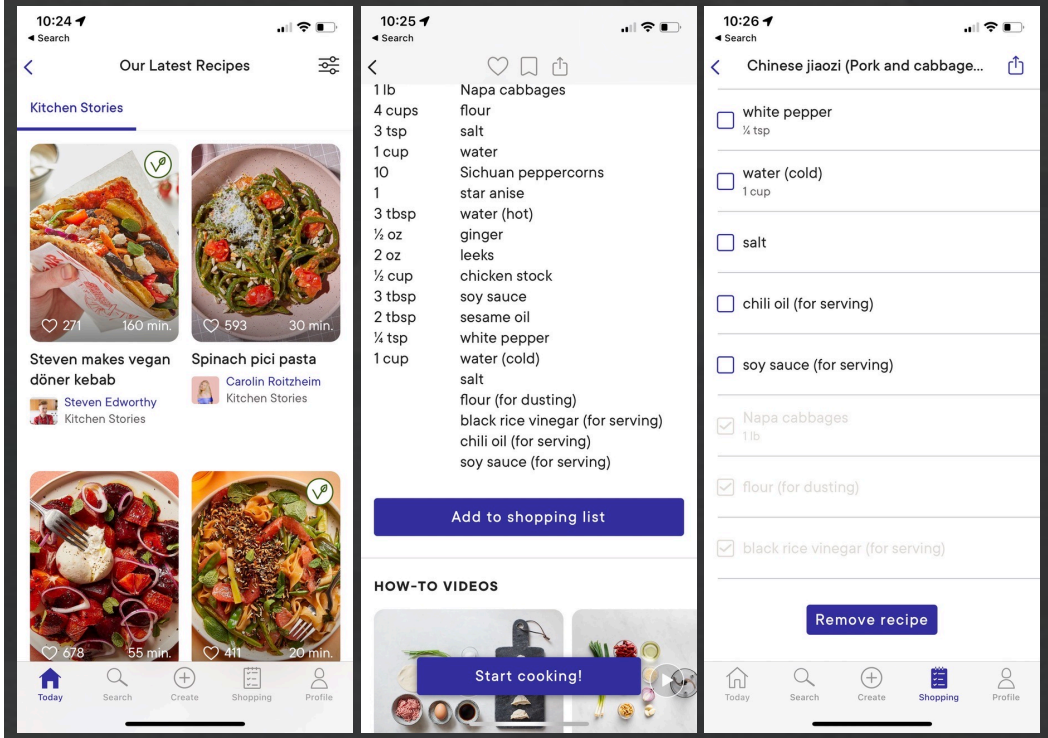
Usability Testing

刘哲明
Prof. James A. Landay
Computer Science Department
Stanford University

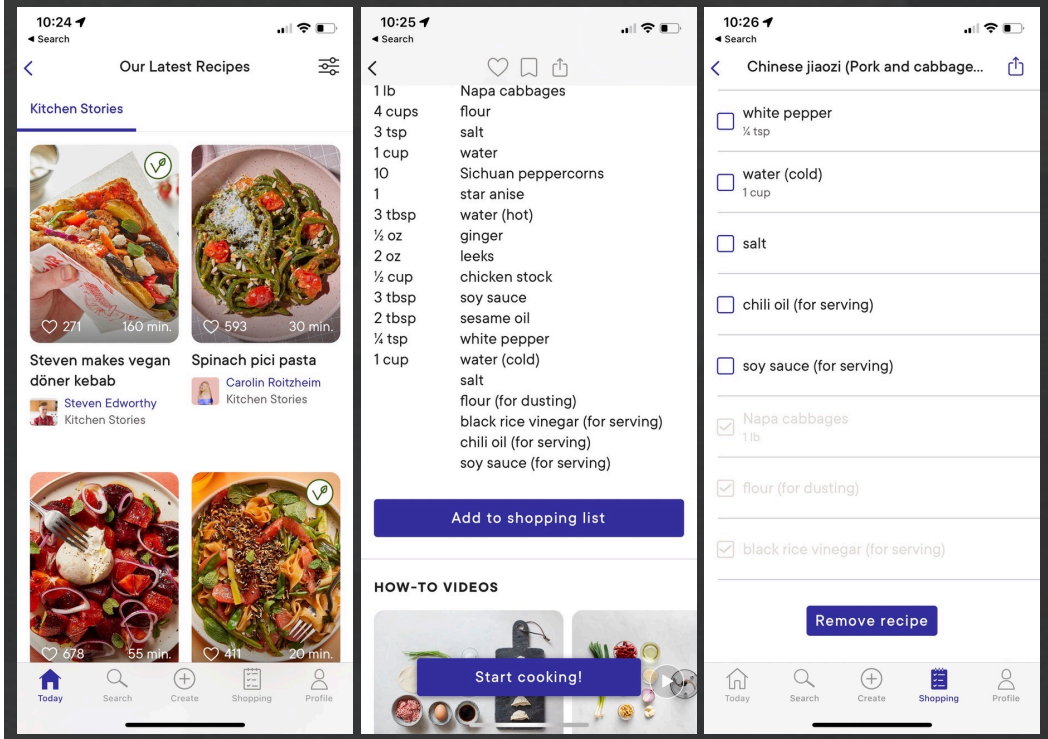
Autumn 2026
November 9, 2026

Hall of Fame or Shame?



- Kitchen Stories

Hall of Fame!



- Kitchen Stories
- Like
  - large pictures of recipes
  - photos & videos
  - shopping list that marks off as you purchase
- Wish
  - fonts hard to read for long list
  - why the tab for “Kitchen Stories”?

Usability Testing

刘哲明
Prof. James A. Landay
Computer Science Department
Stanford University

Autumn 2026
November 9, 2026

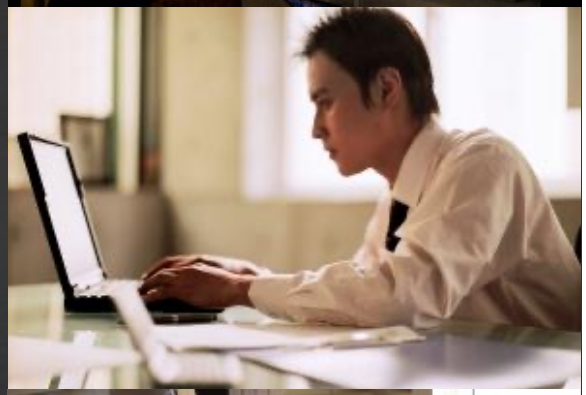
Outline

- Why do usability testing?
- Choosing participants
- Ethical considerations
- Designing & conducting the test
- Using the results
- Experimental options & details

Why do Usability Testing?

- Can't tell how good UI is until?
  - people use it!
