Spring 5-30-2022

Developing a Natural Language Processing Approach for Analyzing Student Ideas in Calculus-Based Introductory Physics

Jon M. Geiger
Seattle Pacific University

Lisa M. Goodhew
Seattle Pacific University

Tor Ole B. Odden
University of Oslo

Developing a natural language processing approach for analyzing student ideas in calculus-based introductory physics

Jon M. Geiger and Lisa M. Goodhew
Department of Physics, Seattle Pacific University, 3307 3rd Ave W, Seattle WA, 98119

Tor Ole B. Odden
Department of Physics, Center for Computing in Science Education, University of Oslo, 0316 Oslo, Norway

Research characterizing common student ideas about particular physics topics has made a significant impact on university-level physics teaching by providing knowledge that supports instructors to target their instruction and by informing curriculum development. This work utilizes a Natural Language Processing algorithm (Latent Dirichlet Allocation, or LDA) to categorize student ideas, with the goal of significantly expediting the process of categorizing student ideas. We preliminarily test the LDA approach by applying the algorithm to a collection of introductory physics student responses to a conceptual question about circuits, specifically attending to whether it is useful for characterizing conceptual resources, or student ideas that may be fruitful for science learning. We find that for a large enough collection of student responses ($N \approx 500$), LDA can be useful for characterizing student resources for conceptual physics questions. We discuss some considerations that researchers may take into account as they interpret the results of the LDA algorithm for characterizing student's physics ideas.

I. INTRODUCTION

Over the last few decades, Artificial Intelligence (AI) has been increasingly useful in day-to-day life. From recommendation algorithms on popular streaming streaming services and e-commerce platforms [1] to the programming backing self-driving cars [2], the application of AI is vast and ever-growing in the twenty-first century. Natural Language Processing (NLP) is a branch of AI which has been defined as: "a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications" [3]. With advancements in computational power, NLP has been utilized to analyze enormous amounts of information in a short amount of time.

One application of NLP is known as Topic Modeling [4], which is used to extract themes or "topics" from large bodies of work. Latent Dirichlet Allocation (LDA) is a popular topic modeling algorithm which takes in a set of documents (called a "corpus" in the language of LDA) and produces clusters of words ("topics") which are commonly used together within those documents. From this output, researchers can then ascribe meaning to each of the topics produced by the algorithm. In the past few years, there have been several studies exploring the utility of NLP in Physics Education [5, 6]; more particularly, LDA has been utilized to characterize topics in Physics Education Research (PER) over the last couple decades as a tool for new PER researchers [7, 8]. As LDA is not limited to PER, we also see examples of its use in other fields such as software and banking [9, 10].

A common research focus in PER is the extensive investigation of common, topic-specific knowledge that students bring to the classroom. This kind of research has important impacts on physics instruction, particularly at the university level: it informs the development of research-based instructional materials (e.g. Tutorials in Introductory Physics, Maryland Open-Source Tutorials, ACORN Physics Tutorials [11, 12]) and it contributes instructors' knowledge of student ideas, an important part of the knowledge that instructors use to teach [13]. In physics, research identifying students' common physics ideas has investigated students' common, incorrect ideas (misconceptions, alternative conceptions, or difficulties [14, 15]) and, less extensively, students' common, potentially-fruitful ideas (p-prims, facets, or conceptual resources [16–19]). Resources are activated in context-sensitive ways (attentive to the particular question at hand, the social context, etc); because of this, research to identify *common* resources attends to variations in the student population and the question asked to elicit them to a greater level of detail than difficulties research does. This makes the work of appropriately characterizing resources very time-intensive, which may limit the impact of resources research.

The time-intensity of characterizing student ideas motivates the work presented here, which aims to support researchers in characterizing students' conceptual resources by investigating the extent to which LDA can be used to identify common, potentially-fruitful student ideas. We propose a method by which we can automate part of the process of characterizing student ideas by applying LDA to a corpus of student responses to a particular conceptual physics question. By inspecting the different words within a topic, then looking at documents which most reflect that topic, we can understand the distinctions between ideas in the corpus. This paper builds on previous research by using this approach to analyze students' responses to physics questions, which often include a mixture of technical and informal language.

The questions which guides our research are: to what extent can LDA be used to characterize patterns in student thinking? Can we produce a useful, time-saving method for researchers which yields "instructionally-useful" student ideas for conceptual resources research?

II. METHODS

A. Student Task

In order to analyze patterns in student thinking, many responses to the same question were needed so that clusters of ideas can be analyzed. For the primary analysis, we looked at a circuits question which primarily elicits an explanation of student reasoning rather than simply asking for a correct answer. In many cases, these questions give the students a scenario and explains what happens, then ask the student why that outcome makes sense to them. We also performed this same analysis on two questions relating to heat and temperature, as well as one question on propagation of waves. These will be included as case studies in the Results section.

B. Data Cleaning

We used data gathered from several universities across the United States by members of the Conceptual Resources team. Student responses were either recorded on paper, scanned, then transcribed into CSV files, or gathered digitally. For our data set on responses to our Temperature question, we were able to combine together responses to the same question from multiple universities in order to increase the sample size and algorithm stability, described in the next section.

We loaded the responses into Python for cleaning. The following steps were used in order to prepare the data for modeling:

1. Removed punctuation (quotes, commas, periods, and parentheses).
2. Removed stopwords (commonly used words such as "a," "the," and "is").
3. *Lemmatized* words down to their roots ("increased" and "increases" would become "increase").
4. Created *bigrams*, (pairs of words such as "potential_difference" or "ohm_law"). This is necessary in

order to distinguish between concepts such as "potential difference" and "potential energy," which have distinct meanings in physics, even though they both contain the word "potential."

5. Filtered out the most and least common words (based on user-defined thresholds). This process is explained below.
6. Created a bag-of-words for the LDA function [20].

The most and least common words were filtered out based on criteria chosen by the researchers. Filtering out the most common words involves choosing a threshold of a percentage of all documents in which a certain word occurs. For example, in a circuits question, the word "current" may appear in 70% of all responses, so we may choose to eliminate all words which appear in more than 50% of all documents. Words found in a majority of documents likely occur in the problem statement itself, and these words may therefore be unhelpful in characterizing distinct student ideas. In the algorithm, this is our "no above" threshold. Filtering out the least common words involves choosing a minimum number of documents in which a word can occur. Words that appear in one or two documents may include fanciful words used by only one student, or misspellings of words ("increaes" rather than "increase"). More common misspellings or typos (such as "becuase") tend to appear in more than just one document. Because of this, and due to the corpus sizes used ($N < 500$), we generally chose to exclude words that are only found in two or three documents. In the algorithm, this is our "no below" threshold.

