

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Virtual Choreographies of Learning: AI-Driven Discovery in K–12 Online American Schools

Eduardo Salé Areias

WORKING VERSION



FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Fernando Cassola Marques

Co-Supervisor: Prof. Dennis Beck

January 31, 2026

Resumo

As *Connections Academy*, apoiadas pela Pearson Online and Blended Learning, oferecem escolas totalmente online do Jardim-de-Infância ao 12.º ano em 32 estados dos EUA, servindo mais de 100 000 alunos, muitos dos quais transitaram do ensino presencial devido a problemas anteriores. Relatórios nacionais indicam piores resultados médios em exames padronizados para alunos de escolas online a tempo inteiro, embora estas escolas matriculem mais estudantes em risco, dificultando comparações diretas. Apesar da escala, existe pouca investigação sobre o que os alunos efectivamente fazem nestes ambientes e como os seus padrões de estudo e interação com docentes/colegas se relacionam com o progresso. Esta dissertação analisa dados das 32 escolas usando o enquadramento de *coreografias virtuais* do INESC TEC para identificar sequências recorrentes de ações de aprendizagem online e a sua ligação ao sucesso. Para tal, explora técnicas de IA como *k*-means, t-SNE/UMAP e modelos supervisionados (árvores de decisão, redes neuronais leves, *random forests*) para descobrir *clusters* e trajetórias comportamentais menos óbvias e enriquecer a compreensão da dinâmica de aprendizagem em escolas K–12 totalmente online.

Abstract

Connections Academy schools, supported by Pearson Online and Blended Learning, provide fully online K–12 education across 32 U.S. states and currently serve over 100,000 students. Many families choose these schools because their children faced specific problems in prior in-person settings. National reports indicate that students enrolled in full-time online schools often perform worse on standardized exams than peers in traditional public schools; however, fully online schools enroll a higher proportion of at-risk students, which complicates fair comparisons.

Despite the scale of this model, there is limited research on what actually happens in these learning environments: Which patterns of study do students follow? How do they engage with teachers and peers? And how are these patterns connected to learning progress? This dissertation addresses these questions by analyzing data from the 32 Connections Academy schools using the *virtual choreographies* framework developed at INESC TEC. The goal is to identify recurring sequences of online learning actions and examine their relevance for student success.

To that end, the work explores Artificial Intelligence techniques to detect and characterize behavior patterns and trajectories, including unsupervised methods such as k -means and dimensionality-reduction techniques (t-SNE/UMAP), alongside supervised models such as decision trees, lightweight neural networks, and random forests. These approaches are expected to surface non-obvious configurations of behavior and to enrich our understanding of student learning dynamics in fully online K–12 schools.

Índice

1	Thesis Manifest	1
1.1	Context	1
1.2	Motivation	1
1.3	Problem	1
1.4	Research Questions	2
1.5	Hypotheses	2
2	Literature Review	3
2.1	Overview	3
2.2	Type of review	3
2.3	Search strategy	4
2.4	Inclusion and exclusion criteria	4
2.4.1	Inclusion criteria	5
2.4.2	Exclusion criteria	5
2.5	Screening and selection process	5
2.5.1	Stage 1: Title and abstract screening	5
2.5.2	Stage 2: Full-text assessment	6
2.5.3	Stage 3: Citation-based refinement	6
2.6	Effectiveness of K-12 Online Learning	6
2.7	Supporting At-Risk Populations	7
2.8	Engagement and Interaction Patterns	7
2.9	Data management	8
2.10	Limitations and threats to validity	8
3	State of the Art	9
3.1	Overview	9
3.2	Educational Data Mining	9
3.3	Modeling Student Behavior	10
3.3.1	Educational Process Mining (EPM)	10
3.3.2	The Virtual Choreographies Framework	10
3.4	Unsupervised Learning Techniques	11
3.4.1	Clustering Algorithms	11
3.4.2	Dimensionality Reduction	11
3.5	Summary	11
4	Proposed Approach	12
4.1	Overview	12
4.2	Methodological Architecture	12

4.2.1	Stage 1: Data Collection and Context	12
4.2.2	Stage 2: Semantic Abstraction (The Virtual Choreography)	13
4.2.3	Stage 3: Pattern Discovery	13
4.3	Experimental Process and Evaluation Design	14
4.3.1	Stability and Robustness (Addressing RQ1 & RQ2)	14
4.3.2	Association with Outcomes (Addressing RQ3)	14
4.3.3	Predictive Utility (Addressing RQ4)	14
4.4	Expected Results	15
4.5	Work Plan	15

