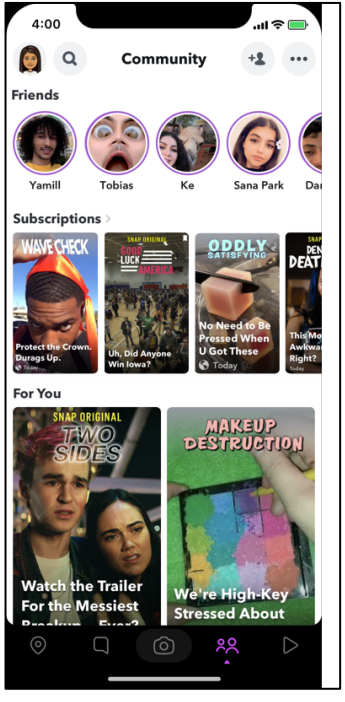
Aalto University

Lecture 8: Analytic Evaluation Methods

ELEC-D7010 Engineering for Humans
May 18, 2021
Antti Oulasvirta
Aalto University

How do we know how to use a UI?



Antti's example: Snapchat


FF
2
EJECT
3
SIDE 1-2
4
COMP
5
BASS
TREB
BAL
FADE
TAPE
CD
AUTOSET
SHUFFLE
6
My brother was upset because his car's "docking station" for his iPhone wasn't working and it was scratching his screen.

Today

Feedback for A3

Analytic evaluation methods

Cognitive walkthrough

Updates

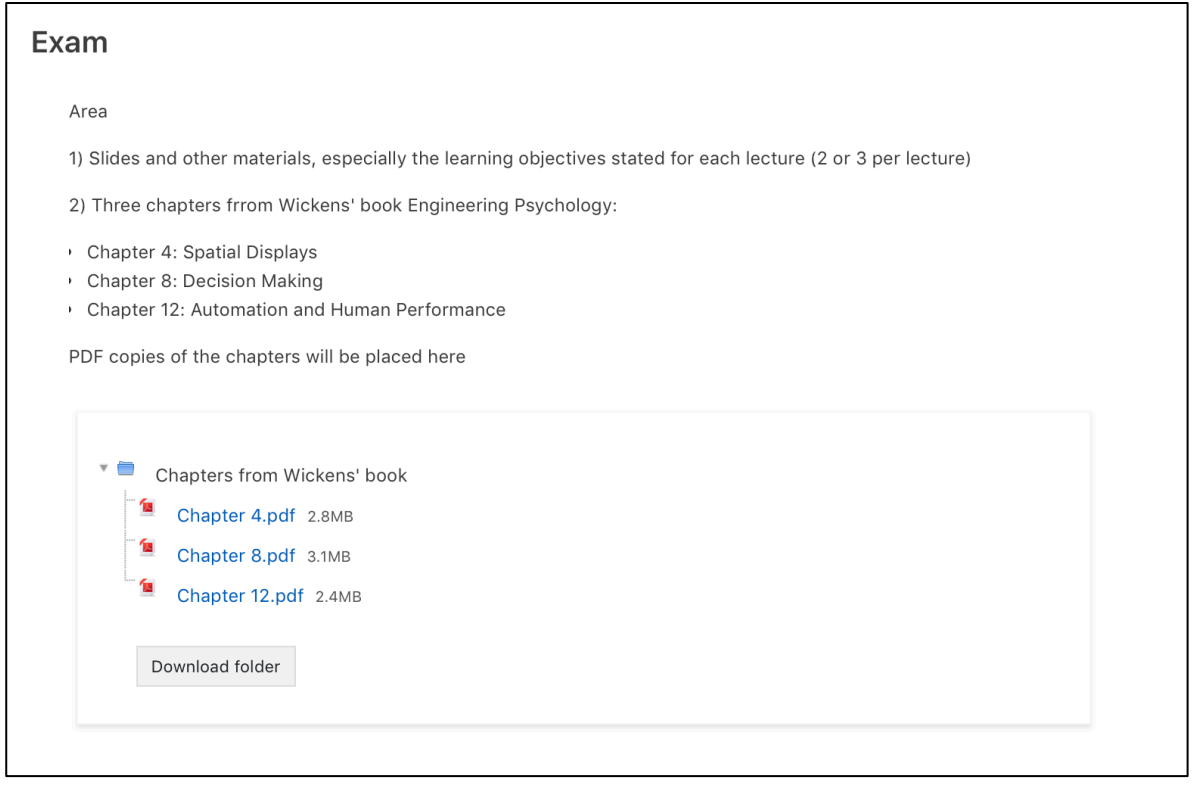
Rehearsal for exam: I will send a practice exam early next week.

Q: Is there be interest for an extra session to prepare for the exam together? May 27 at 12.15pm

Need more assignment points? I will launch an assignment sheet also for Lecture 10 and, if needed, an extra sheet due after the exam

Curious about professionals in this area? What would you like to ask from a human factors professional? I will open a poll of questions to Jari Laarni / VTT.

Exam area


Exam
Area
1) Slides and other materials, especially the learning objectives stated for each lecture (2 or 3 per lecture)
2) Three chapters frrom Wickens' book Engineering Psychology:
Chapter 4: Spatial Displays
Chapter 8: Decision Making
Chapter 12: Automation and Human Performance
PDF copies of the chapters will be placed here
Chapters from Wickens' book
Chapter 4.pdf 2.8MB
Chapter 8.pdf 3.1MB
Chapter 12.pdf 2.4MB
Download folder

Learning objectives today

1. Analytic Evaluation Methods
Which, when, and when not?

2. Cognitive Walkthrough
Ability to apply to a case

Assignment 8: Sneak preview

A8-1: Cognitive walkthrough [5p, recommended]

A8-2: STEM: A KLM modeling workbench [5p, optional]

Aalto University

Feedback: Assignment A3

Aalto University

Based on

Lecture materials by Saul Greenberg, University of Calgary, AB, Canada.
http://saul.cpsc.ucalgary.ca/saul/pmwiki.php/HCIResources/HCILectures

Recap: Overview of Evaluation Methods

By Saul Greenberg / University of Calgary

Four questions for today

Why do we evaluate?

Why should we master different methods?

How can we compare methods?

What methods are there?

Why Do We Evaluate?

