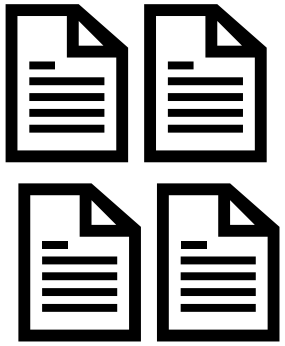
Qualitative analysis: Making sense of coded data

CS798H Semester-II (2023-24)

Last time: Qualitative coding

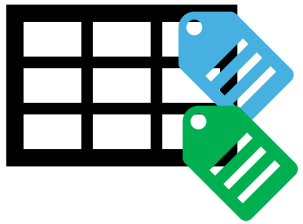


Unstructured, qualitative data

(High variance, same thing said in variety of ways)

+

=



Structured, coded (quantifiable) data (Low variance, we have all codes)

ANALYSIS

HOW??

Results, insights, answers to questions

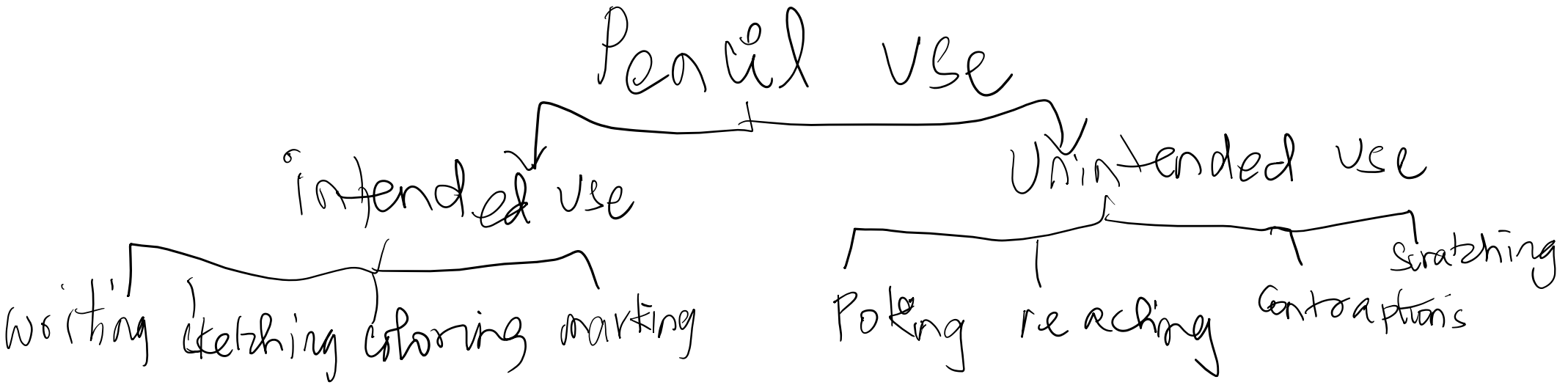
This is still data!

Analyzing coded data: Raw codes

- Example-1: What do users use pencils for?
- Codes: Writing, sketching, colouring, poking, reaching far off objects, marking, contraptions
- Answer to question is: simply, the list of codes.
- This itself is useful, but is less informative!