C. LDA Modeling

Latent Dirichlet Allocation (LDA) is a Natural Language Processing algorithm used for topic modeling [4]. In brief, the algorithm works by taking in a corpus (or collection) of documents, noticing groups of words that commonly occur within certain documents, and picking out those groups of words, labeling them as "topics." Then, each document is assigned a set of weights according to how relevant to each topic that document appears to be. More technically, LDA iteratively "learns" topics by creating and adjusting a probabilistic model for how words are distributed in topics, as well as how topics are distributed among documents.

The LDA modeling process relies on a few key assumptions. Firstly, LDA assumes that the order in which words occur in a document doesn't matter, nor does the part of speech or anything else about the word. Second, we assume that each document is comprised of a percentage mixture of all of the topics. That is, for a three-topic model, a given document could be comprised of 70% Topic 1, 20% Topic 2, and 10% Topic 3. For a seven-topic model, a document could be comprised of 99% Topic 1, with 1% split among the other six topics. This is important for interpreting the results of our model.

There are, in essence, two hyperparameters in the algorithm (α and β) which can be controlled by the modeler. The first parameter, α, ranges from zero to infinity and controls the mixture of topics within a document. A low value of α (less than one) makes the algorithm assign topics to documents fairly exclusively, such that each document is comprised of a small number of topics. A high value of α creates a more heterogeneous mixture, where topic weights within a document can be closer in magnitude than they would be with a high α. The second parameter, β, is similar to α and ranges from zero to infinity, and it controls the mixture of words within a topic. A low value of β makes the algorithm assign words to topics fairly exclusively (similar to α), whereas a higher value of β allows a given word to be assigned to topics more equally. For our models, the values of α and β were chosen automatically by Gensim [21], which we used to implement the LDA algorithm. While the particular details of the mathematics are beyond the scope of this paper, more in-depth descriptions of the mathematics and intuition of LDA can be found in Blei *et al.* [22] and Odden *et al.* [7].

In the context of our research, the corpus of documents was the collection of all student responses for a single question, and a document was a single student response within that corpus. The research we conduct looks at the extent to which a topic can represent a student idea. In this setting of the framework, any given student response would be comprised of a distribution of student ideas, with the heterogeneity of the topic-document distribution controlled by the hyperparameter α. The student ideas might be fairly distinct, implying that a low value of β may be useful. For each of the examples, we built models with a varying range of topics from three to seven, computing coherence values for each model. The coherence is a measure of "the tendency of the top words in the topic to co-occur"[7], where higher coherence values (closer to one) describe more distinct topic distributions. In selecting which topic number to use, we generally chose the topic number which yielded the highest coherence value, though this was not always the case. This was done heuristically, requiring input from the researchers on which topics looked to be the most distinct or aligned with pre-conceived mental models.

Because LDA relies on an initial randomization, it is important to note that between random seeds, there is some run-to-run variation in the different topics which the model converges on. The topics presented in this paper were chosen from just one particular run of the model; results varied when a random seed for the algorithm was not specified.

D. Representative Responses

We chose to display the most representative student response for each modeled topic in order to give a more concrete idea of what each topic actually represents. Without displaying the student responses, there would be no method by which we could convert from a topic (a weighted group of words) to an actual student idea. Due to the assumption of LDA modeling that each document is written on one or more topics, contingent upon the success of our model, this translates to "each student response is comprised of one or more student ideas." Because each response has a set of values associated with it representing the amount that each topic is present in that document, we sorted all of the documents by a given topic, allowing us to see the most prevalent responses in that topic.

Finally, we performed a secondary qualitative data analysis on the results of the LDA analysis to determine the degree to which topics resembled student ideas. This involved looking at the topmost words in a topic, extracting phrases in the student responses which contained these top words, synthesizing those phrases into coherent ideas, and matching those ideas up with existing mental models or student ideas. This qualitative step is crucial to the process of categorizing patterns in student ideas, as the researcher must use their existing physics knowledge in order to assign actual meaning to the model's topics. Examples of this will be shown in each of the case studies.

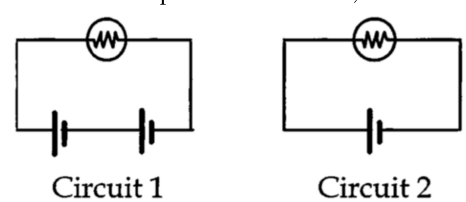
III. RESULTS

As stated above, we used data from four different questions: one from circuits, two from heat & temperature, and one on propagation of waves. In this section, we present the question asked of the students, the parameters used in the modeling process, the topics characterized by the model, and some parts of some representative responses from each topic. These representative topics are then used to formulate the student ideas as outlined in the prior section, with a brief discussion on the final topics.

Case Study: Circuits

Students were tasked with answering the following question (see Figure 1), with $N = 483$ responses:

FIG. 1. Compare Bulbs in Series, Revised



Comparing the brightness of the bulbs in Circuits 1 and 2, we observe that the bulb in Circuit 1 is brighter. Using Ohm's Law, $V = IR$, we know that current and voltage are directly proportional when resistance is the same. Why do you think more voltage leads to more current? What mental models are you using to make sense of this?

For our circuits case study, we chose the following model parameters:

- $k = 5$ topics.
- "No above" = 50%. Words like "current" and "voltage" appeared in more than 50% of all responses, and would not be useful in characterizing student ideas.
- "No below" = 3. Setting this threshold any higher would have excluded important words potentially significant for student ideas, such as "gravitational."

The topics produced by the model are shown in Table I.

This five-topic model yielded a coherence score of 0.4625, which was the highest among the 3-10 topic models. Table II shows some phrases from the most representative student responses for each topic.

