Aalto University

ELEC-E7861

Lecture 2: Computational Modelling

2021
Antti Oulasvirta
Aalto University
userinterfaces.aalto.fi

Schedule

Jan 14: Introduction
Jan 21: Computational modeling
Jan 27: Analytical methods
Feb 4: User research
Feb 11: Literature review
Feb 18: Research strategy
Feb 25: No meeting
Mar 4: Research planning
Mar 11: Study design
Mar 18: Data analysis
Mar 25: No meeting
Apr 1: Scientific writing
April 8: No meeting
Apr 15: Scientific presentation
Independent study period
May 14: Submission of paper (PDF)
May 15: Dress rehearsal
May 16: Final presentations

Today

Sneak peek to AIM

Research problems

Validity and computational models

Assignment 2

Questions you may be asking atm

Is this the right paper for me?

- You can still change!

Can I implement that model or is it available?

- Please find out!

Is there a superior model available?

- Try Google Scholar

Can I formulate a meaningful research problem related to the model?

- We'll discuss this today

Real research projects almost never follow a clean waterfall model like this

Real projects are "messy": many feedback loops and rollbacks. Anticipated results affect early stages

Aalto University

AIM: A sneak peek

AIM:
A sneak peek

```python
class AIMMetricInterface(metaclass=abc.ABCMeta):
    # Dunder methods
    @classmethod
    def __subclasshook__(cls, subclass):
        return (
            hasattr(subclass, "execute_metric")
            and callable(subclass.execute_metric)
            or NotImplemented
        )

    # Abstract methods
    @abc.abstractmethod
    def execute_metric(
        self, gui_image: str, gui_type: int = GUI_TYPE_DESKTOP
    ) -> Optional[List[Any]]:
        """
        Execute the metric.

        Args:
            gui_image: GUI image (PNG) encoded in Base64

        Kwargs:
            gui_type: GUI type, desktop = 0 (default), mobile = 1

        Returns:
            Results (list of measures)

        Raises:
            NotImplementedError: Implementation is missing
        """
        raise NotImplementedError
```

Example

```python
class Metric1(AIMMetricInterface):
    """
    Metric 1: PNG file size.
    """

    # Public methods
    @staticmethod
    def execute_metric(
        gui_image: str, gui_type: int = GUI_TYPE_DESKTOP
    ) -> Optional[List[Any]]:
        """
        Execute the metric.

        Args:
            gui_image: GUI image (PNG) encoded in Base64

        Kwargs:
            gui_type: GUI type, desktop = 0 (default), mobile = 1

        Returns:
            Results (list of measures)
            - PNG file size in bytes (int, [0, +inf))
        """
        # Calculate PNG file size in bytes according to:
        # https://blog.aaronlenoir.com/2017/11/10/get-original-length-from-base-64-string/
        png_file_size_in_bytes: int = int(
            (3 * (len(gui_image) / 4)) - (gui_image.count("=", -2))
        )

        return [
            png_file_size_in_bytes,
        ]
```

Aalto University

Research problems

The definition I gave

"Research problem in HCI" is a stated lack of understanding about some phenomenon in human use of computing, or stated inability to construct interactive technology to address that phenomenon for desired ends.

... Let's hear what you got based on this and then revisit it...

A1: Joni Rautiainen

1 INTRODUCTION

Readability is about how readable the text is visually but also about how readable is the writing, like the structure or the complexity of the text. There are a lot of readability guidelines, which are sometimes contradicting and can be hard to apply since there are so many.

In [1], authors proposed a model which tries to automate the readability guidelines with algorithms and help from design experts. As outcome, the proposed model performs better than humans when evaluating simple guidelines, such as if the text is left-aligned and if there is enough white space. Also they found out that some guidelines are difficult to automate and they require evaluation from humans, such as evaluating if titles are meaningful and if there are too complex words.

The proposed model was not able to evaluate the readability of the used font. In this paper I research how the model could evaluate the readability of the used font.

