Human-Centric Computing

Usability Evaluation

Heuristic Evaluation

Qualitative and Quantitative Methods

Lingyun Yu

Overview

- usability and usability evaluation
- heuristic evaluation techniques
- qualitative evaluation techniques
- quantitative evaluation techniques

Quantitative Evaluation Techniques

Quantitative Evaluation Techniques

- quantitative evaluation
  - precise measurements
  - results in form of numeric values
  - how correct the statements are

- methods
  - user performance data collection
  - controlled experiments

Collecting Performance Data

- people using a system (often lots of data)
- targeted data collection
  - look for specific information, but may miss something
  - e.g., frequency & type of request for online assistance
  - e.g., frequency of use of different parts of the system
  - e.g., **number of errors** and where they occurred
  - e.g., **time** it takes to complete some operation
  - all these tell you something about the usability

Controlled Experiments

- What is controlled experiments?
  - This is when a **hypothesis is scientifically tested**. In a controlled experiment, **an independent variable (the cause)** is systematically manipulated and the **dependent variable (the effect)** is measured.
- striving for
  - removal of experimenter bias
  - clear and testable hypothesis
  - control of variables and conditions
  - quantitative measurement
  - replicability of experiment
  - measurement of confidence in obtained results (statistics)

Removal of Experimenter Bias

- unbiased instructions
- unbiased experimental protocols, for instance, by preparing scripts ahead of time
- unbiased subject selection

novice

expert

Experimental design

- To determine which participants to involve for which conditions in an experiment
- Avoid learning effect: participants who had additional training or knew the testing content have more chances of answering more questions correctly.

Experimental design - Participants

- Three experimental design:
  - Between-subjects: different participants perform in different conditions.
    - No ordering or training effects.
    - large numbers of participants are needed.
  - Within-subjects: all participants perform in all conditions
    - lessen the impact of individual differences and see how performance varies across conditions for each participant.
    - Need to ensure the order in which participants perform tasks for this setup does not bias the results (*counterbalancing*)
  - Pair-wise: participants are matched in pairs based on certain user characteristics such as expertise and gender. Each pair is randomly allocated to each experimental condition.

Experimental design – Collected data

- The data collected to measure user performance on the tasks set in an experiment usually includes:
  - Times to complete a task (efficiency)
  - Number of errors per task (accuracy)

Clear and Testable Hypothesis

- hypothesis: statement about the experiment
- **Specific hypotheses** are tested that make a prediction about the way users will perform with an interface (for instance, one interface feature is easier to understand or faster to use than another)
- Often based on a theory, or previous research finding
- hypotheses tests are used to test the validity of a statement that is made about an interface

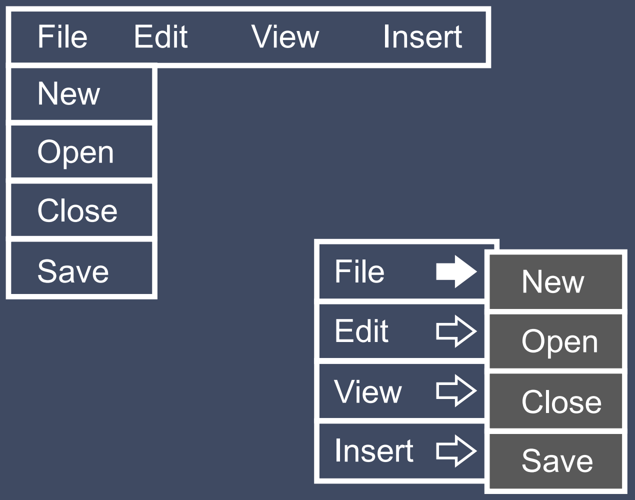
Hypotheses Testing

- A hypothesis involves examining a relationship between two things, called variables.
- Variables can be independent or dependent.

- independent variables: variables that are to be altered, independent of the participants' behavior
- dependent variables: variables that will be measured, depending on the participants' reactions to the independent variables in the experiment, included in hypothesis

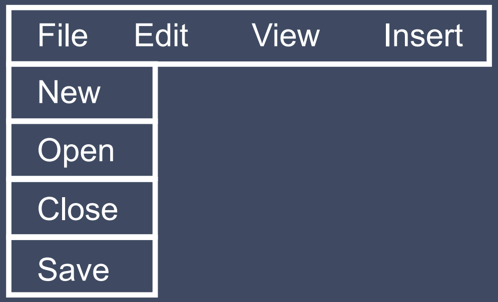
Independent Variables (Factors)

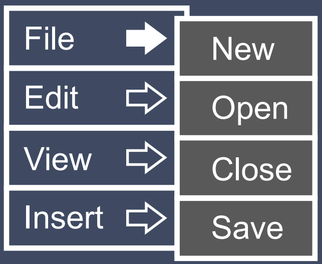
- variables that are to be altered
  - ***independent*** of the participants’ behavior
  - modification to the conditions the participants undergo
  - could also be the classification of participants into groups
- example:
  - menu type: pop-up or pull-down



Dependent Variables (Measures)

- variables that will be measured
  - ***depend*** on the participants' reactions to the independent variables in the experiment, included in hypothesis
  - specific things that will be measured quantitatively
- example: menus
  - time to select an item
  - selection errors made
  - time to learn to use it





Clear and Testable Hypothesis

- examples – valid hypotheses?
  - The Chinese are great table tennis players.
  - The Chinese are better table tennis players than the Germans.
  - The Chinese have won more table tennis competitions than the Germans in the last four years.
  - The Chinese have won more table tennis competitions at the Olympic, world, and European championship levels than the Germans in the last four years.
- hypotheses need to be clear, specific, and testable statements about the world/about our experiment

