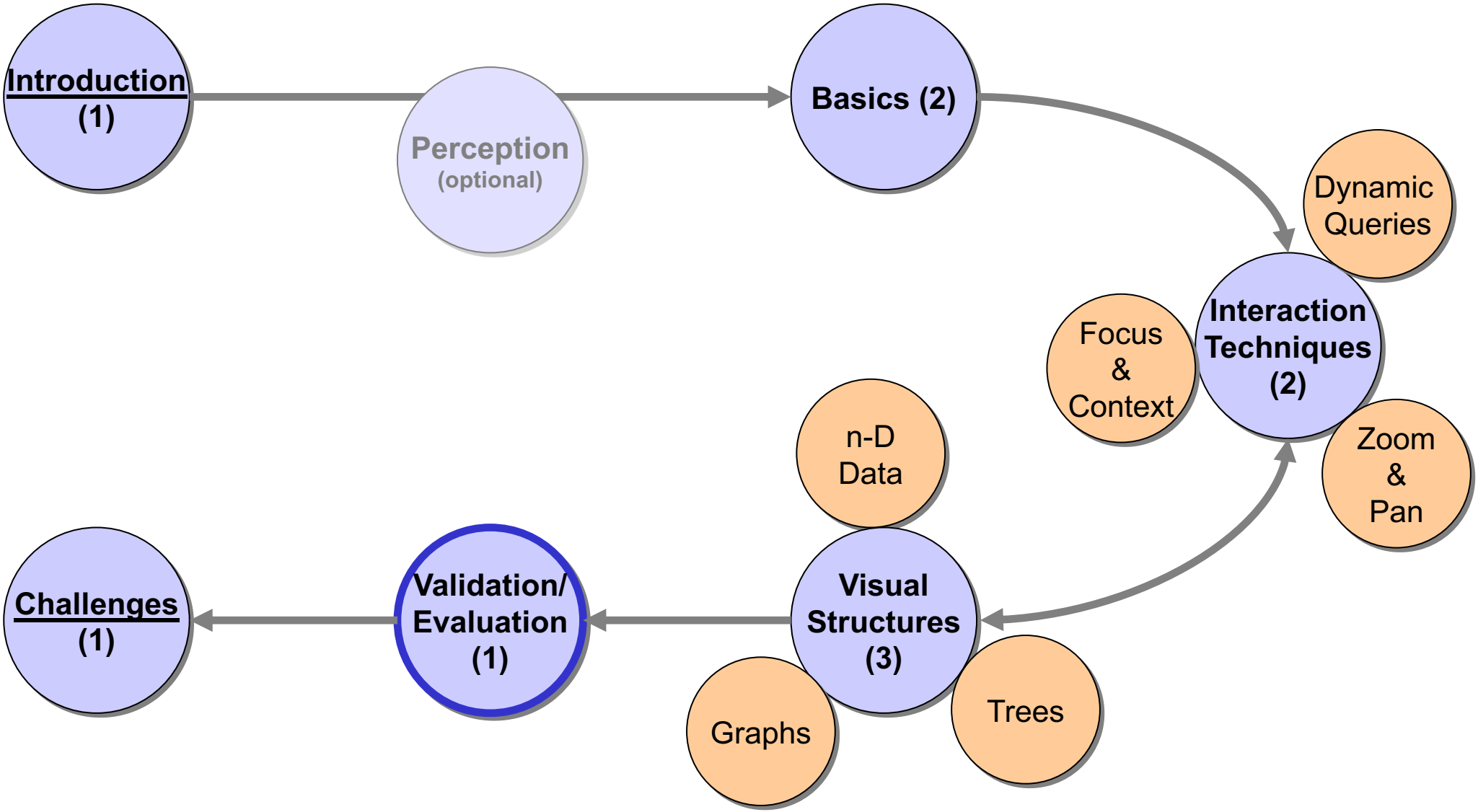
Information Visualization

7. Validation and Evaluation Methods

TNM111: 10 Lectures



7.1 Introduction

- So far, we focused on new techniques and tools for the visualization of abstract data
- It is also important to know whether our efforts are *better* in some sense than previous approaches. We must critically prove our techniques and tools, i.e., we must *validate* and/or *evaluate* them!
- To say it more directly: the nicest and coolest visualization is useless if it doesn't help the user to solve a problem or a specific task

