Automated Annotations

Richard Brath and Martin Matusiak*

Uncharted Software Inc.

ABSTRACT

Annotations on typical charts, graphs and maps draw a viewer's attention to a particular subset of an otherwise information dense display thereby aiding understanding. Instead of manual creation of annotations, we outline approaches to automatically generate annotations. We consider ways to identify potential annotations, such as significant data points; significant regions of the plot context; and associated commentary. We also discuss approaches to organize and sequence these annotations.

Keywords: Visualization annotation, automating commentary.

Index Terms: I.6.9.g Visualization techniques and methodologies: H.2.8.c Data and knowledge visualization.

1 INTRODUCTION

Information visualization and visual analytics are traditionally 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."

Annotation is a broadly used term. In this paper, annotations are graphical 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 data 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. This approach is essentially the Martini Glass narrative structure outlined by Segel and Heer [7] starting with an initial explanation followed with open-ended end-user exploration [8].

Automated annotations are similar to Natural Language Generation of news, whereby a natural language story is generated entirely from datasets. These systems face a similar challenge in "automatically finding the most pertinent meaning in a given dataset." [9].

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 a few cases, the objective is to provide some useful insights when the visualization first appears. This helps guide attention in an otherwise data-dense display.
- In others, a researcher is expected to publish a set of visualizations as part of a report. Some of these researchers are generating many reports per day. Automated annotations provide suggestions which can be adapted (or removed) to suit the larger narrative.
- Finally, some of the visualizations evolved into 24 x 7 live visualizations in public spaces. Automated annotations provide new points for a viewer to focus on when passing the visualization throughout the day. In this case, the data and annotations need to be continuously updated so that the visualization and commentary remain fresh and insightful throughout the day.

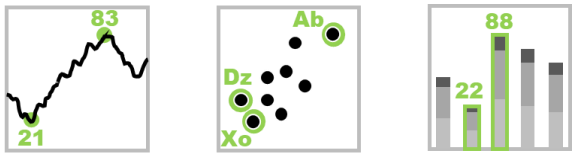
Over time, we have evolved automatic approaches to determine which elements to annotate in fairly straight forward charts (e.g. line charts, bar

charts), scatterplots, graphs and maps. In all our implementations, these annotations are embedded directly in the plot area of the visualization to help the viewer directly associate the annotation with the data. This is assumed to be better than cross-referencing between the plot area and a separate commentary area. These approaches fall into three main categories:

2.1 Significant data points

Identifying significant data points attempts to determine a small subset of highly salient data points in the dataset and/or the visual representation.

A. 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 annotations allow for small countries with high/low values to become visible (e.g. Ireland).

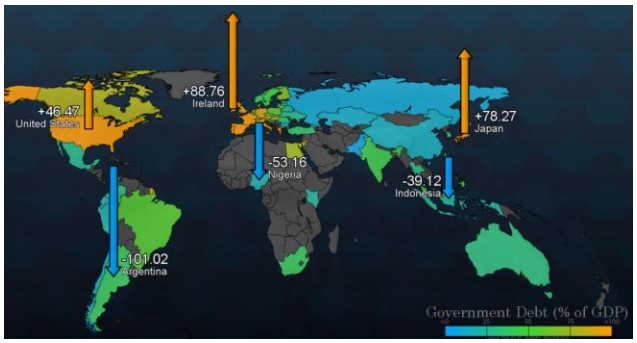


Figure 1: National debt by country, with highest/lowest annotations.

B. Outlier detection. 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.

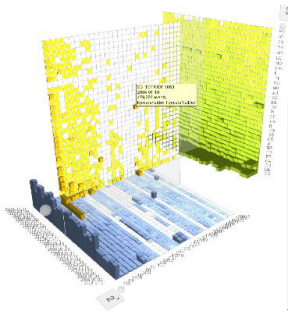


Figure 2: Outlier indicated via an initial sticky tooltip.

C. Reference points. In some datasets, a few well known data points can be used as a reference to act as landmarks to help users orient themselves using prior knowledge, such as large countries in a country dataset; or a specific date.

Figure 3 shows a set of 180 countries where countries with large populations annotated.

