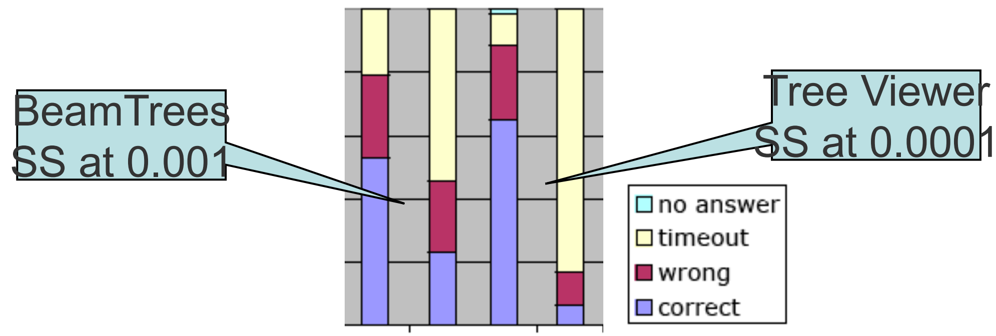
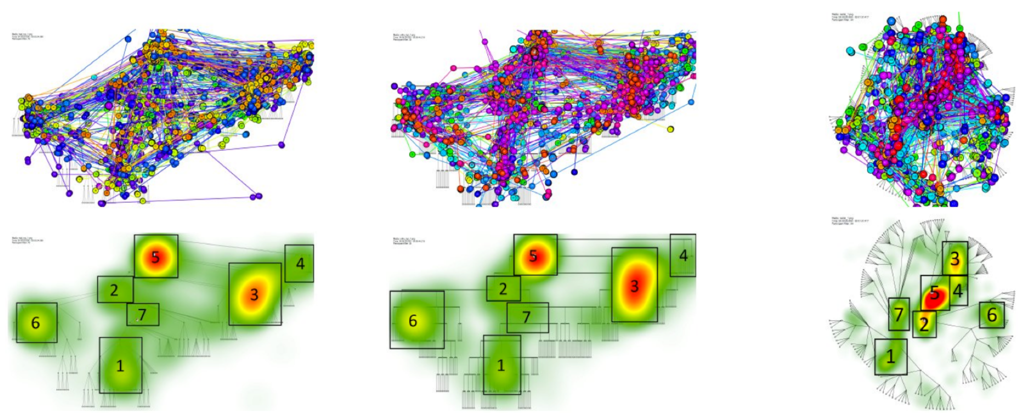


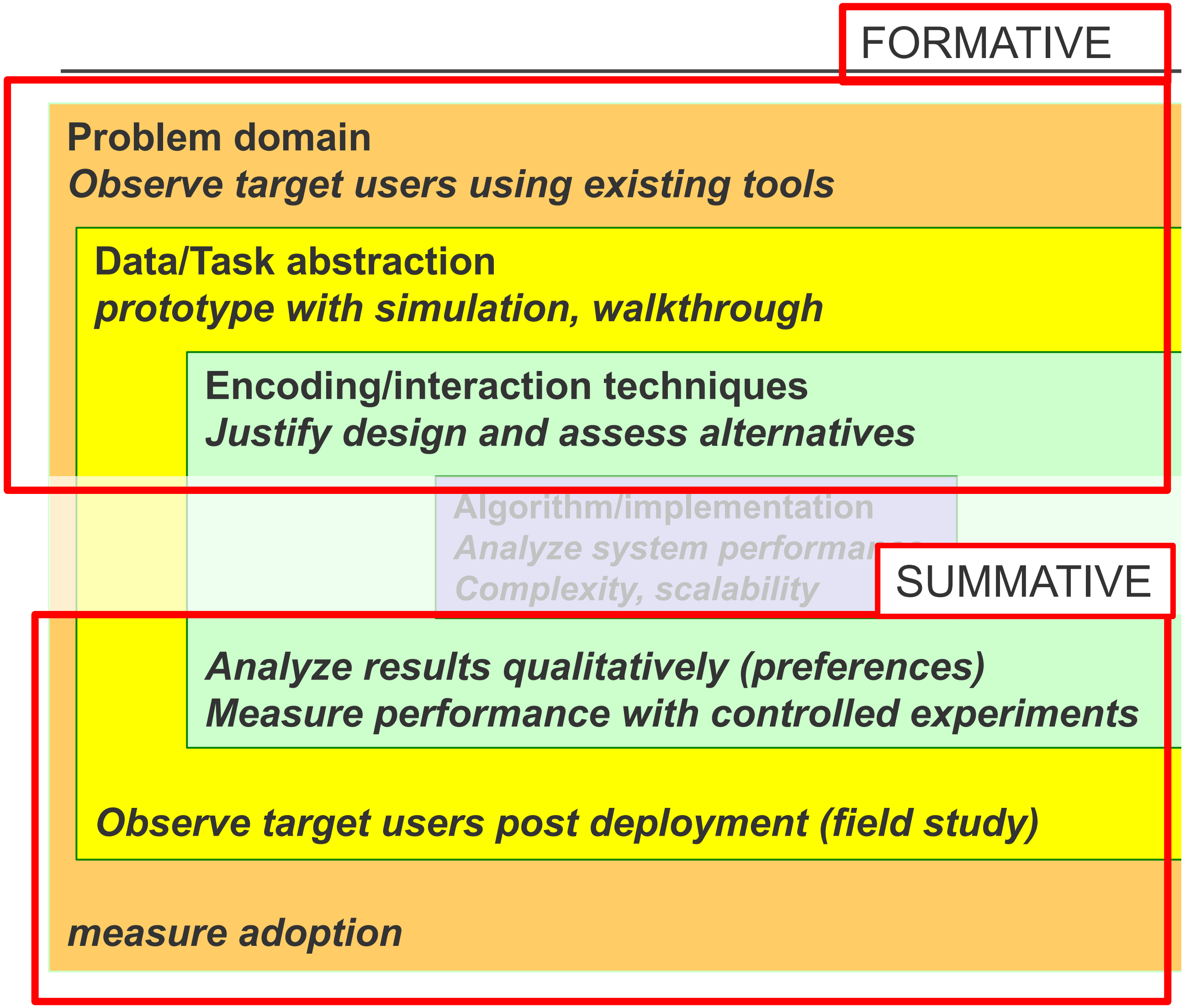
The What, Where, When, Why and How of Evaluating Visualizations

IAT 814



SFU

Munzner's nested model



The general process

- Identify right evaluation questions
- Choose right variables (??) to evaluate
- Pick appropriate tasks, users or data sets
- Choose appropriate method(s)

- Why evaluate?
  - What do we want to find out, and why?
  - How will we use it?

