Geometric models reveal the hidden structure of conceptual knowledge

Paxton C. Fitzpatrick and Jeremy R. Manning*
Dartmouth College
*Corresponding author: jeremy.r.manning@dartmouth.edu

Abstract

We develop a mathematical framework, based on natural language processing models, for tracking and characterizing the acquisition of conceptual knowledge. Our approach embeds each concept in a high dimensional representation space, where nearby coordinates reflect similar or related concepts. We tested our approach using behavioral data collected from a group of college students. In the experiment, we asked the participants to answer sets of quiz questions interleaved between watching two course videos from the Khan Academy platform. We applied our framework to the videos' transcripts, and to text of the quiz questions, to quantify the content of each moment of video and each quiz question. We used these embeddings, along with participants' quiz responses, to track how the learners' knowledge changed after watching each video. Our findings show how a limited set of quiz questions may be used to construct rich and meaningful representations of what each learner knows, and how their knowledge changes over time as they learn.

Keywords: education, learning, knowledge, concepts, natural language processing

Introduction

Suppose that a teacher had access to a complete "map" of everything a student knew. Defining what such a map might even look like, let alone how it might be constructed or filled in, is itself a non-trivial problem. But if a teacher *were* to gain access to such a map, how might that change their ability to teach the student? Perhaps they might start by checking how well the student knew the to-be-learned information already, or how much they knew about related concepts. For some students, they could potentially optimize their teaching efforts to maximize efficiency by focusing primarily on not-yet-known content. For other students (or other content areas), it might be more effective to optimize for direct connections between already-known content and any new material. Observing how the student's knowledge was changing over time, in response to their training, could also help to guide the teacher.

Designing and building procedures and tools for mapping out knowledge touches on deep questions about what it means to learn. For example, how do we acquire conceptual knowledge? Memorizing course lectures or textbook chapters by rote can lead to the superficial *appearance* of understanding the underlying content, but achieving true conceptual understanding seems to require something deeper and richer. Does conceptual understanding entail connecting newly acquired information to the scaffolding of one's existing knowledge or experience [1, 5, 7, 8, 22]? Or weaving a lecture's atomic elements (e.g., its component words) into a structured network that describes how those individual elements are related? Conceptual understanding could also involve building a mental model that transcends the meanings of those individual atomic elements by reflecting the deeper meaning underlying the gestalt whole [14, 16, 21].

The difference between "understanding" and "memorizing," as framed by the researchers in education, cognitive psychology, and cognitive neuroscience [9, 10, 13, 16, 21] has profound analogs in the fields of natural language processing and natural language understanding. For example, considering the raw contents of a document (e.g., its constituent symbols, letters, and words) might provide some information about what the document is about, just as memorizing a passage might be used to answer simple questions about the passage [e.g., whether it might contain words related to furniture versus physics; 2, 3, 15]. However, modern natural language processing models [e.g., 4, 6, 20] also attempt to capture the deeper meaning *underlying* those atomic elements. These models consider not only the co-occurrences of those elements within and across documents, but also patterns in how those elements appear across different scales (e.g., sentences, paragraphs, chapters, etc.), the temporal and grammatical properties of the elements, and other high-level characteristics of how they are used [17, 18]. According to these models, the deep conceptual meaning of a document may be captured by a feature vector in a high-dimensional representation space, where nearby vectors reflect conceptually related documents. A model that succeeds at capturing an analog of "understanding" is able to assign nearby feature vectors to two conceptually related documents, *even when the words contained in those documents have very little overlap*.

Given these insights, what form might the representation of the sum total of a person's knowledge take? First, we might require a means of systematically describing or representing the nearly infinite set of possible things a person could know. Second, we might want to account for potential associations between different concepts. For example, the concepts of "fish" and "water" might be associated in the sense that fish live in water. Third, knowledge may have a critical dependency structure, such that knowing about a particular concept might require first knowing about a set of other concepts. For example, understanding the concept of a fish swimming in water first requires understanding what fish and water *are*. Fourth, as we learn, our "current state of knowledge" should change accordingly. Learning new concepts should both update our characterizations of "what is known" and should also unlock any now-satisfied dependencies of that newly learned concept so that they are "tagged" as available for future learning.

Here we develop a framework for modelling how knowledge is acquired during learning. The central idea is to use text embedding models to define the coordinate systems of two maps: (a) a *knowledge map* that describes the extent to which each concept is currently known and (b) a *learning map* that describes the extent to which each concept could be learned. Each location on these maps represents a single concept, and the geometries are defined such that related concepts are located nearby in space. We use this framework to analyze and interpret behavioral data collected from an experiment that has participants watch and answer conceptual questions about a series of recorded course lectures.

Our primary research goal is to advance our understanding of what it means to acquire deep real-world conceptual knowledge. Traditional laboratory approaches to studying learning and memory (e.g., list learning studies) often draw little distinction between memorization and understanding. Instead, these studies typically focus on whether information is effectively encoded or retrieved, rather than whether the information is *understood*. Approaches to studying conceptual learning, such as category learning experiments, can start to investigate the distinction between memorization and understanding, often by training participants to distinguish arbitrary or random features in otherwise meaningless categorized stimuli. However the objective of real-world training, or learning from life experiences more generally, is often to develop new knowledge that may be applied in *useful* ways in the future. In this sense, the gap between modern learning theories and modern pedagogical approaches and classroom learning strategies is enormous: most of our theories about *how* people learn are inspired by experimental paradigms and models that have only peripheral relevance to the kinds of learning that students and teachers actually seek. To help bridge this gap, our study uses course materials from real online courses to inform, fit, and test models of real-world conceptual learning. We also provide a "proof of concept" demonstration of how our models might be used to construct "maps" of what students know, and how their knowledge changes with training. In addition to helping to visualize knowledge (and changes in knowledge), we hope that such maps might lead to real-world tools for improving how we educate.

