Big Data Visualization Tools *

Nikos Bikakis

1 Synonyms

Visual exploration; Interactive visualization; Information visualization; Visual analytics; Exploratory data analysis.

2 Definition

Data visualization is the presentation of data in a pictorial or graphical format, and a *Data visualization tool* is the software that generates this presentation. Data visualization offers intuitive ways for information perception and manipulation that essentially amplify the overall cognitive performance of information processing, enabling users to effectively identify interesting patterns, infer correlations and causalities, and supports sense-making activities.

3 Overview

Data visualization provides users with intuitive means to interactively explore and analyze data, enabling them to identify interesting patterns, discover

Nikos Bikakis
ATHENA Research Center, Greece

* This article appears in Encyclopedia of Big Data Technologies, 2nd Edition, Springer, 2021

correlations and causalities, and support sense-making activities.[2] This is of great importance, especially given the massive volumes of digital information concerning nearly every aspect of human activity that are currently being produced and collected.

Data visualization and analytics are nowadays one of the corner-stones of Data Science, turning the abundance of Big Data being produced through modern systems into actionable knowledge. Indeed, the Big Data era has realized the availability of voluminous datasets that are dynamic, noisy and heterogeneous in nature. Transforming a data-curious user into someone who can access and analyze that data is even more burdensome now for a great number of users with little or no support and expertise on the data processing part. Thus, the area of data visualization and analysis has gained great attention recently, calling for joint action from different research areas and communities such as information visualization, data management and mining, human-computer interaction, and computer graphics.

Several traditional problems from those communities, such as efficient data storage, querying and indexing for enabling visual analytics, ways for visual presentation of massive data, efficient interaction and personalization techniques that can fit to different user needs, are revisited with Big Data in mind [1, 2, 3, 4, 5, 6].

Given the above, modern visualization systems should effectively and efficiently handle the following aspects:

- *Real-time Interaction.* Efficient and scalable techniques should support the interaction with billion-objects datasets, while maintaining an acceptable system response in in less than a second.
- *On-the-fly Visualization.* Support of on-the-fly visualizations over large and dynamic sets of volatile raw (i.e., not preprocessed) data is required. In several cases, a preprocessing phased is not an option.
- *Visual Scalability.* Provision of effective data abstraction mechanisms is necessary for addressing problems related to visual information overloading (a.k.a. overplotting).
- *User Assistance and Personalization.* Encouraging user comprehension and offering customization capabilities to different user-defined exploration scenarios and preferences according to the analysis needs are important features.

The literature on visualization is extensive, covering a large range of fields and many decades [7, 8]. Data visualization is discussed in a great number of recent introductory-level textbooks, such as [9, 10] Further, surveys of Big Data visualization systems can be found at [2, 3, 4, 5, 11, 12].

[2] Throughout the article, terms *visualization* and *visual exploration*, as well as terms *tool* and *system* are used interchangeably.

Finally, there is a great deal of information regarding visualization tools available in the Web. We mention dataviz.tools[3] and datavizcatalogue[4] which are catalogs containing a large number of visualization tools, libraries and resources.

4 Visualization in Big Data Era

This section discusses the basic concepts related to Big Data visualization. First, the limitations of traditional visualization systems are outlined. Then, the basic characteristics of data visualization in the context of Big Data era are presented. Finally, the major prerequisites and challenges that should be addressed by modern visualization systems are discussed.

Traditional Visualization Systems. Most *traditional visualization systems* perform well for ad-hoc visualizations of small data files (e.g., showing a trend-line or a bar chart) or over aggregated data (e.g., summaries of data points, into which user can zoom in), which can fit in main memory. Hence, they restrict themselves to dealing with *small datasets*, which can be easily handled and analyzed with conventional data management and visual explorations techniques. For larger data files, the conventional systems usually require a *preprocessing phase*, such as loading in a data management system. As a result, they are limited to accessing *preprocessed sets of static data*.

