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?...”
Reflection

• Creation of new techniques is very important but...
  – It’s also important to know that we’re getting better
  – So, it’s important that we evaluate the visualizations being created

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”

Evaluation – Measures?

- How does one judge the quality of work in Information Visualization?
Evaluation – Measures?

- 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
**Broden 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)

**Evaluation – How?**

- What evaluation techniques should we use?
  - (Channel your 3750/6750 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
• What do we measure?
  – Performance, time, errors, ...

• Strengths, weaknesses?
Subjective Assessments

- Find out people’s subjective views on tools
  - Was it enjoyable, confusing, fun, difficult, ...?
- This kind of personal judgment strongly influence use and adoption, sometimes even overcoming performance deficits

Qualitative, Observational Studies

- Watch systems being used (you can learn a lot)
- Is it being used in the way you expected?
- Ecological validity
- Can suggest new designs and improvements
  - (Channel 6455 knowledge)
Running Studies

• Beyond our scope here
• You should learn more about this in 6750 or 6455

Evaluating UI vs. InfoVis

• Seems comparable but...
• What are some differences?
Evaluating UI vs. InfoVis

- 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
Plaisant Paper

• Discuss
• Challenges identified?
• Possible next steps?

Examples

• Let’s look at a few example studies that attempt to evaluate different InfoVis systems
• For multiple examples, see journal issue whose focus is Empirical Studies of Information Visualizations
• Also, now see the BELIV workshops
Commercial Tools Eval

- Empirical study of 3 InfoVis tools
  - Eureka, Spotfire, InfoZoom

- Methodology
  - 3 data sets
  - 83 students
  - Within subjects, 30 minutes per tool

More Methodology

- Tasks
  - Very much from the 10 low-level tasks type of questions (specific, not exploratory)

- Measurements
  - Correctness
  - Time
Results

• Time:
  – InfoZoom – 80 seconds
  – Spotfire – 107 seconds
  – Eureka – 110 seconds

• Correctness
  – Spotfire – 75%
  – Eureka – 71%
  – InfoZoom – 68%

Findings

• Interaction Problems
  – Eureka
    Confusion by hidden labels, problems with 3 or more vars., correlation errors
  – InfoZoom
    Correlations errors
  – Spotfire
    Cognitive set-up costs, scatterplot bias
Findings

- Success depends on
  - Properties of visualization
  - Operations that can be performed on visualization
  - Concrete implementation of paradigm
  - Visualization-independed usability problems

- I would have liked even more discussion on how tools assisted with different classes of user tasks

Space-Filling Hierarchy Views

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

Stasko, Catrambone, Guzdial and McDonald
IJHCS '00
Tools Compared

Treemap
Spring 2009

SunBurst
CS 4460/7450

Hierarchies Used

• Four in total
  Small Hierarchy
  (~500 files)
  Large Hierarchy
  (~3000 files)

• 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 | SB B, TM A | 32 on small hierarchies |
| TM A, SB B | TM B, SB A | 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>Hierarchy A</th>
<th>Tool (n = 8)</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>1</td>
<td>9.88</td>
<td>(1.23)</td>
</tr>
<tr>
<td>SB</td>
<td>1</td>
<td>12.88 (1.96)</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td>2</td>
<td>12.25 (1.75)</td>
<td></td>
</tr>
<tr>
<td>SB</td>
<td>2</td>
<td>12.63 (2.00)</td>
<td></td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td>11.06 (2.79)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td>12.75 (1.91)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hierarchy B</th>
<th>Tool (n = 8)</th>
<th>Phase</th>
<th>Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM</td>
<td>1</td>
<td>11.39 (2.14)</td>
<td></td>
</tr>
<tr>
<td>SB</td>
<td>1</td>
<td>10.38 (1.69)</td>
<td></td>
</tr>
<tr>
<td>TM</td>
<td>2</td>
<td>10.75 (2.77)</td>
<td></td>
</tr>
<tr>
<td>SB</td>
<td>2</td>
<td>11.50 (2.00)</td>
<td></td>
</tr>
<tr>
<td>TM (collapsed across phase)</td>
<td>11.13 (2.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SB (collapsed across phase)</td>
<td>10.94 (1.88)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Correct task completions (out of 16 possible)
Large Hierarchy

Correct task completions (out of 16 possible)

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**

  Li & North
  InfoVis '03

**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 auxiliary control for other vizs

Spring 2009    CS 4460/7450  37
More Recently

- What have been the hot topics in the infovis evaluation community the last few years?

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
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 Functional Genomics</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>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
Results

Spring 2009 CS 4460/7450 45

Workshop focused on this topic

BELIV

Nice locations!

Workshop focused on this topic

Spring 2009 CS 4460/7450 46
MILC

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

Shneiderman & Plaisant
BELIV ’06

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
- Establish schedule of observation & interviews
- Instrument tool to record usage
- Provide attractive log book for comments
- 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
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>

Spring 2009 CS 4460/7450 55

## 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

Spring 2009 CS 4460/7450 56
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

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?!
Upcoming

- Animation
  - Paper
    Heer & Robertson