InfoVis Evaluation

CS 7450 - Information Visualization
December 3, 2012
John Stasko

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, cone trees, hyperbolic displays, tilebars, themescapes, sunburst, jazz, ...
  - “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?

• 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
Evaluating Research

• How does one judge the quality of work in Information Visualization?

Research Evaluation

• Different possible ways
  – Impact on community as a whole, influential ideas
  – Assistance to people in the tasks they care about
Strong View

- Unless a new technique or tool helps people in some kind of problem or task, it doesn’t have any value

Broaden Thinking

- Sometimes the chain of influence can be long and drawn out
  - System X influences System Y influences System Z which is incorporated into a practical tool that is of true value to people

- This is what research is all about (typically)

OK, what has research community done?
Past Review

- Old journal issue whose special topic focus was Empirical Studies of Information Visualizations

- A bit dated now

BELIV

Workshop focused on this topic

Nice locations!
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
  - Hypothesis development
  - Identification of independent variables
  - Control of Independent variables
  - Elimination of complexity
  - Measurement of dependent variables
  - Application of statistics
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 1**

<table>
<thead>
<tr>
<th>Type</th>
<th>Categories</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation goals</td>
<td>Informative (to summarize the effectiveness of an interface), formative (to shape design)</td>
<td>Andrews [5], Ellis and Dix [22], Herbert and Reichle [32]</td>
</tr>
<tr>
<td>Evaluation goals</td>
<td>Predictive (e.g., to compare design alternatives and compare usability methods), observational (e.g., to understand user behavior and performance), participatory (e.g., to understand user behaviors, performance, thoughts, and experience)</td>
<td>Andrews [5], Ellis and Dix [22]</td>
</tr>
<tr>
<td>Evaluation challenges</td>
<td>Quantitative (e.g., type validity, conclusion types I &amp; II errors), construct, correlational, ecological, qualitative (e.g., subjectivity, sample size, analysis approaches)</td>
<td>Carper [10]</td>
</tr>
<tr>
<td>Research strategies</td>
<td>Mixed (generalizability, precision, realism, complementarity, observability) and research strategies (field, experimental, responsive, theoretical)</td>
<td>McGraw [33]</td>
</tr>
<tr>
<td>Research methods</td>
<td>Class (e.g., testing, inspection), type (e.g., log file, analysis, guideline review), association type (e.g., norm, capture), offset level (e.g., regional effort, model development)</td>
<td>Ivory and Hearn [42]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Nested Process Model with four stages (domain problem characterization, design/implementation abstraction, encoding/interaction technique design, algorithm design), each with potential threats to validity and methods of validation</td>
<td>Matzger [54]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Design/Implementation cycle stage associated with evaluation goals (“exploratory”, “hypothesis”, “predictive” with “before implementation”, “formative” with “during implementation”, and “summative” with “after implementation”)</td>
<td>Andrews [2]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Planning &amp; feasibility (e.g., computer analysis), requirements (e.g., user survey), design (e.g., heuristic evaluation), implementation (e.g., style guide), test &amp; metrics (e.g., diagnostic evaluation), and post release (e.g., review evaluation)</td>
<td>Usability.net [88]</td>
</tr>
<tr>
<td>Design stages</td>
<td>Concept design, detailed design, implementation, evaluation</td>
<td>Kaly [16]</td>
</tr>
<tr>
<td>Data and methods</td>
<td>Data collected (qualitative, quantitative), collection method (empirical, analytical)</td>
<td>Burbous and Role [5]</td>
</tr>
<tr>
<td>Data</td>
<td>Data collected (qualitative, quantitative, mixed-methods)</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>Process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. People’s workflow, work practices</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>7</td>
<td>UWP</td>
</tr>
<tr>
<td>2. Data analysis</td>
<td>0</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>13</td>
<td>VDAR</td>
</tr>
<tr>
<td>3. Decision making</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>VDAR</td>
</tr>
<tr>
<td>4. Knowledge management</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>VDAR</td>
</tr>
<tr>
<td>5. Knowledge discovery</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>VDAR</td>
</tr>
<tr>
<td>6. 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>7. Casual information acquisition</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>CTV</td>
</tr>
<tr>
<td>8. Collaboration</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>9</td>
<td>CDA</td>
</tr>
<tr>
<td>Visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. 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>10. 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>11. Usability/effectiveness</td>
<td>25</td>
<td>84</td>
<td>31</td>
<td>18</td>
<td>158</td>
<td>UP/UE</td>
</tr>
<tr>
<td>12. Potential usage</td>
<td>7</td>
<td>1</td>
<td>5</td>
<td>9</td>
<td>22</td>
<td>UE</td>
</tr>
<tr>
<td>13. Adoption</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>5</td>
<td>UE</td>
</tr>
<tr>
<td>14. Algorithm performance</td>
<td>17</td>
<td>37</td>
<td>15</td>
<td>0</td>
<td>69</td>
<td>VA</td>
</tr>
<tr>
<td>15. Algorithm quality</td>
<td>1</td>
<td>10</td>
<td>12</td>
<td>5</td>
<td>28</td>
<td>VA</td>
</tr>
<tr>
<td>Not included as scenarios</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Proposed evaluation methodologies</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>17. Evaluation metric development</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>-</td>
</tr>
</tbody>
</table>
Methods