Chapter 1

Thesis Manifest

1.1 Context

Connections Academy, supported by Pearson Online and Blended Learning, operates fully online public schools serving K–12 students across 32 U.S. states and enrolling over 100,000 learners. Independent national reviews repeatedly find that students in full-time virtual schools, on average, underperform peers in traditional brick-and-mortar public schools on standardized measures [13]. However, the literature suggests these schools often serve a distinct population of “at-risk” students who have opted out of traditional schooling due to medical, social, or academic challenges [3]. These complex demographics make it crucial to understand what students actually *do* in virtual classrooms—beyond simple attendance—and how their behavioral routines relate to learning progress.

1.2 Motivation

Despite the scale of full-time online schooling, there is limited empirical detail about students’ fine-grained learning activities—how they navigate content, participate in live sessions, and interact with teachers and peers—and how these behaviors connect to outcomes [10]. Addressing this evidence gap can inform practice in large online school networks and help target support to the students who need it most, particularly those whose engagement patterns signal early risk of failure.

1.3 Problem

This dissertation investigates *virtual choreographies*—recurring sequences of online learning actions—as a lens to characterize K–12 students’ behavior at scale and to examine how such patterns are associated with academic success. Concretely, the aim is to identify, describe, and relate behavior patterns observed in activity logs to indicators of performance in fully online schools, utilizing the platform-independent framework proposed by Cassola et al. [7].

1.4 Research Questions

RQ1 Identification. Which *virtual choreographies* (recurring sequences of online learning actions) can be reliably discovered in Connections Academy data using unsupervised clustering, and how stable are they over time?

RQ2 Robustness across contexts. To what extent do the discovered choreographies replicate across grades, subjects, schools, and cohorts (i.e., are they consistent behavioral habits rather than school- or cohort-specific artifacts)?

RQ3 Association with outcomes. How are choreography *attributes*—such as temporal regularity, balance between synchronous and asynchronous participation, and frequency of feedback—associated with indicators of student success (e.g., progress, grades), when controlling for student background?

RQ4 Predictive utility. Do choreography-based representations improve early prediction of student success over standard activity metrics, and how early in the semester can such signals be detected?

RQ5 Actionability and Visualization. Which components within the choreographies (specific actions, transitions, or rhythms) most contribute to outcomes, and how can these be visualized to support teacher decision-making?

1.5 Hypotheses

Definition (Virtual Choreography). A *virtual choreography* is the structured set of actions carried out among agents (e.g., student, teacher, platform tools) that unfolds over time and within a specific learning space, abstracting specific platform clicks into semantic behaviors [7].

H1 — Regularity. Choreographies exhibiting stable temporal rhythms (e.g., consistent study times and revisits across the week) are positively associated with academic success.

H2 — Balanced participation. Choreographies that balance synchronous participation (live sessions) and asynchronous work (content study, assignment submission) relate to better outcomes than predominantly one-sided patterns.

H3 — Feedback-centric interaction. Choreographies with frequent teacher–student feedback exchanges (asking questions, reading feedback, revising submissions) are positively associated with success.

H4 — Purposeful navigation. Choreographies characterized by efficient, task-aligned navigation (timely access to resources, limited detours) are associated with stronger progress than fragmented, sporadic engagement.

H5 — Risk signals (inverse). Choreographies marked by long inactivity gaps, last-minute bursts, or minimal interaction with teachers/peers correlate with weaker outcomes.

Chapter 2

Literature Review

2.1 Overview

This chapter describes the methodology used to identify and select the scientific literature relevant to this dissertation. It explains the type of literature review adopted, the search strategy, the inclusion and exclusion criteria, and the process followed to screen and select the final set of papers that compose the document corpus. In line with the thesis manifest, the review focuses on research about K–12 online learning and full-time virtual schools in the United States, with particular emphasis on their effectiveness, challenges and implications for students, teachers and school systems. This corpus provides the conceptual and empirical background needed to frame the later analysis of virtual choreographies of learning in Connections Academy schools.

