Interaction with Machine Learning

MPhil ACS module R230 - 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: deliver "4th wave" of AI

- Four waves according to Hassabis (24 Nov 2017):
  - 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

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 field such as Intelligent User Interfaces, Interactive Intelligent Systems etc
- 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.

Lecture topics

- Week 2 - Labelling (AS)
  - attribution, subjectivity, reliability, consistency
- Week 3 - Mixed initiative interaction (AB)
  - information gain, cognitive ergonomics, agency & control
- Week 4 - Visual analytics (AS)
  - visualisation, tool chains, design case studies –
- Week 5 - Program synthesis (AB)
  - end-user programming, attention investment
- Week 6 - Interpretability (Adrian Weller)
- Week 7 - Risks of AI (Seán Ó hÉigeartaigh)
- 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

- Mini-project report (80%)
  - Based on your practical work
  - Presented as a research paper
- Reflective diary (20%)
  - Summarise guest lectures
  - Document discussions
  - Record development of your own thinking
  - Make 8 weekly entries …
  - … plus a final summative review

Common criteria for assessment

- 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
- 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

Continuous feedback

- 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 undergraduate HCI
  - Cambridge notes online
  - Preece, Rogers and Sharp Interaction Design beyond HCI
- Cambridge guidance on human participants
  - https://www.tech.cam.ac.uk/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**

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"

- 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

Second wave: HCI as social system

- AI models did not result in more usable machines
  - 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 and Contextual Inquiry to understand other ways of seeing the world
- Other stakeholders are integrated into the design process
  - Prototyping and participatory workshops aim to empower users and acknowledge other value systems

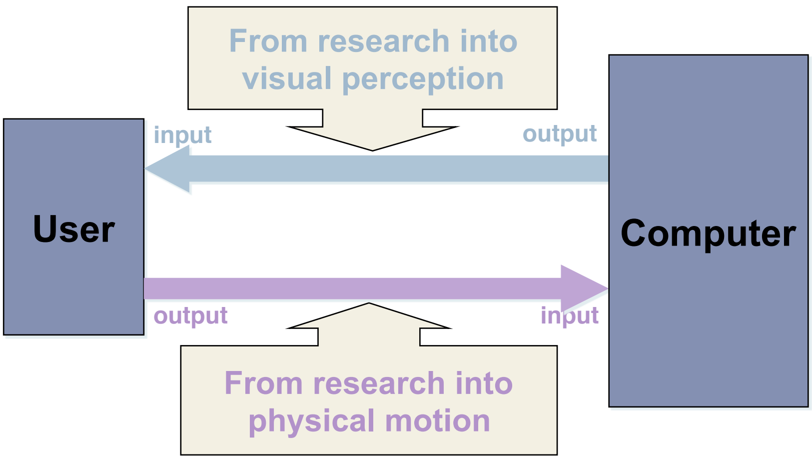
Third wave: HCI as culture and experience

- 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
- Has become pretty much completely divorced from AI
  - But this will change!

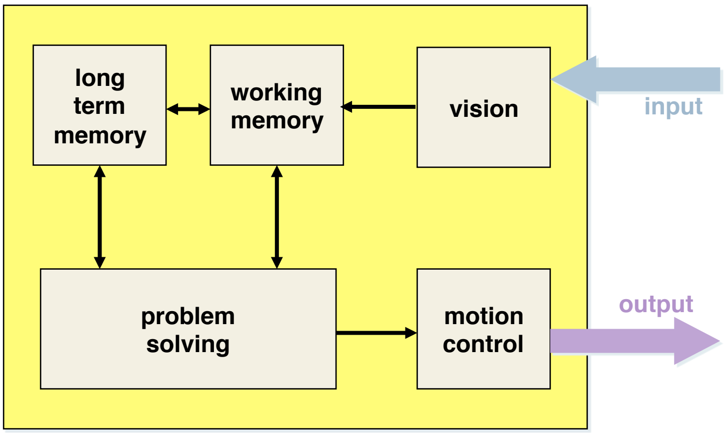
Summary of Cambridge HCI content

- 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 HCI
  - Visual representation. Text and gesture interaction. Inference-based approaches. Augmented reality. Usability of programming languages. Contextual observation, Formative and summative evaluation methods.

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
  - 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'?

The problem of learning

- 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

- 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

Wicked problems (Rittel & Webber)

- Examples of wicked problems
  - Slowing global warming
  - Stopping the spread of antibiotic-resistant diseases
  - Halting nuclear proliferation
  - Avoiding species extinction
  - Providing all citizens with health care (in the USA)
  - Colonizing Mars

Characteristics of wicked problems

1. There is no definitive formulation of a wicked problem
2. Wicked problems have no stopping rule
3. Solutions to wicked problems are not true-or-false, but good-or-bad
4. There is no immediate and no ultimate test of a solution to a wicked problem
5. Every solution to a wicked problem is a "one-shot operation"; because there is no opportunity to learn by trial-and-error, every attempt counts significantly
6. Wicked problems do not have an enumerable (or an exhaustively describable) set of potential solutions, nor is there a well-described set of permissible operations that may be incorporated into the plan
7. Every wicked problem is essentially unique
8. Every wicked problem can be considered to be a symptom of another problem
9. The existence of a discrepancy representing a wicked problem can be explained in numerous ways. The choice of explanation determines the nature of the problem's resolution
10. The planner has no right to be wrong

Intelligent interaction

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
- Turing tests – but what is the (wicked?) objective function?

