Data visualization: ambiguity as a fellow traveler

Vivien Marx

Being sure is good; being uncertain is not necessarily bad. Research teams are working to render uncertainty visual.

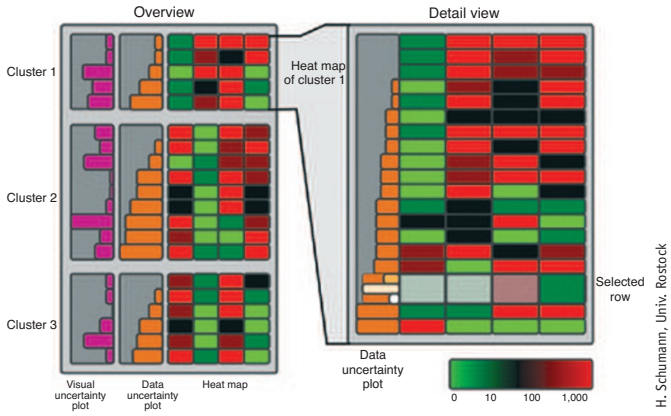
Data from an experiment may appear rock solid. Upon further examination, the data may morph into something much less firm. A knee-jerk reaction to this conundrum may be to try and hide uncertain scientific results, which are unloved fellow travelers of science. After all, words can afford ambiguity, but with visuals, "we are damned to be concrete," says Bang Wong, who is the creative director of the Broad Institute of MIT and Harvard. The alternative is to face the ambiguity head-on through visual means.

Color or color gradients in heat maps, for example, often show degrees of data uncertainty and are, at their core, visual and statistical expressions. "Talking about uncertainty is talking about statistics," says Martin Krzywinski, whose daily task is data visualization at the Genome Sciences Centre at the British Columbia Cancer Agency.

Statistically driven displays such as box plots can work for displaying uncertainty, but most visualizations use more *ad hoc* methods such as transparency or blur. Error bars are also an option, but it is difficult to convey information clearly with them, he says. "It's likely that if something as simple as error bars is misunderstood, anything more complex will be too," Krzywinski says.

He developed the visualization tool Circos, which places data in a circular layout that shows inter-relationships such as those among genes. It is a high-level visualization tool that works well in genomics, says Todd Smith, a research and application scientist at PerkinElmer. But, says Krzywinski, "Circos doesn't have any specific ways to encode uncertainty."

Statistical uncertainty weighs heavily on visualization. Every data point has uncertainty associated with it, Krzywinski says. Adding those statistical data to visualizations can quickly overload them. Despite the potential pitfalls of including uncertainty, the visual cues can remind scientists of their data's ambiguity.



Adding visual features to gene expression heat maps allows visualization of different types of uncertainty in the heat map's individual rows.

Some researchers are working on visualization approaches to tame this ambiguity. "In terms of visualizing ambiguous, uncertain or contentious results, perhaps my favorite method is showing results from multiple experts side by side," says Jim Kent from the University of California at Santa Cruz (UCSC) who co-developed the UCSC Genome Browser, a visualization tool with which scientists navigate genomic information such as reference genomes and assemblies in progress. Another approach, he says, involves layering, in which less-certain findings are displayed behind more-certain data and drawn in lighter shades.

Alternatively, heat maps can give an overview of data quality—for example, DNA alignment quality—and might use a color gradient, says Smith, who cofounded a bioinformatics firm called Geospiza that PerkinElmer acquired in 2011. "You make things dark if they are lower quality and bright if they are higher quality," he says. Likewise, the quality of DNA base assignments from sequencing experiments is sometimes indicated with grayscale or color shading. "The darker the base, the more uncertain it is," Smith says.

Flavors of uncertainty

Uncertainty comes in many flavors. It can arise upon data capture, during analysis or during visualization. It may be due to missing, noisy or imprecise data or to filters that could skew calculations, or there may be too few data to begin with, says Heidrun Schumann, a computer scientist at the University of Rostock who studies uncertainty visualization in many research areas, including the life sciences[1].

