Interaction with Machine Learning

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

1

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

2

Course objective: deliver "4th wave" of AI

- 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
- Practical HCI course:
  - Build, measure and observe
- For: Part III and MPhil ACS (research preparation), Part II (advanced HCI), Digital Humanities researchers (guided methods programme)

3

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	9	9	8	2
ML	6	9	12	1

4

Target outcome

- This is a specialised and focused practical research training course.
- The expected outcome:
  - You will achieve research competence in a 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 assessment will demonstrate competence in delivering and reporting practical experimental work
  - Presentations will be optional
- Digital Humanities students will be briefed separately

5

Lecture topics

- Week 2 - Mixed initiative interaction (AB)
  - information gain, cognitive ergonomics, agency & control
- Week 3 - Labelling (AS)
  - attribution, subjectivity, reliability, consistency
- Week 4 - Program synthesis (AB)
  - end-user programming, attention investment
- Week 5 - Visual analytics (AS)
  - visualisation, tool chains, design case studies
- Week 6 - Addressing data bias (Dr Daniela Massiceti, Microsoft Research)
- Week 7 - Interpretability (Prof Neil Lawrence)
- Week 8 – Your research presentations

6

Practical work plan – for ACS and Part II

- Week 1 - select research question (part II select replication study)
- Week 2 - discuss potential study approaches (part II discuss preparations)
- Week 3 - review and feedback on study proposals (part II may start work)
- Week 4 & 5 - review logistical issues / practical progress
- Week 6 - discuss preliminary findings
- Week 7 - discuss research implications
- Week 8 - final presentation

7

Assessment

- Everyone: final research report (80%)
  - Based on your practical work
  - Presented as a research paper
- Part II: phased submissions of work-in-progress drafts (20%)
  - Flat mark for each submission. *Text can be revised/reused in the final report!*
- ACS: optional (but recommended) work-in-progress drafts (0%)
  - Advisory grades will be provided as feedback, for revision in final report
- ACS: reflective diary (20%)
  - Summarise lectures
  - Document discussions
  - Record development of your own thinking
  - Make 8 weekly entries …
  - … plus a final summative review

8

Common criteria for assessment

- Part II – will be marked to CST Part II unit standards
- ACS – will follow standard ACS criteria:
  - 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
- All - preliminary feedback for guidance
  - 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

9

Continuous feedback

- Week 2 - Research question (200 words) + a sample diary entry for ACS
- Week 3 - Study design (400 words)
- Week 4 - Another sample diary entry for ACS
- 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)

10

Continuous feedback – submit by noon each Tuesday

Week	Part II	ACS
2	Selection of study for replication, with summary of the key research question	Research question (200 words) + a sample diary entry
3	Detailed work plan for data collection and analysis	Study design (400 words)
4	Literature review (summary of original publication, and other work that cites it)	Another sample diary entry
5	Introduction to the replicated study	Draft literature review for final report (400 words)
6	Results and data analysis	Draft introduction to report (200 words)
7	Discussion of results and conclusion	Draft results section for report (400 words)
8		Draft discussion section for report (200 words)

11

Reading suggestions

- Refresh undergraduate HCI
  - Cambridge notes online
  - Preece, Rogers and Sharp Interaction Design beyond HCI
- 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
- **Mostly: Recent research literature**

12

A note about the reading list

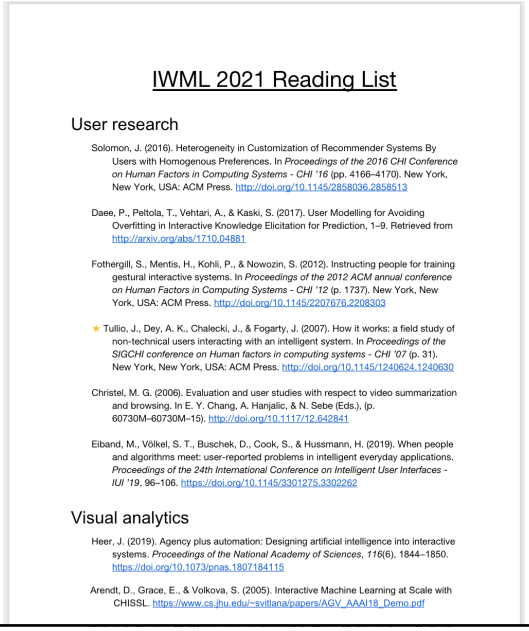
Available on course materials page.

