InfoVis Evaluation

CS 7450 - Information Visualization
December 5, 2016
John Stasko

Agenda

• How do we evaluate visualizations?
  – Different styles and dimensions
• Notable example evaluation projects

• Project preparation and planning
• Grading
Area Focus

- Most of the research in InfoVis that we’ve learned about this semester has been the introduction of a new visualization technique or tool
  - Fisheyes, Tableau, hyperbolic displays, TableLens, themescapes, SunBurst, Wordles, ...
  - “Isn’t my new visualization cool?...”

Evaluation – Why?

- Reasons?
Evaluation – Why?

- Want to learn what aspects of visualizations or systems “works”
- Want to ensure that methods are improving
- Want to insure that technique actually helps people and isn’t just “cool”
- NOT: Because I need that section in my paper to get it accepted ... sigh

Evaluation – How?

- How do we evaluate visualizations?
  - How would you evaluate your project system?
- What do we measure?
  - What data do we gather?
  - What metrics do we use?
- What evaluation techniques should we use?
- (Channel your HCI knowledge)
Evaluation in HCI

• Takes many different forms
  – Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, ...

• So, which ones are best for evaluating InfoVis systems?

Controlled Experiments

• Good for measuring performance or comparing multiple techniques
• Often quantitative in nature
• What do we measure?
  – Performance, time, errors, ...

• Strengths, weaknesses?
Subjective Assessments

- Often observational with interview
- Learn people’s subjective views on tool
  - Was it enjoyable, confusing, fun, difficult, ...?
- This kind of personal judgment strongly influence use and adoption, sometimes even overcoming performance deficits

- Strengths, weaknesses?

Running Studies

- Beyond our scope here
- You should learn more about this in CS 6750 or 6455
Evaluating UI vs. InfoVis

- Seems comparable but...
- What are some differences?

Usability vs. Utility

- Big difference
- Usability is not the same as utility, which seems to be a key factor for InfoVis
- Can think of visualizations that are very usable but not useful or helpful
- More difficult to measure success of an infovis because more domain knowledge and situated use is required
Evaluating InfoVis in General

- Very difficult in InfoVis to compare “apples to apples”
  - Hard to compare System A to System B
  - Different tools were built to address different user tasks
- UI can heavily influence utility and value of visualization technique

BELIV workshop focused on this topic

Nice locations!
Evaluating InfoVis

• Three nice overview papers
  – Plaisant, AVI ‘04
  – Carpendale, book chapter ‘08
  – Lam, et al, TVCG ‘12

Plaisant ‘04

• Discusses challenges, possible next steps, and gives examples from work at Maryland
Evaluation Challenges

- Matching tools with users, tasks, and real problems
- Improving user testing
  - Looking at the same data from different perspectives, over a long time
  - Answering questions you didn’t know you had
  - Factoring in the chances of discovery and the benefits of awareness
- Addressing universal usability

Possible Next Steps

- Repositories of data and tasks
- Case studies and success stories
- The role of toolkits and development tools
Carpendale ‘08

- Challenges in infovis evaluation
- Choosing an evaluation approach

Evaluation Approaches

- Desirable features
  - Generalizability
  - Precision
  - Realism
Quantitative Methods

- Laboratory experiments & studies
- Traditional empirical scientific experimental approach
- Steps
Quantitative Challenges

- **Conclusion Validity**
  - Is there a relationship?

- **Internal Validity**
  - Is the relationship causal?

- **Construct Validity**
  - Can we generalize to the constructs (ideas) the study is based on?

- **External Validity**
  - Can we generalize the study results to other people/places/times?

- **Ecological Validity**
  - Does the experimental situation reflect the type of environment in which the results will be applied?