- When to do it
- What to evaluate

Challenge 1

Evaluation : define problem and metrics

- Identify problem
  - Right tasks, environment?
- Develop new design?
  - New problem?
- Improve existing design or select better candidate?
  - Tools or elements
  - Usability

- The purpose of computing is insight, not numbers - Richard Hamming
  - How do we measure insights?
  - Are some insights better than others?

Choose method(s)

Traditionally

- Usability tests
- Controlled experiments of design elements
- Controlled studies of tools and idioms
- Case studies

Plaisant, C. "The challenge of information visualization evaluation." Proceedings of the working conference on Advanced visual interfaces. ACM, 2004.

Extended by:

- Heuristic/model analysis
- Abstract task evaluation in context
- Grounded evaluation (from grounded theory)
- Scenario based

Isenberg, Petra, et al. "Grounded evaluation of information visualizations." 2008 Workshop on BEyond time and errors: novel evaluation methods for Information Visualization. ACM, 2008.

Lam, Heidi, et al. "Seven guiding scenarios for information visualization evaluation." (2011).

The Problem of Good, Better, Best (usability)

Challenge 2

- There is no "proof of optimality"
- Even if we find all info that we seek – could be faster with another viz
- Maybe another insight was there?
- Settle for "Good" or "Better" rather than "Best"

Depends on

- The data
- The user
  - Domain knowledge
  - Tool knowledge
- The environment
- The tasks
- The methods….

Challenge 4

The Tool vs. the Visualization vs. the context

- Usability ≠ Usefulness
- If viz hard to manage, user may not use
  - Maybe the visualization is perfect for the task.
- Focus is NOT on usability of the viz tool
  - Assume it is easy to learn and use

  ⇒ Apply ALL usability tools/methods!

  ⇒ Or

  ⇒ Assess usability of the tool separately from the utility/effectiveness of the design elements

Challenge 5

Compare Best Apple to Best Orange: the trap of the comparative study

- Compare two visualizations

  => Each must be good as possible
- Great implementation of a low-effectiveness visualization may perform better than a poor implementation of a high-effectiveness visualization

Challenge 6

The Comfort Trap

- Users tend to stick with default settings and visualizations, even though others would be better
  - Example: Kobsa, *An empirical evaluation of three commercial information visualization systems*, InfoVis 2001
  - Seen in other domains as well

- Comfort over-rules performance: users think they are doing well even when they are not [Wakeling 2014]

Purpose of Evaluation: Insight, not Numbers

Challenge 7

- Good to know that A is better than B
- But MUCH better to know WHY A is better than B
  - => Ground experiments in theory
  - => Get inside users' heads

Subjective Evaluation

- The challenges of evaluation
- Subjective evaluation
  - both qualitative and quantitative methods
- Evaluation based on expert models and heuristics
- Experimental evaluation
- Long-term Evaluation
- Conclusions

Subjective Evaluation approaches

- User experience
- Personal preference
- Aesthetic judgment

- Initially: ratings, focus groups, interviews, surveys
  - E.g: animation: people "like it better"

- Currently: grounded methods, Holistic evaluation
  - contextual inquiry
  - Goals, outcomes, strategies

Potential Criteria

- Rationale-based Tasks
  - Expose uncertainty
  - Concretize relationships
  - Formulate cause and effect

Amar and Stasko, *A Knowledge Task-Based Framework for Design and Evaluation of Information Visualizations*,, InfoVis 2004

Isenberg, Petra, et al. "Grounded evaluation of information visualizations." 2008 Workshop on BEyond time and errors: novel evaluation methods for Information Visualization. ACM, 2008.

- Worldview-based Tasks
  - Determination of domain parameters
  - Multivariate explanation
  - Confirm hypotheses

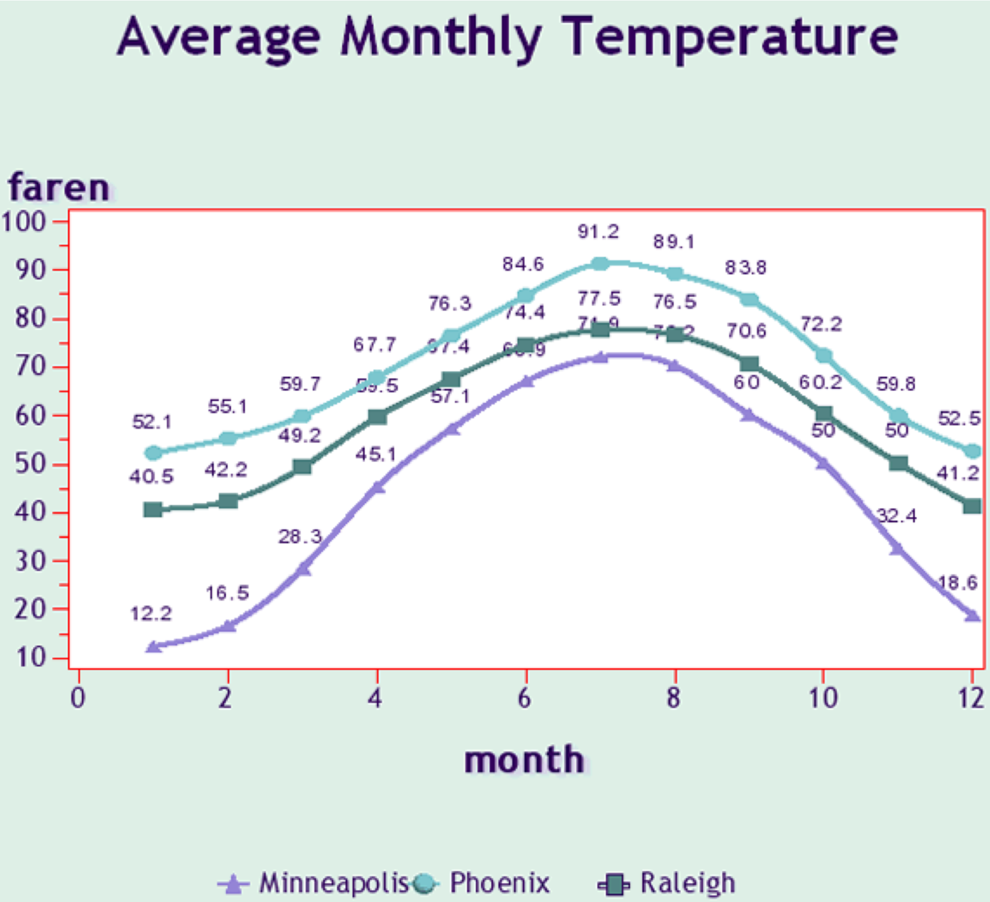
Pros/Cons (+/-)

+ Fast
+ Have rationale basis
- Still based on subjective (but informed) judgments