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 2021 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

Visual analytics

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

Arendt, D., Grace, E., & Volkova, S. (2005). Interactive Machine Learning at Scale with CHISSL. https://www.cs.jhu.edu/~svitlana/papers/AGV_AAAI18_Demo.pdf

13

Theories of interaction

14

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

15

First wave: HCI as engineering "human factors"

- The "user interface" (or MMI "man-machine interface") is a separate module, designed independently of the main system.
- Design goal is efficiency (speed and accuracy) for a human operator to achieve well-defined functions.
- Use methods from cognitive science to model users' perception, decision and action processes and predict usability
  - At this point, relatively closely aligned with AI

16

Second wave: HCI as social system

- AI models did not result in more usable machines (see esp. Lucy Suchman)
  - Resulted in significant intellection challenge to cognitive science and AI
- The design of complex systems is a socio-technical experiment
  - Take account of other information factors including conversations, paper, and physical settings
- Study the context where people work
  - Use 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

17

Third wave: HCI as culture and experience

- 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!

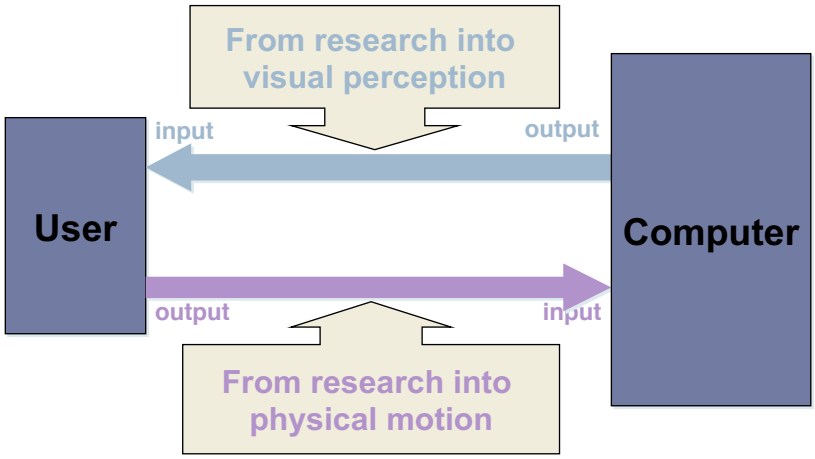
18

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 …

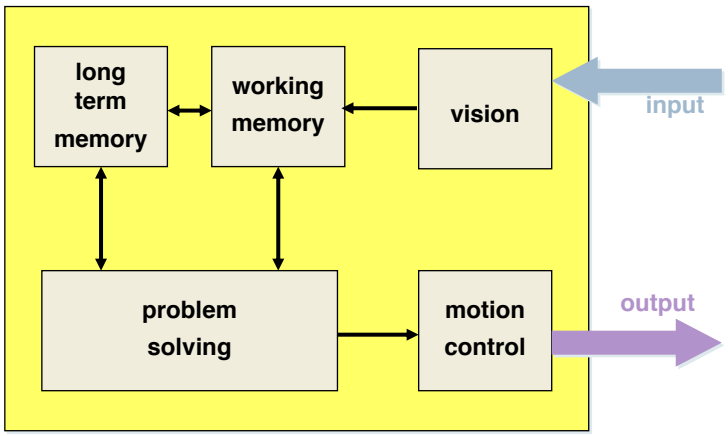
19

Classical cognitive science models of first-wave HCI



20

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



21

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
  - Reaction and 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'?

22

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

23

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

24

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

25

Intelligent interaction

26

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 etc 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
- Programming by example / program synthesis
  - See Lieberman *Watch What I Do*, but also e.g. Microsoft Excel FlashFill
- Human-in-the-loop automation
  - Autopilots, remote-operation, "autonomous" vehicles
- Turing tests – but what is the (wicked?) objective function?