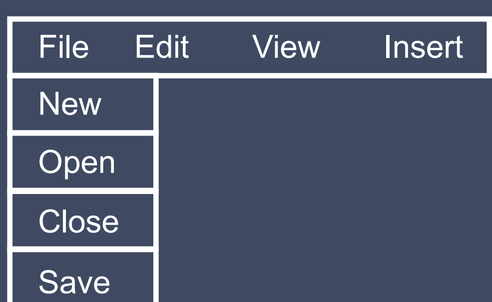
Null Hypothesis

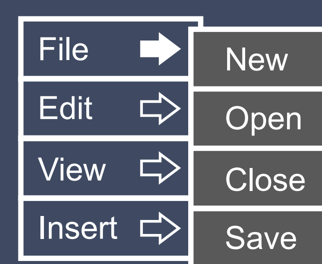
- Null hypothesis
- Alternative hypothesis
- goal: ***disprove*** the null hypothesis

Null Hypothesis

- example:
  There is no difference in user performance (time & error rate) when selecting a single item from a pop-up or pull-down.

- Alternative hypothesis would state that there are differences between the two regarding selection time and error rate.



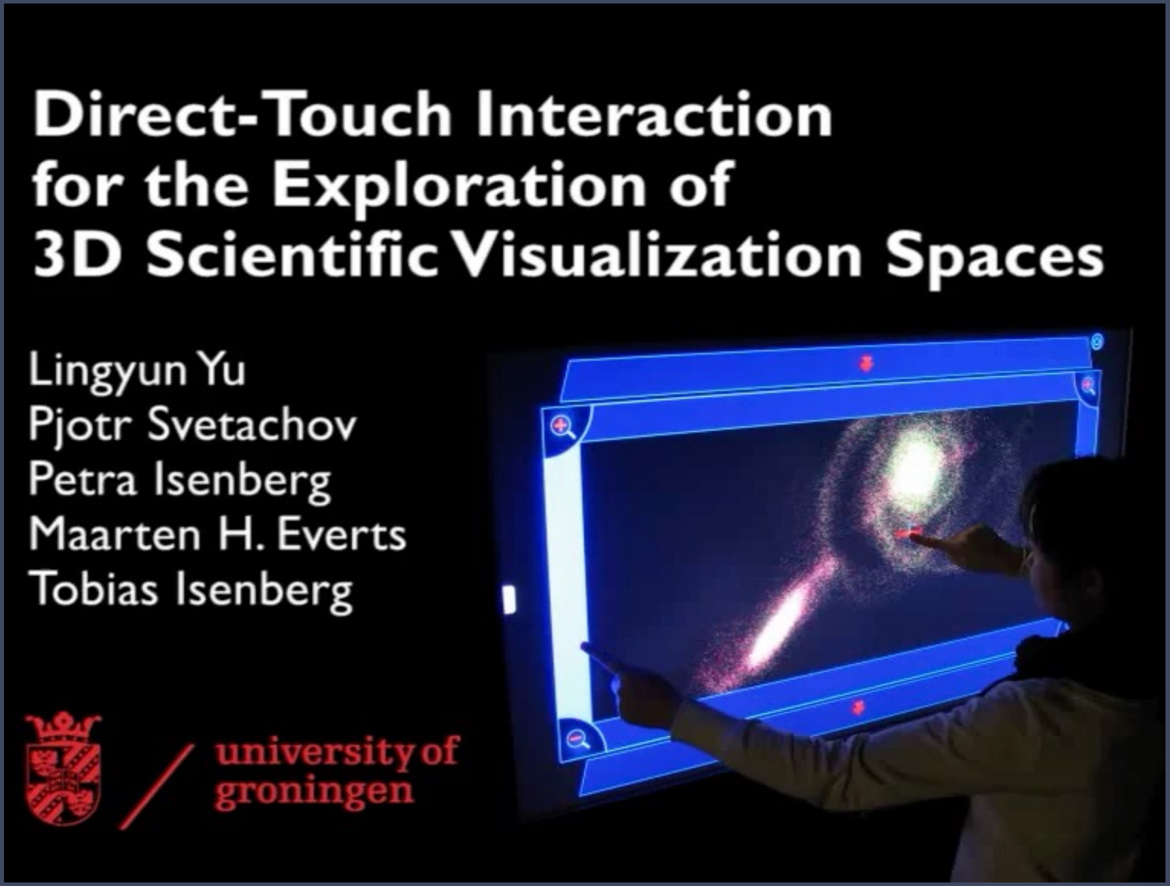


Statistical Methods

- P-value
- T-test
- paired T-test
  (when samples are paired, e.g., single group of users in two experimental conditions)
- one-tailed vs. two-tailed T-test
  (whether one mean is greater than the other is important vs. direction of difference not important)
- ANOVA (analysis of variance):
  compares relationships between many factors
- …

