Announcements

Today after lecture: Team mixer, in this room

Coming up: Project brainstorming ideas due this Friday 9am
Articulating Research Contributions: Common Critiques/Areas to Improve

Clearly describe research question & contribution

  Don’t be vague/abstract
  “We will explore a domain/space” is too vague

Contextualize adequately in existing research

  See if the idea has been pursued (Google Scholar is your friend)
  If you are replicating/Extending, why does your replication/extension matter

Have an evaluation plan
Your paper advances an argument; your methods provide evidence.
Different arguments require different evidence
Figure 2: The strategy circumplex (adapted from Runkel & McGrath).

Maxima for each criterion:
A = Generalizability
B = Precision
C = Realism

Quadrant I: Field Strategies
Quadrant II: Experimental Strategies
Quadrant III: Respondent Strategies
Quadrant IV: Theoretical Strategies
Method triangulation

All methods are have limitations; you can deploy more than one to provide stronger supporting evidence.

E.g., follow up a lab study (abstract, obtrusive) with a field study (concrete, unobtrusive)
E.g., after analyzing log data (large scale, unobtrusive), conduct interviews (small scale, obtrusive)
How do we decide which methods to use?
Determining your methods

Your claims + Standards of evidence in your area = Your methods
Determining your methods

Your claims + Standards of evidence in your area = Your methods
**Common claims – Systems**

<table>
<thead>
<tr>
<th>I built a system that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...solves an entirely new problem</td>
<td>Is the problem important? (How well) does it work?</td>
<td>Field study, lab experiment, technical evaluation</td>
</tr>
</tbody>
</table>
## Common claims – Systems

<table>
<thead>
<tr>
<th>I built a system that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...solves an entirely new problem</td>
<td>Is the problem important? (How well) does it work?</td>
<td>Field study, lab experiment, technical evaluation</td>
</tr>
<tr>
<td>...solves an old problem more effectively</td>
<td>(How well) does it work? How much better is it?</td>
<td>Technical evaluation, lab experiment, field experiment</td>
</tr>
</tbody>
</table>
## Common claims – Systems

<table>
<thead>
<tr>
<th>I built a system that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...solves an entirely new problem</td>
<td>Is the problem important? (How well) does it work?</td>
<td>Field study, lab experiment, technical evaluation</td>
</tr>
<tr>
<td>...solves an old problem more effectively</td>
<td>(How well) does it work? How much better is it?</td>
<td>Technical evaluation, lab experiment, field experiment</td>
</tr>
<tr>
<td>...improves task performance</td>
<td>By how much? Under what circumstances?</td>
<td>Lab experiment, formal theory, judgement study</td>
</tr>
</tbody>
</table>
### Common claims – Systems

<table>
<thead>
<tr>
<th>I built a system that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...solves an entirely new problem</td>
<td>Is the problem important? (How well) does it work?</td>
<td>Field study, lab experiment, technical evaluation</td>
</tr>
<tr>
<td>...solves an old problem more effectively</td>
<td>(How well) does it work? How much better is it?</td>
<td>Technical evaluation, lab experiment, field experiment</td>
</tr>
<tr>
<td>...improves task performance</td>
<td>By how much? Under what circumstances?</td>
<td>Lab experiment, formal theory, judgement study</td>
</tr>
<tr>
<td>...lowers the threshold/raises the ceiling/widens the walls</td>
<td>What can it now make? Who can now make it?</td>
<td>Interviews, Demonstrative applications, long-term deployment</td>
</tr>
</tbody>
</table>
# Common claims – Systems

<table>
<thead>
<tr>
<th>I built a system that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...solves an entirely new problem</td>
<td>Is the problem important?</td>
<td>Field study, lab experiment, technical evaluation</td>
</tr>
<tr>
<td></td>
<td>(How well) does it work?</td>
<td></td>
</tr>
<tr>
<td>...solves an old problem more effectively</td>
<td>(How well) does it work?</td>
<td>Technical evaluation, lab experiment, field experiment</td>
</tr>
<tr>
<td></td>
<td>How much better is it?</td>
<td></td>
</tr>
<tr>
<td>...improves task performance</td>
<td>By how much?</td>
<td>Lab experiment, formal theory, judgement study</td>
</tr>
<tr>
<td></td>
<td>Under what circumstances?</td>
<td></td>
</tr>
<tr>
<td>...lowers the threshold/raises the ceiling/widens the walls</td>
<td>What can it now make?</td>
<td>Interviews, Demonstrative applications, long-term</td>
</tr>
<tr>
<td></td>
<td>Who can now make it?</td>
<td>deployment</td>
</tr>
<tr>
<td>...is more accessible</td>
<td>Who can now use it?</td>
<td>Interviews, field study, field experiment, sample</td>
</tr>
<tr>
<td></td>
<td>How much better is it?</td>
<td>survey</td>
</tr>
</tbody>
</table>
## Common claims – Studies