Qualitative Methods

- **Types**
  - Nested methods
    - Experimenter observation, think-aloud protocol, collecting participant opinions
  - Inspection evaluation methods
    - Heuristics to judge

- **Observational context**
  - In situ, laboratory, participatory
  - Contextual interviews important
Qualitative Challenges

- Sample sizes
- Subjectivity
- Analyzing qualitative data

Lam, et al ‘12

- Meta-review: analysis of 850 infovis papers (361 with evaluation)
- Focus on evaluation scenarios
Evaluation Taxonomies

<table>
<thead>
<tr>
<th>Type</th>
<th>Categories</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation goals</td>
<td>Summative (to summarize the effectiveness of an interface), formative (to inform design)</td>
<td>Jackson [2]</td>
</tr>
<tr>
<td>Evaluation goals</td>
<td>Predictive (e.g., to compare design alternatives and compute usability metrics), observational (e.g., to understand user behavior and performance), participatory (e.g., to understand user behavior, performance, thoughts, and experiences)</td>
<td>Ellis and Dix [22]</td>
</tr>
<tr>
<td>Evaluation challenges</td>
<td>Quantitative (e.g., types validity: conclusion types I &amp; II errors), construct, externalized, internalized, ecological, qualitative (e.g., subjectivity, sample size, analysis approaches)</td>
<td>Reihm and Reihmels [34]</td>
</tr>
<tr>
<td>Research strategies</td>
<td>Mixed (generalizability, precision, relevance, conciseness, robustness) and research situations (field, experimental, exploratory, theoretical)</td>
<td>McGlash [55]</td>
</tr>
<tr>
<td>Research methods</td>
<td>Class (e.g., testing, inspection), type (e.g., log file analysis, guideline reviews), automation type (e.g., non, capture), effort level (e.g., minimal effort, model development)</td>
<td>Irony and Hart [42]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Neural Process Model with four stages (identify problem, characterize data/generation technique, design, algorithm design), each with potential threats to validity and methods of validation</td>
<td>Mumaer [54]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Design/development cycle stage associated with evaluation goals (“exploratory” with “before design”, “predictive” with “before implementation”, “formative” with “during implementation”, and “summative” with “after implementation”). Methods are further classified on feasibility (by usability, quality) or testing (by test users)</td>
<td>Andrews [2]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Planning &amp; feasibility (e.g., competitive analysis), requirements (e.g., user surveys), design (e.g., heuristic evaluation), implementation (e.g., style guide), test &amp; measures (e.g., diagnostic evaluation), and post release (e.g., post-reverse evaluation)</td>
<td>Viability set [88]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Conceptual, Detailed design, implementation, analysis</td>
<td>Kothy et al. [30]</td>
</tr>
<tr>
<td>Data and method</td>
<td>Data collected (qualitative, quantitative), collection method (empirical, analytical)</td>
<td>Buckhans and Rode [2]</td>
</tr>
<tr>
<td>Data</td>
<td>Data collected (qualitative, quantitative, mixed method)</td>
<td>Creswell [17]</td>
</tr>
<tr>
<td>Evaluation scope</td>
<td>Work environment, system, components</td>
<td>Thomas and Cook [82]</td>
</tr>
</tbody>
</table>

Evaluation Scenarios

- Understanding data analysis
  - Understanding environments and work practices (UWP)
  - Evaluating visual data analysis and reasoning (VDAR)
  - Evaluating communication through visualization (CTV)
  - Evaluating collaborative data analysis (CDA)
Evaluation Scenarios

- Understanding visualizations
  - Evaluating user performance (UP)
  - Evaluating user experience (UE)
  - Evaluating visualization algorithms (VA)