2.2 Type of review

Given the objectives of this dissertation, a structured literature review was conducted, following principles of a systematic literature review while allowing some flexibility that is closer to a narrative synthesis. The goal of the review is to identify, select and document existing research on K–12 online learning and virtual schools, so as to support the definition of the dissertation research questions and to provide a solid basis for the later State of the Art chapter and for the empirical analysis of student activity data.

The corpus prioritizes contemporary systematic reviews and empirical studies that reflect the current state of K–12 online learning technologies. Instead of early foundational work, the review relies on recent comprehensive analyses, such as the systematic reviews by Martin et al. [12] and Johnson et al. [10]. It also incorporates critical studies on student outcomes, engagement and school-level implementation of K–12 online programmes, specifically those addressing at-risk populations as explored by Beck [3, 4] and Toppin and Toppin [15], as well as earlier warnings about the maturity of the field [1].

2.3 Search strategy

The literature search was carried out in 2025, in the months leading up to this deliverable. The primary search engine used was Google Scholar, which offers broad coverage of peer-reviewed publications in education and related fields. In addition, two discovery tools: Litmaps and ResearchRabbit were used to expand the initial set of articles through citation-based exploration and visualisation of related work.

The search targeted peer-reviewed journal articles, conference papers and book chapters written in English. Given the rapid evolution of educational technology and the impact of the COVID-19 pandemic, the search prioritized studies published between **2015 and 2025**. This timeframe ensures that the analyzed studies reflect the current technological infrastructure (modern LMS and learning analytics capabilities) and the post-pandemic reality of K-12 online education.

The main search string used in Google Scholar was iteratively refined, but was based on combinations of the following keywords:

- **Core topic terms:** “K-12 online learning”, “virtual schools”, “virtual schooling”, “online teaching”, “online school”, “distance education”.
- **Outcome / focus terms:** “effectiveness”, “student achievement”, “engagement”, “school choice”, “best practices”, “teacher training”, “at-risk students”.
- **Context terms:** “K-12”, “secondary education”, “United States”, “high school”.

An example of a typical query used in Google Scholar is:

“K-12 online learning” OR “virtual schools” OR “virtual schooling” AND (effectiveness OR achievement OR engagement OR “distance education”)

The initial Google Scholar searches generated a broad set of potentially relevant results. These references were exported and then imported into Litmaps and ResearchRabbit. From this initial seed set, both tools were used to:

- identify additional papers that frequently cite or are cited by the seed articles;
- visualise clusters of research focusing on K–12 virtual schools and online programmes;
- surface key works such as large-scale reviews, empirical studies on student outcomes, and research on teacher practices in virtual schools.

Through this process, an initial pool of approximately 150 records was assembled before applying the inclusion and exclusion criteria described below.

2.4 Inclusion and exclusion criteria

To ensure the relevance and quality of the selected literature, explicit inclusion and exclusion criteria were defined and applied during the screening process.

2.4.1 Inclusion criteria

A study was included in the corpus if it met all of the following criteria:

- **IC1** – The study focuses on K–12 online learning, virtual schools or virtual schooling (fully online or primarily online programmes).
- **IC2** – The study is peer-reviewed (journal article, conference paper or book chapter).
- **IC3** – The study is written in English.
- **IC4** – The study was published between **2015 and 2025**.
- **IC5** – The study addresses at least one of the following aspects: effectiveness or outcomes of virtual schooling (e.g., student achievement, completion, equity), teaching and instructional practices in online K–12 settings, student engagement and experiences, or system-level issues such as school choice, policy and implementation of virtual schools.

2.4.2 Exclusion criteria

Studies were excluded if any of the following applied:

- **EC1** – The main focus is higher education, adult education or corporate training, rather than K–12 schooling.
- **EC2** – The publication type is a thesis, dissertation, report, blog post, poster, editorial or other non-peer-reviewed document.
- **EC3** – The study is not available in full text.
- **EC4** – The study deals with general educational technology or blended learning without a clear focus on fully or primarily online K–12 programmes.
- **EC5** – The paper is purely theoretical or opinion-based without sufficient connection to K–12 online or virtual schooling.

2.5 Screening and selection process

The selection process followed three main stages: (i) title and abstract screening, (ii) full-text assessment, and (iii) citation-based expansion and refinement.

2.5.1 Stage 1: Title and abstract screening

After removing duplicates from the combined Google Scholar, Litmaps and ResearchRabbit exports, approximately 150 unique records remained. Titles and abstracts were screened against the inclusion and exclusion criteria. At this stage, studies that clearly focused on higher education, general ICT in education or non-online contexts were removed. This first screening excluded around 90 records, leaving approximately 60 papers for full-text assessment.

