

Position Paper

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Data, Information and Knowledge in Visualization

Min Chen
Swansea University

David Ebert
Purdue University

Hans Hagen
Technical University
of Kaiserslautern

Robert Laramee
Swansea University

Robert van Liere
CWI, Amsterdam

Kwan-Liu Ma
University of
California, Davis

William Ribarsky
University of North
Carolina, Charlotte

Gerik Scheuermann
University of Leipzig

Deborah Silver
Rutgers University

In visualization, **data**, **information** and **knowledge** are three terms used extensively, often in an interrelated context. In many cases, they are used to indicate different levels of abstraction, understanding or truthfulness. For example, ‘visualization is concerned with exploring *data* and *information* [5]; ‘the primary objective in *data* visualization is to gain insight into an *information* space’ [6]; and ‘*information* visualization’ is for ‘*data* mining and *knowledge* discovery’ [4]. In other cases, these three terms are used to indicate data types, for instances, as adnominals in noun phases, such as *data* visualization, *information* visualization and *knowledge* visualization. These examples suggest that *data*, *information* and *knowledge* could be both the input and output of a visualization process, raising questions about the exact role of *data*, *information* and *knowledge* in visualization.

There are many competing definitions of *data*, *information* and *knowledge*, in different aspects of computer science and engineering and in other disciplines such as psychology, management sciences, epistemology (theory of knowledge). The use of the three terms is not consistent, and often conflicting. For instance, in computing, *data* and *information* are often used in an interchangeable manner (e.g., data processing and information processing; data management and information management). From a system perspective, *data* is referred to as bits and bytes stored on or communicated via, a digital medium. Thus, any computerized representations, including *knowledge* representations are types of *data*. On the other hand, from the perspective of knowledge-based systems, *data* is a simpler form of *knowledge*.

In epistemology, ‘ALL AGREE THAT KNOWLEDGE is valuable, but the agreement about knowledge tends to end there. Philosophers disagree about what knowledge is, about how you get it, and even about whether there is any to be gotten.’ Keith Lehrer [5]

Several attempts were made to clarify taxonomically the terminology used in the visualization community (e.g., [3,8,10]). However, the terms of *data*, *information* and *knowledge* remain ambiguous. This article is not another attempt to offer a different taxonomy for visualization. Instead, we present a clarification that differentiates these three terms from the perspective of visualization processes. Furthermore, we examine the current and future role of *information* and *knowledge* in the development of the visualization technology.

Definitions of data, information and knowledge

Since we can read *data*, grasp *information* and acquire *knowledge*, we must first differentiate these three terms in the **perceptual and cognitive space**. Because we can also store *data*, *information* and *knowledge* in the computer, we thereby must also differentiate them in the **computational space**.

Table 1. Ackoff’s definitions of *data*, *information* and *knowledge* in perceptual and cognitive space [1].

Category	Definition
<i>data</i>	symbols
<i>information</i>	data that are processed to be useful, providing answers to ‘who’, ‘what’, ‘where’, and ‘when’ questions
<i>knowledge</i>	application of data and information, providing answers to ‘how’ questions

Table 2. Our definitions of *data*, *information* and *knowledge* in computational space.

Category	Definition
<i>data</i>	computerized representations of models and attributes of real or simulated entities
■ <i>information</i>	<i>data</i> that represents the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings
■ <i>knowledge</i>	<i>data</i> that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or the transcripts of some knowledge acquired by human beings

Perceptual and Cognitive Space

The *Data-Information-Knowledge-Wisdom* (DIKW) hierarchy [1] is a popular model for classifying the human’s understanding in the perceptual and cognitive space. The origin of this hierarchy can be traced to the poet T.S. Eliot [3]. Table 1 shows the definitions of *data*, *information* and *knowledge* given by Ackoff [1].

Let \mathbb{P} be the set of all possible explicit and implicit human memory. The former encompasses the memory of events, facts and concepts, and the understanding of their meanings, context and associations. The latter encompasses all non-conscious forms of memory, such as emotional responses, skills, habits and so on [8]. We can thus focus on three subsets of memory, $\mathbb{P}_{\text{data}} \subset \mathbb{P}$, $\mathbb{P}_{\text{info}} \subset \mathbb{P}$, and $\mathbb{P}_{\text{know}} \subset \mathbb{P}$, where \mathbb{P}_{data} , \mathbb{P}_{info} , and \mathbb{P}_{know} are the sets of all possible explicit and implicit memory about *data*, *information*, and *knowledge*, respectively.