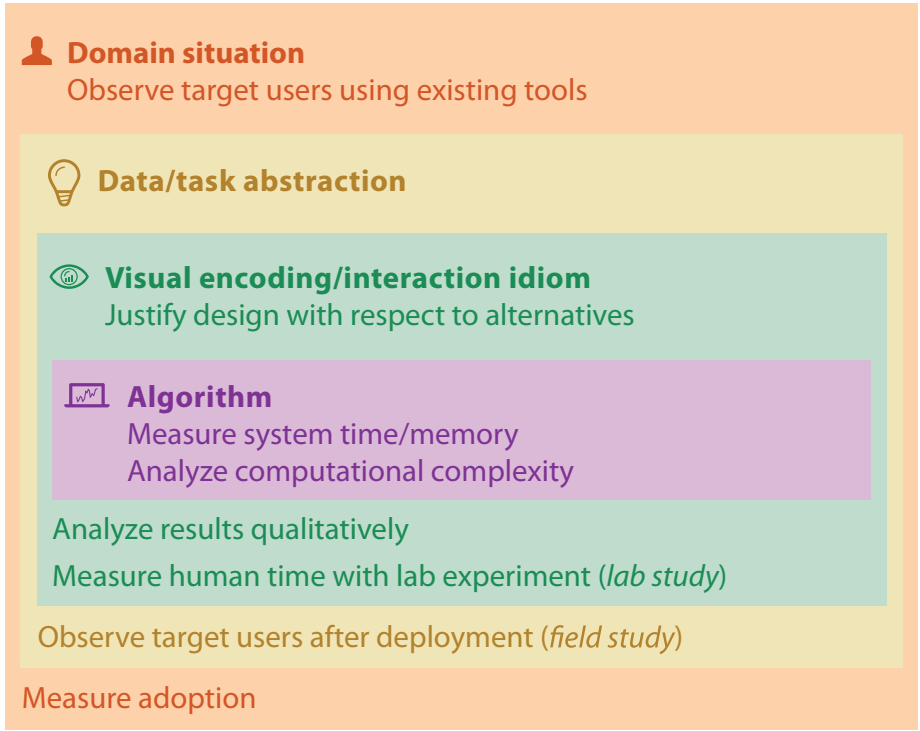
[Inspired by J. Stasko's course]

7.1 Introduction

- However, it is also important to be not too restrictive
  - Often, the development of a visualization that has itself not a huge benefit can lead to a later tool with excellent usability, benefit, and evaluation results
- Definition of **validation**
  - Process of checking whether or not a certain design is appropriate for its purpose, meets all constraints and will perform as expected (→ yes/no answer or attestation)
  - If mathematical rigor is involved, then we speak about *verification* (→ proofs, …)

7.1 Introduction

- Definition of **evaluation**
  - Process of collecting quantitative/qualitative information of some key characteristics of a certain design (→ assessment) to enable reflection
- Let's look into Munzner's nested model
  - Validation is hardly possible on the outer levels; it makes only sense on the innermost level