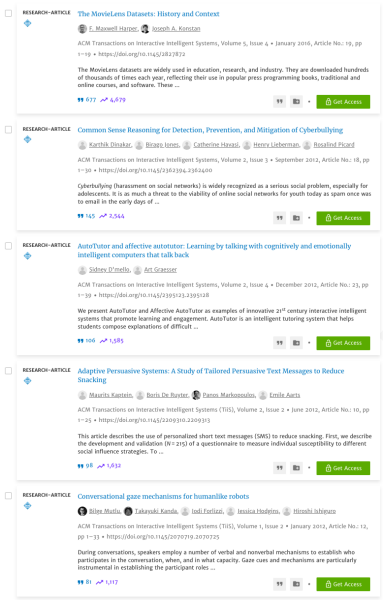
27

Topics at 2021 Intelligent User Interfaces (IUI) conference

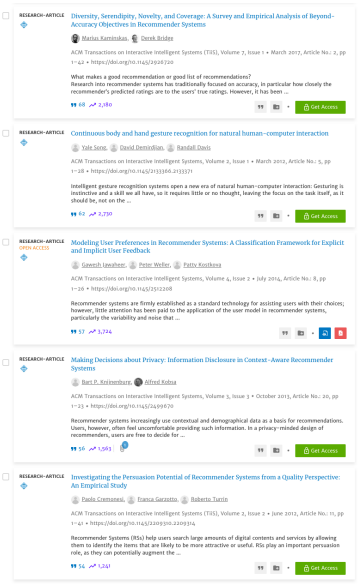
- Evaluations of intelligent user interfaces
  - experiments, studies, meta-analyses
- Modalities
  - agents, multi-modal (speech, gestural, gaze), conversational, tangible, visualization
- Interface paradigms
  - affective, collaborative, AR/VR, wearables/mobile, ubicomp
- Application areas
  - IoT, education, healthcare, social media, info retrieval/search, recommender systems

28

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



Dataset	Evaluation methods; recommender systems; experiments
Social intervention; user modelling	Gestural interaction; ML model
Affective computing; education; controlled experiments	User modelling; recommender systems; human-in-the-loop learning
Conversational interaction; questionnaire; experiments	Trust/transparency; recommender systems; experiment
Gaze interaction; conversational interaction Human-robot interaction; controlled experiments	Trust/transparency; recommender systems; experiment



29

TIIS Special Issues

30

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
- Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300484
- 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
- Alvarado, O., & Waern, A. (2018). Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–12. https://doi.org/10.1145/3173574.3173860
- Chen, N.-C., Suh, J., Verwey, J., Ramos, G., Drucker, S., & Simard, P. (2018). AnchorViz: Facilitating Classifier Error Discovery through Interactive Semantic Data Exploration. *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - IUI '18*, 269–280. https://doi.org/10.1145/3172944.3172950
- 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

31

Research methods

32

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 VPN – part II students send me details based on) https://www.cst.cam.ac.uk/local/policy/ethics/submission-read-only

33

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)