Another possible research problem is to investigate, if the model can be used to evaluate dark theme based websites with equal performance.

A1: Aini Putkonen

1 INTRODUCTION

1.1 Visual search

Visual search refers to the act of locating target items in an environment [1]. This includes a range of tasks from trying to locate a familiar face from a crowd to fixating on a specific item on a computer screen. Several approaches aiming to explain human behaviour in visual search tasks exist. Myopic approaches, those only taking into account the reward from the immediately following action, include saliency-based (e.g. [3]) and optimal state esimation (e.g. [4]) approaches. Even though these types of, perhaps somewhat simplistic, heuristic approaches can explain certain phenomena in visual search tasks, they may fail to extend to more complex tasks involving several actions. These tasks could include, e.g. searching for specific items to buy in a webshop, which we could assume a user aims to do with as few saccades as possible. Approaching these types of scenarios through optimal control approaches (e.g. [1, 2], allowing for planning, seems more appropriate.

1.2 Research problem

My research proposal is to extend the sequential planning model proposed by Hoppe and Rothkopf [2]. Specifically, they model a probabilistic and planning observer through a Partially Observable Markov Decision Process (POMDP). However, the experimental setup they use is fairly abstract, so the performance of the model needs to be further investigated in more naturalistic settings, including those taking into account full peripheral vision. Such a naturalistic environment could be provided by a user interface. In addition, the model could be extended to consider even longer sequences of actions than currently presented, including modelling tasks where the observer fixates on a target for a given time. In the context of AIM, this topic would sit in 'Visual guidance', providing the additional computational challenge of using a POMDP.

A1: Rishabh Kapoor

1 INTRODUCTION

Visual Search modelling is really useful as it enable us to predict the usability of the interface even before it is actually tested on real users, also it helps us develop more scientific understanding of human behavior. This will help us build reliable models that can accurately predict difficulty of the visual search task.

[1] proposes a deep learning approach for the above problem, in this paper they predict human visual search time on large-scale realistic webpages. They claim that this approach can easily accommodate both structured and unstructured data which provides a good generalisation.

However, I feel there are certain things that can be improved in this model. First, I believe a deep learning approach other than CNNs should be considered. The reason being if we use CNNs the structural integrity doesn't play any role i.e. two webpages with same set of features but one with a good structure and other with a hazy one will produce similar results. I think capsule network should be leveraged for this purpose. Also, I believe testing part is also not very appropriate. The testing could be done in batches and then transfer learning approach can be utilized so as to learn across different sort of designs and then use these design to learn more advanced features.

A1: Lena Hegemann

1 INTRODUCTION

Retrieval of relevant designs is the problem of finding closely matching designs based on a query. For instance this could be finished designs based on an artifact created during the process of designing such as a sketch.

The model Swire [1] retrieves user interfaces (UI) of mobile apps based on hand-drawn sketches. For that, two deep convectional neural networks map sketches as well as UIs to an embedding space. To retrieve a UI given a sketch, a nearest neighbor search is performed to find the closest examples of UIs matching the sketch. While Swire was able to find relevant layouts for screens containing common UI elements such as sliding menus, settings or login pages. However, it failed to understand custom elements as well as colorful ones.

In this paper, I look at the problem of incorporating color in a model similar to Swire so that it can do one or several of the following

- distinguish UI elements based on color differences
- retrieve designs based of partially colorized sketches
- retrieve designs relevant with regards of underlying color palettes

Your lessons in A1

Let's do a round of lessons learned

- Hard, because topic was unfamiliar

- Not sure which research problems are feasible

- Model environment

- Unsure about the status of model code / data

Aalto University

Research problems

Properties of a good reseach problem

Contextualized

- States its motivation and objectives clearly
- Acknowledges and builds on existing work

Precise

- What is the type of knowledge that will need to be produced

Important

- The solution of which is important (for who?) and would help them (how?)