Despite the lack of an agreeable set of the definitions of *data*, *information* and *knowledge*, there is a general consensus that *data* is not *information*, and *information* is not *knowledge*. Without diverting from the scope of this article, here we simply assume that \mathbb{P}_{data} , \mathbb{P}_{info} , \mathbb{P}_{know} are not mutually disjoint, and none of them is a subset of another. Without losing generality, we can generalize \mathbb{P}_{know} to include also *wisdom*, and any other high-level of understanding, in the context of DIKW hierarchy.

Computational Space

Let \mathbb{C} be the set of all possible representations in computer memory. Similarly, we may consider three subsets of representations, \mathbb{C}_{data} , \mathbb{C}_{info} , and \mathbb{C}_{know} . However, **data** is an overloaded term in computing. For example, it is common to treat programs as a special class of *data*. In many cases, it is not possible to distinguish programs from other *data*. Applying the same analogy, a computer

Resolving Ambiguity Using the Set Notations

We can resolve the ambiguity in various statements that consist of the terms of *data*, *information* and *knowledge* by tagging such terms using the set notations, P , C and their subsets.

American National Standards Institute, *Directory for Information Systems*, X3.172, 1990:

'Data (C_{data}): a representation of facts, concepts, or instructions in a formalized manner suitable for communication, interpretation, or processing by human beings or by automatic means.'

'Information (P_{info} or C_{info}): the meaning that is currently assigned' (by human beings or computers) 'to data (C_{data}) by means of the conventions applied to those data (C_{data}).'

J. Foley and B. Ribarsky, "Next-generation data visualization tools", in L. Rosenblum *et al.* (eds.) *Scientific Visualization: Advances and Challenges*, Academic Press, 1994:

'A useful definition of visualization might be the binding (or mapping) of data (C_{data}) to representations (C_{visual} , $C_{auditory}$, $C_{tactile}$, etc.) that can be perceived. The types of bindings could be visual, auditory, tactile, etc., or a combination of these.'

R. M. Friedhoff and T. Kiley, "The Eye of the Beholder", *Computer Graphics World*, 13(8):46-59, 1990:

'If researchers try to read the data (C_{data}), usually presented as vast numeric matrices, they will take in the information (P_{info}) at snail's pace. If the information (C_{info}) is rendered graphically, however, they can assimilate it at a much faster rate.'

B. H. McCormick, T. A. DeFanti and M. D. Brown (eds.), "Visualization in Scientific Computing", *Computer Graphics*, 21(6), 1987:

Visualization 'transforms the symbolic (C_{data}) into the geometric (C_{visual}), enabling researchers to observe their simulations and computations'.

S. Card, J. Mackinlay and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999:

Information (C_{info}) visualization is 'the use of computer-supported, interactive, visual representations (C_{visual}) of abstract data (C_{info}) to amplify cognition'.

W. Stallings, *Data and Computer Communications*, (4th ed.), Macmillan, 1994:

'Information (P_{info} or C_{info}) is born when data (C_{data}) are interpreted' (by human beings or computers).

M. J. Usher, *Information Theory for Information Technologists*, Macmillan, 1984:

'Information (P_{info} and C_{info}) has both qualitative and quantitative aspects.' 'The amount of information (P_{info} and C_{info}) conveyed in an event depends on the probability of the event.'

R. A. Frost, *Introduction to Knowledge Based Systems*, Collins, 1986:

'Knowledge (C_{know}) is the symbolic representation of aspects of some named universe of discourse.' 'We define data (C_{facts} or $C_{rawdata}$ but not C_{data} since $C_{know} \subset C_{data}$) as the symbolic representation of simple aspects of some named universe of discourse.' 'The amount of information (P_{info}) obtained by the receiver of a message is related to the amount by which that message reduces receiver's uncertainty about some aspect of the universe of discourse (*Shannon*).'

