* at first) and maintains a confidence interval for each *active* group. Once a confidence interval of a bar has no overlap with other bars, meaning the order of this bar is determined, i.e., this bar is not *active* any more. The algorithm terminates until all bars are not *active*, and the approximate visualization results with ordering guarantee are returned to users.

PFunk-H [99] addresses the approximate visualization as the human perception problem. The basic idea of this work is that to combine sampling-based AQP techniques and human perception limitation together to provide approximate visualization in order to satisfy human perception. PFunk-H is an online sampling algorithm that provides approximate visualizations using perceptual functions from graphical perception. It presents an algorithm that can learn the knowledge of human perception error to provide approximate visualizations with perceptually indiscernible error. Besides, it can provide error bound of approximated query results under the restriction of perceptual function.

3.3 Progressive data visualization

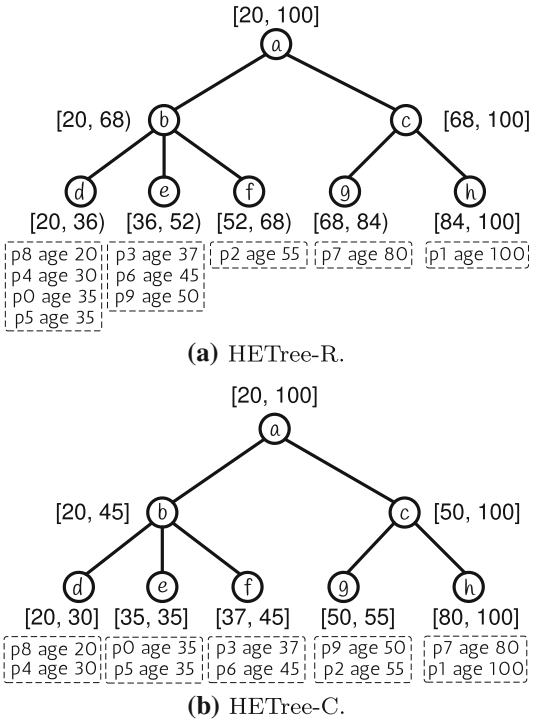
Many works [26–28] in approximate data visualization (Sect. 3.2) produce progressive visualization results to users. Besides, the above incremental sampling-based progressive data visualization, there have also been many works [31, 93,100,101] which provide progressive visualizations by hierarchical aggregation. Generally speaking, they build a hierarchical structure by aggregating the data in different levels, for example, different sizes of bins, different ranges of temporal values, different zones of spatial values. Then, these hierarchical structures are used to support users' progressive visualization exploration.

Range-Based Binning imMens [31] provides visualizations of different resolutions by changing bin sizes. The bins of the same resolution have equal ranges. Multi-dimensional data in imMens are partitioned into data cubes, and cubes are partitioned into tiles of different levels. Users can explore data in different levels and change current explored visualizations' resolution by zooming in or zooming out, and then, the system will change the underlying aggregation bin size correspondingly. imMens bins numeric data by equal ranges, which has some limitations. For example, considering an examination transcript dataset, binning by equal ranges of score is not applicable if teachers want to know the top-10, top 10 to 20, top 20 to 30 students, etc. Also, users cannot change the bin size or number of different resolution in imMens.

Range and Content-Based Binning The work of [100] provides two tree-structures for hierarchical exploration: HETree-R (Range-based HETree) and HETree-C (Content-based HETree). HETree-R is similar with imMens, and the leaf nodes of HETree-R denote data points within equal width ranges, while HETree-C has the same number of data points in all leaf nodes. Thus, the HETree-C can be used in the above examination transcript scenario. Users can explore the data abstraction or data details by a roll-up or drill-down operation to reach the upper or next level. It provides incremental tree construction algorithms based on user interaction and exploration scenarios (top–down or bottom–up). The algorithms automatically determine the proper arguments of the tree, i.e., the height of the tree, the range of the leaf, the number of children, etc. Also, the tree construction algorithms are adaptive and can fit to users' preference selection for parameters of tree to support better user experience. Figure 9 is a binning example of HETree-R and HETree-C.

4 Visualization recommendation

Recall that in Fig. 1, the data visualization process is iterative, and the main pain point of practitioners is that they have to be involved in each step to make some modifications. Naturally, it is highly desirable that there can have some visualization recommendation solutions that make the lives of users easier, by recommending (possibly) good visualizations to them.



(a) HETree-R.

(b) HETree-C.

Fig. 9 A binning example of HETree-R and HETree-C. Ten points (p0–p9) are binned by attribute age. HETree-R bins data by equal age ranges: each leaf is a bin with size 16, and HETree-C bins data by equal number of points: each leaf is a bin with 2 points

The rest of this section will be organized as follows.

- Specification-based recommendations. (Section 4.1)
  - The specification is incomplete, i.e., empty or partial specification of visualization elements. (Section 4.1.1)
  - The specification is treated as a reference. (Section 4.1.2)
- Behavior-based recommendations. (Section 4.2)
- Personalized recommendations. (Section 4.3)

Solution Overview Generally speaking, for solving all the above problems, the visualization recommendation systems need to first enumerate all possible visualizations and then recommend top-ranked visualizations.

Note that the search space of all visualizations is huge, which needs to consider the combination of several factors, such as choosing the columns to be visualized, transforming the data (e.g., group or bin), choosing the right visual encodings including mark types (e.g., bar, line, point), and encoding types for the selected mark chart (e.g., width of bar, position of point).

Pruning Meaningless Visualizations Fortunately, there are many signals (or constraints)—either from the users or from traditional wisdom—that can be used to prune "bad" visualizations.

- *User-specified constraints.* Users can specify interested visualization elements such as columns or data records.
  - SeeDB [22] stipulates that user should specify an interested query Q to obtain the target data before recommendation.
  - In Voyager [41,83,110], users should first specify interested variables too.
- *Expert provided constraints.* Some combinations of variables, transformations, and visual encodings may not generate a valid visualization. For example, the mark-type "pie" cannot combine with the encoding-type "height", and the encoding-type "Y-axis" is not suitable for categorical attributes. These constraints are typically given by experts.
  - Voyager [41,83,110] develops a permitted combination table of different data types, encoding types and mark types.
  - DeepEye [17,21,32,37] defines a set of rules (Table 5) to generate meaningful visualizations. For example, the first transformation rule in Table 5 denotes that if X-axis of a visualization is categorical, and Y-axis is numerical, and then the transformation operation should be grouping by X-axis, and aggregating on Y-axis.
  - Draco [35] has hard constraints and soft constraints, where hard constraints must be satisfied when generating visualizations (e.g., the encoding-type "shape" is not applicable for numeric values), and soft constraints are used to rank visualizations (e.g., it is better to use the encoding-type X-axis for temporal values).

After generating candidate (or valid) visualizations by pruning the entire search space as described above, visualization recommendation systems will then recognize meaningful visualizations based on predefined metrics or rules. Some systems may also rank interesting visualizations or recommend top-k visualizations to users. In the rest of this section, we will discuss these methods to solve the above problems.

