The Value of Information Visualization

Jean-Daniel Fekete, Jarke van Wijk, John Stasko, Chris North

The Value of Information Visualization

Jean-Daniel Fekete[1], Jarke J. van Wijk[2], John T. Stasko[3], and Chris North[4]

[1] Université Paris-Sud, INRIA, Bât 490,
F-91405 Orsay Cedex, France,
Jean-Daniel.Fekete@inria.fr,
WWW home page: http://www.lri.fr/~fekete
[2] Department of Mathematics and Computing Science,
Eindhoven University of Technology, P.O. Box 513,
5600 MB EINDHOVEN, The Netherlands
vanwijk@win.tue.nl,
WWW home page: http://www.win.tue.nl/~vanwijk/
[3] School of Interactive Computing, College of Computing & GVU Center,
Georgia Institute of Technology, 85 5th St., NW,
Atlanta, GA 30332-0760, USA
stasko@cc.gatech.edu
WWW home page: http://www.cc.gatech.edu/~john.stasko
[4] Dept of Computer Science, 2202 Kraft Drive
Virginia Tech, Blacksburg, VA 24061-0106, USA,
north@vt.edu
WWW home page: http://people.cs.vt.edu/~north/

Abstract. Researchers and users of Information Visualization are convinced that it has value. This value can easily be communicated to others in a face-to-face setting, such that this value is experienced in practice. To convince broader audiences, and also, to understand the intrinsic qualities of visualization is more difficult, however. In this paper we consider information visualization from different points of view, and gather arguments to explain the value of our field.

1 Problems and Challenges

This paper provides a discussion of issues surrounding the value of Information Visualization (InfoVis). The very existence of the paper should alert the reader that challenges do exist in both recognizing and communicating the field's value. After all, if the value would be clear and undisputed, there would be no need to write the paper! Unfortunately, the current situation is far from that. By its very focus and purpose, InfoVis is a discipline that makes the recognition of value extremely difficult, a point that will be expanded below.

Why is showing value important? Well, today's research environment places great importance on evaluation involving quantifiable metrics that can be assessed and judged with clarity and accuracy. Organizations sponsoring research and corporations that serve to benefit from it want to know that the monetary investments they make are being well-spent. Researchers are being challenged to

show that their inventions are measurably better than the existing state of the art.

In broad analytic fields, of which we include InfoVis as a member, the existence of a ground truth for a problem can greatly facilitate evaluations of value. For instance, consider the field of computer vision and algorithms for identifying objects from scenes. It is very easy to create a library of images upon which new algorithms can be tested. From that, one can measure how well each algorithm performs and compare results precisely. The TREC [25] and MUC [3] Contests are examples of this type of evaluation.

Even with a human in the loop, certain fields lend themselves very well to quantifiable evaluations. Consider systems that support search for particular documents or facts. Even though different people will perform differently using a system, researchers can run repeated search trials and measure how often a person is able to find the target and how long the search took. Averaged over a large number of human participants, this task yields quantifiable results that can be measured and communicated quite easily. People or organizations then using the technology can make well-informed judgments about the value of new tools.

So why is identifying the value of InfoVis so difficult? To help answer that question, let us turn to what is probably the most accepted definition of InfoVis, one that comes from Card, Mackinlay, and Shneiderman and that actually is their definition for "visualization." They describe visualization as "the use of computer-supported, interactive visual representations of data to amplify cognition." [2] The last three words of their definition communicate the ultimate purpose of visualization, to amplify cognition. So, returning to our discussion above, is the amplification of cognition something with a ground truth that is easily and precisely measurable? Clearly it is not and so results the key challenge in communicating the value of InfoVis.

Further examining the use and purpose of InfoVis helps understand why communicating its value is so difficult. InfoVis systems are best applied for exploratory tasks, ones that involve browsing a large information space. Frequently, the person using the InfoVis system may not have a specific goal or question in mind. Instead, the person simply may be examining the data to learn more about it, to make new discoveries, or to gain insight about it. The exploratory process itself may influence the questions and tasks that arise.

Conversely, one might argue that when a person does have a specific question to be answered, InfoVis systems are often not the best tools to use. Instead, the person may formulate his or her question into a query that can be dispatched to a database or to a search engine that is likely to provide the answer to that precise question quickly and accurately.

InfoVis systems, on the other hand, appear to be most useful when a person simply does not know what questions to ask about the data or when the person wants to ask better, more meaningful questions. InfoVis systems help people to rapidly narrow in from a large space and find parts of the data to study more carefully.

Unfortunately, however, activities like exploration, browsing, gaining insight, and asking better questions are not ones that are easily amenable to establishing and measuring a ground truth. This realization is at the core of all the issues involved in communicating the value of InfoVis. By its very nature, by its very purpose, InfoVis presents fundamental challenges for identifying and measuring value. For instance, how does one measure insight? How does one quantify the benefits of an InfoVis system used for exploring an information space to gain a broad understanding of it? For these reasons and others, InfoVis is fundamentally challenging to evaluate [13].

If we accept that InfoVis may be most valuable as an exploratory aid, then identifying situations where browsing is useful can help to determine scenarios most likely to illustrate InfoVis' value. Lin [8] describes a number of conditions in which browsing is useful:

- When there is a good underlying structure so that items close to one another can be inferred to be similar
- When users are unfamiliar with a collection's contents
- When users have limited understanding of how a system is organized and prefer a less cognitively loaded method of exploration
- When users have difficulty verbalizing the underlying information need
- When information is easier to recognize than describe

These conditions serve as good criteria for determining situations in which the value of InfoVis may be most evident.

