InfoVis Evaluation

Dr. Michele C. Weigle

http://www.cs.odu.edu/~mweigle/CS795-F12/

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, …
  - "Isn't my new visualization cool?"

Content courtesy of John Stasko, Ga Tech

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?

- 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

Evaluation
Measures?

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

- 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)

Content courtesy of John Stasko, Ga Tech

BELIV

Beyond Time and Errors: Novel Evaluation Methods for Vis

Held every 2 years

http://www.beliv.org/

Evaluation
How?

- What evaluation techniques should we use?

Evaluation in HCI

- Takes many different forms
  - Qualitative, quantitative, objective, subjective, controlled experiments, interpretive observations, …
- So, which ones are best for evaluating InfoVis systems?

Experiment Types

Plaisant
AVI '04

- Controlled experiments comparing design elements
  - ex: compare widgets, mappings of data to display
- Controlled experiments comparing tools
- Case studies of tools in realistic settings (observation)
- Usability evaluation (subjective)

Evaluation Types
Controlled Experiments

- Good for measuring performance or comparing multiple techniques or tools
- Metrics
  - task correctness
  - time
- Often also ask for subjective evaluation
  - was it enjoyable, confusing, fun, difficult, …?
  - strongly influences use and adoption, sometimes even overcoming performance deficits

Content courtesy of
John Stasko, Ga Tech

Evaluation Types
Case Studies - Qualitative, Observational

- Watch systems being used (you can learn a lot)
  - is it being used in the way you expected?
- Ecological validity
  - does the setting of the study replicate the real-world environment?
- Can suggest new designs and improvements
- But, results may not be generalizable

Case Studies
Evaluation Methodology

Shneiderman & Plaisant
BELIV '06

- Multi-dimensional In-depth Long-term Case Study (MILC)
- 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

Case Studies
MILC Guidelines

Shneiderman & Plaisant
BELIV '06

- 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

Usability Evaluation
Usability vs. Utility

- Usability
  - the ease of use and learnability of a product
  - measured in a lab
- Utility
  - fitness for some purpose or worth to some end
  - needs demonstration in the real world
- Can think of visualizations that are very *usable* but not *useful*

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
  - domain knowledge and situated use is required
- UI can heavily influence utility and value of visualization technique

Content courtesy of John Stasko, Ga Tech

Why Challenging?

Plaisant AVI '04

What characteristics of info vis make evaluation challenging?

- Users may need to look at the same data from different perspectives over a long time
- Users may need to formulate and answer questions they didn't anticipate having before looking at the visualization
- Discoveries can have a huge impact, but occur rarely

How to Improve Evaluations?

Plaisant
AVI '04

- Repositories of data and tasks
  - visual analytics benchmark repository
    - http://hcil.cs.umd.edu/localphp/hcil/vast/archive/
- Case studies and success stories
- Toolkits and development tools

Running Studies

- Beyond our scope here
- But, this brings up related Psychology courses at ODU
  - PSYCH 662 - Human-Computer Interface Design
  - PSYCH 867 - Human Performance Assessment
  - PSYCH 872 - Methods, Measures, Techniques, and Tools in Human Factors
  - PSYCH 875 - Advanced Visual Perception and Visual Displays

Crowdsourced Evaluation

Heer and Bostock SIGCHI 2010

Kosara and Ziemkiewicz BELIV 2010

- Amazon's Mechanical Turk (MTurk)
  - https://www.mturk.com/mturk/welcome
  - a fake chess-playing machine constructed in the late 18th century, really a person hiding inside
- Post assignments and have people complete the tasks

MTurk

Heer and Bostock SIGCHI 2010

Kosara and Ziemkiewicz BELIV 2010

- *Requesters* post jobs
  - Human Intelligence Tasks (*HITs*)
- Pool of workers, aka *Turkers*
- Each HIT has a *reward*
  - typically 1-10 cents
- Each HIT has a set number of *assignments*, max number of Turkers

MTurk
Example Tasks

Heer and Bostock
SIGCHI 2010

1) Identify the smaller of the two marked values
2) "Make a quick visual judgment" to estimate the percentage the smaller was of the larger

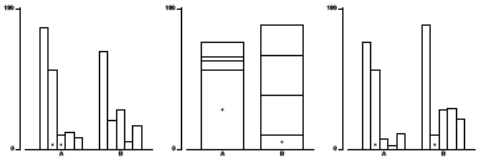


Figure 1: Stimuli for judgment tasks T1, T2 & T3. Subjects estimated percent differences between elements.

