

CREATION AND IMPLEMENTATION OF THE INNOVATION-BASED LEARNING  
FRAMEWORK: A LEARNING ANALYTICS APPROACH

A Dissertation  
Submitted to the Graduate Faculty  
of the  
North Dakota State University  
of Agriculture and Applied Science

By

Lauren Nichole Singelmann

In Partial Fulfillment of the Requirements  
for the Degree of  
DOCTOR OF PHILOSOPHY

Major Department:  
Electrical and Computer Engineering and STEM Education

April 2022

Fargo, North Dakota

# NORTH DAKOTA STATE UNIVERSITY

Graduate School

---

## Title

CREATION AND IMPLEMENTATION OF THE INNOVATION-BASED  
LEARNING FRAMEWORK: A LEARNING ANALYTICS APPROACH

---

## By

Lauren Nichole Singelmann

---

The supervisory committee certifies that this dissertation complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

## SUPERVISORY COMMITTEE:

Daniel Ewert

Chair

---

Benjamin Braaten

---

Warren Christensen

---

Jennifer Momsen

---

Scott Pryor

---

Approved:

April 25, 2022

Date

Benjamin Braaten

Department Chair

## ABSTRACT

To meet the national and international call for creative and innovative engineers, many engineering departments and classrooms are striving to create more authentic learning spaces where students are actively engaging with design and innovation activities. For example, one model for teaching innovation is Innovation-Based Learning (IBL) where students learn fundamental engineering concepts and apply them to an innovation project with the goal of producing value outside the classroom. The model has been fairly successful, but questions still remain about how to best support students and instructors in open-ended innovation spaces.

To answer these questions, learning analytics and educational data mining (LA/EDM) techniques were used to better understand student innovation in IBL settings. LA/EDM is a growing field with the goal of collecting and interpreting large amounts of educational data to support student learning. In this work, five LA/EDM algorithms and tools were developed: 1) the IBL framework which groups student actions into illustrative categories specific to innovation environments, 2) a classifier model that automatically groups student text into the categories of the framework, 3) classifier models that leverage the IBL framework to predict student success, 4) clustering models that group students with similar behavior, and 5) epistemic network analysis models that summarize temporal student behavior. For each of the five algorithms/tools, the design, development, assessment, and resulting implications are presented.

Together, the results paint a picture of the affordances and challenges of teaching and learning innovation. The main insights gained are how language and temporal behavior provide meaningful information about students' learning and innovation processes, the unique challenges that result from incorporating open-ended innovation into the classroom, and the impact of using LA/EDM tools to overcome these challenges.

## ACKNOWLEDGEMENTS

To my committee – Thank you all for making my graduate school journey such a positive one. I feel so incredibly grateful to have a committee that pushes me to be my best and supports me in so many ways. I love this work, and I know that is a direct product of having a team who shows so much curiosity and care. To Dr. Dan Ewert – You have believed in me since the very beginning of my engineering journey, even when others didn't, and even when I couldn't believe in myself. I am so grateful for our many years working together. To Dr. Ben Braaten – Your support of my teaching and research goals and your belief in my abilities has meant so much to me. Your leadership and support are greatly appreciated. To Dr. Warren Christensen – Your teaching and mentorship was integral in helping me identify my discipline-based education research pathway, and I greatly value the insights, advice, and support that you bring. To Dr. Jenni Momsen – You are always pushing me to think about things in new ways and from new angles. Our conversations have helped me grow into a better educator, researcher, and human. To Dr. Scott Pryor – Your trust in me has meant so much. Your leadership has helped make my graduate school experience one that is positive and fulfilling in so many different ways.

To my lab mates – Thank you Enrique Alvarez Vazquez, Ellen Swartz, Mary Pearson, Stan Ng, and Ryan Striker. Your feedback and collaboration has been a huge help as I have been working on this dissertation. I'd especially like to thank Enrique for all of his work on MOOCIBL. Without you, this work would not have been possible.

To the NDSU STEM Education Journal Club – Thank you for letting an engineer stick around for the past few years. I always look forward to our time together on Friday afternoons because I love to learn and connect with you all.

To the North Dakota State University Department of Electrical and Computer Engineering and the College of Engineering – Thank you for the many ways you have supported me: administratively, professionally, and financially.

To my family and friends – Your love and belief in me has meant so much.

To Ben – I am so incredibly lucky to have someone who loves and supports me like you do.

# DEDICATION

To learners.

## PREFACE

“The innovation journey: you can’t control it, but you can learn to maneuver it.”

- Andrew Van de Ven

I came across this paper by Van de Ven well after I started the work presented in this dissertation. It led to a bit of a research identity crisis; isn’t ‘control’ usually a key part of the research process? How are you supposed to measure, compare, and analyze innovation if the process is so unpredictable? These thoughts led to some feelings of doubt about this work; I started to once again worry about “technical rigor” and “soundness of methods”. However, I refused to let these worries convince me that this project wasn’t one worth pursuing. In fact, these questions and challenges came with some added motivation, too. The uncertainty is what makes the project exciting, and I continue to see arguments that complexity is the future of science and engineering. In other words, we will need to find new ways to analyze and interpret highly inter-connected systems such as weather, ecosystems, power grids, and – as seen in this work – classrooms.

However, embracing complexity was in no way easy. Along the way, I knew that we were working within a complex system, but it is so easy to fall back into reductionist ways of thinking. As I was writing the implications and conclusions, for example, I desperately wanted to be able to deliver a set of “rules” for Innovation-Based Learning; make a claim, back it up with evidence, and then state the single implication. However, I eventually recognized that this would not only be a misrepresentation of the findings, but also a disservice to the instructors and students both past and future. Instead, I aimed to use my data and analysis to illustrate the challenges and complexities of teaching, learning, and innovation. Does this mean I could throw out “technical rigor” and “soundness of methods”? No – but it did require me to think differently about how we interpret and share our evidence. Sharing this work in this way has led to nuanced conversations about teaching, learning, and innovation, and these conversations continue to help me grow as a researcher and an educator in ways I didn’t realize were possible. I am thankful for the many people that have engaged in these discussions and believed in this work. We may not be able to control the path ahead, but we are maneuvering it together.

# TABLE OF CONTENTS

ABSTRACT . . . . .	iii
ACKNOWLEDGEMENTS . . . . .	iv
DEDICATION . . . . .	v
PREFACE . . . . .	vi
LIST OF TABLES . . . . .	xi
LIST OF FIGURES . . . . .	xii
LIST OF ABBREVIATIONS . . . . .	xxi
LIST OF APPENDIX TABLES . . . . .	xxii
LIST OF APPENDIX FIGURES . . . . .	xxiii
1. INTRODUCTION . . . . .	1
1.1. The Call for Innovative Engineers . . . . .	1
1.2. Innovation-Based Learning to Answer the Call . . . . .	1
1.3. Dataset . . . . .	2
1.3.1. MOOCIBL . . . . .	2
1.3.2. Cohorts . . . . .	3
1.3.3. Participants . . . . .	4
1.4. Learning Analytics and Educational Data Mining . . . . .	4
1.5. Research Questions . . . . .	6
1.6. Overview of Chapters . . . . .	7
2. LITERATURE REVIEW . . . . .	8
2.1. Introduction . . . . .	8
2.2. Innovation as Simple . . . . .	9
2.2.1. In Industry . . . . .	10
2.2.2. In Education . . . . .	10
2.2.3. In Industry . . . . .	12

2.2.4.	In Education . . . . .	13
2.3.	Innovation as Complex . . . . .	15
2.3.1.	In Industry . . . . .	15
2.3.2.	In Education . . . . .	16
2.4.	Innovation as Chaotic . . . . .	20
2.4.1.	In Industry . . . . .	20
2.4.2.	In Education . . . . .	22
2.5.	Discussion and Relevance . . . . .	23
2.5.1.	Choosing a Domain Lens . . . . .	24
2.5.2.	Choosing Appropriate Methods . . . . .	25
2.5.3.	Choosing Appropriate Metrics of Success . . . . .	26
2.5.4.	Filling the Gap . . . . .	26
3.	DEVELOPMENT OF THE INNOVATION-BASED LEARNING FRAMEWORK . . . . .	28
3.1.	Introduction . . . . .	28
3.2.	Background . . . . .	29
3.2.1.	Motivation for Framework . . . . .	29
3.2.2.	Other Related Frameworks . . . . .	29
3.2.3.	Leveraging the Combination of a Framework and LA/EDM Tools . . . . .	36
3.3.	Research Question 1A: Development of a Framework . . . . .	36
3.3.1.	Methods . . . . .	36
3.3.2.	Results and Analysis . . . . .	38
3.4.	Research Question 1B: Feasibility of Automatic Framework Classification . . . . .	46
3.4.1.	Methods . . . . .	46
3.4.2.	Results and Analysis . . . . .	48
3.5.	Implications for Teaching . . . . .	50
3.6.	Implications for Research . . . . .	50



3.7. Summary . . . . .	52
4. EXTENSION OF CLASSIFICATION AND CLUSTERING IN IBL BY LEVERAGING THE IBL FRAMEWORK . . . . .	53
4.1. Introduction . . . . .	53
4.2. Previous Learning Analytics/Educational Data Mining Work in IBL . . . . .	53
4.2.1. Previous Classification Models . . . . .	54
4.2.2. Previous Clustering Models . . . . .	55
4.3. Research Question 2A: Using the IBL Framework to Improve Classification . . . . .	55
4.3.1. Methods . . . . .	55
4.3.2. Results and Analysis . . . . .	61
4.4. Research Question 2B: Using the IBL Framework to Improve Clustering . . . . .	72
4.4.1. Methods . . . . .	72
4.4.2. Results and Analysis . . . . .	73
4.5. Implications for Teaching . . . . .	78
4.6. Implications for Research . . . . .	80
4.7. Summary . . . . .	81
5. IMPLEMENTATION OF EPISTEMIC NETWORK ANALYSIS IN IBL THROUGH THE IBL FRAMEWORK . . . . .	82
5.1. Introduction . . . . .	82
5.2. Background . . . . .	82
5.2.1. Introduction to Epistemic Network Analysis . . . . .	82
5.2.2. Uses of Epistemic Network Analysis . . . . .	83
5.2.3. Mathematical Theory of Epistemic Network Analysis . . . . .	85
5.2.4. Affordances and Limitations of Epistemic Network Analysis . . . . .	92
5.3. Research Question 3A: Behavior of Students and Teams in Low-Structure IBL . . . . .	94
5.3.1. Methods . . . . .	94
5.3.2. Results and Analysis . . . . .	97

5.4. Research Question 3B: Effect of Added Structure on Student Behavior in IBL . . . .	105
5.4.1. Methods . . . . .	105
5.4.2. Results and Analysis . . . . .	107
5.5. Implications for Teaching . . . . .	112
5.6. Implications for Research . . . . .	113
5.7. Summary . . . . .	115
6. DISCUSSION . . . . .	116
6.1. Insights . . . . .	116
6.2. Limitations . . . . .	118
6.3. Challenges to Implementation . . . . .	120
6.4. Future Directions . . . . .	121
7. CONCLUSION . . . . .	123
REFERENCES . . . . .	124
APPENDIX. WORKED EXAMPLES . . . . .	136
A.1. A Worked Example of Calculating the ROC AUC Metric . . . . .	136
A.2. A Worked Example of Epistemic Network Analysis . . . . .	138
A.3. A Worked Example of Calculating Network Complexity . . . . .	145

## LIST OF TABLES

<u>Table</u>	<u>Page</u>
3.1. Example token for each of the framework categories . . . . .	41
3.2. Categorization errors . . . . .	43
3.3. Percent agreement and Cohen’s Kappa of team of human raters and classifier models . .	48
4.1. Performance metrics for text classifiers intra-set . . . . .	61
4.2. Probability that accuracy is above baseline at various randomness levels . . . . .	64
4.3. Average performance metrics at various randomness levels . . . . .	66
4.4. Performance metrics for text classifiers inter-set . . . . .	66
4.5. Performance metrics for single category text classifiers intra-set . . . . .	68
4.6. Performance metrics for single category text classifiers inter-set . . . . .	69
4.7. Performance metrics for quantitative classifiers intra-set . . . . .	69
4.8. Analysis of each of the eight clusters as labeled in the dendrogram in Figure 4.14. Cluster descriptions were developed by qualitatively comparing the category breakdowns for all members of that cluster and finding commonalities. . . . .	76

## LIST OF FIGURES

Figure	Page
2.1. Cynefin Framework from [25] . . . . .	9
2.2. The Engineering Design Process as published on TeachEngineering [27] . . . . .	12
2.3. Behavior of $X_t$ over time $t$ for the simple logistic difference equation . . . . .	21
2.4. Phase diagrams of $X_t$ , $X_{t-1}$ , and $X_{t-2}$ for the simple logistic difference equation . . . . .	21
2.5. Cycle of innovation presented in [5]. Divergent behaviors lead to the introduction of constraining factors. These factors require behavior to converge, which then leads to enabling factors. The enabling factors lead to new ideas and options, going back to divergent behavior. . . . .	23
3.1. In order to develop the Innovation-Based Learning framework, five areas of literature were explored. Existing frameworks from each of these areas were applied to the data in order to determine strengths and weaknesses, and these five areas were combined to create the final framework. . . . .	30
3.2. Bloom’s Revised Taxonomy from [62] . . . . .	31
3.3. Framework for complex problem solving from [42] . . . . .	32
3.4. Cyclical framework for self-regulated learning from [70] . . . . .	33
3.5. One example of the engineering design process from [26] . . . . .	33
3.6. The Double Diamond for Engineering Design from [72] . . . . .	35
3.7. The Cycle of Divergent and Convergent Behavior in Innovation from [5] . . . . .	35
3.8. The developed IBL framework integrates engineering innovation and learning in engineering education. The framework consists of three diamonds, each consisting of a diverging activity and a converging activity that produce an output. Diamond 1 consists of <i>surveying</i> and <i>defining</i> the problem to develop a gap. Diamond 2 consists of <i>exploring</i> and <i>solving</i> to create an innovative solution. Diamond 3 consists of <i>drafting</i> and <i>sharing</i> , leading to impact of the innovation. The final category is <i>environment</i> , and it covers any other activities related to being a member of a group or class. . . . .	40
3.9. Confusion matrix between the two independent raters. Boxes on the diagonal represent agreement between the two raters. The color relates to the ratio of tokens sorted into a certain category compared to the total number of tokens in that category (as determined by Rater 1). For example, Rater 1 put 193 tokens into the <i>survey</i> category, and Rater 2 agreed for 148 of those tokens, giving a recall value of 0.767. The matrix also identifies areas of discrepancy; these are the brighter green values that are not on the diagonal. . . . .	42

- 3.10. Confusion matrix between the lead rater and the team of secondary raters. Boxes on the diagonal represent agreement between the two raters. The color relates to the ratio of tokens sorted into a certain category compared to the total number of tokens in that category (as determined by Rater 1). For example, Rater 1 put 517 tokens into the *survey* category, and Rater 2 agreed for 465 of those tokens, giving a recall value of 0.90. The matrix also identifies areas of discrepancy; these are the brighter green values that are not on the diagonal. . . . . 45
- 3.11. Graphical representation of methods for RQ1B. First, preprocessing is performed by the research team by splitting all tokens into the categories of the framework. Next, a TFIDF (term frequency-inverse document frequency) matrix created where each row represents a token, each column represents a word, and each entry represents how frequently a word appears in that token compared to all other tokens. To perform five-fold cross-validation, all tokens are split into five folds with tokens from each category balanced among the folds. For each of the five folds, a model is created and trained using four folds and tested on the fifth fold. The results from each fold are combined to get overall performance metrics. Four different algorithms were assessed: support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF), and logistic regression (LR). Finally, a confusion matrix is created for each of the models to determine where discrepancies occurred, and percent agreement and Cohen’s Kappa are calculated to assess performance. 47
- 3.12. Confusion matrices that show discrepancies between a lead human rater and the team of other human raters (left) and a lead human rater and the optimized classification algorithm (right). Entries are colored by comparing the value of the cell to the sum of the values in that row, which corresponds to precision. Preliminary results for RQ1A showed that human raters had sufficient agreement when categorizing text into the framework categories. These new results show that a classifier can also be trained to categorize objectives with higher reliability, consistency, and speed. In addition, it shows that discrepancies are most common between the two triangles that make up a diamond (e.g. *Explore* and *Solve*.) . . . . . 49

4.1.	Graphical representation of methods for RQ2A. First preprocessing is performed by splitting student text into documents using the framework. Each document represents one students' text for one framework category. Next, four different methods for feature engineering are performed to create feature matrices. For the 'No Framework' model, all documents for each student are combined into a single document. These documents are then used to create a TFIDF matrix where each row represents a student and each column represents a word. For the 'Framework' model, each document is converted into its own TFIDF matrix, and these matrices are then concatenated to create a combined TFIDF matrix where each row represents a student and each column represents a word in a specific framework category. For the 'Category' models, the single category TFIDF matrices are all analyzed individually to determine how well text from each category differentiates high and low performance. For the 'Quantitative' model, the number of tokens in each category is counted, and a feature matrix is created with the relative proportions for each token type. Next, students are split into groups. For the cross-validation assessment, students are split into five folds with cohorts evenly represented in each fold. For the prediction assessment, Cohort 1 and 2 are put into the training group, and Cohort 3 is put into the test group. Classification models are then trained and tested, and the following performance metrics are calculated: accuracy, precision, recall, F1 score, and ROC AUC. . . . .	57
4.2.	Sample confusion matrix defining true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) in the context of this study. Low performance was deemed to be the positive case because future work hopes to identify at-risk students. . . . .	58
4.3.	The confusion matrices for the text intra-set classification models assessed using five-fold cross-validation. Both (a) the model trained without sorting text into the IBL framework and (b) the model trained after sorting text into the IBL framework are able to differentiate between lower and higher performers at a level sufficiently higher than a random classifier. Entries are colored by comparing the value of the cell to the sum of the values in that column. This corresponds to recall (the ratio of students at a performance level that were also predicted to be at that performance level). . . . .	62
4.4.	Receiver operating characteristic curve for the text classifiers trained intra-set (with five-fold cross-validation). (a) shows the model trained without sorting text into the IBL framework (originally designed and tested in [22]), and (b) shows the model trained by sorting text into the IBL framework (the new model). The curve for each fold is shown, along with the mean curve $\pm 1$ standard deviation. As expected from the preliminary work, the original model without the framework maintains its strong performance on the larger dataset. The new model trained with the framework maintains this strong performance. . . . .	62

4.5.	Top 150 features for the optimized text classifier that differentiate between lower-performing and higher-performing students. Words are grouped by the framework category that they fell into. Features in green more highly differentiated higher-performing students, and features in red more highly differentiated lower-performing students. The font size represents the relative chi-2 value; larger words were more highly differentiating than smaller words. Because each word in each category is its own feature, some words appear in more than one category. In fact, some words such as ‘learn[ed/ing]’ and ‘website’ appear in red in some categories and green in others. Specific conferences and universities were replaced with [conference] or [university]. . . . .	63
4.6.	Monte Carlo results for the optimized text classifier (a) without the framework and (b) with the framework. 10,000 samples were run for each of the three randomness levels: 5%, 10%, and 20%. The results are plotted using a histogram with 100 bins ranging from accuracy of 0 to accuracy of 1. The y-axis represents the percentage of samples that fell into any given bin. Vertical lines representing the baseline performance and the performance of the classifier with no randomness are also plotted. . . . .	65
4.7.	The confusion matrices for the text inter-set classification models trained using data from Cohorts 1 and 2 and tested on Cohort 3. (a) shows the confusion matrix for the model trained without sorting text into the IBL framework and (b) shows the model trained after sorting text into the IBL framework. Entries are colored by comparing the value of the cell to the sum of the values in that column. This corresponds to recall (the ratio of students at a performance level that were also predicted to be at that performance level). The new model trained with the framework had higher inter-set performance than the original model in [22] without the framework. From the confusion matrices, it is clear that a fault of the model trained without the framework is its prediction that most of the students are high-performing, whereas the model trained with the framework is more balanced. . . . .	67
4.8.	Receiver operating characteristic curve for the text classifiers trained inter-set (trained with Cohort 1 and 2 data and tested with Cohort 3 data). (a) shows the model trained without sorting text into the IBL framework (originally designed and tested in [22]), and (b) shows the model trained by sorting text into the IBL framework (the new model). The old model has some ability to differentiate between classes, but the new model shows improved performance. . . . .	67
4.9.	The confusion matrices for the quantitative intra-set classification models assessed using five-fold cross-validation. (a) shows the model trained without sorting tokens into the IBL framework before counting and (b) shows the model that sorted tokens into the IBL framework before counting. Both models had similar performance in identifying low performing students, but the model with the framework had much better performance differentiating between the two classes overall. . . . .	70

4.10. Receiver operating characteristic curve for the quantitative classifiers trained intra-set (with five-fold cross-validation). (a) shows the model trained without sorting tokens into the IBL framework before counting (originally designed and tested in [22]), and (b) shows the model trained by sorting tokens into the IBL framework before counting (the new model). The curve for each fold is shown, along with the mean curve  $\pm 1$  standard deviation. As expected from the preliminary work, the original model without the framework performs no better than random at separating low- and high-performing. The new model trained with the framework is better able to differentiate between low- and high-performing. . . . . 70

4.11. The top features for the quantitative classifier by framework category. If a triangle is green and has a positive chi-2 value, higher-performing students were more likely to have that type of token compared to lower-performing students. If a triangle is red and has a negative chi-2 value, lower-performing students were more likely to have that type of token compared to higher-performing students. . . . . 71

4.12. Dendrograms of student clusters based on (a) proportions of tokens in each category and (b) raw counts of tokens in each category colored by cohort. Each of the leaves represent a student and are labeled with a letter (L or H) to represent low or high performing and a number (1, 2, or 3) to represent cohort number. The intersection of two branches represents the distance between them. From this, we see that raw token counts cause students to cluster mostly by cohort, whereas student proportions allows us to better compare students across cohorts. . . . . 75

4.13. Dendrogram of student clusters based on proportions of tokens in each category colored by performance. Each of the leaves represent a student and are labeled with a letter (L or H) to represent low or high performing and a number (1, 2, or 3) to represent cohort number. The intersection of two branches represents the distance between them. The dotted line represents the chosen cluster cutoff point; each of the eight places that a branch intersects with the cutoff point represents one of the eight clusters. From this, we see that some types of behaviors are more likely to lead to low performance (e.g. Clusters 1 and 7), and some types of behaviors are more likely to lead to high performance (e.g. Clusters 4 and 5). However, some of the clusters aren't as predictive of performance (e.g. Clusters 6 and 8). . . . . 76

4.14. Dendrograms of student clusters for each of the individual cohorts: (a) Cohort 1, (b) Cohort 2, and (c) Cohort 3. These dendrograms show that the discrepancies between performance and cluster are not a result of analyzing multiple cohorts together; these discrepancies occur within cohorts as well. For example, for Cohort 1, we see a grouping of low performers, a grouping of high performers, and two mixed groupings. . . . . 77



4.15.	Three-dimensional representation of students colored by cluster. Python code was created that automatically plots students based off of token proportions in a three-dimensional space where the x-axis represents proportion of gap tokens, the y-axis represents proportion of solution tokens, and the z-axis represents proportion of impact tokens. This visualization uses color to represent a student’s cluster and shape to represent a student’s performance. By coloring the students by cluster, instructors can still take in some of the information that is otherwise lost by the reduction from seven dimensions to three dimensions. For example, both Clusters 2 and 4 have mostly solution tokens, so they are in similar locations on the graph. However, the color still allows us to differentiate between Cluster 2 (High Explore) and Cluster 4 (High Solve). . . . .	78
5.1.	A visual interpretation of individual networks, individual centroids, average networks, average centroids, and difference networks. For each unit (in this case, student), the number of co-occurrences is counted for each pair of actions. The weighted network for that unit is then created where each node represents a code and each edge represents the relative co-occurrence of that pair of codes. The individual centroids (represented by circles) are plotted corresponding with the edge weights. To calculate the average network for a group of units, the weights of each edge for all networks are averaged. Next, the average centroid (represented by a square) is plotted corresponding with the average network edge weights. Alternatively, a difference network can be created by taking the difference between each set of corresponding edges. . . . .	96
5.2.	ENA representation of all students in Cohort 1. Each circle represents a student; green circles represent a student who was identified as high-performing, and red circles represent a student who was identified as low-performing. The labeled squares represent average centroids for high- and low-performers, and the dotted boxes represent a 95% confidence interval for that group. Note that the confidence intervals do not overlap; this corresponds with the results of the Mann-Whitney Test for a statistical significant difference between the two groups ( $p < 0.00001$ ). It also should be noted that this projection did not use a means rotation – meaning the algorithm did not use a supervised approach to intentionally separate the two groups. Yet, the low and high performers separate themselves across the x-axis (which represents the first singular value decomposition and accounts for 21.0% of the variability within the data). Thus, much of the variability in the data is related to student performance. . . . .	98

5.3. ENA representation comparing the average network for high-performing students in Cohort 1 and low-performing students in Cohort 1. From figure 5.2, low- and high-performing students generally separated across the x-axis. To understand the differences between the two, the difference network between the average high-performing and average low-performing can be plotted. If an edge is green, it is more likely to be an action pair completed by a high performer. If an edge is red, it is more likely to be an action pair completed by a low performer. Edge thickness represents how great the difference was between the two groups for that specific node, allowing us to see which action pairs were more likely to occur for each group. Because the goodness of fit for this visualization is 0.9886 on the x-axis and 0.9546 on the y-axis, we can also interpret the data using the positions of the nodes. *Explore* and *solve* codes appear more on the left (high-performing side), and *environment* codes appear more on the right (low-performing side). Therefore, we can conclude that high-performing students were more likely to iterate between *explore* and *solve* tokens, whereas low-performing students often focused on starting and finishing *environment* tokens. . . . . 99

5.4. Average percentages of each set of code pairs for (a) low-performing and (b) high-performing students. . . . . 100

5.5. ENA representation of all teams in Cohort 1. Each circle represents a team; green circles represent a team who was identified as high-performing, and red circles represent a team who was identified as low-performing. The labeled squares represent average centroids for high- and low-performing teams, and the dotted boxes represent a 95% confidence interval for that group. Note that the confidence intervals do not overlap; this corresponds with the results of the Mann-Whitney Test for a statistical significant difference between the two groups ( $p < 0.01$ ). It should once again be noted that this projection did not use a means rotation – meaning the algorithm did not use a supervised approach to intentionally separate the two groups. Yet, the low and high performers separate themselves across the x-axis (which represents the first singular value decomposition and accounts for 31.2% of the variability within the data). Thus, much of the variability in the data is related to student performance. . . . . 101

5.6. ENA representation comparing the average network for high-performing teams in Cohort 1 and low-performing teams in Cohort 1. From figure 5.5, low- and high-performing students generally separated across the x-axis. To understand the differences between the two, the difference network between the average high-performing and average low-performing can be plotted. If an edge is green, it is more likely to be an action pair completed by a high-performing team. If an edge is red, it is more likely to be an action pair completed by a low-performing team. Edge thickness represents how great the difference was between the two groups for that specific node, allowing us to see which action pairs were more likely to occur for each group. Because the goodness of fit for this visualization is 1.0 on both the x- and y-axis, we can also interpret the data using the positions of the nodes. *Explore* and *solve* codes appear more on the left (high-performing side), and *environment* codes appear more on the right (low-performing side). Therefore, we can conclude that high-performing students were more likely to iterate between *explore* and *solve* tokens, whereas low-performing students often focused on starting and finishing *environment* tokens. . . . . 102

- 5.7. ENA representation of team trajectories for Cohort 1. Each circle represents a team’s aggregated network for a quarter of the semester. Light colored dots represent earlier in the semester, and brighter colored dots represent later in the semester. Green dots represent teams identified as high-performing, and red dots represent teams identified as low-performing. Because performance groups separate on the x-axis (the first SVD), we can deduce that differences in performance account for the first level of variability in the data (16.6%). Because the quarters separate on the y-axis (the second SVD), we can deduce that temporal behavior differences account for the second level of variability in the data (10.0%). . . . . 103
- 5.8. ENA representation comparing the average network for high-performing teams in Cohort 1 and low-performing teams in Cohort 1. If an edge is green, it is more likely to be an action pair completed by a high-performing team. If an edge is red, it is more likely to be an action pair completed by a low-performing team. Edge thickness represents how great the difference was between the two groups for that specific node, allowing us to see which action pairs were more likely to occur for each group. Because the goodness of fit for this visualization is 0.9304 on the x-axis and 0.9578 on the y-axis, we can also interpret the data using the positions of the nodes. Once again, *explore* and *Solve* codes appear more on the left (high-performing side), and *Environment* codes appear more on the right (low-performing side). Temporally, *Survey* and *Define* codes tend to occur closer to the top of the graph (Q1 end), and *Draft* and *Share* codes tend to occur closer to the bottom of the graph (Q4 end). However, the differences between Q1 and Q4 behavior is less than the differences between high- and low-performing teams based off of the assumption that the first SVD accounts for the most variability and aligns with performance. . . . . 104
- 5.9. The plotted ENA centroids for all students across all 3 cohorts. Each circle represents a student, and labeled squares represent an average network for a group of students. The x-axis represents the first singular value decomposition and accounts for 18.2% of the variability in the data. Therefore, by noting that cohorts are grouped along the x-axis (2019 on the left, 2020 on the right, and 2021 in the middle), we can assume that much of the variability in the data is related to differences between the three cohorts. The y-axis represents the second singular value decomposition and accounts for the next 7.8% of variability in the data. On average, lower performing students are grouped near the top, and higher performing students are grouped near the bottom, suggesting that performance level also plays a large role in variability. It is also interesting to note the large distance between low- and high-performing groups for 2019 (whereas the averages for 2020 and 2021 have significant overlap. . . . . 108

5.10. The average networks for (a) Cohort 1 in 2019, (b) Cohort 2 in 2020, and (c) Cohort 3 in 2021. Each cohort varies in the strongest action pairs. Cohort 1’s connections are more evenly distributed; Cohort 2’s connections largely iterate between Survey and Define codes; Cohort 3’s connections have some variability, but largely iterate between start and finish codes of the same action type (e.g. S.Survey is highly connected to F.Survey). These results are quantitatively demonstrated in Figure 5.11, which shows the relative weight of each action pair for each of the three cohorts. We define complexity of innovative activity as the number of difference connections made. Using this definition, we can visually see that Cohort 1 (2019) has the highest complexity, Cohort 3 (2021) has the next highest complexity, and Cohort 2 (2020) has the lowest complexity. These results are quantitatively supported through the results presented in Table X. This change in complexity is not surprising because of the structure of the course for each of these years; 2019 was loosely structured, 2020 was heavily structured because of token “stacking”, and 2021 was somewhere in the middle. . . . . 109

5.11. Average percentages of each set of code pairs for (a) Cohort 1, (b) Cohort 2, and (c) Cohort 3. . . . . 110

5.12. Box and whisker plots that show the ENA network complexity in bits for each of the six sub-groups: Low 2019, High 2019, Low 2020, High 2020, Low 2021, and High 2021. The bars represent statistical difference between groups and subgroups ( $p < 0.01$ ). There is a statistically significant difference between Low 2021 and High 2021, between All 2019 and All 2020, and All 2019 and All 2021. When given less structure (as in 2019), both low- and high-performing students had greater ENA network complexity. When given some open-ended structure (as in 2021), high-performing students had greater ENA network complexity than low-performing students. . . . . 111

## LIST OF ABBREVIATIONS

IBL. . . . .	Innovation-Based Learning
DOK. . . . .	Depth of Knowledge
LA. . . . .	Learning Analytics
EDM. . . . .	Educational Data Mining
DBER. . . . .	Discipline-Based Education Research
RQ. . . . .	Research Question
ECG. . . . .	Electrocardiogram
SVM. . . . .	Support Vector Machine
KNN. . . . .	K-Nearest Neighbors
RF. . . . .	Random Forest
LR. . . . .	Logistic Regression
TFIDF. . . . .	Term Frequency-Inverse Document Frequency
ROC. . . . .	Receiver Operating Characteristic
AUC. . . . .	Area Under Curve
ENA. . . . .	Epistemic Network Analysis
SNA. . . . .	Social Network Analysis
SVD. . . . .	Singular Value Decomposition
STEM. . . . .	Science, Technology, Engineering, Mathematics

## LIST OF APPENDIX TABLES

<u>Table</u>	<u>Page</u>
A.1. A sample list of code vectors for a given student. Each row in the table represents one log line that has been converted into a code vector. . . . .	138
A.2. A sample list of adjacency vectors for a given student using the data in Table A.1. Each row in the table represents one student for one week. The three rightmost columns represent existence of a pair of codes. 'S/D' represents existence of 'Survey' and 'Define', for example. . . . .	139

## LIST OF APPENDIX FIGURES

<u>Figure</u>	<u>Page</u>
A.1. A partially drawn ROC where the large points represent the drawn decision boundary at about 0.45. For the top case, this decision boundary results in a false positive rate of 0 (0/5 negative cases were predicted to be positive) and a true positive rate of 1 (5/5 positive cases were predicted to be positive). For the middle case and bottom case, this decision boundary results in a false positive rate of 0.4 (2/5 negative cases were predicted to be positive) and a true positive rate of 0.6 (3/5 positive cases were predicted to be positive). . . . .	137
A.2. A completed ROC graph for the three classification models. . . . .	137
A.3. A visual representation of the six association vectors in a 3-dimensional space. Blue vectors correspond to students in Group 1, and red vectors correspond to students in Group 2. Each axis represents the number of co-occurrences for a pair of codes. . . . .	140
A.4. A visual representation of the six normalized association vectors plotted in the same 3-dimensional space as A.3. Each vector is now length 1. . . . .	141
A.5. A visual representation of the six normalized association vectors now centered at the origin. Note that the axes ranges have changed from figure A.4. . . . .	142
A.6. The six example students projected into the 2-dimensional space using matrix $R$ . This figure was created using ENA Web Tool, and the values match those calculated by the author using MATLAB (seen above). . . . .	143
A.7. A four-node network with nodes A, B, C, and D for the sake of the worked example. . .	145
A.8. Complexity in bits for a variety of four-node networks . . . . .	147
A.9. Two worked examples of calculating network complexity (entropy) for four-node networks	148

# 1. INTRODUCTION

## 1.1. The Call for Innovative Engineers

The World Economic Forum published its most recent Future of Jobs report that explored what skills will be most important in the workplace in the year 2025. The most valuable skills included creativity, critical thinking, and complex-problem solving, and the number one spot went to analytical thinking and innovation [1]. Similarly, the National Academy for Engineering [2], ABET [3], and hundreds of industry engineers [4] have agreed that innovation is one of the most important skills that engineers should be developing. This means that safe environments for practicing innovation need to be developed [5]. However, questions still remain about how to teach and foster innovation in students. Therefore, best practices for teaching and researching innovation in educational settings still need to be developed.

## 1.2. Innovation-Based Learning to Answer the Call

One model for teaching engineering students how to innovate is Innovation-Based Learning (IBL). In this model, students learn fundamental engineering principles and apply them to create value on an innovation project [7]. Baregeh et al. define innovation as “the multi-stage process whereby organizations transform ideas into new/improved products, service or processes, in order to advance, compete and differentiate themselves successfully in their marketplace” [8], and we used this definition and Baregeh et al.’s corresponding literature review analysis to extract three main ideas that differentiate IBL from other design and project-based engineering experiences: 1) teams must identify a gap (“differentiate themselves”), 2) teams must develop a solution (“transform ideas into new/improved products, service, or processes”), and 3) teams must create impact (“successfully advance”). IBL differs from many other design and project-based engineering experiences because of its focus on not only solution development, but also gap identification and impact creation.

In this work, innovative success is defined as the creation of impact that extends beyond the course in both place and time [9]. Successful students provide evidence of an existing gap, develop a proof-of-concept solution, and create value for an individual or community outside the course

---

Some material in this introduction was drawn directly from [6], a publication co-authored by Lauren Singelmann and Dan Ewert. Lauren Singelmann drafted and revised all versions of this chapter. Dan Ewert served as a reviewer of the content.



that continues beyond the end of the course. For example, high-performing students have created value by publishing research papers, competing in business development competitions, submitting invention disclosures, and more. These activities create value for the scientific community or society as a whole, and because the work is shared publicly, this value remains even after the course is over.

In this study, students were in a cardiovascular engineering course, so each of the innovation projects was related to cardiovascular engineering, and students learned five industry-recommended pillars of cardiovascular engineering: the functional block diagram of the heart, pressure/volume loops and time domain, resistance and compliance, electrocardiogram (ECG), and the arterial system. Students demonstrated their understanding of the five pillars and used their knowledge to support work on innovation projects such as a multiparameter biosensor, a Simulink model of the cardiovascular system, a sensor to detect deep vein thrombosis, an algorithm that could diagnose illness by sound, and more. Students were able to choose their projects and teams [10, 11], and they had the freedom to learn other material to support their project [12]. To pass the course, students needed to show competence in each of the five pillars of cardiovascular engineering. To earn a higher grade, they also needed to demonstrate that they contributed to a project that had external value (or value outside the course) [7]. Many students rose to the challenge and created high external value deliverables, but questions still remain about the factors that lead to student success.

### **1.3. Dataset**

In order to better understand and predict student success in the course, data were collected using MOOCIBL, a custom learning management system for IBL settings. The data spanned three cohorts and 97 students.

#### **1.3.1. MOOCIBL**

MOOCIBL is a custom online learning management system specifically designed for IBL. To demonstrate their learning, students created tokens in an online platform called MOOCIBL [13]. Each token represents a piece of learning and could be in a variety of topics. As students learn course material and work on their innovation projects, they create learning ‘tokens’ that each represent a piece of learning. These tokens are logged in MOOCIBL, the custom course learning management system. Each token has a title, description, and place to link evidence of learning

[14]. MOOCIBL logs every time a token is created, edited, reviewed, or deleted. In total, 3,016 tokens were created across the 97 students.

### 1.3.2. Cohorts

Although the main motivations behind the course remained constant across the three cohorts, there were some key differences/modifications made over time<sup>1</sup>. In 2019, Cohort 1 had a significant amount of freedom in how they uploaded tokens. They created overarching learning objectives and then added deliverables under each learning objective. Students were encouraged to (and often did) add learning objectives and deliverables to MOOCIBL while they were still in progress, but there were few requirements about when or how new information should be added. This lack of structure led to a few challenges; many students did not upload the required pillar concepts, and it was challenging and time-consuming to monitor students' individual progress.

Therefore, in 2020, Cohort 2 was given significantly more structure. The pillars of cardiovascular engineering became more central to the course, and students were required to “stack” their tokens using the Webb’s Depth of Knowledge (DOK) taxonomy [?]. These changes made grading more straightforward; if a student had DOK-1 (recall and reproduce) and DOK-2 (skills and concepts) tokens for all five pillars and at least one DOK-3 (strategic thinking) and DOK-4 (extended thinking) token, they earned an ‘A’ in the course. However, this increased structure also led to more challenges. For students, the “stacking” structure required them to complete actions in a certain order, even when it may not fit the immediate needs of the project or the learner. For instructors, monitoring student progress throughout the semester was still a challenge; all students followed the general provided structure, so it was challenging to use the MOOCIBL data to find meaningful differences among students and teams.

Finally, in 2021, Cohort 3 was given a structure that fell somewhere in the middle. The data from Cohorts 1 and 2 were used to create the IBL Framework (which will be presented in Chapter 3), and students were instructed to use the framework as a guide when working on their project. For Cohort 3, the five pillars were still “stacked”, but this process was separated from the project.

---

<sup>1</sup>This sub-subsection makes some general claims about the student and instructor experience that are not backed up by data and are outside the scope of the work. However, I think it is important to attempt to summarize these observations and claims because it gives the reader a better understanding of why changes were made over time. I have done my best to accurately portray the reasoning of the entire instructional team.

In order to analyze data across cohorts, the final dataset has two main characteristics: 1) the student-written text is concatenated for each token into a single string of text, and 2) tokens from Cohorts 1 and 2 were categorized into the framework categories to be consistent with Cohort 3. Despite the differences in the data, identifying these two unifying characteristics allowed for the same analysis methods to be used from year to year. However, that does not mean the differences across cohorts can or should be ignored. Rather, results should be interpreted within the context of that specific year and cohort.

### **1.3.3. Participants**

Within the 3 cohorts, 97 students participated in the study. 41 were marked as low-performing, and 56 were marked as high-performing (where high-performing is defined as having contributed to a high external value deliverable as defined in [9]). Cohort 1 had 28 students, Cohort 2 had 35 students, and Cohort 3 had 34 students. Of the students who provided demographic information, 64.9% were male and 35.1% were female; 76.3% were white, 17.1% were Asian, and 3.9% were Black/African American; 60.7% were undergraduate students and 39.3% were graduate students.

## **1.4. Learning Analytics and Educational Data Mining**

In order to better support students in IBL contexts, the MOOCIBL data can be analyzed using learning analytics and educational data mining techniques. Learning analytics has been defined as *“the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs”* [15]. Educational data mining has been defined as *“a field that exploits statistical, machine-learning, and data-mining algorithms over different types of educational data... to analyze these types of data in order to resolve educational research issues”* [16].

These fields started to gain popularity in the 2000s when computers became more prominent in learning environments. According to a 2013 review of LA/EDM, some of the original goals of the fields were to predict student performance, make recommendations to students or teachers, and model domains to determine relationships between concepts [17]. Since this original review in 2013, applications of LA/EDM have continued to expand to a variety of contexts (e.g. sentiment discovery, analysis of programming code, and foreign language learning), including some contexts

more directly related to IBL (e.g. including collaborative learning and group work and self-regulated learning) [18].

There are many similarities between LA and EDM, and both have a shared common goal: “improving education quality by analysing huge amounts of data to extract useful information for stakeholders” [19]. However, a few main differences between LA and EDM have also been identified [20]:

- **Discovery:** EDM focuses on automating discovery of information whereas LA focuses on leveraging human judgement for discovery.
- **Reductionism and Holism:** EDM aims to reduce problems to their individual components and compare relationships among them whereas LA aims to understand systems as wholes.
- **Origins:** EDM originated from educational software and LAK originated from the semantic web.
- **Adaptation and Personalization:** EDM places greater focus on automation, whereas LA places greater focus on empowering instructors and learners.
- **Techniques and Methods:** EDM methods commonly include classification, clustering, relationship mining, and visualization whereas LA methods commonly include network analysis, sentiment analysis, discourse analysis, concept analysis, and sensemaking models.

In many ways, the goals of LA align with those of Discipline-Based Education Research (DBER) as defined by the National Academies DBER Report. These goals include “*understand[ing] how people learn the concepts, practices, and ways of thinking of science and engineering, understand[ing] the nature and development of expertise in a discipline, help[ing] identify and measure appropriate learning objectives and instructional approaches that advance students towards those objectives, contribut[ing] to the knowledge base in a way that can guide the translation of DBER findings to classroom practice, and identify[ing] approaches to make science and engineering education broad and inclusive*” [21]. Both LA and DBER place focus on not only discovery, but the *understanding* of that discovery. In addition, both place focus on bringing findings back into the classroom to support both teachers and students, and LA’s focus on a holistic view of systems aligns better with DBER’s goals for inclusivity.

Therefore, the goal of this dissertation is to shift from the more EDM-focused approach taken in previous work to a more LA-focused approach. Previously published work by the author used reductionist methods; the data were simplified to only a few features (words used when writing tokens), and these features were used to create linear classification [22] and clustering [23] models with the goal of automating prediction and discovery of information [24]. This shift to a more LA-focused approach will require qualitative exploration and analysis, new methods that consider the complexity of the data, and a shift from an automated process to one that involves feedback loops between researcher and computer.

### **1.5. Research Questions**

The first step in moving to a more LA-focused approach is identifying or creating a framework for IBL data. After the framework is created, the second step is to determine if and how the framework can be used to improve the methods originally developed by the author in [24]. Finally, the third step is to extend the work to new methods that were not previously feasible before the implementation of the framework.

These three steps align with the three research goals – each with two corresponding research questions.