From the selected phrases of student responses in Table II, we can begin to code some distinct student ideas for each topic, as previously described in the Methods:

1. Current flows through circuits like water through a river or pipe.
2. Voltage "pulls" electrons; more V means more I.

TABLE I. Top words in topics with weights, Circuits

Topic	**Words in Topic** (with weights)
1	water (.090), large (.053), flow (.053), think (.044), energy (.042), pressure(.040), like (.028), great (.026), push (.023), electron (.023)
2	electron (.074), force (.057), great (.044), high (.041), big (.038), mean (.033), faster (.031), potential (.027), push (.023), think (.022)
3	increase (.053), high (.052), charge (.050), electron (.039), great (.033), battery (.031), lead (.029), flow (.027), mean (.027), potential_difference (.026)
4	battery (.081), circuit (.069), power (.069), bulb (.048), increase (.043), brighter (.036), think (.029), push (.028), lead (.024), double (.022)
5	increase (.163), resistance (.083), circuit (.048), constant (.041), equation (.027), mean (.024), bulb (.024), ir (.024), ohm_law (.022), change (.022)

TABLE II. Selected phrases from top responses, Circuits

Topic	Representative Student Responses
1	"**water flows** through a river **like** current **flows** through circuits" "if there is more potential (or, less accurately, **pressure**) ... more **electrons** will **flow**"
2	"the more **potential** ... means the more **force** there is **pulling** [the **electron**] to where it wants to be" "more voltage ... results in a harder **pull**, and therefore the electrons speed up more ... **faster** electrons will result in a **higher** current"
3	"if the voltage of the **battery** increases ... the **potential difference** ... **increases** ... the **battery** must push more charge through it to maintain the voltage" "a **greater** magnitude electric field ... **increases** the amount of **charge** that passes through a ... circuit ... **increasing** the current"
4	"second **battery** ... **doubles** the voltage ... **increasing** the current **increases** the **power**" "a **brighter bulb** represents more **power** and therefore more current"
5	"rearranging the equation $V = \boldsymbol{IR}$ ... in terms of the **equation** ... a change in one value affects the other ... if R is **constant**, I must **increase**."

3. Voltage "pushes" electrons; voltage is potential difference, causing a greater electric field.
4. Batteries increase the power in a circuit. More power means brighter.
5. Voltage is proportional to current. With resistance held constant, increasing the voltage increases the current.

For this particular set of responses on circuits, it appears that LDA worked fairly well in categorizing distinct student ideas about circuits. Topic 1 was the most distinct and stable topic, in that each run of the algorithm produced some topic related to the flow of water. Though there do appear to be some similarities between Topic 2 and Topic 3, we noted that in the top three most representative responses from each topic, students who wrote on Topic 2 exclusively referred to voltage/potential as a "pull," whereas students who wrote on Topic 3 exclusively referred to voltage/potential as a "push." Topic 4 refers to the idea that "batteries provide power," relating the conventional notion of power (i.e. a "power outage") to the circuits definition ($P = IV = I^2R$). Lastly, Topic 5 refers directly to Ohm's Law, in which students will use the equation itself as a mental model for why increasing the voltage should increase the current.

Case Study: Temperature

The student task was to answer the following question ($N = 248$ responses):

> *Imagine you have two room-temperature blocks made of the same metal, but one has more mass than the other. You drop the blocks into equal volumes of 5-degrees-C water, count to five, and then dump them out onto a table. You measure the temperature of both blocks, and the less massive one is colder than the more massive one. That's because the mass of an object matters for how much its temperature changes.*
>
> *What we want to know is why that makes sense or doesn't make sense to you: Why is it that a less massive block changes its temperature more than a more massive block made of the same material and at the same starting temperature?*

For our temperature case study, we used the following model parameters:

- $k = 5$ topics.
- "No above" = 50%. Words like "block," "temperature,", and "change" appeared in more than 50% of all responses, and would not be useful in characterizing student ideas.
- "No below" = 2. Setting this threshold any higher would have excluded important words potentially significant for student ideas, such as "inertia" and "thinner."

The topics produced by the model are shown in Table III.

The coherence score of this particular set of topics was 0.2805. Though this value is significantly lower than the coherence score from the Circuits case study, there was significant run-to-run variation in the coherence scores for this model. We also noted that changing the "no below" threshold drastically changed the coherence scores. Including more words seemed to drive the coherence scores downwards, though further study is required to fully understand this effect. Table IV includes representative responses from these topics.

TABLE III. Top words in topics with weights, Temperature

Topic	**Words in Topic** (with weights)
1	particle (.156), energy (.116), average (.049), molecule (.043), mass (.030), massive (.029), great (.027), kinetic (.025), need (.023), make_sense (.022)
2	mass (.175), object (.112), small (.053), heat (.053), material (.044), make_sense (.033), large (.024), need (.018), mean (.017), cool (.017)
3	massive (.206), water (.040), heat (.040), faster (.039), material (.039), cool (.026), absorb (.025), drop (.024), small (.024), make_sense (.021)
4	volume (0.085), massive (0.059), surface_area (0.051), small (0.051), take (0.048), time (0.041), large (0.041), heat (0.038), cool (0.036), water (0.034)
5	energy (0.167), massive (0.096), thermal (0.054), mass (0.047), atom (0.034), require (0.032), material (0.027), take (0.021), large (0.020), specific_heat (0.020)

TABLE IV. Selected phrases from top responses, Temperature

Topic	Representative Student Responses
1	"temperature is a measurement of the **average energy** of **particles** in an object" "**energy** per **particle** or per **mass** would be less in the more **massive** block ... you will have to change the **energy** of more **particles**"
2	"I know the equation $Q = mc\Delta T$ ... it would take longer to **heat** up something with a bigger **mass** than something with a smaller **mass**" "the only difference between them is their **mass** ... $Q = mc\Delta T$" "In the equation $Q = mc\Delta T$, when **mass** is doubled then **heat** is doubled"
3	"the less **massive** block initially carries less **heat** than the more **massive** block does before going into the **water**" "a less **massive** block ... has less substance ... [and] it can change quicker than the more **massive** block"
4	"**heat** will enter through the *surface* of the block ... **volume** increases faster than **surface area** when increasing the size of the block" "the **surface area** to **volume** ratio of a **smaller** block is higher than that of a larger block, so it **absorbs**/releases more **heat**"
5	"it takes a certain amount of **energy** to change a given **mass** of a material by some temperature" "a less **massive** block has fewer units of matter for which **thermal energy** can be distributed" "the amount of **energy** that a **material** can store as temperature, otherwise known as **specific heat**, is a measure of the **thermal energy** stored per unit **mass**"

1. Temperature is related to the energy of particles in an object.
2. Heat transfer is related to mass ($Q = mc\Delta T$).
3. "Massive-ness" of the object determines the rate at which the object heats up.
4. Heat enters through the surface of an object. Surface area and volume contribute to heat transfer rate.
5. Heat is thermal energy. More massive objects can distribute thermal energy over more material.

Despite the relatively low coherence score of the model, it appears that LDA was able to pick out some relatively distinct topics. Topic 1 appears to be distinctively about the energy of the particles in each of the blocks. The idea of energy comes about as well in Topic 5, but the difference is that Topic 5 discusses thermal energy in relation to heat, whereas Topic 1 relates the amount of energy an object has to the amount of particles it has. Topic 2 is an interesting result: by looking at the words in the topic, there is very little indication that this topic would reference the equation $Q = mc\Delta T$. These responses included this equation written in a variety of different ways, including "`Q=mcdeltaT`," "`Q=mc /\T`," and "`Delta Q = m c delta T`." Because of this, the LDA model didn't pick up on the equations themselves (See Table I word "ir"), but rather students' semantic explanations of these equations.

