Evaluation.

VU Vis 2013.
Michael Sedlmair

Readings

- Carpendale: *Evaluating Information Visualization*, Chapter 2 in "Information Visualization for: Human-Centered Issues and Perspectives", Springer 2008

Overview

- Introduction
  - Motivation
  - Interface Design and Evaluation
- Techniques
  - Algorithmic Performance & Image Quality
  - Usability Testing
  - User Studies
- Conclusion and Discussion
  - New Ways of Evaluation
  - Questions

Motivation

What is Evaluation?

What is Evaluation?

"Evaluation is the systematic assessment of the worth or merit of some object"

(Quasi-standard definition from 50s/60s)

Why Evaluating Visualization?

- To ensure quality in product development
- To compare solutions
- To provide quantitative figures
- To get a scientific statement (instead of personal opinion)
- To convince your audience/reviewers...

Evaluation in Visualization...

Various forms:

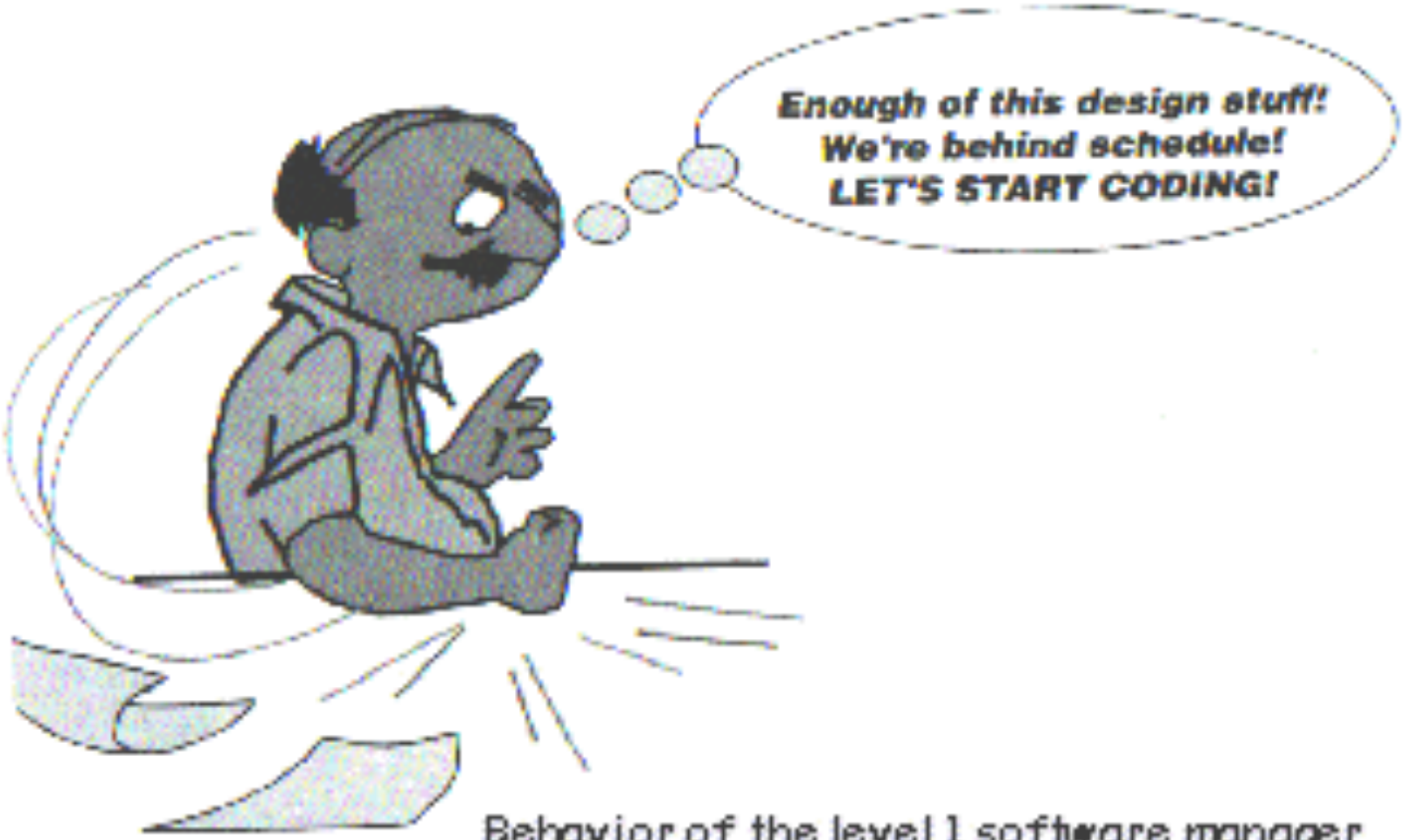
- Who: With Users vs. Without Users?
- Why: Formative vs. Summative?
- How: Quantitative vs. Qualitative?
- Where: Field studies vs. Lab studies?
- When: Before vs. During vs. After Tool Development