Big Data era. On the other hand, nowadays, the Big Data era has made available large numbers of *very big* datasets, that are often *dynamic* and characterized by high *variety* and *volatility*. For example, in several cases (e.g., scientific databases), new data constantly arrive on an hourly basis; in other cases, data sources offer query or API endpoints for online access and updating. Further, nowadays, an increasingly large number of *diverse users* (i.e., users with different preferences or skills) explore and analyze data in a plethora of *different scenarios and tasks*.

Visualization Systems in Big Data era *Modern systems* should be able to efficiently handle *big dynamic datasets*, operating on machines with limited computational and memory resources (e.g., laptops). The dynamic nature of nowadays data (e.g., stream data), hinders the application of a preprocessing phase, such as traditional database loading and indexing. Hence, systems should provide *on-the-fly processing and visualization* over large sets of data.

Further, in conjunction with performance issues, modern systems have to address challenges related to *visual presentation*. Visualizing a large number of data objects is a challenging task; modern systems have to "*squeeze a*

[3] http://dataviz.tools

[4] www.datavizcatalogue.com

billion records into a million pixels" [6]. Even in small datasets, offering a dataset overview may be extremely difficult; in both cases, *information overloading* (a.k.a. overplotting) is a common issue. Consequently, visual scalability a basic requirement of modern systems, which have to effectively support data reduction/abstraction (e.g., sampling, aggregation) over enormous numbers of data objects.

Apart from the aforementioned requirements, modern systems must also satisfy the *diversity of preferences and requirements* posed by *different users and tasks*. Modern systems should provide the user with the ability to customize the exploration experience based on her preferences and the individual requirements of each examined task. Additionally, systems should automatically adjust their parameters by taking into account the *environment setting and available resources*; e.g., screen resolution/size, available memory.

5 Visualization Tools and Techniques

This section presents how state-of-the-art approaches from Data Management and Mining, Information Visualization and Human-Computer Interaction communities attempt to handle the challenges that arise in the Big Data era.

Data Reduction. In order to handle and visualize large datasets, modern systems have to deal with information overloading issues. Offering *visual scalability* is crucial in Big Data visualization. Systems should provide efficient and effective abstraction and summarisation mechanisms. In this direction, a large number of systems use *approximation techniques* (a.k.a. data reduction techniques), in which abstract sets of data are computed. Considering the existing approaches, most of them are based on: (1) *sampling* and *filtering* [13, 14, 15, 16] and/or (2) *aggregation* (e.g., binning, clustering) [17, 18, 19, 20].

Hierarchical Data Exploration. Data reduction techniques are often defined hierarchically [17], allowing users to explore data in different levels-of-detail by, e.g., hierarchical aggregation.

Hierarchical approaches (a.k.a.multilevel) allow the visual exploration of very large datasets in multiple levels (with different level-of-details), offering both an overview, as well as an intuitive and effective way for finding specific parts within a dataset.

Particularly, in hierarchical approaches, the user first obtains an overview of the dataset before proceeding to data exploration operations (e.g., roll-up, drill-down, zoom, filtering) and finally retrieving details about the data. A significant challenge, in large data visualization, is the problem of overplotting. It can effectively be addressed in hierarchical approaches, in which, in

each level, the number of the presented visual elements is controlled by data reduction methods.

Hierarchical techniques have been extensively used in large graphs/network visualization, in order to handle the common problem of overloading, in winch the graph is presented as "hairball". In these techniques the graph is recursively decomposed into smaller sub-graphs that form a hierarchy of abstraction layers. In most cases, the hierarchy is constructed by exploiting clustering and partitioning [21, 22], sampling [23], and edge bundling [24] techniques.

Progressive Data Visualization. Visual data exploration requires real-time system's response. However, computing complete results over large (unprocessed) datasets may be extremely costly, and in several cases unnecessary. Modern systems should progressively return partial and preferably representative results, as soon as possible [25, 26].

