# Intro

Assessments:

- AE1 – Prototype Design & Evaluation (20%)
    - Team Project
    - Description and Deadlines:
        - Usability evaluation (5%) – 15 Oct
        - Design & prototype (5%) – 29 Oct
        - Evaluation plan (5%) – 5 Nov
        - Evaluation report (5%) – 12 Nov
- AE2 – HCI experiment data analysis
    - Individual Project
    - Stats & Visualisations from real-world data set – 3 Dec
    - Submit PDF of Jupyter notebook
- Exam (60%)

Course:

- Readings are essential to succeed in course
- Ability to articulate and defend ideas needed
    - Concepts and terms are crucial to effectively communicate

Examinable:

- All assigned readings as listed on Moodle
- All text of lecture slides

Not examinable:

- Links provided in lecture slides labelled as "For reference"

# Week 2

## Reading

Consistency

- About pleasing others by giving them what they understand and can rely on
- Establishing
    - Setting and maintaining expectation by using elements people are familiar with
- Interpretation factors:
    - Other screens seen in other apps
    - Other screens seen in the same app
    - Location, situation
    - Age, background, experience
- Build consistency by **anticipating expectations**
- Types
    - Internal (in app)

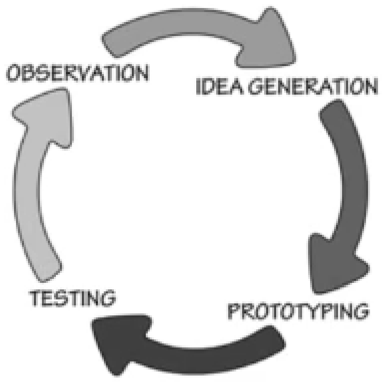
- External (similar to other apps)

## Lecture: HCI History and Visual Usability

HCI:

- Information Security, Speech-language pathology, Computer science, Economics & Human factors, Engineering, Design, Sociology & Social psychology, Ethnography, Cognitive science, Psychology

Iterative Cycle of Human-Centred Design




- Observation
  - Understand problems
- Idea generation
  - Draw on knowledge, conventions
  - Be creative, question everything, break conventions if with compelling reasons
- Prototyping
  - Paper/low-tech is often quickest way, and less risk of becoming ‘attached’ to specific ideas
- Evaluation
  - Many possible methods

Typing Study Case

- General Stats & Info
  - ~3.2 hours a day, typing
  - ~51.56 wpm
    - “Fast” typists - ~89.56 wpm
    - “Fast” typists make fewer errors
    - “Fast” typists use both hands more effectively
- Metrics for evaluating typing:
  - Performance measures
    - wpm
    - uncorrected error rate

        - error corrections
        - keystrokes per character
        - inter-key interval
        - keypress duration
    - Error metrics
        - Substitution error rate
        - omission error rate
        - insertion error rate
- Rollover
    - Rollover typing – previous key is not released before the following key is entered
    - On average, users perform rollover for 25% of keystrokes (SD=17%), and this has a high correlation with performance ($r = .73$, $p < 0.001$)
        - Fast typists perform rollover for 40-70% of keystrokes
    - Idea:
        - Encourage more rollover on soft keyboards, might see faster typing speeds
        - Design a soft keyboard that employs haptic feedback to train and encourage rollover typing behaviours

Designing Interactive Systems

- What is your idea?
- Why do you think it will work?
- Where is the proof?
- Collecting data from potential users
    - Can provide insights, but must be collected in a valid and theoretically sound way

History of HCI

- Tied to history of Computing
    - Colossus (1940s) – Bletchley Park code breaking
    - Programming via punched cards
- Initial computers in research labs, took up full rooms
    - Only ever operated on by specialists/engineers/the people who built them
- As technology progressed, got smaller/more affordable, started appearing in workplaces and homes
    - A need for 'real' people to be able to operate them
    - Thoughts of human efficiency/task completion times/error rates
- A need for a new discipline to study these issues
- Start of HCI
    - Thought of as beginning in early 1980s
        - Conferences began
        - Influential textbooks
        - Emergence of the GUI
    - Trying to create an … psychology of HCI
    - Based on knowledge of human psychology
        - Perception
        - Cognition
        - Motor function

  - For many in software design communities, first exposure to psychology basis
  - Engineering-style theories to give approximate calculations of how efficiently humans would interact
- Graphical User Interfaces
  - Xerox Star – 1981
    - First GUI computer released
    - Bit-mapped display
    - WIMP, WYSIWYG
    - Desktop metaphor
    - Yet not a commercial success
      - Very expensive; network terminal, not 'personal' computer
  - Apple Macintosh – 1984
    - Brought GUI to a wider audience
- Broadening of HCI topics
  - ~1980s: early research often looking at efficiency
    - Measure speed and accuracy
    - Lab-based studies
    - Formal experiments
  - ~1990s: field started to broaden, alongside importance of Internet
    - Emails, Web: topics related to communication
  - ~2000s: mobile/portable computing
    - Real world studies 'in the wild'
    - New technologies: sensors, wearable, XR
    - Study social, emotional, cultural issues
    - "Older" forms of research haven't gone away
- Broadening methods
  - Technology pushed progress here too
    - Eye-tracking studies, EEG, …
    - Large-scale studies, users' own devices
  - From early studies that timed tasks/counted errors
  - Brought in new techniques more from sociology than psychology
    - Ethnography
    - Interviews
    - Case studies
- 3 Waves:
  - First Wave: Psychology and Perception
  - Second Wave: Organisational and Process-Oriented
  - Third Wave: Social and Ubiquitous

Visual Usability

- "Visual communication of any kind, whether persuasive or informative […] should be seen as the embodiment of form and function: the integration of the beautiful and the useful." – Paul Rand, A Designer's Art, p. 3
- The Things You See Around You Today Are Not There by Random Chance

    - The interfaces familiar with us may seem easy to design, but are the result of many attempts and many failed designs
- Multidisciplinary challenges
    - Graphic design alone doesn't help teach us how to create complex IS
    - Usability alone might not teach us how to create the best experiences
- A complex interface might need to convey many messages
    - Should provide order, patterns to help people process info and derive meaning
- Visual Usability
    - Designs grounded in principles of aesthetics and understanding of people
    - We should be able to design and defend a design based on heuristics and best practices
- **Consistency**
    - Establishing consistency means setting and maintaining **expectations**
    - **External**
        - Consistency with **other** similar apps
        - If designing an interface for online shopping, it should be similar to the established look/feel of existing interfaces
    - **Internal**
        - Consistency **within** different parts of an app
        - If designing an interface for online banking, all the views need consistency
    - Internal/External can sometimes clash
        - e.g., suite of apps from same company. Should they primarily look and feel like each other or should each one meet the conventions of that type of app?
    - Types:
        - **Layout**
            - Screens showing similar info should have all elements positioned the same way
            - Spatial relationships should remain consistent
        - **Typography**
            - Use fonts, weights, and sizes meaningfully and consistently
        - **Colour**
            - Use colour to establish and maintain consistency – typically means establishing a defined colour scheme
        - **Imagery**
            - Charts, logos, videos, photography, icons, backgrounds, an anything else that isn't typography
    - Breaking
        - Can break the rules to make a point/**highlight** something
        - e.g., make one piece of content bigger than others if it's the most important/where you want to guide the eye
        - Don't change more than 2 aspects of a single item
- Hierarchy
    - Visual hierarchy is used to communicate structural relationships, and relative importance
    - More important items need more "visual weight"

        - Understanding behaviour of gaze is important when deciding how to effectively give important elements more visual weight
        - Use position, size, colour, groupings, contrast, control types to represent priorities
        - Make sure people notice what they need to – based on identified user priorities
    - Start with black and white wireframes – only vary size and positions
- Layout
    - Screen size
        - The screen gives the frame within which the entire interface sits
        - Core layout principles might apply to all screen sizes – main thing is how elements relate to each other, so layout can flex
    - Position
        - Does the relative position of elements communicate structure?
        - Might need to balance lots of relationships
    - White Space
        - Absence of content is equally important, for example, the sparse design on a Google landing page
        - Trick to create dense but appealing screens is white space to group and establish hierarchy
    - Grid
        - Align items relative to (invisible) horizontal and vertical lines
- Proximity, Scale, and Alignment
    - Proximity
        - Is the relative placement of items arbitrary or meaningful?
    - Scale
        - Is the relative size of elements arbitrary or meaningful?
    - Alignment
        - Is the alignment consistent and used to represent the hierarchy?
- Colour
    - Powerful way of attracting the eye
        - Can create emotional response
        - Enhance usability and appeal
        - Aid understanding by creating connections between related items
    - Choices
        - Can be cultural associations
        - Specific UI conventions
            - e.g., red for error messages (or only if critical)
    - Shouldn't convey anything crucial through colour alone
        - Visually impaired/colour-blind users
    - Properties
        - Hue is a categorical description of the perceived colour
            - red, yellow, green, cyan, blue, violet, magenta, purple
        - Saturation is the purity of colour compared to grey
            - When fully saturated, the 'purest' form of the hue
            - Saturated colours can draw the eye more
        - Brightness – relative amount of light

- Contrast
  - Warm-Cool Contrast
    - warm – red, orange, yellow
    - cool – purple, blue, green
  - Complementary Contrast
    - orange-blue, yellow-purple, red-green
- Hierarchy of Colour
  - Primary, Secondary

