Supplemental Material:
An Evaluation of Visualization Techniques to Illustrate Statistical Deformation Models

J.J. Caban, P. Rheingans, T. Yoo

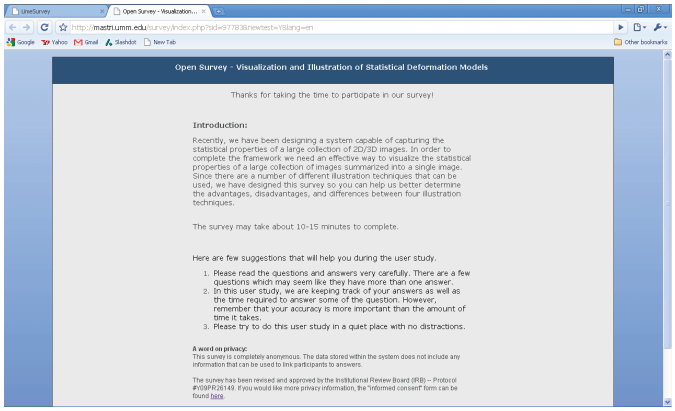


Fig. 1. A web-based user study was conducted to compare and measure the effectiveness of statistical illustration techniques. The first page our survey included a consent form explaining the purpose of the user study.

A User Study

This section provides additional information about the design of our user study, explanation about the datasets as well as sample questions.

A.1 Design

To design a user study that compared different statistical annotation techniques we began with an exploratory pilot study with three subjects. Our preliminary experiments were used to determine the specific information, instructions, and details most users needed to be able to understand each annotation technique and task. The test subjects listed questions that were difficult to understand, images that needed to be resized, and questions that needed additional explanation. All the feedback obtained from our test subjects was also used to guarantee that questions of different visualization techniques had the same level of difficulty.

After a long refinement process and working closely with our test subjects, an online survey was created and over 65 individuals were invited electronically to participate in our study. Note that any of the preferences, results, and comments obtained from our test subjects were not included in our final survey. The online survey was completely anonymous and presented the questions in random order to avoid any order and/or training effect. Figures 1 shows a screenshot of our online survey.

A.2 Structure

The user study consisted of three main sections: *Introduction, User Preferences*, and *User Performance*. First, the *Introduction* included detailed information and explanation about each individual annotation technique. The main purpose of this section was to teach subjects about each of the visualization techniques. Figure 2 shows an example of how *Deformation Grids* was introduced to participants.

Second, the *User Preferences* section captured information about the specific techniques subjects were more inclined to choose. In this section, five different datasets were annotated using each of the four illustration techniques (20 total images) and users were asked to grade each annotation. The preference metrics were captured using a Likert scale with five choices ranging from poor to excellent. Figure 3 shows a sample question about how preferences were captured.

Finally, the third section was the *User Performance* section which measured how well subjects can infer and determine characteristic properties of the input data based on a single annotation. Questions designed to capture the users performance did not provide any information about the original dataset used to generate a given annotation. In those questions, a single illustration was shown and users had to analyze the image and draw conclusions about the statistical deformation properties of the group. Figures 4-6 show sample questions from our user study. To further compare the differences between visualization techniques, the time required to answer each question and the user's confidence in the solution were also captured. All the questions within each section were displayed randomly to avoid any learning effect.

A.3 Datasets

The survey included a number of images, annotations, and illustrations from over 20 different datasets. In particular, four collections were used within the *User Preferences* section, 15 within the *User Performance* section, and two within the introduction. No dataset with the same deformation properties was reused in any part of the user study. A mix of synthetic and real-world medical 2D/3D images was used throughout the survey and within the three primary sections of the survey.

Some of the datasets used in the user study can be described as follows:

1. *Synthetic #1*: consisted of a set of 21 cubes deforming left with minimum deformation -20, maximum deformation of 0, and an average deformation of -10.

2. *Synthetic #2*: consisted of a set of 42 cubes deforming left and right with minimum deformation of -20, maximum deformation of 20, and an average deformation of 0.