One issue we noticed immediately is that in Topic 3, there is one word which is weighted five times more than the second-most prevalent word. The word "massive" appeared in many student responses because it was a word contained within the prompt, and the algorithm seemed to settle that this topic was people who referred to the blocks as being either more or less "massive" than the other. This word also appears in the top ten words of every topic but Topic 2, which implies that perhaps this word should have been excluded in the data cleaning step. Topic 4 is specifically about the surface of the block and how it relates to the rate at which heat is transferred into the block. Many student responses included direct reference to the "surface area to volume ratio," or the "volume increasing faster than the area" or the "square cube law," implying something deeply geometrical about this student idea.

Case Study: Heat

The student task was to answer the following question ($N = 191$ responses):

> *You may have heard that "heat" or "thermal energy" transfers from hot to cold objects, and not the other way around. Why is this the case? How do you make sense of this phenomenon?*

For our heat case study, we used the following model pa-

TABLE V. Top words in topics with weights, Heat

Topic	**Words in Topic** (with weights)
1	thing (.101), hotter (.045), temperature (.036), colder (.030), cool (.025), sense (.024), particle (.023), tend (.022), kinetic (.019), equilibrium (.019)
2	high (.078), particle (.068), low (.065), spread (.040), move (.037), entropy (.035), movement (.033), state (.030), concentration (.021), lack (.016)
3	entropy (.045), molecule (.036), lack (.034), equilibrium (.025), atom (.023), hotter (.022), want (.021), way (.020), gain (.019), reach (.018)
4	equilibrium (.068), go (.033), colder (.030), reach (.030), make_sense (.030), entropy (.029), way (.028), state (.025), thermodynamics (.023), law (.023)

rameters:

- $k = 4$ topics.
- "No above" = 30%. Words like "energy," "cold,", "object," and "hot" appeared in more than 30% of all responses, and would not be useful in characterizing student ideas.
- "No below" = 2. This was enough to remove typos and misspellings, but setting this threshold any higher would have excluded words which carry meaning such as "conserve" and "external."

The topics produced by the model are shown in Table V.

The coherence score of this particular set of topics was 0.5041. Table VI includes representative responses from these topics.

From the phrases picked out from the representative student responses, we code our student ideas as follows:

1. Thermal energy tends to spread out/equilibrate, causing hot objects to cool.
2. Energy moves toward basic states, or diffuses like particles.
3. Hot states are unstable. Energy "wants" to distribute evenly.
4. Systems tend toward equilibrium.

Unlike the questions on temperature and circuits, these topics seemed to blend into each other to a large extent. A large overarching theme in all of these topics is that energy wants to "equilibrate," which is seen in every topic in some way or another. This could be attributed to the relatively low sample size, or perhaps to many students simply thinking about energy going from hot to cold in terms of equilibrium. It does appear that the algorithm picked out responses which referred specifically to "basic states" in reference to entropy, as well as hot states being "unstable" and the system "wanting" to go toward equilibrium.

TABLE VI. Selected phrases from top responses, Heat

Topic	Representative Student Responses
1	"heat/thermal energy is a measure of the **kinetic** energy of each **particle** in a material, and **things** get **hotter** as they gain more energy" "**things tend** to fall into **equilibrium**"
2	"it's the same as diffusion: **particles** will **move** from areas of **high concentration** to **low concentration** ... **entropy** tending to want a **higher** number of basic **states**" "a system tends to **move** toward having the most basic **states**, or the greatest **entropy**" "this fulfills the law of **entropy** ... heat will **spread** in a way that creates the greatest number of basic **states**"
3	"all things **want** to be at lower energy level because that is more stable" "being hot is unstable, so the object '**wants**' to remove its heat to become more stable"
4	"the system always tends toward the **equilibrium state**, the stable **state**" "the heat therefore **goes** to where it can most even out" "energy would travel to an area where there is less energy, for equal distribution of energy"

Case Study: Waves

The student task was to answer the following question ($N = 318$ responses):

> *Consider the following two scenarios: In scenario 1, your Teaching Assistant (TA) creates a pulse by flicking the end of a spring ... In scenario 2, your TA pulls the spring so that it is more taut (i.e., increases the tension in the spring) and then creates a pulse by flicking the end of the spring in the same way. The pulse in scenario 2 travels down the spring **faster** (i.e., has a larger speed) than the pulse in scenario 1.*
>
> *Why would it make sense for a pulse to move faster on a higher-tension spring? (We're trying to understand your intuition, not not whether or not you can remember particular equations. In other words, we want to know how you make sense of this phenomenon.)*

For our waves case study, we used the following model parameters:

- $k = 8$ topics.
- "No above" = 20%. Words like "energy," "cold,", "object," and "hot" appeared in more than 30% of all responses, and would not be useful in characterizing student ideas.

TABLE VII. Top words in topics with weights, Waves

Topic	**Words in Topic** (with weights)
1	frequency (.127), change (.123), wavelength (.066), property (.062), material (.044), pitch (.034), end (.033), guitar (.027), depends (.027), large (.025)
2	length (.060), equation (.053), sqrt (.051), mass (.047), density (.043), root (.038), mu (.038), decrease (.032), unit (.029), square (.027)
3	move (.056), pull (.053), cause (.041), amplitude (.033), tighter (.031), flick (.029), small (.027), medium (.027), scenario (.027), molecule (.023)
4	energy (.099), large (.081), amplitude (.080), pull (.033), taut (.032), molecule (.030), transfer (.026), sense (.026), medium (.022), scenario (.020)
5	rope (.091), point (.047), slack (.045), time (.044), distance (.038), low (.029), pull (.027), take (.025), tight (.024), scenario (.023)
6	particle (.211), pull (.094), time (.062), shorter (.051), mean (.047), motion (.037), energy (.032), period (.027), position (.025), great (.022)
7	equilibrium (.073), restore_force (.056), cause (.040), medium (.032), position (.031), quicker (.030), return (.030), quickly (.029), move (.028), want (.028)
8	medium (.106), great (.093), scenario (.033), act (.026), second (.025), like (.023), move (.022), result (.022), pull (.021), displace (.019)

- "No below" = 3. This was enough to remove typos and misspellings, but setting this threshold any higher would have excluded words which carry meaning such as "kinetic" and "horizontal."