4.1 Specification-based recommendations

4.1.1 Incomplete specification

Visualization recommendation systems with *empty specification* require no user inputs, while recommendation systems with *partial specification* accept users' partial visualization elements specification inputs for desired visualization. For example, APT [88] accepts user's data viewing goals before

Table 5 Constraints in DeepEye (**T** denotes the data type, **AGG** denotes the aggregation function, including **AVG** (average), **SUM** (sum), **CNT** (count), and X, Y denote attributes for X- and Y- axes, respectively). The transformation rules define how to transform the data when the data types of the X- and Y- axes are given. The sorting rules define how to sort the data. And the visualization rules define how to choose the right visualization types for different data types of X- and Y- axes

Transformation Rules
*$\mathbf{T}(X)$ = Categorical, $\mathbf{T}(Y)$ = Numerical → **GROUP**(X), *$\mathbf{AGG}(Y)$.
$\mathbf{T}(X)$ = Categorical, $\mathbf{T}(Y) \neq$ Numerical → **GROUP**(X), **CNT**(Y).
$\mathbf{T}(X)$ = Numerical, $\mathbf{T}(Y)$ = Numerical → **BIN**(X), **AGG**(Y).
$\mathbf{T}(X)$ = Numerical, $\mathbf{T}(Y) \neq$ Numerical → **BIN**(X), **CNT**(Y).
$\mathbf{T}(X)$ = Temporal, $\mathbf{T}(Y)$ = Numerical → **GROUP/BIN**(X), **AGG**(Y).
$\mathbf{T}(X)$ = Temporal, $\mathbf{T}(Y) \neq$ Numerical → **GROUP/BIN**(X), **CNT**(Y).
Sorting Rules
$\mathbf{T}(X)$ = Numerical/Temporal → **ORDER BY**(X).
$\mathbf{T}(Y)$ = Numerical → **ORDER BY**(Y).
Visualization Rules
$\mathbf{T}(X)$ = Categorical, $\mathbf{T}(Y)$ = Numerical → **BAR/PIE**.
$\mathbf{T}(X)$ = Numerical, $\mathbf{T}(Y)$ = Numerical → **LINE/BAR**.
$\mathbf{T}(X)$ = Numerical, $\mathbf{T}(Y)$ = Numerical, (X, Y) correlated → **SCATTER**.
$\mathbf{T}(X)$ = Temporal, $\mathbf{T}(Y)$ = Numerical → **LINE**.
*$\mathbf{T}$ = Type, **AGG** = {**AVG**, **SUM**, **CNT**}

recommendation. Users should first choose one interested column before visualization recommendation in Voyager [41,83,110]. DeepEye [17,21,32,37] accepts users' keyword specification, e.g., the user may input "*show me line charts about electricity*".

The only difference between empty and partial specification is that the latter should prune the search space by the user-specified constraints when enumerating visualization elements to generate visualization candidates. For example, the visualizations which do not contain column c are filtered from the visualization candidates if users specify column c as the interested column in Voyager, and the visualizations which are not line charts or do not contain the column "*electricity*" are filtered from the visualization candidates when users type "*show me line charts about electricity*" in DeepEye. And there is no difference when ranking candidate visualizations.

In the remaining part of this section, we describe two common methods used to rank the visualization candidates: rule-based solution and machine learning-based solution.

Rule-based visualization ranking

Most earlier works (APT [88], SAGE [111], BOZ [112]) on visualization recommendations are rule-based, which are inspired by the work of [88,113–116]. Rule-based recommendation systems rank the visualization candidates by their predefined rules, which are usually human perceptual effectiveness metrics, measured as an effectiveness score s considering data type, statistical information, human visual preference, etc. For example, a pie chart consists of many blocks (e.g., $>$ 500) is not a good visualization by human perception, because it is too messy. The main difference in the rule-based recommendation systems is the definition of s. Voyager [41,83,110], Show Me [39], Polaris [20], DIVE [117], DeepEye [17,21,32,37], Wang et al. [40] develop richer perceptual rules with more data types, mark types, statistical information compared with former works, while Rank-by-feature [38] ranks visualizations by a single statistical metric.

Statistical Rules Rank-by-feature framework [38] is a statistical rule-based recommendation system. It can rank 1D or 2D axis-parallel projection visualizations (histograms, boxplots, and scatterplots) to users by different statistical ranking metrics. The metrics for 1D ranking (histograms and boxplots) include normality or uniformity of the distribution, number of potential outliers or unique values, and size of the biggest gap, and the metrics for 2D ranking (scatterplots) include correlation coefficient, number of potential outliers, uniformity of scatterplots, etc. By discovering these ranked low-dimensional visualizations, users may find complex relations, clusters, outliers, and so on.

Perceptual Rules The Rank-by-feature framework can only rank between the same visualization type (e.g., histograms, boxplots) by a single statistical metric, while Voyager [41,83,110] ranks different visualization types by a perceptual effectiveness score s considering data type, cardinality, human visual preference, and so on. For example, high cardinality variables should not be mapped to color; visualizations with less screen space are preferred. s is a weighted sum of these factors, and the weight of these factors is manually determined through tests and experiments. Table 6 shows some perceptual effectiveness ranking rules used in Voyager; for example, the third row in Table 6 means that if the data types of X- and Y- axes are temporal and numerical, respectively, then line chart is the best choice, and bar, point, text types

Table 6 Ranking rules in Voyager

$T(X)$	$T(Y)$	Mark Type
Categorical	Categorical	point > text
Categorical	Numerical	bar > point > text
Temporal	Numerical	line > bar > point > text
Numerical	Numerical	point > text

T(X) and *T(Y)* denote the data type of X- and Y- axes, respectively, and *Mark Type* denotes the permitted ranked mark types for this data type correspondingly

ranked behind line. And the rules together with users' input for column preference form the ranking metric in Voyager.

DeepEye [17,21,32,37] captures human perceptual effectiveness in richer details than Voyager. DeepEye defines three factors to describe the quality of a visualization and then develops a partial-order-based solution to rank all the valid visualization candidates. The three factors are: ① the matching quality between data and chart; ② the quality of transformations; and ③ the importance of columns. A visualization precedes another if all of the three factors are greater than another. And based on the partial relation, DeepEye can construct a graph $G(V, E)$, where V denotes the all valid visualizations and E denotes the partial orders. Then, DeepEye ranks all the visualizations in a way similar to topological sorting.

The above two works (Voyager [41,83,110] and DeepEye [17,21,32,37]) consider common visualization types, while Wang et al. [40] propose an algorithm to automatically pick line graph or scatter plot for time series. Although people may use line graphs to visualize time series in most cases, scatter plots are better choices sometimes. For example, the scatter chart provides a clearer trend than line chart when there are many outliers in time series. The algorithm first constructs line graph, scatter plot, and a trend curve by LOESS regression [118], then calculates the visual consistency between the trend curve and the line graph or scatter plot, respectively, and picks the visualization type which has bigger visual consistency (i.e., smaller distance) with the trend curve, where the visual consistency is achieved by comparing the consistency (i.e., distance) of visualizations' density fields, which can be calculated by KDE [119] algorithm. Experiments have shown that the choices of the algorithm are consistent with users in most cases.

Machine learning-based visualization ranking

With the rapid development of machine learning and deep learning, more and more systems [21,32,35,84,120] focus on machine learning-based visualization recommendation. Given two visualizations u and v, the systems should determine which is better. Typically speaking, machine learning-based recommendation systems first collect training data, which comes from crowdsourcing or web, then train a ranking model which takes the input space $\mathcal{X}$ as lists of feature vectors, and $\mathcal{Y}$ the output space consisting of grades (or ranks). The model learns a function $F(\cdot)$ from the training examples, such that given two input vectors $\mathbf{x_1}$ and $\mathbf{x_2}$, it can determine which one is better, $F(\mathbf{x_1})$ or $F(\mathbf{x_2})$.

Learning with Soft Constraints Draco [35] expresses preferences by soft constraints, and the soft constraints (e.g., it is better for temporal values to use the encoding type: X-axis) are specified by human perceptions. Each soft constraint has a weight denoting the penalty when a visualization v violates the soft constraint. And the overall cost of V is: $cost(v) = \sum_{i=1}^{n} w_i \cdot n_i, n_i \in \{0, 1\}$, where w_i is the weight of the ith constraint, and n_i denotes that V violates the ith constraint n_i times. Draco prefers visualizations with less cost and formulates the problem of learning weights as a *learning-to-rank* [121] problem using RankSVM model [122]. Draco gets 1110 ranked visualization pairs from crowdsourcing, and the ranking principles are from Kim et al. [123] and Saket et al. [124]. Draco is similar to Voyager, but Draco differs with Voyager in that it learns the weight of constraints by machine learning techniques.

Learning with Examples The constraints in Draco are predefined to the system by users or developers, rather than learned by machines. In contrast, DeepEye [21,32] develops a machine learning-based solution which captures visualization design knowledge automatically by learning from examples besides the above rule-based solution.