2.5.2 Stage 2: Full-text assessment

The full text of the remaining studies was examined in more detail. Papers that mentioned online learning only superficially, or that did not present results or discussion directly related to K–12 virtual schools, were excluded. This stage reduced the set to a core group of studies that provide substantial evidence on K–12 online learning. This includes recent systematic reviews [10, 12] and research on student experiences in online K–12 environments, with a particular focus on support for at-risk students as emphasized by Beck [4, 5]. After full-text assessment, around 20 papers remained.

2.5.3 Stage 3: Citation-based refinement

The core set of papers was then used for backward and forward snowballing. Their reference lists and citation networks were explored using Litmaps and ResearchRabbit. This step led to the identification of additional key works, including highly cited recent reviews that had not appeared in the initial Google Scholar rankings. After applying the same inclusion and exclusion criteria to these additional articles and removing new duplicates, the final corpus consisted of **12 highly relevant papers**. This set represents the state-of-the-art in the field and includes:

- Systematic reviews of K–12 online learning effectiveness and teaching practices, such as those by Martin et al. [12], Johnson et al. [10], and Barbour and Hodges [2];
- Empirical studies on student engagement, particularly regarding at-risk populations and support structures, including recent work by Beck [3, 4, 5] and Curtis and Werth [9];
- Analyses of policy, school choice and the growth of K–12 online learning, as discussed by Molnar et al. [13].

The detailed bibliographic information for these papers is provided in the Bibliography at the end of this document.

2.6 Effectiveness of K-12 Online Learning

A central theme in the selected literature is the comparative effectiveness of full-time virtual schools versus traditional brick-and-mortar settings. Large-scale annual reports, most notably those by the National Education Policy Center (NEPC), consistently highlight significant performance gaps. Molnar et al. [13] report that virtual schools generally exhibit lower graduation rates and lower proficiency rates on standardized assessments when compared to national averages. They argue that the rapid expansion of these schools often outpaces the development of effective regulatory frameworks and instructional quality assurance.

However, recent systematic reviews caution against simple binary comparisons between “online” and “offline” schooling. Martin et al. [12] and Johnson et al. [10] synthesize two decades of research to suggest that effectiveness is not an inherent property of the medium but is highly

dependent on instructional design, teacher preparation, and student support structures. Johnson et al. [10] specifically note that while the *average* performance may be lower, there are specific contexts—such as credit recovery or advanced coursework—where online learning provides essential opportunities that would otherwise be unavailable. Barbour and Hodges [2] further emphasize that the “age of disruptions” (post-COVID) requires a shift from questioning *if* online learning works to determining *under what conditions* it is effective.

2.7 Supporting At-Risk Populations

A critical finding in the literature is that the demographic profile of students in full-time virtual schools differs significantly from traditional public schools. Toppin and Toppin [15] and Beck [3] observe that virtual schools frequently attract “at-risk” students—those who have experienced bullying, medical issues, or academic failure in face-to-face settings. This selection bias complicates direct performance comparisons, as virtual schools are often serving a population with higher pre-existing needs.

Success for these learners relies heavily on support beyond the screen. Beck’s extensive research [4, 5] highlights the pivotal role of the “on-site facilitator” or “advocate”—typically a parent or guardian. In a study of cyber schools during the pandemic, Beck and Levine [5] found that the level of parental engagement and their ability to structure the student’s physical learning environment were stronger predictors of success than many course-level variables. Beck [4] further explores parent perceptions, noting that while families value the safety and flexibility of the online environment, they often feel under-equipped to provide the necessary academic and motivational support.

2.8 Engagement and Interaction Patterns

Student engagement is widely cited as the primary predictor of retention and academic success in online environments. Curtis and Werth [9] identify specific strategies to foster engagement, noting that “transactional distance” can lead to feelings of isolation. They argue that successful online schools must intentionally design for both learner-content and learner-teacher interaction. Kumi-Yeboah et al. [11] similarly found that for minority high school students, interaction with instructors was a key factor in promoting a positive academic self-concept.