Related Areas

- Visualization Evaluation is related to Evaluation in...
  - Computational Performance
  - HCI (e.g., Usability)
  - Perceptual Psychology
  - Cognitive Reasoning/Sense-making
  - Social Science
- ... and of course: Statistics!

Interface Design and Evaluation

Evaluation is required at all stages in system development



Initial Assessment

- **Evaluation is required at all stages in system development**
  - Initial assessments (Domain Problem Characterization):
    - What kind of problems are the system aiming to address?
      (e.g., difficult to analyze a large and complex dataset)
    - Who is your target users?
      (e.g., data analysts)
    - What are the tasks? What are the goals?
      (e.g., to find trends and patterns in the data via exploratory analysis)
    - What are their current practice; what tools do they use?
      (e.g., statistical analysis)
    - Why and how can visualization be useful?
      (e.g., presentation, communication, debugging, speeding up the workflow, hypothesis testing/creation, ...)
    - Talk to the users, and observe what they do
    - Task analysis

Iterative Design Process

- **Evaluation is required at all stages in system development**
  - Initial assessments (Domain Problem Characterization)
  - Iterative design process (data abstraction, encoding, interaction, algorithmic):
    - Does your design address the users' needs?
    - Can they use it?
    - Where are the usability problems?

    - Evaluate without users: cognitive walkthrough, action analysis, heuristics analysis, algorithmic analysis
    - Evaluate with users: usability evaluations—think aloud, performance measurements

Bench-Marking

- **Evaluation is required at all stages in system development**
  - Initial assessments (Domain Problem Characterization)
  - Iterative design process (data abstraction, encoding, interaction, algorithmic)
  - Bench-marking:
    - How does your system compare to existing systems?
    - Numerical (complexity - speed, memory - image quality)
    - Empirical, comparative user studies
      - Ask specific questions
      - Compare an aspect of the system with specific tasks (task taxonomy paper; Ware's appendix C)
      - Quantitative, but limited (see The Challenge of Information Visualization Evaluation)

Deployment

- **Evaluation is required at all stages in system development**
  - Initial assessments (Domain Problem Characterization)
  - Iterative design process (data abstraction, encoding, interaction, algorithmic)
  - Bench-marking
  - Deployment:
    - How is the system used in the wild?
    - Are people using it?
    - Does the system fit in with existing work flow? Environment?

    - Contextual studies, field studies…

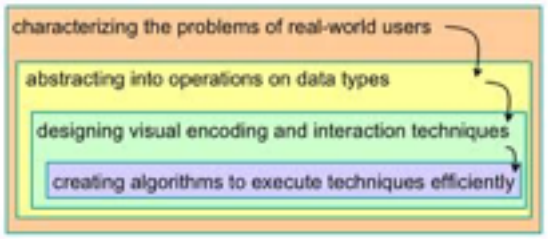
Iterative Design and Evaluation

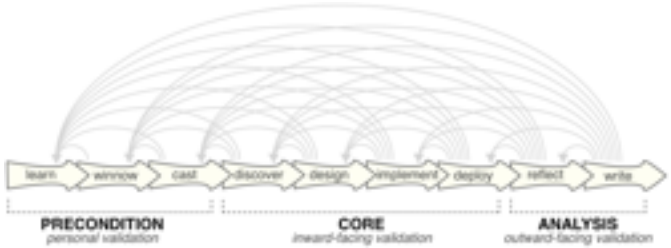
- **Evaluation is required at all stages in system development**
  - Initial assessments (Domain Problem Characterization)
  - Iterative design process (data abstraction, encoding, interaction, algorithmic)
  - Bench-marking
  - Deployment
  - Identify problems and go back to 1, 2, 3, or 4

(Task-Centered User Interface Design, Clayton Lewis and John Rieman, Chapters 0-5.)

