AERA Open
January-December 2022, Vol. 8, No. 1, pp. 1–18
DOI: 10.1177/23328584221126516
Article reuse guidelines: sagepub.com/journals-permissions

Forecasting Undergraduate Majors: A Natural Language Approach

David Lang
Alex Wang
Nathan Dalal
Andreas Paepcke
Mitchell L. Stevens
Stanford University

Committing to a major is a fateful step in an undergraduate education, yet the relationship between courses taken early in an academic career and ultimate major issuance remains little studied at scale. Using transcript data capturing the academic careers of 26,892 undergraduates enrolled at a private university between 2000 and 2020, we describe enrollment histories by using natural-language methods and vector embeddings to forecast terminal major on the basis of course sequences beginning at college entry. We find that (a) a student's very first enrolled course predicts their major 30 times better than random guessing and more than one-third better than majority-class voting, (b) modeling strategies substantially influence forecasting metrics, and (c) course portfolios vary substantially within majors, such that students with the same major exhibit relatively modest overlap.

Keywords: *colleges, decision-making, degree planning, descriptive analysis, higher education, information retrieval, institutional research, Jaccardian similarity, LASSO, network analysis, NLP, observational research, postsecondary education, predictive analytics, regression analyses, textual analysis, word embedding*

Introduction

Committing to a major is a fateful step in an undergraduate education. Major selection influences academic persistence (Leppel, 2001), postcollegiate income (Bleemer & Mehta, 2020), and job satisfaction (Wolniak & Pascarella, 2005). Majors also are outcomes of chains of prior decisions. Under elective curriculums common in the United States, students often declare majors only after a significant portion of coursework has been completed. Yet the iterative academic commitments that sum to majors have only rarely been observed systematically (Baker, 2018). Understanding how and when course selections signal major commitment has practical implications for students and schools. Once equipped with such knowledge, students might better recognize how early course selections are related to downstream commitments, and school leaders might design course sequences and advising supports that make incremental academic decisions more informed and efficient.

Serial research has established that undergraduates often find the course-selection process cognitively daunting (Bailey et al., 2015; Baker, 2018; Chambliss & Takacs, 2014; Chaturapruek et al., 2019; Page & Scott-Clayton, 2016). Time-to-degree delays and major-switching are common phenomena that can be costly for students and schools (Fink et al., 2018; Liu et al., 2021; Witteveen & Attewell, 2021). Providing students with a general sense of the relationship between initial course selections and eventual majors might lighten the cognitive task for students as well as enable more informed consideration of curricular options and their downstream consequences early in college. Academic advisors might use the same information to assist students in appreciating the cumulative implications of individual course choices, while administrators might use it to predict future enrollments and guide their allocation of scarce instructional resources and student-support services.

In theory, academic transcripts detailing the course sequences of former students can be the source of insight to address these problems. Yet until recently, researchers lacked the computational capacity to leverage transcripts for such insight at scale. The work presented here combines newly commonplace computational tools with academic transcript data to forecast undergraduate majors on the basis of early college enrollments among 26,892 undergraduates who moved through a private university between 2000 and 2020. We offer this analysis as an existence proof and illustrative case study in how major forecasting might be carried out at virtually any school that maintains transcripts with course-level academic selections.

Our work makes three contributions. First, borrowing an analytic strategy developed by Pardos and Nam (2020), we use natural language-processing (NLP) techniques to forecast the majors based on the sequenced course selections students make from the very beginning of their academic careers. Second, we illustrate how ablative analyses can be deployed to further specify which, and how many, courses are necessary to forecast a student's terminal major with a specified degree of accuracy. Third, we illustrate how these analytic techniques reveal degrees of variance in academic pathways toward particular majors.

The work presented here uses course-enrollment data exclusively. We focus on course enrollments, exclusive of student grades and demographic characteristics, because enrollment information is an essential component of institutional records and is likely to be readily available to administrators and researchers elsewhere. Because virtually all schools retain and archive official enrollment records, the techniques we develop here are highly portable. Additionally, we offer the current work as an empirical and technical baseline for further study of major selection. Subsequent inquiries incorporating grades and demographic information will surface substantive questions about the sociology of academic pathways and important issues of algorithmic fairness and bias that are necessarily beyond the scope of the present study.

Background and Prior Work

An often-expressed ideal of a liberal-arts education is intellectual exploration and serendipitous discovery (Delbanco, 2014). Important expressions of this ideal are *elective curricula* in which students are allowed and even encouraged to sample a range of coursework before committing to a major. At least in principle, elective curricula make it possible for students to select from a wide variety of academic paths, especially at large research universities, where the range of subjects and potential course sequences are much more extensive than the typically smaller and more regimented academic programs of high schools (McFarland, 2006).

Prior research offers conflicting cues about the potential relationship between early course taking and college major. On the one hand, signals of major choice surface long before students arrive at college. Using data from an online-learning platform for middle-school students, researchers found that clickstream behavior data predicted whether students would major in a science, technology, engineering, and mathematics (STEM) field with 66% accuracy (San Pedro et al., 2014). Other work has shown that professed occupational plans in high school are predictive of a STEM major (Weeden et al., 2020), as are high school courses: For example, enrollment in advanced placement calculus nearly doubles the odds of a student committing to a STEM major in college (Gottfried & Bozick, 2016). Such findings suggest at least some continuity in how commitment to fields of study unfolds throughout young adulthood.

On the other hand, evidence suggests that exposure to, and experiences in, academic fields early in college can substantially influence undergraduate course trajectories. Research leveraging quasi-experimental assignment to courses has found that exposure to courses in disciplines outside students' preferred majors increases the likelihood of majoring in those subjects (Fricke et al., 2018). Other work highlights that attribution bias and availability heuristics also inform major choice. For example, students assigned to early-morning courses are less likely to major in those courses' subject domains (Haggag et al., 2021). Similarly, students who are randomly assigned to a course during the period in which they must declare a major are more likely to major in that course's subject domain (Patterson et al., 2021). A carefully designed qualitative study of undergraduate pathways at a selective college has found that students would eliminate further investigation of entire fields of study on the basis of negative experiences in a single introductory course (Chambliss & Takacs, 2014).

Prior predictive-analytic work has focused on attrition from STEM fields. Chen and colleagues (2018) find that college admission scores, term-level grade point averages (GPAs), and student demographic information can be leveraged to forecast whether and when students will leave an already declared major. Other research indicates that performance in gatekeeping courses, such as calculus, is strongly predictive of attrition from STEM majors (Aulck et al., 2017).

Although empirically generative and conceptually promising, much of the prior predictive work focuses specifically on STEM pathways, to the exclusion of other academic fields. Yet navigating course choices and committing to majors are universal problems in U.S. undergraduate education. For example, work by Aulck and West (2017) suggests that as many as 19% of all college students change majors at least once. Scientific progress on major prediction might yield insight into academic pathways across the entire ecology of academic knowledge.

Our work builds upon that of Beaulac and Rosenthal (2019), who find that random-forest models featurized with subject-level course completions and subject-specific GPA can predict major with 47% accuracy on the basis of courses taken in the first year. Although compelling, their effort is constrained by limitations of data and analytic strategy. First, the authors are obliged to infer major based on the plurality of courses taken rather than the ground truth of degree issuance. Second, the authors model all courses within a subject as equivalent, when, in fact, particular courses may be especially powerful or weak predictors of subsequent course-taking. Third, the random-forest approach makes it difficult for analysts to interpret why a

model predicts an outcome of a particular student—a limitation on transparency that may be particularly problematic in academic settings. NLP techniques mitigate these concerns, yielding better performance and more transparent interpretability.

Pardos and Nam (2020) have adapted techniques from NLP to describe and intervene in how students navigate college coursework. Their key contribution is the powerful idea that a sequence of courses in an academic career is analogous to a sequence of words in a sentence. To the extent that this analogy holds, NLP methods can be used to reveal important structural features of a curriculum, such as prerequisite relations and likely sequences of course-enrollment decisions (Jiang & Pardos, 2020). These researchers and others have used the same approach to generate course equivalencies that can inform articulation agreements between institutions (Pardos et al., 2019a). We extend that work here by applying NLP tools to forecast undergraduate major on the basis of course enrollments early in the academic career.

Conceptual Framework

We begin from the premise that elective curricula provide students with broad yet bounded contexts for pursuing their academic careers. Most schools require students to complete some number of core courses or distribution requirements and a major comprising specific courses or topics. Limits of course supply, time, attention, and tuition place additional constraints on what and how many selections students can make. These constraints facilitate conditions to forecast majors on the basis of early coursework: A large, but ultimately limited, number of possible academic pathways is available to students within any given institution.

These conditions also mean that individual course selections are fateful and become more so as completed courses accumulate. Each course represents an investment that cannot be made in another option. To the extent that the resources students and their families might allocate to degree completion are limited, each selection entails a trade-off (Chaturapruek et al., 2021). And because time, attention, and tuition invested in prior courses cannot be invested again, each passing academic term reduces the remaining resources available to consider alternate academic paths. This situation is why students, parents, administrators, and higher-education researchers are wise to worry about the structure of undergraduate course offerings (Baker, 2018; Rosenbaum et al., 2007) and about how students consider and select courses (Chambliss & Takacs, 2014).

We conceptualize individual courses as the building blocks of academic careers for three reasons: (a) courses represent chunks of academic credit that accumulate to fulfill degree requirements, (b) undergraduate majors are defined as portfolios of courses or course categories, and (c) courses are convenient units of academic increment, leaving clear traces on transcripts. Under this broad conceptual umbrella, our empirical inquiry is focused on three ideas: major forecasting, data ablation, and the comparative composition of majors. We address each of these in turn below.