On this course: problems should relate to 1) computational modeling and 2) HCI

Putting research problems in context

How to construct a *Nature* summary paragraph

Annotated example taken from *Nature* 435, 114–118 (5 May 2005).

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion. Under these circumstances, the length of the paragraph can be up to 300 words. (This example is 190 words without the final section, and 250 words with it).

During cell division, mitotic spindles are assembled by microtubule-based motor proteins[1,2]. The bipolar organization of spindles is essential for proper segregation of chromosomes, and requires plus-end-directed homotetrameric motor proteins of the widely conserved kinesin-5 (BimC) family[3]. Hypotheses for bipolar spindle formation include the 'push–pull mitotic muscle' model, in which kinesin-5 and opposing motor proteins act between overlapping microtubules[2,4,5]. However, the precise roles of kinesin-5 during this process are unknown. Here we show that the vertebrate kinesin-5 Eg5 drives the sliding of microtubules depending on their relative orientation. We found in controlled *in vitro* assays that Eg5 has the remarkable capability of simultaneously moving at ~20 nm s^{-1} towards the plus-ends of each of the two microtubules it crosslinks. For anti-parallel microtubules, this results in relative sliding at ~40 nm s^{-1}, comparable to spindle pole separation rates *in vivo*[6]. Furthermore, we found that Eg5 can tether microtubule plus-ends, suggesting an additional microtubule-binding mode for Eg5. Our results demonstrate how members of the kinesin-5 family are likely to function in mitosis, pushing apart interpolar microtubules as well as recruiting microtubules into bundles that are subsequently polarized by relative sliding. We anticipate our assay to be a starting point for more sophisticated *in vitro* models of mitotic spindles. For example, the individual and combined action of multiple mitotic motors could be tested, including minus-end-directed motors opposing Eg5 motility. Furthermore, Eg5 inhibition is a major target of anti-cancer drug development, and a well-defined and quantitative assay for motor function will be relevant for such developments.

Common problems at early stages

The problem contains a misunderstanding

The problem is underspecified

The problem is not relevant for HCI (but e.g. ML)

The problem ignores some obvious constraints

The problem is too broad for this course

The problem is too easy (”toy problem”)

The problem is “wicked” (no solution can be found)

The problem is already solved (by someone else)

How do researchers formulate research problems?