Themes at Intelligent User Interfaces (IUI) conference

- Interactive labelling
- Information retrieval
- Information visualisation
- Recommender systems
- Personalisation
- Gesture recognition
- Tutoring systems
- User state recognition
- Trust

Themes in ACM TIIS (Trans. Intell. Interactive Systems)

- Creative arts applications
- Brain and body measurement
- Models of trust and persuastion
- Visual analytics
- Gaze and gesture interaction
- Recommender and query systems
- Affect and robot interaction

Themes of CHI 2016 workshop on Interaction with ML

- Gesture tracking
- Debugging
- 'User state' detection, including brain activity
- Interpreting topic models & sense-making
- Interactive visualisation / analytics
- Creativity – art and music
- Crowd-assisted data mining

Selection of full papers at CHI 2017

- UX Design Innovation: Some Challenges for Working with Machine Learning as a Design Material
  - Graham Dove, Jodi L. Forlizzi, Kim Halskov, John Zimmerman (CMU)
- The Trouble with Autopilots: Assisted and Autonomous Driving on the Social Road
  - Barry Brown, Eric Laurier (Stockholm / Edinburgh)
- Assessing Multiple Sclerosis with Kinect: Designing Computer Vision Systems for Real-world Use
  - Cecily Morrison, Kit Huckvale, Robert Corish, Jonas F. Dorn, Peter Kontschieder, Kenton P. O'Hara, Assess MS Team, Antonio Criminisi, Abigail Sellen (Microsoft Cambridge)
- Variolite: Supporting Exploratory Programming by Data Scientists
  - Mary Beth Kery, Amber Imogene Horvath, Brad A. Myers (CMU)
- Revolt: Collaborative Crowdsourcingfor Labeling Machine Learning Datasets
  - Joseph Chee Chang, Saleema Amershi, Ece Kamar (CMU / Microsoft)
- Us vs Them: Understanding Artificial Intelligence Technophobia over the Google DeepMind Challenge Match
  - Changhoon Oh, Taeyoung Lee, Yoojung Kim, SoHyun Park, Sae bom Kwon, Bongwon Suh (Seoul National University)

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 Lab Ethics committee, giving above details.

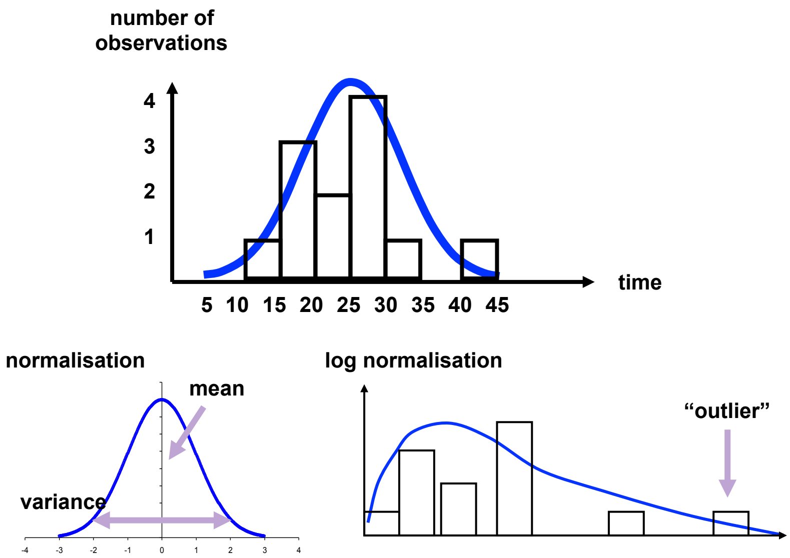
Controlled Experimental Methods

- ***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)

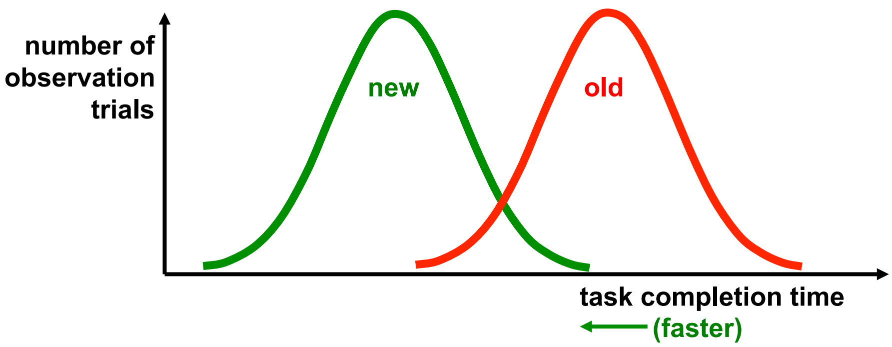
Controlled Experiments in HCI

- 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



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%)

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 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 …

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'

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

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

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

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?
- 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
  - 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

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

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.

Review of feedback timetable

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