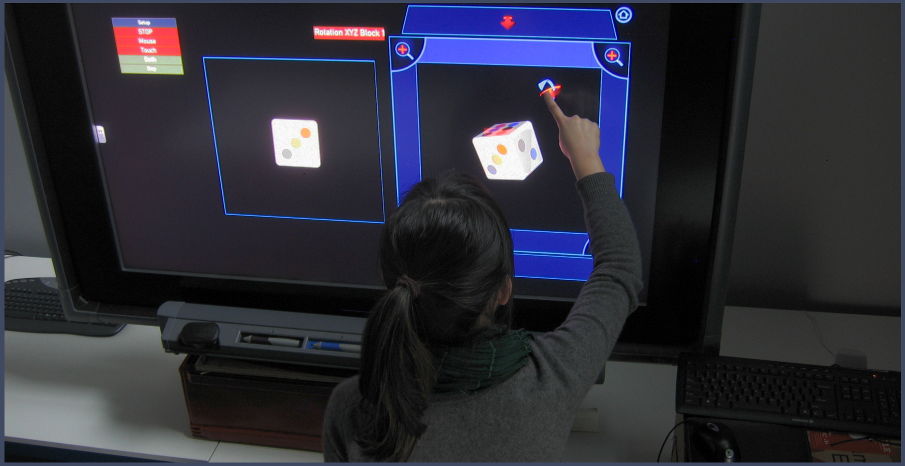
Example

- FI3D - Direct-Touch Interaction for the Exploration of 3D Scientific Visualization Spaces



A comparative study

- Baseline: the standard interface, mouse + keyboard



Participants

- Twelve members (6 male, 6 female) from the local university participated in the study.
- Seven participants reported prior experience with 3D computer games.
- Ages ranged from 19 to 39 (M =27.25; SD=5.29).
- All participants were right-handed.
- Eight participants were students from different disciplines and four non-students.

Apparatus

- The experiment was performed on a 52” LCD screen with full HD resolution (1920 * 1080 pixels, 115.4 cm * 64.5 cm).
- The system ran on Windows 7, and the mouse pointer speed was set to the average speed (half-way between slow and fast).

Tasks

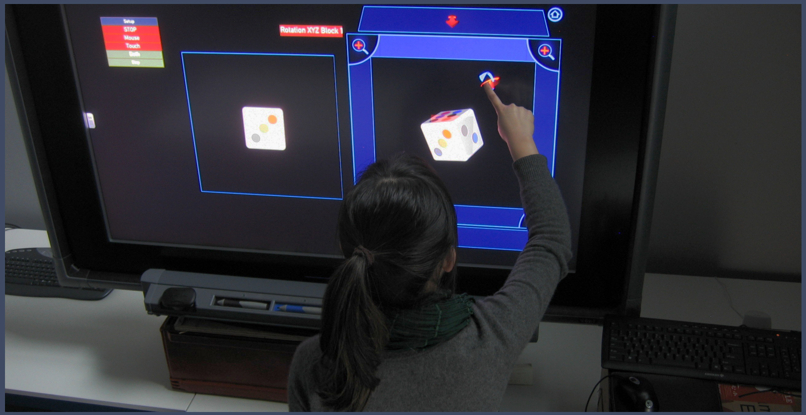
- Eight travel tasks

Table 2. Sequence of 3D interaction tasks per condition.

1	translation *x*/*y*	center object in target area
2	translation *x*/*y*/*z*	center, fill object in target area
3	rotation *z*	rotate object to face screen
4	rotation trackball	rotate object to face screen
5	rotation *x*	rotate object to face screen
6	rotation *y*	rotate object to face screen
7	zoom	zoom object to fill target area
8	rotate-scale-translate (RST)	center, fill object in target area

Tasks

- Design: We used a *repeated-measures design* with the *within subject.*
- *Independent variable*: input device (touch, mouse).
- *Dependent variable*: times to finish each task.
- Each participant performed 4 runs of 4 trials for each input device and task.
- Input devices were *counterbalanced* among participants. The first two runs were discarded as practice runs for the final analysis of the data.

