Visualization Analysis and Design

Tamara Munzner

July 18, 2016

1 Lecture 1

Disclaimer: taking notes for lectures which use slides is harder because I can't take advantage of the time the speaker writing on the board. So this will probably contain a lot less information than the previous one. Fortunately the notes by the speaker is available here.

Lets start with the definition of visualization. Computer-based visualization systems provide visual presentatinos of datasets designed to help people carry out tasks more effectively. When there is fully automated solutions, visualization is not very useful, but when we have an ill-specified problem, we need to use visualization to help us understand the problem better to build the model. It could also help developers and end users of automatic solutions to determine parameters / build trust. Visualization is a way to help us out of our jobs: towards ultimate automatic solutions.

But why do we use an external representation? If we have a table of numbers it is hard to keep track of them in our minds, but if we have a short-cut through the cognitive act of memorizing and analyzing numbers by plotting them, then we can recognize patterns and understand trends much more easily. We don't want to burden our cognition with low-level book-keeping, but save them for much higher level understanding, and that is where visualization comes in.

How does visualization help people do things faster? It summarizes lose information, details matter. We can put things together to confirm expected and find unexpected patterns, etc. The so-called Anscombe's Quartet shows 4 datasets with identical statistics, but look completely differently when plotted. These kind of situations are where it is very useful to show the whole dataset at the same time without losing details.

We can build an analysis framework by putting it under 4 levels and ask 3 questions. The first level is *domain*, i.e. "who are the target users"? There is a vocabulary for a domain (domain specific language), to communicate with the users of that specific domain. Then we come down to the next level which is *abstraction*. Here we want to abstract/translate the specifics of domain to the vocabulary to visualization. We don't usually draw exactly what we are given, but transform them to a new form. This is the question of "What is shown?". We also need to address "Why is the user looking at this".

The next level is *idiom*, i.e. "how to draw the visualization and how to manipulate them?". Then the last level is the *algorithm* level which is separate from the representation levels above.

We have different ways to get things wrong at any level. At the domain level, we could assume the needs of the audience wrong, e.g. assuming the same level of domain-specific knowledge as the speaker from the audience. At the abstraction level we could be showing the wrong data. At the visual level we could be showing things in a way that doesn't work, and finally we may have a code too slow to do the job.

How to do things right? We need to use methods from different fields at each level. At algorithm level we are talking about computer science, which is technique-driven work. At idiom level we are in the regime of design and cognitive psychology. At the higher levels we start to venture into the field of anthropology/ethnography. Here we are talking about problem-driven work.

Lets start by asking the question of *What*. There are three main types of datasets: tables, networks, and spatial (fields or geometry). Tables are easy to understand, with columns and rows. Networks are like graphs with nodes and links between them. Spatial data can come as fields, like a grid of positions with values associated to them. It could also be geometry, like a map.

There are also different types of attributes that we want to represent. We could have categorical data, or those with intrinsic ordering. With the latter we also have the question of which ordering direction (sequential/diverging/cyclic). In situations where the data is cyclic it could represent some interesting challenges.

Lets now look at the question of *Why*. We can split the question into actions and targets. On the action side, we typically want to analyze/query data. Analyzing is a way to consume data, to discover things or present them to others. We could also present new data by annotating older sets or by deriving new things from older things. Querying is changing the amount of data, to identify/compare things, typically reducing the amount of data to process in our head.

We can talk a bit more about deriving. We actually rarely draw things we are given directly! First we need to decide what the right thing to show is, create it with a series of transformations on the original data, and then draw that. A simple example is a graph of exports and imports, we could take the difference and plot the trade balance.

An example of analysis is to derive one attribute from a series of trees/networks, namely the Strahler number. Details in the slides, I don't want to summarize it on the fly...

Now lets talk about the targets. We might want to present trends, outliers, or features, where the last one is an all-encompassing way to say things for example quantitative...

Lets now talk about the important question of *How*. First thing we need to do is to encode data. There are various structures for visually encoding the data. To talk about this, we introduce marks and channels. Marks are geometric primitives, and channels control the appearance of marks, e.g. color, position, size, etc. We can use different channels to encode the same data to bring emphasis to the data, or we could add different channels to encode different content to squeeze in more information.

The different idiom structures of visually encoding the data are combinations of marks and channels. A histogram is using the vertical positions of lines to encode quantities. A scatter plot is using the vertical/horizontal positions of points, etc.

There are two main types of channels, one is magnitude, and the other is identity. The magnitude channel, like position, length, saturation, are suited at showing ordered attributes, whereas the identity channel (spatial region, shape, etc.) is suited at showing categorical attributes. We need to match the channel and the data attributes we want to represent.