1.1 Epistemological Issues

Natural sciences are about understanding how nature works. Mathematics is about truth and systems of verifiable inferences. Human sciences are about understanding Man in various perspectives. Information Visualization is about developing insights from collected data, not about understanding a specific domain. Its object is unique and therefore raises interest and skepticism.

Science has focused on producing results: the goal was essentially the creation and validation of new theories compatible with collected facts. The importance of the process — coined as the Method — was raised by the development of *epistemology* in the 20th century, in particular with the work of Karl R. Popper (1902–1994) [14]. It showed that the Method was paramount to the activity of science.

Karl Popper has explained that a scientific theory cannot be proved true, it can only be *falsified*. Therefore, a scientific domain searches for theories that are as compatible as possible with empirical facts. The good theories are the ones that have been selected by domain experts among a set of competing theories in regard of the facts that they should describe. Popper considers science as a Darwinian selection process among competing theories.

Still, no other scientific domain has argued that generating insights was important for science. Popper does not explain how a new theory emerges; he only

explains how it is selected when it emerges. Furthermore, Popper has demonstrated in an article called "The Problem of Induction" that new theories cannot rationally emerge from empirical data: it is impossible to justify a law by observation or experiment, since it 'transcends experience'.

Information Visualization is still an inductive method in the sense that it is meant at generating new insights and ideas that are the seeds of theories, but it does it by using human perception as a very fast filter: if vision perceives some pattern, there might be a pattern in the data that reveals a structure. Drilling down allows the same perception system to confirm or infirm the pattern very quickly. Therefore, information visualization is meant at "speeding up" the process of filtering among competing theories regarding collected data by relying on the speed of the perception system. Therefore, it plays a special role in the sciences as an insight generating method. It is not only compatible with Popper's epistemology system but it furthermore provides a mechanism for accelerating its otherwise painful Darwinian selection process.

1.2 Moving Forward

It is clear that InfoVis researchers and practitioners face an important challenge in communicating the value of InfoVis. In the remainder of the paper we explore this challenge more deeply and we provide several answers to the questions "How and why is InfoVis useful?". Since there are several audiences to convince, we present a number of different sections that are each facets of argumentation to explain why InfoVis is useful and effective as a mean of understanding complex datasets and developing insights. The contents of the sections are gathered from practitioners who already attested that the arguments developed were convincing. We hope they will be useful to you as well.

2 Cognitive and Perceptual Arguments

Several famous historical figures have argued that the eye was the main sense to help us understand nature.

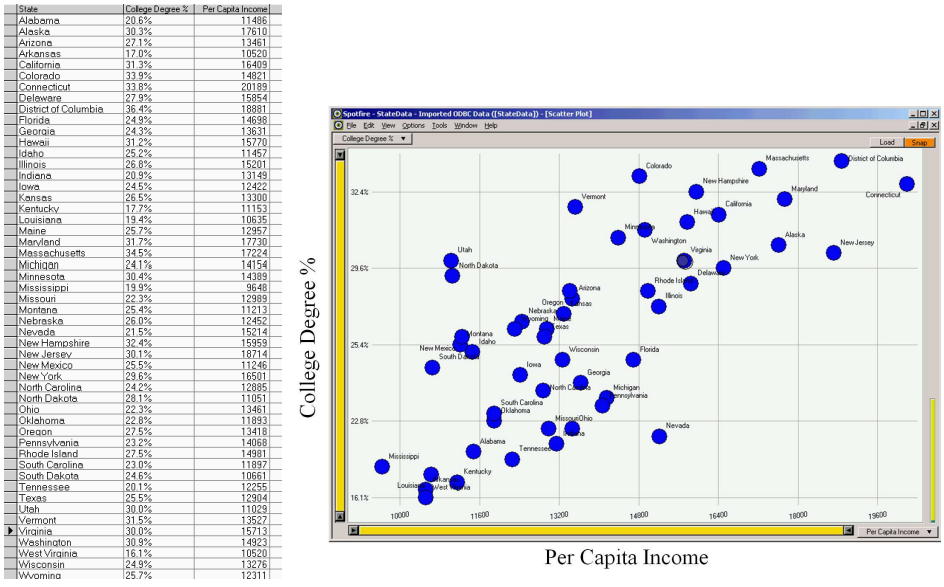
> The eye...
> the window of the soul,
> is the principal means
> by which the central sense
> can most completely and
> abundantly appreciate
> the infinite works of nature.
>
> Leonardo da Vinci (1452 - 1519)

Leonardo's words are inspirational and they are echoed in everyday expressions that we are all familiar with such as, "Seeing is believing" and "A picture is worth a thousand words." Is there actual support for these sentiments, however?

Let us first consider the case of the phrase, "A picture is worth a thousand words." While people may agree or disagree with the sentiments behind that cliché, specific examples can help support the claim. Consider, for instance, the example shown in Figure 1. Part (a) shows a spreadsheet with data for the 50 states and the District of Columbia in the U.S. Also shown are the percentage of citizens of each state with a college degree and the per capita income of the states' citizens.

