Choosing visual properties for successful visualizations
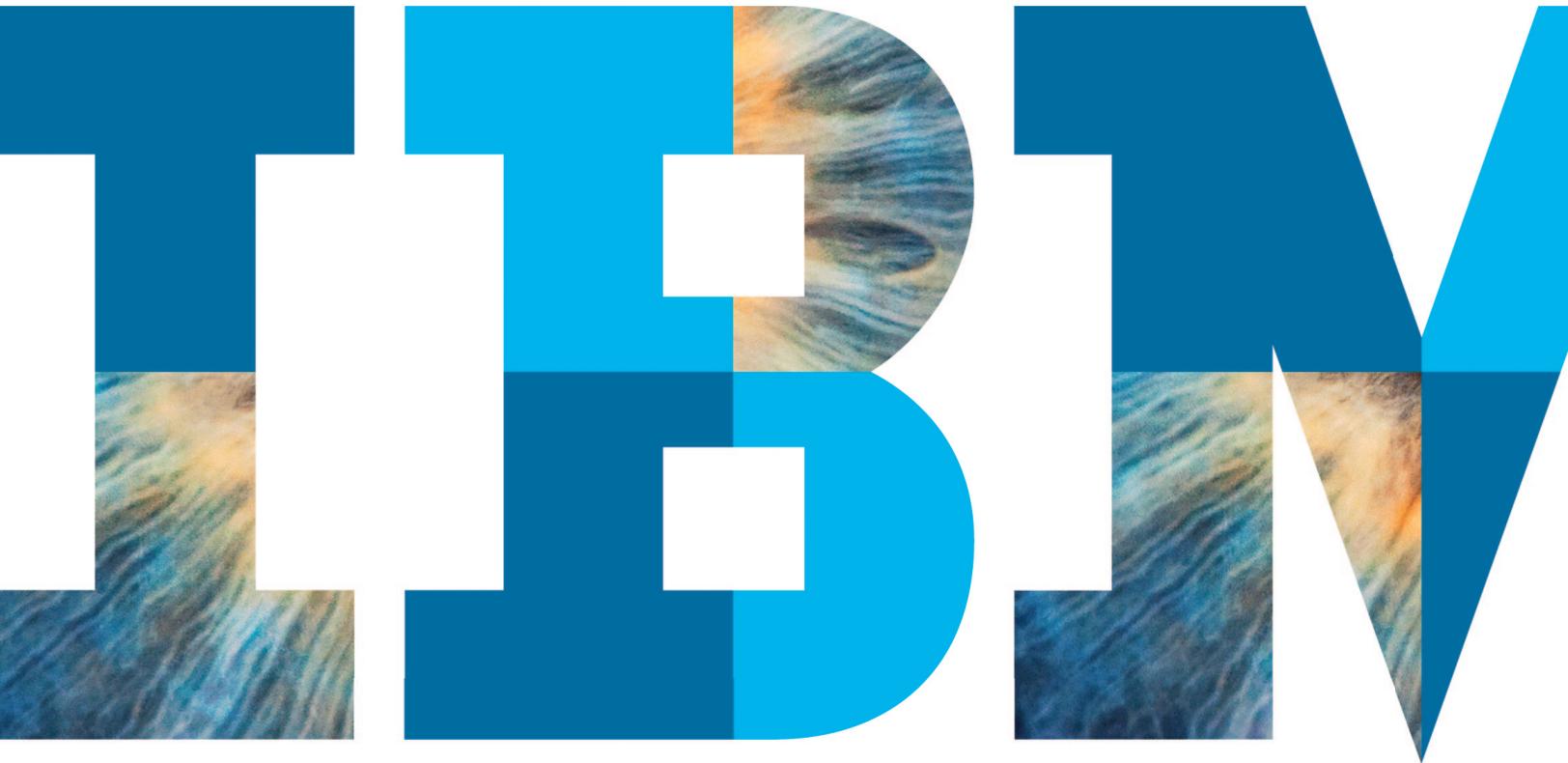
By Noah Iliinsky, IBM Visualization Expert

Contents

Introduction

The world produces more than 2.5 exabytes of data every day. Visualization is one key approach to gaining insight from this mountain of data. It enables you to see the trends and patterns (along with gaps and outliers) in the data that are not as easily identified in rows and columns of numbers.

