perceptual edge

Solutions to the Problem of Over-Plotting in Graphs

Stephen Few, Perceptual Edge
Visual Business Intelligence Newsletter
September/October 2008

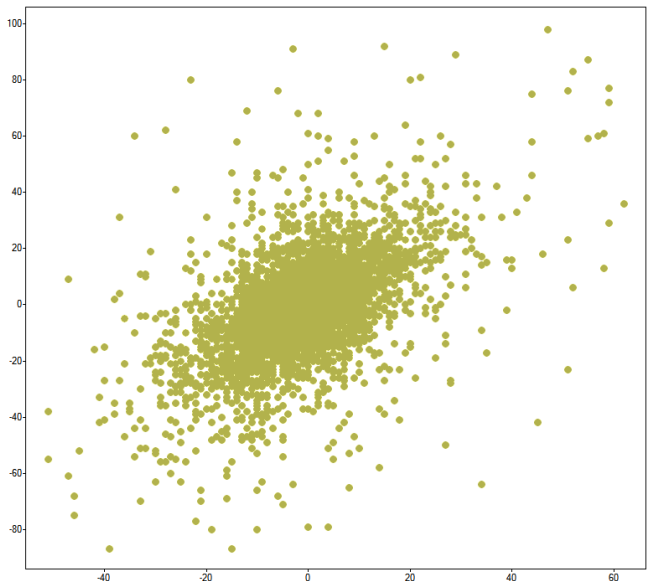
[This is an advance excerpt from the new book *Now You See It* by Stephen Few, scheduled for publication in March, 2009.]

In some graphs, especially those that use data points or lines to encode data, multiple objects can end up sharing the same space, positioned on top of one another. This makes it difficult or impossible to see the individual values, which can undermine analysis. This problem is called *over-plotting*. When it gets in the way, we need some way to eliminate or at least reduce the problem. The information visualization research community has worked hard to come up with over-plotting reduction methods. We'll take a look at the following six:

- Reduce the size of data objects
- Remove fill color from data objects
- Change the shape of data objects
- Jitter data objects
- Make data objects transparent
- Reduce the amount of data

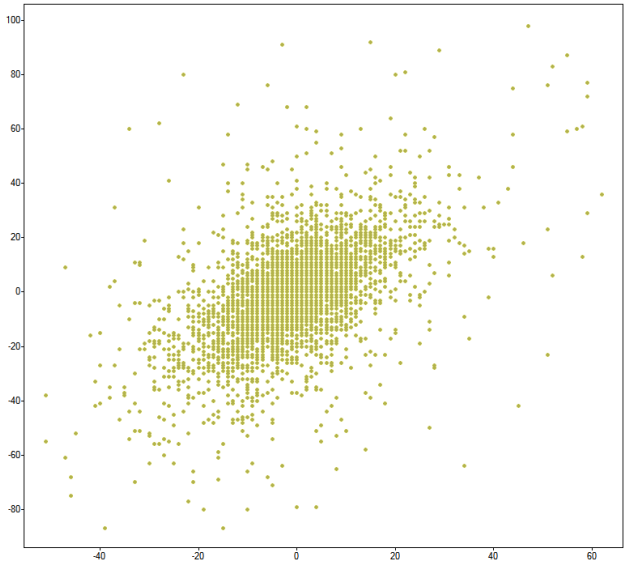
Reduce the Size of Data Objects

Consider the following scatterplot filled with data. Notice that in some areas multiple data points fight for the same location and sit on top of one another as a result.



[Created using Spotfire.]

Sometimes the problem can be adequately resolved simply by reducing the size of the objects that encode the data—in this case the dots. Here is the same data set, but the size of the dots has now been reduced.

