Information Visualization (InfoVis)

by Robert Kosara, 2008-04-06 Article Research

Information Visualization (InfoVis) is a field of research that deals with the visual display of data. By looking at images of the data, we can use the immense power of our visual system to detect patterns or outliers, and quite generally come to a better understanding of our data. To achieve this, the visualization method must be suitable for the data and the task in question.

Making abstract numbers visible is not a new idea. Everybody knows line, bar, and pie charts. These give you a better overview of the data, because they make it possible to see trends or relationships immediately. It takes a very long time to read a column of numbers, and even longer to decide the overall trend of these numbers. Looking at a simple line chart, you can tell immediately. Such qualitative impressions of the data (e.g., does the trend point up or down, which company has the largest market share and how do the others compare, etc.) are often much more useful than the exact numbers. Of course, charts don't replace the numbers, they just show them in a way that is easy to grasp.

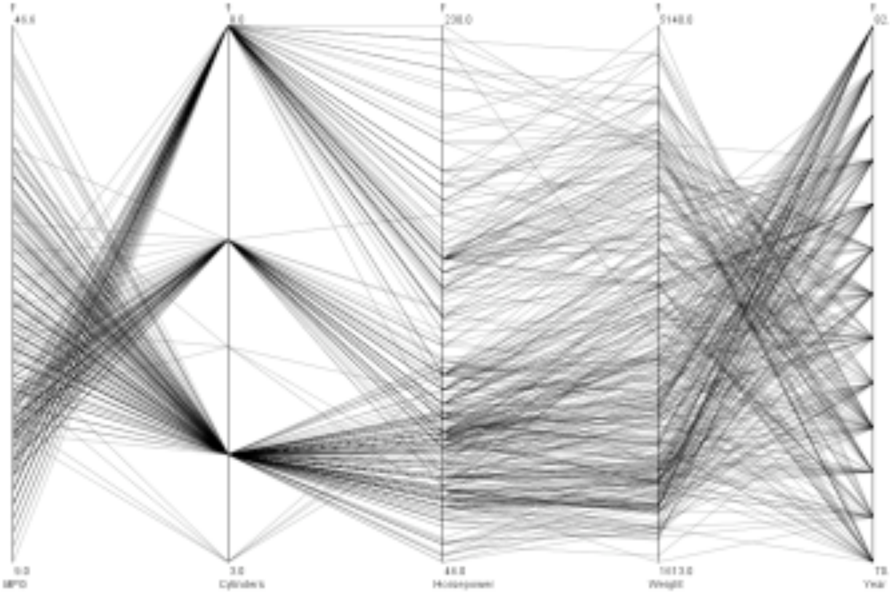
InfoVis goes into a similar direction, but goes much further. It allows you to look at millions of data items at the same time, and to interact with it. It thus makes it possible to visually analyze your data, not just draw a pretty graph of the results that you obtained by statistical (or other) means.

An Example

This example uses a dataset about car models from 1970-82. It contains about 380 records, and about 10 values per record. For this example, only five values are used: MPG (miles per gallon), cylinders, horsepower, weight (in kg), year (two digits). The following table shows you the first five records:

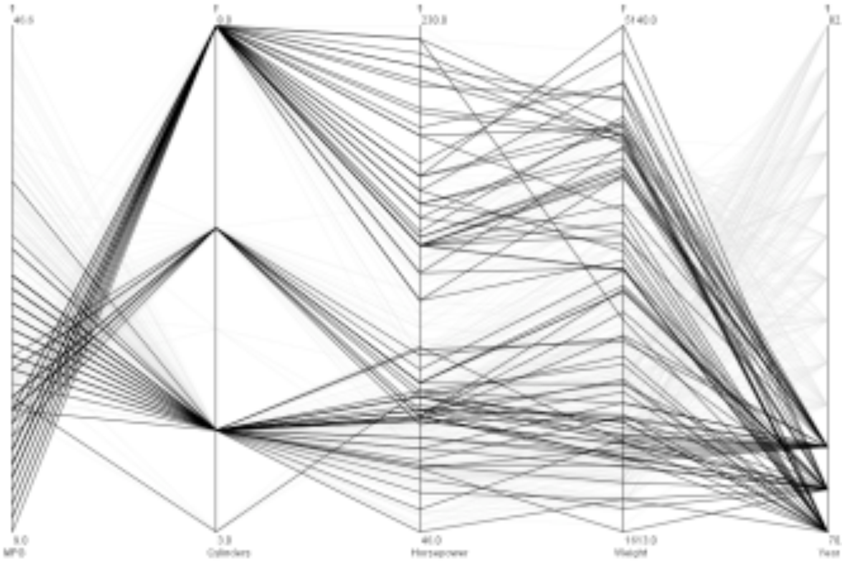
MPG	Cylinders	Horsepower	Weight	Year
18	8	130	3504	70
15	8	165	3693	70
18	8	150	3436	70
16	8	150	3433	70
17	8	140	3449	70
...	...	...	...	...