Visualization can also provide access to huge data sets, such as weather, web traffic, sales and voting records. Data sets of this size have the potential to be overwhelming and inaccessible; a good visualization provides a way to explore, understand and communicate the data, along with actions the data indicate should be taken. For example, there are visualizations that show past prices for airline tickets and prices for future travel dates. From this sort of visualization, you can see whether ticket prices are trending upwards or downwards and you have guidance when considering tickets for different possible travel dates. A list of ticket prices wouldn't be nearly as compelling or useful as showing the price variations with graphs.

New tools now make it possible for all kinds of people to create visualizations to gain insight, illustrate a theory or tell a story. However, when you are creating a visualization, there are a huge number of design decisions to make, many of which have the ability to enable or inhibit the success of the visualization. This paper addresses one key aspect of the design process: how to choose an appropriate visual property (position, shape, size, color and others) to encode

the different types of data that will be presented in the visualization. Although this choice of encoding is often undertaken without much intention or deeper consideration, it has significant impact on the ability of the visualization to communicate knowledge accurately and efficiently.

Although encoding is often undertaken without much intention or deeper consideration, it has significant impact on the ability of the visualization to communicate knowledge accurately and efficiently.

The role of data types

Human brains and visual systems have evolved to be highly sensitive to specific types of visual stimuli. The brain interprets stimuli in different ways and with varying levels of accuracy. For example, human brains can accurately differentiate many different sizes, but only a relatively small number of line weights. They naturally consider lengths to be ordered (ranked), but not patterns. Some arrangements of lines imply groupings, whereas others imply hierarchies. Knowing how to use these properties effectively has a powerful effect on the accessibility and utility of a visualization.

Being familiar with the types of data you work with makes it possible to identify and discuss them relative to various visual properties. The types of data that you can encode in visualizations include categorical, quantitative (or numeric), ordinal, time, location and relationship.

Categorical

Categorical data consists of groupings of things that are alike, but not ranked, ordered or numbered. Think flavors, regions, teams, departments and other collections of similar things.

Quantitative or numeric

The numeric measure is the most important factor in quantitative or numeric data. Think dollars, units shipped, population and distance. Common exceptions include entities with numeric labels, where the number is more of a name or descriptor than an actual measure: minute rice, 700c bicycle wheels, 747 airliner, 12-hour cold medicine, zip codes and others. Other exceptions are entities that are sorted by number, but not actually measured (such as shoes).

Ordinal

This type of data is ordered or ranked, and the sequence matters, but there isn't necessarily a specific value. Examples include steps in processes or sequences of events. This type of data also includes classically ordered data, such as 1st place, 2nd class, 3rd floor, size 8 shoe, bronze medal, blue ribbon. Sometimes this data also has a numeric value; sometimes that actual value matters more than the sequence; sometimes not (for example, a 1st-place finish with a time of 19:43.8).

Time

This type includes anything measured in time: dates, durations, ranges and more. Time data may be treated like numeric data where the size of the interval between data points matters (such as placing events on a calendar or proportional timeline) or like ordinal data (ranked from fastest to slowest or last to first), where only the sequence matters, not the interval.

Location

Location data can be tricky, because it can be one of various other data types. If it's GPS locations, the data is numeric. If it's a list of zip codes or cities or landmarks, the data is categorical. If it's a combination of city, state and country, it can be categories arranged in a hierarchical structure. Street addresses are a messy mix of numeric data (building number), possibly ordinal data (5th avenue) and categorical data (street name, city and others) that you might be able to map onto something useful, such as GPS coordinates.