E. Turban, *Decision Support and Expert Systems*, Prentice-Hall, 1995:

'Knowledge (P_{know}): understanding, awareness, or familiarity acquired through education or experience. Anything that has been learned, perceived, discovered, inferred, or understood. The ability to use information (P_{info} and/or C_{info}).'

'Knowledge base: the assembly of all the information (C_{info}) and knowledge (C_{know}) of a specific field of interest.'

representation of a piece of *information* or *knowledge* is just a particular form of *data*. A computer representation of visualization is also a form of *visual data*.

We hence propose to use the definitions in Table 2 for the following discussions. With such definitions, we have $C_{data} = C$, $C_{info} \subset C_{data}$, and $C_{know} \subset C_{data}$. The definitions in Table 2 can easily be extended to include categories of *raw data* ($C_{rawdata}$), *volume data* (C_{volume}), *flow data* (C_{flow}), *software* ($C_{software}$), *videos* (C_{video}), *mathematical models* ($C_{mathmodel}$), *visual data* (C_{visual}), and so forth. This also makes sense when using the category names as the adnominals in noun phrases, such as *volume visualization* and *software visualization*.

Figure 1 shows a typical visualization process, where instances of *data*, *information* and *knowledge* in both computational space and perceptual and cognitive space are illustrated. Hence the purpose of visualization can be rationalized by the difficulties for humans to acquire a sufficient amount of *information* ($P_{info} \subset P_{info}$) or *knowledge* ($P_{know} \subset P_{know}$) directly from a dataset ($C_{data} \subset C_{data}$). The process of visualization is a function that maps from C_{data} to the set of all imagery data, C_{image} . It transforms a dataset C_{data} to a visual representation C_{image} , which facilitates a more efficient and effective cognitive process for acquiring P_{info} and P_{know} .

A visualization process is a search process

Given a dataset C_{data} , a user first makes some decisions about visualization tools to be used for exploring the dataset. The user then experiments with different controls, such as styles, layout, viewing position, color maps, transfer functions, etc. until a collection of satisfactory visualization results, C_{image} , is obtained. Depending on the visualization tasks, satisfaction can be in many forms. For example, the user may have obtained sufficient information or knowledge about the dataset, or may have obtained the most appropriate illustration about the data to assist the knowledge acquisition process of others.

Such a visualization process is fundamentally the same as a typical *search* process, except that it is usually much more complex than trying out a few keywords with a search engine. In visualization, the tools for the 'search' tasks are usually application-specific (e.g., network, flow, volume visualization). The parameter space for the 'search' is normally huge (e.g., exploring many viewing positions or trying out many different transfer functions). The user interaction for the 'search' sometimes can be very slow, especially in handling very large datasets. This is depicted in Figure 1 by a large interaction box that connects from the user to the control parameters, C_{ctrl} , which are also data.

In fact, over the past two decades, much of the emphasis has been placed on improving the speed of visualization tools, so the user can carry out the interactive 'search' faster, can explore bigger parameter space, and hopefully find satisfactory results quicker.

However, with the growing amount of data and increasing availability of different visualization techniques, the 'search' space for a visualization process is also getting larger and larger. Like the internet search problem, *interactive visualization* alone is no longer adequate.

Information-assisted visualization

In recent years, an assortment of techniques were introduced for visualizing complex features in *data* by relying on *information* abstracted from the *data*. Note that here we consider C_{info} in the computational space as well as P_{info} in the perceptual and cognitive space. Figure 2 illustrates an **information-assisted visualization** process.

There are techniques that make use of information captured in the visualization process to improve the efficiency and effectiveness of visualization. Examples of such information are given in Table 3.

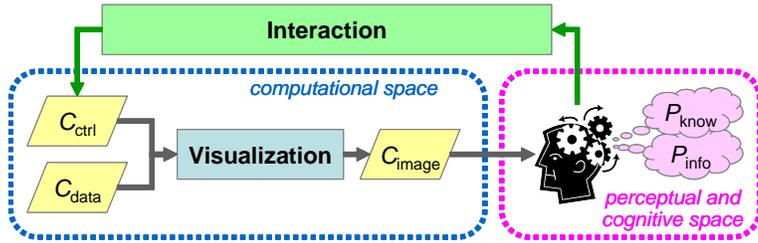


Figure 1. A typical visualization process.