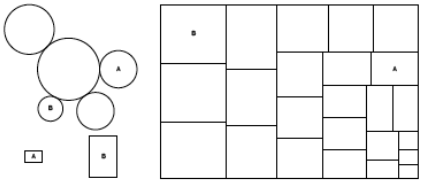


Figure 2: Area judgment stimuli. Top left: Bubble chart (T7), Bottom left: Center-aligned rectangles (T8), Right: Treemap (T9).

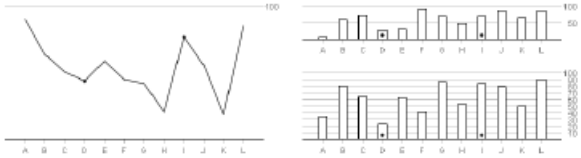


Figure 10: Experiment 3 stimuli varying chart type, chart height, and gridline spacing.

MTurk
Example Tasks

Kosara and Ziemkiewicz
BELIV 2010

1) Enter estimated percentage
2) Indicate their confidence as low, medium, high

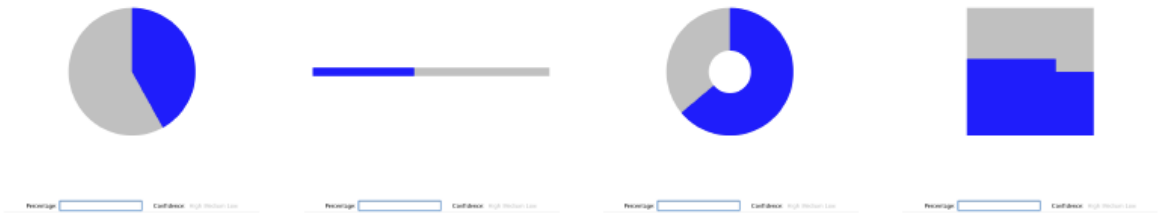


Figure 2: The four visualization types tested in Study II: pie chart, bar chart, donut chart, and square pie chart.

MTurk

Heer and Bostock
SIGCHI 2010

Kosara and Ziemkiewicz
BELIV 2010

- Advantages
  - convenient labor pool
  - wide range of test subjects
  - with careful design and applicable task, can replicate lab experiments
  - feasible to conduct large scale studies
- Disadvantages
  - lack of control
  - technical issues with display of experiment
  - should develop qualification tests to ensure Turkers understand the question

Examples

- Controlled Experiments
  - Treemap vs. SunBurst
  - Animation
- Case Studies
  - SocialAction
  - ManyEyes

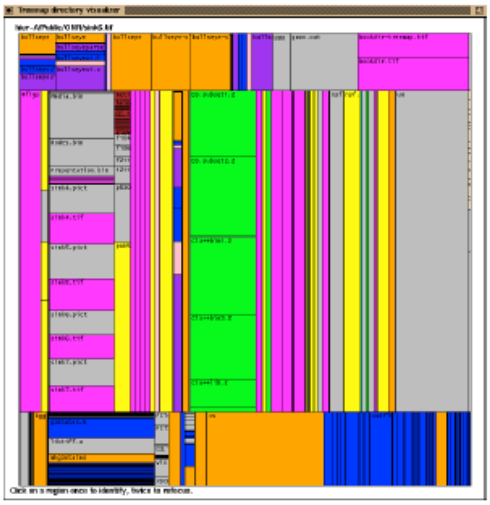
Controlled Experiments
Which Technique is Best?

Stasko et al
IJHCS'00

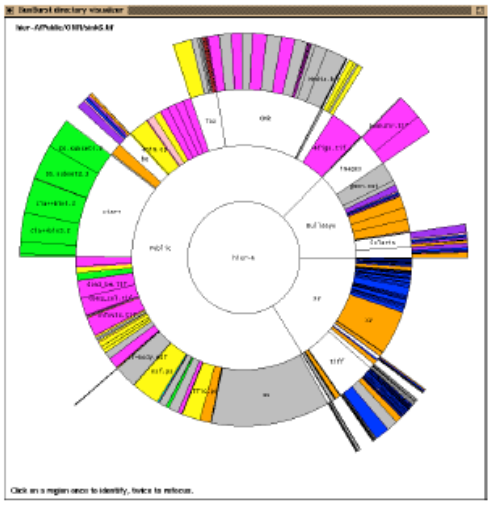
- Space-filling hierarchical views
- Compare Treemap and SunBurst with users performing typical file/directory-related tasks
- Evaluate task performance on both correctness and time

