

How are they different? A quantitative domain comparison of information visualization and data visualization (2000–2014)

Meen Chul Kim¹ · Yongjun Zhu¹ · Chaomei Chen¹

Received: 20 July 2015 / Published online: 27 January 2016
© Akadémiai Kiadó, Budapest, Hungary 2016

Abstract Information visualization and data visualization are often viewed as similar, but distinct domains, and they have drawn an increasingly broad range of interest from diverse sectors of academia and industry. This study systematically analyzes and compares the intellectual landscapes of the two domains between 2000 and 2014. The present study is based on bibliographic records retrieved from the Web of Science. Using a topic search and a citation expansion, we collected two sets of data in each domain. Then, we identified emerging trends and recent developments in information visualization and data visualization, captivated in intellectual landscapes, landmark articles, bursting keywords, and citation trends of the domains. We found out that both domains have computer engineering and applications as their shared grounds. Our study reveals that information visualization and data visualization have scrutinized algorithmic concepts underlying the domains in their early years. Successive literature citing the datasets focuses on applying information and data visualization techniques to biomedical research. Recent thematic trends in the fields reflect that they are also diverging from each other. In data visualization, emerging topics and new developments cover dimensionality reduction and applications of visual techniques to genomics. Information visualization research is scrutinizing cognitive and theoretical aspects. In conclusion, information visualization and data visualization have co-evolved. At the same time, both fields are distinctively developing with their own scientific interests.

Keywords Visual analytics · Domain analysis · Information visualization · Data visualization · Data science · Scientometrics

✉ Yongjun Zhu
zhu@drexel.edu

Meen Chul Kim
meenchul.kim@drexel.edu

Chaomei Chen
chaomei.chen@drexel.edu

¹ Drexel University College of Computing and Informatics, Philadelphia, PA, USA

Introduction

The term information visualization refers to computer generated interactive graphical representations of information (Chen 2010). It aims to convey unclear and abstract ideas to its audience in a visually intuitive and effective way, so that the users can be stimulated for new insights (Thomas and Cook 2005). The field of information visualization has emerged from multidisciplinary research fields such as human–computer interaction, computer science, graphics, visual design, and psychology. A diverse range of domains such as scientific research, data mining, and drug discovery is employing information visualization as a critical component (Bederson and Shneiderman 2003). Thomas and Cook (2005) argue that visual representations and interaction techniques allow users to intuitively see, explore, and understand large amounts of data. Within this context, we can view information visualization as the study of interactive visual representations of abstract data to reinforce human cognition. Data visualization, often regarded as a sub-domain of the former, is the science of visual representation of “data” (Friendly 2008). Data visualization also pursues to make people understand the significance of data by placing it in a visual context such as statistical graphics and graphical display so that the users can make comparison or determine causality (Tufte 1983). Tukey (1977) argues that a visual representation is valuable when it forces us to recognize what we never expect to see. Patterns, trends, and relationships that might go unrevealed in raw data can be recognized easier within a set of visualization techniques and software. Information visualization and data visualization may be viewed as different fields. Figure 1 displays Google Trends when given information visualization (blue dotted line) and data visualization (red line) as search queries. The height of each line represents a frequency of each query appearing in web documents such as news headlines during a specific year. It is obvious that data visualization has been receiving more attention from people than information visualization. In addition, it may indicate that these two domains are distinctively regarded by people.

By the definitions, it is evident that these two domains share some characteristics and discriminate each other from a few perspectives. First of all, both of the domains are concerned with the visual representation in order to communicate values which underlie data (Chen et al. 2008). Intuitive and effective visualizations might help users in analyzing and reasoning about data and evidence. It makes complex data more apprehensible and usable. This shared goal of the domains can be categorized into functional visualization and aesthetic representation. Gaviria (2008) argues that the former aims to convey a message or clarify patterns hidden under the surface of the data by symbols and metaphors

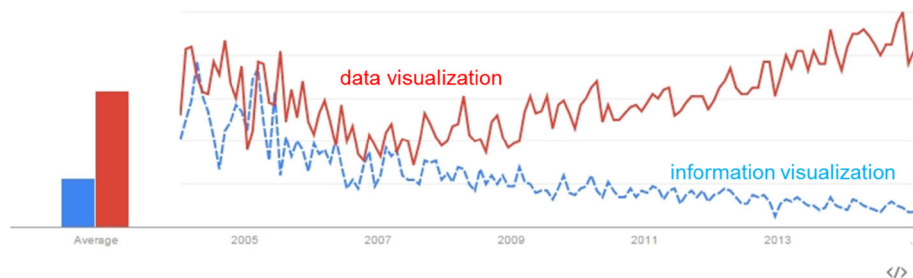
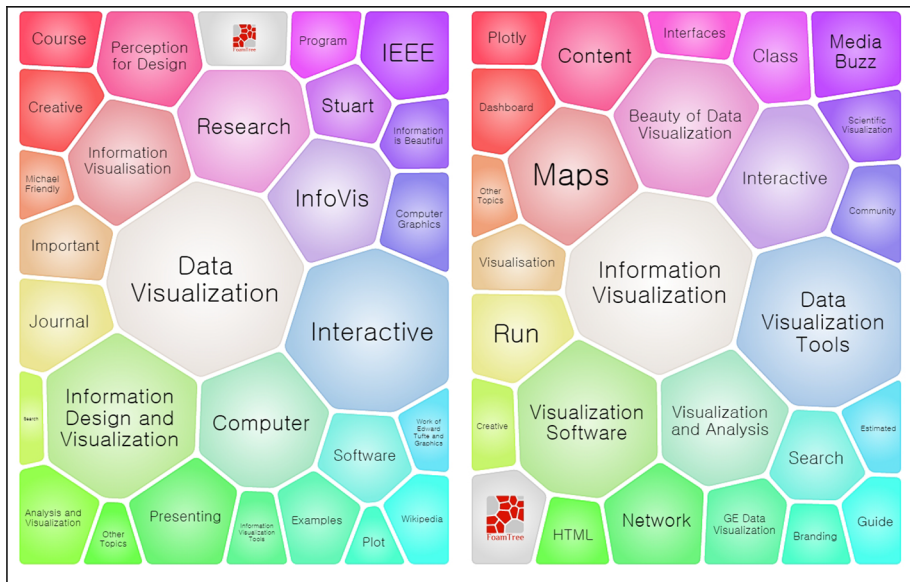


Fig. 1 (Color online) Google trends on the topics of information visualization (blue dotted line) and data visualization (red line). The height of each line indicates a frequency of each query appearing in web documents such as news headlines in a specific year



that users can understand better, whereas the latter is more concerned with generating a subjective impression of a dataset by eliciting a cognitive or emotive response from the user. Therefore, information visualization and data visualization can be regarded as both an art and a science. What makes these fields distinct from each other depends on what they handle to convey messages to their audience. That is, information visualization primarily deals with abstract or non-spatial data whereas its counterpart focuses on processing and displaying numeric or statistical input. Chen (2010) distinguishes information visualization from data or scientific visualization, based on the following concerns: (1) is the raw data numerical or spatial? and (2) if the data is not quantitative in nature, what does transform it to something that is numerical and visual? He argues that creatively transforming such non-spatial data to intuitive and meaningful graphical representations is of fundamental importance of information visualization (Chen 2010). Therefore, the difference between the domains is that in information visualization the spatial representation is generated whilst it is given in data visualization.

Major topics in both domains are shown in Fig. 2. The two visual representations, known as FoamTree, were generated by Carrot search¹ based on the top 100 search results from the Web, given information visualization and data visualization as queries. As depicted in the figure, two domains are cross-referencing to each other and also having their own interests. For instance, “information design”, “research”, and functional aspects of information visualization represented as “interactive” are among the leading topics in information visualization, whereas data visualization focuses more on “data visualization tools”, “visualization software”, and aesthetics of data visualization described as “beauty of data visualization”. The commonly occurring clusters are “visualization and analysis”,

¹ <http://search.carrotsearch.com/carrot2-webapp/search>.

“beauty of visualization”, and “interactive”. It indicates that both information visualization and data visualization aim to support users to intuitively navigate underlying data within a visual context. In doing that, the beauty of visual representations is of importance in both domains. In addition, the concept of interactivity is regarded as fundamental in information visualization and data visualization.

The motivation of the present study lies in our intention to identify the similarity and difference between the domains in a quantitative manner. Toward that end, we explore and compare the intellectual landscapes of information and data visualization research. Thematic patterns, landmark articles, and emerging trends and new developments of the domains are identified. In particular, our comparative study is guided by a computational approach implemented in CiteSpace, a visual analytic system for illustrating emerging trends and critical changes in scientific literature (Chen 2006; Chen et al. 2012, 2014a, b; Kim and Chen 2015). We also employ a dual-map overlay technique (Chen and Leydesdorff 2014) to expand the present study to the investigation of citation patterns at a disciplinary level. In the paper that follows, we aim to quantitatively address following research questions:

- What are the shared contexts between information visualization and data visualization?
- How are emerging trends and recent developments of the domains different from each other?

The rest of the study is organized as follows. We survey preceding literature and introduce the methodology of our study. Then, the intellectual structures of the fields are described and compared. Finally, findings and implications from our investigation are discussed.

Related work

Previous literature has investigated research trends in information visualization and data visualization. Herman et al. (2000) study graph visualization and navigation techniques as used in information visualization. They survey literature from various venues such as Graph Drawing Symposia and Journal of Graph Algorithms and Applications and introduce recent techniques of the domain such as spanning trees, 3D layout, and clustering. Mehlan et al. (2013) also examine various data visualization techniques used in environmental proteomics including parallel plots, microcharts, heat maps, stream graphs, tree graphs, tree maps, and so on. They manually select related works based on their prior domain knowledge and provide detailed explanations on each data visualization technique. Another school of researchers scrutinizes a variety of applications of information and data visualization techniques to other sciences. Kersten-Oertel et al. (2013) explore the state of the art techniques of data visualization employed in mixed reality image guided surgery (IGS). They retrieve 87 publications with the terms “reality” or “virtuality” in the titles along with terms related to surgery. Then, the authors introduce visual analytic systems for clinical use in the operating room. Faisal et al. (2013) present a systematic review of the literature on information visualization techniques within a context of personal healthcare. They present five major application themes of information visualization in healthcare such as treatment planning, examination of patients’ medical records, representation of pedigrees and family history, communication and shared decision making, and life management and health monitoring. Faisal and colleagues regard information visualization as the

interactive visual representation of data generated by computer-supported tools. Some articles explore the deployment of information visualization and data visualization into social sciences. Orford et al. (1999) analyze a list of works on information visualization within the context of social sciences. This study reveals that there is no concerted finding in social sciences, compared to natural sciences. They conclude that the future of visualization in social sciences is likely to be similar as in the past, but more diffusing and ephemeral. Other researchers have focused on investigating research trends in information visualization evaluation. Lam et al. (2011) conduct an extensive review of over 800 visualization publications collected from EuroVis, InfoVis, and VAST conferences as well as IVS journal. They survey evaluation research in information visualization and propose seven evaluation scenarios which can be classified into process and visualization. Isenberg et al. (2013) systematically review 581 papers from ten years of IEEE Visualization conference publications to investigate the state and historic development of evaluation practices. This study reveals that there was a steady increase in evaluation methods that include UX testing. However, these preceding studies cover only a fraction of the published articles in identifying emerging trends and recent developments in information visualization and data visualization. In addition, to our knowledge, there has been little concerted effort to quantitatively understand the similarity and difference between information visualization and data visualization.

In this present study, we aim to investigate and compare the intellectual landscapes of information visualization and data visualization. To this end, we identify thematic patterns, landmark articles, and emerging trends of the domains. In addition, we employ citation indexing to expand the current study to thematically relevant research so that we could identify a more comprehensive landscape in information and data visualization research.

Methodology

This section introduces how we collect the datasets and analyze emerging trends and recent developments of information and data visualization research. Figure 3 displays the research process.

Data collection

The present study aims to investigate and compare the intellectual landscapes, thematic patterns, landmark articles, and emerging trends in information and data visualization research. To this end, we collected four datasets of bibliographic records, namely, “Info_{core}”, “Info_{exp}”, “Data_{core}”, and “Data_{exp}”, from the Web of Science, using a topic search and a citation expansion. The datasets obtained by the topic search are referred as the core datasets while the expanded ones represent a broader context of the core (Chen et al. 2014b). It is possible that the topic search do not include relevant articles if querying terms used for the topic search do not appear in the fields such as titles, abstracts, and keywords.² We assume that if an article cites at least one of the core records, then the article is thematically related to the topics (Chen et al. 2012). This citation expansion method supports our goal to explore much broader literature influencing emerging trends in the domains. A data collection method proposed by preceding literature (Chen et al. 2012, 2014a, b; Kim and Chen 2015) is employed as follows. The datasets are retrieved by a

² In this article, keywords refer to both author keywords and ISI keyword plus.