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

Examples

• Let’s examine a few 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

Tools Compared

Stasko et al. *IJHCS '00*
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>
</tr>
</thead>
<tbody>
<tr>
<td>TM ($n = 8$)</td>
<td>1</td>
<td>9.88 (3.25)</td>
</tr>
<tr>
<td>SB ($n = 8$)</td>
<td>1</td>
<td>12.88 (1.90)</td>
</tr>
<tr>
<td>TM ($n = 8$)</td>
<td>2</td>
<td>12.25 (1.75)</td>
</tr>
<tr>
<td>SB ($n = 8$)</td>
<td>2</td>
<td>12.63 (2.00)</td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td>11.06 (2.79)</td>
<td></td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td>12.75 (1.91)</td>
<td></td>
</tr>
</tbody>
</table>

Correct task completions (out of 16 possible)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<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>10.38 (1.69)</td>
</tr>
<tr>
<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>11.50 (2.00)</td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td>11.13 (2.42)</td>
<td></td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td>10.94 (1.88)</td>
<td></td>
</tr>
</tbody>
</table>

### Large Hierarchy

<table>
<thead>
<tr>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM ($n = 7$)</td>
<td>1</td>
<td>6.71 (1.60)</td>
</tr>
<tr>
<td>SB ($n = 7$)</td>
<td>1</td>
<td>11.43 (1.27)</td>
</tr>
<tr>
<td>TM ($n = 7$)</td>
<td>2</td>
<td>11.57 (1.27)</td>
</tr>
<tr>
<td>SB ($n = 7$)</td>
<td>2</td>
<td>11.00 (2.16)</td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td>10.14 (2.03)</td>
<td></td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td>11.21 (1.72)</td>
<td></td>
</tr>
</tbody>
</table>

Correct task completions (out of 16 possible)

<table>
<thead>
<tr>
<th>Tool</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM ($n = 7$)</td>
<td>1</td>
<td>8.29 (2.14)</td>
</tr>
<tr>
<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>10.86 (1.57)</td>
</tr>
<tr>
<td>SB ($n = 7$)</td>
<td>2</td>
<td>11.00 (2.00)</td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td>9.57 (2.24)</td>
<td></td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td>11.07 (2.27)</td>
<td></td>
</tr>
</tbody>
</table>
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

DQ vs. BH

- Empirical Study
  - Use DataMaps, a geographic (US states) data visualization tool
  - Have participants do different tasks with both methods
    - How many states have pop between x and y in 1970?
    - Given 3 states, which has the lowest median income?
    - What’s the relationship between education and income?
    - List states with pops. 0->x and y->z.
    - What kind of a state is Florida?

We saw this earlier in term
Findings

- Brushing histograms better and more highly rated for more complex discovery tasks
  - Attribute correlation, compare, and trend evaluation
- Dynamic queries better for more simple range specification tasks
  - Single range, multiple ranges, multiple criteria
  - Functioned more as its own infovis tool
  - Functioned more as auxiliary control for other vizs

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)
- Data
  - UN data about countries
- Tasks
  - 24 tasks, 1-3 requires answers per
    Example: Select 2 countries with significant decreases in energy consumption
Results

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

![Accuracy Bar Chart]

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.
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”

Useful Junk?