Relationship

This type of data indicates grouping, hierarchy, influence, correlation or other non-numeric interactions. Examples include military ranks, lists of requirements, influencing factors, flavors that pair well together, notes in a chord, family relations (cousin, grandmother) and more.

The encoding chosen to apply to the data must be interpreted by the brain in ways that are compatible with the way you want that underlying data to be interpreted.

Properties

Given these (occasionally ambiguous) data types and the multitude of visual properties available, how do you pair them in ways that are most effective? The key concept is *compatibility*: the encoding chosen to apply to the data must be interpreted by the brain in ways that are similar to and compatible with the way you want that underlying data to be interpreted. You should not use different shapes to represent quantity, or length to represent different flavors, or groupings to represent sequence, or huge fonts for unimportant things, or bigger to represent less good. However, you can use bigger to represent more bad; brains are subtle and tricky like that.[1]

The data types compatible with a given visual property are determined by two fundamental factors: the number of useful variations or values of that visual property that the brain can perceive and whether or not the brain interprets that visual property as naturally ordered.

Some visual properties have a very large number of useful values, some relatively few. You can differentiate a huge number of different locations; someone will be able to tell if you put one photo 2 mm to the right of the rest of them and will tell you it's not as far out of line as the one that's 4 mm to the right. On the other hand, the ability to accurately differentiate line thickness is very limited; beyond a handful of values, it's very difficult to tell if the line you are looking at is the 7th, 8th or 9th thickest. This is where the word "useful" comes in: you can create arbitrarily large numbers of different values of line thickness (or different patterns or subtle shades of color or incrementally adjusted angles), but if you can't perceive, differentiate or discuss those differences, they are not useful variations.

Some visual properties are naturally ranked or ordered in the brain. It's relatively easy to sort by length, thickness, darkness or angle, because these are all naturally interpreted as ordered. Ordering by icon shape, texture or the existence of enclosure simply doesn't make sense; they may be sorted into groups, but those properties are not naturally ordered in the brain.

When it comes to natural ordering, the trickiest property is color. Color may be ordered if you have a spectrometer, and there are social conventions around that ordering, but color is not considered to be ordered by the brain. Brightness and saturation are ordered, but hue is not. This might seem counterintuitive, but consider: is yellow bigger than blue? Is green more important than purple? These questions are unanswerable based solely on how the human brain interprets color. All color rankings are based on social convention: blue ribbon, red ribbon, white ribbon; yellow alert, orange alert, red alert. There is no ranking of color without social convention, and you cannot depend on social conventions when you care about an accurate and consistent interpretation of your visualization.

A summary of how these properties apply to various encoding is in the following table. You can also view the table at http://complexdiagrams.com/properties.

Example	Encoding	Ordered	Useful Values	Quantitative	Ordinal	Categorical	Relational
	Position, Placement	Yes	Infinite	Good	Good	Good	Good
1, 2, 3; A, B, C	Text Labels	Optional (alphabetic/numbered)	Infinite	Good	Good	Good	Good
	Length	Yes	Many	Good	Good		
	Size, Area	Yes	Many	Good	Good		
	Angle	Yes	Medium/Few	Good	Good		
	Pattern Density	Yes	Few	Good	Good		
	Weight, Boldness	Yes	Few		Good		
	Saturation, Brightness	Yes	Few		Good		
	Color	No	Few (<20)			Good	
	Shape, Icon	No	Medium			Good	
	Pattern, Texture	No	Medium			Good	
	Enclosure, Connection	No	Infinite			Good	Good
	Line Pattern	No	Few				Good
	Line Endings	No	Few				Good
	Line Weight	Yes	Few		Good		

Applying visual properties

After you understand the properties of various visual encodings, you can apply them to the different data types in your visualization in compatible ways. Here are some examples and explanations:

- **Position and placement.** With many naturally ordered and useful values, this property is limited only by the display size and resolution. Location is one of the most powerful encodings because it can be applied to nearly any data type. Consider using placement for your most important relationships or data.
- **Color.** This property is not ordered, and it has somewhere between 12 and 20 useful variations, after which you get into arguments about whether that's light blue, sky blue, robin's egg blue or cyan. The standard list of useful (highly distinguishable, easily described) colors is: red, green, yellow, blue, black, white, pink, cyan, gray, orange, brown and purple. Select from the first group of six before using values from the second group of six. Color is fantastic for labeling categories, terrible for representing quantities or ranks.
- **Pattern density.** This is a naturally ordered property, but it does not have a huge number of useful variations. You could say, "This is more dense than that," but not "This is close to 39 percent saturation by area, and that is more like 36 percent." Pattern density is good for smaller collections of ranked values and passable for numeric data where precision doesn't matter.
- **Enclosure and connection.** Enclosure and connection are not ordered (though *number of connections* might be), and they have many useful values. Similar to position, these properties are limited more by display size and resolution than anything else. They are excellent for hierarchy, influence, grouping (categories) and other similar functions.

- **Text labels.** Strictly speaking, this is encoding with language not with a visual property. Text labels are useful for differentiation and adding specificity to your data points. Add as many as you have room for, but not more, and understand that any extra information on the page becomes noise.

IBM Many Eyes: Bring your data to life

In the era of big data and analytics, visualization holds both prominence and promise as a means to help us discover insights in our data quickly and efficiently. To deliver on that promise, IBM has brought together its research, expertise and technology in the Many Eyes website, an IBM community that connects visualization experts, practitioners, academics and enthusiasts. With more than 100,000 members and thousands of visualizations, Many Eyes is IBM's portal for advanced visualization thought leadership that surfaces compelling insight from IBM visualization luminaries. As a member, you can visit the Many Eyes website to:

- Bring your data to life.
- Create visualizations in a three-step process.
- Learn how to develop effective visualizations.

To learn more and get started, visit **ibm.com/manyeyes**.

Example data set and visualization construction

An example of how to apply this information to the construction of a visualization starts with a data set that contains several data types: a list of the most deadly natural disasters in history.[2]

The data looks like this:

Type	Death toll	Event name	Location	Region	Date
Avalanche	20000	Huascarán avalanche	Peru	South America	1970
Avalanche	4000	Huascarán avalanche	Peru	South America	1962
Avalanche	265	Winter of Terror	Austria-Switzerland	Europe	1951
Avalanche	172	2010 Satang Avalanches	Afghanistan	Asia	2010
Avalanche	125	Kolka-Karmadon rock slide	Russia	Asia	2002
Avalanche	102	2010 Kohistan avalanche	Pakistan	Asia	2010
Avalanche	96	Wellington, Washington, avalanche	United States	North America	1910
Avalanche	90	Frank Slide	Canada	North America	1903
Avalanche	62	1910 Rogers Pass avalanche	Canada	North America	1910
Avalanche	59	1993 Bayburt Űzengili avalanche	Turkey	Asia	1993
Avalanche	57	1954 Blons avalanches	Austria	Europe	1954
Blizzard	4000	1972 Iran blizzard	Iran	Asia	1972
Blizzard	926	2008 Great Afghanistan blizzard	Afghanistan	Asia	2008
Blizzard	400	Great Blizzard of 1888	United States	North America	1888

The column headers and associated data types are:

• Disaster type: Categorical
• Death toll (estimate): Numeric
• Event name: Text
• Location (country, city or state): Subordinate category
• Region (continent or world): Superordinate category
• Date (time or time range): Numeric

Each of these different data types will require different considerations in the process of selecting an appropriate visual encoding.

The very first question is: what should be emphasized? The simplest approach is to list and display the disasters by number of deaths caused. Using the properties table, you can browse down the quantitative column and see that position, length and size are the best properties for encoding numeric data. (Angle and pattern density are also good for numeric data, but don't have as many useful values, so they are temporarily removed from consideration).