Before discussing the topics, something was notable about this set of data in particular. As we ran different models, we noticed that there were repeated responses being output by the model, and that one topic was just one response, repeated over and over. We looked back at our original data, and noticed that our original data set had multiple entries of the same response. In particular, there were two responses which were repeated by between five and fifteen students. One of the repeated responses was repeated between students almost verbatim, and the other was notable because they referred to the spring/string as a "wire," which is something no other students did. Because of this, we removed all of these similar responses from the data set, leaving one of each behind. The rationale for this decision is that several students likely copied off of one person's response, and so it's very likely the case that these repeated responses only represent one student's idea.

After the further *ad-hoc* data cleaning, the number of responses went from $N_{\mathrm{pre}} = 318$ to $N_{\mathrm{post}} = 279$. The topics produced by the model are shown in Table VII.

The coherence score of this particular set of topics was 0.3966. This was the highest coherence score among the 3-10 topic models by a significant margin. Table VIII includes representative responses from these topics.

TABLE VIII. Selected phrases from top responses, Waves

Topic	**Representative Student Responses**
1	"I know strings with higher tension have a higher **pitch** because I play the **guitar**. **Pitch** is determined by **frequency**" "the length of the string is ... increased when tension is applied ... the **wavelength** would also increase"
2	"the speed equals the **square root** of tension/**mu** ... when the tension increases the rope gains a little bit of **length**" "when the spring is pulled and ... stretched out, it not only has a higher tension but also a smaller **mass** per **unit length**"
3	"the tense rope **pulls** forces in the horizontal direction, making the wave **move** faster [in] that direction, rather than making the **amplitude** higher" "the pulse would **move** faster because it was **pulled tighter**"
4	"**energy** is transferred quicker through a tighter **medium** ... on a spring with a **larger** tension, more **energy** is **transferred** horizontally rather than vertically as **amplitude**"
5	"when the string is **tight**, it creates a straight line" "when you **pull** something tighter you are **taking** the **slack** out of it ... you make it easier to manipulate the actual string" "more tension means each small unit of the spring can communicate the movement to the next small unit of spring quicker"
6	"we are **pulling** the **particles** away from each other and this will increase the ability of a **particle** to **pull** its neighbor **particle**" "with **greater** forces between the **particles**, the **shorter** the **time** for the **particle** to complete its **motion**"
7	"if the tension is higher, the string will have a higher **restoring force**, **causing** the individual pieces of the rope to '**want**' to **return** to the **equilibrium** point faster" "higher tension ... larger **restoring forces** ... larger acceleration for the string to **return** to its **equilibrium position**"
8	"wave **moving** through the string on a molecular level, the molecules of the **medium** become **displaced** as it passes through" "a **greater pulling** force causes **greater** tension between the differential elements of the spring"

Due to the large number of topics, we label these topics as student ideas devoid of a concerted effort to ensure distinctness of each topic. After coding the responses, we assess which topics blend into one another.

1. Tension (e.g. in a guitar string) changes the frequency/pitch, which is related to pulse speed.
2. Pulling on the spring changes the length, tension, and mass density. Changing T and μ affects pulse speed.
3. Pulling the string horizontally affects how quickly a pulse propagates.
4. Energy is transferred quicker through a tighter medium.
5. A tight string responds quicker to disturbances.
6. A spring can be modeled as a series of connected particles.
7. Tension affects the restorative force on a part of the spring. Higher restorative force means the spring will equilibrate faster.
8. Pulling creates a tension between small parts of the spring.

Most of these topics represent fairly distinct student ideas about how wave propagation is related to the tension in the spring. Topic 1 clearly outlines the relationship between pulse propagation and frequency/pitch, as in a guitar. Topic 2 seems to convey the idea that by pulling on the spring, the length of the spring changes, thus changing the mass density of the spring. According to students whose responses fell within this topic, pulling on one end of the spring increases the overall length, which decreases the mass per unit length, affecting the speed of the pulse due to it being a "lighter" medium. Topic 4 was very clearly about energy and its transfers through a tight medium, though responses related to this idea also appeared in Topic 5. Topics 6 and 7 also appear to be fairly distinct, with Topic 6 relating to the model of a spring as a series of small particles attached to one another, and Topic 7 relating tension to the restorative force which a small piece of the spring experiences.

While we couldn't decipher a coherent, distinct student idea from Topic 3, the responses which fell under this topic seemed to relate some sort of horizontal force to the spring being pulled tighter. Similarly, Topic 5 was somehow related to the spring/rope being slack, but responses under this topic were generally hard to decipher and it was difficult to pull out any student ideas. The idea that pulling a string taut creates a "straight line" lends itself to a similar notion as we got from topic 4, that something is "transferred quicker" through a tighter medium than through a slack one. Additionally, we can see a relationship between Topic 6 and Topic 8 in the sense that students model a spring as a series of very small particles which interact with one another. Topic 8 in particular seemed to be fairly incoherent by looking at the different student responses; in the top four student responses, the word "medium" only occurred in one response, and the word "great" only occurred in one (different) response.

Prior research and analysis of this question has found three common conceptual resources: "(i) properties of the medium either impede or facilitate the motion of the pulse, (ii) the speed or duration of transverse motion affects pulse speed, and (iii) the speed of the pulse is affected by its kinetic energy" [19]. The student ideas which we have categorized from this set of responses do not reveal the same broad categories of conceptual resources as in Goodhew *et al.*, however we do see similar themes in talking about tension affecting the speed of propagation ((i) with Topics 1–3, 5), the model of a spring as a series of particles ((ii) with Topics 6–8), and energy ((iii) with Topic 4).

IV. CONCLUSION

The research questions guiding this investigation were: *To what extent can LDA be used to characterize patterns in student thinking?* and *Can we produce a time-saving method using LDA that yields instructionally-relevant student ideas?* In this preliminary study, LDA was used to characterize distinct student ideas about circuits, heat & temperature, and propagation of waves. While the model on its own was able to produce relatively distinct topics with relevant physics words, we found that researcher interpretation was necessary in translating the output of the model to disciplinarily meaningful conceptual resources.

A primary goal of this analysis was to propose a method by which investigations of common student ideas can be streamlined by means of automation, rather than to not to propose a representative set of misconceptions or conceptual resources about circuits. Our results suggest that LDA can be used to support researchers in characterizing student ideas, but it does not remove the need for researcher interpretation. To support further development of a semi-automated method for characterizing student ideas about physics topics, we have discussed some ways in which researchers' decisions about the LDA model (e.g. choices about α and β parameters or the number of topics) can affect the topics it yields. We have also described the methodological steps taken to interpret instructionally-relevant student ideas from algorithm-generated topics. These choices about the model and the interpretive steps that follow suggest a framework for how future work can use LDA to characterize student ideas.