Major forecasting is the term we use to describe the use of artificial intelligence–based techniques to predict the likelihood of a major on the basis of some number and sequence of courses taken early in the undergraduate career. The conceptual foundation of major forecasting is that the course-taking patterns of prior students provide insight into how current students are likely to navigate their academic careers. Although we acknowledge that curricula grow and evolve over time, it is also likely that early course selections relate to subsequent major declaration to some empirically observable extent. Thus, the mechanical requirements of majors and the course sequences of prior students should yield at least some valuable empirical insight about how current students will navigate course selection.

Data ablation is a technical term describing the process by which analysts systematically vary the amount and kind of information they incorporate into a predictive model. Data ablation enables analysts to discern the relative predictive power of different data representations. In our work, we systematically vary the number of completed courses used to forecast student major. Doing so enables us to better articulate the trade-offs of predictive accuracy versus timely information. Moreover, by explicitly excluding or including certain pieces of information, such as course type (prerequisite status, for example), we can make credible claims about what kinds of data more and less optimally serve the task of major forecasting.

The comparative composition of majors is another key factor. We define an *undergraduate major* as the set of courses necessary and sufficient to receive a degree in a specified field (e.g., art history, biology, or chemistry). Within this broad definition, institutional documents specifying major requirements make clear that not all majors are defined with the same level of specificity and stringency. Jurisdiction over major requirements typically rests at the department or program level (Abbott, 2010), and practitioners of different fields structure majors in different ways. Such variation has consequences for major forecasting. For example, it is a different task to forecast a major comprising 10 courses within a given subject domain, of which five are specified requirements, than to forecast a major comprising 10 courses within a subject domain but with no specified requirements. The *comparative composition of majors* refers to the task of identifying what kinds of course sequences and course obligations a major entails.

With these goals established, we specify the following research questions.

Research Questions

Each course selection represents a unit of investment of time, tuition, and opportunity costs. In light of this specification, we posit that each course selection contains a bit of information about any given student's academic predilections and preferences. Information accumulates by some observable factor in tandem with the accumulation of course selections. Our four research questions are designed to surface this factor.

Prior research gives us reason to suspect that early course selections carry information about future majors. This may be the case because coursework in high school produces academic orientations that shape initial college preferences (Gottfried & Bozick, 2016; Weeden et al., 2020), because early course-taking in college creates path dependencies that influence subsequent selections (Chambliss & Takacs, 2014), or because of some combination of these two processes. A fortuitous feature of our data is that each individual course selection is time-stamped at the moment of registration. This feature enables us to order course selections not only by academic term but also within each term chronologically. We thus specify our first research question:

RQ 1: Is there a relationship between the first selected college course and terminal major?

Although answering this question with the observational data available from academic transcripts will not by itself enable us to specify causal chains, it can indicate whether early college coursework contains signals about a subsequent major. Particular courses have varying relationships to majors. Specifically, some courses serve as prerequisites for other courses in an academic domain. Prerequisite courses may be formally or functionally necessary to complete a major, or some combination of the two. By *formally necessary*, we mean that academic regulations require majors to complete a course; by *functionally necessary*, we mean that subsequent coursework requires students to deploy knowledge or skills taught in prerequisite courses. Either way, we suspect that prerequisite courses play a distinctive role in the unfolding of course sequences. They may ensure that students are academically prepared to benefit from subsequent course-taking, they may serve to exclude certain students from pursuing or even considering particular majors, or they may do both, as is commonly observed in STEM fields (Slim et al., 2014).

By contrast, courses that are not prerequisites—we refer to these as *non-prerequisites*—may tell us more or less than prerequisites about subsequent course selections. On the one hand, because they require no prior coursework, these courses may carry more information about student preferences than prerequisites and thus contain more signals. On the other hand, because non-prerequisites are less tightly coupled with other courses in sequences leading to majors, they may contain relatively fewer signals. Thus, our second research question:

RQ 2: Are prerequisite courses a better forecasting resource than non-prequisite courses?

Having detailed administrative data is necessary but not sufficient to build useful predictive models. Transcript information typically contains textual and numerical components (e.g., MATH 31, COMPUTER SCIENCE 101), but such data features are highly peculiar to particular schools and may not be applied consistently across departments and programs even within a given school. For these reasons, we disregard course numbers and names in our models and instead focus on four distinct and potentially universal representations of courses:

- One-Hot Encodings: Each course is represented as a binary/dummy variable.
- Subject-Level Encodings: Each course is represented as a running count of how many times a student completed a course in a particular subject listing in the course catalog (e.g., Math: 3, Economics: 1, History: 0).
- Academic-Term Embeddings: Each course is represented by its co-occurrence with other courses in an academic term.
- Academic-Career Embeddings: Each course is represented by its co-occurrence within a student's undergraduate academic career.

We suspect that embedded representations will outperform one-hot encodings. Embedded representations may capture similarities between related courses and might also generate abstract representations of course histories that are more compatible with machine-learning architectures (Erhan et al., 2010; Pardos et al., 2019b). We also suspect that embedded representations at different time windows (e.g., a single academic term vs. academic career) may provide varied utilities for distinct tasks (Caselles-Dupré et al., 2018). Academic-term embeddings may be more beneficial for tasks in which a course history is to be used for term-specific decisions—for example, to suggest a course that complements others being taken (Pardos & Jiang, 2019) or to estimate a student's academic workload during a specific term (Chockkalingam et al., 2021). By contrast, career embeddings may be more useful for forecasting majors because major completion requires the accumulation of many specific courses over academic time. Data-ablation techniques enable us to train our models with different kinds of courses to gain empirical insight on these matters. Thus, our third research question:

RQ 3: Do different representations of courses affect the predictive validity of major forecasts?

We suspect that the relationship between course selections and terminal major will change as selections accumulate. Accumulating courses represent a journey in degree progress and can indicate the likelihood of changing or completing paths at different points in academic careers. Researchers, academic advisors, and administrators would benefit from knowing how the forecasting signals carried by course selections change over academic time. Thus, our fourth research question:

RQ 4: How do accumulating course selections affect major forecasts at critical points in the academic career?

To answer this question, we again use data-ablation techniques to vary the number of courses in a sequence that our models can perceive. We choose the number of courses that roughly correspond to the first academic term ($n = 5$) and the first and second years of academic instruction ($n = 15$ and $n = 30$, respectively).

Setting and Data

This is a case study of undergraduate students moving through a private research university in the United States. We refer to the school by the pseudonym Western University (or "Western" for short). Western has selective admissions and an elective curriculum. It traditionally has strong programs in the applied sciences and attracts many students interested in these fields. Nevertheless, students are admitted to the university as a whole—not to particular programs or schools—and are explicitly encouraged to sample a range of academic coursework before committing to a major near the end of their second year. Western's academic calendar is organized on a quarter system, and undergraduates are expected to be enrolled full-time during each academic year's fall, winter, and spring terms. Courses are offered by administrative units that are either disciplinary departments (e.g., sociology) or interdisciplinary programs (e.g., urban studies). For this analysis, these administrative units are interchangeable, and thus we use the term *department/program* to refer to the offering unit.

Majors at Western vary in scope and intensity, ranging from some that consist of fewer than 60 credit units to others that have complex prerequisite structures and comprise well over 100 credit units. The scope and content of major requirements are set at the department/program level, with only minimal regulation by the higher administrative units. During their first 2 years, students navigate the elective curriculum with the help of several online course-exploration tools as well as human advisors. Once students officially declare a major, human advising shifts from a general set of advisors to staff dedicated to specific programs of study. Especially in programs with very large enrollments (e.g., computer science), students with declared majors may be given preferential access to particular courses.

In contrast with universities that use GPA and other requirements to limit access to particular academic programs (Bleemer & Mehta, 2021), Western places very few conditions or strictures on undergraduate major declarations. Advising websites and other supporting documents consistently encourage students to engage in wide academic exploration across the curriculum.

Major declaration is a strong but imperfect signal of major issuance at the case school. Completing required courses is a necessary but not sufficient criterion to receive a degree in a particular major: Students must also select and declare this major. Our data indicate that for approximately every five majors declared, only four degrees listing that major are issued—indicating the proportion of majors that are dropped or changed subsequent to declaration. The median student declares their major exactly 2 years after initial enrollment, at the start of the third year. Approximately 25% of students do not declare a major until the end of the third year.

Our data set contains the enrollment histories of undergraduates at Western from 2000 to 2020. In the following subsections, we specify additional data features as well as how we configured the overall corpus to represent courses, enrollments, and majors. It includes all the courses completed by each student, specified by term of enrollment, and eventual issued major. Students are identified in the data set only by an anonymous hash code. We limit our analysis to all undergraduate students who enrolled and graduated with an undergraduate degree during the 2-decade observation window, yielding an analytic sample of 26,892 students.[1]

Courses

Students at Western have an "open-enrollment" period in which they can enroll in any class. During this period, at the beginning of each term, students can enroll in the overwhelming majority of available courses with minimal restrictions. The only courses that typically have binding enrollment caps are in a first-year writing program, athletics and wellness courses, studio-art courses, and foreign-language courses.

Our data include information on the relational structure among all selected courses. Leveraging information from Western's registrar and its official course catalog, we identify how courses relate to one another via prerequisites and course requirements. For example, if Mathematics 101 is a prerequisite—a specified requirement for students seeking to enroll in Mathematics 202 or Economics 101—then the dependent courses may be considered *postrequisites*. These

TABLE 1

Course enrollment statistics at Western University, by prerequisite status (2000–2020)

Course type	Courses	Enrollments	Share
Prerequisites	2,271	522,327	37.6%
Non-prerequisites	19,423	37.6%	62.4%
All	21,694	1,390,971	100%

Note. Course corresponds to the number of unique course offerings during the period. *Enrollment* corresponds to the number of student enrollments during the period. *Share* is a student enrollment weighted share between the two course types.

courses share ties of dependency that may influence efforts to forecast majors on the basis of early enrollments. Of the more than 21,694 distinct courses in which students in our analytic sample enrolled, 2,271 have explicitly specified prerequisite relations. Courses that are specified prerequisites represent a disproportionate share of overall enrollments (see Table 1). Although approximately 10% of courses serve as prerequisites, they compose nearly 37% of enrollments.