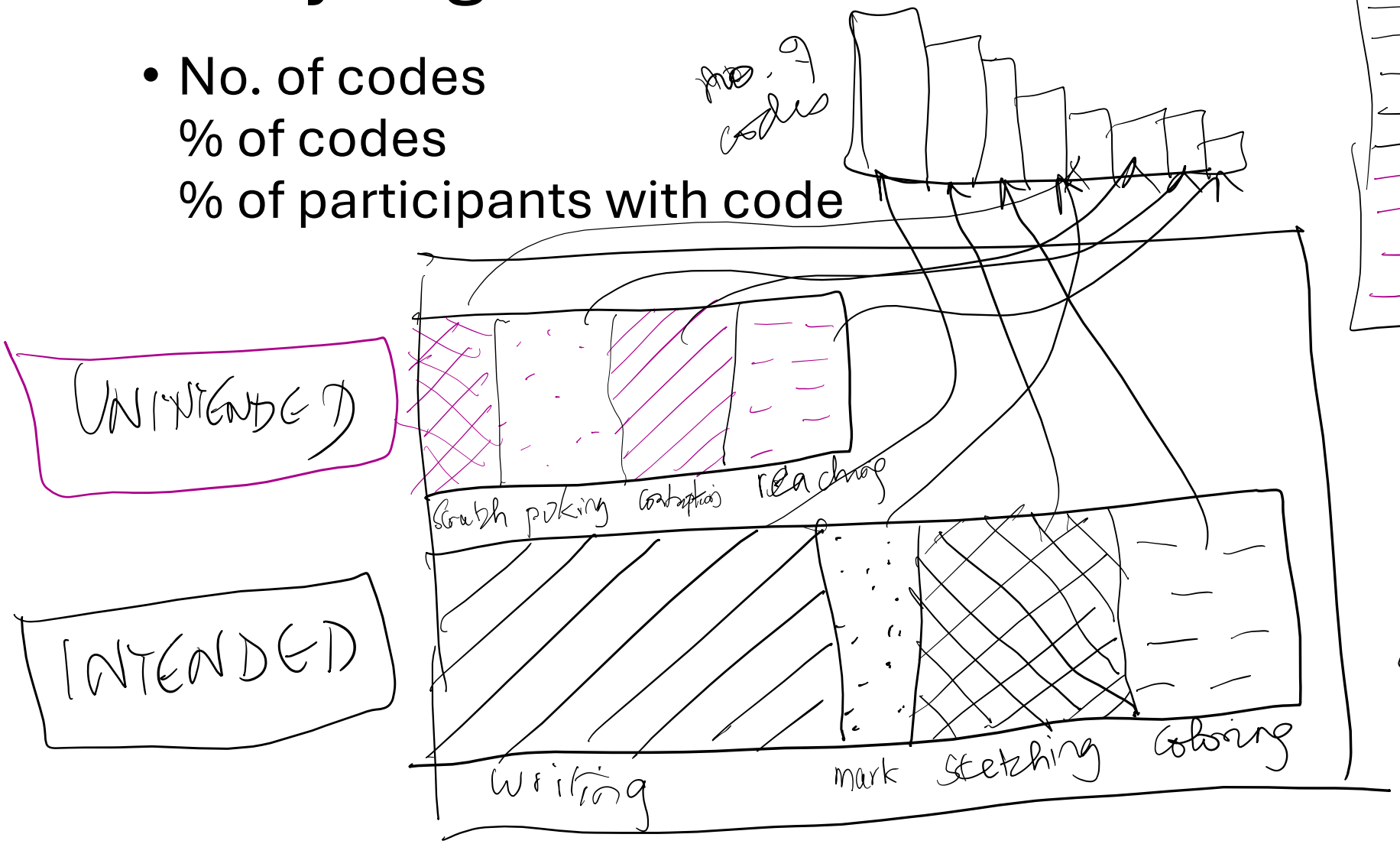
Analyzing coded data: Classification

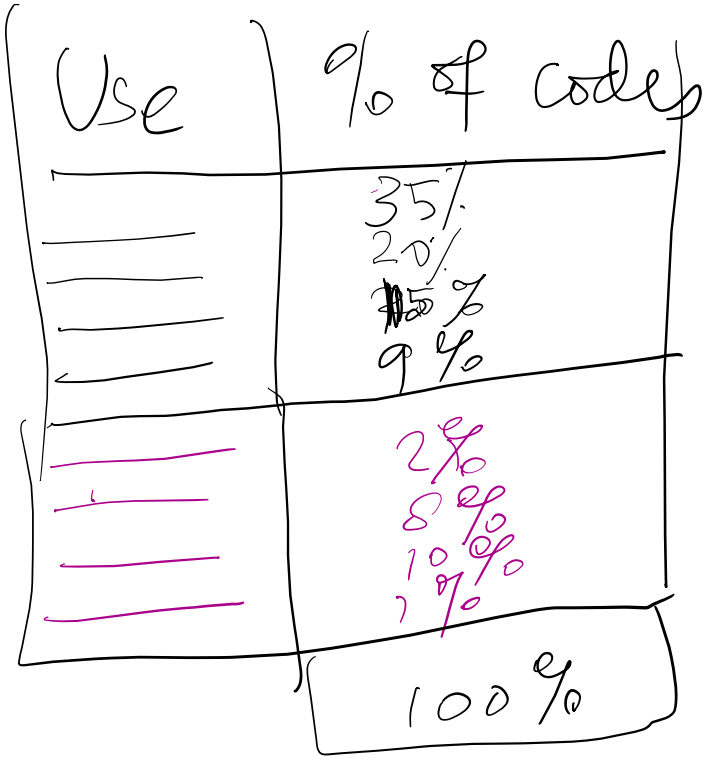
- Simply list of codes are often boring!