Limitations of this approach for categorizing student ideas stem from the relatively small sample sizes, assumptions and hyperparameters of the model, and the heterogeneity of the texts. Typically, when LDA is performed, in order to get any meaningful results, the corpus contains upwards of $N = 1000$ documents, preferring a large amount of small documents over a small amount of large documents [7]. With $N = 483$ responses for our question on circuits, the run-to-run topic variation was not incredibly high, though there were some runs which were noticeably different than others. When we ran this model on a question which had fewer responses such as our heat question ($N = 191$), the run-to-run variation in topics was much higher, with the topics less distinct and tougher to interpret. Additionally, with regards to the case study on waves, despite the 8-topic model scoring the highest coherence value out of any of the 3-10 topic models (by a relatively wide margin of ≈ 0.1), prior research has categorized only three conceptual resources from this particular question (as mentioned in the case study). Because of this, much researcher discretion is required in choosing the optimal number of topics to discover in any given question.

Perhaps a larger limitation, however, is the need for a secondary qualitative analysis in order to match model results with theoretical constructs, such as resources. Researchers should be aware that in order to get meaningful results, a large amount of student responses should be collected. Because this use of LDA is driven by the desire for a time-saving approach for researchers, a smaller sample size of responses would be easier to hand-code, and interpreting model results would be more tedious. For a larger sample, however, hand-coding can take a huge amount of time, but interpreting model results can be much easier.

There is no theoretically-based method choosing the "best" hyperparameters α and β for the LDA algorithm [23]. Rather, these hyperparameters should be chosen based on the research goals and aims. In our model, we allowed these parameters to be estimated automatically, whereas there may be some reason to believe that lower values of α and/or β might be useful to highlight the distinctness of the various topics produced by the model. Future work could include an exploration of how tuning the various model parameters affects the coherence scores or the perceived usefulness of the model output.

One key drawback to having no theoretically-based method for tuning hyperparameters is visible in the representative student responses. For each topic, when we selected the most representative student responses for each topic, it was invariably the case that a given response scored very highly in one topic, with all other topic weights in that document being close to zero. In reality, though, as we interpreted topics as student ideas, we could clearly see that many responses were comprised of multiple student ideas. This LDA modeling process seems like a valid method for determining student ideas which are likely to be common in a similar sample. It does not, however, seem like a valid or useful method of determining what student responses correspond to what topics, nor for determining what fraction of student responses are examples of each idea or resource. This is arguably the most time consuming part of resources research, and further exploration of LDA/NLP methods would be required to develop an effective method for assisting in this part of the work.

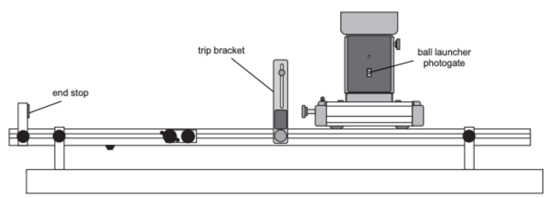
As we mentioned in our discussion of data cleaning, because of the informal nature of student responses to conceptual physics questions, there is significant variability between responses, in terms of diagrams used, spelling, and length of response. The student responses used for this analysis were originally written on paper, by students sitting in a classroom. Many students include diagrams or equations in their responses. A researcher parsing through these responses manually can interpret the meanings of these diagrams; however when the scanned documents are transcribed into CSV files, those diagrams are lost, and the equations sometimes lose their meaning. Additionally, both students or transcribers can misspell words, which makes them lose their meaning in the LDA model. Lastly, student homework responses vary from just a few words in length to several sentences forming a paragraph or two. Future work could include more data cleaning and pre-processing, including spell-checking and splitting responses up into one-sentence chunks which act as documents; this may improve model stability.

The code used for these analyses is available on GitHub as Jupyter Notebooks [24]. In its current state, we hope that this modeling process will be useful for the research done locally at Seattle Pacific University in categorizing student ideas as part of the Conceptual Resources Project.

HONORS RESEARCH SYMPOSIUM PRESENTATION
MAY 21, 2022, SEATTLE PACIFIC UNIVERSITY

To introduce my project, let me begin with an anecdote. You're nine years old, sitting in the back seat of your parents' car, driving through Salt Lake City on a road trip from California to Michigan to visit grandparents. Among the various toys and gizmos you've brought to keep yourself entertained throughout the trip, you've taken special interest in a small, polished rock you typically keep in the side pocket of your cargo shorts. As you're sitting in the back seat of the car, you might toss this rock up, and catch it as it comes back down. As you do this over and over, you start to get pretty good at catching the rock as it falls back down; you start to get a *feel* for it. You may even notice that if you've thrown the rock upwards and your parents speed the car up or slow down, you have a harder time catching it because it may seem to "move forward" or "backward" during the toss. You don't think much of it at the time.

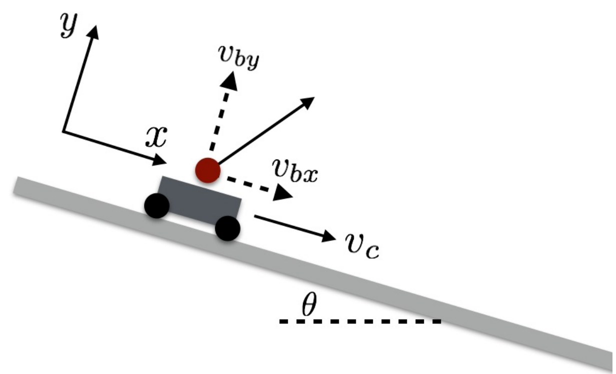
Fast-forward nine more years, and you're a first-year college student taking an introductory physics course. You're studying kinematics—or, the study of the motion of objects—and your professor poses the following question:



A ballistic cart is a smooth-rolling cart which has a vertical ball-launcher and basket attached to the top. The cart will roll along the track at a constant velocity, and when a photogate is triggered, the cart will shoot the ball up into the air. Once the ball is shot up and comes back down, will the ball land in front of, behind, or directly inside of the basket?