We identified 2,879 prerequisite relations among the 2,271 prerequisite courses. This means that specific courses may require a student to take more than one prerequisite. For instance, machine-learning courses may require students to take an introductory programming class as well as linear algebra. Additionally, certain courses may have more than one *postrequisite*. Introductory computer science and introductory statistics have more than 100 unique course postrequisites, indicating that particular courses may have disproportionate importance on curricular pathways. Many of these relations cross department lines, with a course offered in one department (e.g., economics) serving as a prerequisite for a course offered by a different department or program (e.g., public policy).

We aggregated all courses serving as prerequisites and collapsed each department/program domain into a supernode (see Figure 1). Each label corresponds to a department/program offering courses. The number in parentheses indicates how many external departments/programs require students to take courses in that department. Perhaps reflecting its legacy strengths in applied-science fields, computer science (CS) and mathematics (math) are at the center of Western's prerequisite network: More than 20 departments require students to take at least one CS or math course.

Course enrollments tend to follow a fat-tailed distribution: Many courses have sparse enrollments. In Figure 2, we plot the proportion of students enrolled in each course by rank. Courses in linear algebra, introductory computer programming, and introductory statistics had the highest enrollments at Western during the study period. The most enrolled

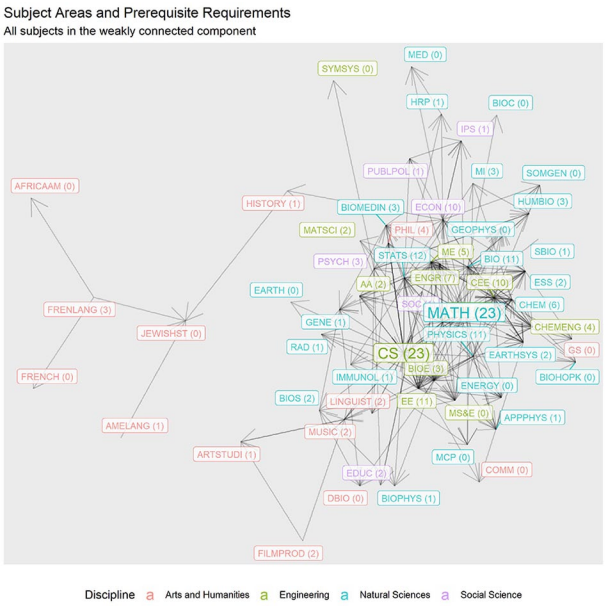

FIGURE 1. *Course requirements, by department.*

Note. Each subject area is colored by discipline. The number associated with each subject area corresponds to the out degree of that subject, meaning how many departments outside the subject area have required courses in that subject area. Arrows indicate the direction of the prerequisite relation. See the online article for the color version of this figure.

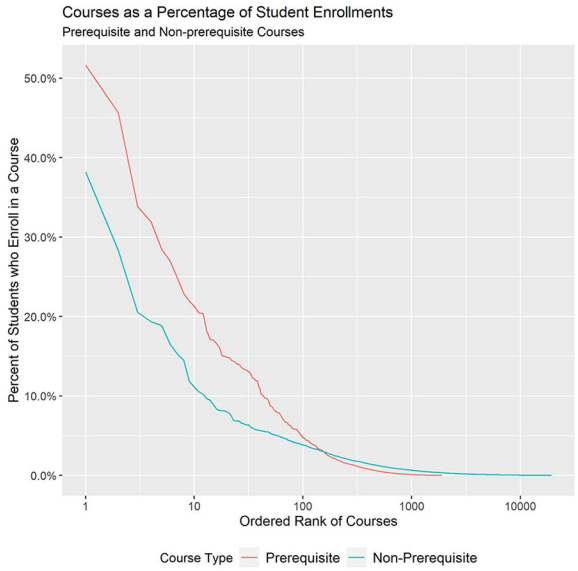

FIGURE 2. *Enrollment share, by course type (2000–2020).*

Note. Courses correspond to the rank-ordered enrollment frequency by each course type. See the online article for the color version of this figure.

course that did not have a prerequisite relation is a writing seminar required of all undergraduates as part of first-year general-education requirements. Generally, prerequisite courses receive larger enrollment shares than nonrequisite courses.

TABLE 2
Enrollment statistics (2000–2020)

Measure	Courses	Courses per quarter	Quarters
P25	47.00	3.83	12.00
mean	51.72	4.23	12.28
median	51.00	4.17	12.00
P75	56.00	4.57	13.00
sd	8.35	0.58	1.51
n	26,892	26,892	26,892

Note. Measures correspond to the first quartile, mean, median, third quartile, standard deviation (*sd*), and observation count (*n*). *Course* corresponds to the number of courses a student takes in their entire academic career. *Courses per quarter* correspond to the number of courses a student takes per academic term. *Quarter* corresponds to the number of academic terms that a student remains enrolled during their time at Western University.

Enrollments

We present summary enrollment data in Table 2. Students typically complete between 47 and 56 courses during their academic careers at Western, enrolling in three to five courses per academic term and remaining matriculated for 12 terms over 4 academic years.

Majors

The 26,892 students in our data set exhibited 81 unique majors. The most frequent major is a program in human biology, with 2,758 students; the least frequent is an independently designed English major, with one student. To simplify analysis and to minimize computational problems from including many very small cells, we condensed these 81 majors into 26 categories, based on their frequency (see Table 3). The first 25 categories are the most highly enrolled majors during the study period, together representing 86% of all students. The 26th category comprises the remaining 56 majors, representing 14% of all degrees awarded during the study period.

We group the 25 most highly enrolled majors into four academic domains:

- Engineering: chemical engineering (CHEME), computer science (CS), electrical engineering (EE), engineering (ENGR), mathematical and computational sciences (MATSC), mechanical engineering (ME), management science and engineering (MGTSC), and symbolic systems and information sciences (SYMBO)
- Natural sciences: biology (BIO), biological sciences (BIOL), chemistry (CHEM), earth systems (EASYS), human biology (HUMBI), mathematics (MATH), and physics (PHYS)
- Social Sciences: economics (ECON), international relations (INTLR), political science (POLSC), psychology (PSYCH), and public policy (PUBPO)
- Arts and Humanities: communications (COMMU), English (ENGL), history (HIST), science technology and society (STS), and American studies (AMSTU)

We designate the academic domains of all other declared majors as "Other."

TABLE 3
Distribution of majors

Degree	Discipline	Share (%)	Students
CHEME	Engineering	1.4	373
CS	Engineering	9.4	2,518
EE	Engineering	2.8	740
ENGR	Engineering	5.2	1,400
MATCS	Engineering	1.5	402
ME	Engineering	3.9	1,038
MGTSC	Engineering	3.7	1,000
SYMBO	Engineering	2.4	640
Subtotal	Engineering	30.2	8,111
BIO	Natural Science	3.5	948
BIOL	Natural Science	2.9	772
CHEM	Natural Science	1.0	278
EASYS	Natural Science	1.8	496
HUMBI	Natural Science	10.3	2,758
MATH	Natural Science	1.9	510
PHYS	Natural Science	1.3	350
Subtotal	Natural Science	22.7	6,112
ECON	Social Science	6.6	1,765
INTLR	Social Science	4.6	1,241
POLSC	Social Science	3.9	1,059
PSYCH	Social Science	4.2	1,121
PUBPO	Social Science	1.4	367
Subtotal	Social Science	20.6	5,553
AMSTU	Arts & Humanities	1.0	278
COMMU	Arts & Humanities	1.7	454
ENGL	Arts & Humanities	3.5	939
HSTRY	Arts & Humanities	3.0	811
STS	Arts & Humanities	3.2	865
Subtotal	Arts & Humanities	12.4	3,347
Other		14.0	3,769
Total		100.0	26,892

Note. Each row corresponds to the number of students graduating in a particular major during the 2000–2020 period at Western University. *Share* corresponds to the fraction of majors as a share of all students during the period.

Methods

Course Embeddings

One of the challenges of analyzing student enrollments at the course level is that many schools, including Western, have thousands or even tens of thousands of courses listed in

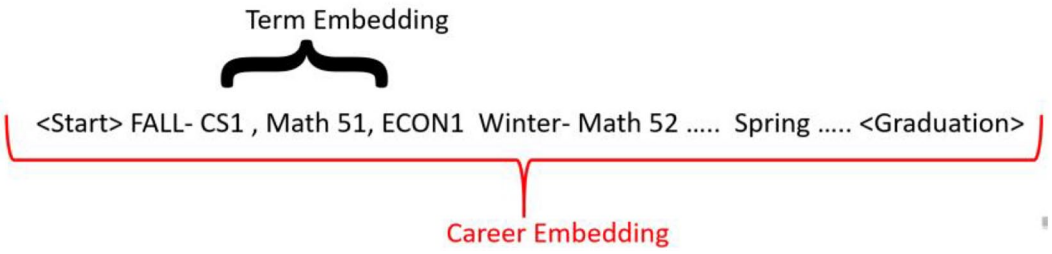

FIGURE 3. *Visualization of a term and academic-career embedding.*
Note. The black bracket corresponds to an academic-term embedding, which only incorporates data from a given academic term. The red bracket corresponds to an academic-career embedding, which is trained on a student's entire course history. See the online article for the color version of this figure.

their catalogs over a 20-year period. Many courses may enroll only a handful of students. Given this sparsity, it is difficult to understand the extent to which small-enrollment courses are associated with major selection. Concurrently, we have strong reason to believe that many courses exhibit similarities in terms of subject matter, content, and importance in regard to major completion. Alternative representations of students' transcripts that aggregate these individually sparse signals may enhance predictive validity. Course embeddings offer an approach to capturing some of this commonality while reducing the dimensionality problem.