Given just the spreadsheet, answering a question such as, "Which state has the highest average income?" is not too difficult. A simple scan of the income column likely will produce the correct answer in a few seconds. More complex questions can be quite challenging given just the data, however. For example, are the college degree percentage and income correlated? If they are correlated, are there particular states that are outliers to the correlation? These questions are much more difficult to answer using only the spreadsheet.

Now, let us turn to a graphical visualization of the data. If we simply draw the data in a scatterplot as shown in part (b), the questions now become much easier to answer. Specifically, there does appear to be an overall correlation between the two attributes and states such as Nevada and Utah are outliers on the correlation. The simple act of plotting the spreadsheet data in this more meaningfully communicative form makes these kinds of analytic queries easier to answer correctly and more rapidly.



State	College Degree %	Per Capita Income
Alabama	20.6%	11486
Alaska	30.3%	17610
Arizona	27.1%	13461
Arkansas	17.0%	10520
California	31.3%	16409
Colorado	33.9%	14821
Connecticut	33.8%	20188
Delaware	27.9%	15854
District of Columbia	36.4%	18881
Florida	24.9%	14638
Georgia	24.3%	13631
Hawaii	31.2%	15770
Idaho	25.2%	11457
Illinois	26.8%	15201
Indiana	20.9%	13149
Iowa	24.5%	12422
Kansas	26.5%	13300
Kentucky	17.7%	11153
Louisiana	19.4%	10635
Maine	25.7%	12957
Maryland	31.7%	17730
Massachusetts	34.5%	17224
Michigan	24.1%	14154
Minnesota	30.4%	14389
Mississippi	19.9%	9648
Missouri	22.3%	12989
Montana	25.4%	11213
Nebraska	26.0%	12452
Nevada	21.5%	15214
New Hampshire	32.4%	15959
New Jersey	30.1%	18714
New Mexico	25.5%	11246
New York	29.6%	16501
North Carolina	24.2%	12885
North Dakota	28.1%	11051
Ohio	22.3%	13461
Oklahoma	22.8%	11893
Oregon	27.5%	13418
Pennsylvania	23.2%	14068
Rhode Island	27.5%	14981
South Carolina	23.0%	11897
South Dakota	24.6%	10661
Tennessee	20.1%	12255
Texas	25.5%	12904
Utah	30.0%	11029
Vermont	31.5%	13627
Virginia	30.0%	15713
Washington	30.9%	14923
West Virginia	16.1%	10520
Wisconsin	24.9%	13276
Wyoming	25.7%	12311

(a) A thousand words

(b) A picture

Fig. 1: "A picture is worth a thousand words"

Note that the spreadsheet itself is a visual representation of the data that facilitates queries as well. Consider how difficult the three questions would be if the data for each state was recorded on a separate piece of paper or webpage. Or worse yet, what if the data values were read to you and you had to answer the questions? In this case, already challenging questions become practically impossible.

2.1 Cognitive Benefits

While the states example illustrates that visualizations can help people understand data better, how do visuals facilitate this process? The core of the benefits provided by visuals seems to hinge upon their acting as a frame of reference or as a temporary storage area for human cognitive processes. Visuals augment human memory to provide a larger working set for thinking and analysis and thus become external cognition aids. Consider the process of multiplying two long integers in your head versus then having a pencil and paper available. The visual representations of the numbers on paper acts as a memory aid while performing the series of multiplication operations.

Performing a multiplication requires the processing of symbolic data, which is arguably different than the processing of visual features and shapes. In "Abstract Planning and Perceptual Chunks: Elements of Expertise in Geometry" [5], Koedinger and Anderson show that experts in geometry effectively use their vision to infer geometrical properties (parallelism, connectivity, relative positions) on diagrams; they solve simple problems quickly and accurately, several of magnitude faster than if they had to apply symbolic inference rules.

Larkin and Simon, in their landmark paper "Why a diagram is (sometimes) worth 10,000 words" [7], discuss how graphical visualization can support more efficient task performance by allowing substitution of rapid perceptual inferences for difficult logical inferences and by reducing the search for information required for task completion. They do note that text can be better than graphics for certain tasks, however.

Don Norman provides many illustrative examples where visuals can greatly assist task performance and efficiency [11]. He argues that it is vital to match the representation used in a visualization to the task it is addressing. The examples he cites show how judicious visuals can aid information access and computation.

Earlier, we noted how the definition of visualization from Card, Mackinlay and Shneiderman [2] focused on the use of visuals to "amplify cognition." Following that definition, the authors listed a number of key ways that the visuals can amplify cognition:

- Increasing memory and processing resources available
- Reducing search for information
- Enhancing the recognition of patterns
- Enabling perceptual inference operations
- Using perceptual attention mechanisms for monitoring
- Encoding info in a manipulable medium

2.2 Perceptual Support

Most lectures on Information Visualization argue about theoretical properties of the visual system or more broadly to the perception abilities of humans. Rational arguments rely on information theory [17] and psychological findings.

According to *Information Theory*, vision is the sense that has the largest bandwidth: 100 Mb/s [26]. Audition only has around 100 b/s. In that respect, the visual canal is the best suited to carrying information to the brain.

According to Ware [26], there are two main *psychological theories* that explain how vision can be used effectively to perceive features and shapes. At the low level, *Preattentive processing* theory [19] explains what visual features can be effectively processed. At a higher cognitive level, the *Gestalt* theory [6] describes some principles used by our brain to understand an image.