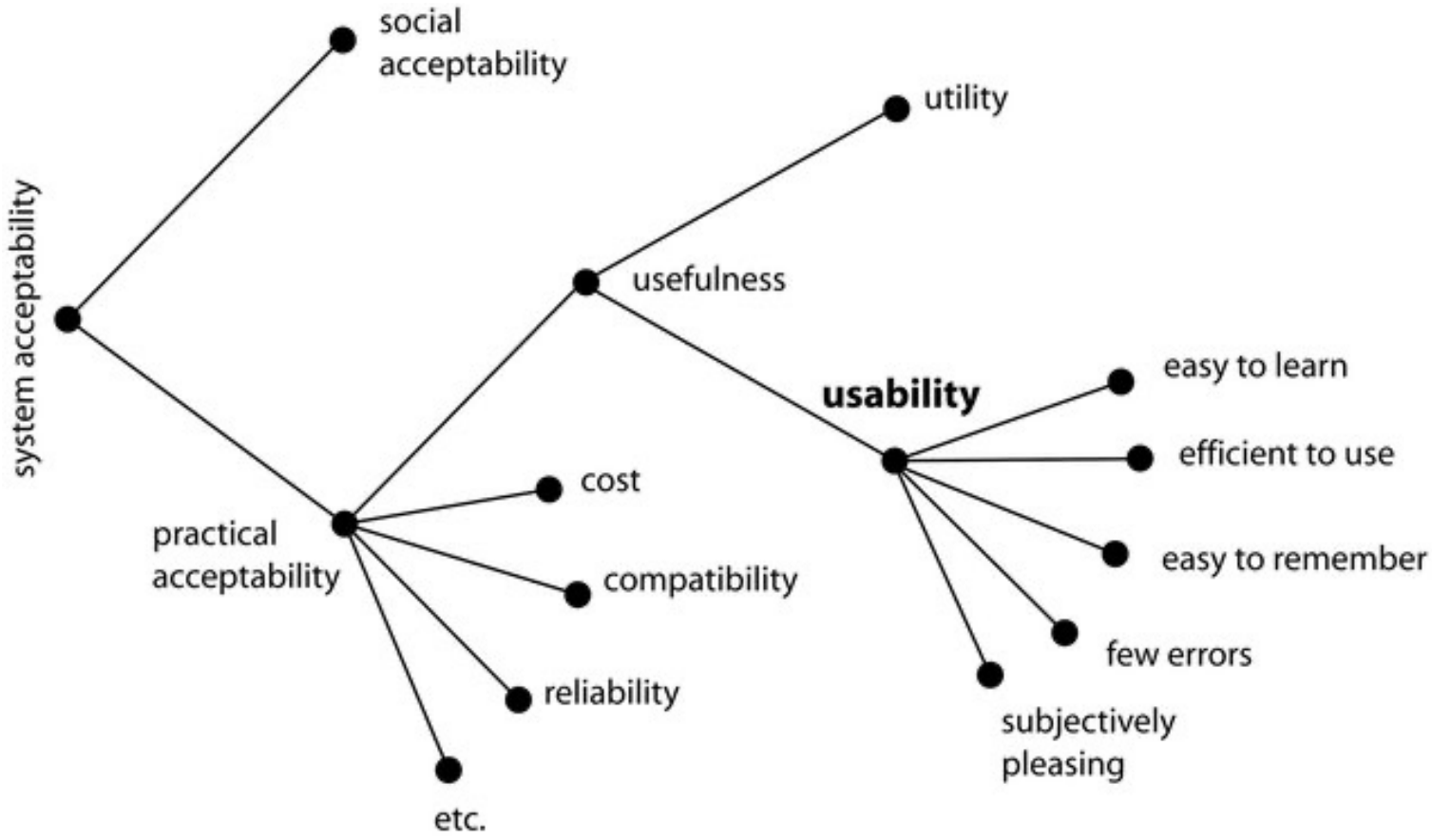


7.1 Introduction

- Thus, for the strongly human-centered approach of InfoVis, the term and definition of *evaluation* is better suited to “validate” more or less complex visual designs
- So, let’s evaluate our interactive visualizations!
- But, how can we perform such evaluations? Are there good techniques to do that?
  - Because of the lack of time, we can only list interesting techniques in this course and cannot go very deep
  - But before, we have to know what is meant with **usability**

7.2 Usability

- System acceptability attributes (Nielsen 1993, p. 25)



7.2 Usability

- Usability is a quality attribute that assesses how easy user interfaces are to use; there are **5 quality components** according to Nielsen

  1) *Learnability*. How easy is it for users to accomplish basic tasks the first time?

  2) *Efficiency*. Once users have learned, how quickly can they perform tasks?

  3) *Memorability*. When users return using the system (after a period of not using it), how easily can they remember how to use it?