- Expert review methods (e.g., HE) are based on evaluators who may?
  - know too much
  - not know enough (about tasks, etc.)
- Hard to predict what real users will do







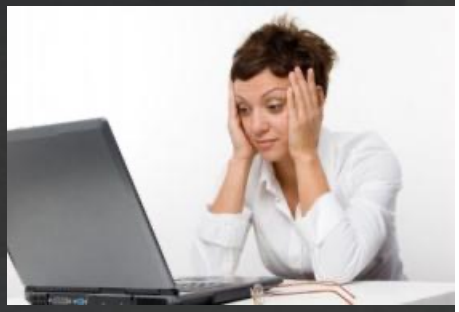
Choosing Participants

- Representative of target users. How so?
  - job-specific vocab / knowledge
  - tasks
- Approximate if needed
  - system intended for doctors?
    - get ~~medical students~~ or nurses
  - system intended for engineers?
    - get engineering students
- Use incentives to get participants
  - t-shirt, mug, free coffee/pizza





Ethical Considerations



- Usability tests can be distressing
  - users have left in tears
- Testing/fieldwork can be coercive if there is a power imbalance (e.g., in under resourced communities)



http://centread.ucsc.edu/CenTREAD%20photos/BrianDowd2.JPG

People may feel no option but to speak to you or give you their time even though they may not get anything of value in return.