3. *Synthetic #3*: a set of 12 X-shaped models deforming in four directions. The right half was deforming left, the left half was deforming right, the bottom half was deforming up, and the upper half was deforming down. The dataset had a minimum deformation of +8, a maximum deformation of +20, and an average deformation of 14.

4. *Synthetic #4*: a set of 21 rectangles with a circular cut-away region in the center deforming right and up, however the data was deforming right at a faster rate than was deforming up.

5. *Medical Images #1*: a collection of 9 segments of the spine from different subjects.

6. Medical Images #2: a longitudinal study with 14 MRI slices representing the deformations caused by a brain tumor.

Note that when the images of a synthetic dataset are presented in order, it could be fairly easy to determine the specific statistical deformation properties of the group. In our user study, any synthetic dataset presented in the *User Performance* section was not shown in order.

A.4 Participants

Once the final version of the study was completed, 65 individuals were invited electronically to participate in our survey. The online survey was accessed by 51 unique subjects. Eleven subjects did not complete the entire survey and their partial answers were not included in our analysis. The survey was not restricted to a particular age group or gender.

A.5 Sample Questions

To more effectively show the design and structure of our user study, this section presents a few sample questions that were included within our survey.

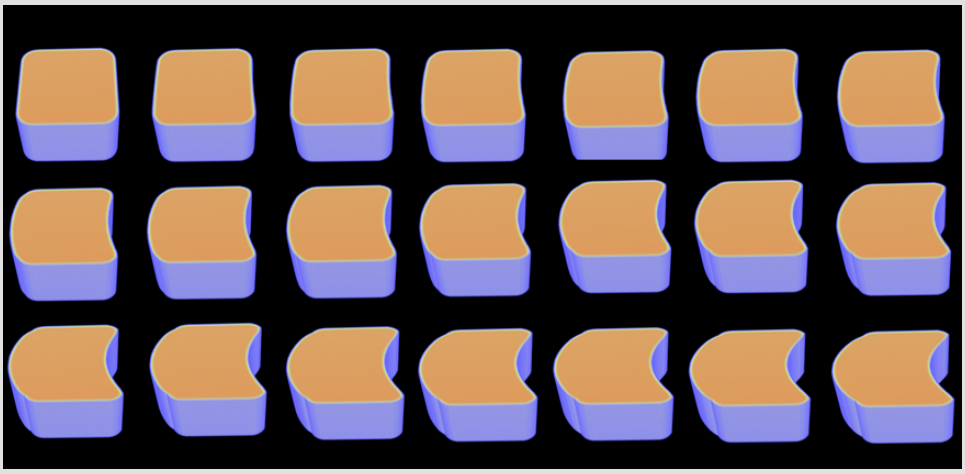
Figure 2 shows how *Deformation Grids* were introduced to participants. Each illustration technique was introduced by showing a dataset of images, some numerical information of the dataset, an illustration for the visualization technique under consideration, and an explanation of the resulting visualization.

Figure 3 shows a sample question used to capture user's preference. For each of the preference questions, a dataset was shown, numerical information of the dataset presented, and illustrations of the four primary annotation techniques were displayed. Users were asked to analyze the dataset and then analyze how each of the resulting annotations was able to capture the statistical deformation properties of the group under consideration. Users were then asked to score each of the illustration techniques.

Figures 4-6 present some of the questions asked in the third section of the survey where we captured user's performance. Each question consisted of an image or collection of images and a multiple-choice question. For instance, Figures 4 and 5 show annotations describing the statistical deformation properties of a collection of images as well as the questions that were asked about each individual illustration. Figure 6 shows a sample question that included real-world medical images. Note that each of the questions also asked the user to score his/her confidence in the answer.

Introduction: Deformation Grids

Instructions: Take a look at the collection of images below. Note how the cubes are deforming to the left.



General Information	
Number of Images	21
Deforming	Left
Minimum Deformation	-20
Maximum Deformation	0
Average Deformation	-10

Explanation:
Given a collection of images like the one above, we would like to generate a SINGLE image that captures the deformation properties of that group. One technique is to generate a ”deformation grid”. The general idea of deformation grids is to overlay a grid pattern over the image, distort straight lines to show the average deformation of the group and create a wavy pattern within the lines to show variability.