1. Development of an Innovation-Based Learning Framework
  - (a) Are there existing frameworks that are appropriate for categorizing Innovation-Based Learning data? If not, what is an appropriate framework?
  - (b) Can a classification model be used to sort student text into the categories of the IBL framework with greater consistency than a human rater?
2. Extension of Classification and Clustering in IBL by Leveraging the IBL Framework
  - (a) Does categorizing student text into framework categories improve the performance of a classifier model that separates between lower and higher performing students?
  - (b) Given student token proportions of each framework category, what types of student clusters form?
3. Implementation of Epistemic Network Analysis in IBL through the IBL Framework

- (a) Do high- and low-performing students and teams have different behaviors in the course in the context of co-occurrence?
- (b) How does the structure of the course change student behavior in the context of co-occurrence?<sup>2</sup>

## 1.6. Overview of Chapters

Chapter 2 summarizes existing methods for modeling, measuring, and understanding innovation from multiple fields. This chapter is adapted from the candidate’s STEM Education Qualifying Exam submission. Chapter 3 defines the creation of a framework for IBL and a classifier model that automatically sorts student text into the categories of the framework. The first part of Chapter 3 is adapted from a conference paper titled “Creating of a framework that integrates technical innovation and learning in engineering,” published at *2021 IEEE Frontiers in Education Conference* and was co-authored by Lauren Singelmann, Ryan Striker, Enrique Alvarez Vazquez, Ellen Swartz, Mary Pearson, Stanley Shie Ng, and Dan Ewert. The second part of Chapter 3 is adapted from a publication titled “Leveraging the innovation-based learning framework to predict and understand student success in innovation,” which has been accepted to the *IEEE Access Education Society Section* and was co-authored by Lauren Singelmann and Dan Ewert. Chapter 4 discusses the use of the IBL Framework to extend previous classification and clustering methods. The first part of Chapter 4 is also adapted from the *IEEE Access* publication. Chapter 5 details the implementation of a new method for analyzing IBL data: epistemic network analysis. The first part of this chapter will be submitted to the *Journal of Learning Analytics* or similar, and the second part of this chapter will be submitted to the journal *Biomedical Engineering Education*, specifically in its special section, *Experiential Learning in Biomedical Engineering*. Chapter 6 will combine insights from all three chapters to discuss implications for teaching and researching innovation, as well as limitations and future directions. Finally, Chapter 7 will summarize the value of the work as a whole.

---

<sup>2</sup>This research question was not one of the originally proposed research questions. After analyzing data from all three cohorts, it became clear that each of the cohorts had very different behaviors. Thus, this question was added to further explore how the various course structures influenced student behavior.

## 2. LITERATURE REVIEW

### 2.1. Introduction

The pathway to innovation is not a straightforward one. Economists, engineers, historians, and psychologists alike have conducted research in order to better understand the innovation process, but questions still remain about how to teach innovation in educational settings and leverage innovation in industry settings. Although there have been interesting and meaningful findings, researchers are still far from identifying unifying theories and best practices. Qualitative studies have led to a richer understanding of factors that play into innovation, but these studies are relatively slow and almost impossible to scale, especially at the current rate of innovation. On the other hand, methods using machine learning in areas such as LA/EDM offer fast and scalable approaches, but they rarely generalize to other contexts and often run the risk of oversimplifying the complex interactions. Although both qualitative and machine learning methods both have weaknesses, they also both have strengths that can be combined and leveraged. Thus, the first step in this process is to explore the wide variety of work that has been done to understand innovation.

The main objective of this literature review is to summarize existing methods for modeling, measuring, and understanding innovation from multiple fields (e.g. engineering, business, education, and psychology). Although the studies presented are from a wide variety of backgrounds and contexts, unifying themes and ideas can still be identified and discussed.

Some of the studies shared in this review are ordered; constraints are put on the system in order to control the possible inputs and outputs (e.g. analyzing how students solve a specific problem in a specific online system). Other studies fall on the unordered side; there are many components interacting in dynamic ways. These studies might be exploring students working in groups on open-ended projects or companies working to develop a new device. In order to compare and discuss ordered and unordered systems, the studies will be discussed in the context of the Cynefin Framework. The Cynefin Framework, pictured in 2.1, is a conceptual framework that sorts situations and contexts into four domains: simple, complicated, complex, and chaotic [25].

The simple domain consists of predictable and straightforward relationships. The simple domain is also sometimes referred to as the obvious domain because these relationships are apparent

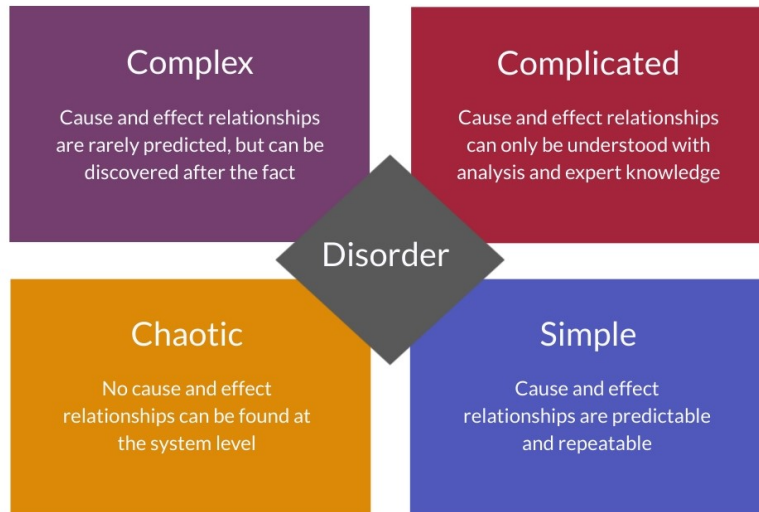


Figure 2.1. Cynefin Framework from [25]

to most observers. The complicated domain also consists of linear relationships, but they require expert knowledge or analysis to understand. The complex domain consists of relationships that interact in nonlinear ways; the sum is more than the parts, meaning behavior is not predictable, but behaviors can still be observed and trends can be found. Finally, the chaotic domain consists of interactions and behaviors where cause and effect relationships are not able to be identified, even after the fact. The simple and complicated domains are ordered systems, whereas the complex and chaotic are unordered systems [25]. These domains give us a framework that illustrates how components interact and suggestions for best research practices. Ultimately, the domain that researchers work in must align with the methods they are using, as well as their philosophical views about the behavior of social systems.

## 2.2. Innovation as Simple

Within the simple domain, relationships are obvious and do not require expert analysis, so very little research falls into this domain. However, many of the models that have been created to illustrate innovation are simple models in that they are linear and predictable. Although these simple models come from the earlier generations of the conceptual innovation models, they are still widely used. Later generations recognize that there are more components interacting both in and outside the system, but simple models often eliminate these exterior factors to keep their models easy to understand and communicate. However, by stripping away the external factors and



complex relationships, you lose the flavor of innovation; instead, you are left with just a component of the innovation process: engineering design<sup>1</sup>. These engineering design process models have been created for both industry and education contexts.

### **2.2.1. In Industry**

In industry contexts, engineering design process models can be used to manage product development, better allocate resources, and increase efficiency [26]. However, by making a simple model of a complex process, you lose accuracy and the ability to be illustrative of all scenarios. Hence, dozens of engineering design process models exist; each is generally easy to understand, but each with its own limitations. One review of existing engineering design process models cited 23 different models of how innovation occurs in industry [26]. Although they all were different, the authors still were able to create 6 categories that summarize stages of most of the existing engineering design models: establishing a need, analysis of task (which focuses on the function of the innovation), conceptual design (which focuses on the behavior of the innovation), embodiment design (which focuses on the structure of the innovation), detailed design, and implementation. The authors note these models (and similar ones) are heavily used and cited, but that they are most appropriate for managing the design process and teaching innovation to new designers in industry settings [26].

### **2.2.2. In Education**

Other simple engineering design process models focus more on the teaching and learning aspect in traditional educational settings (especially K-12). Dozens of these models also exist, but one example is shown in Figure 2.2 [27]. One benefit of modeling innovation in such a simplistic way is that it can allow for straightforward teaching and assessment. For an outreach activity like the “Slender Tower Challenge”, the steps may look like this:

1. Ask: The problem is identified as a class. Students need to create a tower that has the lowest base-to-height ratio. They get 10 pieces of paper and a roll of tape.
2. Research: As a class, we look at pictures of tall buildings and make observations about how they look.

---

<sup>1</sup>This dissertation differentiates between innovation and engineering design by assuming that innovation must consist of three components: identifying a new and unique problem gap, developing a solution to fill this gap, and creating impact. Engineering design usually consists only of developing a solution to a given problem.

3. Imagine: Students get into groups and draw a few different options for their tower.
4. Plan: Students pick out the tower that they want to build.
5. Create: Students get their materials and build a prototype.
6. Test: Students measure their base and height, calculate the ratio, and ensure that their tower is freestanding.
7. Improve: Students get more time to adjust their prototype before the final measurements are taken.

Although students can get creative with how they build their tower, the process they use is made up of clear, linear steps, allowing for straightforward assessment of an engineering activity or program (as seen in [28], for example). *If* you complete your planning stage by drawing your tower on a piece of paper, *then* I will give you your paper and tape so you can create your prototype. These simple models for engineering design work well because of the constraints placed around the activity: groups all get 10 sheets of paper, they all are being assessed the same way, they all get 50 minutes, and they all have a goal that is given to them. These constraints allow for a 50-minute lesson for 20 4th graders, but this one-size-fits-all engineering design process model doesn't give us a look inside the process of innovation or complex problem-solving (both of which are more dynamic processes).

Overall, simple models are great tools for teaching and assessing engineering design in both industry and education settings; they are easy to communicate, and they provide a clear framework for designing and assessing engineering design activities. However, their simplicity comes with constraints that limit the ability to consider the dynamic and interconnected relationships that are found in innovation. There is a place for *Innovation as Simple*, but we'll need to remove some of these constraints to get a better picture of the process of innovation.

The complicated domain consists of if/then relationships that are not obvious, but can be identified by an expert or through analysis. Many of the studies presented in this section involve simplifying complexities into a few specific variables. Even though they are exploring how people solve open-ended problems, there are still constraints in place to keep the problem within the complicated domain. These studies can still lead to new and insightful findings about the

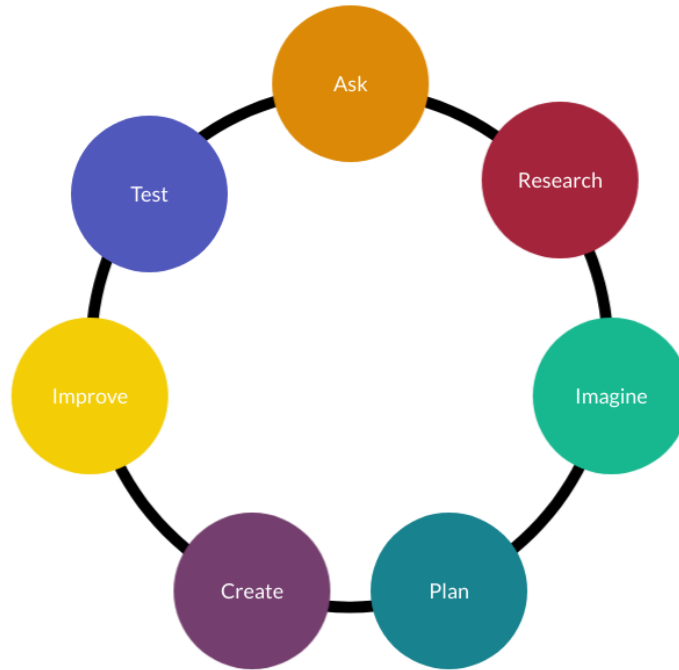


Figure 2.2. The Engineering Design Process as published on TeachEngineering [27]

innovation process, but the methods and framework proposed are still specific to the proposed context and could be missing how other non-measured components play a role. It should be noted that many of the authors mentioned in this section recognize and directly state the complexity of understanding innovation and problem-solving, but their preliminary work arguably falls more within the complicated domain.

### 2.2.3. In Industry

A variety of studies in industry have aimed to find a handful of predictors for innovation. Acs et al., for example, found a correlation between number of patents in a specific area and number of original innovations [29], simplifying innovation into one (somewhat) predictive factor. Other studies have tried to create more involved models with more variables, but these models do not account for interactions between variables and therefore still fall into the complicated domain. For example, Fleuren et al. created the Measurement Instrument for Determinants of Innovations that scores the chance of innovative success by measuring 29 determinants (e.g. client cooperation, financial resources, time available, etc.). Each of these determinants was given a score by a researcher using a Likert scale [30]. Similarly, Kasa took Likert data in categories such as organizational

culture, strategy, and technological modernity and used it to train a neurofuzzy system (a type of machine learning classification algorithm) [31]. Even though this model used more advanced computing techniques, it still falls into the complicated domain because it aims to simplify the prediction of innovation into fundamental rules.

#### **2.2.4. In Education**

Generally, many educational data mining techniques take a complicated approach; lots of data is collected and relationships are mined with the assumption that if enough data is analyzed, fundamental rules can be found that can improve teaching and learning. These techniques are appropriate for tasks such as determining how content in an e-learning platform should be ordered, identifying what material a student doesn't know, or recognizing off-task behaviors [32]. These researchers have data from hundreds of thousands of students, and creating if/then rules for the typical student tends to work well.

Less work, however, has been done in the complicated domain to explore tasks like innovation, where there is more than one right approach. Some, however, have used complicated approaches to explore components of the innovation process. Both [33] and [34], for example, use educational data mining techniques with a computer modeling tool called Energy3D that allows students to design renewable energy projects. [33] used the software to look specifically at engineering design, and [34] looked specifically at experimentation strategies. Using a specific prompt and narrowing the scope to these two ideas placed these into the complicated domain; they are looking for "rules" that are specific to a certain prompt and context. [33] mapped student actions in the software to various stages of design (e.g. adding a solar panel aligns to construction, removing a wall aligns to revision). A classifier model was then used to identify struggling students based off of their click behavior. [34] used qualitative analysis of log data to determine what behaviors are related to student experimentation. This analysis was then used to develop an algorithm that could automatically identify and classify these strategies. The results found that novice designers conducted very few experiments, and experiments became more systematic as designers gained proficiency. These studies are great examples of how work in the complicated domain could lead to work in the complex domain; they are strong because they tie student behaviors into task models that are driven both by student data and the literature (e.g. mapping click sequences to the engineering design process). Although these studies are both specific to a particular context, the task mod-

els could be adapted to other contexts, allowing for more complex relationships to be explored in depth. Currently, these studies are only able to look at small pieces of innovation (experimentation and design), but the quantitative approach allows for consistent and speedy analysis.

Other researchers have taken a more qualitative approach to understanding engineering innovation. These researchers analyzed and measured the engineering design abilities of both engineering students and experts, requiring the creation of a clear framework that could be used to compare each subject's process. [35] had students and experts complete an exercise where they are creating a playground that follows a set of requirements (e.g. safe, cost-effective, inclusive). As the subjects worked on the problem, researchers coded their actions into stages of the engineering design process: defining the problem, gathering information, generating ideas, modeling, feasibility analysis, evaluation, decision, and communication. The authors also noted that the identification of a need and implementation stages were part of their engineering design process framework but were not included in the scope of the study; the need was already identified for the participants, and they did not actually get to build their playground. Because of these exclusions, one might argue that the subjects were not participating in the process of innovation, but there are also many overlaps between the process that the subjects completed and the stages of innovating a new product or service (e.g. generating ideas, evaluation, making decisions, etc.)

Although this study is specific to one prompt and has therefore been sorted into the complicated domain, the results demonstrate that the prompt and processes used are complex. One observation from the timelines of how students and experts transitioned through the design stage shows that students had "choppy" timelines whereas experts had smooth cascading timelines. In addition, it was found that experts spent more time defining the problem and gathering information, and they often returned back to these activities throughout the activity. These results suggest that engineering educators should teach students to scope a design problem before planning details and take time to gather information both at the beginning and throughout the process. The authors end by suggesting that although this was a specific prompt and problem, the results still are useful for engineering educators across other contexts because this can promote meaningful discussion about engineering design. In addition, the authors suggest that this coding scheme could be used in wider settings [35]. Overall, this study is another great example of placing constraints on a

research problem to put it into the complicated domain, but finding ways to extend the methods, the results, and the implications back into the complex domain.

Although the findings of these studies can promote discussion and future research directions beyond their original scope, the research methods and analysis procedures are specific to a very specific context. To get a full picture of the innovation process and how students from different backgrounds and with different goals work together, we'll need to move to the complex domain.

### **2.3. Innovation as Complex**

The complex domain consists of relationships that are more unpredictable than within the complicated domain. The studies mentioned in this section expand the boundaries of problems to allow for more interactions beyond just if/then relationships. Components can interact in non-linear ways, provide feedback to each other, and ultimately lead to emergent properties that were not possible without the interactions. These interacting components can be due to multiple students working together, a more open-ended problem or prompt, or a wider scope in general.

#### **2.3.1. In Industry**

Many authors that study innovation “in the field” have tried to tackle why innovation is so difficult to predict and measure, and most arguments either implicitly or explicitly discuss complexity. One review paper by Dziallas and Blind looked at studies that aimed to measure indicators of successful innovation. They came up with a list of over 80 indicators including how team members are given ownership, project efficiency, time to market, time to implementation, etc. However, they also note that it is not yet possible (or maybe never will be possible) to develop concrete factors because of the interconnectedness of factors and lack of subjective, consistent data. The papers cited in this review look at only a subset of factors of a small piece of the process, so Dziallas and Blink advocate for future work that aims to develop understanding of the relationships between these factors [36]. Another group of researchers states that this complexity is due to the ambiguity of the word innovation. Innovation can be defined as an outcome, a process, and a mindset, but teasing out and measuring components from just one of these factors is oversimplifying the dynamic relationships between them [37]. Therefore, researchers that have aimed to identify characteristics of successful innovators [38] or companies [39, 40] typically use qualitative methods that allow for exploration of a large variety of constructs. [38] conducted interviews to find characteristics of strong innovators and found that there were some commonalities

(e.g. curious, creative, strong systems thinkers), but the relationships found were nonlinear and not always predictable, illustrating the complexity of innovation.

Similarly, [39] looked at important components of small- and medium-sized enterprises (SMEs) and found that successful innovation depended on the constraining and enabling factors ranging from resources, competencies, the company's organizational structure, etc. [40] also found that there are trends in innovative companies, but that these trends vary from industry to industry, leading to even more complexities. In order to tackle these complexities, [39, 40] both created frameworks to help categorize, measure, and find relationships between these factors. Although these studies come from business and are geared for more industry-based settings, the frameworks that were created can bridge the gap between qualitative observations and coding to quantitative analysis practices. Future work in this area may benefit from similar methods, especially those hoping to scale up studies on innovation using tools like machine learning.

### **2.3.2. In Education**

As mentioned in the previous section, most work in EDM/LA has fallen into the complicated domain, but more research is emerging in the complex domain. In fact, arguments have been made that EDM/LA research and the complex domain go hand-in-hand [41]. Berland et al. offer four main reasons for this strong partnership: 1) there are EDM/LA methods that still allow for illustrative qualitative analysis (not just simplistic statistical methods), 2) EDM/LA allows qualitative researchers to scale up their methods, 3) EDM/LA methods can allow for more precision and replicability than some traditional qualitative methods, and 4) EDM/LA allows for realtime feedback for learners and educators (which can be especially useful in complex, nontraditional learning environments). This paper by Berland et al. was published in 2014, and almost of the EDM/LA studies in this section were published after 2014, so work in this area is still very new. Only my own previous work has made an effort to use EDM/LA to explore the whole innovation process, but EDM/LA research about engineering design and complex problem solving are becoming increasingly popular, and the methods and analysis techniques used can be used to grow my current work.

These synergies can be leveraged because researchers measuring complex problem solving face many similar challenges to those measuring innovation as a whole. Complex problem solving can be defined as overcoming dynamic and unfamiliar barriers between a given state and a goal

state [42]. By that definition, innovation is a type of complex problem solving. Compared to the measurement of other cognitive tasks, complex problem solving differs in that it should account for five additional factors: 1) intransparency, 2) complexity, 3) interconnectedness of variables, 4) polytely of the task (meaning multiple goals exist and must be balanced), and 5) dynamics of the system [43]. For those measuring innovation, similar components should be considered. Unlike traditional learning tasks, the research methods should include a way to account for how the innovator is monitoring progress, what they are learning, and how those things are connected together to create a solution and impact. Researchers in EDM/LA have measured these factors through the use of qualitative frameworks, multimodal analysis, and graph/network analysis.

For example, one virtual tool for measuring complex problem solving, MicroDYN, uses a qualitative framework that maps student tasks to each of those five areas. MicroDYN has students monitor input and output variables in a virtual chemistry lab; the problem is complex in that there are interconnected variables and the user must gather information as they go in order to solve the problem. Rather than just measuring information acquisition, MicroDYN collects datastreams related to each of the five factors related to complex tasks. For example, users keep track of information they are gaining over time (intransparency), create a model in the platform (interconnectedness of variables), and compare their target values and achieved values (polytely of the task). These actions can then be measured to assess users' abilities to solve complex problems that don't have a single possible solution [44].

Another platform that maps student log data to a qualitative framework is iRemix, a platform for an English and Language Arts course where students created multimedia to demonstrate their learning. The platform allowed students to engage with course material, post their own work, and connect with other students and their work. The analysis techniques mapped click behavior to a variety of learning activity types derived from the literature including creating and revising work, seeking support, and exploring the community. The use of this framework allowed the research team to automatically give each student a creative production, self-directed learning, and social learning score, potentially leading to student-specific support and/or interventions [45].

Another strategy for modeling complex EDM/LA data is the use of multimodal analysis (the collection of multiple streams of data including clickstream data, body tracking, student text, and eye tracking as detailed in [46]). Multimodal analysis has been used widely across various EDM/LA



applications, but it especially has a place for modeling complex behaviors in environments such as project-based learning. Spikol et al. [47], for example, developed methods for monitoring how students work on a variety of projects including inventing a toy, creating a color sorter machine, and designing an autonomous automobile. Although the projects differ from each other, the authors used flexible datastreams and analysis methods including hand tracking, face tracking, Arduino software data, and noise levels. However, it may be challenging to use methods like this to model innovation because of the different time periods; these sessions lasted just a few hours. Nonetheless, the data fusion techniques that consider multiple modes of data over time could still be useful for longer term projects.

Graph and network analysis can also be used to understand how students approach complex problems. Chen and Zhang developed a tool that allows students to map connections between ideas, highlight material, and add notes. For example, students who are learning about the water cycle read material and create a concept map to answer various questions (e.g. How does water vapor not go higher than the clouds?) [48] Similarly, Giabbanelli et al. created a tool that allowed students to create concept maps of how they think about various ill-structured topics [49]. These concept maps could then be compared with expert concept maps to determine not only if a student knows a concept, but how they think about it and relate it to other concepts.

A similar, but more qualitative, approach is the use of epistemic network analysis (ENA). Epistemic network analysis is a mixed methods analysis technique that takes coded qualitative data and quantitatively calculates the strengths of relationships between various elements, creating a dynamic network that can be used for analysis of a group or comparisons between groups [50]. It is part of a larger body of research methods called quantitative ethnography, an approach that differs from other mixed methods approaches in that the qualitative and quantitative methods are combined into one solution rather than having two separate analysis techniques [51]. ENA varies from traditional code and count strategies because it includes a temporal aspect, allowing the research team to build quantitative representations that account for both frequency of codes and their temporal relationships with other codes [52]. Eagan and Hamilton, for example, used epistemic network analysis to find relationships between constructs seen when people were working in a makerspace. These constructs included ability to see the perspective of others, content confidence, identity, self-awareness, and self-efficacy. The researchers could use the networks that were created

to compare how these constructs were related for each group of students [53]. ENA can also be used to find relationships between action types. Csanadi et al. coded action types as groups of students worked on solving complex problems. Codes included actions such as problem identification, hypothesis generation, and evidence generation [52]. ENA is a relatively new analysis technique that is just starting to pop up in EDM/LA [50], but the data collected in our Innovation-Based Learning course might also benefit from using ENA because it would allow us to see how different groups of students transition between various project stages.

These studies in EDM/LA allow for the measuring/analysis of how students interact in complex learning environments, but unlike the studies in industry, none looked at the entire innovation process (including gap identification, solution development, and the creation of impact). My thesis work was a first attempt at bridging the gap between studies in industry and studies in education by looking at the entire innovation process in educational settings using scalable methods. Students in an Innovation-Based Learning course logged their goals and progress over time in an online portal. This portal then created logs of student behavior, and these logs could then be mined using two methods for machine learning: classification and clustering. The classification work showed that the words that students were using helped differentiate between higher and lower performing students [22], and the clustering showed that different groups of behavior emerged beyond the binary groups of higher performing and lower performing [23]. However, the methods I originally used in my thesis fell slightly more into the complicated domain. Rather than taking a complex approach, I simplified my data down to just the words that students were using over the course of the semester. Therefore, my dissertation work could benefit from a framework similar to the ones seen in the papers from industry and the methods seen in the papers from EDM/LA.

The EDM/LA studies and the business/industry studies presented in this section helped me identify two gaps and two future directions in the study of innovation. Where one field lacks, the other offers a solution. The EDM/LA studies didn't address the entire process of innovation (gap identification to solution development to creation of impact), but the studies in business and industry offer existing frameworks for the entire innovation process that can be adapted to fit innovation in an educational setting. The business/industry settings lacked scalability, but the studies in EDM/LA identify some analysis methods that allow for complexity *and* scalable, realtime analysis (e.g. epistemic network analysis and quantitative ethnography).

Many of the studies that look at innovation are based in complexity, recognizing that although the relationships between people and external factors are dynamic and highly interconnected, these relationships can be found and measured. Or can they? Before we can confidently settle into the complex domain, one more domain must be explored: the chaotic domain.

## 2.4. Innovation as Chaotic

The chaotic domain is arguably the scariest domain for any researcher to tackle because no clear cause and effect relationships are present. However, chaos is hard to avoid because it is found in countless natural and social settings [54]. Arguably, innovation in its true form falls within the chaotic domain; multiple agents are interacting in unpredictable and nonlinear ways. The studies that explicitly aim to understand innovation in the context of chaos involve researching innovation in the “field”, meaning they are collecting data from real companies and enterprises that are aiming to design and market the next big thing.

### 2.4.1. In Industry

Many of the studies that involve researching innovation in the “field” come out of the University of Minnesota and their Minnesota Innovation Research Program [55]. In order to find commonalities about all of the innovations coming out of the program, they developed a framework about the process of innovation:

“Innovation is the invention and implementation of new *ideas*, which are developed by *people*, who engage in *transactions* with others over time within an institutional *context*, and who judge *outcomes* of their efforts and act accordingly.”

These five concepts are the core components that the group used to collect and analyze data. Although the innovations differed greatly, items can be compared by using a consistent framework. In short, the research group observes events and incidents, qualitatively codes them into constructs, and then converts them into bit maps for quantitative time series analysis.

For example, one study out of the group [56] looked at the development of two biomedical innovations to determine if the developmental process was orderly, random, or chaotic. To help visualize the differences between orderly, random, and chaotic, see figures 2.3 and 4.15 which was created by adapting the ideas and code presented in [57].

Each of the 6 plots were created from the simple logistic difference equation:  $X_t = k * X_{t-1}(1 - X_{t-1})$ . Various values of k cause different behaviors. For the ordered plot, k=3.2. For the

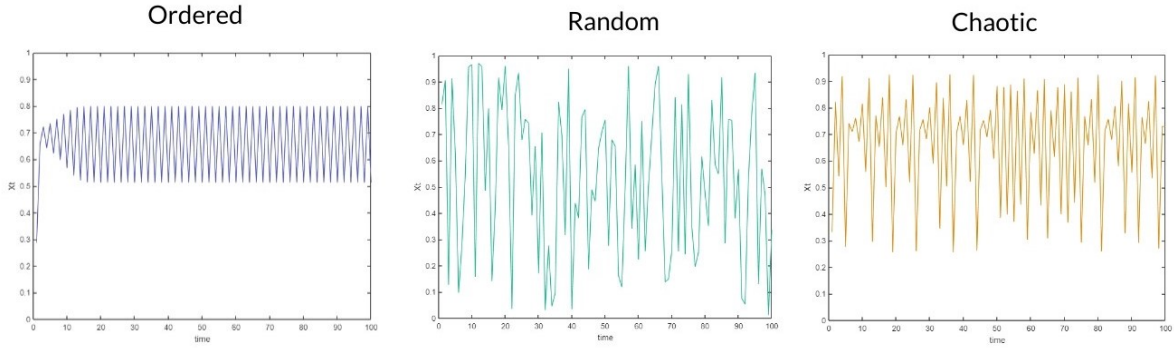


Figure 2.3. Behavior of  $X_t$  over time  $t$  for the simple logistic difference equation

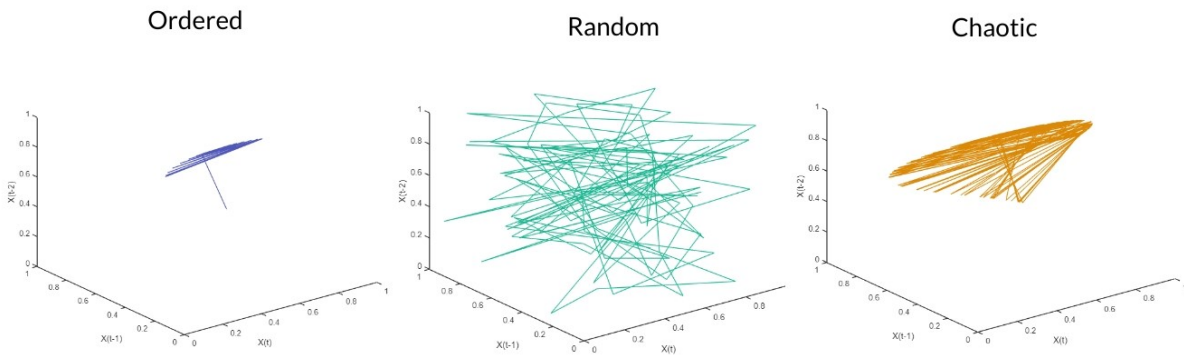


Figure 2.4. Phase diagrams of  $X_t$ ,  $X_{t-1}$ , and  $X_{t-2}$  for the simple logistic difference equation

chaotic plot,  $k=3.7$ . For the random plot, the `rand()` function in MATLAB was used to generate random behavior. The 2D diagrams depict behavior over time, and the 3D diagrams are phase diagrams that plot  $X_t$ ,  $X_{t-1}$ , and  $X_{t-2}$ . Although the 2D plots for chaotic and random might look similar at first glance, the 3D phase plots show that chaotic behavior still consists of patterns, not just random data points. figures were created using code and ideas adapted from [57].

Each month, the number of actions, outcomes, and context events were counted by using qualitative observations. Actions could be coded as either continuation of current actions or a change in action, outcomes could be either good news and accomplishments or bad news and mistakes, and context events were external environmental incidents that did not necessarily fall into the other categories. In order to determine the behavior over time, quantitative analysis was performed to calculate the Lyapunov exponent, correlation dimension, and the BDS statistic. Results showed that the beginning of the innovation process showed chaotic behavior that later

transitioned into ordered behavior. The authors note that the existence of chaotic behavior is especially interesting because it means that two common ideas about innovation are incorrect. Innovation is not an orderly process like what is represented by existing linear models for innovation, but it also doesn't occur out of random behavior. The authors end by suggesting that although challenges exist in being able to understand the process of innovation and learning, chaotic behavior means that a nonlinear model could eventually be created. They also discuss a need to better understand if actions drive outcomes or if outcomes drive actions [56].

Another more recent paper out of this group [5] shows how some of the group's ideas have evolved. They map innovation as a cycle of divergent behavior, constraining factors, convergent behavior, and enabling factors. Ideas may arise which leads to divergent behavior (e.g. learning new things, building relationships, trying out new ideas) that then leads to constraining factors (i.e. realizing that there are obstacles that eliminate some of the options). This causes a need to shift to convergent behavior (e.g. conducting tests, narrowing ideas, and implementing specific strategies). These convergent behaviors can lead to enabling factors that lead to new ideas, cycling back to divergent behaviors. This model is depicted in Figure 2.5. Ultimately, the authors argue that no one can control the success of innovation, but they can increase the odds of success by developing strategies for cycling through convergent and divergent behavior. The authors end by presenting a practical implication: allowing people to practice innovation in low-stakes environments.

#### **2.4.2. In Education**

A second group that is tackling innovation as a chaotic system is the Delft School of Industrial Design in the Netherlands. In [58], the authors present two separate models for innovation: one simple, and one chaotic. These models were created by spending 30 years exploring existing literature and observing and collecting data on small- and medium-sized enterprises (SMEs) that were participating in their Delft Design School. The chaotic model consisted of four stages: product development, strategy formulation, design brief formulation, and product launch/use. These four steps occur in no particular order. In more recent versions of this four-stage model, the authors connect the 4 quadrants with a heart because "leadership, culture, emotion, motivation, risk-taking, and passion are the true ingredients of innovative behavior". During their observation of SMEs, they found that most spent at least some time in each of the 4 areas, but in varying order and duration. The model aims to represent the fact that innovation is abstract and can be approached from

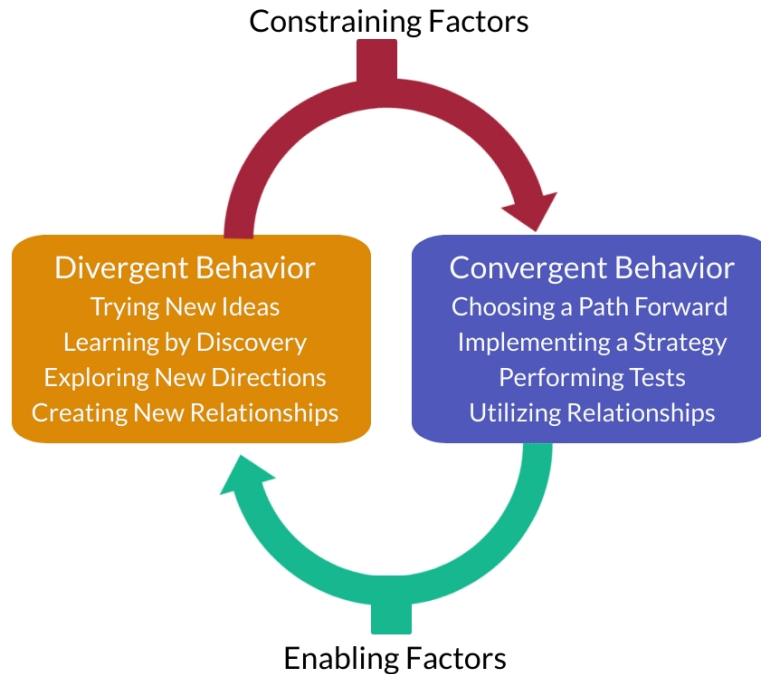


Figure 2.5. Cycle of innovation presented in [5]. Divergent behaviors lead to the introduction of constraining factors. These factors require behavior to converge, which then leads to enabling factors. The enabling factors lead to new ideas and options, going back to divergent behavior.

multiple different strategies. However, the authors also used their experience to create a 26-item process model. This model does have a sequence, but some items can occur in parallel as shown by the model. The authors end by arguing that the best way to teach innovation is to teach both models: one that is concrete and detailed, and another that is abstract and flexible [58]. Students may need a suggestion for a next step, but they also should not be confined to a recipe. This study is an appropriate one to end with because it brings us right back to the simple domain. No domain is better than the others, and all can provide insight to those trying to learn how to be innovators or those trying to support innovators.

## 2.5. Discussion and Relevance

Researchers studying the innovation process can learn from each of the domains: the simple, complicated, complex, and chaotic. The simple domain allows us to communicate, share, and teach new concepts. When teaching students about engineering, instructors often start with the engineering design process; it breaks a problem into step-by-step directions that teach the fundamental components of an engineering problem. By layering on the complicated domain, researchers

are able to explore beyond the “obvious” of the simple domain and identify new and meaningful information. By shifting into the complex domain, highly interactive and dynamic relationships can be explored, allowing for findings to be transferred across contexts. Rather than researching how people solve a specific problem or approach a specific prompt, it opens the door to exploring how people approach creating new and innovative solutions. Finally, the chaotic domain embraces that research in innovation (and education in general) is never perfect, but that doesn’t mean that researching these things is not useful. As discussed in [5], there is no way to predict what innovations will be successful and what ones will not, but there are ways to increase that probability of success. Similarly, instructors can teach the same content in the same way from year to year but get different results each time. Even though there is variability, using best teaching practices still increases the chances of student success.

When choosing a domain lens, it is important to consider our own thoughts about education and social systems, the type of environment that we are working in, the data that we have access to, the existing relationships within the system, and the goals/outcomes of the work. However, choosing a lens does not mean ignoring the others; Ryan Baker and Dragan Gasevic, two of the top researchers in the field of educational data mining and learning analytics, argue that not only do we need researchers working in the various paradigms, but we also need collaboration across each. For example, those working with a reductionist lens (complicated) provide scalability and often greater algorithm performance, whereas those working with an ontological lens (complex) are able to identify dynamic relationships and give data meaning [59]. This literature from each of the domains helps guide the choosing of an appropriate domain lens, methods, and metrics of success for the dissertation work.

### **2.5.1. Choosing a Domain Lens**

Innovation consists of interactions between stakeholders, creators, and industries. Because these interactions are dynamic and unpredictable, it lives on the left side of the Cynefin framework (the unordered side). There is some level of chaos involved in innovation as seen in [56], but by bringing it into the classroom environment, we can put bounds on it to move it into the complex domain. Students are working within the “rules” of the class including the timeline, assessment form, scope, etc. For me, the complex domain is a “sweet spot” for both researching and teaching innovation. Working in the chaotic domain suggests serious challenges for research because of the

lack of any identifiable relationships, and it can be too high-risk or daunting for students and educators. The complicated domain, on the other hand, oversimplifies the complex relationships that researchers are exploring, and it suggests to students that there is a finite number of correct ways to innovate. From a research standpoint, the complex domain places bounds on the system that allow us to assume that there are relationships to be found and improvements to be made, but it also still recognizes that there is no perfect recipe for innovation. From an education standpoint, these bounds allow students to practice innovation in a lowstakes environment (as suggested by [5]), but they still get experience working in a space where there isn't an answer in the back of the book. Therefore, this dissertation will primarily use a lens of complexity.

### **2.5.2. Choosing Appropriate Methods**

When Snowden presented the Cynefin framework, he also presented best practices for acting in each domain. For the complex domain, the best practice is “probe, sense, respond” [25]. Learning analytics methods allow us to collect large amounts of data (probe), analyze in real time (sense), and determine courses of action or new places to place a probe (respond). As mentioned in [41], learning analytics allows for the measuring and monitoring of complex problems because it has the “replicability and methodological rigor” of quantitative methods while allowing for diverse and illustrative pieces of data seen in qualitative methods.

However, most educational data mining methods are most appropriate for work in the complicated domain. Although previous work by the author that used LA/EDM to model innovation [22, 23] viewed innovation as complex, the methods aligned more with the complicated domain and were more similar to “traditional” educational data mining methods. The input of the algorithms that were created was raw student text, meaning all words and phrases were treated equally. This leads to challenges when trying to compare students from one group who are doing a research project with the goal of publishing at an academic conference and students from another group who are working on device development with the goal of submitting an invention disclosure; much of their vocabulary doesn't overlap, and the original methods had no way of aligning activities between the groups if they were not using the same words to describe these activities. Therefore, this dissertation uses other learning analytics methods to work within the complex.

When working in the complex domain, data rarely comes in a form that can be fed directly into an algorithm. When working on open-ended problems, there are virtually infinite pathways



that can be taken, meaning there is little meaning in the log data alone. As seen in many of the studies in the *Innovation as Complex* section, qualitative frameworks can help transform this log data into interpretable features that provide context without losing their ability to illustrate complex tasks.

### **2.5.3. Choosing Appropriate Metrics of Success**

The overarching goal of learning analytics work is to better support learning. Although quantitative results are important to identify trends and ensure these trends are meaningful, the bigger question is how these results can be used to support student learning. Therefore, any models created should balance two things: quantitative performance metrics (e.g. accuracy, F1 score) and meaningfulness. Using a neural network, for example, could increase the model's quantitative performance, but it also is a "black box" model, meaning it is not interpretable. In order to use the information to better support future iterations of the course, features should be able to be extracted and algorithms should be interpretable, and the work should be analyzed and interpreted in the context of implications for engineering education. Even though the bounds around the work were designed to keep it in the complex domain, chaos is still inevitable, meaning perfect accuracy is not realistic. However, that does not mean that the "probe, sense, respond" method is not useful. Van de Ven and team [5] remind us that we cannot perfectly control the innovation process, but we can learn how to increase the odds of success.

### **2.5.4. Filling the Gap**

To increase the odds of success in innovation, there should be more opportunities to practice innovation in low-stakes settings [5], and the Innovation-Based Learning model allows for that to happen. However, in order to improve and scale this model, best practices for teaching and learning innovation must be developed. The key gap that this dissertation addresses is the creation of research methods for studying innovation in educational settings. Currently, researchers in LA/EDM have studied components of innovation, but this work can be better defined as engineering design research because students are not identifying a unique gap and creating a solution with real-world value. Researchers in industry have studied "real innovation", but this work is not scalable or directly applicable to educational settings. This work bridges the gap between these two bodies of research by creating a qualitative framework for innovation in the classroom and implementing learning analytics methods to allow for scalable analysis of student innovation.

More broadly, it contributes to the body of research in learning analytics (and beyond) that is making the shift from complicated to complex. The community of researchers in this area has identified the potential to combine rich qualitative research with scalable and consistent quantitative research to better understand how *all* students grow and learn (not just the majority or the average student). Although this dissertation work is unique in that it looks at using learning analytics methods to understand student innovation, it also will become a part of an emerging body of work that aims to understand the complexities of student learning rather than simplify them.

## 3. DEVELOPMENT OF THE INNOVATION-BASED LEARNING FRAMEWORK

### 3.1. Introduction

The first of the three goals of this dissertation work is to identify or create an appropriate framework for categorizing student actions in IBL. As seen in the literature, most LA/EDM applications working in the complex domain have some way of taking raw data and categorizing it in a way that gives the algorithms “context”. Therefore, the rest of the dissertation work relies on the identification or creation of such a categorization system; Chapter 4 uses the framework to improve existing LA/EDM models in IBL, and Chapter 5 introduces new models only made possible through the existence of such a framework. Because the framework is so central to the work, it is imperative that the framework accurately and illustratively represents the IBL data and can realistically be implemented into LA/EDM work. Thus, this chapter will detail the development and assessment of the framework and a classifier model that automatically groups student text into the categories of the framework.

First, the chapter will provide a background about LA/EDM and frameworks; it will share further motivation for a framework, information about existing frameworks that were explored, and how frameworks and learning analytics mutually support each other. Next, it will describe the methods, results, and analysis for RQ1A which aimed to identify or create an appropriate framework. Then, it will describe the methods, results, and analysis for RQ1B which aimed to assess the possibility of creating a text classifier that automatically groups student text into the categories of the given framework. Finally, it will combine the findings from these two questions to share implications for teaching and research.

---

Some material in this chapter was drawn directly from [60], a publication co-authored by Lauren Singelmann, Enrique Alvarez Vazquez, Ellen Swartz, Ryan Striker, Mary Pearson, Stanley Shie Ng, and Dan Ewert. Lauren Singelmann drafted and revised all versions of this chapter. Other authors served as reviewers of the content and contributed to the interrater reliability testing.

## **3.2. Background**

The background shares motivation for creating a framework, other related frameworks (learning taxonomies, complex problem solving, self-regulated learning, the engineering design process, and diverging and converging behaviors), and the benefits of combining the framework with LA/EDM tools.

### **3.2.1. Motivation for Framework**

Although there is no existing framework that integrates innovation and learning in the engineering classroom, there are other existing qualitative frameworks that strengthen both teaching and research. In education, learning frameworks can help instructors develop course goals, outcomes, and assessments. They also can give learners a unifying vocabulary and a way to practice metacognition and reflect on their own learning process. In research, frameworks can bridge the gap between illustrative qualitative work and objective quantitative work. This mixed methods approach can address the complexities of learning and individual students while still providing reliable measurements across groups and over time [41].