Ethical Considerations

- You have a responsibility to alleviate these issues
  - make voluntary with informed consent (form)
  - avoid pressure to participate
  - let them know they can stop at any time
  - stress that you are *testing the system, not them*
  - make collected data as anonymous as possible
- Often must get human subjects approval (IRB)



https://www.unthsc.edu/north-texas-regional-irb/institutional-review-board-meeting/

Usability Test Proposal

- A report that contains
  - objective
  - description of system being testing
  - task environment & materials
  - participants
  - methodology
  - tasks
  - test measures



- Get approved & then reuse for final report
- Seems tedious, but writing this will help “debug” your test

Selecting Tasks

- Tasks from low-fi design can be used
  - may need to shorten if
    - they take too long
    - require background that test user won't have



- Don't train unless that will occur in real deployment
- Avoid bending tasks in direction of what your design best supports
- Don't choose tasks that are too fragmented ?
  - fragmented = does not represent a complete goal someone would try to accomplish with your application
  - e.g., login/create account is *not* a task

Two Types of Data to Collect

- Process data
  - observations of what users are doing & thinking
  - *qualitative*



http://allazollers.com/discovery-research.php

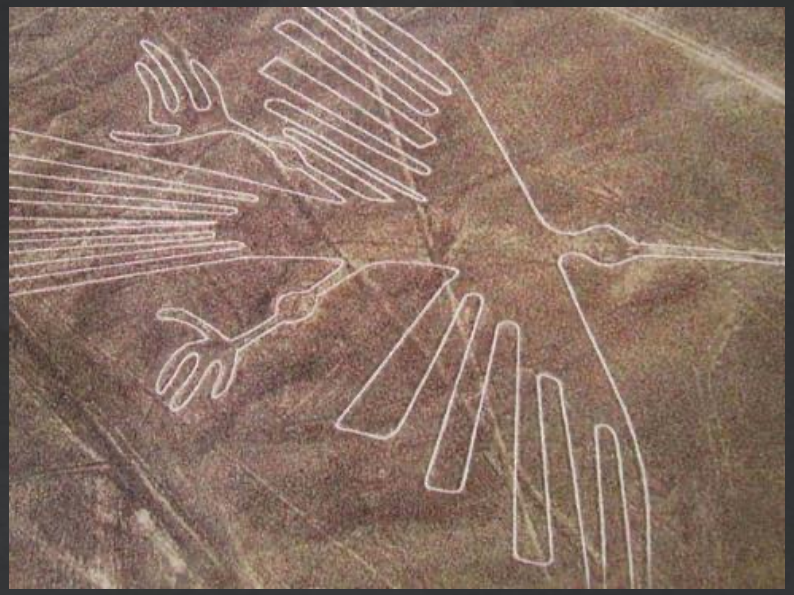
- Bottom-line data
  - summary of what happened
    - time, errors, success
  - i.e., the dependent variables
  - *quantitative*



http://www.fusionfarm.com/content/uploads/2012/10/analyzing-data.jpg

Which Type of Data to Collect?

- Focus on process data first
  - gives good overview of where problems are



http://www.redicecreations.com/ul_img/24592nazca_bird.jpg

Which Type of Data to Collect?

- Focus on process data first
  - gives good overview of where problems are
- Bottom-line data doesn't tell you ?
  - where to fix
  - just says: "too slow", "too many errors", etc.
- Hard to get reliable bottom-line results
  - need many users for statistical significance



http://netbreak-ph.blogspot.com/2012/09/Hitleap-get-traffic-views-on-site.html

The “Thinking Aloud” Method

- Need to know what users are thinking, not just what they are doing

- Ask users to talk while performing tasks
  - tell us *what they are thinking*
  - tell us *what they are trying to do*
  - tell us *questions that arise as they work*
  - tell us *things they read on the screen*



Thinking Aloud (cont.)