DeepEye [21,32] captures human perception by learning from examples and supposes the models learned from former examples can be extended to different domains. DeepEye identifies 12 features: statistical information (cardinality, distinct numbers, max, min, correlation, etc.) and mark type. DeepEye uses a binary classifier (decision tree [125]) to determine whether a visualization is good or bad, which is called *Visualization Recognition*, then use a *learning-to-rank* [121] model (LambdaMART algorithm [126]) to score all good visualizations, which is called *Visualization Ranking*. DeepEye collects 42 real-world datasets from various domains and then picks 285, 236 visualization comparisons over these datasets labeled by 100 students.

Figure 10 shows the architecture of DeepEye. A user can pose a keyword query to VSE (Visualization Search Engine) module (Fig. 10-②), and VSE returns ranked visualizations to users. The VSE first translates the keyword query to multiple visualization candidates (keyword-to-visualizations) by querying database (Fig. 10-③), then discovers good visualizations (visualization transformation) and ranks them (visualization ranking). The crawler (Fig. 10-④) extracts training data for visualization transformation and ranking of VSE. In the client (Fig. 10-①), users can input keyword,

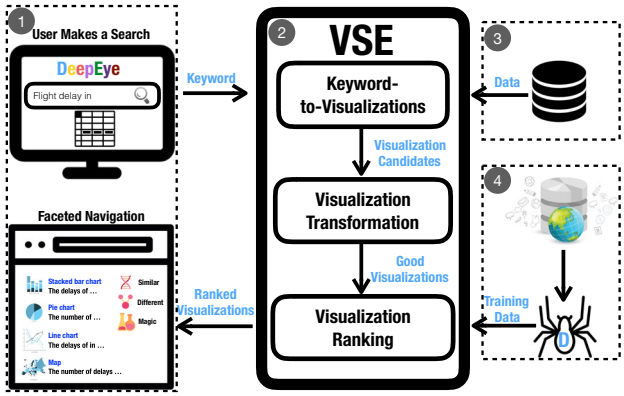


Fig. 10 The architecture of DeepEye [21,32]. User can post a keyword search in ①, and ② generates visualization candidates by querying ③, then ② returns ranked visualizations to user using the model trained by data from ④

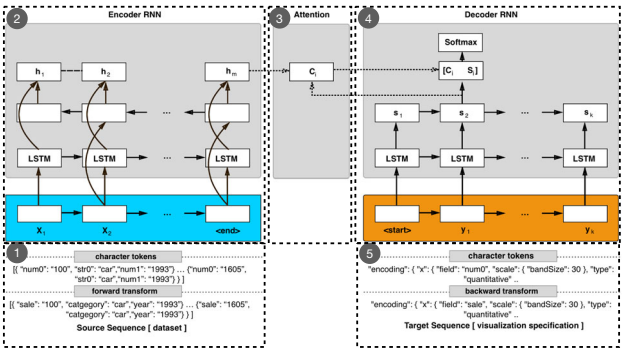


Fig. 11 The architecture of Data2Vis [120]. Data2Vis is built based on sequence to sequence model with the encoder–decoder architecture (② and ④) and the attention mechanism (③). It takes original datasets (①) as input and automatically recommends visualizations (⑤) by given datasets (①)

interact with recommended visualizations, and do faceted navigation, etc.

Data2Vis [120] is an attempt in generating visualizations using RNN (recurrent neural network [127]). As shown in Fig. 11, Data2Vis treats visualization design as a sequence to sequence [128,129] translation problem, where the input string is a dataset in JSON format (Fig. 11-①) and the output string is a Vega-Lite [2] visualization specification (Fig. 11-⑤). Data2Vis trained a model with a 2-layer RNN encoder (Fig. 11-②) and a 2-layer RNN decoder (Fig. 11-④), and both have 256 LSTM (Long Short-Term Memory [130,131]) cells. The training dataset has 4300 training instances [132]. Experiments have shown that Data2Vis can generate visualizations with appropriate mark types (e.g., use scatter for two numeric attributes), transformations (e.g., use means for numeric attributes), selection patterns (e.g., select data by country, gender), etc.

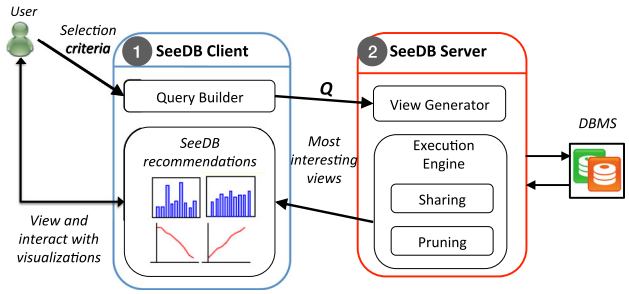


Fig. 12 The architecture of SeeDB [22]. Client (①) accepts users' input, constructs visualization queries and shows recommended visualizations to users. Server (②) generates visualization candidates (view generator) and recommends visualizations to users (execution engine)

4.1.2 Reference-based specification

Some visualization recommendation systems recommend visualizations based on reference data or reference visualizations [18,22,36,85]. Typically, the system would recommend visualizations which are similar to or different from the given reference in certain aspects.

Deviation-based SeeDB [22] recommends visualizations by deviation with some reference visualizations. Before recommendation, the user should specify an interested query Q to obtain the target data which is called D_Q and a reference dataset D_R is also needed. Then, SeeDB enumerates different combinations of same variables, transformation, mark types, encoding types on both D_Q and D_R to get all $V(D_Q)$ and $V(D_R)$, respectively. Finally, SeeDB recommends the top-k $V(D_Q)$ which have the largest value $S(P[V(D_Q)], P[V(D_R)])$, where S is a distance function, and $P[V(D_Q)]$ and $P[V(D_R)]$ are the distribution of $V(D_Q)$ and $V(D_R)$, respectively. Figure 12 shows the architecture of SeeDB.

Anomaly-based Profiler [85] recommends visualizations which can best distinguish anomalies in the primary visualization. The tuples in the primary visualization are classified by some anomaly detection methods: normal points and abnormal points are in different classes, denoted as a column $class$. Suppose $VisToCol(V)$ is a function which returns a column that describes the classes of each tuple in the visualization V, then Profiler recommends the visualization V that minimizes $D(VisToCol(V), class)$. $D(X, Y)$ is a distance function measuring the independence between X and Y:

$$D(X, Y) = 1 - \left(\frac{I(X, Y)}{\max\left(H(X), H(Y)\right)} \right) \tag{1}$$

where $I(X, Y)$ denotes mutual information of X and Y, quantifying the reduction of predicting one variable when another is given, and $H(X)$, $H(Y)$ denote entropies of X, Y, respectively.

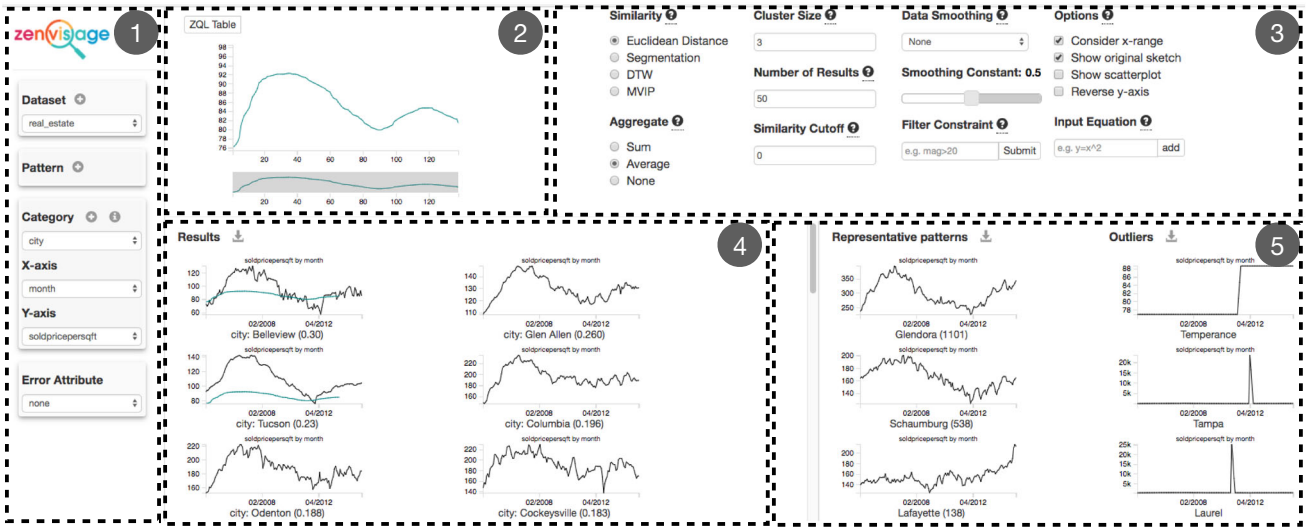


Fig. 13 The front end of Zenvisage [18,36]. The user can upload their datasets and select the X-axis, Y-axis, and category for visualizations in ①. After that, Zenvisage first recommends representative (i.e., typical trends) and outlier trends in ⑤ according to the settings of ①. ④ will also show those visualizations that are similar to the reference one in ②. The user can specify some system parameters, e.g., similarity functions and aggregation functions, in ③

