Interaction with Machine Learning

ACS P230 / Part II unit / CDH - Alan Blackwell & Advait Sarkar

Overview

- Practical experimental course
  - lectures provide overview and sample of current research
- This introduction
  - general principles, research approaches, current trends
- Specialist lectures:
  - six specialist topics
- Design and run your own study
  - discussion and feedback each week
- Final presentation of your results

Course objective

- "Human-Centered AI"
  Ben Shneiderman
  (OUP 13 Jan 2022)

1) Process: HCAI builds on user experience design methods of user observation, stakeholder engagement, usability testing, iterative refinement, and continuing evaluation of human performance in use of systems that employ AI and machine learning.
2) Product: HCAI systems are designed to be supertools which amplify, augment, empower, and enhance human performance. They emphasize human control, while embedding high levels of automation by way of AI and machine learning. Examples include digital cameras and navigation systems, which give humans control yet have many automated features.

But with a technical focus:

- Four waves of AI according to Google DeepMind founder Demis Hassabis:
  - First wave (GOFAI): Expert systems & symbolic reasoning
  - Second wave: Statistical inference
  - Third wave: Deep learning
  - Fourth wave: Intelligent tools
- Our approach:
  - Intelligent tools as advanced HCI
  - Including: Visualisation, Programming, Labelling, Explanation
- A *practical* HCI course:
  - Project work to build, customise, measure, observe …
- For: Part III and MPhil ACS (research preparation), Part II (advanced HCI), Digital Humanities researchers (guided methods programme)

Your background

- 1. Prior HCI experience
- 2. Prior ML/AI experience
- 3. What do you hope to get out of this course?

	None	Casual	Student	Professional
HCI				
ML				

Target outcome

- This is a specialised and focused practical research training course.
- The expected outcome:
  - You will achieve research competence in a recognised academic field such as Intelligent User Interfaces, Interactive Intelligent Systems etc
- ACS assessment will be relative to the international standard of graduate students working in these fields.
  - Written work will be graded relative to typical student publications in the field
  - Presentations will be expected to meet the standard of first-year PhD students in the field, for example at the Doctoral Consortium of a specialised conference.
- Part II and Digital Humanities students will be briefed separately

Lecture topics

- Week 2 - Labelling (AS)
  - attribution, subjectivity, reliability, consistency
- Week 3 - Visual analytics (AS)
  - visualisation, tool chains, design case studies
- Week 4 - Mixed initiative interaction (AB)
  - information gain, cognitive ergonomics, agency & control
- Week 5 - Program synthesis (AB)
  - end-user programming, attention investment
- Week 6 - Explainability (Simone Stumpf, Glasgow)
- Week 7 – Bias and fairness (AB)
  - discrimination, accountability and ethics in hybrid systems
- Week 8 – Your research presentations

Practical work plan

- Week 1 - select research question
- Week 2 - discuss potential study approaches
- Week 3 - review and feedback on study proposals
- Week 4 & 5 - review logistical issues / practical progress
- Week 6 - discuss preliminary findings
- Week 7 - discuss research implications
- Week 8 - final presentation

Assessment for ACS

- Final research report (80%)
  - Based on your practical work
  - Presented as a research paper
- Optional (but recommended) work-in-progress drafts
  - Advisory grades will be provided as feedback, for revision in final report
- Reflective diary (20%)
    - Summarise lectures
    - Document discussions
    - Record development of your own thinking
    - Make 8 weekly entries …
    - … plus a final summative review

Marking criteria

- Standard ACS criteria for final grades
  - 90-100% - Original contribution
  - 80-89% - Significant insight or creativity
  - 75-79% - Demonstrates critical thought
  - 70-74% - Execution basically good
  - 60-69% - Adequate presentation
  - 50-59% - Some serious flaws
  - 40-49% - Work is poor
- Indicative feedback for work in progress
  - A+ excellent - on target for 85-100
  - A very good - on target for 75-85
  - B good - on target for 70-80
  - C acceptable - on target for 60-70
  - D disappointing - risk of fail

Continuous feedback opportunities

- Week 2 - Research question (200 words) + a sample diary entry
- Week 3 - Study design (400 words)
- Week 4 - Another sample diary entry
- Week 5 - Draft literature review for final report (400 words)
- Week 6 - Draft introduction to report (200 words)
- Week 7 - Draft results section for report (400 words)
- Week 8 - Draft discussion section for report (200 words)

Reading suggestions

- Refresh knowledge of undergraduate HCI
  - Cambridge notes online
  - Preece, Rogers and Sharp Interaction Design beyond HCI
- Review Cambridge guidance on human participants
  - https://www.tech.cam.ac.uk/research-ethics/school-technology-research-ethics-guidance
- Cairns and Cox (2008)
  - Research Methods for Human-Computer Interaction
- Carroll (2003)
  - HCI Models, Theories and Frameworks: Toward a multidisciplinary science
- Shneiderman (2022)
  - Human-Centered AI
- **Mostly: Recent research literature**