Preattentive processing theory explains that some visual features can be perceived very rapidly and accurately by our low-level visual system. For example, when looking at the group of blue circles in Figure 2, it takes no time and no effort to see the red circle in the middle. It would be as easy and fast to see that there is no red circle, or to evaluate the relative quantity of red and blue circles. Color is one type of feature that can be processed preattentively, but only for some tasks and within some limits. For example, if there were more than seven colors used in Figure 2, answering the question could not be done with preattentive processing and would require sequential scanning, a much longer process.

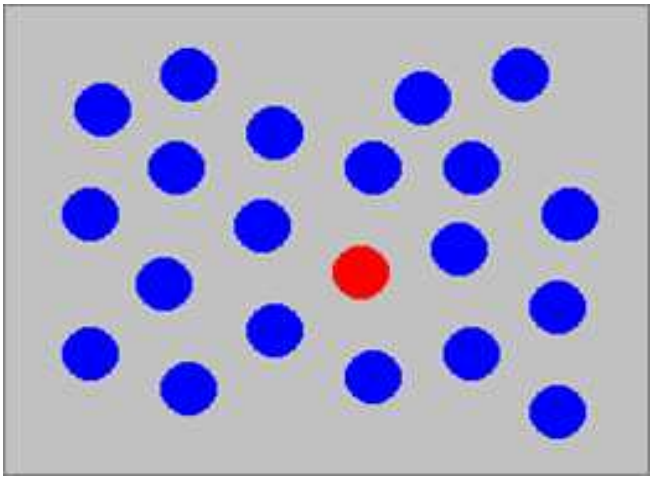


Fig. 2: Example of preattentively processed task: finding if there is a red circle among the blue circles

There is a long list of visual features that can be preattentively processed for some tasks, including line orientation, line length or width, closure, curvature,

color and many more. Information visualization relies on this theory to choose the visual encoding used to display data to allow the most interesting visual queries to be done preattentively.

Gestalt theory explains important principles followed by the visual system when it tries to understand an image. According to Ware [26], it is based on the following principles:

Proximity Things that are close together are perceptually grouped together;
Similarity Similar elements tend to be grouped together;
Continuity Visual elements that are smoothly connected or continuous tend to be grouped;
Symmetry Two symmetrically arranged visual elements are more likely to be perceived as a whole;
Closure A closed contour tends to be seen as an object;
Relative Size Smaller components of a pattern tend to be perceived as objects whereas large ones as a background.

Information Visualization experts design visual representations that try to follow these principles. For example, graph layout algorithms such as [10] designed to find communities in social networks adhere to the *proximity* principle by placing nodes that are connected to a dense group close together and push away nodes that are connected to another dense group. The Treemap algorithm [4] uses the *closure* principle to layout a tree: children of a node are placed inside their parent node.

3 Success Stories

Information Visualization is much easier to explain using demonstrations than words. However, to be understood, the data used should be familiar to the audience and interesting. Preparing demonstrations targeted at all the possible audiences is not possible but there are some datasets that interest most audiences and greatly help make the point. Several striking static examples can be found in Tufte's books [20,21,22].

To better explain the value of visualization, demonstrations should start using a simple question, show that a good representation answers the question at once and then argue about additional benefits, *i.e.* questions the users did not knew they had. From the users perspective, a good representation will confirm what they already know, let them answer at once the question asked and show them several insights, leading to the so-called "a-ha" moments when they feel like they understand the dataset.

3.1 Striking Examples

Static examples used by most InfoVis courses include the map of Napoleon's 1812 March on Moscow drawn in 1869 by M. Minard (Figure 3) and the map of London in 1854 overlaid with marks positioning cholera victims that led John

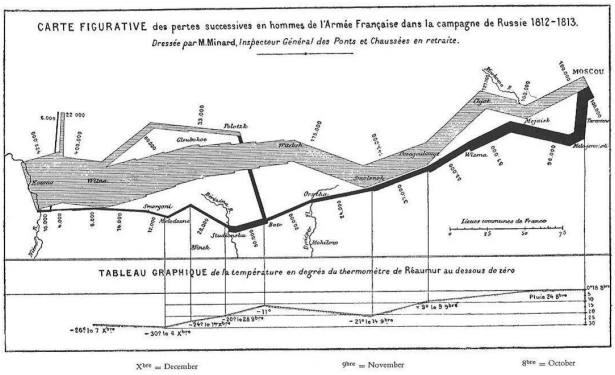


Fig. 3: Napoleon's March on Moscow depicted by M. Minard. Width indicates the number of soldiers. Temperature during the retreat is presented below the map.

Snow to discovering the origin of the epidemic: infected water extracted with a water pump at the center of the marks (Figure 4).

In general, good examples show known facts (although sometimes forgotten) and reveal several unexpected insights at once. Minard's map can help answer the question: "What were the casualties of Napoleon's Russian invasion in 1812?". The map reveals at once the magnitude of casualties (from 400,000 to 10,000 soldiers) as well as the devastating effect of crossing the Berezina river (50,000 soldiers before, 25,000 after). The depiction confirms that Napoleon lost the invasion (a well known fact) and reveals many other facts, such as the continuous death rate due to disease and the "scorched earth" tactics of Russia instead of specific death tolls of large battles.