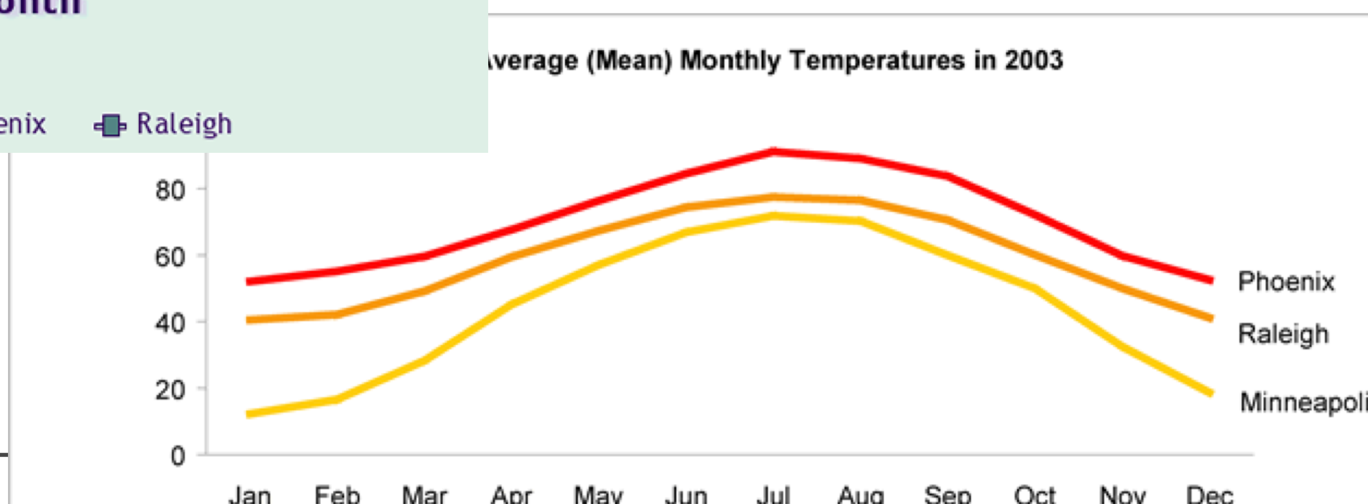
Heuristics and models

- Visualization principles of design
- Performance and cognitive models
  - Fitts' Law of interaction times
  - Lohse, *A Cognitive Model for the Perception and Understanding of Graphics*, CHI 1991)
  - Eye movements (saccades)
  - Word recognition
  - Interaction time

Analysis based on expert heuristics



http://www.perceptualedge.com/example2.php



	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Phoenix	52.1	55.1	59.7	67.7	76.3	84.6	91.2	89.1	83.8	72.2	59.8	52.5
Raleigh	40.5	42.2	49.2	59.5	67.4	74.4	77.5	76.5	70.6	60.2	50.0	41.2
Minneapolis	12.2	16.5	28.3	45.1	57.1	66.9	71.9	70.2	60.0	50.0	32.4	18.6

SFU

Pros/Cons

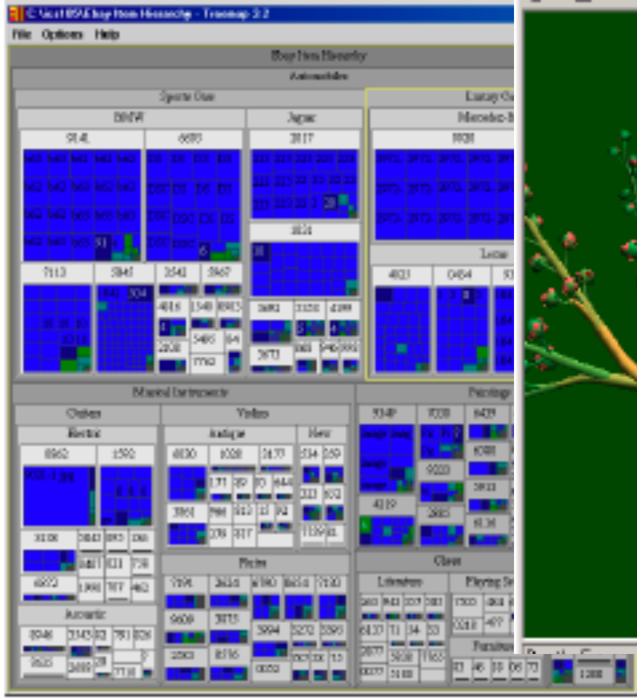
- Cheaper than experiments
- Fast
- Relies on expert
- Relies on standard, or at least well-accepted, models of task and environment and practice

Experimental Evaluation

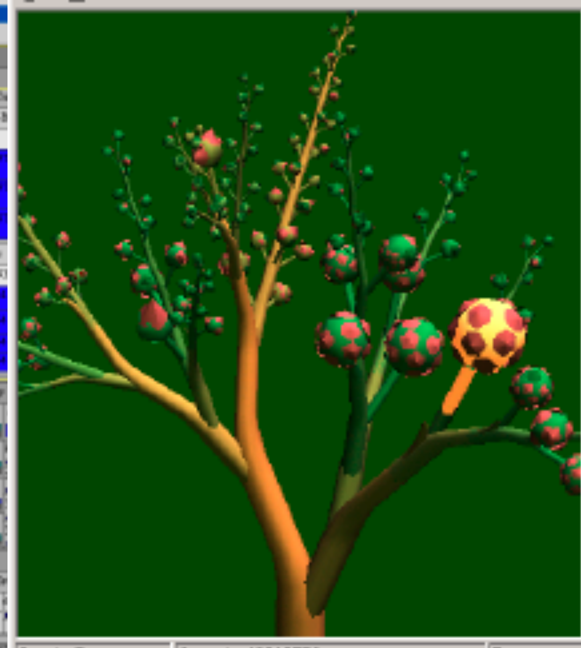
- The challenges of evaluation
- Subjective evaluation
- Evaluation based on model of cognition & perception
- Experimental evaluation
- Long-term Evaluation