Designer:
- user-centered iterative design

Researcher
- developing a knowledge base

Customer
- selecting among systems

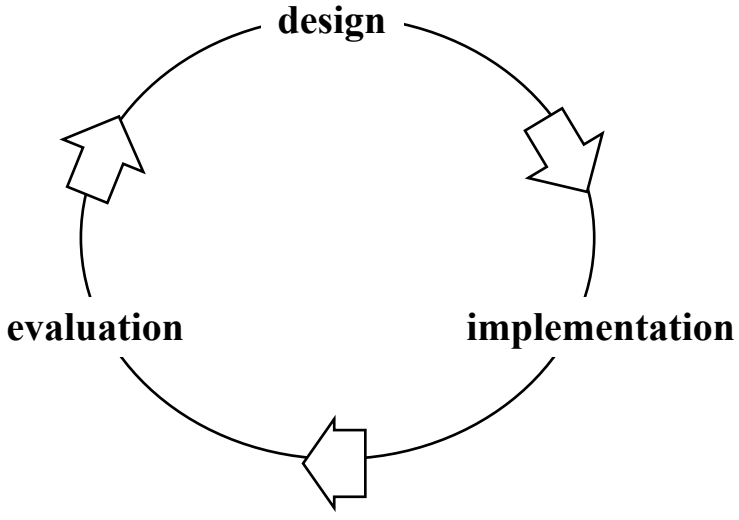
Manager
- assisting effectiveness

Marketer
- building a case for the product

(From Finholt & Olsons CSCW 96 Tutorial)

Why Do We Evaluate?

1. Evaluation is a necessary part of human-centred design



Why Do We Evaluate?

A. Pre-design stage: Evaluate design ideas

- what do people do?
- what is their real world context and constraints?
- how do they think about their task?
- how can we understand what we need in system functionality?
- can we validate our requirements analysis?

Why Do We Evaluate?

B. Initial design stage: Evaluate choices and sketches

- evaluate choices of initial design ideas and representations
- usually sketches, brainstorming exercises, paper prototypes
  - *is the representation appropriate?*
  - *does it reflect how people think of their task*

Why Do We Evaluate?

C. Iterative design stage: Evaluate prototypes

- iteratively refine / fine tune the chosen design / representation
- evolve low / medium / high fidelity prototypes and products
- look for usability bugs
  - *can people use this system?*

Why Do We Evaluate?

D. Post-design stage

- *acceptance test:* did we deliver what we said we would?
  - *verify human/computer system meets expected performance criteria*
  - *ease of learning, usability, user's attitude, time, errors…*
    - e.g., 9/10 first-time users will successfully download pictures from their camera within 3 minutes, and delete unwanted ones in an additional 3 minutes
- *revisions:* what do we need to change?
- *effects:* what did we change in the way people do their tasks?
- *in the field*: do actual users perform as we expected them to?

Why Do We Evaluate?

2. Evaluation to produce generalized knowledge

- generating design principles
- contributing to theories of human behavior and experience
  - *explanatory*
  - *predictive*
- validiting ideas / visions / hypotheses?

Why Do We Evaluate?

Design and evaluation

- Best if they are done **together**
  - *evaluation informs design*
  - *design suggests evaluation*
  - *use evaluation to create as well as critique*

- Design and evaluation methods **must fit** development constraints
  - *budget, resources, time, product cost…*
  - *do triage: what is most important given the constraints?*

- Design usually needs quick approximate answers
  - *precise results rarely needed*
  - *close enough, good enough, informed guesses,…*

Why Use Different Methods?

All methods have trade-offs:

- enable but also limit what can be gathered and analyzed
- are valuable in certain situations, but weak in others
- have inherent weaknesses and limitations
- can be used to complement each other's strengths and weaknesses.

-McGrath (Methodology Matters)

Why Use Different Methods?

Information requirements differ

- pre-design, iterative design, post-design, generalizable knowledge…

Information produced differs

- outputs should match the particular problem/needs

Relevance

- does the method provide information to our question / problem?

Why Use Different Methods?

Cost/benefit of using method

- cost of method should match the benefit gained from the result

Constraints and pragmatics

- may force you to chose quick and dirty discount usability methods

How Can We Compare Methods?

Is the method naturalistic?

- is the method applied in an ecologically valid situation?
  - *observations reflect real world settings*
    - real environment, real tasks, real people, real motivation

Is the method repeatable?

- would the same results be achieved if the test were repeated?

How Can We Compare Methods?

Validity

- External validity:
  - *can the results be applied to other situations?*
  - *are they generalizable?*

- Internal validity:
  - *do we have confidence in our explanation?*

How Can We Compare Methods?

Design-relevance

- Does the test measure something relevant to the usability and usefulness of real products in real use outside of lab?

- Some typical **reliability problems** of testing vs real use
  - *non-typical users tested*
  - *tasks are not typical tasks*
  - *tests usability vs usefulness*
  - *physical environment different*
    - quiet lab vs very noisy open offices vs interruptions
  - *social influences different*
    - motivation towards experimenter vs motivation towards boss

How Can We Compare Methods?

Quickness

- can I do a good job with this method within my time constraints?

Cost

- Is the cost of using this method reasonable for my question?

Equipment

- What special equipment / resources required?

Personnel, training and expertise

- What people / expertise are required to run this method?

How Can We Compare Methods?

Subject selection

- how many do I need, who are they, and can I get them?

Scope of subjects

- is it good for analyzing individuals? small groups? organizations?

Type of information (qualitative vs quantitative)

- is the information quantitative and amenable to statistical analysis?

Comparative

- can I use it to compare different things?

How Can We Compare Methods?

Control

- can I control for certain factors to see what effects they have?

Cross-sectional or Longitudinal

- can it reveal changes over time?

Setting

- field vs laboratory?

Support

- are there tools for supporting the method and analyzing the data?

How Can We Compare Methods?

Routine application

- is there a fairly standard way to apply the method to many situations

Result type

- does it produce a description or explanation?

Metrics

- are there useful, observable phenomena that can be measured

How Can We Compare Methods?

Measures

- can I see processes or outcomes

Organizational

- can they be included within an organization as part of a software development process

Politics

- are there ‘method religion wars’ that may bias method selection?

What methods are there?

Laboratory tests

requires human subjects that act as end users

- Experimental methodologies
  - *highly controlled observations and measurements to answer very specific questions i.e., hypothesis testing*
- Usability testing
  - *mostly qualitative, less controlled observations of users performing tasks*

What methods are there?

Analytic evaluation methods

done by interface professionals, no end users necessary

- Usability heuristics
  - *several experts analyze an interface against a handful of principles*
- Walkthroughs
  - *experts and others analyze an interface by considering what a user would have to do a step at a time while performing their task*

What methods are there?

Field studies

requires established end users in their work context

- Ethnography
  - *field worker immerses themselves in a culture to understand what that culture is doing*
- Contextual inquiry
  - *interview methodology that gains knowledge of what people do in their real-world context*

What methods are there?

Self reporting

requires established or potential end users

- interviews
- questionnaires
- surveys

What methods are there?

Cognitive modeling

requires detailed interface specifications

- Fitt's Law
  - *mathematical expression that can predict a user's time to select a target*
- Keystroke-level model
  - *low-level description of what users would have to do to perform a task that can be used to predict how long it would take them to do it*
- Cognitive models
  - *Computational, multi-level descriptions of what users would have to do to perform a task that can also be used to predict time, errors etc*

Becoming an expert evaluator is a long journey...

