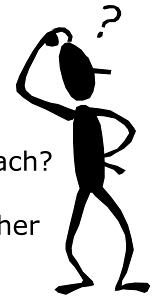
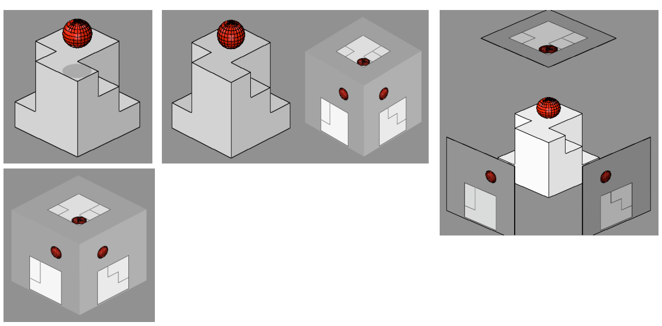
User Centered Design and Evaluation

Overview

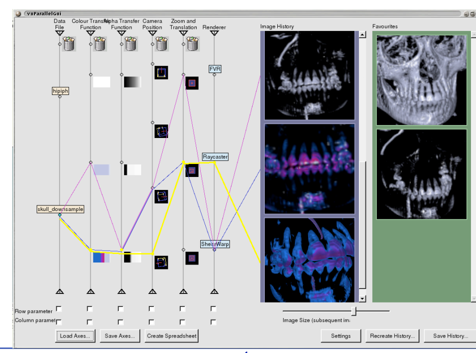


- My evaluation experience
- Why involve users at all?
- What is a user-centered approach?
- Evaluation strategies
  - Examples from "Snap-Together Visualization" paper

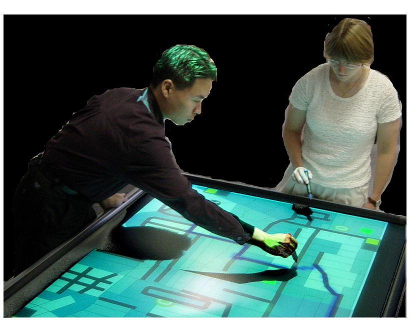
Empirical comparison of 2D, 3D, and 2D/3D combinations for spatial data



Development and evaluation of a Volume visualization interface



Collaborative visualization on a tabletop

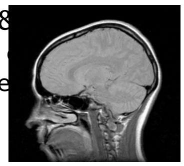


Why involve users?

"Darn these hooves! I hit the wrong switch again! Who designs these instrument panels, raccoons?"

Why involve users?

- Understand the users and their problems
  - Visualization users are experts
  - We do not understand their tasks and information needs
  - Intuition is not good enough
- ment &
  ealistic
  e stake



What is a user-centered approach?

- Early focus on users and tasks
- Empirical measurement: users' reactions and performance with prototypes
- Iterative design

Focus on Users

- Users' characteristics and context of use need to be supported
- Users have varied needs and experience
  - E.g. radiologists vs. GPs vs. patients

Understanding users' work

- Field Studies
  - May involve observation, interviewing
  - At user's workplace
- Surveys
- Meetings / collaboration

Design cycle

- Design should be iterative
  - Prototype, test, prototype, test, …
  - Test with users!
- Design may be participatory

Key point

- Visualizations must support specific users doing specific tasks
- “Showing the data” is not enough!

Evaluation

How to evaluate with users?

- Quantitative Experiments

Clear conclusions, but limited realism

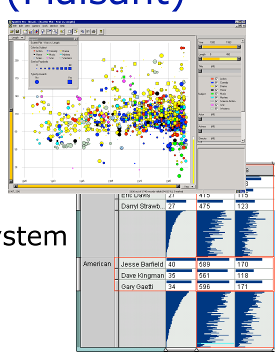
- Qualitative Methods
  - Observations
  - Contextual inquiry
  - Field studies

More realistic, but conclusions less precise

How to evaluate without users?

- Heuristic evaluation
- Cognitive walkthrough
  - Hard – tasks ill-defined & may be accomplished many ways
    - Allendoerfer et al. (InfoVis05) address this issue
- GOMS / User Modeling?
  - Hard – designed to test repetitive behaviour

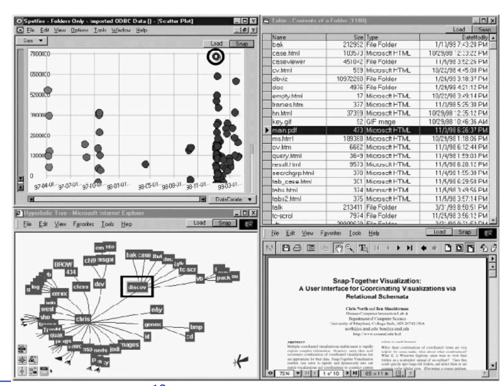
Types of Evaluation (Plaisant)



- Compare design elements
  - E.g., coordination vs. no coordination (North & Shneiderman)
- Compare systems
  - E.g., Spotfire vs. TableLens
- Usability evaluation of a system
  - E.g., Snap system (N & S)
- Case studies
  - Real users in real settings E.g., bioinformatics, E-commerce, security

Snap-Together Vis

Custom coordinated views



Questions

- Is this system usable?
  - Usability testing
- Is coordination important? Does it improve performance?
  - Experiment to compare coordination vs. no coordination

Usability testing vs. Experiment

Usability testing

- Aim: improve products
- Few participants
- Results inform design
- Not perfectly replicable
- Partially controlled conditions
- Results reported to developers

Quantitative Experiment

- Aim: discover knowledge
- Many participants
- Results validated statistically
- Replicable
- Strongly controlled conditions
- Scientific paper reports results to community

Usability of Snap-Together Vis

- Can people use the Snap system to construct a coordinated visualization?
- Not really a research question
- But necessary if we want to use the system to answer research questions
- How would you test this?