Progressiveness can significantly improve efficiency in exploration scenarios, where it is common that users attempt to find something interesting without knowing what exactly they are searching for beforehand. In this case, users perform a sequence of operations (e.g., queries), where the result of each operation determines the formulation of the next operation.

Recently, many systems adopt the *progressive paradigm* attempting to reduce the response time [13, 15, 25, 26, 27, 28]. Progressive approaches, instead of performing all the computations in one step (that can take a long time to complete), splits them in a series of short chunks of approximate computations that improved with time. Therefore, instead of waiting for an unbounded amount of time, users can see the results unfolding progressively. This way the users are able to interrupt the execution and define the next operation, without waiting the exact result to be computed.

In-situ, Incremental and Adaptive Processing. In-situ processing [29, 30, 31] is a recent trend, that aims at enabling the on-the-fly querying over large sets of raw data, by avoiding the (pre)processing (e.g., loading and indexing) overhead of traditional DBMS techniques. Usually, data loading and indexing take a large part of the overall time-to-analysis for both traditional RDBMs and Big Data systems [29]. In-situ query processing aims at avoiding data loading in a DBMS by accessing and operating directly over raw data files. In these systems, *incremental and adaptive* processing and indexing techniques are used, in which small parts of raw data are processed incrementally "following" users' interactions.

Data Structures. In the context of visual exploration, several data structures have been introduced. VisTrees [32] and HETree [18] are tree-based main-memory indexes that address visual exploration use cases; i.e., they offer exploration-oriented features such as incremental index construction and adaptation.

Hashedcubes [33], Nanocubes [34], and SmartCube [35], and are main-memory data structures supporting interactive visualization, They are based

on main-memory variations of a data cube in order to reduce the time needed to generate the visualization.

graphVizdb [36] is a graph-based visualization tool, which employs a 2D spatial index (e.g., R-tree) and maps user interactions into window 2D queries. Finally, tile-based structures are used in several visualization systems, e.g., RawVis [31], ForeCache [37].

Caching and Prefetching. Recall that, in exploration scenarios, a sequence of operations is performed and, in most cases, each operation is driven by the previous one. In this setting, *caching* and/or *prefetching* the sets of data that are likely to be accessed by the user in the near future can significantly reduce the response time [37, 38]. Most of these approaches use prediction techniques which exploit several factors (e.g., user behavior, user profile, use case) in order to determine the upcoming user interactions.

User Assistance. The huge amount of available information makes it difficult for users to manually explore and analyze data. Modern systems should provide mechanisms that assist the user and reduce the effort needed on their part, considering the diversity of preferences and requirements posed by different users and tasks.

Recently, several approaches have been developed in the context of *visualization recommendation* [39]. These approaches recommend the most suitable visualizations in order to assist users throughout the analysis process. Usually, the recommendations take into account several factors, such as data characteristics, environment setting and available resources (e.g., screen resolution/size, available memory), examined task, user preferences and behavior, etc.

Considering data characteristics, there are several systems that recommend the most suitable visualization technique (and parameters) based on the type, attributes, distribution, or cardinality of the input data [40, 41]. Other approaches provide visualization recommendations based on user behavior and preferences [42], using machine learning [43] or similarity-based techniques [44]. In a similar context, some systems assist users by recommending certain visualizations that reveal surprising, interesting data or outliers [45, 46].

6 Examples of Applications

Visualization techniques are of great importance in a wide range of application areas in the Big Data era. The volume, velocity, heterogeneity and complexity of available data make it extremely difficult for humans to explore and analyze data. Data visualization enables users to perform a series of analysis tasks that are not always possible with common data analysis techniques [10].

Major application domains for data visualization and analytics are *Physics* and *Astronomy*. Satellites and telescopes collect daily massive and dynamic streams of data. Using traditional analysis techniques, astronomers are able to identify noise, patterns and similarities. On the other hand, visual analytics can enable astronomers to identify unexpected phenomena and perform several complex operations, which are not are feasible by traditional analysis approaches.