Some visualization approaches try to bring in a global representation of uncertainty for experimental data. Each presentation format calls for its own representation of uncertainty, Schumann says. For example, scientists may choose to express uncertainty by building it into the way data is visualized, such as by using transparency: "The more see-through the data, the more uncertain they are," she says.

An alternative method is to show uncertainty with an image placed next to or near

data. "An additional image demands a bit more from the user," Schumann says. The advantage is that this second image offers information about the data separately, whereas a color gradient encodes information into the data. The choice of which method to use depends on the intended application, but pitfalls abound. "There is no free lunch in these things," she says.

One of the most challenging facets of uncertainty for scientists is visualizing which data are or may be missing. For example, says Smith, in a sequencing experiment, a team might set out to capture a genomic stretch with 60 million bases. The scientists obtain results and a statistical distribution of sequencing coverage across the genome. Some stretches might be sequenced 100-fold, whereas other stretches have lower sequencing depths or no coverage at all. The PerkinElmer visualization platform GeneSifter represents alignments across the genome with a gene list that is accompanied by thumbnail plots with density maps of base coverage for each gene. Clicking these thumbnails takes a user into a genome browser-like visual experience.

But after an alignment, scientists might find they have aligned only 50 million of the sought-after 60 million bases, says Smith. This ambiguity in large data sets due to missing data—in this example, 10 million bases—is a big challenge, he says. He sees opportunity for approaches that show researchers and clinicians what might be missing in their data. This element could help them judge how reliable the results are.

To address uncertainty, Schumann's team worked with researchers from Graz University of Technology to add features to gene expression heat maps that allow visualization of data and visual uncertainty in the heat map's individual rows. One feature visualizes the degree of data uncertainty inherent to the data themselves. Another feature represents visual uncertainty. "It shows where we have more data than we have pixels for in the visual representation," she says.

With large data sets that have millions of data points to be visualized, values pile on top of others, erasing or hiding one another. The lost data become visible only when examined at high resolution. Visual indicators tell the scientists to look more closely at a row "to see the information that had gone missing," she says. Without this flag, a researcher would have no cues for which data to examine more closely.

Tracking uncertainty

One team member on this project with Schumann, Alexander Lex, is building the idea of uncertainty into the data visualization platform Caleydo, a 6-year-old 'omics' data and pathway visualization tool he has co-developed[2]. It is a joint project between the Graz University of Technology, Johannes Kepler University in Linz and Harvard University, where Lex is currently a postdoctoral fellow in the Graphics, Visualization and Interaction group.

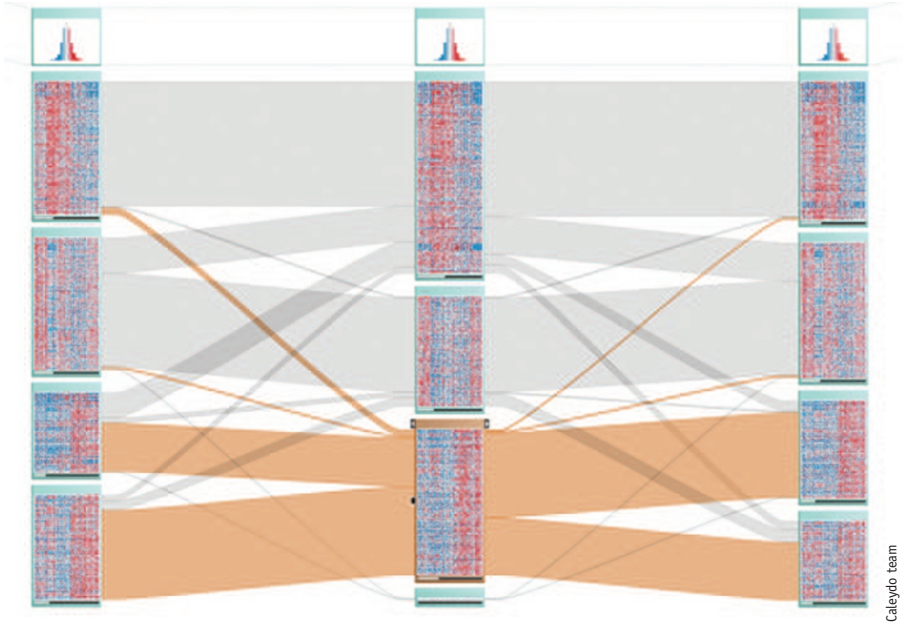
Lex collaborates with Nils Gehlenborg, a postdoctoral fellow in the Center for Biomedical Informatics at Harvard Medical School, who is part of the Caleydo team and also has a visualization and data integration platform called Refinery (http://refinery-platform.org/) in the works. A release for a small circle of researchers is slated in the summer, says Gehlenborg.