A note about the reading list

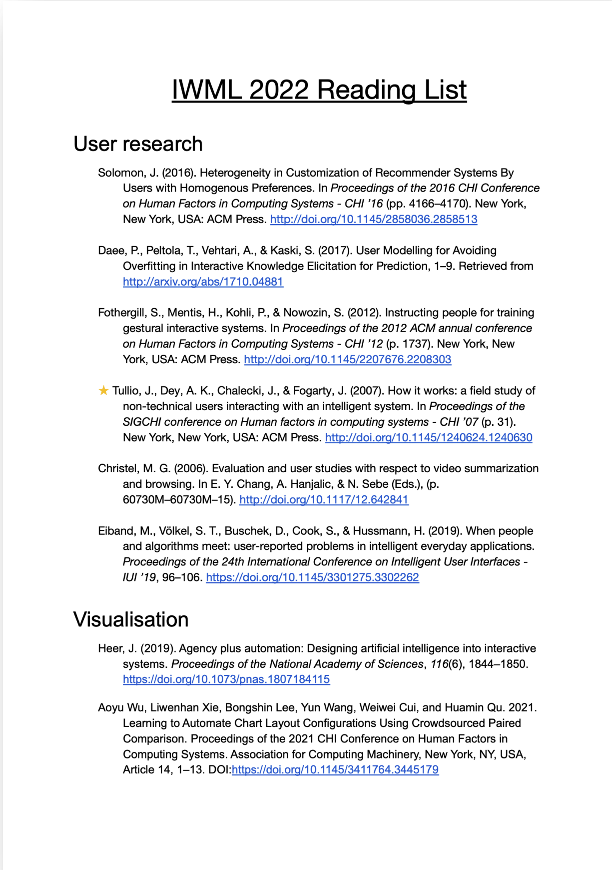
Available on course materials page.

Don’t try to read all of it!

“Starred” entries are particularly good for one or more of the following reasons:
- Influential
- Well-executed research
- Interesting/unique angle

Read at least the abstracts of all of the starred entries.

Use as a basis for your own research question/study design.



IWML 2022 Reading List

User research

Solomon, J. (2016). Heterogeneity in Customization of Recommender Systems By Users with Homogenous Preferences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI ’16* (pp. 4166–4170). New York, New York, USA: ACM Press. http://doi.org/10.1145/2858036.2858513

Daee, P., Peltola, T., Vehtari, A., & Kaski, S. (2017). User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction, 1–9. Retrieved from http://arxiv.org/abs/1710.04881

Fothergill, S., Mentis, H., Kohli, P., & Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI ’12* (p. 1737). New York, New York, USA: ACM Press. http://doi.org/10.1145/2207676.2208303

★ Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI ’07* (p. 31). New York, New York, USA: ACM Press. http://doi.org/10.1145/1240624.1240630

Christel, M. G. (2006). Evaluation and user studies with respect to video summarization and browsing. In E. Y. Chang, A. Hanjalic, & N. Sebe (Eds.), (p. 60730M–60730M–15). http://doi.org/10.1117/12.642841

Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019). When people and algorithms meet: user-reported problems in intelligent everyday applications. *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI ’19*, 96–106. https://doi.org/10.1145/3301275.3302262

Visualisation

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, *116*(6), 1844–1850. https://doi.org/10.1073/pnas.1807184115

Aoyu Wu, Liwenhan Xie, Bongshin Lee, Yun Wang, Weiwei Cui, and Huamin Qu. 2021. Learning to Automate Chart Layout Configurations Using Crowdsourced Paired Comparison. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 14, 1–13. DOI:https://doi.org/10.1145/3411764.3445179

Theories of interaction

Human-Computer Interaction (HCI) - Three waves

- First wave (1980s):
  - Theory from Human Factors, Ergonomics and Cognitive Science
- Second wave (1990s):
  - Theory from Anthropology, Sociology and Work Psychology
- Third wave (2000s):
  - Theory from Art, Philosophy and Design

First wave: HCI as engineering “human factors” (1980s)

- The “user interface” (or MMI “man-machine interface”) was considered to be a separate module, designed independently of the main system.
- Design goal was efficiency (speed and accuracy) for a human operator to achieve well-defined functions.
- Use methods from cognitive science to model the user’s perception, decision and action processes, and predict usability on the basis of that model
  - At this point, relatively closely aligned with AI

Second wave: HCI as social system (1990s)

- AI models did not result in more usable machines (see esp. Lucy Suchman)
  - Resulted in a significant intellectual challenge to cognitive science and AI!
- The design of complex systems is a socio-technical experiment
  - Took account of other information factors including conversations, paper, and physical settings
- Study the context where people work
  - Used ethnography (or "Contextual Inquiry" or "Workplace Studies") to understand other ways of seeing the world and characterise social structures
- Other stakeholders are integrated into the design process
  - Prototyping and participatory workshops aim to empower users and acknowledge other value systems