  4) *Errors*. How many errors do users make, how severe are the errors, and how easily can the users recover from the errors?

  5) *Satisfaction*. How pleasant is it to use the system?

7.2 Usability

- The ISO standard 9241-210 defines usability as
  - the extent to which a product can be used by *specified users* to achieve *specified goals* with *effectiveness*, *efficiency* and *satisfaction* in a *specified context* of use

- Thus, we have 3 usability criteria according to the ISO standard
  1) Effectiveness
     - The accuracy and completeness with which specified users can achieve specified goals in particular environments
     - Measures whether the user's goals were met successfully

7.2 Usability

- Thus, we have 3 usability criteria according to the ISO standard (cont.)
  2) Efficiency
     - The resources expended in relation to the accuracy and completeness of goals achieved
     - Measures the time spent, mental workload, etc. (with accuracy)
  3) User satisfaction
     - The comfort and acceptability of a system
     - Measures the user's attitudes, preferences, etc.
- So, finally, which evaluation techniques can we use to gather data about the usability of a visualization?
  - To determine its merits, value, … and to provide feedback

7.3 Techniques

- Generally, no technique was developed especially for InfoVis. They were derived from other areas, e.g., from HCI, cognitive psychology, statistics, …
- We distinguish between different types, such as
  - Qualitative and quantitative evaluations respectively,
  - Controlled experiments, subjective interviews,
  - Interpretation of observations, usability tests, …
- Now, one must choose which technique is suitable for the aspect to be analyzed

7.3.1 The Right Choice

- The following is based on Carpendale's paper [S. Carpendale. Evaluating Information Visualizations. In A. Kerren et al. (Eds.): Information Visualization, LNCS 4950, pp. 19–45, 2008]
- "Controlled experiments remain the workhorse of evaluation but there is a growing sense that information visualization systems need new methods of evaluation, from longitudinal field studies, insight based evaluation and other metrics adapted to the perceptual aspects of visualization as well as the exploratory nature of discovery" [CFP of BELIV 2006]
- Desirable factors when conducting a study
  - generalizability (G),
  - precision (P), and
  - realism (R)

7.3.2 Methodologies

- Field Study
  - conducted in the actual situation; the observer tries as much as possible to be unobtrusive
  - R++, P−−, G−−
- Field Experiment
  - conducted in a realistic setting; experimenter trades some degree of unobtrusiveness in order to obtain more precision in observations
  - R+, P○, G−
- Laboratory Experiment
  - experimenters fully design the study
  - R−, P+, G−