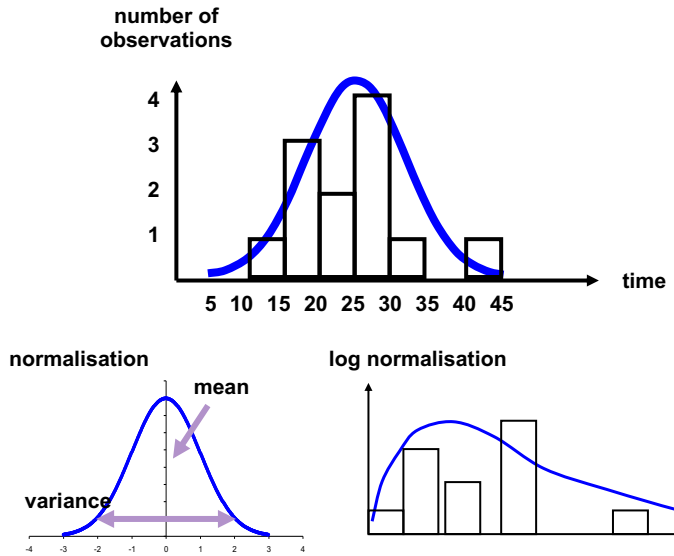
34

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

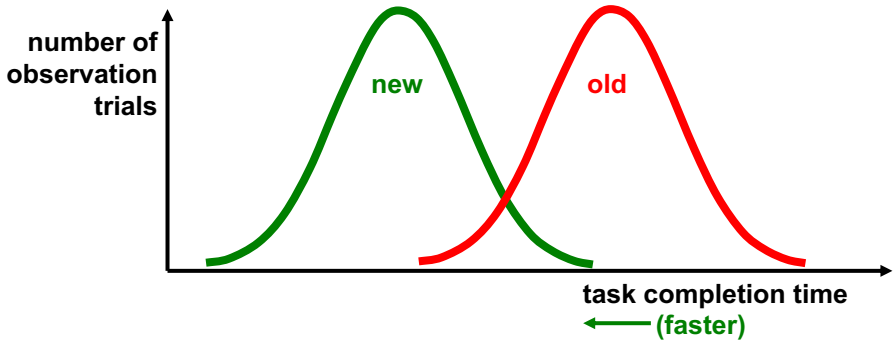
35

Sample Distribution



36

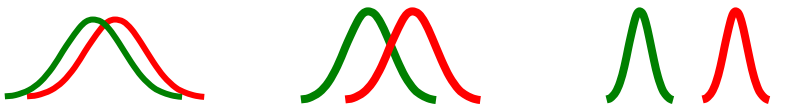
Effect Size



37

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

38

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?

39

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

40

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 inaccurate
  - Suffer from 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?
- Alternative is to collect 'richer' data …

41

Think-aloud

- "Tell me everything you are thinking"
  - 'concurrent verbalisation'
- Problems:
  - Hard tasks become even harder while speaking aloud
  - During the most intense (interesting) periods, participants simply stop talking,
- Alternative:
  - make video recording, or eye-tracking trace
  - playback for participant to narrate
  - 'retrospective verbal report'

42

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

43

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

44

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

45

Evaluation

- Rather than testing hypothesis, or comparing treatments
  - ask 'is my system usable'?
- More typical of commercial practice, for short-term goals, rather than general understanding
  - Formative evaluation assesses options early in design process
  - Summative evaluation identifies usability problems in a system you have built
  - Repeated for iterative refinement in user-centred design
- Weaker research, because no direct contribution to theory
  - However applied research venues require evidence of any claims made for new tools

46

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 analysed using grounded theory approaches

47

Planning your study

48

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?
- applications research
  - who is the intended user?
  - what will they be trying to achieve?

49

Representative tasks and measures

- Identify user activities you plan to observe
  - assigned tasks (controlled experiment)
  - or user's goal (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?
  - From task, measure or analysis

50

Review of study design options

- Do you wish to carry out a comparison, an evaluation, or an open exploratory study?
- If you plan to conduct a controlled experiment, will it be possible to use a within-subjects design?
- 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

51

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 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.

52

Techniques for remote studies

- Surveys and questionnaires
- Interviews
- 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
- Paid recruitment tools: UserTesting.com, Amazon Mechanical Turk, Microworkers, Prolific.
- Free recruitment tools: r/SampleSize, friends and family, this class (beware bias)!
- Survey/questionnaire deployment tools: Microsoft Forms, Google Forms, Survey Monkey

53

Review of feedback timetable (submit by noon each Tuesday)

Week	Part II	ACS
2	Selection of study for replication, with summary of the key research question	Research question (200 words) + a sample diary entry
3	Detailed work plan for data collection and analysis	Study design (400 words)
4	Literature review (summary of original publication, and other work that cites it)	Another sample diary entry
5	Introduction to the replicated study	Draft literature review for final report (400 words)
6	Results and data analysis	Draft introduction to report (200 words)
7	Discussion of results and conclusion	Draft results section for report (400 words)
8		Draft discussion section for report (200 words)

+ don't forget diary entries every week

54