

University of New England
DUNE: DigitalUNE

All Theses And Dissertations


Theses and Dissertations

5-1-2017

Mining Helpdesk Databases For Professional Development Topic Discovery

Joel T. Lowsky
University of New England

Follow this and additional works at: <http://dune.une.edu/theses>

 Part of the [Databases and Information Systems Commons](#), [Educational Assessment, Evaluation, and Research Commons](#), and the [Educational Leadership Commons](#)

© 2017 Joel Lowsky

Preferred Citation

Lowsky, Joel T., "Mining Helpdesk Databases For Professional Development Topic Discovery" (2017). *All Theses And Dissertations*. 113.
<http://dune.une.edu/theses/113>

This Dissertation is brought to you for free and open access by the Theses and Dissertations at DUNE: DigitalUNE. It has been accepted for inclusion in All Theses And Dissertations by an authorized administrator of DUNE: DigitalUNE. For more information, please contact bkenyon@une.edu.

MINING HELPDESK DATABASES FOR PROFESSIONAL
DEVELOPMENT TOPIC DISCOVERY

By

Joel T. Lowsky

B.A. (McGill University) 2001
M.A. (Pepperdine University) 2012

A DISSERTATION

Presented to the Affiliated Faculty of
The College of Graduate and Professional Studies
at the University of New England

In Partial Fulfillment of Requirements
For the Degree of Doctor of Education

Portland and Biddeford, Maine

March, 2017

Copyright 2017 by Joel T. Lowsky

MINING HELPDESK DATABASES FOR PROFESSIONAL
DEVELOPMENT TOPIC DISCOVERY

Abstract

This single-site, instrumental case study created and tested a methodological road map by which academic institutions can use text data mining techniques to derive technology skillset weaknesses and professional development topics from the site's technical support helpdesk database. The methods employed were described in detail and applied to the helpdesk database of an independent, co-educational boarding high school in the northeastern United States. Standard text data mining procedures, including the formation of a wordlist (frequently occurring terms), and the creation and application of clustering (automated data grouping) and classification (automated data labeling) models generated meaningful and revealing themes from the helpdesk database. The results of the text mining procedures were bolstered and analyzed using human interpretation and spreadsheet-based summaries. Major findings included the discovery of four prominent technologies that warranted professional development at the site and a universally-applicable approach to undertaking successful helpdesk data mining endeavors. The case study's conclusions included a call to action for researchers to leverage the methodology at other locations. Future data mining studies may yield practical and applicable knowledge at research sites. Shared methods, approaches, and findings from such studies will advance the field of helpdesk data mining used to glean professional development topics for the very people who have submitted technological support requests to helpdesk providers.

University of New England

Doctor of Education
Educational Leadership

This dissertation was presented by

Joel T. Lowsky

It was presented on
March 28, 2017
and approved by:

Brianna Parsons, Ed.D.
Lead Advisor
University of New England

Michael Patrick, Ed.D.
Secondary Advisory
University of New England

Richard Sell, Ph.D.
Affiliate Committee Member
Northfield Mount Hermon School

ACKNOWLEDGEMENTS

The researcher wishes to thank his father, Michael Lowsky, for his endless support and bottomless confidence.

Much appreciation goes to Dr. Brianna Parsons, for her passion, commitment, honesty, and lifesaving turnaround times.

Additional thanks go to Corey Lowsky, Jane Mellow, Frances Winterstein, Kate Majewski, and Emily Majewski for their support and understanding.

Thank you to Brian Tvenstrup of Lindon Ventures for his remarkable work ethic, clear communication, and student-friendly pricing.

This dissertation would not have happened without the aid of Dr. Kathleen Davis, Dr. Michelle Collay, and the late Dr. Pamela Flood. This dissertation is dedicated to Dr. Flood's memory.

For my mother, Beverly Lowsky.

TABLE OF CONTENTS

LIST OF TABLES.....	xi
LIST OF FIGURES.....	xii
CHAPTER 1: INTRODUCTION.....	1
The Origin and Context of the Study.....	2
Statement of the Problem.....	4
Purpose of the Study.....	5
Rationale and Significance.....	6
Rationale.....	9
Research Questions.....	10
Conceptual Framework.....	10
Assumptions, Limitations, and Scope.....	12
Limitations.....	12
Delimitations.....	13
Scope.....	14
Definitions of Key Terms.....	14
Conclusion.....	16
CHAPTER 2: LITERATURE REVIEW.....	18
Educational Technology Integration.....	19
Defining Educational Technology Integration.....	19
Barriers.....	22
Assessing or Evaluating Integration.....	26

Assessment Risks.....	28
Recommendations and Actions.....	30
Professional Development Models.....	30
Professional Development Design.....	31
Optimizing Professional Development.....	33
Summary: The State of Educational Technology Integration.....	33
Educational Data Mining.....	34
An Overview of Educational Data Mining.....	35
Stakeholders.....	45
EDM Focus Areas.....	47
EDM Trends.....	49
EDM Methods.....	52
Text Data Mining.....	55
EDM Process.....	56
Automating the Process.....	58
Data Sources.....	59
Helpdesk Data Mining.....	61
HDDDB Mining to Support Customer Service.....	62
HDDDB Mining Concluding with Human-Assisted Classification.....	63
HDDDB Mining Beginning with Human-Assisted Classification.....	64
HDDDB Mining for Simple Technology Problems.....	65
HDDDB Mining for Relationships.....	68
HDDDB Mining and People.....	69

HDDDB Mining for Prediction.....	70
HDDDB Mining Summary.....	72
A New Framework.....	72
CHAPTER 3: METHODOLOGY.....	76
Study Purpose and Research Questions.....	76
General Research Methodology.....	77
The Setting of the Case Study.....	79
Sampling and Participants.....	81
Overview of Data Collection, Preprocessing, and Analysis.....	83
Detailed Methodology.....	85
Data Preprocessing Prior to Input into RapidMiner.....	86
Data Preparation in RapidMiner.....	91
Clustering Methodology in RapidMiner.....	102
Classification Methodology in RapidMiner and Excel.....	108
Methodology Conclusion.....	115
Ethical Concerns and Participant Rights.....	116
Consent.....	118
Conflict of Interest.....	119
Assumptions, Limitations, and Scope.....	119
Limitations.....	120
Delimitations.....	121
Scope.....	122
Conclusion.....	122

CHAPTER 4: ANALYSIS, RESULTS, AND FINDINGS.....	123
Analysis.....	124
Additional Wordlist Analysis.....	124
Additional Clustering Analysis.....	128
Findings and Results.....	132
Results: Wordlist.....	132
Results: Clustering Analysis.....	134
Results: HD Records with Cluster Number.....	136
Results: Clusters with Meaningful Tokens.....	138
Results: Summary of Meaningful Tokens in all Clusters.....	145
Results: Help Request Data and Classification Labels.....	149
Results: Clustering and Classification Combined.....	150
Conclusion.....	153
CHAPTER 5: CONCLUSIONS, RECOMMENDATIONS, AND IMPLICATIONS.....	156
Interpretation of Findings.....	156
Question 1.....	157
Question 2.....	164
Question 3.....	168
Implications of the Study.....	171
Implications for Practice.....	172
Implications for Leadership.....	177
Recommendations for Action.....	181
Recommendations for Action: Help Desks.....	182

Other Recommendations for Action.....	184
Recommendations for Further Study.....	187
Conclusion.....	190
REFERENCES.....	192
APPENDIX A. Wordlist Tokens in 10% or More Unique Helpdesk Requests.....	200
APPENDIX B. Top 50 Tokens by Average Term Frequency for Each Cluster.....	202
APPENDIX C. Bottom 50 Tokens by Average Term Frequency for Each Cluster.....	204
APPENDIX D. Prevalent Meaningful Tokens with Count of Clusters.....	206
APPENDIX E. Sample of Classification Label Predictions and Confidences.....	207
APPENDIX F. Wordlist Tokens Deemed Meaningful, or Contextually Relevant.....	208

LIST OF TABLES

Table 1. The Most and Least Frequently Occurring Tokens by Count of Unique HD Request Appearances.....	98
Table 2. Number of Tokens in Percentage Ranges of Appearances in Unique HD Requests.....	99
Table 3. Meaningful Tokens that Appeared in 10% or More of All Helpdesk Records.....	125
Table 4. Summary of Meaningful Tokens by Category and Subcategory.....	127
Table 5. Meaningful Tokens in the Top 100 of Cluster 1 – Hardware.....	140
Table 6. Meaningful Tokens in the Top 100 of Cluster 2 – Software, Accounts, and Technology Services.....	142
Table 7. Meaningful Tokens in the Top 100 of Cluster 3 – Miscellaneous.....	143
Table 8. Meaningful Tokens in the Top 100 of Cluster 4 – Google Groups.....	144
Table 9. Meaningful Tokens that Appeared in the Top 100 Tokens for Two or More Clusters.....	146
Table 10. Summary of Classification Label Assignments by Cluster.....	151

LIST OF FIGURES

Figure 1. Number of appearances in HD records by category.....	133
--	-----

CHAPTER 1

INTRODUCTION

The use of educational technologies has permeated nearly every facet of contemporary independent boarding high school education. Many such schools require faculty and staff to use technological tools for a variety of applications, from classroom pedagogy to dorm room management, and their technological skills or competencies must be sufficient to adequately and consistently operate the particular technologies leveraged at a given school. A lack of technical skills was often regarded as a major contributor to failed educational technology undertakings (Inan & Lowther, 2010b; Ertmer, Ottenbreit-Leftwich, Sadik, Sendurur, & Sendurur, 2012; Pilgrim & Berry, 2014).

When teachers encounter technological problems, the typical path towards resolution includes reaching out (often via email or phone) to an Information Technology (IT) Helpdesk (HD) where technology professionals offer support in an attempt to help the teacher overcome the problem. HD agents often store information about support requests (also referred to as tickets) as records in a helpdesk database (HDDB). The organization of requests in an HDDB allows the HD office to assign an appropriate agent to a problem, store or research similar problems and solutions, and to record information about problematic technologies. In addition to storing descriptive information of support tickets, the data contained within the HDDB might occasionally be analyzed or summarized for exploratory purposes.

HDDB analysis studies have typically had a standard investigative goal. Often, these analyses focused exclusively on understanding situational technological usage or improving HD operations (Hui & Jha, 2000; Błafladt, Johansen, Eide, & Sandnes, 2004; Andrews & Lucente, 2014). Further, HDDB studies rarely if ever focused on characteristics of those submitting the

support requests. The case study under review in this paper attempted to mine HDDDB data for topic detection for technological professional development (PD) for teachers to understand the potential usefulness of data mining techniques applied to secondary school helpdesk databases.

At the time of the study, several methods existed for professional development topic identification. A commonly used method to identify areas of strength and weakness for technology integration and skills was the usage of assessment surveys or other measurement instruments (Conrad & Munro, 2008; Palmieri, Semich, & Graham, 2009; Davies, 2011). A prominent assessment design was the self-assessment, however regarding technology skills, these were found to be unreliable (Kopcha & Sullivan, 2007; Reinhart, Thomas, & Toriskie, 2011; Maderick, 2013; Maderick, Zhang, Hartley, & Marchand, 2015). As an organization's HDDDB contains historical information regarding technology problems which were unique to that institution, extracting accurate professional development topics from this data source might be both expedient and practical.

The Origin and Context of the Study

The site of the HDDDB data mining case study under review was a 650-student boarding high school in the northeastern United States. At the time of the study, the researcher conducting the study had been the Director of Educational Technology at the school for four years. During this time, the Director had observed that many disruptions to teachers' work were the result of perceived technology failures (PTFs). PTFs were not genuine technological problems such as broken hardware or corrupted software. Rather, PTFs were problems that were interpreted as technology-based and insurmountable but were, in fact, a result of a lack of technology understanding or competence. PTFs might be solved or avoided if users had the skills to bypass or overcome them.

There was no limit to the number of PTFs that a teacher or institution could experience. The PTFs were context-relevant, and the contexts varied based on the individual, the technology at hand, and the desired outcome of the technology usage. Five examples of common PTFs, generated by the researcher, follow:

- A teacher typically launched Microsoft Word from a shortcut on their desktop. One day the shortcut seems to have disappeared. The teacher believed that they could no longer launch and use Microsoft Word.
- A teacher was accustomed to connecting their computer to their classroom audio/video projector via a wireless connection. One day the teacher couldn't find their projector in the list of available projectors and thus couldn't use the projector for her class.
- A teacher relied on a wireless network connection to use a web browser and the school's web-based email platform. One day, the teacher's network connection failed, and the teacher could not access any web pages.
- A teacher installed a software update to his computer's operating system as suggested by a pop-up window. As a result, many programs no longer function, and the teacher could not use his computer purposefully.
- A teacher wrote a document containing an assignment and stored the file in a web-based file storage platform. The next day the teacher could not find the document and thus could not use it in her class.

At the site of the case study, when a teacher encountered a PTF, the teacher often sought assistance from the helpdesk, and a helpdesk agent created a ticket and provided support until a solution was found and implemented. The Director of Educational Technology, who had access

to the HD ticket queue, had observed that the problems submitted to the HD and the solutions found were often forgotten or overlooked after the immediate problem was solved. Topics or trends were never revisited with the goal of providing pre-emptive support to the groups encountering the PTFs. Finally, the Director observed that many tickets contained identical or similar issues as teachers encountered similar PTFs over time. Often, the PTFs were a result of a lack of basic technological skills. The Director, assuming the role of researcher, determined that analysis of the HDDB might provide valuable insight into the trends of technological weaknesses and that PD could be developed to address these shortcomings.

Statement of the Problem

A rich account of teachers' technological (in)competencies often resides in a school's technology support helpdesk database, a platform used to store user-reported IT problems. Historically, researchers have performed studies on site-specific data, such as HDDBs, in both educational and corporate environments. These studies took the form of in-depth and thorough statistics-based database analysis; a practice referred to as data mining (DM). HDDB studies typically focused on examining and improving IT practices rather than on developing the skills of those who had submitted the help requests. The researcher identified a problem in the research and practice fields of EDM and mining of HDDBs. Two gaps existed in contemporary EDM research: mining of HDDBs was not used to identify trends of technical weakness amongst support requesters, nor was it used to determine PD opportunities based on the topics identified.