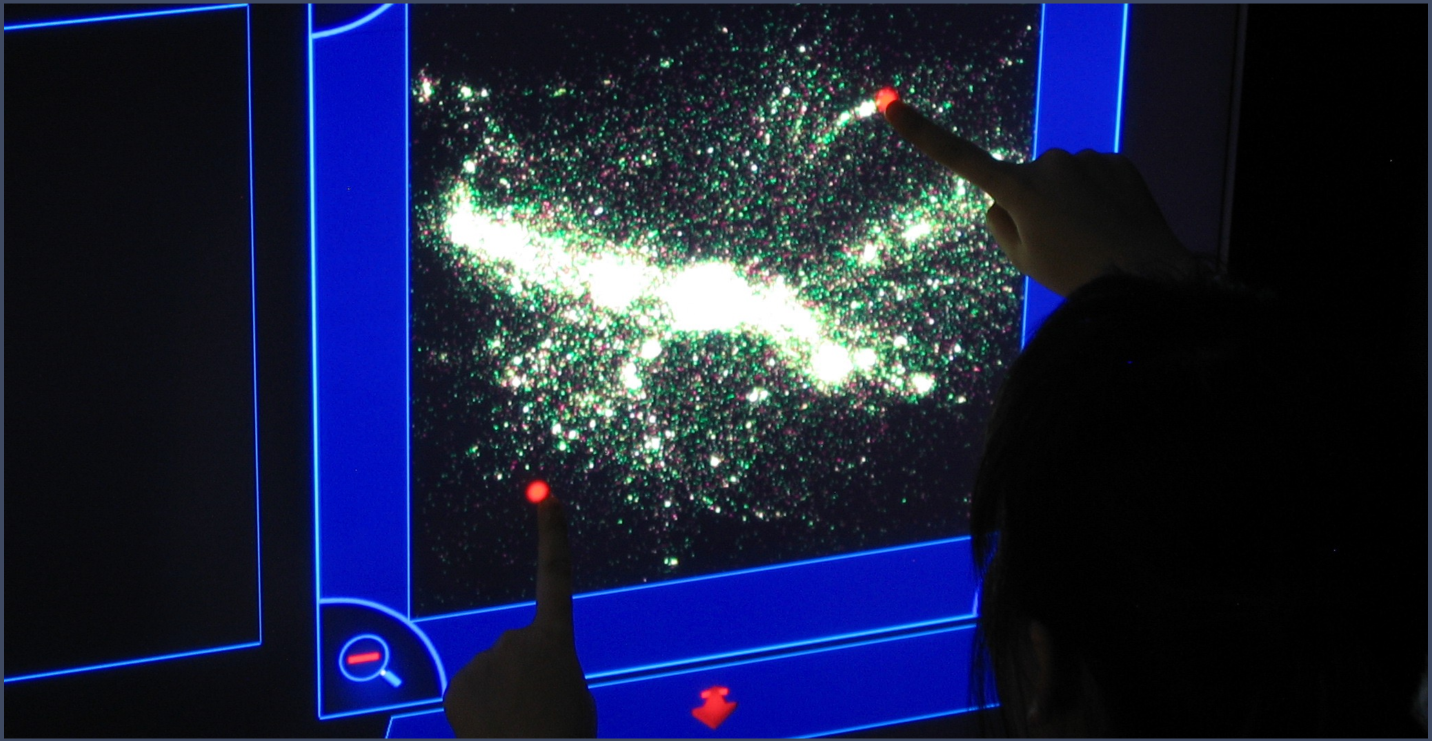


Procedures

- Before each task for each input device: introduction + a set of practice trials.
- Real experiments.
- After completing all tasks with one input device, participants were asked to complete a *questionnaire* to rate the usability of the technique in terms of ease of use, ease of remembering, precision, efficiency, and difficulty on a *seven step Likert scale*.
- After the second condition, participants were also asked to *compare* both techniques, voice their preference, and give *qualitative feedback*.

Tasks

- One longer wayfinding task



Procedures

- During the wayfinding task, we collected only qualitative feedback and took notes on participants' interactions.
- Finally, after finishing the wayfinding task, participants were asked to fill in the final part of the questionnaire to comment on which technique they preferred and why, and whether this technique allowed them to explore the data as they desired.
- Also, they filled in their background information and were asked to provide additional feedback on their experience of which we took notes.

Results

- time was analyzed using a repeated-measures ANOVA (touch, mouse).
  - Translation: no significant effect of input device on the translation time for x-/y-translation ($F(1,11) = .075$; $p = .79$) with mean completion times increasing from 5.49 s for the touch condition to 5.86 s for the mouse.
  - Both techniques scored highly with a median of 6 (agree) for the mouse and 6.5(agree–strongly agree) for touch.
  - Zooming: The analysis of task completion time for the zooming task showed a significant difference between both input types ($F(1,11) = 64.70$; $p < .001$) with mouse being significantly faster.

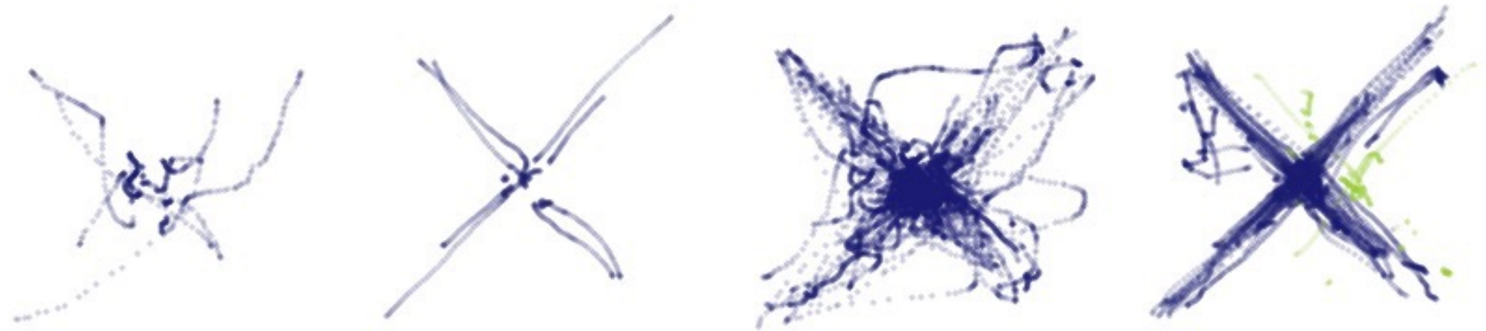


Fig. 9. Translation for mouse and touch for a single participant (left two images, resp.) and for all participants overlaid (right two images, resp.).

Example 2 – selection techniques

- CAST family

A comparative study

- Goal: In order to understand **efficiency** and **accuracy** with our new selection methods in different situations we conducted a comparative quantitative user study. We compared the Cast methods with the traditional selection methods.

Participants

- Twenty people (14 male, 6 female) participated in the study. 15 participants were students from different disciplines and five non-students. Eight of them had at least a Bachelor's degree.
- 16 participants reported prior experience with 3D computer games with playing games up to two times per day, with ten participants reporting at least weekly experience.
- Ages ranged from 18 to 35 years ($M = 23.4$; $SD = 4.8$).
- All participants reported to be right-handed.

Apparatus

- The experiment was performed on a Microsoft Surface Pro 2 (1280 * 800 pixels). The surface recognizes touch as well as a pen input.

