Automated Annotations

Richard Brath*

Uncharted Software Inc.

ABSTRACT

Annotations on typical charts, graphs and maps can be used to draw a viewer's attention to a particular subset of an otherwise information dense display thereby aiding understanding. Automated annotations can generate insightful content on these dense display and a number of techniques can be used to identify data points and associated content to automatically promote into annotations on top of a visualization.

Keywords: Visualization annotation, automating commentary.

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 INTRODUCTION

Information visualization and visual analytics have traditionally been designed for use by people with familiarity with specific datasets. Instead, visualization intended for communication may need to provide understanding and insight to a community unfamiliar with the data of interest.

Even traditional representations such as line charts, maps, bar charts, and so on may be disorienting to a viewer if there are many data points. Viewers may comment "I have no idea where to look", or "There is too much competing for my attention."

Annotations is a broadly used term. In this paper, annotations are overlaid information, such as graphical markers (such as arrows and trend lines) and/or text (such as data values or commentary), on top of a visualization to add contextual information regarding the information in the plot, such as seen in Heer et al [1] or Tufte's layering and separation [2].

This is different than annotation where data patterns are automatically identified and recorded as additional data -- such as finding gene patterns in genetic sequences [3]; or a face detector in image processing software [4]). The output of these algorithms, however, can be used to add graphical annotations on top of e.g. a 3D gene viewer, or a photograph.

One tempting approach might be to simplify the data – for example, reducing a daily timeseries to a weekly timeseries to remove 80% or more of the data points while still retaining the macro data [5]. On the other hand, a Tufte-like approach would strongly suggest not removing data "What is sought is clear portrayal of complexity" [6, p191].

Instead, attention can be directed using annotations. A call-out, such as text and/or a marker, can help the viewer focus on a particular point or key message while the full detailed contextual data remains.

2 EXAMPLES

We have been involved in the design and implementation of annotation subsystems for eight different visualization systems.

Some of these have evolved to create (semi)-automated annotations:

- In some cases, the objective is to provide some useful insights when the visualization first appears.
- In some cases, the researcher is expected to publish some set of visualizations as part of the report. Some of these researchers are generating many reports per day: automated annotations provide a suggestion which can be deleted or modified to suit the larger narrative.
- In one case, the visualization evolved into a 24 x 7 live visualization in public space. Automated annotations provide new points for a viewer to focus on when passing the visualization throughout the day.

Over time, we have evolved some automatic approaches to determine which elements to annotate in fairly straight forward charts (e.g. line charts, bar charts), scatterplots, graphs and maps. These include:

High value/low value. The simplest annotations call out the highest and lowest points in a particular data representation.

There can be many variants on this pattern. In Figure 1 below, arrows indicate the highest and lowest values of national debt in the Americas, EMEA and Asia-Pacific over top a thematic map. Arrow base and text indicates the country, while arrow length indicates magnitude of the debt. Note that small countries do not have much presence in thematic maps, but these annotation will make small countries with high/low values visible (e.g. Ireland).

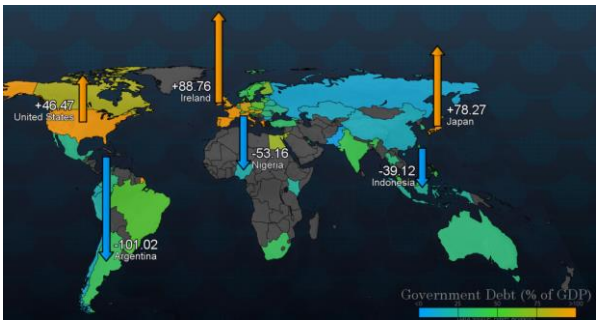


Figure 1: National debt by country, with highest/lowest annotations.

Outlier. A similar annotation is an outlier, which is not necessarily a high or low value. For example, an outlier can be determined by the distance to a moving average in a timeseries or a regression line in a scatterplot.

In the example in Figure 2, a single outlier for the yellow grid is determined by difference to surrounding cells.

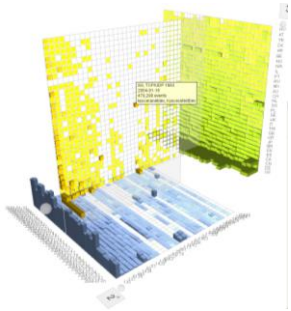


Figure 2: Outlier indicated via an initial sticky tooltip.

Reference points. In some datasets, a few well known data points can be used as a reference to indicate to the user datapoints that they already know to help them orient themselves using prior knowledge, such as large countries in a country dataset; or a waypoint along a race course.