Further Literature

- Munzner: *A Nested Model for Visualization Design an Validation* (InfoVis 2009)
- Sedlmair, Meyer, Munzner: *Design Study Methodology: Reflections from the Trenches and the Stacks* (InfoVis 2012)





II. Methods

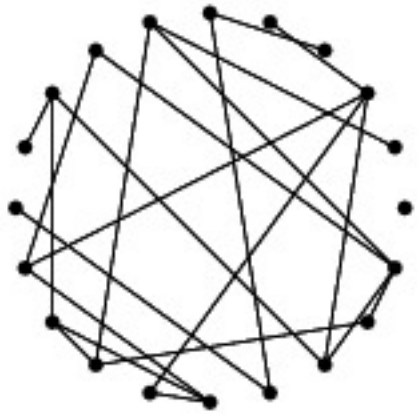
A. Algorithmic Performance & Image Quality

Algorithmic Performance

- Complexity
  - measured in terms of size of input problem
  - e.g. input size of a volume is not N, but N^3
  - is it NP hard?
- Scalability
  - closely related to Complexity
  - reason about interactivity (speed)
  - reason about resource constraints (memory)
- Image Quality

Image Quality

- **Metrics**
- E.g., metrics for graph drawing
  - line crossings?
  - area?
  - sum of edge length?
  - uniform edge length?
  - ...



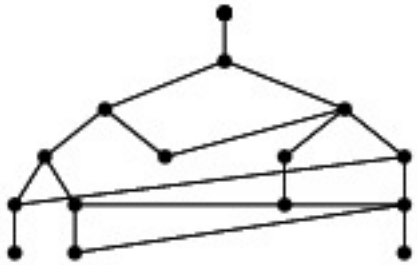
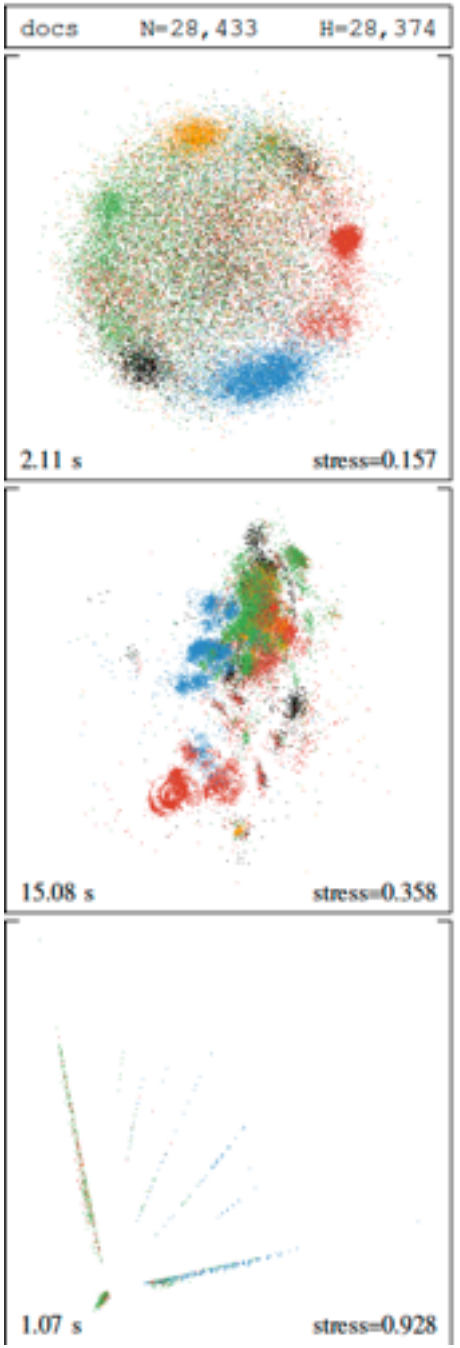


