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


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 e *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 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** **4**
- 2.1 Overview . . . 4
- 2.2 Type of review . . . 4
- 2.3 Search strategy . . . 5
- 2.4 Inclusion and exclusion criteria . . . 5
  - 2.4.1 Inclusion criteria . . . 6
  - 2.4.2 Exclusion criteria . . . 6
- 2.5 Screening and selection process . . . 6
  - 2.5.1 Stage 1: Title and abstract screening . . . 6
  - 2.5.2 Stage 2: Full-text assessment . . . 7
  - 2.5.3 Stage 3: Citation-based refinement . . . 7
- 2.6 Summary of Selected Studies . . . 7
- 2.7 Effectiveness of K-12 Online Learning . . . 7
- 2.8 Supporting At-Risk Populations . . . 7
- 2.9 Engagement and Interaction Patterns . . . 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 . . . 10
  - 3.4.1 Clustering Algorithms . . . 10
  - 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
4.3.4 Actionability (Addressing RQ5) . . . . . 15
4.4 Expected Results . . . . . 15
4.5 Work Plan . . . . . 15

# List of Tables

2.1 Summary of the core literature corpus. . . . . 7

4.1 Examples of mapping raw logs to Semantic Actions. . . . . 13
4.2 Work plan for the Spring 2026 semester. . . . . 16

# List of Acronyms

**AI** Artificial Intelligence

**ARI** Adjusted Rand Index

**EDM** Educational Data Mining

**EPM** Educational Process Mining

**FEUP** Faculdade de Engenharia da Universidade do Porto

**IEP** Individualized Education Program

**K-12** Kindergarten to 12th Grade

**LA** Learning Analytics

**LMS** Learning Management System

**ML** Machine Learning

**NEPC** National Education Policy Center

**RQ** Research Question

**SHAP** SHapley Additive exPlanations

**t-SNE** t-Distributed Stochastic Neighbor Embedding

**UMAP** Uniform Manifold Approximation and Projection

**URL** Uniform Resource Locator

# 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 [10]. 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 [2]. 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 [8]. 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. [6].

## 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 [6].

**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.

**H_Robustness — Context Stability (Addressing RQ2).** We hypothesize that choreographies associated with highly structured subjects (e.g., Mathematics) will exhibit higher stability across cohorts than those in open-ended or creative subjects, and that core behavioral habits will remain consistent across adjacent grade levels.

**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. [9] and Johnson et al. [8]. 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 [2, 3] and Toppin and Toppin [12], 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. While discipline-specific databases like ERIC or Scopus are valuable, Google Scholar was selected as the primary source for its broader indexation of interdisciplinary research, capturing relevant studies at the intersection of Computer Science (Educational Data Mining) and Education that might otherwise be fragmented across distinct databases.

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, and surface key works. 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, teaching practices, student engagement, or system-level issues such as school choice and policy.

### 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.
- **EC2** – The publication type is a thesis, dissertation, report, blog post, 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, approximately 150 unique records remained. Titles and abstracts were screened against the criteria. Studies clearly focusing on higher education or non-online contexts were removed, leaving approximately 60 papers.

### 2.5.2 Stage 2: Full-text assessment

The full text of the remaining studies was examined. Papers that mentioned online learning only superficially were excluded. This stage reduced the set to a core group of studies providing substantial evidence, including recent systematic reviews [8, 9] and research on at-risk students [3, 4]. Around 20 papers remained.

### 2.5.3 Stage 3: Citation-based refinement

The core set was used for backward and forward snowballing using Litmaps. After applying the same criteria to these additional articles, the final corpus consisted of **12 highly relevant papers**.

## 2.6 Summary of Selected Studies

To provide a clear overview of the state-of-the-art, Table 2.1 summarizes the key papers selected for the final corpus.

| Author (Year) | Type | Focus | Key Insight |
|---|---|---|---|
| Martin et al. (2020) | Review | Effectiveness | Effectiveness depends on design/support, not just the medium. |
| Johnson et al. (2022) | Review | Teaching Practices | Importance of teacher facilitation in online settings. |
| Beck (2023, 2024) | Empirical | At-Risk Students | At-risk students perform better with strong "advocate" support. |
| Curtis & Werth (2015) | Empirical | Engagement | Transactional distance leads to isolation; needs interaction. |
| Molnar et al. (2023) | Report | Policy | Virtual schools often lag in graduation rates compared to traditional. |
| Toppin & Toppin (2016) | Analysis | Demographics | Virtual schools attract students with prior academic/social issues. |

Table 2.1: Summary of the core literature corpus.