- Prompt the user to keep talking
  - "tell me what you are thinking"
- Only help on things you have pre-decided
  - keep track of anything you do give help on
- Make a *recording* & take good notes
  - make sure you can tell what they were doing
  - use a digital watch/clock
  - record audio & video
    - or even event logs



http://jennycham.co.uk/wp-content/uploads/2011/08/20091122125022548 1.jpg

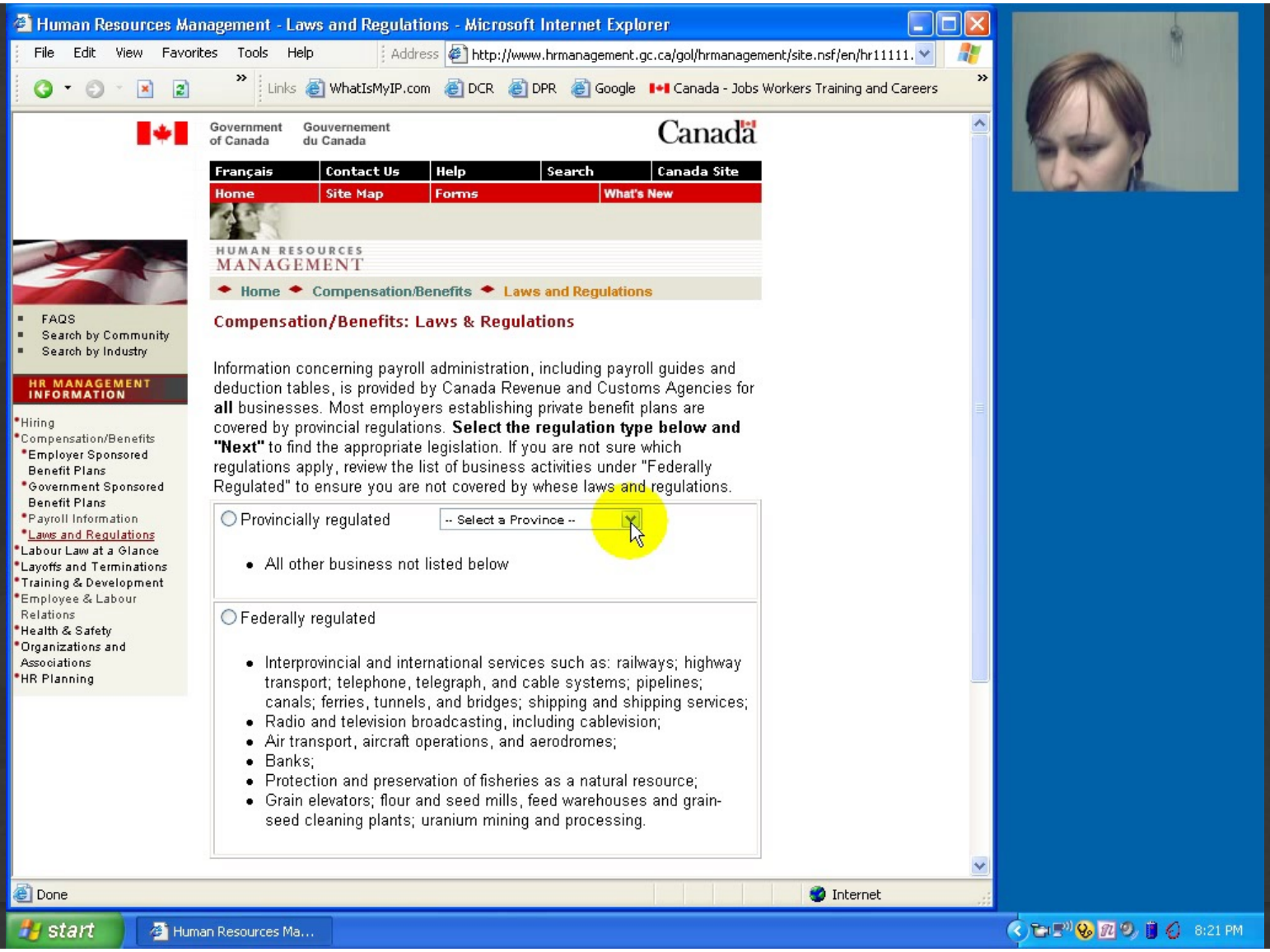
Will Thinking Aloud Give the Right Answers?