Experimental Comparison of Tree Visualization Systems

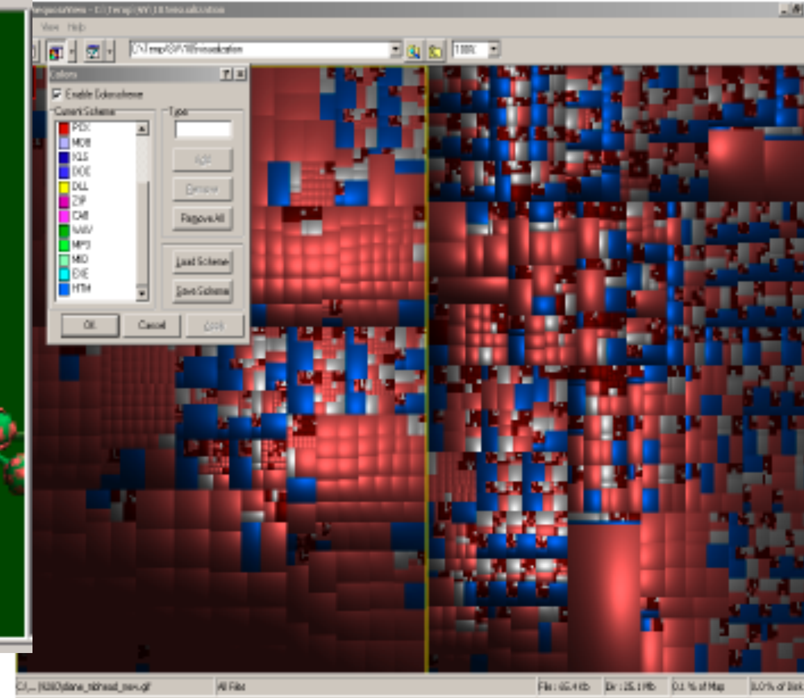
- Kobsa, *User Experiments with Tree Visualization Systems*, InfoViz 2004
- Compares 5 Tree Visualization Systems + Windows file browser
- Quantitative - measure results
- Qualitative - to (partially) understand results