Methods

- Coded each paper with tags

<table>
<thead>
<tr>
<th>Paper Tags</th>
<th>EuroVis</th>
<th>InfoVis</th>
<th>IVS</th>
<th>VAST</th>
<th>Total</th>
<th>Scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>People’s workflow, work practices</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>UWP</td>
</tr>
<tr>
<td>Data analysis</td>
<td>0</td>
<td>5</td>
<td>9</td>
<td>5</td>
<td>13</td>
<td>VDAR</td>
</tr>
<tr>
<td>Decision making</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>VDAR</td>
</tr>
<tr>
<td>Knowledge management</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>VDAR</td>
</tr>
<tr>
<td>Knowledge discovery</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>VDAR</td>
</tr>
<tr>
<td>Communication, learning, teaching, publishing</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>CTV</td>
</tr>
<tr>
<td>Causal information acquisition</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>CTV</td>
</tr>
<tr>
<td>Collaboration</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>CDA</td>
</tr>
<tr>
<td>Visualization-analytical operation</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>13</td>
<td>UP</td>
</tr>
<tr>
<td>Perception and cognition</td>
<td>17</td>
<td>24</td>
<td>15</td>
<td>3</td>
<td>62</td>
<td>UP</td>
</tr>
<tr>
<td>Useability/efficacy</td>
<td>25</td>
<td>84</td>
<td>31</td>
<td>18</td>
<td>158</td>
<td>UP/UI</td>
</tr>
<tr>
<td>Potential usage</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>22</td>
<td>UE</td>
</tr>
<tr>
<td>Adoption</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>UE</td>
</tr>
<tr>
<td>Algorithm performance</td>
<td>17</td>
<td>37</td>
<td>15</td>
<td>0</td>
<td>69</td>
<td>VA</td>
</tr>
<tr>
<td>Algorithm quality</td>
<td>1</td>
<td>0</td>
<td>12</td>
<td>5</td>
<td>28</td>
<td>VA</td>
</tr>
<tr>
<td>Not included in scenarios</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>

Fall 2016  CS 7450  29
Methods

• For each category the authors describe
  – Goals and outputs
  – Evaluation questions
  – Methods and examples

Example

• UWP - Understanding Environments and Work Practices
  – Elicit formal requirements for design
  – Study people for which a tool is being designed and the context of use
  – Very few infovis papers on this topic
**UWP 1**

- **Goals and Outputs**
  - **Goals**: Understand the work, analysis, or info processing practices by a given group of people with or without software in use
  - **Outputs**: Design implications based on a more holistic understanding of current workflows and work practices, the conditions of the working environment, and potentially current tools in use

**UWP 2**

- **Evaluation questions**
  - What is the context of use of visualizations?
  - In which daily activities should the visualization tool be integrated?
  - What types of analyses should the visualization tool support?
  - What are the characteristics of the identified user group and work environments?
  - What data is currently used and what tasks are performed on it?
  - What kinds of visualizations are currently in use? How do they help to solve current tasks?
  - What challenges and usage barriers can we see for a visualization tool?
UWP 3

- Methods and Examples
  - Field observation
  - Interviews
  - Laboratory observation
  - (with example projects cited)

Examples

- Let’s examine example studies utilizing different goals and styles
Which Technique is Best?

- Space-filling hierarchical views
- Compare Treemap and Sunburst with users performing typical file/directory-related tasks
- Evaluate task performance on both correctness and time

Stasko et al. IJHCS '00

Tools Compared

Treemap

SunBurst

Fall 2016 CS 7450 37
Hierarchies Used

- Four in total

Small Hierarchy
(~500 files)

A          B

Large Hierarchy
(~3000 files)

A          B

- Used sample files and directories from our own systems (better than random)

Methodology

- 60 participants
- Participant only works with a small or large hierarchy in a session
- Training at start to learn tool
- Vary order across participants

SB A, TM B
TM A, SB B
SB B, TM A
TM B, SB A

32 on small hierarchies
28 on large hierarchies
Tasks

- Identification (naming or pointing out) of a file based on size, specifically, the largest and second largest files (Questions 1-2)
- Identification of a directory based on size, specifically, the largest (Q3)
- Location (pointing out) of a file, given the entire path and name (Q4-7)
- Location of a file, given only the file name (Q8-9)
- Identification of the deepest subdirectory (Q10)
- Identification of a directory containing files of a particular type (Q11)
- Identification of a file based on type and size, specifically, the largest file of a particular type (Q12)
- Comparison of two files by size (Q13)
- Location of two duplicated directory structures (Q14)
- Comparison of two directories by size (Q15)
- Comparison of two directories by number of files contained (Q16)