Image Quality

- Metrics
- **Qualitative Discussion of results of an algorithm**



B. Usability Testing

Goals

- Is the tool usable?
- Improve product design
- 5 E's: Is the interface…
  - Effective?
  - Efficient?
  - Engaging?
  - Easy to learn?
  - Equally usable by different groups?

with thanks to Melanie Tory

Different Methods

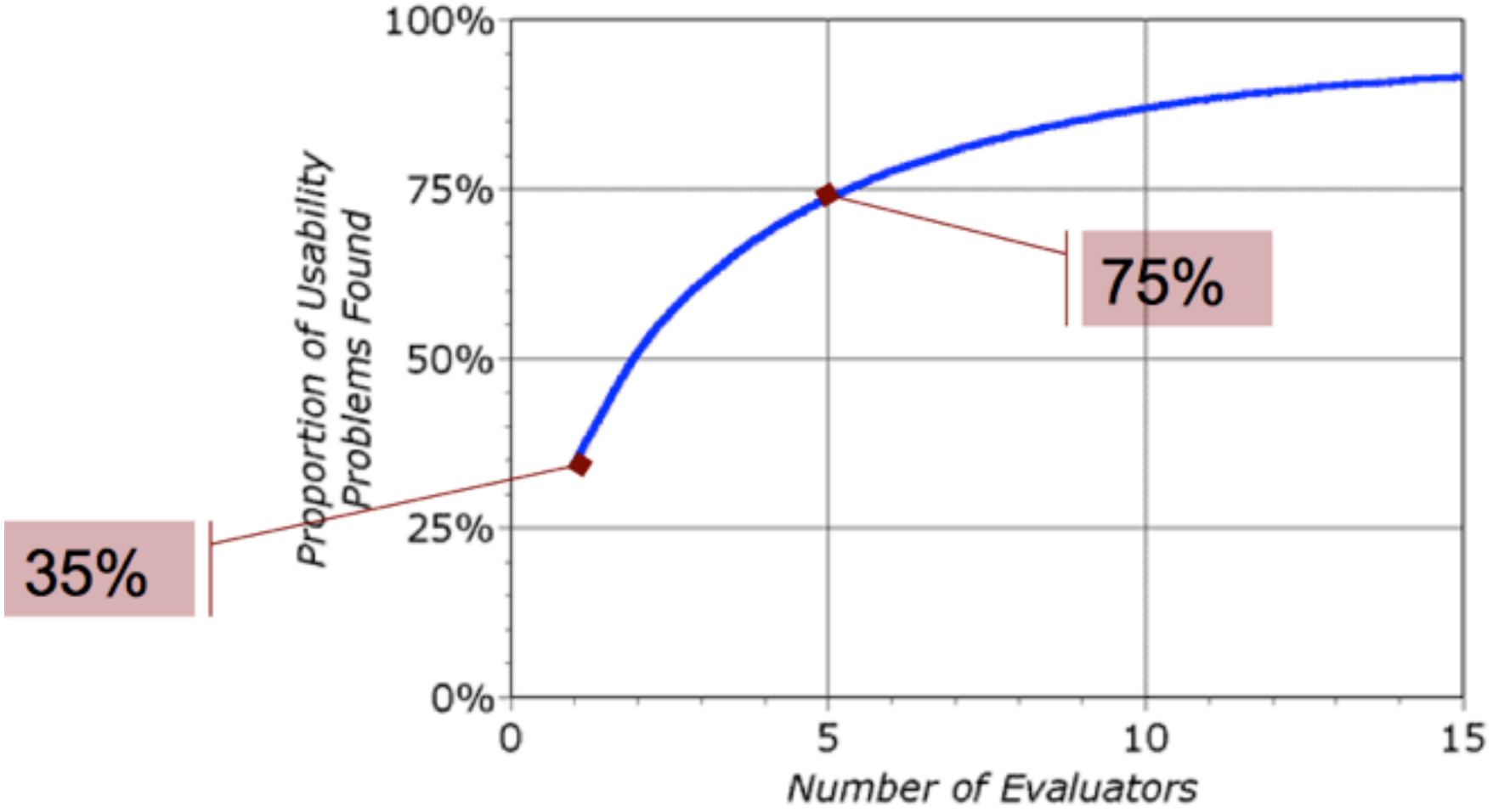
- Usability Inspection (without users)
  - Heuristics
  - Cognitive Walkthrough
- User Testing (with users)
  - Think Aloud Protocol
  - ... User Studies