However, the literature also reveals a methodological gap in *how* engagement is measured. Most existing studies rely on self-reported surveys or coarse outcome metrics (e.g., course completion). There is limited research that utilizes fine-grained log data to understand the *temporal patterns* of engagement—what this thesis defines as “virtual choreographies.” This gap underscores the need for the data-driven approach proposed in this dissertation, which aims to move beyond static indicators of engagement to uncover the dynamic daily routines that characterize successful learning in virtual schools.

2.9 Data management

All selected papers were organised using a combination of reference management and note-taking tools. Bibliographic information for each article (title, authors, year, venue, DOI and URL) was stored in a BibTEX file associated with the Overleaf project, ensuring consistency between the reference list and the citations used in this dissertation. In addition, the PDFs and summaries of the papers were uploaded to a NotebookLM notebook dedicated to the dissertation. For each paper, short notes were created including:

- a brief summary of the research questions, methods and main findings;
- tags indicating the main focus (e.g., effectiveness/outcomes, teaching practices, student engagement, policy and school choice);
- information about the context (e.g., U.S. state or region, subject area, grade levels).

This structure transforms the selected articles into a manageable “stack of documents” that can be systematically analysed in the next stages of the project and will support the development of the State of the Art chapter.

2.10 Limitations and threats to validity

The review process is subject to several limitations. First, although Google Scholar provides broad coverage, it does not index all relevant education databases, so some studies may have been missed. Second, the search was limited to publications in English, which may exclude research on K–12 online learning conducted in other languages and contexts. Third, the focus on K–12 virtual schools in the United States means that results from other countries were not systematically considered, even when they appeared in the search results. Finally, despite the use of explicit inclusion and exclusion criteria, subjective judgement was involved when assessing the relevance of titles, abstracts and full texts. These limitations will be taken into account when interpreting the findings of the selected studies in the subsequent State of the Art chapter.

Chapter 3

State of the Art

3.1 Overview

While the previous chapter focused on the educational context of K–12 online learning and the needs of at-risk populations, this chapter reviews the technological landscape. Specifically, it addresses the field of Educational Data Mining (EDM) and the computational methods available for discovering patterns in student behavior. The chapter begins by defining the scope of EDM and distinguishing it from Learning Analytics. It then contrasts traditional approaches, such as Educational Process Mining, with the proposed framework of *Virtual Choreographies*. Finally, it reviews the specific unsupervised machine learning algorithms (clustering and dimensionality reduction) that will be used to implement this framework in the Connections Academy dataset.

3.2 Educational Data Mining

Educational Data Mining (EDM) is defined as the area of scientific inquiry centered on the development of methods for making discoveries within the unique kinds of data that come from educational settings [14]. In their comprehensive survey, Romero and Ventura [14] distinguish EDM from Learning Analytics (LA) based on their primary focus:

- **Learning Analytics (LA):** Often focuses on human-led decision making (e.g., visual dashboards for teachers) and relies on statistics.
- **Educational Data Mining (EDM):** Emphasizes automated discovery and the development of algorithms to find hidden patterns without human intervention.

In the context of K–12 online schools, EDM is typically applied to three main tasks [14]:

1. **Prediction:** Using historical data (grades, login frequency) to forecast student outcomes or dropout risk.
2. **Structure Discovery:** Finding underlying structures in data, such as grouping students with similar learning strategies.

3. **Relationship Mining:** Identifying relationships between variables, such as which specific activities lead to higher engagement.

This dissertation focuses primarily on *Structure Discovery*, using unsupervised learning to find patterns that are not immediately obvious to teachers.

3.3 Modeling Student Behavior

To analyze student engagement beyond simple metrics (e.g., “total time online” or “number of clicks”), researchers use advanced techniques to model the *sequence* of student actions.

3.3.1 Educational Process Mining (EPM)

A dominant approach in the literature is Educational Process Mining (EPM). As detailed in the survey by Bogarín et al. [6], EPM applies process mining techniques—originally designed for business workflow analysis—to educational data. EPM treats the learning process as a series of events (e.g., Quiz Start \rightarrow Quiz End) and attempts to discover the “process model” or flow that students follow.

While effective for structured tasks (like a specific exam), Bogarín et al. [6] note that EPM can face challenges in more flexible open-ended learning environments. In full-time virtual schools where students have autonomy over their schedule, EPM often results in “spaghetti models”—complex, tangled diagrams that are difficult to interpret because students do not follow a strict linear path.