Hypothesis

- Treemap will be better for comparing file sizes
  - Uses more of the area
- Sunburst would be better for searching files and understanding the structure
  - More explicit depiction of structure
- Sunburst would be preferred overall
# Small Hierarchy

<table>
<thead>
<tr>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM ($n = 8$)</td>
<td>1</td>
<td>9.88 (3.23)</td>
<td>TM ($n = 8$)</td>
<td>1</td>
<td>11.50 (2.14)</td>
</tr>
<tr>
<td>SB ($n = 8$)</td>
<td>1</td>
<td>12.88 (1.96)</td>
<td>SB ($n = 8$)</td>
<td>1</td>
<td>10.38 (1.69)</td>
</tr>
<tr>
<td>TM ($n = 8$)</td>
<td>2</td>
<td>12.25 (1.75)</td>
<td>TM ($n = 8$)</td>
<td>2</td>
<td>10.75 (2.77)</td>
</tr>
<tr>
<td>SB ($n = 8$)</td>
<td>2</td>
<td>12.63 (2.00)</td>
<td>SB ($n = 8$)</td>
<td>2</td>
<td>11.50 (2.00)</td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td></td>
<td>11.06 (2.79)</td>
<td>TM (collapsed across phase)</td>
<td></td>
<td>11.13 (2.42)</td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td></td>
<td>12.75 (1.91)</td>
<td>SB (collapsed across phase)</td>
<td></td>
<td>10.94 (1.88)</td>
</tr>
</tbody>
</table>

Correct task completions (out of 16 possible)

Fall 2016    CS 7450    43

# Large Hierarchy

<table>
<thead>
<tr>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM ($n = 7$)</td>
<td>1</td>
<td>8.71 (1.60)</td>
<td>TM ($n = 7$)</td>
<td>1</td>
<td>9.29 (2.14)</td>
</tr>
<tr>
<td>SB ($n = 7$)</td>
<td>1</td>
<td>11.43 (1.27)</td>
<td>SB ($n = 7$)</td>
<td>1</td>
<td>11.14 (2.67)</td>
</tr>
<tr>
<td>TM ($n = 7$)</td>
<td>2</td>
<td>11.57 (1.27)</td>
<td>TM ($n = 7$)</td>
<td>2</td>
<td>10.86 (1.57)</td>
</tr>
<tr>
<td>SB ($n = 7$)</td>
<td>2</td>
<td>11.00 (2.16)</td>
<td>SB ($n = 7$)</td>
<td>2</td>
<td>11.00 (2.00)</td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td></td>
<td>10.14 (2.03)</td>
<td>TM (collapsed across phase)</td>
<td></td>
<td>9.57 (2.24)</td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td></td>
<td>11.21 (1.72)</td>
<td>SB (collapsed across phase)</td>
<td></td>
<td>11.07 (2.27)</td>
</tr>
</tbody>
</table>

Correct task completions (out of 16 possible)

Fall 2016    CS 7450    44
Performance Results

- Ordering effect for Treemap on large hierarchies
  - Participants did better after seeing SB first
- Performance was relatively mixed, trends favored Sunburst, but not clear-cut
  - Oodles of data!

Subjective Preferences

- Subjective preference: SB (51), TM (9), unsure (1)
- People felt that TM was better for size tasks (not borne out by data)
- People felt that SB better for determining which directories inside others
  - Identified it as being better for structure
Strategies

• How a person searched for files etc. mattered
  – Jump out to total view, start looking
  – Go level by level

Animation Helpful?