Treemap vs. SunBurst

Stasko et al
IJHCS'00



Treemap

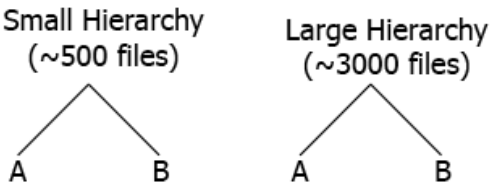


SunBurst

Treemap vs. SunBurst
Hierarchies Used

Stasko et al IJHCS'00

- Four in total



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

Content courtesy of John Stasko, Ga Tech

Treemap vs. SunBurst
Methodology

Stasko et al IJHCS'00

- 60 participants
- Participant only works with a small or large hierarchy in a session
  - 32 on small hierarchies
  - 28 on large hierarchies
- Training at beginning to learn tool
- Vary order across participants
  - SunBurst A, TreeMap B
  - TreeMap A, SunBurst B
  - SunBurst B, TreeMap A
  - TreeMap B, SunBurst A

Content courtesy of John Stasko, Ga Tech

Treemap vs. SunBurst

Stasko et al IJHCS'00

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)

Content courtesy of John Stasko, Ga Tech

Treemap vs. SunBurst

Stasko et al IJHCS'00

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

Content courtesy of John Stasko, Ga Tech

Small Hierarchy

Correct Task Completions (out of 16)

Stasko et al IJHCS'00

Hierarchy A

Tool	Phase	Correct
TM ($n = 8$)	1	9.88 (3.23)
SB ($n = 8$)	1	12.88 (1.96)
TM ($n = 8$)	2	12.25 (1.75)
SB ($n = 8$)	2	12.63 (2.00)
TM (collapsed across phase)		11.06 (2.79)
SB (collapsed across phase)		12.75 (1.91)

Hierarchy B

Tool	Phase	Correct
TM ($n = 8$)	1	11.50 (2.14)
SB ($n = 8$)	1	10.38 (1.69)
TM ($n = 8$)	2	10.75 (2.77)
SB ($n = 8$)	2	11.50 (2.00)
TM (collapsed across phase)		11.13 (2.42)
SB (collapsed across phase)		10.94 (1.88)

Content courtesy of John Stasko, Ga Tech

Large Hierarchy

Correct Task Completions (out of 16)

Stasko et al IJHCS'00

Hierarchy A

Tool	Phase	Correct
TM ($n = 7$)	1	8.71 (1.60)
SB ($n = 7$)	1	11.43 (1.27)
TM ($n = 7$)	2	11.57 (1.27)
SB ($n = 7$)	2	11.00 (2.16)
TM (collapsed across phase)		10.14 (2.03)
SB (collapsed across phase)		11.21 (1.72)

Hierarchy B

Tool	Phase	Correct
TM ($n = 7$)	1	8.29 (2.14)
SB ($n = 7$)	1	11.14 (2.67)
TM ($n = 7$)	2	10.86 (1.57)
SB ($n = 7$)	2	11.00 (2.00)
TM (collapsed across phase)		9.57 (2.24)
SB (collapsed across phase)		11.07 (2.27)

Content courtesy of John Stasko, Ga Tech

Treemap vs. SunBurst
Performance Results

Stasko et al IJHCS'00

- 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!

Content courtesy of John Stasko, Ga Tech

Treemap vs. SunBurst
Subjective Preferences

Stasko et al IJHCS'00

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

Content courtesy of John Stasko, Ga Tech

Examples

- Controlled Experiments
  - Treemap vs. SunBurst
  - Animation
- Case Studies
  - SocialAction
  - ManyEyes

Is Animation Really Good?