3.3.2 The Virtual Choreographies Framework

To address the limitations of rigid process maps, this dissertation adopts the concept of **Virtual Choreographies**. This framework was developed within INESC TEC to identify behavioral patterns in complex environments. Cassola et al. [7] define a *Virtual Choreography* as a platform-independent representation of actions, interactions, and events that unfold over time.

Although originally applied to domains such as Virtual Reality [8] and energy consumption behavior [7], the core mathematical concept is highly applicable to education. Unlike raw clickstreams, a Virtual Choreography maps technical events into **semantic actions** (e.g., mapping URL:/math/quiz/1 to Assessment). This abstraction allows for:

- **Platform Independence:** The analysis focuses on the *behavior* (studying), not the *system* (clicking a specific button ID).
- **Routine Discovery:** By analyzing these choreographies, we can identify clusters of students who share similar daily routines (e.g., “Late Night Crammers” vs. “Steady Workers”), regardless of the specific course they are taking.

3.4 Unsupervised Learning Techniques

The discovery of these choreographies relies on unsupervised machine learning algorithms to group similar behaviors without pre-defined labels.

3.4.1 Clustering Algorithms

Clustering is the task of grouping a set of objects so that objects in the same group (cluster) are more similar to each other than to those in other groups. In EDM, the most widely used algorithm is **k-Means**, due to its computational efficiency and interpretability [14]. However, k-Means requires the number of clusters (k) to be specified in advance, often necessitating the use of validation metrics (such as the Elbow Method or Silhouette Score) to determine the optimal number of student profiles.

3.4.2 Dimensionality Reduction

Student behavioral data is often high-dimensional (e.g., counts of 50 different action types over 30 weeks). To visualize these clusters effectively, dimensionality reduction techniques are required. **t-SNE** (t-Distributed Stochastic Neighbor Embedding) and **UMAP** (Uniform Manifold Approximation and Projection) are state-of-the-art non-linear techniques used to project high-dimensional data into 2D or 3D space. These visualizations allow researchers to “see” the separation between different student groups and validate whether the identified choreographies represent distinct behavioral patterns.

3.5 Summary

Current research in K–12 online learning (Chapter 2) highlights the need for better support systems for at-risk students. Technologically, while Educational Process Mining offers tools to analyze sequences, it can be overly rigid for the flexible nature of online schooling. By adapting the *Virtual Choreographies* framework [7] and combining it with standard EDM clustering techniques [14], this thesis aims to discover interpretable patterns of student behavior that can inform the educational strategies discussed in the literature review.

Chapter 4

Proposed Approach

4.1 Overview

Building upon the research context defined in the Thesis Manifest (Chapter 1) and the technical state of the art (Chapter 3), this chapter details the methodological approach adopted to identify and analyze Virtual Choreographies in K–12 online schools.

The proposed solution is a data-driven pipeline designed to transform raw interaction logs from the Connections Academy Learning Management System (LMS) into interpretable behavioral patterns. The approach combines the semantic abstraction of the *Virtual Choreographies* framework [7] with the unsupervised machine learning techniques reviewed in Chapter 3.

The ultimate goal is to move from low-level clickstream data to high-level educational insights that allow us to distinguish successful learning routines from those that signal disengagement, with a particular focus on the “at-risk” populations identified in the literature [4].

4.2 Methodological Architecture

The research will follow a four-stage pipeline: (1) Data Collection and Preprocessing, (2) Semantic Abstraction, (3) Pattern Discovery (Clustering), and (4) Evaluation and Association.

4.2.1 Stage 1: Data Collection and Context

The dataset for this dissertation is provided by Pearson Online and Blended Learning, encompassing activity logs from 32 Connections Academy schools across the United States.

- **Population:** The study focuses on K–12 students. To address the concerns raised by Beck [3], the data processing will preserve metadata that allows us to distinguish between general enrollment and students identified as “at-risk” (due to prior academic failure or medical/social challenges).
- **Data Granularity:** The raw data consists of server-level logs capturing timestamped events (e.g., login times, page accesses, quiz submissions, live lesson attendance).

- **Privacy:** All data is pseudonymized to ensure student privacy. Personal identifiers are replaced with unique hash keys before the analysis begins.

4.2.2 Stage 2: Semantic Abstraction (The Virtual Choreography)

A core challenge identified in Section 3.3 is the “spaghetti model” problem inherent in analyzing flexible online learning environments [6]. To overcome this, we apply the *Virtual Choreography* abstraction layer.