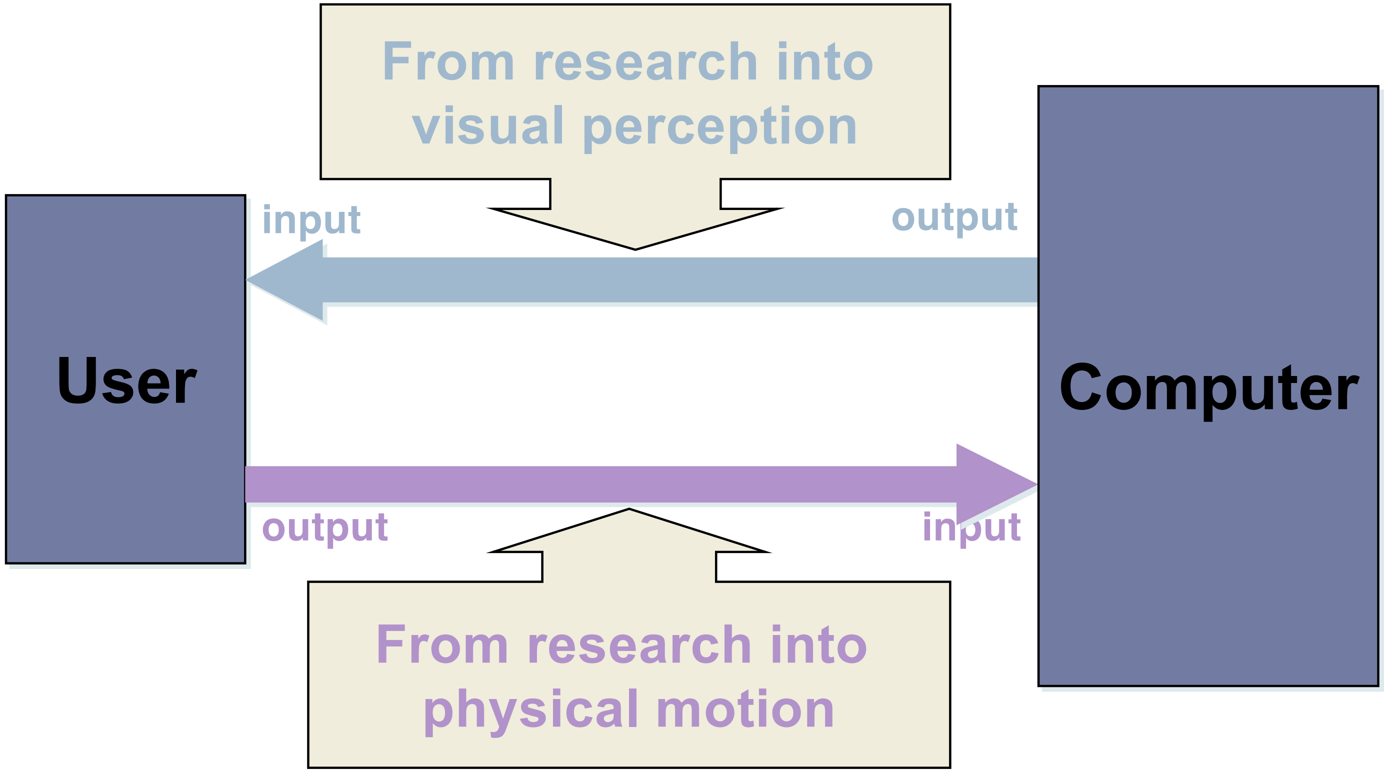
Third wave: HCI as culture and experience (2000s)

- Ubiquitous computing affects every part of our lives
  - It mixes public (offices, lectures) and private (bedrooms, bathrooms)
- Outside the workplace, efficiency is not a priority
  - Usage is discretionary
  - User Experience (UX), includes aesthetics, affect,
- Design experiments are speculative and interpretive
  - Critical assessment of how this is meaningful
- Was until 2018 pretty much completely divorced from AI
  - But this is changing very rapidly, as critical AI studies mature!

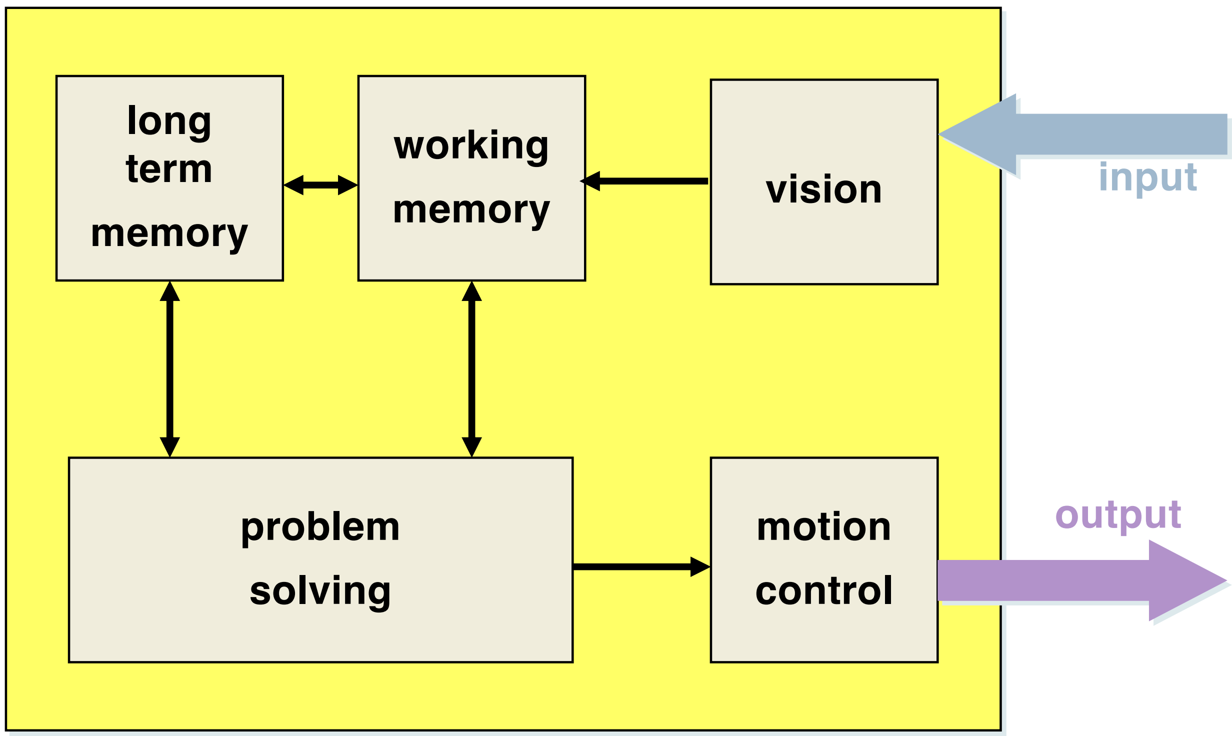
Summary of Cambridge HCI content

- Textbooks
  - Preece, Sharp & Rogers
  - Carroll
- Part 1a Interaction Design
  - Requirements analysis and design process, data collection (observation, interviews, focus groups) and analysis. Design and prototyping, personas, storyboards and task models. Principles of good design. Human cognition. Usability evaluation.
- Part 1b Further HCI
  - Theory driven approaches. Design of visual displays. Goal-oriented interaction. Designing smart systems. Designing efficient systems. Designing meaningful systems. Evaluating interactive system designs. Designing complex systems.
- Part 2/3
  - Affective Computing, Computer Music, Advanced Graphics …