Heuristic Evaluation

- A type of usability inspection
- Vis/HCI experts (sometimes also domain experts) review an interface design with respect to a set of predefined heuristics
- Heuristics:
  - Usability Heuristics (Nielsen/Norman)
  - Collaboration Heuristics
  - Visualization Heuristics

with thanks to Melanie Tory

How many inspectors?



with thanks to Melanie Tory

Cognitive Walkthrough (CW)

- A type of usability inspection where experts ‘walk’ through an interface following a specified set of tasks.
- Step through the task. At each step, ask yourself four questions.

with thanks to Melanie Tory

Cognitive Walkthrough (CW)

1. Will users be try to achieve the effect that the subtask has?
2. Will users see the control (button, menu, switch, etc.) for the action?
3. Once users find the control, will they recognize that it produces the effect they want?
4. After the action, will users understand the feedback they get so they can go on to the next action with confidence?

with thanks to Melanie Tory

Reports for Designers.

- Ranked List of Usability problems.

Usability Inspection: Pros and Cons

Usability inspection…

- Is quick and inexpensive
- May miss important problems or identify false ones
- Complements user testing

with thanks to Melanie Tory

User Testing

- Let the user test your software!
- May find domain specific problems
- Methods:
  - Think Aloud Protocols
  - User Observation
  - Interviews
  - etc., etc., etc. ... User Studies!

with thanks to Melanie Tory

C. User Studies

Goals

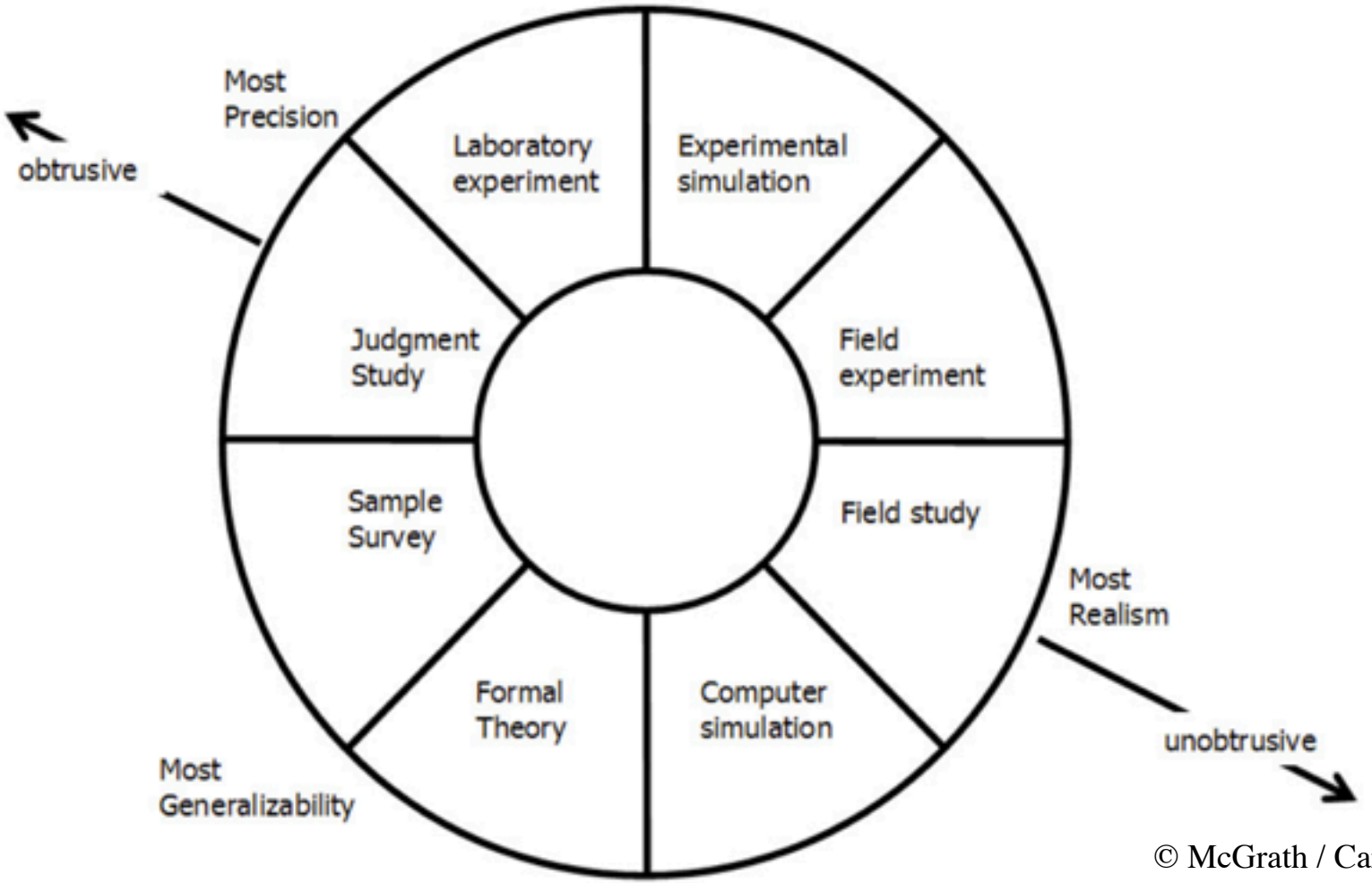
- Discover Knowledge
  - How are interfaces used?
  - Do they generate insights?
- Prove concepts
  - Is your novel technique actually useful?
  - Is your novel technique better than another technique? (faster? / less errors? / more insights? ...)