As a result of the gaps in research and practice, teachers, administrators, and institutions might have missed opportunities to improve teachers' technology capabilities with the particular tools employed at that school. A lack of analysis of HDDBs might thus have equated to a lack of detailed understanding of educators' technical skills *in situ*, as well as a lack of information on

how to reduce frequent problems that might have resulted from weaknesses in technological skill sets. Further, if the HDDBs of independent boarding schools went unanalyzed, other as-yet-undetermined trends or revealing information might be overlooked. Finally, administrations, institutions, helpdesk operations, and teachers stood to benefit from a detailed and organic understanding of teachers' technological weaknesses, as well as the technologies employed that most often lead to perceived technological failures and disruptions. The missed opportunities inherent in the aforementioned gaps did not reduce disruptions and work stoppages experienced by teachers and could not contribute to the ongoing advancement of technology-supported teaching and learning.

Purpose of the Study

The purpose of examining gaps in the body of research, as manifested by an HDDB mining study, is that findings might contribute to fewer technology-based disruptions or improved technology-enhanced instruction, and these changes could benefit teachers, students, institutions, and instruction. The purpose of this case study was to employ data mining of an underused data source to identify and examine areas of improvement for the technology skills of faculty members at an independent boarding high school in the United States in the recent past. The study's overarching goal was to determine the potential and value inherent in the application of data mining techniques to an HDDB. Data mining procedures, applied to a specific subset of a school's complete HDDB (for example, help requests from faculty members) might determine areas of weakness (and, potentially, improvement) for the very people whose technology help requests had created the content within the HDDB. The analysis of genuine, *in situ*, and organic technology failure information may highlight trends in users' weaknesses might be a novel approach to professional development topic determination. By determining the common areas of

technological weakness prevalent in hundreds or thousands of technology support help requests, a professional development model might be devised and intended solely and precisely to address those weaknesses. Thus, mining of the HDDB might be a vitally important method for discovering frequent and pervasive technology weaknesses that inevitably lead to work stoppages and frustration. This method might be used alone or alongside other assessment instruments to determine potential areas for improvement that, if addressed, might lead to substantial beneficial changes in a workplace.

Rationale and Significance

As the Director of Educational Technology at the site of the study, the researcher was uniquely positioned to perform an in-depth analysis of HDDB data with the goal of discovering topics that warranted PD attention. Knowledge gleaned from the data mining might have assisted the Director in enacting change at the study site by highlighting specific problematic technologies currently used by the constituents. The Director's change efforts to date had consisted mostly of implementing new educational technologies, but HDDB mining was anticipated to identify training opportunities with existing hardware and software.

A primary advantage of the approach under review was the leveraging of pre-collected data. As the study focused on data collected organically over time within the HDDB, the researcher, and others in similar roles, need not collect new data in their quest to explore technological skill set weaknesses of a faculty body. The HDDB contained all data relevant to the study. This data included the most significant information: a clear, text-based description of the technology used, the problem, and the solution; as well as potentially relevant data such as the teacher's department, classroom, and other contextual characteristics. Many schools employ a helpdesk and many HDs store information within a database. Therefore, this case study and

other studies that it might inspire, could leverage a data source that likely already exists in near analysis-ready form.

A study that attempted to use the HDDB as a data source for determining areas of weakness and potential growth in teachers' technological competencies had great potential for positive individual and group transformation. If mining of the HDDB could successfully highlight prevalent trends and areas of weakness shared amongst faculty members, then professional development might be conceived to address these problematic topics specifically. By gathering the issues from context-based technology problems (as opposed to teachers' self-assessments or administrators' observations or educated guesses), the potential for valuable, pragmatic, and applicable professional development was high.

If teachers could improve their technological skills on topics that have caused them difficulty, frustration, or work stoppages in the past, it was likely that focused training could reduce similar instances from occurring in the future or provide teachers with the skills necessary to overcome the disruptions without reliance on IT support. Once teachers begin to solve certain technology issues independently without interruption or work stoppages, improvements in their work lives including smooth-flowing technology-bolstered lessons might follow, as well as an increase in confidence regarding their ability to use technology tools productively. As teachers gain confidence and minimize disruptions, instruction quality might improve, and students might benefit. The potential value of PD based on the organic and context-relevant data contained within the HDDB might, therefore, benefit both teachers and students. There was potential value in empowering teachers to solve common technology problems as identified by HDDB mining. The researcher determined that a case study, which scrutinized HDDB data for PD topics, was important and worthwhile.

Beneficiaries. A successful HDDB mining case study that accurately identifies one or more topics or areas of improvement (potentially through PD) might inspire other researchers to leverage mining of HDDB data for similar purposes. Other researchers might conduct similar studies to determine areas of technological weakness within their faculty bodies. Eventually, meta-studies that assemble and identify trends of teachers' technical skill weaknesses across geographic regions, school districts, demographics, or even users of the same software might be conducted to highlight large-scale trends.

Additionally, educational technology practitioners might find the method and results of this study useful. If mining of an independent school's HDDB were to lead to topic identification and PD development, administrators responsible for PD at other institutions might replicate the study to determine their own organization's unique technological areas of improvement. Individuals or groups responsible for developing and offering PD might appreciate a new approach for topic identification. PD developers might choose to leverage the new approach in addition to or in replacement of their current methods. Many learning institutions maintain some form of HDDB or storage system to record technology support requests. If mining the HDDB could provide useful information on professional development topics, and training on these subjects was understood to be useful, then mining the HDDB could be a low-cost way to determine topics for PD accurately.

Other stakeholders may benefit from a faculty body participating in professional development on topics generated directly from their technical support history. If successful PD follows successful topic detection, Information Technology helpdesk operations might experience a reduction in help requests for PTFs. This reduced workload for IT offices might allow those departments to focus attention and energy on other projects, such as software

upgrades and infrastructure improvements. Similarly, budget management within IT departments might shift funding from providing basic support to other projects. Thus, if the workload were to decrease for IT departments as a result of a reduction in PTFs due to focused PD, these offices might be positioned to improve their services, with potential benefits for the institution and the community.

Just as IT departments might benefit from the proposed study, policymakers might also derive value from a successful HDDDB mining undertaking. Decisions about PD but guided by policy could be made more carefully with the results of an HDDDB mining project in hand. Policies might dictate the topics to cover in PD and when to address those topics. However, if PD themes were generated directly from the very group of individuals participating in the PD, policies could be customized to allow for more relevant training offerings. For example, a high school might have a policy that all teachers partake in technology-oriented PD for a certain number of hours per year. If an HDDDB mining analysis informed the policy makers, they could decide on the most relevant, pragmatic, and immediately applicable topic they will address during those hours.

Rationale

By successfully completing a study centered upon mining of the unique and underused data contained with the HDDDB, a new methodology for determining areas of technological improvement for faculty members might be devised. This new approach might inspire others to perform similar analyses to enact conceptually similar changes. Thus, the justification for undertaking this case study was not only to potentially improve conditions at the study site but also to embolden others to question what valuable information might hide within their institution's existing helpdesk database. Indeed, successful mining and analysis of an HDDDB

might inspire researchers to consider mining other existing but potentially underused datasets. By leveraging an existing database, only a data mining inquiry stands between a practitioner and potential new topics and approaches to PD, as well as other unidentified benefits.

Research Questions

The action at the core of this undertaking was the data mining of a helpdesk database. The study's purpose was to determine the information that this data mining might reveal with regards to teachers' technological skills or lack thereof. Professional development might be devised based on the identification of the skills that teachers lacked and the perceived technology failures that could have been a result of these shortcomings. In support of this case study, the researcher formulated several key research questions. These questions accentuated and guided the purpose of the study:

1. In what ways could data mining be leveraged to best extract the desired information from the HDDB?
2. What does data mining of the HDDB reveal about gaps in teachers' technology skill sets?
3. How could data mining of the HDDB be used to determine topics and plan technology-related professional development for teachers?

Conceptual Framework

Academic researchers have cited a lack of necessary technological competencies as a contributing factor to underwhelming educational technology integration (Inan & Lowther, 2010b; Ertmer, Ottenbreit-Leftwich, Sadik, Sendurur, & Sendurur, 2012; Pilgrim & Berry, 2014). When teachers encountered disruptive perceived technology failures (PTFs, as opposed to authentic hardware or software malfunctions), they often sought the services of an onsite information technology helpdesk. The HD staff stored information about each support request in

a helpdesk database and assisted the teacher seeking aid per the procedures unique to each organization. Often, the information in the HDDB remained un- or understudied. Most contemporary studies of HDDB data were situated in the enterprise sphere and addressed the improvement of IT procedures (Hui & Jha, 2000; Blaafldt, Johansen, Eide, & Sandnes, 2004; Andrews & Lucente, 2014). The act of analyzing institutional databases was commonly referred to as data mining and was practiced and studied extensively in the corporate sphere.

At the time of this case study, educational data mining (the practice of analyzing educational institution databases to understand teaching and learning practices) was a rapidly growing research field (Baker & Yacef, 2009; Romero & Ventura, 2013). Despite the growth, however, there was a distinct lack of EDM studies that addressed HDDB analysis. This trend was surprising when contrasted with another EDM trend: that EDM studies increasingly leveraged untapped data sources, predominantly existing data sources rather than data collected specifically for an EDM study (Baker & Yacef, 2009). The methods employed for EDM studies were widely varied, but prominent areas of focus included text mining and distillation of data for human judgment, two approaches commonly used in studies aiming for topic detection and discovery within a large, text-based dataset. Classification and clustering procedures were amongst the most common techniques of text data mining.

Existing text data mining practices from EDM research as well as from enterprise DM could be applied to a school's HDDB to glean valuable information about teachers' technological struggles and (in)competencies. By mining the HDDB, a researcher might be able to highlight the common areas of weakness of a faculty, and perhaps develop professional development to address these areas. The use of data mining practices on the HDDB with the aim of discovering and eventually addressing faculty technological capability shortcomings

represented a new application of EDM, applied to an under-studied data source. The procedures and results of such a study could encourage other researchers to mine institutional HDDBs to search for areas of improvement for their employees.

Assumptions, Limitations, and Scope

Several concepts were held to be true in the undertaking of this study. As a research project, the identification of assumptions was both inevitable and required as these assumptions would bolster and contribute to both the expectations and conclusions of the study. Though the study anticipated reliability for these assumptions as the study progressed, the study might have revealed that certain assumptions were inaccurate or unwarranted.

The study assumed that the HDDB contained a sufficient amount of valid and minable information for topic detection. Similarly, the issues discovered in the HDDB were assumed to be an accurate depiction of a faculty body's technology skill set deficiencies. Teachers at the study site were presumed to use the services of the helpdesk as expected and that helpdesk agents adequately performed the task of storing detailed information about each help request. A final assumption was that data mining methods could effectively be applied to analyze and summarize the data as desired. These assumptions were believed to be true as of the initiation of this research undertaking.

Limitations

A number of limitations were in place that could have restricted the scope of this case study. An important limitation pertained to the employees who have submitted helpdesk requests. While the HDDB at the research site contained requests from teachers, staff, employees, administrators, and students, the study only considered requests submitted by teachers. This limitation was a result of the restrictions placed upon the researcher in his role as

the Director of Educational Technology. The Director's primary group of colleagues and constituents was the faculty, and the Director was empowered to provide professional development to the faculty body. Thus, while a similar study might have revealed useful information about students' and employees' technological skill sets, this study focused on the group with which the Director had the most contact and administrative influence.

An additional limitation of this study was concerned with the transferability and reproducibility of the research. Nearly all HDDBs at any organization store similar data in different structures. While the data mining practices leveraged in this study might have worked in principle at other sites and in other studies, in practice the particular investigative procedures and conclusions drawn as a result of data mining were limited to this site and study. The study did not intend to generate results that were wide-ranging representations or generalizations.

As the study leveraged archived data, any findings might have represented the state of the faculty's technological weaknesses during the timeframe of the study. The use of archival data as a sole data source in any study might yield information about how a situation *was* during the study timeframe, rather than how a situation was at the current moment (Abowitz & Toole, 2009). As the study's timeframe ended in September 2016, and the completed analysis, findings, and results were not anticipated until spring 2017, it was possible that the trends identified in the study were no longer accurate. However, as the study's goal was to determine the value of data mining techniques applied to the HDDB, the content of the trends was not as important to the study's findings as was the capability of revealing those trends in as much detail as possible.

Delimitations

While the HDDB contained many years of support request data, the study only leveraged data collected within the aforementioned three-year period in this study. This limitation was a

result of the application of data distillation for human judgment as a data mining method. As the study required human judgment for topic detection, it was important that the human judge was familiar with the concepts contained within the data. The researcher was employed at the site for the three years identified in the study and could thus competently distil data collected during this timeframe into logical and accurate groupings, trends, or topics.

Scope

The scope of this study was deliberately narrowed to ensure feasibility. The case study focused on the technological support requests placed by teachers at an independent boarding high school within the aforementioned three year period. The data examined in the study was mined using the commercial data mining tool RapidMiner Studio by a certified, professional data miner. The scope of the analysis was limited to topic detection within the HDDB.

Definitions of Key Terms

Actual technology failure (ATF): A problem with a technology that stopped, disrupted, or interrupted work and was attributable to a genuine technological problem such as broken hardware or corrupted software.

Data mining (DM): The act of analyzing databases with the goal of summarizing or analyzing data to inform decisions (Fayyad, Piatetsky-Shapiro, & Smyth, 1996 in Heathcote & Dawson, 2005).

Classification: A prediction data mining method used to automatically categorize data based on categories manually assigned by a human to a subset of the data (Merceron & Yacef, 2005).

Clustering: A data mining method that collects similar data points into groups, and separates dissimilar data points amongst groups (Merceron & Yacef, 2005).

Cluster: A group of similar data points, representing, in this case, a collection of thematically-linked support requests. Cluster and Group are used interchangeably.

Distillation of data for human judgment: An EDM method that does not deliver conclusive answers but, rather, delivers information by which a human participant could derive conclusions (Baker, 2010).

Educational data mining (EDM): Data mining practices enacted upon data generated within academic institutions and with an outcome intended to inform teaching- or learning-related decision making (Merceron & Yacef, 2005).

Helpdesk (HD): An information technology office or department that handles users' technology problems and provides resolution services (Jha & Hui, 2000).

Helpdesk database (HDDB): A software platform for storing and archiving information about HD requests and solutions.

