studies and success stories, and the strengthening of the role of toolkits.

4.1 Repositories of data and tasks

One way to improve evaluation is to create benchmark datasets and tasks. Sample datasets have been made available (e.g. for graphs and time series) but user testing requires benchmark tasks as well. To promote this approach we organized the InfoVis contest [21]. Our goal was to initiate the development of the benchmarks, establish a forum to promote evaluation methods, and also create a new interesting event at the conference. The first contest took place in 2003 and the 2004 contest is now underway [22]. We invited submissions of case studies of the use of information visualization for the analysis of tree structured data, in particular to look at differences between pairs of similar trees. Three pairs of datasets were provided is a standard format, along with a taxonomy of general tasks (about 40 tasks in 11 categories). For each dataset the application domain of the dataset was described (phylogenies, taxonomies and file systems, with about 60, 200,000 and 70,000 nodes respectively), and open-ended domain specific tasks were provided to guide the analysis. After five months we received eight entries (a small number, but satisfactory for a first year). The main finding was that it was difficult to compare systems even with specific datasets and tasks. We had hoped to focus the attention of submitters on tasks and results (insights), but the majority of the materials we received focused on descriptions of system features. Little information was provided on how users could accomplish the tasks and what the results meant, making it very difficult for the judge to compare. The systems presented were extremely diverse, each using different approaches to visualize the data. Each tool addressed only a subset of the tasks, for a subset of the datasets. The phylogeny, which consisted of a small binary tree was not used, probably because the tasks were complex and required working with biologists (i..e. chosen to be realistic).

There were three first-place entries (see [21, 23] for more information on all entries). TreeJuxtaposer [24] submitted the most convincing description of how the tasks could be conducted and results interpreted. Zoomology [25] demonstrated how a custom design for a single dataset could lead to a useful tool that addressed many of the tasks satisfactorily. InfoZoom [26] was the most surprising entry. This tool was designed for manipulating tables and not trees. However the authors impressed the judges by showing that they could perform most of the tasks, find errors in the data and provide insights in the data. The three second-place entries showed promises but provided less information to the judges on how the tasks were conducted and what the results meant. EVAT [27] demonstrated that powerful analytical tools complementing the visualization could assist users to accomplish their tasks. Taxonote [28] demonstrated that labeling is an important issue that makes textual displays attractive. The submission from Indiana University [29] illustrated the benefits of toolkits by quickly preparing an entry combining several tools, each accomplishing different tasks. All entries were given a chance to revise their materials after the contest. We required participants to fill a structured form with screenshots and explanations for each task. That information is now archived in the Information Visualization Benchmark Repository [23].

With the InfoVis2003 contest we attempted to provide real data and tasks while trying to narrow the problem to one data type (trees) and three representative tree types. The contest taught us that the problem was still too large for a contest and that the vague nature of the tasks made it impossible to compare answers effectively. Our next step - the 2004 contest - will only have one dataset, much fewer tasks and a more structured reporting format. Nevertheless, we anticipate that the open-ended nature of realistic tasks and the diversity of approaches will still make judging a challenge. The contest also illustrated the difficulty of presenting convincing evidence. Demonstrating the power of a tool is difficult. Researchers are trained to describe their tools' novel features more than illustrating them with convincing examples using real data.

Contests are an artificial testing situation where the opinion of judges reflects the quality of the submitted materials, opposed to the actual merits exhibited when tools are tested interactively and discussed with designers. The impact of contests may be limited but the datasets and tasks remain available after the contests. They can be used by developers to exercise their tools and identify missing features, and by evaluators to enrich their testing procedures with complex tasks. We hope that more specific lists of tasks can be added to the repository and used in controlled experiments.

4.2 Case studies and success stories

Case studies report on users in their natural environment doing real tasks. They can describe the entire discovery process, collaborations among users, the frustrations of data cleansing and the excitement of seeing the first overview of the data. They can report on frequency of use and benefits gained. The disadvantage is that results may not be replicable and generalizable. Case studies have the potential to convince users and managers that information visualization is a legitimate investment. They can be extremely convincing for potential adopters working in the same application domain as the one addressed by the study. Others will need to be able to extrapolate the results to their domains and imagine how the tool could be useful for them.

Case studies may use entirely different measures of success that the ones traditionally used in user studies. E-commerce criteria might the percentage of completed sales, not time or error. In the example of Figure 1 Peets Coffee goals are to sell coffee and tea. What matters is that users complete the order and come again. User satisfaction remains a likely indicator of success, hopefully correlated with order completion. What makes Peets coffee's interface more satisfying remains unverified, even though we may have some hypothesis... Users might feel more in control because they can review all the choices at once? The novelty of the interface might entice expert computer users who are also more able to work thru the interface? Longitudinal studies will be helpful.

The recording of usage data with appropriate privacy protections is a powerful method to glean information about user behavior. It can inform designers about the frequency of feature use, guide screen layouts to speed interaction, pinpoint possible sources of dissatisfaction and inform the revision of help materials. It can also provide evidence of success when users