Previous work to explore innovation in engineering classrooms could also benefit from such a framework. Studies have been done using learning analytics to explore what differentiates successful students in the IBL model, but this work was not able to compare students at the system or process level [22]. With a unifying framework, students could be compared at specific stages of the innovation and learning processes. This could lead to earlier and more specific interventions for struggling students.

### **3.2.2. Other Related Frameworks**

In order to create a framework that integrates innovation and learning, five areas of literature were explored: learning taxonomies, complex problem solving, self-regulated learning, the engineering design process, and divergent and convergent actions. Figure 3.1 illustrates each of these areas of literature coming together to create the final framework. Each of these areas has its own frameworks that group actions into categories. Information about existing frameworks in each of these areas is presented below. In addition, studies that have successfully used these frameworks are also presented, illustrating how qualitative frameworks can transform complex data into illustrative and meaningful results.

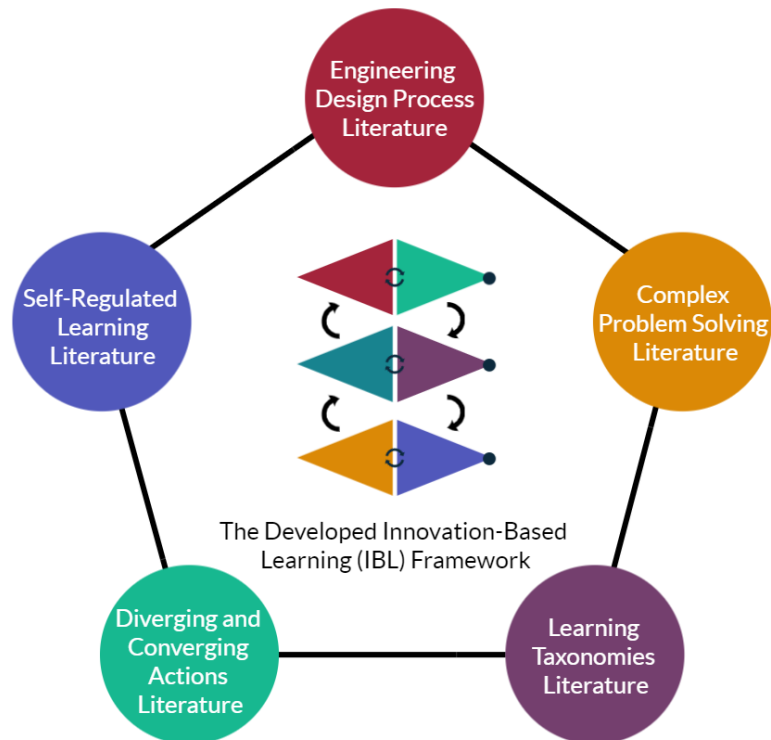


Figure 3.1. In order to develop the Innovation-Based Learning framework, five areas of literature were explored. Existing frameworks from each of these areas were applied to the data in order to determine strengths and weaknesses, and these five areas were combined to create the final framework.

### 3.2.2.1. Learning Taxonomies

Learning taxonomies are frameworks created to scaffold various actions and behaviors in the learning process [61]. The most commonly known learning taxonomy is Bloom's Taxonomy, but other taxonomies include Webb's Depth of Knowledge, the Structure of Observed Learning Outcomes (SOLO) taxonomy, Fink's taxonomy, and a handful of variations of Bloom's Taxonomy (e.g. Bloom's Revised Taxonomy shown in Figure 3.2 [62]). The stages of each taxonomy vary, but most are presented as a hierarchy; lower level learning builds to higher level learning.

Learning taxonomies are helpful frameworks because they can be used to help instructors build course materials that are well scaffolded and assessments that properly align. For example, questions from exams in both Electronics [63] and Computer Science [64] were mapped to Bloom's Taxonomy in order to better understand what level of learning the exams were assessing. This research can help instructors develop more appropriate assessments based off of course goals.

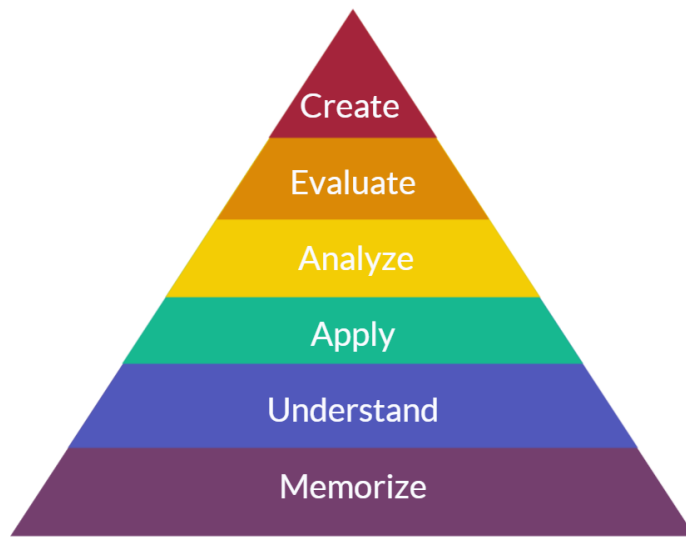


Figure 3.2. Bloom's Revised Taxonomy from [62]

Mapping student actions to learning taxonomies can also be done at a much larger scale by using learning analytics tools that use machine learning on educational data. For example, the Cognitive Operation Framework for Analytics (COPA) automatically maps the language used in course learning objectives, assessments, etc. to the various stages of Bloom's Revised Taxonomy. The results can then be shared with instructors in order to help them ensure that their goals, outcomes, and assessments are well aligned [65].

#### **3.2.2.2. Complex Problem Solving**

Although there are many definitions of complex problem solving, a general definition is the ability to overcome barriers between a given state and a goal state. The problem is considered complex because these barriers are usually unfamiliar to the problem solver and changing over time [42]. In order to better understand how problem solvers approach complex tasks, researchers have created a theoretical framework (Figure 3.3) to illustrate the components of complex problem solving. Not only does it include the task itself, but also the problem solver and the environment. The problem solver component considers prior knowledge, experience, and affect; the environment component considers any group interactions, feedback, or expectations from others [42]. Therefore, various studies in complex problem solving have focused not only on the completion of a task, but

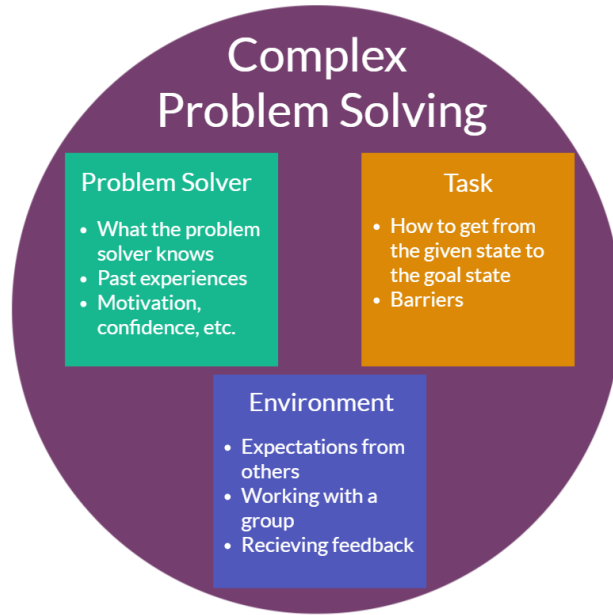


Figure 3.3. Framework for complex problem solving from [42]

also the background of the problem solver, affect, and the environment [66, 67]. The use of this framework strengthens work in this area by encouraging the consideration of variables in all areas.

### 3.2.2.3. Self-Regulated Learning

Self-regulated learning is the act of monitoring and directing one’s learning, including strategies, information, and self [68]. Multiple frameworks for self-regulated learning exist (see [69]), each with its own empirical evidence to support it. One of the most common frameworks for self-regulated learning consists of three components: forethought, performance, and self-regulation (shown in Figure 3.4). Forethought consists of planning out how the learner will approach the problem or take in information, performance consists of actually completing the task and maintaining focus and motivation, and self-regulation consists of reflecting on what was learned and assessing how well the task was completed and what strategies were successful [70].

By mapping student actions to the stages of self-regulated learning, information can be gained about how students guide and support their own learning. For example, this model has been used to explore how engineering students monitor behavior in design tasks in a 3D design and simulation platform by mapping clickstream data to the stages of the self-regulated learning framework. The clickstream data on its own is lacking the context needed to understand how students monitor their learning, but when context is introduced through the self-regulated learning

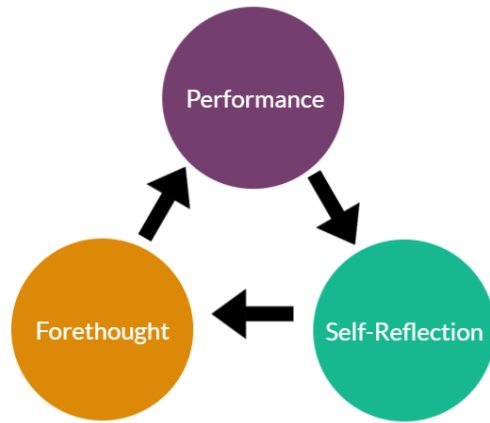


Figure 3.4. Cyclical framework for self-regulated learning from [70]

framework, algorithms were then able to cluster students into illustrative categories: reflective-oriented, adaptive, and minimally self-regulated learners [71].

#### 3.2.2.4. Engineering Design Process

The engineering design process has dozens of variations, but all consist of actions related to solving a problem. Although each of the variations has different names and stages, many consist of similar overarching categories. One review found 23 different versions of the engineering design process and created six stages that best represent all models: establishing a need, analysis of task (which focuses on the function of the innovation), conceptual design (which focuses on the behavior of the innovation), embodiment design (which focuses on the structure of the innovation), detailed design, and implementation [26]. These steps are shown in Figure 3.5.

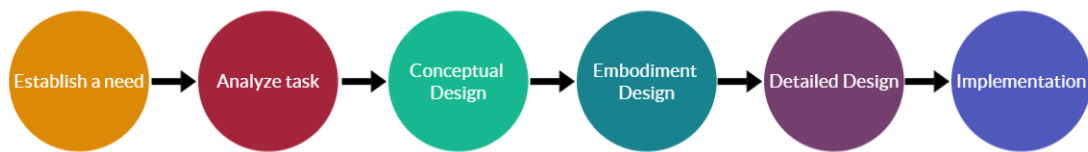


Figure 3.5. One example of the engineering design process from [26]

The engineering design process is a valuable framework for both assessment and research purposes. For assessment, understanding and applying the stages of the engineering design process



has been a proposed metric when evaluating the quality of K-12 engineering education activities [28]. For research, the engineering design process has been used as a framework for a variety of learning analytics studies. The researchers map student actions (often from clickstream data) to the various stages of their chosen version of the engineering design process. Because there are countless combinations of clicks in these virtual environments, a qualitative framework allows researchers to simplify their data without losing the ability to illustrate student behavior. For example, [33] looks at clickstream data in a 3D design and simulation platform and maps strings of actions to stages of the engineering design process (e.g. adding a wall maps to the construction phase, and running an energy test maps to the testing phase). Similarly, [34] uses the same platform to explore how students conduct experiments when doing an engineering design task in a 3D design and simulation platform. By mapping the clickstream data to stages of the engineering design process, usable quantitative features are created that can be used as inputs for classification, clustering, sequence analysis, or other machine learning tasks.

#### **3.2.2.5. Diverging and Converging Behaviors**

The final area of literature is divergent and convergent behaviors. Divergent behaviors involve exploring multiple ideas, whereas convergent behaviors involve honing in on a specific problem or solution. Divergent and convergent behaviors appear in a variety of contexts. For example, one offshoot of the engineering design process, the Double Diamond Model for Design (shown in Figure 3.6) consists of two diamonds, both made up of a diverging action and a converging action [72]. In order to create a solution, first a problem must be discovered and defined; then a solution must be developed and delivered. Discover and develop are considered diverging activities, meaning different options are being explored. Define and deliver are considered converging activities, meaning solutions are being chosen.

Similarly, one group that studies innovation in industry settings maps innovation as a cycle of divergent behavior, constraining factors, convergent behavior, and enabling factors. Ideas may arise which leads to divergent behavior (e.g. learning new things, building relationships, trying out new ideas) that then leads to constraining factors (i.e. realizing that there are obstacles that eliminate some of the options). This causes a need to shift to convergent behavior (e.g. conducting tests, narrowing ideas, and implementing specific strategies). These convergent behaviors can lead to enabling factors that lead to new ideas, cycling back to divergent behaviors. This model is

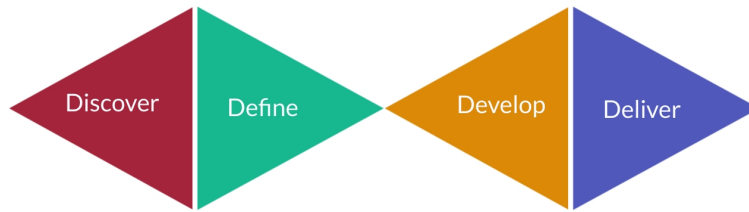


Figure 3.6. The Double Diamond for Engineering Design from [72]

called the Cycle of Divergent and Convergent Behavior in Innovation, and it is depicted in Figure 3.7 [5] (and was previously presented in Section 2.4 *Literature Review: Innovation as Chaotic*). The creation of this framework for innovation allows the researchers to categorize their qualitative observations, leading to data that can then be analyzed quantitatively to better understand how companies approach the innovation process. As seen in the other literature areas, the use of the qualitative framework translates complex data into features that can be analyzed and interpreted without losing its illustrative components, showing the need for development of strong frameworks.

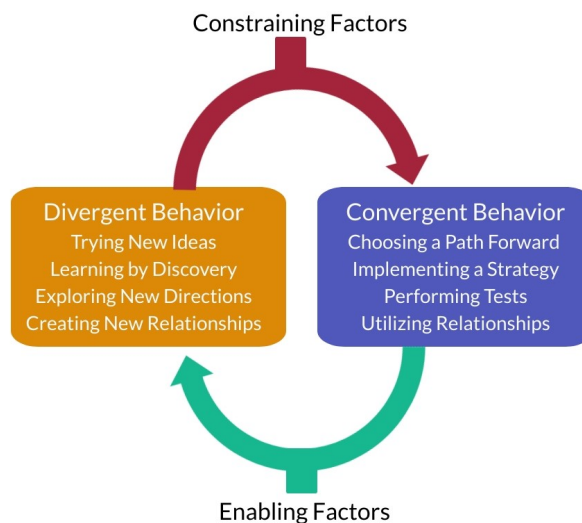


Figure 3.7. The Cycle of Divergent and Convergent Behavior in Innovation from [5]

### **3.2.3. Leveraging the Combination of a Framework and LA/EDM Tools**

After the framework is developed (RQ1A), the second step will be creating a text classifier with the goal of automatically categorizing student text into the framework categories (RQ1B). Combining the qualitative framework and LA/EDM tools has a variety of benefits, and this combination has shown promise in other similar lines of research. For example, one framework for complex and open-ended learning tasks that closely relates to IBL is constructionism, or the process of enabling learners to create artifacts [73]. Although the complexities of these open-ended learning environments offer new challenges for LA/EDM work, researchers have also identified four main benefits for using LA/EDM tasks to understand learning with a constructionist lens: 1) LA/EDM does not require researchers to abandon qualitative analysis, 2) LA/EDM analyses methods can complement existing methods to provide new insights, 3) LA/EDM can support methodological rigor and replicability, and 4) LA/EDM can be used to provide data to students and instructors in real time [41]. For these reasons, the IBL context may benefit from further establishing LA/EDM methods.

Qualitative work is more illustrative and holistic, but the trade-offs can include time and replicability. A machine learning classifier can group hundreds of pieces of student text in seconds, allowing for real-time analysis. In addition, using a classifier guarantees that any given piece of text is classified the same way every time. Although categorizing actions into framework categories will always involve some level of subjectivity and human judgement, the use of a classifier can ensure equal comparisons across all students and groups. If successful, the combination of the framework and the text classifier will allow for new methods that consider the complexities of innovation and allow for comparison of students across a variety of project types.

## **3.3. Research Question 1A: Development of a Framework**

RQ1A asked: Are there existing frameworks that are appropriate for categorizing Innovation-Based Learning data? If not, what is an appropriate framework?

### **3.3.1. Methods**

This section will define the alternate templates strategy that was used to create the IBL framework and the assessment process to determine interrater reliability of the framework.

### 3.3.1.1. Alternate Templates Strategy

In order to identify an appropriate model for the data, the alternate templates strategy was used, which is common in qualitative research dealing with complex process data [74]. Alternate templates strategy consists of qualitatively reviewing the data and the literature, finding and/or creating appropriate models, assessing how well that model fits the data, and continuing to reiterate that process until a final framework is selected. Each of the five areas of literature mentioned in the background were explored: learning taxonomies, complex problem solving, self-regulated learning, engineering design process, and converging and diverging actions. The ways that each of these models do and do not fit the data are presented in the *Results and Analysis* section.

### 3.3.1.2. Assessment of the Framework

A qualitative framework should balance simplicity, generalizability, and accuracy [74]. In order to assess the framework, the research team balanced the number of categories, how well it fit all students from both cohorts, and how well the data fit the framework. To maintain consistent categorization, a code book was created that gave definitions and examples of each framework category. To quantitatively assess how well the framework fits the data at the token level, interrater reliability was measured. Two rounds of interrater reliability assessment were conducted. First, two members of the instructional team categorized a subset of 853 tokens, and percent agreement and Cohen's Kappa were then calculated to assess for interrater reliability.

The observed percent agreement is calculated  $P_o$  (Equation 3.1). By also calculating the probability of chance agreement  $P_e$  (Equation 3.3), Cohen's Kappa can be calculated (Equation 3.4).

These calculations use a  $k$  by  $k$  confusion matrix ( $k = 7$  in the case of the framework categorization), in which an element  $f_{ij}$  defines the number of cases that the first rater assigned a particular item to category  $i$  and the second to  $j$ . So,  $f_{jj}$  is the number of agreements for category  $j$ . Then (from [75]):

$$P_o = \frac{1}{N} \sum_{j=1}^k f_{jj} \quad (3.1)$$

$$r_i = \sum_{j=1}^k f_{ij} \forall i \text{ and } c_j = \sum_{i=1}^k f_{ij} \forall j \quad (3.2)$$

$$P_e = \frac{1}{N^2} \sum_{i=1}^k r_i c_i \quad (3.3)$$

where  $P_o$  the observed percent agreement,  $r_i$  and  $c_j$  the row and column totals for category  $i$  and  $j$ , and  $P_e$  the expected proportion of agreement.

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \quad (3.4)$$

A confusion matrix was then created to determine where discrepancies were occurring. These discrepancies were then analyzed and grouped into categories to further refine the categories and make recommendations for future token creation. These error categories were then added to the code book to help raters differentiate between categories for common discrepancies, and the definitions and examples for each category were further refined where appropriate.

The second round of interrater reliability was conducted with the entire set of 2,116 tokens and among more members of the instructional team. For each of the tokens, the lead rater was assigned as the primary rater and two of the six other members of the instructional team were assigned as the secondary or tertiary rater. Cohen's Kappa and percent agreement were calculated between the primary rater and the secondary rater for all tokens, and another confusion matrix was created to show discrepancies. In order to decide final classes for all tokens, the tertiary rater or direct language from the codebook were used to break ties.

### 3.3.2. Results and Analysis

This section will assess how existing frameworks align with the IBL data, describe the final created framework, and share results from the two inter-rater reliability tests.

#### 3.3.2.1. Comparison to Existing Frameworks

In order to create the final framework, each of the existing areas of literature were explored to assess how well they fit the IBL data; explanations of the affordances and limitations of using each of the five areas of literature is presented below.

Learning Taxonomies: Learning taxonomies can be a great introduction to what behaviors are incentivized in IBL because the taxonomies illustrate the difference between lower level learning (e.g. memorization and understanding) and higher level learning (e.g. analysis and synthesis). However, the hierarchical structure of most learning taxonomies does not lend itself well to our

data. When trying to explain student learning in the context of any of the learning taxonomies, the taxonomies fail in being able to explain the nonlinear pathways that students take. For example, Bloom's taxonomy (which is commonly used in engineering fields to measure ABET requirements [76]) is set up in a way that suggests you are constantly moving up the pyramid from lower level learning to higher level learning. This hierarchy fails to illustrate that students who are creating and innovating must often return to the lower levels of the pyramid when ideas or prototypes do not work. This does not mean that students did not meet the objectives of the lower levels and were ill prepared to tackle the higher levels; it is simply part of the process.

Complex Problem Solving: Complex problem solving frameworks align nicely with the course because they include information about the problem solver, the task, and the environment – all of which play a role in both innovation and learning. In the IBL data, many tokens are related to the student being part of the class and of the group. Hence, the final IBL framework will include an “environment” category as inspired by the literature on complex problem solving. However, complex problem solving assumes that the problem has already been stated, so it does not account for the problem identification tasks that students complete in IBL.

Self-Regulated Learning: Students in the course clearly participate in self-regulated learning processes during the semester. Creating tokens aligns with forethought, completing tokens aligns with performance, and updating tokens aligns with self-reflection. However, the actions being defined by each token do not map to the categories in self-regulated learning frameworks. In addition, these frameworks fail to address the outcomes of the innovation and learning processes.

Engineering Design Process: Variations of the Engineering Design Process illustrate many of the actions that students are completing during IBL, but it misses out on the learning component. For example, it assumes that students know how to choose the best option and design it. In addition, most versions are presented with a specific order of steps. Even though some illustrate the cyclical nature of design, none fully illustrate the nonlinear processes that the data shows that the students took.

Diverging and Converging Behaviors: The Double Diamond for Design and the Cycle of Divergent and Convergent Behavior in Innovation both illustrate the idea of transitioning between divergence (e.g. trying new ideas, learning new things) and convergence (e.g. choosing a specific path, implementing a specific strategy). This idea of convergence and divergence appears in the

IBL data; students write about content and skills that they are exploring, and they write about actions they are taking. However, both models do not directly address the creation of impact. In addition, the Double Diamond for Design model suggests that innovation is a linear process where you start at stage 1 and end at stage 4. The Cycle of Divergent and Convergent Behavior in Innovation eliminates this linear process model, but it does not focus on the outcomes of the process.

### 3.3.2.2. Description of Final Framework

The developed IBL framework (shown in Figure 3.8) implements the concept of convergent and divergent behaviors but extends it by creating three diamonds, each with a converging category, a diverging category, and an end product.

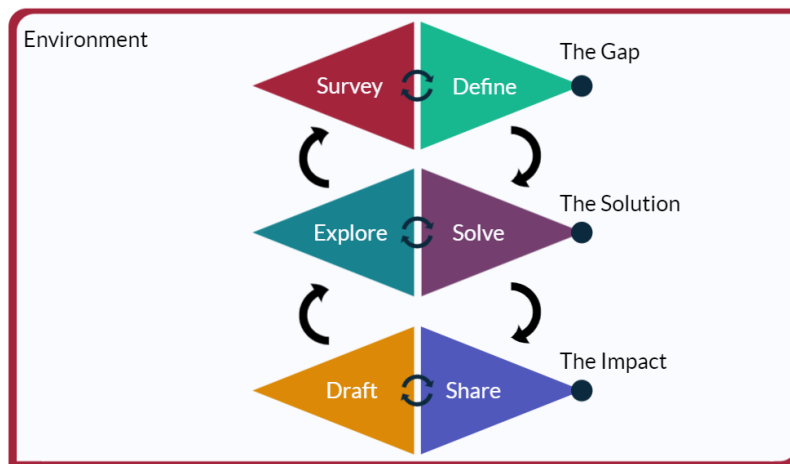


Figure 3.8. The developed IBL framework integrates engineering innovation and learning in engineering education. The framework consists of three diamonds, each consisting of a diverging activity and a converging activity that produce an output. Diamond 1 consists of *surveying* and *defining* the problem to develop a gap. Diamond 2 consists of *exploring* and *solving* to create an innovative solution. Diamond 3 consists of *drafting* and *sharing*, leading to impact of the innovation. The final category is *environment*, and it covers any other activities related to being a member of a group or class.

The first diamond consists of *survey* and *define*, with the end product being the identification of a gap that needs to be filled (e.g. a question that still remains in the research or a space in the market that has not been filled). *Surveying* consists of learning about the problem space (including concepts learned in the course), and *defining* consists of narrowing the scope of the project to hone in on a specific project goal. The second diamond consists of *explore* and *solve*, with the ends

product being the solution to the gap. *Exploring* consists of learning the tools or concepts that will be necessary to create a solution, and *solving* consists of making design choices and building a solution to the problem. The third diamond consists of *draft* and *share*, with the end product being the creation of impact. Students create impact by publishing their work, submitting invention disclosures, or sharing elsewhere. *Drafting* consists of learning how to navigate this process and growing in their communication skills, and *sharing* consists of choosing an outlet and creating the impact. Finally, the *environment* covers any activity that does not directly lead to developing a gap, a solution, or impact. The *environment* can include activities like team meetings, completing activities for class, or doing any housekeeping items.

Table 3.1 shows the list of framework categories and an example token from the dataset that falls under each of the categories.

Table 3.1. Example token for each of the framework categories

<b>Category</b>	<b>Example</b>
Survey	Understand the cardiovascular system and the applications of tissue engineering
Define	Narrow research topic to inform class
Explore	Learning about ECG signals and feature extraction
Solve	Apply preprocessing functions to data collected with team’s hardware and revise functions as necessary
Draft	Obtain feedback from class for revisions of symposium poster
Share	Final paper to submit for provisional patent
Environment	Group meetings and presentations to evaluate progress and provide updates

Overall, the framework is built on three main tenets:

- An innovation consists of three components: an existing gap, a new and unique solution, and developed impact.
- The process of innovation consists of iterating between converging and diverging behaviors (explore options, make decisions, repeat).



- There is not a linear pathway from start to end; the problem gap, solution, and intended impact may be refined and adjusted over time.

### 3.3.2.3. Assessment of Final Framework

For the first round of assessing interrater reliability between the two raters, the raters had 71.2% agreement and a Cohen’s Kappa of 0.664, which is considered moderate agreement [75]. The confusion matrix comparing the raters’ categorizations is shown in Figure 3.9.

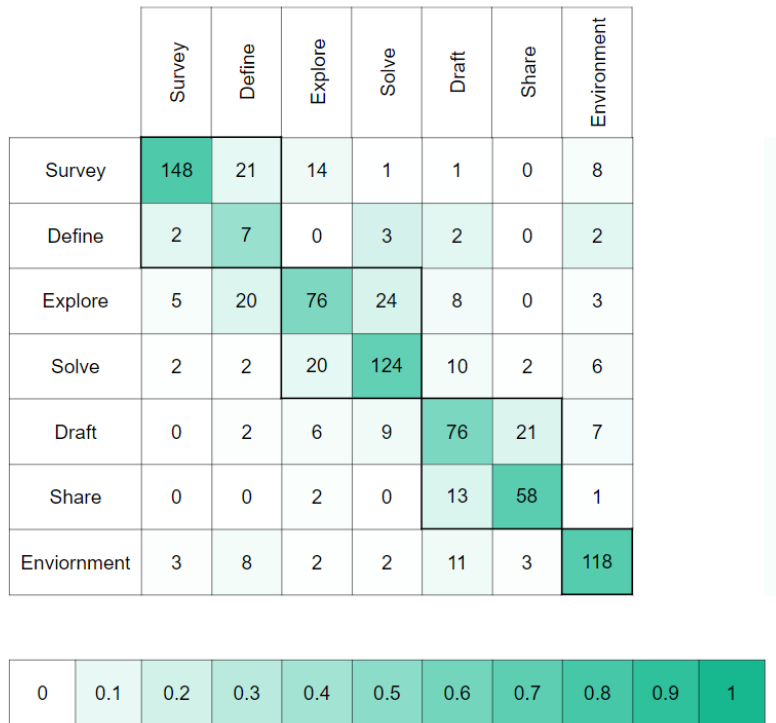


Figure 3.9. Confusion matrix between the two independent raters. Boxes on the diagonal represent agreement between the two raters. The color relates to the ratio of tokens sorted into a certain category compared to the total number of tokens in that category (as determined by Rater 1). For example, Rater 1 put 193 tokens into the *survey* category, and Rater 2 agreed for 148 of those tokens, giving a recall value of 0.767. The matrix also identifies areas of discrepancy; these are the brighter green values that are not on the diagonal.

The discrepancies between the two raters were then analyzed and grouped into error categories: social learning, level of impact, practice vs. solution, content vs. solution, draft vs. submit, specific learning, unclear language, and websites. These error types, their descriptions, examples, and the number of occurrences of each type is shown in Table 3.2.

Table 3.2. Categorization errors

Error Type	Error Description	Example	Occurrences
Social Learning	Social activities often fell under multiple categories. If someone is teaching a classmate about a tool they learned, the token could fall under <i>explore</i> , <i>share</i> , or <i>environment</i> .	“Teach a lab coworker how to use iWorks hardware/software”	50
Level of Impact	Some tokens provide “secondary impact”, meaning they are not the main impact deliverable. These tokens led to discrepancies between <i>solve</i> and <i>share</i> because they do solve a problem and create impact, but they are not part of the team’s main solution or impact.	“Create public Jupyter notebook with tutorial on how to preprocess ECG signals.”	33
Practice vs. Solution	The boundary between converging and diverging was sometimes unclear when it came to developing the solution. When using sample data or code, the student is still exploring possible options, but they have honed in on a more specific potential solution.	“Perform bivariate analysis with sample data.”	33
Content vs. Solution	Some activities fell somewhere between learning content that is related to the problem and learning skills that are needed to develop a solution, leading to discrepancies between <i>survey</i> and <i>explore</i> .	“Understand how genes play an important role in the cardiovascular system.”	19
Draft vs. Submit	Some tokens combined both the drafting component and the submitting component of producing impact.	“Manuscript preparation/abstract submission”	25
Specific Learning	If students hone in on a specific thing they need to learn in order to refine the gap they are solving, the learning could arguably fall under <i>survey</i> or <i>define</i> .	“Understand the cardiovascular system and the applications of tissue engineering.”	23
Unclear Language	Unclear language occurred when token titles and descriptions did not match, or when the tokens were vaguely written.	“Gain access to data. Abstract development.”	22
Websites	If the deliverable is a website, this led to confusion about how to classify components of the website. Because the website was designed to be both a solution and an impact, it was hard to separate components of the website into categories.	“Create a working website.”	19

Some errors could be reduced by more clearly defining the framework categories, but others are caused by the way that the specific token was written. These discrepancies led to takeaways for both token raters and token writers.

Takeaways for those categorizing tokens include:

- If a student is “just trying something”, err on the “diverging” side. If students are not being specific about the content or skills they are developing, or if they use words like “practicing”, or “looking into”, they have not transitioned to the converging stage yet. This addresses the “Practice vs. solution” and “Specific Learning” error types.
- Teaching another student can be part of the learning process. For example, teaching a student about Simulink would still be considered part of learning about Simulink, putting the token into the *explore* category. This addresses the “Social Learning” error type.
- If students are learning about a process or skill, or if they are exploring the market space, *explore* is the most appropriate category. This addresses the “Content vs. Solution” error type.

Takeaways for those writing tokens include:

- Proofread your tokens to ensure that you are communicating clearly and that your title and description match. This addresses the “Unclear Language” error type.
- Clarify the purpose of your token. Are you learning more about the problem, or are you learning material in an attempt to assess possible solutions? This addresses the “Content vs. Solution” error type.
- Clarify the outcome of your token. Is it still in progress, or is it mostly finalized? This addresses the “Practice vs. Solution” and “Draft vs. Submit” error types.
- Carefully define your problem, solution, and form of impact. This addresses the “Level of Impact” and “Websites” error types.

The framework showed great promise for balancing simplicity, generalizability, and accuracy in the first round, but these tips can even further improve performance by reducing the number of tokens that are unclear or fall into more than one category.

For the second round of assessing interrater reliability, percent agreement was 69.1% and Cohen’s Kappa was 0.627, which was also deemed moderate agreement [75]. The confusion matrix comparing the interrater reliability for the second round can be show in Figure 3.10.

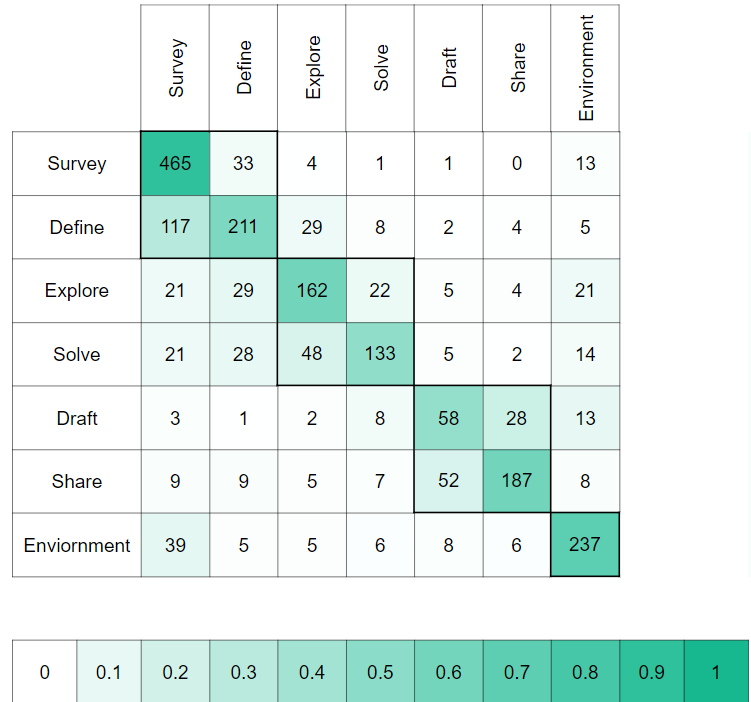


Figure 3.10. Confusion matrix between the lead rater and the team of secondary raters. Boxes on the diagonal represent agreement between the two raters. The color relates to the ratio of tokens sorted into a certain category compared to the total number of tokens in that category (as determined by Rater 1). For example, Rater 1 put 517 tokens into the *survey* category, and Rater 2 agreed for 465 of those tokens, giving a recall value of 0.90. The matrix also identifies areas of discrepancy; these are the brighter green values that are not on the diagonal.

The slight decrease in interrater reliability from Round 1 was most likely due to more raters being introduced. When looking at specific secondary raters, the interrater reliability with the primary rater ranged from 61.2% to 77.1% (or  $\kappa = 0.519$  to  $\kappa = 0.727$ ). Although these results still show moderate agreement (especially with seven categories), it is imperative to reduce the inconsistencies in categorization. Because many of the errors involved missing a handful of specific rules, it was hypothesized that a machine learning classifier could categorize the tokens with greater consistency. Although there is no way to eliminate the subjectivity of the categorizations, it is

possible to ensure that similar tokens are always categorized in the same way. This consistency is especially important if further analyses will rely on these categorizations.

### **3.4. Research Question 1B: Feasibility of Automatic Framework Classification**

RQ1B asked: Can a classification model be used to sort student text into the categories of the IBL framework with greater consistency than a human rater?

#### **3.4.1. Methods**

To answer RQ1B, four models were developed and assessed, and success was measured by comparing the agreement levels of the algorithm and the lead human rater versus the lead human rater and the other human raters. The four models included support vector machine (SVM) with a linear kernel, K-Nearest Neighbors (KNN), random forest (RF) and logistic regression (LR). These models were chosen because they allow for quantitative data (not just categorical), they are appropriate for small sample sizes, and they allow for feature extraction so an instructor or researcher could have an intuitive understanding of how the model is sorting tokens into categories [77]. The sample set included 2,116 tokens that had been labeled by the research team (step 1 in Figure 3.11). For each token, the title and description were combined into a single document, and these documents were then tokenized to create unigram TFIDF (term frequency-inverse document frequency) matrices (step 2 in Figure 3.11). Code was written in Python [78], and the sci-kit learn module [79] was used for training and testing models.

Five-fold cross-validation was used for assessment, meaning the tokens were split into five folds of equal size (step 3 in Figure 3.11), and five models were created — each tested on a different fold and trained with the other four folds (step 4 in Figure 3.11). The performance of the five models is then averaged to get final performance indicators, and the actual and predicted classes are counted and compiled into a confusion matrix (step 5 in Figure 3.11). Five-fold cross-validation was used because it is representative of the whole dataset but also gives indication of how well the model will perform on future data. Percent agreement and Cohen's Kappa were used as performance metrics to compare to the results seen in RQ1A.

To test for statistical significance, 100 random testing and training sets were created (80% train, 20% test). All four algorithms were tested on each of the 100 cases, and a Wilcoxon Signed Rank test was performed on two algorithms at a time. This statistical test assumes paired data and does not require any assumptions about the distribution of the data [80]. However, it should be

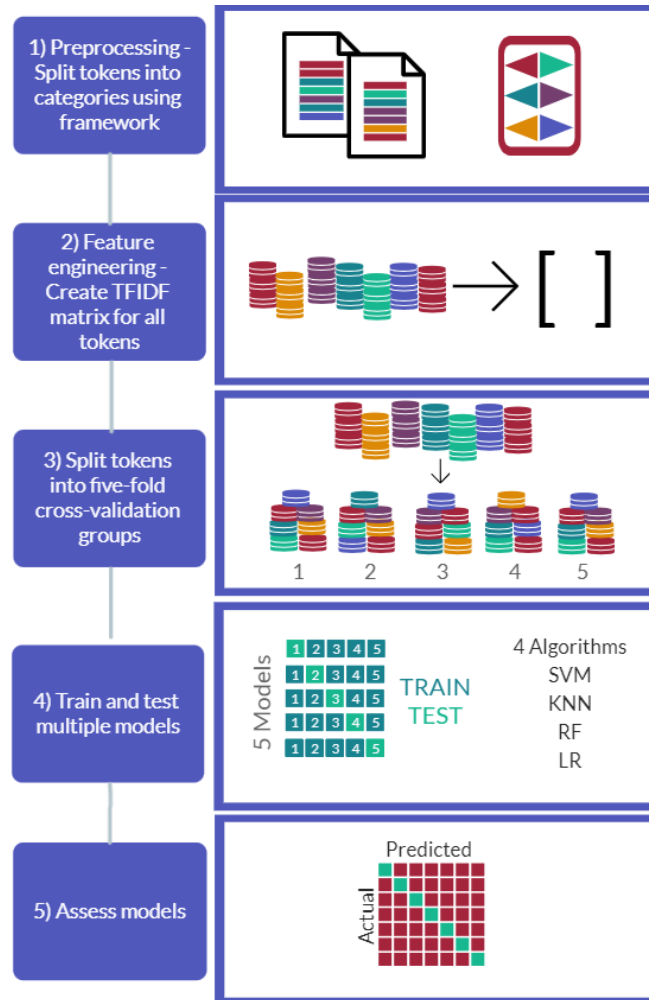


Figure 3.11. Graphical representation of methods for RQ1B. First, preprocessing is performed by the research team by splitting all tokens into the categories of the framework. Next, a TFIDF (term frequency-inverse document frequency) matrix created where each row represents a token, each column represents a word, and each entry represents how frequently a word appears in that token compared to all other tokens. To perform five-fold cross-validation, all tokens are split into five folds with tokens from each category balanced among the folds. For each of the five folds, a model is created and trained using four folds and tested on the fifth fold. The results from each fold are combined to get overall performance metrics. Four different algorithms were assessed: support vector machine (SVM), k-nearest neighbors (KNN), random forest (RF), and logistic regression (LR). Finally, a confusion matrix is created for each of the models to determine where discrepancies occurred, and percent agreement and Cohen’s Kappa are calculated to assess performance.

noted that this test assumes independence of samples, and our 100 cases have overlapping training and testing data. Although a 5x2 test (five different two-fold cross-validation tests) reduces this problem of having overlapping training data because there are only two folds [81], our sample size is too small and the tokens are too variable to hold out half of the data for testing each time.

Therefore, the Wilcoxon Signed Rank test was deemed to be most appropriate for our data, but still has some limitations.

### 3.4.2. Results and Analysis

Percent agreement and Cohen’s Kappa were calculated to measure agreement between the lead human rater and the team of other human raters, as well as between the lead human rater and the four different classification algorithms. These performance metrics can be seen in Table 3.3.

Table 3.3. Percent agreement and Cohen’s Kappa of team of human raters and classifier models

Model	% Agreement	Cohen’s Kappa
Team of Human Raters	0.691	0.627
Support Vector Machine Model	<b>0.799*</b>	<b>0.760**</b>
k Nearest Neighbors Model	0.705	0.643
Random Forest Model	0.674	0.608
Logistic Regression Model	0.793	0.752

\* denotes a statistically significant result ( $p < 0.05$ )

\*\* denotes an extremely statistically significant result ( $p < 0.005$ )

The models created using SVM, KNN, and LR all had higher agreement with the human rater than the team of human raters, and the SVM performed the strongest with a percent agreement of 79.9% and Kappa of 0.760 (compared to the human raters who had a percent agreement of 69.1% and a Kappa of 0.627). The performance difference between the SVM and the other models was also deemed to be statistically significant using the Wilcoxon Ranked sum test. Compared to the second best model (the LR), the SVM had better percent agreement than the LR ( $p < 0.05$ ) and a better Kappa ( $p < 0.005$ ).

To determine where discrepancies were occurring, the confusion matrix from the original interrater reliability test (first presented in Figure 3.10) was compared with the confusion matrix from the highest performing classification model (SVM). These confusion matrices are shown in Figure 3.12. For both the team of human raters and the optimized classification algorithm, discrepancies were most likely to occur between the triangles of the framework that make up the same diamond (e.g. between *explore* and *solve*). However, the classification algorithm was able to reduce other discrepancies that were caused by human errors such as missing one of the rules in the framework codebook.

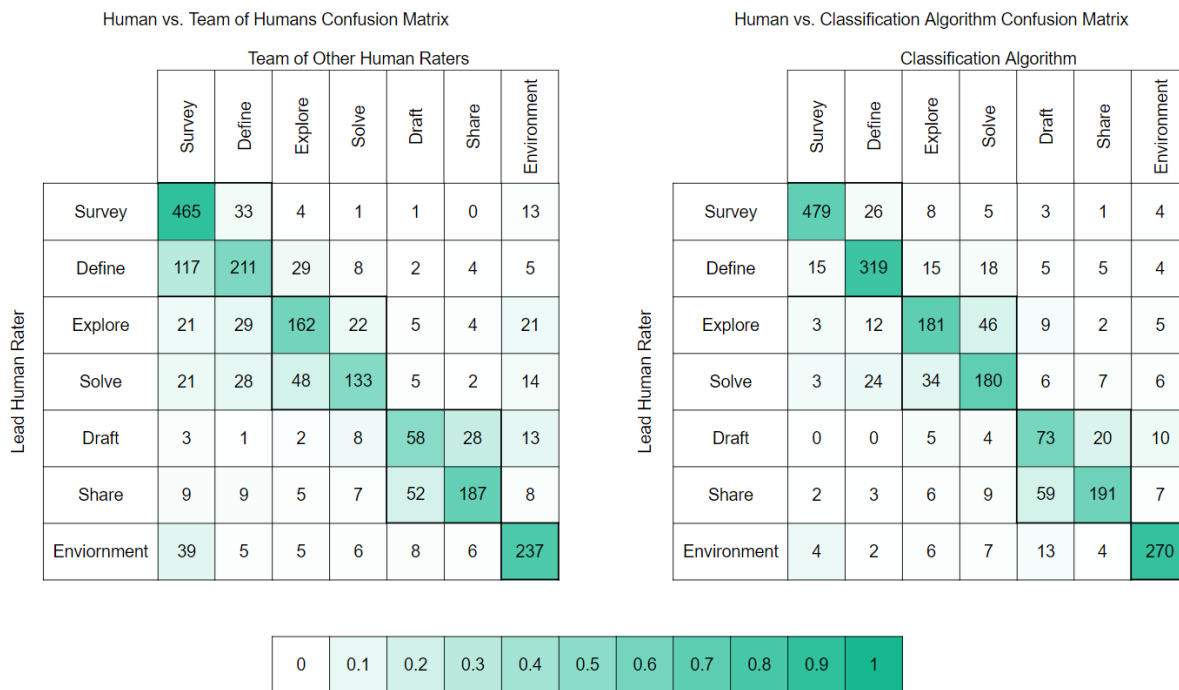


Figure 3.12. Confusion matrices that show discrepancies between a lead human rater and the team of other human raters (left) and a lead human rater and the optimized classification algorithm (right). Entries are colored by comparing the value of the cell to the sum of the values in that row, which corresponds to precision. Preliminary results for RQ1A showed that human raters had sufficient agreement when categorizing text into the framework categories. These new results show that a classifier can also be trained to categorize objectives with higher reliability, consistency, and speed. In addition, it shows that discrepancies are most common between the two triangles that make up a diamond (e.g. *Explore* and *Solve*.)