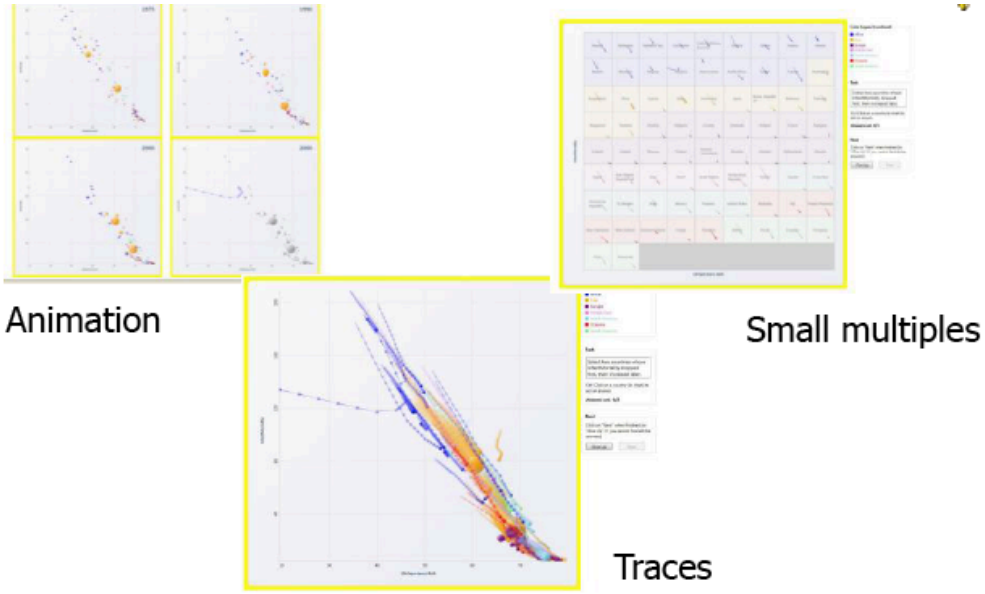
Robertson et al
TVCG (InfoVis) '08

- 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

Content courtesy of
John Stasko, Ga Tech

Animation
Visualizations Studied

Robertson et al
TVCG (InfoVis) '08



Content courtesy of
John Stasko, Ga Tech

Animation
Experiment Design

Robertson et al
TVCG (InfoVis) '08

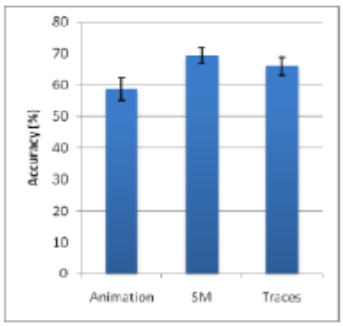
- 3 (animation types) x 2 (data size: small & large) x 2 (presentation vs. analysis)
  - analysis - full use of interaction techniques
  - presentation - passive, ala conference presentation
- Data
  - UN data about countries
- Tasks
  - 24 tasks, 1-3 required answers per
    - example: select 2 countries with significant decreases in energy consumption

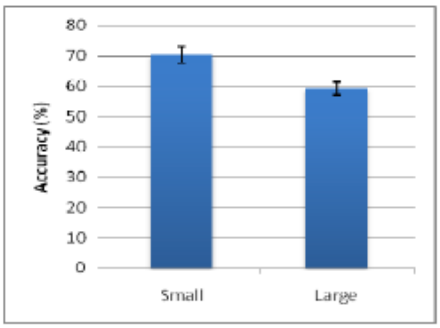
Content courtesy of
John Stasko, Ga Tech

Results
Accuracy

Robertson et al
TVCG (InfoVis) '08

Measured as percentage correct
65% overall (pretty tough)





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

Content courtesy of
John Stasko, Ga Tech

Results
Speed

Robertson et al
TVCG (InfoVis) '08

- Presentation
  - animation faster than small multiples and traces
  - 15.8 secs vs. 25.3 secs vs. 27.8 secs.
- Analysis
  - animation slower than small multiples and traces
  - 83.1 secs. vs. 45.69 secs. vs. 55.0 secs.

Content courtesy of
John Stasko, Ga Tech

Animation
Discussion

Robertson et al
TVCG (InfoVis) '08

- People rated animation more fun, but small multiples was more effective
- As data grows, accuracy becomes an issue
  - traces and animation get cluttered
  - small multiple gets tiny
- Animation
  - "fun", "exciting", "emotionally touching"
  - confusing, "the dots flew everywhere"

Content courtesy of
John Stasko, Ga Tech

Examples

- Controlled Experiments
  - Treemap vs. SunBurst
  - Animation
- Case Studies
  - SocialAction
  - ManyEyes