Datasets

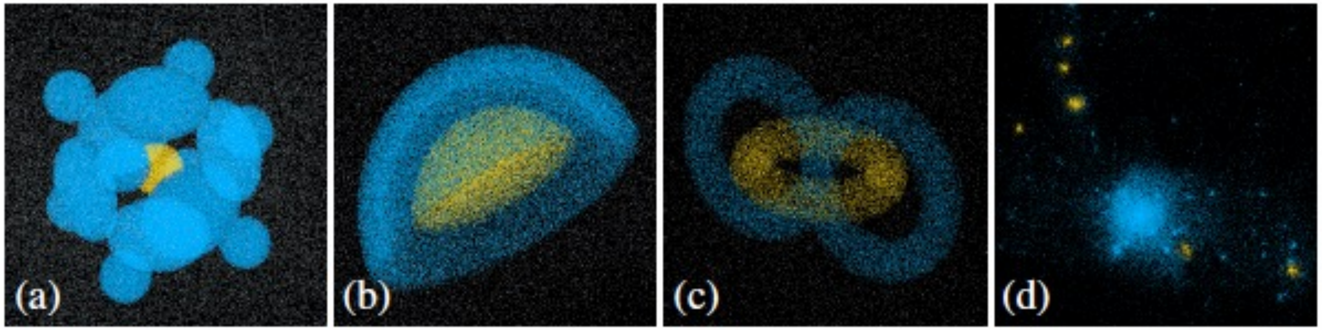


Fig. 4. Four datasets: (a) multiple particle clusters, (b) semispherical shell with half-ball inside, (c) three rings, and (d) an N-body mass simulation; named *clusters*, *shell*, *rings*, and *simulation* in the study description.

Design

- We used a repeated-measures design with the *within-subject*
- *Independent variables:* selection technique (CylinderSelection, CloudLasso, SpaceCast, TraceCast, and PointCast) and dataset (Clusters, Shell, Rings, Simulation).
- *Dependent variables*: accuracy and time
- Each technique was used for each dataset in 3 repetitions. In summary, the design consisted of 20 participants * 5 methods * 4 datasets * 3 repetitions = 1200 trials in total.

Tasks

- Before the actual experiment, participants practiced with three additional *training* datasets to get accustomed to the selection techniques.
- We asked participants to perform their selections as quickly and accurately as possible but we did not tell them when they had accomplished the selection goal.
- We also *counter-balanced* the order of selection methods presented to the participants such that each possible starting-ending pair of methods was used exactly once and that no participant saw the same progression order.

- We allowed participants to undo/redo the five most recent operations. Once they felt that they accomplished the selection goal or that they were not able to improve the result, they could press a finish button to advance to the next trial.

Results

Table 1. Mean task completion times, accuracy scores, and their 95% confidence intervals per technique.

Technique	Time	CI	F1	CI	MCC	CI
Cylinder	88s	[73,106]	.96	[.95,.97]	.96	[.95,.96]
CloudLasso	43s	[38,48]	.96	[.96,.97]	.96	[.95,.97]
SpaceCast	23s	[20,27]	.97	[.94,.97]	.96	[.93,.97]
PointCast	16s	[13,19]	.97	[.93,.98]	.97	[.93,.98]
TraceCast	18s	[16,20]	.97	[.97,.98]	.97	[.97,.98]

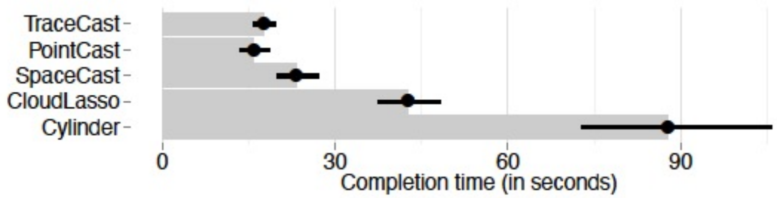
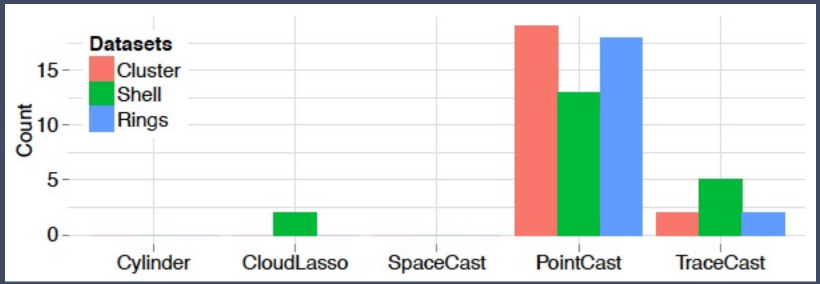


Fig. 5. Mean completion time (in seconds) across all participants for each selection technique. Error bars show 95% confidence intervals. Means are geometric means.



Quantitative Evaluation: Summary

- participants – task - data – hypothesis – measure – statistics
  - participants: sample (the group of individuals who participate in your study), population (the broader group of people to whom your results will apply)
  - task: datasets, design, apparatus
  - data: measured from sample to learn about the population
  - hypothesis: statement
  - statistics: analysis of results, confidence, hypothesis support? Reject?