Achieving perfect agreement is not a realistic performance goal, but using a classification algorithm showed higher agreement and consistency, a key reason that learning analytics is becoming increasingly important in work involving complex learning environments [41]. This consistency is imperative when framework classification is the first step in completing other learning analytics tasks using this data. The algorithm is trained with classifications made subjectively using human judgement, so it is not possible to eliminate subjectivity by implementing a classification algorithm. However, it does ensure that tokens are always categorized in the same way.

These results show a positive result for RQ1B; a machine learning classifier can be trained to group pieces of student text into the categories of the IBL framework with greater consistency (and speed) than a human rater. Although it is important to continue to evaluate and validate the



results of the classification model, these results show that real-time learning analytics work can be performed on the IBL data.

### **3.5. Implications for Teaching**

The creation of the framework offers a variety of benefits for teaching and learning innovation that help develop shared language, expectations, and assessment.

The framework creates shared language between students and instructors to be able to communicate progress, talk about their challenges, and ask questions. Not only does it give them language to talk about the gap, solution, and impact, but it also helps them explain where they are getting stuck. If they are struggling to transition from a diverging action to a converging action, other students or instructors may be able to provide advice about choosing a path forward. If they are struggling to transition from converging to diverging, other students or instructors can help brainstorm new ideas or provide advice about where to explore next.

Similarly, the framework helps set shared expectations. Before the creation of the framework, success in the course was defined by “the creation of external value”. The framework breaks this idea into sub-components: the gap, solution, and impact, and it shows that learning and developing is involved in all three of those components. Students gain a better understanding of what makes something innovative, and they have a clearer idea of how they will be assessed.

Finally, the framework helps both students and instructors monitor progress. Students and instructors can track how much work is being done towards each of the three components: gap, solution, and impact, as well as how much work is being done looking at new possibilities (diverging) and moving forward with specific ideas (converging).

The presented implications above are driven mostly by the first two tenets of the framework: 1) innovation has three components (gap, solution, impact), and 2) each of those components consists of cycles of converging and diverging actions. However, it should be noted that questions still remain about how to help students embrace tenet 3: the nonlinear and complexity of innovation. Instructors should be careful to present the framework as a guideline, not a recipe to be followed.

### **3.6. Implications for Research**

The creation of the framework also leads to various impacts on engineering education and learning analytics research.

It is the first published framework that integrates technical learning and engineering innovation, and it helps further differentiate innovation-based learning from project-based learning and engineering design activities. The use of alternate templates strategy allowed for a qualitative deep dive guided by actual student behavior that better illustrated these differences. The steps of the engineering design process, for example, did not appropriately map to the actions that students were completing in IBL. IBL places equal focus on gap, solution, and impact, whereas the engineering design process focuses overwhelmingly on solution development. Similarly, the engineering design process places little importance on the learning process compared to the IBL framework. Innovation requires learning content and skills in order to drive an innovation forward, and the engineering design process generally assumes that these skills have already been developed. In addition, the student behavior showed how complex and nonlinear the process of innovation is; because it involves identifying and shaping new problems and solutions, actions in one category might lead to ideas for actions in another category; rather than moving through the steps in the process, innovation creates new and unique pathways.

In addition, the framework provides a lens for researchers that allows for meaningful comparisons across all students and projects. Categorizing student actions can be challenging because different students are working on different components of different projects, but this framework groups actions in a way that transcends project type. Previous categorization schemes were either too broad (e.g. adding a token, editing a token, deleting a token), or too specific (e.g. literature review, experimental design, professional development). If the categorization scheme has too broad of categories, information is lost. If the categorization scheme has too specific of categories, it becomes increasingly challenging to make comparisons across students because they are working in different categories from each other. The framework's seven categories serve as a middle ground; the categories are illustrative, but apply to all projects and students.

Finally, the framework allows for extension of previously used learning analytics methods and implementation of new ones. For the classification and clustering approaches, rather than treating all text equally, the framework allows algorithms to consider the words used in the context of the framework category. In addition, because student actions are now sorted into categories that are both illustrative and generalizable, new methods can be used. For example, time- and

trajectory-based analyses methods can now be used because there is an appropriate number of action types that are consistent across all projects.

### 3.7. Summary

This work detailed the development and assessment of a framework for categorizing student actions in IBL and a classifier that automatically sorts student text into the categories of the framework.

To answer RQ1A, the alternate templates strategy was used to assess how well existing frameworks fit the collected IBL data. These frameworks came from literature about learning taxonomies, complex problem solving, self-regulated learning, the engineering design process, and diverging and converging behaviors. Although no existing framework by itself sufficiently covered the variety of activities that students complete in IBL, these areas of literature were combined to create the final IBL framework with seven categories: *survey*, *define*, *explore*, *solve*, *draft*, *share*, and *environment*.

To answer RQ1B, a variety of text classification models were trained and tested to assess the viability of sorting student text into the categories of the framework automatically. Because there was both human error and human differences in judgement when there were multiple human raters, an automated classifier could support more consistent categorization of student text. There will always be some subjectivity while categorizing text, but the key metric is consistency; if this classification occurs at the beginning of the research workflow, it is imperative that the same type of action is put into the same category every time. The trained SVM model had the highest percent agreement and Cohen's Kappa, making it the strongest trained classifier. However, this does not mean that a researcher should rely only on using the automatic classifier. The automatic classifier is meant to be a tool for real-time analysis, but a researcher should verify the results before performing any final analysis of the IBL data. This researcher/algorithm partnership will allow the classifier to continue to improve as more labeled data becomes available, and it should increase the consistency of the classification process overall.

The creation of the IBL framework is key to the rest of the presented work; it allows for the extension of existing LA/EDM methods and the implementation of new ones.

## 4. EXTENSION OF CLASSIFICATION AND CLUSTERING IN IBL BY LEVERAGING THE IBL FRAMEWORK

### 4.1. Introduction

The second goal of the dissertation is to assess if and how the IBL framework can be used to improve previously developed classification and clustering models in IBL. The previous chapter detailed the development of the framework, and the next chapter will use the framework to implement new analysis methods that are more aligned with complex system modeling. Therefore, this chapter serves as an intermediate step; by combining the IBL framework with the existing methods that have already shown promise [24], we can assess if and how the IBL framework provides any added benefit. Although classification and clustering are more appropriate for work in the complicated domain, these results identify meaningful patterns and trends and assess the IBL framework in action.

First, the chapter will provide a background about LA/EDM work that has previously been done in IBL settings. Next, it will describe the methods, results, and analysis for RQ2A which aimed to determine if the implementation of IBL framework improves performance of classification models when compared to the original models designed and tested in [22]. Then, it will describe the methods, results, and analysis for RQ2B which aimed to determine which clusters emerge when student text is categorized using the IBL framework. Finally, it will combine the findings from these two questions to share implications for teaching and research.

### 4.2. Previous Learning Analytics/Educational Data Mining Work in IBL

According to [82], there are three overarching categories of methods for LA/EDM: prediction (predicting one pre-determined variable using other variables), structure discovery (finding structure within data without an a priori idea), and relationship mining (finding variables that are somehow correlated with each other). Prediction is usually considered a supervised approach, meaning the algorithm is trained to find patterns in labeled data. For example, classification is a su-

---

Some material in this chapter was drawn directly from [6], a publication co-authored by Lauren Singelmann and Dan Ewert. Lauren Singelmann drafted and revised all versions of this chapter. Dan Ewert served as a reviewer of the content.

ervised method because the algorithm is told which cases in the training data were low-performing and which were high-performing. The algorithm then finds common features that separate these two groups and uses those features to predict new cases. Structure discovery and relationship mining are usually unsupervised, meaning the algorithm is finding patterns within unlabeled data. Clustering is an example of an unsupervised method because it finds similar samples and groups them together without considering any existing groups. Previous LA/EDM work in IBL has been done in both prediction (classification) and structure discovery (clustering).

#### 4.2.1. Previous Classification Models

Previous work in [22] created classification models that were designed to predict student success. Two feature sets were assessed: 1) text features that came directly from the text that students were writing and 2) quantitative features (e.g. number of tokens, number of deleted tokens, etc.) Models trained with text features performed significantly higher than baseline performance, whereas the models trained with quantitative features performed no better than baseline. This suggests that *what* students write is more important than *how often* they are writing it. Ten-fold cross-validation was used when testing the models, which suggested that the highest performing models could be used to predict performance for future cohorts. However, because the input of these models is student text, success of the models largely depends on users using the same words in the same context. This leads to issues because words like ‘write’, ‘journal’, and ‘create’ can signal different things at different stages of the innovation process. For example, a student can be ‘writing’ notes about a ‘journal’ publication they read, or they can be ‘writing’ a publication to submit to a ‘journal’. The former is an important step in the process, but does not necessarily translate to student success. The latter demonstrates an example of creating a high external value deliverable. Although both use similar words, their meanings differ because of the context. Therefore, it was hypothesized that the classification algorithm could be improved by sorting student text into framework categories that give each piece of text more context. In addition, it was hypothesized that sorting student text into the framework categories could create a better performing quantitative classifier. Rather than using total numbers of tokens as features, this new quantitative classifier would count the number of student tokens in certain categories.

### 4.2.2. Previous Clustering Models

Previous clustering work in [23] used student text to group students with similar behavior. Four clusters emerged from the data, the text that differentiated between clusters was extracted, and the results were qualitatively explored in order to name each cluster. The four clusters included surface-level, surveyors, learners, and innovators. Student success was not an input of the model, but yet the unsupervised models differentiated between low- and high-performing students. Learners and innovators were significantly more likely to be high performers in the course, and surface-level and surveyors were significantly more likely to be low performers. These results allowed for analysis beyond a binary classification of low- and high-performing. For example, innovators and learners were both likely to be high-performing, but innovators focused on writing about actions they were doing whereas learners focused on writing about things they were learning.

### 4.3. Research Question 2A: Using the IBL Framework to Improve Classification

RQ2A asked: Does categorizing student text into framework categories improve the performance of a classifier model that separates between lower and higher performing students?

#### 4.3.1. Methods

To answer RQ2A, classification models were developed to differentiate between lower and higher performing students. In our model, higher performing was defined as contributing to a high external value deliverable. It should be noted that performance level does not directly align with the final grade in the course (which also considered performance on pillar concepts) or the quality of the innovation. For example, if a team creates a device that not end up being marketable, they can still create impact and value by sharing the unique insights gained about why their solution did *not* work. This distinction was made so students were safe to take risks, but were still encouraged to use their findings to create value.

Previous work in [22] determined that text classifiers could predict student success with adequate performance whereas quantitative classifiers could predict student performance only slightly better than random. However, in further experiments, neither text nor quantitative models performed sufficiently on data from new cohorts. Therefore, it was hypothesized that sorting student work into the framework categories would improve both the text and quantitative models because the framework considers the context of the language used and work completed.

The work in [22] used data from only one cohort, but this work looked at three cohorts, allowing for both intra- and inter-dataset performance to be assessed for both text and quantitative models. Intra-dataset performance refers to performance ‘within’ the dataset; in this case, data from all cohorts are combined and assessed using five-fold cross-validation. Because the dataset is relatively small, intra-dataset performance makes use of all data available while still avoiding overfitting by using cross-validation. Inter-dataset performance refers to performance ‘between’ datasets; in this case, data from the first two cohorts was used as training data, and the models were assessed using the third cohort as test data.

For each student, documents were created that consisted of all of that student’s tokens in a specific category (step 1 in Figure 4.1). To compare models with various feature sets, these documents were processed in four different ways (step 2 in Figure 4.1). For the “No Framework” model, the documents were all combined into one single document that was then tokenized into a single TFIDF matrix that did not account for framework category. For the “Framework” model, each of the seven documents for each student was tokenized separately into a TFIDF matrix, and these matrices were then concatenated. For example, if a student used the word ‘create’ in both *Define* and *Solve* tokens, the relative document frequency of ‘create’ in the *Define* category would be a separate feature than the relative document frequency of ‘create’ in the *Solve* category. Feature selection was then performed by calculating chi-2 values for each of the features. The optimized feature numbers were 300 for the ‘No Framework’ model and 350 for the ‘Framework’ model. For the ‘Category’ model, TFIDF matrices were created for each of the seven framework categories. Performance for these models were compared to determine which framework category most highly differentiated between lower and higher performing students. For each of the text classifiers, four models were tested: SVM, KNN, RF, and LR. Only linear models were tested due to the small data size. A support vector machine with a linear kernel had the best preliminary results in [22] and was consistently a top performer in this work, so results were only reported for this model. Finally, the ‘Quantitative’ model simply counted the number of tokens each student had in each category and created a feature matrix with the relative proportions.

All models were tested intra- and inter-set. Intra-set testing (cross-validation assessment) treated all students as one dataset, and this set was split into five folds for cross-validation. Inter-set testing (prediction assessment) separated the students by cohort (Step 3 in Figure 4.1). For the

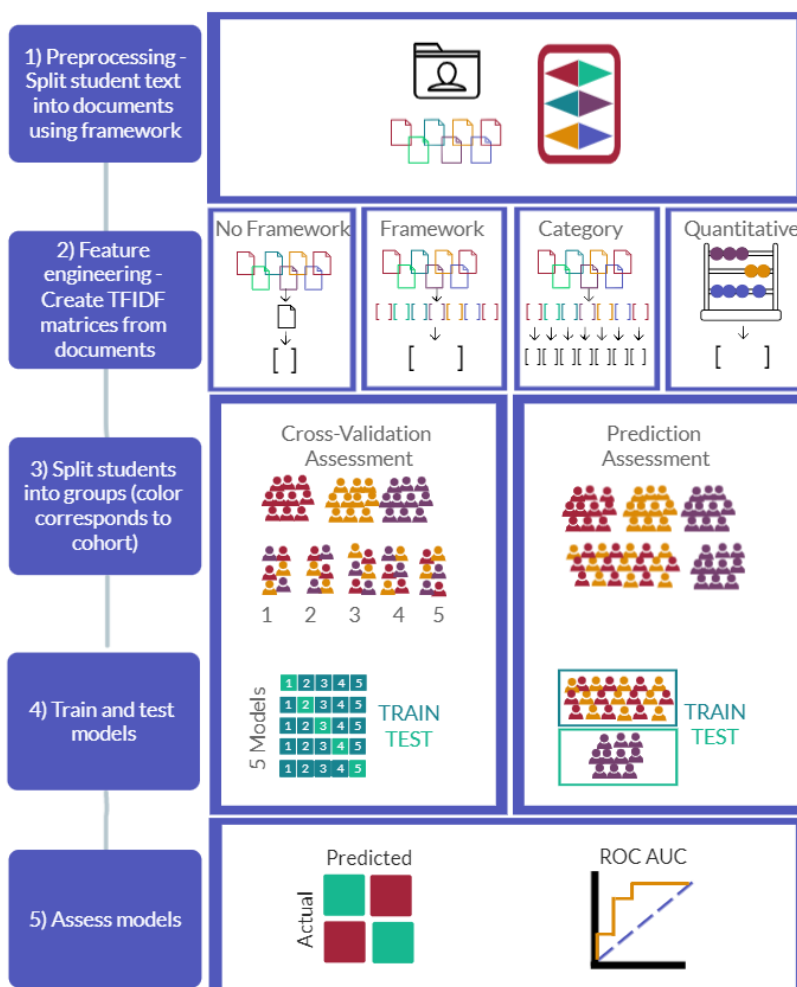


Figure 4.1. Graphical representation of methods for RQ2A. First preprocessing is performed by splitting student text into documents using the framework. Each document represents one students' text for one framework category. Next, four different methods for feature engineering are performed to create feature matrices. For the 'No Framework' model, all documents for each student are combined into a single document. These documents are then used to create a TFIDF matrix where each row represents a student and each column represents a word. For the 'Framework' model, each document is converted into its own TFIDF matrix, and these matrices are then concatenated to create a combined TFIDF matrix where each row represents a student and each column represents a word in a specific framework category. For the 'Category' models, the single category TFIDF matrices are all analyzed individually to determine how well text from each category differentiates high and low performance. For the 'Quantitative' model, the number of tokens in each category is counted, and a feature matrix is created with the relative proportions for each token type. Next, students are split into groups. For the cross-validation assessment, students are split into five folds with cohorts evenly represented in each fold. For the prediction assessment, Cohort 1 and 2 are put into the training group, and Cohort 3 is put into the test group. Classification models are then trained and tested, and the following performance metrics are calculated: accuracy, precision, recall, F1 score, and ROC AUC.



cross-validation assessment, five models were created where each fold was used as test data once. For the prediction assessment, Cohorts 1 and 2 were used as training data, and Cohort 3 was used as testing data (Step 4 in Figure 4.1).

To assess each of the classification models that aim to differentiate between low- and high-performing students, four performance metrics were calculated: accuracy, precision, recall, F1 score, and ROC AUC. In the context of this work, accuracy is the percentage of students who were correctly classified by the model (Equation 4.1) where TP, FP, FN, and TN correspond to the labeled confusion matrix in Figure 4.2. Precision is the percentage of students that were predicted to be low performance and were actually low performance (Equation 4.2). Recall is the percentage of students that were low performance and were predicted to be low performance (Equation 4.3). F1 score is the harmonic mean of precision and recall (Equation 4.4). ROC AUC is the area under the receiver operating characteristic curve, and this metric shows how well the two classes separate. The receiver operating characteristic curve plots false positive rate versus true positive rate at any given decision threshold. Rather than comparing predicted and actual class (as with the other metrics), ROC AUC ranks each item by the class probability. ROC AUC can be interpreted as a measure of the probability that a random positive sample has a higher ranking than a random negative sample. More information about calculating ROC AUC can be found in [83], and a worked example of calculating ROC AUC can be found in Appendix Section A.1.

Sample Confusion Matrix

		Actual Performance	
		Low	High
Predicted Performance	Low	TP	FP
	High	FN	TN

Figure 4.2. Sample confusion matrix defining true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) in the context of this study. Low performance was deemed to be the positive case because future work hopes to identify at-risk students.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4.1)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4.2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.3)$$

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FP} + \text{FN}} \quad (4.4)$$

Although it is important to look at all five performance metrics, accuracy, recall, and ROC AUC were deemed the most important. Because the classes are relatively balanced, accuracy provides an overall look at how well the algorithm is classifying students. Recall is also a helpful metric because it gives a measure of how many students who were actually at risk were predicted to be high risk. It is better to accidentally flag a high-performing student as being at risk than to not flag a student who is at risk, so recall was deemed a key performance metric over precision. ROC AUC is also a valuable metric because it takes into account the class probabilities. For example, if an instructor chooses to prioritize checking in with students that have the highest probability of being “low-performance”, a higher ROC AUC score would best support that strategy because it determines how well the classes separate using these probabilities.

For tests that were run intra-set (using five-fold cross-validation), the Wilcoxon Signed Rank test (previously described in Section 3.4.1) was performed to test for statistical significance. Because the inter-set tests aimed to understand performance for one specific case, (Cohort 1 and 2 as training data, Cohort 3 as testing data), no tests for statistical significance were performed. Instead, the inter-set results should be interpreted practically. They determine how well the model would have performed if it had been used during the third iteration of the course as a prediction tool.

Next, because linear models were used, features of importance were extracted from the optimized models. For the text models, these features are words that differentiated between lower- and higher-performing students. For the quantitative models, the features are categories where

students' numbers of tokens differentiated between lower- and higher-performing students. To extract these features, a chi-2 test was performed on each of the features to determine its weight. In the calculation for chi-2 (Equation 4.5),  $O_i$  represents the observed values for each feature and  $E_i$  represents the expected values for each feature.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4.5)$$

The coefficients of the linear models were extracted to determine if the feature was associated with high performance or low performance. For the sake of more easily interpreting the chi-2 value, the weight was reported as a negative value if it corresponded with a low performance and a positive value if it corresponded with high performance.

Finally, because there is some lack of certainty in categorizing students by performance, the final models were also tested for robustness to that type of variability. This robustness analysis uses a Monte Carlo simulation to determine how the performance levels change when various levels of randomness are injected into the class labels. A Monte Carlo simulation uses random processing to allow for stochastic modeling of complex algorithms/processes [84]. It is not realistic to mathematically determine how human errors in classification effect our final classifier performance. However, a Monte Carlo simulation lets us run thousands of different scenarios so we can see how various levels of randomness effect the outcomes. For each of the final text models (one without the text sorted into framework categories and one with the text sorted into framework categories), robustness was assessed at a variety of randomness levels (5%, 10%, and 20%). The algorithm is trained and tested using the methods detailed previously in the section, but different levels of randomness are injected into the student classes. For example, at the 5% randomness level, each student has a 5% chance of getting their label switched. Performance metrics are then reported for that sample. This process is run 10,000 times, leading to many different combinations of student classifications. The performance metrics for each of the 10,000 samples can then be visualized using a histogram plot.

### 4.3.2. Results and Analysis

#### 4.3.2.1. Text Classifiers Assessed Intra-Dataset

First, the performance of the text classifiers with and without the framework were compared both intra-dataset (or within a single dataset). Students from all cohorts were grouped into five folds, and each fold had students from all three cohorts. The intra-dataset performance using five-fold cross-validation is shown in Table 4.1.

Table 4.1. Performance metrics for text classifiers intra-set

Model	Accuracy	Precision	Recall	F1	ROC AUC
Random Classifier	0.492	0.476	0.323	0.385	0.500
Without framework	<b>0.701</b>	<b>0.625</b>	0.732	0.674	0.779
With framework	0.691	0.600	<b>0.805**</b>	<b>0.688*</b>	<b>0.806</b>

\* denotes a statistically significant result ( $p < 0.05$ )

\*\* denotes an extremely statistically significant result ( $p < 0.005$ )

The confusion matrices can be seen in Figure 4.3 where Figure 4.3a shows the results without sorting text into the IBL framework, and Figure 4.3b shows the results after sorting text into the IBL framework. Figure 4.4 shows the ROC AUC both without and with sorting text into the framework categories. The ROC AUC for each of the five folds is plotted, as well as the mean ROC AUC  $\pm 1$  standard deviation.

These results show little differentiation between the models trained with and without the framework; the differences between the two models was not statistically significant for accuracy, precision, and ROC AUC. The model with the framework did have slightly higher recall and F1 ( $p < 0.005$  and  $p < 0.05$ , respectively). Both models performed substantially better than a random classifier, which was expected because of the promising preliminary results from [22]. Although the model trained with the framework did not lead to substantial improvement in all performance categories, our hypothesis was supported; the model trained with the framework maintained the performance of the original model with some slight performance increases in recall and F1.

To better understand which words in which categories were most likely to differentiate lower performing and higher performing students, the top 150 features were extracted from the classifier and sorted in Figure 4.5 by IBL framework category. The font size of the word corresponds to

its chi-2 value, and the color corresponds to its association with higher performance or lower performance.

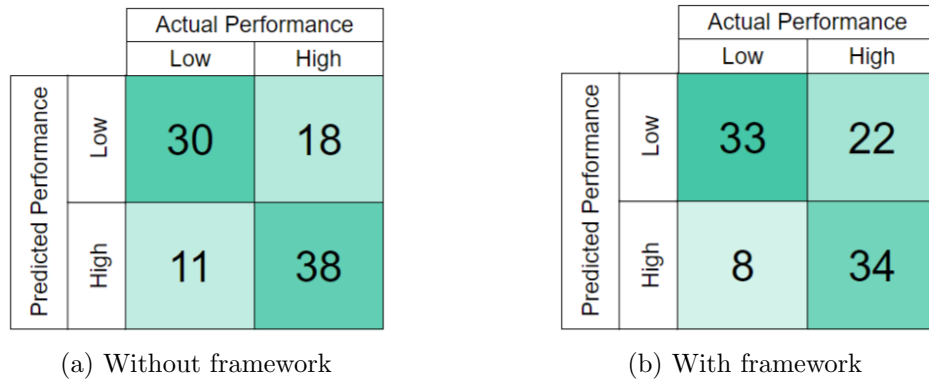


Figure 4.3. The confusion matrices for the text intra-set classification models assessed using five-fold cross-validation. Both (a) the model trained without sorting text into the IBL framework and (b) the model trained after sorting text into the IBL framework are able to differentiate between lower and higher performers at a level sufficiently higher than a random classifier. Entries are colored by comparing the value of the cell to the sum of the values in that column. This corresponds to recall (the ratio of students at a performance level that were also predicted to be at that performance level).

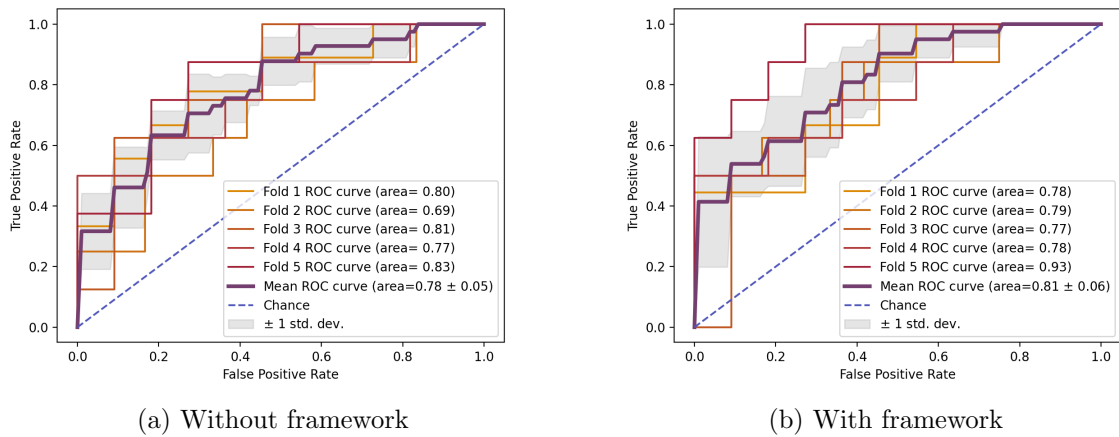


Figure 4.4. Receiver operating characteristic curve for the text classifiers trained intra-set (with five-fold cross-validation). (a) shows the model trained without sorting text into the IBL framework (originally designed and tested in [22]), and (b) shows the model trained by sorting text into the IBL framework (the new model). The curve for each fold is shown, along with the mean curve  $\pm 1$  standard deviation. As expected from the preliminary work, the original model without the framework maintains its strong performance on the larger dataset. The new model trained with the framework maintains this strong performance.

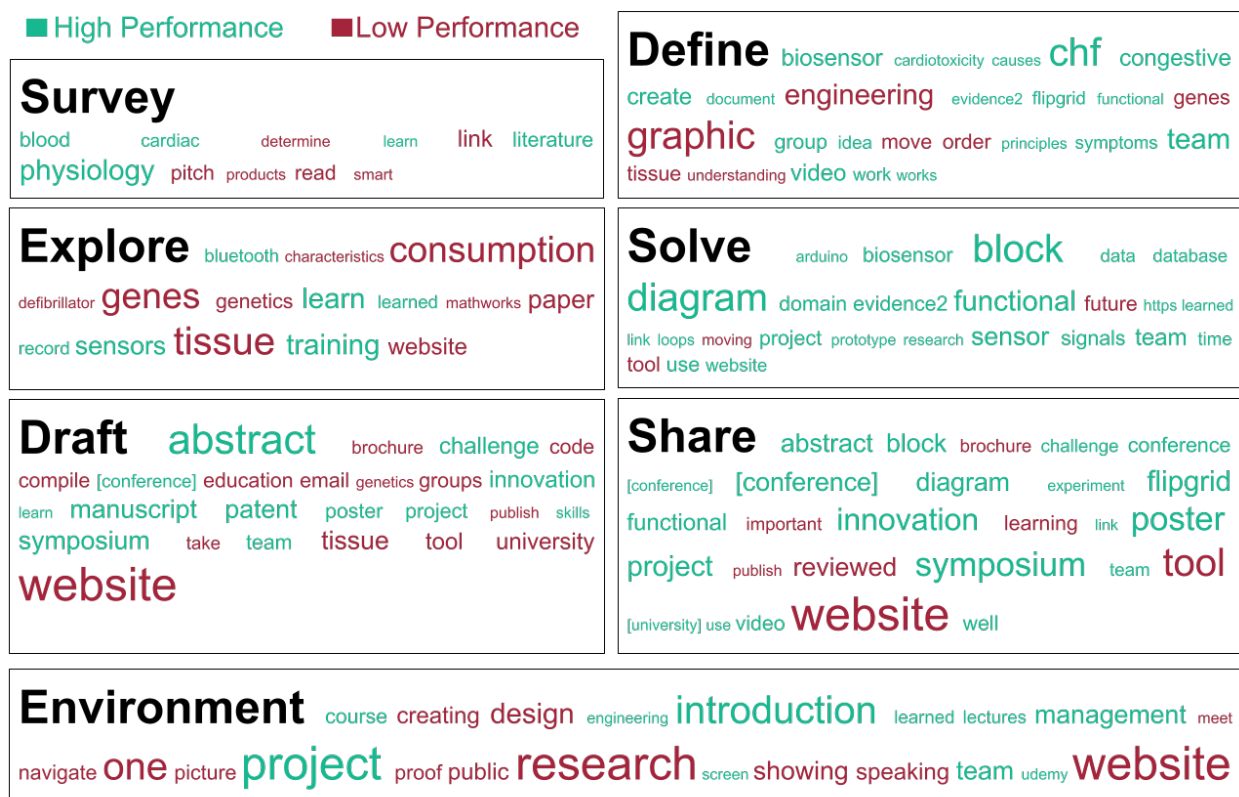


Figure 4.5. Top 150 features for the optimized text classifier that differentiate between lower-performing and higher-performing students. Words are grouped by the framework category that they fell into. Features in green more highly differentiated higher-performing students, and features in red more highly differentiated lower-performing students. The font size represents the relative chi-2 value; larger words were more highly differentiating than smaller words. Because each word in each category is its own feature, some words appear in more than one category. In fact, some words such as ‘learn[ed/ing]’ and ‘website’ appear in red in some categories and green in others. Specific conferences and universities were replaced with [conference] or [university].

Because each word in each category is its own feature, words can appear in more than one of the categories of the IBL framework. In fact, some words such as ‘learn[ed/ing]’ and ‘website’ were associated with high performance in some categories and low performance in other categories. We would expect a high-performing student to be reporting about their ‘learning’ during the top stages of the IBL process (e.g. *survey* or *explore*, and we see that ‘learn[ed/ing]’ was associated with high performance in these categories. However, during the *share* stage, we would expect students to be reporting on how they are creating value, not on what they are learning. The data supports this claim; ‘learning’ in the *share* category was associated with low performance. These differences

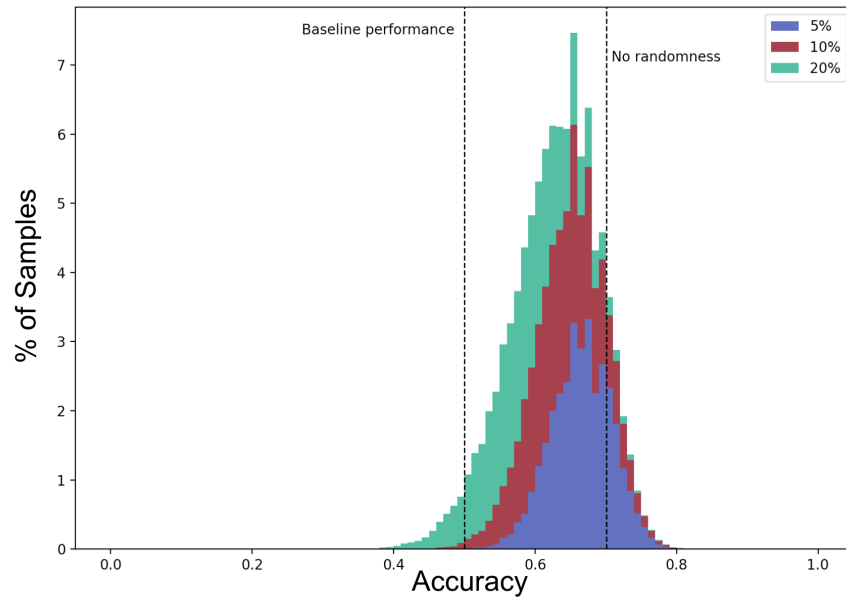
across categories suggests why the classifier with the framework might perform better than the model without; words are more highly differentiating when they are paired with a specific context – in this case, a category of the IBL framework.

Finally, to assess the robustness of each of the optimized text classifiers (with and without the framework), Monte Carlo simulations were run with various levels of randomness interjected in the student classes to simulate the possibility of the instructors incorrectly categorizing a student by performance. The accuracy results of the Monte Carlo simulation for the classifier without the framework are shown in Figure 4.6a, and the accuracy results for the classifier with the framework are shown in Figure 4.6b. For each of the models at each of the randomness levels, the probability that the accuracy is less than baseline is reported in Table 4.2. The average values for accuracy and all of the other performance metrics for each randomness level are listed in Table 4.3.

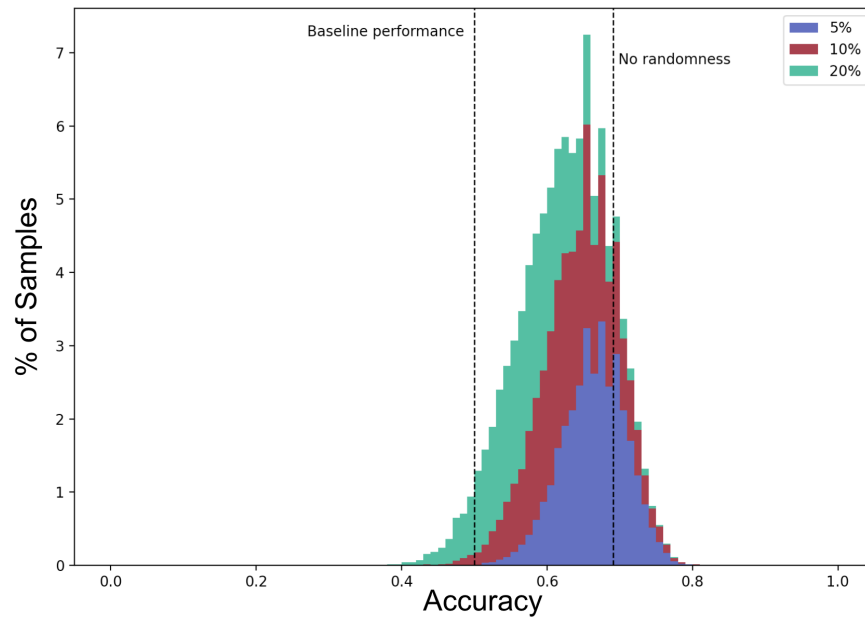
From these results, we see that injecting randomness into the model can lead to greater variability in performance. However, even with 20% randomness added, the probability that accuracy is greater than baseline is over 90% for both the models without and with the framework. It also should be noted that the models with no added randomness performed better than average in all categories. This shows that the original groupings of high-performing and low-performing students agreed upon by the instructors is more highly separable than most other possible groupings – suggesting that the instructors’ observations also aligned with differences in text written. If the observations did *not* align with differences in text written, we would expect that the performance of the model with no randomness would be of about average performance; about half of the samples would have higher performance than the original model, and about half of the samples would have lower performance than the original model.

Table 4.2. Probability that accuracy is above baseline at various randomness levels

<b>Model</b>	<b>Randomness p(Acc&gt;Baseline)</b>	
Without Framework	5%	99.98%
	10%	99.49%
	20%	91.28%
With Framework	5%	99.94%
	10%	98.97%
	20%	90.74%



(a) Without framework



(b) With framework

Figure 4.6. Monte Carlo results for the optimized text classifier (a) without the framework and (b) with the framework. 10,000 samples were run for each of the three randomness levels: 5%, 10%, and 20%. The results are plotted using a histogram with 100 bins ranging from accuracy of 0 to accuracy of 1. The y-axis represents the percentage of samples that fell into any given bin. Vertical lines representing the baseline performance and the performance of the classifier with no randomness are also plotted.



Table 4.3. Average performance metrics at various randomness levels

Model	Randomness	Accuracy	Precision	Recall	F1
Without Framework	0%	0.701	0.625	0.732	0.674
	5%	0.665	0.614	0.693	0.638
	10%	0.637	0.590	0.654	0.607
	20%	0.584	0.544	0.585	0.550
With Framework	0%	0.691	0.600	0.805	0.688
	5%	0.663	0.589	0.783	0.666
	10%	0.632	0.566	0.753	0.640
	20%	0.580	0.529	0.690	0.591

#### 4.3.2.2. Text Classifiers Assessed Inter-Dataset

Next, inter-dataset performance was assessed to determine how well the models perform across datasets. Because factors change from year to year, it is important to assess if it is possible to use data from old cohorts to predict performance for new cohorts. The inter-dataset performance of the classifier trained using Cohorts 1 and 2 and tested on Cohort 3 is shown in Table 4.4. The confusion matrices can be seen in Figure 4.7 where Figure 4.7a shows the results without sorting text into the IBL framework, and Figure 4.7b shows the results after sorting text into the IBL framework. Figure 4.8 shows the ROC AUC both without and with sorting text into the framework categories.

Table 4.4. Performance metrics for text classifiers inter-set

Model	Accuracy	Precision	Recall	F1	ROC AUC
Random Classifier	0.529	0.647	0.524	0.579	0.500
Without framework	0.441	<b>1.000</b>	0.095	0.174	0.640
With framework	<b>0.676</b>	0.778	<b>0.667</b>	<b>0.718</b>	<b>0.710</b>

These results show that the model with the framework was significantly better at predicting performance for a new cohort. The model without the framework identified 32 of the 34 students as high performance, causing low accuracy and recall. This could be improved by adjusting the decision threshold, but the relatively low ROC AUC score of 0.640 shows that this may make only marginal improvements. In addition, without knowing the final performance of the students, there is no way to know what the decision threshold should be. Therefore, an instructor may have

assumed that things were going very well in the class because so many students were predicted to be high-performance.

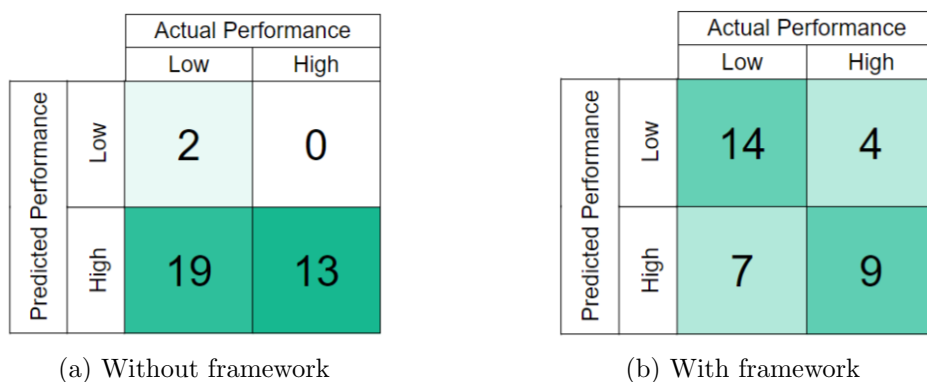


Figure 4.7. The confusion matrices for the text inter-set classification models trained using data from Cohorts 1 and 2 and tested on Cohort 3. (a) shows the confusion matrix for the model trained without sorting text into the IBL framework and (b) shows the model trained after sorting text into the IBL framework. Entries are colored by comparing the value of the cell to the sum of the values in that column. This corresponds to recall (the ratio of students at a performance level that were also predicted to be at that performance level). The new model trained with the framework had higher inter-set performance than the original model in [22] without the framework. From the confusion matrices, it is clear that a fault of the model trained without the framework is its prediction that most of the students are high-performing, whereas the model trained with the framework is more balanced.

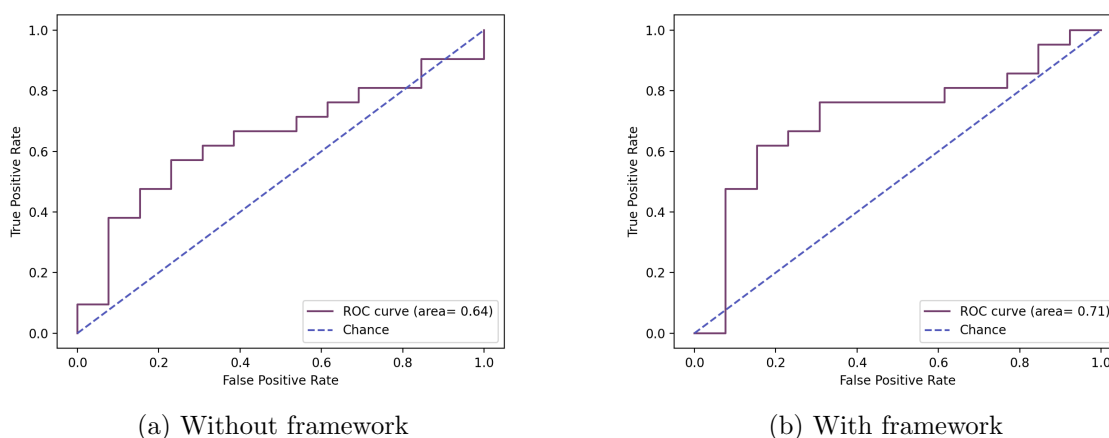


Figure 4.8. Receiver operating characteristic curve for the text classifiers trained inter-set (trained with Cohort 1 and 2 data and tested with Cohort 3 data). (a) shows the model trained without sorting text into the IBL framework (originally designed and tested in [22]), and (b) shows the model trained by sorting text into the IBL framework (the new model). The old model has some ability to differentiate between classes, but the new model shows improved performance.

One possible explanation for the high number of predicted high-performing students with the old model could be that students were doing the right types of tasks and submitting the right types of deliverables as encouraged by the instructors, but not in the right context. Tasks and deliverables look differently at each level of the framework, and the original text model cannot differentiate between these contexts. The model with the framework, on the other hand, can use the categorizations of student text to differentiate between contexts.

Because there are a variety of factors that change from semester to semester including instructors, groups, and projects, it is not surprising that inter-set performance is lower than intra-set performance. However, the results of the classifier using the IBL framework show promise for being able to continue to predict student performance for new cohorts. The strong ROC AUC scores also suggest that the probability results are sufficiently separating the classes.

#### 4.3.2.3. Single Category Classifiers Assessed Intra-Set

To determine which categories' text most strongly differentiates between lower and higher performing students, intra-set performance metrics were calculated for each of the individual framework categories. These results can be seen in Table 4.5. Text in the *survey*, *define*, and *explore* categories differentiated between lower and higher performing students no better than a random classifier. Text in the *solve* and *share* categories were more highly differentiating than the other categories, but using text from all categories combined still had the best performance. These results suggest that the words that students are writing in the *solve* and *share* categories are the most highly differentiating.

Table 4.5. Performance metrics for single category text classifiers intra-set

Model	Accuracy	Precision	Recall	F1	ROC AUC
Random Classifier	0.492	0.476	0.323	0.385	0.500
<i>Survey</i>	0.442	0.272	0.400	0.310	0.522
<i>Define</i>	0.621	0.530	0.956	0.681	0.524
<i>Explore</i>	0.650	0.585	0.761	0.647	0.523
<i>Solve</i>	0.629	0.536	0.853	0.658	0.739
<i>Draft</i>	0.577	0.594	0.747	0.549	0.608
<i>Share</i>	0.727	0.569	0.783	0.653	0.727
<i>Environment</i>	0.608	0.548	0.636	0.570	0.687
All categories	0.691	0.600	0.805	0.688	0.806

#### 4.3.2.4. Single Category Classifiers Assessed Inter-Set

Table 4.6 shows the performance metrics for the single category text classifiers inter-set. Interestingly, the *Explore* and *Solve* categories alone predicted performance for Cohort 3 better than combining all categories. However, because using all categories performed better on the intra-set cross-validation sets, this appears to be specific to the train/test split of Cohorts 1 and 2 versus Cohort 3.