with thanks to Melanie Tory

Goals

- Generalizability
  - Results can be applied to other people
- Precision
  - We measured what we wanted to measure (controlling factors that were not intended to study)
- Realism
  - Study context is realistic

... usually **trade-off** between them!



The selection of a research method depends on the research question and the object under study!

C1. Quantitative Evaluation

Focus

- Generalizability (?)
- Precision

- Reveal: Cause-Effect relationships, e.g. smoking --> cancer

Quantitative Methods

1. Hypothesis Development
   - A precise problem statement
   - E.g., Participants will be faster with VisMethod A than with VisMethod B for finding paths
2. Identification of **Independent variables**
   - Factors to be studied
   - E.g., interface, …
3. Control of Independent Variable
   - Levels: The number of variables in each factor
   - Limited by the length of the study and the number of participants

Quantitative Methods

4. Control Environment
   - In order not to distort your study
5. Measurement of **Dependent Variables**
   - Performance indicators: task completion time, error rates, mouse movement, (insights)…
   - Subjective participant feedback: satisfaction ratings, closed-ended questions, interviews…
   - Observations: behaviors, signs of frustrations…
6. Application of Statistics
   - Significance: How sure can we be that our result could (or could not) have happened by chance!
   - p-Value: < 0.05?
   - Methods: T-Test, ANOVA, ...
   - Should know how to analyze the main results/hypotheses BEFORE the study

Subjective Ratings as Dependent Variables

- Rating Scales
  - Commonly used to solicit subjective feedback
  - E.g., NASA-TLX (Task Load Index) to assess mental workload
  - E.g., Likert scale Good==1 .... Bad==5
  - E.g., "It is frustrating to use the interface"

  Strongly Disagree | Disagree | Neutral | Agree | Strongly Agree
- Can also be statistically analyzed