Results

At its core, our main modeling approach is based around a simple assumption that we sought to test empirically: all else being equal, knowledge about a given concept is predictive of knowledge about similar or related concepts. From a geometric perspective, this assumption implies that knowledge is fundamentally "smooth." In other words, as one moves through a space representing someone's knowledge (where similar concepts occupy nearby coordinates), their "level of knowledge" should change relatively gradually throughout that space. To begin to test this smoothness assumption, we sought to track our participants' knowledge and how it changed over time in response to training.

We asked our participants to answer questions from several multiple choice quizzes and watch two lecture videos from the *Khan Academy* platform (Fig. 1). One lecture video, entitled *Four Fundamental Forces*, was about the four fundamental forces in physics: gravity, strong and weak interactions, and electromagnetism. The second lecture video, entitled *Birth of Stars*, provides an overview of our current understanding of how stars form. We selected both lessons to be (a) accessible to a broad audience, e.g., by minimizing prerequisite knowledge, (b) largely independent of each other, e.g., so that the two videos focused on different material and did not depend on each other, and (c) related to each other, e.g., so that both videos contained at least *some* similar or overlapping content. The two videos we selected are introductory, about different primary concepts, but also touch on "physics" and "astronomy" themes. We also wrote a set of multiple choice quiz questions that would enable us to test participants' knoweldge about each individual video and about related content not specifically presented in either video (Tab. S1). Participants answered questions randomly drawn from each content area (lecture 1, lecture 2, and general physics knowledge) across each of three quizzes. Quiz 1 was intended to assessed participants' knowledge before training; quiz 2 assessed knowledge after watching the Four Fundamental Forces video (i.e., lecture 1); and quiz 3 assessed knowledge after watching the Birth of Stars video (i.e., lecture 2).

We trained a text embedding model using sliding windows of text from the two videos' transcripts (see *Constructing text embeddings of multiple videos and questions*). We also used the same model (i.e., trained on the videos' transcripts) to embed the text of each question in our pool. This yielded, for each second of each video, and for each question, a single topic vector– i.e., a coordinate in a text embedding space (Fig. 6). Intuitively, each dimension of the embedding space corresponds to a "theme" or "topic" reflected in some part(s) of the videos (Tab. S2), and the coordinates in embedding space denote the blend of themes reflected by a particular excerpt of text (e.g., from

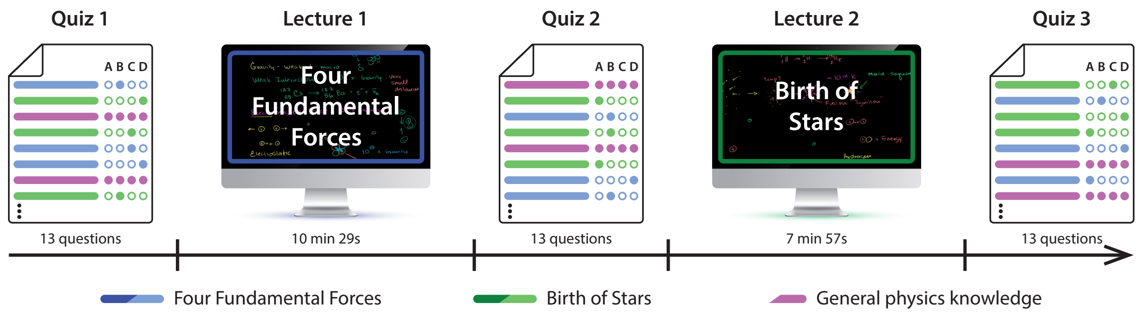

Figure 1: Experimental paradigm. Participants alternate between answering 13-question multiple choice quizzes and watching two Khan academy videos. Each quiz contains a mix of 5 questions about lecture 1, 5 questions about lecture 2, and 3 general physics knowledge questions. The specific questions reflected on each quiz, and the orders of each quiz's questions, were randomized across participants.

part of a video's transcript, from a question, etc.).

Although a single lecture may be organized a single broad theme at a coarse scale, at a finer scale each moment of a lecture typically covers a narrower range of content. We wondered whether a text embedding model trained on the lectures' transcripts might capture some of this finer scale content. For example, if a particular question asks about the content from one small part of a lecture, we wondered whether our text embedding model could be used to automatically identify the "matching" moment(s) in the lecture. When we correlated each question's topic vector with the topic vectors for each second of the lectures, we found some evidence that each question is temporally specific (Fig. 2). In particular, most questions' topic vectors were maximally correlated with a well-defined range of timepoints from their corresponding lectures, and the correlations fell off sharply outside of that range. We also examined the best-matching intervals for each question qualitatively by comparing the text of the question to the text of the most-correlated parts of the lectures. Despite that the questions were excluded from the text embedding model's training set, in general we found a close correspondence between the conceptual content that each question covered and the content covered by the best-matching moments of the lectures. Two representative examples are shown at the bottom of Fig. 2.