Following Pardos and Nam (2020), we define *course embedding* as a high-dimensional vector representation of courses, analogous to word vectors in NLP. Analogizing course catalog numbers and codes as words (e.g., ECON1, ENGLISH10), we posit that the sequences of course enrollments represented in college transcripts are amenable to the same analytic strategies that enable computational "reading" of word sequences in linguistic texts. We describe the construction of these embeddings in Appendix A.

These embedded representations perform well not only on such tasks as word analogy (Mikolov et al., 2013a) but also in a wide variety of contexts, from genetics to the production of recommendations (Barkan & Koenigstein, 2016; Zou et al., 2019). These techniques have been used extensively in academic domains by Pardos and colleagues on a host of tasks ranging from course recommendation to course articulation to degree planning (Pardos & Jiang, 2019; Pardos et al., 2019a; Shao et al., 2021).

A common way to validate embeddings is by assessing their efficacy at these analogy tasks (Arthurs & Alvero, 2020). The analogies can be syntactic or semantic. Examples of syntactic analogy tasks in academic contexts include whether one course is an honors version of a similar course or whether a course is a component of a sequence. Examples of semantic analogy tasks are whether courses are combinations of domains (e.g., biology + information sciences = bioinformatics; Pardos & Nam, 2020).

We focus on a prerequisite identification task to select embeddings for our forecasting exercise. Our goal is to construct a vector space in which courses with a prerequisite relationship are close to each other. What we are building is akin to a search engine for prerequisite relations. For example, assume that single-variable calculus (MATH 19) is a prerequisite for multivariable calculus (MATH 20). Ideally, if we typed *MATH 20* into our search engine, *MATH 19* would be near the top of the list, and unrelated courses (e.g., art history) would be toward the bottom of the list.

Our procedure for constructing an embedding appropriate for prerequisite discovery is as follows: We take our true postrequisite course for each of our 2,849 prerequisite relations and generate all (2,271) possible candidate pairs of courses in the prerequisite course network. We then compute cosine similarity for each pair of courses. We sort the candidate pairs by cosine similarity and examine the corresponding rank of the true relation. For this task, we suspect that institutional knowledge of the university can inform how to train vector representations. Specifically, we suspect that students are not likely to take a prerequisite and its associated postrequisite in the same academic term. To test this intuition, we consider two distinct window types for course-enrollment sequences. The first type we refer to as *academic-term embeddings*. These are embeddings trained only on courses that co-occur in the same academic term (see Figure 3). We also consider embeddings that can contain courses from a student's initial enrollment to their final course as an undergraduate at Western. We describe these latter embeddings as *academic-career embeddings*. We have chosen this prerequisite task because we believe that it highlights the relative affordances of the two representations. We anticipate that students typically will not take a prerequisite course and its postrequisite within the same term. This anticipated result will be reflected in term embeddings exhibiting relatively poor performance.

Our focal metric is the mean reciprocal rank (MRR), a standard metric in information retrieval and search tasks. In this case, Q corresponds to the number of prerequisite relations, as specified by the equation below:

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i}. \quad (1).$$

TABLE 4
Embedding performance at prerequisite detection task

Representation	Embedding size	Window	MRR	Mean rank	Median reciprocal rank	Median rank
Academic career	256	∞	0.163	6.14	0.034	29.00
Academic term	256	∞	0.031	31.80	0.0024	423.00

Note. Each row corresponds to the best model from hyperparameter search results for each embedding type. *MRR* corresponds to the mean reciprocal rank at the prerequisite detection task.

We explored a variety of embedding dimensions $\{32, 64, 128, 256, 512\}$ and window sizes $\{3, 5, 10, 20, 40, \infty\}$ for each embedding type. We found that an embedded representation of 256 dimensions worked best for term and academic-career embeddings. In both cases, we found that an infinite window, a model that could potentially retain all of a student's prior and subsequent history, generated the best performance for our prerequisite detection task. Academic-career embeddings performed substantially better at the prerequisite detection task. The MRR of the academic-career embeddings is 0.163 (mean rank 6.14) (see Table 4), suggesting that the cosine similarity with academic-career embeddings between two courses can identify a course's true relation within the first six or seven courses.

In contrast, academic-term embedding has a reciprocal rank of 0.031 and takes 31.80 courses on average to identify the true course-prerequisite pair. To put this information in a more familiar context, if we used this approach as a search engine for course dependencies, academic-career embeddings would generally return the right answer on the first page, while academic-term embeddings would return the right answer on the fourth page. Alternative measures, such as the median reciprocal rank, present a qualitatively similar picture.

Forecasting Major

We use a shallow-learning algorithm such that incremental changes in course selections can be readily associated with changes in the probability of a particular major issuance. At the same time, we wish to minimize the complexity associated with potentially keeping track of tens of thousands of possible enrollments. We use the following technique to forecast a student's terminal major by using course-enrollment history. The specific method is a multinomial logistic Least Absolute Shrinkage and Selection Operator (LASSO) with the following loss function:

$$-\left[\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} y_{il}(\beta_{0k} + x_i^T\beta_k) - \log\left(\sum_{\ell=1}^{K} e^{\beta_{0\ell} + x_i^T\beta_\ell}\right)\right)\right] + \lambda\sum_{j=1}^{p}\|\beta_j\|. \quad (2)$$

In this equation, x_i corresponds to a vector of course-enrollment data for a given student. The β parameter corresponds to a vector of coefficients for each course representation. The term K corresponds to the number of majors. The term λ corresponds to the regularization parameter that we tune as a hyperparameter. One affordance of a LASSO model compared to other regression-based approaches is that it shrinks coefficients toward zero, leaving a relatively small number of parameters for investigation while preserving predictive validity (Tibshirani, 1996). To ensure that our reported model performance is accurate and generalizable, we split the data into three sets. The first is a *training set*, which allows researchers to freely generate and test various machine-learning models and assess their predictive power. The second is the *dev set*, which comprises data excluded from model training and is used to better understand how the model performs on sample data to inform hyperparameter selection. Last is the *testing set*, which is used to provide a measure of model performance on data that neither the machine-learning model nor the researcher has examined. We allocated our individual-level student data into an 80/10/10 train-dev-test split. All preprocessing steps, including construction of the embeddings, were completed by using only the training set. We trained models by using student course data from the training set, chose hyperparameters based on the dev set, and reported errors from the test set for all analyses. In what follows, we report only performance on the test set.

These data splits were generated by random sampling over the 20-year observation period. We provide analysis in Appendix B on the performance of this modeling approach when our train and test sets come from distinct time periods. We also report supplementary results following our shuffling of the order of course enrollments within an academic term.

We trained major forecasting models after a student had taken n courses, where $n \in \{1, 2, 3, 4, 5, 10, 15, 20, 30, 40, 50, 60\}$. For embedded representation of courses, we represented course history by computing the mean of the vector embeddings.

Results

Our findings indicate that course enrollment can forecast a student's major with reasonable accuracy early in the undergraduate career. We present the results of our analyses for each of the research questions below in turn.

TABLE 5
Major classification accuracy at the time of first course

Major	Discipline	One-hot (%)	Subject (%)	Term (%)	Academic-Term (%)	Guesser (%)	*n*
CHEME	Engineering	0.00	0.00	5.88	5.88	0.02	34
CS	Engineering	24.71	43.97	45.49	45.10	0.90	255
EE	Engineering	5.63	12.00	5.63	7.04	0.07	71
ENGR	Engineering	0.00	0.00	0.00	0.00	0.26	138
MATCS	Engineering	0.00	0.00	0.00	0.00	0.02	41
ME	Engineering	0.94	6.00	6.60	9.43	0.16	106
MGTSC	Engineering	0.00	0.00	0.00	0.00	0.13	96
SYMBO	Engineering	0.00	0.00	0.00	0.00	0.07	69
subtotal	Engineering	8.40	15.68	15.93	16.30	0.38	810
BIO	Natural science	0.00	0.00	4.12	1.03	0.13	97
BIOL	Natural science	0.00	0.00	3.49	5.81	0.10	86
CHEM	Natural science	0.00	0.00	0.00	0.00	0.01	32
EASYS	Natural science	0.00	2.13	2.63	2.63	0.02	38
HUMBI	Natural science	15.35	23.63	40.25	43.57	0.80	241
MATH	Natural science	30.77	0.00	30.77	30.77	0.04	52
PHYS	Natural science	25.00	24.32	41.67	41.67	0.02	36
subtotal	Natural science	10.65	11.43	23.37	24.57	0.38	582
ECON	Social science	4.26	6.82	9.57	7.98	0.49	188
INTLR	Social science	0.00	0.87	2.96	5.93	0.25	135
POLSC	Social science	0.00	7.58	1.11	7.78	0.11	90
PSYCH	Social science	0.00	12.07	8.41	9.35	0.16	107
PUBPO	Social science	0.00	0.00	0.00	0.00	0.02	38
subtotal	Social science	1.43	6.04	5.73	7.17	0.28	558
AMSTU	Arts and humanities	0.00	0.00	0.00	0.00	0.01	32
COMMU	Arts and humanities	0.00	0.00	0.00	0.00	0.03	43
ENGL	Arts and humanities	0.00	8.33	0.00	0.98	0.14	102
HSTRY	Arts and humanities	0.00	0.00	0.00	3.57	0.10	84
STS	Arts and humanities	5.06	0.00	3.80	7.59	0.09	79
subtotal	Arts and humanities	1.18	2.50	0.88	2.94	0.09	340
Other	Other	86.72	64.23	54.39	49.62	2.20	399
Total		18.15	18.30	19.23	19.45	0.59	2689

Note. Each row corresponds to forecasting accuracy for each major by using just the first course a student enrolled in. This table only corresponds to performance on the out-of-sample test set.