- Tufte claimed that graphs loaded with chartjunk are no good
- Is that really so?
- How could you test this?
Comparing

Methodology

- Two versions of each chart
- Participant sees one
  - Asked immediate interpretation accuracy questions
  - Asked similar questions again 5 minutes or 2-3 weeks later
**Results**

- No significant difference in immediate interpretation accuracy, or after 5 minute gap
- After 2-3 week gap, recall of chart topic and details was significantly better for chartjunk graphs
- Participants found the chartjunk graphs more attractive, enjoyed them more, and found them easiest and fastest to remember

**Caveats**

- Small datasets
- “Normal” charts were really plain
- No interaction
- How about other added interpretations from the flowery visuals?
- Be careful reading too much into this
More Complex Task Eval

- Consider investigative analysis tasks involving sensemaking, awareness, and understanding
- Research questions
  - How do people use systems?
  - What characteristics matter?
  - What should we measure/observe?
- Exploring methods for utility evaluation

System Examined - Jigsaw
Study Design

• Task and dataset
  – 50 simulated intelligence case reports
    Each a few sentences long
    23 were relevant to plot
  – Identify the threat & describe it in 90 minutes

Source: doc017
Date: Oct 22, 2002

Abu H., who was released from custody after the September 11 incidents and whose fingerprints were found in the U-Haul truck rented by Arnold C. [see doc033] holds an Egyptian passport. He is now known to have spent six months in Afghanistan in the summer of 1999.

Study Design - Settings

1: Paper

2: Desktop

3: Entity

4: Jigsaw
Performance Measures

- Task sheets (like VAST Contest)
  - Three components (relevant people, events, locations)
  - +1 for correct items, -1 for a misidentified items

- Summary narrative
  - Subjective grading from 1 (low) to 7 (high)

- Two external raters
- Normalized, each part equal, mapped to 100-point scale
## Results

<table>
<thead>
<tr>
<th></th>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>22.87</td>
<td>65.00</td>
<td>24.26</td>
<td>90.00</td>
</tr>
<tr>
<td>P2</td>
<td>62.08</td>
<td>67.13</td>
<td>42.13</td>
<td>58.07</td>
</tr>
<tr>
<td>P3</td>
<td>67.13</td>
<td>42.13</td>
<td>29.41</td>
<td>75.20</td>
</tr>
<tr>
<td>P4</td>
<td>Very good</td>
<td>Good</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P5</td>
<td>Very good</td>
<td>Good</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P6</td>
<td>Very good</td>
<td>Good</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P7</td>
<td>Very good</td>
<td>Good</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P8</td>
<td>Very good</td>
<td>Good</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P9</td>
<td>52.23</td>
<td>15.00</td>
<td>29.26</td>
<td>81.19</td>
</tr>
<tr>
<td>P10</td>
<td>Good</td>
<td>Poor</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P11</td>
<td>Good</td>
<td>Poor</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P12</td>
<td>Good</td>
<td>Poor</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>P13</td>
<td>Excellent</td>
<td>Good</td>
<td>Very good</td>
<td>Excellent</td>
</tr>
<tr>
<td>P14</td>
<td>Excellent</td>
<td>Good</td>
<td>Very good</td>
<td>Excellent</td>
</tr>
<tr>
<td>P15</td>
<td>Excellent</td>
<td>Good</td>
<td>Very good</td>
<td>Excellent</td>
</tr>
<tr>
<td>P16</td>
<td>Excellent</td>
<td>Good</td>
<td>Very good</td>
<td>Excellent</td>
</tr>
<tr>
<td>Final Score</td>
<td>49.80</td>
<td>50.19</td>
<td>44.42</td>
<td>79.59</td>
</tr>
<tr>
<td>Performance</td>
<td>Fair</td>
<td>Very good</td>
<td>Fair</td>
<td>Excel- lent</td>
</tr>
<tr>
<td>Average Score</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Daily Score</td>
<td>49</td>
<td>31</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>Documents Viewed</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td># of Queries</td>
<td>19</td>
<td>16</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>First Query</td>
<td>40:49</td>
<td>19:55</td>
<td>2:47</td>
<td>12:41</td>
</tr>
<tr>
<td>Amount of Notes</td>
<td>Many</td>
<td>None</td>
<td>Many</td>
<td>Some</td>
</tr>
<tr>
<td>First Note Taking</td>
<td>9:07</td>
<td>0:05</td>
<td>0:16</td>
<td>1:53</td>
</tr>
<tr>
<td>First Task Sheet</td>
<td>43:20</td>
<td>32:53</td>
<td>70:13</td>
<td>3:25</td>
</tr>
<tr>
<td>Fall 2012</td>
<td>CS 7450</td>
<td>65</td>
<td>66</td>
<td></td>
</tr>
</tbody>
</table>
Jigsaw Usage Patterns