Perceived technology failure (PTF): A problem with a technology that stopped, disrupted, or interrupted work but was not attributable to an actual technological problem such as broken hardware or corrupted software. PTFs were problems that were interpreted as being technology-based and insurmountable but were, in fact, a misinterpretation of a potentially solvable user experience error.

Record: An entry in a database representing, in this case, a single technology support request.

Support request: A single instance of a user requesting technical assistance from a technology helpdesk, manifested by a record in an HDDB. Used interchangeably with record or case.

Text data mining: A DM or EDM method intended to summarize and analyze textual and human-readable information (Romero & Ventura, 2007).

Token: Data mining synonym for word or term. Word, term, and token are used interchangeably.

Topic detection: A DM or EDM approach that aimed to glean subjects or trends prevalent in many records within a dataset (Wartena & Brussee, 2008).

Wordlist: A human-readable list of all unique words in a dataset. Includes frequency of occurrence.

Word vector: A computer-readable version of the wordlist that includes information about each word's appearance in each database record.

Conclusion

As an educator with nearly a decade of professional experience in educational technology integration and a Master's degree in learning technologies, the researcher had dedicated many thousands of hours to the processes involved in improving teachers' usage of technology tools. Years of anecdotal observation of failures and successes in these undertakings inspired an initiative to review teachers' basic or fundamental technology skills, and the researcher identified the helpdesk database as an unbiased source of information about these skills. As the data contained within the HDDB represented genuine, in-context perceived technology failures, this data source was selected for extensive mining, with a goal of extracting topics or areas of technological weakness. Professional development could then be generated to address these technical shortcomings.

By identifying trends of technical weakness amongst support requesters, a practitioner might devise PD opportunities based on the topics identified and for the very audience that

submitted the requests. The results of this case study might have, in turn, lead to improvements in educational technology usage at the study site, and could perhaps inspire other researchers to undertake similar data mining endeavors. This investigation leveraged the research histories of numerous fields of study and practice. Academic technology integration, including barriers and challenges to successful integration, as well as data mining, educational data mining, and HD data mining have extensive histories, and these histories molded the study at hand. The ensuing literature review chapter presents the contemporary status of and trends within these fields.

CHAPTER 2

LITERATURE REVIEW

This literature review will provide an overview of several key facets of educational technology integration undertakings, situated in the current body of scholarly research. Additionally, educational data mining (EDM) and helpdesk database data mining (HDDDB DM) practices were reviewed in depth to highlight important trends in those fields. The primary purpose of this review was to create a solid knowledge base upon which to build a new framework that leverages data mining to address educational technology integration problems. Following recurring themes found in nearly all literature about obstacles to educational technology integration, the approach proposed emphasizes technological skills, capability, and competencies required for teachers to successfully implement technology tools in their teaching. The planned methodology employed text data mining of an underused database to identify technology skill topic areas for teacher training.

The framework and methodology developed and tested for this study outlined summarily in the final sections of this review (and in detail in Chapter 3), considers the research pertaining to definitions of educational technology integration, current barriers to integration, and contemporary methods for analyzing or assessing integration efforts, as well as common recommendations for improving integration. Additionally, the framework was situated within the modern state of data mining in education, which was explored in depth, and numerous case studies were highlighted to underscore the importance and relevance of this nascent field. Finally, research pertaining specifically to helpdesk data mining was surveyed to determine the potential of these methods within a new framework. This literature review addressed these important topics in turn.

Educational Technology Integration

At the time of this research undertaking, many schools pursued educational technology integration as a goal to achieve, a practice to master, or a standard towards which to strive. To best understand educational technology integration, the researcher conducted a multifaceted review of the field's pertinent and contemporary literature. Important areas of study included generating definitions for educational technology integration, reviewing the barriers that prevented or thwarted educational technology integration efforts, integration assessment methods, and the risks inherent in those assessment practices.

Defining Educational Technology Integration

To properly situate educational technology integration within the scope of the study, the researcher deemed it useful to review other scholars' definitions or descriptions of the topic. Inan and Lowther (2010a) provided a brief survey of accepted definitions of technology integration. The authors categorized integration into three general sets wherein technology was used either for instructional preparation, instructional delivery, or as a teaching and learning instrument. Inan and Lowther (2010a) provided descriptive details of each category, but the primary difference between the categories was the affected user base and the degrees to which those users employ the technology. The preparation category of educational technology integration was likely to be undertaken solely by teachers while delivery of instructional content could be acted upon by teachers (for example, presenting content via a projector) and consumed passively by students (watching projected content). The third category, technology as a learning instrument, typically involved both teachers and students actively engaging with the technology. Furthermore, they accepted technology integration as any of the three categories of technology

use, with the caveat that the technology selected successfully supported instructional goals, as opposed to, for example, organizational goals.

In a similar vein to Inan and Lowther (2010a), Davies (2011) attempted to categorize educational technology integration, with emphasis on teachers' behaviors with regards to technological tools. Davies noted that teachers' increased awareness of the technological tools at their disposal was an essential accomplishment in the path toward integration. The author noted that that advanced educational technology integration behavior included the use of technology effectively while discerning the appropriate pedagogic usage of specific tools. In summation, Davies (2011) succinctly posited, "the goal of technology integration in education was the wise and competent use of technology to facilitate learning" (p. 50).

Other researchers have embraced similar definitions of educational technology integration. Ertmer, Ottenbreit-Leftwich, Sadik, Sendurur, and Sendurur (2012) categorized the integration of technology as a collection of roles. Technology integrated into teaching should play one or more roles, including a delivery platform for content and skills, a complement or enrichment for curricular material, and a transformational agent for teaching and learning. Technology integration might thus be considered to have occurred when teachers regularly use technological tools in any (or all) of these roles.

Zhao and Bryant (2006), preceding Inan and Lowther (2010a) and Ertmer et al. (2012), contributed an additional dimension to the idea of educational technology integration. Zhao and Bryant categorized teachers' technology integration into five categories: technology incorporation into curricular standards, using an array of technological tools, and leveraging the tools for new teaching methodologies, classroom management practices, and pedagogical enhancements. Though the authors measured these categories in their qualitative study of

teachers' technology integration practices, it was the frequency with which teachers engaged in behaviors within these categories that served as a primary motivation for the study. Educational technology integration, therefore, might not be just a matter of actions taken by teachers, but the frequency with which teachers took those actions.

Bauer and Kenton (2005) also highlighted the consistent use of technology tools, or lack thereof, as an essential element of technology integration. The authors proposed *use* as a simple alternative term to *integration* but noted a striking difference between the two. Citing Hooper and Rieber (1999), Bauer and Kenton (2005) clarified that integration implied and required "full-time, daily operation [of technology] within lessons" (p. 535). This powerful statement situated educational technology integration firmly in both the scheduling and execution of pedagogy.

While Bauer and Kenton (2005) and Zhao and Bryant (2006) considered time as a critical element of educational technology integration, Schibeci et al. (2008) contributed other dimensions to the concept of educational technology integration. Schibeci et al. focused on educational technology integration as being directly affected by personality traits of the teachers involved and the institutional culture in which the integration was occurring. Considering technology integration systematically, the authors noted that integration progress "should include policies and practices that simultaneously consider the processes by which teachers learn technical skills as well as the cultural, social and historical contexts of learning" (Schibeci et al., p.314). This description of technology integration thus infused the collective definition of the integration process with components that were decidedly non-technological.

Schibeci et al. (2008) highlighted the importance of additional factors in the technology integration process. Okojie, Olinzock, and Okojie-Boulder (2006) addressed the process from an alternate perspective, postulating that true integration only occurs when the technology use had a

direct and tangible impact upon pedagogy. Indeed, Okojie, Olinzock, and Okojie-Boulder (2006) placed little emphasis on other factors that might guide or impact educational technology integration, preferring to focus on the relationship between technology and teaching and learning. In contemporary education, pedagogy and technology, together as educational technology integration, were nearly inseparable (Okojie, Olinzock, & Okojie-Boulder, 2006). The authors clarified this perspective explicitly when they wrote, “technology should not be treated as a separate entity, but should be considered as an integral part of instructional delivery” (p. 67). Thus, educational technology integration, as a concept, might be perceived as the deliberate combination of pedagogic practices with technological resources.

With a link between educational technology and pedagogy and other perceptions of educational technology integration in mind, it was useful to evaluate both the barriers that affect educational technology integration and the factors that impact it. If the research highlighted that a lack of technological skills was a persistent barrier towards integration, then a method to identify and address these skill weaknesses might be a useful contribution to the field. Similarly, a recurring theme of technological weakness as a factor inhibiting educational technology integration could provide evidential support for undertaking this case study.

Barriers

There was significant research into the barriers that adversely impact educational technology integration. Some barriers were found consistently in the academic literature, and the recurrence of these barriers implicitly highlighted their importance. Earlier research (Bauer & Kenton, 2005) placed emphasis on hardware and software accessibility issues, referring to the availability of educational technologies. However, during the first and second decades of the 21st century, technology achieved a remarkable level of ubiquity, resulting in a substantial reduction

in many of the initial barriers to technology integration (Ertmer et al., 2012). The barriers were classified as “first-order” and “second-order,” with the former pertaining to barriers that were external to the teacher, and the latter referring to barriers that were internal to the teacher (p. 423). First-order barriers, such as access to computers, were largely overcome in contemporary K12 education in North America. Internal barriers, which remain as obstacles for teachers to overcome, could include teachers’ capability or competence with technology, or their attitudes and beliefs regarding technology and pedagogy and the interplay between the two (Ertmer et al.). This literature review will focus primarily on internal barriers to educational technology integration.

Conrad and Munro (2008) identified numerous internal barriers to technology integration. These factors, particularly self-efficacy (which the authors explain as one’s beliefs in their capabilities), attitudes, and anxiety were found to have significant impacts on the success of educational technology integration efforts. Self-efficacy, linked by the authors with technological competence, might have a positive or negative impact upon attitudes and anxiety, depending on the teacher’s level of capability. Thus, while teachers’ attitudes and anxiety were found to have an impact on educational technology integration, improved self-efficacy might improve the other barriers, and thus improve integration undertakings.

Groff and Mouza (2008) studied external barriers, which they situated within realms such as the school or district, the project, and the students involved, but also focused on obstacles that were teacher-dependent. These hindrances included technology proficiency or lack thereof. These obstacles also included teachers’ attitudes or pedagogic beliefs that were not conducive to successful technology implementations.

Channeling Conrad and Munro (2008), Ertmer and Ottenbreit-Leftwich (2010) returned to the topic of self-efficacy, including it amongst other important internal barriers that might affect a teacher's educational technology integration efforts. In addition to self-efficacy, Ertmer and Ottenbreit-Leftwich (2010) studied pedagogical beliefs, teachers' responses to their school or district's culture, and teacher knowledge as categories of barriers that could impact technology integration. The researchers emphasized teachers' knowledge as this category included numerous types of understanding, such as topic and content comprehension. Furthermore, a unique type of knowledge was required for educational technology integration as, "first, the teachers need knowledge of the technology itself" (p. 259). This statement directly addresses the importance of a lack of knowledge as a barrier to educational technology integration.

Güneş, Gökçek, and Bacanak (2010) studied the impacts of teachers' confidence and attitudes regarding technology usage and noted a link between a lack of confidence and negative attitudes, as well as the potential negative impacts to an educational technology undertaking as a result of negative attitudes. Technology competence was found to have an impact on confidence, and thus upon attitudes (Schibeci et al., 2008; Conrad & Munro, 2008; Groff & Mouza, 2008). They also noted, however, that technology competence could vary widely across different categories of technology usage, including, for example, word processing and database operation.

While teachers' technology competency, attitudes, and beliefs continued to recur as barriers in the literature, occasionally new barriers were identified. Inan and Lowther (2010a) highlighted "teacher readiness" as a potential barrier (p. 141). This factor was explained as a measure of capability or self-efficacy, but emphasized the ability to integrate contemporary technologies in contemporary circumstances. Not surprisingly, significant relationships were found between computer proficiency and readiness, as well as between computer proficiency and

many other salient factors in the study. Indeed, the authors found that readiness (itself informed by proficiency) had the highest effect on technology integration, while computer proficiency had the third highest effect. An external barrier, technical support (or lack thereof), had the second highest effect. Inan and Lowther (2010b) asserted their findings in a second study in 2010, concluding, “teacher readiness and teacher beliefs were the most important factors with the highest direct effect” on laptop integration (p. 941). Clearly, internal barriers continued to have significant impacts on educational technology integration.

Numerous researchers have contributed to the body of literature regarding the impact of teachers’ feelings, knowledge, attitudes, and beliefs on educational technology integration. While Buabeng-Andoh’s (2012) conclusions echoed those of Güneş, Gökçek, and Bacanak (2010), the researcher also noted that teachers’ attitudes might affect not only their usage of educational technologies but also their acceptance of the usefulness of the technology and its integration into pedagogy. Channeling previous scholars, the Buabeng-Andoh highlighted the value of technology self-efficacy and competence, the impact of these factors upon attitudes, and thus the impact of these internal factors upon a teacher’s willingness to accept technology as a useful tool for teaching. Similarly, Kopcha (2012), acknowledged the value in teachers’ acceptance of the usefulness of technology, and also identified an additional belief-based barrier that might impact educational technology integration.

Kopcha (2012) further highlighted teachers’ vision for technology use as having a significant impact on integration projects. As attempts were made to improve technology skills, often through professional development (an external barrier in its own right), teachers’ visions of technology in their classrooms might be affected. Both Kopcha and Buabeng-Andoh (2012) contributed to the field of belief-based barriers to educational technology integration.

A final barrier to technology integration identified in the literature focused on the technologies themselves. Pilgrim and Berry (2014), observed that integration of technology into the contemporary North American K12 classroom depended on teachers having the correct sets of technological skills for the tools they might employ. The authors noted, “technology integration requires practice with technology tools used in the 21st century” (p. 138). The authors concluded that having experience with technology in general might not be adequate and that educational technology integration projects might be disrupted if technological competence or capability was not pertinent to the technologies at hand.

Clearly, there were many potential barriers that could affect educational technology integration. While there were numerous manners in which practitioners could identify barriers, researchers often identified similar barriers in unrelated studies. An appropriate research path pursuant to the documentation of barriers preventing educational technology integration was an examination of the tools used to measure technology integration.