RQ1: Using First Enrolled Course to Forecast Major

We trained models with one-hot representations, subject-level representations, and embedding representations to predict major solely on the basis of the first course. We identified the first course by the transaction time in the enrollment database. Accuracy for each major's performance in the test set is displayed in Table 5. The third column corresponds to a model that used the one-hot encoded representation of courses. The fourth column corresponds to subject representations. The fifth and sixth columns correspond to model accuracy for the academic-term-embedded and academic-career-embedded representation of courses, respectively. The seventh column shows the corresponding accuracy rates of a stratified random model that guesses each major randomly according to its frequency in the test set. The eighth column corresponds to the individual counts of students by major in the test set.

Our findings indicate that the first enrolled course carries signals of a student's predilection toward a field of study. Using only that first course to forecast subsequent major increases the accuracy of our model more than 30 times over random guessing (0.59%). A one-hot encoding model identifies a major with more than 18% accuracy. Using our academic-career-embedded representation increases accuracy while also making the forecasts become more diverse: The one-hot encoding and subject model selects fewer than half of 26 categories in the test set, while the academic-career-embedded representation selects 18 of the 26 categories. These findings suggest that embeddings can aid performance

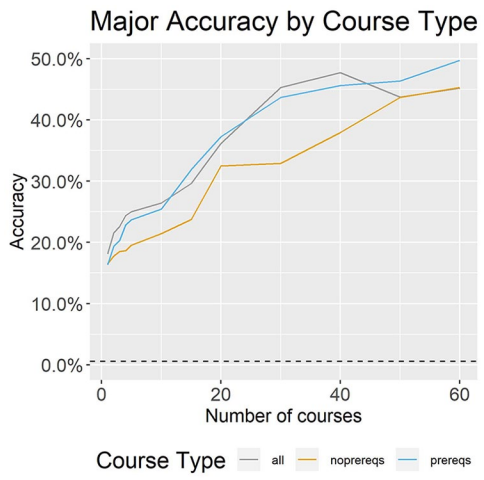

FIGURE 4. *Model accuracy, by course type.*
Note. Each line corresponds to models trained by using one-hot encodings and a subset of courses based on prerequisite type. Accuracy metrics are computed solely on the test set. See the online article for the color version of this figure.

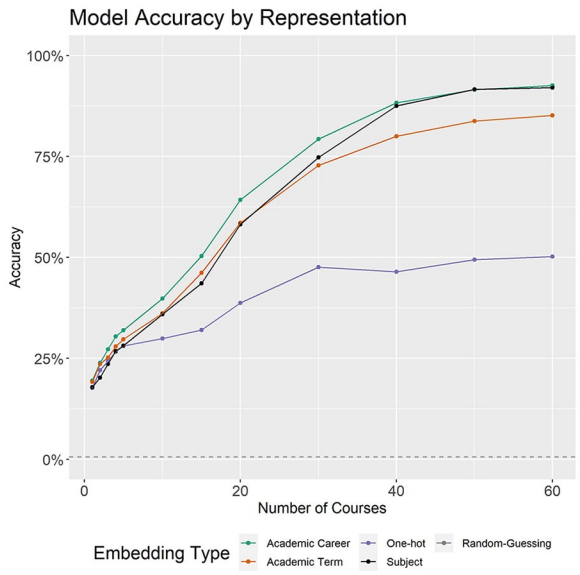

FIGURE 5. *Model accuracy, by representation type.*
Note. Each line corresponds to models trained by using distinct strategies of representing course history. Accuracy metrics are computed solely on the test set. See the online article for the color version of this figure.

on a major-forecasting task, even when using only a single course.

RQ2: Comparing the Predictive Power of Prerequisite and Non-prerequisite Courses

Results are displayed in Figure 4 for models trained with prerequisite courses, non-prerequisite courses, and all courses.[2] Prerequisites outperform non-prequisite courses at all points in the academic career after the first course. The performance difference ranges from –0.1% to 10.7%. Notably, prerequisite courses outperform models trained with all courses later in the academic career.

RQ3: Representation Matters

Our third research question addresses the relative utility of different course representations for the task of major forecasting. We explore model performance with respect to different representations of the data via one-hot encoding, subject dummies, term embedding, and academic-career embedding in Figure 5. The *x*-axis plots the number of courses that each student took before the model was trained and analyzed. The *y*-axis plots overall forecast accuracy aggregated across all categories. The dashed line plots the performance by random guessing. Model performance is consistent with our contention that academic-career embedding would outperform term embedding and one-hot encoding. One-hot encoding performs substantially worse than an embedded representation of courses, at best generating just under 48% accuracy for 40 courses. One-hot encoding also exhibits decreasing performance as the number of courses increases beyond 40, indicating—however counterintuitively—that using more course history to forecast majors does not necessarily improve model accuracy. We do not see this same phenomenon with either the embedded representations or subject dummies.

Academic-career embeddings have between 6% and 90% greater accuracy than one-hot encoding. The gains from switching from term to academic-career embedding are less dramatic, with academic-career embedding improving accuracy by 1% to 10% relative to term embedding. Although comparably small, these differences become statistically significant by the time students have taken their fifth course. Using an exact-binomial paired test between academic-career and academic-term embeddings at the fifth course, we find an associated *p*-value of 0.046.

The simplest approach, using subject dummies, performs somewhere in the middle between these two extremes. Subject dummies perform somewhat more poorly than embedded representation early in the academic career but reach parity with embedded representation once the model learns approximately 40 courses.

We also explore how model performance varies across academic domains. Are these forecasting techniques more and less accurate in engineering, the natural sciences, the social sciences, and arts/humanities? To investigate this question, we plot the performance of term- and academic-career embeddings in Figure 6. We find that engineering and natural sciences are much easier to classify early in academic careers relative to social sciences and arts/humanities. Academic-career embeddings are the strongest performers across all four domains. Striking is that career embedding seems to exhibit the strongest relative gains for arts/

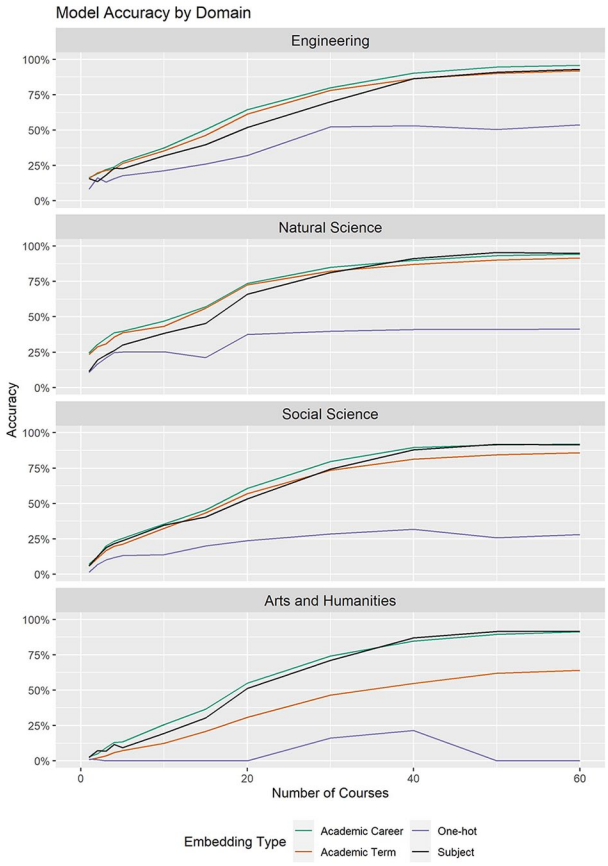

FIGURE 6. *Major classification accuracy, by academic domain. Note.* Each pane corresponds to an academic domain/discipline. Each line corresponds to models trained by using distinct strategies of representing course history. Accuracy metrics are computed solely on the test set. See the online article for the color version of this figure.

humanities majors, for whom progressing through a major is typically thought to be less structured or linear. The gains of career-embedded representation for this category range from 2 to more than 30 percentage points. These findings lend credence to the notion that choice of representation can have a substantial cumulative effect on the major forecasting task.

We speculate that career embeddings outperform at this task for arts/humanities majors because many students take courses in this domain to satisfy general-education requirements. Arts/humanities courses have high co-occurrence with many other courses across disciplines, particularly early in academic careers. By measuring co-occurrence on a longer time horizon, career embeddings are likely better able to distinguish course taking for general-education requirements from course taking for major completion.

Subject dummies tend to perform more poorly than career embeddings across the four academic domains. One-hot encoded representation performed the worst in arts/humanities, where it had an accuracy rate of zero for many of the points we evaluated in academic time.

RQ4: Model Accuracy Across Academic Time

To assess the performance of models at different points of significance in a student's academic career, we present accuracy rates for the 26 major categories of our test set, based on models trained with our academic-career embeddings (see Table 6). We chose points roughly corresponding to the end of the first academic term (five courses), the end of the first academic year (15 courses), and the end of the second academic year, when majors are typically declared (30 courses). We see in our test set that by the end of the first term, only three majors perform at or below levels of a stratified random guesser. Each of these majors (American studies, mathematical and computational sciences, and public policy) has an interdisciplinary focus and comprises at most 2.5% of the overall student population at our case school. At the end of the first year, models for all majors strongly outperform the random-guessing baseline.