7.3.2 Methodologies

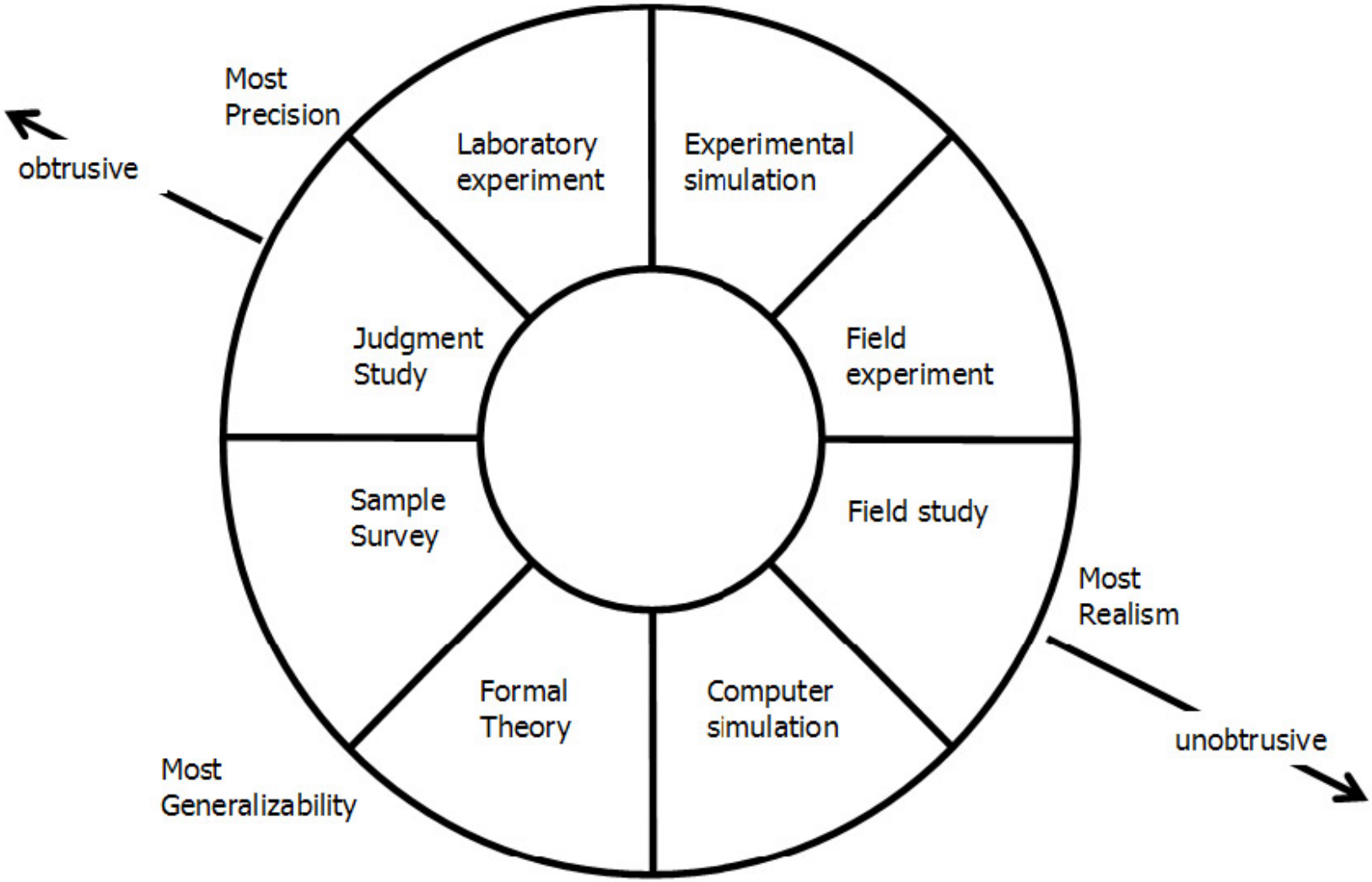
- Experimental Simulation
  - experimenter tries to keep as much of the precision as possible while introducing some realism via simulation
  - R○, P○, G−
- Judgment Study
  - the purpose is to gather a person's response to a set of stimuli in a situation where the setting is made irrelevant (much attention is paid to creating 'neutral conditions').
  - R−, P+, G○
- Sample Survey
  - experimenter is interested in discovering relationships between a set of variables in a given population
  - R−, P○, G+

7.3.2 Methodologies

- Formal Theory
  - not a separate experimental methodology but an important aspect of all empirical research that can easily be overlooked
  - R−, P−, G+
- Computer Simulation
  - develop a computer simulation that has been designed as logically complete
  - R○, P⌐, G⌐ (⌐ = not applicable)

"Mr. Paxburton is out of town this week. Would you like to speak to a computer simulation of him?"