We will map raw URLs and system events into a finite set of **Semantic Actions**. As emphasized in the literature review regarding student isolation [9], this mapping will prioritize not only content consumption but also **social and interactive behaviors**.

Table 4.1 illustrates proposed examples of this mapping strategy:

Raw Log Pattern (Example)	Semantic Action	Category
URL:/math/algebra/quiz/start	Assessment_Start	Evaluation
URL:/live-lesson/room-101	Synchronous_Participation	Social/Instruction
URL:/mail/compose/teacher	Teacher_Interaction	Social
URL:/feedback/view	Feedback_Review	Metacognition
URL:/content/pdf/view	Content_Study	Asynchronous

Table 4.1: Examples of mapping raw logs to Semantic Actions.

This abstraction creates a sequence of events $S = \{a_1, a_2, \dots, a_n\}$ for each student per day, which constitutes the basis of a daily choreography.

4.2.3 Stage 3: Pattern Discovery

To answer **RQ1 (Identification)**, we will employ unsupervised learning to group similar daily sequences.

1. **Vectorization:** Student daily activities will be converted into numerical vectors (e.g., using n-gram frequency or sequence embeddings) to represent the “rhythm” of their day.
2. **Clustering:** As discussed in Section 3.4, we will apply the **k-Means** algorithm to identify distinct clusters of behavior. The optimal number of clusters (k) will be determined using the Elbow Method and Silhouette Analysis.
3. **Visualization:** To inspect the robustness of these clusters (**RQ2**), we will use **UMAP** (Uniform Manifold Approximation and Projection) to project the high-dimensional data into 2D space. This will allow for visual inspection of the separation between behavioral groups, ensuring the identified choreographies are distinct and interpretable.

4.3 Experimental Process and Evaluation Design

The evaluation is designed to directly address the Research Questions and test the Hypotheses (H1–H5) defined in the Thesis Manifest.

4.3.1 Stability and Robustness (Addressing RQ1 & RQ2)

Once clusters are identified, we will validate them by:

- **Temporal Stability:** Analyzing if students remain in the same behavioral cluster over weeks or if they transition between “types” of learners (e.g., does a student transition from a “Steady Worker” to a “Fragmented” profile?).
- **Cross-Context Validation:** Checking if the same choreographies emerge across different grades (e.g., 5th grade vs. 10th grade) to ensure the patterns are universal behavioral habits rather than course-specific artifacts.

4.3.2 Association with Outcomes (Addressing RQ3)

To validate H1 (Regularity) and H2 (Balanced Participation), we will perform statistical correlation analysis between the identified choreographies and student success metrics.

Given the specific needs of the at-risk population discussed in Section 2.7, “Success” will be defined by two complementary metrics:

1. **Academic Performance:** Final course grades and standardized test scores.
2. **Retention:** The student’s ability to maintain enrollment and avoid dropout. As noted by Beck [4], for many at-risk students, retention is a critical indicator of educational success.

We will use regression models to quantify the impact of adopting a specific choreography on these outcomes, controlling for demographic background.

4.3.3 Predictive Utility (Addressing RQ4)

We will train supervised classification models (specifically Random Forests, due to their interpretability) to predict student success at different snapshots of the semester (e.g., Week 4, Week 8). We will compare two conditions:

1. **Baseline:** Prediction using only standard metrics (login count, total time online).
2. **Experimental:** Prediction using standard metrics + **Choreography Cluster ID**.

This will determine if the “shape” of behavior adds predictive power beyond the simple volume of activity.

4.4 Expected Results

We anticipate identifying 4–6 distinct behavioral choreographies. Based on the hypotheses presented in Chapter 1:

- We expect to confirm that **Regularity** (H1) is a stronger predictor of success than total time spent.
- We anticipate uncovering a “risk” profile characterized by high asynchronous activity but low synchronous interaction (validating H5), supporting the concerns about isolation raised by Curtis and Werth [9].
- The results will provide a new layer of analytics for Connections Academy, enabling teachers to intervene based on *how* a student is learning, not just *if* they are logging in.

4.5 Work Plan

The following timeline details the schedule for the completion of the dissertation, ensuring all deliverables are met by the final deadline in June/July 2026.