Classical cognitive science models of first-wave HCI



Classical cognitive science model of the user ('boxology')



Engineering models of human I/O, memory, CPU

- Seeks "impedance match" of computer with computational user model
  - Extend principles of human factors and ergonomics
  - Psychophysical perception
  - Speed and accuracy of movement at keystroke level
  - Measure reaction time (and infer decision time?)
  - Include working memory capacity
    - 7 +/- 2 'chunks'
    - Single visual scene
  - GOFAI-planner style Goals Operators Methods Selection
- Is intelligent task design a matter of 'cognitive ergonomics'?

The problem of learning (Clayton Lewis, Jack Carroll, Mary Beth Rosson …)

- Classical models assumed users would be *made* to read the manual
- In contrast, *discretionary usage* systems require exploratory learning models because users can (and do) walk away
  - Focus on minimal instruction, immediate progress toward user goals
  - Now taken for granted (but only after long battle with usability advocates)
- Cognitive walkthrough review methods allowed system designers to anticipate usability problems, based on model of situated learning rather than cognitive model of planning

The sticky problem of viscosity (Thomas Green)

- Deciding what to do is often harder than doing it
  - But HCI models assume a ‘correct’ sequence of actions
- How do you change your mind if something goes wrong?
  - problem solving
  - planning
  - knowledge representation
- External representations are often required
  - But did the designers anticipate people making mistakes?
- Many systems and visual representations make it hard to change your mind, or to engage in exploratory design
  - Complex systems can be regarded as interaction spaces

Wicked problems (Rittel & Webber)

- Formulated in reaction to promotion of AI/cybernetic methods (e.g. optimization, goal-directed search) in business schools and public policy
- Wicked problems have:
  - no definitive formulation
  - no stopping rule
  - no true-or-false outcome: only good-or-bad
  - no ultimate test of a solution
  - no set of permissible operations
  - essentially unique

The scope of IWML research

Established paradigms of interacting with ML

- Perfect information games (toy worlds, chess, go, videogames)
  - Not considered particularly interesting
- Recommender systems
  - Once a major research area, now familiar - Amazon, Spotify, YouTube, Netflix, etc.
- Dialogue models: diagnostics, FAQ retrieval, interactive query refinement
  - An early example was "metaFAQ" from Cambridge company Transversal
  - But also familiar – consider usage of Google results, autocomplete, image search
  - Voice assistants
- Programming by example, program synthesis
  - See Lieberman *Watch What I Do*, but also e.g. Microsoft Excel FlashFill
  - Advances in code generation: codex, github copilot
- Human-in-the-loop automation
  - Autopilots, remote-operation, "autonomous" vehicles
- Generative AI as a creative assistant
  - Art, creative writing, music
  - 'Filters' in social media

Topics at 2021 Intelligent User Interfaces (IUI) conference

- Human-centred AI methods and approaches
  - e.g., explainability, persuasive technologies, privacy and security, knowledge-based approaches to user interface design, user modelling, personalization, crowd computing
- Computational innovation
  - e.g., machine learning methods, human-in-the-loop machine learning
- Interface modalities
  - e.g., affective and aesthetic interfaces, collaborative interfaces, speech-based interfaces, AR/VR, wearable and mobile interfaces, ubiquitous smart environments.
  - e.g., embodied agents, virtual assistants, multi-modal interfaces, conversational interfaces, tangible interfaces, intelligent visualization.
- Evaluations
  - e.g., user experiments and studies, reproducibility (including benchmarks, datasets, and challenges), meta-analysis, mixed-methods evaluations.
- Application areas
  - e.g., education, health, assistive technologies, social media and other Web technologies, mobile applications, intelligent assistants, conversational agents, Information retrieval, search, and recommendation system, internet of things (IoT).

Top cited papers in ACM TIIS (Trans. Intelligent Interactive Systems)

RESEARCH-ARTICLE
The MovieLens Datasets: History and Context
F. Maxwell Harper, Joseph A. Konstan
ACM Transactions on Interactive Intelligent Systems, Volume 5, Issue 4 • January 2016, Article No.: 19, pp 1–19 • https://doi.org/10.1145/2827872
The MovieLens datasets are widely used in education, research, and industry. They are downloaded hundreds of thousands of times each year, reflecting their use in popular press programming books, traditional and online courses, and software. These ...
677 4,679 Get Access

Dataset

Evaluation methods; recommender systems; experiments

RESEARCH-ARTICLE
Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems
Marius Kaminskas, Derek Bridge
ACM Transactions on Interactive Intelligent Systems (TiiS), Volume 7, Issue 1 • March 2017, Article No.: 2, pp 1–42 • https://doi.org/10.1145/2926720
What makes a good recommendation or good list of recommendations?
Research into recommender systems has traditionally focused on accuracy, in particular how closely the recommender's predicted ratings are to the users' true ratings. However, it has been ...
68 2,180 Get Access

