The Tufts Institute for Research on Learning and Instruction

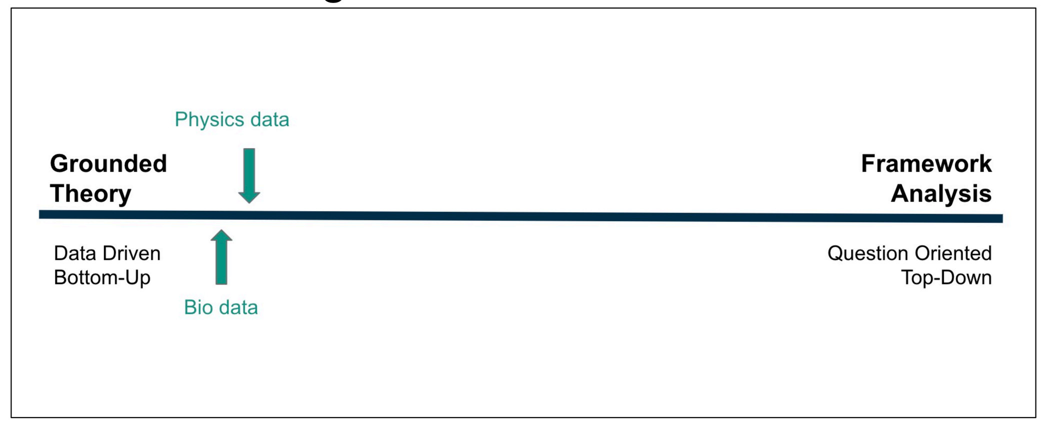
Is AI a learning scientist's new best friend? How natural language processing of qualitative coding can revolutionize educational research

Audrey McGlothlen; Alec Plano; Seixas Aldrich
Tufts University ; RAISE

The Tufts Institute for Research on Learning and Instruction

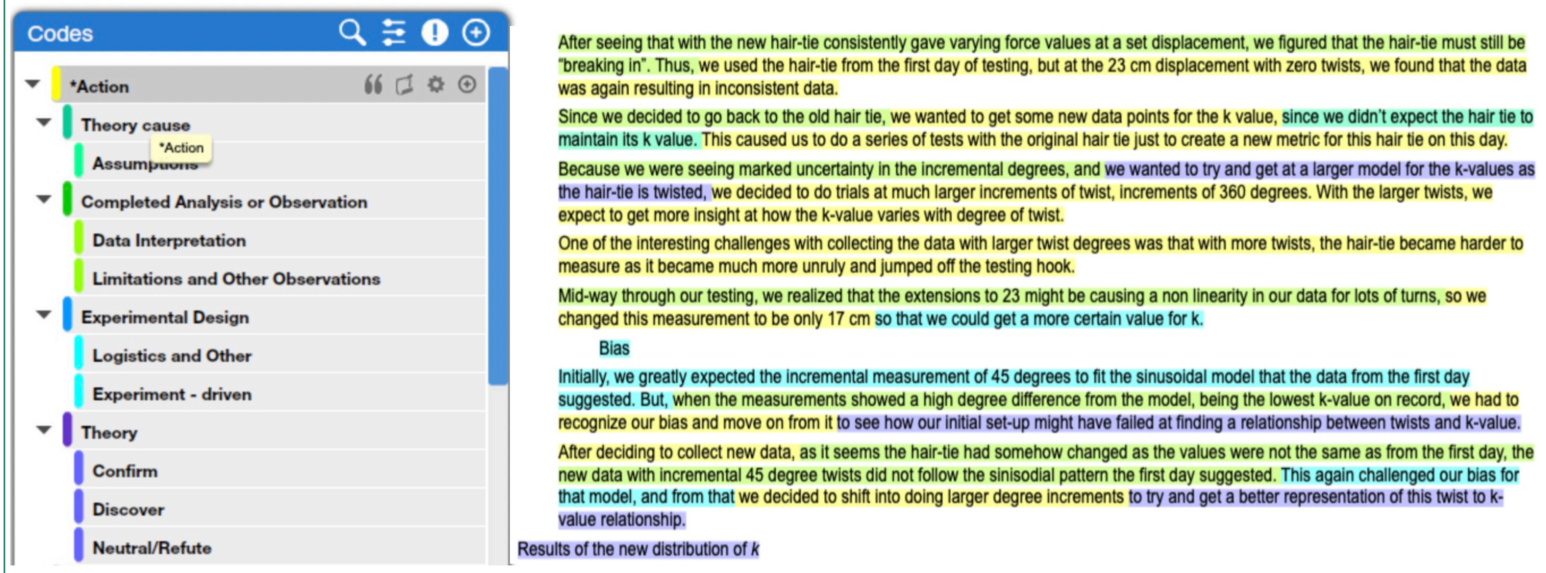
Background

- We were interested in seeing how machine learning could be applied to educational research contexts. This collaborative project brought together cognitive and learning sciences, STEM education, and machine learning experts from both Tufts University and Cornell University to experiment in the use of machine learning algorithms in understanding student's lab reports and lab notes.
- Benefits of this novel utilization of machine learning include being able to greatly expand the scope of learning science's research by sifting through significantly more data at higher rates.
- Analysis of student's reports rooted in literature on grounded theory and framework analysis. While very data-driven and largely bottom-up, there were key questions that informed what the researchers looked for in their qualitative analysis such as focusing on justification (physics) and instances of uncertainty (biology) when students constructed arguments.