Professionals need to learn the full toolbox of evaluation methods

They need to:

- investigate, compare and contrast many existing methodologies
- understand how each methodology fits particular interface design and evaluation situation
- practice several of these methodologies on simple problems
- gain first-hand experience with a particular methodology by designing, running, and interpreting a study.

Recap: You know now

Why we evaluate

Why we use different methods

How we can compare methods

What methods there are

Aalto University

Analytic evaluation methods

Define “Analytic evaluation methods”

= A class of reasoning-based methods where the goal is to expose probable usability problems by analyzing a design in a structured manner

These methods build on (1) some method of systematically describing interaction, (2) reasoned with, and (3) the output of which is then compared against a set of criteria.

Non-empirical: No empirical research is needed, although some methods use an expert evaluator as a proxy for real participants

Role of analytic methods

1. In design, identify potential usability problems so that they can be rectified before deployment;

2. In evaluation, identify potential usability problems to compare against a baseline design or assess how ready a design is for deployment

3. In accident investigation, identify causes for potential errors.

Pros and cons

Appealing because

of their cost-efficiency; savings can be remarkable in comparison to an empirical study

See next slides

However,

they often have a high false positive rate and low reliability

they often fail to identify usability problems

success rate depends on the expertise of the evaluator.

→ They are best treated a <u>complement</u> to empirical evaluation that, when applied correctly, can decrease the cost of design.

Important analytic evaluation methods

Heuristic evaluation (Lecture 1)

Keystroke-level modelling (Lecture 3)

Task analysis (Lecture 4)

Cognitive walkthrough (this Lecture)

Note: Cognitive models also belongs to this category (not discussed in this course)

Nielsen's usability heuristics

Recap of heuristic evaluation

1. **Visibility of system status**
   - Keep users informed about system status.
2. **Match between system and the real world**
   - The system should speak the users' language.
3. **User control and freedom**
   - Give users clear "emergency exits" to leave the unwanted states. Support undo & redo.
4. **Consistency and standards**
   - Give users standard set of words, situations, and actions.
5. **Error prevention**
   - Careful design is better than good error messages.
6. **Recognition rather than recall**
   - Minimize the user's memory load. Make objects, actions, and options visible.
7. **Flexibility and efficiency of use**
   - Give accelerators to speed up the interaction for the expert users.
8. **Aesthetic and minimalist design**
   - Do not give information that is irrelevant or rarely needed.
9. **Help users recognize, diagnose, and recover from errors**
   - Error messages should be expressed in plain language.
10. **Help and documentation**
    - Documentation should be easy to search and focused on the user's task.

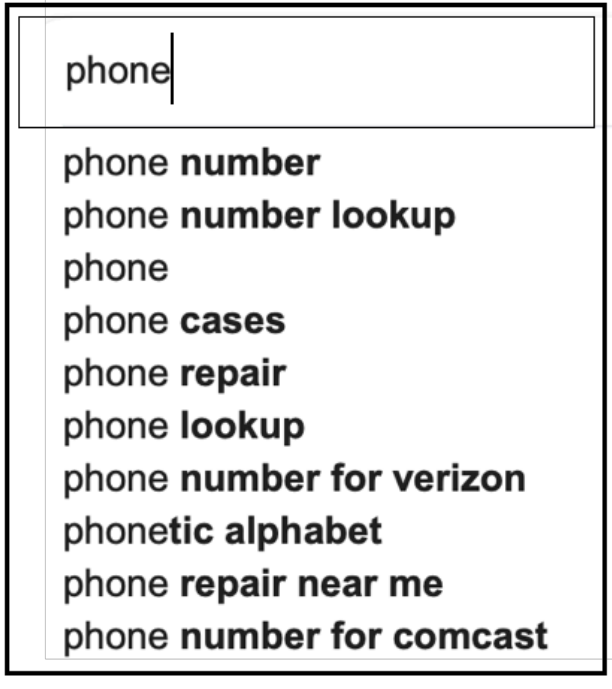
Human-AI evaluation heuristics

Do you know an intelligent UI that fails any of these?

1. Make clear what the AI system can do
2. Make clear how well it can do what it does
3. Time services based on context
4. Show contextually relevant information
5. Match relevant social norms
6. Mitigate social biases
7. Support efficient invocation
8. Support efficient dismissal
9. Support efficient correction
10. Scope services when in doubt
11. Make clear why system did what it did
12. Remember recent interactions
13. Learn from user behavior
14. Update and adapt cautiously
15. Encourage granular feedback
16. Convey the consequences of user actions
17. Provide global controls
18. Notify users about changes

Example

Heuristic evaluation of an autocomplete feature



Violations of Amershi et al.'s AI evaluation heuristics

Show contextually relevant information
* Suggestions can be irrelevant to task

Mitigate social biases
* Suggestions can be inappropriate

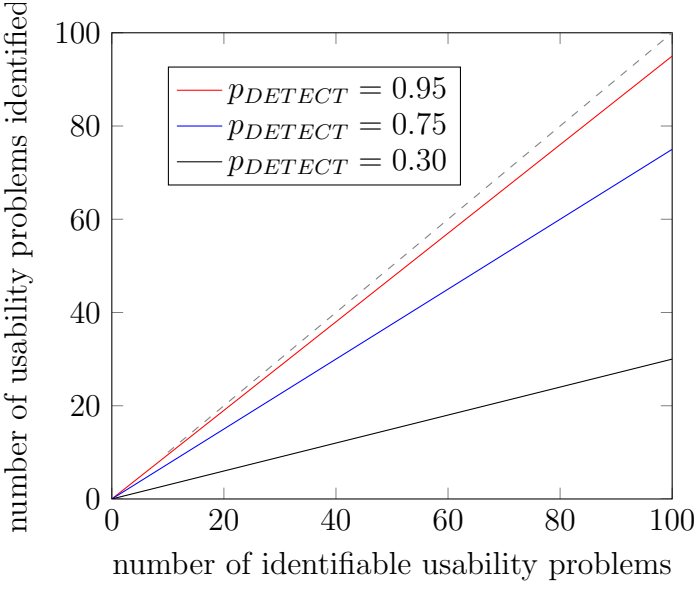
Support efficient dismissal
* Must go to settings to turn off