Assessing or Evaluating Integration

Having reviewed definitions of educational technology integration and completed a review of barriers that affect teachers’ integration efforts, a review of the literature on integration evaluation options could provide insight into the manners by technology endeavors could be evaluated. Hancock, Knezek, and Christensen (2007) attempted to create a powerful and far-reaching analysis instrument by combining three popular instruments. The product provided by the authors indicated that a combined assessment instrument, comprised of the “Stages of Adoption of Technology, CBAM Level of Use, and ACOT stages of technology evolution” instruments, could reliably and validly measure technology integration (p. 20). The authors noted that the combination of the three instruments, which collectively measured a vast array of

internal and external factors, provided “a quick and easy self-appraisal of level of technology integration” (p. 20).

While Hancock, Knezek, and Christensen (2007) created an assessment instrument by combining existing instruments, Schibeci et al. (2008) devised a flowchart outlining the advancement towards technology integration. The flowchart was useful as a tool to measure individual teachers’ progression towards technology integration. Schibeci et al. provided examples of classroom behaviors at each of three stages in the progression, and these examples might be used to identify a given teacher’s integration level at any time. The stages outlined in the authors’ flowchart were each comprised of three categories of skills, briefly summarized as pedagogy, technology use and confidence, and classroom management. The authors concluded that the instrument could be used to identify current levels of technology integration, and provided suggestions for encouraging teacher progress towards higher levels.

Researchers have taken divergent approaches to creating or employing assessment instruments. Conrad and Munro (2008) undertook a quantitative study in the development of a new technology usage measurement instrument. The instrument, titled *The Computer Use Scale (CTUS)*, was deemed capable of evaluating many factors related to educational technology integration, including computer self-efficacy, attitudes towards technology, and technology-related anxiety (Conrad & Munro, 2008).

Groff and Mouza (2008) developed a framework through which practitioners could predict the success of technology integration endeavors. The framework could also predict potential barriers to project success. In the development of their instrument, the authors identified multiple tiers at which obstacles to integration might be present, including the situational context of the project, the project itself, the teacher, the students involved, and the

technology selected for the undertaking. The results provided by the instrument, though untested at the time of publication, were intended to increase the likelihood of project success.

Employing an approach similar to Schibeci et al. (2008), Inan and Lowther (2010a) devised a path model for advancing towards educational technology integration. In the development of the model, the study examined both the effects of teachers' characteristics with regards to educational technology integration as well as the effects of external factors such as institutional culture. They concluded, "the model could provide practitioners and stakeholders guidance on how to resourcefully invest money [for technology integrations] by identifying and prioritizing critical factors" (p. 6).

While many studies focused on the development of instruments intended for use in the educational realm, Rosen, Whaling, Carrier, Cheever, and Rokkum (2013) reported on the development of a general-use scale that measured technology and multimedia usage. The instrument included a number of categories of competencies, themselves organized by numerous specific skills. The authors sought to devise an instrument that was based upon prior self-assessment models but emphasized frequency of technology use as opposed to time spent using the technologies. Rosen et al. (2013) created their instrument, the *Media and Technology Usage and Attitudes Scale*, using a granular approach in which the instrument measured sixty specific technology capabilities. The researchers designed the instrument so that subsections could be used to measure competency within a specific domain of media or technology use.

Assessment Risks

While there was significant scholarly work regarding the development and usage of technology assessment strategies, it was appropriate to consider how these instruments and methods could fail. For example, Kopcha and Sullivan (2007) concluded that data gleaned from

teachers' responses in self-reporting technology usage surveys were "inaccurate" (p. 640). The authors found that teachers often inaccurately portrayed their usage of educational technology compared to that of their peers.

Echoing Kopcha and Sullivan (2007), Reinhart, Thomas, and Toriskie (2011) reached a similar conclusion. They conducted an extensive study to determine the differences in educational technology usage across varied demographics. The study procedures included a comparison of teachers' perceived technology skills with their actual capabilities. Shedding doubt upon the accuracy of assessment tools, commenting "it was interesting to note that teachers' general perception of their technology skills [...] exceeds the means found in a detailed analysis of the actual technology skills they employ" (p. 190). This conclusion might serve as an important reminder of the risks inherent in assessment instruments.

Maderick's (2013) findings echoed those of his predecessors. The researcher stalwartly warned against using self-assessments for technology skills as a standalone gauge. Upon completion of an in-depth study of the validity of self-assessments, the author noted that when self-assessments were "utilized as the only measure of digital competence, there was a strong possibility that [the assessment] was inaccurate" (p. 115). The author elaborated further, noting that teachers who tended to misrepresent themselves in self-assessments were often amongst those with the weakest technology skill sets. Maderick, joined by Zhang, Hartley, and Marchand (2015), confirmed the researcher's earlier findings.

This review of the literature highlighted many facets of technology integration, with emphasis on the barriers that threaten integration efforts. The next section of this literature review will focus on actions that might be taken to promote educational technology integration.

The tactics may be either proactive or reactive and might be manifested in different ways. One approach, however, appeared recurrently in the literature.

Recommendations and Actions

To address challenges related to educational technology integration, particularly challenges that pertain to technology competence or capability, researchers have suggested numerous actionable approaches. A repeated theme within these recommendations was the use of professional development. Researchers differed, however, on their recommendations for the type, style, delivery method, and other elements of professional development.

Professional Development Models

In a section addressing historical problems with technology integration, Okojie, Olinzock, and Okojie-Boulder (2006) provided useful insight into technology training methodology. Citing Means (1994), the authors noted that professional development in the form of technology training must go beyond an emphasis on the procurement of technical capabilities. Okojie et al. (2006) advised that these training sessions devote time and energy to the strategies required to integrate the technical skills into the learning process. The authors recommended adapting instructional technologies to align with course objectives based on student needs, selecting technological methods that were relevant to learning goals, choosing evaluation methods that were relevant to the objectives, the modes of instruction, and to the technologies in use. Additional recommendations for professional development topics for teachers included designing follow-up activities and course enrichment materials that leverage technology in a manner that was accessible to students and supports the initial learning goals. Finally, the authors advised that training sessions encourage teachers to leverage interactive and engaging technology in their development of a dynamic classroom (Okojie et al., 2006).

While Okojie et al. (2006) provided useful guidance for teachers, Zhao and Bryant (2006) offered methodologies and techniques for administrators charged with devising professional development programs, concluding that technology training alone was insufficient in encouraging technology integration into the classroom. The researchers touted mentoring from a technology-savvy peer as an excellent and beneficial professional development model. The authors noted, “with mentor support, [teachers] were able to utilize technology resources regularly to move toward more student-centered learning” (p. 59). Indeed, mentor-based training as a follow-up to professional development was deemed an essential element of the technology integration process and advised administrators to plan for and support this model.

Palmieri, Semich, and Graham (2009) also guided administrators. The authors determined that additional professional development could enhance technological competence and thus technology integration and that the content covered in these sessions could be selected more effectively. Professional development planners could accomplish this latter goal through analysis of the data collected from assessment instruments such as Flowers and Algozzine’s (2000) *Basic Technology Competencies for Educators Inventory*. They advised administrators to employ data collection instruments, use the information gathered to gain an understanding of technological realms in which their faculty lack skills, and then devise professional development to address these problematic areas.

Professional Development Design

Other researchers contributed recommendations for professional development design. Ertmer et al. (2010) provided a simple but powerful strategy for teacher training. Based on the results of their study, the researchers suggested that administrators and organizers of professional “should be utilizing the same technology tools for professional development as the teachers were

able to use in their classrooms” (p. 434). Employing the same technologies in training and teaching could provide a training experience that was both contemporary and authentic (Ertmer et al.).

Davies (2011) provided useful but divisive recommendations for the design of professional development. The author noted the importance of teachers gaining proficiency with specific technologies as well as opportunities to select the educational technology tools that complemented their learning technology goals. Davies noted, however, that technology training was beneficial mostly as a method for helping teachers to gain new technology skills. Per the author, “[technology training] typically involves only the lower levels of literacy” (p. 51). Professional development programs intended to encourage practical competency (beyond simple technology operation) must consider the pedagogic explanations of the technologies addressed in training sessions (Davies, 2011).

Kopcha (2012) considered an additional contemporary approach to professional development. The scholar employed communities of practice, or groups of individuals working together to advance towards a learning-centered goal (Wenger & Lave, 1991) as well as mentor-mentee groupings as professional development models. Though both communities of practice and mentoring were found to be effective, Kopcha (2012) noted that mentoring was more effective in advancing educational technology integration practices. Specifically, “teachers reported that the mentor helped improve their beliefs in their own ability to plan and implement technology-integrated lessons” (p. 1118). The author, channeling Zhao and Bryant (2006), concluded that mentoring was an effective and powerful contributor to a professional development approach aimed at improving educational technology integration.

Optimizing Professional Development

Regarding effective professional development offerings, numerous researchers have provided guidance by which professional development planners might maximize the impact and value of this training approach. Ertmer and Ottenbreit-Leftwich (2012), citing Hew and Brush (2007) noted that content of the training must include technology knowledge and skills, pedagogical knowledge and skills supported by technology, and technology-related classroom. Ertmer and Ottenbreit-Leftwich (2012) purported that professional development designed with these principles in mind was likely to be effective.

Tasir, Abour, Halim, and Harun (2012) provided important guidance to professional development administrators and planners. The researchers found that teachers' satisfaction with professional development programs had a direct and positive impact on their confidence and competency with technology. The authors advised that the overall professional development experience should be considered carefully during the design phases and noted that training program decision makers "must formulate strategies that not only might increase teachers' satisfaction but also exceed their expectations of the acquired knowledge that they might gain at the end of the course" (p. 143). This finding had repercussions that might apply to both the content covered and delivery methods used in professional development.

Summary: The State of Educational Technology Integration

There was extensive research on the topic of educational technology integration, and this research was used to inform and devise a new approach towards addressing the challenges related to this complex but essential contemporary practice. Implicitly or explicitly, all definitions of educational technology integration included a need for teachers to possess an acceptable level of technology skills and competency. Further, researchers have identified many

barriers which could prevent or disrupt integration, but nearly all scholars included competence and capability in their writings. Similarly, academics have created a number of assessment instruments that specifically focused on measuring technology competence and deemed many as accurate predictors or gauging tools for technology integration. Finally, this review highlighted numerous recommendations to address and increase technological. This study addresses the identification of educational technology integration challenges rooted in a lack of technical skills by mining data extracted from helpdesk databases.

Educational Data Mining

The proposed methodology and framework, which employed data mining of helpdesk databases (HDDBs) to discover technology topics for professional development for teachers, was situated in a contemporary research field. This area of study, Educational Data Mining (EDM, with sibling fields Knowledge Discovery in Databases, KDD, and Learning Analytics, LA), was a relatively young research domain (Romero, & Ventura, 2006; Romero, Ventura, & García, 2008; Baker, & Yacef, 2009; Koedinger, D'Mello, McLaughlin, Pardos, & Rosé, 2015). Indeed, the first conference, journal, and textbook dedicated to the field were all established within the past decade (Romero & Ventura, 2013). The field was considered beyond infancy (Koedinger et al., 2015) but within adolescence (Romero & Ventura, 2010) and not yet mature (Mohamad & Tasir 2013; Romero & Ventura, 2013), so the number of studies conducted to date and the number of researchers in the field was comparatively low, and the number of domain experts even more so. Several of the researchers already cited were principal architects in the development of the aforementioned conference and publications. The works by these researchers were at the forefront of this nascent field.

This review of the literature provides a detailed synopsis of the state and practice of EDM. An overview of the field is followed by an examination of the key features of the domain. These components include the stakeholders in EDM undertakings, primary focus areas, goals, and trends within EDM research. Next, educational data mining methods will be explored, with emphasis on two mining techniques. As the new framework leverages an underused data source, this review also includes a survey of the data sources currently employed in EDM research, and addresses the current (limited) usage of helpdesk databases. This section of the literature review thus provides a detailed description of the field of research and practice which will house the new EDM-supported technology integration framework.

An Overview of Educational Data Mining

Data mining (DM) in the corporate realm, as a field, predates educational data mining (EDM), though the practices of the former helped shaped the latter (Merceron & Yacef, 2005). Heathcote & Dawson (2005, citing Fayyad, Piatetsky-Shapiro, & Smyth, 1996) provided a concise and authoritative definition of DM, noting that DM was a method for pattern discovery in databases. Leaders and administrators could leverage these patterns in business decision-making processes. Simultaneously, other researchers noted the existence of commercial products designed to analyze data from website server logs (Monk, 2005). At the time, the most common practice of such analysis was to glean information from e-commerce websites to improve sales (Romero & Ventura, 2006, 2010).

Despite rapid growth in corporate DM, the advancement of DM applications in education lagged behind. In 2005, while leveraging business data mining techniques to study information gleaned from a learning management system, Merceron and Yacef referred to the usage of data mining methods in education as “scarce” (p. 467). The researchers described data mining as

leveraging a wide variety of algorithms with distinct methodologies and purposes intended to explore, explain, and visualize trends and results conveniently. Merceron and Yacef concluded that the practice of data mining in education had the potential to inform pedagogical decisions. Monk (2005) elaborated further, noting that EDM practices applied to e-learning tools provided a method to understand learners in a way which might inform decisions regarding teaching practices, content delivery methods, and even infrastructure expenditures. These suggestions were amongst the first detailed explanations of the real-life applicability of EDM.

In 2006, Romero and Ventura published what was considered a seminal paper in the annals of EDM research. The paper presented a survey of EDM practices from 1995-2005. Without providing differentiation between DM and knowledge discovery in databases (KDD), the authors, citing Klösgen and Zytkow (2002), referred to both practices as the automated extraction of patterns from large databases. Romero and Ventura then noted that DM or KDD methods had already been applied to inform the learning process, create student models, to glean information from learning portfolios, and to evaluate and improve e-learning systems.

Interestingly, and denoting the first in a decades-long pattern, Romero and Ventura (2006) do not elaborate on the potential for EDM to provide insight into the errors or problems that teachers or learners might encounter while using e-learning systems. Indeed, the authors provided a detailed list of educational domains in which EDM practices could be beneficial. These included the traditional classroom, as well as technologically enhanced educational areas such as distance education, web-based courses, learning management systems, and intelligent tutoring or adaptive educational systems. There was no mention of the potential for the application of EDM practices to data collected when users experience technical errors with these

systems. The authors did, however, provide compelling insight into the transformative power of successful EDM.