- Try: classification



Analyzing coded data: Numbers!

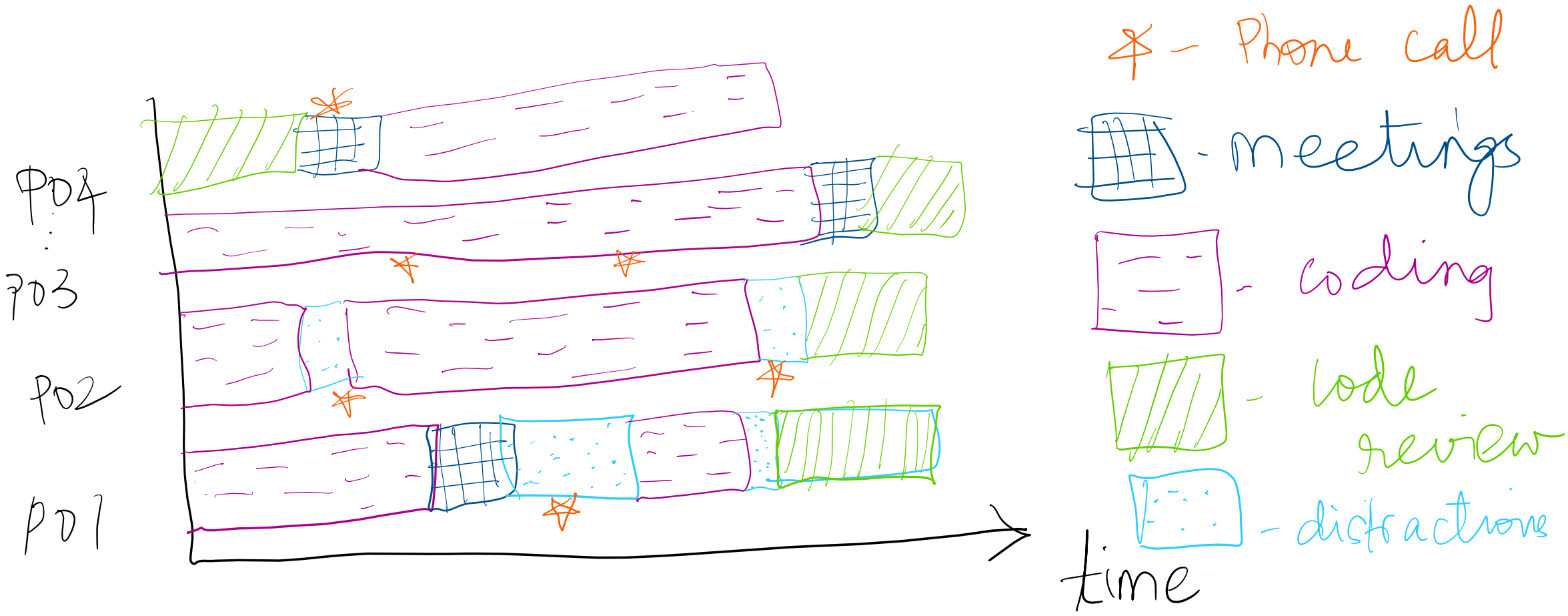
- No. of codes
  % of codes
  % of participants with code