The ability to quantify how much each question is "asking about" each moment of the lectures could enable high-resolution insights into participants' knowledge. Traditional approaches to

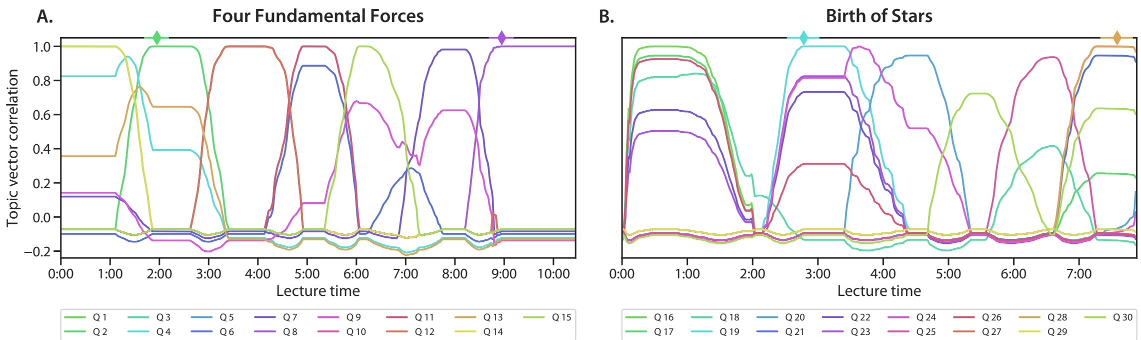

Figure 2: Which parts of each lecture are captured by each question? Each panel displays timeseries plots showing how each question's topic vector correlates with each video timepoint's topic vector. The left panel displays these correlations for the *Four Fundamental Forces* lecture and associated questions, and the right panels displays these correlations for the *Birth of Stars* lecture and associated questions. The colors denote question identities. The diamonds in each panel denote the moment of peak correlation between the indicated questions, in the indicated lectures. The associated questions' text, and snippets of the lectures' transcripts in the best-matching sliding windows, are displayed at the bottom of the figure.

estimating how much a student "knows" about the content of a given lecture entail computing the proportion of correctly answered questions. But if two students receive identical scores on an exam, might our modeling framework help us to gain more naunced insights into the *specific* content that each student has mastered (or failed to master)? For example, a student who misses three questions that were all about the same concept (e.g., concept *A*) will have gotten the same *proportion* of questions correct as another student who missed three questions about three *different* concepts (e.g., *A*, *B*, and *C*). But if we wanted to fill in the "gaps" in the two students' understandings, we might do well to focus on concept *A* for the first student, but to also add in materials pertaining to concepts *B* and *C* for the second student.

We developed a simple formula (Eqn. 1) for using a participant's responses to a small set of multiple choice questions to estimate how much the participant "knows" about the concept reflected by any arbitrary coordinate, x, in text embedding space (e.g., the content reflected by any moment in a lecture they had watched; see *Estimating dynamic knowledge traces*). Essentially, the estimated knowledge at the coordinate is given by the weighted average proportion of quiz questions the participant answered correctly, where the weights reflect how much each question is "about" the content at x. When we apply this approach to estimate the participant's knowledge

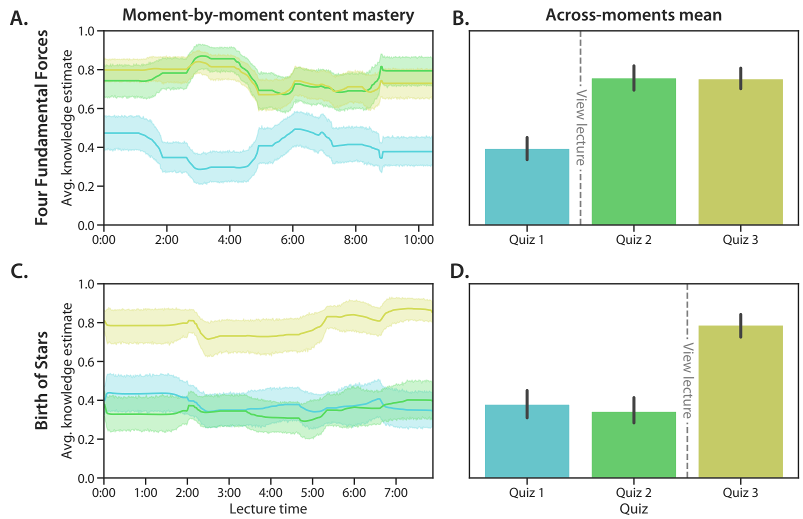

Figure 3: Estimating moment-by-moment knowledge acquisition. A. Moment-by-moment knowledge about the *Four Fundamental Forces*. Each trace displays the weighted proportion of correctly answered questions about the content reflected in each moment of the lecture (see *Estimating dynamic knowledge traces*), using responses from one quiz (color). The traces are averaged across participants. **B. Average estimated knowledge about the Four Fundamental Forces.** Each bar displays the across-timepoint average knowledge, estimated using the responses to one quiz's questions. **C. Moment-by-moment knowledge about the *Birth of Stars*.** The panel is in the same format as Panel A, but here the knowledge estimates are for the moment-by-moment content of the *Birth of Stars* lecture. **D. Average estimated knowledge about the *Birth of Stars*.** The panel is in the same format as Panel B, but here the knowledge estimates are for the content of the *Birth of Stars* lecture. All panels: error ribbons and error bars denote 95% confidence intervals, estimated across participants.

about the content presented in each moment of each lecture, we can obtain a detailed timecourse describing how much "knowledge" the participant has about any part of the lecture. As shown in Figures 3A and C), we can also apply this approach separately for the questions from each quiz the participants took throughout the experiment. From just 13 questions per quiz, we obtain a high-resolution snapshot (at the time each quiz was taken) of what the participants knew about any moment's content, from either of the two lectures they watched (comprising a total of 1106 samples across the two lectures).

Of course, even though the timecourses in Figure 3A and C provide detailed *estimates* about participants' knowlege, those estimates are only *useful* to the extent that they accurately reflect what participants actually know. As one sanity check, we anticipated that the knowledge estimates should show a content-specific "boost" in participants' knowledge after watching each lecture.