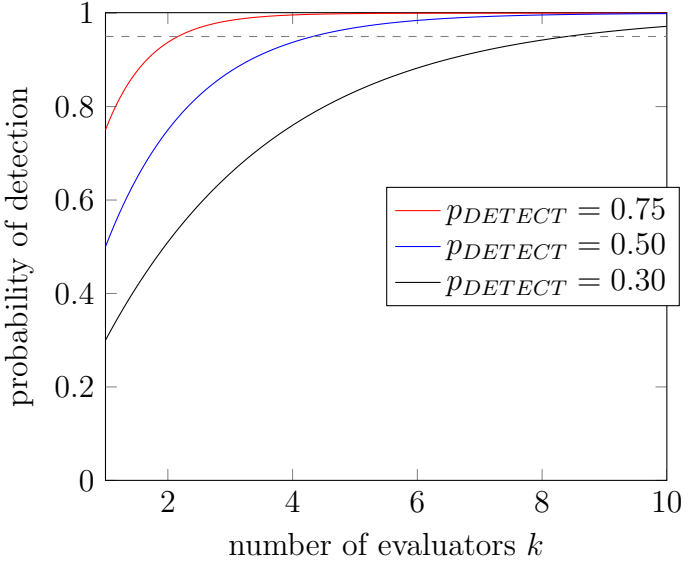
Support efficient correction
* No way to correct or edit suggestions

Make clear why system did what it did
* Unclear where the suggestions come from

How many evaluators? A statistics viewpoint

Think about an evaluator applying a heuristic as a Bernoulli trial (coin flip)





Conclusion

High coverage of problems is hard to achieve but possible

by increasing number of expert evaluators (red line)

Evaluations by a single evaluator are inherently unreliable

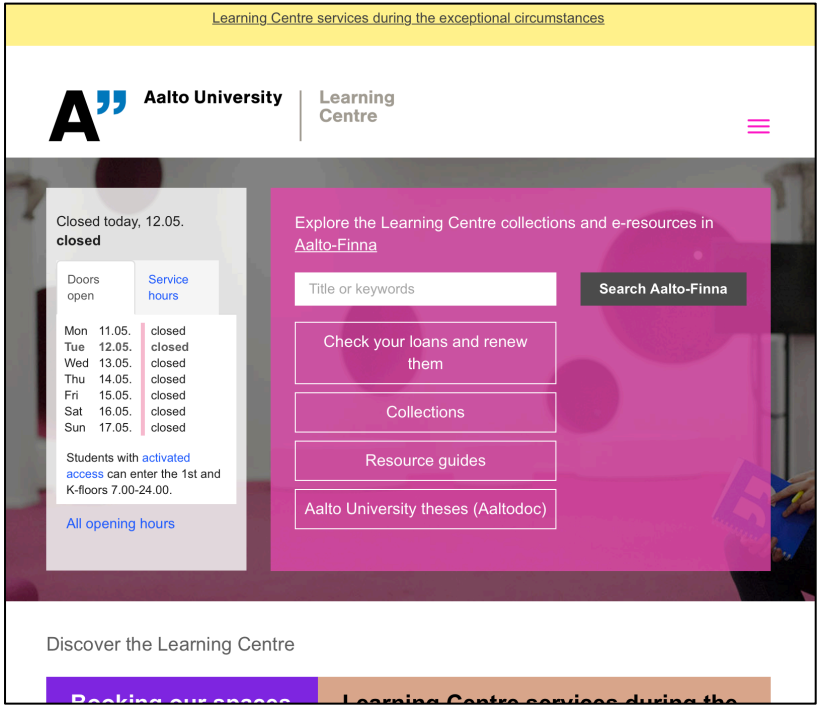
Even an expert evaluator will miss an unacceptably large proportion of ’obvious usability problems’ (blue)

On the other hand, even a poor evaluator will find *some* usability problems

Aalto University

Cognitive walkthrough

How do we know what to do when using a UI for the first time?



Cognitive walkthrough in a nutshell

https://www.youtube.com/watch?v=Edqjao4mmxM

(6 mins)

The theory of cognitive exploration
[Polson and Lewis 1990]

Novice users generate hypotheses on the goal-structures that help them solve a task with a UI

Idea generation is cued by the UI, effortful, and error-prone

First, a user must set a relevant *main goal* for the task:

How do I know what goals can be accomplished here?

They must then set a *goal structure* that consists of *subgoals* related to achieving the goal.

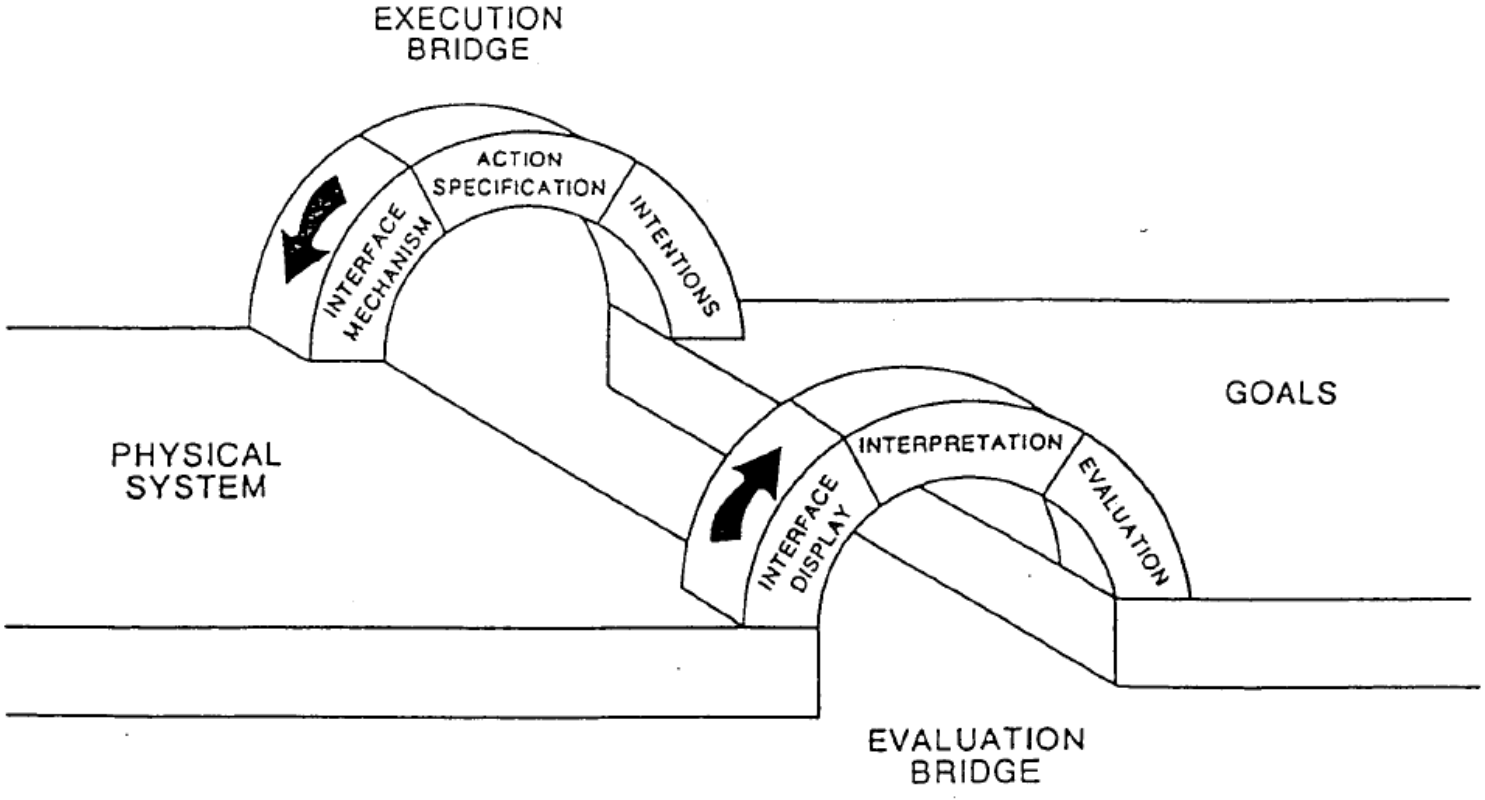
For each subgoal, they must solve:

How do I recognize what actions are available?

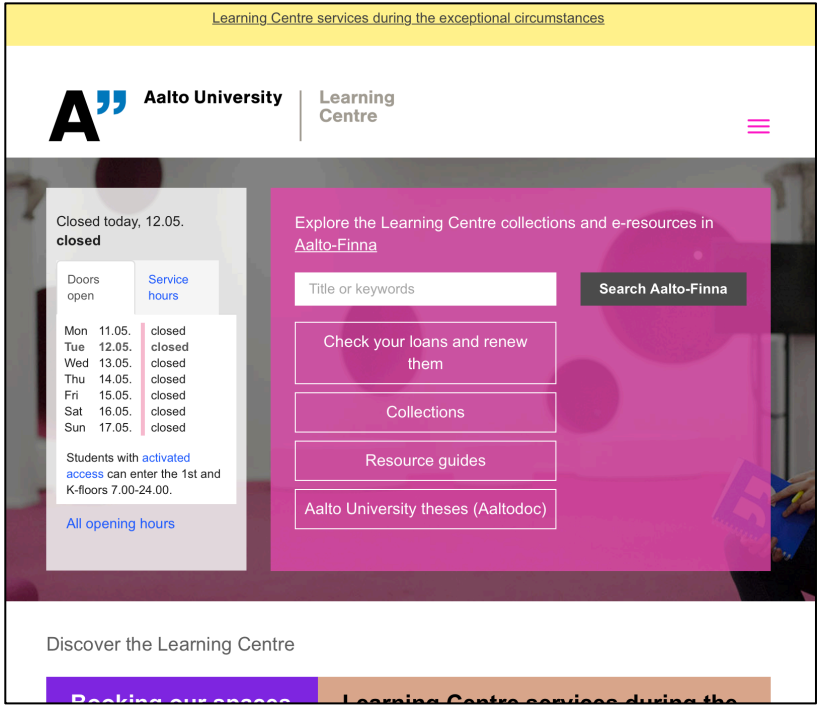
How do I know this action is what I want?

After executing, how do I know the action had the right effect?

Two metaphors: The gulfs of evaluation and execution



CW: Four questions to consider



1. Will the user try to achieve the right effect?
2. Will the user notice the availability of the correct action?
3. Will the user associate the correct action with the intended effect?
4. If the correct action is carried out, will the user be aware that the task is progressing as intended?

Cognitive walkthrough

CW is an instance of a broader class of walkthrough methods used across engineering disciplines, for example architectural walkthroughs and code walkthroughs.

An analytical evaluation method based on structured mental simulation how users think.

An artefact is inspected systematically, in a step-by-step manner, and evaluated against criteria.

What makes cognitive walkthrough special is that evaluation criteria are related to thinking and cognition

The question CW answers is this:

How might a novice user succeed or fail in interaction?

CW is a method for understanding ease-of-use

Exposes usability problems related to the ease-of-use of a system. It is recommended for understanding how *novice users* may figure out how to use a system.

There is evidence that it can predict a significant part of related usability problems:

- In the original study by Clayton Lewis and colleagues [1990], cognitive walkthrough detected almost 50% of usability problems exposed in an empirical user study. Similar results with 30% to 70% detection rate have been obtained across user interface types.

Preparations

The inputs to the method are

(1) the user interface,

(2) a task scenario that tells what the users are supposed to accomplish,

(3) assumptions about users and the contexts of use, and

(4) a sequence of actions that complete the tasks. Task analysis is needed to prepare this point.

Walkthrough procedure in more detail

For each user task, ask the following questions:

1. **Will the user try to achieve the effect that the subtask has?**

- Does the user understand that this subtask is needed to reach the user's goal?

2. **Will the user notice that the correct action is available?**

- E.g. is the button visible?

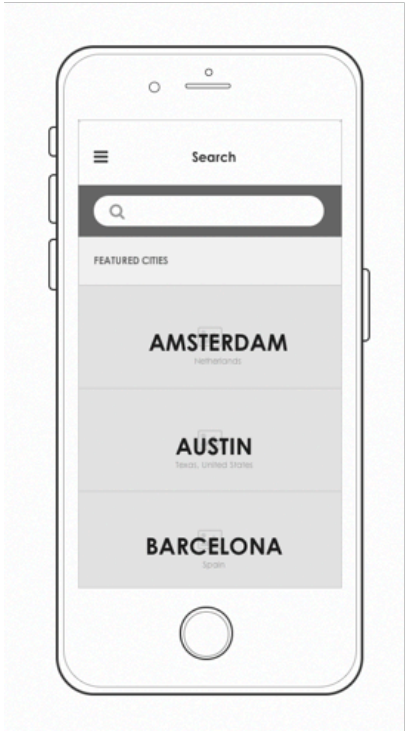
3. **Will the user understand that the wanted subtask can be achieved by the action?**

- E.g. the user does not understand a button and will not click on it

4. **Does the user get appropriate feedback?**

- Will the user know that they have done the right thing after performing the action?

Example



Task: Find hotels in Helsinki

Walkthrough questions

1. Will the user try to achieve the right effect?
2. Will the user notice the availability of the correct action?
3. Will the user associate the correct action with the intended effect?
4. If the correct action is carried out, will the user be aware that the task is progressing as intended?