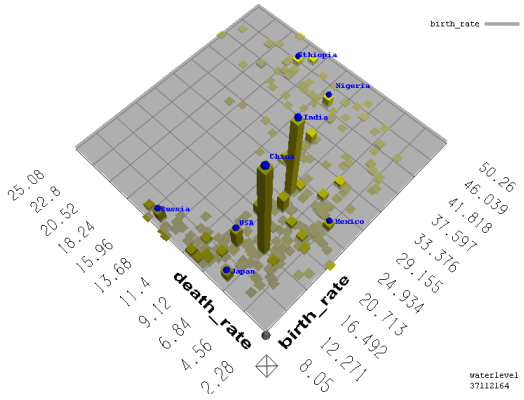
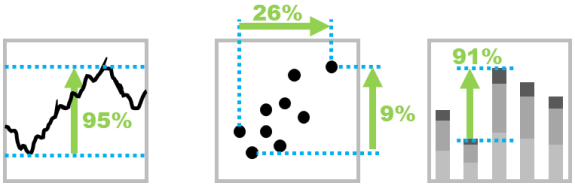


Figure 3: Scatterplot of country birth rate vs. death rate with high population countries annotated.

2.2 Significance across the plot

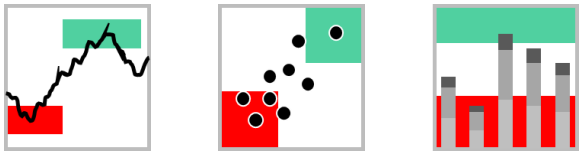
Trends across the broader dataset, and in relation to the plot area, can be identified and annotated on top of the plot.

D. Range. 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. This can be made explicit with an annotation, particularly useful if the range is different than the historical norm.

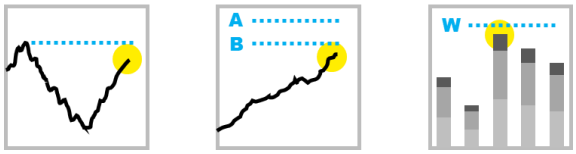


E. Trend. Trend lines can work well – they are familiar in scatterplots can also be used in line charts and bar charts on ordered data series. Similarly, averages, standard deviation and other simple statistics translate well into annotations.

F. 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 the 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. In the diagram below, the green explicitly indicates the region of the plot where the data values are in a desirable state.



G. Round Numbers and Thresholds: 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 historic data, metadata or supplemental data.



2.3 Associated commentary

Longer strings such as sentences or titles may make highly useful annotations:

H. Commentary: Explicit descriptive text may be generated (e.g. using NLG) to explain interesting patterns. Alternatively, the dataset may have long text strings, such as document titles, headlines or descriptive text which provide much more useful information than simply indicating values, regions or labels. These longer text elements may need to be located in a relatively open portion of the plot, with a leader line, arrow or other means of associating the annotation with the target data.

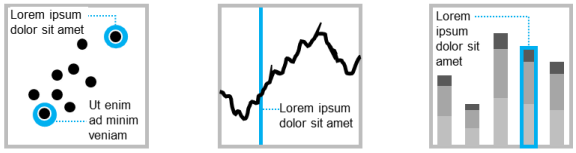


Figure 4 shows 500 stocks organized as a graph, (based on relationships), and color-coded by performance. Top performers in each quadrant are labeled with the headline for the most popular news story referencing that security.

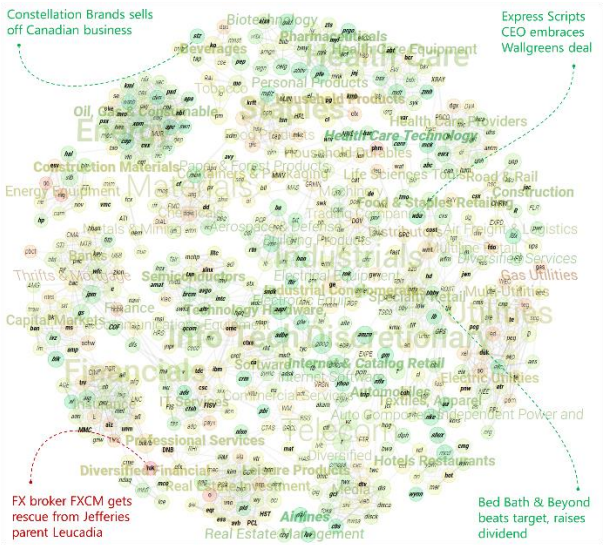


Figure 4: Graph of stocks with headlines for top performers indicated.

2.4 Broader narrative flow

These annotation elements can be used together to fit into a broader narrative, such as paragraphs of narrative text or animations:

Cross-references. 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 annotations can be cross-referenced in a broader narrative. For example, narrative text such as this paragraph can refer to annotations such as the red headline in Figure 4 or the tallest orange bar in Figure 1. Future work could consider techniques for automating which annotation techniques may work better with broader narratives.