Following what others do

Curiosity

Heuristics

Taxonomies

Empirical discoveries

Anomalies, gaps

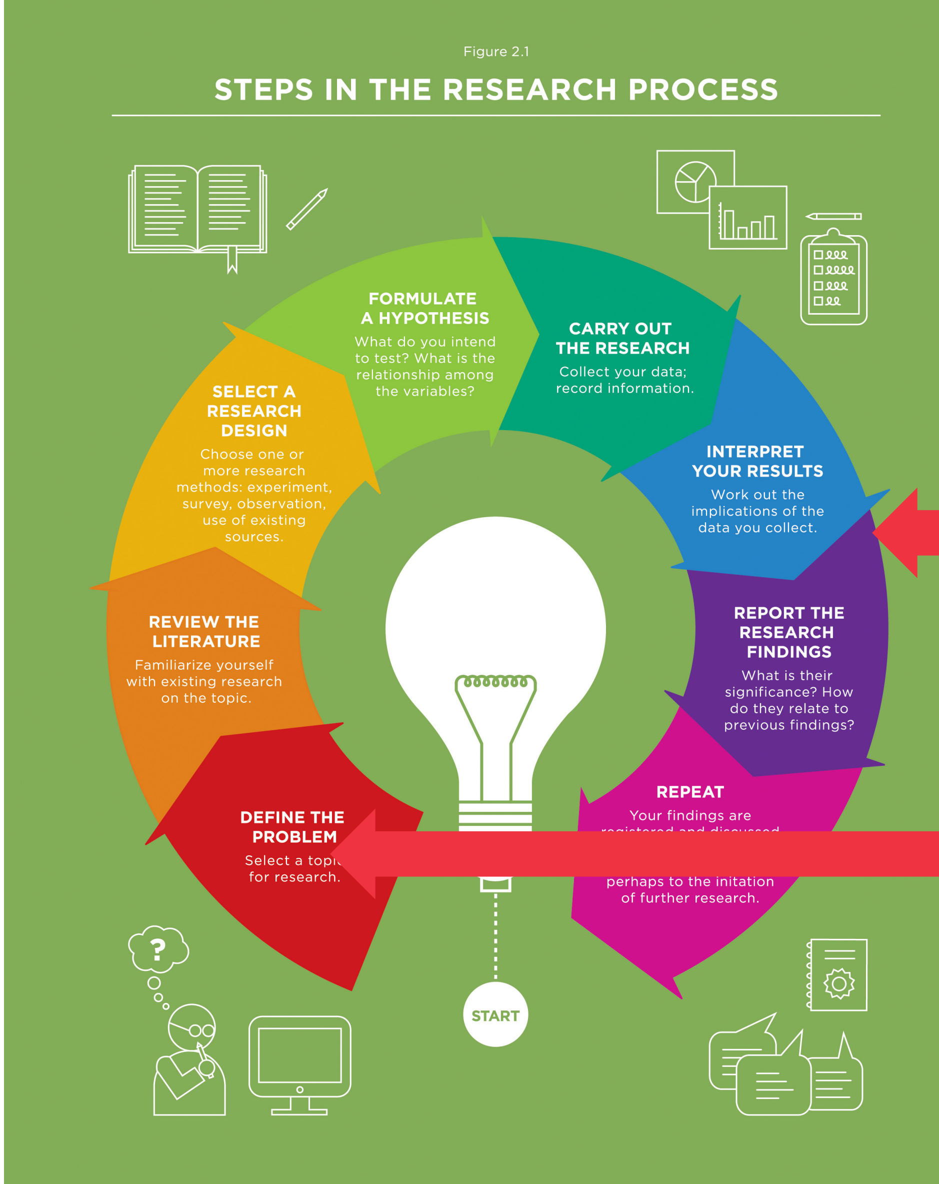
Revisiting a foundational stance

Heilmeier’s chatechism

1. What are you trying to do? Articulate your objectives using absolutely no jargon.
2. How is it done today, and what are the limits of current practice?
3. What's new in your approach and why do you think it will be successful?
4. Who cares?
5. If you're successful, what difference will it make?
6. What are the risks and the payoffs?
7. How much will it cost?
8. How long will it take?
9. What are the midterm and final "exams" to check for success?

A different view

Problem-solving perspective



Start from here...

... to know what needs to be done here

Larry Laudan's philosophy of science

A very different perspective: start from the end (from the solution not from the problem)

Key term: "Problem-solving capacity" (PSC)

- Our capability to solve important problems efficiently thanks to your research

Scientific progress = increasing PSC

“Progress and its problems”

In appraising the merits of research, it is more important to ask whether they produce **adequate solutions to significant problems** than it is to ask whether they are true, corroborated, well-confirmed or otherwise justifiable within the framework of contemporary epistemology

(Laudan, 1977, p. 14).

Qualities of the solution determine problem-solving capacity

Significance

- It solves a problem that is significant

Effectiveness

- It solves the problem effectively

Efficiency

- It solves the problem efficiently

Transfer

- It has the potential to solve many new problems

Confidence

- It can be executed with high reliability and little risk

What type of research outcome would maximize PCS in your project?

Summary: Qualities of a great research problem

1. **Relevance**
- If you solve your research problem, will it significantly help your audience apply the model?

2. **Preciseness**
- Is the problem formulated in a clear and precise way?

3. **Feasibility**
- Will you have the necessary skills, equipment, and time to solve the problem?

4. **Novelty**
- Has this problem been solved already by others?

5. **“Problem-solving capacity”**
- How will your solution increase our field’s (or your customer’s) capability to solve important problems?

Validity in the case of computational models

!