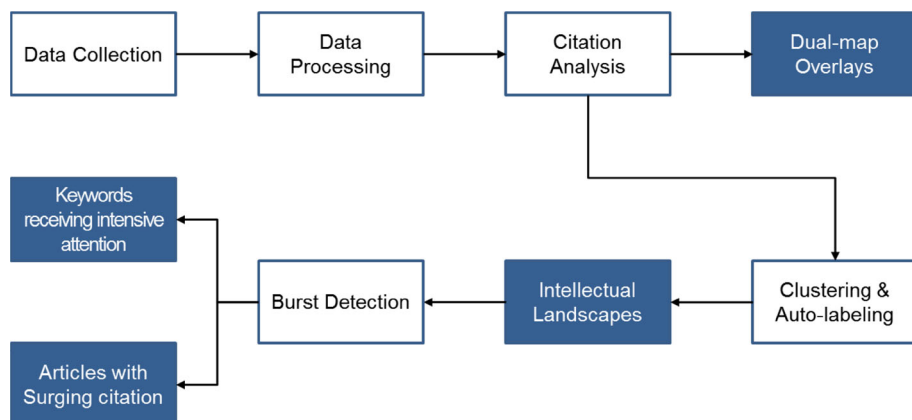


Fig. 3 Research procedure

topic search in the Web of Science from 2000 to 2014 inclusively. The search query for each domain is as follows: “information visual*” and “data visual*”. The wildcard “*” captivates relevant variations of a word such as “information visualisation”, “data visualizing”, etc. Queries returned 1019 (Info_{core}) and 1524 (Data_{core}) records of original research articles and review articles as of 31 March 2015. Then, through the citation expansion, we collected articles and reviews citing the core literature. In both core and expanded datasets, we excluded conference proceedings since the Web of Science’s coverage of conference proceedings is not adequate for our purposes. Instead of including partially represented conference proceedings, we chose to exclude them. On top of this, documents without keywords do not contribute to the subsequent analysis of keywords. Of all 23,795 records, however, 835 documents (3.5 %) do not have any keywords. We argue that this gap is not very significant. Finally, we have 286,736 keywords and they are not preprocessed to check and merge semantically similar keywords. Table 1 describes the brief statistics of the datasets.

Figure 4 illustrates a set diagram of the core datasets. As depicted in the figure, the overlapping records between two datasets are 67 which take only 6.6 % of Info_{core} and 4.4 % of Data_{core}. It indicates that information visualization and data visualization have been regarded as distinct domains in academia. We argue that this characteristic of the datasets helps us investigate domain-specific intellectual landscapes and research trends.

Figures 5 and 6 depict the yearly number of published papers retrieved by the topic search and the number of papers per year citing the literature in information visualization and data visualization respectively. As shown in the figures, both of the domains have received increasing attention from the community over time.

Exploring the intellectual landscapes in information visualization and data visualization

In this scientometric study, the intellectual landscapes of the domains are depicted by a network of a variety of entities such as publication statistics, cited references, and dual-map overlays. Particularly, we focus on document co-citation networks and networks of co-occurring keywords to explore past and present trends in information and data

Table 1 Brief statistics of the datasets

Dataset	Duration	Total	Articles	Reviews	Authors	Refs.	Keywords
Info _{core}	2000–2014	1019	1000	19	3414	36,108	8454
Info _{exp}	2000–2015	4740	4349	391	18,045	236,343	50,570
Data _{core}	2000–2014	1524	1463	61	6240	47,362	14,352
Data _{exp}	2000–2015	16,512	14,930	1582	86,385	900,114	213,360

Fig. 4 Venn diagram of the collected core datasets

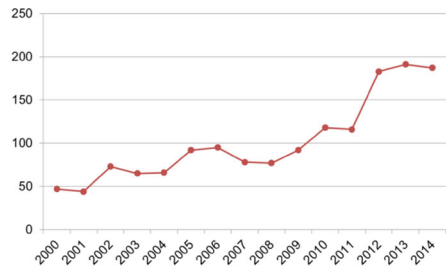
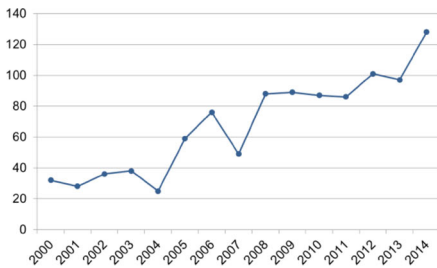
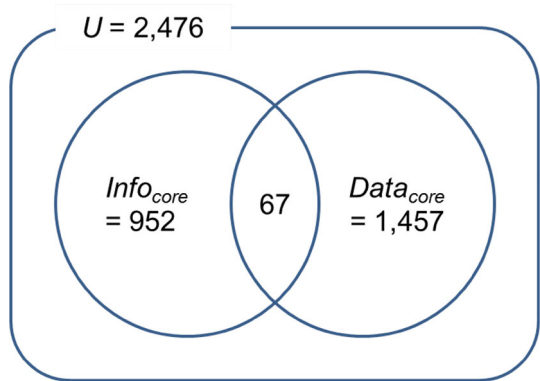


Fig. 5 Yearly number of published articles in InfoViz (left) and DataViz (right) retrieved by topic search from the Web of Science

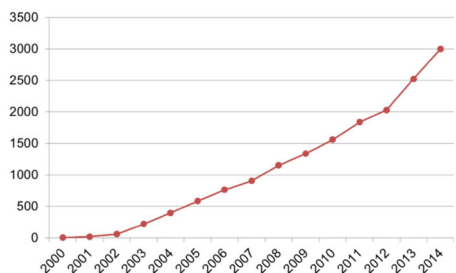
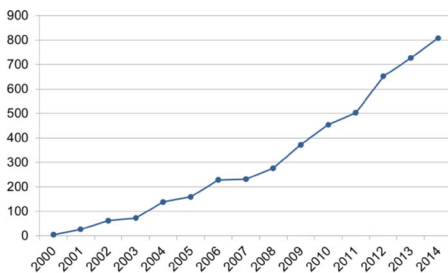


Fig. 6 Yearly number of citations to InfoViz (left) and DataViz (right) retrieved by citation expansion from the Web of Science

visualization research. The key features of our investigating the intellectual landscapes in the domains are threefold: (1) highly cited articles, (2) articles with citation bursts, and (3) frequently occurred keywords (Chen et al. 2014b). Measuring an article's impact on a domain by counting the cumulative frequency of citations may not properly capture its significance if the manuscript only receives intensive attention during a specific span of time. To address this issue, we employ Kleinberg's (2002) burst detection algorithm which models a burst of activity with certain features rising sharply in frequency. If an article has a strong surge of citation from others during a particular moment in time, then the article can be regarded as having an increasing attention to the underlying work. It overcomes the limitation of just considering the cumulative number of received citations as a measure of an impact of an article. For the purpose of the study, we employ CiteSpace. CiteSpace is a scientometric software to generate and analyze networks of co-cited references based on bibliographic records retrieved from the Web of Science (Chen 2006). The input for the software is a collection of scientific literature relevant to a specific topic. This tool computationally captivates and visualizes thematic patterns and emerging trends from the collected dataset. In addition, CiteSpace provides a visual representation called a dual-map overlay which depicts citation patterns at a domain level (Chen and Leydesdorff 2014). We use the dual-map representations to explore how far information visualization and data visualization diverge from each other in the global perspective.

Scientometric landscapes from the core datasets

Categorical trends of literature

A subject category of an article is one of the evidence showing a thematic concentration of the writing. We begin with examining which categories are the most popular in the literature of Info_{core} and Data_{core}. In doing so, we aim to captivate overarching research themes in the core datasets at a disciplinary level. Tables 2 and 3 list the Web of Science

Table 2 Statistics of the literature in Info_{core} in the Web of Science category

Rank	WoS category	No. of records	Percentage of records
1	Computer science software engineering	480	47.105
2	Computer science information systems	167	16.389
3	Information science library science	107	10.500
4	Computer science artificial intelligence	106	10.402
5	Computer science theory methods	98	9.617
6	Computer science interdisciplinary applications	55	5.397
7	Computer science cybernetics	45	4.416
8	Engineering electrical electronic	35	3.435
9	Medical informatics	33	3.238
10	Ergonomics	30	2.944

(1) Dark gray—the same categories with the same ranks in both datasets, (2) medium gray—the same categories with different ranks in each dataset, (3) light gray—categories that only appear in one dataset but not in the other

Table 3 Statistics of the literature in Data_{core} in the Web of Science category

Rank	WoS category	No. of records	Percentage of records
1	Computer science software engineering	197	12.952
2	Computer science artificial intelligence	178	11.703
3	Computer science interdisciplinary applications	171	11.243
4	Computer science information systems	134	8.810
5	Biochemical research methods	125	8.218
6	Statistics probability	114	7.495
7	Mathematical computational biology	104	6.838
8	Computer science theory methods	98	6.443
9	Biotechnology applied microbiology	96	6.312
10	Engineering electrical electronic	91	5.983

(1) Dark gray—the same categories with the same ranks in both datasets, (2) medium gray—the same categories with different ranks in each dataset, (3) light gray—categories that only appear in one dataset but not in the other

subject categories assigned to each record of the datasets. Each category is ranked by the number and proportion of articles out of the total number of literature. In the tables, the total numbers of records are slightly different from the statistics described in Table 1. It is due to the fact that a bibliographic record can be assigned to multiple categories in the Web of Science. In each table, categories are colored differently: (1) dark gray: the same categories with the same ranks in both datasets, (2) medium gray: the same categories with different ranks in each dataset, and (3) light gray: categories that only appear in one dataset but not in the other.

As shown in Tables 2 and 3, “Computer science software engineering” is the only subject category ranked at the same position in both records (rank #1). In addition, five additional categories relevant to computer science and engineering are assigned to both domains’ articles (see medium gray colored rows in each table). It indicates that computational techniques and applications are the main research themes in both domains. However, categorical distributions of the records show different patterns. As observed in the third column of Table 2, “Computer science software engineering” is the dominant category in Info_{core} (47.1 %), and the categories from rank #1 to rank #5 take up more than 84 % of the entire records. Contrary to this, the subject categories in Data_{core} are almost equally distributed (see the % of records column in Table 3). We argue that Data_{core} is evolving into a variety of applications such as biomedicine, statistics, etc. while Info_{core} largely focuses on computational aspects and human–computer interaction (“Ergonomics”).

Keywords as indicators of emerging trends and new developments

Tables 4 and 5 describe the top seven keywords with the strongest bursts. As shown in the tables, a variety of issues in the domains has received an intensive degree of attention from the community. Based on the beginning years of bursts, research trends in the domains can be grouped into two phases. It is obvious that Info_{core} and Data_{core} scrutinize algorithmic concepts underlying the domains during the first phase as represented in such words “algorithms”, “data mining”, “neural networks”, and “volume rendering”. Then, two domains

Table 4 Top 7 bursting keywords in Info_{core}

Phase	Keywords	Burst	Begin	End	2000–2014
Phase-1	Algorithms	2.927	2000	2004	
	Visual data mining	3.682	2002	2005	
Phase-2	Collaboration	2.943	2007	2009	
	Visual analytics	2.970	2008	2009	
	Visualization	3.148	2009	2009	
	Social networks	3.047	2013	2014	
	Analytics	2.874	2013	2014	

start to digress from each other during the second phase. In this phase, Info_{core} focuses on analytic perspectives and social networks whereas “dimensionality reduction” is among the leading research theme in Data_{core}.

Document co-citation networks

Figures 7 and 8 visualize the comprehensive research landscapes in Info_{core} and Data_{core}. The topologies consist of 100 highly cited articles within a given slice of time in each dataset. Each node represents a cited reference. The size of a node is proportional to its cited frequency. References with citation bursts are rendered with rings in red. Landmark articles cited more than or equal to 50 times are labeled in black. The more references the clusters have, the higher rankings they are given. The color legend at the top indicates that cooler colors are closer to the year of 2000 whereas hotter ones represent close years to 2014. Referring to the color, we can keep track of the thematic trends and temporal developments in Info_{core} and Data_{core}.

Landscape of information visualization (Info_{core})

It is evident that cluster #0 has a high concentration of nodes with citation bursts, which indicates that it is the most recently formed cluster. Considering both the size of member nodes and citation bursts to references, clusters #0 and #1 are the emerging ones in the domain with most of the recent publications. We assume that this is because this domain is relatively new and still developing. We focus on these clusters for further analysis. Table 6 summarizes the clusters in terms of the number of references included, labels extracted from titles and abstracts, and average year of cited articles.

The largest cluster, #0 *interaction*, contains 178 references and the mean year of publications is 2002. As discussed in Introduction, the interactive nature is one of the most fundamental themes in information visualization research. This cluster’s another label by TFIDF is “wordle” which is a word cloud generation tool. This technique gives greater

Table 5 Top 7 bursting keywords in Data_{core}

Phase	Keywords	Burst	Begin	End	2000–2014
Phase-1	Data mining	4.915	2000	2005	
	Exploratory data analysis	3.489	2000	2005	
	Neural networks	3.370	2000	2006	
	Volume rendering	3.155	2001	2005	
Phase-2	Dimensionality reduction	3.875	2008	2011	
	Framework	4.348	2011	2014	
	Exploration	3.529	2011	2012	

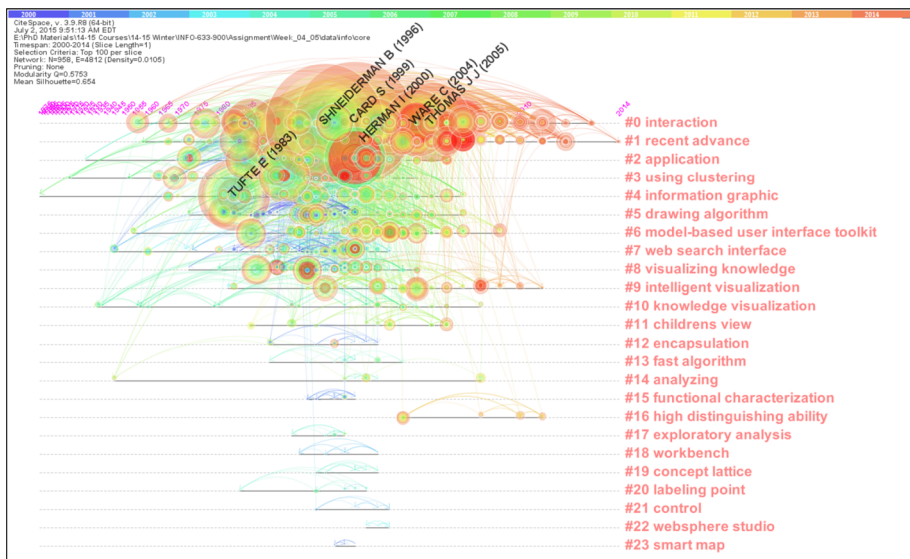


Fig. 7 (Color Online) Timeline visualization (Info_{core})

prominence to words which frequently appear in the given text. Users can adjust their outputs in a visually appealing way with different fonts, layouts, and color schemes. This observation indicates that investigating the concept of interaction is the most important and recent in Info_{core}.