In fact, we see an interesting pattern across academic domains: Predicting majors that straddle conceptual areas proves most challenging. The majors mathematics and computation science (MATCS), symbolic systems (SYMBO), public policy (PUBPO), and American studies (AMSTU) are poorly predicted on the basis of courses completed in the first term. It may be that those students who end up in "hybrid" majors spanning multiple academic domains take required and introductory-level courses in more stochastic sequences across their academic careers.

Error Analysis: Variation in the Similarity of Courses Within Majors

Our work thus far has demonstrated that NLP techniques can be powerful tools for forecasting undergraduate major on the basis of courses taken early in the academic career. In a manner analogous to computational "reading" of word sequences in linguistic utterances, NLP models trained on the course sequences of prior students can be used to forecast the ultimate majors of subsequent students. Yet it also is the case that course sequences leading to majors vary across academic domains. To extend the analogy with language, we might imagine that majors may have different "grammars" or required sequence structures. For example, some majors may oblige students to complete sequences of specific courses, while other majors may instead require selections from menus of course offerings. For this reason, efforts to forecast majors would be well served by observing for variation in the composition (or "grammar") of majors within a given institution.

To do this for our case school, we use the one-hot encoded representation of courses and compute Jaccardian similarity between students' course histories. *Jaccardian similarity* is a measure of the intersection over the union of two sets. As

a toy example, consider two students who each took three courses. Student A took $\{ECON1, CS106A, MATH51\}$. Student B took $\{ENGLISH101, PHYSICS10, MATH51\}$. In total, these two course schedules comprise five distinct courses and share one course in common. In this example, Jaccardian similarity is

$$\frac{|A \cap B|}{|A \cup B|} = \frac{|\{MATH51\}|}{|\{ECON1, CS106A, MATH51, ENGLISH101, PHYSICS10\}|} = \frac{1}{5} \quad (3)$$

We deploy this technique across our entire data corpus, computing pairwise Jaccardian similarity between all students' course histories and then averaging them within and across major by academic domain. Findings indicate that, on average, students in the same major experience relatively little of the same academic curriculum (see Figure 7). Chemical-engineering majors exhibit the most similarity, on average having an overlap of approximately one in four courses with other students in that major on their final academic transcripts. In arts/humanities, history exhibits the least self-similarity: On average, a student who graduates with a history major has at most one or two courses in common with others pursuing history majors during the same 20-year period.

Several mechanisms might account for why Western's students exhibit such low course similarity within majors. First, this university has a very large course catalog, which means that many courses may satisfy the major requirements. Second, courses may be renamed or reclassified over time, such that comparing students across a 20-year period may erode forecasting accuracy. Third, our findings are consistent with the notion that majors at Western may have few explicitly required courses in comparison with other schools. Students may have substantial discretion in course selection generally, even while the extent of this discretion varies across majors.

To inform the understanding of the relatively high performance of embedded representation of courses, we plot the T-SNE representation of career-embedded academic history upon a student's graduation (see Figure 8). *T-SNE* is a dimension-reduction algorithm that preserves physical closeness between points (Van der Maaten & Hinton, 2008). Each point in the plot corresponds to a student and their academic history. Although the x- and y-axes should not be interpreted, points that are close in space on the reduced plot can be considered close in space in the higher dimensional space. In each facet, we plot each student's academic career embedding as represented by the T-SNE algorithm. Through this representation, we see well-defined clusters of majors, as instrumented by their shared course enrollments. Courses cluster into clear domains of engineering, social sciences, natural sciences, and arts/humanities. We also observe that conceptually similar topics are found where boundaries are ambiguous. For example, political science and international relations are entangled. Perhaps the most surprising finding is that students who have identical or near-identical embedded representations can have different majors. For example, a political-science major and an international-relations major may have very similar course histories but select different degrees. Such a representation broadly comports with the common imagery of a university as a "city of intellect" (Brint, 2002; Kerr, 2001; Pardos & Nam, 2020), with different precincts of scholars and students sharing common concerns and activities that overlap at the borders between precincts.

These error analyses complement our earlier inquiries: They suggest that part of the reason one-hot encoding is a weaker technique for forecasting majors is that it cannot capture variability in the courses that might compose majors. Few majors at our case school do not offer multiple sequences through even their most commonly enrolled courses. The high dimensionality of this search space, together with the large number of infrequently enrolled courses, make hyperparameter tuning difficult. Although we used the same sets of hyperparameters in our embedded and one-hot encoded results, additional hyperparameter tuning could close only some of the gap between the different representations of course history.

Discussion and Future Work

Elective curricula offer the benefit of choice and flexibility, but these benefits come with potentially costly uncertainty for students, advisors, and administrators. The work presented here illustrates that newly ubiquitous computational techniques can be used to produce insight about current students' academic pathways through concise observation of the course selections of prior students. Borrowing tools honed to observe patterns of sequences of words in large corpora of linguistic texts, we found that early course selections can be powerful tools for forecasting subsequent major issuances. Even a single course—specifically, the very first course a student selects for registration—carries signals about eventual major issuance, while courses taken in the first academic term can be leveraged to forecast eventual major with 32.0% percent accuracy.

There are at least two potential explanatory mechanisms for these findings. The first is that students' academic preferences are substantially set by the time they arrive at college. The second is that early college courses initiate path

TABLE 6
Major accuracy over time

Major	Domain	1st term (%)	1st year (%)	2nd year (%)	Guesser (%)	*n*
CHEME	Engineering	11.8	52.9	85.3	0.02	34
CS	Engineering	62.4	72.5	90.2	0.90	255
EE	Engineering	18.3	63.4	90.1	0.07	71
ENGR	Engineering	5.8	25.4	57.2	0.26	138
MATCS	Engineering	0.0	17.1	41.5	0.02	41
ME	Engineering	30.2	52.8	84.9	0.16	106
MGTSC	Engineering	8.3	43.8	89.6	0.13	96
SYMBO	Engineering	1.4	30.4	76.8	0.07	69
Subtotal	Engineering	27.8	50.5	80.0	0.38	810
BIO	Natural science	18.6	40.2	76.3	0.13	97
BIOL	Natural science	39.5	67.4	86.0	0.10	86
CHEM	Natural science	3.1	25.0	81.2	0.01	32
EASYS	Natural science	18.4	42.1	81.6	0.02	38
HUMBI	Natural science	55.2	69.3	93.8	0.80	241
MATH	Natural science	40.4	44.2	67.3	0.04	52
PHYS	Natural science	50.0	58.3	77.8	0.02	36
Subtotal	Natural science	39.9	57.0	84.9	0.38	582
ECON	Social science	46.8	63.3	87.8	0.49	188
INTLR	Social science	17.0	40.0	74.8	0.25	135
POLSC	Social science	18.9	36.7	75.6	0.11	90
PSYCH	Social science	12.1	38.3	80.4	0.16	107
PUBPO	Social science	0.0	15.8	65.8	0.02	38
Subtotal	Social science	25.3	45.3	79.7	0.28	558
AMSTU	Arts and humanities	0.0	9.4	62.5	0.01	32
COMMU	Arts and humanities	7.0	41.9	81.4	0.03	43
ENGL	Arts and humanities	10.8	44.1	82.4	0.14	102
HSTRY	Arts and humanities	14.3	39.3	78.6	0.10	84
STS	Arts and humanities	25.3	31.6	59.5	0.09	79
Subtotal	Arts and humanities	13.5	36.5	74.1	0.09	340
Other		54.1	59.1	73.7	2.20	399
Total		32.0	50.4	79.3	0.59	2689

Note. Each row corresponds to forecasting accuracy for each major by using just the career-embedded representation and a number of courses that correspond roughly to the end of the first academic term (five courses), the end of the first academic year (15 courses), and the end of the second academic year (30 courses). *Guesser* corresponds to a stratified random-guessing strategy. This table only corresponds to performance on the test set.

dependencies that substantially influence major choice. These explanations are not exclusive; indeed, prior work on academic pathways suggests some combination of these processes (Chambliss & Takacs, 2014; Page & Scott-Clayton, 2016; San Pedro et al., 2014). Regardless of the mechanism, our findings suggest that students, advisors, and academic administrators working in schools with elective curricula should take students' initial course selections very seriously. Whether the concern is the extent to which students "funnel" toward a small number of popular majors (Binder et al., 2016) or the best allocation of scarce instructional resources 1 or 2 years down the line, early course selections are informative data points.

An important empirical limitation of our work—and, indeed, any inquiries relying solely on administrative data—is that it cannot capture how students make sense of the tasks of considering, choosing, and sequencing courses and deciding on major fields of study. An ideal next step would be to link information derived from transcripts with qualitative methods (interviews, open-ended surveys, or focus groups, for example) designed to capture how students think about their course and major selections *in situ*, over time. We are happy to report that work in this vein is already underway (Harrison et al., 2022).