RESEARCH-ARTICLE
Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying
Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, Rosalind Picard
ACM Transactions on Interactive Intelligent Systems, Volume 2, Issue 3 • September 2012, Article No.: 18, pp 1–30 • https://doi.org/10.1145/2362394.2362400
Cyberbullying (harassment on social networks) is widely recognized as a serious social problem, especially for adolescents. It is as much a threat to the viability of online social networks for youth today as spam once was to email in the early days of ...
145 2,544 Get Access

Social intervention; user modelling

Gestural interaction; ML model

RESEARCH-ARTICLE
Continuous body and hand gesture recognition for natural human-computer interaction
Yale Song, David Demirdjian, Randall Davis
ACM Transactions on Interactive Intelligent Systems, Volume 2, Issue 1 • March 2012, Article No.: 5, pp 1–28 • https://doi.org/10.1145/2133366.2133371
Intelligent gesture recognition systems open a new era of natural human-computer interaction: Gesturing is instinctive and a skill we all have, so it requires little or no thought, leaving the focus on the task itself, as it should be, not on the ...
62 2,730 Get Access

RESEARCH-ARTICLE
AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back
Sidney D'mello, Art Graesser
ACM Transactions on Interactive Intelligent Systems, Volume 2, Issue 4 • December 2012, Article No.: 23, pp 1–39 • https://doi.org/10.1145/2395123.2395128
We present AutoTutor and Affective AutoTutor as examples of innovative 21st century interactive intelligent systems that promote learning and engagement. AutoTutor is an intelligent tutoring system that helps students compose explanations of difficult ...
106 1,585 Get Access

Affective computing; education; controlled experiments

User modelling; recommender systems; human-in-the-loop learning

RESEARCH-ARTICLE
OPEN ACCESS
Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback
Gawesh Jawaheer, Peter Weller, Patty Kostkova
ACM Transactions on Interactive Intelligent Systems, Volume 4, Issue 2 • July 2014, Article No.: 8, pp 1–26 • https://doi.org/10.1145/2512208
Recommender systems are firmly established as a standard technology for assisting users with their choices; however, little attention has been paid to the application of the user model in recommender systems, particularly the variability and noise that ...
57 3,724

RESEARCH-ARTICLE
Adaptive Persuasive Systems: A Study of Tailored Persuasive Text Messages to Reduce Snacking
Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, Emile Aarts
ACM Transactions on Interactive Intelligent Systems (TiiS), Volume 2, Issue 2 • June 2012, Article No.: 10, pp 1–25 • https://doi.org/10.1145/2209310.2209313
This article describes the use of personalized short text messages (SMS) to reduce snacking. First, we describe the development and validation ($N = 215$) of a questionnaire to measure individual susceptibility to different social influence strategies. To ...
98 1,632 Get Access

Conversational interaction; questionnaire; experiments

Trust/transparency; recommender systems; experiment

RESEARCH-ARTICLE
Making Decisions about Privacy: Information Disclosure in Context-Aware Recommender Systems
Bart P. Knijnenburg, Alfred Kobsa
ACM Transactions on Interactive Intelligent Systems, Volume 3, Issue 3 • October 2013, Article No.: 20, pp 1–23 • https://doi.org/10.1145/2499670
Recommender systems increasingly use contextual and demographical data as a basis for recommendations. Users, however, often feel uncomfortable providing such information. In a privacy-minded design of recommenders, users are free to decide for ...
56 1,563 1 Get Access

RESEARCH-ARTICLE
Conversational gaze mechanisms for humanlike robots
Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, Hiroshi Ishiguro
ACM Transactions on Interactive Intelligent Systems (TiiS), Volume 1, Issue 2 • January 2012, Article No.: 12, pp 1–33 • https://doi.org/10.1145/2070719.2070725
During conversations, speakers employ a number of verbal and nonverbal mechanisms to establish who participates in the conversation, when, and in what capacity. Gaze cues and mechanisms are particularly instrumental in establishing the participant roles ...
81 1,117 Get Access

Gaze interaction; conversational interaction
Human-robot interaction; controlled experiments

Trust/transparency; recommender systems; experiment

RESEARCH-ARTICLE
Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: An Empirical Study
Paolo Cremonesi, Franca Garzotto, Roberto Turrin
ACM Transactions on Interactive Intelligent Systems (TiiS), Volume 2, Issue 2 • June 2012, Article No.: 11, pp 1–41 • https://doi.org/10.1145/2209310.2209314
Recommender Systems (RSs) help users search large amounts of digital contents and services by allowing them to identify the items that are likely to be more attractive or useful. RSs play an important persuasion role, as they can potentially augment the ...
54 1,241 Get Access

TIIS Special Issues

September 2012, Issue 3

Special Issue on Common Sense for...

Special Issue on Common Sense for Interactive Systems

January 2015, Issue 4

Special Issue on Activity Recognition...

Special Issue on Activity Recognition for Interaction and Regular Article

November 2020, Issue 3

Special Issue on Data-Driven...

Special Issue on Data-Driven Personality Modeling for Intelligent Human-Computer Interaction

March 2018, Issue 1

Special Issue on Interactive Visual...

Special Issue on Interactive Visual Analysis of Human Crowd Behaviors and Regular Paper

January 2016, Issue 4

Regular Articles and Special issue on Ne...

Regular Articles and Special issue on New Directions in Eye Gaze for Interactive Intelligent Systems (Part 1 of 2)

April 2014, Issue 1

Special Issue on Interactive...

Special Issue on Interactive Computational Visual Analytics

July 2013, Issue 2

Special issue on interaction with...

Special issue on interaction with smart objects, Special section on eye gaze and conversation

December 2016, Issue 4

Special Issue on Human Interaction...

Special Issue on Human Interaction with Artificial Advice Givers

March 2012, Issue 1

Special Issue on Affective Interactio...