The second largest cluster is #1 *recent advance* containing 105 articles with the average year of 2002. The most active citer to the cluster is Liu et al. (2014) titled “a survey on information visualization: recent advances and challenges”. In this review, Liu and

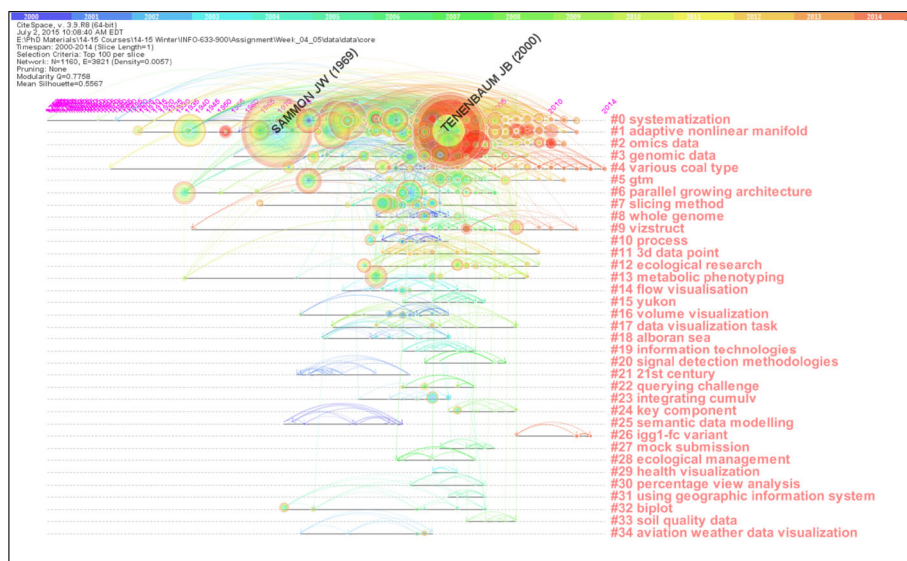


Fig. 8 (Color Online) Timeline visualization ($Data_{core}$)

Table 6 Brief summary of the emerging clusters in $Info_{core}$

Cluster	No. of refs.	Label ^a			Ave. year of citees
		TFIDF	LLR	MI	
0	178	Wordle	Interaction	Interaction constraint	2002
1	105	Error	Recent advance	Exploring flow	2002

^a Cluster labeling techniques—LLR log-likelihood ratio, MI mutual information

colleagues argue that empirical methodologies, user interactions, visualization frameworks, and applications are the main themes fast-rising in the domain.

Landscape of data visualization ($Data_{core}$)

Figure 8 displays the timeline representation derived from $Data_{core}$. As shown in the figure, there are a few recently emerging clusters. Cluster #1 has a high concentration of nodes in the light of citation bursts. Clusters #0 and #2 appear to have recent publications. Again, we argue that it is due to the fact data visualization is relatively new and still emerging. We focus on these clusters for further analysis. Table 7 summarizes the clusters.

The largest cluster, #0 *systematization*, contains 165 publications and the average year of references is 1993. As discussed in the section “*Categorical trends of literature*”, implementing visual techniques into biomedicine and other domains is one of the major research trends in $Data_{core}$. This observation is also justified by cluster #2. It indicates that the application of visualization technique to the domain of biomedicine is the emerging trend and recent development in $Data_{core}$.

The second largest cluster, #1 *adaptive nonlinear manifold*, has a high concentration of nodes with citation bursts. Nonlinear dimensionality reduction is an essential step in the

Table 7 Brief summary of the emerging clusters in Data_{core}

Cluster	No. of refs.	Label ^a			Ave. year of citees
		TFIDF	LLR	MI	
0	165	Frequency-based visualization	Systematization	Bioinformatic progress	1993
1	133	Pattern recognition	Adaptive nonlinear manifold	Spatial quantile	1995
2	81	Systems biology	Omics data	Processing	2005

^a Cluster labeling techniques—LLR log-likelihood ratio, MI mutual information

field of pattern recognition (Yin and Huang 2010). As discussed in the section “[Keywords as indicators of emerging trends and new developments](#)”, dimensionality reduction is one of the emerging research themes for effective data processing and visualization.

References with surging citation

The emerging trends and new developments in sciences can be captivated by manuscripts receiving surging citations during a particular span of time. A citation burst indicates that the scientific community has paid or is paying intensive attention towards the underlying work (Chen et al. 2014b).

Landmark articles in Info_{core}

Table 8 lists the top 20 references with the strongest citation bursts in the Info_{core} dataset. Papers are categorized into two phases, based on the beginning years of bursts. In introducing these references, we first selected papers with the strongest bursts (colored gray in the table) from articles that started to burst at the same time. Then, we excluded an article if it is a review paper, since we assume that they do not affect thematic patterns and emerging trends in the domain. We also do not introduce later articles published by same authors and handbooks in the domain. Five references selected in this way are summarized in Table 9.

Morse et al. (2000) propose a visual taxonomy for evaluating information visualization systems. To address the lack of evaluation criteria for visual representation, the authors create a new visual taxonomy which is domain-independent and allows testing of the visualization in isolation from the rest of the systems. Feteke (2004)’s writing introduces a visual analytic tool called InfoVis Toolkit. The system aims to simplify the management of rich data structures and the creation and extension of various visualizations. Ellis and Dix (2007) examine a large number of clutter reduction methods and propose a taxonomy which can be used as a guide to match techniques to specific problems in display. Cleveland and McGill (1984) aim to establish a scientific foundation of information visualization for systematic data analysis and representation, examining a theory called graphical perception. The authors provide a guideline for grounded graph construction. Yi et al. (2007) frame seven categories of interaction techniques in information visualization: (1) select, (2) explore, (3) reconfigure, (4) encode, (5) abstract/elaborate, (6) filter, and (7) connect. The implication of the study is that the proposed categories can be used in designing, deploying, and evaluating interaction techniques.

Table 8 Top 20 references with strongest citation burst in Info_{core}

Phase	References	Burst	Begin	End	2000–2014
Phase-1	Card et al. (1999)	18.5834	2000	2006	
	Ware (2000)	9.1484	2000	2005	
	Schneiderman (1998)	5.0212	2000	2005	
	Chen (1999)	4.5637	2000	2005	
	Tufte (1990)	4.5569	2000	2007	
	Robertson et al. (1993)	4.4051	2000	2001	
	Robertson et al. (1991)	3.6873	2000	2003	
	Spence (2000)	3.6992	2002	2005	
	Morse et al. (2000)	3.8901	2003	2007	
	Leung et al. (1994)	3.9304	2004	2008	
Phase-2	Tufte (1990)	3.6474	2007	2008	
	Fekete (2004)	4.0649	2008	2009	
	Isenberg (2007)	3.8152	2008	2012	
	Ware (2004)	8.0773	2009	2012	
	Ellis et al. (2007)	4.1019	2010	2014	
	Cleveland et al. (1984)	4.9463	2011	2014	
	Segel (2010)	3.8566	2011	2014	
	Tory (2005)	3.7391	2011	2014	
	Yi et al. (2007).	4.0176	2012	2014	
	Elmqvist et al. (2011)	3.7151	2012	2014	

The exploration of the landmark literature from the Info_{core} dataset reveals that there has been concerted effort to build up a science of information visualization. In other words, the collection of articles which have strong citation bursts have tried to establish scientific foundations in various aspects of information visualization such as elemental taxonomy,

Table 9 References with the strongest citation bursts (Info_{core})

References	Phase	Citation burst		
		Strength	Begin	End
Morse et al. (2000)	1	3.8901	2003	2007
Fekete (2004)	2	4.0649	2008	2009
Ellis and Dix (2007)	2	4.1019	2010	2014
Cleveland and McGill (1984)	2	4.9463	2011	2014
Yi et al. (2007)	2	4.0176	2012	2014

and perceptual theory. Furthermore, we found that the concept of interactivity is of fundamental importance to the field. This argument coheres with the finding from the section “[Landscape of information visualization \(Info_{core}\)](#)”. It is also acknowledged that the community has tried to integrate a variety of visual techniques and data structures into a single framework.

Landmark articles in Data_{core}

Table 10 lists the top 20 references with the strongest citation bursts in Data_{core}. We applied the same selection criteria clarified in the section “[Landmark articles in Info_{core}](#)” and discuss four papers. The selected literature is summarized in Table 11.

Törönen et al. (1999) investigate an unsupervised neural network learning algorithm called self-organizing map (SOM) and deploy it into analyzing yeast gene expression. Inselberg and Dimsdale (1990) present a methodology for visualizing analytic and synthetic geometry, employing the concept of parallel coordinates which induces a non-projective mapping between N-dimensional and two-dimensional sets. Shannon et al. (2003)’s article proposes a system called Cytospace where large databases of protein–protein, protein–DNA, and genetic interactions are processed and visualized with expression profiles, phenotypes, and other molecular states. Tenenbaum et al. (2000) suggest an approach to solving the problem of dimensionality reduction in human brain research, measuring local metric information to learn the underlying global geometry of a dataset.

It is often acknowledged that researchers working with massive datasets often confront the problem in finding the meaningfully small number of relevant features latent in their original dimensions. The influential manuscripts in data visualization have proposed novel approaches seeking an important class of data manifolds. Research trends identified by bursting keywords in the first phase (see Table 5) and emerging clusters (see Table 7) support this observation. On the other hand, a neural network was the algorithm receiving strong attention during the first phase of Data_{core} as addressed in the section “[Keywords as indicators of emerging trends and new developments](#)”. The surging citation to the literature is also consistent with that finding. Finally, we found that the application of data visualization techniques to biomedicine and other fields are emerging trends and recent developments in the domain.

Dual-map overlay

The dual-map overlay, proposed by Chen and Leydesdorff (2014), presents the disciplinary concentrations of given articles through their citation paths. The global base map depicts the interconnections of over 10,000 scientific journals and these journals are grouped into

Table 10 Top 20 references with strongest citation burst in Data_{core}

Phase	References	Burst	Begin	End	2000–2014
Phase-1	Card et al. (1999)	3.2746	2000	2005	
	Ripley (1996)	3.1160	2000	2004	
	Törönen et al. (1999)	3.1449	2002	2004	
	Swayne et al. (1998)	3.0141	2002	2003	
	Inselberg et al. (1990)	3.3895	2003	2006	
	Kohonen (1997)	3.1301	2003	2005	
Phase-2	Duda et al. (2001)	3.7706	2007	2010	
	Shannon et al. (2003)	3.4787	2010	2014	
	Yin (2008)	3.1486	2010	2011	
	Card et al. (1999)	3.6331	2011	2014	
	Jolliffe (2002)	3.5664	2011	2014	
	van der Maaten et al. (2008)	3.2457	2011	2014	
	Gehlenborg et al. (2010)	2.8707	2011	2014	
	Tenenbaum et al. (2000)	4.5767	2012	2014	
	Venna et al. (2010)	4.0443	2012	2014	
	Lee et al. (2007).	3.4448	2012	2014	
	Krzywinski et al (2009).	3.2522	2012	2014	
	Belkin et al. (2003)	3.0017	2012	2014	
	Torgerson (1952)	2.9525	2012	2014	
	Rauber et al. (2002)	2.8566	2012	2014	

regions that represent publication and citation activities at a domain level (Chen et al. 2014a). Given a dual-map overlay, we can identify patterns of how specific domains (citing articles) are influenced by other domains (cited literature). Figures 9 and 10 display the dual-map overlays of Info_{core} and Data_{core}. The left clusters, i.e., research front, represent

Table 11 References with the strongest citation bursts (Data_{core})

References	Phase	Citation burst		
		Strength	Begin	End
Törönen et al. (1999)	1	3.1449	2002	2004
Inselberg and Dimsdale (1990)	1	3.3895	2003	2006
Shannon et al. (2003)	2	3.4787	2010	2014
Tenenbaum et al. (2000)	2	4.5767	2012	2014

where the literature in each dataset publishes, while the right ones, i.e., intellectual base, indicate where they cite. Regions are marked by common terms found in the underlying journals. The thickness of lines is proportional to a z -score-scaled frequency of citation.

As displayed in the figures, there are six main stream citation paths in Info_{core} and five dominant citation relationships in Data_{core}. Table 12 briefly summarizes these paths with citing and cited region names. The relationships are sorted by the z -scores in descending order where the values are rounded to the nearest thousandth. Each row is colored in grayscale by the same color of the corresponding path shown in the dual-map overlays. As described in Table 12, it is evident that the domain the most frequently covering Info_{core} and Data_{core} is “mathematics, systems, mathematical” and the literature is mostly affected by “systems, computing, computer”. We also found that the discipline “molecular, biology, genetics” affects Data_{core} more than Info_{core}. These findings cohere with the thematic trends of publications identified by the subject category assignment, scientometric landscapes, and landmark literature of the domains.