Although the work presented here is a case study of students in a single school, it provides compelling evidence that

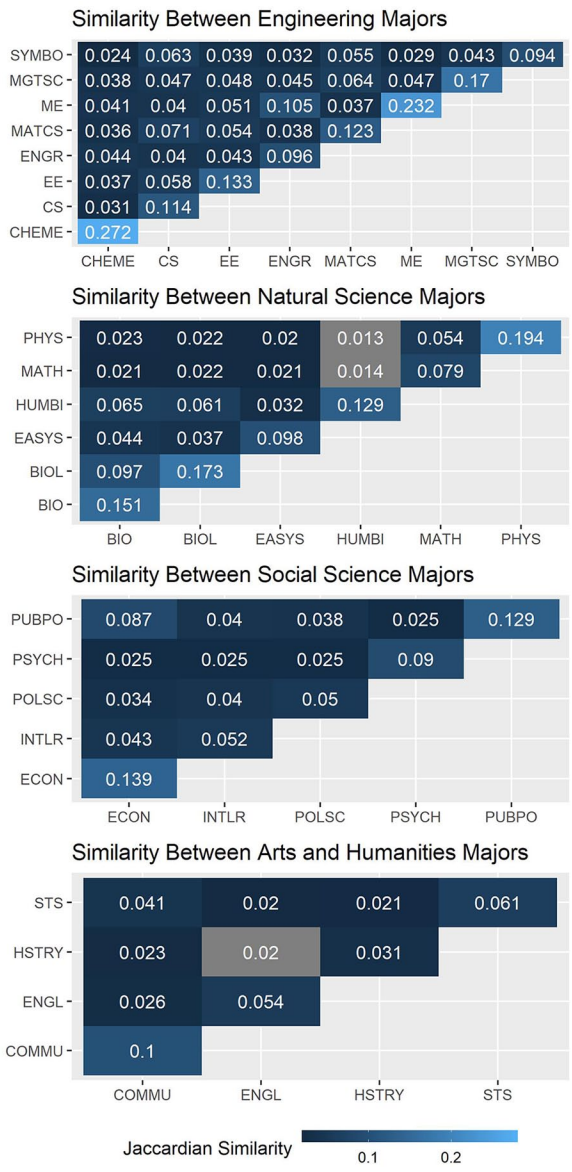

FIGURE 7. *Jaccardian similarity of majors, by domain.*
Note. Each cell in the figure computes the Jaccardian similarity across all possible pairs of students within each corresponding major. The right-most cell of each row corresponds to the average similarity of each student within its own major. See the online article for the color version of this figure.

academic transcripts can yield useful insights for students, academic advisors, and administrators. Students may have little understanding of the path dependencies they might be setting in motion when selecting courses early in their academic careers. Students may not appropriately weigh the signals or appreciate that sparsely offered courses and prerequisite-laden majors can exclude entire academic paths (Slim et al., 2014; Thompson, 2021). Software applications deploying data and analytic techniques such as those presented here could help students see the potential consequences of early course choices and encourage them to be more purposeful in their consideration of each course. Academic advisors might use this information to caution students about the potential narrowing or foreclosing of academic options as their selections accumulate. Administrators might use these same forecasts to inform the allocation of scarce instructional resources across programs.

As these technologies are incorporated into administrative practices in higher education, several practical and ethical considerations should be noted. First, it will be important to understand the extent to which algorithms such as those developed here perform relative to human advisors, who almost surely rely on their own experiences and local heuristics to make predictions about trajectories into majors. Second, we have presented models that have varying levels of interpretability and scrutability. The extent to which students and administrators trust each type of model warrants future study (Kizilcec, 2016). Finally, to the extent that academic exploration and discovery are fundamentally human endeavors, it is worth considering how novel computational technologies can augment—not replace—the work of human advisors in guiding students on their academic paths (Snyder et al., 2022).

We recognize that our case school is, in many ways, an outlier in the national postsecondary ecology. It is a residential campus with a very high graduation rate, and most of its students remain continuously enrolled full-time from entry to graduation. These students are presented a plethora of courses to choose from and have very few constraints on their choices. At the same time, Western's peculiarities make it a good site for an initial case study of major forecasting: The complexity of Western's curriculum, coupled with the relative similarity in how its students transit from entry to graduation, makes it amenable to assessing the general viability of this approach to modeling academic paths to majors. We encourage replication of our general analytic strategy across a variety of schools with different curricular structures and characteristic undergraduate careers.

Our study of the comparative composition of majors also raises interesting questions about the character and meaning of the declared major itself as a representation of course history. Similarity analyses indicate—at least at our case school—that a declared major provides relatively little information about any given student's specific course history. Even in majors exhibiting the most similar course compositions, students have less than a quarter of their courses in common. Further, the great extent to which students with similar course histories can commit to different majors suggests that course choices and major choices may be decoupled—that is, students may choose majors for reasons other than the portfolios of coursework they imply. In any case, we hope that our finding spurs more research on major composition across institutions and over time. Given sufficiently granular transcript data, researchers might potentially decompose the earning power of a degree between its

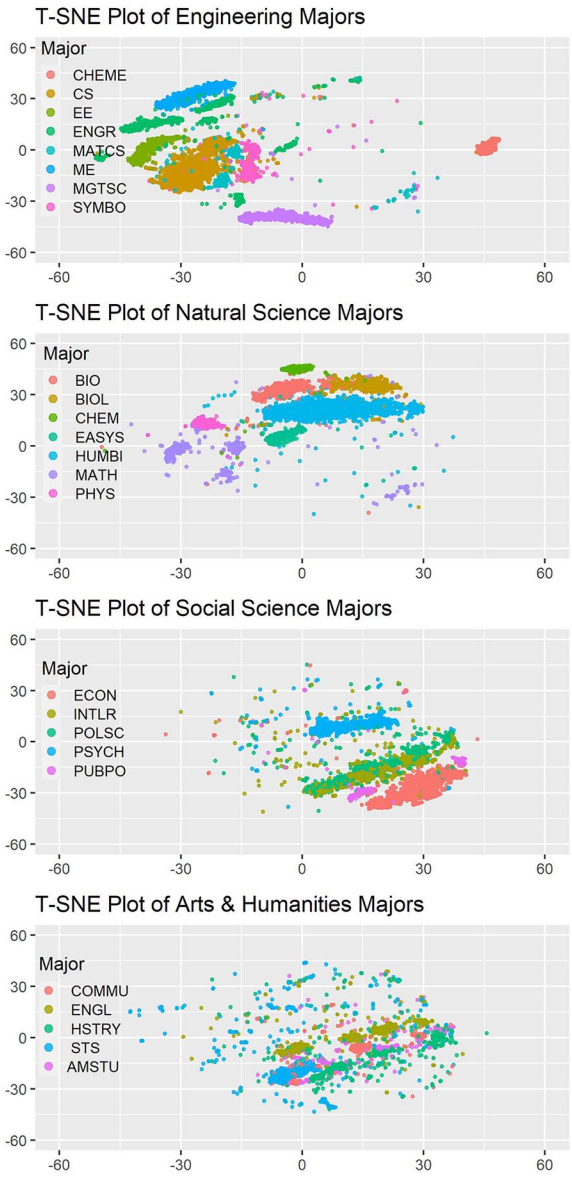

FIGURE 8. *T-SNE plot of majors, by domain.*
Note. Each point corresponds to the embedded representation of a student's academic history at the end of their academic career. Each point is colored to correspond to the student's terminal major. See the online article for the color version of this figure.

signaling value and its human-capital contributions (Arteaga, 2018; Spence, 1973). Historical analysis of course-transcript data might provide additional insight into the college-wage premium and skill-biased technological change over the past several decades (Goldin & Katz, 2007).

We have pursued this work as a first step toward developing a portfolio of computational tools for modeling undergraduate academic careers. Although illuminating, the methods we have used here cannot capture the temporal character of student decision-making. An important next step in this work would be to augment our machine-learning algorithm with some form of sequential modeling, such as recurrent neural networks. Such techniques would enable researchers to incorporate additional features of undergraduate careers, such as major-switching. Future work should also incorporate student demographic data and course grades to better capture the complexity of student decision-making (Carrell et al., 2010; Harrison et al., 2022; Owen, 2010).

Finally, we see great promise in using NLP and neural-network models in tandem with other technical methods to describe academic pathways. Others have found valuable affordances with network analysis to describe the relationships between major requirements (Baker & Huntington-Klein, 2018; Slim et al., 2014) and between students via co-enrollments (Weeden & Cornwell, 2020). Developing a comparative applied science of undergraduate academic careers may enable researchers and academic advisors to formulate alternative representations of curricula that are more navigable, reducing the cognitive load of course selection and perhaps also time to degree (Alkaoud & Pardos, 2019). Given the wage premiums associated with college completion, pressures to reduce college costs, and a growing interest in alternatives to college, ongoing efforts to leverage insight from ubiquitous academic data are work worth doing.

Authors' Note

We are grateful to the administrative leadership of Western University for their crucial enabling of this work. Ben Domingue, Nick Haber, Arik Lifschitz, and Ben Listyg provided input on earlier drafts, as did serial audiences of the Stanford Pathways Lab (pathwayslab.stanford.edu). The work was much improved by comments from AERA editors and external reviewers. The research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B140009 to the Board of Trustees of the Leland Stanford Junior University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

ORCID iD

David Lang https://orcid.org/0000-0001-5415-0125

Open Practices

The analysis code and data availability statement can be found at https://doi.org/10.3886/E175541V1 on ICPSROPEN

Notes

1. Data from 2020 overwhelmingly contain course selections made before the case university made administrative responses to the COVID-19 pandemic in March of that year.

2. In examining the differences of prediction power from using prerequisite versus unrelated courses, we worry that the use of an embedded representation may unintentionally leak data about students' course history across prerequisite and non-prerequisite course sets. As such, we use one-hot encoding for this analysis.

References

Abbott, A. (2010). *Chaos of disciplines*. University of Chicago Press.

Alkaoud, M., & Pardos, Z. A. (2019). *Degree curriculum contraction: A vector space approach*. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren & R. Luckin (Eds.), Artificial Intelligence in Education (pp. 14–18). Springer.

Arteaga, C. (2018). The effect of human capital on earnings: Evidence from a reform at colombia's top university. *Journal of Public Economics*, *157*, 212–225.

Arthurs, N., & Alvero, A. J. (2020). Whose truth is the "Ground Truth"? College admissions essays and bias in word vector evaluation methods. In A. Rafferty, J. Whitehill, C. Romero & V. Cavalli-Sforza. (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)* (pp. 342–349). International Educational Data Mining Society.