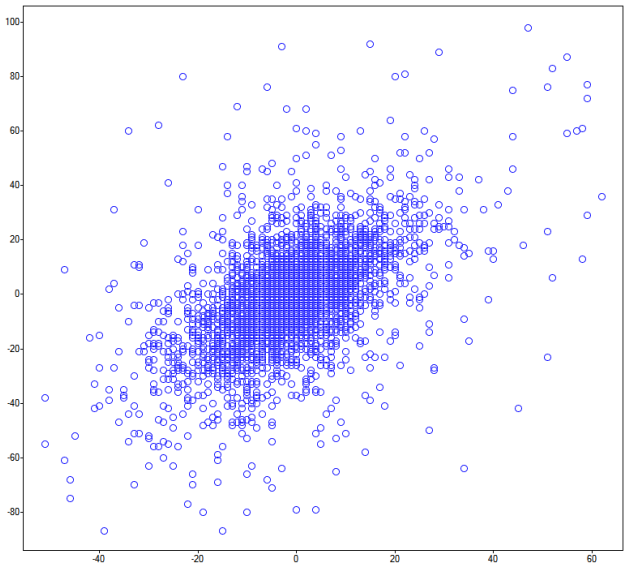


[Created using Spotfire.]

In this particular case, reducing the size of the objects alone doesn't overcome the over-plotting problem sufficiently, but we can certainly see a great deal more than we could before. However, there are still too many dots occupying some of the same locations. Nevertheless, when the problem of over-plotting is relatively minor, reducing the size of objects can often do the job.

Remove Fill Color from Data Objects

Another simple method that can be used in an attempt to remedy the over-plotting problem involves removing the fill color from the objects that encode the data to reveal how the objects overlap one another. In this next example, I have enlarged the dots a little and have removed the fill color from them. I also changed the color to something that stands out more clearly against the white background, even when the amount of color in each dot has been reduced by removing it from the dot's interior.

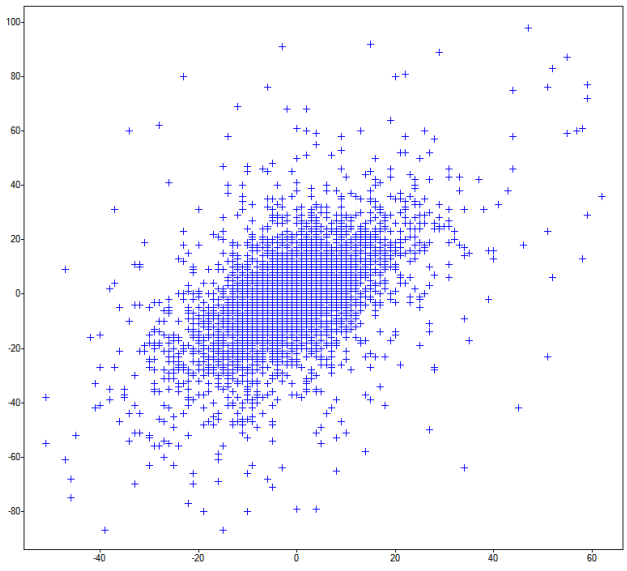


[Created using Spotfire.]

Once again, even though this method is often quite useful, in this particular case it hasn't done the job.

Change the Shape of Data Objects

Another simple technique that often helps involves changing the shape of the data objects from one that functions as a container with an interior (such as a circle or a square), which requires a fair amount of space, to one that isn't shaped like a container (such as a plus sign or an X).

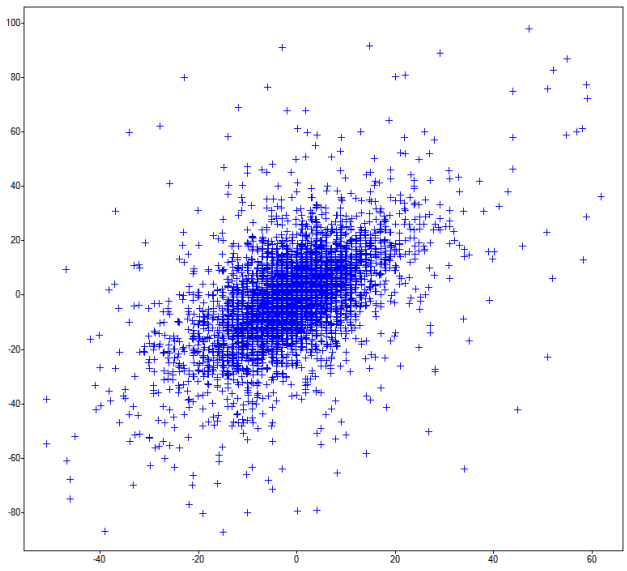


[Created using Spotfire.]