Romero and Ventura (2006) praised the results of EDM methods. The researchers noted that data mining techniques applied to appropriate data could assist in the formative evaluation of educators. Information unearthed via EDM could be used in these assessments to help educators determine a pedagogical basis for decision-making as they modified or designed a learning environment or teaching approach. The authors artfully elucidated this concept, stating that EDM methodologies were capable not only of “turning data into knowledge, but also filtering mined knowledge for decision making” (p. 136). Regarding potential beneficiaries from EDM undertakings, Romero and Ventura asserted that EDM had the potential to benefit teachers, students, administrators; “each actor of the learning process” (p. 137). A detailed review of EDM stakeholders follows this section.

While Romero and Ventura (2006) touted the potential for EDM to guide decisions that might affect an array of educational participants, the authors repeatedly noted that EDM as a research and practice field was still in its early stages of development. Indeed, the authors referred to EDM as a “promising,” “upcoming,” “very recent,” and “young research area” within a single page of their survey (p. 144). Looking towards the future, Romero and Ventura encouraged researchers to expand the usage of EDM and reminded researchers that a tenet of EDM practices was the need to consider pedagogical aspects of the learning process. This consideration, the researchers noted, was the primary distinction between DM and EDM. The authors also commented that EDM studies might leverage existing DM practices as long as the study incorporated an educational context or component. Finally, Romero and Ventura

recommended continued and increased application of educational data mining, and called for the standardization of EDM methods and techniques.

Following their recommendations, Romero and Ventura, joined by García (2007), conducted an EDM case study, applying EDM techniques to Moodle, a popular learning management system (LMS). Early in their report, Romero, Ventura, and García postulated a practical and pragmatic driving purpose for EDM. Quite simply, e-learning systems such as LMSs and other educational software tools generated massive amounts of data each day, and this quantity of data was difficult to manage manually. The researchers qualified this further, adjusting the developing definition of EDM to include the “discovery of non-obvious valuable patterns from a large collection of data” (p. 376). To explain the discovery process, Romero, Ventura, and García described and employed numerous mining techniques, including “statistics, visualization, clustering, classification, rule mining, sequential pattern mining, text mining, etc.” (p. 369). The study presented an easily reproducible approach to an EDM undertaking, highlighting the general EDM process. This process begins with the collection of data, followed by pre-processing the data (preparing it for analysis). Subsequent steps include the application of data mining techniques and evaluation, interpretation, and eventually taking informed action based on the results. Romero, Ventura, and García engaged in each step and provided a high-level overview of several standard EDM methods. The model proposed by Romero, Ventura, and García in 2008 foreshadowed the approaches employed in many future studies, and the authors concluded their report as before: calling for the standardization of EDM practices and highlighting the young age of the field.

In 2009, Baker and Yacef published a highly-regarded survey of EDM to date, with projections towards the future. This survey represented one of Baker’s earliest important

contributions to EDM study, though he would become a prominent researcher within and champion of the field. Published in the first issue of the *Journal of Educational Data Mining* in the fall of 2009 and highlighting the relative youth of the field, the article refers to the EDM research community as “nascent” within the first six words of the publication (Baker & Yacef, p. 3). The authors described the emerging field as interested in the development of methods for examining data unique to educational environments and leveraging those methods to understand learners and learning contexts. Further, the authors echoed previous definitions of EDM and KDD, highlighting that the aim was to discover new and useful information from large datasets. Situating the field within a particular realm, the authors noted that EDM “emerged from the analysis of logs of student-computer interaction” (p. 4). More specifically, the authors highlighted e-learning platforms as a prominent source for the origins of modern EDM. The report by Baker and Yacef, leveraged here to provide an overview of the origins and the state of EDM at the time of publication, will be revisited in later sections of this review. Additional survey research reviews the continuing development of the EDM field.

Baker (2010) continued to participate in the expansion of EDM as a research field by contributing an overview of data mining for education to *International Encyclopedia of Education*. The researcher described the area in similar terms as he had used in the past, however, in this paper, he emphasized the use of EDM to investigate scientific questions within educational research. While the field continued to develop into a scientific area of inquiry, the practice of EDM continued to rely upon data generated uniquely by educational systems, and analysis intended to inform education-related decision-making. Echoing previous contributions to the field, Baker included prediction, clustering, relationship mining, and discovery with models as primary EDM approaches. Baker added another method: distillation of data for human

judgment. While a later section of this literature review will address many of these methods, it was beneficial to this project to clarify Baker's new approach. The goal of distilling data for human judgment as an educational data mining methodology was to analyze and summarize data so that a person could identify and organize features of the data. Researchers commonly used this method when looking for patterns in learner behavior.

In addition to EDM methodology, Baker (2010) also introduced the main applications of EDM. These included improving student models, understanding the nature and design of knowledge structures, reviewing the pedagogical support presented by education-oriented software systems, and contributing to scientific discovery about cognition and learners. The publication of this article, which provided an overview of EDM and its history, as well as conventional methods and applications, in a well-known journal could have brought EDM practices closer to the forefront of scientific inquiry in education.

Romero and Ventura (2010) continued the expansion of EDM into new realms when, in 2010, they published an additional overview of the state of EDM research, incorporating their previous publications as well as those by other researchers, including Baker and Yacef (2009). This article served as an update to Romero and Ventura's 2006 paper, and thus much of the content was reiterated. Romero and Ventura (2010) accepted Baker's (2010) contributions and included his methods and applications in their publication. Citing Castro (2007), Romero and Ventura expanded on Baker's EDM applications to include: assessment of students' learning performances, providing course and learning adaptations and recommendations based on learning behavior, evaluating learning material and digital education services, providing feedback to both teachers and learners in e-learning programs, and the development of systems

to highlight undesirable student behavior in e-learning. The close relationship between EDM and digital learning tools was increasingly apparent and important.

Indeed, Romero and Ventura (2010) provided significant commentary on the relationship between digital learning software, web technologies, and EDM. The authors touted the rapid rise of educational software and the trend towards storing information from these systems in large databases as being highly influential on the growth of EDM. Similarly, Romero and Ventura highlighted that the use of the Internet in education, often referred to as e-learning or web-based learning, had generated enormous amounts of accessible data. The authors referred to this store of information as “a gold mine of educational data” (p. 601). The researchers presented EDM as the means for mining the data cache.

Increasing data sources. While Romero and Ventura (2010) highlighted the value and importance of data generated by teachers and learners as they used digital education tools, the authors made no mention of the data generated when those involved in education encountered issues or problems with the technical tools they relied upon increasingly. This phenomenon, however, might not be surprising, as the focus of EDM remained on analyzing education data to contribute towards pedagogic and educational decision making (Romero & Ventura, 2010). Nevertheless, the data sources leveraged to date pertained almost exclusively to teacher and student interactions with digital learning platforms.

Baker and Siemens (2010) were among the first to address a lack of diversity in data sources. Baker, by now a notable EDM scholar, joined Siemens (a learning analytics and knowledge, LA or LAK, proponent) in writing a brief paper aimed to unite the two similar fields. Regarding data sources, the authors noted, “increasingly, very large data sets were available from students’ interactions with educational software and online learning - among other sources” (p.

252). Though the authors did not elaborate upon the additional sources, the very mention of the possibility of data sources aside from those associated with learning software and e-learning platforms represented a shift from EDM norms.

Romero and Ventura (2013) noted that EDM practitioners could analyze data created by any electronic information system in use at learning institutions, either by learning professionals, by learners, or by other participants in educational efforts. In contrast to past publications, the authors observed that the data sources “are not restricted to interactions of individual students with an educational system” (p. 12). Indeed, Romero and Ventura acknowledged that data sources could include text chats amongst students, data collected by school administrators, demographic information, and other data sources. Though the authors did not mention helpdesk databases by name, the trend towards new and previously untapped data sources was becoming evident.

In line with Romero and Ventura’s (2013) descriptions of the growth of data sources, Baker and Siemens (2014) acknowledged the progress of the field from this perspective. Noting that early work in EDM examined data from intelligent tutoring systems and e-learning platforms, Baker and Siemens cited student collaboration data, simulation data, teacher communications, and grading systems as viable new data sources. (Helpdesk databases, however, remain absent from lists or descriptions of the growing number of data sources leveraged in EDM research.) Aside from the mention of new data sources, much of the content of Baker and Siemens’s (2014) contribution echoed previous work, with regards to methods, practices, procedures, and applications. The authors, however, as Romero and Ventura (2013) had before them, noted that EDM studies and contributions continued to increase and gain in complexity and scope.

As the methods of EDM began to stabilize, the applications continued to grow. In 2015, Koedinger et al., relative newcomers to EDM, described EDM as a rapidly growing field that combined numerous areas of study. Regarding the application of EDM to pedagogic psychology, the authors endorsed the use of data generated by educational technologies, referring to this data as “ecologically valid data on student learning” (p. 333). Koedinger et al. proceeded to discuss the application of EDM methods towards the learning sciences and concluded with resounding praise for the field. Noting that the field had progressed since its infancy, and calling for more EDM research going forward, the researchers confirmed that “EDM techniques have now become essential tools for researchers interesting in modeling affect, motivation, and metacognition in digital learning contexts” (p. 350). This resounding praise for a maturing field highlighted the confirmation of EDM as an important, contemporary area of scientific inquiry that affected many people involved in the education process. Interestingly, concurrent with the increase in applications and data sources, the research area itself was branching out and evolving.

EDM and LAK. Baker and Siemens (2010) provided a useful comparison of learning analytics and knowledge (LAK) and EDM, highlighting similarities and differences between the two fields. The authors noted that both areas share the goals of improving education by improving analysis of large educational datasets and supporting both research and educational practice. Regarding the differences, two of the most prominent distinctions between LAK and EDM were the focus of each field and the manners in which practitioners leveraged technological tools. Baker and Siemens wrote: “EDM had a considerably greater focus on automated discovery, and LAK had a considerably greater focus on leveraging human judgment” (p. 253). Regarding the use of technological tools, the EDM field relies on technology for automated discovery, while the LAK field leverages technology to condense data for human

consumption. Though the authors emphasized a prominent overlap between the fields, they noted that “LAK models were more often designed to inform and empower instructors and learners” (p. 253). Similarly, LAK researchers might place a comparatively greater focus on addressing the needs of a variety of stakeholders, whereas EDM researchers might concentrate on deriving generalizations or hypotheses.

As the lines between EDM and LA continued to blur, scholarly research often referred to them interchangeably and simultaneously. Regarding the primary differences between EDM and LA, Romero and Ventura (2013) offered several distinctions. Chief among these was the difference in emphasis between the fields: LA emphasized the descriptions of the data and the analysis results, while EDM emphasized the description and categorization of the methods used. Further, echoing Baker and Siemens (2010), Romero and Ventura noted that in LA, automated discovery was a tool intended to support and contribute to human judgment. Conversely, in EDM, automated discovery was the primary goal, and human judgment was considered a tool used in the process.

EDM and LAK methods for this undertaking. Despite these noteworthy differences, Baker and Siemens (2010) acknowledged that the fields were similar enough that the lines between them might be blurred. The authors clearly noted that researchers who identify as LAK practitioners regularly undertake projects that fall into the EDM domain, and vice versa. Indeed, when discussing the project of mining helpdesk databases for technological topic discovery, with regards to the differences between EDM and LAK, Baker commented: “I don't think these differences really matter that much for your project” (R.S.J.d. Baker, personal communication, July 23, 2016). The scholar then encouraged this researcher to leverage the elements from both

communities that best supported the undertaking and the participants or interested parties who could benefit.

Stakeholders

The project under review intended to use educational data mining approaches to analyze helpdesk databases with the goal of discovering technical topics that were problematic for teachers. As the researcher, also a school administrator at the site, conducted the study, and the participant audience was teachers, it was important to consider who could be affected or influenced by EDM undertakings, and how those effects might manifest.

Romero, Ventura, and García (2008), in their EDM study of the Moodle LMS, noted that all levels of education, from schools to universities to businesses, employed LMSs and other online learning environments. The authors applied EDM techniques to data extracted from Moodle and concluded that the data mining processes were beneficial in providing valuable decision-making information to improve Moodle courses as well as student learning. Within an appropriate context, EDM techniques could be used to assist in decision-making processes for learners of all ages.

The authors elaborated further on the usability of knowledge discovered from EDM (Romero, Ventura, & García, 2008). Romero, Ventura, and García noted that both educators and students could make use of information gleaned by EDM, and thus the EDM process could be adjusted to derive information pertinent to a particular audience. EDM practitioners might apply their processes towards learners with the goal of recommending activities, resources, directions, or materials that might assist in the learning process. Similarly, EDM processes might be oriented towards educators to provide feedback, evaluate the structure and effectiveness of a course, categorize learners for differentiated instruction, or for finding commonly made

mistakes. Further, EDM approaches might be adjusted to address the needs or curiosities of administrators to understand challenges such as efficiently using infrastructure and resources. Finally, the authors noted that academics might leverage EDM methods as they sought to understand learning and the educational process.

Romero and Ventura eloquently summarized the applicability of EDM to various stakeholders in their 2013 review of the field. Citing Hanna (2004), they wrote, “different groups [of users or participants] look at educational information from different angles according to their own mission, vision, and objectives for using data mining” (p. 602). The authors once again provided a list of stakeholders who could benefit from EDM undertakings. As before, this list included learners, educators, administrators, and academics. However, Romero and Ventura improved on the previous list by adding additional stakeholders. These included course developers, who might use EDM to evaluate courses or course software or to review course content and structure. Additionally, Romero and Ventura added learning organizations (such as universities or private training companies) as potential stakeholders who could leverage EDM practices. These organizations might employ EDM methods to assist in decision-making about course suggestions or students qualified for graduation or advancement.

Romero and Ventura again provided lists of potential stakeholders or beneficiaries of EDM studies in 2010 and 2013. In each iteration, the researchers identified new and exciting applications of EDM. While the variety of applications of EDM seemed to grow over time, the list of stakeholders remained relatively consistent. Thus, while the list of interested parties might increase as new studies occurred, the predominant stakeholders remain learners, educators, course developers, organizations, administrators, and scholars. Just as the list of stakeholders had evolved, so had the focus of EDM studies.