Similarity/Distance-based Zenvisage [18,36] tries to find other interesting visualizations when the users provide their desired trends, patterns, or insights. Users can draw their desired trends or patterns as a visualization V, then the system recommends visualizations V' by their similarity or dissimilarity (specified by users) with V, i.e., recommends V' with largest or smallest $S(V, V')$, where S is a distance function. Thus, the definition of the distance function is of great importance. The distance functions used by Zenvisage are Euclidean distance and Dynamic Time Warping [133]. Figure 13 depicts the front end of the Zenvisage.

4.2 Behavior-based recommendations

Behavior-based recommendation systems capture users' current behavior as inputs, then infer users' intended task and recommend useful visualizations based on their tasks.

HARVEST [134] is a behavior-driven visualization recommendation system. It recommends visualizations based on the tasks of users which are inferred by their behavior. Since it is difficult for a user to describe her intent clearly and the task of a user evolves as the process of the exploration, HARVEST guesses users' intent by their behavior. As shown in Fig. 14, when the user interacts with HARVEST in the front end, it captures user's action α_i and sends it to the back-end. There are several common atomic actions studied by HARVEST: `inspect`, `filter`, and `bookmark`. The atomic actions form a complex pattern, which usually indicates specific intents. Next, the Action Tracker module (Fig. 14-②) analyzes and outputs the user's task context c_i and pattern p_i based on accepted actions. More concretely, c_i denotes

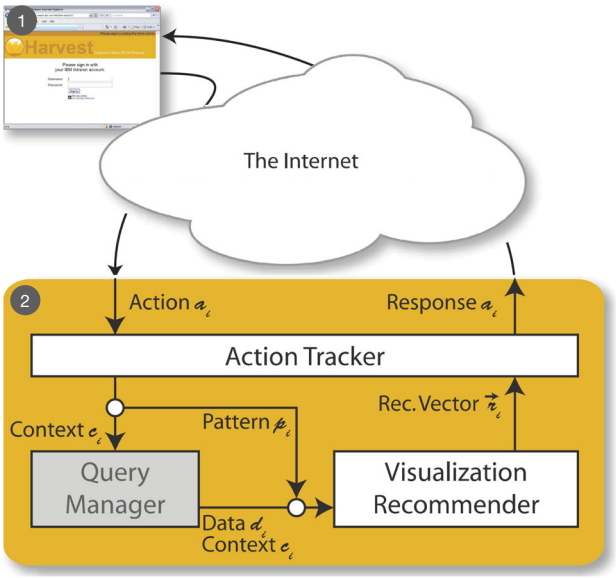


Fig. 14 The architecture of HARVEST [120]. The front end is a web-based user interface (①). The user can specify the dataset and create data visualizations on the front end. The back-end (②) accepts users' actions, detects patterns, and recommends relevant visualizations by users' task intents, which is inferred by their behavior.

the constraints on current data and p_i indicates user's task intents. HARVEST defines 4 patterns: `scan`, `flip`, `swap`, and `drill-down`. Then, HARVEST can recommend visualizations by the inferred patterns of users. For example, if a user iteratively `inspects` hotel price of different regions in a map, HARVEST can detect a `scan` pattern, which means that users want to compare some attributes between some

similar objects, thus HARVEST may recommend a bar chart showing the comparison of hotel price of different regions.

4.3 Personalized recommendations

Personalized recommendation systems capture users' historical behavior as inputs to recommend personalized interesting visualizations.

Linear Model VizDeck [84] provides personalized visualization recommendation results by training a linear model for each user using their historical behavior. VizDeck provides a new interface design which displays top-k recommended visualizations to users in a grid. The elements displayed in the grid are called vizlets. Users can browse, drop, promote or reorder the vizlets, and the final selected vizlets will be displayed on an interactive dashboard. VizDeck can get users' visualization preference by their historical behavior during the exploration, extract features for these vizlets, and then train a linear model which score vizlets for future recommendation.

Collaborative Filtering Besides training a model for each user, there are many other techniques [135–137] in personalized recommendation systems. For example, collaborative filtering (CF) [135] is a widely used personalized recommendation algorithm. Based on CF, VizRec [138] proposes three methods for personalized visualization recommendation.

1. *Collaborative Filtering.* VizRec constructs an $m \times n$ matrix A, where $A[i][j]$ denotes the rating (e.g., 1 to 7) of the user i on the visualization j in the past. And for a given user u, VizRec first calculates the top-k users who are the most similar (i.e., their ratings for different visualizations are similar.) with u by Pearson correlation coefficient, denoted as $u_1, u_2, ..., u_k$. Then, for a visualization candidate v, the rating of u on v is calculated as:

$$r_u^v = \hat{r_u} + \frac{\sum_{i=1,2,...,k} sim(u, u_i)(A[u_i][v] - \hat{r_{u_i}})}{\sum_{i=1,2,...,k} sim(u, u_i)} \quad (2)$$

   where $\hat{r_u}$ and $\hat{r_{u_i}}$ denote the average rating of user u and u_i, respectively.
2. *Content-based Filtering.* For the users who are new to the system, CF-based recommendation is not applicable. Thus, VizRec also develops a content-based recommendation. VizRec defines many features (e.g., attributes of given datasets and mark types) to characterize users and visualizations and uses the frequency of features to construct user and visualization profiles. VizRec constructs the profile of a user by her current (together historical for old users) annotations of visualizations. And VizRec builds the visualization profile by the aggregation of user profiles for this visualization. Then, the user and visualization profiles are transformed to vectors by TF-IDF (Term Frequency-Inverse Document Frequency). And for a given user u and a visualization recommendation candidate v, the similarity of u's and v's vectors can be considered as the ranking score between u and v.
3. *Hybrid Filtering.* A hybrid method of the above two methods will bring a host of benefits (e.g., the algorithm becomes adaptive when the users' interest changed). VizRec uses the weighted sum of the normalized scores of the above two methods as the hybrid filtering score.

4.4 A summary

Table 7 shows a summary of the supported visualization types, input, and ranking metric of the above visualization recommendation systems.

Visualization recommendation systems with empty specification, such as Draco [35], Data2Vis [120] and Rank-by-feature [38], are helpful for users to quickly explore the data when the users are not very familiar with data and desired visualizations. Most of the existing recommendation systems require partial specification, because they permit users' specification for desired visualizations as inputs, e.g., keyword specification in DeepEye [17,32]. Rule-based solution is in line with person's intuitive understanding of visualizations, but it does not make a complete understanding of human perceptions, just focusing on several interested metrics. Machine learning-based solutions need to collect training data, and the results are hard to interpret, but it may well capture human's cognitive knowledge about visualization effectiveness. And the learning model will become smarter when more training data are collected.

Users should specify the reference data or desired pattern in the reference-based visualization recommendation, which may be difficult for users who are not familiar with the original data and want to explore the data with the help of the recommendation systems. The advantage is that it is easy to develop such a system, and convenient when users are clear about their needs, e.g., find a line chart with desired trend with Zenvisage [18,36].