Different Designs

- Within-Subjects
  - Everybody does all the conditions (interface A, task 1…9; interface B, task 1…9, interface C, task 1…9)
  - Can account for individual differences and reduce noise (and therefore might statistically more powerful)
  - Limits the number of conditions, and even types of tasks tested (may be able to workaround by having multiple sessions)
  - Can lead to ordering/learning effects
- Between-Subjects
  - Divide the participants into group, and each group does some of the conditions
  - Costly: More participants
  - How to ensure both groups are equal?
- Do Within-Subjects if you can ... [Andy Field]

based on Heidi Lam's notes

Careful study design

- Number of participants
  - Depends on effect size and study design--power of experiment
  - Usually 12-20 (per group)
- Possible confounds?
  - **Learning effect**: Did everybody use the interface in a certain order? If so, are people faster because they are more practiced, or because of the effect of the interface?
  - **Biasing**: "compare my interface vs. this other interface",
- Pilot studies
  - Should test the study setup for possible problems
- Other factors
  - Cross-cultural: Does it hold for all humans?
  - Child studies: Unbiased with vis. conventions!

based on Heidi Lam's notes

Quantitative Challenges

- Errors:
  - Type I: False positives (usually more critical)
  - Type II: False negatives
- Internal Validity
  - Are there alternate causes?
- External Validity
  - Can we generalize the study?
- Ecological Validity
  - Reflects realistic environment?
  - (usually low for lab studies)
- Comparing visualization tools?
  - What are the actual causes for tool A being ‘better’ than tool B?

Internal Validity: Storks deliver babies!?

- R. Matthews, "Storks Deliver Babies". Journal of Teaching Statistics, vol. 22, issue 2, pages 36-38, 2001;
- There is a correlation coefficient of r=0.62 (reasonably high)
- A statistical test can be employed that shows that this correlation is in fact significant **(p = 0.008)**
- What are the flaws?

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

C2. Qualitative Evaluation

Focus

- Realism

- Reveal: “a richer understanding by using a more holistic approach” (Carpendale, 08)

Qualitative Techniques

- Observation Techniques
  - In Situ Observations (fly-on-the-wall)
  - Participatory Observations
  - Laboratory Observational Studies
- Interview Techniques
  - Contextual Interviews
  - Focus Groups

Often, qualitative methods in addition to other studies ...

- Experimenter Observations
- Think-Aloud Protocol
- Collecting Participants Opinions

Helpful for...

- Usability Improvement (cf. section B)
- New insights, explanation of unforeseen results, new questions
- Can help to confirm results

Qualitative Methods as Primary

- Pre-design studies
  - Rich understanding of a complex domain
  - Problems, challenges, domain language
- During-, Post-design studies
  - Case studies/ Field studies

Helpful for...

- holistic understanding

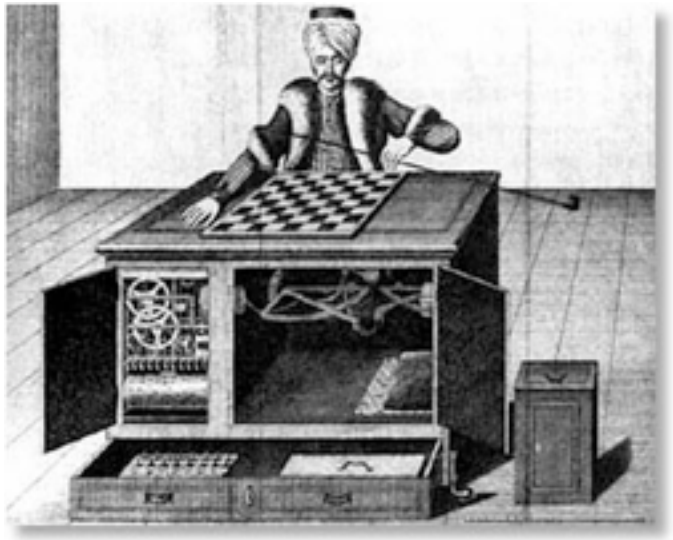
Qualitative Challenges

- Sample Sizes
  - Less participants: Doing intensive studies with a lot of participants?
  - Time? Data produced?
- Subjectivity
  - Social relationship?
  - Transferability, not generalizability
- Analyzing the data
  - Open coding, based on previous work, theory-driven coding
- Complexity of Data Analysis Environments
  - Very large datasets
  - High-Level Domain Task vs. Low-Level Vis Task
  - Often: Large amount of domain knowledge

III. Conclusion and Questions

New Ways of Evaluation

- Measuring Insights
- Mechanical Turk
- ...



Questions?