EDM Focus Areas

One of the earliest prominent applications of EDM was the analysis of databases generated from web-based educational platforms (Merceron and Yacef, 2005). These databases stored information about how students navigated through the systems, including how often they accessed resources and how much time students spent on any given page. A statistical review of these databases could provide an overview of class progress (Merceron and Yacef, 2005). Merceron and Yacef noted that many early EDM applications mined web-learning platform databases to control digital learning objects, externalize and visually present data about student progress, predict student performance, or to link common mistakes. The authors considered these tasks as comparatively basic but noted that the quantity of data was vast and deserved a more nuanced analysis method. By leveraging EDM methods, Merceron and Yacef determined that LMS data could be analyzed to provide evaluative input into learning and teaching practices, and thus the data was useful beyond simple analyses that described the usage of the platform.

Heathcote and Dawson (2005) performed a similar study and arrived at similar conclusions. The researchers analyzed data from an LMS but noted that the data was limited and mainly technical in nature. For example, Heathcote and Dawson highlighted page hits, counts of unique user visits, uptime and downtime, the number and type of resources integrated into the site, and the file sizes of resources and the LMS itself. The authors noted that the characteristics of the data made it mostly applicable to analyses by information technology support staff, who might leverage the analyses for planning and monitoring. As with Merceron and Yacef (2005), it was the vast quantity of data, and the potential for discovery therein, that encouraged Heathcote and Dawson to apply more advanced data mining techniques to the LMS databases. Similarly,

Heathcote and Dawson's conclusions echoed those of Merceron and Yacef, as Heathcote and Dawson found pedagogic meaning and trends within the technical data.

As EDM advanced as a field, scholars continued to attempt to extract increasingly meaningful data from educational databases. Indeed, by the publication of Romero and Ventura's seminal 2006 EDM survey, the authors had noted that "data mining techniques could discover useful information that could be used in formative evaluation to assist educators establish a pedagogical basis for decisions when designing or modifying an environment or teaching approach" (p. 136). Thus, despite the relative youth of the field, practitioners of EDM were already regularly succeeding in extracting useful and powerful pedagogic information from educational databases. With regards to students, EDM practices aimed to recommend learning plans, activities, and resources that might improve learning. With regards to educators, EDM practices could offer objective feedback, evaluate course content, and discover learner behaviors that might warrant modification of courses. With regards to administrators, EDM practices aimed to help school administrators better organize resources, review their educational offerings, and enhance programs. Interestingly, Romero and Ventura (2006) also noted that institutions continued to gain useful information technology feedback from EDM practices, including how to design networks and deploy network resources. Thus, technical feedback and review remained a focus of EDM projects, even as the authors noted that "the objective of data mining in e-learning was to improve the learning" (p. 136).

By 2010, the list of key areas of application of EDM no longer included information technology department improvements (Baker, 2010). Baker listed four areas of EDM application that were studied most commonly in contemporary EDM research. These included the improvement of student models, the discovery or development of knowledge structures within an

institution, reviewing the pedagogical support offered by digital learning tools, and scientific discovery about learners and cognition. Baker based his areas of concentration upon counts of research topics, and his efforts highlighted that the trends away from EDM applications for information technology benefits and towards pedagogy were becoming apparent.

EDM Trends

In the early years of EDM as a research field, the most prominent field-wide trend was growth. As illustrated, 2005 to 2008 featured a flurry of survey or overview articles intent on introducing researchers to the burgeoning field. Practitioners were also beginning to expand the applications of EDM processes, beyond its roots within e-learning and basic, technology department-friendly statistics. By 2009, however, domain-wide trends began to appear. Baker and Yacef (2009) identified a noteworthy change in the applications of EDM. While EDM research efforts before Baker and Yacef's contribution to the first issue of *Journal of Educational Data Mining* leveraged EDM methods for relationship mining, the authors noted a decline in relationship mining studies and an increase in the use of EDM for prediction. Relationship mining efforts typically aimed to determine accords, connections, or affiliations between trends in the data. These efforts were important undertakings in the early years of EDM, but the trend towards prediction indicated a shift towards more sophisticated analyses. EDM used for prediction attempted to analyze existing data to envision future trends. A typical example of EDM for prediction was estimating a student's performance and grades based on data about past academic achievements.

While Baker and Yacef (2009) noted the trend towards more complex data mining, the researchers identified another significant trend in the field. Baker and Yacef highlighted that during the infancy of the field, from 1995-2005, data mining practitioners typically had to collect

the data they intended to mine. As the field expanded, a nearly domain-wide trend became apparent: researchers began to use existing databases rather than collecting their own data (Baker and Yacef, 2009). This change was ushered in by increased usage of LMSs and other learning technologies that were capable of storing significant amounts of data in easily accessible databases. One of the primary benefits of this trend, Baker and Yacef noted, was the increasing ease by which any particular study could be replicated, validated, or refuted.

In 2010, Baker elaborated further upon the trend towards leveraging existing data sources. Baker highlighted the creation and popularity of education institution-sponsored repositories of data. The creation of these widely available educational data repositories further facilitated data mining process for practitioners, and, again, invited scholars to perform replication studies for comparative purposes. Finally, the expansion of the use of LMSs (and the data stored within) provided researchers with similarly-styled data from a wide variety of educational contexts. These similar datasets provided the opportunity for researchers to study the impacts of contextual influences upon learning. Baker clearly highlighted the changes in data usage within educational data mining through 2010.

Aligning with societal trends. Just as Baker (2010) highlighted shifts in the availability and consistency of data, Mohammad and Tasir (2013) identified broad societal changes and the impacts those might have on EDM. In their review of the EDM field, the authors acknowledged the ongoing trend of situating EDM studies within e-learning but addressed a timely shift for future research. Mohammad and Tasir called for a change from mining e-learning data towards mining the data generated by social media, blogs, and online relationship cultivation systems. The authors noted that “these applications already gained high popularity among students and [were] suitable to be used to engage the students with collaborative learning” (Mohammad &

Tasir, 2013, p. 323). Though previous scholars (Romero & Ventura, 2010) had mentioned social network analysis, many EDM publications from 2013 onwards would begin to include social network analysis as an application of EDM practices. By considering trends outside of the most popular EDM realms, Mohammad and Tasir helped popularize a new data source that could be mined to provide insight into learning.

While Mohammad and Tasir (2013) associated EDM trends with greater societal trends, Baker and Siemens (2014) reviewed the trends that propelled the growth of the EDM field. These trends were the increase in the quantity of minable data, an improvement in the organization and format of these datasets, improvement in the capabilities of computing hardware, and improvement in the functionality of data mining software. Of these, perhaps the most impactful were the significant increase in the quantity of minable data, as more learning technologies provided the capability to store student-system interaction data. Echoing Romero and Ventura (2013), Baker and Siemens highlighted this trend as being a prominent contributor to the rise of EDM as a field of scientific inquiry.

Baker and Siemens (2014) noted two new trends in the EDM domain. The first was the emergence of a positive feedback cycle between research and practice in EDM. As the EDM field matured, EDM procedures were applied to aid in decision-making processes. Upon implementation of these decisions, new data could be collected, mined, and analyzed to review the impacts of the decisions. Similarly, as researchers developed new ideas in EDM study, the methods were applied to EDM practice, and the results, once again, could be measured and reviewed. In turn, researchers might then develop new areas of investigation. Thus, as the field grew, a positive relationship between theory and practice developed.

The second trend highlighted by Baker and Siemens (2014) was the application of EDM methods in an ever-increasing list of fields. From its beginnings in analyzing e-learning system transaction data, EDM methods had grown to find meaningful use in areas such as gaming, argumentation, virtual worlds, and teacher learning. The expansion and applicability of EDM practices had thus grown rapidly, and as the field gained maturity, it also gained widespread relevance.

Missing from EDM trends. Despite significant growth in the field, a trend not seen in the EDM literature was the mining of helpdesk databases to determine topics for professional development that might help those submitting requests to the helpdesk. Administrators who planned professional development might find this unique teacher-technical skill feedback useful. However, helpdesk data remained absent from the list of data sources. This case study posited that knowledge and understanding could be gained by leveraging standard DM methods applied to HDDBs. The methods leveraged in EDM practices by EDM practitioners have stayed relatively consistent. These methods, discussed in detail in the following sections, were applicable nearly universally to any EDM undertaking and their potential applicability to HDDBs was at the core of the case study under review.

EDM Methods

There were numerous methods available for mining educational data. Many of these methods involved advanced statistical analysis, the details of which were beyond the scope of this literature review. An overview of the most common analysis methods, however, was essential to understanding EDM and its practices. Many of the methods have remained largely unchanged throughout the young history of EDM. However, the execution of these processes had improved due to the advances in technology-supported analysis (Baker & Siemens, 2014).

Numerous methods were reviewed, with particular attention paid to those that were most applicable to the study at hand.

As early as 2005, researchers had established the most common methods for EDM analyses. In their early case study of EDM practices, Merceron and Yacef (2005) highlighted four common mining approaches, including data exploration and visualization, clustering, classification, and association rules. Data exploration and visualization were used to present the data in a human-consumable manner, typically via charts, graphs, or histograms. Clustering, a common analysis approach, involved combining or grouping similar or analogous data points. Classification, a prediction method, was used to forecast unknown values based on the values of a known variable. Association rules were used to find relationships between items. Per Merceron and Yacef, these four approaches represented the core methodologies employed in EDM research.

Merceron and Yacef (2005) also highlighted an essential method in the act of executing data mining: data preparation or preprocessing. The authors often reiterated the important step of priming and normalizing the data for analysis in the documentation about successful data mining. The researchers noted that the preparation steps varied based on the analyses in use, and data preparation practices could be very different amongst studies.

Regarding the steps carried out in data preprocessing, Romero and Ventura (2007) elaborated in greater detail. While the preprocessing steps will vary by project, most undertakings included data cleaning, a nearly ubiquitous action. Cleaning was used to normalize data for consistency, and to remove incorrect or incomplete data points or records. Many preprocessing steps (such as user identification, session identification, and path completion) were performed automatically by web-based platforms. These steps identified the type of user, what

they were doing within the online tool, and the routes they navigated within the platform.

Additional preprocessing steps included data transformation, which generated new data attributes from existing attributes; data integration, which bolstered the existing dataset with data from another source; and data reduction, intended to minimize the size of the data elements for ease of use. Romero and Ventura also described other preprocessing steps, including anonymizing data as needed, and filtering data to remove data points that were irrelevant to the study at hand. Not all studies included the same preprocessing steps, though all studies typically required some data preprocessing.

Beyond the necessary preprocessing steps, Romero and Ventura (2007) expounded upon the EDM methods introduced by Merceron and Yacef (2005). Recalling the work of their predecessors, Romero and Ventura listed statistics and visualization, association rule mining, clustering, and classification as prominent methods. The authors further bolstered their list, adding web mining, outlier detection, sequential pattern mining, and text mining as useful analysis approaches. Outlier detection was grouped with clustering and classification, as all three methods were employed to bundle and organize data points. Romero and Ventura grouped sequential pattern mining with association rule mining as the process of finding patterns was similar to that of finding trends that link variables. Romero and Ventura introduced web mining and text mining as new EDM methods. Web mining involved the use of data mining techniques applied to data gathered from web-based tools. The process of web mining might focus on extracting information from the content, the structure, or the usage of websites. Text mining was related to web content mining but applicable to any text source. Indeed, with a goal of mining an HDDB, the researcher of the case study under review determined that text mining was especially relevant and this topic warranted closer investigation.

Text Data Mining

Romero and Ventura (2007) described text mining as an amalgam of several fields that included machine learning, data mining, statistics, information retrieval, and language processing. Text mining had universal applications, and practitioners might leverage this approach in studies of organized or disorganized data sources. The authors noted that practitioners might apply text mining to an email archive, for example, which contained disorganized data stored in an orderly manner. Another application of text mining included the grouping of documents based on content similarity, a process that employed clustering techniques as well. Regarding the use of text mining for pedagogic purposes, the authors provided the example that text mining could be leveraged to help an educator gauge the quality of students' online discussions.

Elaborating further on the role of text mining within an EDM study, Romero, Ventura, and García (2008) provided a useful overview of the text mining process. The authors highlighted that the process typically begins with structuring the data, often by parsing out invalid or unnecessary information. This step was in line with the aforementioned data preprocessing procedures. Once practitioners had properly structured the data, text mining could be applied to derive patterns from the data and present the patterns for interpretation. The authors noted that satisfactory discovery from text data mining typically yielded unique or novel topics, trends highlighting significance or relevance, or was capable of bringing to light issues of interest.

Romero and Ventura (2013) complemented their earlier procedural description of the text mining process. The authors noted that the text mining task might contain similar processes as other EDM methods. For example, text mining could include text categorization or clustering, as

well as idea extraction from a body of text. The text mining process could also include more complicated steps such as sentiment analysis, relationship modeling, or document summarization. By 2013, Romero and Ventura had found that researchers had used educational text mining had on discussion board data, chat logs, web forums, web pages, documents and other (unnamed) data sources.

Wu He (2013) published a case study on the application of text data mining methods to typed questions and chat messages recorded within a video streaming system. The researcher highlighted several challenges of using text mining and other EDM methods on this type of data. These challenges included the limitation of the clustering-based text mining tool's ability to only identify one pattern or trend for a given data point when the point might, in fact, contain text that was organized into multiple clusters. An additional challenge was that text mining could successfully identify patterns, but might not inform the researcher on the significance of those trends. Finally, He observed that text data mining was not useful in determining causality between behaviors and variables. Despite the drawbacks of the text mining methods He employed, the scholar concluded his research with powerful praise for the methodology. He wrote, "the study clearly shows the value of educational data mining and text mining as an alternative analytical approach in gaining insights from large amounts of untapped textual data" (p. 100). The author's inspiring work succeeded in using text mining methods to extract useful and informative information from an unused data source. Text mining was one of the many valuable tools that a practitioner could implement in an EDM workflow.

EDM Process

The EDM process itself had a distinct methodology. Romero, Ventura, and García (2008) highlighted the four fundamental steps in EDM application. The first step pertained to the

collection of data for mining. The next step involved preparing or preprocessing the data. The third step involved the application of data mining methods or techniques, and the final step was the interpretation of the results. This basic process was nearly ubiquitous in EDM projects, regardless of the data source, analysis methodology, or application. This procedure might be repeated several times throughout an EDM undertaking, as interpretations of the results might lead to decision making and changes in practice. While the process remains consistent, the data mining techniques varied by project but researchers typically employed methods from the aforementioned list of methods.