Table 4.6. Performance metrics for single category text classifiers inter-set

Model	Accuracy	Precision	Recall	F1	ROC AUC
Random Classifier	0.529	0.647	0.524	0.579	0.500
<i>Survey</i>	0.382	0.5	0.048	0.087	0.667
<i>Define</i>	0.559	0.667	0.571	0.615	0.498
<i>Explore</i>	0.706	0.824	0.667	0.737	0.707
<i>Solve</i>	0.765	0.760	0.905	0.826	0.788
<i>Draft</i>	0.382	1.000	0.000	0.000	0.615
<i>Share</i>	0.676	0.667	0.952	0.784	0.570
<i>Environment</i>	0.382	0.500	0.095	0.160	0.606
All categories	0.676	0.778	0.667	0.718	0.710

#### 4.3.2.5. Quantitative Classifiers Assessed Intra-Set

Finally, the performance of the quantitative classifiers with and without the framework were compared both intra- and inter-dataset. The intra-dataset performance using five-fold cross-validation is shown in Table 4.7. The confusion matrices for the models without and with the framework are shown in Figure 4.9, and the ROC AUC curves are shown in Figure 4.10.

Table 4.7. Performance metrics for quantitative classifiers intra-set

Model	Accuracy	Precision	Recall	F1	ROC AUC
Random	0.492	0.476	0.323	0.385	0.500
W/o framework	0.536	0.464	0.634	0.536	0.491
W/ framework	<b>0.691**</b>	<b>0.628**</b>	<b>0.659</b>	<b>0.643**</b>	<b>0.782**</b>

\* denotes a statistically significant result ( $p < 0.05$ )

\*\* denotes an extremely statistically significant result ( $p < 0.005$ )

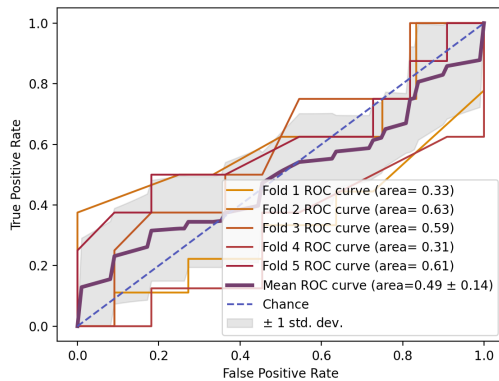
		Actual Performance	
		Low	High
Predicted Performance	Low	26	30
	High	15	26

(a) Without framework

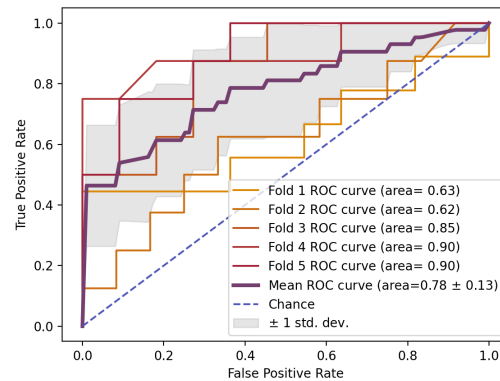
		Actual Performance	
		Low	High
Predicted Performance	Low	27	16
	High	14	40

(b) With framework

Figure 4.9. The confusion matrices for the quantitative intra-set classification models assessed using five-fold cross-validation. (a) shows the model trained without sorting tokens into the IBL framework before counting and (b) shows the model that sorted tokens into the IBL framework before counting. Both models had similar performance in identifying low performing students, but the model with the framework had much better performance differentiating between the two classes overall.



(a) Without framework



(b) With framework

Figure 4.10. Receiver operating characteristic curve for the quantitative classifiers trained intra-set (with five-fold cross-validation). (a) shows the model trained without sorting tokens into the IBL framework before counting (originally designed and tested in [22]), and (b) shows the model trained by sorting tokens into the IBL framework before counting (the new model). The curve for each fold is shown, along with the mean curve  $\pm 1$  standard deviation. As expected from the preliminary work, the original model without the framework performs no better than random at separating low- and high-performing. The new model trained with the framework is better able to differentiate between low- and high-performing.

Without the framework, classification performance is no better than random. Preliminary results in [22] showed that quantitative classification models performed poorly on one cohort, but the performance was even worse when performed on three cohorts here. However, when sorting

student tokens into the framework categories, performance improves; accuracy, precision, F1, and ROC AUC for the model with the framework were all higher than without the framework with extreme statistical significance ( $p < 0.005$ ). If logging more work in general led to student success, we would expect to see high performance when using the model without the framework. However, the poor performance of the model without the framework shows that this was not the case. The *type* of work also mattered, and the model trained with the framework begins to account for this.

The feature extraction results from the quantitative classifier are shown in Figure 4.11. These results show that students who were higher performing were more likely to have more *solve* and *share* tokens, whereas lower performing students were more likely to have more *survey* and *environment* tokens. However, we should be careful in interpreting these results; for example, it is inappropriate to assume that doing less *surveying* is actually helpful for students. Even so, as instructors, these metrics can still be helpful when mentoring teams because they can help teams monitor their time and efforts and ensure that enough time is being spent on other categories like *solve* and *share*.

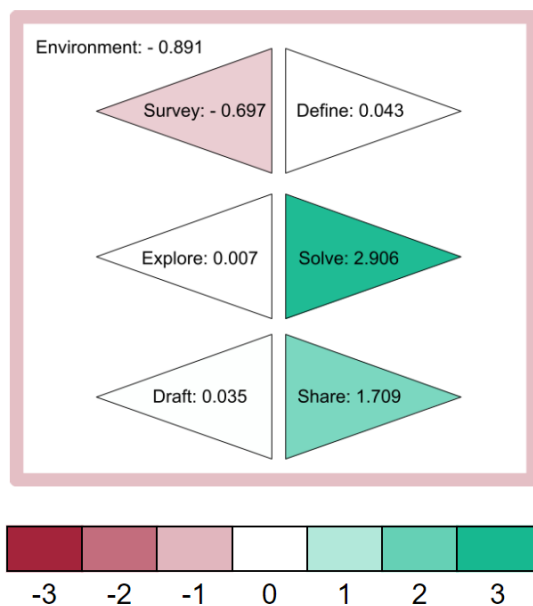


Figure 4.11. The top features for the quantitative classifier by framework category. If a triangle is green and has a positive chi-2 value, higher-performing students were more likely to have that type of token compared to lower-performing students. If a triangle is red and has a negative chi-2 value, lower-performing students were more likely to have that type of token compared to higher-performing students.

#### 4.3.2.6. Quantitative Classifiers Assessed Inter-Set

However, when the quantitative classifiers were tested inter-set, all students were predicted to be high performance both with and without using the framework, leading to little information gained from either classifier. This result could be due to the instructor team introducing students to the IBL framework; because the students in Cohort 3 knew about all categories, they were more likely to create tokens in each of the categories. However, participating in this “desired behavior” that led to predicted high performance did not transfer to actual high performance.

#### 4.4. Research Question 2B: Using the IBL Framework to Improve Clustering

RQ2B asked: Given student token proportions of each framework category, what types of student clusters form?

##### 4.4.1. Methods

To answer RQ2B, hierarchical clustering was implemented to find similarities in quantitative token behavior without considering student performance. The number of tokens in each category was calculated for each student. Both raw token counts and token proportions were then used to create two clustering schemes. Each student is represented by a vector with seven components – one for each of the seven categories. Using the Scipy library [85], a linkage matrix is then created that groups similar students hierarchically. The distances between each cluster are calculated using Ward’s distance, which is calculated:

$$\Delta(A, B) = \sum_{i \in A \cup B} \|\vec{x}_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|\vec{x}_i - \vec{m}_A\|^2 - \sum_{i \in B} \|\vec{x}_i - \vec{m}_B\|^2 \quad (4.6)$$

where  $m_j$  is the center of cluster  $j$  and  $x_i$  is the  $i$ th point in the given set. In other words,  $\Delta$  calculates the merging cost of combining clusters  $A$  and  $B$ , and the merging cost is the difference between the sum of squares distances in the combined clusters and the sum of squares distances in the original separate clusters [86]. When the algorithm begins, each cluster is an individual student, so it aims to find students that are the most similar. After it groups the two most similar students, it continues to iterate through to find other pairs of students or groups that minimize the merging cost. This continues until all clusters are grouped, forming a hierarchical structure.

These results are then plotted in a dendrogram, which illustrates this hierarchical structure with a tree diagram. Each leaf represents a student, and branches form to show which students

are most closely related. These branches are then connected to other branches, showing which *groups* of students are most similarly related. This hierarchical structure continues until the groups of clusters are combined into one single cluster that includes all data points. One benefit of hierarchical clustering is that a researcher can visually identify similarities at any level. From the same figure, the researcher is able to zoom in and see small clusters that are made up of two students, or zoom out and see large clusters that are made up of many students.

After the dendrogram for the raw token counts and for the token proportions were plotted, they were analyzed in the context of cohort. This was done to determine which metric would allow for better comparison across cohorts. If there are large differences in behaviors from year to year, we would expect that students from the same cohort would be clustered together. However, because we want to analyze student behavior *across* cohorts and not *between* cohorts, we want to find a clustering scheme that is less dependent on cohort. In other words, we are looking for a clustering scheme where Cohort 1 students are grouped with Cohort 2 and 3 students, not just with other Cohort 1 students.

After an appropriate clustering scheme is chosen, the clusters were qualitatively analyzed to determine which characteristics are similar across students in any given cluster. In addition, these clusters were analyzed in the context of student performance to determine if certain quantitative behavior patterns align with student success. This performance analysis was performed for the dataset as a whole, as well as for each of the individual cohorts.

Finally, a three-dimensional visualization tool was created to allow an instructor to visualize a student's token proportions in real time. The students can be categorized by cluster, cohort, or performance level – allowing instructors or researchers to quickly gain quantitative information from an easily interpreted three-dimensional space. The tool and its affordances and limitations will be shared.

#### **4.4.2. Results and Analysis**

First, dendrograms for each of the two clustering schemes were plotted. The proportions of student tokens in each category is shown in Figure 4.12a, and the raw counts of student tokens in each category is shown in Figure 4.12b. These figures were colored to show which cohort each of the students fell into. Because the raw count clusters aligned very closely with cohort, that clustering scheme was deemed to not fit the goals of this experiment. Because the leftmost cluster is made up



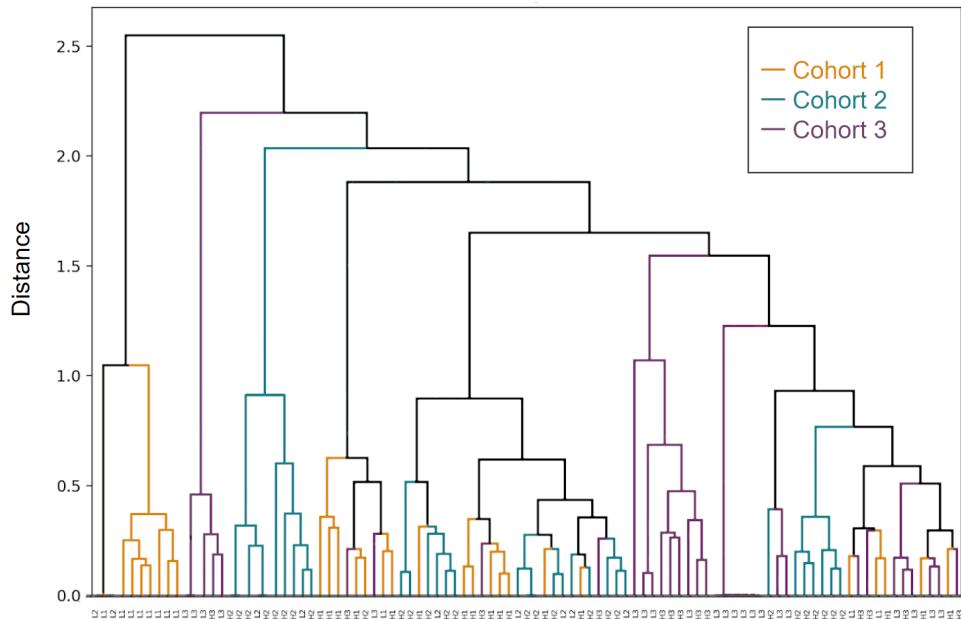
of mostly Cohort 3 students and the rightmost cluster is made up of mostly Cohort 1 students, the characteristics of the clusters were most likely due to differences from year to year – not differences from student to student. The proportions clusters, on the other hand, had some cohort grouping (e.g. the cluster of Cohort 1 students on the far left), but there was more interweaving of students from different cohorts, allowing for a better comparison of student behaviors across the various cohorts.

Next, the dendrogram of the chosen clustering scheme (token proportions) was re-plotted to show student performance (shown in Figure 4.13). From this dendrogram, we identified eight clusters and plotted a distance line to show where the cluster breaks occur. A description and performance breakdown for each of the labeled clusters is shown in Table 4.8.

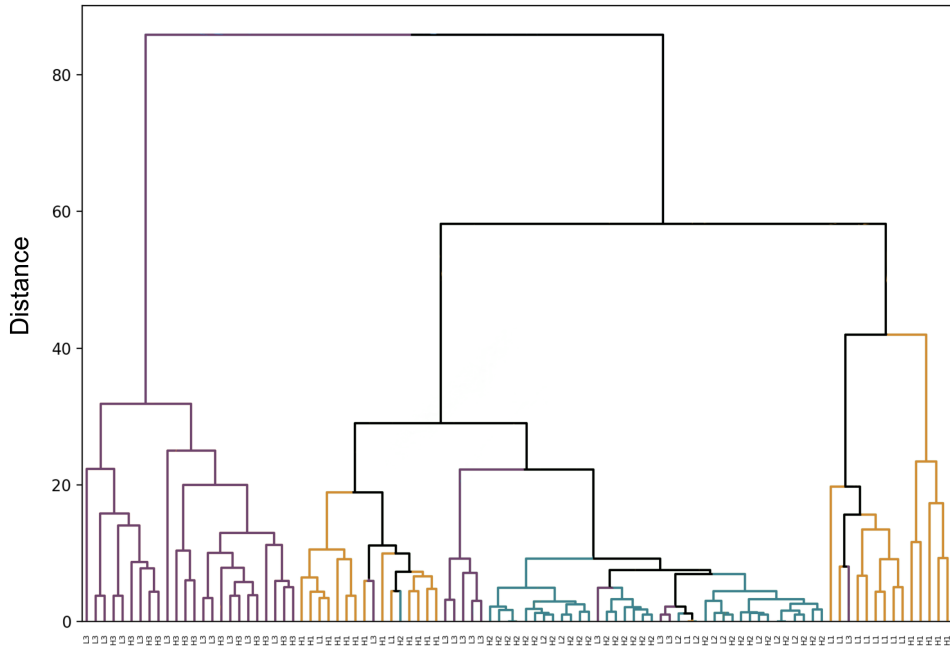
From the cluster analysis, we see that two types of student behavior always led to low student performance: Cluster 1 (High *Environment*) and Cluster 7 (No Project Activity). These results are not particularly surprising because little effort on the project is not likely to lead to a successful innovation. However, no cluster always led to high performance. In fact, some pairs of students with very similar behavior ended up having different performance outcomes.

The question then emerged if these differences were due to differences between cohorts. In other words, if Student A and Student B had similar behavior but different outcomes, is this because Student A was from one cohort and Student B was from another cohort? To explore this question, the dendrograms for each of the individual cohorts were plotted. These results are shown in Figure 4.14. Although some of the clusters are now made up of only high performing students, we still see a handful of pairs of students that have very similar behavior but different end results.

These clustering methods allow researchers to quickly identify similar students or groups of students. However, the only way to interpret why these differences are occurring is to go back and analyze the original data. Therefore, both researchers and instructors could benefit from a way to visually represent token breakdown that provides more meaning than a number of tokens. To meet this need, a three-dimensional visualization tool was developed to be able to plot student token proportions in real time. The x-axis represents proportion of gap tokens, the y-axis represents proportion of solution tokens, and the z-axis represents proportion of impact tokens. An example of this clustering visualization tool is shown in Figure 4.15.



(a) Student token proportion clusters



(b) Student token raw count clusters

Figure 4.12. Dendrograms of student clusters based on (a) proportions of tokens in each category and (b) raw counts of tokens in each category colored by cohort. Each of the leaves represent a student and are labeled with a letter (L or H) to represent low or high performing and a number (1, 2, or 3) to represent cohort number. The intersection of two branches represents the distance between them. From this, we see that raw token counts cause students to cluster mostly by cohort, whereas student proportions allows us to better compare students across cohorts.

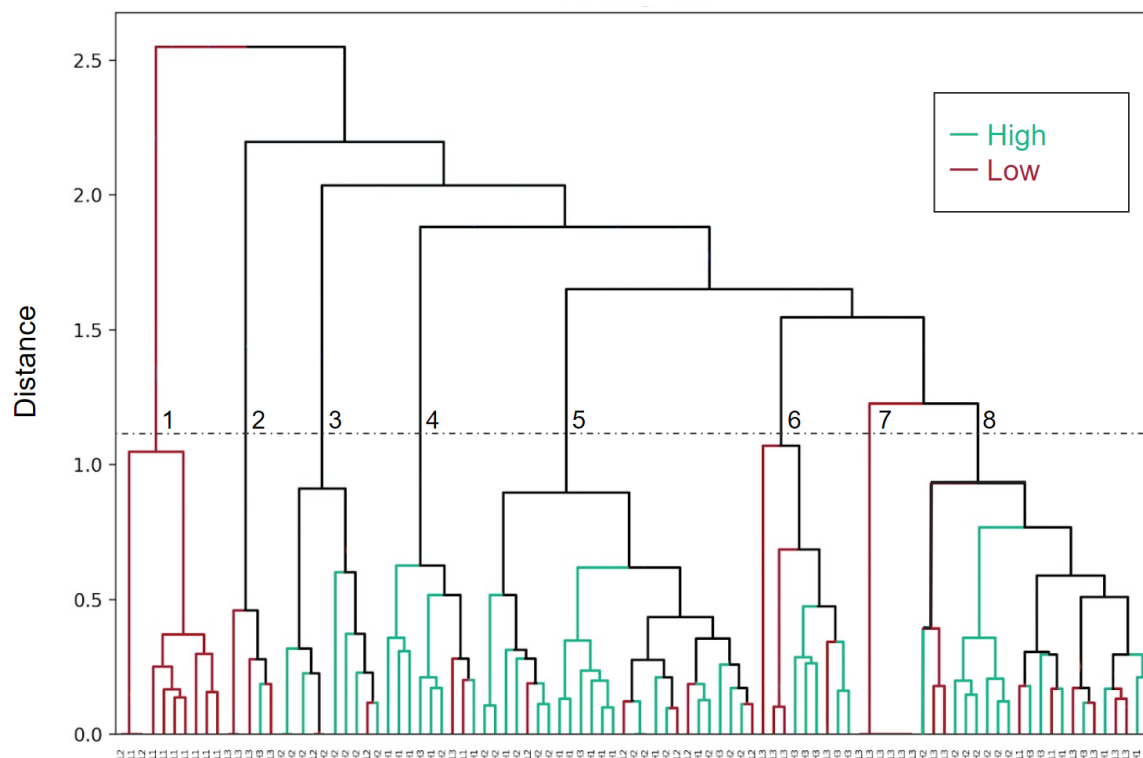


Figure 4.13. Dendrogram of student clusters based on proportions of tokens in each category colored by performance. Each of the leaves represent a student and are labeled with a letter (L or H) to represent low or high performing and a number (1, 2, or 3) to represent cohort number. The intersection of two branches represents the distance between them. The dotted line represents the chosen cluster cutoff point; each of the eight places that a branch intersects with the cutoff point represents one of the eight clusters. From this, we see that some types of behaviors are more likely to lead to low performance (e.g. Clusters 1 and 7), and some types of behaviors are more likely to lead to high performance (e.g. Clusters 4 and 5). However, some of the clusters aren't as predictive of performance (e.g. Clusters 6 and 8).

Table 4.8. Analysis of each of the eight clusters as labeled in the dendrogram in Figure 4.14. Cluster descriptions were developed by qualitatively comparing the category breakdowns for all members of that cluster and finding commonalities.

Cluster	Cluster Description	High Performers	Low Performers
1	High Environment	0	10
2	High Explore	4	1
3	High Share	8	2
4	High Solve	7	2
5	Explore/Solve/Share	21	5
6	Survey/Explore	5	4
7	No Project Activity	0	6
8	Wide Mix	14	8

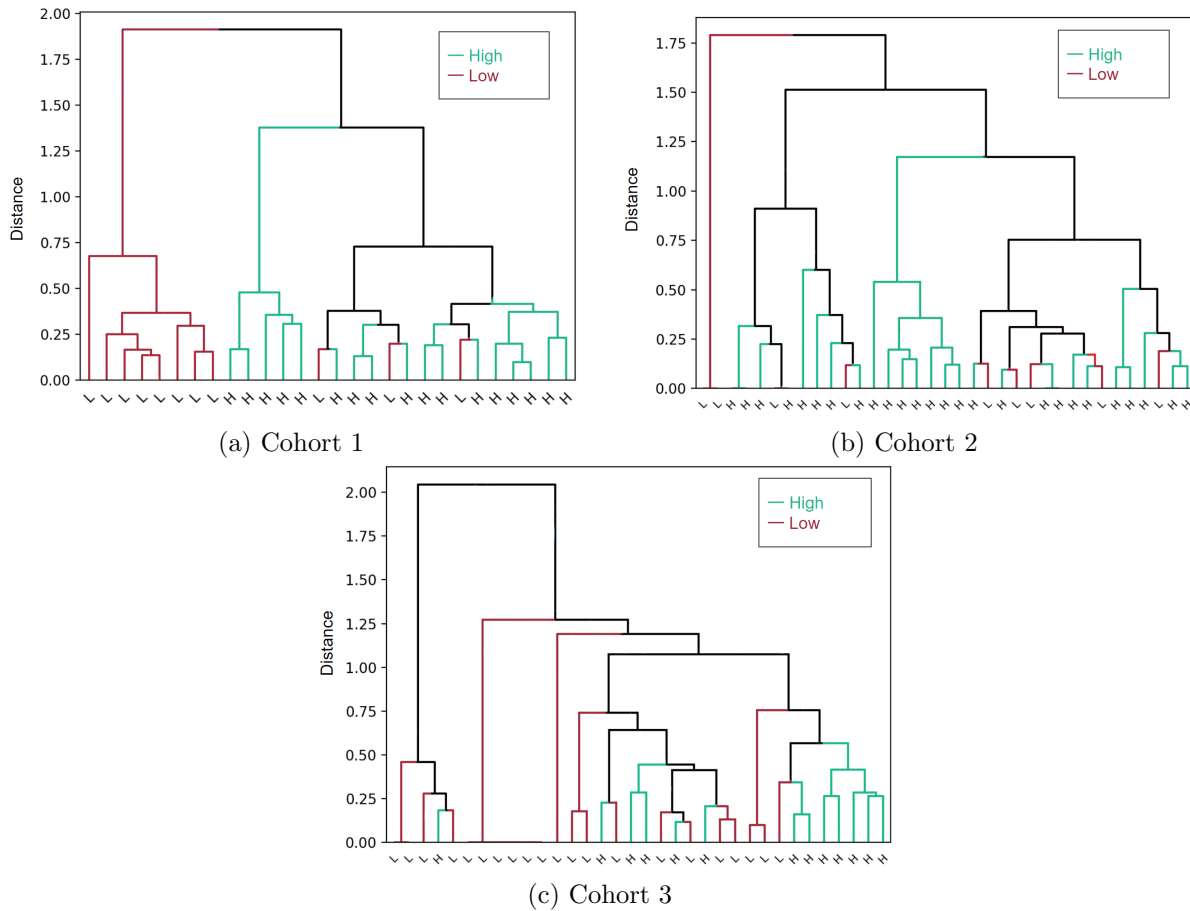


Figure 4.14. Dendrograms of student clusters for each of the individual cohorts: (a) Cohort 1, (b) Cohort 2, and (c) Cohort 3. These dendrograms show that the discrepancies between performance and cluster are not a result of analyzing multiple cohorts together; these discrepancies occur within cohorts as well. For example, for Cohort 1, we see a grouping of low performers, a grouping of high performers, and two mixed groupings.

Although it does not plot all seven categories separately, the reduced three-dimensional space allows for greater interpretability. With a quick glance, an instructor can identify students who may be most at risk (e.g. those who are near Cluster 1). In addition, by coloring each point by cluster, differences can still be identified. For example, Clusters 2 and 4 are in similar positions because Cluster 2 is High Explore and Cluster 4 is High Solve, but the colors help differentiate between the two different behavior types. The instructor can also toggle between different color views to see teams, cohorts, performance, or clusters. For example, by toggling on a specific team, the instructor can see where they are focusing their time: gap, solution, impact, or a combination of the three.

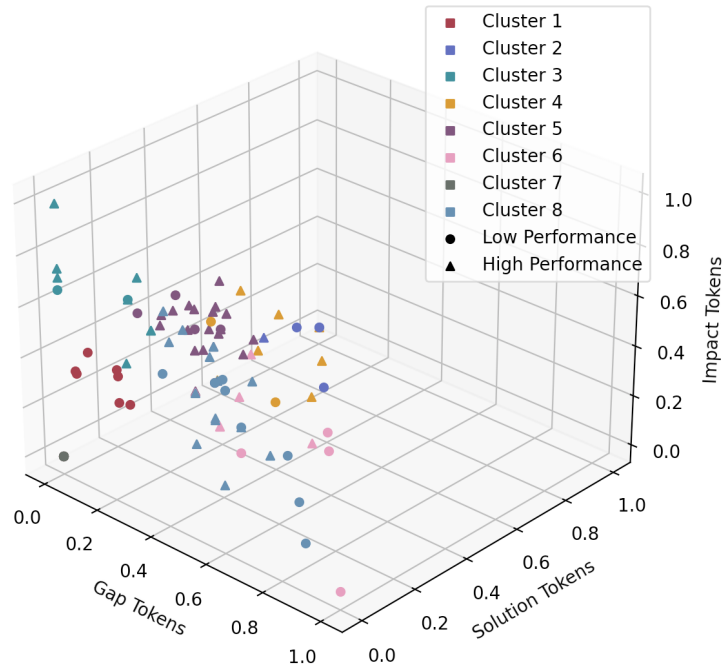


Figure 4.15. Three-dimensional representation of students colored by cluster. Python code was created that automatically plots students based off of token proportions in a three-dimensional space where the x-axis represents proportion of gap tokens, the y-axis represents proportion of solution tokens, and the z-axis represents proportion of impact tokens. This visualization uses color to represent a student’s cluster and shape to represent a student’s performance. By coloring the students by cluster, instructors can still take in some of the information that is otherwise lost by the reduction from seven dimensions to three dimensions. For example, both Clusters 2 and 4 have mostly solution tokens, so they are in similar locations on the graph. However, the color still allows us to differentiate between Cluster 2 (High Explore) and Cluster 4 (High Solve).

#### 4.5. Implications for Teaching

These results in classification and clustering offer various implications for the teaching and learning of innovation.

Because sorting text into the categories IBL framework improved text classifier models, we can conclude that words written by students are further contextualized by the framework category they fall into. This suggests that the same types of words can have different meanings and contexts depending on the framework category considered. This is further demonstrated by the feature extraction results; some words differentiated high performers in one category and low performers in another category. Therefore, the teaching and learning process should help students better identify these different contexts and expectations. For example, if students hear the word ‘paper’,

they might not immediately differentiate between writing a paper to detail their learning about existing pacemakers (*survey*) or submitting a paper to a journal about their research contributions to pacemaker technology (*share*).

The assessment of text classifiers trained using the individual categories of the IBL framework shows that some categories of the framework are more highly differentiating than others. Therefore, instructors may benefit from spending more instructional time focusing on communicating expectations in these categories and supporting teams. For example, the *solve* category was found to be highly differentiating, so instructors could spend time helping students develop their ability to simulate, model, and prototype. Similarly, the *share* category was found to be highly differentiating, so instructors could use class time having students brainstorm ways to create impact with their work and discussing what sorts of activities are high impact.

Finally, the quantitative classification and clustering models offer two interesting insights about assessment in the course: the danger of assessment using strictly quantitative measures and the danger of assessment using outcomes that are unpredictable and sometimes uncontrollable. On the classification side, although the ratio of tokens in each category was somewhat helpful in differentiating between low- and high-performing students intra-set, it had no predictive power on a new dataset. Knowing the number of tokens in each category can be a helpful tool for students and instructors as they track project progress, but having certain token counts or ratios does not predict student success or lead to better student outcomes. After the introduction of the IBL framework for the Cohort 3 students, more students were predicted to be higher performing when using the quantitative models, but these models did not actually predict student performance or improve student outcomes. Therefore, there are dangers in assessing students using these quantitative measures; quantity does not necessarily lead to quality. On the clustering side, however, we see the dangers of assessment using unpredictable outcomes. Students with very similar behavior sometimes ended up with very different final performance assessments. Although this could be a question of quality over quantity, it could also be an issue of the unpredictable nature of innovation; a student or team could do everything right and still end up with a product that does not have external value. Questions still remain about the most equitable and empowering ways to assess students in IBL, but these results demonstrate the nuance of this issue and promote further discussion about holistic

ways to assess students (e.g. “Ungrading” [87]) that support quality learning while eliminating some of the uncertainty that comes in the innovation space.

#### **4.6. Implications for Research**

This work also has a variety of implications for research, including demonstrating the viability of using the IBL framework for LA/EDM analysis, the robustness of these models to sources of uncertainty, and the benefits of linear models.

Previous LA/EDM work has been shown to benefit from the implementation of frameworks that classify actions into certain categories, but this is the first work that demonstrates the benefits of incorporating the IBL framework. By placing actions into the context of the framework, text models improved their ability to predict success for students in new cohorts, and quantitative models improved their ability to separate low- and high-performers intra-set. In addition, the framework allows for meaningful quantification of student performance – beyond just a single token count. This in turn leads to new clustering methods and visualization tools. This chapter speaks to the added benefit of implementing the IBL framework; methods that already had shown some promise were extended and improved with the integration of the framework.

This was also the first work that demonstrated the level of robustness to sources of uncertainty. The results of the Monte Carlo simulation showed that even if some students were classified incorrectly, the models are still able to differentiate between low and high performers. This shows that even though there is some subjectivity involved in the performance classifications, models can still be successfully trained and tested. Very little work in LA/EDM has considered the affects of interrater reliability on model performance, so this could even lead to further research directions that explore the impact of uncertainty for other variables (e.g. classification of student actions into framework categories).

Finally, this work shows that linear models can still provide meaningful insights – even in complex environments. Because linear models allow for feature extraction, the researcher can interpret the feature extraction results to better understand the complex environment. Because these models are reductionist, we should be careful when consider why and how we will be using them. For example, it is not appropriate to use these models to assess students because the use of specific words does not perfectly align with success. In addition, we should be monitoring the evolution of these models over time and assessing not only their performance, but also their

fairness and equity. Although these challenges must be addressed, these models can still be helpful in identifying students at risk in real time.

#### 4.7. Summary

This work shows that learning analytics combined with the Innovation-Based Learning framework can provide meaningful insights about the process of innovation in the classroom.

To answer RQ2A, a variety of text and quantitative classifiers were created and assessed both intra- and inter-set to determine if the IBL framework could improve performance. Compared to previous text classifiers developed in [22], the model trained with the IBL framework had better inter-set performance than the old model while maintaining the strong intra-set performance seen in the old model. In addition, the quantitative classifier trained with the IBL model had stronger intra-set performance than the quantitative classifier developed in [22]. However, quantitative inter-set performance is still a challenge. These results support the hypothesis that the IBL framework can be used to improve classification models.

To answer RQ2B, hierarchical clustering was performed to group students with similar token category ratios. This work showed that there are multiple pathways to student success, but it also demonstrates that the quantitative pathway alone cannot predict success. In fact, some students with very similar behaviors were marked at different performance levels. This work demonstrates the nuance of supporting and assessing IBL and the stochastic nature of predicting student success. Some pathways may have lower or higher probability of success, so we should nudge students towards pathways of higher probability while still giving them freedom to fail.

Supporting the teaching and learning of innovation can be challenging, but these improved models could be used to help instructors identify students and teams at-risk at scale. In addition, the results can be used to better understand the factors of success for innovation in the classroom – leading to the development of evidence-based practices in this area. However, classification and clustering still oversimplify the innovation process; they are reductionist in nature and do not consider a time element. Therefore, the next chapter will aim to tackle the analysis of innovation with a more holistic method that considers the connections between actions and student and team trajectory.



## 5. IMPLEMENTATION OF EPISTEMIC NETWORK ANALYSIS IN IBL THROUGH THE IBL FRAMEWORK

### 5.1. Introduction

Whereas the goal of Chapter 4 was to find ways to understand innovation by reducing its complexities, the goal of Chapter 5 is to understand innovation through embracing its complexities. Rather than creating reductionist linear models, the work in this chapter takes advantage of network analysis, a popular method for modeling complex systems [88]. Specifically, the work uses epistemic network analysis (ENA), a method for identifying temporal relationships between coded data. This method was not originally feasible because it requires data to be sorted into an appropriate amount of meaningful categories or codes.

First, the chapter will provide a background about ENA, including its applications, mathematical foundations, affordances, and limitations. Next, it will describe the methods, results, and analysis for RQ3A which aimed to determine if epistemic network analysis differentiates between low- and high-performing students and teams. Then, it will describe the methods, results, and analysis for RQ3B which aimed to determine how the implementation of different structures in the course changes student behavior. Finally, it will combine the findings from these two questions to share implications for teaching and research.

### 5.2. Background

This background will introduce ENA, discuss some of its current applications, detail its mathematical theory, and assess its affordances and limitations.

#### 5.2.1. Introduction to Epistemic Network Analysis

ENA was originally introduced in 2009 as a method for modeling epistemic frames, or the way that various skills, knowledge, identity, values, and epistemology are related within a specific community or context [89]. Rather than looking at each of these items individually, ENA and the epistemic frame hypothesis make the assumption that “the structure of connections among cognitive elements is more important than the mere presence or absence of those elements in isolation” [50]. In social network analysis (SNA), nodes represent people, and edges represent the

connections between people. With ENA, on the other hand, nodes represent ideas, and edges represent a temporal relationship between those two ideas. This temporal relationship (also called co-occurrence), can be two ideas expressed in the same chat message, two actions that both occurred within a specific time frame, etc. In the case of the IBL data, temporal co-occurrence represents actions in MOOCIBL that were completed within the same week, so ENA analyzes how students and teams transition between various tasks related to their innovation project. Innovation can be defined as the result of combining multiple ideas in new ways [90], and this ultimately aligns with the epistemic frame hypothesis about the importance of connections over individual elements.

ENA is uniquely designed to explore these connections through 3 components: quantification, visualization, and interpretation. ENA allows us to *quantify* data by creating vectors that represent the strength of connections between all variables. In our case, these vectors have a size of 91, and each entry relates to how often a pair of MOOCIBL actions occurred in the same week. From there, specific action pairs can be analyzed across students and groups. In addition, overall behavior of students and groups can be compared by taking these vectors of size 91 and reducing them into vectors of size 2 using singular value decomposition (SVD). This allows researchers to quantify the overall differences between groups and determine statistical significance. ENA allows us to *visualize* data by plotting student vectors in a 2-dimensional space and overlaying the resulting networks. Users can toggle between the networks of individual units or entire groups, and the positions of students on the graph allow the user to see which students are most closely related. Finally, ENA allows us to *interpret* data by placing network nodes in a way that provides the viewer with additional information about the nature of the differences between units or groups. Nodes (each representing a MOOCIBL action) are placed closest to the students that had greater connectedness to that node. By combining the quantification, visualization, and interpretation capabilities of ENA, we are able to uniquely pull meaningful information from the network graphs and confirm these observations quantitatively.

### 5.2.2. Uses of Epistemic Network Analysis

The earliest publication about ENA in 2009 branded it as a “prototype for 21st Century learning and assessment” [89]. Since then, work has been done to show its usefulness and effectiveness in a wide variety of contexts.

For example, one fundamental paper directly showed the advantages of ENA over traditional code-and-count techniques in the context of complex problem solving [52]. Individuals and pairs of pre-service teachers were instructed to verbalize their problem-solving process as they worked on a provided pedagogical problem [91], and these differences were explored using ENA [52]. The transcripts were coded using eight categories: problem identification, questioning, hypothesis generation, generating solutions, evidence generation, evidence evaluation, communicating and scrutinizing, and drawing conclusions. This work was especially impactful because it clearly demonstrated that 1) the connections between codes using ENA provided more information than traditional code-and-count strategies, and 2) the information gained was a direct result of temporality of the data. For example, the ENA results showed that for pairs of teachers, evidence evaluation was central to many other points of discussion. However, when the order of the events was randomized, evidence evaluation was no longer central, showing that temporality mattered in the dataset.

Even within education and learning environments, ENA has become increasingly popular in new contexts. The method was originally designed to find connections in discourse, but a 2022 review of ENA in educational research noted that over half of the papers analyzed in the review did not use student interactions or conversations as a main data source, showing that ENA is expanding to other types of data.

ENA is gaining some traction as a method in engineering education as well. For example, [92] uses ENA to analyze the experiences of ten middle school girls in a 60-hour engineering outreach experience. The five codes used in analysis were skills (participating in actions such as brainstorming or keeping documentation), knowledge (referencing a STEM concept such as center of mass or cross bracing), identity (aligning tasks with their personal image of an engineer), values (implementing big-picture engineering ideas such as understanding and meeting clients' needs), and epistemology (making a justified argument using engineering judgement). Researchers coded interview transcripts and engineering notebooks with these five categories, and these results were then used to explore how central these categories were over time. For example, skills and knowledge were consistently mentioned over time, whereas engineering identity was mentioned less as the experience went on. Values and epistemology, on the other hand, fluctuated depending on the type of activity being completed. If the girls were focused more on client-facing work, values and epistemology were more central to their interview responses and notebook documentation. The authors noted that

ENA allowed them to get a more complete representation of the highly interconnected ideas; their dataset was very qualitatively rich, and ENA allowed them to maintain this richness while reporting quantitative trends.

At the college engineering level, [93, 94] used ENA to explore how student teams progress through the stages of engineering design during a virtual internship. As students worked on their projects, they chatted with other members in an online platform. These chat logs were then coded with the following categories: problem definition, planning, management, information gathering, feasibility analysis and evaluation, selection/decision, and documentation. ENA was then used to determine how team conversation transitioned between categories. Design quality was scored by counting the number of stakeholder requirements that the design met, allowing for teams to be sorted into low- and high-quality design groups. The ENA results were then compared in the context of these groups. The main finding was that the networks of teams that produced high-quality designs had stronger connections to the management code. This code related to setting goals and deadlines. The authors note that ENA uniquely allows for standardization of engineering design assessment without requiring oversimplification of the design process, suggesting that ENA could also be an appropriate method for analyzing the IBL data.

Although originally designed with learning and education in mind, ENA has gained popularity in a variety of other contexts as well. New applications range from surgical errors [95]<sup>1</sup> to Donald Trump's tweets about the COVID-19 [96]<sup>2</sup> to teacher agency in relation to inclusive pedagogy [97]<sup>3</sup>. Although still in its relatively early stages, ENA is gaining popularity because of its unique ability to consider complex connections.

### **5.2.3. Mathematical Theory of Epistemic Network Analysis**

One recent publication gives a complete overview of the mathematics behind ENA, including how data is used to create networks and how the networks are then displayed [98]. Because ENA is relatively new and may be unfamiliar to the reader, this subsection will summarize the original

---

<sup>1</sup>Low-performing surgeons were more likely to only identify and make a plan to fix motor errors, whereas high-performing surgeons were more likely to also identify and make a plan to fix cognitive errors (e.g. doing something in the wrong order) or visuospatial errors (e.g. misjudging where something is).

<sup>2</sup>There were large variations in messaging between Trump and the Center for Disease Control at the beginning of the pandemic, and this messaging became more consistent over time

<sup>3</sup>Teachers acting as agents of change (those proactively taking action) were more focused on increasing student capacity, whereas teachers acting as role implementers (those carrying out their expected responsibilities) were more focused on reducing student barriers.

overview and add additional context about how the concepts should be understood for the IBL data specifically<sup>4</sup>. A worked example of a four-node network can be found in Appendix Section A.2.

### 5.2.3.1. Lines and Codes

Let  $\Theta = (\theta_1, \dots, \theta_n)$  be a sequence of lines (MOOCIBL actions) and  $A = \{\alpha_1, \dots, \alpha_m\}$  be a set of codes (IBL action types). For each line  $\theta_i$ ,  $i = 1, \dots, n$ , an associated code vector  $w_i = (w_1^i, \dots, w_m^i)$  can be created where  $w_j^i = 1$  if line  $\theta_i$  falls under code  $\alpha_j$ .

### 5.2.3.2. Organize Lines and Codes By Student/Group and Week

In order to determine co-occurrence between codes, the unit of co-occurrence (conversations) is defined to be one week. In other words, two actions are said to be temporally related if they occur during the same week. Therefore,  $\Theta$  should be divided into subsequences that each represent one conversation  $c^k$  where  $k = 1, \dots, N$ .  $\Theta$  is also divided into another set of subsequences called units that represent individual students or groups.  $\Upsilon = \{v_1, \dots, v_M\}$  represents the set of  $M$  students/groups where  $v^k = (\theta_{v1}, \dots, \theta_{vk})$  for  $\theta_v \in \Theta$ . To find the set of all lines for unit  $v^y$  and conversation  $c^x$ ,  $\lambda^{xy}$  is defined as  $v^y \cap c^x$  for  $x = 1, \dots, N$  and  $y = 1, \dots, M$ .  $\Lambda$  is then defined as the set  $\{\lambda^{xy} : x = 1, \dots, N \text{ and } y = 1, \dots, M\}$ . In other words,  $\lambda$  represents the set of all code vectors (actions) for student/group  $y$  during week  $x$ .

### 5.2.3.3. Optional: Create Groups to Be Compared

$\Upsilon$  can also be divided into discrete subsets (groups of units) to be compared (e.g. low- and high-performing students). The set of groups  $\Xi = \xi^i, \dots, \xi^n$  is defined such that  $v^n \in \xi^\psi$ .

### 5.2.3.4. Create Adjacency Vectors

Next, a set of adjacency vectors  $A$  are defined for all units and conversations where  $a^{xy} = \sum_{\lambda \in \lambda^{xy}} \lambda$ . In other words, the set of all coding vectors for a specific student/group  $x$  during a given week  $y$  are summed to get a single adjacency vector.

---

<sup>4</sup>Having an understanding of the mathematical foundations of ENA is of the utmost importance for those hoping to analyze and draw conclusions from ENA results. In fact, some of the limitations of ENA that will be discussed in subsection 5.2.3 are related to misunderstandings of how to interpret some of the mathematical topics presented here (e.g. network placement, goodness of fit, and statistical comparison of groups). However, it should be noted that most of the mathematics done in this section is automated through the use of the ENA Web Tool, a software for ENA. Therefore, this summary is not meant to be a complete step-by-step tutorial for getting results, but rather a deep explanation of the mathematical foundations. For a more detailed and complete tutorial, the reader should refer to [98].

### 5.2.3.5. Create Co-Occurrence Matrices

From the adjacency vectors, co-occurrence matrices  $H^{xy}$  are created to denote co-occurrence for codes  $i$  and  $j$  for each unit  $v_y$  and conversation  $c_x$ . These co-occurrence matrices are defined:

$$H_{i,j}^{x,y} = \begin{cases} 1 & \text{if } a_i^{xy} > 0 \text{ and } a_j^{xy} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (5.1)$$

### 5.2.3.6. Create Association Matrices and Vectors

From these co-occurrence matrices, association matrices are created for each unit. Each of these matrices are calculated  $\Omega^y = \sum_{x=1}^N \sum_{H \in H^{xy}} H$  where  $\Omega^y$  is an  $m \times m$  matrix where the diagonal elements are set to zero. In other words, for each unit (student/team), the adjacency vectors from all conversations (weeks) are summed together, showing how many times each pair of codes co-occured throughout the entire time frame (semester).