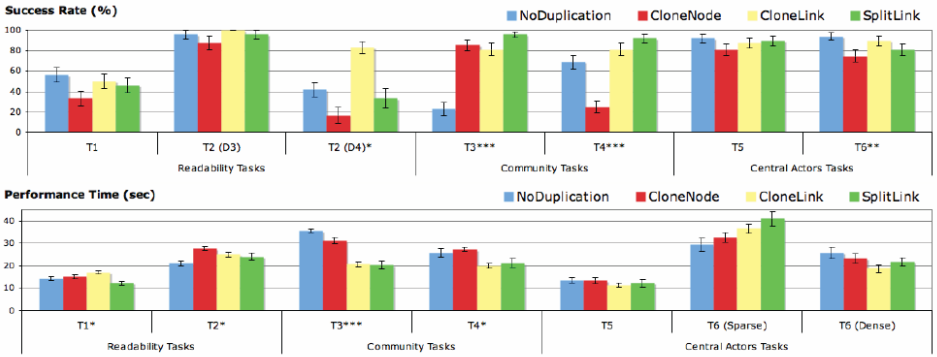
7.3.2 Methodologies



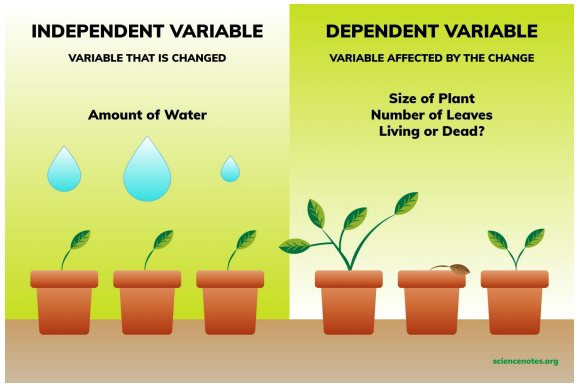
[adapted, simplified from McGrath, 1995]

7.3.3 Quantitative Evaluation

- Mostly laboratory experiments or studies



- Methodology
  - Hypothesis development
  - Identification of the independent variables (= what is manipulated?)
  - Control of the independent variables
  - Elimination of complexity
  - Measurement of the dependent variables
  - Application of statistics

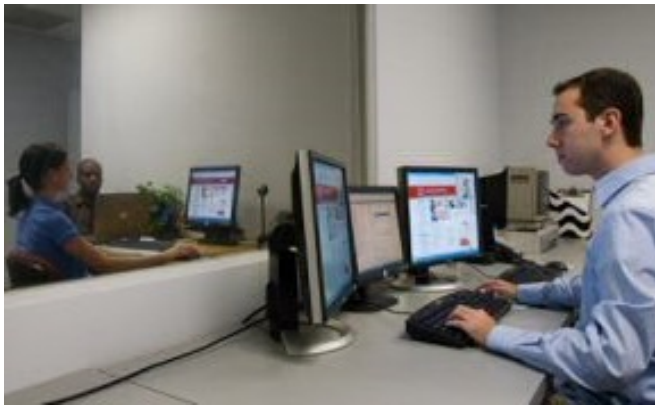


7.3.3 Quantitative Evaluation

- Challenges
  - Conclusion Validity
    - Is there a relationship between the independent and the dependent variables?
  - Type I (false negative) and Type II (false positive) Errors
  - 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

7.3.4 Qualitative Evaluation

- A more holistic approach that considers the interplay among factors that influence visualizations, their development, and their use
- Methods
  - Observation techniques
  - Interview techniques
- Types of qualitative methodologies
  - Nested qualitative methods
    - Experimenter observations
    - Think-aloud protocol
    - Collecting participant opinions



7.3.4 Qualitative Evaluation

- Types of qualitative methodologies (cont.)
  - Inspection evaluation methods
    - Usability heuristics
    - Collaboration heuristics
    - Information visualization heuristics
- Qualitative methods as primary
  - can be used at any time in the development life cycle
- Challenges
  - Sample sizes
  - Subjectivity
  - Analyzing qualitative data

7.3.5 Conclusions

- All techniques have their specific advantages and drawbacks
  - For example, the testing of the usability of a visualization technique gives no clear statement about the intrinsic quality of the visualization
- Exactly that is the main problem
  - It is hard to compare different systems because too many aspects play an important role, e.g., interaction possibilities, specific purposes, …
- The evaluation of InfoVis techniques and tools is and remains (for now) one of the big challenges

7.3.5 Conclusions

- All techniques have their specific advantages and drawbacks
  - For example, the testing of the usability of a visualization technique gives no clear statement about the intrinsic quality of the visualization
- Exactly that is the main problem
  - It is hard to compare different systems because too many aspects play an important role, e.g., interaction possibilities, specific purposes, …
- The evaluation of InfoVis techniques and tools is and remains (for now) one of the big challenges

7.4 Plan, Design, Conduct

- Do not do too much
  - A simple but good user study is better than a complex but faulty/risky one
- What method to choose?
  - One that is appropriate for the aim of your evaluation
  - One that generates interesting results and conclusions
  - One that you feel will give you reliable and valid results
  - One that you feel you can master well
  - One that fits your resources (time, availability of participants, equipment, etc.)