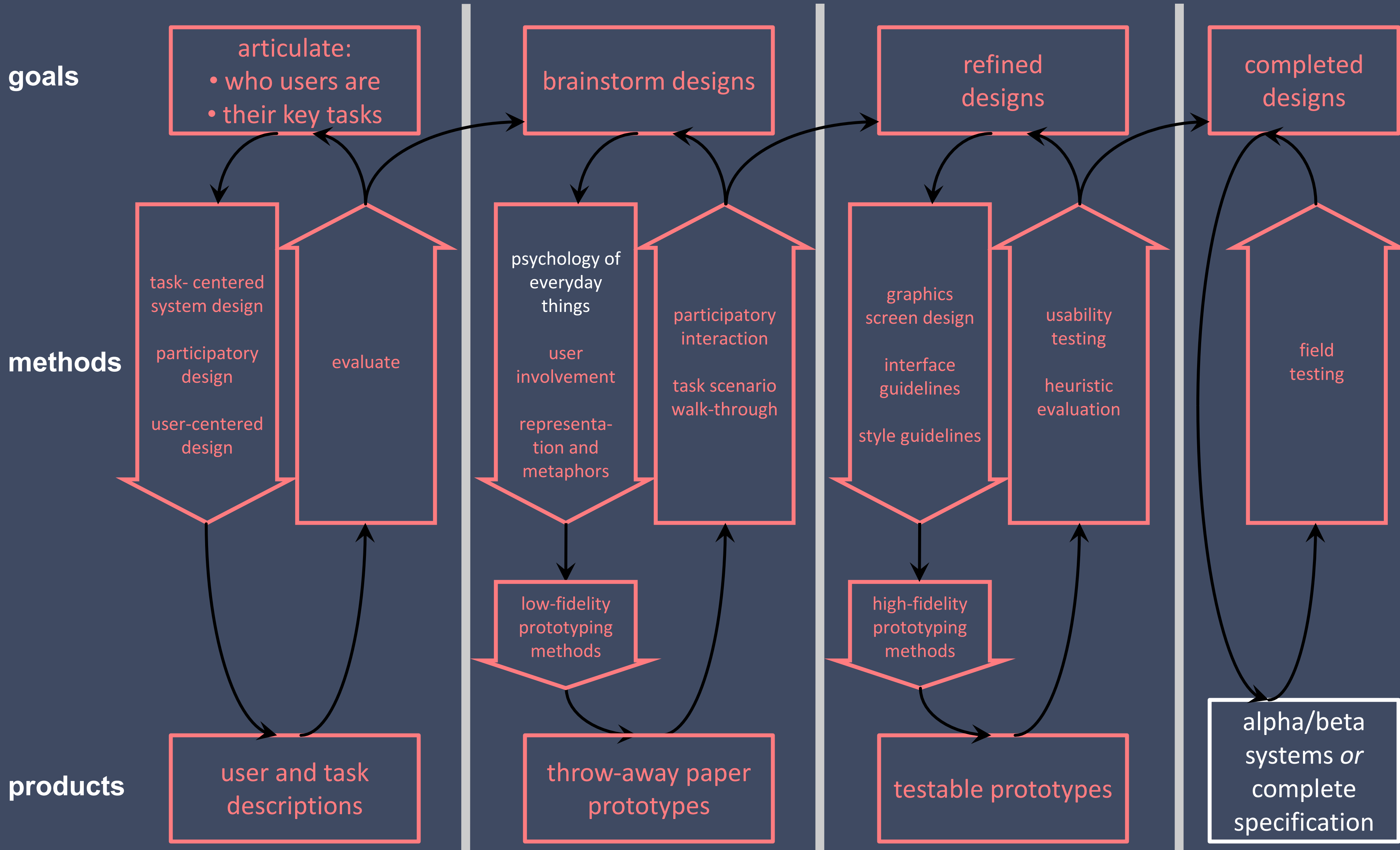
Usability Evaluation: Summary

- heuristic evaluation
  - check interface against design principles & heuristics
  - avoid more or less obvious problems
- qualitative evaluation
  - walk-throughs, interviews, direct observation …
  - evaluate non-numeric aspects of interfaces
- quantitative evaluation (controlled studies)
  - measurements and statistical evaluation
  - learn about specific aspects, numeric results

- go back, refine, and repeat … ☺

Some Final Words and Evaluation of the Class

An Interface Design Process
(we have discussed all the red parts)



Class Contents and Structure

- what are we learning here?
  - human-centric design principles & practice
  - overview of visualization techniques
  - overview of novel interaction techniques & platforms
  - unique challenges & opportunities of these
  - experience of going through a design process
- teaching structure
  - Human-centric design topics that are covered
  - connections between lecture and presentation
  - presentations and implementations

Class Contents and Structure

- what did you like best/not as much?
- what was easy, what was hard?
- other comments?
- MQ