As you can see, however, while it often does the trick, this method does not reduce over-plotting when multiple objects encode the same exact value, because they continue to occupy the same exact space.

Jitter Data Objects

One of the best ways to reduce over-plotting when data objects have the same exact value is to change something about the data, rather than the appearance of the object that encodes the data. *Jittering* is a technique that slightly alters the actual values so they are no longer precisely the same, which results in moving them to slightly different positions. In the scatterplot below, compared to the one above, we can now see more detail in the cluster of values near the center. For instance, we can see that there are more values in the center of the cluster, which was not apparent in the previous example.

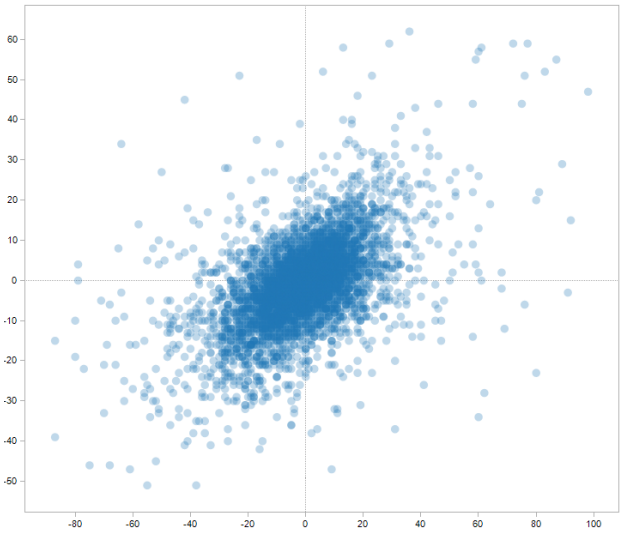


[Created using Spotfire.]

We have now made significant progress in reducing the over-plotting. It is much easier to see the differences in the density of data where before it was a mass of undifferentiated clutter. Jittering wouldn't be practical if we had to change the values manually. Fortunately, many good visual analysis products support jittering as a function that can be turned on and off as needed and can be adjusted by degrees, from slightly jittered to greatly jittered. If you are fortunate enough to use one of these products, just don't get carried away. If you jitter data too much, you will produce patterns that don't actually exist in the data.

Make Data Objects Transparent

A newer method, which in many cases works even better than jittering and does so without altering the values or changing the shape of the data objects, does the job by making the objects partially transparent. The proper degree of transparency allows us to see through the objects to discern differences in the amount of over-plotting as variations in color intensity. The following scatterplot allows us to easily detect differences between the dense center of the cluster, which is intensely blue, versus surrounding areas of progressively less concentration (less intensely blue). By using a slider control to vary the degree of transparency, a display that suffers from over-plotting can be quickly and easily adjusted to reveal nuance in the midst of clutter.