Investigative Strategies

1. Overview, filter and detail (OFD)
2. Build from detail (BFD)
3. Hit the keyword (HTK)
4. Find a clue, follow the trail (FCFT)

P16: “I like this people-first approach. Once I identify key people, then things that are potentially important come up, too. I’m an impatient person and don’t want to read all documents chronologically.”
## Results by Strategy

<table>
<thead>
<tr>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>P5</td>
<td>P9</td>
<td>P13</td>
</tr>
<tr>
<td>P2</td>
<td>P6</td>
<td>P10</td>
<td>P14</td>
</tr>
<tr>
<td>P3</td>
<td>P7</td>
<td>P11</td>
<td>P15</td>
</tr>
<tr>
<td>P4</td>
<td>P8</td>
<td>P12</td>
<td>P16</td>
</tr>
</tbody>
</table>

### Strategy Used

<table>
<thead>
<tr>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFD</td>
<td>OFD</td>
<td>OFD</td>
<td>OFD</td>
</tr>
<tr>
<td>OFD</td>
<td>OFD</td>
<td>OFD</td>
<td>OFD</td>
</tr>
<tr>
<td>OFD</td>
<td>OFD</td>
<td>FCFT</td>
<td>FCFT</td>
</tr>
<tr>
<td>OFD</td>
<td>OFD</td>
<td>FCFT</td>
<td>FCFT</td>
</tr>
<tr>
<td>OFD</td>
<td>OFD</td>
<td>FCFT</td>
<td>FCFT</td>
</tr>
</tbody>
</table>

### Performance

<table>
<thead>
<tr>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair</td>
<td>Very</td>
<td>Good</td>
<td>Excel</td>
</tr>
<tr>
<td>Fair</td>
<td>Very</td>
<td>Good</td>
<td>Excel</td>
</tr>
<tr>
<td>Fair</td>
<td>Very</td>
<td>Good</td>
<td>Excel</td>
</tr>
<tr>
<td>Fair</td>
<td>Very</td>
<td>Good</td>
<td>Excel</td>
</tr>
</tbody>
</table>

### Documents Viewed

<table>
<thead>
<tr>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>49</td>
<td>31</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>31</td>
<td>50</td>
<td>46</td>
<td>23</td>
</tr>
</tbody>
</table>

---

Fall 2012  | CS 7450  | 69
---

---

Fall 2012  | CS 7450  | 70
## Results by Strategy

<table>
<thead>
<tr>
<th></th>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy Used</td>
<td>OFD</td>
<td>OFD</td>
<td>BFD</td>
<td>OFD</td>
</tr>
<tr>
<td>Performance</td>
<td>Fair</td>
<td>Very good</td>
<td>Fair</td>
<td>Excellent</td>
</tr>
<tr>
<td>Documents Viewed</td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>
Results by Strategy

<table>
<thead>
<tr>
<th>Strategy Used</th>
<th>Paper</th>
<th>Desktop</th>
<th>Entity</th>
<th>Jigsaw</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OFD</td>
<td>OFD</td>
<td>OFD</td>
<td>OFD</td>
</tr>
<tr>
<td>Performance</td>
<td>Fair</td>
<td>Very good</td>
<td>Fair</td>
<td>Excelent</td>
</tr>
<tr>
<td>Documents Viewed</td>
<td>50</td>
<td>50</td>
<td>49</td>
<td>31</td>
</tr>
</tbody>
</table>