Behavior-based recommendations can recommend visualizations based on inferred tasks, but are limited to the predefined behavior patterns, making it not flexible for users' random behavior.

Personalized recommendations perform differently for different users, because personalized recommendations are customized for different users by their historical behavior.

Besides the above works, there is also a preliminary design of a framework [110] which uses a language, CompassQL, to describe different ranking metrics. CompassQL is a general framework, aiming to describe the ranking metrics of SeeDB [22], Voyager [41,83,110], VizDeck [84], etc. But, there is no implementation of CompassQL yet.

Table 7 A summary of visualization recommendation systems, where *Visualization Types* is the supported visualization recommendation types; *Input* is the input of this recommendation system; *Ranking Metric* is the main ranking strategy of this recommendation system

Category	Visualization Recommendation System	Visualization Types				Input	Ranking Metric
		Bar	Pie	Line	Scatter		
Specification-based							
Incomplete Specification	Draco [35]	√	√	√	√	–	RankSVM Model
	Data2Vis [120]	√	√	√	√	–	RNN
	APT [88]	√	√	√	√	Keyword	Perceptual Rules
	SAGE [111]	√	√	√	√	Keyword	Perceptual Rules
	Voyager [83,110]	√	√	√	√	Columns	Perceptual Rules
	Rank-by-feature [38]	√	×	×	√	–	Statistical Rules
	Polaris [20]	√	√	√	√	Columns	Perceptual Rules
	Show Me [39]	√	√	√	√	Columns	Perceptual Rules
	DeepEye [17,32]	√	√	√	√	Keyword	LambdaMART Algorithm
	Wang et al. [40]	×	×	√	√	Time Series	Perceptual Rules
Reference-based Specification	SeeDB [22]	√	√	√	√	Query	$S(P[V(D_Q)], P[V(D_R)])$
	Profiler [85]	√	√	√	√	Visualization	$D(VisToCol(v), class)$
	Zenvisage [18,36]	√	√	√	√	Visualization	$S(V, V')$
Behavior-based	HARVEST [134]	√	√	√	√	Current Behavior	Task Driven
Personalized	VizDeck [84]	√	√	√	√	Historical Voting	Linear Model
	VizRec [138]	√	×	√	√	Historical Rating	CF

5 Other research directions

In this section, we will discuss other research topics that are also relevant to data management issues, but are not yet well studied.

5.1 Data preparation for data visualization

Real-life data are typically dirty, and visualizing dirty data may mislead users. This phenomenon has been known for a long time as one type of biased visualizations from the data visualization community. For example, a dataset that is integrated from multiple sources may contain duplicates. Naturally, the data being visualized should be cleaned, such as value normalization, deduplication, missing value imputation, and outlier detection. Tableau has integrated Trifacta for data preparation over the entire dataset. The following studies have been conducted, from both database community and visualization community, to investigate the impact of dirty data on data visualization.

- *What-if Analysis for Outliers*: Scorpion [139] allows users to manually pinpoint the outliers from the result of an aggregation query. It then tries to find and remove the predicate that causes such outliers, without affecting the other non-outliers. The problem was formulated as an *influential predicates problem* and was solved by using techniques from *sensitivity analysis*. As a result, Scorpion can automatically move away outliers identified by users.
- *Evaluating Visualizations with Missing Data*: [140] did a crowdsourced study to measure factors influencing response accuracy, data quality, and confidence in interpretation for time series data with missing values. In particular, it tries a combination of three imputation methods (1) zero-filling, (2) marginal mean, and (3) linear interpolation with four ways of showing the imputed values (i) highlight, (ii) downplay, (iii) annotation, and (iv) information removal. The evaluation over two real-world datasets with 300+ crowd users partially verifies the following hypotheses: (I) Perceived data quality and response accuracy will both degrade as the amount of missing data increases. (II) Highlighting methods will generate higher perceived data quality than downplaying and information removal methods. (III) Linear interpolation will lead to higher perceived confidence and data quality than marginal means or zero-filling as it takes into account local trends in dataset. (IV) Imputed values will lead to higher perceived data quality than removed values.

Research Opportunities

- *Detecting biased visualizations.* A seemingly good visualization might actually be biased; hence, it requires to detect such visualizations automatically. Many people have approached this problem from a statistics perspective. However, it is also important to study this problem, from the angle of dirty data.
- *Task-aware data cleaning.* Intuitively, it is easier to clean a dataset if the targeting task is known, such as only a small part of data needs to be cleaned, which is cheaper than cleaning the entire dataset in the conventional way.

5.2 Data visualization benchmarks

Like ImageNet or the classic TPC benchmarks, it is important to develop benchmarks for performance and recommendation. The benchmarks should be faithful to the visual analysis tasks, provide reusable traces and data, and in the case of recommendation, have high coverage and quality of its labels. There is an emerging focus on developing benchmarks for performance measures [141–143].

- A research work VizNet [144] has presented a large-scale corpus of over 31 million datasets compiled from open data repositories and online visualization galleries. It provides the necessary common baseline for comparing visualization design techniques, and developing benchmark models and algorithms for automating visual analysis.

Naturally, more needs to be done.

Research Opportunities

- *Categorization of visualizations.* For ImageNet, it is easy to set categories, such as "balloon" or "strawberry", because the classification task is easier. It is not clear about how to define similar categories for visualizations in a conceptual level, such as "trend" or "distribution".
- *Training data.* Assuming the categories can be provided, there remains a daunting task to label visualizations, and each visualization may have multiple labels. Afterward, it remains a hard problem on how to use these labeled data, e.g., using which machine learning or deep learning model to predict a good visualization for a given task.

5.3 Data visualization for database-related applications

As mentioned earlier in Sect. 1, data visualization also plays an important role in database-related applications, such as Excel [9], Google Sheets [10], Oracle Data Visualization Desktop [11], IBM Db2 [12], Amazon Quicksight [13],

Microsoft Power BI [5], and many others. Naturally, with the rapid development of visualization techniques, there are more opportunities about using data visualization for database-related applications.

Research Opportunities

- *Data visualization for data discovery.* Data discovery, the problem of finding interesting datasets for a certain application from a data lake with thousands or millions of data silos, remains a hard problem to solve [145]. One roadblock is to quickly understand the discovered datasets. Practically, browsing each dataset is time-consuming. Intuitively, data visualization that provides a high-level understanding can help in this important problem.
- *Data visualization for data debugging.* One problem that was recently raised by the Data Civilizer system [146,147] is data debugging, where the output of a data analytics workflow is wrong not because of bugs in programs, but in the data such as erroneous input or wrong parameters. Although [146] has some initial attempt to combine data visualization for data debugging, the solution for data debugging is far from being mature, and evidently, data visualization can help for more effective data debugging.

6 Conclusion

Data visualization is a fast growing field with a great many new research results and novel systems developed recently. Research and practitioners from many fields have contributed to the remarkable success of data visualization, which is driven by most (if not all) domains and applications.

This article mainly surveys recent data visualization works, from data management perspective. In particular, we have comprehensively described the works in visualization specifications, efficient methods for data visualization, and visualization recommendation. As mentioned earlier, most commercial data visualization systems are good at ease-of-use in terms of data visualization specifications. However, many practitioners are still suffering from the efficiency and recommendation issues of these systems. Hence, we also discuss several open problems that database researchers can make significant contribution to advance the field of data visualization.

Acknowledgements Funding was provided by 973 Program of China (Grant No. 2015CB358700) and National Natural Science Foundation of China (Grant Nos. 61632016, 61521002, 61661166012).