Of the cited domains, publication venues receiving intensive citations during a specific span of time are listed in Tables 13 and 14. Evidently, the literature in Data_{core} is influenced by more various domains than Info_{core}. The publication venues with surging citations in Data_{core} include computer science, information visualization, statistics, and graphics engineering whereas Info_{core} is mostly affected by research published in widely recognized venues in information visualization.

Emerging trends and recent developments in a broader context

In this section, we extend the present study to investigating intellectual landscapes of published literature influenced by the core datasets. It is argued that the citation expansion allows us to explore the shared contexts and domain-specific themes of the disciplines in a much comprehensive way. To this end, we integrated core datasets with expanded ones and relabeled them as Info_{exp} and Data_{exp} for the rest of the study.

Investigation of bursting keywords

The investigation of the datasets obtained by the citation expansion gives an understanding of how the core literature has influenced successive research. Keyword co-occurrence networks are displayed in Figs. 11 and 12. We let the links between nodes transparent. In Info_{exp}, keywords on biomedicine start to appear and burst, forming a distinct cluster from general keywords. Compared to this, biomedical keywords in Data_{core} co-occur with ones in information studies. We assume that researchers in Info_{exp} distinguish the application of

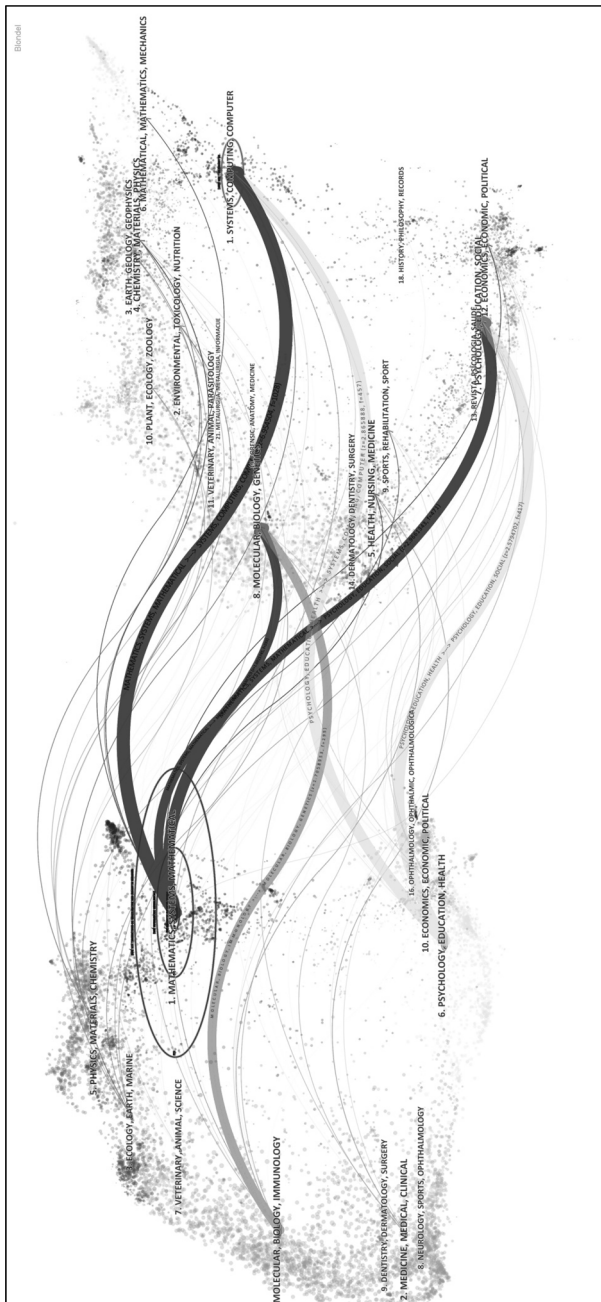


Fig. 9 Dual-map overlay of Info_{core} (the *left* clusters represent citing domains while the *right* ones are cited domains)

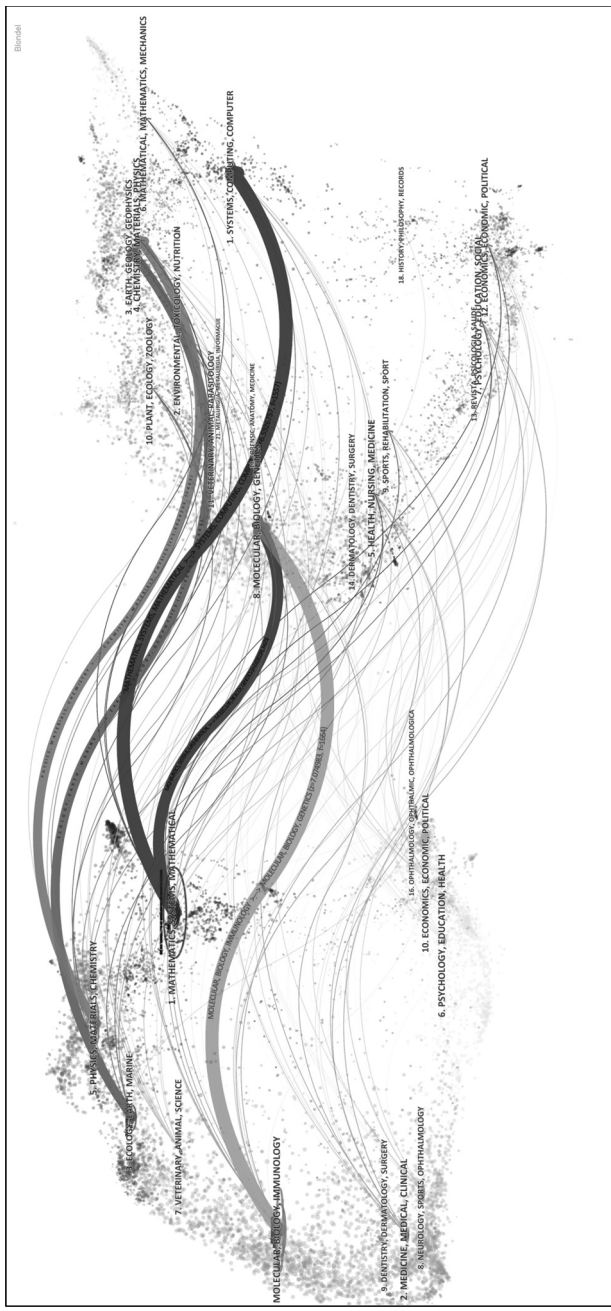


Table 12 Citation trends in Info_{core} and Data_{core} at a domain level

Dataset	Citing domain	Cited domain	z-score
Info _{core}	Mathematics, systems, mathematical	Systems, computing, computer	6.954
	Mathematics, systems, mathematical	Psychology, education, social	2.966
	Psychology, education, health	Systems, computing, computer	2.866
	Psychology, education, health	Psychology, education, social	2.579
	Molecular, biology, immunology	Molecular, biology, genetics	1.706
Data _{core}	Mathematics, systems, mathematical	Molecular, biology, genetics	1.692
	Molecular, biology, immunology	Molecular, biology, genetics	7.075
	Mathematics, systems, mathematical	Systems, computing, computer	6.593
	Mathematics, systems, mathematical	Molecular, biology, genetics	2.608
	Ecology, earth, marine	Earth, ecology, geophysics	1.661
	Physics, materials, chemistry	Chemistry, materials, physics	1.661

Table 13 9 Journals and conferences with citation bursts (Info_{core})

Phase	Journal/conference	Burst	Begin	End	2000–2014
Phase-1	Readings in Information Visualization	22.3851	2000	2006	
	IEEE Symposium on Information Visualization	17.4958	2000	2006	
	Proceedings of Information Visualization '99	8.0107	2000	2005	
	Designing the User Interface	6.781	2001	2007	
Phase-2	IEEE Symposium on Visual Analytics Science and Technology	6.0124	2008	2009	
	IEEE Symposium on Information Visualization	19.4805	2012	2014	
	IEEE Symposium on Visual Analytics Science and Technology	8.0436	2012	2014	
	CHI 2010	8.0281	2012	2014	
	CHI 2008	7.6903	2012	2014	

information visualization to biomedicine from the preceding stream research, whereas the application is regarded as an extension of data visualization research in the counterpart community.

Table 14 9 Journals and conferences with citation bursts (Data_{core})

Phase	Journal/conference	Burst	Begin	End	2000–2014
Phase-1	SIAM Journal on Scientific and Statistical Computing	6.4280	2000	2005	
	IEEE Symposium on Information Visualization	4.7992	2000	2005	
	Journal of the American Statistical Association	5.8286	2001	2004	
	IEEE Computer Graphics and Applications	4.5839	2001	2003	
	Self-Organizing Maps	10.8472	2002	2006	
	Journal of Computational and Graphical Statistics	6.0443	2002	2005	
	Computational Statistics	4.7500	2002	2005	
	IEEE Transactions on Computers	4.3602	2002	2006	
Phase-2	PLoS ONE	13.3335	2012	2014	

Then, we extracted the top 20 keywords with the strongest bursts from the expanded datasets (see Tables 15, 16). They show bipartite bursting trends having several years of pause between phases. With a few exceptions such as “cluster” (Info_{exp}), “information-retrieval” (Info_{exp}), “data visualization” (Data_{exp}), “data mining” (Data_{exp}), “database” (Data_{exp}), and “information resource” (Data_{exp}), most of the dominant keywords during the first phase of Info_{exp} and Data_{exp} are from the field of biomedicine. It indicates the deployment of information and data visualization techniques is the major research theme in expanded data visualization studies. In addition to that, it is obvious that information visualization is mostly employed by molecular biology research whereas genomics uses techniques from data visualization in modeling gene expression and DNA sequence. In phase-2, information visualization research has affected a wider range of topics such as “computer graphics”, “management”, “social media”, and “electronic tongue” than data visualization.

Intellectual landscapes

Document co-citation network in Info_{exp}

Figures 13 and 14 display topology view and timeline visualization of the co-citation network from Info_{exp}. The topologies consist of 200 highly cited articles within a given

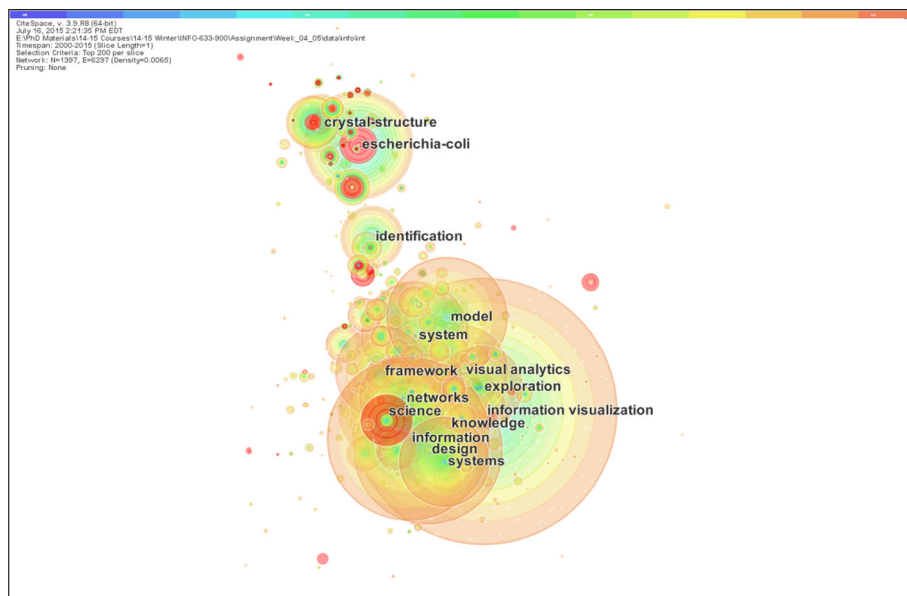


Fig. 11 Keyword co-occurrence network from Info_{exp} (keyword threshold: 100 times)

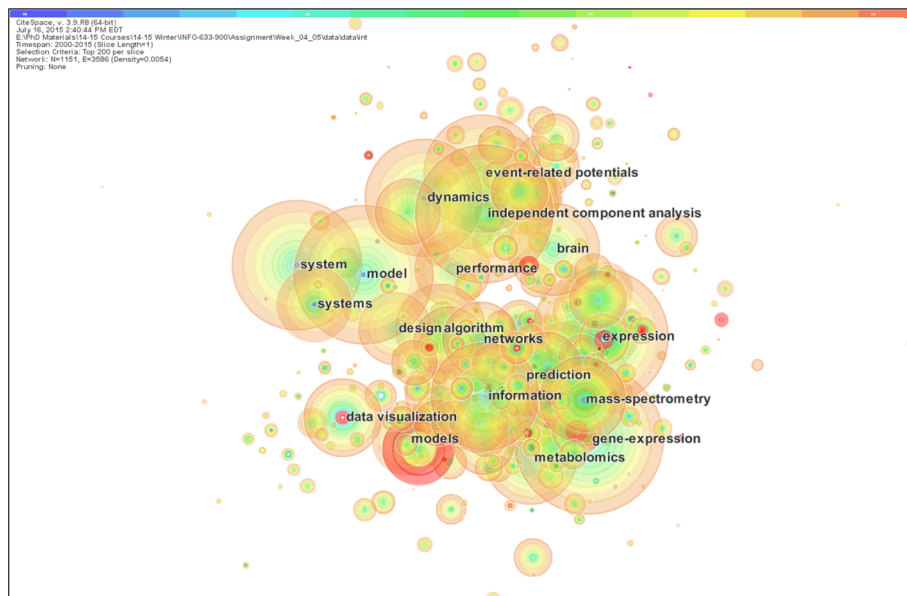


Fig. 12 Keyword co-occurrence network from Data_{exp} (keyword threshold: 300 times)

slice of time in each dataset. The edges are processed as transparent. Landmark articles cited more than or equal to 150 times are labeled in black. The manuscripts by Card, Schneiderman, Herman, and Thomas are still influential in the broader context. The most impacting article is by Sofia et al. (2001) on employing information visualization

Table 15 20 Bursting keywords in Info_{exp}

Phase	Keywords	Burst	Begin	End	2000–2015
Phase-1	<i>Escherichia-coli</i>	25.3059	2001	2006	
	Protein	16.3327	2001	2007	
	s-Adenosylmethionine	10.0154	2001	2006	
	Lysine 2	9.9207	2001	2005	
	Pyruvate formate-lyase	9.5703	2001	2005	
	Anaerobic ribonucleotide reductase	8.5371	2001	2007	
	Activating enzyme	8.0448	2001	2007	
	Iron–sulfur center	7.9160	2001	2005	
	Electron-paramagnetic-resonance	6.2284	2001	2007	
	Biotin synthase	5.8789	2001	2004	
	Cluster	5.7090	2001	2005	
	Radical mechanisms	10.1094	2002	2006	
	Biosynthesis	6.3286	2002	2007	
	Information-retrieval	5.3847	2002	2008	
	Crystal-structure	6.7024	2004	2008	
	Protein–protein interactions	5.3428	2005	2009	
Phase-2	Computer graphics	7.2612	2012	2013	
	Management	5.7650	2012	2015	
	Social media	8.3863	2013	2015	
	Electronic tongue	6.8881	2013	2015	

techniques to discover a novel protein superfamily. References which have close linkage with this manuscript also receive intensive citations. This observation is consistent with the finding from the investigation of bursting keywords.