# Week 3

## Lab 1 Feedback

Observations:

- Most people used Excel
  - Encouraged options: Jupyter/matplotlib/seaborn
- Many possibilities for what to plot
- Don't use the same axis for different units

## Lecture: Human Perception and Capability

Studying the Human

- HCI – Human-Computer Interaction
- Early HCI work took findings/approaches from Psychology to apply interactions with computers
  - Perception
  - Cognition
  - Motor function
- Used to guide sys dev
- Continue to measure, refine, experiment

Time Scale of Human Action

| Scale (sec) | Time Units | System | World (theory) |
|---|---|---|---|
| $10^7$ | Months | | **SOCIAL BAND** |
| $10^6$ | Weeks | | |
| $10^5$ | Days | | |
| $10^4$ | Hours | Task | **RATIONAL BAND** |
| $10^3$ | 10 min | Task | |
| $10^2$ | Minutes | Task | |
| $10^1$ | 10 sec | Unit task | **COGNITIVE BAND** |
| $10^0$ | 1 sec | Operations | |
| $10^{-1}$ | 100 ms | Deliberate act | |
| $10^{-2}$ | 10 ms | Neural circuit | **BIOLOGICAL BAND** |
| $10^{-3}$ | 1 ms | Neuron | |
| $10^{-4}$ | 100 µs | Organelle | |

- Bands:
  - Social Band
    - Days, weeks, months
      - Activities such as workplace habits, social networking, online dating, privacy
      - Require development of social bonds or establishing norms/standards
    - Ex: study on how people develop relationships in online dating https://dl.acm.org/citation.cfm?id=2702417
      - Interviews with members of the community
      - Participation/observation in active forums
    - Qualitative methods dominate
      - Although often opportunity for mixed methods studies/data analytics
  - Rational Band
    - Minutes or hours
      - Tasks, like web site use, user search strategies, OS navigation
      - User must experience and interface and make decisions about their next actions
    - Ex: user search behaviour https://dl.acm.org/citation.cfm?id=2124322
      - How often do users “branch” their search results?
      - How many “branches” do users generate during a typical search?
      - Why do users establish a new “branch”?
  - Cognitive Band
    - 100 milliseconds to 10 seconds
      - Pointing devices, selection techniques, text entry, gestural input
      - Times based on reaction times and biomechanical properties
    - Ex: multitouch rotation gestures https://dl.acm.org/citation.cfm?id=2481423
      - Does the angle of rotation impact performance?
      - Do users pivot from the thumb or rotate multiple touchpoints?
      - Does the starting angle impact performance?
  - Biological Band
    - Less relevant for most HCI research/practice

Model of HCI

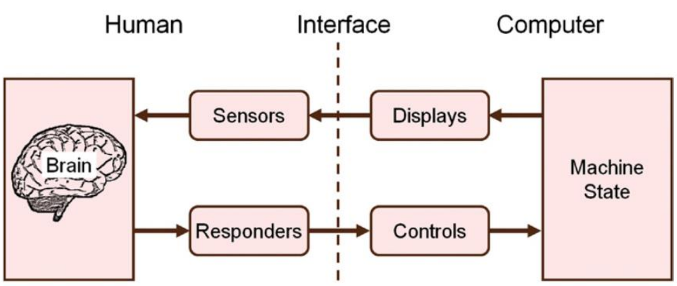
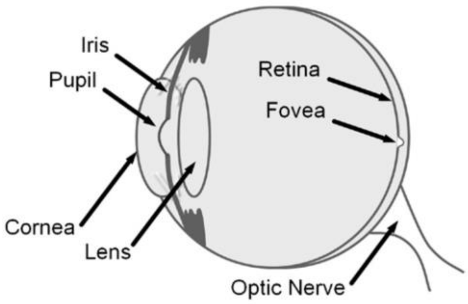



**FIGURE 2.2**

Human factors view of the human operator in a work environment.

*(After Kantowitz and Sorkin, 1983, p. 4)*

Human Senses

- Purely physiological (perception involves brain processing)
- Vision
    - 

    - Biology:
        - Light passes through the lens
        - The lens focuses light into an image projected into the retina
        - The retina converts visible light into neurological signals
        - The centre of the retina, the fovea, processes details
    - Properties:
        - Frequency of visible light
        - Intensity
            - Eye light sensitivity varies by wavelength
        - Fixations and Saccades
            - Fixations process detail while the eyes are still
            - Saccades – rapid movements (30-120 ms) of the eyes to a new position
- Hearing
    - Sounds are perceived from cyclic fluctuation of pressure
        - Typically, in air
    - Loudness
        - Subjective perception of sound pressure level
    - Pitch
        - Subjective perception of frequency
    - Timbre
        - Harmonic structure to be described as richness/brightness
    - Envelope
        - Changes in the subjective properties over time
- Touch
    - Touch / haptic
        - Through vibration, air, and ultrasound
        - https://dl.acm.org/citation.cfm?id=2663280
    - Temperature
        - https://dl.acm.org/citation.cfm?id=1979316
    - Pain
        - Try to avoid in HCI
    - Proprioception
        - The ability to sense the position of your body and limbs

- Smell
    - Olfaction
        - The ability to perceive odours
    - HCI has explored scent through scent ‘cubes’
        - fans that disperse scent, and pressurised delivery systems
    - Olfoto: tagging photos with smells vs text
        - https://dl.acm.org/doi/abs/10.1145/1124772.1124869
- Taste
    - Chemical reception of sweet, salty, umami, bitter, and sour
    - TastyFloats
        - Levitate food onto user’s tongue
        - https://dl.acm.org/citation.cfm?id=3134123
- Multi-sense interactions
    - https://dl.acm.org/citation.cfm?id=3134123

Human Responses

- Limbs
    - Input for systems is primarily achieved by moving the limbs in 3D space
        - Typing, using a mouse, using a trackpad
        - Use limbs to generate a signal that is interpreted as input
- Voice
    - Speech recognition has come a long way, but we still face challenges of segmentation (separating intended input, like talking to somebody else), recovery from errors, and information throughput
- Eyes
    - Selection based in Gaze is a common approach in VR, and becoming more common in less instrumented environments as well
        - For example, consider common phone unlocking techniques
    - Most info probably also coming in through vision, so eyes doing double tasks

Human Brain

- Perception
    - First stage of processing in the brain
        - Associations and meanings take shape
    - Just Noticeable Difference
        - Below what threshold can humans no longer perceive difference?
    - Ambiguity
        - Illusions work when our perception fills the gaps in ambiguous stimuli
            - Ponzo lines demonstrate how our depth perception changes how we look at 2 black lines
        - There are illusions that can trick our visual, aural, and haptic senses
- Cognition
    - Human process of conscious intellectual activity
        - Thinking, reasoning, deciding
- Memory
    - Ability to store, retain, and recall information
    - Short term memory capacity: $7 \pm 2$

        - Has often been used to guide UI design, e.g., number menu items
            - Might be misunderstanding the original intent
            - Shorter menus probably still good!

Human Performance

- Speed Accuracy Trade-off – tasks completed faster are more error-prone
    - People often prioritise speed or accuracy differently based on context
- Most of early HCI measured this, but still important and studied today
- e.g., performance with various input devices
    - Also augment overall human performance, e.g., find answers to questions with visualisation tool vs looking at raw numbers
- Reaction Time
    - Different sensory modalities have different reaction times
        - 150ms audio
        - 200ms visual
        - 300ms smell
        - 700ms pain
    - Visual search is another example of reaction which includes more complex cognition than simply responding to stimuli
- Skilled Behaviour
    - In most tasks beyond simple responses, human performance can increase with practice
    - Can involve training sensor and motor or mental skills, can involve both
- Attention
    - Task requires attention – When task performance degrades while performed simultaneously with another
    - Divided attention – concentrating/performing more than one task at a time
        - Typically, this will degrade performance, which is not an option in safety-critical contexts like driving
    - Focused attention – attending to 1 task to the exclusion of all others
        - The ability to ignore external events not always possible or feasible
        - In a noisy room, you might be able to have a conversation but are likely to be distracted
    - Sensory modalities are often thought of as channels, but not so simple in practice

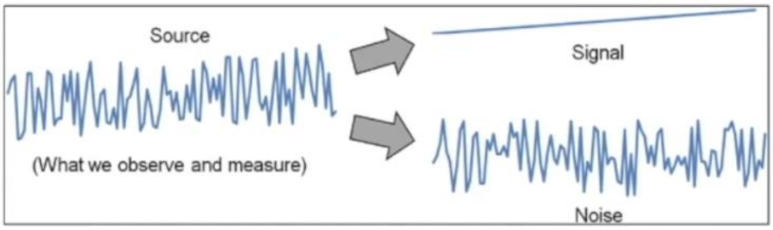
Human Error

- Error – a discrete event in a task where the outcome is incorrect or deviates from the desired outcome
- In practice, this of coarse measure of error provides a limited understanding
    - Consider the Key Stroke dataset - “error” isn’t reported
- Often trying to measure something more complicated than % of errors

# Week 4

## Lecture: Quantitative Methods

Methodology – the way an experiment is designed and carried out

- Sound methodology is critical to allow us to understand what is really going on (signal) in a noisy and messy world (noise)
- 


Reasons to care:

- Will help run studies in 3rd year group projects and 4th year dissertation
- Helpful for publishing scientific papers testing hypotheses
- Need to know how data was collected to handle it
- Critical thinking

Validity

- Internal
    - Is effect observed due to varied condition(s)?
- External
    - Are experimental results generalisable to other people/situations?
        - Sampling
        - Realistic conditions
- Often tension between internal vs external: one at expense of other?