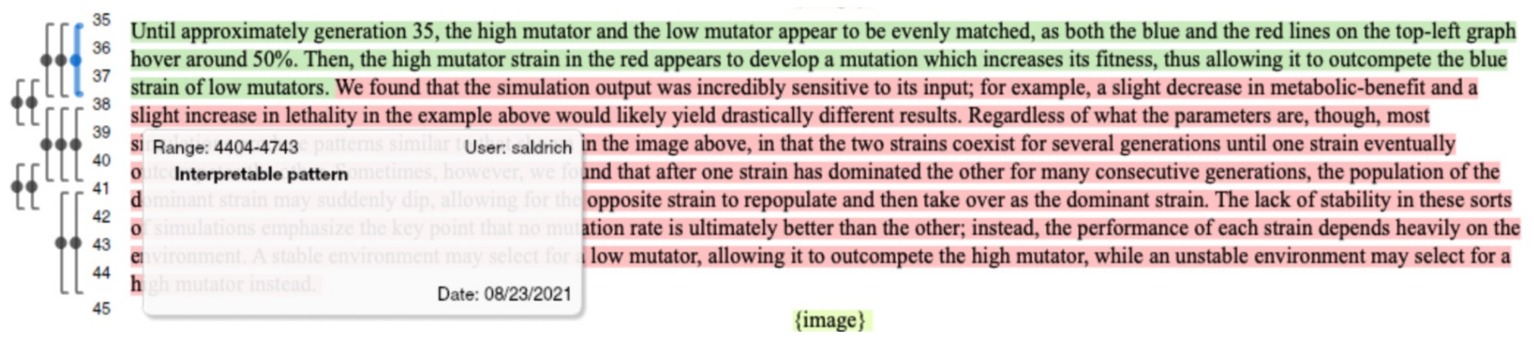


Examples

- Physics Lab Notes



- Biology Lab Reports



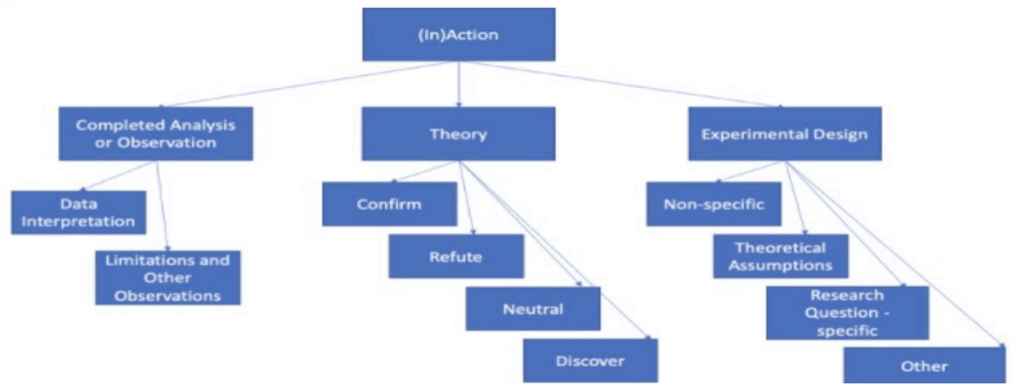
Methods for Physics Data

Student agency

Research Questions: How often are experimentation decisions justified in the lab notes? How often are these decisions based on past data analysis or existing observations? How often are decisions based on theoretical assumptions? When students consider how theory impacts their experimental decisions, are they attempting to confirm or refute a model, do they have a more neutral stance, or are they perhaps attempting to discover a new model? How do these forms of justifications vary across units and time?