References

1. Michael, B., Vadim, O., Jeffrey, H.: D3: Data-driven documents. TVCG **17**(12), 2301–9 (2011)
2. Satyanarayan, A., Moritz, D., Wongsuphasawat, K., Heer, J.: Vega-lite: a grammar of interactive graphics. TVCG **23**(1), 341–350 (2016)
3. Hanrahan, P.: Vizql: a language for query, analysis and visualization. In: SIGMOD, p. 721 (2006)
4. Tableau. https://www.tableau.com. Accessed 31 Dec 2018
5. Power bi: Interactive data visualization bi tools. https://powerbi.microsoft.com. Accessed 31 Dec 2018
6. Hyper: A hybrid oltp and olap high performance dbms. https://hyper-db.de. Accessed 31 Dec 2018
7. Neumann, T., Mühlbauer, T., Kemper, A.: Fast serializable multi-version concurrency control for main-memory database systems. In: SIGMOD, pp. 677–689 (2015)
8. Neumann, T.: Efficiently compiling efficient query plans for modern hardware. PVLDB **4**(9), 539–550 (2011)
9. Microsoft excel. https://products.office.com/en-us/excel. Accessed 31 Dec 2018
10. Google sheets: Free online spreadsheets for personal use. https://www.google.com/sheets/about/. Accessed 31 Dec 2018
11. Oracle data visualization desktop. https://docs.oracle.com/en/middleware/bi/data-visualization-desktop/tutorials.html. Accessed 31 Dec 2018
12. Ibm db2. https://www.ibm.com/analytics/db2. Accessed 31 Dec 2018
13. Amazon quicksight: Cloud based business intelligence. https://aws.amazon.com/quicksight/. Accessed 31 Dec 2018
14. Vega: A visualization grammar. https://vega.github.io/vega/. Accessed 31 Dec 2018
15. Wickham, H.: ggplot2–elegant graphics for data analysis. J Comput. Graph. Stat. **19**(1), 3–28 (2009)
16. Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., Zu, M., Chen, W.: ECharts: A declarative framework for rapid construction of web-based visualization. Vis. Inform. **2**, 136–146 (2018)
17. Luo, Y., Qin, X., Tang, N., Li, G.: DeepEye: towards automatic data visualization. In: ICDE, pp. 101–112 (2018)
18. Siddiqui, T., Lee, J., Kim, A., Xue, E., Yu, X., Zou, S., Guo, L., Liu, C., Wang, C., Karahalios, K., Parameswaran, A.G.: Fast-forwarding to desired visualizations with zenvisage. In: CIDR (2017)
19. Kalinin, A., Cetintemel, U., Zdonik, S.: Interactive data exploration using semantic windows. In: SIGMOD, pp. 505–516 (2014)
20. Stolte, C., Hanrahan, P.: Polaris: a system for query, analysis and visualization of multi-dimensional relational databases. In: INFOVIS, pp. 5–14 (2000)
21. Qin, X., Luo, Y., Tang, N., Li, G.: DeepEye: an automatic big data visualization framework. Big Data Min. Anal. **1**(1), 75–82 (2018)
22. Vartak, M., Madden, S., Parameswaran, A., Polyzotis, N.: Seedb: automatically generating query visualizations. PVLDB **7**(13), 1581–1584 (2014)
23. Vartak, M., Rahman, S., Madden, S., Parameswaran, A.G., Polyzotis, N.: SeeDB: efficient data-driven visualization recommendations to support visual analytics. PVLDB **8**(13), 2182–2193 (2015)
24. Ding, B., Huang, S., Chaudhuri, S., Chakrabarti, K., Wang, C.: Sample + seek: approximating aggregates with distribution precision guarantee. In: SIGMOD, pp. 679–694 (2016)
25. Moritz, D., Fisher, D., Ding, B., Wang, C.: Trust, but verify: optimistic visualizations of approximate queries for exploring big data. In: CHI, pp. 2904–2915 (2017)
26. Kim, A., Blais, E., Parameswaran, A.G., Indyk, P., Madden, S., Rubinfeld, R.: Rapid sampling for visualizations with ordering guarantees. PVLDB **8**(5), 521–532 (2015)
27. Fisher, D., Popov, I., Drucker, S., Schraefel, M.: Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster. In: CHI, pp. 1673–1682 (2012)

28. Rahman, S., Aliakbarpour, M., Kong, H.K., Blais, E., Karahalios, K., Parameswaran, A., Rubinfield, R., Rahman, S., Aliakbarpour, M., Kong, H.K.: I've seen "enough": incrementally improving visualizations to support rapid decision making. PVLDB **10**(11), 1262–1273 (2017)
29. Wesley, R.M.G., Eldridge, M., Terlecki, P.: An analytic data engine for visualization in tableau. In: SIGMOD, pp. 1185–1194 (2011)
30. Wang, Z., Ferreira, N., Wei, Y., Bhaskar, A.S., Scheidegger, C.: Gaussian cubes: real-time modeling for visual exploration of large multidimensional datasets. TVCG **23**(1), 681–690 (2016)
31. Liu, Z., Jiang, B., Heer, J.: imMens: real-time visual querying of big data. In: Eurographics Conference on Visualization, pp. 421–430 (2013)
32. Luo, Y., Qin, X., Tang, N., Li, G., Wang, X.: DeepEye: creating good data visualizations by keyword search. In: SIGMOD, pp. 1733–1736 (2018)
33. Wu, E., Psallidas, F., Miao, Z., Zhang, H., Rettig, L.: Combining design and performance in a data visualization management system. In: CIDR (2017)
34. Doshi, P.R., Rundensteiner, E.A., Ward, M.O.: Prefetching for visual data exploration. In: DASFAA, pp. 195–202 (2003)
35. Moritz, D., Wang, C., Nelson, G.L., Lin, H., Smith, A.M., Howe, B., Heer, J.: Formalizing visualization design knowledge as constraints: actionable and extensible models in draco. TVCG **25**(1), 438–448 (2019)
36. Siddiqui, T., Kim, A., Lee, J., Karahalios, K., Parameswaran, A.G.: Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. PVLDB **10**(4), 457–468 (2016)
37. Qin, X., Luo, Y., Tang, N., Li, G.: DeepEye: visualizing your data by keyword search. In: EDBT Vision (2018)
38. Seo, J., Shneiderman, B.: A rank-by-feature framework for interactive exploration of multidimensional data. IV **4**(2), 96–113 (2005)
39. Mackinlay, J.D., Hanrahan, P., Stolte, C.: Show me: automatic presentation for visual analysis. TVCG **13**(6), 1137–1144 (2007)
40. Wang, Y., Han, F., Zhu, L., Deussen, O., Chen, B.: Line graph or scatter plot? Automatic selection of methods for visualizing trends in time series. TVCG **24**(2), 1141–1154 (2018)
41. Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J.D., Howe, B., Heer, J.: Voyager: exploratory analysis via faceted browsing of visualization recommendations. TVCG **22**(1), 649–658 (2016)
42. Kandel, S., Paepcke, A., Hellerstein, J., Heer, J.: Wrangler: interactive visual specification of data transformation scripts. In: CHI, pp. 3363–3372 (2011)
43. Von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J.J., Fekete, J.-D., Fellner, D.W.: Visual analysis of large graphs: state-of-the-art and future research challenges. Comput. Graph. Forum **30**, 1719–1749 (2011)
44. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: a survey. TVCG **6**(1), 24–43 (2000)
45. Beck, F., Burch, M., Diehl, S., Weiskopf, D.: A taxonomy and survey of dynamic graph visualization. Comput. Graph. Forum **36**(1), 133–159 (2017)
46. Bikakis, N., Sellis, T.: Exploration and visualization in the web of big linked data: a survey of the state of the art. arXiv preprint arXiv:1601.08059 (2016)
47. Marie, N., Gandon, F.: Survey of linked data based exploration systems. In: IESD (2014)
48. Dadzie, A.-S., Pietriga, E.: Visualisation of linked data-reprise. Semant. Web **8**(1), 1–21 (2017)
49. Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., Giannopoulou, E.: Ontology visualization methods'a survey. ACM Comput. Surv. (CSUR) **39**(4), 10 (2007)
50. Liu, S., Maljovec, D., Wang, B., Bremer, P.-T., Pascucci, V.: Visualizing high-dimensional data: advances in the past decade. TVCG **3**, 1249–1268 (2017)
51. Wohlfart, E., Aigner, W., Bertone, A., Miksch, S.: Comparing information visualization tools focusing on the temporal dimensions. In: IV, pp. 69–74 (2008)
52. Mei, H., Ma, Y., Wei, Y., Chen, W.: The design space of construction tools for information visualization: A survey. J. Vis. Lang. Comput. **44**, 120–132 (2018)
53. Diamond, M., Mattia, A.: Data visualization: an exploratory study into the software tools used by businesses. J. Instr. Pedag. **17**, 1–7 (2017)
54. Ghosh, A., Nashaat, M., Miller, J., Quader, S., Marston, C.: A comprehensive review of tools for exploratory analysis of tabular industrial datasets. Vis. Inform. **2**(4), 235–253 (2018)
55. Keim, D.A., Lee, J.P., Thuraisinghaman, B., Wittenbrink, C.: Database issues for data visualization: supporting interactive database exploration. In: Workshop on Database Issues for Data Visualization, pp. 12–25 (1995)
56. Idreos, S., Papaemmanouil, O., Chaudhuri, S.: Overview of data exploration techniques. In: SIGMOD, pp. 277–281 (2015)
57. Bikakis, N.: Big data visualization tools. arXiv:1801.08336 (2018)
58. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. TKDE **6**, 734–749 (2005)
59. Burke, R.: Hybrid recommender systems: survey and experiments. User Model. User-Adapted Interact. **12**(4), 331–370 (2002)
60. Sharma, L., Gera, A.: A survey of recommendation system: research challenges. IJETT **4**(5), 1989–1992 (2013)
61. Bobadilla, J., Ortega, F., Hernando, A., Gutiérrez, A.: Recommender systems survey. Knowl. Based Syst. **46**, 109–132 (2013)
62. Christi, J.R., Premkumar, K.: Survey on recommendation and visualization techniques for QoS-aware web services. In: ICICES, pp. 1–6 (2014)
63. Guy, I., Zwerdling, N., Carmel, D., Ronen, I., Uziel, E., Yogev, S., Ofek-Koifman, S.: Personalized recommendation of social software items based on social relations. In: RecSys, pp. 53–60 (2009)
64. Wei, K., Huang, J., Fu, S.: A survey of e-commerce recommender systems. In: ICSSSM, pp. 1–5 (2007)
65. Heer, J., Card, S.K., Landay, J.A.: prefuse: a toolkit for interactive information visualization. In: CHI, pp. 421–430 (2005)
66. Flare. http://flare.prefuse.org. Accessed 31 Dec 2018
67. Bostock, M., Heer, J.: Protovis: a graphical toolkit for visualization. TVCG **15**(6), 1121–8 (2009)
68. Satyanarayan, A., Russell, R., Hoffswell, J., Heer, J.: Reactive vega: a streaming dataflow architecture for declarative interactive visualization. TVCG **22**(1), 659–668 (2015)
69. Khan, M., Khan, S.S.: Data and information visualization methods, and interactive mechanisms: a survey. Int. J. Comput. Appl. **34**(1), 1–14 (2011)
70. Wilkinson, L.: The Grammar of Graphics. Springer, Berlin (2005)
71. Wickham, H.: A layered grammar of graphics. J. Comput. Graph. Stat. **19**(1), 3–28 (2010)
72. VanderPlas, J., Granger, B.E., Heer, J., Moritz, D., Wongsuphasawat, K., Satyanarayan, A., Lees, E., Timofeev, I., Welsh, B., Sievert, S.: Altair: interactive statistical visualizations for python. https://altair-viz.github.io. Accessed 31 Dec 2018
73. Echarts. http://echarts.baidu.com. Accessed 31 Dec 2018
74. Shneiderman, B.: Direct manipulation: a step beyond programming languages. IEEE Comput. **16**(8), 57–69 (1983)