Your intuition tells you that the ball will land inside of the basket, but you can't quite reconcile this claim with any reasoning other than, *it just makes sense*. You recall how you used to toss rocks up in the car when you were a child, and they would fall back into your hand. If the cart acts like a car, the basket like your hand, and the ball like your small, polished rock, why wouldn't it fall back into the basket? While discussing the problem with your group mates, your professor sets up the experiment in the front of the classroom. They gather everyone back together, and everyone anxiously awaits to see what will happen to the ball. The experiment is performed, and the ball lands back in the basket. People seem generally unsurprised, and your professor heads to the whiteboard and goes through the calculations necessary to show that the ball should, in fact, fall back into the basket. But you already knew this. You may have even gone through the calculations yourself with your group, but you were more sure of it than just that; you knew *deep down* that this was the only logical outcome of the experiment, provided that the experiment was set up properly and experimental error was accounted for.

But that was the easy question. Your professor then asked, "what would happen if the track were on an incline?"



Suddenly, the answer isn't so clear. You have distinct memories of taking car trips, throwing small rocks and catching them, but you can't think of a time you tried this while going down a hill allowing for free acceleration. You discuss with your group, and they also can't intuit quite what would happen. You mess around with the calculations a little bit, and before you're able to finish and come up with a prediction, your professor sets up the experiment and everyone watches with bated breath. The cart begins to roll, the photogate triggers, the ball is launched. The ball comes back down, and lands... directly in the basket. Some people are upset because they predicted the opposite; others are happy because they happened to guess correctly with a one-third chance. Once you see this result, it seems obvious to the people around you. Some people may say, "why wouldn't this happen?" But this result deeply bothers you, and you can't quite figure out why. You're not bothered at the result of the experiment, but rather at your inability to use quick intuition in the problem-solving process. You were able to use your intuition to answer the first question with a hundred and ten percent confidence; why weren't you able to have the same confidence in answering the second question? You leave the class with your mind racing and your brow furrowed. Why did this bother you so much?

Fast-forward to my senior year of college. I've been a Physics Learning Assistant for three years now, helping students understand how to connect their intuitions with physical scenarios such as this one, all the while diving into the mathematics behind how we model these situations and how they can be useful to us. What is most fascinating to me, however, is not necessarily the physics itself, but how students perceive physics in the context of their own personal experience. Within the discipline of physics, learners gather knowledge first and foremost through our experiences with the physical world, and secondarily in the classroom where experiments can award us a more diverse scope of how the world physically operates. These chunks of knowledge gained through lived experience then translate to more concrete ideas through the process of learning physics, allowing a student to make meaning of the plethora of small experiences that have accrued throughout their life.

A. Project Description

While I could have chosen to do an entire research project on theories of learning and how they relate to the lived experiences of physics learners, I decided to take a different, perhaps more utility-focused approach. Let me frame this project with even more context which comes directly from this paper:

> A primary research focus in Physics Education is the investigation of common, topic-specific knowledge that students bring to the classroom. This kind of research has important impacts on physics instruction, particularly at the university level. It informs the development of research-based instructional materials... and it contributes instructors' knowledge of student ideas, which is an important part of the knowledge that instructors use to teach. Research identifying students' common physics ideas has investigated common, incorrect ideas, and, less extensively, students' common, potentially-fruitful ideas, known as "resources."

In essence, students have great ideas coming into the classroom, and Physics Education Researchers want to figure out what those great ideas are, and how they can create instructional material—like worksheets, labs, and homework assignments—that build upon those strengths, those seeds of knowledge. This allows students to form cohesive mental models, build stronger intuition, and construct a framework for effectively solving problems as it relates to their unique discipline. While this all sounds quite riveting, "the work of appropriately characterizing resources is very time-intensive, which can limit the extent and impact of this type of research."

To combat this time intensity, this project proposes a method by which researchers can automate part of the process of characterizing resources which students express when they answer conceptual physics questions, using Artificial Intelligence. To hearken back to the beginning, my project is titled "Developing a natural language processing approach for analyzing student ideas in calculus-based introductory physics." To clear up some terminology before diving into the specifics of my project, natural language processing (or, NLP) is a branch of Artificial Intelligence which has been briefly defined as a " range of computational techniques for analyzing and representing ... texts ... for the purpose of achieving human-like language processing for a range of applications." In particular, this project utilizes a topic modeling algorithm known as Latent Dirichlet Allocation (or, LDA for short) in order to group various clusters of words commonly used together within student responses. Let's now dive into the nitty-gritty of the process of using LDA as a tool for interpreting student responses to a conceptual question in physics.

The first step to our process was selecting a question to use. We decided to first use student responses to a question regarding DC circuits, with the following prompt:

> *Comparing the brightness of the bulbs in Circuits 1 and 2, we observe that the bulb in Circuit 1 is brighter.. Why do you think more voltage leads to more current? What mental models are you using to make sense of this?*

One thing to immediately note is that there is no incorrect answer to this question. This type of question is aimed at eliciting explanations and rationalizations from students, with the goal of understanding how they have constructed meaning from the knowledge they have gained in or outside of the classroom. We gathered 483 responses to this question in a digital format, stored in a CSV file.

Our data cleaning process then worked as follows. From the responses, we removed punctuation and stopwords, which are commonly used words such as "a," "the," and "is." We then lemmatized each of the words used in the response, which is the process of breaking down words to their roots, so words like "increased" and "increasing" become "increase." We also created bigrams, which are pairs of words commonly found together such as "potential difference" or "ohm's law." We then filtered the most and least commonly used words based on thresholds which included only words that would be useful to us in our analyses. We then created a bag-of-words out of our responses, which is a type of data structure consisting of a document-word matrix in which word order is ignored.

Once our data was cleaned and our words were properly filtered, we could begin the process of constructing the LDA model. In brief, the algorithm works by taking in a collection of documents, noticing groups of words that commonly occur together within certain documents, labeling these groups of words as "topics." These words are weighted by their prevalence in the topic, relating to how often those words co-occur with other words in that topic. One key assumption of LDA modeling is that each document (in our case, each student response) is composed of a percentage mixture of all of the topics.

The mathematics of LDA is founded upon the Dirichlet Distribution, which can be thought of as a multivariate generalization of the Beta Distribution which constructs a high-dimensional simplex, or generalized tetrahedron, to organize the topic learning. I know, sounds fascinating. For the sake of everyone's sanity, I'll refrain from talking about the mathematics any further. Let's get into results. For this circuits question, we chose to look at a five-topic model in an attempt to characterize five distinct student ideas from the set of responses we gathered. Topic 1, "water, large, flow, think, energy, pressure, like, great, push, electron." Topic 2, "electron, force, great, high, big, mean, faster, potential, push, think," et cetera. To an outsider, these topics don't appear to make any sense whatsoever. And to a physics researcher, they also don't appear to make a whole lot of sense. In their current form, these topics reveal only the key words used in responses which these topics fell under, and if our modeling worked as we hoped it would, each one of these topics output by the model should correspond to one distinct student idea.