Aulck, L., Aras, R., Li, L., L'Heureux, C., Lu, P., & West, J. (2017). Stem-ming the tide: Predicting STEM attrition using student transcript data. *arXiv:1708.09344*

Aulck, L., & West, J. (2017). Attrition and performance of community college transfers. *PloS One*, *12*(4), e0174683.

Bailey, T., Jaggars, S. S., & Jenkins, D. (2015). *What we know about guided pathways: Helping students to complete programs faster. Research overview*. Community College Research Center, Teachers College, Columbia University.

Baker, R. (2018). Understanding college students' major choices using social network analysis. *Research in Higher Education*, *59*(2), 198–225.

Baker, R., & Huntington-Klein, N. (2018). *Student preference for guidance and complexity in college major requirements*. CEPA Working Paper no. 18-06. Stanford Center for Education Policy Analysis.

Barkan, O., & Koenigstein, N. (2016). *Item2vec: Neural item embedding for collaborative filtering*. In F. Palmieri, A. Uncini, K. Diamantaras & J. Larsen (Eds.), 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP) (pp. 1–6). IEEE.

Beaulac, C., & Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, *60*(7), 1048–1064.

Binder, A. J., Davis, D. B., & Bloom, N. (2016). Career funneling: How elite students learn to define and desire "prestigious" jobs. *Sociology of Education*, *89*(1), 20–39.

Bleemer, Z., & Mehta, A. (2020, April). Will studying economics make you rich? A regression discontinuity analysis of the returns to college major. *American Economic Journal: Applied Economics*, *14*(2), 1–22.

Bleemer, Z., & Mehta, A. (2021). College major restrictions and student stratification. (EdWorkingPaper No. 21-502). Annenberg Institute for School Reform at Brown University.

Brint, S. G. (2002). *The future of the city of intellect: The changing American university*. Stanford University Press.

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *Quarterly Journal of Economics*, *125*(3), 1101–1144.

Caselles-Dupré, H., Lesaint, F., & Royo-Letelier, J. (2018). *Word2vec applied to recommendation: Hyperparameters matter*. In Proceedings of the 12th ACM Conference on Recommender Systems (pp. 352–356). October 2–7, Vancouver, BC, Canada.

Chambliss, D. F., & Takacs, C. G. (2014). *How college works*. Harvard University Press.

Chaturapruek, S., Dalberg, T., Thompson, M. E., Giebel, S., Harrison, M. H., Johari, R., Stevens, M. L., & Kizilcec, R. F. (2021). Studying undergraduate course consideration at scale. *AERA Open*. https://doi.org/10.1177/2332858421991148

Chaturapruek, S., Johari, R., Kizilcec, R., Stevens, M., Bernstein, M., Cina, S., Harrison, M., Lifschitz, A., Mitchell, J., Paepcke, A., et al. (2019). *Understanding undergraduate course consideration*. https://ssrn.com/abstract=3432748

Chen, Y., Johri, A., & Rangwala, H. (2018). *Running out of STEM: A comparative study across STEM majors of college students at-risk of dropping out early*. In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (pp. 270–279).

Chockkalingam, S., Yu, R., & Pardos, Z. A. (2021, April 12–16). *Which one's more work? Predicting effective credit hours between courses*. In LAK21: 11th International Learning Analytics and Knowledge Conference (pp. 599–605). Irvine, CA, USA.

Delbanco, A. (2014). *College: What it was, is, and should be—updated edition*, *vol. 82*. Princeton University Press.

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). *Why does unsupervised pre-training help deep learning?* In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (pp. 201–208).

Fink, J., Jenkins, P. D., Kopko, E. M., & Ran, X. (2018). *Using data mining to explore why community college transfer students earn bachelor's degrees with excess credits*. CCRC Working Paper no. 100. Community College Research Center.

Fricke, H., Grogger, J., & Steinmayr, A. (2018). Exposure to academic fields and college major choice. *Economics of Education Review*, *64*, 199–213.

Goldin, C., & Katz, L. F. (2007). *The race between education and technology: The evolution of U.S. educational wage differentials, 1890 to 2005*. NBER Working Paper 12984. National Bureau of Economic Research.

Gottfried, M. A., & Bozick, R. (2016). Supporting the stem pipeline: Linking applied stem course-taking in high school to declaring a stem major in college. *Education Finance and Policy*, *11*(2), 177–202.

Haggag, K., Patterson, R., Pope, N. G., & Feudo, A. (2021). Attribution bias in major decisions: Evidence from the United States Military Academy. *Journal of Public Economics*, *200*. https://doi.org/10.1016/j.jpubeco.2021.104445

Harrison, M. H., Hernandez, P. A., & Stevens, M. L. (2022). Should I start at math 101? Content repetition as an academic strategy in elective curriculums. *Sociology of Education*, *95*(2). https://doi.org/10.1177/00380407221076490

Jiang, W., & Pardos, Z. A. (2020). Evaluating sources of course information and models of representation on a variety of institutional prediction tasks In: A. Rafferty, J. Whitehill, C. Romero & V. Cavalli-Sforza. (Eds.), *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)* (pp. 115–125). virtual conference.

Kerr, C. (2001). *The uses of the university*. Harvard University Press.

Kizilcec, R. (2016). *How much information? Effects of transparency on trust in an algorithmic interface*. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 2390–2395).

Leppel, K. (2001). The impact of major on college persistence among freshmen. *Higher Education*, *41*(3), 327–342.

Liu, V., Mishra, S., & Kopko, E. M. (2021). Major decision: The impact of major switching on academic outcomes in community colleges. *Research in Higher Education*, *62*(4), 498–527.

McFarland, D. A. (2006). Curricular flows: Trajectories, turning points, and assignment criteria in high school math careers. *Sociology of Education*, *79*(3), 177–205.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv:1301.3781*

Owen, A. (2010). Grades, gender, and encouragement: A regression discontinuity analysis. *Journal of Economic Education*, *41*(3), 217–234.

Page, L. C., & Scott-Clayton, J. (2016). Improving college access in the United States: Barriers and policy responses. *Economics of Education Review*, *51*, 4–22.

Pardos, Z. A., Chau, H., & Zhao, H. (2019a). *Data-assistive course-to-course articulation using machine translation*. In Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale (pp. 1–10).

Pardos, Z. A., Fan, Z., & Jiang, W. (2019b). Connectionist recommendation in the wild: On the utility and scrutability of neural networks for personalized course guidance. *User Modeling and User-Adapted Interaction*, *29*(2), 487–525.

Pardos, Z. A., & Jiang, W. (2019). Combating the filter bubble: Designing for serendipity in a university course recommendation system. *arXiv preprint arXiv:1907.01591*.

Pardos, Z. A., & Nam, A. J. H. (2020). A university map of course knowledge. *PloS One*, *15*(9), e0233207.

Patterson, R. W., Pope, N. G., & Feudo, A. (2021). Timing matters: Evidence from college major decisions. *Journal of Human Resources*. https://doi.org/10.3368/jhr.59.2.0820-11127R1

Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, *3*(2).

Rosenbaum, J. E., Deil-Amen, R., & Person, A. E. (2007). *After admission: From college access to college success*. Russell Sage Foundation.

San Pedro, M. O., Ocumpaugh, J., Baker, R. S., & Heffernan, N. T. (2014). *Predicting STEM and non-STEM college major enrollment from middle school interaction with mathematics educational software*. In Proceedings of the 7th International Conference on Educational Data Mining (pp. 276–279).

Shao, E., Guo, S., & Pardos, Z. A. (2021). *Degree planning with PLAN-BERT: Multi-semester recommendation using future courses of interest*. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 14920–14929).

Slim, A., Kozlick, J., Heileman, G. L., Wigdahl, J., & Abdallah, C. T. (2014). *Network analysis of university courses*. In Proceedings of the 23rd International Conference on World Wide Web (pp. 713–718).

Snyder, C., Keppler, S., & Leider, S. (2022). *Algorithm reliance under pressure: The effect of customer load on service workers*. http://dx.doi.org/10.2139/ssrn.4066823

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*(3), 355–374.

Thompson, M. (2021). Grade expectations: The role of first-year grades in predicting the pursuit of STEM majors for first-and continuing-generation students. *Journal of Higher Education*, *92*, 961–985.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Weeden, K. A., & Cornwell, B. (2020). The small-world network of college classes: implications for epidemic spread on a university campus. *Sociological Science*, *7*, 222–241.

Weeden, K. A., Gelbgiser, D., & Morgan, S. L. (2020). Pipeline dreams: Occupational plans and gender differences in stem major persistence and completion. *Sociology of Education*, *93*(4), 297–314.

Witteveen, D., & Attewell, P. (2021). Delayed time-to-degree and post-college earnings. *Research in Higher Education*, *62*(2), 230–257.

Wolniak, G. C., & Pascarella, E. T. (2005). The effects of college major and job field congruence on job satisfaction. *Journal of Vocational Behavior*, *67*(2), 233–251.

Zou, Q., Xing, P., Wei, L., & Liu, B. (2019). Gene2vec: Gene subsequence embedding for prediction of mammalian N^6-methyladenosine sites from mRNA. *RNA*, *25*(2), 205–218.

Authors

DAVID LANG obtained his doctorate in the economics of education at Stanford University. His research interests include applications of NLP, machine learning, and causal inference to problems in higher education and learning analytics.

NATHAN DALAL has a master's in artificial intelligence and computer science from Stanford University. He is broadly interested in educational applications and machine learning.

ALEX WANG has a master's in artificial intelligence and computer science from Stanford University. He is broadly interested in NLP and algorithmic fairness.

ANDREAS PAEPCKE is a senior research scientist at Stanford University. His interests include user interfaces and systems for teaching and learning.

MITCHELL L. STEVENS is a professor of education at Stanford, where he co-directs the Pathways Lab (pathwayslab.stanford.edu).