Lessons from Pangloss: User Encounters with Uncertainty

Dominik Moritz
University of Washington
Seattle, WA 98195, USA
domoritz@cs.uw.edu

Danyel Fisher
Microsoft Research
Redmond, WA 98052, USA
danyelf@microsoft.com

Abstract
Pangloss implements "Optimistic Visualization", a method that gives analysts confidence to use approximate results for exploratory data analysis. In this position paper, we outline some ways in which user experiences with an approximate visualization system did not match analysts' intuitions. These observations have implications for the design of future systems that expose uncertainty to users.

Author Keywords
Uncertainty; approximate query processing; data visualization; optimistic visualization.

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation: UI;
H.2.4. Database Management: Systems

Introduction
Approximate Query Processing (AQP) is a database paradigm that allows users to exchange precision for query time on very large datasets: it returns rapid, uncertain answers to aggregate queries. This is invaluable for exploratory visualization systems, which struggle to deliver results quickly when the data is sufficiently large that precise database queries take too long to execute.

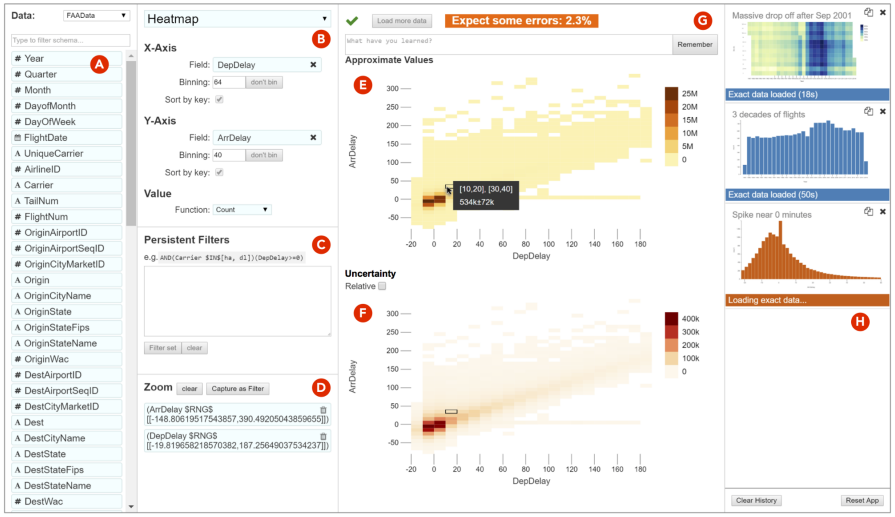

Figure 1: The Pangloss UI, exploring a flight delay dataset, with a list of draggable fields (A), which can be placed on the chart specification (B), filters (C), and zoom specification (D) to describe the chart. The chart shows both an approximate visualization (E) and a visualization of the uncertainty (F). The user may press the "remember" button (G) to store the chart in the history (H) and request an offline compute of the precise result. Two precise results are ready (in blue), while a third is loading (in orange).

In Pangloss [1], we approach challenges of interacting with approximate queries from a user experience perspective. Pangloss is a two-phase big data system: it allows users to explore their data through fast, approximate queries; users can then request precise responses with slow queries over the full data. In the first phase, the engine that drives Pangloss, called "Sample+Seek" [2], returns approximate results in interactive time with an overall uncertainty level.

In this paper, we summarize some of our insights with designing the user experiences for Pangloss and what we learned from analysts using our system. The more detailed conference paper describes the user experience and design of Pangloss; here, we describe some general issues and challenges we saw with users exploring data and working with an AQP system.

Brief Overview of Pangloss

Pangloss as shown in Figure 1, is an exploratory visualization system like familiar tools such as Tableau. Users can create 1D or 2D histograms and bar charts by dragging and dropping fields from the schema onto the chart specification forms. The system initially shows approximate visualizations; the analyst may request that Pangloss compute the precise query result in the background. The analyst can then open the view of the precise data to confirm or challenge their observation. This is the main idea behind *Optimistic Visualization*. It gives analysts the confidence that they will know if the approximation was significantly different from the precise result.

The Sample+Seek system, that is used to compute the approximations, is designed to be highly responsive for aggregation queries over a single table. It incrementally loads more rows into a sample until either the time to execute a query reaches a time threshold (typically a few

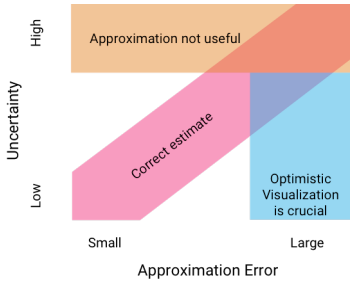

Figure 2: Relationship between uncertainty (estimated approximation error) and approximation error (true error of the approximation). When the uncertainty is high, analysts cannot draw useful conclusions from the results. When approximation error is larger than expected, approximation failed and analysts need to be informed.