Critique of Snap-Together Vis Usability Testing

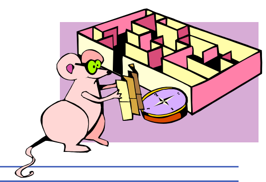
+ Focus on qualitative results
+ Report problems in detail
+ Suggest design changes
- Did not evaluate how much training is needed (one of their objectives)
- Results useful mainly to developers

Summary: Usability testing

- Goals focus on how well users perform tasks with the prototype
- May compare products or prototypes
- Techniques:
  - Time to complete task & number & type of errors (quantitative performance data)
  - Qualitative methods (questionnaires, observations, interviews)
  - Video/audio for record keeping

Controlled experiments

- Strives for
  - Testable hypothesis
  - Control of variables and conditions
  - Generalizable results
  - Confidence in results (statistics)



Testable hypothesis

- State a testable hypothesis
  - this is a precise problem statement
- Example:
  - (BAD) 2D is better than 3D
  - (GOOD) Searching for a graphic item among 100 randomly placed similar items will take longer with a 3D perspective display than with a 2D display.

Controlled conditions

- Purpose: Knowing the cause of a difference found in an experiment
  - No difference between conditions except the ideas being studied
- Trade-off between control and generalizable results

Confounding Factors (1)

- Group 1
  Visualization A in a room with windows
- Group 2
  Visualization B in a room without windows

What can you conclude if Group 2 performs the task faster?

Confounding Factors (2)

- Participants perform tasks with Visualization A followed by Visualization B.

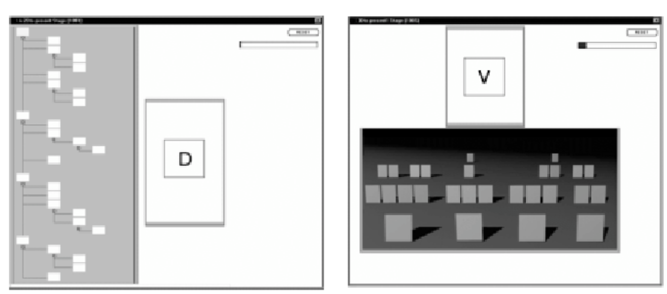
What can we conclude if task time is faster with Visualization A?

Confounding Factors (3)

- Do people remember information better with 3D or 2D displays?
- Participants randomly assigned to 2D or 3D
- Instructions and experimental conditions the same for all participants

Tavanti and Lind (Infovis 2001)

What are the confounding factors?



2D Visualization **3D Visualization**

What is controlled

- Who gets what condition
  - Subjects randomly assigned to groups
- When & where each condition is given
- How the condition is given
  - Consistent Instructions
  - Avoid actions that bias results (e.g., "Here is the system I developed. I think you'll find it much better than the one you just tried.")
- Order effects

Order Effects

Example: Search for circles among squares and triangles in Visualizations A and B

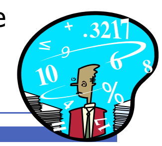
1. Randomization
   - E.g., number of distractors: 3, 15, 6, 12, 9, 6, 3, 15, 9, 12…
2. Counter-balancing
   - E.g., Half use Vis A 1^{st}, half use Vis B first

Experimental Designs

	Between-subjects	Within-subjects
No order effects?	+	-
Participants can compare conditions?	-	+
Number of participants	Many	Few

Statistical analysis

- Apply statistical methods to data analysis
  - confidence limits:
    - the confidence that your conclusion is correct
    - "p = 0.05" means:
      - a 95% probability that there is a true difference
      - a 5% probability the difference occurred by chance



Types of statistical tests

- T-tests (compare 2 conditions)
- ANOVA (compare >2 conditions)
- Correlation and regression
- Many others

Snap-Together Vis Experiment

- Are both *coordination* AND *visual overview* important in overview + detail displays?

- How would you test this?

Critique of Snap-Together Vis Experiment

+ Carefully designed to focus on factors of interest
- Limited generalizability. Would we get the same result with non-text data? Expert users? Other types of coordination? Complex displays?
- Unexciting hypothesis – we were fairly sure what the answer would be

How should evaluation change?

- Better experimental design
  - Especially more meaningful tasks
- Fewer "Compare time on two systems" experiments
- Qualitative methods
- Field studies with real users

Take home messages

- Talk to real users!
- Learn more about HCI!