For example, in the image below we can see that vertical straight lines are distorted (bended) to the left to show the average deformation of the group (-10), the wavy patterns represent the variability of those regions, and the horizontal straight lines show that no deformation occurs from the top or bottom of the image.

Illustration Technique: Deformation Grids

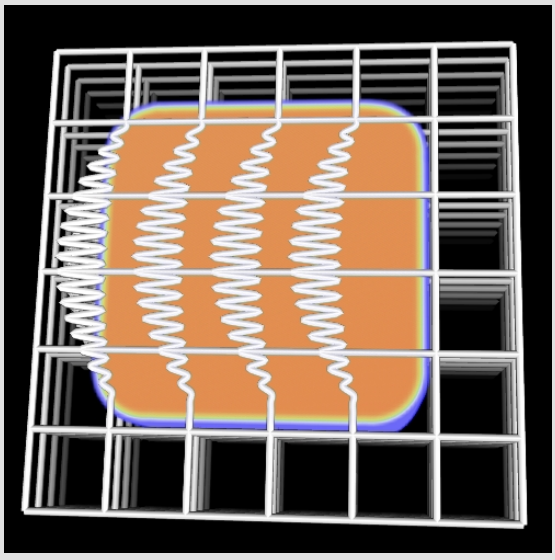
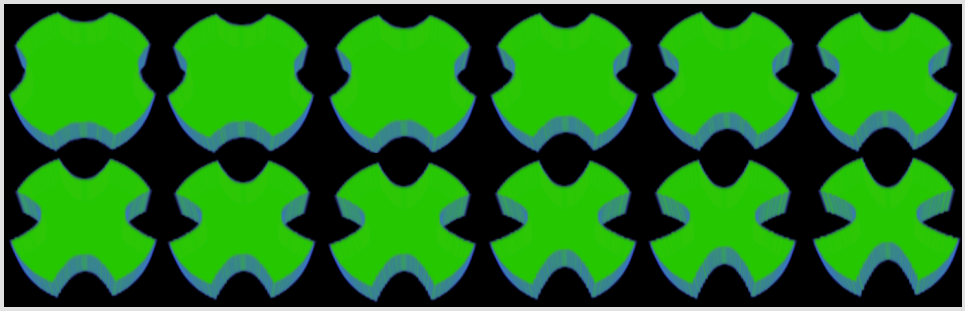


Fig. 2. The first section of the user study introduced each of the visualization techniques by showing a sample collection of images, the corresponding illustration, and an explanation of the resulting visualization.

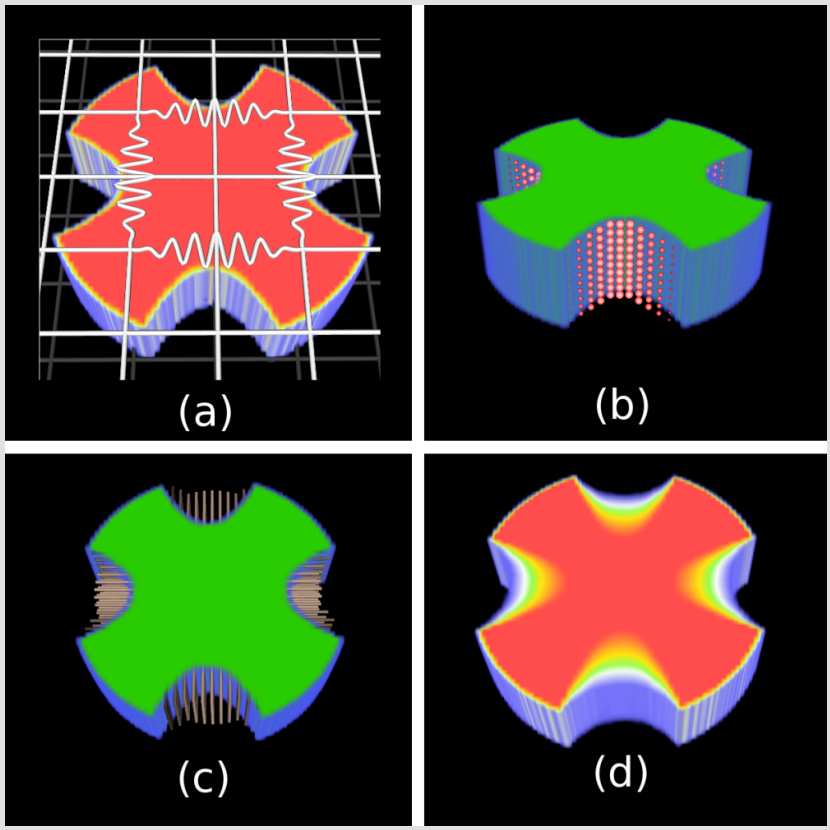
Preference Question:

Intro: Take a look at the collection of images below. Note that there are 12 images in which each side is being deformed to a different direction.



General Information	
Number of Images	12
Deforming	Left (right half), Right (left half), Up (bottom half), Bottom (upper half)
Minimum Deformation	8
Maximum Deformation	20
Average Deformation	14

Instructions: Please analyze the following annotations generated for the dataset shown above.



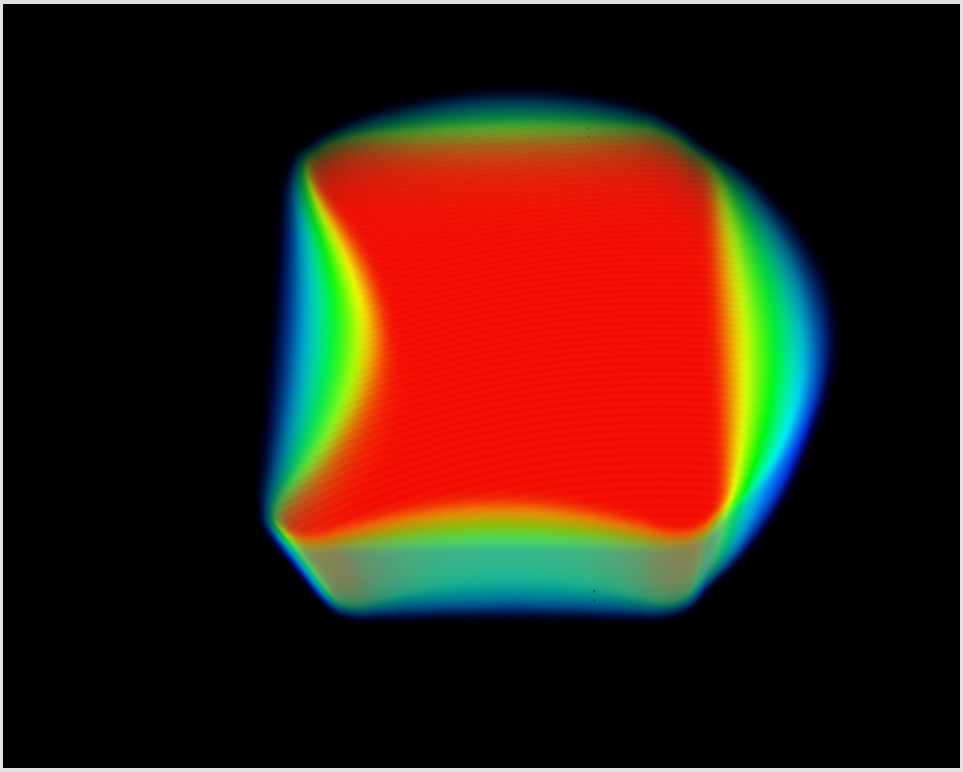
Question: How well do these annotations seem to capture the deformation properties of the group under consideration? **Note:** There's no right or wrong answer, this question is just to better determine your preferences. Please rank each of the illustration techniques.

