CS 293/EDUC 473

Measurement

Announcements & reminders

- Project rationale due tomorrow at midnight
  - **Who does not yet have a project partner?**
- HW2 due next Tuesday
  - start early especially if you feel uncomfortable with running machine learning models!
- Project pitches are in 2 weeks: 4 minutes per team
  - Rubric & example shared on Canvas
- At the end of the week, you'll be getting a survey about what you want to do in class
- Extra office hours
  - Dora: 2:30-3:30pm on Friday
  - Rose: email her if you want to talk (rewang@stanford.edu)

Revisiting the “Getting to know you” survey

What are you most excited to learn about?

Ethics

Large language models & education

Teacher Feedback

End-to-end research

Fairness & bias

Collaborating & Engaging with peers

Pitching project to educators

Causal estimation with text

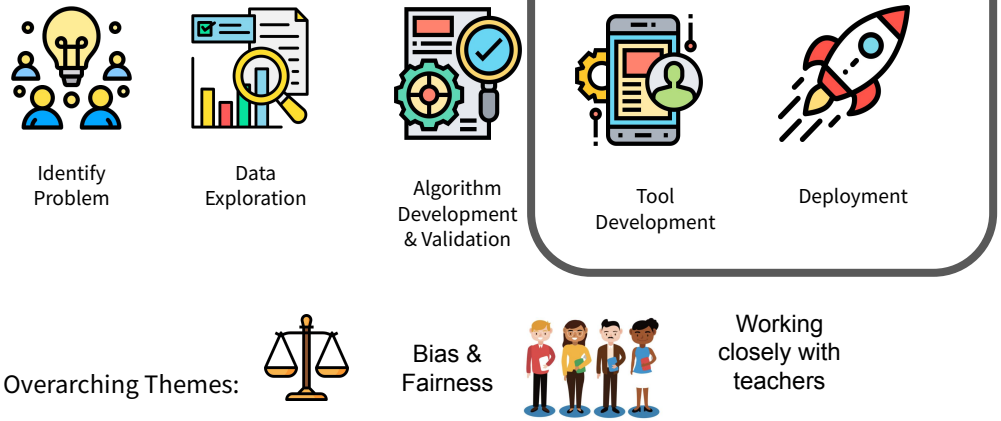
Upcoming Survey

What else would you like to do?

Plan:

- Rose's lecture on LLMs for teacher feedback
- Design & Deployment
  - Visit by Rakiya Brown (TeachFX)
- Experimental Design
- Guest lecture by Diane Litman

~3 classes have room for flexibility, so we can make adjustments (e.g. to continue existing conversations; practice final pitches)



Today’s class

- Measurement intro
- Measurement discussion
- Paper discussion led by Joy and Tanmay
- (Likely for next class:) Case study on unsupervised measurement

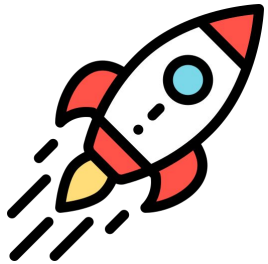
Where are we?

Identify Problem

Data Exploration

Algorithm Development & Validation

Tool Development



Deployment

Overarching Themes:

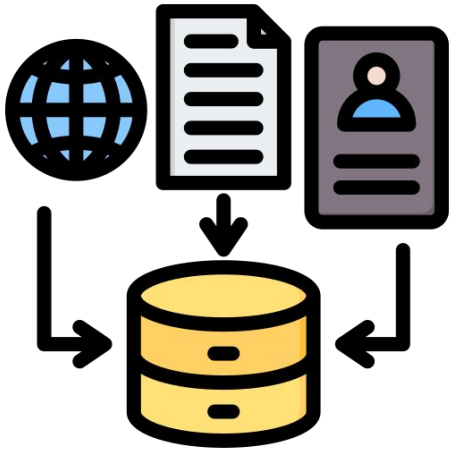
Bias & Fairness

Working closely with teachers

How do we define measurement?