Treemap 3.2

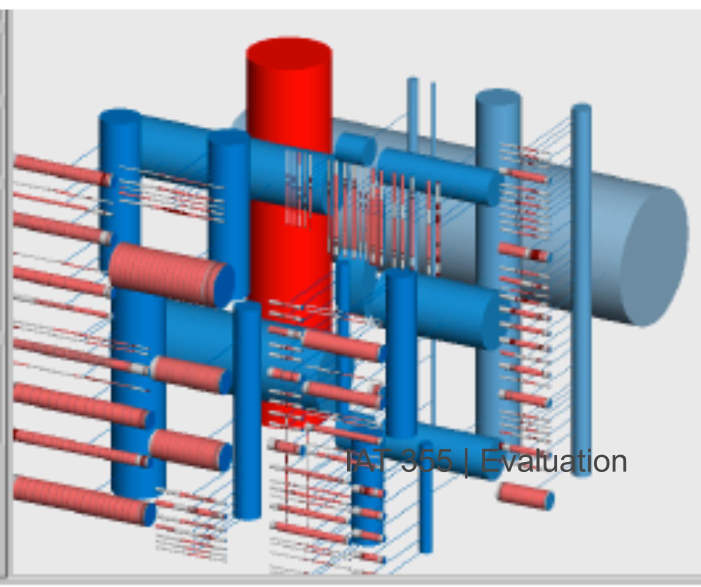


Tree Viewer

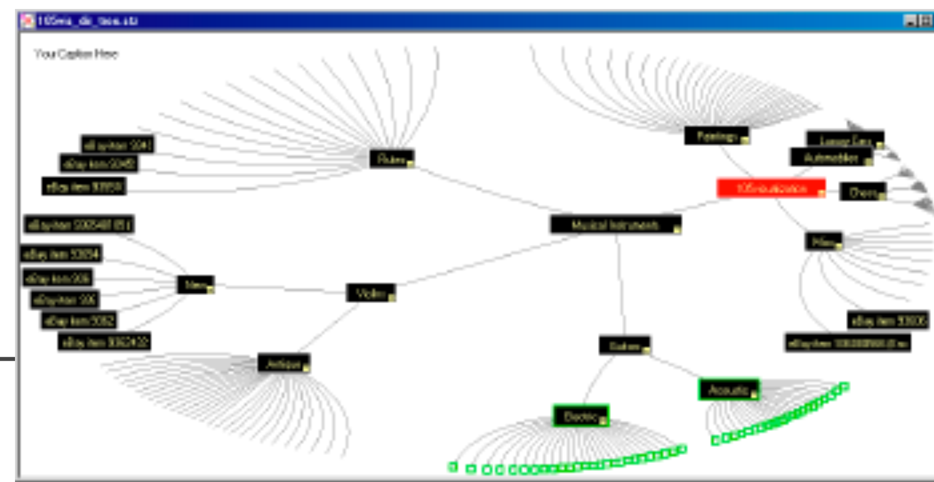


SequoiaView

BeamTrees



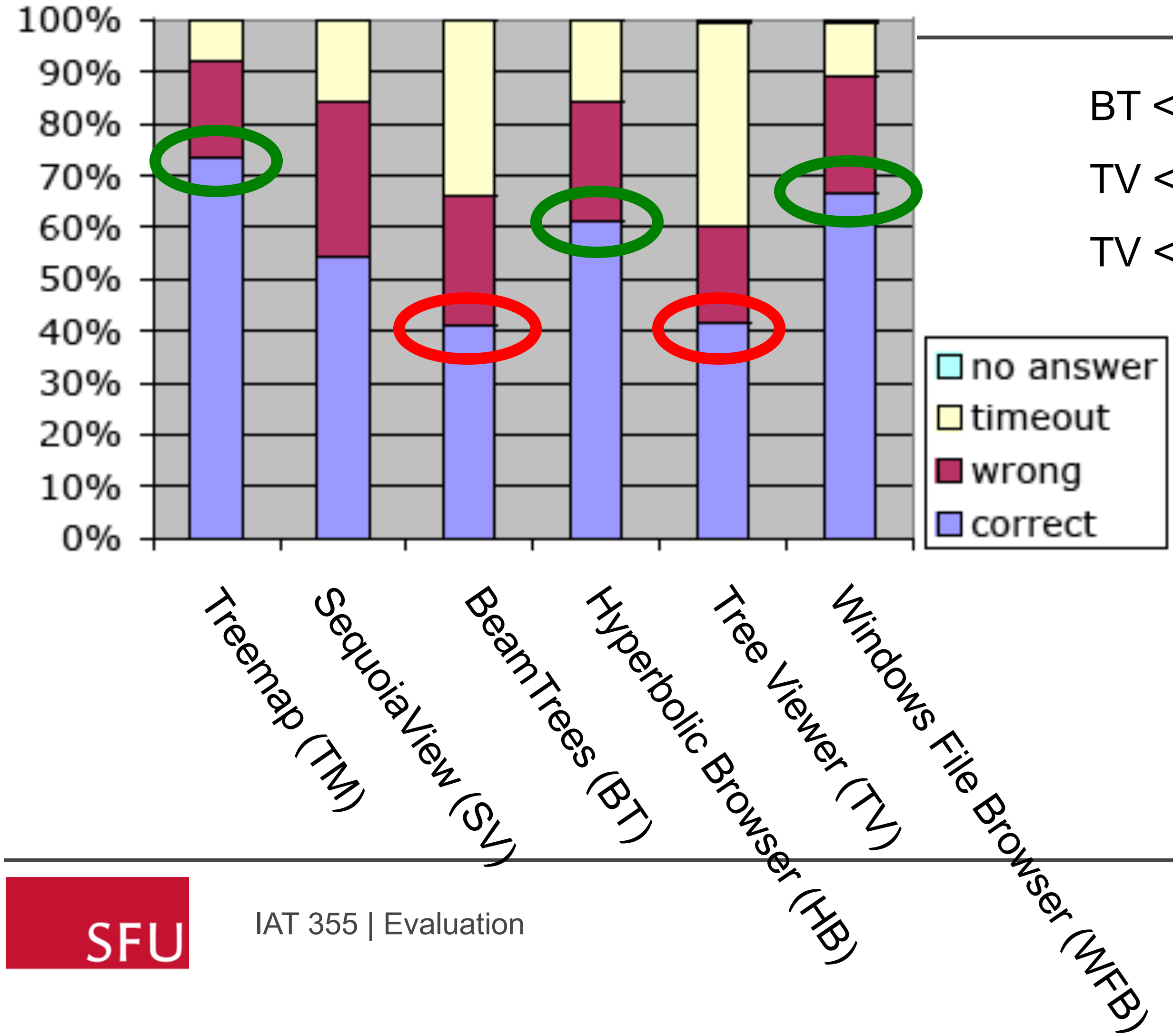
Hyperbolic Browser



Controlled Experimental Design

- Each subject performs 15 tasks
  - 9 tasks concern directory structure
    - Maximum depth of hierarchy?
    - Tree balanced or unbalanced?
  - 6 tasks concern file or directory attributes
    - Find file with name= xxxx
    - Find name of largest file

Results - Correct Answers

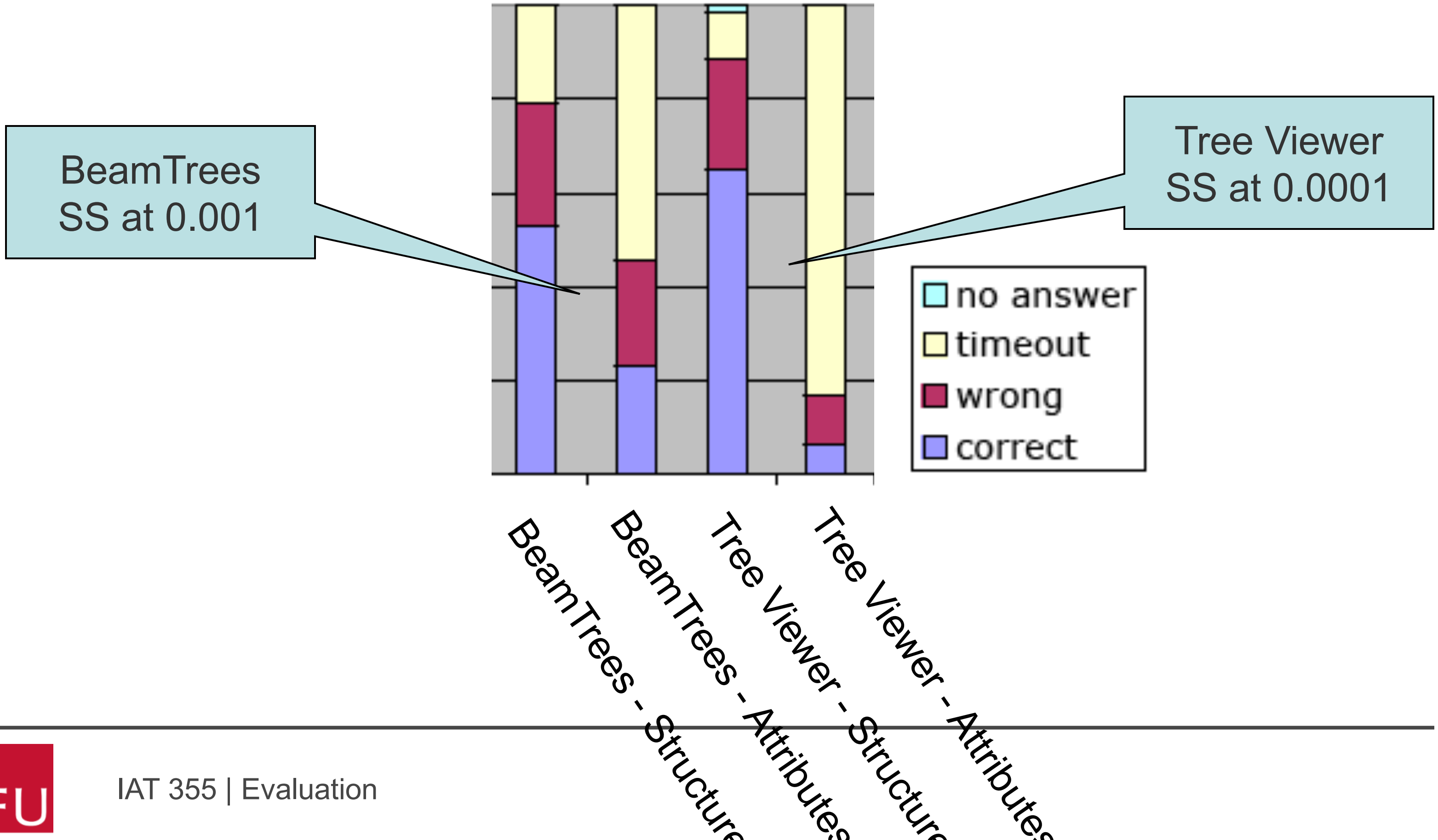


BT << TM, SV, HB, WFB (1%)

TV << TM, HB, WFB (1%)

TV < SV (5%)