Ethics

- Often borrow from psychology research
- **Informed consent**
    - Nature of research
    - Methodology
    - Risks/benefits
    - Right not to participate or to withdraw
    - Right to anonymity and confidentiality
- Issues in HCI work can involve recruitment of **vulnerable groups** (e.g., when investigating assistive technology), or **deception** that might be involved during a study

Independent Variables (Factors)

- Experiments with independent variables are often called *factorial experiments*
- Naturally occurring (age) or directly manipulated by experimenter
- **Characteristics**
    - e.g., of computer interface (input device, feedback modality, display size)
    - e.g., of participants (gender, handedness, expertise)
- **Circumstances**
    - e.g., background noise, room lighting
- Levels: each test condition (mouse, trackpad) are **levels** of independent variable (input device)

Dependent Variables

- Any observable, measurable behaviour
    - typing speed, eye movements, ‘negative facial expressions’, ‘read text events’, how to respond to questionnaire

Effect numbers

- More IVs -> more comparisons
    - Increase rapidly
- Limit IVs to 1-3

Control Variables

- Not under investigation, but might influence participant behaviour (DVs)
    - keyboard angle, chair height, display size
- Experimenters control these variables to prevent their influence by setting up their study in a controlled environment and recruiting with strict inclusion/exclusion criteria
    - e.g., right-handed only, experienced users only
- Increases internal validity but reduces external

Random Variables

- Often better to allow some variables to vary randomly to generalise results
    - e.g., height, hand/finger size, social disposition
- Each study will require judgment about the trade off between control and allowing random variation
    - e.g., using questionnaire of motion sickness and recruiting only those under a threshold
        - In a study investigating the acceptability of 2 in-car interaction techniques for the general population
        - In a study comparing the use of a VR headset in a moving vehicle with different VR conditions to mediate sickness level

Confounding Variables

- A circumstance/condition will change systematically with the independent variable
    - e.g., practice, different types of measurements for levels of the IV, prior experience with an interface (e.g., when comparing Google to anything)
- Such variables are confounding because they prevent the possibility of a cause-and-effect relationship being inferred from the results. **Need to be controlled for**

Participants

- To correctly assume that research results apply to people other than those recruited:
    - Recruit people from the population you want to **investigate**
    - Recruit **enough** participants
- Recruitment methods
    - Ideal = participants drawn **at random** from a population
    - In practice = **convenience sampling**
- How many?
    - More is better

    - Balance between representation and practical nature, sometimes ethical considerations
    - Practical: Not much time to recruit (e.g., in a student project), population difficult to access, more testing delays product going to market
    - Ethical: Study puts participants under burden; continuing study beyond necessary delays useful intervention/technique that can improve access
- Central Limit Theorem
    - As sample size increases to >=30, it becomes approximately normally distributed
- Within/Between subjects
    - Within-subjects (repeated measures) (preferred in HCI)
        - All participants use both sides of experiments, then compare average difference in performance
    - Between-subjects
        - Split the participants in 2 groups, compare performance of each group
- Order Effects
    - Interference between test conditions
    - Learning effects
        - People's performance improves as the study goes on, this makes it hard to know if observed differences are due to the IV or confounding variable (learning)
    - Fatigue effects
        - Possible that people get worse due to tiredness/loss of interest
- Counterbalancing
    - Balance the order in which participants do each level
    - Latin square – n x n table that allows conceptualisation (and generating) of counterbalanced conditions
        - Assign users to groups. Each group gets different orders of conditions
        - Ensure equal number of people in each group
        - Use Latin squares to ensure that each condition in a text entry experiment is presented first for each group
        -

| | swype | standard |
|---|---|---|
| Group 1 | 1 | 2 |
| Group 2 | 2 | 1 |

| | swype | standard | prediction |
|---|---|---|---|
| Group 1 | 1 | 2 | 3 |
| Group 2 | 2 | 3 | 1 |
| Group 3 | 3 | 1 | 2 |

| | swype | standard | prediction | voice |
|---|---|---|---|---|
| Group 1 | 1 | 2 | 3 | 4 |
| Group 2 | 2 | 3 | 4 | 1 |
| Group 3 | 3 | 4 | 1 | 2 |
| Group 4 | 4 | 1 | 2 | 3 |

    - Necessary only when order effects are confounding variables
    - Group size matters when order effect is an IV in the study
- Randomisation
    - When Latin Squares become needlessly complicated/impractical, randomisation mitigates order effects

Longitudinal studies

- Investigate learning effects over time
- Important considering the ubiquity of technology use in everyday life

- Crossover
    - In the case of a new product, it may be that performance on the traditional system will start off better, but that this crosses over after long term use

Running the Experiments

- Pilot study before running it
    - Technical issues, no one can understand task instructions, takes too long to finish
- Use consent forms
- Be consistent
    - Neutral manner, use a script if needed
- Be aware of bias

# Week 5

## Lecture: Surveys, Focus Groups, Qualitative Methods

Surveys

- In general:
    - Allow researchers, designers, and devs to capture high-level info about user experiences, attitudes, and perceptions
    - Paper/phone/email/website
    - Low cost, fast, broad reach
    - A well-designed and analysed survey can provide useful insights; a poorly designed and analysed survey is noise (and waste of time for researchers and respondents)
- As a Topic of Research
    - Population sampling
    - Optimisation of data collection
        - e.g., return rates
    - Reduction of biases in questions
    - Question order effects
    - Computer vs paper-based
        - Study (1983) found less socially desirable responses & longer open-ended responses in digital survey
    - Tourangeau's 4 cognitive steps to survey responses (1984)
        - **Comprehension** of the question, instructions, and answer options
        - **Retrieval** of specific memories to aid with answering the question
        - **Judgement** of the retrieved info and its applicability to the question
        - **Mapping** of judgement onto the answer options
- **Good** for:
    - Attitudes
    - Intent
    - Task Success
    - UX Feedback
    - User Characteristics
    - Interactions with Technologies

    - Awareness
    - Comparisons
- Can survey regularly to assess changes over time
- **Bad** for:
    - Precise Behaviours
        - Log data can often give more accurate info
            - Not infallible: log data can fail to record, might need to stream back data over unreliable network connection, might record someone else using the device, etc
    - Underlying Motivations
    - Usability Evaluations
- Pitfalls
    - Surveys need to consider experimental design and confounding factors
        - Common issue – ask about multiple dependent variables in a single question
            - e.g., One a scale of 1 to 5, rate how usability and enjoyable your experience was
    - Low completion rates
        - Repetitive questions, poor usability, poor design
    - Noisy data from bad question design
        - Vague/ambiguous questions provide noisy data
    - Biased Questions
- How to develop a good survey:
    - Research Goals
        - Articulate goal, identify user **constructs** (factors)
            - Only crucial constructs: too many makes the survey too long, might increase drop-out rate
        - Cognitive pretesting: are respondents interpreting constructs as intended?
            - Test: think aloud with a few users
    - Population and Sampling
        - As in labs, survey respondents need to be recruited from the target population
        - To ensure intended population is represented, might develop inclusion criteria
            - e.g., only respondents with >20 hours of gameplay in a game
    - Questionnaire Design
        - Open-ended Questions – Free response text
            - Use when:
                - impossible to determine all possible answers in advance
                - list of options would be unusably long
                - capturing numerical data (can always be grouped later)
                - qualitative aspects of user experience
        - Closed-ended Questions – Predefined answers
            - Use when there are small number of possible answers
            - Using rating scales (Likert scale: Agree to Disagree) and ordinal data
            - Unipolar construct: 0 to extreme amount
                - Importance, usefulness

        - Labels: “Not at all”, “Slightly”, “Moderately”, “Very”, “Extremely” – shown to be semantically equidistant
      - Bipolar construct: Extreme negative to extreme positive
        - Ease of use (difficult to easy), visual appeal, …
        - Labels: “Extremely”, “Moderately”, “Slightly”, “Neither nor”, “Slightly”, “Moderately”, “Extremely”
      - Can have single/multiple choice, ranking, rating
      - Implications on how to analyse the data and what kind of statistics can be completed
    - Measurement error: deviation of answers from true values on the measures
      - Can come from respondent
        - Lack of motivation
        - Lack of comprehension
        - Lies
      - Can come from instrument
        - Working/design flaws
        - Technical/interaction flaws
      - No opportunities for clarification
      - Usually only deploy once
        - Can’t revise halfway through, then consider all results when respondents have answered slightly different questions
    - Bias
      - Each question must be carefully checked for bias
        - A bias introduced in 1 question can even affect subsequent ones
      - 5 types:
        - Satisficing (Tiring out)
          - Surveys require focus, motivation, and/or cognitive load
            - Respondents don’t put in effort => fail to follow >=1 of Tourangeau’s 4 cognitive steps
          - Weak: might pick answer that’s OK but not optimal
          - Strong: pick answer completely randomly
          - Avoid options such as “no opinion” or “n/a” if you want to force a choice
            - Can avoid by offering even number of possible responses on scale
          - Avoid repetitive questions that use the same scale (‘straight-lining’)
          - Avoid overly long questionnaires
          - Communicate importance of survey to increase motivation
          - ‘Trap’ questions
        - Acquiescence
          - Respondents more likely to agree than disagree
          - Avoid yes/no questions, phrase questions neutrally
        - Social Desirability