John Snow's map was made to answer the question: "What is the origin of the London cholera epidemics?". Contrary to the previous map, the answer requires some thinking. Black rectangles indicate location of deaths. At the center of the infected zone lies a water pump that John Snow found to be responsible for the infection. Once again, choosing the right representation was essential for finding the answer. As a side-effect, the map reveals the magnitude of the epidemic.

Figure 1 answers the question: "Is there a relationship between income and college degree?" by showing a scatter plot of income by degree for each US state. The answer is the obvious: yes, but there is much more. There seems to be a linear correlation between them and some outliers such as Nevada (likely due to Las Vegas) and Utah do exist, raising new unexpected questions.

Information Visualization couples interaction and visual representation so its power is better demonstrated interactively. The simplest demonstration suited to the largest audience is probably the Dynamic HomeFinder[5] [28] . It shows the map of the Washington D.C. area overlaid with all the homes for sale (Figure 5).

[5] `http://www.cs.umd.edu/hcil/pubs/dq-home.zip`

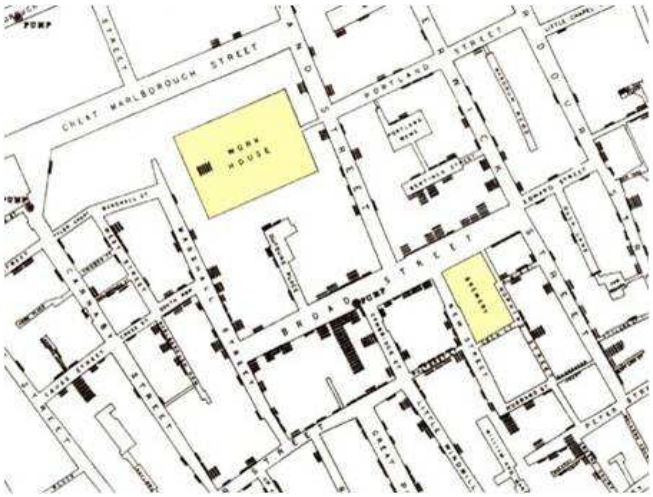


Fig. 4: Illustration of John Snow's deduction that a cholera epidemic was caused by a bad water pump, circa 1854. Black rectangles indicate location of deaths.

Dynamic queries implemented by sliders and check-boxes interactively filter-out homes that do not fit specific criteria such as cost or number of bedrooms.

Using the interactive controls, it becomes easy to find homes with the desired attributes or understand how attributes' constraints should be relaxed to find some matching homes. Unexpectedly, the Dynamic HomeFinder also reveals the unpopular neighborhoods around Washington D.C. since they are places where the homes are cheaper, and the wealthy ones where the houses are more expensive.

Many more examples can be found to demonstrate that InfoVis is effective. The Map of the Market[6], represented by a squarified treemap, is interesting for people holding stocks or interested by economic matters. InfoZoom video on the analysis of Formula 1 results[7] is interesting for car racing amateurs. The video[8] comparing two large biological classification trees is interesting to some biologists. The Baby Name Wizard's NameVoyager[9] is useful for persons searching a name for their baby to come and a large number of other persons as witnessed by [27].

With the advent of Social InfoVis through web sites such as Swivel[10] or IBM's Many-Eyes[11], more examples can be found to convince specific audiences. Still,

[6] http://www.smartmoney.com/marketmap/

[7] http://www.infozoom.com/enu/infozoom/video.htm

[8] http://www.fit.fraunhofer.de/~cici/InfoVis2003/StandardForm/Flash/InfoZoomTrees.html

[9] http://babynamewizard.com/namevoyager/

[10] http://www.swivel.com

[11] http://www.many-eyes.com

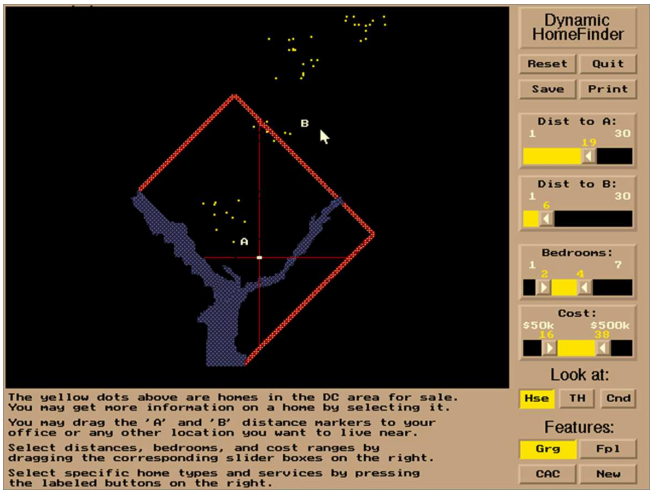


Fig. 5: Dynamic HomeFinder showing the Washington D.C. area with homes available for sale and controls to filter them according to several criterion.

the process of explaining how InfoVis works remains the same: ask a question that interests people, show the right representation, let the audience understand the representation, answer the question and realize how many more unexpected findings and questions arise.

3.2 Testimonials

One effective line of argumentation about the value of InfoVis is through reporting the success of projects that used InfoVis techniques. These stories exist but have not been advertised in general scientific publications until recently [16,12,9]. One problem with trying to report on the success of a project is that visualization is rarely the only method used to reach the success. For example, in biological research, the insights gained by an InfoVis system can lead to an important discovery that is difficult to attribute mainly to the visualization since it also required months of experimentation to verify the theory formulated from the insights. In fact, most good human-computer interaction systems allow users to forget about the system and focus on their task only, which is probably one reason why success stories are not so common in the InfoVis literature.