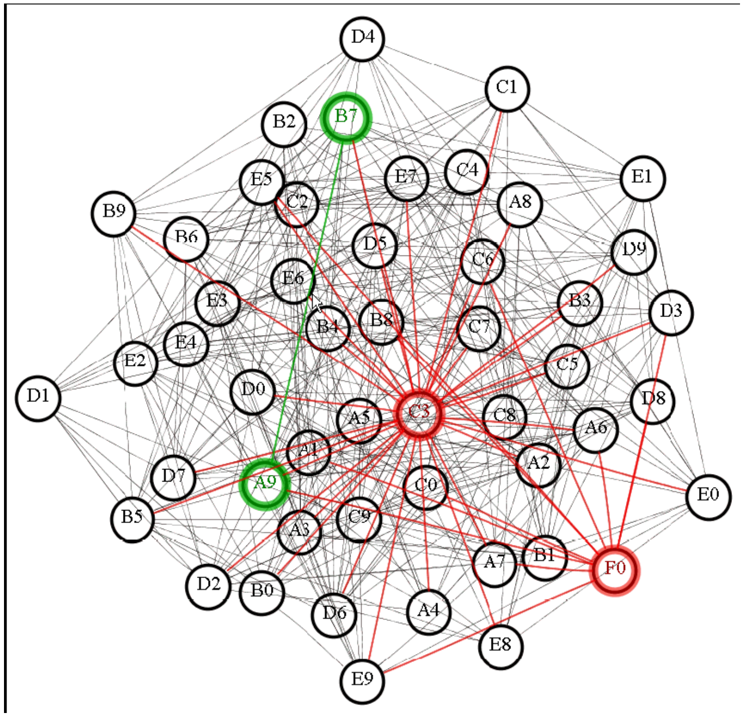
Results - Effect of Task Type

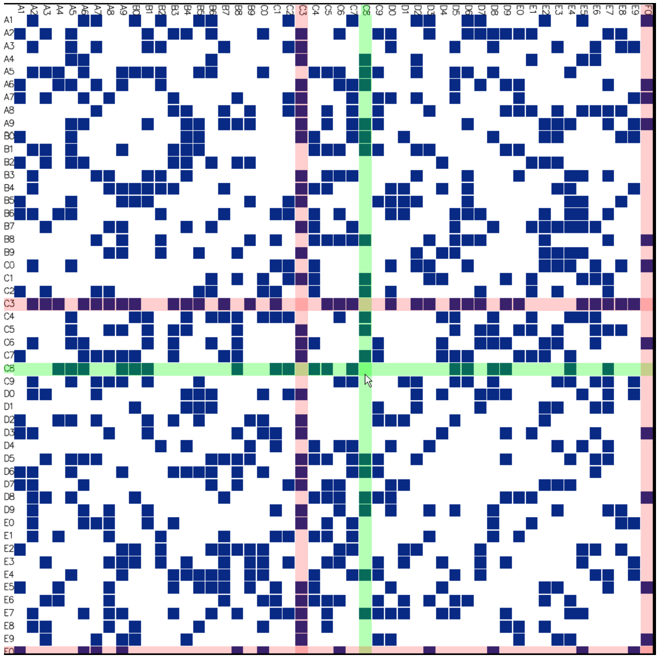


Qualitative Results

- Treemaps
  - Color coding and filtering heavily used
  - Needs “Detail on Demand”
- Hyperbolic Browser
  - Cartesian distance ≠ distance from root
    - Tree depth not what HB intended for!!
- Lots of usability problems
  - From analyzing videos of users

Study Example





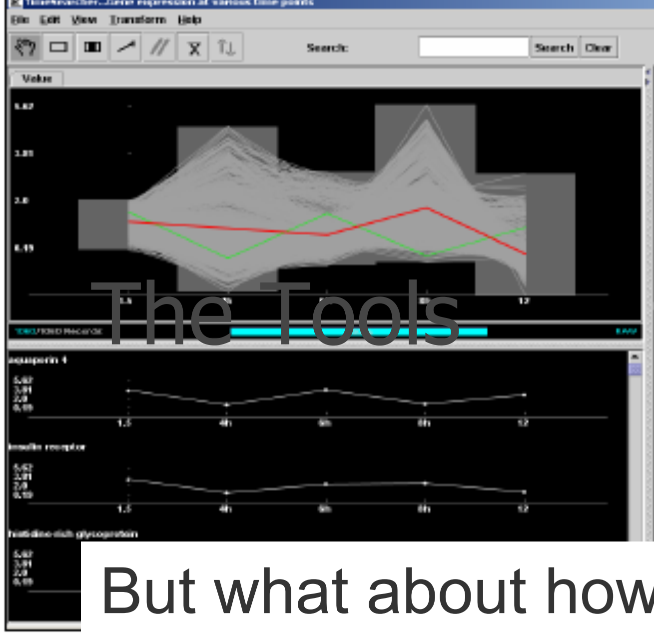
Mohammad Ghoniem, Jean-Daniel Fekete, Philippe Castagliola. A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. InfoVis 2004, Austin, TX, Oct 2004. IEEE

Nodes & Links vs. Matrix

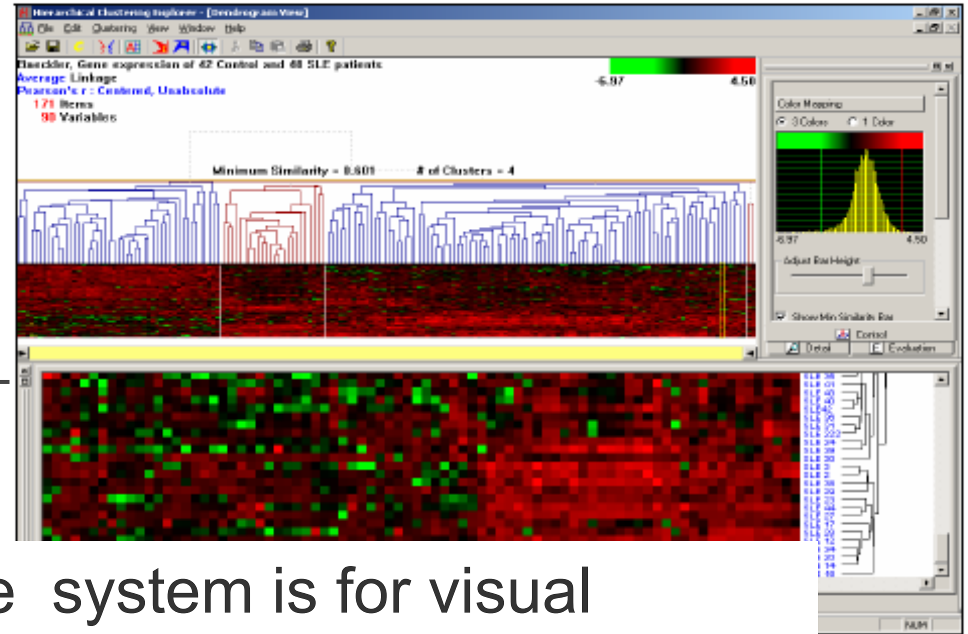
- Initial studies led to a design modification
  - Added in color highlighting of moused-over and selected nodes

- Looked at a set of typical graph operations
- Varied graphs by # of nodes and connectivity
- Found that matrix better for all tasks except path following
  - Better here means faster and higher accuracy on average