	Poor	Weak	Good	Very Good	Excellent
(a) Deformation Grids	☐	☐	☐	☐	☐
(b) Spherical Glyphs	☐	☐	☐	☐	☐
(c) Line Glyphs	☐	☐	☐	☐	☐
(d) Likelihood Volumes	☐	☐	☐	☐	☐

Fig. 3. Sample question of the second section of the user study that was design to capture preferences.

Performance Question:

Input: Observe the following "Likelihood Volume" generated from 21 images.



Question: What can we say about the deformation of the group represented by this illustration?

(a) Only deforming to the right
(b) Deforming more to the right than down
(c) Equally deforming right and up
(d) Only deforming to the right and up
(e) None of the above

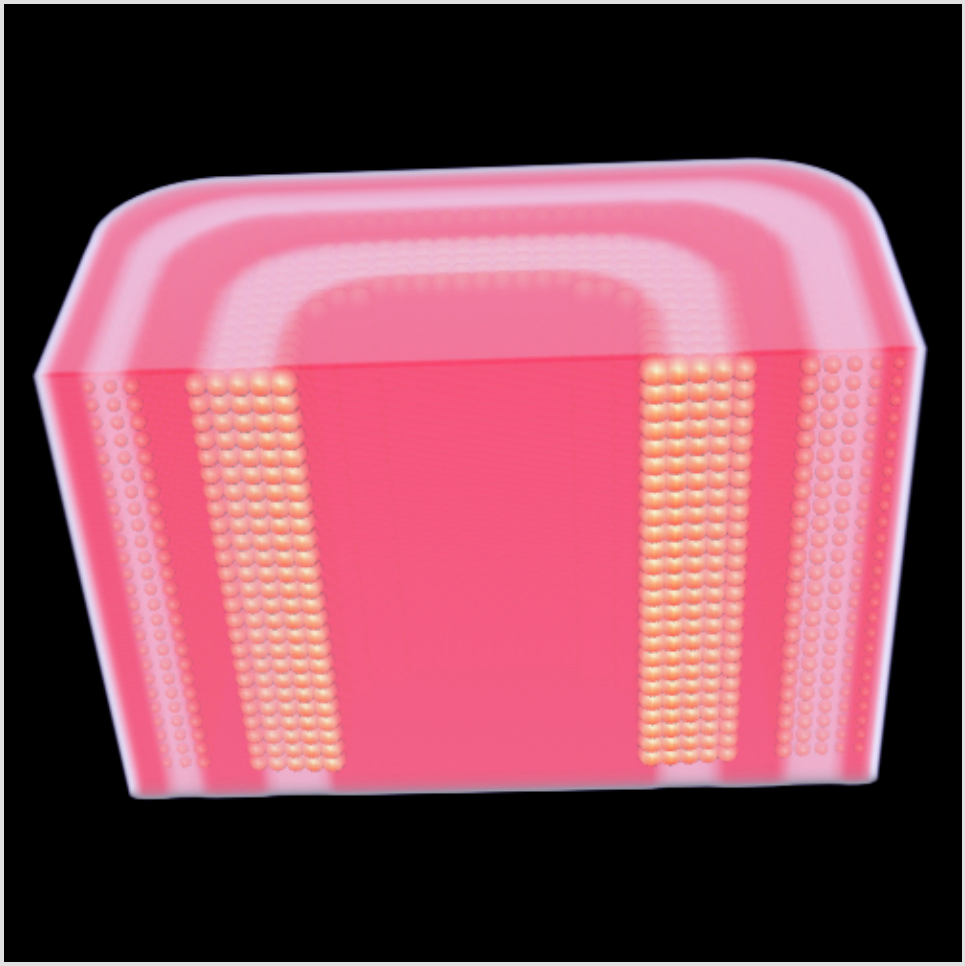
Question: How confident are you with your answer?

Extremely Confident	Very Confident	Confident	Unconfident	Very Unconfident
☐	☐	☐	☐	☐

Fig. 4. Sample question of the third section of the user study where performance was captured. Together with each answer, the user was asked to provide a confidence measurement for the answer.

Performance Question:

Input: Observe the following annotation generated from a collection of 21 images.