Because  $\Omega_{ij} = \Omega_{ji}$ , the matrix can be simplified into a vector that eliminates redundant components. If the elements of  $\Omega$  are defined as

$$\Omega_y = \begin{bmatrix} \Omega_{11}^y & \Omega_{12}^y & \Omega_{13}^y & \cdots & \Omega_{1m}^y \\ \Omega_{21}^y & \Omega_{22}^y & \Omega_{23}^y & \cdots & \Omega_{2m}^y \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \Omega_{m-1,1}^y & \Omega_{m-1,2}^y & \Omega_{m-1,3}^y & \cdots & \Omega_{m-1,m}^y \\ \Omega_{m1}^y & \Omega_{m2}^y & \Omega_{m3}^y & \cdots & \Omega_{mm}^y \end{bmatrix},$$

then the association vector  $z^y = (\Omega_{12}^y, \Omega_{13}^y, \Omega_{23}^y, \dots, \Omega_{1m}^y, \dots, \Omega_{(m/2)-1,m}^y)$  where  $z$  is the concatenation of the upper diagonal matrix by column.

### 5.2.3.7. Normalize and Center Association Vectors

The association vectors are then normalized by the length of the vector to create unit vectors that have the same proportions of each code pair as the original association vector.

$$N^y = \frac{z^y}{\|z^y\|} \quad (5.2)$$

To center these vectors, the mean of the normalized association vectors is calculated and subtracted from each of the individual association vectors:

$$N^y = N^y - \bar{N} \text{ where } \bar{N} = \frac{\sum_{y=1}^M N^y}{M} \quad (5.3)$$

### 5.2.3.8. Projecting the Multidimensional Vectors into a 2D Space

Next, a projection is created to place  $N$  in a lower-dimensional space. This can be done in one of two ways: 1) performing a singular value decomposition where each dimension accounts for as much variability between units as possible, or 2) performing a means rotation to separate two groups and then performing a singular value decomposition to account for the remaining variance.

Option 1 is performed when there are more than two groups to compare or when an unsupervised approach is preferred. Option 1 displays the networks with the goal of illustrating the greatest separation of all points – regardless of group. For this option, singular value decomposition is performed on the matrix  $N$  by factoring  $N$  as

$$N = UDV' \quad (5.4)$$

where  $U$  is a matrix whose columns are the eigenvectors of  $NN'$ ,  $V$  is a matrix whose columns are the eigenvectors of  $N'N$ , and  $D$  is a diagonal matrix whose elements are the non-zero eigenvalues of  $NN'$  (which are also equal to the eigenvalues of  $N'N$ ) arranged in descending order. The reduced matrix  $R$  can then be calculated  $R = N' * U$ . For each of the values in  $R$ , item  $R^{ij}$  corresponds to the  $i$ th unit of the  $j$ th dimension of the SVD. In most cases, the 1st and 2nd dimensions of the SVD would be plotted on the x- and y-axis, respectively, so only the first two columns of the SVD would be used for visualizing the 2-dimensional space. The percentage variance for the  $n$ th dimension SVD can be calculated

$$\frac{var(R_{n*})}{\sum_{j=1}^K var(N_{*j})} \quad (5.5)$$

where  $R_{n*}$  is the vector representing the entire  $n$ th column of  $R$ ,  $N_{*j}$  is the vector representing the entire  $j$ th row of  $N$ ,  $K$  is the total number of dimensions in the SVD, and  $var()$  is the variance  $V$

of vector  $A$  of length  $M$  calculated

$$V = \frac{1}{M-1} \sum_{i=1}^M |A_i - \mu|^2 \text{ where } \mu = \frac{1}{M} \sum_{i=1}^M A_i \quad (5.6)$$

Option 2 is performed when there are exactly two groups to compare and when a supervised approach is preferred. This option finds the connections that separate the two groups and displays the network with the goal of illustrating this separation. For this option, a means rotations is performed to separate the two groups first. A vector  $\vec{u}$  is calculated

$$\vec{u} = \frac{\bar{\xi}_1 - \bar{\xi}_2}{\|\bar{\xi}_1 - \bar{\xi}_2\|} \text{ where } \bar{\xi}^{\psi} = \frac{\sum_{v \in \xi^{\psi}} N^v}{|\xi^{\psi}|} \quad (5.7)$$

To find the values for each unit, multiply  $N'\vec{u}$  to get a vector where the  $m$ th value corresponds to the position of unit  $m$  on the x-axis. The percentage variance due to the means rotation dimension can be calculated

$$\frac{\text{var}(N'\vec{u})}{\sum_{j=1}^K \text{var}(N_{*j})} \quad (5.8)$$

In this case, rather than performing SVD on matrix  $N$ , the variance from the means rotation must be removed. Matrix  $\tilde{N}$  is calculated by removing each point's projected component on  $\vec{u}$  to get:

$$\tilde{N} = N - N\vec{u}\vec{u}' \quad (5.9)$$

SVD is then performed on  $\tilde{N}$  as seen in Option 1.

#### 5.2.3.9. Placement of Nodes

In order to place each of the nodes in the lower dimensional space, matrix  $B$  must be found where  $B_{ij}$  corresponds to code  $i$  for dimension  $j$ .

To find  $B$ , first minimize the sum of square distances between the actual centroids  $c^k$  for each unit and the projected points  $R^k$  for each unit:

$$\sum_{k=1}^{|U|} (R^k - c^k)'(R^k - c^k) \quad (5.10)$$

where:



$$c^k = B\tilde{w}^k \quad (5.11)$$

and for unit  $k$ :

$$\tilde{w}_i^k = \frac{1}{2} \left( \frac{\sum_{j=1}^m \Omega_{ij}^k}{\sum_{i=m, j=m} \Omega_{ij}^k} \right) \quad (5.12)$$

In other words,  $\tilde{w}_i^k$  represents half of the sum of all weights of the connections to node  $i$  normalized by the sum of all weights in the network for unit  $k$ .

Because  $\tilde{w}^k$ ,  $c^k$ , and  $R^k$  are all known,  $B$  can be found using linear least squares. After  $B$  has been found, it can then be used to determine the position of each node where  $B_{ij}$  corresponds to code  $i$  for dimension  $j$ .

Intuitively, nodes are placed to help better visualize differences between groups of units. If a node is placed closer to the origin of the graph, that node is less differentiating between groups. If a node is placed farther from the origin of the graph and is on the negative x-axis, it highly differentiates data on the first dimension and is more important to units that also are placed on the negative x-axis.

#### 5.2.3.10. Goodness of Fit of Visualization

Because the 2D plots are network reductions, it is important to measure how well the 2D plot actually represents the original high-dimension networks. Therefore, goodness of fit should be measured for each dimension individually. Goodness of fit can be defined as the Spearman correlation between the pairwise directed difference vector of the actual centroids and the pairwise directed difference vector of the projected centroids.

To calculate the pairwise directed difference vectors, the pairwise directed difference matrices are calculated using function  $F$ :

$$f_{i,k}^j(M) = m_{i,j} - m_{k,j} \quad (5.13)$$

In other words, function  $F$  creates a matrix where entry  $M_{i,k}^j$  is the distance between units  $i$  and  $k$  along dimension  $j$ .  $F^j(C)$  returns the pairwise directed difference matrix of the actual centroids, and  $F^j(R)$  returns the pairwise directed difference matrix of the projected centroids. Because  $f_{i,k}^j(M) = f_{k,i}^j(M)$ , these matrices can be converted to vectors that eliminate any redundant components (as originally seen in sub-subsection 5.2.2.5). For each dimension, the goodness of fit is then calculated by taking the Spearman correlation between the vector for the actual centroids and the vector for the projected centroids.

If the goodness of fit for a specific case is close to 1, this means that the distances between the projected centroids for all pairs of units and the distances between the actual centroids for all pairs of units are strongly correlated. Thus, the visualization is mathematically consistent with the summary statistics of the network, and the positions of nodes can be used to interpret the meaning of the dimensions.

#### 5.2.3.11. Mean network and confidence interval

To determine the position of the mean network centroid for a group of units, the average values on the x- and y-axes are taken for the group, and these values are plotted. Confidence intervals are then found using the equation:

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}} \quad (5.14)$$

where  $\bar{x}$  is the mean of the values on that dimension,  $z$  corresponds to the confidence level ( $z = 1.96$  for 95% confidence),  $s$  corresponds to the standard deviation of the values on that dimension, and  $n$  represents the number of samples.

#### 5.2.3.12. Statistically Comparing Groups

To compare two groups, a Mann-Whitney test is performed on both the x- and y-axes. For the IBL data, the ENA Web Tool calculates the Mann-Whitney test automatically. To perform the Mann-Whitney test by hand, all units are ranked from smallest to largest. The ranks are then added for the units in each group where the sum of rankings for group 1 is  $R_1$  and the sum of rankings for group 2 is  $R_2$ .

$U_1$  is then given by:

$$U_1 = R_1 - \frac{n_1(n_1 + 1)}{2} \quad (5.15)$$

and  $U_2$  is given by:

$$U_2 = R_2 - \frac{n_2(n_2 + 1)}{2} \quad (5.16)$$

The smaller value for  $U$  can then be used to consult a significance table [99] to get a p-value and determine if the differences between the two groups are statistically significant.

#### 5.2.4. Affordances and Limitations of Epistemic Network Analysis

Although ENA is still an emerging method, it offers various affordances over other more common methods. As mentioned in the previous chapter, popular learning analytics tasks include classification, clustering, association analysis, sequence mining, and process mining. Therefore, sequence and process mining were the first two methods explored when considering the temporal relationships between IBL actions. However, these methods were deemed to be inappropriate for the current IBL dataset. Because the number of possible sequences is so large, sequence and process mining methods require large amounts of data in order to find patterns [52]. This problem is further exacerbated due to the complexity of the innovation process where there are many valid variations. Sequence and process mining may be appropriate for procedural problems with hard boundaries (e.g. where Action B always starts after Action A ends), but the results of these methods lose meaning as the context becomes more complex. Therefore, methods that look for relationships and interactions are more appropriate than methods that rely heavily on the same action sequences occurring across cases.

In addition to ENA, other methods exist for exploring relationships in data, such as multivariate analysis methods and other network analysis methods (e.g. social network analysis). However, compared to these other methods, ENA is uniquely positioned to handle the size and characteristics of the IBL data. If a researcher wants to look at interactions between ideas or codes, the number of connections increases exponentially as more ideas or codes are added. For example, in the case of the IBL data, there are 14 codes (7 categories of the IBL framework times 2 action types: start or finish), meaning there are 91 possible interactions. Because our sample size is small, limited conclusions can be drawn from traditional multivariate analyses that consider all of these relationships. In addition, the results of these analysis methods place little importance on the interactions between *many* elements. Network analysis methods, on the other hand, are

designed specifically to explore these interactions. However, most network analysis methods are designed for hundreds, thousands, or even millions of nodes, so specific relationships are not able to be visualized and interpreted. ENA is designed for data that falls somewhere between multivariate analysis and other network analysis techniques; there are too many interacting elements to draw conclusions from traditional statistical methods, but there are few enough elements that visualization and qualitative interpretation is possible.

However, limitations of ENA must also be considered. A 2022 review of 76 empirical studies using epistemic network analysis in educational settings took a critical look at ENA and its uses [100]. These limitations were framed in the context of “unmet promises” of ENA, including analysis of quantitative relationships, visualization of large networks, easy interpretability, automation and scalability, and ability to map trajectories. Each of these five unmet promises is described in detail below.

- *Analysis of quantitative relationships:* ENA was originally presented as a mixed methods approach, but in many studies, little to no quantitative data is reported. Statistical significance of differences between groups, weight of connections, and goodness of fit are rarely all reported. The differences between groups and weight of connections can be qualitatively extracted (Group A and Group B are “fairly” different; Connection C is “somewhat” stronger than Connection D), but these claims could easily be strengthened by providing quantitative data.
- *Visualization of large networks:* ENA was designed to allow for exploration of relationships when multivariate parametric techniques are inappropriate. However, some studies in the review had so many nodes (more than 40) that it was almost impossible to draw conclusions from the network visualization. Therefore, those using ENA should be careful to ensure that their specific dataset is appropriate. Although none of the core ENA papers [89, 50, 101, 102, 98] give an exact range of appropriate number of nodes, the most cited tutorials and worked examples contain between 6 and 18 nodes.
- *Easy interpretability:* Even if the number of nodes is appropriate, interpreting the meanings of networks has still been limited. Although most studies presented network visualizations that showed differences between students or groups, the meaning of that difference in the

context of the educational environment was rarely reported. For example, few studies gave the x- and y-axes meaning as was presented in [101, 102]. Even if a practitioner can visualize the location of a student on the plot, most studies did not explain what that position meant or what interventions might be appropriate. In order to decrease the gap between research and practice, researchers using ENA should consider the goals and needs of practitioners when presenting their results.

- *Automation and scalability:* ENA was originally presented as a way to analyze interactions in real time. However, most of the articles analyzed in Elmoazen et al.'s 2022 review still required a human to code all pieces of data. In fact, only about one in ten articles reported an automatic coding process. Without an automatic coding process, research is limited to manageable sample sizes and is slowed by the coding process. Therefore, in order to provide the automation and scalability that ENA originally promised, the entire workflow of the analysis process should be carefully considered. Not only is there a need for automated coding, but also automated preprocessing and analysis to reduce the time between data collection and final delivery of results.
- *Ability to map trajectories:* In a popular ENA tutorial [50], a method for presenting trajectories of students or groups was presented as a way to indicate changes in connections over time. However, the authors of the 2022 review paper did not see this method used elsewhere. They also note that most of the trajectory models use aggregated networks over certain time periods. Rather than treating the data as a continuous piece of yarn illustrating student trajectory, this method treats each time period as its own individual bead on a string. The beads can be compared, but the connections within and across beads are lost.

In Section 5.6, these limitations will be revisited and evaluated in the context of this study.

### **5.3. Research Question 3A: Behavior of Students and Teams in Low-Structure IBL**

RQ3A asked: Do high- and low-performing students and teams have different behaviors in the course in the context of co-occurrence?

#### **5.3.1. Methods**

In order to compare low- and high-performing students and teams, ENA Web Tool (version 1.7.0) [103] was used. Three experiments were conducted: 1) comparing students by performance,

2) comparing teams by performance, and 3) comparing team trajectories by performance. It should be noted that RQ3A looks only at Cohort 1 (N=28) to explore how students progress through the innovation process when given little structure.

For all three experiments, MOOCIBL actions are categorized into 14 categories. The prefix of the category (S or F) corresponds to the stage of the action (either Start or Finish), and the suffix of the category (Survey, Define, Explore, Solve, Draft, Share, Or Environment) corresponds to the category of the IBL framework. For example, when a student creates a new token in the *explore* category, that line would be coded ‘S.Explore’.

For each unit of analysis, a network model is created by calculating the number of times each pair of codes appears in the same stanza (in our case, a week). For experiment 1, the unit is defined to be a single student. For experiment 2, the unit is defined to be a single team. For experiment 3, the unit is a single team during a single quarter of the semester (i.e. four networks are created to represent each team).

These network models are then normalized, and the resulting network is then reduced using singular value decomposition, which produces orthogonal dimensions that maximize the variance explained by each dimension. Refer to subsection 5.2.3 for a complete mathematical description. It should be noted that none of the experiments used a means rotation. In other words, student success level was *not* considered when placing the network points (an unsupervised approach).

The results is two coordinated representations for all units of analysis: 1) a plotted centroid that represents the location of the unit and 2) a weighted network graph for that unit. The location of the plotted centroid is found by taking the first and second SVD of the network; the first SVD represents the x-coordinate, and the second SVD represents the y-coordinate. The weighted network graph is made up of 14 nodes (one for each of the codes), and each edge weight represents the relative co-occurrence of that pair of codes. In order to determine if these 2-dimensional representations are representative of original data, goodness of fit is calculated. If goodness of fit is close to 1, the positions of centroids and nodes can be used to interpret results. For example, by qualitatively looking at which types of nodes appear more in each quadrant, it is possible to give the axes some intuitive meaning.

In order to compare groups by performance, sets of units are created by categorizing the units as low- or high-performing. By creating these sets, the units in each set can be aggregated

to create one average network. In addition to average networks, difference networks can also be calculated. For each edge, the difference between the weights of the two networks is calculated. Thinner lines represent smaller differences, and the color represents which network had a higher weight for that specific connection.

By toggling between the various views, individual centroids, individual networks, average centroids, average networks, and difference networks can be visualized. Figure 5.1 is designed to give readers an intuitive idea of how each of these representations are created.

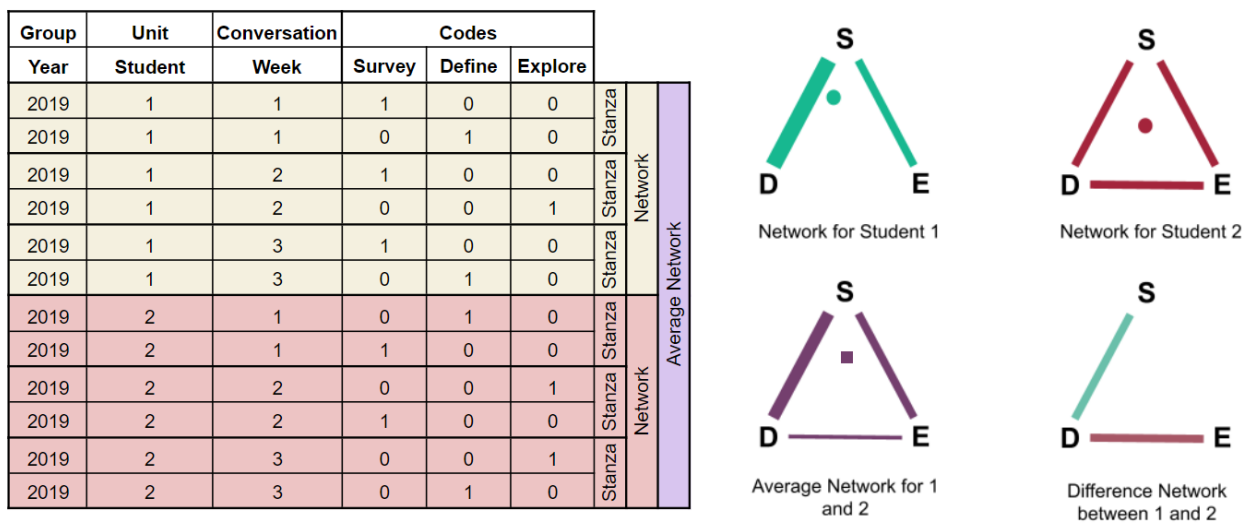


Figure 5.1. A visual interpretation of individual networks, individual centroids, average networks, average centroids, and difference networks. For each unit (in this case, student), the number of co-occurrences is counted for each pair of actions. The weighted network for that unit is then created where each node represents a code and each edge represents the relative co-occurrence of that pair of codes. The individual centroids (represented by circles) are plotted corresponding with the edge weights. To calculate the average network for a group of units, the weights of each edge for all networks are averaged. Next, the average centroid (represented by a square) is plotted corresponding with the average network edge weights. Alternatively, a difference network can be created by taking the difference between each set of corresponding edges.

Plots of all 2019 students, all 2019 teams, and team trajectories will be reported using two views. The first will show the centroids for each unit, the averages for low and high-performing, and the confidence interval bound. The second will show the positions of the nodes and the difference network between low- and high-performing units.

In order to show which action pairs were most common among low- and high-performing students, tables were also created showing the relative proportion of each action code for each of the two performance groups.

### 5.3.2. Results and Analysis

#### 5.3.2.1. Experiment 1: Comparing Students by Performance

First, all students from Cohort 1 were plotted using ENA. For readability, the individual centroids (circles) and average centroids (squares) with confidence intervals are plotted in Figure 5.2, and the difference network between high and low performing students is plotted in Figure 5.3. Both have the same axes and scale. The x-axis corresponds to the first SVD and 21.0% of the overall variability in the data, and the y-axis corresponds to the second SVD and the next 10.3% of variability in the data.

From Figure 5.2, it is clear that there is high separability between the low-performing and high-performing groups of students – even with an unsupervised approach. Because the confidence intervals do not overlap on the x-axis, we can visually see that there is significant separation on the x-axis. These results are confirmed by the results of the Mann-Whitney Test ( $p < 0.00001$ ).

From Figure 5.3, it becomes clearer what types of differences occur between the two groups. First, by looking at the network difference graph, we can see which edges were more likely to be prominent in higher-performing students (green) and lower-performing students (red). Qualitatively, we can see that many of the green connections are connected to *explore* and *solve* codes, and many of the red connections are connected to *environment* codes. These observations can be supported quantitatively by looking at the reported strengths of each connection in Figure 5.4.

In addition, because the groups separate on the x-axis and the goodness of fit was 0.9886 on that axis, we were able to extract meaning from the positions of codes in relation to their position on the x-axis. F.Explore, S.Solve, and F.Solve all appear on the left side of the graph (with S.Explore right in the middle). Therefore, the negative x-axis was labeled “Explore/Solve”. S.Environment and F.Environment both appear on the right side of the graph, so the positive x-axis was labeled “Environment”. A point that falls on the left hand side is more likely to have *explore/solve* codes as central codes in their network, and a point that falls on the right hand side is more likely to have *environment* codes as central codes in their network. Because there are other codes involved, this “rule” is not perfect, but it gives some level of visual interpretation (overcoming one of the



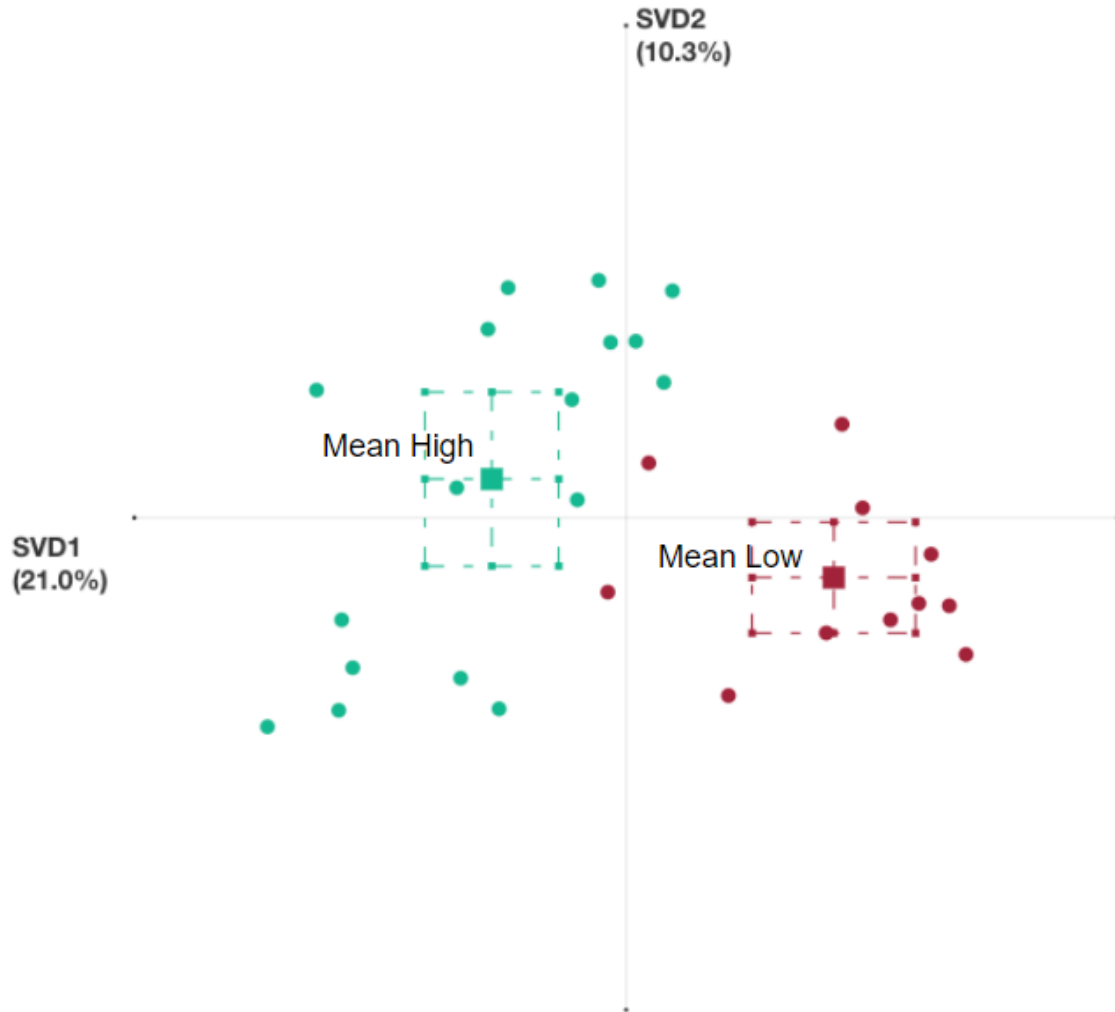


Figure 5.2. ENA representation of all students in Cohort 1. Each circle represents a student; green circles represent a student who was identified as high-performing, and red circles represent a student who was identified as low-performing. The labeled squares represent average centroids for high- and low-performers, and the dotted boxes represent a 95% confidence interval for that group. Note that the confidence intervals do not overlap; this corresponds with the results of the Mann-Whitney Test for a statistical significant difference between the two groups ( $p < 0.00001$ ). It also should be noted that this projection did not use a means rotation – meaning the algorithm did not use a supervised approach to intentionally separate the two groups. Yet, the low and high performers separate themselves across the x-axis (which represents the first singular value decomposition and accounts for 21.0% of the variability within the data). Thus, much of the variability in the data is related to student performance.

limitations presented in the 2022 review of ENA [100]). Because there was no separation of groups or types of codes on the y-axis, it was not labeled for this experiment.

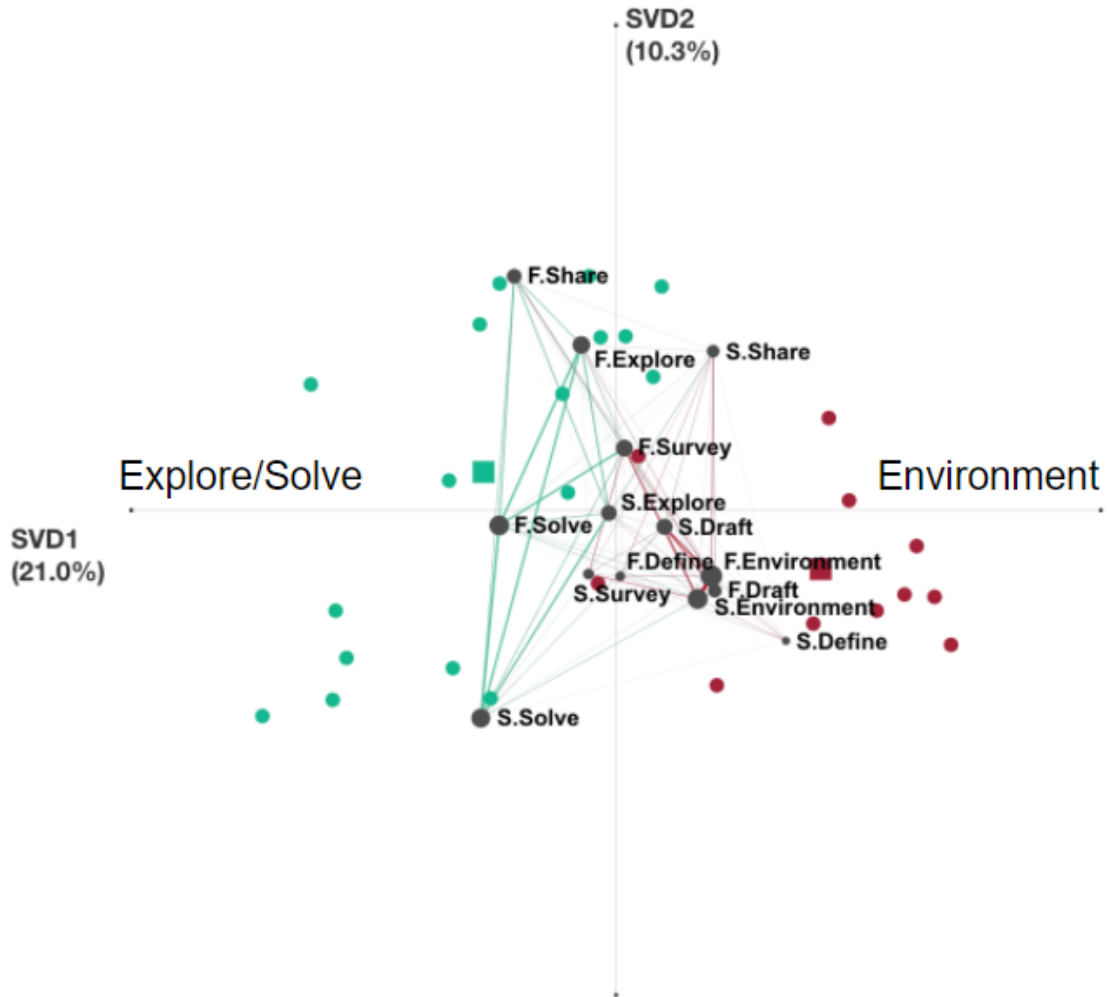


Figure 5.3. ENA representation comparing the average network for high-performing students in Cohort 1 and low-performing students in Cohort 1. From figure 5.2, low- and high-performing students generally separated across the x-axis. To understand the differences between the two, the difference network between the average high-performing and average low-performing can be plotted. If an edge is green, it is more likely to be an action pair completed by a high performer. If an edge is red, it is more likely to be an action pair completed by a low performer. Edge thickness represents how great the difference was between the two groups for that specific node, allowing us to see which action pairs were more likely to occur for each group. Because the goodness of fit for this visualization is 0.9886 on the x-axis and 0.9546 on the y-axis, we can also interpret the data using the positions of the nodes. *Explore* and *solve* codes appear more on the left (high-performing side), and *environment* codes appear more on the right (low-performing side). Therefore, we can conclude that high-performing students were more likely to iterate between *explore* and *solve* tokens, whereas low-performing students often focused on starting and finishing *environment* tokens.

### 5.3.2.2. Experiment 2: Comparing Teams by Performance

Next, all teams from Cohort 1 were plotted using ENA. In this case, the analysis considers how team members may be influencing each other; one team member might complete a *Define*



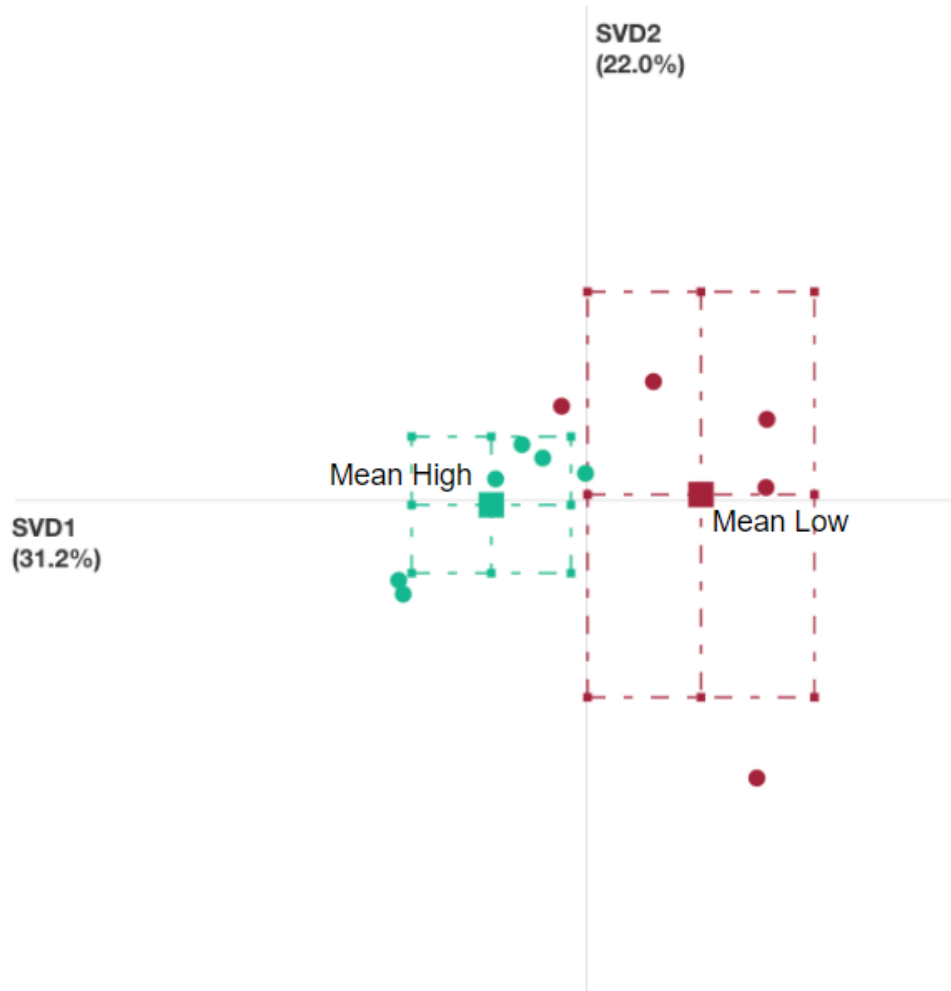


Figure 5.5. ENA representation of all teams in Cohort 1. Each circle represents a team; green circles represent a team who was identified as high-performing, and red circles represent a team who was identified as low-performing. The labeled squares represent average centroids for high- and low-performing teams, and the dotted boxes represent a 95% confidence interval for that group. Note that the confidence intervals do not overlap; this corresponds with the results of the Mann-Whitney Test for a statistical significant difference between the two groups ( $p < 0.01$ ). It should once again be noted that this projection did not use a means rotation – meaning the algorithm did not use a supervised approach to intentionally separate the two groups. Yet, the low and high performers separate themselves across the x-axis (which represents the first singular value decomposition and accounts for 31.2% of the variability within the data). Thus, much of the variability in the data is related to student performance.

that axis meaning based off of the positions of the nodes (explore/solve on the left, environment on the right). These results do not provide much additional information beyond Experiment 1, but it does show that teams can also be modeled using ENA.

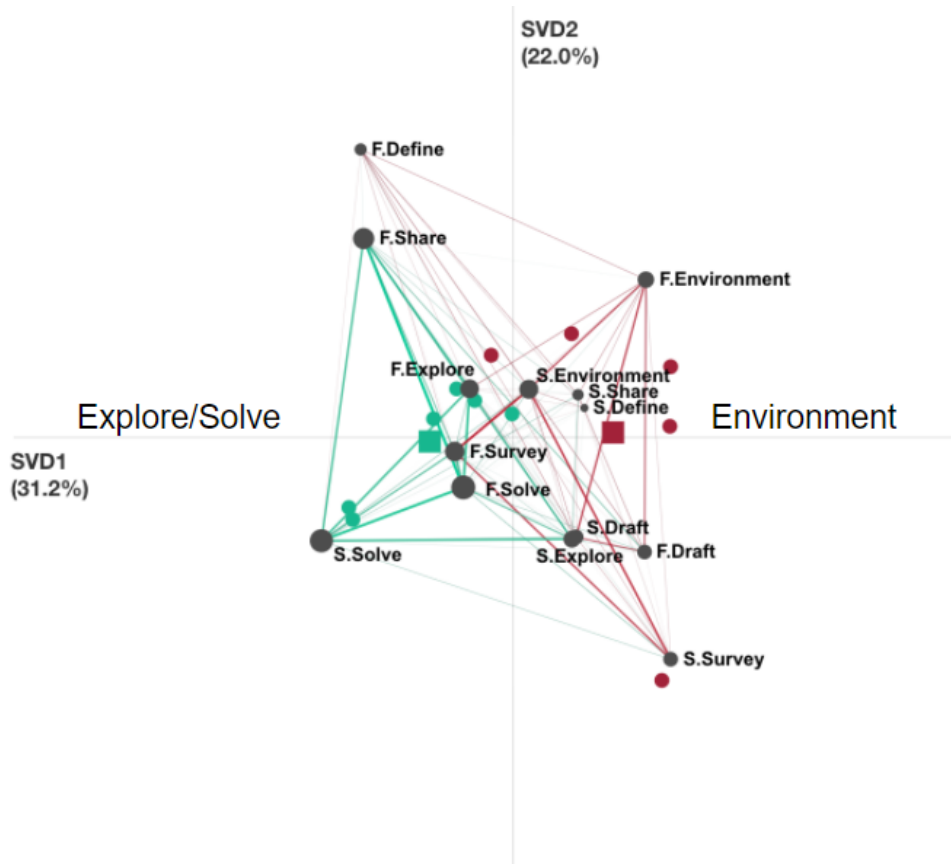


Figure 5.6. ENA representation comparing the average network for high-performing teams in Cohort 1 and low-performing teams in Cohort 1. From figure 5.5, low- and high-performing students generally separated across the x-axis. To understand the differences between the two, the difference network between the average high-performing and average low-performing can be plotted. If an edge is green, it is more likely to be an action pair completed by a high-performing team. If an edge is red, it is more likely to be an action pair completed by a low-performing team. Edge thickness represents how great the difference was between the two groups for that specific node, allowing us to see which action pairs were more likely to occur for each group. Because the goodness of fit for this visualization is 1.0 on both the x- and y-axis, we can also interpret the data using the positions of the nodes. *Explore* and *solve* codes appear more on the left (high-performing side), and *environment* codes appear more on the right (low-performing side). Therefore, we can conclude that high-performing students were more likely to iterate between *explore* and *solve* tokens, whereas low-performing students often focused on starting and finishing *environment* tokens.

### 5.3.2.3. Experiment 3: Comparing Team Trajectories by Performance

Finally, team trajectory was also mapped. Team behavior from the semester was split into four quarters, and each of the quarters for each team was plotted. Once again, two visualizations are shown: 5.7 shows individual team centroids (circles) and average centroids (squares) for each quarter, and 5.8 shows the difference network between high and low performing teams. Both have

the same axes and scale. The x-axis corresponds to the first SVD and 16.6% of the variability in the data, and the y-axis corresponds to the second SVD and the next 10.0% of the variability in the data.

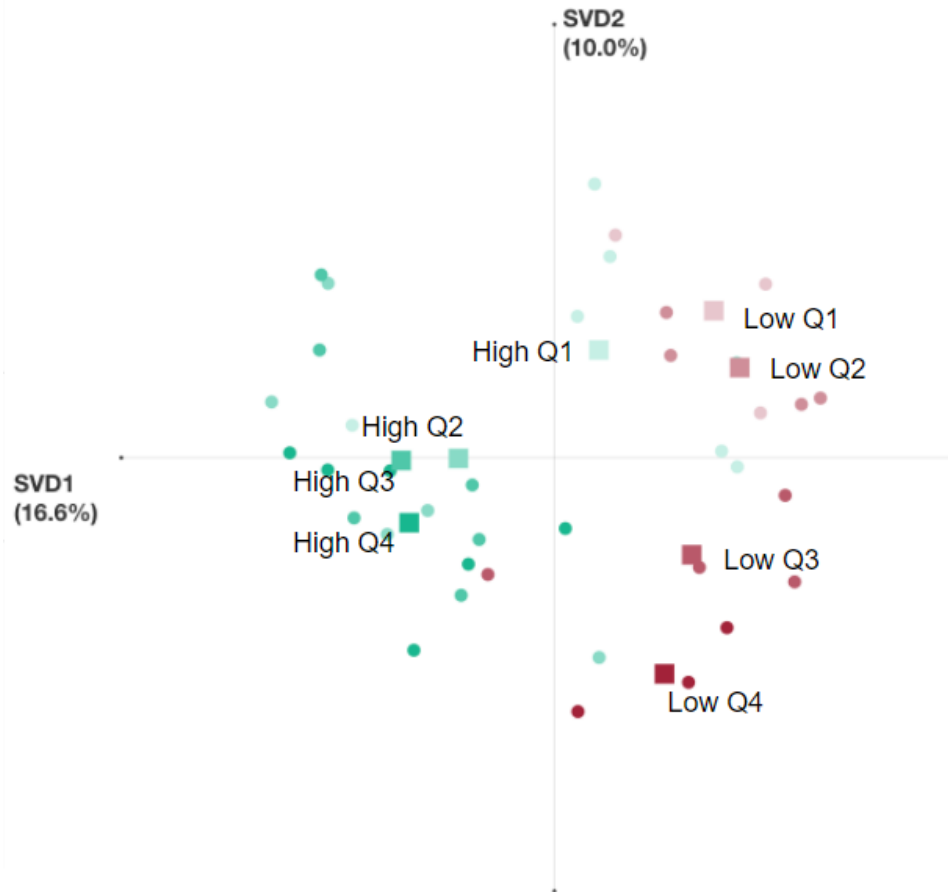


Figure 5.7. ENA representation of team trajectories for Cohort 1. Each circle represents a team’s aggregated network for a quarter of the semester. Light colored dots represent earlier in the semester, and brighter colored dots represent later in the semester. Green dots represent teams identified as high-performing, and red dots represent teams identified as low-performing. Because performance groups separate on the x-axis (the first SVD), we can deduce that differences in performance account for the first level of variability in the data (16.6%). Because the quarters separate on the y-axis (the second SVD), we can deduce that temporal behavior differences account for the second level of variability in the data (10.0%).

From Figure 5.7, we can see that team behavior generally starts at the top and moves down with high-performing on the left and low-performing on the right. Because there is higher variability in behavior over the shorter periods of time, these differences are not statistically significant, but

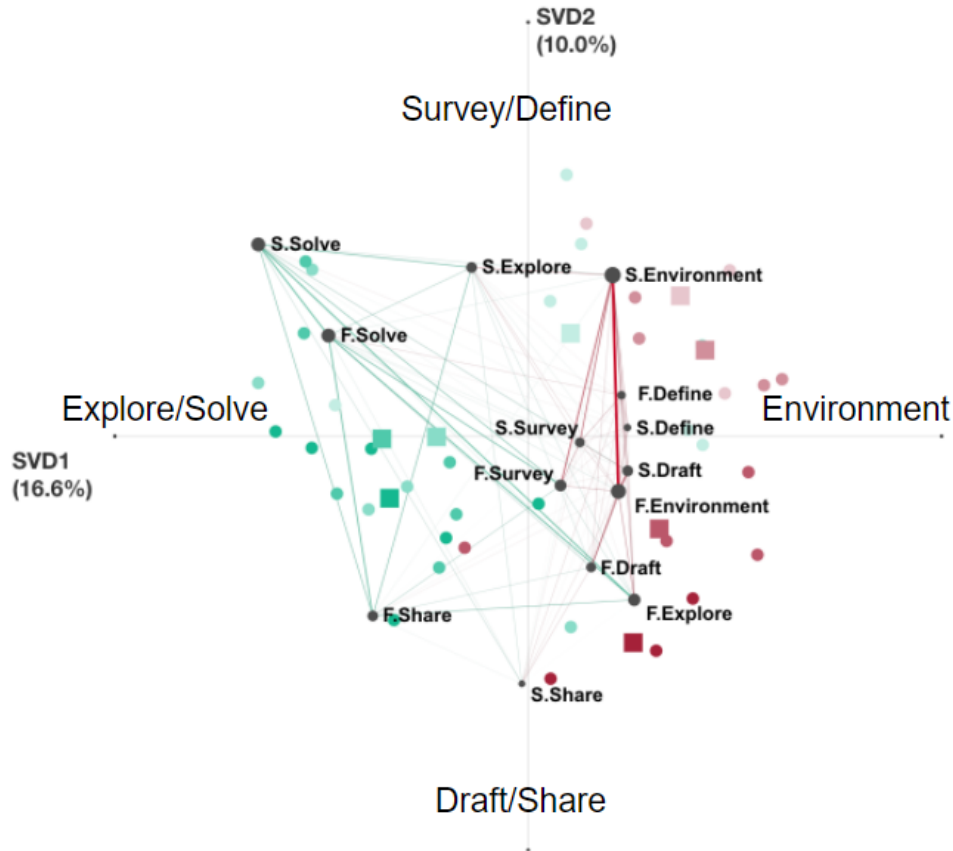


Figure 5.8. ENA representation comparing the average network for high-performing teams in Cohort 1 and low-performing teams in Cohort 1. If an edge is green, it is more likely to be an action pair completed by a high-performing team. If an edge is red, it is more likely to be an action pair completed by a low-performing team. Edge thickness represents how great the difference was between the two groups for that specific node, allowing us to see which action pairs were more likely to occur for each group. Because the goodness of fit for this visualization is 0.9304 on the x-axis and 0.9578 on the y-axis, we can also interpret the data using the positions of the nodes. Once again, *explore* and *Solve* codes appear more on the left (high-performing side), and *Environment* codes appear more on the right (low-performing side). Temporally, *Survey* and *Define* codes tend to occur closer to the top of the graph (Q1 end), and *Draft* and *Share* codes tend to occur closer to the bottom of the graph (Q4 end). However, the differences between Q1 and Q4 behavior is less than the differences between high- and low-performing teams based off of the assumption that the first SVD accounts for the most variability and aligns with performance.

we see that by quarter 2, high- and low-performing teams are fairly separated. It is interesting to note that most of the variability in the data is still due to the differences between low- and high-performing teams (hence the separation between groups on first SVD axis), whereas the next dimension accounts for differences between earlier and later in the semester (hence the separation between quarters on the second SVD axis).