The Tools

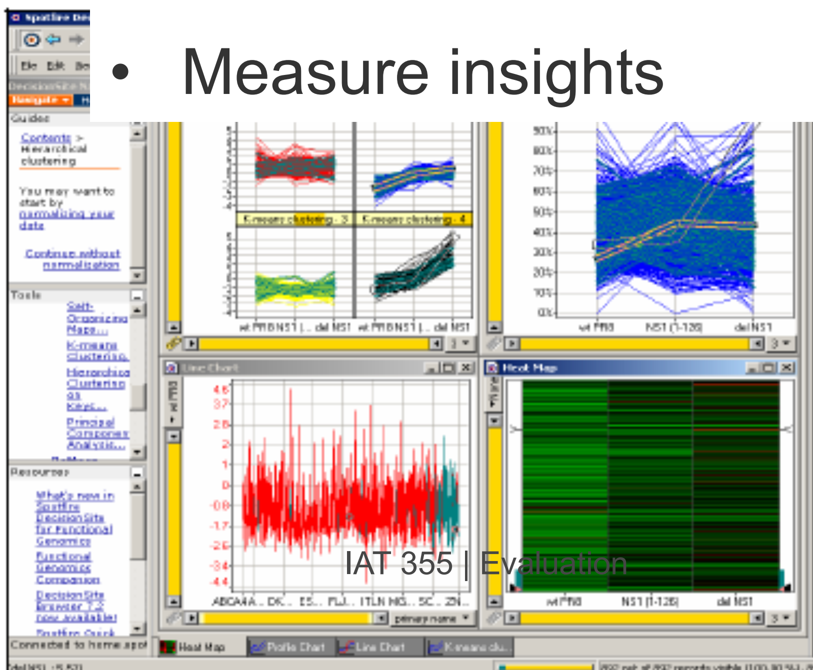


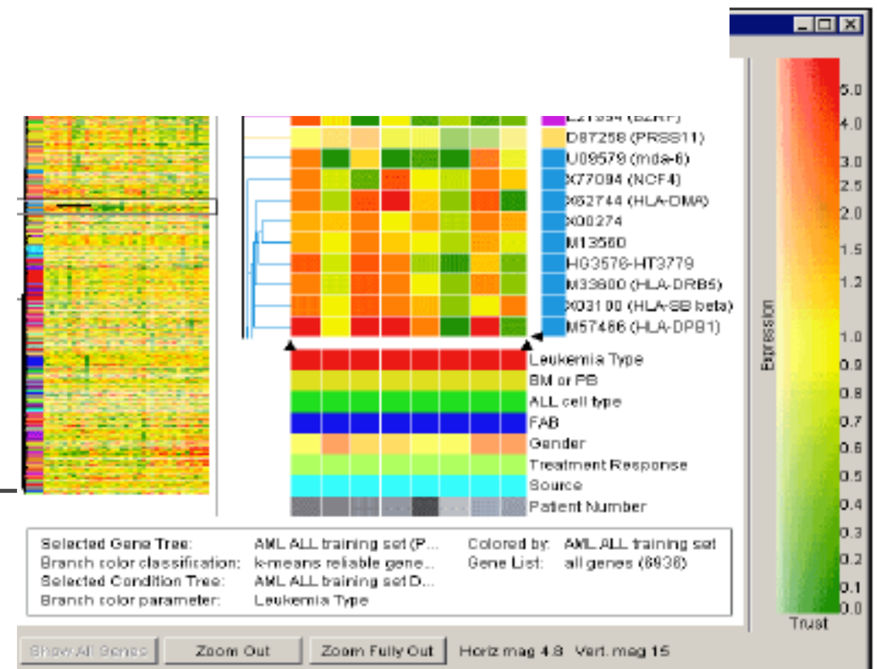
Cluster / Treeview, aka Clusterview (not shown)



But what about how useful the system is for visual analysis and discovery?

- Measure insights





Measure *insight*

- Users instructed to list some questions they might want to ask, and to then use tool until could learn nothing more
- For each insight, record / assess
  - The actual insight
  - Time to reach insight
  - Importance of the insight
  - Correctness of the insight
  - Sought for vs serendipitous insight
  - Etc

Results

- Insight Value
  - Spotfire® (66) > GeneSpring® (40)
- User's perception of how much learned
  - Spotfire® > HierClusExplr, Clusterview
- Time to first insight
  - Clusterview (4.6) < all others (7 to 16) :-)
  - GeneSpring® (16) > all others (4.6 to 14) s :-(

Pros/Cons (+/-)

+ ***Very important methodology!***

  Gets at users' real objectives!

- Wide variation in tool capabilities

  TimeSearcher did really well with the time series data - but not with other types of data

- Only 10 domain experts
- Short-term use – just a few hours
- Users not 100% motivated

Long-term Evaluation

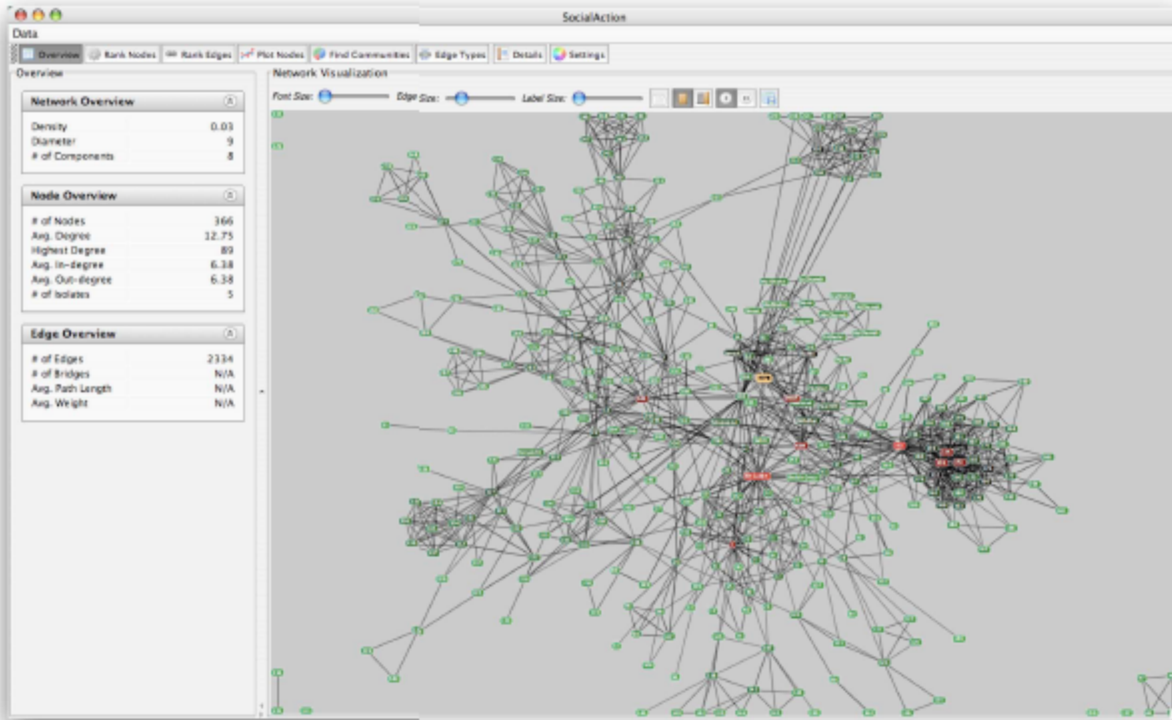
- The challenges of evaluation
- Subjective evaluation
- Evaluation based on model of cognition & perception
- Experimental evaluation
- Long-term Evaluation