Table 16 20 Bursting keywords in Data_{exp}

Phase	Keywords	Burst	Begin	End	2000–2015
Phase-1	Data visualization	26.3725	2000	2004	
	Data mining	13.0723	2000	2005	
	Gene expression	11.8781	2000	2008	
	Saccharomyces-cerevisiae	17.0809	2002	2006	
	Microarray	24.6587	2003	2007	
	Sequence	14.5957	2003	2006	
	Gene-expression	11.0185	2003	2006	
	Messenger-RNA	10.7802	2003	2007	
	Galaxy:center	10.5628	2003	2008	
	Infrared:stars	8.8014	2003	2008	
	Functional genomics	27.1472	2004	2008	
	Mouse	21.0223	2004	2007	
	Resource	18.3577	2004	2007	
	Database	13.3774	2004	2005	
	Thaliana	13.1200	2004	2008	
	Biology	12.5103	2004	2006	
	Information resource	9.2435	2004	2008	
	Transcriptomes	9.3928	2005	2007	
Phase-2	RNA-seq	15.4237	2013	2015	
	Connectivity	8.8701	2013	2015	

Considering both the size of member nodes and citation bursts to references, clusters #0 through #5 are the emerging ones in the broader context. Table 17 summarizes these clusters.

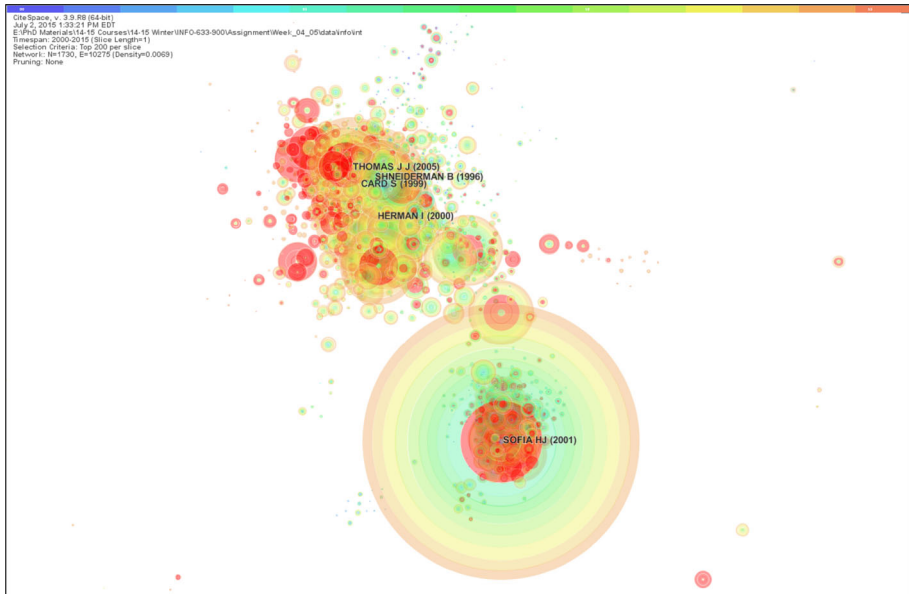


Fig. 13 Document co-citation network in Info_{exp}

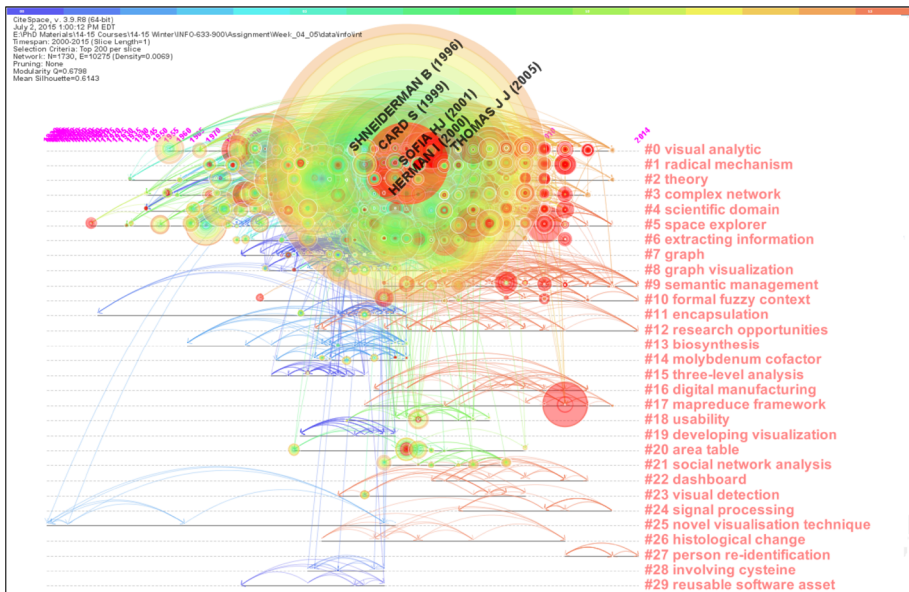


Fig. 14 Timeline visualization in Info_{exp}

The largest cluster, #0 *visual analytic*, contains 313 cited articles and the average year of publications is 2002. This cluster is also labeled as information visualization novice by TFIDF. It indicates that designing and developing visual components for cognitive and instructional support is among the leading topics in the broader context.

Table 17 Brief summary of the emerging clusters in Info_{exp}

Cluster	No. of refs.	Label			Ave. year of citees
		TFIDF	LLR	MI	
0	313	Information visualization novice	Visual analytic	Analyzing mutual information	2002
1	249	Reaction	Radical mechanism	Auxiliary iron–sulfur cluster	1999
2	189	Data exploration	Theory	<i>Arabidopsis thaliana</i>	1995
3	185	Community structure	Complex network	Choreographic information	2000
4	149	Essential science structure	Scientific domain	Bibliometric investigation	1997
5	125	Electronic tongue	Space explorer	Adaptive privacy-preserving visualization	1997

Clusters #1, #2, and #5 consist of references on applying information visualization to biomedicine. As discussed above, the deployment of theory and practice in information visualization into biomedical research is the prominent research trend.

The fourth largest cluster is #3 *complex network* containing 185 cited references with the average year of 2000. Complex networks often occur in real-world graphs modelling systems. In complex networks, the detection of community structure is to identify a homogeneous set of entities. The study of complex networks is currently an emerging and active area of scientific research. Cluster #4, labeled as *scientific domain* by LLR, contains 149 articles. Scientometrics is the quantitative study of science. It aims to understand the structure and dynamics of a research topic or a scientific field. Information visualization enables complex networks and intellectual structure of science to be intuitively communicated to their audience.

Document co-citation network in Data_{exp}

Figures 15 and 16 illustrate intellectual landscapes from Data_{exp}. Landmark articles cited more than or equal to 300 times are labeled in black. In Fig. 15, the writing by Tenebaum is still influential in the broader context (see Fig. 8). Delorme and Makeig (2004) and Su et al. (2002) are among the most influential manuscripts. They receive 2420 and 971 citations respectively. Delorme and Makeig (2004) develop a graphic user interface to explore electrical activity in the brain. Su et al. (2002) investigate human and mouse transcriptomes with data visualization. This observation is consistent with the finding from the investigation of bursting keywords. In Figs. 15 and 16, we let these two nodes invisible since the size of the nodes hinders from recognizing the entire structure.

As depicted in Fig. 16, a variety of topics are emerging in the broader context of data visualization. It is obvious that clusters #0, #1, #2, #3, #4, #10, and #12 are leading the research trends. Table 18 summarizes these clusters.

The largest cluster, #0 *application*, contains 219 references and the mean year of publications is 1995. In data visualization, most of the algorithms aim to find a true projection of data from a high-dimensional space for multivariate data exploration and knowledge discovery. As discussed in sections “[Keywords as indicators of emerging trends and new developments](#)” and “[Document co-citation networks](#)”, dimensionality

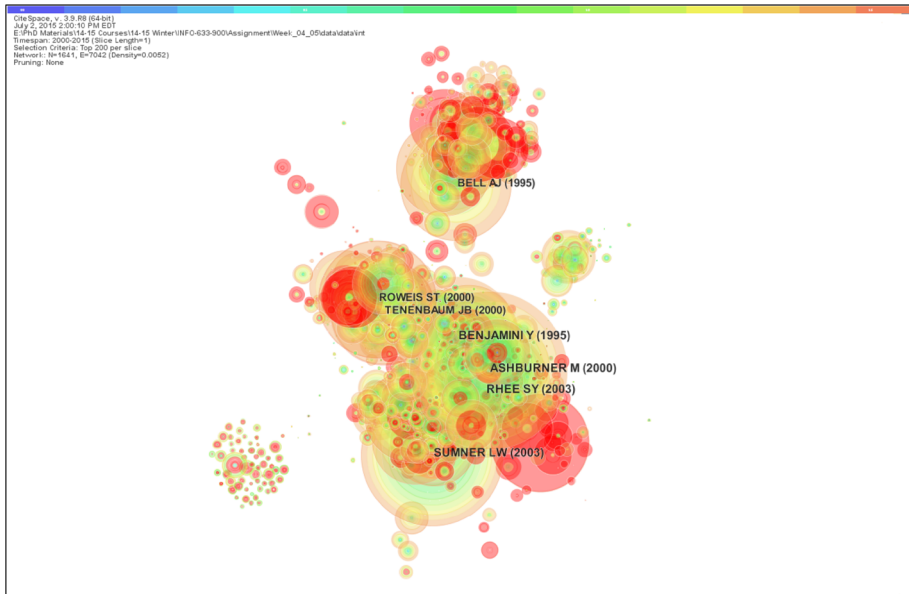


Fig. 15 Document co-citation network in Data_{exp}

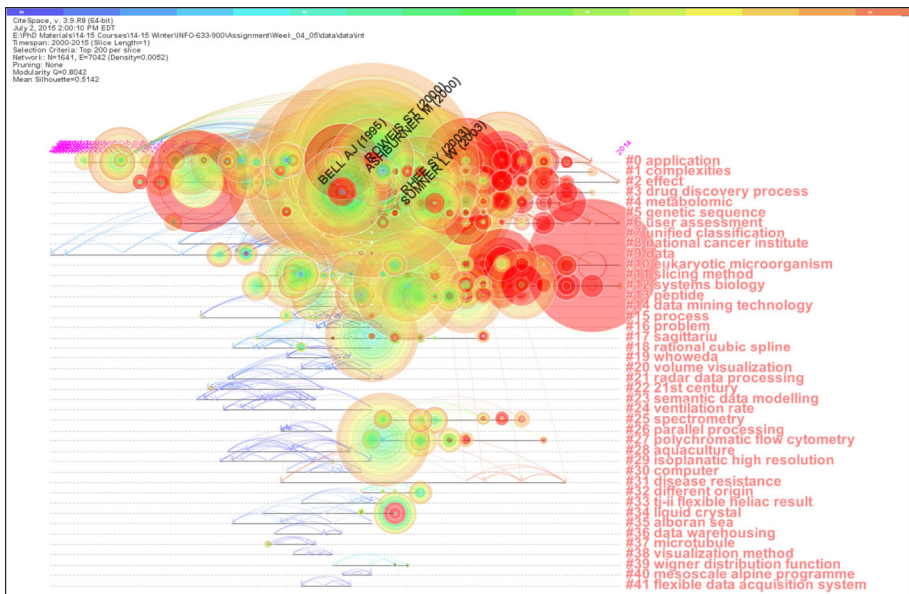


Fig. 16 Timeline visualization in Data_{exp}

reduction is of importance in data visualization research (see Tables 5, 7). The investigation into the cluster indicates that data processing is also among the leading topics in the broader context.