Data

Structured (e.g. likert scale responses) / Unstructured (e.g. language)

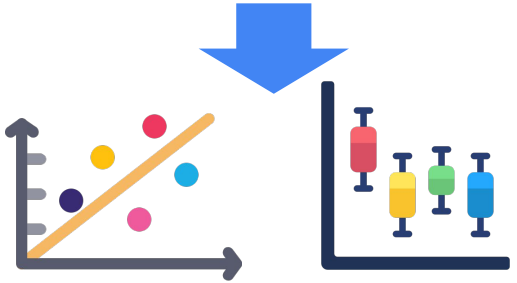


Score / Label

Measuring a target construct

Binary: Is this utterance on task?
Continuous: To what extent does the student feel empowered in this classroom?
Categorical: What is the topic of this lesson?

You can use these in quantitative analyses:



Most aspects of a quantitative research project / intervention / tool require measurement

- Identifying & analyzing teaching practices
- Evaluating fairness & bias
- Identifying need for intervention
- Understanding teachers' and students' perceptions of a tool
- Measuring outcomes
- …

NONE OF THESE ARE TRIVIAL TO MEASURE

Most aspects of a quantitative research project / intervention / tool require measurement

- Identifying & analyzing teaching practices
- Evaluating fairness & bias
- Identifying need for intervention
- Understanding teachers' and students' perceptions of a tool
- Measuring outcomes (e.g. student learning)
- …

Ex. Challenges and Issues

Subjectivity and context-dependence

Sparse data & oversimplification of demographic categories

High stakes

Low response rate & self-reporting bias

Choice of outcomes are often the most controversial

What type of NLP measures does your final project require?

Nobody has responded yet.

Hang tight! Responses are coming in.

Do you need to identify the measurement target?

E.g., type of classroom practice, dimension for user attitude (e.g. difficulty).

No: Skip to next step

Yes: Pick the target at the intersection of **promise & feasibility**

- Lit review
- Talk to people
- Look at data

Example brainstorming spreadsheet (list of discourse practices relevant to math ed)

Category	Feature Associated with Better Learning Outcomes	What to Measure	Examples
General pedagogical	student participation	number of words uttered by students/minute; and/or length of student utterances	
General pedagogical	test-orientedness	measure the number of references to standardized testing	MCAS, DCCAS
General pedagogical	instructional time	amount of instructional time vs off task time	procedural talk vs instructional talk; noise in the classroom
General pedagogical	wait time after questions	amount of wait time after questions	
General pedagogical	check-in on students	number of times teacher asks questions that check in to see if students are following along	“make sense?” “Any questions”? Thumbs up? Hold your boards up; “student .. looks puzzled”
Math specific	use of math terms	density of math terms from teachers & students; measure to what extent teachers press students to use such terms	angle, fraction
Math specific	use of sloppy (math) terms	density of sloppy terms from teachers and from students	borrowing, top and bottom, cancelling
Math specific	use of proofs, mathematical reasoning/explanation	presence of proofs, mathematical reasoning/explanation in teacher & student talk	
Math specific	degree of direct instruction & focus on memorization	estimate the degree to which the teacher is doing direct instruction	teacher talk with short student answers interspersed; words like remember, recall; first thing you do when you...what do you do next..we’re also going to have to do what?
Math specific	cognitive demand of (math) questions	estimate the degree of cognitive demand of questions (teachers & students)	why, explain, what does that mean, different, difference, compare, what’s missing, how do these relate
Math specific	teachers' evaluation of student contributions	see whether and how the teacher remediates students' misunderstandings;	correction, reformulation, repetition, praise
Math specific	uptake	degree to which teacher uses students' mathematical contributions in subsequent instruction; students' uptake of other students' ideas (with minimal teacher orchestration)	
Classroom climate	positive references to students	degree to which teacher uses student names in a positive way	positive: “Geoffrey’s idea” or “Marie, tell us what you are thinking” “I think Nonie solved the problem in the same way”; negative: Geoffrey!
Classroom climate	affirmation of knowledge and skill	degree to which teacher encourages students	“You totally understand this, you just need to tweak what you’re saying a little bit”
Classroom climate	broad regard	degree to which the teacher shows interest in the students' lives	asking non-academic questions

Does an NLP measure exist already for what you want to do?