<table>
<thead>
<tr>
<th>I hypothesize that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...people behave in accordance with model X</td>
<td>How do you know? What other factors might be at play?</td>
<td>Field study, formal theory, experimental simulation, field experiment</td>
</tr>
</tbody>
</table>
## Common claims – Studies

<table>
<thead>
<tr>
<th>I hypothesize that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...people behave in accordance with model $X$</td>
<td>How do you know?</td>
<td>Field study, formal theory, experimental simulation, field experiment</td>
</tr>
<tr>
<td></td>
<td>What other factors might be at play?</td>
<td></td>
</tr>
<tr>
<td>...we can get better outcomes using mechanism $Y$</td>
<td>How can you be sure?</td>
<td>Lab experiment, field experiment, sample survey, experimental simulation</td>
</tr>
<tr>
<td></td>
<td>How much better?</td>
<td></td>
</tr>
</tbody>
</table>
## Common claims – Studies

<table>
<thead>
<tr>
<th>I hypothesize that</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>...people behave in accordance with model X</td>
<td>How do you know? What other factors might be at play?</td>
<td>Field study, formal theory, experimental simulation, field experiment</td>
</tr>
<tr>
<td>...we can get better outcomes using mechanism Y</td>
<td>How can you be sure? How much better?</td>
<td>Lab experiment, field experiment, sample survey, experimental simulation</td>
</tr>
<tr>
<td>...dimension X impacts people’s interactions with system Y</td>
<td>How do you know? What other factors?</td>
<td>Field study, field experiment, sample survey</td>
</tr>
</tbody>
</table>
## Common claims – Studies

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Likely questions</th>
<th>Possible methods</th>
</tr>
</thead>
</table>
| ...people behave in accordance with model X    | How do you know?  
What other factors might be at play? | Field study, formal theory,  
experimental simulation, field experiment |
| ...we can get better outcomes using mechanism Y | How can you be sure?  
How much better? | Lab experiment, field experiment, sample survey,  
experimental simulation |
| ...dimension X impacts people’s interactions with system Y | How do you know?  
What other factors? | Field study, field experiment, sample survey |
| ...understanding system X can inform us about broader problem Y | Why do you think the two are sufficiently similar? | Field study, formal theory, field experiment |
Determining your methods

Your claims + Standards of evidence in your area = Your methods
Determining your methods

Your claims + Standards of evidence in your area = Your methods
Standards of evidence

Every field has a set of standards that are agreed upon as being sufficient or appropriate for proving a point. For example,

**Medicine:** Double-blind randomized controlled trials

**Economics:** Mathematical modeling

**Math:** Formal proofs

**History:** Close reading of primary source documents

(To dive deeper, check out work on evidence in Philosophy of Science.)
Standards of evidence

In computing, the standard of evidence varies by subfield

Your goal: convince an expert *in your area*
(Use the methods that those experts expect)
Don’t reinvent the wheel

Unless your contribution is to propose a new method, most research will use already-established methods. This is good!

Your nearest-neighbor paper (“model paper”), and the rest of your literature search, has likely already introduced you to methods that can be adapted to your purpose.

To start, figure out what the norms are, and maybe tweak them.
Designing an evaluation
What NOT to do

“We built a thing/ran an experiment. Now how would we evaluate this?”

Claiming the idea is not separate from designing the evaluation—they’re part of the same process.