Computational models in HCI

What: Computer programs that connect three types of variables:

1. **Those describing initial conditions (inputs)**
2. **Predictions for outcomes and process (outputs)**
3. **Free parameters** (free = determined empirically)

Why: Accurate-but-practical models to inform practical decisions

Applications:

- **Decision support**
- **Evaluation**
- **Computational design**
- **Adaptation**

Engineering models try to find a pragmatic trade-off between validity and applicability

A threat to validity

= Basically anything that can go wrong and threatens your ability to draw solid conclusions

In empirical research, established taxonomies for threats

In computational modelling,

Validity of computational models

Parameter recovery, model recovery

Theoretical plausibility

Descriptive accuracy

Explanatory accuracy

Predictive accuracy, cross-validation accuracy

Errors

Anomalies

Threats to Experimental Validities ala Cook & Campbell, 1979

Statistical Conclusion Validity	**Internal Validity**	**Construct Validity of Putative Causes and Effects**	**External Validity**
Is there a relationship between the two variables?	*Given that there is a relationship, is it plausibly causal from one operational variable to another?*	*Given that the relationship is plausibly causal, what are the particular cause and effect constructs involved in the relationship?*	*Given that there is probably a causal relationship from construct A to construct B, how generalizable is this relationship across persons, settings, and times?*
1. Low Statistical Power. The lower the power of the statistical test, the lower the likelihood of capturing an effect which does in fact exist.	***1. History.*** The purported treatment effects may in fact be due to nontreatment events occurring between pre and posttesting.	***1. Inadequate Preoperational Explication of Constructs.*** A precise explication of constructs is vital for the linkage between treatments and outcomes. For example, attitudes are usually defined in terms of stable predispositions to respond. Thus a self-report scale administered on a single occasion may be an inadequate operational definition.	***1. Interaction of selection and treatment.*** People who agree to participate in a particular experiment may differ substantially from those who refuse, thus results obtained on the former may not be generalizable to the latter.
2. Violated Assumptions of Statistical Tests. The particular assumptions of a statistical test must be met if the analysis results are to be meaningfully interpreted.	***2. Maturation.*** The purported treatment effects may in fact be due to nontreatment events occurring between pre- and post-testing.	***2. Mono-Operation Bias.*** Single operational definitions of causes and/or effects (e.g., one counselor administering treatment and/or one outcome measure) both under-represent the constructs and contain irrelevancies.	***2. Interaction of Setting and Treatment.*** Results obtained in one setting may not be obtained in another (e.g., factory, military camp, university, etc.).
3. Fishing and the Error Rate Problem. The probability of making a Type I error on a particular comparison in a given experiment increases with the number of comparisons to be made in that experiment.	***3. Testing.*** Improved scores on the second administration of a test can be expected even in the absence of treatment.	***3. Mono-Method Bias.*** Multiple operational definitions of causes and/or effects may still contain irrelevancies or preclude generalization, if single methods are employed (e.g., videotaped young, male, WASP counselors administering treatment, and *self*-report devices exclusively representing outcome).	***3. Interaction of History and Treatment.*** Causal relationships obtained on a particular day (December 7, 1941 as an extreme example) may not hold up under more mundane circumstances.
4. The Reliability of Measures. Measures of low reliability may not register true changes.	***4. Instrumentation.*** Changes in the calibration of the measuring instrument over time or changes in personnel making ratings may result in spurious criterion differences that masquerade as treatment effects.	***4. Hypothesis Guessing within Experimental Conditions.*** If subjects are aware of the hypotheses, the effects of a treatment may be confounded with the subject's desire to conform to the hypotheses.	
5. The Reliability of Treatment Implementation. When treatments are not administered in a standard fashion (e.g., different administrators and/or the same administrator behaving differently on different occasions) error variance will increase and the chance of obtaining true differences will decrease.	***5. Statistical Regression.*** Individuals selected on the basis of extreme scores, high or low, on a particular test will regress toward the mean on a second test administration. Thus a group of low-scoring individuals will "improve" without treatment. Conversely, high-scoring individuals might deteriorate in spite of it.	***5. Evaluation Apprehension.*** Apprehension about being evaluated may result in attempts by respondents to depict themselves as more competent or psychologically healthy than is in fact the case.	
6. Random Irrelevancies in the	***6. Selection.*** Unless experimental and	***6. Experimenter Expectancies.*** The data in	