Another application domain is *atmospheric sciences* like *Meteorology* and *Climatology*. In this domain high volumes of data are collected from sensors and satellites on a daily basis. Storing these data over the years results in massive amounts of data that have to be analyzed. Visual analytics can assist scientists to perform core tasks, such as climate factors correlation analysis, event prediction, etc. Further, in this domain, visualization systems are used in several scenarios in order capture real-time phenomena, such as, hurricanes, fires, floods, and tsunamis.

In the domain of *Bioinformatics*, visualization techniques are exploited in numerous tasks. For example, analyzing the large amounts of biological data produced by DNA sequencers is extremely challenging. Visual techniques can help biologist to gain insight and identify interesting "areas" of genes on which to performs their experiments.

In the Big Data era, visualization techniques are extensively used in the *business intelligence* domain. *Finance markets* is one application area, where visual analytics allow to monitor markets, identify trends and perform predictions. Besides, *market research* is also an application area. Marketing agencies and in-house marketing departments analyze a plethora of diverse sources (e.g., finance data, customer behavior, social media). Visual techniques are exploited to realize task such as, identifying trends, finding emerging market opportunities, finding influential users and communities, optimizing operations (e.g., troubleshooting of products and services), business analysis and development (e.g., churn rate prediction, marketing optimization).

7 Cross-References

- Visualization
- Visualization Techniques
- Visualizing Semantic Data
- Graph Exploration and Search

References

1. G. L. Andrienko, N. V. Andrienko, S. M. Drucker, J. Fekete, D. Fisher, S. Idreos, T. Kraska, G. Li, K. Ma, J. D. Mackinlay, A. Oulasvirta, T. Schreck, H. Schumann, M. Stonebraker, D. Auber, N. Bikakis, P. K. Chrysanthis, G. Papastefanatos, and M. A. Sharaf, "Big data visualization and analytics: Future research challenges and emerging applications," in *Proceedings of the International Workshop on Big Data Visual Exploration and Analytics (BigVis)*, 2020.
2. X. Qin, Y. Luo, N. Tang, and G. Li, "Making data visualization more efficient and effective: a survey," *Journal on Very Large Data Bases (VLDBJ)*, vol. 29, no. 1, 2020.
3. S. Idreos, O. Papaemmanouil, and S. Chaudhuri, "Overview of Data Exploration Techniques," in *ACM Conference on Management of Data (SIGMOD)*, 2015.
4. M. Behrisch, D. Streeb, F. Stoffel, D. Seebacher, B. Matejek, S. H. Weber, S. Mittelstaedt, H. Pfister, and D. Keim, "Commercial Visual Analytics Systems-Advances in the Big Data Analytics Field," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 25, no. 10, 2019.
5. P. Godfrey, J. Gryz, and P. Lasek, "Interactive Visualization of Large Data Sets," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 28, no. 8, 2016.
6. B. Shneiderman, "Extreme Visualization: Squeezing a Billion Records into a Million Pixels," in *ACM Conference on Management of Data (SIGMOD)*, 2008.
7. D. Rees and R. S. Laramee, "A survey of information visualization books," *Computer Graphics Forum*, vol. 38, no. 1, 2019.
8. L. McNabb and R. S. Laramee, "Survey of surveys (sos) - mapping the landscape of survey papers in information visualization," *Computer Graphics Forum*, vol. 36, no. 3, 2017.
9. M. O. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications, Second Edition*. A. K. Peters, Ltd., 2015.
10. D. A. Keim, J. Kohlhammer, G. P. Ellis, and F. Mansmann, *Mastering the Information Age - Solving Problems with Visual Analytics*. Eurographics Association, 2010.
11. L. Po, N. Bikakis, F. Desimoni, and G. Papastefanatos, *Linked Data Visualization: Techniques, Tools, and Big Data*. Synthesis Lectures on the Data, Semantics, and Knowledge, Morgan and Claypool, 2020.
12. N. Bikakis and T. Sellis, "Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art," in *6th Intl. Workshop on Linked Web Data Management (LWDM)*, 2016.
13. D. Fisher, I. O. Popov, S. M. Drucker, and M. C. Schraefel, "Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster," in *Conference on Human Factors in Computing Systems (CHI)*, 2012.
14. Y. Park, M. J. Cafarella, and B. Mozafari, "Visualization-aware Sampling for Very Large Databases," in *IEEE Intl. Conf. on Data Engineering (ICDE)*, 2016.
15. S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica, "Blinkdb: Queries with Bounded Errors and Bounded Response Times on Very Large Data," in *European Conference on Computer Systems (EuroSys)*, 2013.
16. L. Battle, M. Stonebraker, and R. Chang, "Dynamic Reduction of Query Result Sets for Interactive Visualizaton," in *IEEE Conf. on Big Data (BigData)*, 2013.
17. N. Elmqvist and J. Fekete, "Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 16, no. 3, 2010.
18. N. Bikakis, G. Papastefanatos, M. Skourla, and T. Sellis, "A Hierarchical Aggregation Framework for Efficient Multilevel Visual Exploration and Analysis," *Semantic Web Journal*, vol. 8, no. 1, 2017.
19. U. Jugel, Z. Jerzak, G. Hackenbroich, and V. Markl, "VDDa: Automatic Visualization-driven Data Aggregation in Relational Databases," *Journal on Very Large Data Bases (VLDBJ)*, 2015.