Special Issue on Affective Interaction in Natural Environments

July 2015, Issue 2

Special Issue on Behavior...

Special Issue on Behavior Understanding for Arts and Entertainment (Part 1 of 2)

October 2014, Issue 3

Special Issue on Multiple Modalities ...

Special Issue on Multiple Modalities in Interactive Systems and Robots

April 2013, Issue 1

Special section on internet-scale hum...

Special section on internet-scale human problem solving and regular papers

Recent papers at CHI (and elsewhere)

- Useful overview papers:
  - Dudley, J. J., & Kristensson, P. O. (2018). A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems*, *8*(2), 1–37. https://doi.org/10.1145/3185517
  - Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–18. https://doi.org/10.1145/3173574.3174156

- Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 95, 1–9. https://doi.org/10.1145/3411764.3445740

- Minhyang (Mia) Suh, Emily Youngblom, Michael Terry, and Carrie J Cai. 2021. AI as Social Glue: Uncovering the Roles of Deep Generative AI during Social Music Composition. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 582, 1–11. https://doi.org/10.1145/3411764.3445219

- Eiband, M., Völkel, S. T., Buschek, D., Cook, S., & Hussmann, H. (2019). When people and algorithms meet: User-reported Problems in Intelligent Everyday Applications. *Proceedings of the 24th International Conference on Intelligent User Interfaces, Part F1476*, 96–106. https://doi.org/10.1145/3301275.3302262

- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, *81*, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

- Yang, Q., Suh, J., Chen, N.-C., & Ramos, G. (2018). Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*, 573–584. https://doi.org/10.1145/3196709.3196729

Research methods

Ethical Issues in Research

- Review the Cambridge Technology Ethics guide
  - What kind of study are you planning?
  - What potential concerns might there be?
  - What will you do to address them?
- Submit a proposal to the Computer Science Ethics committee, giving above details.
  - https://dbwebserver.cl.cam.ac.uk/Administration/Ethics/EthicsRequest.aspx
  - (accessible from department VPN, using department login not Raven)

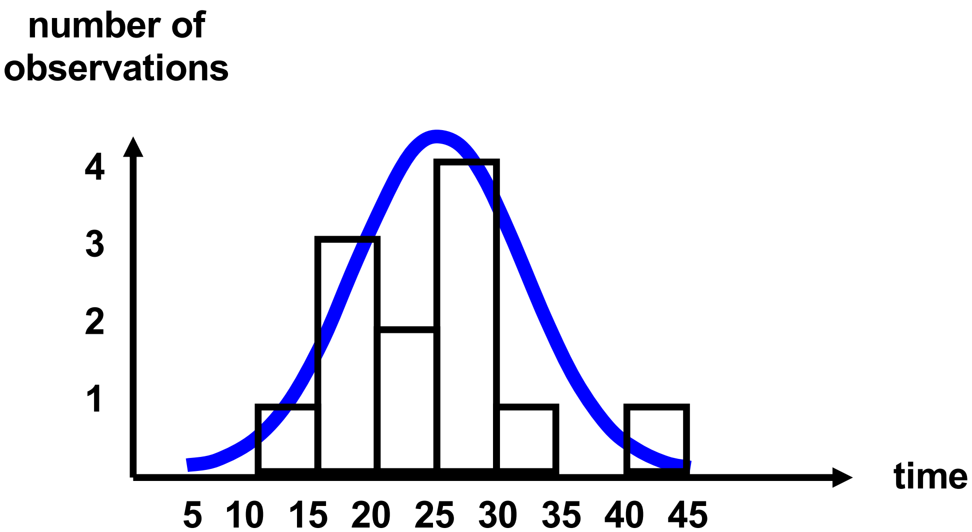
Controlled Experimental Methods

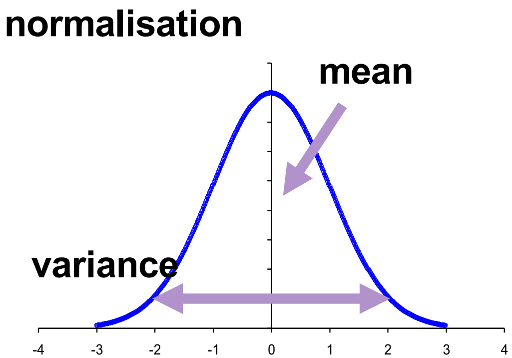
- ***Participants*** (subjects), potentially in ***groups***
- Experimental ***task***
- Performance ***measures*** (speed & accuracy)
- Trials
- ***Conditions*** / Treatments / Manipulations
  - modify the system
  - use alternative systems
  - Use different features of the system
- ***Effect*** of treatments on sample means
  - Within-subjects (each participant uses all versions)
  - Between-subjects (different groups use different versions)

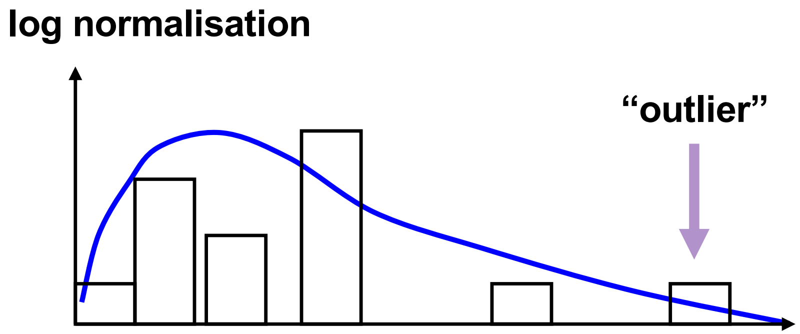
Controlled Experiments in HCI

- Based on a number of observations:
  - How long did Fred take to complete this task?
  - Did he get it right?
- But every observation is different.
- So we compare averages:
  - over a number of trials
  - over a range of people (experimental subjects)
- Results often have a normal distribution