In 2009, Baker and Yacef reiterated the most commonly used EDM methods. These included the methods already discussed: statistics and visualization, web mining, clustering, relationship mining, and classification. Baker and Yacef reorganized the list of methods so that classification, for example, was now highlighted as a subcategory within prediction methods. Importantly, Baker and Yacef included two new methods for mining data: discovery with models, and distillation for human judgment. Discovery with models leveraged prediction methods but mined data with the intention of creating a model of some phenomenon. Often the model was used for additional data mining. The method of distillation for human judgment was especially relevant for this case study and warranted closer investigation.

Distillation of data for human judgment. Regarding the EDM practice of distillation of data for human judgment, Baker (2010) provided valuable insight into the approach. Baker wrote “in some cases, human beings could make inferences about data, when it was presented appropriately, that were beyond the immediate scope of fully automated data mining methods” (p. 116). This method thus explicitly incorporates human interaction into the data mining process, using data mining techniques to prepare data for human consideration. Per Baker,

typically, the data was summarized so that humans might identify or classify trends, patterns, or even the raw data itself. Specifically, Baker (2010) noted that when data miners distilled information for identification, the extracted data was available to human interpreters in a way that would allow for recognition of patterns that might be known but were difficult to identify or express formally. Peer researchers embraced Baker's suggestion of data mining for distillation of data for human judgment.

Romero and Ventura (2013) acknowledged the importance of distillation of data for human judgment when they included it with their list of EDM methods. The list included all of the methods that this review has addressed. Romero and Ventura situated the role of the distillation of data for human judgment process when they noted its purposeful role in helping educators visualize and summarize student course activities. The researchers praised this method for its capability of making large amounts of data accessible to human interpreters who would interpret the patterns in a way that an automated system might not.

Automating the Process

The most recent and impactful changes to EDM methodology directly addressed the use of automated systems. The tools by which researchers performed EDM methods and processes began to mature rapidly. Baker and Siemens (2014) noted that modern computer systems had experienced massive increases in processing and computation power and that these systems were capable of performing data mining tasks on large datasets quickly and accurately. The authors listed more than ten readily available data mining software systems that could be employed to execute data mining procedures. These included open-source tools (such as the R-project), commercial tools (such as RapidMiner and IBM Cognos), and even data mining plugins for commonly used software such as Microsoft Excel. In some cases, the companies producing these

tools offered training courses and certification exams. Some of these systems, such as RapidMiner, were available free of charge for scholars. Thus, EDM practitioners could now apply many of the most common and important EDM methods to large datasets conveniently and quickly.

Having established a baseline of data mining methods and the applications for which data mining was useful, an important topic for review was the sources of data. Without databases to analyze, there would be no potential in EDM practice, and without the mining of new data sources, the field's growth might be minimal.

Data Sources

Romero and Ventura (2013) provided a concise overview of the data sources most commonly found in educational settings. In traditional classroom-based education where student-teacher and student-student interactions represented the most prominent forms of interaction, frequently stored data includes attendance, grades, curricular goals, and individualized academic plans. Educational institutions were also likely to store administrative information, such as courses, schedules, and student information. Additional information for these organizations might be gathered from their website or online course offerings and stored in a database. Practitioners might have used any of these datasets in an EDM study; the authors do not draw a finite line marking the end of where minable educational data might be stored. On the contrary, by adding “and so forth” to their list of data sources, Romero and Ventura openly called for new data sources suitable for analysis (p. 16).

Computer- or technologically-supported educational systems could store different or additional information compared to traditional institutions (Romero & Ventura, 2013). E-learning platforms, learning management systems, intelligent tutoring systems, adaptive systems,

and testing or assessment systems were all likely to store valuable data in a minable database. Romero and Ventura noted that for both types of educational institution (traditional or technology-oriented), practitioners could leverage nearly any system employed by the organization which was capable of generating and storing data as an EDM source. The authors expanded the list of potential sources, noting that data could come from sources as diverse as administrative work, observations, questionnaires, or measurements from experiments. Thus, while different institutions might generate (or not) various types of data, nearly every organization generated data of some sort and practitioners could leverage these data sources for analysis and decision-making.

Koedinger et al. (2015) confirmed this trend in their recent review of the potential contributions of EDM to the study of learning and cognition. They noted, “the data that makes educational data mining possible was coming from an increasing variety of sources and was being used to address a variety of questions” about learning and the evaluation and improvement of learning (p. 333). Other contemporary researchers had supported the growth in data source variety, including an unexpected data source.

In 2012, Shum, Knight, and Littleton published a policy briefing for *UNESCO Institute for Information Technologies in Education*, which listed common data sources, such as learning management systems and student information systems, and noted that some of these systems incorporated basic analytical reporting. Additionally, more advanced analytical systems could also integrate data from “other university systems, (e.g. Helpdesk calls)” (Shum, Knight, & Littleton, p. 3). It seemed inevitable that helpdesk data would eventually appear as a potential data source in a field wherein the range of minable data sources continued to expand.

Helpdesk Data Mining

Though largely absent from educational data mining endeavors, helpdesk database (HDDDB) analyses were the subject of data mining undertakings in the enterprise realm since the turn of the millennium, and before. A review of the literature about academic HDDDB analysis projects highlighted certain trends which were prominent in the field. One of the primary trends was the group of users who submit requests to the HD in a given study. In nearly all HDDDB mining studies, this group was typically clients of a company, who have purchased goods or services from that company and now require after-purchase support, or employees of the company, contacting internal technical support for assistance with problems that interfere with their work. While this study (which will call upon the body of work in HDDDB mining) focused on HD tickets submitted by employees, there was value in researching HDDDB mining studies that addressed both groups of users. The primary difference between these groups in the enterprise realm was the source of funds that paid for support services. Fundamentally, the goals and operations of helpdesks, to provide quick and reliable support, were consistent regardless of the audience submitting requests for help.

The other primary trend observed in the domain of HDDDB mining publications was that the overwhelming focus of HDDDB mining studies to date was the improvement of HD or technical support operations. The goals of each study varied; some addressed minimizing resolution turnaround time while others focused on automatic problem resolution, but the underlying purpose was to improve support procedures to benefit business practices. To the one, publications in the field of enterprise HDDDB mining do not focus on leveraging HDDDB analyses to aid in the training of those who submitted help requests and relied upon the HD. While the audience might vary, those same studies unanimously do not address the potential for HDDDB

analysis to change the behaviors of the support requesters. The studies reviewed in this section of the literature review, which varied by the audience but not fundamentally in their goals, were indicative of the state of the field.

HDDB Mining to Support Customer Service

Jha and Hui (2000) applied data mining techniques to a corporate HDDB with the goal of supporting customer service. The authors described the steps performed by the HD agents in receiving phone based support requests, which, along with the problem resolutions, were then stored in an HDDB. Highlighting the importance and potential of the HDDB, Jha and Hui (2000) noted that the database “serves as a repository of invaluable information and knowledge” (p.2) that could be used by the department providing technical support. To glean this information, the authors leveraged existing data mining methods, including classification, clustering, and prediction, and commonly available data mining tools. Interestingly, the authors noted that existing tools lacked optimization for analyzing the textual data stored in their HDDB, and thus had to leverage newer text mining methods as well. Having established the data source and its value, as well as the methods to be employed, the authors commenced their data mining.

The mining process as undertaken by Jha and Hui (2000) in a corporate environment echoed the methods used in EDM, including data selection, preprocessing, transformation, mining, and results evaluation. The authors also highlighted an additional preliminary step: establishing the goals or intentions of the mining project. Jha and Hui (2000) highlighted numerous business-oriented goals, including marketing and resource management objectives, but the authors particularly addressed customer support. The customer service goals included providing an outstanding customer service experience that considered contextual factors, such as the product for which the customer sought support, the type of problem the user was

experiencing, and the user's geographical location. The customer service HDDDB mining goals did not include highlighting prevailing trends in customer support requests to determine areas for potential customer skill improvement, which might have taken the form of training offering or support documents.

Jha and Hui (2000) concluded that mining their company's customer technical support database bore fruit for numerous operational aspects. The HDDDB mining process yielded valuable information with regards to marketing (including highlighting customers who could be candidates for additional sales), improved customer support based on given features of the client and the issue at hand, and improved resource allocation. Further, the authors also noted that HDDDB mining provided insight into diagnosing problems or issues with hardware. Clearly, the textual data contained within an HDDDB had great potential for informing business decisions once mined.

HDDDB Mining Concluding with Human-Assisted Classification

In 2004, Blaaffladt, Johansen, Eide, and Sandnes identified adverse impacts of technical difficulties experienced by employees, including frustration, dissatisfaction, and costly work stoppages during which employees anxiously awaited support. Further, company-wide issues might be identified based on communication between employees and HD offices, as early support requests might be used to highlight system anomalies or problems. These concerns, among others, prompted the researchers to use data mining to help create an automated technical support email classifying system to speed up the support seeking and giving processes.

Part of the classification scheme included identifying the relative importance (to the business, not to the individual) of any given support request. In determining the importance ranking scheme, the authors grouped messages from users into four categories: automated

messages from user systems, spam messages from user accounts, general questions, and urgent questions. General questions, typified by a *how do I do this?* format, were given a low priority. The authors noted that these questions were often asked repeatedly by the same or different users. Blaaffladt et al. (2004) pointed out that in these cases, support providers could reuse an existing response from an earlier support case.

Additionally, Blaaffladt et al. (2004) provided a valuable historical context for HDDB data mining undertakings intended to improve corporate technical support operations. Past endeavors in the field included efforts to leverage HDDB information to create a decision-based problem-solving tool (Zhao, Leckie, & Rowles, 1996), and studies intended to automate the retrieval of answers to users' questions (Foo, Hui, & Leong, 2001; Wilbur & Sirotkin).

The leveraged clustering of text data to build their automated classification tool was a success. The researchers determined that their system could successfully classify one out of every two messages when the tool ran autonomously. Interestingly, the authors noted that the tool had a much higher success rate when it ran under human supervision. While touting the usefulness of automated classification, and the great potential of such systems in environments where the volume of email was on the rise, the authors might have foreshadowed future trends in educational data mining. Automatic classification systems had "great potential for reducing the workload, but to ensure quality there should be a human in the loop" (p. 10). Perhaps the authors were calling for data mining aimed at distillation for human judgment as an effective method when mining HDDBs for grouping and classification of problems.

HDDB Mining Beginning with Human-Assisted Classification

While Blaaffladt, Johansen, Eide, and Sandnes (2004) noted the value of human involvement in HDDB mining, Forman, Kirshenbaum, and Suermondt (2006) furthered the goal

of minimizing human involvement through analyzing customer technical support phone call data with the intention of quickly classifying calls. The overall business objective was to allocate engineering resources appropriately and to provide optimal problem identification, support, and documentation for common issues. Interestingly, the concept of mining HDDB data to eventually provide documentation represents the closest an HDDB mining project had come to addressing the technical skills of the users who submitted the requests. The authors, however, did not elaborate upon the creation and distribution of this documentation. This lack of elaboration was due in large part to the study's emphasis on the methodology employed in the research.

Forman, Kirshenbaum, and Suermondt (2006) leveraged existing data mining methods, including clustering, classification, and quantification of text data in the development of their approach. These tools were incorporated into the data mining process, although the authors noted that no existing approach perfectly satisfied their needs. Indeed, the authors developed a new text clustering approach that emphasized the creation of clusters based on text data rather than on the assignment of elements to clusters. Their process employed the use of humans in critical decision-making steps early in the process to aid in cluster creation. The authors concluded that standard clustering methods were more effective on organized and well-written text such as newspaper articles and that the ad-hoc nature of HDDB data could require both fine-tuned mining methodology and human involvement. However, even with these conditions, the authors noted that they were successful in leveraging HDDB mining to develop a classification system that ran autonomously once configured, and addressed the desired business goals.

HDDB Mining for Simple Technology Problems

Marom and Zukerman (2009) took HDDB mining for automation purposes further when they developed a system for automating replies for email inquiries to the HD. The authors noted

that email inquiries to helpdesks often focused upon common topics and that HD operators often addressed similar problems repeatedly. Further, many submissions to the HD were often of a low level of technical complexity or difficulty. In a statement of purpose, the authors commented that “organizations and clients would benefit if an automated process was employed to deal with the easier problems, and the efforts of human operators were focused on difficult, atypical problems” (Maron & Zukerman, p. 597). The authors employed a unique analytical approach to mining a very large collection of independent, text-based documents.

Marom and Zukerman (2009) situated their helpdesk automated response research within the existing body of work. Their tactic varied somewhat from traditional methods, as their method relied on a “corpus-based approach,” that is, analysis of a large body of documents (p. 598). Rooted in the history of helpdesk data mining and response automation, Marom and Zukerman’s work was a powerful example of modern HDDB mining practices. In line with the work of their predecessors, the researchers leveraged clustering and prediction techniques as HDDB mining methods. The complex analyses and system design employed these techniques to develop an understanding of the prevailing trends within their helpdesk data. These trends were then used in an automated process to generate responses to support requesters. While the authors deemed the study a success, an important takeaway might be to focus on the early phase of the data mining process in which data was parsed out to create a smaller, more manageable database for mining. Despite the large size of their dataset, the authors concluded that careful data mining was indeed capable of highlighting repetitive, or minimally-technical issues within HD data. As was the trend, Marom and Zukerman (2009) successfully achieved their business goal of more accurate support resource allocation.

HDDB mining and an additional dimension. While considering the business goals of HDDB data mining undertakings, the work of Joshi, Joshi, and Yesha (2011) deserves attention. The authors leveraged HDDB data mining to develop a measurement and tracking system of a complex virtual service delivery system within a university. While the authors succeeded in developing a system that could predict the quality of an electronic service delivery system, the authors also presented a noteworthy contribution to the HDDB mining practice. This study called attention to the cyclicity of certain types of HD support requests. For example, the number of incidents and support requests was observed to increase dramatically at the start of each semester as new students used and encountered problems with technological systems. Similarly, HD support requests increased significantly during and after the launching of new software systems. This information contributed to the development of the tracking platform.