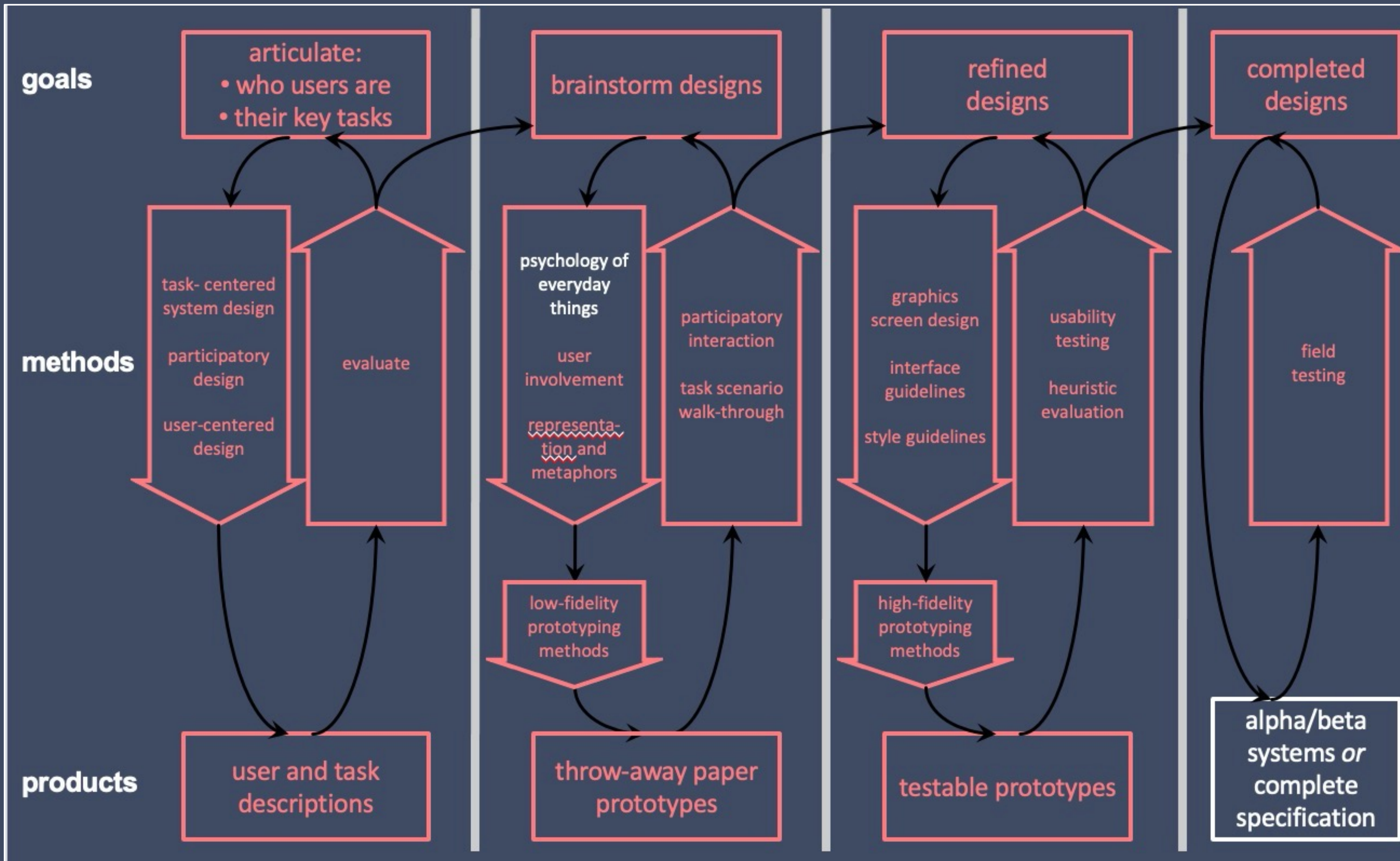
What will the exam be about?

- what we talked about in class: principles & practice
  - design life cycle
  - task-centered design techniques
  - user-centered design, participatory design & prototyping techniques
  - design principles & heuristics
  - visualization
  - user interaction
  - Implementation (Java swing)
  - evaluation techniques (heuristic, qualitative, quantitative)
- use slides & teaching material to study

What will the exam be about?

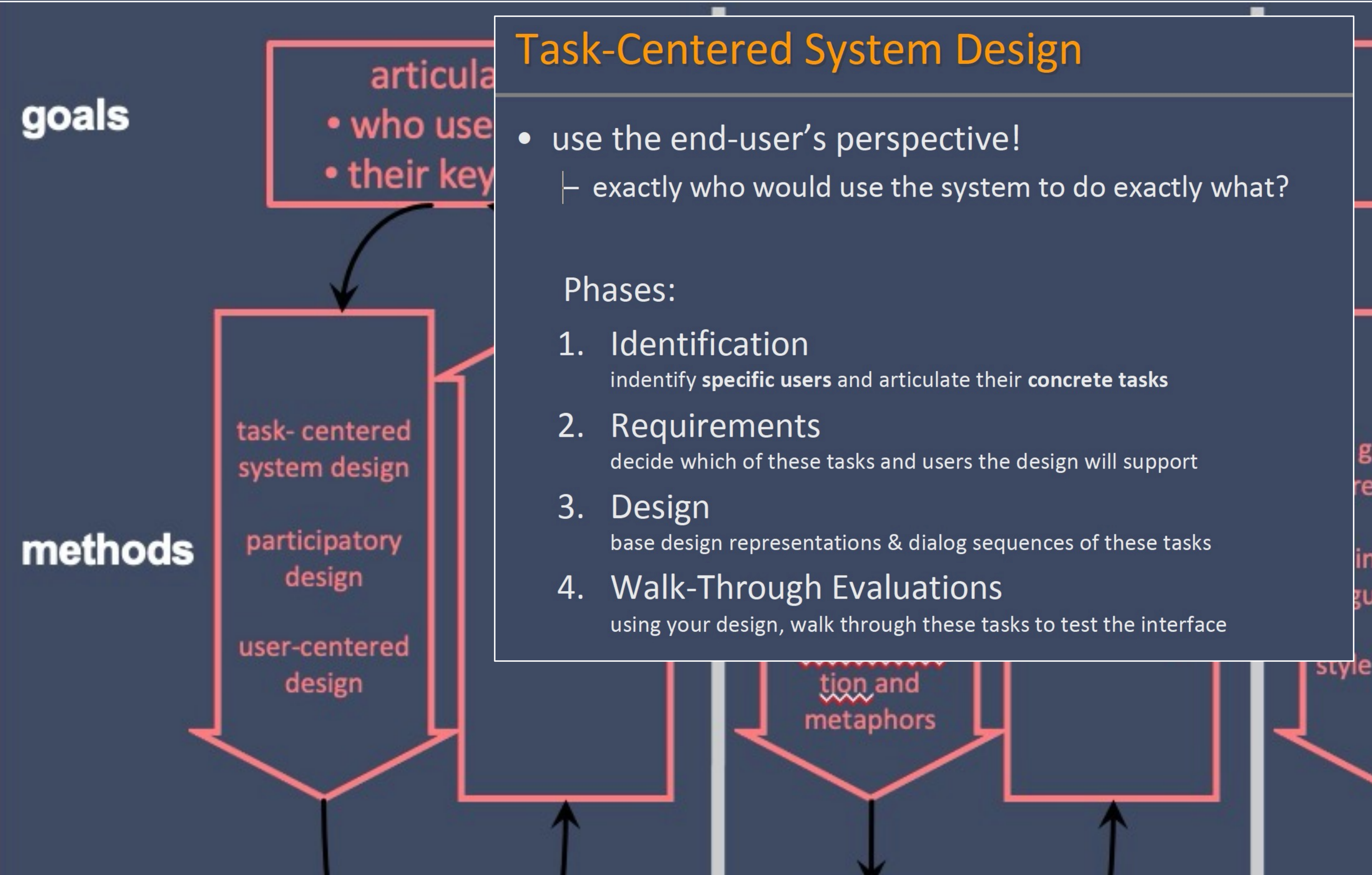
- overview and detail
  - know the overview of each of the topics
  - know details for some of the techniques
  - know examples