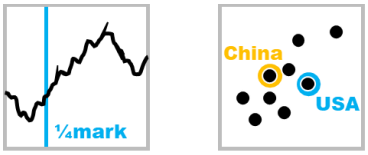


Figure 3 shows a set of 180 countries with a number of countries annotated.

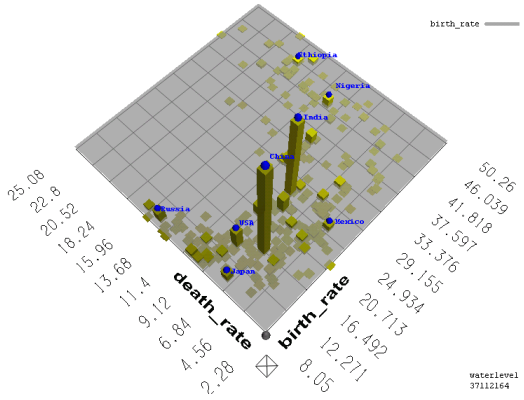
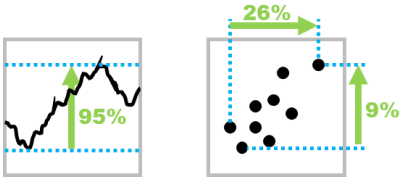


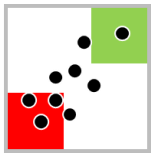
Figure 3: Scatterplot of country birth rate vs. death rate with high population countries annotated.

Range. Given that baselines on scatterplots, bar charts, line charts, etc may be non-zero, a user interested in the range of values is required to mentally compute the range whereas this can be made explicit with an annotation.

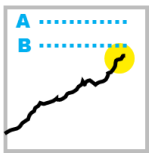


Trend. Trend lines can work well – they are familiar in scatterplots and can also be used in line charts and bar charts on ordered data series.

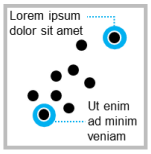
Data relative to plot area interpretation. In some types of plots, the desirable and undesirable areas of the plot may not be obvious, particularly if the viewer is unfamiliar with dataset and the plot type used. Explicitly calling out which parts of the plot and data are favorable or unfavorable (e.g. green and red boxes) aid the user in interpreting the semantics of the plot area.

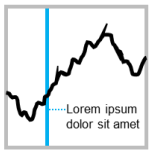


Round Numbers and Records: Data aligning to a round number (e.g. 100), approaching a threshold (e.g. pressure limit), or approaching a record (e.g. star athlete's all-time performance) may be of high interest. Explicitly identifying and labeling this threshold may have high value. Ideally, these should be configured by metadata or supplemental data.



Commentary: Explicit descriptive text may be generated to explain some of these patterns, or the dataset may have long text strings, such as document titles, headlines or descriptive text which provide much more useful information than simply indicating the values explicitly represented by the axes. In this case, the addition of arrows or leader lines may be required to locate the block of text in a relatively open portion of the plot with the a leader line or other means of associating the text with the target datapoint.





As indicated at the beginning of this section, we have implemented these in different types of applications such as automated 24 x 7 information walls, interactive publications and authored research. Any of the above can be used in together with other content, such as a narrative cross-reference and inserted into a narrative flow – future work could consider techniques for automating which annotation technique to use based on the narrative text.

For information displays and interactive publications, animation and/or interaction can be added to introduce and remove the annotations. These can also be triggered by interaction with narrative content, such as scrolling through a research article.

3 Conclusion

The notion of automated annotations is an area worth further investigation and this paper only introduces the topic and suggests potential techniques. The design space for automated annotations is likely much larger than the items discussed here. Furthermore, the approach here has the potential to generate varying annotations, but doesn't provide a framework for weighting and choosing amongst many possible annotations within a given display.

References

[1] J. Heer, , F. B. Viégas, and M. Wattenberg. "Voyagers and voyeurs: supporting asynchronous collaborative information visualization." In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1029-1038. ACM, 2007.

[2] E. Tufte. *Envisioning Information,* Graphics Press, Cheshire, CT. 1990.

[3] A. Marchler-Bauer and S. H. Bryant. CD-Search: protein domain annotations on the fly. *Nucleic Acids Research*, 32, 2L W327-331. 2004.

[4] L. Bourdev and J. Malik. Poselets: : Body Part Detectors Trained Using 3D Human Pose Annotations, In Computer Vision, 2009 IEEE 12th International Conference on, pp. 1365-1372. IEEE, 2009.

[5] J. Bertin. *Semiology of Graphics*. University of Wisconsin, 1993.

[6] E. Tufte. *Visual Display of uantitive Information,* Graphics Press, Cheshire, CT. 1983.