• Examine whether animated bubble charts (a la Rosling and GapMinder) are beneficial for analysis and presentation
• Run an experiment to evaluate the effects of animation

Robertson et al
TVCG (InfoVis) '08
Visualizations Studied

- Animation
- Small multiples
- Traces

Experiment Design

- 3 (animation types) x 2 (data size: small & large) x 2 (presentation vs. analysis)
  - Presentation vs analysis – between subjects
  - Others – within subjects

- Animation has 10-second default time, but user could control time slider
Experiment Design

• Data
  – UN data about countries

• Tasks
  – 24 tasks, 1-3 requires answers per
    Select 3 countries whose rate of energy consumption was faster than their rate of GDP per capita growth
    Select 2 countries with significant decreases in energy consumption
    Which continent had the least changes in GDP per capita

Conditions

• Analysis – straightforward, interactive

• Presentation
  – 6 participants at a time
  – Presenter described a trend relevant to task, but different
  – No interaction with system
    In animation condition, participants saw last frame of animation (no interaction)
Results

• Accuracy  Measured as percentage correct
  65% overall (pretty tough)

Significant:
SM better than animation
Small data size more accurate than large

Results

• Speed
  – Presentation
    Animation faster than small multiples & traces
    15.8 secs vs. 25.3 secs vs. 27.8 secs.
  – Analysis
    Animation slower than small multiples & traces
    83.1 secs. vs. 45.69 secs. vs. 55.0 secs.
Results

Table 3. Average ratings for seven questions for each visualization. * indicates significant differences (p<.05).

<table>
<thead>
<tr>
<th>Question</th>
<th>Animation</th>
<th>SM</th>
<th>Traces</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. I found the animation view enjoyable</td>
<td>4.4 *Traces</td>
<td>4.2</td>
<td>4.1</td>
</tr>
<tr>
<td>Q2. I found the small dataset view enjoyable</td>
<td>4.6 *SM</td>
<td>4.2</td>
<td>4.5</td>
</tr>
<tr>
<td>Q3. I found the small dataset view enjoyable</td>
<td>3.6 *Traces</td>
<td>3.4</td>
<td>2.3</td>
</tr>
<tr>
<td>Q4. I found this visualization exciting</td>
<td>4.3 *SM</td>
<td>3.7</td>
<td>3.5</td>
</tr>
<tr>
<td>Q5. I found the small dataset view enjoyable</td>
<td>4.3 *SM</td>
<td>3.1</td>
<td>3.0</td>
</tr>
<tr>
<td>Q6. I found the small dataset view enjoyable</td>
<td>1.8</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Q7. I found the large dataset view enjoyable</td>
<td>4.4 *SM</td>
<td>3.8</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Subjective

Table 4. Average ratings for a few general questions.

<table>
<thead>
<tr>
<th>Question</th>
<th>Presentation</th>
<th>Analysis</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1. I found the Traces view enjoyable</td>
<td>3.8</td>
<td>2.9</td>
<td>3.4</td>
</tr>
<tr>
<td>G3. I found the Small Multiples view enjoyable</td>
<td>4.1</td>
<td>3.4</td>
<td>3.7</td>
</tr>
<tr>
<td>G5. I found the Animation view enjoyable</td>
<td>4.6</td>
<td>3.0</td>
<td>4.8</td>
</tr>
<tr>
<td>G7. The animation went too fast for me</td>
<td>3.2</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>G8. The animation went too slow for me</td>
<td>1.6</td>
<td>1.3</td>
<td>1.4</td>
</tr>
<tr>
<td>G9. I lost track of some data points as they moved</td>
<td>4.9</td>
<td>4.0</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Likert: 0-strongly disagree, 6-strongly agree

Results

G13: Which visualization did you PREFER for the small dataset?
G14: For the large?

Presentation, small: Animation (9) > SM (6) > Traces (3)
Presentation, large: Traces (8) > SM (6) > Animation (4)
Analysis, small: Animation (7) > SM (6) > Traces (5)
Analysis, large: Animation (8) > SM (6) > Traces (4)
Discussion

• People rated animation more fun, but small multiples was more effective

• As data grows, accuracy becomes an issue
  – Traces & animation get cluttered
  – Small multiple gets tiny

• Animation:
  – “fun”, “exciting”, “emotionally touching”
  – Confusing, “the dots flew everywhere”

Insight

• Isn’t one of the key ideas about InfoVis that it helps generate insights?