A very simple bubble graph, created on the IBM Many Eyes website and displaying only the death toll dimension with size (area) and conveying nothing else, is shown in Figure 1.
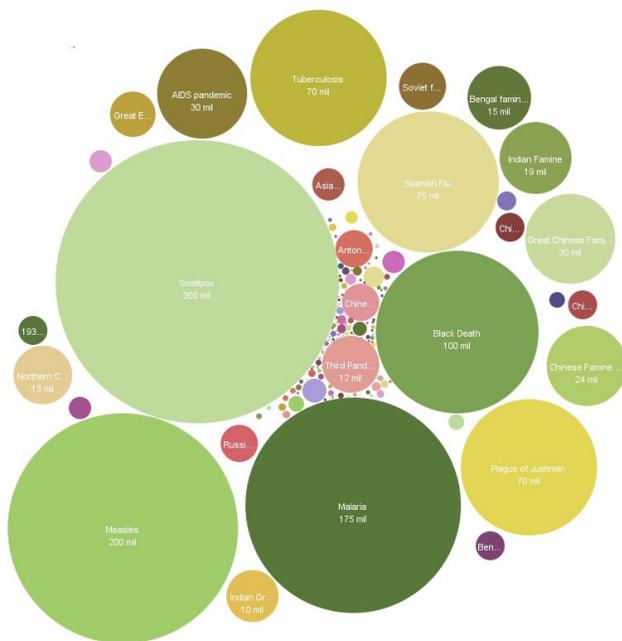

Figure 1: Bubble chart that shows the death toll of natural disasters. You may also view this graph at **ibm.com**/manyeyes/v/267467.

Each bubble has an area that corresponds to mortality and each bubble is labeled, but that's all that is represented.

So, what happens when you add a second data dimension (mortality being the first)? You get a bar chart. You can use length (or height against the y-axis) for mortality, and pick other data dimensions to represent other encodings. The simplest encoding option is to use the horizontal position along the x-axis to sort from greatest to least number of deaths, optionally grouping by type (Figure 2), region or single event. To get fancier, you could choose another dimension of categorical data that isn't already being displayed and use color to encode it.
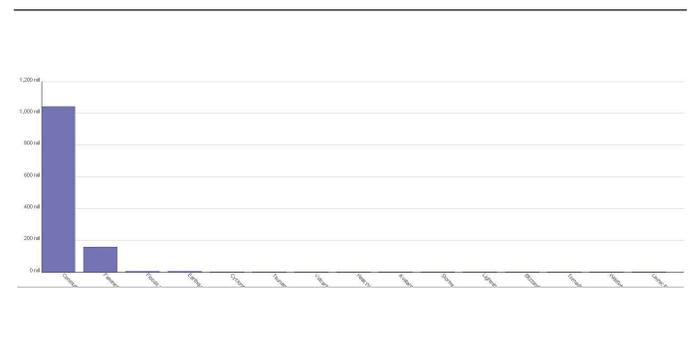
Figure 2: Bar graph that shows death toll and grouping by disaster type. You may also view this graph at **ibm.com**/manyeyes/v/267452.

Finally, you can choose to encode several of these dimensions all at once. It's quite reasonable to want to see one display that represents:

- Mortality of each given event
- Mortality of all events of similar type
- Mortality by specific location
- Mortality by region

To create this display, you should select good encodings for these four different data dimensions instead of two or three. And it gets more complicated than that. Single events are subordinate values of a given event type, and single locations are subordinate values of a given region, so there are hierarchical relationships to represent. All four of these stated data dimensions are numeric, so quantitative representations that can also be displayed hierarchically are required. This sounds challenging.