Table 18 Brief summary of the emerging clusters in Data_{exp}

Cluster	No. of refs.	Label			Ave. year of citees
		TFIDF	LLR	MI	
0	219	Manifold	Application	Hierarchical finite mixture model	1995
1	190	Gene expression	Complexities	Cancer tissue sample	2001
2	164	Oscillation	Effect	Canonical correlation analysis	2000
3	137	Genome	Drug discovery process	Molecular evolution rate	2001
4	131	Metabolomic	Metabolomic	Relationship	2003
10	40	Sequencing	Eukaryotic microorganism	Functional motif	2008
12	32	Systems biology	Systems biology	Hierarchical aggregation	2006

The third largest cluster is labeled as *effect* by LLR. It consists of 164 cited articles and the average year of citees is 2000. Oscillation pattern analysis and canonical correlation analysis are descriptive spatio-temporal statistical methods. They are often employed by systems biology research. With the exception of cluster #0, therefore, the dominant research theme in the broader context is the application of data visualization techniques to biomedical research. These clusters are younger than the former, meaning they are recently formed. These clusters cover a variety of research topics such as modelling gene expression, metabolic process, and microorganism.

Landmark articles with strong citation bursts

Influential manuscripts in Info_{exp}

Table 19 lists the top 50 references with the strongest citation bursts in Info_{exp}. We discuss four articles described in Table 20.

Sofia et al. (2001)'s study uses information visualization techniques to discover a novel protein superfamily with over 600 members named Radical SAM. The sequence similarity relationships in the superfamily are visualized, using the mBE measure of cluster cohesion to filter a dendrogram produced by hierarchical clustering of a distance matrix. Walsby et al. (2002b) employ information visualization methods to investigate an anchoring role for Fe–S clusters. Rzhetsky et al. (2004) propose a visual analytic system which integrates tasks such as automated selection of relevant articles and information extraction, information visualization, and generation of specialized knowledge bases. Moreland et al. (2005) propose a new software toolkit-MBT which is designed for protein analysis and visualization.

The investigation of the landmark literature from the Info_{exp} dataset reveals that successive research of the core literature has made concerted effort to apply information visualization techniques to discovering novel knowledge in biomedicine. It is also argued that the community has tried to integrate a variety of visual techniques into a single system. This observation coheres with the findings from sections “[Investigation of bursting keywords](#)” to “[Intellectual landscapes](#)”.

Table 19 50 Articles with strongest citation burst in the broader context of InfoViz

Phase	References	Burst	Begin	End	2000–2015
Phase-1	Card et al. (1999)	29.0447	2000	2006	
	Ware (2000)	14.3399	2000	2005	
	Chen (1999)	8.0122	2000	2005	
	Cheek et al. (2001)	18.9041	2001	2006	
	Ollagnier-de-Choudens et al. (2000)	12.9162	2001	2004	
	Duin et al. (1997)	11.6591	2001	2004	
	Ollagnier et al. (1997)	11.3737	2001	2006	
	Cosper et al. (2000)	10.1923	2001	2005	
	Miller et al. (2000)	9.4906	2001	2007	
	Bui et al. (1998)	9.0675	2001	2005	
	Lieder et al. (1998)	8.5951	2001	2006	
	Bui et al. (1999)	7.6792	2001	2005	
	Henshaw et al. (200)	7.6261	2001	2005	
	Guianvarc'h et al. (1997)	7.6261	2001	2005	
	Külzer et al. (1998)	7.4658	2001	2005	
	Bradford (1976)	7.0574	2001	2004	
	Sofia et al. (2001)	19.3604	2002	2006	
	Frey (2001)	15.8627	2002	2007	
	Marquet (2001)	11.2621	2002	2006	
	Walsby et al. (2002a)	10.3787	2002	2005	

Table 19 continued

Phase	References	Burst	Begin	End	2000–2015
	Fontecave et al. (2001)	10.1507	2002	2007	
	Hewitson et al. (2002)	9.3537	2002	2006	
	Ugulava et al. (2001a)	8.8126	2002	2005	
	Ollagnier-de-Choudens et al. (2002b)	8.6813	2002	2005	
	Sanyal et al. (1994)	8.5221	2002	2005	
	Ugulava et al. (2001b)	7.5235	2002	2005	
	Fabrikant et al. (2001)	7.3071	2002	2007	
	Walsby et al. (2002b)	8.2733	2003	2005	
	Ollagnier-de-Choudens et al. (2002a)	7.9042	2003	2007	
	Ugulava et al. (2002)	7.0652	2003	2007	
	Frey et al. (2003)	11.5888	2004	2007	
	Jarrett (2003)	10.6754	2004	2007	
	Berkovitch et al. (2004)	10.5411	2004	2007	
	Layer et al. (2003)	8.5501	2004	2007	
	Jenssen et al. (2001)	8.3272	2004	2007	
	Rzhetsky et al. (2004)	18.6260	2005	2008	
Phase-2	Moreland et al. (2005)	10.0234	2008	2011	
	Ware (2004)	12.4933	2009	2012	
	Frey et al. (2008)	8.6123	2010	2012	
	Vey et al. (2011)	7.8418	2011	2014	
	Stasko et al. (2008)	7.7145	2011	2014	
	Yan et al. (2011)	7.5994	2011	2012	

Table 19 continued

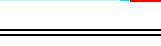
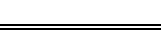
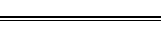
Phase	References	Burst	Begin	End	2000–2015
	Grove et al. (2010)	7.1686	2011	2014	
	Riul Jr. et al. (2010)	14.5788	2012	2014	
	Cobo et al. (2011)	9.5084	2012	2014	
	Aigner et al. (2011)	8.8844	2012	2014	
	Andronis et al. (2011)	8.2255	2012	2014	
	Blei et al. (2003)	7.6844	2012	2014	
	Munzner (2009)	7.6491	2012	2014	
	Del Valle (2010)	7.1313	2012	2014	

Table 20 Articles with surging citations (Info_{exp})

References	Phase	Citation burst		
		Strength	Begin	End
Sofia et al. (2001)	1	19.3604	2002	2006
Walsby et al. (2002b)	1	8.2733	2003	2005
Rzhetsky et al. (2004)	1	18.6260	2005	2008
Moreland et al. (2005)	2	10.0234	2008	2011

Influential manuscripts in Data_{exp}

Table 21 shows the top 50 articles with surging citations in Data_{exp} . We discuss 10 articles described in Table 22.

Velculescu et al. (1995) propose a method called serial analysis of gene expression (SAGE). The proposed method provides a broadly applicable means for the quantitative cataloging and comparison of expressed genes in a variety of normal, developmental, and disease states. A 2001 paper by Lander et al. aims to produce and make freely available a draft sequence of the human genome. Various statistical analyses are conducted and reported with various charts (e.g., frequency chart, 3D frequency chart, scatter plot, pie chart, line chart, etc.). Eisen et al. (1998) propose a system used for cluster analysis of genome-wide expression data from microarray hybridization. Su et al. (2002) analyze the human and mouse transcriptomes, investigating gene expressions from 91 humans and mice. To illustrate methods of extracting knowledge from the data, they employ data visualization techniques. Rhee et al.'s article published in 2003 investigates an online organism database called Arabidopsis Information Resource (TAIR). TAIRs SeqViewer supports various functions such as nucleotide view (shows 10 kb of sequence), close up

Table 21 50 Articles with strongest citation burst in the broader context of DataViz

Phase	References	Burst	Begin	End	2000–2015
Phase-1	Velculescu et al. (1995)	13.4274	2000	2006	
	Lander et al. (2001)	23.2995	2001	2006	
	Schena et al. (1995)	17.8931	2001	2006	
	Venter et al. (2001)	17.8911	2001	2006	
	Lockhart et al. (1996)	16.8352	2001	2006	
	Scherf et al. (2000)	14.5765	2001	2006	
	Eisen et al. (1998)	38.1939	2002	2006	
	Rabow et al. (2002)	22.9156	2002	2007	
	Brazma et al. (2001)	21.3195	2002	2006	
	Penterman et al. (2002)	15.2917	2002	2007	
	Tusher et al. (2001)	14.0306	2002	2007	
	Su et al. (2002)	109.3584	2003	2008	
	Chinwalla et al. (2002)	18.4255	2003	2008	
	Rhee et al. (2003)	46.8372	2004	2008	
	Sumner et al. (2003)	35.9021	2004	2007	
	Kaul et al. (2000)	21.6963	2004	2008	
	Roessner, et al. (2001)	20.3127	2004	2008	
	Roessner et al. (2000)	18.7853	2004	2008	
	Fiehn et al. (2000)	17.5907	2004	2008	
	Fiehn (2002)	17.1450	2004	2008	

Table 21 continued

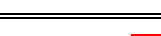
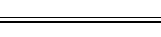
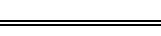
Phase	References	Burst	Begin	End	2000–2015
	Raamsdonk et al. (2001)	15.351	2004	2007	
	Zhang et al. (2004)	33.6849	2005	2009	
	Su et al. (2004)	21.7271	2005	2007	
	Binoemail et al. (2004)	16.9301	2005	2008	
	Zimmermann et al. (2004)	15.9069	2005	2008	
Phase-2	Salomonis et al. (2007)	16.8535	2008	2011	
	Tao et al. (2008)	21.8191	2010	2013	
	Asuncion et al. (2007)	13.7073	2010	2011	
	Gehlenborg et al. (2010)	29.5472	2011	2013	
	Maris et al. (2010)	20.3707	2011	2015	
	Wicks et al. (2010)	14.7330	2011	2013	
	Sugiyama (2007)	19.9878	2012	2015	
	Langmead et al. (2009)	15.8380	2012	2015	
	Luck (2005)	15.8193	2012	2015	
	Thorvaldsdóttir et al. (2013)	108.5528	2013	2015	
	Li et al. (2009b)	38.9412	2013	2015	
	Li et al. (2009a)	34.6253	2013	2015	
	Tautenhahn et al. (2012)	32.7960	2013	2015	
	Robinson et al. (2011)	31.1043	2013	2015	
	Pérez et al. (2007)	22.5490	2013	2015	
	Oostenveld et al. (2011)	22.4635	2013	2015	

Table 21 continued

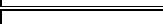
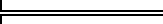
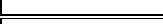
Phase	References	Burst	Begin	End	2000–2015
	Polich (2007)	20.4386	2013	2015	
	Davis et al. (2013)	19.6801	2013	2015	
	McKenna et al. (2010)	19.2065	2013	2015	
	Kanehisa et al. (2012)	17.9089	2013	2015	
	ENCODEProject Consortium(2012)	17.6698	2013	2015	
	Vaas et al. (2012)	16.9759	2013	2015	
	Francisco et al. (2012)	15.3296	2013	2015	
	Hunter (2007)	13.9839	2013	2015	
	Anders et al. (2010)	13.9645	2013	2015	

Table 22 Articles with surging citations ($Data_{exp}$)

References	Phase	Citation burst		
		Strength	Begin	End
Velculescu et al. (1995)	1	13.4274	2000	2006
Lander et al. (2001)	1	23.2995	2001	2006
Eisen et al. (1998)	1	38.1939	2002	2006
Su et al. (2002)	1	109.3584	2003	2008
Rhee et al. (2003)	1	46.8372	2004	2008
Zhang et al. (2004)	1	33.6849	2005	2009
Salomonis et al. (2007)	2	16.8535	2008	2011
Tao et al. (2008)	2	21.8191	2010	2013
Sugiyama (2007)	2	19.9878	2012	2015
Thorvaldsdóttir et al. (2013)	2	108.5528	2013	2015

view (shows a 40 kb region surrounding a gene), and genome view (shows the five Arabidopsis chromosomes). Zhang et al. (2004) develop a web-based tool called GOTree machine (GOTM), employed for analyzing and visualizing an array of genes. GOTree Machine generates a tree-like structure to navigate through the Gene Ontology Directed Acyclic Graph for given gene sets. A 2007 paper by Salomonis and colleagues presents a software suite called Gene Map Annotator and Pathway Profiler (GenMAPP). A new mode of data visualization for supporting analysis of complex data, including time-course, single nucleotide polymorphism (SNP), and splicing, is introduced. Tao et al. (2008) propose a new subspace selection algorithm which is designed for pattern classification and data visualization. Especially, they scrutinize the geometric mean for subspace selection.

Sugiyama (2007) proposes local fisher discriminant analysis (LFDA), a new linearly supervised dimensionality reduction method. Thorvaldsdóttir et al. (2013)’s study proposes Integrative GenomicsViewer, supporting not only the exploration of array-based and next-generation sequencing data but the integration of clinical and phenotypic data. It also provides various visual components such as reference genome selector, search box, zoom bar, chromosome ideogram, genomic coordinates, data tracks, and reference gene track.

In the broader context of data visualization research, we identified that there are two main streams of dominant trends. First, the collection of articles has focused on proposing integrated visual analytic systems with an aim of biomedical knowledge discovery. They employ a variety of visual components for the effective modelling and navigation of biomedical entities and mechanisms. Second, there is concerted effort to improve the capability of data visualization by finding the true dimensions of dataset. These arguments are consistent with the findings in the previous sections.

Discussion

Shared context of the domains

The quantitative investigation of the intellectual landscapes, landmark articles, bursting keywords, and citation trends in information and data visualization research reveals following shared contexts between the domains. First, computational techniques and applications, and mathematics are the main research themes in both fields. On top of that, it is obvious that information and data visualization research has scrutinized algorithmic concepts underlying the domains, particularly during the first phase as represented by a set of keywords such as “algorithms” (Info_{core}), “data mining” (Info_{core}, Data_{core}), “neural networks” (Data_{core}), and “volume rendering” (Data_{core}). Second, there are a few emerging research themes in the domains. We argue that this is because both domains are relatively new, still emerging, and developing. Third, the most influential discipline to the domains is “systems, computing, computer”. In Info_{exp} and Data_{exp}, successive research citing the core literature covers a variety of topics within the context of biomedicine. The exploration through the expanded datasets indicates that the application of information and data visualization techniques to the investigation of biomedical and biochemical data has been receiving increasing attention from the community. Introducing visual analytic approaches to tackling challenges in those studies are of significant and widely recognized concerns. It is also acknowledged that the community has aimed to integrate a variety of visual representations and data structures into a single system. Based on these findings, we contend that information visualization and data visualization have co-evolved.