            - Responses given because the respondent thinks it will be looked upon favourably
            - e.g., on a scale of 1-5, how important is climate change research
        - Response Order
            - Respondents being more likely to choose responses at the beginning or end
            - Primacy/recency effects
            - Categorical answers should be presented in a random order
            - Rating scales (positive->negative, negative->positive) can be counterbalanced (give one to half, other to other half)
        - Question Order
            - Same as experimental design, order effects should be mitigated where order effects may occur
            - Might keep demographic questions in the same order at the beginning of a survey, but randomise the remaining questions
            - If randomisation doesn't apply, organise by:
                - starting with more general questions and finishing with specific questions
                - starting with easy questions and finishing with difficult questions – limit amount of dropout
                - starting with most important questions – limit impact of dropout
    - Questions to Avoid:
        - Broad questions
            - Provide noisy data and confuse respondents
        - Leading questions
            - Influence respondents and add noise/bias to data
        - Double-barrelled questions
            - Conflate multiple constructs and make clear conclusions impossible
        - Recall questions
            - Self-report from the past is inaccurate and noisy
        - Prediction questions
            - Self-prediction is very susceptible to bias
    - Existing standards (widely tested, validated, accepted):
        - SUS (System Usability Scale)
            - One of the most commonly used
            - 10 questions (efficiency, satisfaction), yielding single score
        - NASA TLX (Task Load Index)
        - UEQ (User Experience Questionnaire)
- Review and Testing

        - Running a pilot to determine realistic completion times and check bugs/configuration issues/etc are fixed before public launch
            - Minor tweaks to survey configuration can mean all the data collected thus far is invalid/incomparable to data collected moving forward
    - Implementation and Launch
    - Data Analysis and Reporting
        - Can learn certain findings, e.g., user attitudes, from interviews, but surveys can get statistically reliable metrics
        - Quantitative and Qualitative analysis possible
- Types:
    - Experience Sampling
        - 'Ecological momentary assessment'
        - Regularly fill out several brief questionnaires
            - Daily/several times a day
            - At specific times, and/or by responding to alerts
        - Sampling regularly, don't know participant circumstances => limit burden
            - Closed-ended, fast, few questions
        - Ask about current activities and feelings
            - NOT recall: reduce cognitive biases of memory-based self-report methods
    - Intercept Survey
        - Deploy while person is using the technology
        - e.g., popup in an app while in use
        - Real-time data capture – minimise issue of imperfect recall
        - Can be triggered by particular behaviour of interest
        - Might be very annoying
        - Design the timing – maybe not while using feature, but some time soon after
            - Balance precision of recall/getting in users' way
    - With other methods
        - Combine larger and smaller scale
            - Use a survey
                - Captures high-level info from a broad group of users
            - along with a focus group/lab study/interview study
                - Captures detailed info from a smaller group of users
            - Is data representative or anecdotal? What are the reasons for large-scale trends?
        - Keep a record of all user interactions
            - Then can compare survey responses when user has done X or Y – create user groups
                - e.g., impressions of product – those who skipped tutorial and those who didn't

Focus Groups

- In general:
    - Involve bringing together a group of participants for a group discussion

        - Can be of various sizes: 3-6 common in HCI, sometimes 6-12+
    - Video/audio is recorded and analysed using qualitative methods
    - Develop protocol/script
- Can be efficient way of getting many viewpoints
    - e.g., 4 hour-long groups of 6
    - Participants can debate issues among themselves
- Cons
    - Shy people
    - Someone monopolising conversation
    - Solution: split group
        - more time for each person to contribute
    - 'Groupthink' and conformity
        - Can be spotted if separate groups give opposite consensuses
- Experience Prototype
    - Prototypes can range in fidelity, but give devs, designers, and potential users hands-on experience with a prototype
        - Focus on creating an experience, especially during the early stages of design when a fully functional prototype doesn't exist
- Keep/Lose/Change
    - In groups, facilitating positive, negative, and creative feedback can be achieved
    - After experiencing a prototype/app/demonstration, ask:
        - What would you keep?
        - Lose?
        - Change?
    - Often works best with large printouts of interface views that participants can mark up and annotate with post-its as a group

# Week 6

## Lecture: Ethnography, Interviews, Qualitative Methods

Ethnography

- Understanding and describing social and cultural scene from insider's perspectives
- Roots in anthropology
    - Studies of non-Western cultures
    - Attempt to develop deep understandings of unfamiliar civilisations
    - Local people as pursue daily lives in own communities
- Fieldwork
    - In general:
        - Dispassionate observer insufficient – engage directly with people in everyday lives
        - First-hand encounters to gain understanding
        - Deeply embedded perspective to get insights otherwise impossible
    - Being there, observe, ask insightful questions
    - Interviewing: "ethnographer's most important data gathering technique"

        - Explain and put into context everything seen/experienced
        - Study every word for subcultural connotations
    - Document everything seen/heard
        - Notepads, audio, video, photo, survey
            - Analysis at various stages – field notes, reports
    - Info gathered can be subjective and misleading
        - Cross-check, compare, triangulate before use as a basis of knowledge
    - Classical ethnographies might spend 6 months to 2 years on fieldwork or 2 weeks every few months
- Innovation by Chicago School
    - e.g., urban sociology
    - Local, maybe familiar settings
    - Still based on immersion in context/community/culture
        - Understand how people go about everyday business
        - How organised
        - Standards and norms
- Ethnographic perspectives
    - Focus on predictable, daily patterns of human thought and behaviour
    - Interpret observed behaviour in culturally relevant context
    - Allow multiple interpretations of data/reality
    - Open-minded approach – allow exploration of rich sources of data not mapped out in research design
    - Ethnographer – human instrument
        - Senses, thoughts, feelings; very sensitive and perceptive data gathering tool
    - Bias
        - All researchers have bias; make it explicit
        - e.g., choice of what to study is biased. Controlled can focus and limit research effort; uncontrolled can undermine research quality
- In HCI
    - Combination of observation, interviews, participation
    - Computer use as communication/collaboration
        - Use in existing groups (work, education), or purely virtual (forums, communities)
        - Norms and dynamics that might be important to study
    - How systems are used
        - How design affects the way they're used
    - Just understanding
        - Groups, communication, new technologies, etc
    - Different stages of design cycle
        - Early stages – to gain deep understanding of system requirements
        - Later stages – to gain deep understanding of how a product is being used (particular setting/group), so can redesign to better support users
        - Study combination of range of technologies in a particular setting
    - Ex: designing a new system in an unfamiliar domain
        - Need to understand system requirements
        - Can be rooted in context of how target users work and interact
            - Organisational concerns

            - Work practices
            - People's values
            - Types of interactions between people
        - Don't assume users are 'just like you'
        - Could use surveys/interviews instead?
            - Maybe – certainly easier and cheaper
            - If early stages and unfamiliar area, don't know what to ask
            - People's descriptions of what they do are often inaccurate
                - Poor at explaining
                - Misremember
                - Don't realise what they do
                - Bias (e.g., socially acceptable answers)
        - Site visits
            - Potentially for days/weeks
            - Observe
            - Interview
            - 'Shadow' them
            - As start to understand how they work and what they need, can begin listing requirements & designing
                - Discuss with potential users – for approval or to correct misperceptions
                - Try with different users, possibly in different setting
    - **Participant** Observation for Design Inspiration
        - Participate while observing
- In online communities
    - ex: Analysed collaborative play in WoW (https://dl.acm.org.citation.cfm?id=1180898)
        - Authors performed a lot of gameplay – active participants
        - Wanted to know what players were experiencing
        - Make recommendations to improve

Observation Techniques

- Observation
    - Passive observation of everyday activities without active participation/intervention
    - Maybe not integrating into any community – just watching public spaces
- Participant (participatory) Observation
    - Combines participation in the lives of those being studies with appropriate professional distance
    - Forms:
        - Complete participant
            - Become part of community as much as possible
            - May take years
            - Risk losing ability to be detached – "going native"
            - Covert observation – don't tell community you're a researcher. Ethically challenging
        - Complete observer
            - Don't interact directly

            - Could also be ethically problematic – not as much info gained/help given as possible
    - Can integrate quicker into 'own' culture – already an 'insider'
        - But if too familiar, can take events for granted and leave important data unrecorded

Ethnography Challenges

- Requires a lot of skills
    - Skills in conversation
    - Data interpretation
    - What to pay attention to
    - Whom to talk to
    - Reconcile contradictory data
- Expensive
    - Often used in 3 contexts:
        - Users not well understood
        - Tasks not well understood
        - Safety-critical systems

Interview Techniques

- Types:
    - Structured Interviews
        - Each participant answers same questions
        - Verbal approximation of questionnaire
        - Maybe appropriate when there's explicit research goals
    - Semi-structured Interviews
        - Each participant answers the same questions, but additional questions and follow-up questions can be added as needed
    - Unstructured Interviews
        - Interview may have little/no set structure
        - Could be tool for early evaluation, where there's no firm idea of specific research questions
- Designing
    - Types of questions:
        - Survey questions
            - Designed to elicit a broad picture of the participant's experience
            - Good for building rapport and establishing scope
        - Specific questions
            - Designed to gather feedback on specific categories, attributes, and themes
        - Open- and close-ended questions
            - Balance of structures and unstructured responses
    - Many issues (like with surveys)
        - e.g., recall bias – if asking about past behaviour, do it soon after
    - Interview (possibly) > survey
        - Probably longer open-ended answers
        - Can ask follow-up questions