Tool Design Implications

- Support finding starting points/clues
- Guide the analyst to follow the right trail
- Support different strategies of SM process
- Support smooth transition between SM stages
- Provide a workspace
- Allow flexibility in organizing
- Support to find next steps when dead-end
- Facilitate further exploration
Jigsaw’s Influence

- Supporting different strategies
- Showing connections between entities

- Helping users find the right clue
- Helping users focus on essential information

- Reviewing hypotheses
- Increasing motivation

Evaluation Recommendations

- Compare system usage to traditional methods
- Collect qualitative data, support with quantitative data
- Consider questions to be answered
- Possible metrics
  - Number of documents viewed
  - When note-taking initiated
  - The quantity of representations created
  - Amount of time and effort in organizing
  - Time spent in reading/processing relevant information
How to Evaluate Many Eyes?

- Two main evaluation papers written about system
- Studied use of system, visualizations being created, discussions about system, etc.

---

Paper 1

- Case study of early use
- System uses
  - Visual analytics
  - Sociability
  - Generating personal and collective mirrors
  - Sending a message

Viégas et al
HICSS ’08
Use Characteristics

<table>
<thead>
<tr>
<th>Data Topic/Area</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Society</td>
<td>14.0</td>
</tr>
<tr>
<td>Economics</td>
<td>12.7</td>
</tr>
<tr>
<td>Obscured/Anon</td>
<td>12.4</td>
</tr>
<tr>
<td>Art &amp; culture</td>
<td>10.8</td>
</tr>
<tr>
<td>Web &amp; new media</td>
<td>10.3</td>
</tr>
<tr>
<td>Science</td>
<td>10.0</td>
</tr>
<tr>
<td>Test data</td>
<td>9.5</td>
</tr>
<tr>
<td>Politics</td>
<td>7.4</td>
</tr>
<tr>
<td>Technology</td>
<td>6.6</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comment Type</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td>46.3</td>
</tr>
<tr>
<td>Question</td>
<td>15.8</td>
</tr>
<tr>
<td>Affirmation</td>
<td>13.7</td>
</tr>
<tr>
<td>Hypothesis</td>
<td>11.6</td>
</tr>
<tr>
<td>Socializing</td>
<td>11.6</td>
</tr>
<tr>
<td>System design</td>
<td>11.6</td>
</tr>
<tr>
<td>Data integrity</td>
<td>9.5</td>
</tr>
<tr>
<td>Testing</td>
<td>4.2</td>
</tr>
<tr>
<td>Tips</td>
<td>4.2</td>
</tr>
<tr>
<td>To do</td>
<td>4.2</td>
</tr>
</tbody>
</table>

Fall 2012

Paper 2

- Interview-based study
- Individual phone interviews with 20 users
  - Lots of quotes in paper
- Bloggers vs. regular users
- Also includes stats from usage logs
  - 3069 users
  - 1472 users who uploaded data
  - 5347 datasets
  - 972 users who created visualizations
  - 3449 visualizations
  - 222 users who commented
  - 1268 comments

Danis et al
CHI '08
Findings

- User motivations
  - Analyzing data
  - Broadening the audience, sharing data
- Lots of collaborative discussion
  - Much off the ManyEyes site
- Concerns about data and other eyes

Specific to Infovis?

- How about evaluation techniques specifically focused on infovis?
**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

---

*Start*
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, Clustered parallel coordinates</td>
<td>Brushing, Zooming</td>
</tr>
</tbody>
</table>

### Tools Images

- **Cluster/Treeview**
- **TimeSearcher**
- **GeneSpring**
- **Spotfire**
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
Reconsidering Insight

- Insight with visualization
  - Is not spontaneous “aha!” moments (e.g., in cognitive science)
  - Is knowledge-building and model-confirmation
    Like a substance that people acquire with the aid of systems

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
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
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?!

Project

• Demos on Thursday
• 15 minutes, plan appropriately, leave open-ended exploration time
• Bring page or two describing project

• Video due next Monday
Upcoming

- Review & recap
  - Reading
    Few chapter 13
    Heer et al ’10