Divergent themes, trends, and developments

The bursting keywords in Info_{core} and Data_{core} indicate that two domains started to digress from each other during the second phase. In this phase, Info_{core} focuses on analytic perspectives whereas “dimensionality reduction” is among the leading research theme in Data_{core}. This diverging trend is also justified by the investigation of the landmark articles since the manuscripts with citation bursts in Data_{core} have proposed novel approaches finding the small number of relevant features hidden in high-dimensional dataset. The emerging clusters in Data_{core} are consistent with this observation. On the other hand, the

exploration of the landmark literature in Info_{core} reveals that there has been concerted effort to establish a scientific ground of the domain. This finding is supported by the fact that the collection of articles with surging citations has scrutinized fundamental and cognitive aspects of information visualization such as elemental taxonomy and perceptual theory. In addition, we identified that the concept of interactivity is of fundamental importance to the field. In the broader context, successive research digresses from each other, although the deployment of visual techniques into biomedicine and biochemistry is the common research theme (see Tables 15, 16). In particular, information visualization is employed to analyze biomolecules whereas data visualization targets studying genomics. In addition, researchers influenced by information visualization seem to distinguish the application of information visualization to biomedicine from the general information research, whilst the application is regarded as an extension of data visualization in the counterpart research. In the meantime, there were few overlaps in bursting keywords, emerging clusters, and landmark articles between Info_{exp} and Data_{exp}. To sum up, information visualization and data visualization are developing with distinct concerns.

Conclusion

In the present study, we aimed to answer the motivational question, “How is information visualization similar with or different from data visualization?” Toward that end, we investigated and compared the bursting keywords, intellectual landscapes, thematic patterns, and citation trends. Based on the findings from the study, we articulate that information visualization has co-evolved with data visualization and they are developing with distinctive interests at the same time. For example, both of them have topics such as computational techniques, algorithms, and mathematics in common. In addition, both domains have special interests in applications to biomedicine and biochemistry. While investigating scientific grounds and concepts such as interactivity and cognitive aspects is more prominent in information in visualization than in data visualization, developing and improving data processing techniques has been addressed more frequently in data visualization than information visualization.

There may be alternative ways to study this topic from different perspectives, for instance, based on a content analysis, a survey of domain experts, or a demographic study of researchers. On the one hand, qualitative approaches provide valuable insights. On the other hand, quantitative approaches have advantages such as having a relatively higher reproducibility and lower cost for conducting studies at a larger scale, especially as the number of publications in information visualization and data visualization is fast-growing.

The methodology of the present study provides advantages in answering those questions as follows. First, we comprehensively collected the datasets to be analyzed. Conventional domain analysis studies tend to cover only a fraction of published articles in the literature. The combination of topic search and citation indexing in our method provides a systematic way to broaden the coverage of a knowledge domain, thus provides a much broader context to compare overlapping domains such as information visualization and data visualization. Second, we investigated the two domains from a multi-faceted point of view. Bursting keywords, document co-citation networks, emerging thematic patterns, and citation trends were identified and compared in this study. In particular, we used the concept of burstness. For instance, an article’s citation count is a significant indicator reflecting its impact on a community. However, we cannot identify the paper’s influence or density of that impact

during a particular span of time if only considering the cumulative measure. Instead, we argue that the concept of citation and keyword bursts is much more convenient in understanding the advent and decline of scientific trends. Finally, the analytic tool used in the present study enabled us to explore time-specific research trends in the domains. Similar questions may arise to differentiate how other scientific fields differ. To this end, the scientometric approach and tool employed in the present study can support comprehensive data collection and scalable data analytics. In addition, one can conduct domain analysis of his/her concern as frequently as needed without prior domain knowledge.

There are also several challenges despite the fact that there are several methodological advantages in our scientometric analysis. First, it is still likely that the topic search does not include thematically relevant literature if a set of terms used for the topic search do not appear in the titles, abstracts or keywords. Generally, the vocabulary mismatch problem presents a challenge for keyword-based search (Deerwester et al. 1990). Second, the Web of Science as our source of data may have underrepresented publications in conference proceedings in these areas. Additional sources such as Scopus are recommended for future refinements of this type of analysis. Unfortunately, at the time of the study, we did not have access to the counterpart data from Scopus. In generating the intellectual landscapes, we selected 100 and 200 highly cited references within the core and the expanded contexts respectively. This approach can be strengthened if we choose cited articles based on more refined criteria such as h-index. It may be worth a separated study of the theoretical implications of using h-based or top-100 or other conceivable selection criteria. We also plan to apply the scientometric approach to much comprehensive papers that cover a various type of publication materials.

Acknowledgments This work is in part supported by the NSF I/UCRC Center for Visual Decision and Informatics (NSF IIP-1160960).

References

- Aigner, W., Miksch, S., Schumann, H., & Tominski, C. (2011). *Visualization of time-oriented data*. London: Springer.
- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11, R106.
- Andronis, C., Sharma, A., Virvilis, V., Deftereos, S., & Persidis, A. (2011). Literature mining, ontologies and information visualization for drug repurposing. *Briefings in Bioinformatics*, 12(4), 357–368.
- Asuncion, A., & Newman, D. (2007). UCI machine learning repository. <http://www.ics.uci.edu/~mllearn/MLRepository.html>. Accessed 30 June 2015.
- Bederson, B., & Shneiderman, B. (2003). *The craft of information visualization: Readings and reflections*. San Francisco, CA: Morgan Kaufmann.
- Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 1373–1396.
- Berkovitch, F., Nicolet, Y., Wan, J. T., Jarrett, J. T., & Drennan, C. L. (2004). Crystal structure of biotin synthase: An *s*-adenosylmethionine-dependent radical enzyme. *Science*, 303(5654), 76–79.
- Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., et al. (2004). Potential of metabolomics as a functional genomics tool. *Trends in Plant Science*, 9(9), 418–425.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993–1022.
- Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, 72(1–2), 248–254.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., et al. (2001). Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data. *Nature Genetics*, 29, 365–371.

- Bui, B. T. S., Florentin, D., Fournier, F., Ploux, O., Méjean, A., & Marquet, A. (1998). Biotin synthase mechanism: On the origin of sulphur. *FEBS Letters*, 440(1–2), 226–230.
- Bui, B. T. S., Florentin, D., Marquet, A., Benda, R., & Trautwein, A. X. (1999). Mössbauer studies of *Escherichia coli* biotin synthase: Evidence for reversible interconversion between [2Fe–2S](2+) and [4Fe–4S](2+) clusters. *FEBS Letters*, 459(3), 411–414.
- Card, S. K., Mackinlay, J. D., & Schneiderman, B. (1999). *Readings in information visualization: Using vision to think*. San Francisco, CA: Morgan Kaufmann.
- Cheek, J., & Broderick, J. B. (2001). Adenosylmethionine-dependent iron–sulfur enzymes: Versatile clusters in a radical new role. *Journal of Biological Inorganic Chemistry*, 6(3), 209–226.
- Chen, C. (1999). *Information visualisation and virtual environments*. London: Springer.
- Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359–377.
- Chen, C. (2010). Information visualization. *Wiley Interdisciplinary Review: Computational Statistics*, 2(4), 387–403.
- Chen, C., Dubin, R., & Kim, M. C. (2014a). Emerging trends and new developments in regenerative medicine: A scientometric update (2000–2014). *Expert Opinion on Biological Therapy*, 14(9), 1295–1317.
- Chen, C., Dubin, R., & Kim, M. C. (2014b). Orphan drugs and rare diseases: A scientometric review (2000–2014). *Expert Opinion on Orphan Drugs*, 2(7), 709–724.
- Chen, C. H., Härdle, W. K., & Unwin, A. (Eds.). (2008). *Handbook of data visualization (Springer handbooks of computational statistics)*. Santa Clara, CA: Springer.
- Chen, C., Hu, Z., Liu, S., & Tseng, H. (2012). Emerging trends in regenerative medicine: A scientometric analysis in CiteSpace. *Expert Opinions on Biological Therapy*, 12(5), 593–608.
- Chen, C., Ibekwe-SanJuan, F., & Hou, J. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology*, 61(7), 1386–1409.
- Chen, C., & Leydesdorff, L. (2014). Patterns of connections and movements in dual-map overlays: A new method of publication portfolio analysis. *Journal of the American Society for Information Science and Technology*, 65(2), 334–351.
- Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420, 520–562.
- Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 521–554.
- Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402.
- Cosper, N. J., Booker, S. J., Ruzicka, F., Frey, P. A., & Scott, R. A. (2000). Direct Fe–S cluster involvement in generation of a radical in lysine 2,3-aminomutase. *Biochemistry*, 39(51), 15668–15673.
- Davis, A. P., Murphy, C. G., Johnson, R., Lay, J. M., Lennon-Hopkins, K., Saraceni-Richards, C., et al. (2013). The comparative toxicogenomics database: Update 2013. *Nucleic Acids Research*, 41(D1), D1104–D1114.
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.
- DellValle, M. (2010). Electronic tongues employing electrochemical sensors. *Electroanalysis*, 22(14), 1539–1555.
- Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1), 9–21.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. London: Wiley-Interscience.
- Duin, E. C., Lafferty, M. E., Crouse, B. R., Allen, R. M., Sanyal, I., Flint, D. H., et al. (1997). [2Fe–2S] to [4Fe–4S] cluster conversion in *Escherichia coli* biotin synthase. *Biochemistry*, 36(39), 11811–11820.
- Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25), 14863–14868.
- Ellis, G., & Dix, A. (2007). A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1216–1223.
- Elmqvist, N., Moere, A. V., Jetter, H. C., Cernea, D., Reiterer, H., & Jankun-Kelly, T. J. (2011). Fluid interaction for information visualization. *Information Visualization*, 10(4), 327–340.
- Fabrikant, S. I., & Buttenfield, B. P. (2001). Formalizing semantic spaces for information access. *Annals of the Association of American Geographers*, 91(2), 263–280.

- Faisal, S., Blandford, A., & Potts, H. W. (2013). Making sense of personal health information: Challenges for information visualization. *Health Informatics Journal*, 19(3), 198–217.
- Fekete, J. (2004). The infovis toolkit. In *INFOVIS 2004 IEEE symposium on information visualization* (pp. 167–174).
- Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171.
- Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., & Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nature Biotechnology*, 18, 1157–1161.
- Fontecave, M., Mulliez, E., & Ollagnier-de-Choudens, S. (2001). Adenosylmethionine as a source of 5'-deoxyadenosyl radicals. *Current Opinion in Chemical Biology*, 5(5), 506–512.
- Francisco, A. P., Vaz, C., Monteiro, P. T., Melo-Cristino, J., Ramirez, M., & Carriço, J. A. (2012). PHYLOViZ: Phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*, 13, 87.
- Frey, P. A. (2001). Radical mechanisms of enzymatic catalysis. *Annual Review of Biochemistry*, 70, 121–148.
- Frey, P. A., Hegeman, A. D., & Ruzicka, F. J. (2008). The radical SAM superfamily. *Critical Reviews in Biochemistry and Molecular Biology*, 43(1), 63–88.
- Frey, P. A., & Magnusson, O. T. (2003). *s*-Adenosylmethionine: A wolf in sheep's clothing, or a rich man's adenosylcobalamin? *Chemical Reviews*, 103(6), 2129–2148.
- Friendly, M. (2008). *Milestones in the history of thematic cartography, statistical graphics, and data visualization*. http://www.math.usu.edu/~symanzik/teaching/2009_stat6560/Downloads/Friendly_milestone.pdf. Accessed 7 February 2015.
- Gaviria, A. R. (2008). When is information visualization art? Determining the critical criteria. *Leonardo*, 41(5), 479–482.
- Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., et al. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7(3 Suppl), S56–S68.
- Grove, T. L., Benner, J. S., Radle, M. I., Ahlum, J. H., Landgraf, B. J., Krebs, C., et al. (2011). A radically different mechanism for *s*-adenosylmethionine—Dependent methyltransferases. *Science*, 332(6029), 604–607.
- Guianvarc'h, D., Florentin, D., Bui, B. T. S., Nunzi, F., & Marquet, A. (1997). Biotin synthase, a new member of the family of enzymes which uses *s*-adenosylmethionine as a source of deoxyadenosyl radical. *Biochemical and Biophysical Research Communications*, 236(2), 402–406.
- Henshaw, T. F., Cheek, J., & Broderick, J. B. (2000). The [4Fe–4S]¹⁺ cluster of pyruvate formate-lyase activating enzyme generates the glycy radical on pyruvate formate-lyase: EPR-detected single turnover. *Journal of the American Chemistry Society*, 122(34), 8331–8332.
- Herman, I., Melancon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1), 24–43.
- Hewitson, K. S., Ollagnier-de-Choudens, S., Sanakis, Y., Shaw, N. M., Baldwin, J. E., Münck, E., et al. (2001). The iron–sulfur center of biotin synthase: Site-directed mutants. *Journal of Biological Inorganic Chemistry*, 7(1–2), 83–93.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, 9(3), 90–95.
- Inselberg, A., & Dimsdale, B. (1990). Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *Visualization '90 proceedings of the first IEEE conference on visualization* (pp. 361–378).
- Isenberg, P. (2007). Interactive tree comparison for co-located collaborative information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1232–1239.
- Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., & Moller, T. (2013). A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2818–2827.
- Jarrett, J. T. (2003). The generation of 5'-deoxyadenosyl radicals by adenosylmethionine-dependent radical enzymes. *Current Opinion in Chemical Biology*, 7(2), 174–182.
- Jenssen, T., Lægreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28, 21–28.
- Jolliffe, I. T. (2002). *Principal component analysis*. Secaucus, NJ: Springer.
- Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1), D109–D114.
- Kaul, S., Koo, H., Jenkins, J., et al. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814), 796–815.
- Kersten-Oertel, M., Jannin, P., & Collins, D. L. (2013). The state of the art of visualization in mixed reality image guided surgery. *Computerized Medical Imaging and Graphics*, 37(2), 98–112.