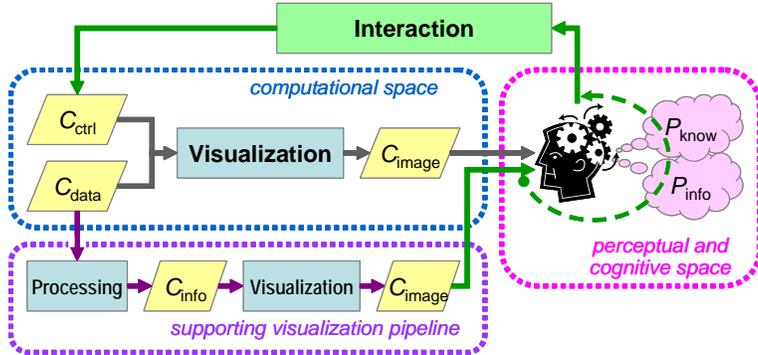


Figure 2. Information-assisted visualization.

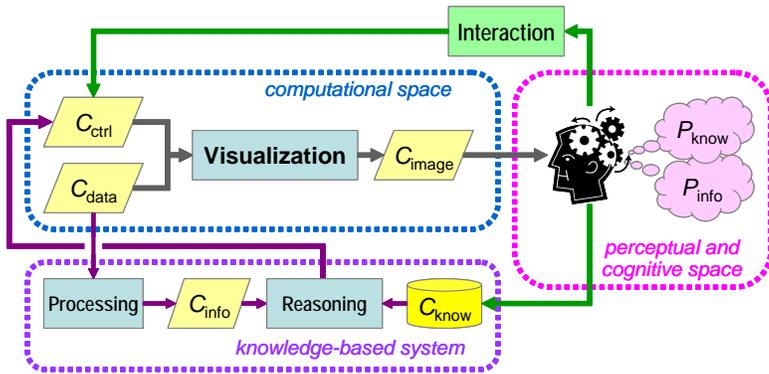


Figure 3. Knowledge-assisted visualization with acquired knowledge representations.

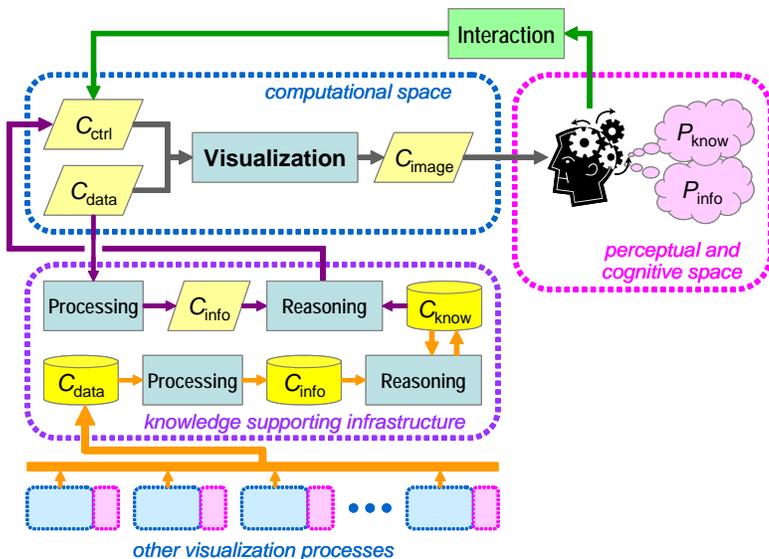


Figure 4. Knowledge-assisted visualization with simulated cognitive processing.

Table 3. Examples of information used in visualization.

information categories	examples
information about the input dataset	
■ abstract geometric and temporal characteristics	skeletons, features, events
■ topological properties	contour tree for volume data, vector field topology, tracking graph for time-varying data
■ statistical indicators and information measurements	histogram, correlation, importance, certainty, entropy, mutual information, local statistical complexity
information about the results	color histogram, level of cluttering
information about the process	interaction patterns, provenance
Information about users' perception	response time, accuracy

In information-assisted visualization, the user is provided with a second visualization pipeline (see Figure 2), which typically displays the information about the input dataset, but can also present attributes of the visualization process, the properties of the results, or characteristics of the user's perceptual behaviors. The user uses such information to reduce the 'search' space for optimal control parameters, hence making the interaction much more cost-effective.