## 2.7 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 reports consistently highlight significant performance gaps [10], noting lower graduation rates. However, recent reviews [8, 9] caution against binary comparisons, suggesting effectiveness is dependent on instructional design and support. Johnson et al. [8] note that while average performance may be lower, online learning provides essential opportunities for credit recovery and advanced coursework.

## 2.8 Supporting At-Risk Populations

A critical finding is that the demographic profile of virtual schools differs significantly from traditional ones. Toppin and Toppin [12] and Beck [2] observe that virtual schools frequently attract

"at-risk" students—those with prior bullying, medical issues, or academic failure. Success for these learners relies heavily on support beyond the screen. Beck's research [3, 4] highlights the pivotal role of the "on-site facilitator" (typically a parent). Beck and Levine [4] found that parental engagement was a stronger predictor of success than many course-level variables during the pandemic.

## 2.9 Engagement and Interaction Patterns

Student engagement is cited as the primary predictor of retention. Curtis and Werth [7] identify strategies to foster engagement, noting that "transactional distance" leads to isolation. However, a methodological gap exists: most studies rely on surveys. There is limited research utilizing fine-grained log data to understand *temporal patterns* of engagement. This gap underscores the need for the data-driven approach proposed in this dissertation.

## 2.10 Limitations and threats to validity

The review is subject to limitations. First, Google Scholar does not index all databases, potentially missing some studies. Second, the search was limited to English publications. Third, the focus on U.S. virtual schools excludes international contexts. Finally, subjective judgement was involved in screening. These limitations will be considered in the State of the Art analysis.

# 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 defines the scope of EDM, contrasts traditional Educational Process Mining with the proposed *Virtual Choreographies* framework, and reviews the unsupervised machine learning algorithms selected for this study.

## 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 [11]. In their comprehensive survey, Romero and Ventura [11] distinguish EDM from Learning Analytics (LA):

- **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 [11]:

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.

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

## 3.3 Modeling Student Behavior

To analyze student engagement beyond simple metrics, researchers use techniques to model the *sequence* of student actions.

### 3.3.1 Educational Process Mining (EPM)

A dominant approach is Educational Process Mining (EPM). As detailed by Bogarín et al. [5], EPM applies process mining techniques to educational data, treating learning as a series of events (e.g., `Quiz Start` → `Quiz End`). While effective for structured tasks, Bogarín et al. [5] note that EPM faces challenges in flexible environments. In full-time virtual schools, EPM often results in "spaghetti models"—complex, tangled diagrams that are difficult to interpret. **This lack of readability makes such models unsuitable for teachers who need clear, actionable insights (addressing RQ5), necessitating a higher-level abstraction like Virtual Choreographies.**

### 3.3.2 The Virtual Choreographies Framework

To address the limitations of rigid process maps, this dissertation adopts the concept of **Virtual Choreographies**. Cassola et al. [6] define a *Virtual Choreography* as a platform-independent representation of actions, interactions, and events that unfold over time. 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, not the system.
- **Routine Discovery:** It enables the identification of clusters of students who share similar daily routines (e.g., "Late Night Crammers" vs. "Steady Workers") regardless of the specific course.

## 3.4 Unsupervised Learning Techniques

The discovery of these choreographies relies on unsupervised machine learning algorithms.

### 3.4.1 Clustering Algorithms

Clustering groups objects so that objects in the same group 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 [11]. However, k-Means requires the number of clusters ($k$) to be specified, necessitating the use of validation metrics such as the Elbow Method or Silhouette Score.