Besides these stories that are empirical evidence of the utility of information visualization, there are strong theoretical arguments to how and why information visualization works.

4 Information Visualization vs. Automatic Analysis

Several scientific domains are concerned by understanding complex data. Statistics is the oldest, but Data Mining — a subfield of Artificial Intelligence — is

also concerned with automatically understanding the structure of data. Therefore, InfoVis practitioners frequently need to explain what InfoVis can do that statistics and data mining cannot.

4.1 Statistics

Statistics is a well grounded field but is composed of several subfields such as descriptive statistics, classical statistics (also called *confirmatory* statistics), Bayesian statistics and Exploratory Data Analysis. Information Visualization is sometimes considered as a descendant and expansion of Exploratory Data Analysis.

The differences between these subfields are the methods and the nature of the answer they seek. All of them start with a problem and gathered data that is related to the problem to solve. Classical analysis starts by designing a *model* of the data, then uses mathematical analysis to test whether the model is refuted or not by the data to conclude positively or negatively. The main challenge for classical statistics is to find a model.

Exploratory Data Analysis performs an analysis using visual methods to acquire insights of what the data looks like, usually to find a model. It uses visual exploration methods to get the insights.

So why is visualization useful before the modeling? Because, there are cases when we have no clear idea on the nature of the data and have no model.

To show why visualization can help finding a model, Anscombe in [1] has designed four datasets that exhibit the same statistical profile but are quite different in shape, as shown in Figure 6. They have the following characteristics[12]:

- mean of the x values = 9.0
- mean of the y values = 7.5
- equation of the least-squared regression line is: y = 3 + 0.5x
- sums of squared errors (about the mean) = 110.0
- regression sums of squared errors (variance accounted for by x) = 27.5
- residual sums of squared errors (about the regression line) = 13.75
- correlation coefficient = 0.82
- coefficient of determination = 0.67.

Visualization is much more effective at showing the differences between these datasets than statistics. Although the datasets are synthetic, Anscombe's Quartet demonstrates that looking at the shape of the data is sometimes better than relying on statistical characterizations alone.

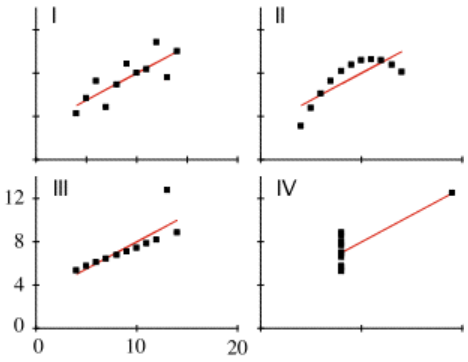
4.2 Data Mining

More than statistics, the goal of data mining is to automatically find interesting facts in large datasets. It is thus legitimate to wonder whether data mining, as a competitor of InfoVis, can overcome and replace the visual capacity of humans.

[12] See `http://astro.swarthmore.edu/astro121/anscombe.html` for details

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

(a) Four datasets with different values and the same statistical profile



(b) Dot Plot of the four datasets

Fig. 6: Anscombe's Quartet

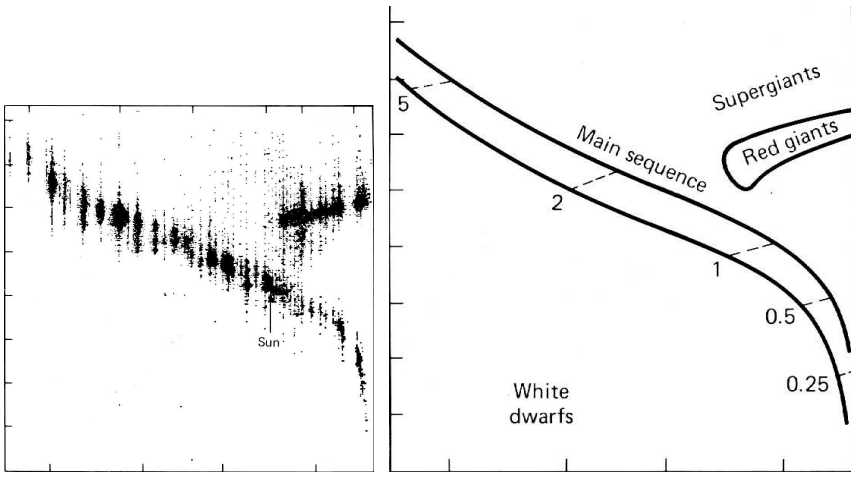
This question has been addressed by Spence and Garrison in [18] where they describe a simple plot called the Hertzsprung Russell Diagram (Figure 7a). It represents the temperature of stars on the X axis and their magnitude on the Y axis. Asking a person to summarize the diagram produces Figure 7b. It turns out that no automatic analysis method has been able to find the same summarization, due to the noise and artifacts on the data such as the vertical bands.

Our vision system has evolved with the human specie to help us survive in a hostile world. We train it to avoid obstacles since we learn how to walk. It remains remarkably effective at filtering-out noise from useful data, a very important capability for hunters in deep forests to distinguish the prey moving behind leaves. We have relied on it and trained it to survive for millennia and it still surpasses automatic data mining methods to spot interesting patterns. Data mining still needs to improve to match these pattern matching capabilities..