Report

1. Yes, assuming familiarity with the concept of a search box
2. Yes, assuming that the icon is large enough
3. Yes, assuming familiarity with search boxes
4. Yes, the changing of the display will be clear

Aalto University

Modeling Workbenches

Good to know

Example: CogTool

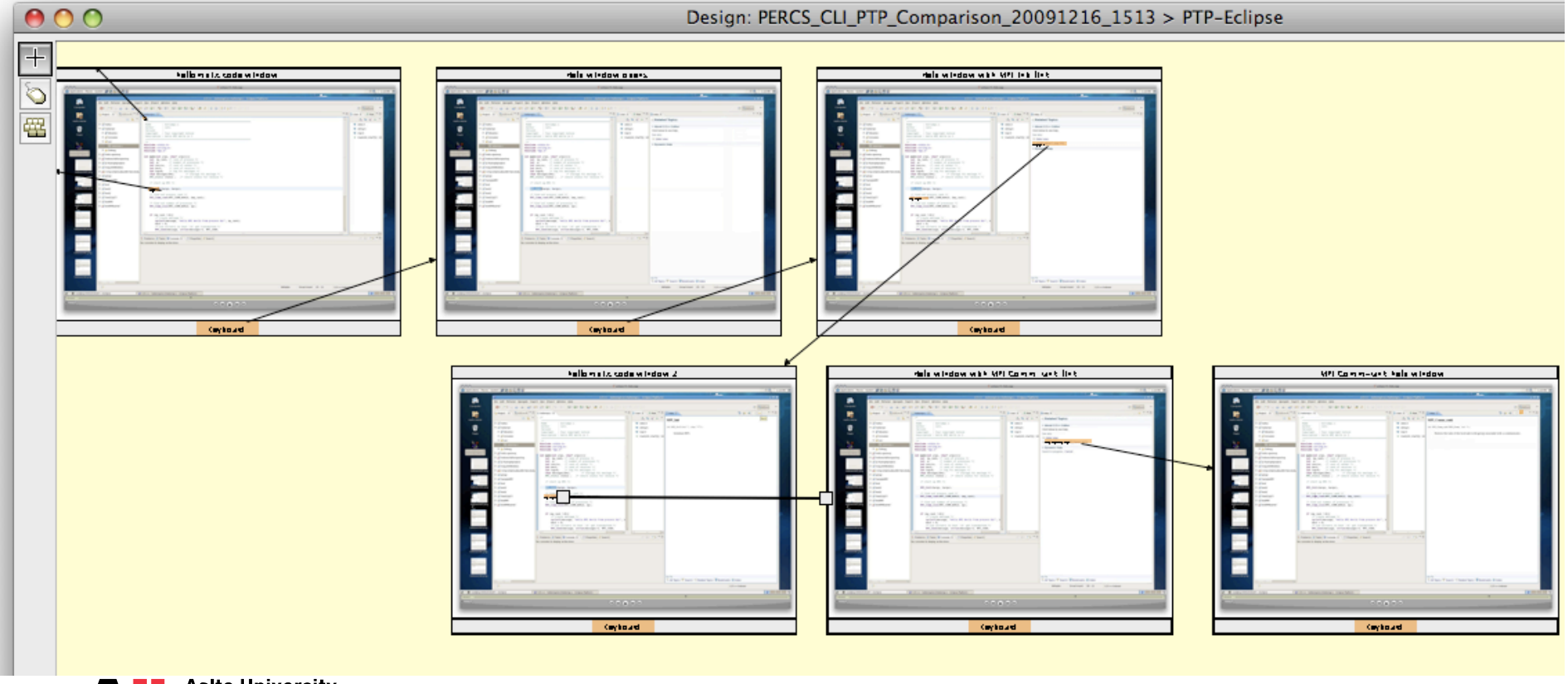
A modeling workbench for GOMS (”the godfather of KLM”)