Question: What can we tell about the deformation properties of the group used to generate this annotation?

(a) Objects are deforming down
(b) White regions have the highest variability
(c) Red regions have the highest variability
(d) A and B are correct
(e) None of the above

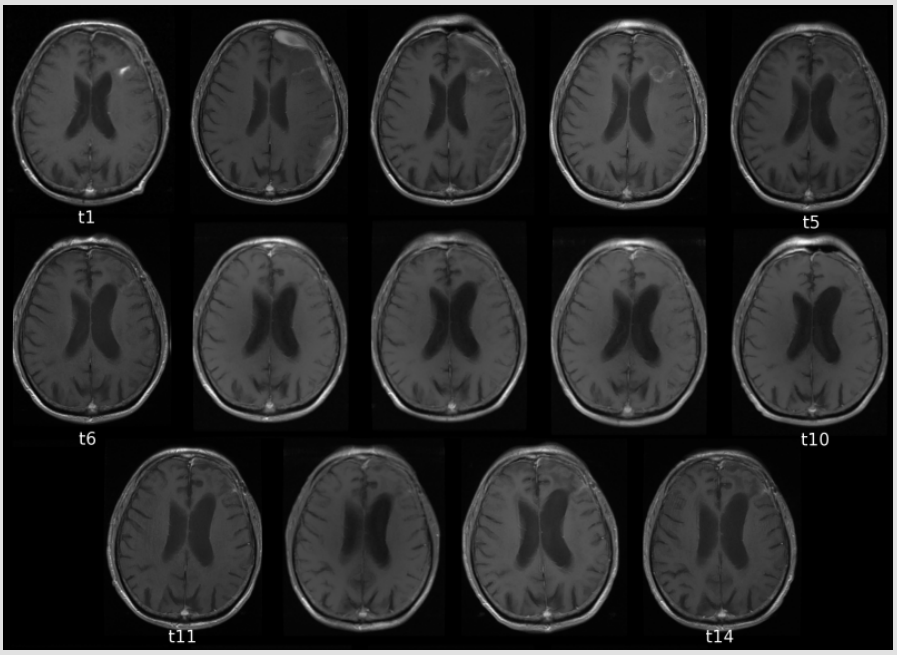
Question: How confident are you with your answer?

Extremely Confident	Very Confident	Confident	Unconfident	Very Unconfident
☐	☐	☐	☐	☐

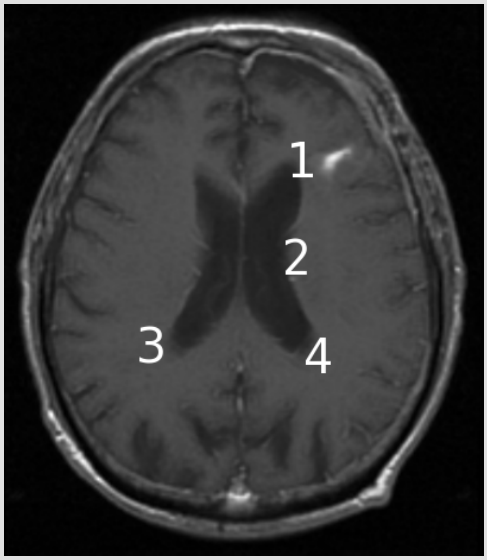
Fig. 5. Sample question of the third section of the user study where performance was captured.

Performance Question:

Input: Observe the following longitudinal study of a patient with a high-grade glioma brain tumor. The images were taken every two months for over two years and show the same anatomical region.



Please, analyze the following regions.



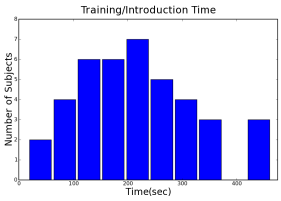
Question: Which of the regions highlighted in the previous annotation present the greatest amount of deformation over time?

Region #1	Region #2	Region #3	Region #4
☐	☐	☐	☐

Question: How confident are you with your answer?