75. Qlik: Data analytics for modern business intelligence. https://www.qlik.com/us. Accessed 31 Dec 2018
76. Gonzalez, H., Halevy, A.Y., Jensen, C.S., Langen, A., Madhavan, J., Shapley, R., Shen, W., Goldberg-Kidon, J.: Google fusion tables: web-centered data management and collaboration. In: SIGMOD, pp. 1061–1066 (2010)
77. Ren, D., Höllerer, T., Yuan, X.: iVisDesigner: expressive interactive design of information visualizations. TVCG **20**(12), 2092–2101 (2014)
78. Satyanarayan, A., Heer, J.: Lyra: An interactive visualization design environment. https://idl.cs.washington.edu/projects/lyra/. Accessed 31 Dec 2018
79. Yalçın, M.A., Elmqvist, N., Bederson, B.B.: Keshif: Rapid and expressive tabular data exploration for novices. TVCG **24**(8), 2339–2352 (2018)
80. Liu, Z., Thompson, J., Wilson, A., Dontcheva, M., Delorey, J., Grigg, S., Kerr, B., Stasko, J.: Data illustrator. http://www.zcliu.org/di/. Accessed 31 Dec 2018
81. Liu, Z., Thompson, J., Wilson, A., Dontcheva, M., Delorey, J., Grigg, S., Kerr, B., Stasko, J.T.: Data illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring. In: CHI, p. 123 (2018)
82. Warren, L.: The visual display of quantitative information. Yale J. Biol. Med. **44**(4), 400–400 (1986)
83. Wongsuphasawat, K., Qu, Z., Moritz, D., Chang, R., Ouk, F., Anand, A., Mackinlay, J.D., Howe, B., Heer, J.: Voyager 2: augmenting visual analysis with partial view specifications. In: CHI, pp. 2648–2659 (2017)
84. Key, A., Howe, B., Perry, D., Aragon, C.R.: Vizdeck: self-organizing dashboards for visual analytics. In: SIGMOD, pp. 681–684 (2012)
85. Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J.M., Heer, J.: Profiler: integrated statistical analysis and visualization for data quality assessment. In: AVI, pp. 547–554 (2012)
86. Elzen, S.V.D., van Wijk, J.J.: Small multiples, large singles: a new approach for visual data exploration. Comput. Graph. Forum **32**(3pt2), 191–200 (2013)
87. Wilkinson, L., Anand, A., Grossman, R.: Graph-theoretic scagnostics. In: IEEE Symposium on Information Visualization, 2005. IEEE, Minneapolis, MN, USA (2005)
88. Mackinlay, J.: Automating the design of graphical presentations of relational information. ACM Trans. Graph. **5**(2), 110–141 (1986)
89. Setlur, V., Battersby, S.E., Tory, M., Gossweiler, R., Chang, A.X.: Eviza: A natural language interface for visual analysis. In: UIST, pp. 365–377 (2016)
90. Hoque, E., Setlur, V., Tory, M., Dykeman, I.: Applying pragmatics principles for interaction with visual analytics. TVCG **24**(1), 309–318 (2017)
91. Wu, E., Battle, L., Madden, S.R.: The case for data visualization management systems: vision paper. PVLDB **7**(10), 903–906 (2014)
92. Wu, E., Psallidas, F., Miao, Z., Zhang, H., Rettig, L., Wu, Y., Sellam, T.: Combining design and performance in a data visualization management system. In: CIDR (2017)
93. Lins, L., Klosowski, J.T., Scheidegger, C.: Nanocubes for real-time exploration of spatiotemporal datasets. TVCG **19**(12), 2456–2465 (2013)
94. Pang, Z., Wu, S., Chen, G., Chen, K., Shou, L.: FlashView: an interactive visual explorer for raw data. PVLDB **10**(12), 1869–1872 (2017)
95. Zoumpatianos, K., Idreos, S., Palpanas, T.: Indexing for interactive exploration of big data series. In: SIGMOD, pp. 1555–1566 (2014)
96. Piringer, H., Tominski, C., Muigg, P., Berger, W.: A multi-threading architecture to support interactive visual exploration. TVCG **15**(6), 1113–1120 (2009)
97. Chan, S.-M., Xiao, L., Gerth, J., Hanrahan. P.: Maintaining interactivity while exploring massive time series. In: VAST, pp. 59–66 (2008)
98. Battle, L., Chang, R., Stonebraker, M.: Dynamic prefetching of data tiles for interactive visualization. In: SIGMOD, pp. 1363–1375 (2016)
99. Alabi, D., Wu, E.: PFunk-H: approximate query processing using perceptual models. In: HILDA@SIGMOD, pp. 10–16 (2016)
100. Bikakis, N., Papastefanatos, G., Skourla, M., Sellis, T.: A hierarchical aggregation framework for efficient multilevel visual exploration and analysis. Semant. Web **8**(1), 139–179 (2017)
101. Elmqvist, N., Fekete, J.D.: Hierarchical aggregation for information visualization: overview, techniques, and design guidelines. TVCG **16**(3), 439–454 (2010)
102. Pahins, C.A., Stephens, S.A., Scheidegger, C., Comba, J.L.: Hashedcubes: simple, low memory, real-time visual exploration of big data. TVCG **23**(1), 671–680 (2016)
103. Moritz, D., Howe, B., Heer, J.: Falcon: balancing interactive latency and resolution sensitivity for scalable linked visualizations. In: CHI, p. 694 (2019)
104. Tauheed, F., Heinis, T., Shrmann, F., Markram, H., Ailamaki, A.: SCOUT: prefetching for latent feature following queries. PVLDB **5**(11), 1531–1542 (2012)
105. Yesilmurat, S.: Retrospective adaptive prefetching for interactive web gis applications. Geoinformatica **16**(3), 435–466 (2012)
106. Dong, H.L., Kim, J.S., Kim, S.D., Kim, K.C., Yoosung, K., Park, J.: Adaptation of a neighbor selection markov chain for prefetching tiled web GIS data. In: ADVIS, pp. 213–222 (2002)
107. Fua, Y.H., Ward, M.O., Rundensteiner, E.A.: Structure-based brushes: a mechanism for navigating hierarchically organized data and information spaces. TVCG **6**(2), 150–159 (2000)
108. Tao, W., Liu, X., Demiralp, Ç., Chang, R., Stonebraker, M.: Kyrix: Interactive visual data exploration at scale. In: CIDR (2019)
109. Broy, M., Denert, E., Bayer, R., McCreight, E.: Organization and maintenance of large ordered indexes. In: Broy, M., Denert, E. (eds.) Software Pioneers. Springer, Berlin, Heidelberg (2002)
110. Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J.D., Howe, B., Heer, J.: Towards a general-purpose query language for visualization recommendation. In: HILDA@SIGMOD, pp. 4–9 (2016)
111. Roth, S.F., Kolojejchick, J., Mattis, J., Goldstein, J.: Interactive graphic design using automatic presentation knowledge. In: CHI, p. 207 (1994)
112. Casner, S.M.: Task-analytic approach to the automated design of graphic presentations. ACM Trans. Graph. **10**(2), 111–151 (1991)
113. Bertin, J.: Semiology of graphics - diagrams, networks, maps. ESRI. ISBN: 978-1-58948-261-6. http://esripress.esri.com/display/index.cfm?fuseaction=display&websiteID=190&moduleID=0 (2010)
114. Cleveland, W.S., McGill, R.: Graphical perception: theory, experimentation, and application to the development of graphical methods. ASA **79**(387), 531–554 (1984)
115. Shepard, R.N.: Toward a universal law of generalization for psychological science. Science **242**(4880), 1317–1323 (1988)
116. Lewandowsky, Stephan, Spence, Ian: Discriminating strata in scatterplots. ASA **84**(407), 682–688 (1989)
117. Hu, K.Z., Orghian, D., Hidalgo, C.A.: DIVE: a mixed-initiative system supporting integrated data exploration workflows. In: HILDA@SIGMOD, pp. 5:1–5:7 (2018)
118. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. ASA **74**(368), 829–836 (1979)
119. Silverman, B.W.: Density estimation for statistics and data analysis. Springer, pp. 1–158 (1986). https://doi.org/10.1007/978-1-4899-3324-9