Example



goals
methods
products
articulate:
• who users are
• their key tasks
brainstorm designs
refined designs
completed designs
task- centered system design
participatory design
user-centered design
evaluate
psychology of everyday things
user involvement
representa- tion and metaphors
participatory interaction
task scenario walk-through
graphics screen design
interface guidelines
style guidelines
usability testing
heuristic evaluation
field testing
low-fidelity prototyping methods
high-fidelity prototyping methods
user and task descriptions
throw-away paper prototypes
testable prototypes
alpha/beta systems *or* complete specification

Example



Task-Centered System Design

- use the end-user's perspective!
  - exactly who would use the system to do exactly what?

Phases:

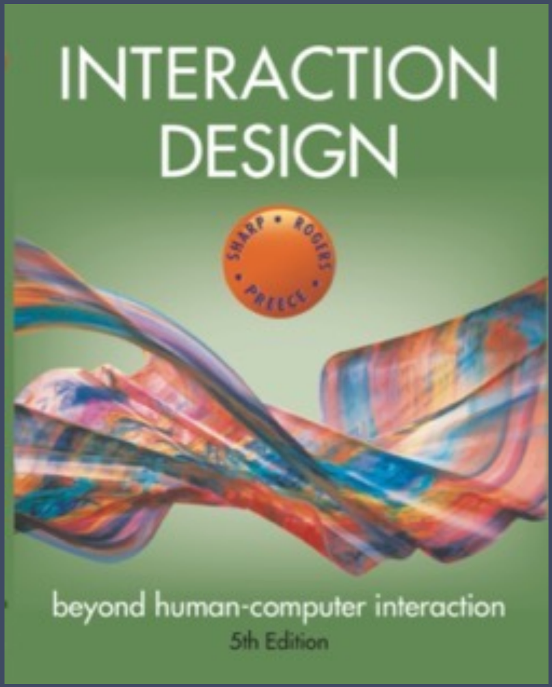
1. Identification
   indentify **specific users** and articulate their **concrete tasks**
2. Requirements
   decide which of these tasks and users the design will support
3. Design
   base design representations & dialog sequences of these tasks
4. Walk-Through Evaluations
   using your design, walk through these tasks to test the interface

Top 10 best ways to lose points

10 I always give answers without explanation.
9 I don’t care about the motivation of the work.
8 I choose qualitative evaluation, because it is easier to do, comparing to the quantitative user study.
7 I always choose the coolest HCI technology, just because it is cool.
6 I completely ignore the design principles, I think they are not useful at all.
5 I never make prototypes. I am a good programmer.
4 I never do evaluations, that’s for losers.
3 I know my users very well, so I don’t need to ask them all the time.
2 I tried very hard to solve the problems that I created.
1 I never care who the users are. I completely trust my design.

If you are interested to read more…

- the book that I suggest is "**Interaction Design: Beyond Human-Computer Interaction**" by Helen Sharp, Yvonne Rogers, and Jenny Preece (www.id-book.com)
  - Chapter 2 (the process of interaction design), for Design Life Cycle
  - Chapter 11 (discovering requirements), for Task-centered Design
  - Chapter 12 (design, prototyping and construction), for User-centered Design
  - Chapter 7 (interfaces), for Visualization and interaction techniques.
  - Chapter 14 (introducing evaluation) and Chapter 15 (evaluation studies), for Usability evaluation.



Finally…

- Thank you!
- MSc Human-Computer Interaction (HCI)
- VIS, HCI, VR and AR
- https://yulingyun.com (Lingyun.Yu@xjtlu.edu.cn)