If you can precisely articulate your idea and your claim, that articulation entails one (or more) appropriate evaluation(s).
What to do instead:
Derive evaluation from thesis
Step 1: Articulate your thesis

**Setup**
Adding a high-quality musical underlay to audio stories is a tedious task best left to expert audio editors

The best remembered gestural interactions result from the careful planning of an expert designer

**Punchline**
We can simplify parts of the underlay process so that non-experts can easily add them but still produce high-quality results

Elicitation from non-expert users can produce gesture sets that are easier to remember
Implicit in most thesis is one of a few claim structures, including:

\(x > y\) Approach \(x\) is better than approach \(y\) at solving the problem

\(\exists x\) It is possible to construct an \(x\) that satisfies some criteria, whereas it was not known to be possible before

bounding \(x\) Approach \(x\) only works given certain assumptions (i.e. has limitations)
x > y? ∃x? bounding x?

Setup
Adding a high-quality musical underlay to audio stories is a tedious task best left to expert audio editors.

The best remembered gestural interactions result from the careful planning of an expert designer.

Punchline
We can simplify parts of the underlay process so that non-experts can easily add them but still produce high-quality results.

Elicitation from non-expert users can produce gesture sets that are easier to remember.

Claim
∃ x: we develop semi-automated tools based on structure extracted from the music that simplify high-quality underlay construction.

x > y: gestures elicited from non-technical users will be easier for end-users to remember and agreement than those designed by experts.
Step 3: Claims imply an evaluation design

Each claim structure implies an evaluation design:

- $\exists x$  Demonstrate that your approach achieves $x$

- $x > y$  Given a representative task or set of tasks, test whether $x$ in fact outperforms $y$ at the problem

- bounding $x$  Demonstrate bounds inside or outside of which approach $x$ fails
### Punchline
We can simplify parts of the underlay process so that non-experts can easily add them but still produce high-quality results.

### Claim
1. ∃ x: we develop semi-automated tools based on structure extracted from the music that simplify high-quality underlay construction.
2. x > y: gestures elicited from non-technical users will be easier for end-users to remember and agreement than those designed by experts.

### Implied evaluation
Demonstrate high-quality underlay results and show that non-expert end-users can easily create them.

Compare coverage and agreement scores of gesture sets elicited from non-technical users and those designed by experts.
Guess the evaluation!
Guess the evaluation

**Punchline**
We can encourage users to lead more active lifestyles via an ambient interface which detects physical activity and displays progress through a calm narrative.

**Claim**
\( \exists x: \) an activity-sensing wearable device can accurately classify and present an ambient summary of users’ recent activity levels, such that users feel encouraged to adopt healthier habits.

**Implied evaluation**

1) “Can accurately classify” – Validate classification of user activity by comparing it to a manually recorded activity log.
2) “Feel encouraged to adopt healthier habits” – Survey users’ attitudes towards the interface and observe their exercise habits over a time period.
Architecture of an evaluation
Four constructs that matter

- Dependent variable
- Independent variable
- Task
- Threats
What's the outcome you're measuring?

Efficiency, accuracy, performance, satisfaction, trust, etc.

*Should be based on your thesis/claims*

How to operationalize your DV?

- Measure a small number (one to three) of central DVs
- Choose DVs that best speak to your claim (don’t fall into the trap of measuring things that are easily quantifiable just because they’re easy to measure)
What are you manipulating in order to cause the change in the dependent variable?

The IV leads to *conditions* in your evaluation. Examples include:

- Algorithm (e.g., if proposing a new algorithm compare new alg. to old alg.)
- Dataset used (e.g., when training a model)
- Version of an interface (or other stimulus shown to participants)
Task

What is the specific routine being followed in order to manipulate the independent variable and to measure the dependent variable?

E.g. “Participants will have thirty seconds to identify each article as disinformation or not, randomizing across interfaces. The data will be analyzed within-subjects.”
Threats

What might threaten the validity of your evaluation?

• Might your participants feel experimenter demand?
• Are your participants biased toward healthy young technophiles?
• Do your participants always see the best interface first?
• Is there some other variable (confound) responsible for differences you see (e.g. one interface is easier to use)?
Threats

Handling threats:
1) Control – equalize across groups (stratification or randomization)
2) Measure – record the confound to later account for it statistically
3) Manipulate – turn it into an IV
4) Argue as irrelevant – yes, that bias might exist, but it’s not conceptually important to the phenomenon you’re studying and is unlikely to strongly effect the outcome or make the results less generalizable
Announcements

Now: Team mixer, in this room

Coming up: Project brainstorming ideas due this Friday 9am