In other words, if participants learn about each lecture's content when they watch each lecture, the knowledge estimates should reflect that. After watching the *Four Fundamental Forces* lecture, participants should show more knowledge for the content of that lecture than they had before, and that knowledge should persist for the remainder of the experiment. Specifically, knowledge about that lecture's content should be relatively low when estimated using Quiz 1 responses, but should increase when estimated using Quiz 2 or 3 responses (Fig. 3B). Indeed, we found that participants' estimated knowledge about the content of the *Four Fundamental Forces* was substantially higher on Quiz 2 versus Quiz 1 ($t(49) = 8.764, p < 0.001$) and on Quiz 3 versus Quiz 1 ($t(49) = 10.519, p < 0.001$). We found no reliable differences in estimated knowledge about that lecture's content on Quiz 2 versus 3 ($t(49) = 0.160, p = 0.874$). Similarly, we hypothesized (and subsequently confirmed) that participants should show more estimated knowledge about the content of the *Birth of Stars* lecture after (versus before) watching it (Fig. 3D). Specifically, since participants watched that lecture after taking Quiz 2 (but before Quiz 3), we hypothesized that their knowledge estimates should be relatively low on Quizzes 1 and 2, but should show a "boost" on Quiz 3. Consistent with this prediction, we found no reliable differences in estimated knowledge about the *Birth of Stars* lecture content on Quizzes 1 versus 2 ($t(49) = 1.013, p = 0.316$), but the estimated knowledge was substantially higher on Quiz 3 versus 2 ($t(49) = 10.561, p < 0.001$) and Quiz 3 versus 1 ($t(49) = 8.969, p < 0.001$).

If we are able to accurately estimate a participant's knowledge about the content tested by a given question, the estimated knowledge should have some predictive information about whether the participant is likely ot answer the question correctly or incorrectly. For each question in turn, for each participant, we used Equation 1 to estimate (using all *other* questions from the same quiz, from the same participant) the participant's knowledge at the held-out question's embedding coordinate. For each quiz, we aggregated these estimates into two distributions: one for the estimated knowledge at the coordinates of each *correctly* answered question, and another for the estimated knowledge at the coordinates of each *incorrectly* answered question (Fig. 4). We then used independent samples *t*-tests to compare the means of these distributions of estimated knowledge.

For the initial quizzes participants took (prior to watching either lecture), participants' estimated

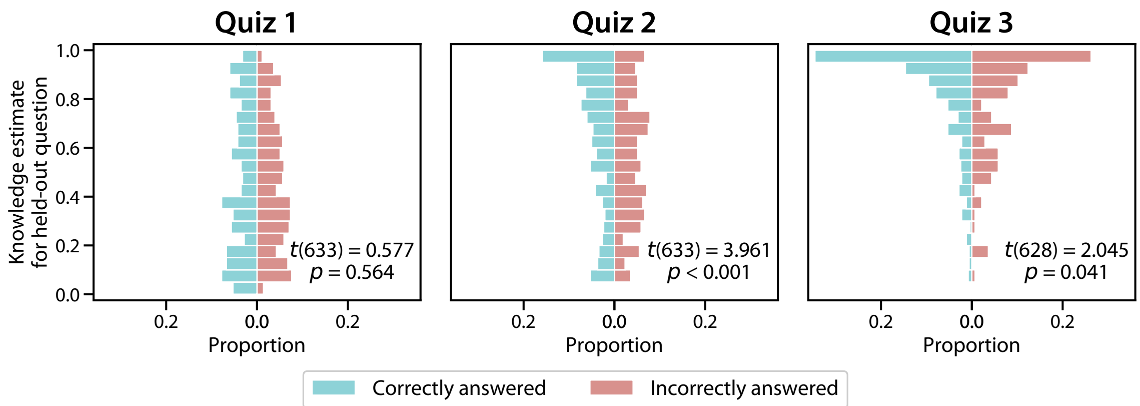

Figure 4: Estimating knowledge at the embedding coordinates of held-out questions. Separately for each quiz (panel), we plot the distributions of predicted knowledge at the embedding coordinates of each held out correctly (blue) or incorrectly (red) answered question. The t-tests reported in each panel are between the distributions of estimated knowledge at the coordinates of correctly versus incorrectly answered held-out questions.

knowledge tended to be low overall, and relatively unstructured (Fig. 4, left panel). When we held out individual questions and estimated their knowledge at the held-out questions' embedding coordinates, we found no reliable differences in the estimates when the held-out question had been correctly versus incorrectly answered ($t(633) = 0.577, p = 0.564$). After watching the first video, estimated knowledge for held-out correctly answered questions (from the second quiz; Fig. 4, middle panel) exhibited a positive shift relative to held-out incorrectly answered questions ($t(633) = 3.961, p < 0.001$). After watching the second video, estimated knowledge (from the third quiz; Fig. 4, right panel) for *all* questions exhibited a positive shift. However, the increase in estimated knowledge for held-out correctly answered questions was larger than for held-out incorrectly answered questions ($t(628) = 2.045, p = 0.041$).

Knowledge estimates need not be limited to the content of the lectures. As illustrated in Figure 5, our general approach to estimating knowledge from a small number of quiz questions may be applied to *any* content, given its text embedding coordinate. To visualize how knowledge "spreads" through text embedding space to content beyond the lectures participants watched, we first fit a new topic model to the lectures' sliding windows with $k = 100$ topics. We hoped that increasing the number of topics from 15 to 100 might help us to generalize the knowledge predictions. (Aside from increasing the number of topics from 15 to 100, all other procedures and model parameters were carried over from the preceding analyses.) We resampled each lecture's topic trajectory to 1 Hz and also projected each question into the same 100-dimensional topic space.