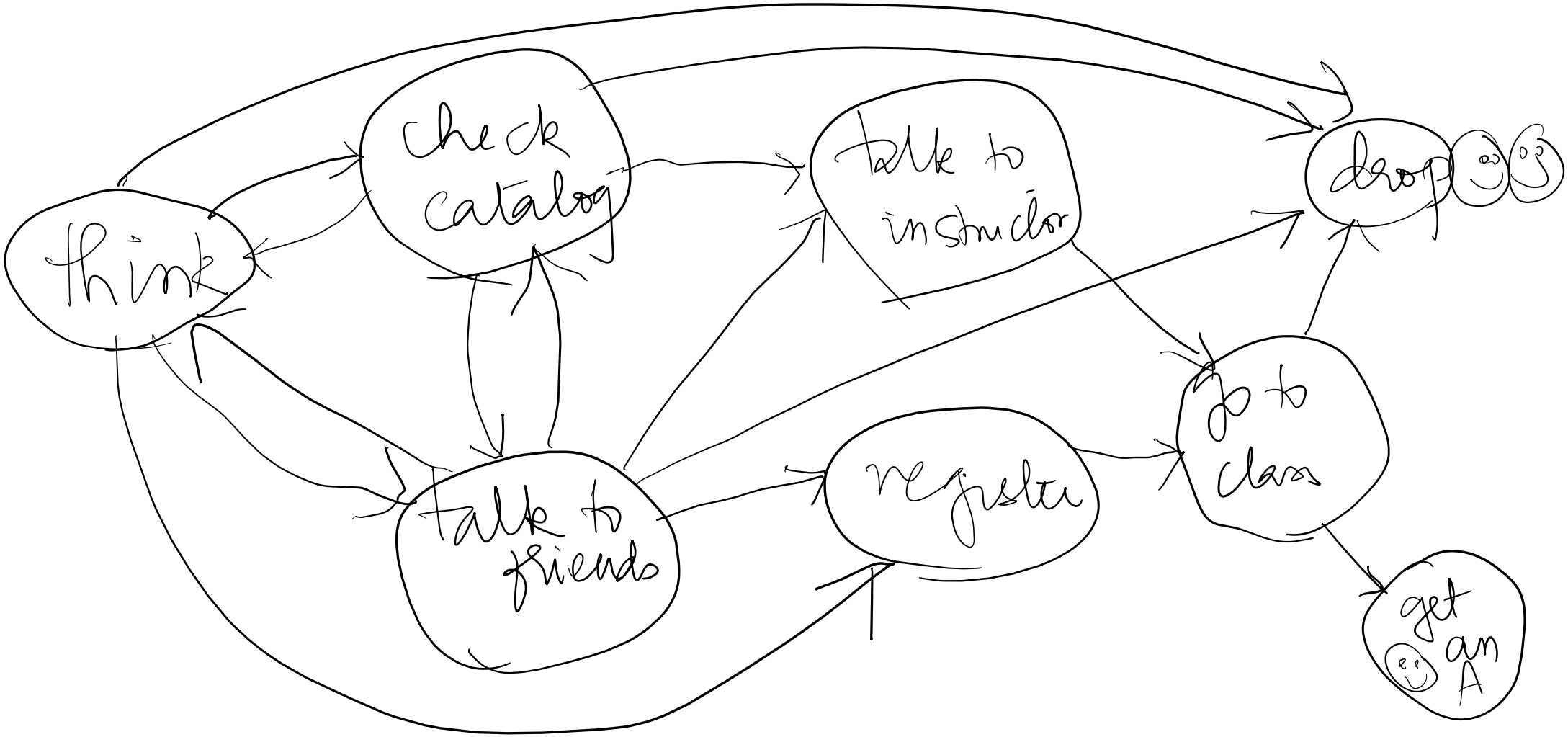


Several ways of viz with or without classification

Patterns / orders

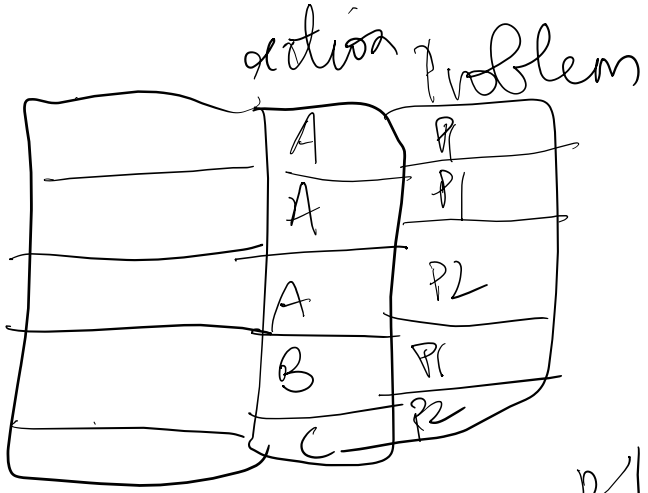


Flow of events



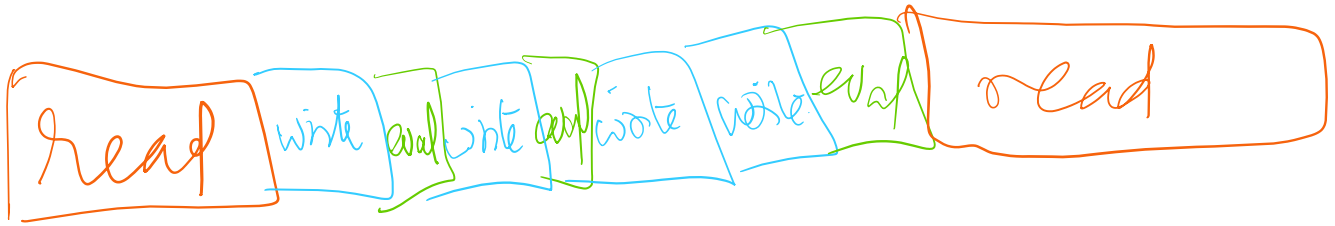
Others:

- Code co-occurrences
- Sequence of actions



P1 happened during A & B

P1 Co-occurred with A&B



write - eval repeats more than read - write - eval

Chain of evidence

- Data + analysis = facts / results

(e.g., 80% of codes are A, A happens together with B and C, etc.)

- Facts + interpretation (happens in head!) = claims / hypotheses / beliefs
- E.g., One reason 80% of codes are A is because A is most frequent; another possibility is that B happened, but manifested as A because…

Alternative interpretations

- Claims + evidence = proposition / conclusions
- Claims without evidence = unsubstantiated claims

Every conclusion should be grounded in evidence, and evidence should be easily traceable. This is called the "chain of evidence"

Evidence vs. proof

- Proofs – definitive, conclusive, “proves”, “confirms” a statement (as in math)
- Evidence – not conclusive but “supports”, “substantiates”, or “shows” a statement or an assertion (as in law, history)
  - You can always find more evidence that shows this incorrect, at a later time!
  - In some fields, you can have a lot of evidence for something, and no evidence contradicting, and so you can call it proven (e.g., proven guilty!)

- In empirical work (as in HCI), we never say “prove”, we always use “suggest/ show/indicate”.

What counts as evidence in HCI?

- Literature (other prior studies)
- Facts from current study (figures, codes, results of statistical comparisons)
- Facts + logical reasoning
- Participant utterances (e.g., P01 said “Pingala is frustrating!”)
- Occasionally, theory that has been proven correct (through various sources of evidences, hasn’t been falsified, etc).

Good hygiene when drawing conclusions

- Look for multiple evidences indicating the same thing
  - “A said so” vs. “70% of participants mentioned so”.
  - “70% of participants said so” vs. “70% of participants said so, as did 80% of participants in X’s study”.
  - “60% of participants said so in interviews”, and was confirmed by another 40% of participants in survey.
  - Participant said so, and on prodding further provided further reasoning for why he said so.