### 3.4.2 Dimensionality Reduction

Student behavioral data is often high-dimensional. To visualize clusters effectively, dimensionality reduction is required. **t-SNE** and **UMAP** are state-of-the-art non-linear techniques used to project high-dimensional data into 2D 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 highlights the need for better support systems for at-risk students. Technologically, while EPM offers tools to analyze sequences, it can be overly rigid for online schooling. By adapting the *Virtual Choreographies* framework [6] and combining it with standard EDM clustering techniques [11], 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 [6] 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 [3].

## 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 [2], 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, special education status, 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 [5]. 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 [7], this mapping will prioritize not only content consumption but also **social and interactive behaviors**.

To ensure the reliability of this abstraction, we will conduct a **rule-based mapping audit**. A random sample of raw events will be manually reviewed against the generated semantic labels to verify that the mapping logic captures the educational intent correctly (e.g., ensuring a "Discussion Board" click is classified as interaction, not just passive viewing).

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 $S = \{a_1, a_2, ..., a_n\}$ 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 to represent the "rhythm" of their day. Specifically, we will utilize **action frequency vectors** combined with **time-of-day bins** (e.g., Morning, Afternoon, Evening) to capture not just *what* students do, but *when* they do it.

2. **Targeted Segmentation (At-Risk Analysis):** While we will perform clustering on the general population, a key methodological step is the **segmented analysis of at-risk students**. We will run the clustering algorithms specifically within the sub-population of at-risk students. This ensures that subtle behavioral patterns unique to this group are not "washed out" by the dominant behaviors of the general student body.

3. **Clustering:** As discussed in Section 3.4, we will apply the **k-Means** algorithm. The optimal number of clusters ($k$) will be determined using the Elbow Method and Silhouette Analysis.

4. **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.

## 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 (RQ1):** We will measure the consistency of student cluster membership over time using the **Adjusted Rand Index (ARI)** between weeks. This will determine if students possess a stable "learning type" or if their behavior is volatile.

- **Cross-Context Validation (RQ2):** We will test the hypothesis that choreographies are more stable in highly structured subjects (e.g., Mathematics) compared to open-ended ones (e.g., Creative Arts). We will verify if the same patterns emerge across different grades (5th vs. 10th grade) to ensure they are universal habits.

### 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 [3], "Success" will be defined by:

1. **Academic Performance:** Final course grades and standardized test scores.

2. **Retention:** The student's ability to maintain enrollment (non-dropout).

We will use regression models to quantify the impact of adopting a specific choreography on these outcomes. Crucially, we will control for student background using available variables such as **Free/Reduced Lunch status** (socioeconomic proxy) and **Special Education (IEP) status**.

### 4.3.3 Predictive Utility (Addressing RQ4)

We will train supervised classification models to predict student success. We selected **Random Forests** as the primary model due to their ability to provide feature importance (interpretability), avoiding the "black box" nature of complex neural networks.

**Validation Strategy:** To prevent data leakage, we will use a **time-based split**: training the models on data from the first half of the semester and testing on the second half (or using a previous cohort year for training and the current year for testing).

We will compare two conditions:

1. **Baseline:** Prediction using only standard institutional metrics (login count, total time online). These are the current "state-of-the-art" metrics available to teachers in most LMS dashboards.

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.

### 4.3.4 Actionability (Addressing RQ5)

To determine which components drive success, we will analyze the **Feature Importance** (e.g., Gini importance or SHAP values) derived from the Random Forest models. This will allow us to tell teachers specifically which transitions (e.g., "Reviewing Feedback → Resubmitting") are the strongest signals of success.

## 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 [7].

- 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.

| Month | Activity |
|---|---|
| **February** | **Data Preprocessing & Abstraction:** Cleaning the Connections Academy dataset; implementing rule-based mapping audit; separating "at-risk" vs. general cohorts. |
| **March** | **Clustering & Discovery (RQ1, RQ2):** Running k-Means on both general and at-risk populations; refining the number of clusters; characterizing the identified choreographies. |
| **April** | **Association Analysis (RQ3):** Statistical analysis linking clusters to Grades and Retention (controlling for IEP/Lunch status). |
| **May** | **Predictive Modeling (RQ4) & Visualization (RQ5):** Training Random Forest models with time-based splitting; designing visualizations. |
| **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] 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.

[3] 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).

[4] 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/.

[5] 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.

[6] 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.

[7] 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/.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.