The following image shows the dataset using a visualization method called parallel coordinates (click it to get a bigger image). Imagine the following steps for constructing the image from the table: replace each column by a vertical line, which represents the whole range of values for "its" column. Then, for each row of the table (i.e., each record), draw a point on each of the axes representing its values, and connect all the points belonging to the same record with lines. The result looks like this:

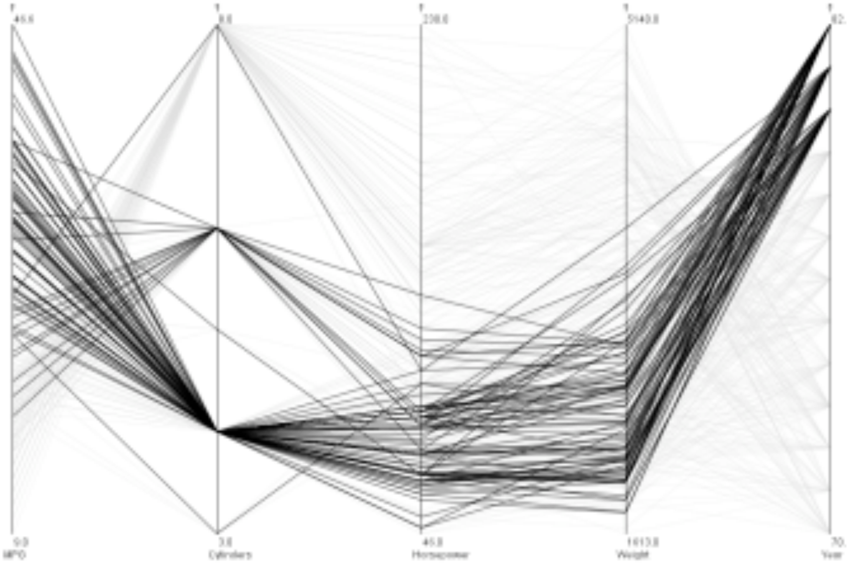


At first glance, this is just some line chaos. But when you look closer, you can see some structure already: there are axes with many different values, and axes with just a few. The *cylinder* axis only has five different values on it, and the *year* axis has a thirteen. For the others, you can get an idea of the distribution of the values (even though this is not very accurate, because lines can cover other lines). Especially on the *MPG* axis, you can see three large groups of values that seem to correspond with certain values on the *cylinders* axis.

But what makes parallel coordinates (and, in fact, most InfoVis techniques) useful, is interaction. You can zoom into parts of some of the axes, rearrange them, throw some of them out and bring other information in (e.g., the country of origin). Perhaps the most useful and most direct interaction is called *brushing*. The idea is that you mark certain values as interesting, and then look for other properties of the selected data. The following images show the results of brushing the above data set based on the year. In the left image, cars that were introduced from 1970-72 are brushed, while in the right image, the years brushed are 1980-82.



Brushing 1970-72



Brushing 1980-82

Even at first glance, the two images look quite different. On closer inspection, several interesting facts can be seen in the data. First, in the 70s (left image), the weight of cars was spread over a much wider range than in the 80s (right image): cars in the 80s were in the lower half of the weight range of the 70s. The same is also true of the engine power (*horsepower* axis). Looking at the MPG scale, you can also see that cars in the 70s had a much lower mileage than in the 80s (for Europeans: low values are bad here, because the MPG gives you the number of miles you can drive with one gallon of gas, as opposed to the amount of gas the car uses per kilometer).

An interesting detail is that in the 1980-82 range, there was only one car model with eight cylinders. If you follow the line from the *cylinders* to the *horsepower* axis, you can see that there is another line leading to the same value. Following that line back to the *cylinders* axis, we find a four-cylinder car. So the last remaining eight-cylinder had only as much power as one of the four-cylinders, and definitely needed more gas than that car (this is not really visible in this image without some more interaction). The eight-cylinder was also much heavier than the four-cylinder of the same power (this, again would need more interaction).

This was just a very simple example, but (hopefully) one that was easy to follow. InfoVis can in fact do much more, with larger data, more dimensions, and higher data complexity.

Why is InfoVis interesting?

InfoVis brings together several interesting aspects. First of all, it is graphical. That in itself is much more interesting than statistics ;). And that also means that much of the research and experience from perceptual psychology can be used to understand why some visualization methods are better than others. Examples are the Gestalt laws and preattentive vision: we see objects as groups and in certain constellations because of these phenomena.

Another field that is relevant to InfoVis are the visual arts. Visual communication was not invented by InfoVis people, and we certainly can learn a lot about how to use colors, etc. Some work has already been done in building new visualization methods on ideas from the arts, like the layering used in oil painting. Much more can still be done, though.

All this is not to say that InfoVis does not also pose technical challenges - even though this is a particularly weak spot of most InfoVis research (quite in contrary to volume and flow visualization). Especially when dealing with large datasets on the order of magnitude of one million items, it becomes crucial to design systems with speed in mind. To be really useful, InfoVis methods need to respond quickly to user input - only then, interactivity makes sense.