Such techniques provide an intrinsic interface between the scientific visualization and information visualization communities. With the increasing size and complexity of data, the use of information to aid visualization will inevitably become a necessity rather than an option.

Knowledge-assisted visualization

In a visualization process, the *knowledge* of the user is an indispensable part of visualization. For instance, the user may assign specific colors to different objects in visualization according to certain domain knowledge. The user may choose certain viewing positions because the visualization results can reveal more meaningful information or a more problematic scenario that requires further investigation.

Meanwhile, the lack of certain *knowledge* by the user is often a major obstacle in deploying visualization techniques. The user may not have received adequate training about how to specify transfer functions. The user may not have sufficient time or navigation skills to explore all possible viewing positions.

Both scenarios suggest the need for **knowledge-assisted visualization**. The objectives of knowledge-assisted visualization include sharing domain knowledge among different users, and reducing the burden upon users to acquire knowledge about complex visualization techniques. It also enables the visualization community to learn and model the best practice, and to develop powerful visualization infrastructures evolutionarily.

In fact, some general or domain *knowledge* has already been incorporated into various visualization systems, intentionally or unintentionally. For example, a default transfer function in a volume visualization system may capture the domain knowledge about a specific modality. If one could collect a large repository of such knowledge, it would be possible for a visualization system to choose an appropriate transfer function according to the information about the input datasets. Figure 3 shows a visualization pipeline supported by a knowledge base (C_{know}), which stores the knowledge representations captured from expert users. *Rule-based reasoning* can be utilized to establish an appropriate set, or several optional sets, of control parameters, which can significantly reduce the 'search' space, especially for inexperienced users. The system component for *reasoning* is commonly referred to as an

inference engine in knowledge based systems (or expert systems).

The shortcomings of such a system include the difficulties in specifying comprehensively what knowledge to capture and the inconvenience in collecting knowledge from experts. This constrains the deployment of such a system to specific application domains.

An alternative approach is to establish a visualization infrastructure, where data about visualization processes are systematically collected, processed and analyzed. Using *case-based reasoning*, knowledge can be inferred from cases of successes and failures, the common associations between datasets and control parameters, and many other patterns exhibited by the systems, the users and the interactions. Such knowledge may include a popular approach, commonly-used parameter sets, the best practice, an optimization strategy, and so forth. Figure 4 shows such an infrastructure.

Such an infrastructure is general-purpose, and can support multiple application domains. It can potentially enable applications to benefit from the best practice and software developed for other applications. The development of such an infrastructure can be built upon the advances in other areas of computing technologies, including semantic computing, autonomic computing, knowledge-based systems, data warehousing, machine learning, and search engine optimization.

Conclusions

Similar to the development of many other computing technologies, for example, speech processing, computer vision, web technology, one likely development path for visualization is

- from *offline visualization*
- to *interactive visualization*,
- to *information-assisted visualization*,
- then to *knowledge-assisted visualization*.

Interactive visualization has reached a matured status. There is a significant amount of ongoing development currently in *information-assisted visualization*. With a large amount of information collected locally and globally, it is inevitable that there will be a transition to *knowledge-assisted visualization*.

As a discipline, visualization has thrived on helping application users to transfer *data* (C_{data}) in the computational space to *information* (P_{info}) and *knowledge* (P_{know}) in the perceptual and cognitive space. As a discipline, we need infrastructures to collect our own *data* about visualization processes, and to transfer such data to *information* and *knowledge*, which helps further our understanding as well as enhance the visualization technology.

Acknowledgement

We would like to thank Heike Jänicke and Professor Gerhard Brewka at University of Leipzig, and Dr. Phil Grant and Dr. John Sharp at Swansea University for their advices and comments on terminologies used in information theory, knowledge-based systems, and other aspects of computing.