[Inspired by C. Forsell's lecture]

7.4 Plan, Design, Conduct

- Participants
  - Choose representative (typical) participants
  - Collect relevant background information
    - Age, gender, occupation, experience, frequency of use, attitudes, etc.
    - Why?
      - To provide relevant descriptions of participants
      - To explain unexpected results
    - Use a questionnaire or a structured interview guide
- Tasks (or free exploration) and/or questions
  - Should be meaningful and representative (consider time!)
  - Measure/assess towards a usability goal
    - Errors, response times, attitudes, etc.

[Inspired by C. Forsell's lecture]

7.4 Plan, Design, Conduct

- Procedure / Study Design
  - How to record responses?
    - Program records key press, mouse click, etc.
    - Written (one person interviews, one person take notes), voice recorder (must be transcribed)
    - Questionnaires
  - Choose an intuitive method
  - Oral/written instructions for the participants?
  - Demonstration of system and/or equipment?
  - Training?
  - As evaluator
    - Script your own actions to enable a smooth evaluation!!!

[Inspired by C. Forsell's lecture]

7.4 Plan, Design, Conduct

- Before and during evaluation
  - Welcome the participants
  - Describe the session and its parts: how long it will take
  - Your (and other's) role: observing or not, taking notes, clarification of tasks/assistance or not, …
  - The participants role, assignment into groups, etc.
  - The method (think aloud, etc.)
  - The task(s) (training, feedback, execution, responses)
  - The apparatus and other material
  - Breaks?
  - Questions?

[Inspired by C. Forsell's lecture]

7.4 Plan, Design, Conduct

- Analysis of the results
  - Quantitative evaluations
    - Simple calculations and summaries of data
      - error rates, task times, frequencies, mean/median values
    - Illustration by using tables, plots, etc.
    - Statistical analyses (t-test, ANOVA, …)
  - Qualitative evaluations
    - Group/structure data (texts) in a meaningful way
- Report the user study
  - Explain all taken decisions on study design, participants, etc.
  - Show and interpret the results

[Inspired by C. Forsell's lecture]

7.4 Plan, Design, Conduct

- Ethical issues, such as anonymity of the results, avoiding of biases, etc.
- Clearly state to each participant
  - It is the system that we are evaluating with your help – it is *not* you that is evaluated through it
  - If you experience the tasks as difficult, it is again the system that is difficult to understand and use! (i.e., not you who is too stupid)
  - Do not feel any pressure to respond positively if you know/like the administrator of the evaluation and/or the developer of the visualization!
- If something goes wrong blame the technique, yourself, but not the participant

[Inspired by C. Forsell's lecture]

Information Visualization

Oral Exam Information

Oral Exam Info

- Will take place in Weeks 11+12 in Rooms KO121-KO123
- Each group (2 students) chooses an open time slot in Doodle (announced via Lisam)
  - We will block the individual exams via the two assignment groups A+B (thus only one TA needs to attend)
  - For instance, a block on 21/3 afternoon might be for student groups assigned to A, other blocks are for groups assigned to B, etc.
- An exam takes 30 minutes
  - Take care that you both have a command on the course content! Both students get (normally) the same grade.

Oral Exam Info

- I will ask about anything discussed in the lectures and assignments
- Grading criteria document is available on Lisam, but roughly and sloppy spoken:
  - You need to have a good overview of the majority of the content to pass the oral exam (grade 3)
  - If you master all aspects in breadth and depth (grade 5)
- The oral exam grade is the grade of the course; you get the grade immediately after the oral exam

QUESTIONS ???