Yes: Skip to validation (on your domain)

No: **Develop a measure** (in most cases) following the standard paradigm → next slide

Standard NLP measure development workflow

1. **Create high quality validation set**
   a. With sufficient # of examples to capture relevant variation (rule of thumb: at least 1k examples for a relatively straightforward measure, 2k for more subjective ones)
   b. When possible, create a held-out test set too (that you only evaluate on at the very end)
2. **Iteratively develop & validate model**
   a. Supervised paradigm: label training data → train classification/regression model
   b. Unsupervised/self-supervised paradigm: leverage unlabeled data

Standard NLP measure development workflow

1. **Create high quality validation set**
    a. With sufficient # of examples to capture relevant variation (rule of thumb: at least 1k examples for a relatively straightforward measure, 2k for more subjective ones)
    b. When possible, create a held-out test set too (that you only evaluate on at the very end)

WHAT IF CREATING A VALIDATION SET IS NOT AT ALL TRIVIAL BECAUSE THE CONSTRUCT IS HIGHLY SUBJECTIVE?

What to do if your **interrater agreement is fair to moderate**?

Even when working with domain experts & doing several rounds of rater training and discussion

First ask: why is agreement low?

Potential cause	Potential solutions
Poorly defined construct	Improve definition & coding scheme!
Context-dependence of construct	• (When possible) Add more context • (When appropriate) Pre-define context
Intersubjectivity (diff. people might perceive or react to the same thing differently)	This is important variation that you want to keep

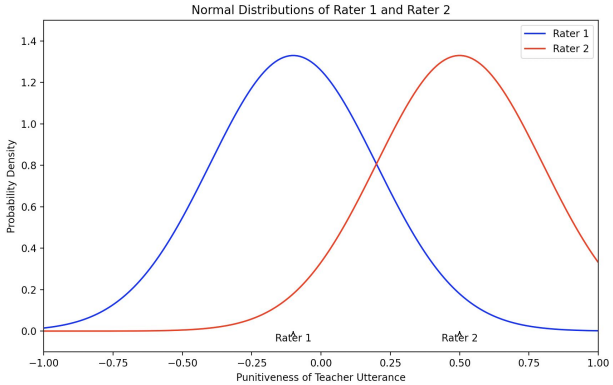
How to handle **inherently subjective** constructs?

During annotation

- Have multiple annotators (the more the better) for each example.

Processing Annotations

- Z-score judgments before aggregating
- Create different subsets of the data (subjective subset; less subjective subset) for evaluation



How to handle **inherently subjective** constructs?

During annotation

- Have multiple annotators (the more the better) for each example.

Processing Annotations

- Z-score judgments before aggregating
- Create different subsets of the data (subjective subset; less subjective subset) for evaluation

Modeling

- Incorporate confidence into the measurement
  - e.g. build a model that predicts rater agreement as a proxy for confidence
- Train representations separately for each rater and then combine them into a shared representation (Davani et al., 2022)

Validation

- Check if results are robust to variations in data or modeling decisions
  - e.g. leave-out validation
- Don't rely too heavily on your "ground truth" values
  - Correlate your measure with other relevant variables (e.g. overall instruction quality) to understand if it relates to positive or negative outcomes
  - Estimate the impact of applying your measure to address a specific issue

Application

- Don't use for making high stakes decisions!

Supervised vs un/self-supervised modeling for measure development

(((Regardless of choice, you still need labeled data for your validation set!)))

Supervised models	Un/self-supervised models
Pros: ● Tends to perform better when sufficient labeled training data is available	Pros: ● Does not need labeled data ● Tends to transfer better across domains
Cons: ● Model performance tends to correlate directly with amount of labeled data, which in turn is expensive to collect ● Performance often generalizes less across domains	Cons: ● Does not need labeled data ● Not available / gets complicated for many high-inference constructs

Language models leverage both approaches!!

Supervised modeling: LLMs or smaller models?

Smaller models (RoBERTa, BERT, etc.)	LLMs
Resources: https://simpletransformers.ai/; https://huggingface.co/docs/transformers/index	GPT-3.5; Llama 2; GPT-4 (instruct tuning)
Pros: ● Downloadable → more transparency & control ● Needs little compute ● Can achieve similar performance to LLMs when sufficient labeled data is available	Pros: ● Very good at few shot learning ● Can be tuned with instructions
Cons: ● Require more training data ● Can't be tuned with instructions or via interacting with the model	Cons: ● Most cannot be downloaded ● Many models can't be finetuned (e.g. GPT-4, Claude)

What is your experience with using smaller models vs LLMs for measurement?

Should we watch the example pitch (8 mins) and go over the rubric *during class*?