Both scientists are exploring ways to build uncertainty visualization into their platforms and, in particular, ways to highlight for scientists what might be missing in their data and their analysis steps. Information is often missing about where data originated and how they were processed, says Gehlenborg, who was interviewed jointly with Lex. "We keep track of where the data come from," chronicling the analysis steps and retaining the metadata, too, he says. His idea for Refinery is to allow scientists to toggle through their data and analysis steps. The software visually heightens awareness of what is missing at a given step, such as after a workflow has run through the open-source genomic analysis platform Galaxy.

The new visualization platform Refinery will keep track of analyses performed on a data set, says Nils Gehlenborg.

Refinery will keep track of analyses performed on a data set. Without such tracking, uncertainty information can get buried in analysis steps that cannot easily be teased apart, he says. In the rounds of data analysis, raw data acquire quality measures along with uncertainty scores, Gehlenborg says, which all need to be captured and propagated through the analysis pipelines.

For example, microarray analysis results from several probes deliver a signal, and these signals are aggregated into a single value for a gene displayed in a heat map. The expression level of each probe used to generate the overall gene expression level might be very different. "But that information gets lost, so that is, in a way, uncertainty," Gehlenborg says.



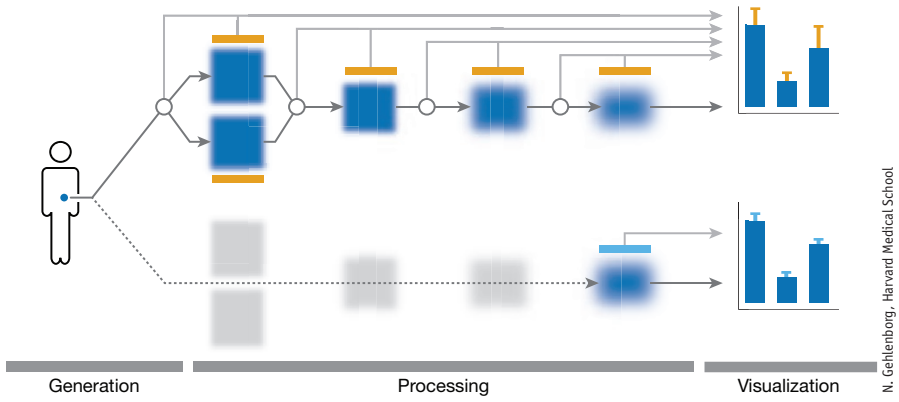
The visualization platform Caleydo has a functionality called StratomeX that reveals uncertainty factors in data analysis. Shown here, different clustering algorithms used to slice through the same brain-tumor gene expression data lead to different results. Algorithms: non-negative matrix factorization method (left), consensus hierarchical clustering method (center) and consensus hierarchical clustering method after manual curation of the data (right).