4.3 Automating or Not?

Is there a competition between confirmatory, automated and exploratory methods? No, they answer different questions. When a model is known in advance or expected, using statistics is the right method. When a dataset becomes too large to be visualized directly, automating some analysis is required. When exploring a dataset in search of insights, information visualization should be used, possibly in conjunction with data mining techniques if the dataset is too large.

Furthermore, combining data mining with visualization is the central issue of *Visual Analytics*, described by the paper *Visual Analytics: Definition, Process, and Challenges* in this book.



(a) Hertzsprung Russell Diagram (b) Standard interpretation of the Hertzsprung Russell Diagram

Fig. 7: Hertzsprung Russell Diagram and its standard interpretation

5 An Economical Model of Value

One important question is how to assess the value of visualization, ranging from the evaluation of one specific use-case to the discipline in general. If we know how to do this, then this might lead to an assesment of the current status as well as the identification of success factors. An attempt was given by Van Wijk [23] and is summarized here. After a short overview of his model, we discuss how this model can be applied for InfoVis.

Visualization can be considered as a technology, a collection of methods, techniques, and tools developed and applied to satisfy a need. Hence, standard technological measures apply: Visualization has to be effective and efficient. To measure these, an economic point of view is adopted. Instead of trying to understand why visualization works (see previous sections), here visualization is considered from the outside, and an attempt is made to measure its profit. The profit of visualization is defined as the difference between the value of the increase in knowledge and the costs made to obtain this insight. Obviously, in practice these are hard to quantify, but it is illuminating to attempt so. A schematic model is considered: One visualization method V is used by n users to visualize a data set m times each, where each session takes k explorative steps. The value of an increase in knowledge (or insight) has to be judged by the user. Users can be satisfied intrinsically by new knowledge, as an enrichment of their understanding of the world. A more pragmatic and operational point of view is to consider if the new knowledge influences decisions, leads to actions, and,

hopefully, improves the quality of these. The overall gain now is $nm(W(\Delta K))$, where $W(\Delta K))$ represents the value of the increase in knowledge.

Concerning the costs for the use of (a specific) visualization V, these can be split into various factors. Initial research and development costs C_i have to be made; a user has to make initial costs C_u, because he has to spend time to select and acquire V, and understand how to use it; per session initial costs C_s have to be made, such as conversion of the data; and finally during a session a user makes costs C_e, because he has to spend time to watch and understand the visualization, and interactively explore the data set. The overall profit now is

$$F = nm(W(\Delta K) - C_s - kC_e) - C_i - nC_u.$$

In other words, this leads to the obvious insight that a great visualization method is used by many people, who use it routinely to obtain highly valuable knowledge, while having to spend little time and money on hardware, software, and effort. And also, no alternatives that are more cost-effective should be available.

In the original paper a number of examples of more or less successful visualization methods are given, viewed in terms of this model. One InfoVis application was considered: SequoiaView, a tool to visualize the contents of a hard disk, using cushion treemaps [24]. The popularity of this tool can be explained from the concrete and useful insights obtained, as well as the low costs in all respects associated with its application.

When we consider InfoVis in general, we can also come to positive conclusions for almost all parameters, and hence predict a bright future for our field. The number of potential users is very large. Data in the form of tables, hierarchies, and networks is ubiquitous, as well as the need to get insight in these. This holds for professional applications, but also for private use at home. Many people have a need to get an overview of their email, financial transfers, media collections, and to search in external data bases, for instance to find a house, vacation destination, or another product that meets their needs. Methods and techniques from InfoVis, in the form of separate tools or integrated in custom applications, can be highly effective here to provide such overviews. Also, many of these activities will be repeated regularly, hence both n and m are high. The growing field of Casual InfoVis [15] further illustrates how InfoVis techniques are becoming more common in people's everyday lives.

The costs C_e that have to be made to understand visualizations depend on the prior experience of the users as well as the complexity of the imagery shown. On the positive side, the use of graphics to show data is highly familiar, and bar-charts, pie-charts, and other forms of business graphics are ubiquitous. On the other hand, one should not overestimate familiarity. The scatterplot seems to be at the boundary: Considered as trivial in the InfoVis community, but too hard to understand (if the horizontal axis does not represent time) by a lay-audience, according to Matthew Ericson, deputy graphics director of the New York Times in his keynote presentation at IEEE InfoVis 2007. Visual literacy is an area where more work can be done, but on the other hand, InfoVis does have a strong edge compared to non-visual methods here. And, there are examples of

areas where complex visual coding has been a great success, with the invention of the script as prime example.

The costs C_s per session and C_u per user can be reduced by tight integration with applications. The average user will not be interested in producing visualizations, her focus will be on solving her own problem, where visualization is one of the means to this end. Separate InfoVis tools are useful for specialists, which use them on a day-to-day basis. For many other users, integration within their favourite tool is much more effective. An example of an environment that offers such a tight integration is the ubiquitous spreadsheet, where storage, manipulation, and presentation of data are offered; or the graphs and maps shown on many web sites (and newspapers!) to show data. From an InfoVis point of view, the presentations offered here can often be improved, and also, the interaction provided is often limited. Nevertheless, all these examples acknowledge the value of visualization for many applications.