Fortunately, help is available. The properties table says that enclosure is good for display of hierarchies. It also says that area is good for display of quantities. When you combine these concepts, you can use enclosure of several related smaller areas to represent the accumulated values with a larger area, which enables you to show both the hierarchy and total values at each hierarchical level together. In practice, this kind of layout is called a tree map and looks like Figure 3.

This particular tree map implementation has a useful feature that enables you to rearrange the hierarchies so that you can easily highlight whatever matters most to you: type, region, specific locations and so on. It provides a compelling view of the data, clearly showing which disasters and individual types of events have been most deadly. For example, Figure 4 shows a closer look at the death toll from natural disaster events not related to disease and region.

Figure 4: Treemap drill down

Figure 3: Tree map that shows hierarchy and total values at each hierarchical level together. You may also view this tree map at:
ibm.com/manyeyes/v/267324.

Finally, if you had chosen to prioritize a view that revealed how these events had occurred at various points in history, with groupings of type and geography being secondary, you would end up with a graphic very much like an excellent example from the New York Times, which you can view at: http://www.nytimes.com/imagepages/2011/11/06/opinion/06atrocities_timeline.html

As you examine it, you can consider how many dimensions of data it displays and how each is encoded. For example, by one count, there are five visually represented data dimensions, and various types of information (rank, title, mortality) encoded in the text labels:

- Time in history, numeric, encoded by position
- Duration, numeric, encoded by length
- Mortality, numeric, encoded by area
- Type of event, categorical, encoded by vertically separate groupings
- Location, encoded by color

Conclusion

There are endless ways to visualize your data, so selecting and designing the right visualization is key to gaining insight from your data. Human brains and visual systems are highly sensitive to specific types of visual stimuli. Therefore, knowing how to use these visual properties effectively has a powerful effect on the accessibility and utility of your visualization. To create a visualization from complex, multidimensional data, you should familiarize yourself with the different types of data and visual properties. Then you should consider which of these data types are in your data source, decide which are most important to reveal and determine their affinities with specific visual properties. By choosing compatible encodings, you will improve your ability to communicate the important data effectively.

About IBM Business Analytics

IBM Business Analytics software delivers data-driven insights that help organizations work smarter and outperform their peers. This comprehensive portfolio includes solutions for business intelligence, predictive analytics and decision management, performance management, and risk management.

Business Analytics solutions enable companies to identify and visualize trends and patterns in areas, such as customer analytics, that can have a profound effect on business performance. They can compare scenarios, anticipate potential threats and opportunities, better plan, budget and forecast resources, balance risks against expected returns and work to meet regulatory requirements. By making analytics widely available, organizations can align tactical and strategic decision-making to achieve business goals. For further information please visit **ibm.com**/business-analytics.

Request a call

To request a call or to ask a question, go to **ibm.com**/business-analytics/contactus. An IBM representative will respond to your inquiry within two business days.
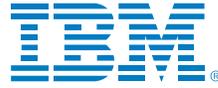
For more information

To learn more about IBM Many Eyes and to meet IBM's visualization luminaries, visit:

You can continue to build your knowledge of the progressing capabilities of advanced data visualization at IBM Many Eyes. On Many Eyes you can read additional visualization perspectives from Noah Iliinsky and other IBM visualization luminaries and create your own visualization in three steps. To get started, visit: **ibm.com**/manyeyes.

About the author

Noah Iliinsky is currently a Visualization Expert at the IBM Center for Advanced Visualization.

Noah strongly believes in the power of intentionally crafted communication. He has spent the last several years thinking, writing and speaking about best practices for designing visualizations, informed by his graduate work in user experience and interaction design. He is a frequent speaker in industry and academic contexts. He is the co-author of *Designing Data Visualizations* and technical editor of and a contributor to *Beautiful Visualization*.

1   For more about compatibility see Kosslyn, Stephen M., Graph Design for the Eye and Mind, Oxford University Press, 2006, pp. 14-16

2   The source of this data can be found at:
http://buzzdata.com/datafacts/natural-disasters#!/data.

Please Recycle