References

1. R. L. Ackoff, "From data to wisdom", *Journal of Applied Systems Analysis*, vol. 16, 1989, pp. 3-9.
2. S. K. Card, J. D. Mackinlay and B. Shneiderman, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann Publishers, San Francisco, 1999.
3. E. H. Chi, "A taxonomy of visualization techniques using the data state reference model", *Proc. IEEE Symposium on Information Visualization*, 2000, pp. 69-75.
4. U. Fayyad, G. G. Grinstein and A. Wierse, *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann Publishers, San Francisco, 2002.
5. K. Lehrer, *Theory of Knowledge*, Westview Press, 1990.
6. G. Scott, G. Domik, T.-M. Rhyne, K. W. Brodrie and B. S. Santos, *Definitions and Rationale for Visualization*, <http://www.siggraph.org/education/materials/HyperVis/visgoals/visgoal2.htm>, (accessed in April 2008).
7. N. Sharma, *The Origin of the "Data Information Knowledge Wisdom" Hierarchy*, http://www-personal.si.umich.edu/~nsharma/dikw_origin.htm, (accessed in April 2008).
8. B. Shneiderman, "The eyes have it: a task by data type taxonomy for information visualizations", *Proc. IEEE Symposium on Visual Languages*, 1996, pp. 336-343.
9. E. E. Smith and S. M. Kosslyn, *Cognitive Psychology: Mind and Brain*, Pearson Prentice-Hall, Upper Saddle River, New Jersey, 2007.
10. M. Tory and T. Moller, "Rethinking visualization: A high-Level taxonomy", *Proc. IEEE Symposium on Information Visualization*, 2004, pp.151-158.

Examples

There are many examples of *information-assisted visualization*. On the other hand, the development of *knowledge-assisted visualization* is very much in its infancy. Here we selectively describe several examples of information-assisted visualization in the literature, whilst accentuating the use, or potential use, of *knowledge* in a few visualization systems. These examples are intended to reinforce the viewpoints of this article, rather than to provide a comprehensive survey.

Examples of information-assisted visualization

Curve-skeleton

Curve-skeletons are 1D geometrical representations abstracted from 3D objects in an input dataset. Such information can be used to aid visualization tasks, including virtual navigation, reduced-model formulation, visualization improvement, and animation. For example, in virtual endoscopy, curve-skeletons are used to specify collision free paths for navigation through human organs [2].

Isosurface topology

Isosurface topology, which is typically represented as a contour tree, provides an abstract insight into the structural relationship and connectivity between isosurfaces in a dataset. In volume visualization, such information can assist users in distinguishing features in different topological zones, comprehending complex relationships between isosurfaces, and designing effective transfer functions [8].

Local statistical complexity

Local statistical complexity (LSC) is an information-theoretic measure, which tells how much information from the local past is required to predict the dynamics in the local future. Given a time-varying dataset, we can assign each data point an LSC value. Higher LSC values indicate regions that feature an extraordinary temporal evolution, whereas, lower values indicate temporal patterns that occur frequently in the dataset [5]. As demonstrated in Figure 5, such information can assist users in generating a visualization that highlights temporally-important features.

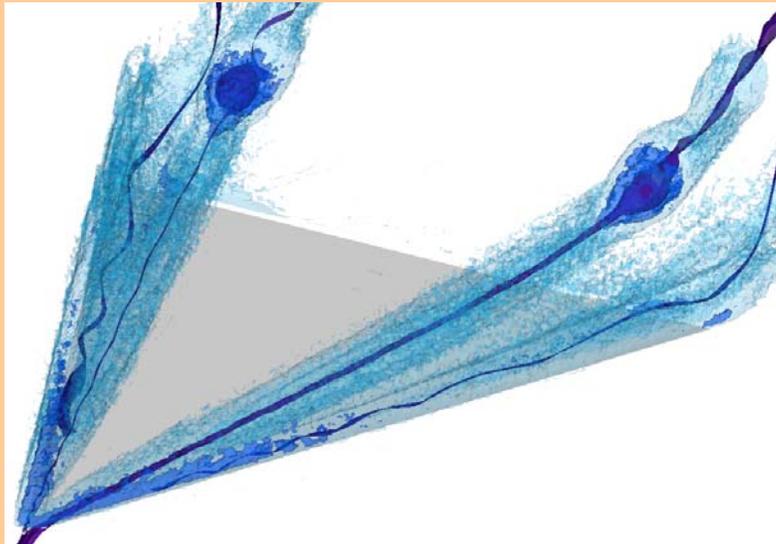


Figure 5. The local statistical complexity (LSC) of a flow around a delta wing (gray triangle). Four streamsurfaces indicate the vortices on top of the delta wing. The two isosurfaces in blue and light blue separate regions that hold LSC values within the range [14.7;15] and [11;15] respectively. High LSC values point the user to distinctive regions that may feature significant temporal events. The image is provided by Heike Jänicke, University of Leipzig [5].