        - Disadvantage: much more time-consuming
- Running
    - Respect for the context the interviewee is coming from
    - Respect for the interviewee
    - Strategies:
        - Be honest, be yourself
        - Focus on learning from participants
        - Be perceptive, know when to press and when to let go
        - Understand silence and use it
    - Being a good interviewer comes with experience

Key Actors

- In ethnographic setting, “some people are more articulate and culturally sensitive than others”
    - Some users respond better to given ideas, provide more useful feedback, and act as “star users”
- Balance star users/key actors with the dataset
- Over-reliance can be dangerous
    - Cross-check with others to ensure they’re providing reliable information

Ethics Checklist [see in lecture]

Analysing Qualitative Data

- Qualitative – interviews, focus groups, open-ended questionnaire responses
- Transcribe any audio data
- Familiarise yourself with all data
- Coding:
    - Deductive
        - *A priori* codes search for; clear pre-existing questions
    - Inductive
        - Find all the themes in the data
- Inductive approaches:
    - Thematic analysis
    - Grounded theory
        - No preconceived theories; open mind
        - Theory eventually ‘emerges’ from the process
- Qualitative Coding – loosely separated ‘stage ???’
    - Can verify with multiple coders at various stages
    - [example in lecture for VR study]

1. Open coding
    a. Identify distinct pieces of info; assign open code
    b. In-vivo coding: use participants’ own words to define codes
    c. Size/scope of pieces determined by researcher’s interpretive process
2. Axial coding
    a. Organise open codes into set of concepts/categories
    b. Think about relationships between concepts
    c. Don’t need to all be same level of specificity, or need even numbers of codes assigned

3. Selective coding
    a. In grounded theory, combing concepts into main theory
    b. Re-code original transcripts using new concept framework
- Reporting Results
    - If, e.g., a thematic analysis uncovers 5 main themes
        - 5 subsections, explaining each issue
        - “You can provide participant quotes” [p12]
            - Can relate to a user summary table – 1 row per user and info provided on age, level of experience, job, etc
    - Discuss overall findings
        - Put in context of related work – reinforce other findings, contradict, expand scope, consider different factors, etc
        - Might lead to implications for future designs

# Week 7

## Lecture: Analysis Techniques, Statistics

Analysing Data from User Studies

- Providing “descriptive statistics” is the bare minimum
    - Average, distribution, standard deviation
- Making claims, inferring causal relationships, in terms of a hypothesis test
    - Have you shown that your product is “better” than existing approaches?

Measurement Scales

- Ratio, Interval, Ordinal, Nominal
    - (sophisticated – crude)
- Nominal / categorical
    - Labels/names
        - Some numbers (without any possible computations), like random IDs
- Ordinal
    - Can put the values in a ranking, but not equally spaced
    - ex: ordered list of favourite films
    - Can do < or > comparisons, but not valid to calculate means
- Interval
    - Equal distances between adjacent values, but no absolute zero
    - Can compute mean
    - e.g., Celsius scale
        - Can take mean value of week’s temperature, but, e.g., 20°C is not “twice as hot” as 10°C
    - e.g., Liker scale
        - Sometimes treated as Ordinal. Important to know which if you want to compute means. Treating as Interval OK if options are equally spaced and centred at neutral value
- Ratio

- Do have absolute zero
- Support many calculations
  - add, divide, mean, standard deviation
- e.g., time, distance, counts of events

Evaluations and Measurements

- Before doing anything, need to plan well and measure the right dependent variable by collecting the right kind of data
- Types of Data
  - Think about data in terms of qualitative vs. quantitative
  - Think about quantitative data in a spectrum from continuous to discrete

Descriptive Statistics

- Measures of central tendency: Mean, Median, Mode
  - Mean – simple to calculate, but also provide little (or potentially misleading) information
    - Typically only useful if normally distributed data
  - Median – may differ significantly from the mean, can insight into the "shape" of the data
- Standard deviation describes the spread of the data
  - Estimate of average difference of values from the mean
  - Empirical Rule
    - 1 std from mean contains 68.2% of values
    - 2 std from mean contains 95.4% of values
    - 3 std from mean contains 99.7% of values
  - With human participants, the data is typically not normal distributed
  - Central Limit Theorem
    - As the sample size approaches infinity, the **distribution of sample means** will follow a normal distribution regardless of how parent population is distributed
    - Often said for sample size to be > 30 (even smaller for interval data)
    - Applies even to binary data (0 or 1 for completion of a task)
    - Implications
      - Many statistical hypothesis tests (e.g., t-test) assume normal distribution of data
        - If data is non-normally distributed (e.g., skewed), will these tests be invalid?
        - If sample size is large enough, CLT says that the distribution of sample means approximate a normal distribution
        - And so, we use these hypothesis tests!
- Standard Error (www.youtube.com/watch?v=A82brFpdr9g)
  - Example: weighing 5 mice
    - Perpendicular line – average (mean) of values measured
    - Parallel line – standard deviation on both sides of mean
      - Quantifies how much the data is spread out
    - Doing experiment 5 times in total, each time with 5 different mice
    - Each experiment has its own mean

        - **Standard error** – standard deviation of the means
    - Use multiple samples, not experiments
        - Estimate = sample std dev / sqrt(sample size)
- Plotting distributions tells you much more than simple values
- t-distribution
    - Can't know from experiments about distributions, means, std deviations of *population*, only sample
        - Student's t-distribution, t-scores rather than z-scores

Hypothesis Testing

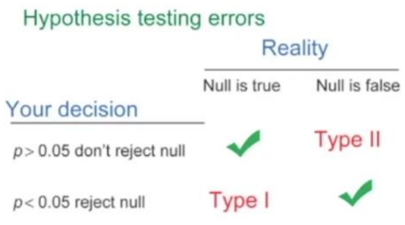
- Null Hypothesis – simple hypothesis against the intuitive hypothesis, e.g., mouse and trackpad are the same,
    - Rejecting the null Hypothesis
- Why is research done this way?
    - Very hard to prove something scientifically
    - Much easier to disprove
- Consider following statements:
    - Every software project has errors
    - Software projects never have errors
- Probably looking for sufficient evidence (instead of definitive proof)
- What are stats tests testing? **How likely is it that 2 samples come from the same distribution?**
- Also interested in:
    - How confident are we that they're different?
    - By how much are they different?
- Ex: comparing mouse to trackpad
    - Null hypothesis: There is no difference between user performance is using these 2 input devices for an object selection task
        - If we reject the null hypothesis, we can analyse the data to present results arguing for where differences occur and what gains this may have for interaction
- Types:
    - > 1 dependent variable
        - 1/2/more levels -> interval & normal one-way MANOVA
        - 2+ independent variables -> interval & normal multivariate multiple linear regression
    - 1 dependent variable
        - 2 independent variables
            - **Between** relationship between samples
                - Nature of DV:
                    - interval & normal -> t-test
                    - ordinal or interval -> Mann Whitney test
                    - categorical -> Chi-square test
            - **Within** relationship between samples
                - Nature of DV:
                    - interval & normal -> paired t-test
                    - ordinal or interval -> Wilcoxon-Signed Rank test

            - categorical -> McNemar test
    - >2 IVs
        - **Between** relationship between samples
            - Nature of DV:
                - interval & normal -> one-way repeated measures ANOVA
                - ordinal/interval -> Kruskal-Wallis
        - **Within** relationship between samples
            - Nature of DV:
                - interval & normal -> one-way ANOVA
                - ordinal/interval -> Freidman
                - categorical -> Chi-square test
- Specific:
    - t-test and paired t-test
        - Developed by chemist William Gosset working at Guinness in 1908, quantitatively measuring quality of beers
        - Assumptions:
            - Data follows a normal distribution
            - Data drawn from interval/ratio data
        - Can be completed on dependent (within subjects) datasets with paired t-test, or independent datasets
    - Friedman and Wilcoxon Tests
        - Friedman
            - Participants rate quality of n different wines
                - Null Hypothesis: There is no difference between the wines
        - Wilcoxon
            - Used for pairwise comparison, can provide results describing which wines are rated *significantly* better than others
            - Signed comparison: better/worse?
        - Tests for a difference in related samples (**within** subjects)
        - Used for ordinal data or interval data that is not normally distributed
    - Mann-Whitney and Kruskall-Wallis Tests
        - Kruskall-Wallis like Friedman but for independent samples (**between** subjects)
        - Mann-Whitney like Wilcoxon, but for independent samples
        - Ex: wines
            - Null hypothesis: Participants are unable to discern the difference between wines
            - Kruskall-Wallis test will say if there is variance across participants (e.g., by grouping participants by experience with wine tasting) and Mann-Whitney will provide pairwise comparisons to compare each group

How to Present Statistical Results

- Each test produces a *p* value (the probability that the samples come from the same distribution) and a test statistic
  - Can choose a target for p; often say $p<0.05$ means statistically significant
  - Test statistics are interpreted differently for each test
- Most tests would also be presented with an effect size
  - Ranges from 0 to 1 and describes how "visible" the effect is
- Reference: [www.statisticsdonewrong.com](www.statisticsdonewrong.com)

