Evaluation
Today’s OhYay

We are doing small discussions throughout today, so get into a row with 4 people total.

Try to find a room with your small group + another small group!
Administrivia

Assignment 4 going live today, due next week (Week 7)
  Proposed Solution (written)
  Evaluation Plan (outline only)

Reminders:
  Assignment 5, Draft Proposal (2-week assignment) due Week 9
  Final Presentation + Final Proposal due Week 10
“But how would we even evaluate that?”

People often rush to this question early on in ideation.

Today’s goal is to provide scaffolding for how to answer it.
Today’s big idea: evaluation

How do we get precise about what we need to evaluate for our project?

How do we design an appropriate evaluation?

How do we analyze our evaluation results?
Why perform evaluation in research?
Idea Shark Tank

Recall from Week 1 that research introduces a new idea into the world.

So...how do we know if that idea is worth adopting or paying attention to?

Option 1 (“The Simon Cowell Solution”): Academia’s Got Talent, Shark Tank

Option 2: Construct an evaluation to test the idea fairly

Let’s do this one: the goal isn’t advocacy; it’s an understanding of the idea’s strengths and limits
Standards of evidence

Every field has an accepted standard of evidence — a set of methods that are agreed upon for proving a point

- Medicine: Double-blind randomized controlled trial
- Philosophy: Rhetoric
- Math: Formal proof
- Applied Physics: Measurement
Standards of evidence

In computing, because areas use different methods, the standard of evidence differs based on the area.

Your goal: convince an expert in your area.

So, use the methods appropriate to your area.
Designing an evaluation
In-class activity

Today, you will outline how to evaluate your idea and bit flip.

The expectation is **not** that you are already ready to evaluate your system!

Some fields naturally have constant, continual evaluation baked into the research process. Others have front-loaded design or engineering costs.

However, knowing how you will evaluate your project will help you better determine what is core and what is periphery in what you build.

So…think big and high-level for this activity!

You will submit this outline as part of Assignment 5.
Problematic point of view

“But how would we evaluate this?”

Why is this point of view problematic?

Implication: “I believe the idea is right, but I don’t believe that we can prove it.”

Implication: “The thread of designing the evaluation is different than the process Evaluation is distinct from the validity of the idea.”

Neither implication is correct. If you can precisely articulate your idea and your bit flip, then you can design an appropriate evaluation.

If you can’t precisely articulate your idea and your bit flip, then you can’t design an appropriate evaluation.
Step 1: articulate your thesis

A much more productive approach is to derive an evaluation design directly from your idea.

What is the main thesis of your work?

(Lucky for you, you came up with this when writing the Introduction of your paper. It’s the topic sentence of your bit flip paragraph.)
Recall:

Bit

Network behaviors are defined in hardware, statically.

Code compilers should utilize smart algorithms to optimize into machine code.

A minimum graph cut algorithms should always return correct answers.

Flip

If we define the behaviors in software, networks can become dynamic and more easily debuggable.

Code compilers will find more efficient outcomes if they just do monte carlo (random!) explorations of optimizations.

A randomized, probabilistic algorithm will be much faster, and we can still prove a limited probability of an error.
Recall:

**Bit**

Network behaviors are defined in hardware, statically.

Code compilers should utilize smart algorithms to optimize into machine code.

A minimum graph cut algorithms should always return correct answers.

**Flip**

If we define the behaviors in software, networks can become dynamic and more easily debuggable.

Code compilers will find more efficient outcomes if they just do monte carlo (random!) explorations of optimizations.

A randomized, probabilistic algorithm will be much faster, and we can still prove a limited probability of an error.

[3 min] Step 1. On your slide, write your thesis as A) a bit and B) flip.

https://docs.google.com/presentation/d/1k_0blqOHXXxeQzH-uer7RlYwtniNaw-FkMe76cvTiZg/edit?usp=sharing
Step 2: map your thesis onto a claim

There are only a small number of claim structures implicit in most theses:

- **$x > y$:** 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 possible before

- **bounding $x$:** approach $x$ only works given certain assumptions
Bit
Network behaviors are defined in hardware, statically.

Flip
If we define the behaviors in software, networks can become dynamic.

Claim
∃ x: software-defined behaviors can be changed on the fly

> y: monte carlo exploration will produce more optimized code than hand-tuned compilers

> y: a randomized graph cut algorithm is faster and has bounded error
<table>
<thead>
<tr>
<th>Bit</th>
<th>Flip</th>
<th>Claim</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network behaviors are defined in hardware, statically.</td>
<td>If we define the behaviors in software, networks can become dynamic.</td>
<td>$\exists \ x$: software-defined behaviors can be changed on the fly</td>
</tr>
<tr>
<td>Code compilers should utilize smart algorithms to optimize into machine code.</td>
<td>Code compilers will find more efficient outcomes if they just do monte carlo (random!) explorations of optimizations.</td>
<td>$x &gt; y$: monte carlo exploration will produce more optimized code than hand-tuned compilers</td>
</tr>
<tr>
<td>A minimum graph cut algorithms should always return correct answers.</td>
<td>A randomized, probabilistic algorithm will be much faster, and we can still prove a limited probability of an error.</td>
<td>$x &gt; y$: a randomized graph cut algorithm is faster and has bounded error</td>
</tr>
</tbody>
</table>

[3 min] Step 2. On your slide, A) identify your claim type, and B) write your claim.
Step 3: claims imply an evaluation design

Each claim structure implies an evaluation design:

\( x > y \): given a representative task or set of tasks, test whether \( x \) in fact outperforms \( y \) at the problem

\( \exists x \): demonstrate that your approach achieves \( x \)

bounding \( x \): demonstrate bounds inside or outside of which approach \( x \) fails
If we define the behaviors in software, networks can become dynamic.