Data abstraction quality

Measuring the quality of visualization results, such as visual density and clutter, provides users with useful guidance in synthesizing the most effective visualization. One of such measurements is data abstraction quality, measuring the degree to which the visualization results convey the original dataset. Such information enables users to determine the optimal abstraction level for a given visualization task, and to compare different visualization methods in terms of their capability of maintaining dominant characteristics of the original dataset while reducing the size and detail of the data [3].

Examples of knowledge-assisted visualization

Viewpoint mutual information

From Figures 2 and 3, we can observe that one transition path of *information-assisted visualization* to *knowledge-assisted visualization* is to automate the process of reasoning about the information abstracted from the input data. A classical example of such a transition is [7], where viewpoint mutual information (VMI) that measures the dependence or correlation between a set of viewpoints and a set of objects in a dataset is used to determine the optimal viewpoint. The fundamental difference between this approach and the above-mentioned examples is that users do not make decision according to the processed VMI. Instead, a relatively simple rule for minimizing VMI is used to determine viewpoint transformation automatically. Such a rule can be seen as a piece of knowledge hard-coded in the system.

Pre-determined ranking

In [6], a noticeable amount of generic knowledge is captured as ranks of different visualization designs. This enables the visualization system to automatically take users through a design process for creating a visualization. The stored ranks and ranking conditions are essentially a collection of expert knowledge.

Ontology mapping

The determination of visualization designs and parameters should depend on the input data. One approach is to extract semantic information from the input data, and try to find the best match with the semantic information of visualization designs (e.g., treemaps, graphs) and the associated parameters (e.g., size, axes). In [4], three ontologies, which are knowledge representations, are used to store (a) the domain-specific semantics about a class of input data, (b) the semantics about available visualization designs, and (c) the ontological mapping from (a) to (b). With these three ontologies, different visualization designs are dynamically ranked according to the input data, and a set of highly-ranked visualization designs are presented to the user automatically.

Workflow management

VisTrails is a visualization infrastructure that provides users with workflow management [1]. It is capable of capturing and storing a huge amount of data about input datasets, user interaction and visualization results in visualization processes. *VisTrails* exhibits some of the primary characteristics of the knowledge supporting infrastructure shown in Figure 4, though it currently has limited automated reasoning capability. Such an infrastructure has great potential to be developed into an infrastructure for *knowledge-assisted visualization*.

References

1. L. Bavoil, S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva and H. T. Vo, "VisTrails: enabling interactive multiple-view visualizations", *Proc. IEEE Visualization*, 2005, pp. 135-142.
2. N. D. Cornea, D. Silver, P. Min, "Curve-skeleton properties, applications, and algorithms", *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 3, 2007, pp. 530-548.
3. Q. Cui; M. Ward, E. A. Rundensteiner and J. Yang, "Measuring data abstraction quality in multiresolution visualizations", *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, 2006, pp. 709-716.
4. O. Gilson, N. Silva, P.W. Grant and M. Chen, "From web data to visualization via ontology mapping", *Computer Graphics Forum* (special issue for *EuroVis2008*), vol. 27, no. 3, 2008, pp. 959-966.
5. H. Jänicke, A. Wiebel, G. Scheuermann and W. Kollmann, "Multifield visualization using local statistical complexity", *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, 2007, pp. 1384-1391.
6. J. D. Mackinlay, P. Hanrahan and C. Stolte, "Show Me: automatic presentation for visual analysis", *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, 2007, pp. 1137-1144.
7. I. Viola, M. Feixas, M. Sbert and M. E. Gröller, "Importance-driven focus of attention", *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, 2006, pp. 933-940.
8. G. H. Weber, S. E. Dillard, H. Carr, V. Pascucci, B. Hamann, "Topology-controlled volume rendering", *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 2, 2007, pp. 330 - 341.