20. Z. Liu, B. Jiang, and J. Heer, "*imMens*: Real-time Visual Querying of Big Data," *Comput. Graph. Forum (CGF)*, vol. 32, no. 3, pp. 421–430, 2013.
21. J. F. R. Jr., H. Tong, J. Pan, A. J. M. Traina, C. T. Jr., and C. Faloutsos, "Large Graph Analysis in the GMine System," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 25, no. 1, 2013.
22. C. Tominski, J. Abello, and H. Schumann, "Cgv - an Interactive Graph Visualization System," *Computers & Graphics*, vol. 33, no. 6, 2009.
23. S. Sundara, M. Atre, V. Kolovski, S. Das, Z. Wu, E. I. Chong, and J. Srinivasan, "Visualizing Large-scale RDF Data Using Subsets, Summaries, and Sampling in Oracle," in *IEEE Intl. Conf. on Data Engineering (ICDE)*, pp. 1048–1059, 2010.
24. E. R. Gansner, Y. Hu, S. C. North, and C. E. Scheidegger, "Multilevel Agglomerative Edge Bundling for Visualizing Large Graphs," in *IEEE Pacific Visualization Symposium (PacificVis)*, 2011.
25. M. Angelini, G. Santucci, H. Schumann, and H. Schulz, "A review and characterization of progressive visual analytics," *Informatics*, vol. 5, no. 3, 2018.
26. E. Zgraggen, A. Galakatos, A. Crotty, J. Fekete, and T. Kraska, "How Progressive Visualizations Affect Exploratory Analysis," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 23, no. 8, 2017.
27. D. Moritz, D. Fisher, B. Ding, and C. Wang, "Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data," in *Conference on Human Factors in Computing Systems (CHI)*, 2017.
28. S. Rahman, M. Aliakbarpour, H. Kong, E. Blais, K. Karahalios, A. G. Parameswaran, and R. Rubinfeld, "I've Seen "enough": Incrementally Improving Visualizations to Support Rapid Decision Making," *VLDB Endowment (PVLDB)*, vol. 10, no. 11, 2017.
29. S. Idreos, I. Alagiannis, R. Johnson, and A. Ailamaki, "Here Are My Data Files. Here Are My Queries. Where Are My Results?," in *Conf. on Innovative Data Systems Research (CIDR)*, 2011.
30. I. Alagiannis, R. Borovica, M. Branco, S. Idreos, and A. Ailamaki, "Nodb: Efficient Query Execution on Raw Data Files," in *ACM Conference on Management of Data (SIGMOD)*, 2012.
31. N. Bikakis, S. Maroulis, G. Papastefanatos, and P. Vassiliadis, "In-situ visual exploration over big raw data," *Information Systems, Elsevier*, vol. 95, 2021.
32. M. El-Hindi, Z. Zhao, C. Binnig, and T. Kraska, "Vistrees: Fast Indexes for Interactive Data Exploration," in *HILDA*, 2016.
33. C. A. de Lara Pahins, S. A. Stephens, C. Scheidegger, and J. L. D. Comba, "Hashedcubes: Simple, Low Memory, Real-time Visual Exploration of Big Data," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 23, no. 1, 2017.
34. L. D. Lins, J. T. Klosowski, and C. E. Scheidegger, "Nanocubes for Real-Time Exploration of Spatiotemporal Datasets," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 19, pp. 2456–2465, 2013.
35. C. Liu, C. Wu, H. Shao, and X. Yuan, "Smartcube: An adaptive data management architecture for the real-time visualization of spatiotemporal datasets," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 26, no. 1, 2020.
36. N. Bikakis, J. Liagouris, M. Krommyda, G. Papastefanatos, and T. Sellis, "graphVizdb: A Scalable Platform for Interactive Large Graph Visualization," in *IEEE Intl. Conf. on Data Engineering (ICDE)*, 2016.
37. L. Battle, R. Chang, and M. Stonebraker, "Dynamic Prefetching of Data Tiles for Interactive Visualization," in *ACM Conference on Management of Data (SIGMOD)*, 2016.
38. F. Tauheed, T. Heinis, F. Schürmann, H. Markram, and A. Ailamaki, "SCOUT: Prefetching for Latent Feature Following Queries," *VLDB Endowment (PVLDB)*, vol. 5, no. 11, 2012.
39. M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. G. Parameswaran, "Towards Visualization Recommendation Systems," *SIGMOD Record*, vol. 45, no. 4, 2016.