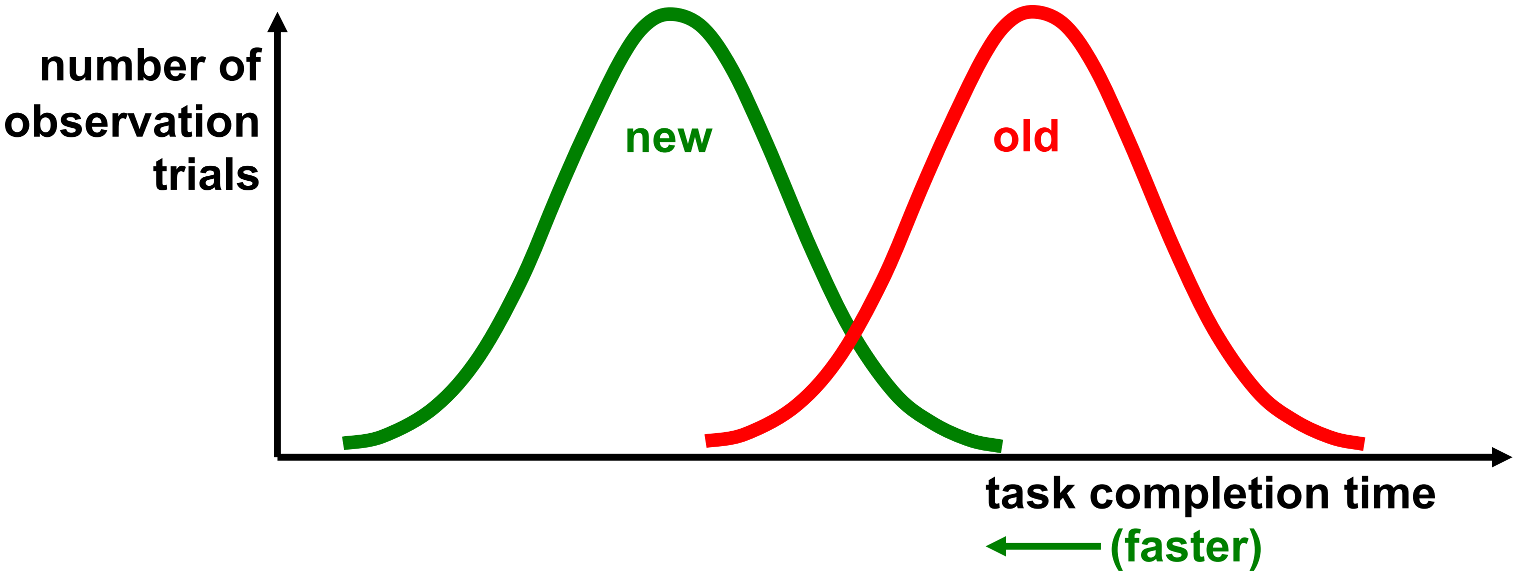
Sample Distribution





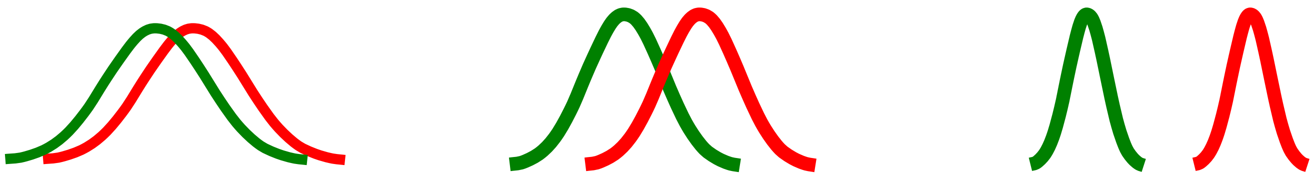


Effect Size


number of observation trials
new
old
task completion time
(faster)

Significance testing

- What is the likelihood that this amount of difference in means could be random variation between samples (null hypothesis)?
- Hopefully very low ($p < 0.01$, or 1%)



only random variation observed

observed effect probably does result from treatment

very significant effect of treatment

Experimental Manipulations

- Compare productivity gains (effect size) of version with new feature to one without?
  - Will system work without the new feature?
  - Will the experimental task be meaningful if the feature is disabled?
  - Must new feature be presented second in a within-subjects comparison (order effect)
  - Is your system sufficiently well-designed for external validity of productivity measure?
- Is full implementation necessary?
  - Can you simulate features with Wizard of Oz technique?

Measurement

- Speed (classically 'reaction time')
  - Time to complete task
- Accuracy (number of (non)errors).
  - Is outcome as expected
- Trade-off between speed and accuracy?
  - Or poor performance on both?
  - Check correlation between them
- Task completion:
  - Stop after a fixed amount of time (ideally < 1 hour)
  - Measure proportion of the overall task completed

Self-Report

- Did you find this easy to use? (Likert scale)
  - applied value: appeal to customers
  - theoretical value: estimate ‘cognitive load’
- Danger of bias
  - Subjective impressions of performance are often inaccurate
  - Reports may be influenced by experimental demand
    - Participants want to be nice to the experimenter
    - Should disguise which manipulation is the novel one
- May be necessary to capture affect measures:
  - Did you enjoy it, feel creative/enthusiastic, experience a ‘flow’ state?
- Alternative is to collect ‘richer’ data …