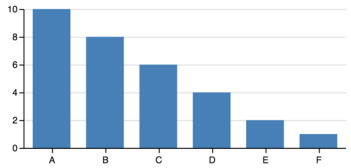

Figure 3: A bar chart with six groups. Without approximation "B, C, and D" is the same as "not A, E, and F". When the data is the result of a query over samples, the two statements are not necessarily equivalent.

hundred milliseconds) or the uncertainty has reached a low-enough value.

Uncertainty in Pangloss

Unlike many other systems, Pangloss uses *distribution uncertainty*. It is defined as the expected distance between the normalized distributions of the approximate answer and the precise one. Distribution uncertainty is different from familiar confidence intervals: it is a metric of uncertainty across *all* groups in the result. Using a distribution uncertainty captures the fact that uncertainties in different groups are not independent: the system instead claims that collectively, the errors are not large. One implication is that the uncertainty that could be drawn with confidence intervals is always higher or equal to the distribution uncertainty.

Understanding Approximation Error

From our first experiences with the prototype, we found that the uncertainty estimates are most effective when approximation errors exist but are reasonably-sized. The approximation error is the true difference between the estimated value and the precise data. When the estimate is very close to the true value, approximation errors are below the perceptual threshold: users will draw only correct conclusions from the approximation. This is, of course, the ideal case. In contrast, when approximation errors are large, analysts would draw incorrect conclusions from the data and would probably prefer to use precise results.

Of course, we cannot know the approximation error until after the precise answer has been computed. Uncertainty is meant to be a predictor of the approximation error: we would hope that high uncertainty would help see cases where approximation error is likely to be high. However, in a recent study, Agarwal et al. examined logs of 70,000 approximate queries from Facebook and found a large fraction had error estimates that were too wide or too narrow [3].

In Figure 2, we show the relationship between uncertainty and approximation error. When the uncertainty is large, analysts cannot make confident decisions, and so the approximation is less useful. The most dangerous area is the lower right of this chart, where the true error is large but the analyst expected small errors because the uncertainty was low. Optimism pays off in this area: those areas are prone to false negatives, where an analyst believes that their results are correct, even when they are not. Running a two-round query can assure analysts that they have a good result.

User Experiences with Uncertainty

Space limitations prevent us from articulating the issues we encountered in any detail. In overview, however, we found that using a sampling scheme meant that visual consistency was tremendously important: any navigation or modification of the view could cause samples to be recomputed.

For example, adding a filter might mean that the domain of values could change: a different sample might be used, which could have more groups. Users found these changing domains to be startling at first. This means, too, that filter in and out change in semantics. In a sample-based system, if I see columns A through F on screen, and filter out A & B, I might suddenly learn about "G", too! Thus, filters must be carefully written to be appropriately restrictive: do I want "everything except A-B" (which includes G), or do I want "only C-D-E-F"? (Figure 3).

Users had trouble choosing appropriate filters. For example, if a user wanted to see the precise query for "the top three items," they would sometimes filter down to the top three items in an approximate view. Unfortunately, those items might not be "top three" anymore when we compute the precise result.

Matching data domains. It is hard to directly compare two query results when the data domain has changed, which can happen as a more precise result may have additional groups or the order of groups may change. Thus, to create difference visualizations, we need to design visual cues to represent these changes. In addition, in binned charts, we need to stabilize the bins between rounds, ensuring that the binning offset and bucket size are reused for the precise version.

Importance of annotations. Pangloss allows the user to take notes on the observations when they "remember" a view. Our participants felt it helped them to remember what exactly was the thing they wanted to check. A richer vocabulary of annotations—not just textual, but linked to the visualization—might help users express what they care about in a visualization. These annotations can also help the software to know whether a more targeted query is appropriate and whether the changes in a precise result are relevant.

Conclusions

Approximate query processing systems could soon become a tool that data analysts use when exploring massive datasets. Trading off a bit of accuracy against a massive performance boost is promising but there remain many open challenges: current UI tools make assumptions that do not hold for approximations and data analysts are not used to working with uncertain data.

Researchers need to continue to investigate the user experience issues that analysts will face when AQP systems become widely available. We need systematic evaluations of uncertainty visualizations for complex visualizations such as heatmaps. In our uncertainty models, we need to consider error metrics that are not just per group errors but also distribution uncertainty or qualitative differences such as the probability of new groups appearing.

Working on these challenges requires close collaboration of user researchers with the visualization designers, database and statistical experts now working on SQL. These collaborations will be invaluable for the development of tools that best benefit the users of future AQP systems.

References

1 Moritz, Dominik, Fisher, Danyel, Ding, Bolin, and Wang, Chi. Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data. In *Proc. CHI* (Denver 2017).

2 Ding, Bolin, Huang, Silu, Chaudhuri, Surajit, Chakrabarti, Kaushik, and Wang, Chi. Sample + Seek: Approximating Aggregates with Distribution Precision Guarantee. In *Proc. SIGMOD* (San Francisco 2016).

3 Agarwal, Sameer, Milner, Henry, Kleiner, Ariel et al. Knowing when You'Re Wrong: Building Fast and Reliable Approximate Query Processing Systems.. In *Proc. SIGMOD* (2014).