Long-term Studies

- Perer & Shneiderman, *Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis*, CHI '08
- Domain:
  - Social network analysis – political influence, etc
- Problem:
  - Network visualizations by themselves not so useful
  - Need statistics about the networks
- SocialAction InfoVis system
  - Combines visualization and statistics
- Question:
  - Does SocialAction improve researcher's capabilities?

Saraiya, North, et al, *An Insight-Based Longitudinal Study of Visual Analytics*, IEEE TVCG 2006

SocialAction



Statistics

Users choose from statistical algorithms to find important nodes, detect clusters and more.

Network Visualization

The visualization is integrated with the statistics. Nodes are colored according to their ranking, with red nodes being the most statistically important.

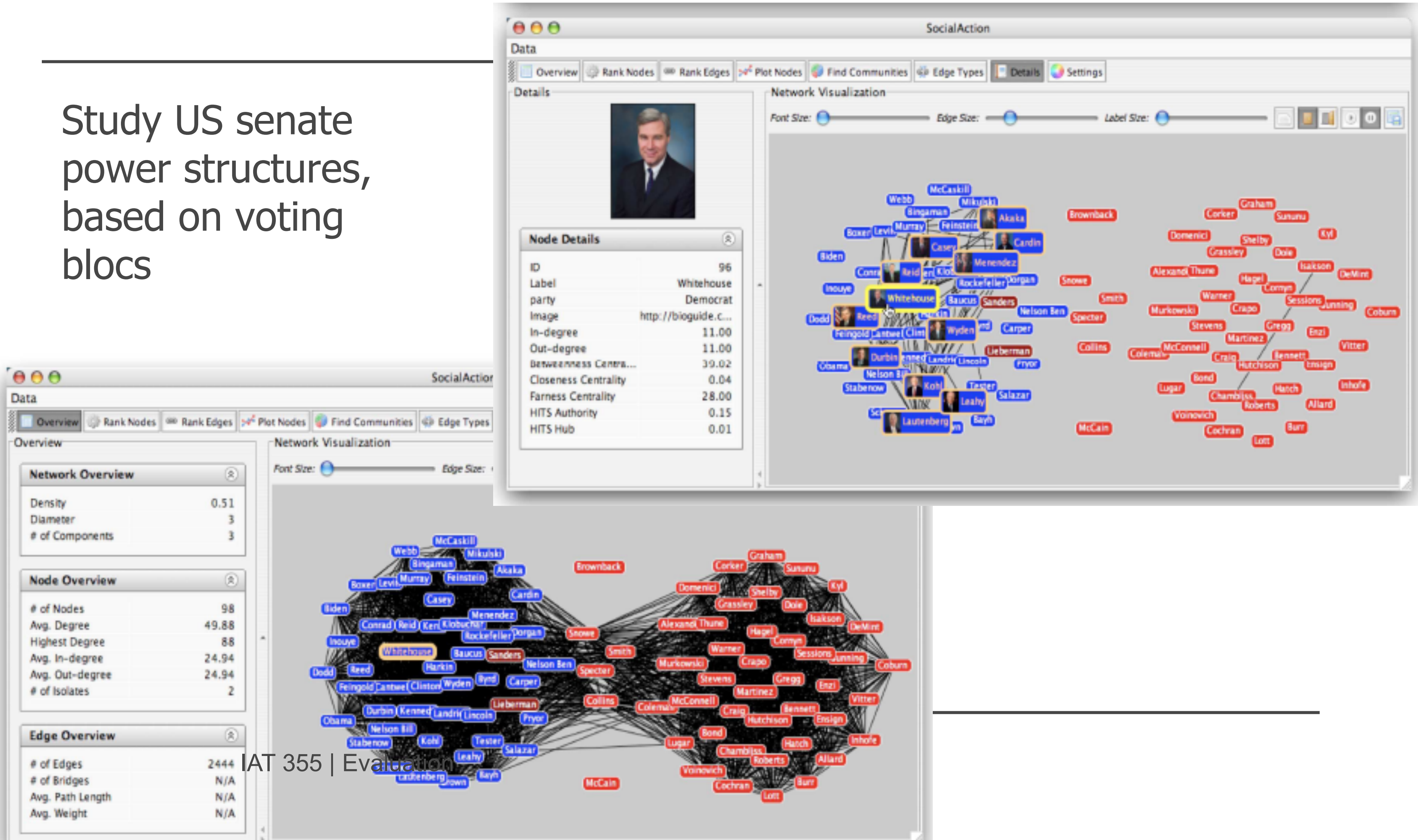
SFU

Methodology – for each user

- Interview (1 hour)
  - Understand their goals
- Training (2 hours)
  -
- Early use (2-4 weeks)
  - With user's data
  - Weekly visits
  - Requested (and feasible) software enhancements
  - Provide help
- Mature use (2-4 weeks)
  - No software improvements
  - Weekly visits
  - Provide help
- Exit Interview (1 hour)
  - How did system impacted research?
  - How well were goals met?

One User: A Political Analyst

Study US senate power structures, based on voting blocs



Findings

- With all four users
  - improved ability to find insights
- One user in depth:
  - Found interesting patterns using the capability to rank all nodes, visualize outcome and then filter out the unimportant
  - The betweenness centrality statistic helped find "centers of gravity"
  - Found geographic alliances.

Pros/Cons

- Does not compare two tools
- Focus on how well a tool helps real users do real work – very authentic, not lab study
  - Not enough users to get s.s. results
  - Users 110% motivated
  - Expensive to do!!

Scenarios [Lam et al 2011]

Data analysis

- work environments and practice
- Visual data analysis and reasoning
- Communication through visualization
- Collaborative data analysis

Visualization

- User performance
- User experience
- Visualization algorithms

Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. IEEE Transactions on Visualization and Computer Graphics,18(9), 1520-1536.

But what about

- Personal visualization
- Ambient visualization
- Public (situated) visualization
- Informative art viz