Code compilers will find more efficient outcomes if they just do monte carlo (random!) explorations of optimizations.

A randomized, probabilistic algorithm will be much faster, and we can still prove a limited probability of an error.

Demonstrate that behaviors propagate, and which kind of behaviors can be authored.

Compare runtime of generated machine code against known best approaches.

Prove runtime for randomized algorithm (vs. prior algorithm) and probability of error.
<table>
<thead>
<tr>
<th>Flip</th>
<th>Claim</th>
<th>Implied evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>If we define the behaviors in software, networks can become dynamic.</td>
<td>$\exists \ x$: software-defined behaviors can be changed on the fly</td>
<td>Demonstrate that behaviors propagate, and which kind of behaviors can be authored</td>
</tr>
<tr>
<td>Code compilers will find more efficient outcomes if they just do monte carlo (random!) explorations of optimizations.</td>
<td>$\times &gt; \ y$: monte carlo exploration will produce more optimized code than hand-tuned compilers</td>
<td>Compare runtime of generated machine code against known best approaches</td>
</tr>
<tr>
<td>A randomized, probabilistic algorithm will be much faster, and we can still prove a limited probability of an error.</td>
<td>$\times &gt; \ y$: a randomized graph cut algorithm is faster and has bounded error</td>
<td>Prove runtime for randomized algorithm (vs. prior algorithm) and probability of error</td>
</tr>
</tbody>
</table>

[3 min] Step 3. On your slide, write the high-level evaluation design.
[5 min] Share out these ideas with your row.

(next, we’ll make this evaluation more concrete, so familiarize yourself with other people’s projects so that you can give feedback)
Architecture of an Evaluation
Four constructs that matter

To develop your evaluation plan, you need to get precise about four components of your evaluation:

- Dependent variable
- Independent variable
- Task
- Threats
DV: dependent variable


The choice of this quantity should be clearly implied by your thesis. It's often tempting to measure many DVs, and I'm not against doing so. However, one should be your central outcome, and the others auxiliary.
In other words, what determines what x and y are? What are you manipulating in order to cause the change in the dependent variable?

The IV is the construct that leads to conditions in your evaluation. Examples might include:

- Algorithm
- Dataset size or quality
- Interface
Lisa’s Example

**Bit**
Understanding student coding process through intermediate work is extremely complex

**Claim**
\[ \exists x \]
An ML system can do the flip

**Flip**
The majority of student work can be categorized into intermediate steps towards the solution

**Implied evaluation**
Demonstrate that intermediate work can be classified by ML
Lisa’s Example

Bit  Understanding student coding process through intermediate work is extremely complex

Claim  ∃x  An ML system can do the flip

DV  accuracy

Flip  The majority of student work can be categorized into intermediate steps towards the solution

Implied evaluation  Demonstrate that intermediate work can be classified by ML

IV  model type labeled data

[2 min] In your slide’s speaker notes, write your 1) dependent variable and 2) independent variable.
Task

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

We will perform one-shot prediction of classes at the 25th percentile of popularity in ImageNet according to Google search volume.

Participants will have thirty seconds to identify each article as disinformation or not, within-subjects, randomizing across interfaces.

We will run a performance benchmark drawn from Author et al. against each system.
Lisa’s Example

**Bit**  Understanding student coding process through intermediate work is extremely complex

**Claim**  An ML system can do the flip

**DV**  accuracy

**IV**  identified buckets of student work model complexity labeled data

**Flip**  The majority of student work can be categorized into intermediate steps towards the solution

**Implied evaluation**  Demonstrate that intermediate work can be classified by ML

**Task**  We will train an ML classifier to categorize student work into buckets, where each bucket pertains to a common step towards the solution.

[2 min] In your slide’s speaker notes, write your 3) your task.
Threats

What are your threats to validity? In other words, what might bias your results or mean that you’re telling an incomplete story?

Might your selection of which classes to predict influence the outcome?

Are you running on particular cloud architectures that are amenable to, or not amenable to, your task?

Are your participants biased toward healthy young technophiles?

Do your participants always see the best interface first?
Threats

There are typically three ways to handle these kinds of issues:

1) 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 affect the outcome or make the results less generalizable

2) Stratify: re-run your evaluation in each setting to see whether the outcomes change

3) Randomize: explicitly randomize (e.g., people) across values of the control variable. For example, randomize the order in which people see the interface.
The majority of student work can be categorized into intermediate steps towards the solution.

**Claim**

We will train an ML classifier to categorize student work into buckets, where each bucket pertains to a common step towards the solution.

**Threat**

- How to pick initial buckets?
  - Argue as irrelevant: Can bring in teaching expert for initial labeling

- Not feasible to label all of student data?
  - Stratify (sort of): run on unlabeled data, and evaluate randomly picked labels by hand

[5 min] With your row, discuss threats and how to address them. Write down in speaker notes.
Find your Patronus

There's no need to start from scratch on this.

Your nearest neighbor paper, and the rest of your literature search, has likely already introduced evaluation methods into this literature that can be adapted to your purpose.

Start here: figure out what the norms are, and tweak them. Talk to your mentor if helpful.

Lisa to link some example papers later tonight
Assignment 5

Proposed Solution:

Describe the high-level and low-level ideas of your system

Examples on website, but write the minimum needed for a reader to understand the thesis in your Evaluation Plan

Written (400-600 words)

Evaluation Plan:

Thesis, Claim, Evaluation Design (i.e., what we did in class today)

Outline only

Due in 1 week
Evaluation

Slide content shareable under a Creative Commons Attribution-NonCommercial 4.0 International License.