Because the goodness of fit is 0.9304 on the x-axis and 0.9578 on the y-axis, we can also label the axes based off of the node positions shown in Figure 5.8. Once again, *explore/solve* codes fall on the left side of the graph and environment codes fall on the right side of the graph, but we now see a temporal shift along the y-axis. *Survey/define* codes generally fall near the top of the graph, and *draft/share* codes generally fall near the bottom of the graph. Even though teams were working in multiple diamonds at any given time, gap tokens were generally more central at the beginning of the semester, and impact tokens were generally more central at the end of the semester.

These results show that we do not need to wait until the end of the semester to compare team behavior; we can track differences throughout the semester, meaning we have the potential to implement interventions for teams whose behavior is trending more towards the low-performing side.

#### **5.4. Research Question 3B: Effect of Added Structure on Student Behavior in IBL**

RQ3B asked: How does the structure of the course change student behavior in the context of co-occurrence?

##### **5.4.1. Methods**

RQ3B aimed to explore how the structure of the course changed student behavior. The same methods detailed in section 5.3 were used to plot the ENA results for students from all three cohorts. Averages were also plotted for the three cohorts and six sub-groups: Low-Performance 2019 (Cohort 1), High-Performance 2019 (Cohort 1), Low-Performance 2020 (Cohort 2), High-Performance 2020 (Cohort 2), Low-Performance 2021 (Cohort 3), and High-Performance 2021 (Cohort 3). To determine what action pairs differentiate between each of the three cohort groups and six performance sub-groups, the weights of the average networks were extracted and plotted.

During the analysis, it became clear that increased structure caused students to have more similar behavior. In order to quantify this observation, weighted network complexity (also called entropy) was calculated for all students. In the context of the IBL data, weighted network complexity relates to the diversity of action transitions. In other words, if you have just finished action type A, do you almost always progress to action type B (low weighted network complexity), or are you equally likely to progress to action type B, C, D, or E (high weighted network complexity)? Weighted network complexity has not been used in an ENA context, but it is commonly used



in other network analyses. This method was chosen over network density (more commonly used in ENA) because of how each of these methods are interpreted in the context of the IBL data. Weighted network density is calculated:

$$W = \sum_{i,j} \frac{(\Omega_{i,j})^2}{2}$$

where  $\Omega$  represents the association matrix of the network. This means that network density is greater when there are strong connections between a few nodes rather than weak connections between many nodes [89, 104]. However, for the IBL data, we want to quantify the *diversity* of connections. Thus, weighted network complexity is a more representative measure of the behavior we are looking for.

To calculate weighted network density [105], we start with the normalized association vector  $N$ . From  $N$ , we can calculate the entropy  $H$  of each node  $i$  of the network:

$$H(v_i) = - \sum_{j=1}^{d_i} p_{ij} \log_2(p_{ij})$$

where  $p_{ij}$  is calculated:

$$p_{ij} = \frac{w(v_i v_j)}{\sum_{j=1}^{d_i} w(v_i v_j)}$$

where  $w(v_i v_j)$  is the corresponding entry of the weight vector  $W$  and  $d$  is the dimension of the node (i.e. the number of other nodes it is connected to). If any of the edge weights are 0, those edges can simply be ignored in the calculation to prevent an undefined value for  $\log_2(p_{ij})$ .

To calculate entropy  $I$  of the entire network, sum the entropy of each of the network nodes:

$$I(G, w) = \sum_{k=1}^N H(v_k)$$

A worked example of calculating network complexity can be found in Appendix Section A.3.

### 5.4.2. Results and Analysis

First, all students from all three cohorts were plotted using ENA (Figure 5.9). Each circle represents the centroid for an individual student, and each square represents the centroid of the average network for a group (2019, 2020, 2021) or sub-group (Low 2019, High 2019, Low 2020, High 2020, Low 2021, High 2021). From Figure 5.9, it is clear that most of the variability in the data is linked to differences between the three cohorts because each of the cohorts is grouped in a similar location on the x-axis, which accounts for the first 18.2% of the variability in the data. The next chunk of variability seems to have some association with performance because low performers tend to be separated from high performers along the y-axis, which accounts for the next 7.8% of the variability in the data. However, these differences in performance are clear for Cohort 1, but much less so for Cohorts 2 and 3. This leads to some further questions: why are there such large differences in behavior among the cohorts, and why do performance groups separate only for Cohort 1?

To further explore the differences between each of the three cohorts, the average networks for each cohort were plotted in Figure 5.10 with Cohort 1 in Figure 5.10a, Cohort 2 in Figure 5.10b, and Cohort 3 in Figure 5.10c. Figure 5.11 shows the strengths of each of these connections quantitatively; the relative weight of each code is plotted for each pair of actions for Cohort 1 in Figure 5.11a, Cohort 2 in Figure 5.11b, and Cohort 3 in Figure 5.11c.

From these results, we can see that the connections between codes for Cohort 1 are fairly evenly distributed because no set structure was suggested to the students. For Cohort 2, over 80% of the co-occurrences involved a Survey or Define code (compared to just over 35% for Cohort 1 and just over 50% for Cohort 3) because students needed to “stack” their tokens starting with Survey and Define actions. Cohort 3 fell somewhere in between; there were more Survey and Define codes than seen in Cohort 1, but there were more diverse connections than seen in Cohort 2. This is most likely because Cohort 3 was introduced to the framework; Cohort 1 may not have been as focused on the gap because it was not explicitly stated that they should be *Surveying* and *Defining*.

However, unlike with Cohort 1, Cohort 3 students were more likely to start one type of action and then finish that type of action (e.g. *S.Survey* and *F.Survey*), or finish a type of action and then start the next type of action (e.g. *F.Explore* and *S.Solve*). In other words, Cohort 3’s

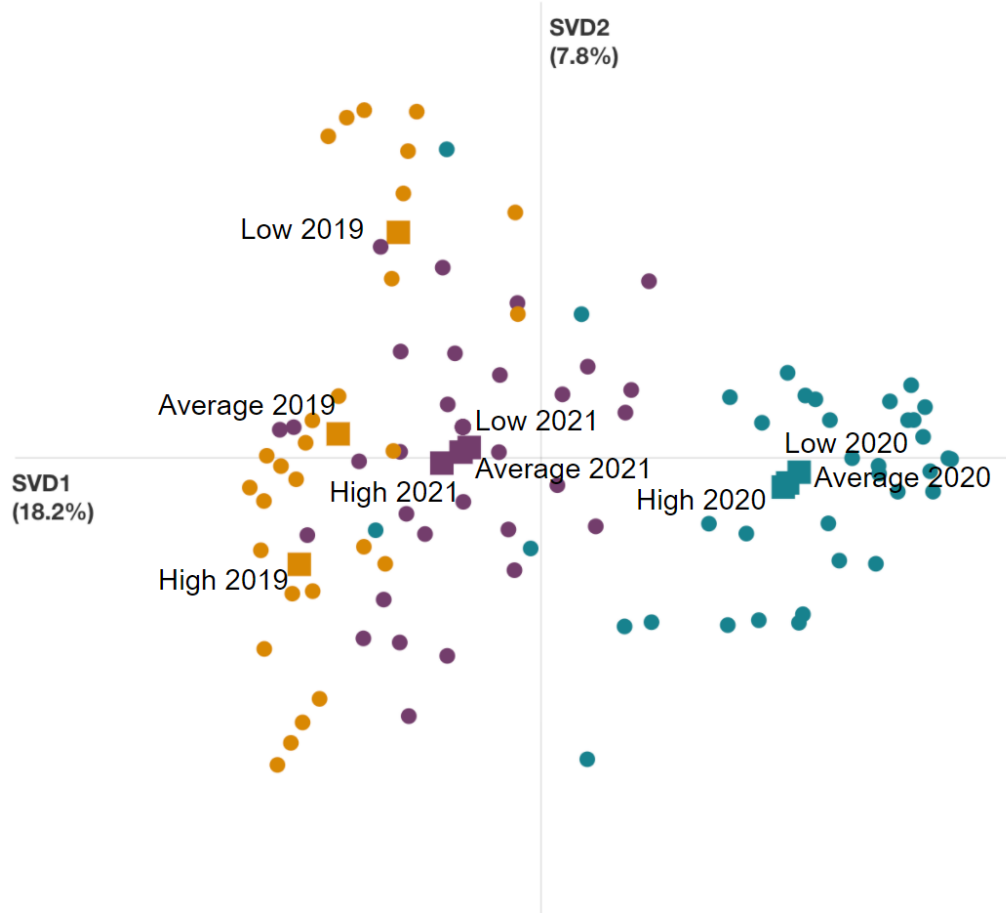
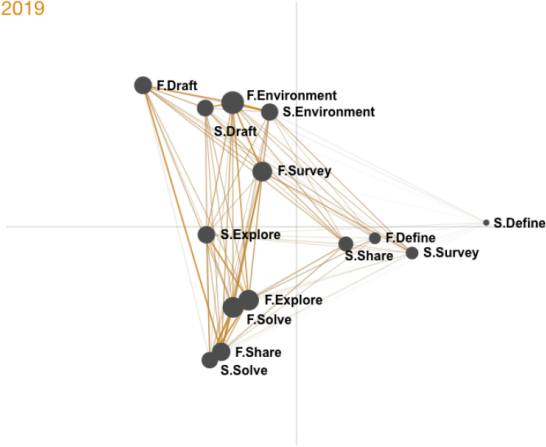


Figure 5.9. The plotted ENA centroids for all students across all 3 cohorts. Each circle represents a student, and labeled squares represent an average network for a group of students. The x-axis represents the first singular value decomposition and accounts for 18.2% of the variability in the data. Therefore, by noting that cohorts are grouped along the x-axis (2019 on the left, 2020 on the right, and 2021 in the middle), we can assume that much of the variability in the data is related to differences between the three cohorts. The y-axis represents the second singular value decomposition and accounts for the next 7.8% of variability in the data. On average, lower performing students are grouped near the top, and higher performing students are grouped near the bottom, suggesting that performance level also plays a large role in variability. It is also interesting to note the large distance between low- and high-performing groups for 2019 (whereas the averages for 2020 and 2021 have significant overlap).

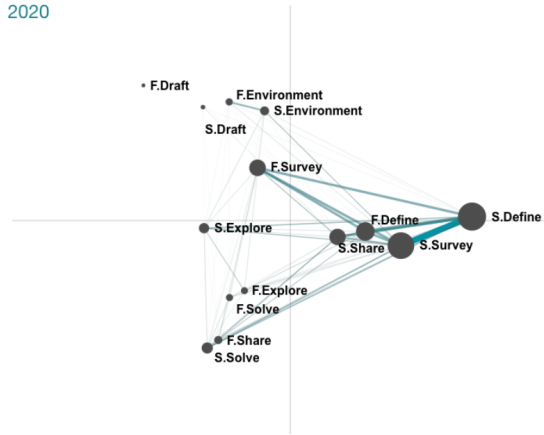
action pairs were more heavily weighted along the top diagonal of the tables shown in Figure 5.11. For Cohort 3, these types of action pairs accounted for about 25.0% of a student's connections on average, whereas they only accounted for about 13.7% of a student's connections on average for Cohort 1 ( $p < 0.05$ ). This may be because Cohort 3 students used the framework to more heavily structure their innovation process. However, it should be noted that this does *not* appear to be a

2019



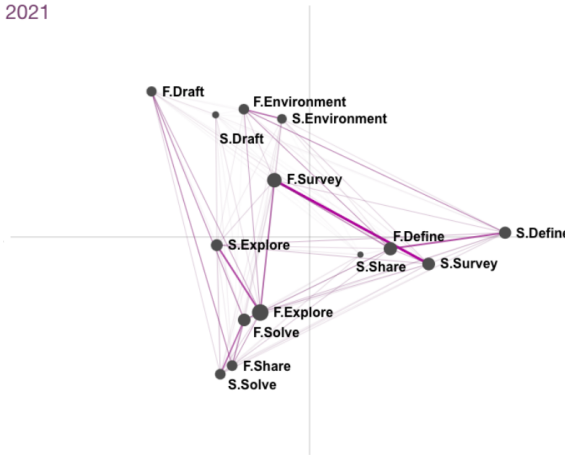
(a) Average Cohort 1 (2019)

2020



(b) Average Cohort 2 (2020)

2021



(c) Average Cohort 3 (2021)

Figure 5.10. The average networks for (a) Cohort 1 in 2019, (b) Cohort 2 in 2020, and (c) Cohort 3 in 2021. Each cohort varies in the strongest action pairs. Cohort 1’s connections are more evenly distributed; Cohort 2’s connections largely iterate between Survey and Define codes; Cohort 3’s connections have some variability, but largely iterate between start and finish codes of the same action type (e.g. S.Survey is highly connected to F.Survey). These results are quantitatively demonstrated in Figure 5.11, which shows the relative weight of each action pair for each of the three cohorts. We define complexity of innovative activity as the number of difference connections made. Using this definition, we can visually see that Cohort 1 (2019) has the highest complexity, Cohort 3 (2021) has the next highest complexity, and Cohort 2 (2020) has the lowest complexity. These results are quantitatively supported through the results presented in Table X. This change in complexity is not surprising because of the structure of the course for each of these years; 2019 was loosely structured, 2020 was heavily structured because of token “stacking”, and 2021 was somewhere in the middle.



Finally, to quantify the diversity of connections across cohorts and performance levels, network complexity was plotted using a box and whisker plot. These results are shown in Figure 5.12. From these results, we see that 2019 students had higher network complexity on average than 2020 and 2021 students ( $p < 0.01$ ). We also see that in 2021, high performers have higher network complexity than low performers ( $p < 0.01$ ). However, there is high variability among all groups, so complexity alone does not seem to predict performance.

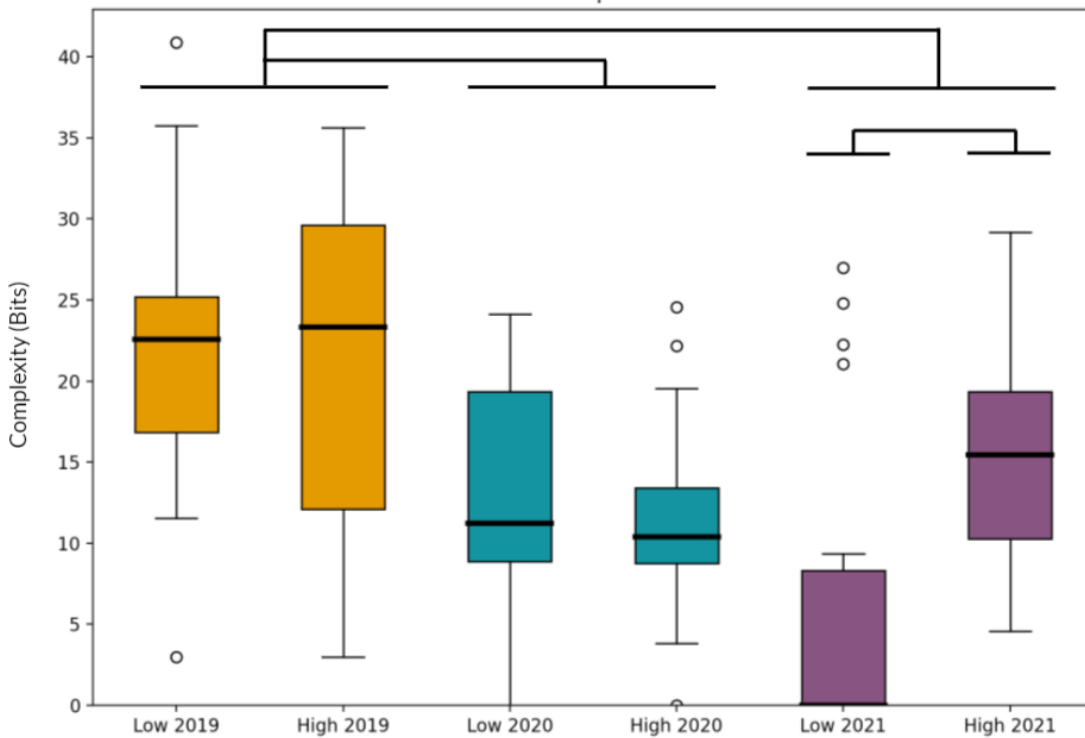


Figure 5.12. Box and whisker plots that show the ENA network complexity in bits for each of the six sub-groups: Low 2019, High 2019, Low 2020, High 2020, Low 2021, and High 2021. The bars represent statistical difference between groups and subgroups ( $p < 0.01$ ). There is a statistically significant difference between Low 2021 and High 2021, between All 2019 and All 2020, and All 2019 and All 2021. When given less structure (as in 2019), both low- and high-performing students had greater ENA network complexity. When given some open-ended structure (as in 2021), high-performing students had greater ENA network complexity than low-performing students.

These results can be summarized into three main findings: 1) when there is little structure, there are large differences in behavior between low and high performers (especially in regards to solution development), 2) when there is more structure, behavior becomes more uniform to meet the given structure (less complexity), and these differences do not differentiate between low and

high performers, and 3) when given medium structure, high performers tend to have more complex behavior than low performers. These results have some interesting alignment with research in organizational behavior about the effect of structure on creativity. [106] found that intuitive problem solvers (those that follow their gut) tend to be more creative than structured problem solvers (those that follow a specific plan or instructions) when given an open-ended problem. However, when offered a more structured problem, both intuitive and structured problem solvers were generally more creative, and the differences between the groups was reduced. However, the authors were careful to note that they were not stating that *all* structure is beneficial. [107] continued along this vein and found that structure can promote or inhibit creativity; it can help guide the creative process in some cases, but it can also cause cognitive fixation. Therefore, the continued challenge is to ensure that problems are given enough structure, but few requirements are placed on the problem-solving process.

### **5.5. Implications for Teaching**

Based off of the ENA results and the existing literature on creativity and structure, we offer a few suggestions for those teaching innovation: 1) consider what behaviors you want students to take part in, 2) examine how the structure of your course and your communication of that structure either encourage or discourage these desired behaviors, and 3) help students structure their problem, but not necessarily their problem-solving process.

By looking at the behavior of high-performing students, we can see that *explore* and *solve* codes are much more core to networks when compared to low-performing students. The high centrality of these nodes not only means that high performers completed these actions more, but they also returned to these actions often. Work in the “gap” diamond and “impact” diamond often led to new actions in the “solution” diamond. Therefore, it is important to encourage students to continue to cycle through the *explore* and *solve* triangles and consider how their other findings can help drive their solution process forward.

Next, it is important to carefully consider how the structure of the course and the way you communicate that structure either encourage or discourage these desired behaviors. Cohort 1 showed us that solution development was more central to the networks of high performers, so our course structure should encourage this behavior. However, the structure implemented for Cohort 2 failed to do this; the structure focused heavily on gap identification and connecting it to the pillars,

and this incentivized behavior is clearly seen in the resulting student behavior networks. Although students may have still been participating in solution development, those actions are now invisible to the instructors because they are not being logged in the platform. The implementation of the framework for Cohort 3, on the other hand, increased visibility of all stages of the framework by encouraging students to spend time and effort in all seven action types.

However, the framework carefully straddles the line between helping students structure their problem and over-structuring their problem-solving process. We argue that the framework can be used as an appropriate form of structure according to the findings of [106] and [107] because it provides scoping and goals for the students (identify a gap, develop a solution, and create impact). However, if a student assumes that the framework is an ordered process, that could lead to inappropriate structure according to the findings of [106] and [107] because it limits the students to a specific plan and increases cognitive fixation on the next step. This could be an explanation for the differences in network complexity seen between low performers and high performers; low performers had lower network complexity on average because they treated the framework as an ordered recipe whereas high performers had higher network complexity on average because they treated the framework as an unordered guide. However, high variability in complexity is present for both performance groups; a student can still have a very structured process and create high value. However, we argue that it is imperative to not over-prescribe structure; it should be available for students to use as a resource, but it should not create a required pathway.

Although these findings and suggestions align with existing literature on creativity, we recognize the complexity of the relationships between structure and innovation. There are dangers both in having low structure and high structure, so more work will need to be done to find the structure “sweet spot”, whether it is a “one-size-fits-all” approach or a more personalized approach.

## **5.6. Implications for Research**

This work also offers various implications for research in innovation – while also addressing the five limitations of ENA identified by [100] and presented in Section 5.2.4.

The first contribution of the work is that it successfully bridges qualitative and quantitative results to measure and understand differences between groups (both by performance and by cohort). The 14 action types led to 91 action pairs, and this high dimension space was successfully reduced to a representative two-dimensional space, addressing the visualization of large networks limitation



identified by [100]. The quantitative relationships limitation identified by [100] was addressed by reporting quantitative results where appropriate (e.g. statistical significance, goodness of fit, and the relative straights of each action pair). The interpretability limitation identified by [100] was addressed by labeling the axes of the ENA representations, allowing a reader to extract meaning from the plots. For example, for the trajectory plots, a point in the bottom-left quadrant represents higher importance placed on *explore* and *solve* codes, as well as *draft* and *share* codes. This interpretability could be even further improved by implementing a tool similar to the one presented in [108]; rather than using the SVD to plot points, nodes are plotted further apart and in a way that still separates groups of importance. Because the nodes are not moving based on new data, teachers can gain meaningful information with a quick glance.

The second contribution of this work is that it is a scalable solution that can be analyzed in real time. Because of the automatic framework classifier developed in Chapter 3 and the code developed to convert raw MOOCIBL logs to ENA files, a researcher or instructor could visualize student or team behavior in seconds. This could support instructors as they aim to identify which teams to check in with first throughout the semester. In addition, it allows them to monitor how changes in the course are affecting student behavior. Many ENA applications still require manual coding, but this process addresses the automation and scalability limitation identified by [100].

Finally, ENA allowed us to consider new lenses that were not possible before. Rather than just considering individual students and their behavior over the whole semester, ENA allows for team comparisons and trajectory comparisons. [100] identified trajectory as being a limitation of ENA because it is very rarely reported in the ENA literature, but this work shows that trajectories can be modeled in IBL settings. However, it should be noted that this limitation was not fully addressed in this work (or by the ENA research community as a whole). The trajectory process still relies on aggregated networks of certain periods (i.e. each quarter of the semester is its own network). This reduces the full longitudinal aspect of the data.

ENA has shown to be an appropriate method for analyzing IBL data, and this work carefully considered and addressed previously identified limitations of ENA. In addition, this work also shows the power of the IBL framework. Without the framework, ENA would not have been possible.

## 5.7. Summary

This chapter detailed the implementation of epistemic network analysis for analysis of IBL data to better understand student success and how the structure of the course impacts student behavior.

To answer RQ3A, epistemic network analysis was performed on students, teams, and team trajectories for Cohort 1 (the semester with little structure). These networks were then aggregated to compare the average networks for both low- and high-performing students and teams. When students were given little structure, epistemic network analysis greatly differentiated between those marked low-performing and those marked high-performing. Students and teams that were high-performing generally focused on solution development and students and teams that were low-performing generally focused on *environment* tokens.

To answer RQ3B, epistemic network analysis was performed on students from all three cohorts, and it was found that most of the variability in the data was because of changes from year to year. By analyzing the differences in behavior between each year, we were able to tie the structure of the course that year to student behavior. Cohort 1 saw high variability because of the low structure, Cohort 2 saw low variability and a focus on *survey* and *define* because of the high structure, and Cohort 3 saw some variability but some increased structure because of the implementation of the IBL framework.

Although the changes from year to year and the complex nature of this course make it challenging to draw any hard conclusions, this work demonstrates that epistemic network analysis can be used to analyze student behavior and determine the effects of implementing various changes in the course.

## 6. DISCUSSION

Chapters 3, 4, and 5 took three different approaches to understanding innovation. Chapter 3 used qualitative analysis driven by literature, Chapter 4 used linear models to reduce innovation to key words and features, and Chapter 5 took a mixed methods approach to understand innovation as a holistic process. Even with the very different approaches, recurring themes emerge across each of the three studies. Overarching insights, limitations, challenges to implementation, and future directions will be shared.

### 6.1. Insights

The results across Chapters 3, 4, and 5 lead to three main overarching insights: the ability of language and temporal behavior to provide meaningful information about students' approaches to innovation, the challenges of researching and teaching innovation, and the demonstrated potential for using LA/EDM methods to overcome these challenges in IBL environments.

The first major insight is further understanding of how language and temporal behavior provide information about students' approaches to innovation. From the classification results, we saw that the words that students choose when writing titles and descriptions for their tokens can help predict their success level. The performance further improves when the language is placed within the context of the IBL framework. This increased performance and the feature extraction results show that words can take on different contexts and implications depending on the category of the IBL framework. This finding can help instructors better understand how students' perceptions of the categories and deliverables might differ from those of the instructors and promote meaningful conversation about how to make project progress and create value. From the clustering work, we saw that certain quantitative behavior patterns can be identified – some that are more likely to lead to low performance, some that are more likely to lead to high performance, and some that are not predictive of performance. Before implementing the IBL framework, the total number of tokens that a student had provided little information about student progress. By looking at category breakdowns rather than total number of tokens, instructors can gain more meaningful information about how a student is doing. For example, without the framework, an instructor would see two students with ten tokens each; with the framework, an instructor would see one student with a mix

of tokens from all categories and one student with mostly *environment* tokens. This differentiation can help an instructor better understand student progress and provide more targeted feedback and support. From the epistemic network analysis work, we are able to see how the structure of the course can change students' temporal behavior. When little structure was put in place, students exhibited a wide variety of action pairs (high complexity). When more structure was put in place, behavior shifted to meet this structure; students were more likely to finish an action they started before continuing to the next action, and they often progressed through these actions in a linear way (e.g. moving from one framework category to the next). Because students adjust their logged actions to follow the course structure, instructors should carefully consider what desired behavior looks like and design the structure to fit those goals in a way that guides and supports without limiting student freedom.

Although these results provided new insights about student behavior, they also illustrated the challenges that come with researching and teaching innovation. When creating classification models, the IBL framework improved the prediction models, but they were still far from perfect, especially when trying to predict performance on new cohorts. The clustering models gave even further insight into the challenges of prediction; students with very similar quantitative behavior could end up with different performance outcomes. Finally, when reviewing the results of the epistemic network analysis work, we saw just how much the year-to-year changes can impact student behavior. However, questions still remain about the full implications of these differences. The breakdown of student performance in Cohort 1 and Cohort 2 was almost identical, even though Cohort 1 had very little structure and Cohort 2 had very high structure; does that mean both of these approaches were equally justified? Was either approach more equitable than the other? Did either approach lead to improved student experience? Which created a more authentic learning experience? On the other hand, there were significantly fewer high performers in Cohort 3. Was this due to the introduction of the framework? Was it due to the continued impacts of COVID-19 on education? Or was it due to something else altogether, or some combination of all of the above? Although we are continuing to collect data to try to get to the core of these questions, they rely heavily on the interconnectedness of various social factors and contextual factors that arise with each new cohort.

Although researching and teaching IBL brings unique challenges, this work also further illustrates the demonstrated potential of using LA/EDM tools to better understand these challenges. The benefits of using LA/EDM tools in this context include scalability and speed, interpretability and visualization, and a blend of qualitative and quantitative work. The developed code is made up of a variety of code building blocks that can be put together to explore many different views with only minor changes. Even though there were changes to the MOOCIBL system each year, a workflow was developed that converts these unique logs into pre-processed data files that can then be input into a variety of other functions for classification, clustering, and epistemic network analysis. This workflow provides researchers with both scalability and speed. Next, the LA/EDM work placed focus on interpretability and visualization capabilities to better support data-driven instruction. For classification, linear models were chosen over black-box models so features could be extracted and interpreted by researchers or instructors. Clustering and epistemic network analysis allow us to reduce highly dimensional datasets into two- or three-dimensional datasets that are able to be visualized, but also interpreted. With clustering, the dendrograms give researchers information about similar students, and the three dimensional cluster plot give instructors information about quantitative student behavior. With epistemic network analysis, both researchers and instructors can use the two-dimensional space to identify similar students and teams and use the axes to give the two-dimensional space interpretable meaning. By combining the scalability and speed of these algorithms with the understanding and expertise of an instructor or discipline-based education researcher, we can use LA/EDM methods to blend both qualitative and quantitative work.

## **6.2. Limitations**

Although this work has shown promising results and insights, there are still a variety of limitations in scope related to the dataset, the study population, and the methods/tools.

One limitation of the work is the nature of the current dataset. Because students self-report their learning in the platform, we are only able to analyze intentional, conscious actions that students choose to log. Research in psychology and cognitive science suggests that conscious actions and ideas only play a small role in the process of developing new and innovative ideas; subliminal thought and even serendipitous mistakes have also shown to be key in finding innovative success [109, 110]. Not only does MOOCIBL offer few opportunities to gain insights into these subliminal

thoughts and events of serendipity, it is also limited to the conscious acts and behaviors that students choose to record. Students generally use MOOCIBL as a way to communicate progress with their instructor, so the recorded actions summarize work in a way that is curated to what they believe the instructor wants to see. Because student perception about what the instructor wants can change from student to student, and even more so from year to year, it becomes almost impossible to use MOOCIBL behavior as a direct representation of the authentic innovation process, leading to challenges in drawing conclusions. A similar limitation is the time scale that the work focuses on. Human behavior takes place across a variety of time scales: the biological band (milliseconds or tens of milliseconds level), the cognitive band (seconds or tens of seconds), the rational band (hours or days), and the social band (weeks or months) [111]. The work presented largely relies on discovery in the rational band; MOOCIBL tracks tasks that students log every few days. However, to get a full picture of IBL, other bands should be explored. The lower bands (biological and cognitive) could answer questions about where ideas come from, the neuroscience of creativity, and how serendipity plays a role in the development of new innovations. The higher band (social) could further our understanding of team dynamics and how ideas and culture emerges from an entire social system.

Another limitation of the work is the study population and the time frame that the study was conducted during. Although there was some racial and gender diversity, the sample size was too small to draw any conclusions about how race or gender may impact student experience. In addition, the participants were almost entirely upper division undergraduate or graduate students who chose to take the course as an elective. Therefore, it is still unclear how these findings would transfer to new populations and settings (e.g. lower division students, core engineering courses, etc.) Similarly, the context of the time frame of the study should not be ignored, especially the impacts of the COVID-19 pandemic. Cohort 1 took the course in fall 2019, Cohort 2 in fall 2020, and Cohort 3 in fall 2021.

Finally, the work is limited because we are trying to work on complex problems while still trying to fill our toolbox. Although advances have been made in modeling complex systems both on the information science [112] and educational research [113] fronts, this area of work has not had time to grow and mature like many traditional qualitative and quantitative modeling methods. The field of complexity science is still working to overcome challenges such as the nature

of complex causality, the evolution and durability of new knowledge, and the defining of meaningful and appropriate boundaries. For example, one of our current tools – feature extraction – does poorly with the nature of complex causality because it makes a reductionist assumption that certain words are more likely to lead to low performance. Although a human understands that using that word is not inherently unproductive to innovation, we lack LA/EDM tools that are similarly able to draw only appropriate conclusions about causality from complex processes. In regards to the evolution and durability of new knowledge, we lack tools that account for changes as they occur (e.g. the COVID-19 pandemic). Because this was an unprecedented condition, there was no way to adjust our model to account for this change in the context of the course. Finally, in regards to defining meaningful and appropriate boundaries, we still lack tools that can account for the interconnected and hierarchical relationships in the course. Using epistemic network analysis at the team level instead of the student level started to account for the relationships between students, but we still lack tools that allow us to fully understand the dynamic and interconnected nature of the relationships between team members and classmates.

### **6.3. Challenges to Implementation**

In addition to the presented limitations, there are also a variety of challenges that must be overcome in order to integrate this work into the real-time experience of students and instructors in the course. These models have the potential for instructors to run classroom reports in real time to identify students at risk, gauge progress, and identify emerging trends in behavior. However, for this to be a practical solution, these models will need to be integrated directly into the learning management system to ensure that reports can be run quickly and easily.

Similarly, to support instructors using these tools, it will be important to design ways to interface with the data that are intuitive and helpful. The tools have been designed with some of these ideas in mind (e.g. feature extraction for classification, visualization tools for clustering, and defining the axes for epistemic network analysis), but more will need to be done to promote buy-in and trust from instructors.

Another challenge of implementation is considering how these models will be updated. The use of the framework creates a consistent lens to use across students, teams, and years, but continuing to adapt our models will be a challenge. As the world changes, so will our students and the projects they work on. Therefore, the algorithms will need flexibility, and questions still

remain about the best way to implement this. For example, should recent data be considered more strongly when making predictions for new students? Does data become obsolete? How will these models continue to be adapted and verified over time?

In addition, it will continue to be important to carefully consider the ethics and equity of implementing machine learning in the classroom. Although this work can be a helpful tool for instructors, it is imperative that we do not rely solely on the algorithm to identify students at risk or assess students, and we must continue to assess the validity and reliability of our results. It is also important to use a research lens that allows us to identify if any specific types of students are being categorized in an inequitable way.

#### **6.4. Future Directions**

To overcome these limitations and challenges, we propose a variety of future research directions both in the short and long term.

In the short term, this work can be applied to new cohorts as IBL continues to go to other institutions. Not only will this lead to a greater sample size, but it also can increase the diversity of the contexts and populations represented. Short term future work could also include integrating this work directly into the MOOCIBL platform to support new students and instructors. By identifying the wants and needs of each of these stakeholders, we can leverage the presented work to create “plug and play” tools that meet these wants and needs. Currently, MOOCIBL has a dashboard that shows number of tokens in each stage of the process, but this could be expanded to include other information from the classification, clustering, or epistemic network analysis methods. Similarly, this work could also be used to support evidence-based design choices for both the structure of the course and MOOCIBL. For example, we should aim to create a classroom structure and MOOCIBL interface that incentivizes progress in all of the categories of the framework while still allowing for student flexibility.

In the long term, there are many other lenses that could be used to better understand how to support teaching and learning innovation. There is a greater push in engineering education and beyond to promote innovation skills in students, but questions still remain about how to scaffold and support the learning process. Future research directions could include interviews to better understand the affordances and limitations of various forms of student support, for example. In addition, as mentioned in the limitations, most of the work in this area has been focused on the



rational time band, but other bands should be considered to get a fuller picture of the science of innovation. This future work could involve moving down the time scale to better understand where innovative ideas emerge from or moving up the time scale to better understand how students, teams, and instructors interact – as well as how the world outside the classroom interacts with the course. Many instructors have noticed large changes in the classroom throughout the COVID-19 pandemic, but we are still struggling to understand all of the underlying interactions between the state of our world and the classroom. Finally, arguably the biggest future direction and challenge will be to identify and create methods for modeling and extracting meaning from complex systems – both in education environments and beyond.

## 7. CONCLUSION

This work used learning analytics and educational data mining techniques to further understand the complexity of student learning and innovation. As mentioned in the literature review, Snowden argues that those working in complex environments should “probe, sense, and respond”, which allows us to embrace the complexity that emerges.

The creation of the IBL framework allowed us to “probe”. The categories of the framework give us a lens that allows us to compare across students, teams, and cohorts – even when students are working on different components of different projects with different deliverables and different pieces of learning.

The use of LA/EDM tools allowed us to “sense”. Every student can be placed into one of several high-dimension spaces, whether the dimensions represent words written, types of actions, or team trajectories. The LA/EDM tools allow us to reduce those high-dimensional spaces into ones that we can interpret and provide context to. With classification, we were able to extract performance metrics and features of importance; with clustering, we were able to extract groupings and hierarchical visualizations; with epistemic network analysis, we were able to extract measures of complexity and network visualizations. Each of these studies allows us to paint a fuller picture of the innovation process and student experience.

So now we “respond”. We have heard the calls from the World Economic Forum, ABET, and the National Academy of Engineering for innovative engineers, and this work illustrates just some of the challenges we face to meet this call. However, it also illustrates how the LA/EDM tools that were developed can help us overcome these challenges. The findings can continue to promote meaningful discussion about creating classroom spaces for innovation, and the tools can be used to collect and analyze data for years to come. The work presented in this dissertation has laid important groundwork, and its impact is only just beginning.

## REFERENCES

- [1] World Economic Forum, “The future of jobs report 2020.” World Economic Forum, Geneva, Switzerland, 2020.
- [2] National Academy of Engineering, *The engineer of 2020: Visions of engineering in the new century*. National Academies Press, 2004.
- [3] ABET, “Criteria for accrediting engineering programs.” [Online]. Available: <https://www.abet.org/accreditation/accreditation-criteria/>
- [4] H. J. Passow and C. H. Passow, “What competencies should undergraduate engineering programs emphasize? a systematic review,” *Journal of Engineering Education*, vol. 106, no. 3, pp. 475–526, 2017.
- [5] A. H. Van de Ven, “The innovation journey: you can’t control it, but you can learn to maneuver it,” *Innovation*, vol. 19, no. 1, pp. 39–42, 2017.
- [6] L. N. Singelmann and D. L. Ewert, “Leveraging the innovation-based learning framework to predict and understand student success in innovation,” *IEEE Access*, 2022.
- [7] M. Pearson, R. Striker, E. M. Swartz, E. A. Vazquez, L. Singelmann, and S. S. Ng, “Innovation-based learning: A new way to educate innovation,” in *2021 ASEE Virtual Annual Conference Content Access*, 2021.
- [8] A. Baregheh, J. Rowley, and S. Sambrook, “Towards a multidisciplinary definition of innovation,” *Management decision*, 2009.
- [9] E. M. Swartz, R. Striker, L. Singelmann, E. A. Vazquez, M. Pearson, and S. S. Ng, “Innovating assessment: Using innovative impact as a metric to evaluate student outcomes in an innovation-based learning course,” in *2021 ASEE Virtual Annual Conference Content Access*, 2021.

- [10] E. A. Vazquez, M. Pearson, L. Singelmann, R. Striker, and E. Swartz, “Federal funding opportunity announcements as a catalyst of students’ projects in mooc environments,” in *2019 IEEE Learning With MOOCS (LWMOOCS)*. IEEE, 2019, pp. 79–83.
- [11] R. Striker, E. A. Vazquez, M. Pearson, L. Singelmann, and E. M. Swartz, “A scalable approach to student team formation for innovation-based learning,” in *2020 ASEE Virtual Annual Conference Content Access*, 2020.
- [12] L. Singelmann, E. Alvarez Vazquez, E. Swartz, M. Pearson, and R. Striker, “Student-developed learning objectives: A form of assessment to promote professional growth,” in *American Society for Engineering Education Annual Conference*. ASEE, 2020.
- [13] E. A. Vazquez, R. Striker, L. Singelmann, M. Pearson, E. M. Swartz, S. S. Ng, and D. Ewert, “The moocibl platform: a custom-made software solution to track the innovation process with blockchain learning tokens,” in *2021 ASEE Virtual Annual Conference Content Access*, 2021.
- [14] L. Singelmann, E. Swartz, M. Pearson, R. Striker, and E. Alvarez Vazquez, “Design and development of a machine learning tool for an innovation-based learning mooc,” in *6th International Conference on Learning with MOOCs*. IEEE, 2019.
- [15] *1st International Conference on Learning Analytics*, 2011.
- [16] C. Romero and S. Ventura, “Educational data mining: a review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.
- [17] —, “Data mining in education,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 3, no. 1, pp. 12–27, 2013.
- [18] —, “Educational data mining and learning analytics: An updated survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.
- [19] L. Calvet Liñán and Á. A. Juan Pérez, “Educational data mining and learning analytics: differences, similarities, and time evolution,” *International Journal of Educational Technology in Higher Education*, vol. 12, no. 3, pp. 98–112, 2015.

- [20] G. Siemens and R. S. d. Baker, “Learning analytics and educational data mining: towards communication and collaboration,” in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 252–254.
- [21] National Research Council and others, *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. National Academies Press, 2012.
- [22] L. Singelmann, E. Alvarez Vazquez, E. Swartz, R. Stiker, M. Pearson, and D. Ewert, “Predicting and understanding success in an innovation-based learning course,” in *Educational Data Mining 2020 Conference*. IEEE, 2020.
- [23] —, “Innovators, learners, and surveyors: Clustering students in an innovation-based learning course,” in *Frontiers in Education 2020 Conference*. IEEE, 2020.
- [24] L. N. Singelmann, “Using classification and clustering to predict and understand student behavior in an innovation-based learning course,” Ph.D. dissertation, North Dakota State University, 2020.
- [25] D. J. Snowden and M. E. Boone, “A leader’s framework for decision making,” *Harvard business review*, vol. 85, no. 11, p. 68, 2007.
- [26] T. J. Howard, S. J. Culley, and E. Dekoninck, “Describing the creative design process by the integration of engineering design and cognitive psychology literature,” *Design studies*, vol. 29, no. 2, pp. 160–180, 2008.
- [27] “Engineering design process.” [Online]. Available: <https://www.teachengineering.org/design/designprocess>
- [28] T. J. Moore, A. W. Glancy, K. M. Tank, J. A. Kersten, K. A. Smith, and M. S. Stohlmann, “A framework for quality k-12 engineering education: Research and development,” *Journal of pre-college engineering education research (J-PEER)*, vol. 4, no. 1, p. 2, 2014.
- [29] Z. J. Acs, L. Anselin, and A. Varga, “Patents and innovation counts as measures of regional production of new knowledge,” *Research policy*, vol. 31, no. 7, pp. 1069–1085, 2002.

- [30] M. A. Fleuren, T. G. Paulussen, P. Van Dommelen, and S. Van Buuren, "Towards a measurement instrument for determinants of innovations," *International Journal for Quality in Health Care*, vol. 26, no. 5, pp. 501–510, 2014.
- [31] R. Kasa, "Approximating innovation potential with neurofuzzy robust model," *Investigaciones Europeas de Dirección y Economía de la Empresa*, vol. 21, no. 1, pp. 35–46, 2015.
- [32] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM— Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
- [33] W. Xing, B. Pei, S. Li, G. Chen, and C. Xie, "Using learning analytics to support students' engineering design: the angle of prediction," *Interactive Learning Environments*, pp. 1–18, 2019.
- [34] C. Vieira, M. H. Goldstein, Ş. Purzer, and A. J. Magana, "Using learning analytics to characterize student experimentation strategies in the context of engineering design," *Journal of Learning Analytics*, vol. 3, no. 3, pp. 291–317, 2016.
- [35] C. J. Atman, R. S. Adams, M. E. Cardella, J. Turns, S. Mosborg, and J. Saleem, "Engineering design processes: A comparison of students and expert practitioners," *Journal of engineering education*, vol. 96, no. 4, pp. 359–379, 2007.
- [36] M. Dziallas and K. Blind, "Innovation indicators throughout the innovation process: An extensive literature analysis," *Technovation*, vol. 80, pp. 3–29, 2019.
- [37] K. B. Kahn, "Understanding innovation," *Business Horizons*, vol. 61, no. 3, pp. 453–460, 2018.
- [38] B. A. Vojak, R. L. Price, and A. Griffin, "Serial innovators: How individuals create and deliver breakthrough innovations in mature firms," *Research-Technology Management*, vol. 55, no. 6, pp. 42–48, 2012.
- [39] T. Edwards, R. Delbridge, and M. Munday, "Understanding innovation in small and medium-sized enterprises: a process manifest," *Technovation*, vol. 25, no. 10, pp. 1119–1127, 2005.