• OK, well let’s count/measure insights

• What challenges do you see in this?
Problem Domain

- Microarray experiments: Gain insight into the extremely complex and dynamic functioning of living cells
- Systems-level exploratory analysis of thousands of variables simultaneously
- Big data sets

Insight

- Insight: An individual observation about the data by the participant, a unit of discovery
- Characteristics
  - Observation
  - Time
  - Domain Value
  - Hypotheses
  - Directed vs Unexpected
  - Category
Insight Characteristics

- **Complex**
  - Involving large amounts of data in a synergistic way
- **Deep**
  - Builds over time, generates further questions
- **Qualitative**
  - Can be uncertain and subjective
- **Unexpected**
  - Often unpredictable, serendipitous
- **Relevant**
  - Deeply embedded in data domain, connecting to existing domain knowledge

Experiment Design

- **Data: Timeseries, Virus, Lupus**

<table>
<thead>
<tr>
<th>Tool</th>
<th>Visual Representations</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster/ Treeview</td>
<td>Heat-map, Clustered heat-map</td>
<td>O+D</td>
</tr>
<tr>
<td>Time-Searcher</td>
<td>Parallel coordinates, line graph</td>
<td>Brushing, O+D, DQ</td>
</tr>
<tr>
<td>HCE</td>
<td>Cluster dendrogram, parallel coordinates, heat-map, scatterplot, histogram</td>
<td>Brushing, Zooming, O+D, DQ</td>
</tr>
<tr>
<td>Spotfire® 7.2</td>
<td>Parallel coordinates, heat-map, scatterplots (2D/3D), histogram, bar/pie chart, tree view, spreadsheet view, Clustered parallel coordinates</td>
<td>Brushing, Zooming, O+D, DQ</td>
</tr>
<tr>
<td>Functional Genomics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GeneSpring® @ 5.0</td>
<td>Parallel coordinate, heat-map, scatterplots (2D/3D), histogram, bar chart, block view, physical position view, array layout view, pathway view, spreadsheet view, compare gene to gene, Clusteresed parallel coordinates</td>
<td>Brushing, Zooming</td>
</tr>
</tbody>
</table>
Results
Discussion

• Methodology difficulties
  – Labor intensive
  – Requires domain expert
  – Requires motivated subjects
  – Training and trial time

• Weakness: Short session time (2 hours) when long-term use more desirable

Rethinking Methodology

• Do controlled lab experiments really tell us very much in information visualization?
MILC Technique

- Multi-dimensional
  - observations, interviews, surveys, logging
- In-depth
  - intense engagement of researchers with domain experts so as to almost become a partner
- Long-term
  - longitudinal use leading to strategy changes
- Case Study
  - detailed reporting about small number of people working on their own problems in their own domain

Influences

- Ethnography
  - Preparation
  - Field study
  - Analysis
  - Reporting

Shneiderman & Plaisant
BELIV '06
Guidelines

- Specify focused research questions & goals
- Identify 3-5 users
- Document current method/tool
- Determine what would constitute professional success for users
- Establish schedule of observation & interviews
- Instrument tool to record usage data
- Provide attractive log book for comments, problems, and insights
- Provide training
- Conduct visits & interviews
- Encourage users to continue using best tool for task
- Modify tool as needed
- Document successes and failures

SocialAction

- Evaluation inspired by MILC ideas goals
  - Interview (1 hour)
  - Training (2 hours)
  - Early use (2-4 weeks)
  - Mature use (2-4 weeks)
  - Outcome (1 hour)

Perer & Shneiderman
CHI '08
Methodology

- Four case studies
  - Senatorial voting patterns
  - Medical research knowledge discovery
  - Hospital trustee networks
  - Group dynamics in terrorist networks
- Named names
  - I like it!
- Tell what they did with system

My Reflections

- Nice paper
- Stark contrast to comparative, controlled experiments
- We likely need more of this in InfoVis
Value & Evaluation