In the secondary, qualitative analysis, we looked at responses which scored highest in each of these topics in order to see what the algorithm believes each topic should actually represent. By looking through top student responses, we extracted key phrases which include topic words. These student responses are critical to our ability to interpret what the model spits out as topics. We noticed that topic one turns from a jumbled mess of words into a few phrases which clearly highlight a distinct student idea, that "water flows through a river like current flows through circuits." As physics researchers, we synthesized these phrases together into coherent physics ideas guided by our "professional vision" as physics instructors. This final interpretive step is crucial for assigning disciplinary meaning to the model's topics.

By taking this qualitative intervention step of assigning actual meaning to the output of a model, we were able to construct distinct, useful physics ideas from this set of responses to a question on DC circuits. Topic 1 is one of the most prolific mental models about the flow of current through a wire, and is the reason why electrical current is called what it is. Another notable student idea comes from Topic 4, which relates the idea of conventional power, as in a power outage, to the electrical description of power, which is current times voltage. Adding a battery to a circuit would add power, which makes the bulb brighter.

Those are the specifics of the process by which we modeled topics as student ideas to semi-automate some of the research process for identifying student ideas and conceptual resources in physics. One great thing about writing a fully-reproducible framework for an algorithm is that a researcher should be able to input a new data set, tune some parameters, choose a number of topics to characterize, apply physics instructor knowledge to representative responses, and get similar results with distinct student ideas. And that's exactly what we did. We were able to apply this algorithm to data from two questions related to heat and temperature, and one question related to the propagation of waves. We noticed that due to the smaller size of those data sets, the topics weren't quite as distinct as those for the circuits question, but nonetheless the model provided interesting information about the responses to those three additional questions.

B. Knowledge Construction

I want to begin concluding this talk with a quote commonly attributed to George Box, that "essentially, all models are wrong, but some are useful." This is a common quip in statistics, physics, and other disciplines which rely heavily on the construction of models to interpret phenomena. What we have done through this project is construct a model by which we can tell a computer a specific set of instructions, give it a set of documents, and it spits back out a jumbled mess of words. A jumbled mess of words, fundamentally, is not a distinct student idea or mental model. Constructing meaning from knowledge in the sphere of Artificial Intelligence requires the intervention of those who build and utilize the algorithms, and without such interpretation, the models are useless. Researcher interpretation of the jumbled messes of words produced by the algorithm creates some sort of utility, some bit of information which the researchers may not have picked up on before. This is not to say that the algorithm has constructed knowledge, and it is certainly not to say that the algorithm has constructed meaning. The algorithm is a human creation, and I choose to believe that true knowledge construction and meaning making are fundamentally human activities.

Thus the disciplinary knowledge construction related to this project is twofold: that of students making sense of their intuitions, and that of researchers interpreting machine output for the purpose of streamlining a research process. Nine-year-old me had no idea that by tossing stones up and catching them on a road trip to Michigan, I was building a strong intuitive foundation by which I could later understand kinematics at a deep, personal level. Likewise, when these students were writing their responses to a circuits question, they were using a myriad of ideas they had gained in and out of the classroom. While an individual experience may be shared by many, the collection of an individual's experiences are unique. By engaging with the world around us in our own distinct ways, we continue to form our identities as unique knowledge constructors, and by reflecting upon this engagement with the world, we make meaning of the knowledge embedded within us to form an essential component of humanity.

As I continue to walk through life, I aim to consider every moment a moment of learning, of gaining knowledge in some capacity, because I recognize that humans are uniquely blessed with that ability. And as I reflect on my journey through life gaining a physical intuition for the world around me works, I continue to be open to things I may have never noticed before. And I encourage everyone to do the same.

[1] I. Portugal, P. Alencar, and D. Cowan, Expert Systems with Applications **97**, 205 (2018).
[2] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali, and W. Cao, Applied Sciences **10**, 2749 (2020).
[3] E. D. Liddy, Encyclopedia of Library and Information Science, 2nd Ed. (2001).
[4] I. Vayansky and S. A. Kumar, Information Systems **94**, 101582 (2020).
[5] J. Munsell, N. S. Rebello, and C. Rebello, in *Physics Education Research Conference 2021*, PER Conference (Virtual Conference, 2021) pp. 295–300.
[6] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, Physical Review Physics Education Research **15**, 020120 (2019).
[7] T. O. B. Odden, A. Marin, and M. D. Caballero, Phys. Rev. Phys. Educ. Res. **16**, 010142 (2020).
[8] T. O. B. Odden, A. Marin, and J. L. Rudolph, Science Education **105**, 653 (2021).
[9] E. Linstead, C. Lopes, and P. Baldi, in *2008 seventh international conference on machine learning and applications* (IEEE, 2008) pp. 813–818.
[10] S. Moro, P. Cortez, and P. Rita, Expert Systems with Applications **42**, 1314 (2015).
[11] L. C. McDermott, P. S. Shaffer, *et al.*, *Tutorials in introductory physics* (Prentice Hall Upper Saddle River, NJ, 2002).
[12] R. M. Goertzen, E. Brewe, L. H. Kramer, L. Wells, and D. Jones, Physical Review Special Topics-Physics Education Research **7**, 020105 (2011).
[13] L. S. Shulman, Educational researcher **15**, 4 (1986).
[14] J. P. Smith III, A. A. DiSessa, and J. Roschelle, The journal of the learning sciences **3**, 115 (1994).
[15] P. Heron, in *Getting Started in PER*, Vol. 2 (2018) 4th ed.
[16] A. A. diSessa, Cognition and Instruction **10**, 105 (1993).
[17] D. Hammer, American Journal of Physics **68**, S52 (2000).
[18] D. Hammer, A. Elby, R. E. Scherr, and E. F. Redish, Transfer of learning from a modern multidisciplinary perspective **89** (2005).
[19] L. M. Goodhew, A. D. Robertson, P. R. L. Heron, and R. E. Scherr, Phys. Rev. Phys. Educ. Res. **15**, 020127 (2019).
[20] Y. Zhang, R. Jin, and Z.-H. Zhou, International Journal of Machine Learning and Cybernetics **1**, 43 (2010).
[21] R. Řehůřek, P. Sojka, *et al.*, Retrieved from genism. org (2011).
[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, Journal of machine Learning research **3**, 993 (2003).
[23] C. P. George, H. Doss, *et al.*, J. Mach. Learn. Res. **18**, 5937 (2017).
[24] https://github.com/jonmgeiger/honors-project.