Different channels have different effectiveness! We are very good at judging spatial positions, but not very good at judging saturation or luminance. We want to encode the most important attributes with the highest ranked channels. Spatial position ranks the highest for both magnitude and identity channels, so it should be the first thing we think about when we determine the way to encode data. One thing to keep in mind is that there are people who respond less well to color hues (colorblindness).

Where does the channel ranking come from? We have the Steven's Psychophysical Power Law: $S = I^N$. The perceived sensation of length has index $N = 1$, with area/depth at $N = 0.7/0.67$. There were visualization experiments done by Cleveland & McGills, and experiments using crowdsourcing.

In addition to the channel to choose, we need to consider how many usable steps we have at each channel. We need to have a channel with enough different levels to show the different data we need. We also can't use all the channels at once, because sometimes channels step on each other. Refer to slide page 32 for a comparison when we have channels that interfere with each other.

One thing we need to consider is "Popout". When we have multiple clues for popout the data basically presents itself, but when we need to do a conscious search on the data representation it takes a much longer time on our cognitive system to recognize the pattern.

We can represent grouping between features. We can use containment or connection, or just put related things close to each other.

One last word about representation is relative vs. absolute judgments. The perceptual system mostly operates with relative judgments, not absolute one. Therefore to represent the comparison of data, accuracy increases with common frame/scale and alignment. The ratio of increment vs. common background determines the ease of judgment.

Now lets to talk how to use space to encode each of the data types. The first is *tables*. Lets use the computer science words key and values to denote the table. The key is an independent attribute to uniquely index the items we want to look up. A simple table has one key, whereas multi-dimensional table has multiple keys. Values are simply the data we want to represent.

As an example, the scatter plot is used to express values only, with no keys. We simply represent two quantitative attributes on each dimension. It scales very well since we can fit in thousands of points on a plot. A bar chart has one key and one value. The key is a category attribute and the value is a quantitative attribute. It is good at comparing or looking up values. A line chart also has one key and one value, with point as the mark. Now the key can be a quantitative attribute.

When should we should bar charts vs. line charts? The answer depends on the type of the key attribute. For female/male key (categorical) it is better to use bar charts, and for a number key (quantitative) it is better to use line chart.

Lets consider heatmap. We have two keys, x-y position, and one value as color. The two keys can be categorical attributes, and the value is quantitative. It can be useful at finding clusters and outliers. However we are limited by how much display can resolve on the screen, and we are capped at $\sim 1\,\mathrm{M}$ items. We are however very limited in the quantitative attribute levels since we only have so many colors/regions to play with without interfering with each other.

We can combine different idioms together using for example scatterplot matrix, where we have a matrix of scatter plots put together. We can also have parallel coordinates that put different line charts together. Either of these we are limited to dozens of attributes.

We also have pie charts or polar area charts. In the former area marks with angle channel to represent data, and in the latter we have length channel for the area marks. Both are better replaced by bar charts. One argument for pie charts is to show parts-to-whole idea, but we could fix that by plotting a normalized stacked bar chart.

There are interesting cases where one could use radial orientation to represent data. One example is glyphmaps to represent cyclic data.

Let's talk about spatial data now. We often want to use directly the data we are given, since the data is encoded directly in the position they are given. For example the choropleth map is to color-code the different regions of a map to represent a quantitative attribute on a geographic geometry. One important issue here however is the normalization of data.

We could also have a topographic map, which is plotting a scalar spacial field on geographic geometry using contours. Similar idea is to plot isosurfaces or directly volume rendering which is to plot scalar

spatial field directly in a 3D space by rendering the isosurface of the value. Due to the 3D nature often we often use opacity as well as colors to encode the desired data.

There are a family of ways to show vector and tensor fields. The challenge is that there are many attributes to show per cell. We could have flow glyphs where we just show local directions. We could also have geometric flow that trace the trajectory of particles. We could have texture flow and feature flow. In general there is no single best way but it depends on the task we want to carry out.

The last data type is networks, which is when we add links to the dataset. One way is simply use node/lines to directly show the networked graph. This is good for exploring topology, locating paths and clusters, etc. It is tricky however to consider spatial positions and proximity semantics, since we may be conveying unintended information. The scalability of this kind of graph is limited by the number of edges. This kind of graph only makes sense when we have number of nodes N less than 4 times the number of edges $4E$.

We could also transform the network into a heatmap where we encode two categorical attributes with x-y positions and use color to represent link. However we need to train ourselves to recognize topological structure. The strength of this representation lies in the scalability.

We could also use a radial node-link tree. It is easy to understand topology and to follow paths.

This feels like a shitty essay to type out. I'm not going to take notes for the second part...