40. A. Key, B. Howe, D. Perry, and C. R. Aragon, "Vizdeck: Self-organizing Dashboards for Visual Analytics," in *ACM Conference on Management of Data (SIGMOD)*, 2012.
41. H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis, "Muve: Efficient Multi-objective View Recommendation for Visual Data Exploration," in *IEEE Intl. Conf. on Data Engineering (ICDE)*, 2016.
42. B. Mutlu, E. E. Veas, and C. Trattner, "Vizrec: Recommending Personalized Visualizations," *ACM Transactions on Interactive Intelligent Systems (TIIS)*, vol. 6, no. 4, 2016.
43. K. Z. Hu, M. A. Bakker, S. Li, T. Kraska, and C. A. Hidalgo, "VizML: A Machine Learning Approach to Visualization Recommendation," in *Conference on Human Factors in Computing Systems (CHI)*, p. 128, 2019.
44. Y. Kim, K. Wongsuphasawat, J. Hullman, and J. Heer, "Graphscape: A Model for Automated Reasoning about Visualization Similarity and Sequencing," in *Conference on Human Factors in Computing Systems (CHI)*, 2017.
45. M. Vartak, S. Madden, A. G. Parameswaran, and N. Polyzotis, "SEEDB: Automatically Generating Query Visualizations," *VLDB Endowment (PVLDB)*, vol. 7, no. 13, 2014.
46. K. Wongsuphasawat, D. Moritz, A. Anand, J. D. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory Analysis Via Faceted Browsing of Visualization Recommendations," *IEEE Trans. Vis. Comput. Graph. (TVCG)*, vol. 22, no. 1, 2016.