- [40] G. Gellatly and V. Peters, “Understanding the innovation process: Innovation in dynamic service industries,” *Statistics Canada Working Paper*, no. 127, 1999.
- [41] M. Berland, R. S. Baker, and P. Blikstein, “Educational data mining and learning analytics: Applications to constructionist research,” *Technology, Knowledge and Learning*, vol. 19, no. 1-2, pp. 205–220, 2014.
- [42] P. A. Frensch and J. Funke, *Complex problem solving: The European perspective*. Psychology Press, 2014.
- [43] J. Funke, “Dynamic systems as tools for analysing human judgement,” *Thinking & reasoning*, vol. 7, no. 1, pp. 69–89, 2001.
- [44] S. Greiff and A. Fischer, “Measuring complex problem solving: An educational application of psychological theories,” *Journal for Educational Research Online= Journal für Bildungsforschung Online*, vol. 5, pp. 34–53, 2013.
- [45] C. K. Martin, D. Nacu, and N. Pinkard, “Revealing opportunities for 21st century learning: An approach to interpreting user trace log data.” *Journal of Learning Analytics*, vol. 3, no. 2, pp. 37–87, 2016.
- [46] P. Blikstein and M. Worsley, “Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks,” *Journal of Learning Analytics*, vol. 3, no. 2, pp. 220–238, 2016.
- [47] D. Spikol, E. Ruffaldi, L. Landolfi, and M. Cukurova, “Estimation of success in collaborative learning based on multimodal learning analytics features,” in *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 2017, pp. 269–273.
- [48] J. Zhang and B. Chen, “Analytics for knowledge creation: Towards epistemic agency and design-mode thinking,” 2016.
- [49] P. J. Giabbanelli, A. A. Tawfik, and V. K. Gupta, “Learning analytics to support teachers’ assessment of problem solving: A novel application for machine learning and graph algorithms,” *Utilizing learning analytics to support study success*, pp. 175–199, 2019.

- [50] D. W. Shaffer, W. Collier, and A. R. Ruis, “A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data,” *Journal of Learning Analytics*, vol. 3, no. 3, pp. 9–45, 2016.
- [51] R. Kaliisa, K. Misiejuk, G. A. Irgens, and M. Misfeldt, “Scoping the emerging field of quantitative ethnography: Opportunities, challenges and future directions,” in *International Conference on Quantitative Ethnography*. Springer, 2021, pp. 3–17.
- [52] A. Csanadi, B. Eagan, I. Kollar, D. W. Shaffer, and F. Fischer, “When coding-and-counting is not enough: using epistemic network analysis (ena) to analyze verbal data in cscl research,” *International Journal of Computer-Supported Collaborative Learning*, vol. 13, no. 4, pp. 419–438, 2018.
- [53] B. Eagan and E. Hamilton, “Epistemic network analysis of an international digital makerspace in africa, europe, and the us,” in *annual meeting of the American education research association*, 2018.
- [54] R.-A. Thietart and B. Forgues, “Chaos theory and organization,” *Organization science*, vol. 6, no. 1, pp. 19–31, 1995.
- [55] A. H. Van de Ven and M. S. Poole, “Methods for studying innovation development in the minnesota innovation research program,” *Organization science*, vol. 1, no. 3, pp. 313–335, 1990.
- [56] Y.-T. Cheng and A. H. Van de Ven, “Learning the innovation journey: order out of chaos?” *Organization science*, vol. 7, no. 6, pp. 593–614, 1996.
- [57] G. Boeing, “Visual analysis of nonlinear dynamical systems: chaos, fractals, self-similarity and the limits of prediction,” *Systems*, vol. 4, no. 4, p. 37, 2016.
- [58] J. Buijs, “Modelling product innovation processes, from linear logic to circular chaos,” *Creativity and innovation management*, vol. 12, no. 2, pp. 76–93, 2003.
- [59] R. S. Baker, D. Gašević, and S. Karumbaiah, “Four paradigms in learning analytics: Why paradigm convergence matters,” *Computers and Education: Artificial Intelligence*, p. 100021, 2021.



- [60] L. Singelmann, R. Striker, E. A. Vazquez, E. Swartz, M. Pearson, S. S. Ng, and D. Ewert, “Creation of a framework that integrates technical innovation and learning in engineering,” © 2021 IEEE. Reprinted, with permission.
- [61] G. O’Neill and F. Murphy, “Guide to taxonomies of learning,” <https://www.ucd.ie/t4cms/media,52160,en..pdf>, 2010.
- [62] D. R. Krathwohl and L. W. Anderson, *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. Longman, 2009.
- [63] A. J. Swart, “Evaluation of final examination papers in engineering: A case study using bloom’s taxonomy,” *IEEE Transactions on Education*, vol. 53, no. 2, pp. 257–264, 2009.
- [64] N. N. Khairuddin and K. Hashim, “Application of bloom’s taxonomy in software engineering assessments,” in *Proceedings of the 8th WSEAS International Conference on Applied Computer Science*, 2008, pp. 66–69.
- [65] A. Gibson, K. Kitto, and J. Willis, “A cognitive processing framework for learning analytics,” in *Proceedings of the fourth international conference on learning analytics and knowledge*, 2014, pp. 212–216.
- [66] S. M. Ranade and A. Corrales, “Teaching problem solving: Don’t forget the problem solver (s),” *European Journal of Engineering Education*, vol. 38, no. 2, pp. 131–140, 2013.
- [67] J. Andrews-Todd and C. M. Forsyth, “Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task,” *Computers in human behavior*, vol. 104, p. 105759, 2020.
- [68] M. Boekaerts, “Self-regulated learning: Where we are today,” *International journal of educational research*, vol. 31, no. 6, pp. 445–457, 1999.
- [69] E. Panadero, “A review of self-regulated learning: Six models and four directions for research,” *Frontiers in psychology*, vol. 8, p. 422, 2017.
- [70] B. J. Zimmerman, “Self-regulated learning and academic achievement: An overview,” *Educational psychologist*, vol. 25, no. 1, pp. 3–17, 1990.

- [71] S. Li, G. Chen, W. Xing, J. Zheng, and C. Xie, “Longitudinal clustering of students’ self-regulated learning behaviors in engineering design,” *Computers & Education*, vol. 153, p. 103899, 2020.
- [72] The Design Council, “The design process: What is the double diamond,” <https://www.designcouncil.org.uk/news-opinion/design-process-what-double-diamond>, accessed: 2018-11-21.
- [73] S. Papert and I. Harel, “Situating constructionism,” *Constructionism*, vol. 36, no. 2, pp. 1–11, 1991.
- [74] A. Langley, “Strategies for theorizing from process data,” *Academy of Management review*, vol. 24, no. 4, pp. 691–710, 1999.
- [75] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.
- [76] R. M. Felder and R. Brent, “Designing and teaching courses to satisfy the abet engineering criteria,” *Journal of Engineering Education*, vol. 92, no. 1, pp. 7–25, 2003.
- [77] W. Hämmäläinen and M. Vinni, “Classifiers for educational data mining,” *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, pp. 57–71, 2011.
- [78] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.
- [79] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [80] S. Taheri and G. Hesamian, “A generalization of the wilcoxon signed-rank test and its applications,” *Statistical Papers*, vol. 54, no. 2, pp. 457–470, 2013.

- [81] T. G. Dietterich, “Approximate statistical tests for comparing supervised classification learning algorithms,” *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.
- [82] R. S. Baker and P. S. Inventado, “Educational data mining and learning analytics,” in *Learning analytics*. Springer, 2014, pp. 61–75.
- [83] J. Huang and C. X. Ling, “Using auc and accuracy in evaluating learning algorithms,” *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.
- [84] R. L. Harrison, “Introduction to monte carlo simulation,” in *AIP conference proceedings*, vol. 1204, no. 1. American Institute of Physics, 2010, pp. 17–21.
- [85] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, 2020.
- [86] J. H. Ward Jr, “Hierarchical grouping to optimize an objective function,” *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.
- [87] S. D. Blum and A. Kohn, *Ungrading: Why Rating Students Undermines Learning (And What to Do Instead)*. West Virginia University Press, 2020.
- [88] S. Thurner, R. Hanel, and P. Klimek, *Introduction to the theory of complex systems*. Oxford University Press, 2018.
- [89] D. W. Shaffer, D. Hatfield, G. N. Svarovsky, P. Nash, A. Nulty, E. Bagley, K. Frank, A. A. Rupp, and R. Mislevy, “Epistemic network analysis: A prototype for 21st-century assessment of learning,” *International Journal of Learning and Media*, vol. 1, no. 2, 2009.
- [90] D. K. Simonton, “Scientific creativity: Discovery and invention as combinatorial,” *Frontiers in Psychology*, p. 3603, 2021.

- [91] A. Csanadi, I. Kollar, and F. Fischer, “Scientific reasoning and problem solving in a practical domain: Are two heads better than one?” Singapore: International Society of the Learning Sciences, 2016.
- [92] G. N. Svarovsky, “Exploring complex engineering learning over time with epistemic network analysis,” *Journal of Pre-College Engineering Education Research (J-PEER)*, vol. 1, no. 2, p. 4, 2011.
- [93] G. Arastoopour, N. C. Chesler, D. W. Shaffer, and Z. Swiecki, “Epistemic network analysis as a tool for engineering design assessment,” in *2015 ASEE Annual Conference & Exposition*, 2015, pp. 26–679.
- [94] G. Arastoopour, D. W. Shaffer, Z. Swiecki, A. Ruis, and N. C. Chesler, “Teaching and assessing engineering design thinking with virtual internships and epistemic network analysis,” *International Journal of Engineering Education*, vol. 32, no. 3, pp. 1492–1501, 2016.
- [95] A. Ruis, A. A. Rosser, C. Quandt-Walle, J. N. Nathwani, D. W. Shaffer, and C. M. Pugh, “The hands and head of a surgeon: Modeling operative competency with multimodal epistemic network analysis,” *The American Journal of Surgery*, vol. 216, no. 5, pp. 835–840, 2018.
- [96] W. Hobbs, M. Pepper, D. Sanchez, and S. Zörgő, “Network analysis of covid-19 tweets between donald trump and centers for disease control twitter,” in *Second International Conference on Quantitative Ethnography: Conference Proceedings Supplement*, 2021, p. 27.
- [97] N. Pantić, S. Galey, L. Florian, S. Joksimović, G. Viry, D. Gašević, H. Knutes Nyqvist, and K. Kyritsi, “Making sense of teacher agency for change with social and epistemic network analysis,” *Journal of Educational Change*, pp. 1–33, 2021.
- [98] D. Bowman, Z. Swiecki, Z. Cai, Y. Wang, B. Eagan, J. Linderoth, and D. W. Shaffer, “The mathematical foundations of epistemic network analysis,” in *International Conference on Quantitative Ethnography*. Springer, 2021, pp. 91–105.
- [99] H. B. Mann and D. R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” *The annals of mathematical statistics*, pp. 50–60, 1947.

- [100] R. Elmoazen, M. Saqr, M. Tedre, and L. Hirsto, “A systematic literature review of empirical research on epistemic network analysis in education,” *IEEE Access*, 2022.
- [101] D. Shaffer and A. Ruis, “Epistemic network analysis: A worked example of theory-based learning analytics,” *Handbook of learning analytics*, 2017.
- [102] D. W. Shaffer, “Epistemic network analysis: Understanding learning by using big data for thick description,” *International handbook of the learning sciences*, pp. 520–531, 2018.
- [103] C.L. Marquart, C. Hinojosa, Z. Swiecki, B. Eagan, and D.W. Shaffer, “Epistemic network analysis version 1.7.0.” [Online]. Available: <http://app.epistemicnetwork.org>
- [104] A. A. Rupp, Y. Choi, M. Gushta, R. Mislevy, E. Bagley, P. Nash, D. Hatfield, G. Svarowski, and D. Shaffer, “Modeling learning progressions in epistemic games with epistemic network analysis: Principles for data analysis and generation,” in *Proceedings from the learning progressions in science conference*, 2009, pp. 24–26.
- [105] Z. Chen, M. Dehmer, F. Emmert-Streib, and Y. Shi, “Entropy of weighted graphs with random weights,” *Entropy*, vol. 17, no. 6, pp. 3710–3723, 2015.
- [106] L. Sagiv, S. Arieli, J. Goldenberg, and A. Goldschmidt, “Structure and freedom in creativity: The interplay between externally imposed structure and personal cognitive style,” *Journal of Organizational Behavior*, vol. 31, no. 8, pp. 1086–1110, 2010.
- [107] E. F. Rietzschel, J. M. Slijkhuis, and N. W. Van Yperen, “Task structure, need for structure, and creativity,” *European Journal of Social Psychology*, vol. 44, no. 4, pp. 386–399, 2014.
- [108] T. Herder, Z. Swiecki, S. S. Fougat, A. L. Tamborg, B. B. Allsopp, D. W. Shaffer, and M. Misfeldt, “Supporting teachers’ intervention in students’ virtual collaboration using a network based model,” in *Proceedings of the 8th international conference on learning analytics and knowledge*, 2018, pp. 21–25.
- [109] R. M. Roberts, *Serendipity: Accidental discoveries in science*. Wiley, 1989.

- [110] C. R. Aldous, “Creativity, problem solving and innovative science: Insights from history, cognitive psychology and neuroscience.” *International Education Journal*, vol. 8, no. 2, pp. 176–187, 2007.
- [111] A. Newell, *Unified theories of cognition*. Harvard University Press, 1994.
- [112] H. Benbya, N. Nan, H. Tanriverdi, and Y. Yoo, “Complexity and information systems research in the emerging digital world,” *Mis Quarterly*, vol. 44, no. 1, pp. 1–17, 2020.
- [113] M. J. Jacobson, J. A. Levin, and M. Kapur, “Education as a complex system: Conceptual and methodological implications,” *Educational Researcher*, vol. 48, no. 2, pp. 112–119, 2019.

## APPENDIX. WORKED EXAMPLES

### A.1. A Worked Example of Calculating the ROC AUC Metric

This worked example will show how to calculate ROC AUC for three different classifications of ten samples. When training a classifier, the class probabilities can be extracted to know how certain the algorithm was that the sample belongs to a specific class. For most classifier models, the decision boundary is 0.5; if the probability of a positive case is greater than 0.5, predict a '1', and if the probability of a positive case is less than 0.5, predict a '0'. However, an ROC considers *all* possible decision boundaries between 0 and 1. Figures A.1 and A.2 show three different examples of class probability graphs and their resulting ROCs. The top has high separation because the positive cases and negative cases are perfectly separated (good classifier). The bottom has low separation because the positive cases and negative cases are interwoven (bad classifier).

From these class probability graphs, we can graph the ROC where false positive rate is on the x-axis, and true positive rate is on the y-axis. The false positive rate is the number of negative cases that were misclassified as positive cases, and the true positive rate is the number of positive cases that were correctly classified as positive cases. All ROCs have endpoints of (0,0) and (1,1) representing a decision boundary of 1 and 0, respectively. If the decision boundary is 1, we assume that all samples are the negative class, leading to a true positive rate of 0 and a false positive rate of 0. If the decision boundary is 0, we assume that all samples are the positive class, leading to a true positive rate of 1 and a false positive rate of 1. The points between (0,0) and (1,1) each represent the false positive rate and true positive rate at a different decision boundary. For example, the large points in A.1 represent the marked decision boundary of about 0.45.

If we keep plotting at all possible decision boundaries, we get the ROC graph in Figure A.2. From this, we can calculate the area under the curve for each of the three cases. For the top case, the area under the curve is equal to 1; for the middle case, the area under the curve is equal to 0.8; for the bottom case, the area under the curve is equal to 0.6. An ROC AUC of 1 corresponds to perfect separability, an ROC AUC of 0.5 corresponds to random classification, and an ROC AUC of 0 represents the inverse of a perfect classifier. Although the top classifier has a perfect ROC AUC, it is interesting to note that it would *not* have perfect accuracy if a typical decision boundary

is used. Usually, the decision boundary is 0.5, so only 4 of the 5 positive cases would have been classified as positive (leading to an accuracy measure of 9/10 or 0.9).

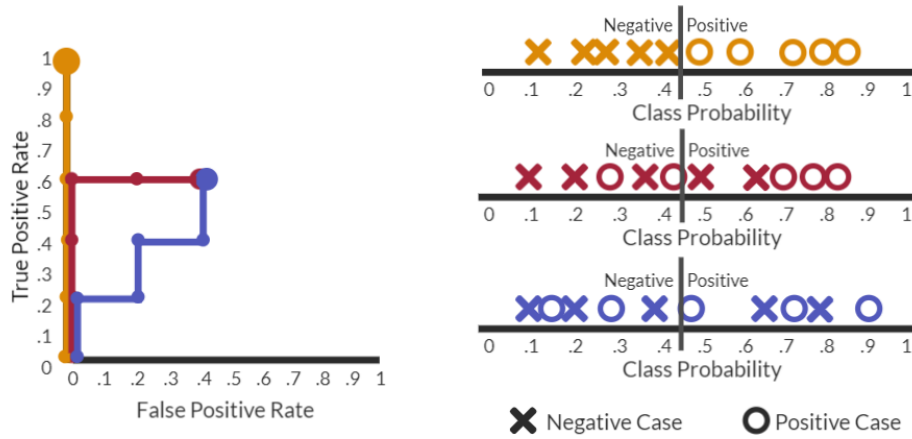


Figure A.1. A partially drawn ROC where the large points represent the drawn decision boundary at about 0.45. For the top case, this decision boundary results in a false positive rate of 0 (0/5 negative cases were predicted to be positive) and a true positive rate of 1 (5/5 positive cases were predicted to be positive). For the middle case and bottom case, this decision boundary results in a false positive rate of 0.4 (2/5 negative cases were predicted to be positive) and a true positive rate of 0.6 (3/5 positive cases were predicted to be positive).

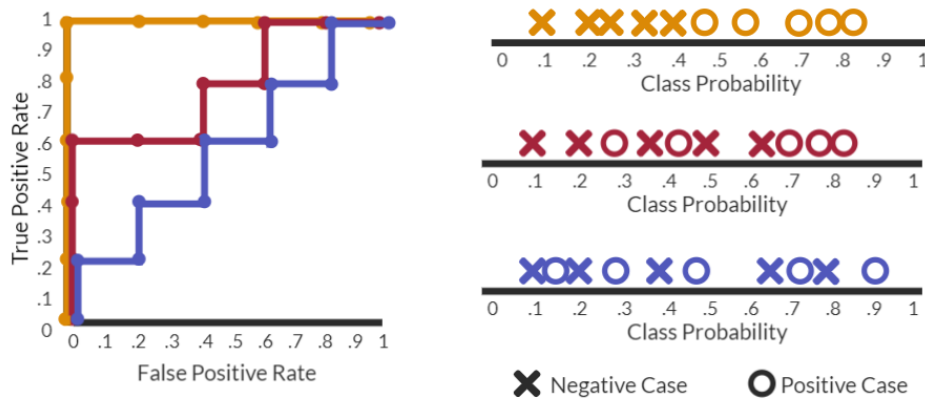


Figure A.2. A completed ROC graph for the three classification models.



## A.2. A Worked Example of Epistemic Network Analysis

This worked example will go through epistemic network analysis for six students and three codes (survey, define, and explore). For the sake of the example, all students complete two actions each week for four weeks, but it should be noted that this is not representative of the actual student behavior. Only three codes are used because it allows the reader to more easily visualize a 3-dimensional space being reduced to a 2-dimensional space.

1. **Organize lines and codes:** First, the log data is converted into code vectors where each code is represented by a column. A ‘1’ represents a presence of that code in that line, and a ‘0’ represents a lack of presence of that code in that line. These vectors are then organized by unit (student) and conversation (week). The entire set of code vectors is represented  $A$ . See Table A.1 for an example for one unit. This is the form of the data that is uploaded to ENA Web Kit.

Table A.1. A sample list of code vectors for a given student. Each row in the table represents one log line that has been converted into a code vector.

		$A$		
<b>Student</b>	<b>Week</b>	<b>Survey</b>	<b>Define</b>	<b>Explore</b>
1	1	1	0	0
1	1	0	1	0
1	2	1	0	0
1	2	0	1	0
1	3	1	0	0
1	3	0	0	1
1	4	0	1	0
1	4	0	0	1

2. **Create adjacency vectors:** Next, the code vectors are converted into adjacency vectors  $H$  where

$$H_{i,j}^{x,y} = \begin{cases} 1 & \text{if } a_i^{xy} > 0 \text{ and } a_j^{xy} > 0 \\ 0, & \text{otherwise} \end{cases}$$

In other words, if codes  $i$  and  $j$  both appear for student  $x$  in week  $y$ , the corresponding value of  $H$  is set equal to 1. From the data in Table A.1, a set of adjacency vectors can be created. These adjacency vectors for Student 1 are listed in Table A.2.

Table A.2. A sample list of adjacency vectors for a given student using the data in Table A.1. Each row in the table represents one student for one week. The three rightmost columns represent existence of a pair of codes. ‘S/D’ represents existence of ‘Survey’ and ‘Define’, for example.

Student	Week	$H$		
		S/D	S/E	D/E
1	1	1	0	0
1	2	1	0	0
1	3	0	1	0
1	4	0	0	1

3. **Create association matrix:** Next, for each student, an association matrix  $\Omega$  is created where

$$\Omega^y = \sum_{x=1}^N \sum_{H \in H^{xy}} H$$

In other words,  $\Omega^y$  represents the sum of each of the  $N$  adjacency vectors that correspond to student  $y$ . The rows and columns of  $\Omega$  represent the codes, and the elements of  $\Omega$  correspond to the number of co-occurrences for that set of codes. For our example student 1,

$$\Omega = \begin{bmatrix} \Omega_{SS} = 0 & \Omega_{SD} = 2 & \Omega_{SE} = 1 \\ \Omega_{DS} = 2 & \Omega_{DD} = 0 & \Omega_{DE} = 1 \\ \Omega_{ES} = 1 & \Omega_{ED} = 1 & \Omega_{EE} = 0 \end{bmatrix}$$

Note that the diagonal entries of the matrix are all equal to 0 because a code cannot co-occur with itself.

4. **Create association vector:** Because many of the elements of the association matrix are redundant, an association vector  $z$  is then created with only the non-redundant elements concatenated by row. For student 1,

$$z_1 = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix}$$

At this point, the other five students will also be introduced into the worked example. Students 1, 2, and 3 are in Group 1, and Students 4, 5, and 6 are in Group 2.

$$\text{Group 1: } z_1 = \begin{bmatrix} 2 & 1 & 1 \end{bmatrix} \quad z_2 = \begin{bmatrix} 2 & 2 & 0 \end{bmatrix} \quad z_3 = \begin{bmatrix} 3 & 1 & 0 \end{bmatrix}$$

$$\text{Group 2: } z_4 = \begin{bmatrix} 0 & 3 & 1 \end{bmatrix} \quad z_5 = \begin{bmatrix} 0 & 2 & 2 \end{bmatrix} \quad z_6 = \begin{bmatrix} 0 & 1 & 3 \end{bmatrix}$$

Because these are vectors of size 3, they can be visualized in a 3-dimensional space where each axis represents a pair of codes. See A.3 for a representation of the six association vectors.

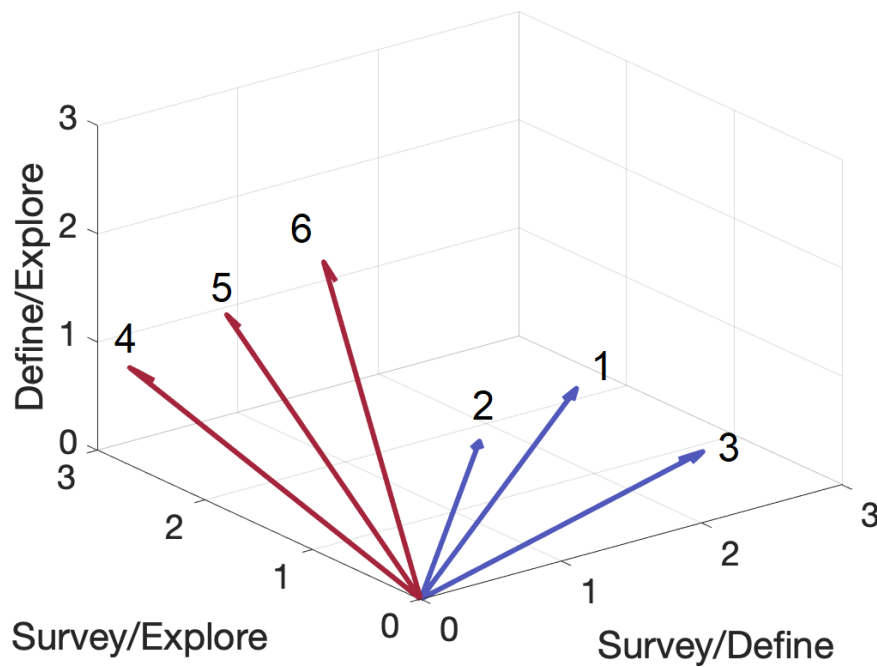


Figure A.3. A visual representation of the six association vectors in a 3-dimensional space. Blue vectors correspond to students in Group 1, and red vectors correspond to students in Group 2. Each axis represents the number of co-occurrences for a pair of codes.

- 5. Normalize association vector:** Next, the association vectors are normalized by dividing them by their magnitude:

$$N^y = \frac{z^y}{\|z^y\|}$$

Each element of the vector now represents the relative co-occurrence of that pair of codes rather than the raw count. The normalized vectors are plotted in Figure A.4.

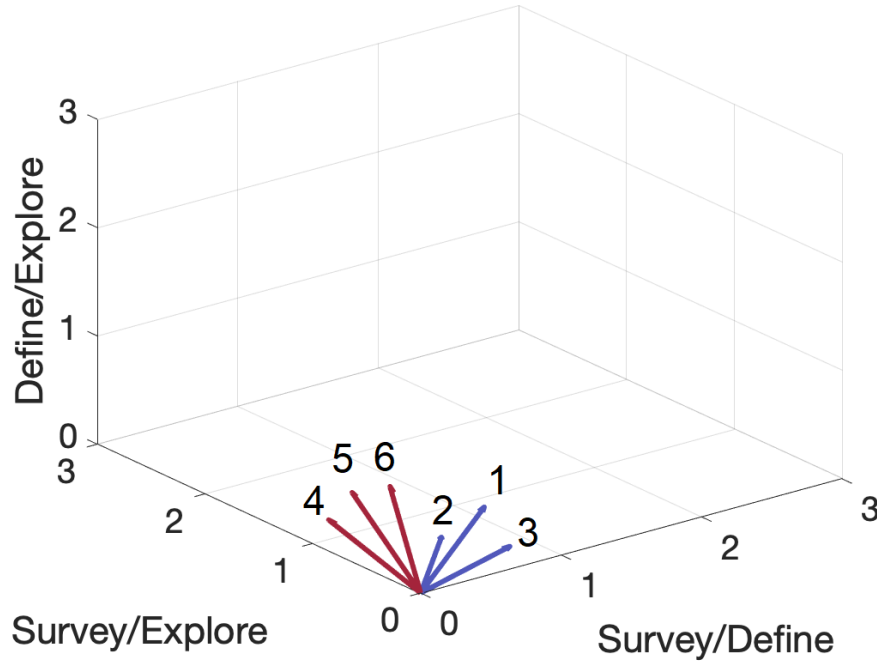


Figure A.4. A visual representation of the six normalized association vectors plotted in the same 3-dimensional space as A.3. Each vector is now length 1.

6. **Center association vector:** Next, all vectors are centered at the origin by subtracting the mean of the  $M$  normalized association vectors  $N$  from each of the individual normalized association vectors:

$$N^y = N^y - \bar{N} \text{ where } \bar{N} = \frac{\sum_{y=1}^M N^y}{M}$$

The vectors retain their information but now span multiple quadrants as seen in Figure A.5.

7. **Project into 2D space:** Next, to determine the coordinates of each of the points when projected into a 2-dimensional space, the eigenvectors of  $NN'$  are calculated by factoring  $N$  as:

$$N = UDV'$$

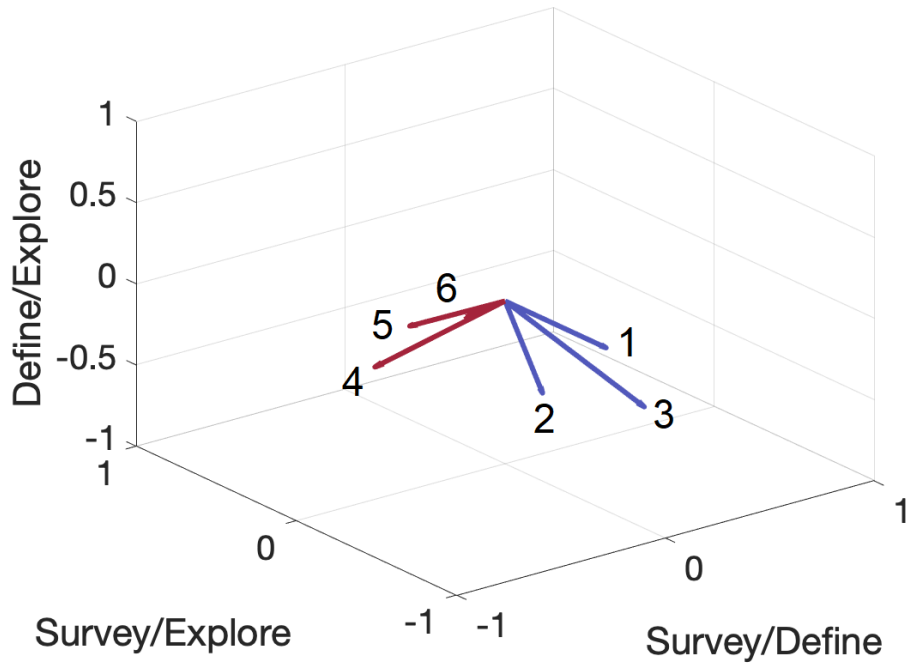


Figure A.5. A visual representation of the six normalized association vectors now centered at the origin. Note that the axes ranges have changed from figure A.4.

where  $U$  is a matrix whose columns are the eigenvectors of  $NN'$ ,  $V$  is a matrix whose columns are the eigenvectors of  $N'N$ , and  $D$  is a diagonal matrix whose elements are the non-zero eigenvalues of  $NN'$ . These eigenvectors  $U$  are then used to find the reduced matrix  $R$ :

$$R = N' * U$$

where  $R^{ij}$  corresponds to the  $i$ th unit of the  $j$ th dimension of the SVD. Using MATLAB to calculate these values gives us:

$$R = \begin{bmatrix} 0.3147 & 0.2270 & -0.3459 \\ 0.4283 & -0.2780 & -0.3492 \\ 0.6657 & 0.0604 & -0.5500 \\ -0.3529 & -0.4946 & -0.3165 \\ -0.5516 & -0.0803 & -0.3234 \\ -0.6490 & 0.3256 & -0.5154 \end{bmatrix}$$

These results correspond with the results plotted by the ENA Web Kit. Each row represents a different student, and ENA Web Kit plots the first column on the x-axis and the second column on the y-axis. For example, student 1.1 (Group 1, Student 1) is plotted at (0.3147, 0.2270), and student 2.5 (Group 2, Student 5) is plotted at (-0.5516, -0.0803).

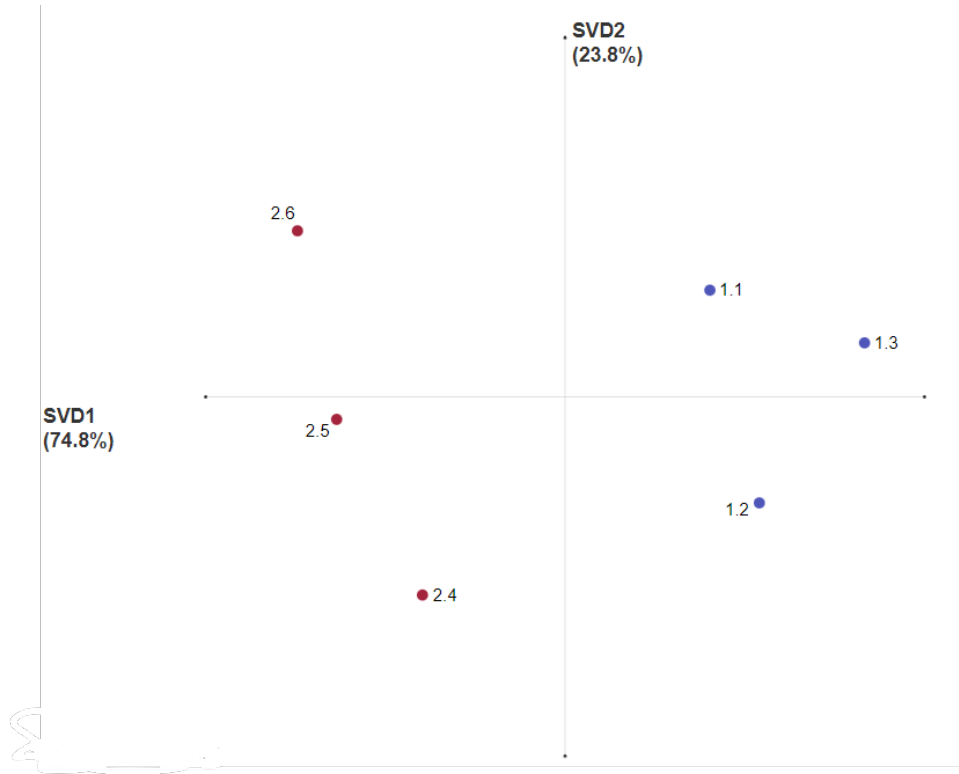


Figure A.6. The six example students projected into the 2-dimensional space using matrix  $R$ . This figure was created using ENA Web Tool, and the values match those calculated by the author using MATLAB (seen above).

8. **Determine SVD variance:** Next, to determine the amount of variance in the data represented by each of the dimensions of the singular value decomposition, the variance of the data for the  $n$ th SVD is divided by the variance of the original data.

$$\text{Variance represented by } n\text{th SVD} = \frac{\text{var}(R_{n*})}{\sum_{j=1}^K \text{var}(N_{*j})}$$

where  $R_{n*}$  is the vector representing the entire  $n$ th column of  $R$ ,  $N_{*j}$  is the vector representing the entire  $j$ th row of  $N$ ,  $K$  is the total number of dimensions in the SVD, and  $\text{var}()$  is the

variance  $V$  of vector  $A$  of size  $M$  calculated

$$V = \frac{1}{M-1} \sum_{i=1}^M |A_i - \mu|^2 \text{ where } \mu = \frac{1}{M} \sum_{i=1}^M A_i$$

Using MATLAB, the variance of the 1st SVD is 0.7464, and the variance of the 2nd SVD is 0.2279. By referring back to A.6, we see that these values match with those given by the ENA Web Tool.

- 9. Calculate difference between groups:** Finally, to calculate the difference between the two groups, the Mann-Whitney test can be performed. Using MATLAB,  $p=0.1$  on the x-axis, and  $p=1$  on the y-axis. These results also match the values given by the ENA Web Tool. Visually, these results make sense; on the x-axis, there is some difference between Group 1 and Group 2, but on the y-axis, there is significant overlap between the two groups (as seen in A.6).

### A.3. A Worked Example of Calculating Network Complexity

This worked example will calculate network complexity for a network with four nodes.

1. **Create association vectors:** Assume  $E_{ij}$  is the thickness of the edge connecting nodes  $i$  and  $j$ . For our example network with four nodes shown in Figure A.7, a corresponding association vector can be written:

$$\left[ E_{AB} \quad E_{AC} \quad E_{AD} \quad E_{BC} \quad E_{BD} \quad E_{CD} \right]$$

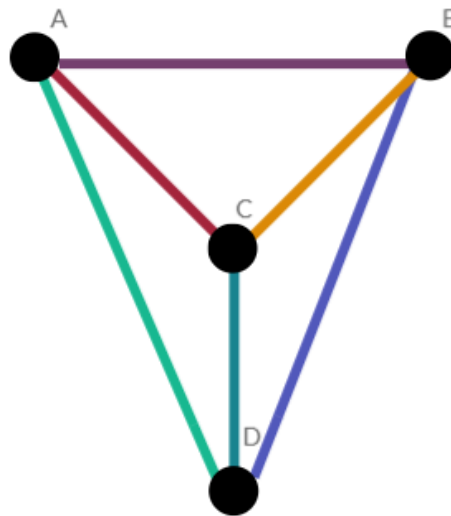


Figure A.7. A four-node network with nodes A, B, C, and D for the sake of the worked example.

2. **Calculate edge weights:** In order to determine the weights of each edge, normalize the association vector  $E$  to get the weight vector  $W$ :

$$W = \frac{E}{\|E\|}$$

3. **Calculate entropy of each node:** Next, the entropy  $H$  of each node  $i$  is calculated:

$$H(v_i) = - \sum_{j=1}^{d_i} p_{ij} \log_2(p_{ij})$$



where  $p_{ij}$  is calculated:

$$p_{ij} = \frac{w(v_i v_j)}{\sum_{j=1}^{d_i} w(v_i v_j)}$$

where  $w(v_i v_j)$  is the corresponding entry of the weight vector  $W$  and  $d$  is the dimension of the node (i.e. the number of other nodes it is connected to). In the context of our epistemic network analysis problem,  $p_{ij}$  can be interpreted as the probability that a student also completes action  $j$  if they complete action  $i$  in the same week. If a student only completes action  $j$  in the same week as completing action  $i$ ,  $p_{ij}$  would be 1, and the entropy of node  $i$  would be 0. Note that  $p_{ij}$  does not always equal  $p_{ji}$  because the denominator of the expression depends on the nodes that are connected to the node indexed first; action  $i$  might co-occur with many actions ( $p_{ij}$  is high), whereas action  $j$  might only co-occur with a few actions ( $p_{ji}$  is low).

For our 4-node example network, assuming each edge connected to A has a weight greater than 0, the entropy of node A is calculated:

$$H(v_A) = -(p_{AB} \log_2(p_{AB}) + p_{AC} \log_2(p_{AC}) + p_{AD} \log_2(p_{AD}))$$

where

$$p_{AB} = \frac{w_{AB}}{w_{AB} + w_{AC} + w_{AD}} \quad p_{AC} = \frac{w_{AC}}{w_{AB} + w_{AC} + w_{AD}} \quad p_{AD} = \frac{w_{AD}}{w_{AB} + w_{AC} + w_{AD}}$$

If any of the edge weights are 0, those edges can simply be ignored in the calculation to prevent an undefined value for  $\log_2(p_{ij})$ .

4. **Calculate entropy of network:** To calculate entropy  $I$  of the entire network, sum the entropy of each of the network nodes:

$$I(G, w) = \sum_{k=1}^N H(v_k)$$

In the case of the example network,

$$I(G, w) = H(v_A) + H(v_B) + H(v_C) + H(v_D)$$

In Figure A.8, five four-node networks are shown in order of increasing complexity to help the reader gain further intuition about the factors that impact network complexity (number of connections, weight of connections, and position of connections in relation to other connections). In addition, two worked examples can be seen in Figure A.9.

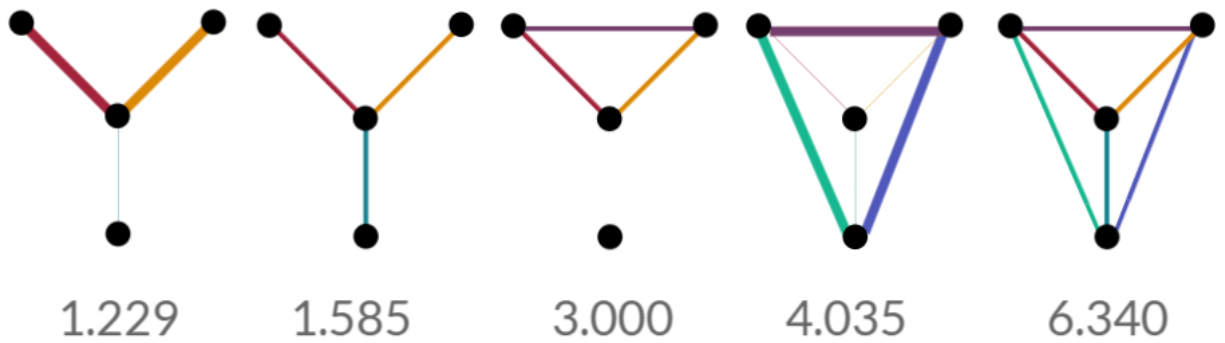


Figure A.8. Complexity in bits for a variety of four-node networks

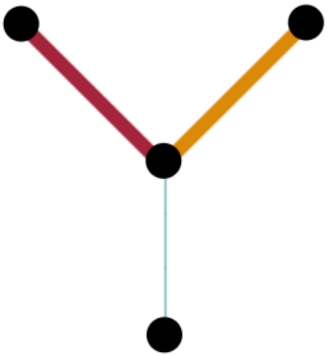
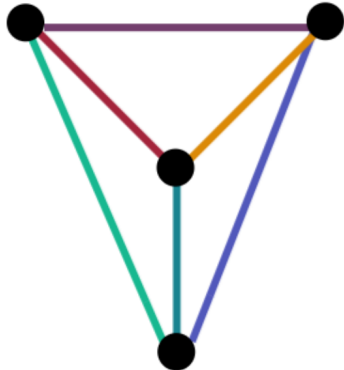
<p>Network Depiction</p>  <p>A ●      B ●</p> <p>    C ●</p> <p>    D ●</p>		
<p>Association Vector</p> <p><small><math>[E_{AB} \ E_{AC} \ E_{AD} \ E_{BC} \ E_{BD} \ E_{CD}]</math></small></p>	$[0 \ 10 \ 0 \ 10 \ 0 \ 1]$	$[5 \ 5 \ 5 \ 5 \ 5 \ 5]$
<p>Normalized Weight Vector</p> $W = \frac{E}{  E  }$	$\left[0 \ \frac{10}{\sqrt{201}} \ 0 \ \frac{10}{\sqrt{201}} \ 0 \ \frac{1}{\sqrt{201}}\right]$	$\left[\frac{5}{\sqrt{150}} \ \frac{5}{\sqrt{150}} \ \frac{5}{\sqrt{150}} \ \frac{5}{\sqrt{150}} \ \frac{5}{\sqrt{150}} \ \frac{5}{\sqrt{150}}\right]$
<p>Edge Probabilities</p> $p_{ij} = \frac{w(v_i v_j)}{\sum_{j=1}^{d_i} w(v_i v_j)}$	$p_{AB} = 0 \quad p_{AC} = 1 \quad p_{AD} = 0$ $p_{BA} = 0 \quad p_{BC} = 1 \quad p_{BD} = 0$ $p_{CA} \approx 0.476 \quad p_{CB} \approx 0.476 \quad p_{CD} \approx 0.048$ $p_{DA} = 0 \quad p_{DB} = 0 \quad p_{DC} = 1$	$p_{AB} = \frac{1}{3} \quad p_{AC} = \frac{1}{3} \quad p_{AD} = \frac{1}{3}$ $p_{BA} = \frac{1}{3} \quad p_{BC} = \frac{1}{3} \quad p_{BD} = \frac{1}{3}$ $p_{CA} = \frac{1}{3} \quad p_{CB} = \frac{1}{3} \quad p_{CD} = \frac{1}{3}$ $p_{DA} = \frac{1}{3} \quad p_{DB} = \frac{1}{3} \quad p_{DC} = \frac{1}{3}$
<p>Node Entropy</p> $H(v_i) = -\sum_{j=1}^{d_i} p_{ij} \log_2(p_{ij})$	$H_A = 0 \quad H_B = 0 \quad H_C \approx 1.23 \quad H_D = 0$	$H_A \approx 1.58 \quad H_B \approx 1.58 \quad H_C \approx 1.58 \quad H_D \approx 1.58$
<p>Network Entropy</p> $I(G, w) = \sum_{k=1}^N H(v_k)$	$I \approx 1.23$	$I \approx 6.34$

Figure A.9. Two worked examples of calculating network complexity (entropy) for four-node networks