120. Dibia, V., Demiralp, Ç.: Data2Vis: Automatic generation of data visualizations using sequence to sequence recurrent neural networks. CoRR, abs/1804.03126 (2018)
121. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.: Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 89–96 (2005)
122. Herbrich, R., Graepel, T., Obermayer, K.: Support vector learning for ordinal regression. In: ICANN, vol. 1, pp. 97–102 (2002)
123. Kim, Y., Heer, J.: Assessing effects of task and data distribution on the effectiveness of visual encodings. Comput. Graph. Forum **37**(3), 157–167 (2018)
124. Saket, B., Endert, A., Demiralp, C.: Task-based effectiveness of basic visualizations. TVCG **PP**(99), 1–1 (2017)
125. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**(1), 81–106 (1986)
126. Wu, Q., Burges, C.J., Svore, K.M., Gao, J.: Ranking, boosting, and model adaptation. Technical report, Microsoft Research (2008)
127. Epelbaum, T.: Deep learning: technical introduction. CoRR, arXiv:1709.01412 (2017)
128. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. NIPS **4**, 3104–3112 (2014)
129. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. Comput. Sci. arXiv preprint arXiv:1409.0473 (2014)
130. Sundermeyer, M., Schlüter, R., Ney, H.: LSTM neural networks for language modeling. In: Interspeech, pp. 601–608 (2012)
131. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
132. Poco, J., Heer, J.: Reverse-engineering visualizations: recovering visual encodings from chart images. Comput Graph Forum **36**(3), 353–363
133. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**(1), 159–165 (1990)
134. Gotz, D., Wen, Z.: Behavior-driven visualization recommendation. In: IUI, pp. 315–324 (2009)
135. Schafer, J.B., Konstan, J., Riedl, J.: Recommender systems in e-commerce. In: World Automation Congress, pp. 158–166 (1999)
136. Liu, R.R., Jia, C.X., Zhou, T., Sun, D., Wang, B.H.: Personal recommendation via modified collaborative filtering. Physica A Stat. Mech. Appl. **388**(4), 462–468 (2012)
137. Soboroff, I., Nicholas, C.: Combining content and collaboration in text filtering. In: IJCAI, pp. 86–91 (1999)
138. Mutlu, B., Veas, E., Trattner, C.: VizRec: recommending personalized visualizations. TIIS **6**(4), 31 (2016)
139. Wu, E., Madden, S.R.: Scorpion: explaining away outliers in aggregate queries. In: PVLDB, pp. 553–564 (2013)
140. Song, H., Szafir, D.A.: Where's my data? Evaluating visualizations with missing data. IEEE Trans. Vis. Comput. Graph. **25**(1), 914–924 (2019)
141. Battle, L., Angelini, M., Binnig, C., Catarci, T., Eichmann, P., Fekete, J., Santucci, G., Sedlmair, M., Willett, W.: Evaluating visual data analysis systems: a discussion report. In: HILDA@SIGMOD, pp. 4:1–4:6 (2018)
142. Battle, L., Chang, R., Heer, J., Stonebraker, M.: Position statement: the case for a visualization performance benchmark. In: DSIA, pp. 1–5 (2017)
143. Jiang, L., Rahman, P., Nandi, A.: Evaluating interactive data systems: workloads, metrics, and guidelines. In: SIGMOD, pp. 1637–1644 (2018)
144. Hu, K.Z., Gaikwad, S.N.S., Hulsebos, M., Bakker, M.A., Zgraggen, E., Hidalgo, C.A., Kraska, T., Li, G., Satyanarayan, A., Demiralp, Ç.: Viznet: Towards A large-scale visualization learning and benchmarking repository. In: CHI, pp. 662 (2019)
145. Valizadegan, H., Jin, R., Zhang, R., Mao, J.: Learning to rank by optimizing NDCG measure. In: NIPS, pp. 1883–1891 (2009)
146. Rezig, E.K., Cao, L., Stonebraker, M., Simonini, G., Tao, W., Madden, S., Ouzzani, M., Tang, N., Elmagarmid, A.K.: Data civilizer 2.0: a holistic framework for data preparation and analytics. PVLDB **12**(12), 1954–1957 (2019)
147. Rezig, E.K., Cao, L., Simonini, G., Schoemans, M., Madden, S., Ouzzani, M., Tang, N., Stonebraker, M.: Dagger: a data (not code) debugger. In: CIDR (2020)