Extremely Confident	Very Confident	Confident	Unconfident	Very Unconfident
☐	☐	☐	☐	☐

Fig. 6. Sample question used to measure the effectiveness of analyzing a collection of images. Note that each section of the survey included synthetic and real-world medical images.



Introduction: Time	
Mean	221s
Standard Deviation	106s
Skewness	0.535
Minimum	20s
Maximum	465s

Fig. 7. (left) Histogram of the time users spent reading and learning about the different statistical illustration techniques. Note that the histogram follows a normal distribution. (right) On average users spent 221 seconds reading about the different annotation techniques
.

Preferences Across Population		
Visualization Technique	**Mean Difference**	**P Value**
Deformation Grids	0.83333	0.012*
Likelihood Volumes	0.46667	0.199
Line Glyphs	0.30000	0.406
Spherical Glyphs	0.50000	0.132

Table 1. Results of our statistical analysis test when comparing preferences across population. Note that the P-value shows a statistically significant difference between both groups when *deformation grids* were used.

B Results

This section provides additional results obtained from our statistical analysis that can be used to better understand or verify our findings.

By analyzing the first section of the user study, we found that the time used in the introduction was highly variable; from a subject spending as little as 20 seconds to a subject who spent 465 seconds. Figure 7(left) shows a histogram of the time participants spent within the introduction of our survey. After analyzing the data, we did not find any significant different between the time used by computer scientists and non-computer scientists. Similarly, we did not find any significant difference between the time spent in each individual technique. Such findings might suggest that each illustration technique or the explanation of each approach had similar complexities.

When comparing the user preferences between computer scientists and non-engineers we found a significant difference in how much they preferred *Deformation Grids*. Table 1 shows some of the results of our statistical analysis. Note that a significance value $p < 0.02$ was found when comparing *Deformation Grids* between the two groups.

When analyzing user's performance, we found that *Deformation Grids* and *Line Glyphs* were significantly better than *Likelihood Volumes*, *Spherical Glyphs* as well as analyzing the raw images. Table 2 shows the post-hoc comparison between the different illustration techniques. From this table we can see that multiple visualization techniques can be used to effectively illustrate the statistical deformation properties of a group of images.

We further analyzed the overall performance across population. Figure 8 shows some of our results. First, we found that non-computer scientists tend to perform slightly better with *deformation grids* and *spherical glyphs* than computer scientists. However, given the large variability observed with both groups, a definite conclusion of the significant difference between the performance among computer scientists and non-engineers cannot be reached. Overall, we found that, regardless their background, subjects performed much better with *deformation grids* than any other technique.

When analyzing the confidence levels with each illustration technique, we found that on average users were more confident with questions involving *likelihood volumes* than with any other annotation technique. In addition, we found that users were least confident with questions involving *spherical glyphs* and raw data. Overall, there was a significant confidence difference between using *deformation grids*, *likelihood volumes*, and *line glyphs* than when using *spherical glyphs* and

Performance Comparison			
Vis Technique (I)	Vis Technique (J)	Mean Diff	P-value
Deformation Grids	LV	-0.21250	0.041*
	LG	0.15000	0.284
	SG	0.4000	0.000*
	RAW	0.7000	0.000*
Likelihood Volumes	DG	-0.21250	0.041*
	LG	-0.62500	0.927
	SG	0.18750	0.099
	RAW	0.48750	0.000*
Line Glyphs	DG	-0.15000	0.284
	LV	0.06250	0.927
	SG	0.25000	0.008*
	RAW	0.55000	0.000*
Spherical Glyphs	DG	-0.40000	0.000*
	LV	-0.18750	0.099
	LG	-0.25000	0.008*
	RAW	-0.30000	0.001*
Raw Images	DG	-0.70000	0.000*
	LV	-0.48750	0.000*
	LG	-0.55000	0.000*
	SG	0.30000	0.001*

Table 2. Results of our statistical analysis tests between each annotation technique. Note that analyzing the raw set of images without any annotation always perform worst than any other illustration technique.