We projected the resulting 100-dimensional topic vectors (for each second of video and for each question) into a shared 2-dimensional space (see *Creating knowledge and learning map visualizations*). Next, we sampled points evenly from a 100×100 grid of coordinates that evenly tiled a rectangle enclosing the 2D projections of the video trajectories and questions. We used Equation 4 to estimate participants' knowledge at each of these 10K sampled locations, and we averaged these estimates across participants to obtain an estimated average *knowledge map* (Fig. 5). Intuitively, the knowledge map constructed from a given quiz's responses provides a visualization of how "much" participants know about any content expressible by the fitted model.

Several features of the resulting knowledge maps are worth noting. The average knowledge map estimated from Quiz 1 responses (Fig. 5, leftmost map) shows that participants tended to have relatively little knowledge about any parts of the text embedding space (i.e., the shading is relatively dark everywhere). The knowledge map estimated from Quiz 2 responses shows a marked increase in knowledge on the left side of the map (around roughly the same range of coordinates covered by the *Four Fundamental Forces* lecture, indicated by the dotted blue line). In other words, participants' estimated increase in knowledge is localized to conceptual content that is nearby (i.e., related to) the content from the lecture they watched prior to taking Quiz 2. This localization is non-trivial: the knowledge estimates are informed only by the embedded coordinates of the *quiz questions*, not by the embeddings of either lecture. Finally, the knowledge map estimated from Quiz 3 responses shows a second increase in knowledge, localized to the region surrounding the embedding of the *Birth of Stars* lecture participants watched prior to taken Quiz 3.

Another way of visualizing these content-specific increases in knowledge (apparently driven by watching each lecture) is displayed in Figure 5B. Taking the point-by-point difference between the knowledge maps estimated from responses to a successive pair of quizzes yields a *learning map* that describes the *change* in knowledge estimates from one quiz to the next. These learning maps highlight that the estimated knowledge increases we observed across maps were specific to the

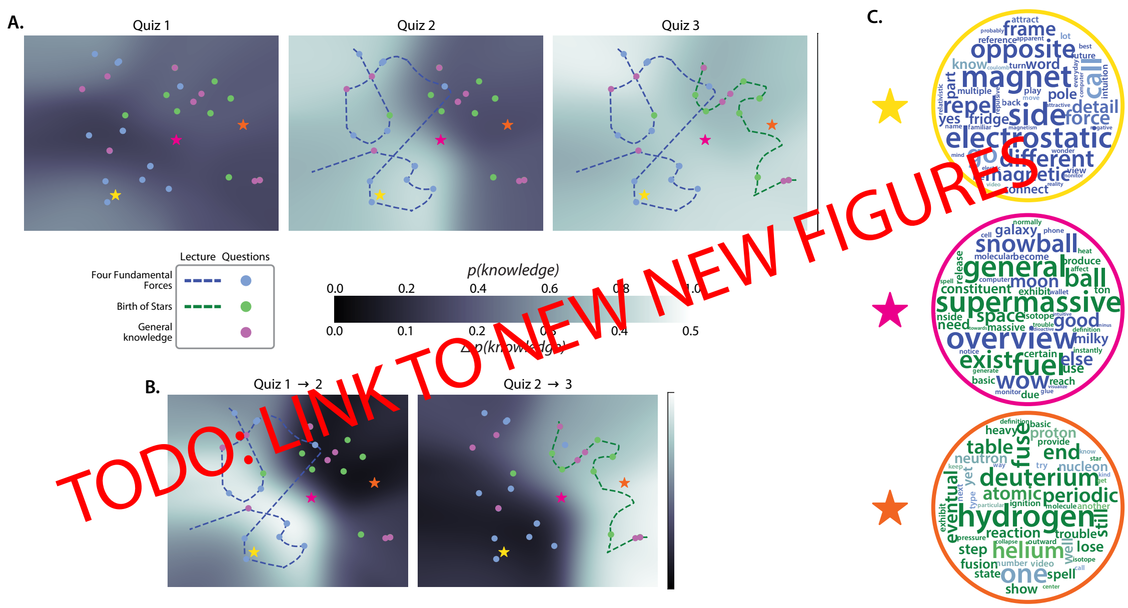

Figure 5: Mapping out the geometry of knowledge and learning. A. Average "knowledge maps" estimated using each quiz. Each map displays a 2D projection of the estimated knowledge about the content reflected by *all* regions of topic space (see *Creating knowledge and learning map visualizations*). The topic trajectories of each lecture and the coordinates of each question are indicated by dotted lines and dots. Each map reflects an average across all participants. For individual participants' maps, see Figures S1, S2, and S3. **B. Average "learning maps" estiamted between each successive pair of quizzes.** The learning maps are in the same general format as the knowledge maps in Panel A, but each coordinate in the learning maps indicates the *difference* between the corresponding coordinates in the indicated *pair* of knowledge maps– i.e., how much the estimated knowledge "changed" across the two quizzes. Each map reflects an average across all participants. For individual participants' maps, see Figures S4 and S5. **C. Word clouds for sampled points in topic space.** Each word cloud displays the relative weights of each word reflected by the blend of topics represented at the locations of the stars in the maps. The words' colors indicate how much each word is weighted on average across all timepoints' topic vectors in the *Four Fundamental Forces* (blue) and *Birth of Stars* (green) videos, respectively.

regions around the embeddings of each lecture in turn.