- Many small, controlled experiment user studies don’t adequately assess true utility of a visualization
- Alternative: Detailed usage scenarios with identification of system’s value along four dimensions

Value Definition

\[ V_{\text{value}} = T + I + E + C \]

T – Ability to minimize the total time needed to answer a wide variety of questions about the data

I – Ability to spur and discover insights or insightful questions about the data

E – Ability to convey an overall essence or take-away sense of the data

C – Ability to generate confidence and trust about the data, its domain and context
**Recommendation**

- Provide one or more case studies that illustrate how a system/technique contributes along each of these four dimensions
- Explain how the system will provide value and utility in data analysis situations

**UX Attributes**

Survey of evaluations measuring
- Memorability
- Engagement
- Enjoyment

When are they important?

Saket, Endert, Stasko
BELIV '16
Summary

- Why do evaluation of InfoVis systems?
  - We need to be sure that new techniques are really better than old ones
  - We need to know the strengths and weaknesses of each tool; know when to use which tool

Challenges

- There are no standard benchmark tests or methodologies to help guide researchers
  - Moreover, there’s simply no one correct way to evaluate
- Defining the tasks is crucial
  - Would be nice to have a good task taxonomy
  - Data sets used might influence results
- What about individual differences?
  - Can you measure abilities (cognitive, visual, etc.) of participants?
Challenges

- Insight is important
  - Great idea, but difficult to measure

- Utility is a real key
  - Usability matters, but some powerful systems may be difficult to learn and use

- Exploration
  - InfoVis most useful in exploratory scenarios when you don’t know what task or goal is
    So how to measure that?!

Learning Objectives

- Understand the different styles of evaluations for visualizations
  - Enumerate the different dimensions of evaluation
  - Describe the benefits and limitations of each style and dimension

- Explain the challenges and difficulties in evaluating visualizations

- Know where to look for assistance and help in designing an evaluation
  - Carpendale '08 & Lam et al '12 papers

- Provide examples of thoughtful evaluation projects and papers

- Describe different visualization evaluation methodologies and +/- of each
  - Comparative study, Insight-based, MILC

- Be able to choose an appropriate evaluation methodology for a visualization system you have created
Project Tips

• Explain the visual mapping
• Labels, legends, etc., are your friends
• If your domain/problem is a little different, spend a little more effort explaining it

Final Project

• Demos Thu-Fri at the Vis Lab (near my office)
  – 20 minutes per session, be on time
  – Sign-ups in t-square – Let’s select now
  – Show/describe for a little, let us try for a while
  – Important: Bring 3 copies of a summary sheet - member names, paragraph overview, image
• Final exam slot video session on Friday@ 2:00pm
  – TSRB Auditorium
  – Show your video, then answer questions
  – Make the file available to me (thumbdrive, web, ...) by noon that day - Requirement
• Questions?
Video Advice

* Script
  - Introduce problem
  - Describe visualization & system
  - Walk through usage scenario
  - (OK to be creative and have a little fun)

* You’ve seen examples all semester
  - eg, our class webpage’s Schedule page,
  http://www.cc.gatech.edu/gvu/ii/videos.html

Video Advice

* Use Camtasia

* Process
  - 1. Develop script (rehearse timing)
  - 2. Record script
  - 3. Capture video of demo to script
  - 4. Add effects
Team Survey/Self-Assessment

- Copies distributed here and in t2
- Only I will read these
  - Be honest
- Return at demo or video showcase

Grades

- Components
  - HWs
  - Project
  - Quizzes
  - Participation
  - Exam

- Items will be posted in t-square this & next week

- Calculation
Course Survey

- Take a few minutes to complete CIOS/TAOS
  - Info: http://www.cetl.gatech.edu/cios
  - Surveys: http://gatech.smartevals.com
    (and from t-square homepage)

Take a few minutes to complete

- Good thing to do and could win an iPod!

InfoVis Gospel

- Hopefully, course has increased your awareness of topic and you can become an advocate

- Keep me posted as your use these ideas in your career