Errors in Statistical Testing

- Type 1: False Positive
- Type 2: False Negative
- 


# Week 8

## Lecture: Theories of HCI and Models of Interaction

Creating charts (practically)

- Tutorials on Moodle for Matplotlib and Seaborn
  - Matplotlib "tries to make easy things easy and hard things possible"
  - Seaborn tries to make a well-defined set of hard things easy
    - Not always 'well-defined' for your needs
- Plotly is an option (maybe too complicated)
- Often good strategy to browse galleries

### Theories of HCI

Reading: The Design of Everyday Things by Don Norman

Products should not need instruction manual

- e.g., Push/Pull on doors

Affordance

- Possible interactions between people and environment
- **The relationship** between physical object and person
  - Not a **property** of an object
  - Objects convey important info about how people could interact with them

    - Presence of affordance jointly determined by object's properties and person with capabilities that determine how it could be used
        - A chair affords sitting
        - A chair affords lifting to some people

Anti-Affordance

- Prevention of interaction
- e.g., glass (might make people think path is free and bump into glass)
- To be effective, affordance and anti-affordances must be discoverable
    - If it can't be perceived, need to signal its presence with a **signifier**
        - Signifier
            - Communicate behaviour
            - Image/text/sound/…
                - Make an affordance apparent
            - Deliberate
                - e.g., labels
            - Emergent
                - e.g., Paths worn onto ground
                - e.g., Queues of people

John H Williamson: Shoogle – Physical Affordances in a Digital Interaction

- Keys in a pocket. The user carries the phone in a pocket while walking. Motion from the gait of the user is sensed by the accelerometers. As messages arrives, objects begin jangling around in the user's pocket, in a manner similar to loose change/keys

Knowing what to do

- How can you make unfamiliar situations feel familiar?
    - Knowledge in the world
        - Perceived affordances, controls & their actions
    - Knowledge in our heads
        - Experience,
        - Conceptual models,
        - Constraints
            - Physical – rely on properties of the physical world
                - e.g., can only insert the correct way (USB-A, bank cards)
            - Cultural – rely on socially learned behaviours
                - e.g., Moodle relies on roles that make sense to us because we know how a course is run
            - Semantic – rely on intrinsic meaning
                - e.g., When added all items to buy, look for control for checkout screen
            - Logical – rely on trial and reasoning
                - e.g., An online form won't submit. Even if it doesn't highlight required fields, we can scan through and see if we left one empty – that one's probably the problem
            - Imposing these constraints prevents errors and guides users towards correct/desired/useful behaviour

- Guiding Interaction
    - Forcing Functions – Preventing action until certain requirements are met
        - Interlocks – Requiring actions to occur in sequence
            - e.g., web app doesn't offer functionality until logged in
        - Lock ins – Keeps an action active, preventing action from stopping
            - e.g., Gmail checks if an attachment is attached before sending email (if attachment mentioned)
        - Lock Outs – Prevents an action from occurring (typically in safety context)
            - e.g., Operators of x-ray machine cannot enter dangerous values, fire safety gate in front of basement entrance (for people not to go down in fire emergency)
    - Where?
        - ex: ATMs [in lecture]
    - Forcing Functions and Usability
        - Balance error prevention with frustration

Convention

- Design consistency is virtuous
    - Lessons learned from one system transfer to others
- If can't put knowledge in the design, put it into a cultural constraint
- Standardisation
    - Maybe a last resort; when no other solution seems possible, at least do everything the same way
- vs Progress
    - People don't like change
        - New learning is required
        - Which is 'better' design is irrelevant – the change is upsetting
        - Better to be consistent if new way is only slightly better than old?
        - If change to a new system, everyone has to change – mixed systems confusing
    - Standards simplify life, but can hinder future development

## Modelling Interaction

State machines can be used to model interaction

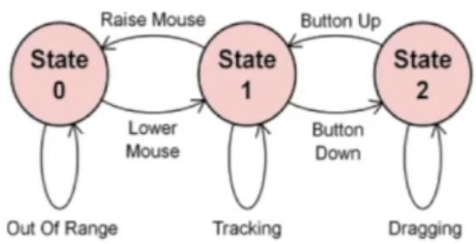
- 

FIGURE 7.9
Buxton's three-state model for graphical input.
(Adapted from Buxton, 1990)
    - (mouse, raising and dragging)

Fitts' Law

- Model to predict the speed of people's movements
    - One of few 'Laws' in HCI
    - Very widely used, proven to hold on many forms of interaction devices
- Paul Fitts, 1954
    - Psychologist, work predates HCI
- in HCI
    - Most used form adapted for HCI by Scott MacKenzie
    - Ease to acquire a target function of size of and distance to target
    - Equation:

$$MT = a + b \cdot log_2\left(\frac{D}{W} + 1\right)$$

        - MT – time to complete a movement
        - D – distance from start point to target
        - W – width of target (how accurate you need to be on arrival)
            - Measured on axis of motion
            - Pragmatically, often measured using the minimum of width, height
        - a & b – constants determined by cognition, hand-eye coordination, often different for different device type
            - Out of control in terms of designing on-screen positioning within interfaces
        - ID – index of **task difficulty** (in bits) (non-constant part – everything in log2, including)
        - $IP = \frac{ID}{MT}$. IP – index of performance/**throughput** (in bits per sec)
- Purposes
    - Many hundreds of studies have confirmed Fitts' Law holds with different devices, input methods (e.g., mouse vs trackpad vs touchscreen)
    - Can be used in:
        - Predicting movement time (if a and b are known)
        - Comparing 2 devices (by comparing their IP values)

- Guiding design choices
  - e.g., login button big and close to end of input, right click window shouldn't be too big
  - Easiest places to reach
    - Easiest place: where we are right now
      - Right-click menu – pop-up in place
    - Screen edge – can't overshoot, don't have to be accurate
      - Effectively a target of infinite width in a pointer-based interface
      - Corners are especially good
      - e.g., MacOS menus always bound to top
      - Have to decide when to make use of this behaviour, e.g., Windows X close button (don't always want to close, sometimes irreversible action)

# Week 9

## Lecture: Large-Scale Studies

A/B Testing (web experiments/…)

- Randomly split traffic among different app versions
  - A/Control: usually current live version
  - B/Treatment: new idea
- Collect metrics and analyse
- Any design has huge impact on conversion rates
- Examples:
  - Amazon – Shopping cart recommendations
    - Seemed unlikely, but wildly successful
  - Microsoft
    - MSN Real Estate
    - Office Online
- Experimentation > theory because data > intuition
- Ramp-Up
  - To detect an effect, you need to expose a certain number of users to the treatment (based on power calculations)
  - Fastest way to achieve exposure – run equal-probability variants
    - e.g., 50/50 for A/B
    - But that's too risky
  - Ramp-up – start low, do simple analyses, increase until equal
- Advantages
  - Tests for **causal** relationships, not just correlations
  - Reduces effect of external factors
    - e.g., history/seasonality impact both A and B in the same way
  - Overcome poor intuition, especially with novel ideas
    - Less data => stronger the opinions

        - Get data through experimentation
- Disadvantages
    - Organisation has to agree on OEC (Overall Evaluation Criterion)
        - Hard, but provides clear direction and project alignment
    - Quantitative metrics may not explain **why** a treatment is better/worse
        - => May not help designers solve problems/know where to go in next design iteration
    - Primacy effect
        - Changing app/site may degrade UX (temporarily) even if design is better
    - Multiple experiments
        - Statistical variance increases, making it harder to get statistically significant results
    - Consistency, contamination
        - Assignment to A/B usually cookie-based, but people may use multiple machines/erase cookies
    - Hard to do proper randomisation

Large-Scale Mobile HCI Studies

- Mobile HCI studies in many forms:
    - e.g., text entry, gestures, AR, usage studies, privacy
- Quantitative analysis
    - e.g., time taken to complete task/error rates
- Qualitative analysis
    - e.g., interviews, ethnography, opinions of experience
- Into the wild
    - Early/'traditional' experiments all done in lab
        - Easy to observe, control, eliminate confounding variables
        - Possibly unrepresentative of technology's eventual intended context of use
    - More recent studies performed in more realistic settings
    - Forms:
        - Direct observation
            - Videos, interviews
        - Often still using evaluator-supplied hardware
    - Challenge of Space & Duration Trials
        - The longer and requiring more space, the more difficult to exercise experimental control
- Research via app stores
    - Put software to study on app stores
    - Benefits:
        - Participants using own devices
            - Already experts with hardware => no training
            - No extra device to carry, already with them always
            - No fixed end date
        - Potentially very large number of users, globally
            - Chosen to use app => more representative (?)
    - Drawbacks:
        - Don't meet users

    - Can’t directly observe users
    - Qualitative data?
    - Internal vs external validity
    - Additional ethical challenges?
  - Ex – Hungry Yoshi
    - Game using Wi-Fi access points as game resource
    - Investigating use of app stores in running mobile HCI trials
    - > 300,000 downloads
    - Global user base
      - Only have locations from those users who agree to supply it
    - Data logging
      - Recorded (“logged”) ingo on use while apps are running
    - Data Visualisations
    - Qualitative evaluation
      - Questionnaires
        - Answers with radio buttons/typed
        - Tasks, like become FB friend
        - 19% responded
      - Server-side, so instant updates
      - Paid telephone interviews
  - Ex – Hit It!
    - Android game: touch objects on screen
    - Found that touch positions are skewed
    - Could create function that shifts touch input to compensate
    - Updated game to use compensation
      - Error reduced by 7.8%