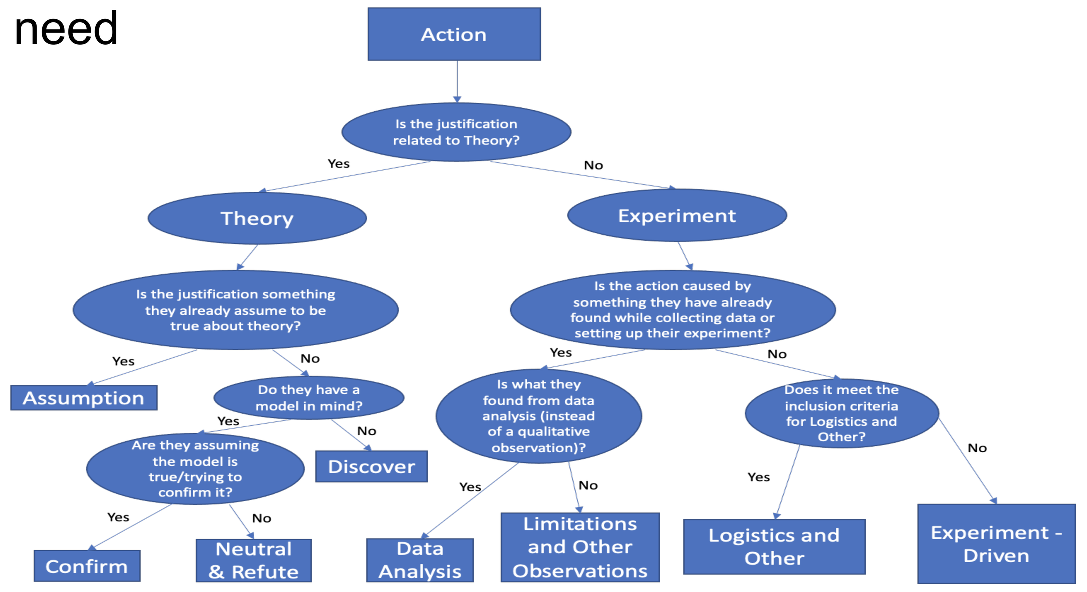
Development of Coding Scheme:

- Started with three categories: procedural decision making, epistemic decision making, and other (e.g. emotion and hedging).
- We cut the other category and focused on students' justifications of procedural decisions and interpretations of results.
- Finally, we cut the interpretation of the results section, although it was interesting, the scheme was too comprehensive for one round of analysis.



Implementing Coding Scheme:

- Made note of explicit actions students carried out in their experiments.
- Characterized justifications connected to those actions as
  - Assuming theory: A justification based on one or more assumption about how a model works
  - Explaining theory: Justifications for the aim to confirm a model, refute a model, or develop a novel representation
  - Based on previous data: An argument is made using evidence that has already been observed or analyzed
  - Based on experimental design: A justification that a particular action was taken to fulfill some experimental purpose or logistical need



Methods for Biology Data

Student Uncertainty

Research Questions: what do students do when met with uncertainty? Is there a correlation with uncertainty and how much conceptual diversity they cover?

Development of Coding Scheme:

- Started from data looking at the different concepts student chose to research, specificity of claims, where they got their data, and how they evaluated the quality of their conclusions.
- Initially hoped to characterize "what students' were up to" by following how the research question developed across the paper but the scope became unreasonable to capture in one scheme.
- Finally focused on uncertainty as a topic of inquiry, analyzing the source of the uncertainty and how students chose to account for it.

Implementing Coding Scheme:

- Characterized types of uncertainty/certainty:
  - Uninterpretable vs interpretable pattern
  - Contrary vs consistent to hypothesis/theory/expectation
  - Misalignment vs alignment between multiple replicates or data sources
- Characterized reactions to uncertainty:
  - Proposing an interpretation/explanation
  - Articulates expectation
  - Articulate research question
  - Concludes by revising an idea
  - Concludes by supporting an idea
  - Concludes by rejecting an idea
  - Identify possible source of error
  - Proposes new experiment or modification

Results

- Coded 160 physics notes
- Initial findings demonstrate the ability of the machine to cluster the physics notes by unit and structure

Future Aims

- Use machine learning algorithms (to be developed) to analyze physics data further
- Finalize bio codes and run it through machine
- Run unsupervised clustering of bio data

Tufts UNIVERSITY

References

1. [Cheuk, 2021] Can AI be racist? Color-evasiveness in the application of machine learning to science assessments
2. [Chen et al., 2018] Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity
3. [Corbin & Strauss, 1990] Grounded Theory Research: Procedure, Canon, and Evaluative Criteria
4. [Kelly & Takao, 2002] Epistemic levels in argument An analysis of university oceanography students' use of evidence in writing
5. [Phillips, Watkins, & Hammer, 2017] Problematizing as a scientific endeavor
6. [Robertson, McKagan, & Scherr, 2018] Selection, Generalization, and Theories of Cause in Physics Education Research
7. [Bartlett, 1932] Chapter X A Theory of Remembering
8. [Tannen, 1979] What's in a Frame
9. [Scherr & Hammer, 2009] Student Behavior and Epistemological Framing Examples from Collaborative Active-Learning Activities in Physics