The initial costs C_i for new InfoVis methods and techniques roughly fall into two categories: Research and Development. Research costs can be high, because it is often hard to improve on the state of the art, and because many experiments (ranging from the development of prototypes to user experiments) are needed. On the other hand, when problems are addressed with many potential usages, these costs are still quite limited. Development costs can also be high. It takes time and effort to produce software that is stable and useful under all conditions, and that is tightly integrated with its context, but here also one has to take advantage of the large potential market. Development and availability of suitable middleware, for instance as libraries or plug-ins that can easily customized for the problem at hand is an obvious route here.

One intriguing aspect here is how much customization is needed to solve the problem concerned. On one hand, in many applications one of the standard data types of InfoVis is central (table, tree, graph, text), and when the number of items is not too high, the problem is not too hard to solve. On the other hand, for large numbers of items one typically has to exploit all a priori knowledge of the data set and tune the visualization accordingly; also, for applications such as software visualizations all these data types pop up simultaneously, which also strongly increases the complexity of the problem. So, for the time being, research and innovation will be needed to come up with solutions for such problems as well.

In conclusion, graphics has been adopted already on a large scale to communicate and present abstract data, which shows that its value has been acknowledged, and we expect that due to the increase in size and complexity of data available, the need for more powerful and effective information visualization methods and techniques will only grow.

6 Conclusion

In this paper we have described the challenges in identifying and communicating the value of InfoVis. We have cited and posed a number of answers to the

questions, "How and why is InfoVis useful?" Hopefully, the examples shown in the paper provide convincing arguments about InfoVis' value as an analytic tool. Ultimately, however, we believe that it is up to the community of InfoVis researchers and practitioners to create techniques and systems that clearly illustrate the value of the field. When someone has an InfoVis system that they use in meaningful and important ways, this person likely will not need to be convinced of the value of InfoVis.

References

1. F.J. Anscombe. Graphs in statistical analysis. *American Statistician*, 27(1):17–21, February 1973.
2. Stuart K. Card, Jock Mackinlay, and Ben Shneiderman, editors. *Readings in Information Visualization – Using Vision to Think*. Morgan Kaufmann, 1998.
3. N. Chincor, D. Lewis, and L. Hirschman. Evaluating message understanding systems: An analysis of the third message understanding conference (MUC-3). *Computational Linguistics*, 19(3):409–449, 1993.
4. Brian Johnson and Ben Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *VIS '91: Proceedings of the 2nd conference on Visualization '91*, pages 284–291, Los Alamitos, CA, USA, 1991. IEEE Computer Society Press.
5. Kenneth R. Koedinger and John R. Anderson. Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14(4):511–550, 1990.
6. Kurt Koffa. *Principles of Gestalt Psychology*. Routledge & Kegan Paul Ltd., London, 1935.
7. Jill H. Larkin and Herbert A. Simon. Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science*, 11:65–100, 1987.
8. Xia Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54, 1997.
9. Peter McLachlan, Tamara Munzner, Eleftherios Koutsofios, and Stephen North. Liverac: Interactive visual exploration of system management time-series data. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*. ACM Press, 2008.
10. Andreas Noack. Energy-based clustering of graphs with nonuniform degrees. In Patrick Healy and Nikola S. Nikolov, editors, *Proceedings of the 13th International Symposium on Graph Drawing (GD 2005)*, pages 309–320, Limerick, Ireland, 2005. Springer-Verlag.
11. Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1993.
12. Adam Perer and Ben Shneiderman. Integrating statistics and visualization: Case studies of gaining clarity during exploratory data analysis. In *SIGCHI Conference on Human Factors in Computing Systems (CHI 2008)*. ACM Press, 2008.
13. Catherine Plaisant. The challenge of information visualization evaluation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.
14. Karl R. Popper. *The Logic of Scientific Discovery*. New York, Basic Books, 1959.

15. Zachary Pousman, John Stasko, and Michael Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.
16. P. Saraiya, Chris North, V. Lam, and K. Duca. An insight-based longitudinal study of visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1511–1522, 2006.
17. Claude E. Shannon and Warren Weaver. *A Mathematical Theory of Communication*. University of Illinois Press, Champaign, IL, USA, 1963.
18. I. Spence and R. F. Garrison. A remarkable scatterplot. *The American Statistician*, pages 12–19, 1993.
19. A. Triesman. Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156–177, August 1985.
20. Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Box 430, Cheshire, CT 06410, USA, 1983.
21. Edward R. Tufte. *Envisioning Information*. Graphics Press, Box 430, Cheshire, CT 06410, USA, 1990.
22. Edward R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Box 430, Cheshire, CT 06410, USA, 1997.
23. Jarke J. van Wijk. The value of visualization. In *Proceedings IEEE Visualization 2005*, pages 79–86, 2005.
24. Jarke J. van Wijk and Huub van de Wetering. Cushion treemaps. In *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*, pages 73–78. IEEE CS Press, 1999.
25. Ellen Voorhees and Donna Harman. Overview of the sixth Text Retrieval Conference. *Information Processing and Management*, 36(1):3–35, 2000.
26. Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.
27. Martin Wattenberg and Jesse Kriss. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, 2006.
28. Christopher Williamson and Ben Shneiderman. The dynamic homefinder: evaluating dynamic queries in a real-estate information exploration system. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 338–346, New York, NY, USA, 1992. ACM.