Because the 2D projection we used to construct the knowledge and learning maps is (partially) invertable, we may gain additional insights into the estimates by reconstructing the original high-dimensional topic vectors for any point(s) in the maps we are interested in. For example, this could serve as a useful tool for an instructor looking to better understand which content areas a student (or a group of students) knows well (or poorly). As a demonstration, we show the top-weighted words from the blends of topics reconstructed from three locations on the maps (Fig. 5C): one point near the *Four Fundamental Forces* embedding (yellow); a second point near the *Birth of Stars* embedding (orange), and a third point somewhere in between the two lectures' embeddings (pink). As shown in the word clouds in the Panel, the top-weighted words at the coordinate near the *Four Fundamental Forces* embedding also tended to be weighted heavily by the topics expressed in that lecture. Similarly, the top-weighted words at the coordinate near the *Birth of Stars* embedding tended to be weighted most heavily by the topics expressed in *that* lecture. And the top-weighted words at the coordinate between the two lectures' embeddings show a roughly even mix of words most strongly associated with each lecture.

Discussion

Materials and methods

Participants

We enrolled a total of 50 Dartmouth undergraduate students in our study. Participants received course credit for enrolling. We asked each participant to fill out a demographic survey that included questions about their age, gender, native spoken language, ethnicity, race, hearing, color vision, sleep, coffee consumption, level of alertness, and several aspects of their educational background and prior coursework.

Participants' ages ranged from 18 to 22 years (mean: 19.52 years; standard deviation: 1.09

years). A total of 15 participants reported their gender as male and 35 participants reported their gender as female. A total of 49 participants reported their native language as "English" and 1 reported having another native language. A total of 47 participants reported their ethnicity as "Not Hispanic or Latino" and three reported their ethnicity as "Hispanic or Latino." Participants reported their races as White (32 participants), Asian (14 participants), Black or African American (5 participants), American Indian or Alaska Native (1 participant), and Native Hawaiian or Other Pacific Islander (1 participant). (Note that some participants selected multiple recial categories.)

A total of 49 participants reporting having normal hearing and 1 participant reported having some hearing impairment. A total of 49 participants reported having normal color vision and 1 participant reported being color blind. Participants reported having had, on the night prior to testing, 2 – 4 hours of sleep (1 participant), 4 – 6 hours of sleep (9 participants), 6 – 8 hours of sleep (35 participants), or 8+ hours of sleep (5 participants). They reported having consumed, on the same day and leading up to their testing session, 0 cups of coffee (38 participants), 1 cup of coffee (10 participants), 3 cups of coffee (1 participant), or 4+ cups of coffee (1 participant).

No participants reported that their focus was currently impaired (e.g., by drugs or alcohol). Participants reported their current level of alertness, and we converted their responses to numerical scores as follows: "very sluggish" (-2), "a little sluggish" (-1), "neutral" (0), "fairly alert" (1), and "very alert" (2). Across all participants, a range of alertness levels were reported (range: -2 – 1; mean: -0.10; standard deviation: 0.84).

Particpants reported their undergraduate major(s) as Social Sciences (28 participants), Natural sciences (16), Professional (e.g., pre-med or pre-law; 8 participants), Mathematics and engineering (7 participants), Humanities (4 participants), or Undecided (3 participants). Note that some participants selected multiple categories for their undergraduate major. We also asked participants about the courses they had taken. In total, 46 participants reported having taken at least one Khan academy course in the past or being familiar with the Khan academy, and 4 reported not having taken any Khan academy courses. Of the participants who reported having watched at least one Khan academy course, 1 participant declined to report the number of courses they had watched; 7 participants reported having watched 1–2 courses; 11 reported having watched 3–5 courses; 8

reported having watched 5–10 courses; and 19 reported having watched 10 or more courses. We also asked participants about the specific courses they had watched, categorized under different subject areas. In the "Mathematics" area participants reported having watched videos on AP Calculus AB (21 participants), Precalculus (17 participants), Algebra 2 (14 participants), AP Calculus BC (12 participants), Trigonometry (11 participants), Algebra 1 (10 participants), Geometry (8 participants), Pre-algebra (7 participants), Multivariable Calculus (5 participants), Differential Equations (5 participants, Statistics and Probability (4 participants), AP Statistics (2 participants), Linear Algebra (2 participants), Early Math (1 participant), Arithmetic (1 participant), and other videos not listed in our survey (6 participants). In the "Science and engineering" area participants reported having watched videos on Chemistry, AP Chemistry, or Organic Chemstry (21 participants); Physics, AP Physics I, or AP Physics II (15 participants); Biology, AP Biology; or High school Biology (15 participants); Health and Medicine (1 participant); or other videos not listed in our survey (20 participants). We also asked participants if they had specifically seen the videos used in our experiment. When we asked about the *Four Fundamental Forces* video, 45 participants reported not having watched it before, 1 participant reported that they were not sure if they had watched it before, and 4 participants declined to respond. When we asked about the *Birth of Stars* video, 46 participants reported not having watched it before and 4 participants declined to respond. When we asked participants about non-Khan academy online courses, they reported having watched or taken courses on Mathematics (15 participants), Science and engineering (11 participants), Test preparation (9 participants), Economics and finance (3 participants), Arts and humanities (2 participants), Computing (2 participants), and other categories not listed in our survey (18 participants). Finally, we asked participants about in-person courses they had taken in different subject areas. They reported taking courses in Mathematics (39 participants), Science and engineering (38 participants), Arts and humanities (35 participants), Test preparation (27 participants), Economics and finance (26 participants), Computing (15 participants), College and careers (7 participants), or other courses not listed in our survey (6 participants).