- Kim, M. C., & Chen, C. (2015). A scientometric review of emerging trends and new developments in recommendation systems. *Scientometrics*, 104(1), 239–263.
- Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *KDD '02 Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 91–101).
- Kohonen, T. (1997). *Self-organizing maps*. Secaucus, NJ: Springer.
- Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9), 1639–1645.
- Külzer, R., Pils, T., Kappl, R., Hüttermann, J., & Knappe, J. (1998). Reconstitution and characterization of the polynuclear iron–sulfur cluster in pyruvate formate-lyase-activating enzyme. *The Journal of biological chemistry*, 273(9), 4897–4903.
- Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2011). Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9), 1520–1536.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860–921.
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10, R25.
- Layer, G., Moser, J., Heinz, D. W., Jahn, D., & Schubert, W. (2003). Crystal structure of coproporphyrinogen III oxidase reveals cofactor geometry of radical SAM enzymes. *The EMBO Journal*, 22, 6214–6224.
- Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Berlin: Springer.
- Leung, Y. K., & Apperley, M. D. (1994). A review and taxonomy of distortion-oriented presentation techniques. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 1(2), 126–160.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14), 1754–1760.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079.
- Lieder, K. W., Booker, S., Ruzicka, F. J., Beinert, H., Reed, G. H., & Frey, P. A. (1998). *s*-Adenosylmethionine-dependent reduction of lysine 2,3-aminomutase and observation of the catalytically functional iron–sulfur centers by electron paramagnetic resonance. *Biochemistry*, 37(8), 2578–2585.
- Liu, S., Cui, W., Wu, Y., & Liu, M. (2014). A survey on information visualization: Recent advances and challenges. *The Visual Computer*, 30(12), 1373–1393.
- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., et al. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13), 1675–1680.
- Luck, S. J. (2005). *An introduction to the event-related potential technique*. Cambridge, MA: The MIT Press.
- Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1), 177–190.
- Marquet, A. (2001). Enzymology of carbon–sulfur bond formation. *Current Opinion in Chemical Biology*, 5(5), 541–549.
- McKenna, A., Hanna, B., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20, 1297–1303.
- Mehlan, H., Schmidt, F., Weiss, S., Schüler, J., Fuchs, S., Riedel, K., et al. (2013). Data visualization in environmental proteomics. *Proteomics*, 13(18–19), 2805–2821.
- Miller, J. R., Busby, R. W., Jordan, S. W., Cheek, J., Henshaw, T. F., Ashley, G. W., et al. (2000). *Escherichia coli* LipA is a lipoyl synthase: In vitro biosynthesis of lipoylated pyruvate dehydrogenase complex from octanoyl–acyl carrier protein. *Biochemistry*, 39(49), 15166–15178.
- Moreland, J. L., Gramada, A., Buzko, O. V., Zhang, Q., & Bourne, P. E. (2005). The molecular biology toolkit (MBT): A modular platform for developing molecular visualization applications. *BMC Bioinformatics*, 6, 21.
- Morse, E., Lewis, M., & Olsen, K. A. (2000). Evaluating visualizations: Using a taxonomic guide. *International Journal of Human-Computer Studies*, 53(5), 637–662.
- Munzner, T. (2009). A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6), 921–928.
- Ollagnier, S., Mulliez, E., Schmidt, P. P., Eliasson, R., Gaillard, J., Deronzier, C., et al. (1997). Activation of the anaerobic ribonucleotide reductase from *Escherichia coli*. *The Journal of biological Chemistry*, 272(39), 24216–24223.

- Ollagnier-De-Choudens, S., Mulliez, E., Hewitson, K. S., & Fontecave, M. (2002a). Biotin synthase is a pyridoxal phosphate-dependent cysteine desulfurase. *Biochemistry*, 41(29), 9145–9152.
- Ollagnier-de-Choudens, S., Sanakis, Y., Hewitson, K. S., Roach, P., Baldwin, J. E., Münck, E., et al. (2000). Iron–sulfur center of biotin synthase and lipoate synthase. *Biochemistry*, 39(14), 4165–4173.
- Ollagnier-de-Choudens, S., Sanakis, Y., Hewitson, K. S., Roach, P., Münck, E., & Fontecave, M. (2002b). Reductive cleavage of *s*-adenosylmethionine by biotin synthase from *Escherichia coli*. *The Journal of biological chemistry*, 277(16), 13449–13454.
- Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011, Article ID 156869.
- Orford, S., Harris, R., & Dorling, D. (1999). Geography: Information visualization in the social sciences. *Social Science Computer Review*, 17(3), 289–304.
- Penterman, R., Klink, S. I., de Koning, H., Nisato, G., & Broer, D. J. (2002). Single-substrate liquid–crystal displays by photo-enforced stratification. *Nature*, 417, 55–58.
- Pérez, F., & Granger, B. E. (2007). IPython: A system for interactive scientific computing. *Computing in Science and Engineering*, 9(3), 21–29.
- Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, 118(10), 2128–2148.
- Raamsdonk, L. M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., et al. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19, 45–50.
- Rabow, A. A., Shoemaker, R. H., Sausville, E. A., & Covell, D. G. (2002). Mining the national cancer institute's tumor-screening database: Identification of compounds with similar cellular activities. *Journal of Medical Chemistry*, 45(4), 818–840.
- Rauber, A., Merkl, D., & Dittenbach, M. (2002). The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13(6), 1331–1341.
- Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., et al. (2003). The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Research*, 31(1), 224–228.
- Ripley, B. D. (1996). *Pattern recognition and neural networks*. Cambridge, MA: Cambridge University Press.
- Riul, A., Jr, Dantas, C. A. R., Miyazaki, C. M., & Oliveira, O. N., Jr. (2010). Recent advances in electronic tongues. *Analyst*, 135, 2481–2495.
- Robertson, G. G., Card, S. K., & Mackinlay, J. D. (1993). Information visualization using 3D interactive animation. *Communications of the ACM*, 36(4), 57–71.
- Robertson, G. G., Mackinlay, J. D., & Card, S. K. (1991). Cone trees: Animated 3D visualizations of hierarchical information. In *CHI '91 Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 189–194).
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29, 24–26.
- Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L., et al. (2001). Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *The Plant Cell*, 13(1), 11–29.
- Roessner, U., Wagner, C., Kopka, J., Trethewey, R. N., & Willmitzer, L. (2000). Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *The Plant Journal*, 23(1), 131–142.
- Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P., Morris, M., et al. (2004). GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics*, 37(1), 43–53.
- Salomonis, N., Hanspers, K., Zamboni, A. C., Vranizan, K., Lawlor, S. C., Dahlquist, K. D., et al. (2007). GenMAPP 2: New features and resources for pathway analysis. *BMC Bioinformatics*, 8, 217.
- Sanyal, I., Cohen, G., & Flint, D. H. (1994). Biotin synthase: Purification, characterization as a [2Fe–2S] cluster protein, and in vitro activity of the *Escherichia coli* bioB gene product. *Biochemistry*, 33(12), 3625–3631.
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235), 467–470.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., et al. (2000). A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 23(3), 236–244.
- Schneiderman, B. (1998). *Designing the user interface: Strategies for effective human–computer interaction*. Boston, MA: Addison-Wesley.

- Segel, E. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 1139–1148.
- Shannon, P., Markiel, A., Ozier, W., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13, 2498–2504.
- Sofia, H. D., Chen, G., Hetzler, B. G., Reyes-Spindola, J. F., & Miller, N. E. (2001). Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information visualization methods. *Nucleic Acids Research*, 29(5), 1097–1106.
- Spence, B. (2000). *Information visualization*. New York, NY: ACM Press.
- Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2), 118–132.
- Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., et al. (2002). Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 99(7), 4465–4470.
- Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(16), 6062–6067.
- Sugiyama, M. (2007). Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8, 1027–1061.
- Sumner, L. W., Mendes, P., & Dixon, R. A. (2003). Plant metabolomics: Large-scale phytochemistry in the functional genomics era. *Phytochemistry*, 62(6), 817–836.
- Swayne, D. F., Cook, D., & Buja, A. (1998). XGobi: Interactive dynamic data visualization in the X window system. *Journal of Computational and Graphical Statistics*, 7(1), 113–130.
- Tao, D., Li, X., Wu, X., & Maybank, S. J. (2008). Geometric mean for subspace selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2), 260–274.
- Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS online: A web-based platform to process untargeted metabolomic data. *Analytical Chemistry*, 84(11), 5035–5039.
- Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A Global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500), 2319–2323.
- The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74.
- Thomas, J. J., & Cook, K. A. (2005). *Illuminating the path: The R&D agenda for visual analytics*. USA: National Visualization and Analytics Center.
- Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, 14(2), 178–192.
- Torgerson, W. S. (1952). Multidimensional scaling: I. *Theory and method*. *Psychometrika*, 17, 401–419.
- Törönen, P., Kolehmainen, M., & Castrén, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451(2), 142–146.
- Tory, M. (2005). Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications*, 25(5), 8–11.
- Tufte, E. (1983). *The visual display of quantitative information*. Cheshire, Connecticut: Graphics Press.
- Tufte, E. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.
- Tukey, J. (1977). *Exploratory data analysis*. Boston, MA: Addison-Wesley.
- Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9), 5116–5121.
- Ugulava, N. B., Gibney, B. R., & Jarrett, J. T. (2001a). Biotin synthase contains two distinct iron–sulfur cluster binding sites: Chemical and spectroelectrochemical analysis of iron–sulfur cluster interconversions. *Biochemistry*, 40(28), 8343–8351.
- Ugulava, N. B., Sacanell, C. J., & Jarrett, J. T. (2001b). Spectroscopic changes during a single turnover of biotin synthase: Destruction of a [2Fe–2S] cluster accompanies sulfur insertion. *Biochemistry*, 40(28), 8352–8358.
- Ugulava, N. B., Surerus, K. K., & Jarrett, J. T. (2002). Evidence from Mössbauer spectroscopy for distinct [2Fe–2S]²⁺ and [4Fe–4S]²⁺ cluster binding sites in biotin synthase from *Escherichia coli*. *Journal of the American Chemistry Society*, 124(31), 9050–9051.
- Vaas, L. A. I., Sikorski, J., Michael, V., Göker, M., & Klenk, H. (2012). Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS ONE*, 7(4), e34846.

- van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9, 2579–2605.
- Velculescu, V. E., Zhang, L., Vogelstein, B., & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science*, 270(5235), 484–487.
- Venna, J., Peltonen, J., Nybo, K., Aidos, H., & Kaski, S. (2010). Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *The Journal of Machine Learning Research*, 11, 451–490.
- Venter, J. C., Adams, M. D., & Myers, E. W. (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351.
- Vey, J. L., & Drennan, C. L. (2011). Structural insights into radical generation by the radical SAM superfamily. *Chemical Reviews*, 111(4), 2487–2506.
- Walsby, C. J., Hong, W., Broderick, W. E., Cheek, J., Ortillo, D., Broderick, J. B., et al. (2002a). Electron–nuclear double resonance spectroscopic evidence that *s*-adenosylmethionine binds in contact with the catalytically active $[4\text{Fe-4S}]^+$ cluster of pyruvate formate-lyase activating enzyme. *Journal of the American Chemistry Society*, 124(12), 3143–3151.
- Walsby, C. J., Ortillo, D., Broderick, W. E., Broderick, J. B., & Hoffman, B. M. (2002b). An anchoring role for FeS clusters: Chelation of the amino acid moiety of *s*-adenosylmethionine to the unique iron site of the $[4\text{Fe-4S}]$ cluster of pyruvate formate-lyase activating enzyme. *Journal of the American Chemistry Society*, 124(38), 11270–11271.
- Ware, C. (2000). *Information visualization: Perception for design*. San Francisco, CA: Morgan Kaufmann.
- Ware, C. (2004). *Information visualization: Perception for design*. San Francisco, CA: Morgan Kaufmann.
- Wicks, P., Massagli, M., Frost, J., Brownstein, C., Okun, S., Vaughan, T., et al. (2010). Sharing health data for better outcomes on PatientsLikeMe. *Journal of Medical Internet Research*, 12(2), e19.
- Yan, F., LaMarre, J. M., Röhrich, R., Wiesner, J., Jomaa, H., Mankin, A. S., et al. (2010). RlmN and Cfr are radical SAM enzymes involved in methylation of ribosomal RNA. *Journal of the American Chemistry Society*, 132(11), 3953–3964.
- Yi, J. S., Kang, Y. A., Stasko, J. T., & Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1224–1231.
- Yin, H. (2008). On multidimensional scaling and the embedding of self-organising maps. *Neural Networks*, 21(3), 160–169.
- Yin, H., & Huang, W. (2010). Adaptive nonlinear manifolds and their applications to pattern recognition. *Information Sciences*, 180(14), 2649–2662.
- Zhang, B., Schmoyer, D., Kirov, S., & Snoddy, J. (2004). GOTree machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, 5, 16.
- Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., & Gruissem, W. (2004). GENEVESTIGATOR: Arabidopsis microarray database and analysis toolbox. *Plant Physiology*, 136(1), 2621–2632.