Animation/Sequencing: For live information displays and interactive publications, animation can be used to dynamically add and remove the annotations. These can also be triggered by interaction with narrative content, such as scrolling through a research article.

In the example in Figure 5 below, we used the GDELT news headline data service to collect top global headlines every 15 minutes; plot news locations as a heatmap over a map; then divide the screen with a grid and label the most recent trending headline in each grid cell with a dot and headline text. This is still a very information dense display with up to 40 headlines.

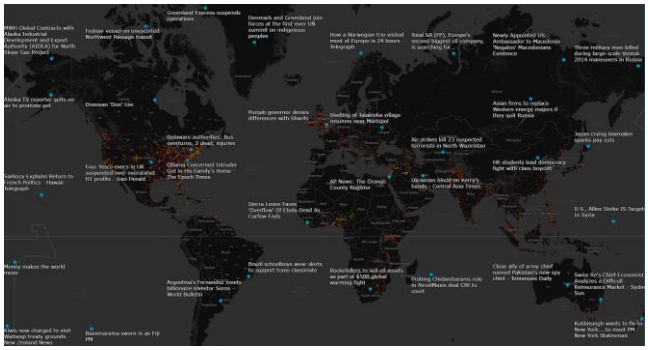


Figure 5: Map of top emerging global news headlines.

As part of a 24 x 7 ambient display, we then automatically zoom in to different continents, and then animate in the first paragraph (and photo if available) of the top two stories in that region (Figure 6). The viewer can touch any headline of interest to override the automation and display any story of interest.

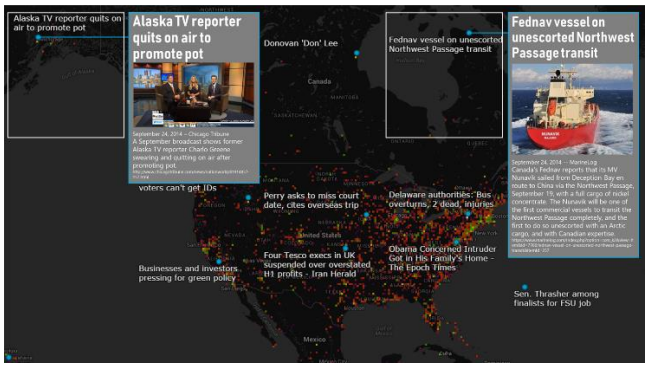


Figure 6: Zoom on North America with two stories animated in.

In the future these annotations could be sequenced with other introductory narrative elements, such as an explanation of the visualization technique or a tutorial pointing at live annotations rather than static data. For example, the BBC's interactive election map in Figure 7 has an overlaid video of a presenter walking through a few key features of the visualization prior to leaving the viewer to find their own insights [10]. Automated annotations could provide a logical progression in the narrative sequence between the introduction and unguided exploration.

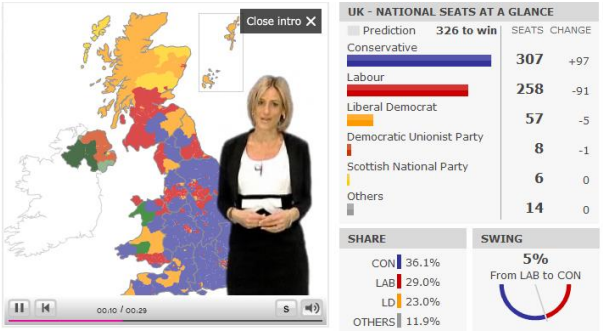


Figure 7: The animated host explains the visualization but leaves it to the viewer to explore to find any insights.

3 Conclusion

The notion of automated annotations is an area worth further investigation. 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.

[7] E. Segel and J. Heer. "Narrative visualization: Telling stories with data." IEEE transactions on visualization and computer graphics 16, no. 6 (2010): 1139-1148.

[8] A. Thudt, J. Walny, T. Gschwandtner, J. Dykes and John Stasko. "Exploration and Explanation in Data-Driven Storytelling." In *Data-Driven Storytelling*, pp. 77-102. AK Peters/CRC Press, 2018.

[9] A. Wright. "Algorithmic Authors", in Communications of the ACM, November, 2015.

[10] BBC.com. Election 2010 Results. URL: news.bbc.co.uk/2/shared/election2010/results/