Month	Activity
February	Data Preprocessing & Abstraction: Cleaning the Connections Academy dataset; implementing the mapping from URLs to Semantic Actions; separating cohorts.
March	Clustering & Discovery (RQ1, RQ2): Running k-Means and UMAP; refining the number of clusters; characterizing the identified choreographies.
April	Association Analysis (RQ3): Statistical analysis linking clusters to Grades and Retention; testing Hypotheses H1–H3.
May	Predictive Modeling (RQ4) & Visualization (RQ5): Training Random Forest models; designing visualizations for the findings.
June	Writing & Revision: Finalizing the “Results” and “Discussion” chapters; integrating all chapters; final review with supervisors.
July	Submission & Defense: Final submission of the dissertation document and preparation for the public defense.

Table 4.2: Work plan for the Spring 2026 semester.

Bibliography

- [1] Michael K. Barbour. Virtual education: Not yet ready for prime time? In William J. Mathis and Tina Trujillo, editors, *The Test-Based Education Reforms: Lessons from a Failed Agenda*, pages 407–429. Information Age Publishing, Charlotte, NC, 2016.
- [2] Michael K. Barbour and C. B. Hodges. Preparing teachers for effective k-12 online learning in the age of disruptions. *Open Praxis*, 16(4):583–594, 2024. doi: 10.55982/openpraxis.16.4.678.
- [3] Dennis Beck. Be prepared: Online school experience and student achievement during the pandemic. *Frontiers in Education*, 8:1161003, 2023. doi: 10.3389/feduc.2023.1161003.
- [4] Dennis Beck. At-risk and online: Parent perceptions of at-risk learner’s supports in a fully online school. *Journal of Research on Technology in Education*, 2024. doi: 10.1080/15391523.2023.2285497. (Ahead-of-print).
- [5] Dennis Beck and J. Levine. The role of the advocate in cyber schools during the covid-19 pandemic. *Journal of Online Learning Research*, 6(3):195–217, 2020. URL <https://www.learntechlib.org/primary/p/217172/>.
- [6] Alejandro Bogarín, Rebeca Cerezo, and Cristóbal Romero. A survey on educational process mining. *WIREs Data Mining and Knowledge Discovery*, 8(1):e1230, 2018. doi: 10.1002/widm.1230.
- [7] Fernando Cassola, Leonel Morgado, et al. Using virtual choreographies to identify office users’ behaviors to target behavior change. *Energies*, 15(12):4354, 2022. doi: 10.3390/en15124354.
- [8] Fernando Cassola et al. Design and evaluation of a choreography-based virtual reality authoring tool for experiential learning. *IEEE Transactions on Learning Technologies*, 15(5): 526–539, 2022. doi: 10.1109/TLT.2022.3206307.
- [9] Heidi Curtis and Loredana Werth. Fostering student success and engagement in a k-12 online school. *Journal of Online Learning Research*, 1(2):163–190, 2015. URL <https://www.learntechlib.org/primary/p/151187/>.

- [10] Carla C. Johnson, Janet B. Walton, Lacey Strickler, and Jennifer B. Elliott. Online teaching in k-12 education in the united states: A systematic review. *Review of Educational Research*, 93(3):353–411, 2022. doi: 10.3102/00346543221105550.
- [11] Alex Kumi-Yeboah, J. Dogbey, and Guangji Yuan. Exploring factors that promote online learning experiences and academic self-concept of minority high school students. *Journal of Research on Technology in Education*, 50(1):1–17, 2018. doi: 10.1080/15391523.2017.1365669.
- [12] Florence Martin, T. Sun, and C. D. Westine. A systematic review of research on k-12 online teaching and learning: Comparison of research from two decades 2000 to 2019. *Journal of Research on Technology in Education*, 52(3):353–374, 2020. doi: 10.1080/15391523.2019.1605825.
- [13] Alex Molnar, Gary Miron, Michael K. Barbour, Luis Huerta, Steven R. Shafer, Jennifer King Rice, Angela Glover, Heather Knight, and Jennifer Key. *Virtual Schools in the U.S. 2023*. National Education Policy Center, Boulder, CO, 2023. URL <https://nepc.colorado.edu/publication/virtual-schools-annual-2023>.
- [14] Cristobal Romero and Sebastian Ventura. Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3):e1355, 2020. doi: 10.1002/widm.1355.
- [15] Ian N. Toppin and Sheila M. Toppin. Virtual schools: The changing landscape of k-12 education in the u.s. *Education and Information Technologies*, 21(6):1571–1581, 2016. doi: 10.1007/s10639-015-9402-8.