Think-aloud

- "Tell me everything you are thinking"
  - 'concurrent verbalisation'
- Problems:
  - Hard tasks become even harder while speaking aloud
  - During the most intense (i.e. interesting) periods, participants simply stop talking
- Alternative:
  - make a screen recording (showing cursor, or even eye-tracking trace?)
  - play this back for participant to narrate
  - 'retrospective verbal report'

Qualitative Data

- Protocol analysis methods, e.g.
  - verbal protocol – transcript of recorded verbal data
  - video protocol – recording of actions
- Hypothesis-, or theory-driven
  - Create ‘coding frame’ for expected/hypothetical categories of behaviour
  - Segment the protocol into episodes, utterances, phrases etc
  - Classify these into relevant categories (considering inter-rater reliability)
  - Compare frequency or order statistically
- Grounded theory
  - Open coding, looking for patterns in the data
  - Stages of thematic grouping and generalization
  - Constant comparison of emerging framework to original data
  - More interpretive, danger of subjective bias

Experiment Design

- Arrangement of participants, groups, tasks, trials, conditions, measures, and hypothesized effects of treatments
- Within-subjects designs are preferred
  - because so much variation between individuals
- This leads to order effects:
  - first condition may seem worse, because of learning effect
  - last condition may suffer from fatigue effect
  - task familiarity – can't use the same task twice
- Precautions:
  - Prior training to reduce learning effects
  - Minimise experimental session length to reduce fatigue effects
  - Use different tasks in each condition, but 'balance' with treatment and order
- These are typically combined in a 'latin square' where each participant gets a different combination

Analysis

- For an easy life, plan your analysis before collecting data!
- Will quantitative data be normally distributed?
  - t-test to compare two groups
  - ANOVA to compare effect of multiple conditions (which include latin square of task and order)
  - Pearson correlation to compare relationship between measures
- Distributions of task times are often skewed:
  - a small number of individuals complete the task quite slowly
  - don't exclude 'outliers' who have difficulty with your system
  - log transform of time is usually found to be normally distributed
- Subjective ratings are seldom normally distributed
  - chi-square test of categories
  - non-parametric comparison of means

Usability evaluation

- Rather than testing hypothesis, or comparing treatments
  - ask 'is my system usable' (a.k.a. fit for purpose, in a user-centric project)?
- More typical of commercial practice, for short-term rectification of immediate problems, rather than general understanding of design principles
  - Formative evaluation assesses alternatives early in the design process
  - Summative evaluation identifies usability problems in a system you have built
  - Repeated for iterative refinement in user-centred design processes
- Weaker as research, because no direct contribution to theory
  - However, applied research venues do require evidence of any claims made for new tools

Field Study Methods

- Laboratory studies are not adequate for:
  - organizational context of system deployment
  - interaction within a user community
- Typical methods:
  - ‘contextual inquiry’ interviews
  - ‘focus group’ discussions
  - ‘case studies’ of projects or organisations
  - ‘ethnographic’ field work as participant-observer
- All result in qualitative data, often transcribed, and in HCI research often analysed using grounded theory approaches

Planning your study

Candidate interactive systems / intelligent tools

- your own personal research
  - e.g. MPhil dissertation
- other research
  - other research in Cambridge
  - recent product releases
  - research prototypes developed elsewhere
- theoretical models
  - including topics introduced in our specialist lectures
  - is there a (well-articulated) user model to challenge?
- (user-centred) applications research
  - who is the intended user?
  - what will they be trying to achieve?

Representative tasks and measures

- Identify user activities you plan to observe
  - *either* assigned tasks (controlled experiment)
  - *or* toward the user's own goals (observational study)
- Will these explore an interesting research *question*?
- What *measures* are relevant to that question?
- Will *qualitative* data analysis be necessary?
- Will there be a threat to external validity?
  - Potentially resulting from choice of task, choice of measure or approach to analysis

Review of study design options

- Do you wish to carry out a comparison between systems, a (usability) evaluation of one system, or an open exploratory study – perhaps with no existing system?
- If you plan to conduct a controlled experiment, will it be possible to use a within-subjects design to reduce uncertainty resulting from variation between participants?
- What data analysis method will you use?
- What would you need to do in order to complete a pilot study?
- What ethical issues are raised by your planned research?

- A good starting point is to choose a published study that you would like to emulate / replicate

Theoretical goal

- What do you expect to learn from conducting your study?
- What contribution will it make to the research literature in interaction with machine learning?
- Where (venue, track) would you publish the results?

- A good starting point is to review contributions that were made in published studies you would like to emulate
  - Warning – be careful of studies done without prior training in HCI, and not published in peer-reviewed HCI venues.

Techniques for remote studies, if required by pandemic

- Surveys and questionnaires
- Interviews (e.g. by Zoom, potentially recorded)
- Instrumented remote prototypes (i.e. telemetry)
- Diary studies & experience sampling (see https://www.microsoft.com/en-us/research/project/meetings-during-covid-19/ for a recent example)
- Things that don't work well:
  - prototypes requiring a complicated software setup or low latency interaction
- Paid recruitment tools: UserTesting.com, AMT, Microworkers, Prolific, Gorilla, Sona
- Free recruitment tools: r/SampleSize, friends and family, this class (beware bias)!
- Survey/questionnaire deployment tools: Microsoft Forms, Google Forms, Survey Monkey

Review of feedback timetable (submit by noon each Tuesday)

- Week 2 - Research question (200 words) + a sample diary entry
- Week 3 - Study design (400 words)
- Week 4 - Another sample diary entry
- Week 5 - Draft literature review for final report (400 words)
- Week 6 - Draft introduction to report (200 words)
- Week 7 - Draft results section for report (400 words)
- Week 8 - Draft discussion section for report (200 words)

- + don't forget diary entries every week