Confidence Comparison			
Vis Technique (I)	Vis Technique (J)	Mean Diff	P-value
Deformation Grids	LV	-0.31250	0.258
	LG	-0.0250	1.000
	SG	0.88750	0.000*
	RAW	0.62500	0.000*
Likelihood Volumes	DG	0.31250	0.258
	LG	0.28750	0.341
	SG	1.20000	0.000*
	RAW	0.93750	0.000*
Line Glyphs	DG	0.02500	1.000
	LV	-0.28750	0.341
	SG	0.91250	0.000*
	RAW	0.65000	0.000*
Spherical Glyphs	DG	-0.88750	0.000*
	LV	-1.20000	0.000*
	LG	-0.91250	0.000*
	RAW	-0.26250	0.436
Raw Images	DG	-0.62500	0.001*
	LV	-0.93750	0.000*
	LG	-0.6500	0.000*
	SG	0.26250	0.436

Table 4. Results of our statistical analysis tests when comparing the user's confidence levels. Note that users were always more confident when using deformation grids, likelihood volumes, and line glyphs.

		Correlation: Learning Time vs. Performance			
		Performance DG	**Performance LV**	**Performance LG**	**Performance SG**
Intro Time - DG	Correlation	-0.051	0.115	0.097	-0.149
	P-Value	0.753	0.482	0.550	0.358
Intro Time - LV	Correlation	-0.251	0.070	-0.005	-0.163
	P-Value	0.118	0.668	0.977	0.315
Intro Time - LG	Correlation	-0.206	0.099	0.084	-0.034
	P-Value	0.203	0.543	0.608	0.834
Intro Time - SG	Correlation	-0.366	0.020	0.126	0.209
	P-Value	0.020*	0.904	0.440	0.195

Table 3. This table presents the correlation results obtained when analyzing if the time used to read about each technique had any effect on the overall user's accuracy.

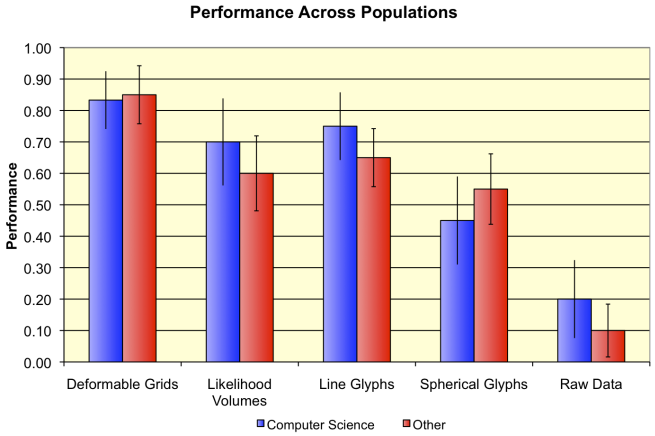


Fig. 8. Plot of the performance results across population. Note that on average, there was not a significant difference between both groups.

Correlation: Time to Answer and Performance		
Visualization Technique	**Correlation**	**P-Value**
Deformation Grids	0.076	0.640
Likelihood Volumes	-0.119	0.463
Line Glyphs	0.121	0.456
Spherical Glyphs	0.013	0.935
Raw Data	-0.085	0.605

Table 5. This table includes the correlation estimates and p-values obtained when analyzing the relationship between time to answer a question and performance.

analyzing the raw data. Table 4 summarizes some of our results. No statistical differences were found between groups.

From our results, we also extracted other information and hidden correlations. For instance, we wanted to explore if the time taken to read about each annotation technique in the introduction had any effect with the overall performance. Table 3 shows some of our results. It seems that the time used to read, understand, and analyze each annotation technique in the introduction did not have any significant effect on the overall performance. From Figure 1 we can see that the variability of the training time was very large; from one subject spending around 20s while another subject used 465s. However, overall it seems that the time did not have a direct effect of the overall performance.

We also compared and analyzed the correlation between the time to answer a question and the accuracy of the answer. We found that on average, there is no significant correlation between the variables and the more time users spent analyzing the questions did not translate into more accurate answers. Table 5 shows some of the results.