As scientists integrate, aggregate and summarize their large, heterogeneous data sets, the risk of losing data is great, says Lex, which is "the most critical part of uncertainty." His idea is to equip Caleydo at each stage with ways to visualize whether the performed aggregation and abstraction are trustworthy, he says.

For example, in 100 patients, scientists might collect data from a pathway with 50 genes for which gene expression data exist. The summary of these data, such as average gene expression across all patients, can hide crucial differences between patients, Lex says.

As scientists aggregate and summarize their large, heterogeneous data sets, the risks of ambiguity can rise, says Alexander Lex.

"When you compute these summary statistics, you end up with one value, which is nice for your visualization, but what you get rid of is essentially the variability across the patients," Gehlenborg says. It is all too common in labs to work with the summaries and the results, without a way to return to the raw data and all of their variability, which have scientific value.



The team developing the visualization platform Refinery (top row) is testing how to let users track uncertainty levels (orange) that arise in each data analysis step. Traditionally, (bottom row) uncertainty accompanies the tasks of integrating and summarizing heterogeneous data sets. The resulting errors are difficult to track.

Lost and found

As Lex explains, Caleydo users can drill through the levels of summarized results to see the variation in the data and understand what data might be absent. One Caleydo feature, StratomeX, combs different types of molecular data. The method helps to segment and stratify patients and is currently being used, for example, to slice data sets in The Cancer Genome Atlas, a consortium funded by the US National Institutes of Health, to sequence and analyze cancer genomes on a large scale.

The 'stratome', a term Gehlenborg coined, is how the Caleydo team describes the set of all stratifications. StratomeX allows explorations of different molecular stratifications, such as along a gene expression pattern for a set of genes. It can help discover and characterize subtypes within tumors, aligning different data types with each other. Not only can this approach reduce uncertainty, "it can show you uncertainty," says Gehlenborg.

Statistics itself can deliver a degree of uncertainty. Different clustering algorithms yield different results when applied to a growing data set, a common situation in consortia-led efforts when data are constantly being collected, says Lex. Although that result should not be surprising, it adds to the uncertainty about which clusters are ultimately trustworthy.

But, Lex says, applying two different clustering algorithms to the same data set can also lead to differing results. If the clustering results were "really true and certain," the clusters should not differ, he says. However, differences are noticeable between these analyses. Referring to two clustering algorithm results, he says, "It looks like this one clustering algorithm has a totally different opinion from the other one." Visual representations have to do justice to this uncertainty to indicate meaningful results and increase trust in the data.

The challenge with this variability sits in the math. For example, each clustering algorithm assigns patients to one cluster, but the data are actually large and multivariate. "Not every patient will fit into a single cluster," Lex says. Other approaches with a different statistical method, so-called fuzzy clustering, can assign patients to multiple clusters; the Caleydo team members are exploring ways to integrate this variety of statistical tools into their platform.

Finding missing data offers rewards for researchers working on visualization methods. Perhaps they may even find surprises, which could lead to a new kind of uncertainty metric. "One uncertainty metric might be 'How many published algorithms would disagree?'" says Krzywinski. "In biology, many!" A measure of uncertainty does not have to be a negative factor: it could be a measure more like a "surprise factor," he says. The less likely something is to happen, the greater the level of surprise. "We want to make sure that we're paying attention to the surprising results," he says, because they can be important indicators in experimental findings.

Visualization that integrates statistics in layered ways can help researchers use numbers to reach tangible, biologically meaningful results, says Lex. Visualization methods have to keep up with large data sets that are big, complex and noisy, but they cannot replace statistics. Visualization partnered with statistics stands to become a powerful part of biology-related data analysis, says Gehlenborg.

1. Holzhüter, C. *et al. Proc. SPIE* **8294**, 82940 (2012).
2. Lex, A. *et al. Comput. Graph. Forum* **31**, 1175–1184 (2012).

Vivien Marx is technology editor for *Nature* and *Nature Methods* (v.marx@us.nature.com).