- Not always

- If you ask, people will always give an answer, even if it has nothing to do with facts
  - panty hose example



→Try to avoid specific questions (especially that have binary answers)


Human Resources Management - Laws and Regulations - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.hrmanagement.gc.ca/gol/hrmanagement/site.nsf/en/hr11111.
Links WhatIsMyIP.com DCR DPR Google Canada - Jobs Workers Training and Careers
Government of Canada
Gouvernement du Canada
Canada
Français
Contact Us
Help
Search
Canada Site
Home
Site Map
Forms
What's New
HUMAN RESOURCES MANAGEMENT
Home
Compensation/Benefits
Laws and Regulations
FAQS
Search by Community
Search by Industry
HR MANAGEMENT INFORMATION
Hiring
Compensation/Benefits
Employer Sponsored Benefit Plans
Government Sponsored Benefit Plans
Payroll Information
Laws and Regulations
Labour Law at a Glance
Layoffs and Terminations
Training & Development
Employee & Labour Relations
Health & Safety
Organizations and Associations
HR Planning
Compensation/Benefits: Laws & Regulations
Information concerning payroll administration, including payroll guides and deduction tables, is provided by Canada Revenue and Customs Agencies for all businesses. Most employers establishing private benefit plans are covered by provincial regulations. Select the regulation type below and "Next" to find the appropriate legislation. If you are not sure which regulations apply, review the list of business activities under "Federally Regulated" to ensure you are not covered by whese laws and regulations.
Provincially regulated
-- Select a Province --
All other business not listed below
Federally regulated
Interprovincial and international services such as: railways; highway transport; telephone, telegraph, and cable systems; pipelines; canals; ferries, tunnels, and bridges; shipping and shipping services;
Radio and television broadcasting, including cablevision;
Air transport, aircraft operations, and aerodromes;
Banks;
Protection and preservation of fisheries as a natural resource;
Grain elevators; flour and seed mills, feed warehouses and grain-seed cleaning plants; uranium mining and processing.
Done
Internet
start
Human Resources Ma...
8:21 PM

Using the Test Results

- Summarize the data
  - make a list of all critical incidents (CI)
    - positive & negative
  - include references back to original data
  - try to judge why each difficulty occurred
- What does data tell you?
  - UI work the way you thought it would?
    - users take approaches you expected?
  - something missing?



https://cdn.careerfoundry.com/en/wp-content/uploads/old-blog-uploads/data-analyst-skills-and-tasks.jpeg

Using the Results (cont.)

- Update tasks & rethink design
  - rate severity & ease of fixing CIs
  - fix both severe problems & make the easy fixes



http://www.thetomorrowplan.com/exchange/policies-prairie-chickens-and-parking/

Measuring Bottom-Line Usability



- Situations in which numbers are useful
  - time requirements for task completion (e.g., ticket machine or ATM)
  - successful task completion % (e.g., checkout)
  - compare two designs on speed or # of errors
- Ease of measurement
  - time is easy to record
  - error or successful completion is harder
    - define in advance what these mean
- Do not combine with thinking-aloud. Why?
  - talking can affect speed & accuracy

Stats: Analyzing the Numbers

- Example: trying to get task time ≤ 30 min.
  - test gives: 40, 5, 20, 90, 10, 15
  - mean (average) = 30
  - median (middle) = 17.5
  - looks good!