Experiment

We hand-selected two roughly 10-minute course videos from the Khan Academy platform: *The Four Fundamental Forces* (an introduction to gravity, electromagnetism, the weak nuclear force, and the strong nuclear force; duration: 10 minutes and 29 seconds) and *Birth of Stars* (an introduction to how stars are formed; duration: 7 minutes and 57 seconds). We hand-wrote 39 multiple choice questions: 15 about the conceptual content of *The Four Fundamental Forces*, another 15 about the conceptual content of *Birth of Stars*, and 9 other questions that tested for general conceptual knowledge about basic physics (covering material that was not presented in either video). The full set of questions may be found in Table S1.

Participants began the main experiment by answering a battery of 13 randomly selected questions (chosen from the full set of 39). Then they watched the *The Four Fundamental Forces* video. Next, they answered a second set of 13 questions (chosen at random from the remaining 26 questions). Fourth, participants watch the *Birth of Stars* video, and finally they answered the remaining 13 questions. Our experimental procedure is diagramed in Figure 1. We used the experiment to develop and test our computational framework for estimating knowledge and learning maps.

Analysis

Constructing text embeddings of multiple videos and questions

We extended an approach developed by [12] to construct text embeddings for each moment of each lecture, and of each question in our pool. Briefly, our approach uses a topic model [3], trained on a set of documents, to discover a set of k "topics" or "themes." Formally, each topic is defined as a set of weights over each word in the model's vocabulary (i.e., the union of all unique words, across all documents, excluding "stop words."). Conceptually, each topic is intended to give larger weights to set of words that appear conceptually related or that tend to co-occur in the same documents. After fitting a topic model, each document in the training set, or any *new* document that contains at least some of the words in the model's vocabulary, may be represented as a k-dimensional vector describing how much the document (most probably) reflects each topic. (Unless, otherwise noted,

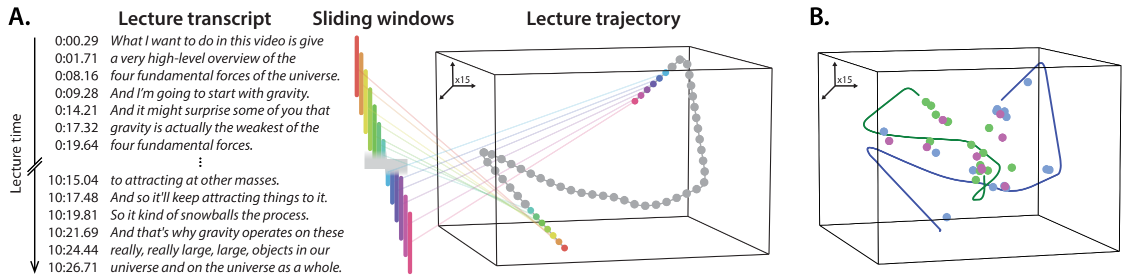

Figure 6: Constructing video content *trajectories*. A. Building a document pool from sliding windows of text. We decompose each video's transcript into a series of overlapping sliding windows. The set of transcript snippets (across all windows) may be treated as a set of "documents" for training a text embedding model. After training a text embedding model using the two videos' sliding windows, along with the text from each question in our pool (Tab. S1), we construct "trajectories" through text embedding space by joining the embedding coordinates of successive sliding windows from each video. **B. Embedding multiple videos and questions.** Applying the same text embedding approach to each video, along with the text of each question, results in one trajectory per video and one embedding coordinate (dot) per question (blue: Four Fundamental Forces; green: Birth of Stars; pink: general physics knowledge). Here we have projected the 15-dimensional embeddings into a 3D space using Uniform Manifold Approximation and Projection [UMAP; 19].

we used $k = 15$ topics.)

As illustrated in Figure 6A, we start by building up a corpus of documents using overlapping sliding windows that span each video's transcript. Khan Academy videos are hosted on the YouTube platform, and all YouTube videos are run through Google's speech-to-text API [11] to derive a timestamped transcript of any detected speech in the video. The resulting transcripts contain one timestamped row per line, and each line generally corresponds to a few seconds of spoken content from the video. We defined a sliding window length of (up to) $w = 30$ transcript lines, and we assigned each window a timestamp according to the midpoint between its first and last lines' timestamps. These sliding windows ramped up and down in length at the very beginning and end of the transcript, respectively. In other words, the first sliding window covered only the first line from the transcript; the second sliding window covered the first two lines; and so on. This insured that each line of the transcript appeared in the same number (w) of sliding windows. We treated the text from each sliding window as a single "document," and we combined these documents across the two videos' windows to create a single training corpus for the topic model. The top words from each of the 15 discovered topics may be found in Table S2.

After fitting a topic model to each videos' transcripts, we could use the trained model to transform arbitrary (potentially new) documents into k-dimensional topic vectors. A convenient property of these topic vectors is that documents that reflect similar blends of topics (i.e., documents that reflect similar themes, according to the model) will yield similar (in terms of Euclidean distance, correlation, etc.) topic vectors. In general, the similarity between different documents' topic vectors may be used to characterize the similarity in content between the documents.

We transformed each sliding window's text into a topic vector, and then used linear interpolation (independently for each topic dimension) to resample the resulting timeseries to once per second. This yielded a single topic vector for each second of each video. We also used the fitted model to obtain topic vectors for each question in our pool (Tab. S1). Taken together, we obtained a *trajectory* for each video, describing its path through topic space, and a single coordinate for each question (Fig. 6B). Embedding both videos and all of the questions using a common model enables us to compare the content from different moments of videos, compare the content across videos, and estimate potential associations between specific questions and specific moments of video.