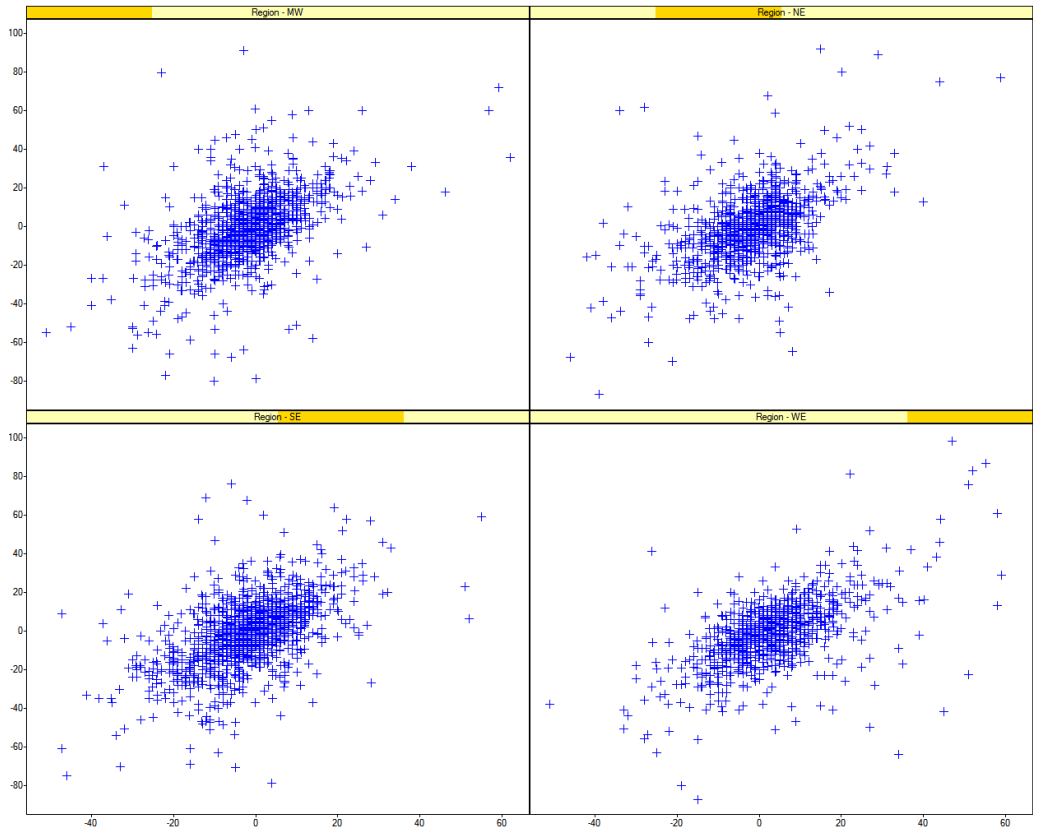
[Created using Tableau.]

Reduce the Number of Values

The remaining methods don't involve any changes to the objects that encode the data or to the values; they involve reductions in the number of values that are displayed in the graph. The four most useful methods of this type are:

- *Aggregate the data.* This can be done when we really don't need to view the data at its current level of detail, but can accomplish our analytical objective by viewing it at a more summarized level.
- *Filter the data.* This is a simple solution that can be used when some of the values that reside in the graph are not necessary—simply remove them.
- *Break the data up into a series of separate graphs.* When we cannot aggregate or filter the data any further without losing important information, sometimes we can reduce over-plotting by breaking the one graph into a trellis or visual crosstab display.
- *Statistically sample the data.* This technique involves reducing the total data set to a subset, using statistical sampling techniques to produce a subset that still manages to represent the whole. This is a promising method for the reduction of over-plotting, but it is relatively new and still under development at this time. Hopefully this will become a standard feature of visual analysis software not too far into the future.

Trellis and visual crosstab displays can often solve the problem quite easily. Here's the same data that we've been looking at in the previous graphs, this time broken into four graphs—one per region.



[Created using Spotfire.]

The software that you use for data analysis might not support all of these methods for the reduction of over-plotting. Few do. In time, however, I have little doubt that these methods will reside in all of the products that survive the increasing and justified demands of customers, plus other methods that have yet to be developed.

Information visualization software should support over-plotting reduction in the following ways:

- Provide the means to easily change the size of data objects, such as through the use of a slider control.
- Provide the means to remove the fill color from data objects that have interiors, such as circles (dots), squares, triangles, and diamonds.
- Provide the means to select from an assortment of simple shapes for encoding data points.
- Provide the means to jitter data objects, including a simple way to vary the degree of jittering.
- Provide the means to make data objects transparent.
- Provide the means to aggregate and filter data.
- Provide the means to break data up into a series of graphs in the form of a trellis or visual crosstab display.
- Provide the means to reduce the amount of data through statistical sampling.

About the Author

Stephen Few has worked for over 20 years as an IT innovator, consultant, and teacher. Today, as Principal of the consultancy Perceptual Edge, Stephen focuses on data visualization for analyzing and communicating quantitative business information. He provides training and consulting services, writes the monthly *Visual Business Intelligence Newsletter*, speaks frequently at conferences, and teaches in the MBA program at the University of California, Berkeley. He is the author of two books: *Show Me the Numbers: Designing Tables and Graphs to Enlighten* and *Information Dashboard Design: The Effective Visual Communication of Data*. You can learn more about Stephen's work and access an entire library of articles at www.perceptualedge.com. Between articles, you can read Stephen's thoughts on the industry in his blog.