- Did we achieve our goal?
- No, we are not certain of anything!
- Factors contributing to our uncertainty?
  - small number of test users (n = 6)
  - results are very variable (standard deviation = 32)
    - std. dev. measures dispersal from the mean

Stats: Analyzing the Numbers (cont.)

- This is what basic statistics can be used for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
  - note: 95% (or alpha=5%) is standard for HCI

Stats: Analyzing the Numbers (cont.)

Web Usability Test Results						
Participant #		**Time (minutes)**				
1		20				
2		15				
3		40				
4		90				
5		10				
6		5				
	number of participants	**6**				
	mean	**30.0**				
	median	**17.5**				
	std dev	**31.8**				
	standard error of the mean	**= stddev / sqrt (#samples)**			**13.0**	
	typical values will be mean +/- 2*standard error			**--> 4 to 56!**		
	what is plausible? = confidence (alpha=5%, stddev, sample size)	**25.4**	**--> 95% confident between 4.6 & 55.4**			

Stats: Analyzing the Numbers (cont.)

- This is what basic statistics can be used for
- Crank through the procedures and you find
  - 95% certain that typical value is between 5 & 55
- Usability test data is *highly variable*
  - need lots to get good estimates of typical values
  - 4x as many tests will only narrow range by 2x
    - breadth of range depends on sqrt of # of test users
  - this is when online methods become useful
    - easy to test w/ large numbers of users

Measuring User Preference

- How much users like or dislike the system
  - can ask them to rate on a scale of 1 to 10
  - or have them choose among statements
    - "best UI I've ever…", "better than average"…
  - hard to be sure what data will mean
    - novelty of UI, unrealistic setting …
- If many give you low ratings → trouble



- Can get some useful data by asking
  - what they liked, disliked, where they had trouble, best part, worst part, etc.
  - redundant questions are OK

Comparing Two Alternatives

- *Between groups* experiment
  - two groups of test users
  - each group uses only 1 of the systems
- *Within groups* experiment
  - one group of test users
    - each person uses both systems (cheaper)
    - can't use the same tasks or order (learning)
  - best for low-level interaction techniques
    - e.g., new mouse, new swipe interaction, …



Comparing Two Alternatives

- Between groups requires many more participants than within groups
- See if differences are statistically significant
  - assumes normal distribution & same std. dev.
- Online companies can do large AB tests
  - look at resulting behavior (e.g., buy?)

Instructions to Participants

- Describe the purpose of the evaluation
  - "I'm testing the product; I'm not testing you"
- Tell them they can quit at any time
- Demonstrate the equipment
- Explain how to think aloud
- Explain that you will not provide help
- Describe the task
  - give written instructions
  - one task at a time



Reporting the Results

- Report what you did & what happened
- Images & graphs help people get it!
- Video clips can be quite convincing





Heuristic Evaluation vs. User Testing

- HE is much faster
  - 2-4 hours each evaluator vs. days-weeks
- HE doesn't require interpreting user's actions
- User testing is far more accurate (by def.)
  - takes into account actual users and tasks
  - HE may miss problems & find "false positives"

- Good to alternate between HE & user testing
  - each technique may find different types of problems
  - don't waste participants

Summary

- User testing is important, but takes time/effort
- Use ????? tasks & ????? participants
  - *real tasks* & *representative* participants
- Be ethical & treat your participants well
- Want to know what people are doing & why? collect
  - process data
- Bottom line data requires ???? to get statistically reliable results
  - *more participants*
- Difference between between & within groups?
  - between groups: each subject participates in only one of n conditions
  - within groups: everyone participates in multiple conditions

Administrivia

- Group project grades (60% of your overall grade)
  - your project grade can be impacted by your contribution to the team → pull your weight!
  - midway hi-fi due studio week 9 (show >= 1 task working!)
- Web sites now linked from class projects page
  - let Lindsay @popowski know if you updated your tagline or project name (check)
- Assignments due the final week
  - Poster & pitch slide drafts (Mon) & finals (Wed at noon)
  - Video demo (Wed)
  - Prototype (Fri)
  - Report (Sun)