Good hygiene when drawing conclusions

Look for alternative interpretations / hypotheses

- This could happen due to one of three reasons, A, B and C; we are yet to find why!
- This could happen due to one of three reasons, A, B and C; but we only found evidence of A in the survey. This suggests A is perhaps the most common reason for something.
- This could happen due to one of three reasons, A, B and C; but we only found evidence of A in the survey. One possibility is that our participants haven't been exposed to B and C, as some participants mentioned "what is that, never heard before".
- One reason for X is A and another reason is B. To test which of this is correct, we conducted PCA, and found that A was a moderator whereas B had no effect. This suggests that A influences X in Y way.

- Seek further data / evidence in favor of each alternative hypothesis

Chain of evidence: Good habits

- Never make a statement, without grounding in evidence
- Never make a statement with “adjectives”, use concrete values (most participants → 80% of participants)
- Do not over-generalize
  - Say “participants in our studies”, not “users of X”. (Remember, we didn’t do all users!)

Threats to validity

- We draw a conclusion from studies: what are potential reasons we draw an incorrect conclusion?

- Four broad kinds:
  - Construct validity: Is our measure of a “construct” right?
  - Statistical validity: Are the statistical tests right for the data we have?
  - Internal validity: Are there confounds? (Is any cause and effect relationship potentially influenced by other factors?)
  - External validity (or generalizability): Will the results generalize to other tasks/situations/populations outside the study sample?

Example: threats to construct validity

- Question: How usable is system A?
  - Usability can be operationalized in a variety of ways
  - Usability = score on a scale of 1-10 as rated by participants.
  - Usability = score on all scales of system usability survey
  - Usability = no. of errors and frustrations verbalized by participant

- Effectiveness of tool = clock-time to complete a task
  - Effectiveness is operationalized as clock time (end time – start time)
  - Problem 1: With talk aloud; could be slower
  - Problem 2: What about interruptions?
  - Problem 3: What about correctness of task?

- Others: biases in questions

- What to do?
  - Look for standard tests
  - Look for alternative measures, and see they confirm
  - Never make leaps (look for all dimensions, and be specific). E.g., instead of using clock-time as a measure of effectiveness, simply say task times. Or, say, effectiveness is a measure of task time, correctness and no. of tasks attempted=and use all three.

Examples: threats to internal validity (Includes all bias / confounds)

- Question: How usable is system A?
  - Unusually low internet stability on the day of study?
  - Think aloud increases distraction and therefore affects usability

- Task completion faster with New version compared to old version.
  - What if learning for first time, so slow?
  - What if participants were stressed because of time constraint?
  - What if participants were nervous because of recording?
  - What if task for old was easier than task for New.
  - What if all participants given New were more experienced than those who were given Old.

- What to do?
  - Additional checks to ensure factors like experience, task difficulty are comparable
  - Where control not possible, use the same = give 10 min in both new and old, so pressure effect cancels out.
  - Balance: some people do task 1 with new, some do task 1 with old and likewise for task 2.
  - Randomize: Assign participants randomly / deliberately to balance (as needed!)
  - Add dimensions: use 3 tasks instead of 1, to reduce effects of chance

Examples: threats to external validity (Includes all sampling-related issues)

- Consider if results are generalizable to:
  - Other tasks
  - Other participants in same population
  - Other populations
  - Other tools in category
- Self-selection bias

- How?
  - Triangulate with other methods (e.g., survey, diaries, etc.)
  - Replication (same study with other population / task / etc.)
  - Use multiple tasks (e.g., instead of one task, use 3 tasks!)
  - Report demographics!

Examples: threats to statistical validity (Includes all data-related issues)

- Later, but…
  - Sample size suitable for conclusions
  - Some tests make assumptions on data distribution (e.g., normality)
  - Some tests want pure random data, but recruitment not entirely so!

- Pick right tests, triangulate, report everything!!!