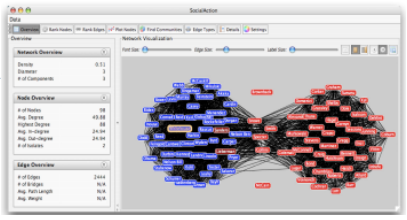
SocialAction

Perer & Shneiderman
CHI '08

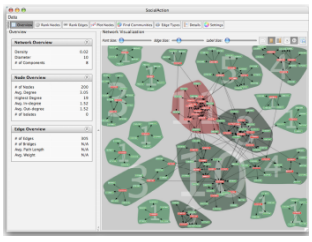
- Evaluation inspired by MILC ideas goals
  - Interview (1 hour)
    - understand intention of participant
  - Training (2 hours)
  - Early use (2-4 weeks)
    - use own data in their workplace, interviewed each week, developers may modify software to meet needs
  - Mature use (2-4 weeks)
    - hands-off observation, no modification of software
  - Outcome (1 hour)
    - exit interview, how software impacted research

Content courtesy of John Stasko, Ga Tech

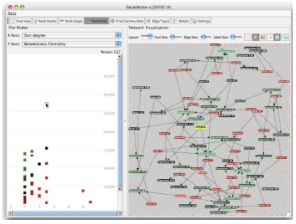
Methodology
Four Case Studies



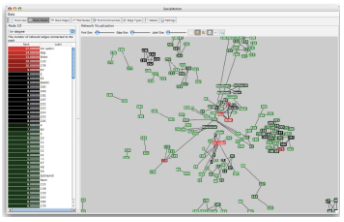
- Senatorial voting patterns



- Medical research knowledge discovery



- Hospital trustee networks



- Group dynamics in terrorist networks

Content courtesy of John Stasko, Ga Tech

SocialAction
Outcomes

Perer & Shneiderman
CHI '08

- Senatorial voting patterns
  - tight integration of statistics and visualization allowed user to uncover findings and communicate them
- Medical research knowledge discovery
  - users better able to understand their retrieval algorithm
- Hospital trustee networks
  - users had better understanding of hospital network, but lacked certain features, such as additional statistical measures, map-editing for nodes, saving edits
- Group dynamics in terrorist networks
  - allowed exploration in new, interesting ways, integrating SocialAction into their online global terrorism database

Examples

- Controlled Experiments
  - Treemap vs. SunBurst
  - Animation
- Case Studies
  - SocialAction
  - ManyEyes

How to Evaluate Many Eyes?

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

ManyEyes
Paper 1

Viégas et al
HICSS '08

- Case study of early use
- System uses
  - visual analytics
  - sociability
  - generating personal and collective mirrors
  - sending a message

ManyEyes
Evaluation

Viégas et al
HICSS '08

- Quantitative, objective
- 1895 posts as of March 2007
- Wide variety of topics of visualizations and motivations for creating visualizations
- Does seem to be fostering discussion

Content courtesy of John Stasko, Ga Tech

ManyEyes
Use Characteristics

Viégas et al
HICSS '08

Data Topic/Area	Percentage
Society	14.0
Economics	12.7
Obscured/Anon	12.4
Art & culture	10.8
Web & new media	10.3
Science	10.0
Test data	9.5
Politics	7.4
Technology	6.6
...	

Comment Type	Percentage
Observation	46.3
Question	15.8
Affirmation	13.7
Hypothesis	11.6
Socializing	11.6
System design	11.6
Data integrity	9.5
Testing	4.2
Tips	4.2
To do	4.2

Content courtesy of John Stasko, Ga Tech

ManyEyes
Paper 2

Danis et al.
CHI '08

- 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

Content courtesy of John Stasko, Ga Tech

ManyEyes
Evaluation

Danis et al.
CHI '08

- Qualitative, subjective
- In-depth interviews with some ME users
- Visualizations used largely for communication and collaboration (not necessarily analysis)
  - privacy and audience management a concern
- Highlights a number of interesting, non-expected uses of the technology

Content courtesy of John Stasko, Ga Tech

ManyEyes
Findings

Danis et al.
CHI '08

- User motivations
  - analyzing data
  - broadening the audience, sharing data
- Lots of collaborative discussion
  - much off the ManyEyes site
- Concerns about data and other eyes

Content courtesy of
John Stasko, Ga Tech

Examples

- Controlled Experiments
  - Treemap vs. SunBurst
  - Animation
- Case Studies
  - SocialAction
  - ManyEyes

InfoVis Evaluation
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

InfoVis Evaluation
Challenges

- 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?

InfoVis Evaluation
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
  - info vis most useful in exploratory scenarios when you don't know what task or goal is
    - so how to measure that?!

Content courtesy of John Stasko, Ga Tech

Coming Up

- Nov 29 - WWW
- Dec 4 - Big Data
- Dec 6 - Course Wrap-Up
- Fri, Dec 7 - Final Report due
- Sat, Dec 8 - Final Demos - 3:45-6:45pm