Large-Scale Trials: Difficulties

- Verification of user info
  - Are people telling truth?
    - Age, gender, opinions, etc
- Trial software installed on large variety of devices
  - Android
  - OS, CPU, screen sizes, etc
  - Potential confounding variables
- Collecting qualitative data is difficult
  - Very short questionnaire answers
  - Solution: Phone / online calls?
    - Most users probably don’t speak language
    - Time zones
- Mass of quantitative data; harder to get qualitative
  - **What**, not **why**

Potential solution: Hybrid Methodology

- Hybrid approach: combining ‘mass participation’ and local deployments
- Concurrent large-scale and small-scale studies
- App released to general public and local users recruited via poster adverts

- Some aspects of trial best suited to each group
- More solid ethical practice
- Ex: Predictor
  - World Cup Predictor app
  - Released 1 week before football World Cup
    - => 11 locally-recruited users
    - => 10,806 through app store
- Benefits
  - Use the Small to Explain the Large
  - Use the Large to Verify the Small
    - If a system is trialled among small group of locally recruited participants,
      - Do results generalise to population at large?
      - 'Outlier' users
        - Are there users showing unusual behaviour?
        - They could skew results of study
    - Experimenter effects
      - Subtle conscious/unconscious cues an experimenter gives participants
      - Could affect users' performance in the trial
      - Less likely in large-scale trial?
        - User interaction with evaluators generally far lower
    - Ex: looking at 1 feature of app – head-to-head challenge
      - Local users
        - Head-to-head uptake: 73%
        - Average number ofH2Hs by those using: 5.2
      - Global users:
        - Head-to-head uptake: 0.8%
        - Average number of H2Hs by those using: 1.5
      - Local trial alone would have led to misleading results

Ethical Challenges of Large-Scale Trials

- Capturing a lot of info on people
- Never meet participants
- **Informed consent**
- No opportunity to **debrief** participants
  - Can't tell the last time a user will launch the app
- Solution: Terms & Conditions page
  - Page often shown on first launch
  - Provides info on experiment
    - About authors
    - About study
    - About info captured and reasons
  - Often have to be explicitly agreed to before using app
  - Opt-out mechanisms
  - Multiple languages

Hybrid Approach: Levels of Engagement

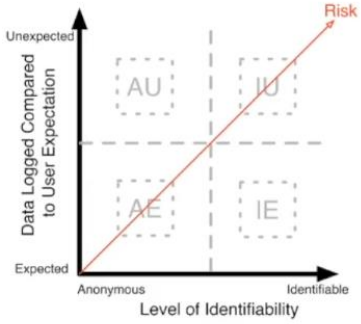
- Problem:
    - All participants agree to same T&Cs
    - But difference in confidence with which researchers have gained informed consent
        - Ease of deception
        - Inability to debrief
- Solution:
    - Framework of levels of engagement
        - Local users: studied in detail
        - Remote users: looked at aggregate data
        - Types of questions asked
            - Ability to converse sensitively at a distance
    - Compromise: getting useful info, but not exploiting users as a resource just because they tick T&Cs

Ethical Challenges in Detail

- Informed Consent
    - Do people know what we're doing?
        - That the app is part of University research?
        - The purpose of experiment?
        - What info is being recorded?
        - What we do with this info?
        - How to opt out?
    - Solution: T&Cs
        - State purpose of study – URL to project site
        - All logging explained and must be explicitly agreed to before app usage
        - Store/transmit data securely
        - Email address opt out at any time on request – have all collected data destroyed
        - Multiple languages
        - Disadvantages:
            - No one reads it
                - ex: Hungry Yoshi studies
                    - Did you read T&Cs?
                        - In-app questionnaire: out of 1226 responses, yes was 20%, no was 80%
                        - Telephone interviews: out of 11 responses, yes was 0, no was 11
                    - Opening
                        - 2% opened
                        - None spend >60 secs reading the 842 words

Researching Ethics

- Interpreting existing guidelines to cover large-scale mobile HCI
    - Human trials in Psychology: BPS & APA
        - Autonomy, Dignity, Self-Determination
        - Concern for Others' Welfare
        - Social Responsibility

        - Scientific Value, Integrity, Competence
    - Internet-Mediated Research
    - General Guidelines
        - Restrict age of users where stores allow
            - Graphics, icon sets, descriptive language
        - T&Cs in store description AND in-app
        - Historic log data not on externally visible server
        - Privacy-preserving data publishing techniques
- Framework categorising trials based on participant 'risk'
    - 2 dimensions of participant 'risk':
        - Anonymous vs identifiable
        - User expectation of app's data access
    - 


      4 Quadrants

      **AE**: Anonymous, Expected

      **AU**: Unexpected, anonymous

      **IE**: Identifiable, expected

      **IU**: Identifiable, unexpected
        - AE
            - e.g., aggregate download/usage stats
            - e.g., logging data that is integral to app usage, but cannot be used to identify user
            - Generally low risk
            - Advice:
                - General guidelines sufficient (T&Cs, etc)
        - AU
            - e.g., a game looking at 'unnecessary' data: how many contacts, contents of media in library
            - Advice:
                - Pop-ups to gain explicit consent for each new data type captured
                    - Mobile Oss now incorporate this
        - IE
            - e.g., location-sharing apps, social media apps
            - Advice:
                - Provide functionality to browse data and delete specific parts
                - Effectively allowing 'opt out' at any time
        - IU
            - e.g., a game looking at 'unnecessary' data that could identify user, e.g., location
            - Highest risk
            - Advice:

        - Actively interrupt users to show them examples of recorded data
    - Interruptions
        - Alternative to T&Cs (since they're never read)
        - Visual representation of log data
        - Delayed presentation of info
        - Personalised with user's own data
        - User Study: Yoshi
            - 1007 users; between-groups design
                - Hash function on device's unique ID to randomly assign to a condition
                - Some shown map, some shown text only
            - Further Results
                - More concerned users stopped using the app around twice as quickly
                - Difference of showing the map more pronounced for non-English speakers
                - Also looked at age, gender: no significant differences
            - Discussion
                - Look beyond current common practice of T&Cs
                - Majority of users seem relaxed
                - Small number of concerned users, who we should be going further to support
                    - Personalised visual representations of data
                        - More users reported concern
                        - Stopped using the app sooner
- Advice for how to run each type of trial in ethically sound manner
- Experiments on new ethical procedures

Discussion beyond Yoshi study:

- Can be extended to many forms of data
- Collect data only locally on device for a short period at start
- Interrupt user with visual depiction of their own data
    - If they agree to participation, upload all collected so far and keep logging
    - If they disagree, destroy collected data without it ever leaving the device
- Should be more engagement of users generally
    - Ethics as active area of research (not just box to tick)

# Week 10

## Lecture: Information Visualisation

Definition

- To visualise
    - "To form a mental image/vision of …"
        - Not just immediate perception, but fitting what's seen and interacted with into a mental model… and so updating that model
- Information Visualisation
    - "The use of computer-supported, interactive, visual representations of data to amplify cognition"
        - Reading in Information Visualization: Using ???

InfoVis in general

- Forming mental model to gain insight
- 'Offloading' cognition
    - Reduce load on working memory
    - Using recognition rather than recall
    - e.g., analogous to long multiplication in head
        - Much easier if you can write notes (workings)
- What it isn't:
    - Scientific visualisation and cartography
        - Usually physical data about objects & spaces
        - Based on inherent/'natural' dimensions
- About abstract data
    - How best to present a data set?
        - Type of data?
            - Column types – Numerical? Ordinal? Dates?
        - Who's analysing it?
        - What are they looking for?
        - Who's looking at it?
    - Correlations, clusters, outliers
- vs Information Retrieval
    - IR: Absolutes
        - Maximum, average, exact query match
        - Formalised, suited for a command language
    - Info Visualisation: Relatives
        - Overview, trends, patterns
        - Distributions and outliers, 'sense'
        - Difficult to formalise
        - Suited for interaction, browsing and exploration
            - Built up over time via successive interactions
- Been about for 30 years

Examples

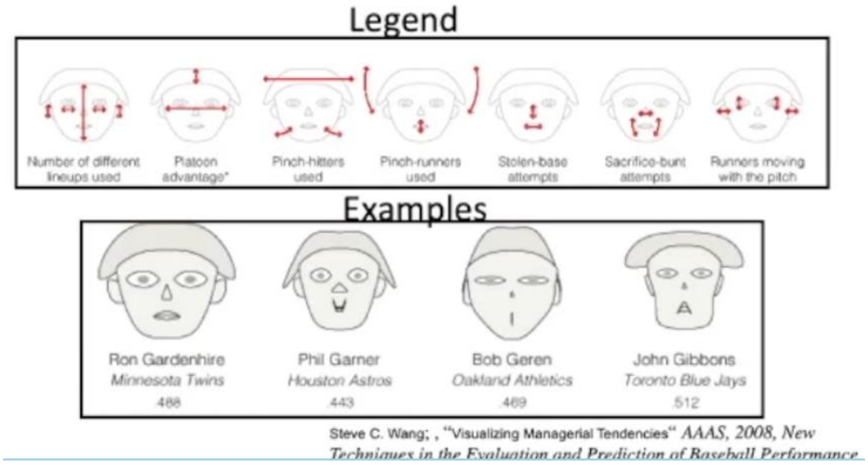
- Earlier – e.g., London underground
    - Harry Beck, 1933
    - 'Circuit board' design
    - Abstraction
        - Aid clarity
    - Still used today