Validity of experimental research
Cook & Campbell 1979

Modeling workflow

Constructing a model

Generate a model (done in our case)

Test model-generated data against existing findings

Test parameter recovery

Validating a model

Assess parameter inference

Validate against one-to-one data from human subjecs

Applying a model

Integrate to a practioner tool or computational design algorithm

Test with practitioners or end users

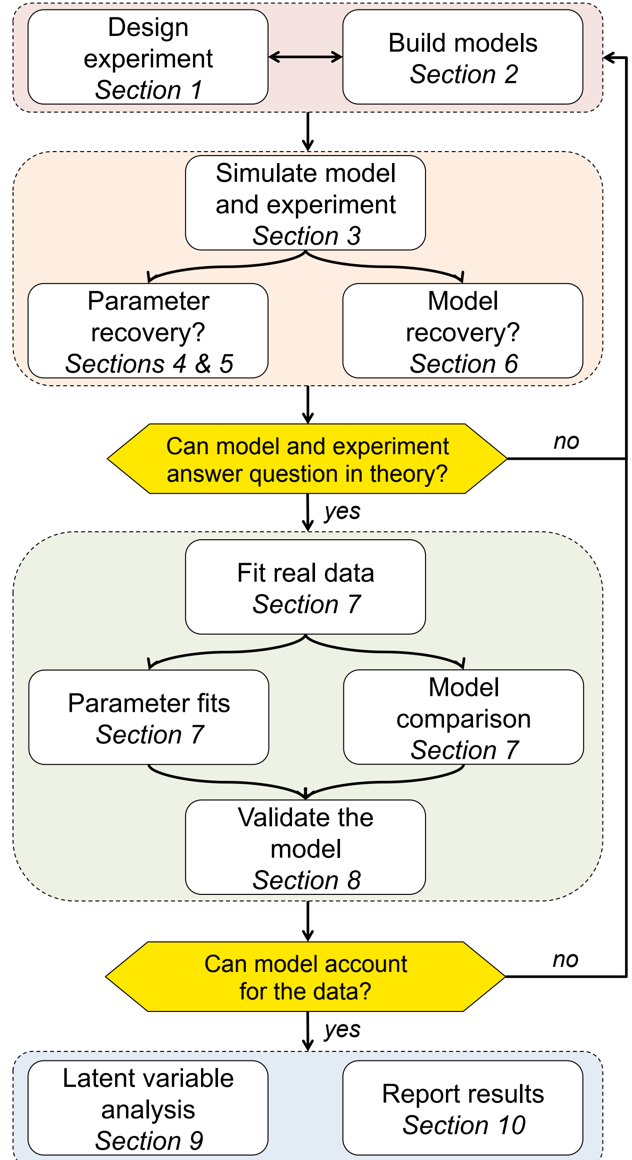
10 step workflow

Wilson & Collins 2019

(PDF to be added to MyCourses)

Takeaways:

- Model validity is tested “in silico” prior to validation against human data
- Modeling is iterative: several modelling ideas are tried out



In-class exercise

!

Paired exercise in Zoom (6 mins)

You: Introduce your paper brieflyi to your pair

Together: Brainstorm potential validity issues

Aalto University

Assignment 2

Assignment 2

1. Revise your research problem

- **Rethink and redefine your research problem *using terminology* discussed in this lecture**
- **Changes are fine and even expected!**

2. Write title + abstract → Discussed next week

3. The topic of next week's meeting is "analytical methods". I will send a method for each of you to try out (something lightweight!)

A2 will be released by EOD