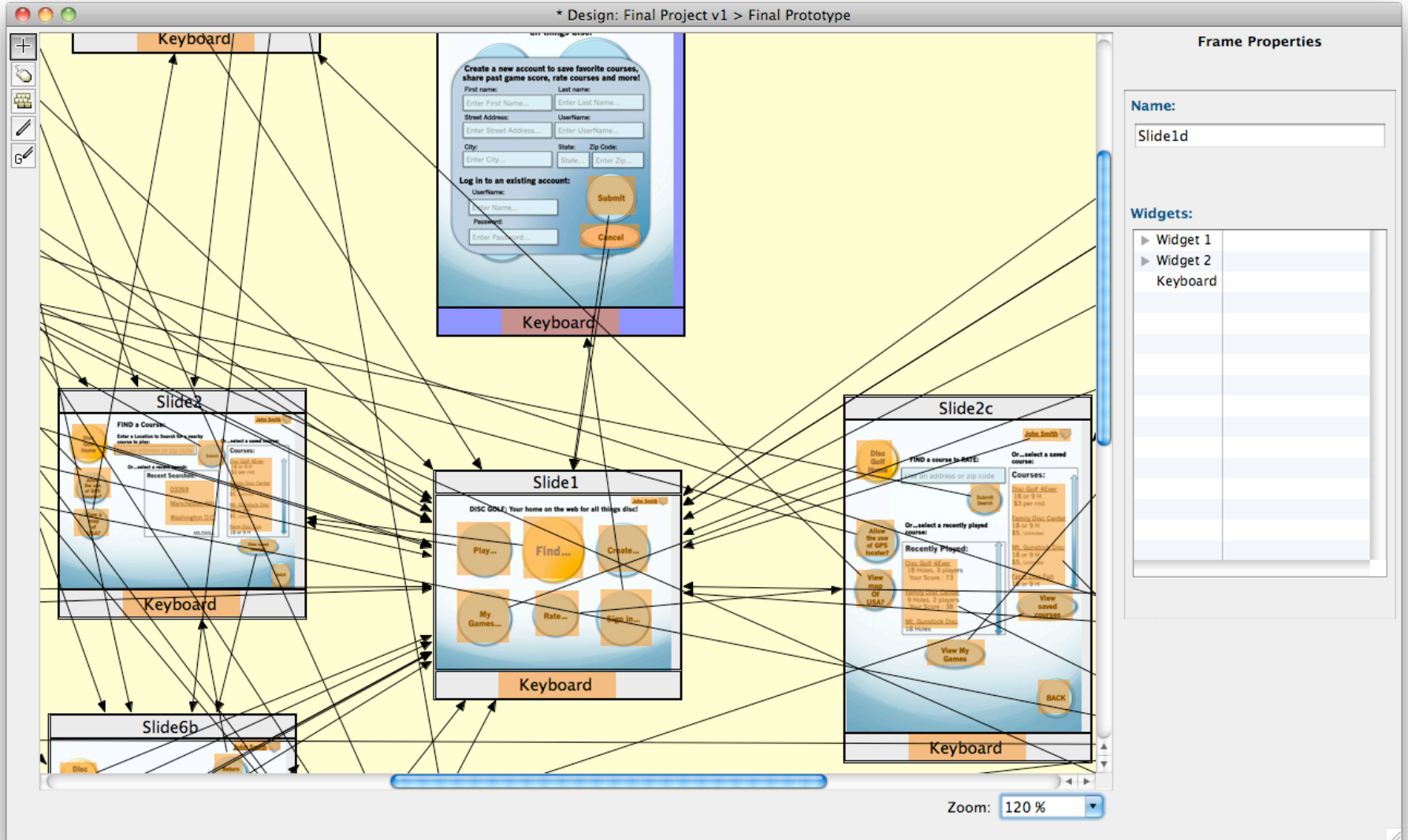
Storyboard to define UI sequences

Tasks defined by demonstration

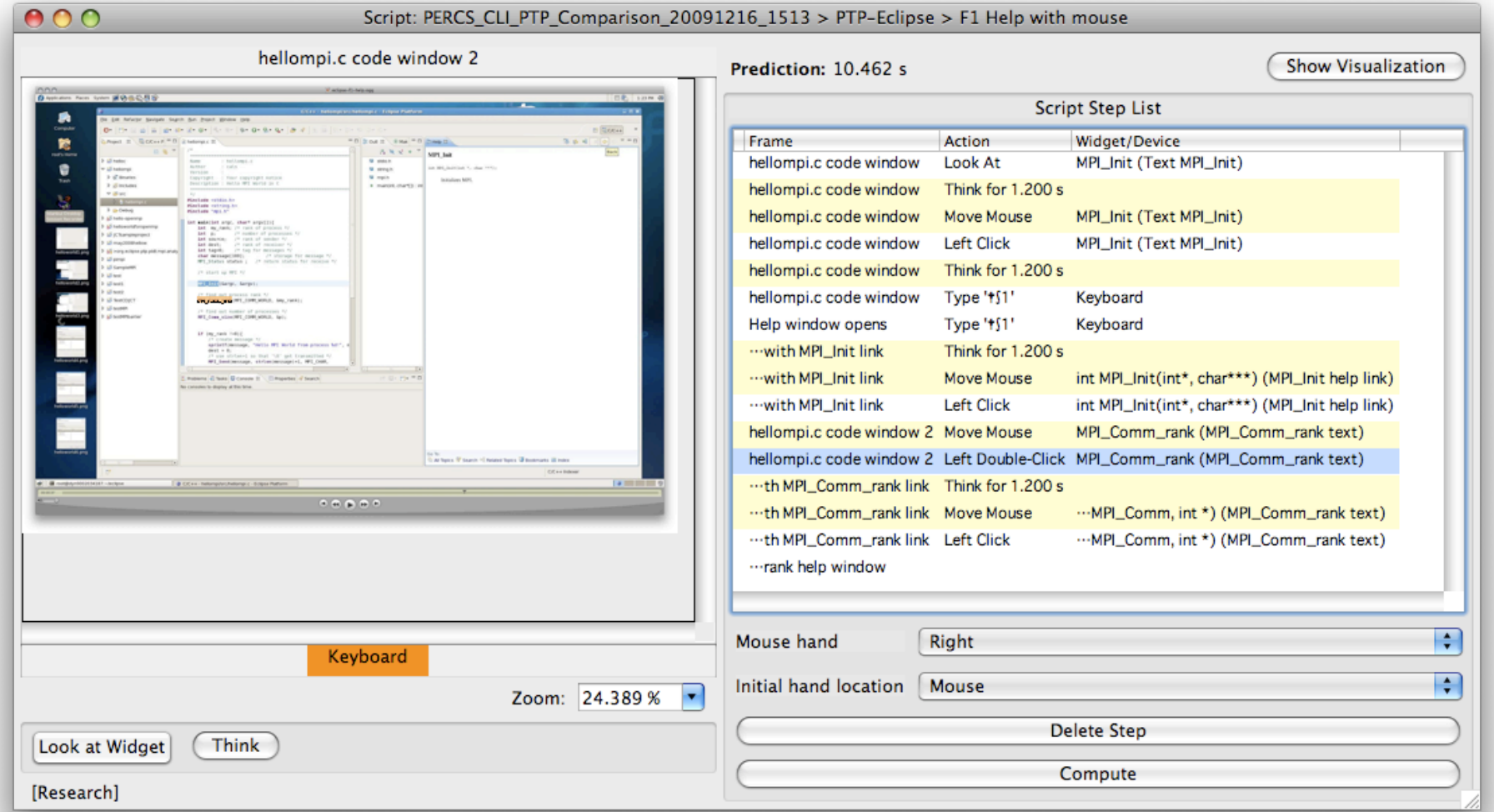
Model gives predictions for performance

A storyboard environment



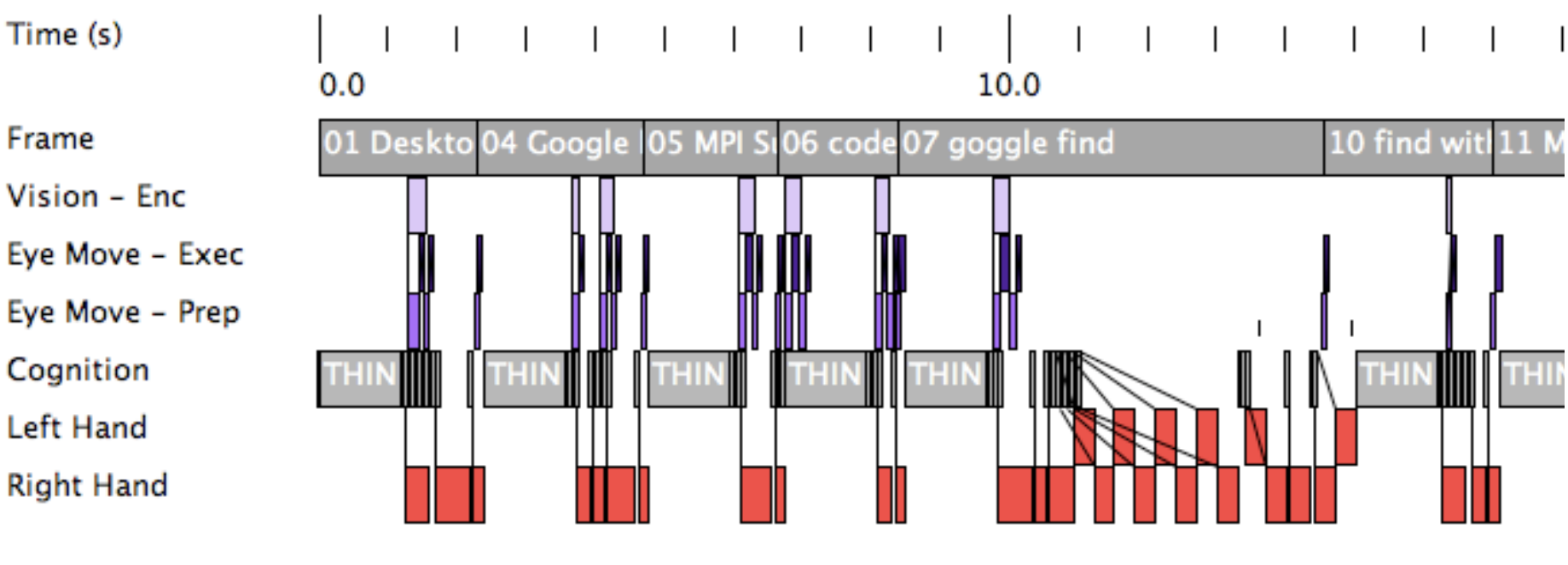

* Design: Final Project v1 > Final Prototype
Keyboard
Create a new account to save favorite courses, share past game score, rate courses and more!
First name:
Last name:
Enter First Name...
Enter Last Name...
Street Address:
UserName:
Enter Street Address...
Enter UserName...
City:
State:
Zip Code:
Enter City...
State...
Enter Zip...
Log in to an existing account:
UserName:
Password:
Submit
Cancel
Keyboard
Slide2
Keyboard
Slide1
DISC GOLF: Your home on the web for all things disc!
Play...
Find...
Create...
My Games...
Rate...
Sign In...
Keyboard
Slide2c
Keyboard
Slide6b
Frame Properties
Name:
Slide1d
Widgets:
Widget 1
Widget 2
Keyboard
Zoom:
120 %

Demonstrate a task

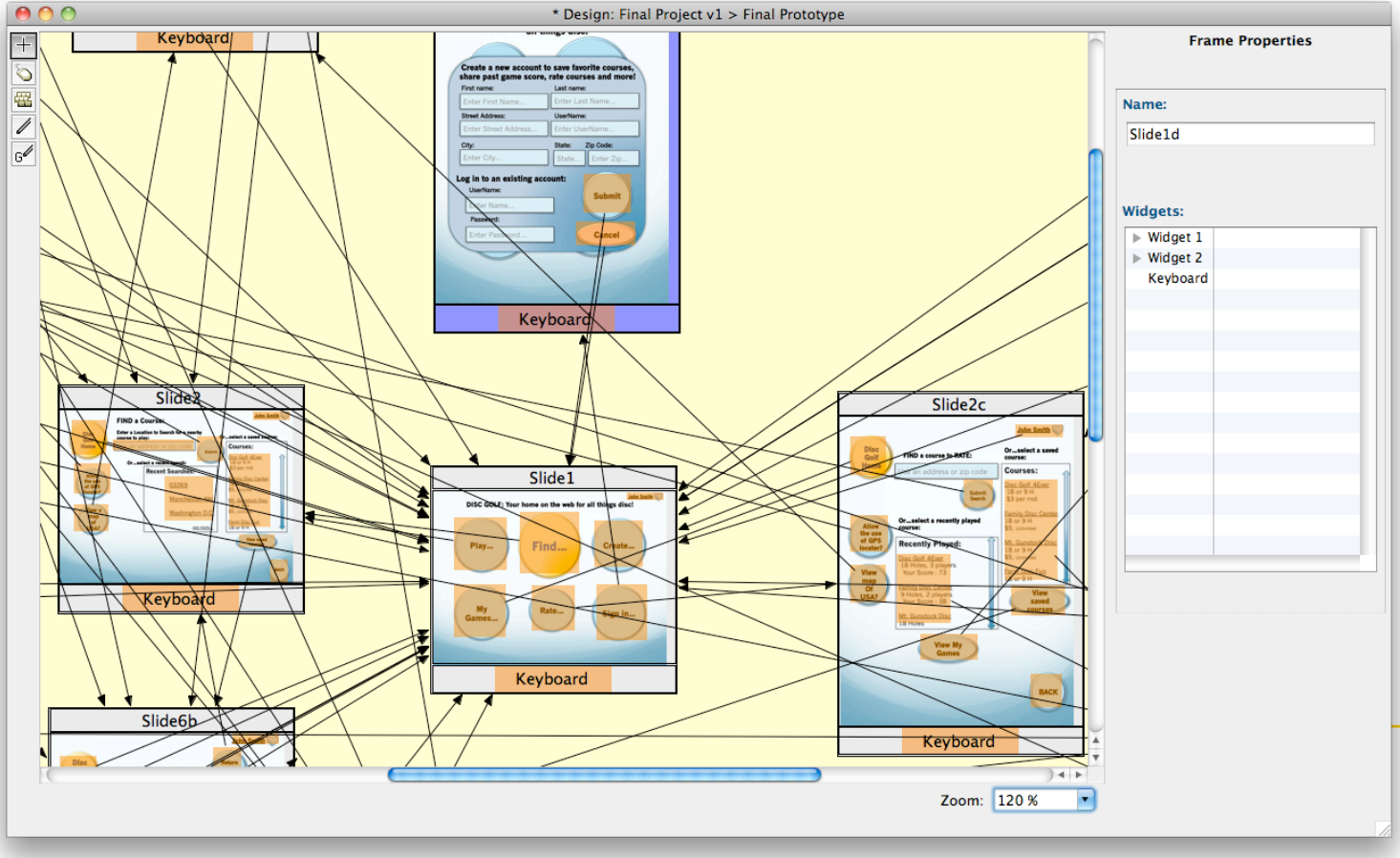



Start in top left corner
Canon law .Canon law is an ecclesiastical law or code of laws established by a church council. Canon law is usually the body of legislation of various Christian churches dealing with matters of constitution or discipline. Although all religions have regulations, the term applies mainly to the formal systems of the Roman Catholic, Orthodox and Anglican communions. It is distinguished from civil or secular law, but conflict can arise in areas of mutual concern (for example, marriage and divorce).
Encyclopedia 32 Topics
Look at link nearest to the point of visual attention
2
Evaluate the link's information scent to the goal
Decide to continue to look at and read another link
4
Click!
3
Decide to satisfice and choose the best link read so far
Plants
Musicians & Composers
Theology & Practice
U.S. States, Territories, & Regions
Economics & Business
Education
Time, Weights, & Measures
Religions & Religious Groups
Chemistry
Anthropology
Regions of the World
Countries
Music
Writers & Poets
Paleontology
Invertebrate Animals
History of the Americas
Ancient History
Religious Figures
Scripture
Artists
Cinema, TV, & Broadcasting
Political Science
People in European History
Canadian Provinces & Cities
Theater
Painting, Drawing, & Graphic Arts
Literature & Writing
Birds

Visualization of cognitive, motor, and perceptual performance



Predicting task completion time from processing constraints



CogTool

Aalto University

Assignment 8

Learning objectives today

1. Heuristic Evaluation Methods
When to choose which

2. Cognitive Walkthrough
Ability to apply to a case

Assignment 8

A8-1: Cognitive walkthrough [5p, recommended]

A8-2: STEM: A KLM modeling workbench [5p, optional]