Estimating dynamic knowledge traces

We used the following equation to estimate each participant's knowledge about timepoint t of a given lecture, $\hat{k}(t)$:

$$\hat{k}(t) = \frac{\sum_{i \in \text{correct}} \text{ncorr}(t, i)}{\sum_{j=1}^{N} \text{ncorr}(t, j)}, \tag{1}$$

where

$$\text{ncorr}(x, y) = \frac{\text{corr}(x, y) - \text{mincorr}}{\text{maxcorr} - \text{mincorr}}, \tag{2}$$

and where mincorr and maxcorr are the minimum and maximum correlations between any lecture timepoint and question, taken over all timepoints and questions across both lectures and all three question sets.

Intuitively, $\text{ncorr}(x, y)$ is the correlation between two topic vectors (e.g., the topic vector from one timepoint in a lecture, x, and the topic vector for one question, y), normalized by the minimum and maximum correlations (across all timepoints and questions) to range between 0 and 1, inclusive. Equation 1 then computes the weighted average proportion of correctly answered questions about the content presented at timepoint t, where the weights are given by the normalized correlations between timepoint t's topic vector and the topic vectors for each question. The normalization step (i.e., using ncorr instead of the raw correlations) insures that every question (except the least-relevant question) contributes some non-zero amount to the knowledge estimate.

Creating knowledge and learning map visualizations

An important feature of our approach is that, given a trained text embedding model and participants' quiz performance on each question, we can estimate their knowledge about *any* content expressable by the embedding model– not solely the content explicitly probed by the quiz questions. To visualize these estimates (Figs. 5, S1, S2, S3, S4, and S5), we used UMAP [19] to define a 2D projection of the text embedding space. Sampling the original 100-dimensional space at high resolution to obtain an adequate set of topic vectors spanning the embedding space would be computationally intractable. However, sampling a 2D grid is much more feasible. We defined a rectangle enclosing the 2D projections of the lectures' and quizzes' embeddings, and we sampled points from a regular 100×100 grid of coordinates that evenly tiled the enclosing rectangle. We sought to estimate participants' knoweldge (and learning– i.e., changes in knowledge) at each of the resulting 10000 coordinates.

To generate our estimates, we placed a set of 39 radial basis functions (RBFs) throughout the embedding space, centered on the 2D projections for each question (i.e., we included one RBF for each question). At coordinate x, the value of an RBF centered on a question's coordinate μ, is given by:

$$\mathrm{RBF}(x, \mu, \lambda) = \exp\left\{-\frac{\|x-\mu\|^2}{\lambda}\right\}. \tag{3}$$

The λ term in the RBF equation controls the "smoothness" of the function, where larger values of λ result in smoother maps. In our implementation we used $\lambda = 50$. Next, we estimated the “knowledge” at each coordinate, x, using:

$$\hat{k}(x) = \frac{\sum_{i \in \text{correct}} \text{RBF}(x, q_i, \lambda)}{\sum_{j=1}^{N} \text{RBF}(x, q_j, \lambda)}. \tag{4}$$

Intuitively, Equation 4 computes the weighted proportion of correctly answered questions, where the weights are given by how nearby (in the 2D space) each question is to the x. We also defined *learning maps* as the coordinate-by-coordinate differences between any pair of knowledge maps. Intuitively, learning maps reflect the *change* in knowledge across two maps.

References

[1] Blaye, A., Bernard-Peyron, V., Paour, J.-L., and Bonthoux, F. (2006). Category flexibility in children: distinguishing response flexibility from conceptual flexibility; the protracted development of taxonomic representations. *European Journal of Developmental Psychology*, 3(2):163–188.

[2] Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the International Conference on Machine Learning*, pages 113–120, New York, NY. Association for Computing Machinery.

[3] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

[4] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *arXiv*, 2005.14165.

[5] Caramazza, A. and Mahon, B. Z. (2003). The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, 7(8):354–361.

[6] Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, 1803.11175.

[7] Constantinescu, A. O., O'Reilly, J. X., and Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468.

[8] Deacon, D., Grose-Fifer, J., Yang, C. M., Stanick, V., Hewitt, S., and Dynowska, A. (2004). Evidence for a new conceptualization of semantic representation in the left and right cerebral hemispheres. *Cortex*, 40(3):467–478.

[9] Gallagher, J. J. (2000). Teaching for understanding and application of science knowledge. *School Science and Mathematics*, 100(6):310–318.

[10] Hall, R. and Greeno, J. (2008). *21st century education: A reference handbook*, chapter Conceptual learning, pages 212–221. Sage Publications.

[11] Halpern, Y., Hall, K. B., Schogol, V., Riley, M., Roark, B., Skobeltsyn, G., and Bäuml, M. (2016). Contextual prediction models for speech recognition. In *Interspeech*, pages 2338–2342.

[12] Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral and neural signatures of transforming naturalistic experiences into episodic memories. *Nature Human Behavior*, 5:905–919.

[13] Katona, G. (1940). *Organizing and memorizing: studies in the psychology of learning and teaching*. Columbia University Press.

[14] Kintsch (1970). *Learning, memory, and conceptual processes*. Wiley.

[15] Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

[16] Maclellan, E. (2005). Conceptual learning: the priority for higher education. *British Journal of Educational Studies*, 53(2):129–147.

[17] Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook of Human Memory*. Oxford University Press.

[18] Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave" function? *Psychological Review*, 128(4):711–725.

[19] McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 1802(03426).

[20] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, 1301.3781.

[21] Scott, P., Asoko, H., and Leach, J. (2007). *Handbook of research on science education*, chapter Student conceptions and conceptual learning in science. Routledge.

[22] Simon, M. A., Tzur, R., Heinz, K., and Kinzel, M. (2004). Explicating a mechanism for conceptual learning: elaborating the construct of reflective abstraction. *Journal for Research in Mathematics Education*, 35(5):305–329.