Key Principles

- Abstraction
    - Replace many objects by one representative object
- Start with overview -> Support zooming & filtering -> Only then show details on demand
- Direct Manipulation
    - Objects output on screen take input too
- Dynamic Queries
    - e.g., move a slider up and down, linked graph changes too
- Immediate Feedback
    - GUI interaction triggers response straight away
- Linking & Brushing
    - Views linked so that selections match in all
- Focus & Context
    - Show key objects in detail, but in the setting of the wider data set
- Animate transitions and change of focus
    - Don't jump so harshly that context is lost
- Output is input
    - Anything one can use to show data can be used to select data too
- Colour with care
    - Be aware of colour blindness, (non)linear perception, visual overload

Representation

- Data encoded by:
    - Location
        - Spatial location on display conveys value
        - e.g., X-Y plots
            - e.g., scatterplot
        - Can encode 2/3 variables this way
    - Size
        - Size of points represent value
        - Can run out of room very quickly
            - Occlusion: big points hide smaller ones
        - Negative values?
    - Colour
        - Colour Scales: many to choose from
        - Careful: RGB is a non-linear colour system
            - e.g., 100/100/100 is not twice as bright as 50/50/50
            - Stick to a small simple palette: use highlight colours cautiously
            - Use a perceptually linear colour scale
        - Minimum size at which visible

- Perceptually linear colour scales
  - Arrays of RGB values scaled to better fit with average human perception
  - Colour at index 100 generally perceived as 2x as bright as at 50
  - More limited range: may have to avoid the many dark values at low indices
- Shape
  - 'Glyphs' are used to visually represent multiple dimensions of data by combining them into a single pictorial representation
  - Ex (most famous): Chernoff Faces
    - 

  - Usually need the legend to understand them
- Texture
  - Easy to tell difference, e.g., Tweed & Silk
  - The finer the texture, the closer we have to be to the graph to understand it

- Ranking Visual Attributes
  - -> Increased accuracy for quantitative data (1984) ->
    - Colour -> Size -> Angle, Slope -> Length -> Position
  - Guideline:
    - Map more important data attributes to more accurate visual attributes

Focus & Context (Principle)

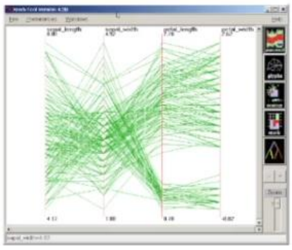
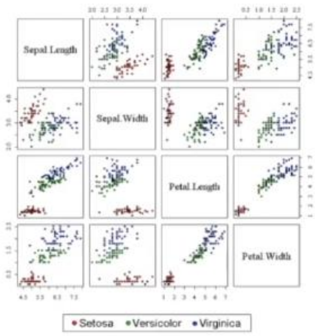
- Show detail as well as 'big picture'
- e.g., Maximise usage of available screen real estate
  - Overview & detail
    - Area of detail and (usually smaller in screen size) overview covering larger area of data
    - Separate views (big map and mini map in corner showing big map in context)
      - Can often interact with both
  - Distortions
    - e.g., blurring, fisheye
    - Single view of the data
    - Focus in high-detail, surroundings much less
    - Normal vision involves perspective (things further away gradually get smaller)

            - Fisheyes exaggerate the same effect
            - But still use smoothly increasing distortion
        - Example metric:
            - DOI (b|a) = API(b) – D(a,b)
                - DOI (b|a) – degree of interest in point b, given current focus a
                - API(b) – a priori importance of b
                - D(a,b) – distance from a to b
            - 'Information suppression' function
            - General idea applicable in 1D, 2D, etc

Sheiderman's Taxonomy:

- 7 types of data
    - 1D data
        - e.g., single column of numbers/text
        - List with scroll bar
            - People often only look at top of list
            - Can't easily see/move further down
            - But how to explore a long list/column?
        - InfoVis techniques
            - Distortions
        - Edit Wear and Read Wear
            - 'Wear marks' show pattern of where file has been most used
            - Rectangles can show where each person is currently working
            - Showing accumulated history of use
            - Worn by use, like well-thumbed book pages, paths in grass, old stone steps
            - Can extend well-known representations
                - e.g., fit into simple scroll bar
            - or fit into newer designs
        - Experiment applying fisheye effect to 1D data
            - Fisheye list faster to use than traditional for drag & drop tasks
            - No difference for selection tasks
            - Error rates same
            - Users preferred fisheye
    - Temporal Data
        - Has time attribute
        - Very common: records, logs, databases
        - Can 'stack' dimensions, sharing time axis
        - Can use 1D techniques (e.g., fisheye, distortion)
        - Ex: Ebb and Flow of Movies (in nytimes)
        - Ex: timeline slider
            - 'playback' or can query specific times
            - See pattern across country
            - But: hard to compare temporally distant data
            - No overview over time
                - Pattern detection more difficult

        - See everything – external cognition
        - Memory – internal cognition
- 2D Data
    - Scatterplots – plot x vs y
    - Ex: maps (geographical data)
    - Techniques: fisheye, focus & context
- 3D Data
    - Appeal to the '3D is natural' idea
    - Often think of the world as a 3D shape
        - But don't treat it that way
        - e.g., how wide is a city? How high?
        - e.g., can only see surfaces of most objects
    - Often invites occlusion problems
        - Nearby objects block distant ones
        - e.g., can only see one half of a sphere
    - 2D vs 3D: easy to use vs aesthetically pleasing
- Hierarchical Data
    - Trees difficult to handle
    - Basic problem – fan-out to many objects (can't show all tree in detail at the same time)
    - Hard to show many objects and lots of structure at the same time
    - Can't avoid having to move around and explore
    - Focus & context, fisheye
        - e.g., hyperbolic tree
    - Debate over glitz vs utility
    - Experiments and design continue
    - Alternative: **tree maps**: convert tree to **rectangles**
        - Area proportional to, e.g., node size
        - Split space horizontally and vertically in turn
        - ex: SpaceSniffer, WinDirStat, Disk Inventory X
- Graph Data
    - Nodes and edges
    - Aesthetics. 'Appealing' layout
        - Subjective?
        - Generally accepted desirable properties:
            - Minimise edge crossings
            - Uniform edge lengths
            - Evenly spaced nodes
            - Symmetry
    - Even more difficult to handle than trees
        - Links can go anywhere: may be no regular order/structure
    - Optimisation algorithm
        - e.g., find positions that minimise edge length & crossings
        - Closely related to algorithms for multidimensional data
    - Sometimes better to make simpler
        - Reduce to simpler type, e.g., tree: choose root, lift up, cut off/hide excess links

- Multidimensional Data
  - Cleveland and McGill: humans best equipped to make judgements when data is encoded by position
  - Strategies for visualising:
    - Non-orthogonal display of dimensions, e.g., Parallel Coordinates
      - 
      - Each object a single polygonal line
        - Intersects each 'axis' at appropriate value
      - See patterns, clusters, etc
        - 'Iris' data set: 150 objects, 3 natural clusters
      - Good for correlations, if adjacent
        - Might need to rearrange dimensions
      - Hard to follow a single object's line left to right
        - Worse with bigger datasets
        - Interactive controls can help
          - Mouse-over to highlight a single line
    - Numerous Paired Combinations, e.g., Scatterplot Matrix
      - 

      - x-y scatterplots of every pair of dimensions
      - Good for seeing correlations in pairs of dimensions
        - Position irrelevant
      - Duplication in grid: can just show 'triangle' either side of diagonal
      - Screen space requirement rises quadratically with dimensionality
        - $\frac{d^2-d}{2}$ plots
        - d – dimensions
      - No overview of all the data
      - Interaction can make more powerful
        - Brushing and Linking
          - Linking together multiple views, so that 'brushing' a selection in one view colours matching objects in other views

- Dimensional reduction to create single scatterplot, e.g., Force-Directed Placement
    - Single Plot Visualisation
        - Create single scatterplot showing overall structure of data
        - Compare objects: rows of the spreadsheet
            - Treat inter-object similarity as high-dimensional distance
            - Find a low-dimensional layout that retains as much of the relative distances between objects as possible
                - Similar objects close together in the layout, and dissimilar objects far apart
            - General approach often called **Dimensional Reduction / Multidimensional Scaling (MDS)**
                - Matrix methods (e.g., PCA) spring models
        - 'Reduce dimensionality' – to 2D/3D
        - Force-based models
            - 'Spring model' to position objects
            - Consider a spring between each pair of objects
                - Ideal relaxed length of spring proportional to difference between objects
                - i.e., if A & B are similar, C more different: AB short, BC long, AC long

                1. Start from random positions (some springs too stretched, others too squashed)
                2. The springs then iteratively push and pull objects until the layout reaches equilibrium
            - Strengths:
                - Scatterplot layout positions show global relationships
                    - Neighbours on layout are usually high-D neighbours
            - Weaknesses:
                - All dimensions are combined in 2D layout
                    - Not for exploring individual dimensions, unlike other techniques

    - Can be very slow – often O(n^3) overall
        - May be unable to lay out large/complex datasets, e.g., many millions
- Ex: JavaScript library for visualisation on the Web
    - HTML, SVG, DOM manipulation
- Ex: https://bl.ocks.org
- Ex: bservable – Jupyter-style notebooks