For a study in which the researcher intended to apply data mining methods to an HDDB with the goal of extracting topics for professional development, the contribution of cyclicity from Joshi et al. (2011) could provide valuable insight. For example, a practitioner could apply data mining methods to a subset of an academic institution's HDDB, limited to new faculty support requests during the first three months of each school year, over a period of years. Armed with this information, a human might devise a list of topics upon which new faculty of the current year should receive training. This practice might successfully reduce HD requests by new faculty at the start of the school year while also empowering new faculty. Similarly, as schools often employ specific technologies during exams or as the end of the year approaches, an HDDB data mining undertaking focused on these periods and aimed at determining common problems to be addressed via training might result in a smoother workflow during these times. Thus, HD

database mining with an objective of generating technical and operational improvement suggestions should consider timing and cyclicity.

HDDB Mining for Relationships

Despite the apparent abundance of HDDB data mining studies conducted through 2014, Andrews and Lucente (2014) opined that HDDB analyses aimed towards generating recommendations for HD improvements were still rare. The enhancements desired by the authors referred to augmentations and adjustments to helpdesk operations. These desired improvements were clearly linked - within the very first paragraph - to work stoppages that were attributable to employees' technical problems. The authors posited that data mining of the HDDB at a corporation could improve HD processes which, in turn, might minimize the number and duration of costly business cessations. Referring to the potential held within an HDDB, Andrews and Lucente (2014) wrote: "the dynamic phenomena of a helpdesk environment in an industrial setting result in opportunities to gain practical knowledge from historical information" (p. 94). The authors noted that information about the resolution of employees' problems, stored in the DB along with details of the problems themselves, could be leveraged to enhance HD operations. Mining of the HDDB might potentially highlight relationships between characteristics or features of similar HD requests, and understanding of these relationships might increase the productivity of a helpdesk support operation. Indeed, the authors noted that no tools for analyzing behaviors amongst groups of incidents existed at the time of publication, and the authors sought to leverage data mining to create one.

To facilitate the process of HD operation enhancements, Andrews and Lucente (2014) leveraged two data sources; the HD agents themselves, who provided the researchers with knowledge about helpdesk operations and the larger, quantitative archival information contained

with the HDDB. The authors noted that practitioners had already leveraged this data source for trend analyses which, in turn, might contribute to the making of business decisions. Through the application of a process called Principal Component Analysis (PCA), essentially a sophisticated form of clustering, the authors succeeded in identifying relationships between characteristics or attributes of helpdesk incident reports. The establishment of these relationships led, in turn, to the formulation of recommendations for HD support process improvements.

HDDB Mining and People

Interestingly, Andrews and Lucente (2014) highlighted the importance of human involvement in their data mining undertaking. The authors presented both their methodology and their interpretations of the data to helpdesk managers, who confirmed their validity. Underlining the value of human involvement even further, the authors noted that additional involvement of helpdesk technicians in both the assessment and recommendation phases would likely increase the validity of the results. The researchers thus successfully leveraged data mining of the HDDB in their creation of a model capable of providing operational enhancements to the helpdesk. The model worked best when human involvement complemented the advanced mining methods employed.

While Andrews and Lucente (2014) extolled the value of human involvement in a relationship discovery mining project, Povoda, Arora, Singh, Burget and Dutta (2015) leveraged data mining to examine a core facet of the human side of technical support requests. Povoda et al. leveraged data mining of the HDDB to recognize emotions contained within messages sent to the HD. From these emotions, the researchers hoped to classify, categorize, and prioritize the HD support requests. While the goal of the project was to automate helpdesk services to distinguish between high and low priority requests based on the emotions encountered in request messages,

the underlying objective of this use of Hddb data mining remained consistent with the studies in this field. The undertakings of Hddb mining to date were intended to improve helpdesk services. Povoda et al. confirmed this in their review of related work when they noted that the greatest challenge in technology-based helpdesk evaluation was “to increase customer satisfaction and lower the costs” (p. 311). Thus, the authors intended to contribute to this field by creating a system that could expedite operations by analyzing emotions in stored text-based messages between users and support agents.

The primary challenges encountered by Povoda et al. (2015) pertained to the structure and content of the data they sought to mine. Emotion analysis, the researchers noted, depended largely on the vernacular used by those involved in the study, and analyses grew increasingly difficult as the size of the database increased. By using a small dataset, however, and leveraging text-based classification and combination methods, the authors succeeded in developing a system that could identify a range of emotions with an accuracy exceeding 70%. The authors thus concluded that it was indeed possible to create a system intended to enhance HD operations by using data mining to discover emotions and to establish a priority sequence from those emotions. This unique application of data mining methods highlighted one of several new directions in which researchers were advancing Hddb mining practices.

Hddb Mining for Prediction

While Hddb mining studies might invariably be intended to improve HD operations in general, some studies through 2016 addressed leveraging data mining for improving specific processes. Andrews, Beaver, and Lucente (2016) used data mining of an employee-serving technical helpdesk to predict problems and the effort required to address those problems. The researchers began their report by affixing a dollar value to the resolution of each helpdesk

resolution and made a pragmatic case for the advantages of improving helpdesk operations. Noting numerous factors, including the cyclical nature of increased and decreased support requests, the authors posited that a prediction method would be beneficial for staffing purposes at the helpdesks of North American businesses. Proper staffing would contribute to quicker problem resolution while also keeping costs to a minimum.

As in Andrews and Lucente's (2014) previous study, Andrews et al. (2016) leveraged Principal Components Analysis (PCA) in their prediction-oriented data mining undertaking. They described the PCA method as being an advanced form of clustering. For this study, PCA was used to select an item from a cluster that best represented the entire cluster. In the context of their study, clusters contained products on which employees had contacted the helpdesk for support. Through advanced statistical analysis, the authors succeeded in identifying which products were most likely to require support, and when. The success of this endeavor, in which data mining of the HDDB played a crucial role, led the researchers to interesting conclusions.

While Andrews et al. (2016) touted the capability of their system to predict helpdesk related costs to within acceptable margins of error, the authors highlighted a noteworthy trend that they discovered in the data. They noted that "20% of the products generated 80% of the incidents" (p. 447). This remarkable statistic was presented as an interesting discovery while highlighting the statistical importance of an adequately sized product portfolio. It was worth noting that the authors do not mention training employees on the most common problems encountered when using these commodities. Given the relationship between a small number of products and a large number of HD support requests, it would seem as though training was a viable solution if the goal were a reduction in support instances. However, the authors did not seek a reduction in support cases attributable to behavior changes of those submitting the

requests. Rather, they succeeded in developing an HDDB-informed tool that could predict costs associated with HD ticket submissions by product and timing.

HDDB Mining Summary

There was clearly an abundance of historical and contemporary HDDB data mining research. These studies, to the one, were concerned with improving the operation of, or costs associated with helpdesk support services. Many of these undertakings employed HDDB mining as a means to an end; the process of mining HDDBs was often used as a method of analysis within a larger study and towards an overarching goal. While numerous studies leveraged HDDB mining as a tool in the pursuit of a goal, the purpose of determining technical topics for professional development was conspicuously absent from HDDB mining undertakings.

This case study attempted to address this absence. The fundamental logic behind the project was straightforward. A helpdesk database was mined to determine the technical errors or problems most commonly encountered by faculty employees of a medium-sized independent boarding high school. By mining the HDDB for topic detection, professional development upon these issues might be devised and provided for teachers. If the professional development was successful, there might be a reduction in helpdesk requests, which will benefit both teachers and helpdesk operations. Teachers might feel empowered and experience fewer work stoppages, and helpdesk agents might process fewer requests or have more time to focus on complex problems. The potential benefits of HDDB mining for professional development were tangible and easily discernable.

A New Framework

A researcher could devise a potentially momentous educational technology integration and assessment framework out of the literature. With an understanding that the context of an

integration effort, such as the classroom, course, school, or district, will have unique impacts on an integration project, a framework should consider the unique idiosyncrasies of the project setting. This unique context should include the specific technology integration goals (including the definition of integration in that environment), as well as the barriers specific to that setting or context. The proposed framework would leverage the use of data mining methods to measure technology competency barriers to integration and potentially identify areas of opportunity for professional development. Where this framework might differ from existing frameworks, models, paths, or procedures was in its usage of a potentially untapped and unbiased data source.

Existing instruments, even several used in combination, might not accurately represent every facet of a unique educational technology integration project and its corresponding barriers. This lack of reliability places a potential limitation on the value of these tools as weakness or topic identifiers. Perhaps information mined from helpdesk requests might be more indicative of users' technical skills (or lack thereof) than measurement instruments. The new framework methodology proposes incorporating a text data mining analysis of Information Technology helpdesk support requests and solutions, stored in a minable database, to inform and complement professional development topic selection. Many institutions encourage or require faculty members to submit helpdesk support request tickets when seeking technology aid for a particular tool, context, content, or scenario. Over time, the database that contains this data could become an expansive resource of technology integration information that portrays an organic and accurate display of the range of technology competencies of the body of individuals submitting help requests.

Institutions might use text data mining for topic discovery within the helpdesk ticket database as a data source for identifying barriers, obstacles, and planning approaches to

encourage educational technology integration. This data was collected gradually over time, directly from the teachers who were the potential beneficiaries of targeted and situated professional development. The helpdesk database accurately describes genuine problems and challenges faced by teachers in the precise context where educational technology integration was hoped to occur. Analysis of helpdesk ticket data might identify which users needed what types of content-based professional development and might also highlight integration barriers or trends across academic departments or demographics. Similarly, a lack of tickets submitted by certain users might highlight those users as potential mentor candidates. In brief, the database of submitted helpdesk tickets could be an untapped resource capable of accurately informing and shaping educational technology integration efforts. This database must be mined and analyzed carefully with an eye on discerning professional development topics that might enhance the capabilities of the people submitting helpdesk requests.

At the time of the case study's undertaking, there was a gap in the body of research: few scholars were analyzing helpdesk databases with the goal of improving the skill set of those requesting assistance from the helpdesk. A framework that leveraged existing research about definitions of educational technology integration and common barriers but employed the mining of the unique and highly customized data stored in helpdesk tickets had the potential to contribute to the body of work regarding educational technology integration. Further study and analyses were required to determine both the legitimacy of using helpdesk tickets and the mining procedures by which this information could be understood and then incorporated into educational technology integration projects. Though the value and usefulness of helpdesk ticket data with regards to technology integration and training remain to be studied, it appears that this

understudied data source had the potential to change the way integration projects were examined, measured, and undertaken.

CHAPTER 3

METHODOLOGY

In both educational and corporate realms, data mining (DM) practitioners analyze large databases such as HDDBs for in-depth analysis. In education, this practice is known as educational data mining (EDM), learning analytics (LA), or learning analytics and knowledge (LAK). There are many goals of EDM practices and many applicable methods. A common EDM objective is topic detection, and two established fields of practice are text data mining and distilling of data for human judgment (Baker, 2010). By applying text data mining and distillation methods (a practice typically performed by an expert using a powerful computer and commercial data-mining software) to the information contained with an HDDB, topics for professional development might be determined. The methods leveraged in this case study will be used to address the study purpose and answer the research questions.

Study Purpose and Research Questions

The purpose of this case study was to apply data mining practices to an underused data source to identify and examine areas of improvement for the technology skills of faculty members at an independent boarding high school in the United States in the recent past. It was the intention of the study to examine the following research questions:

1. In what ways could data mining be leveraged to best extract the desired information from the HDDB?
2. What does data mining of the HDDB reveal about gaps in teachers' technology skill sets?
3. How could data mining of the HDDB be used to determine topics and plan technology-related professional development for teachers?

General Research Methodology

The researcher undertook this investigation as a case study, with a goal of determining the feasibility and usefulness of applying data mining techniques to the HDDB to identify areas of technological weakness within a faculty body. At the time the study was executed, many similarly designed case studies existed within the fields of educational data mining and helpdesk data mining. Indeed, case studies were a prevalent form of qualitative research in the realm of educational data mining, as the area of study continued to advance towards a mature domain. Case studies were essential for the ongoing development of the field since the studies presented to researchers the manners in which their peers were applying the principles and concepts of educational data mining. This study employed data mining techniques to a new and underused data source in the same vein as other relevant case studies (Merceron & Yacef, 2005; Romero, Ventura, & García, 2008; Abdous & He, 2009; and He, 2013). These undertakings served to introduce data mining “to all users interested in this new research area,” and this study hoped to provide a similar service (Romero, Ventura, & García, p. 368). A commonality amongst these published case studies and the case study at hand was that the practitioners used archived data that standard business practices had generated. This case study, like many others, leveraged data mining practices applied to archival data to test the potency and capability of those practices.

This case study focused on information technology help requests submitted by teachers throughout a three-year period at an independent boarding high school. The boundaries of the case study were clear: analysis focused on support requests by faculty members collected by IT helpdesk agents between September 2013 and September 2016. The study excluded help requests by teachers before and after this period or requests by non-teachers during this period.

This case study was thus a “within-site” study as the entirety of the study was confined to a single site (Creswell, 2013, p. 97).

While many case studies typically rely on multiple sources of information (Creswell, 2013), this study was atypical in that it was a case study specifically focused on extracting value from and determining the value of one archival data source. The intent of the study was to understand the potential value contained within the HDDB, and thus the HDDB employed at this site was central to the study. This case study was an “instrumental” case study as it examined the information contained within technology help requests at a specific site to advance general understanding of the process of mining HDDBs (Creswell, p. 98). The execution and results of this study were not intended to present a generalizable conclusion, but, rather, to serve as a valid and reliable example of HDDB mining designed to understand and aid those submitting the help requests. This single-site instrumental case study audited the benefits of mining archived data within an HDDB with a goal of developing PD topics for improving the technology skills of the people whose help requests had populated the HDDB. The given site, participants, and period provided reasonable confinements or boundaries for the case so that the case study best illustrated the ideas under review.

This chapter provides an in-depth review of the characteristics of the case study, including the study’s setting and participants. This review also includes the sampling method used to select participants and the data that was collected and analyzed. Additionally, the case study’s participant rights, informed consent practices, and other ethical issues are identified and reviewed in detail. Finally, this section addresses potential limitations of the study’s methodology.

The Setting of the Case Study

The researcher undertook the study at an independent boarding high school located in the northeastern United States. The faculty members of this site have adopted many new technologies during the three-year period upon which this case study focused. During this time, the Director of Educational Technology, also the scholar-practitioner who undertook this study, was charged with overhauling the educational technologies in use at the school. The most prominent new technologies implemented during the three-year timeframe included Google Apps for Education (an expansive web-based communication, creation, and collaboration suite), MacBook Air computers (administered to the faculty by the IT department), and modern classroom audio/visual systems (including high definition projectors and AppleTV devices for wireless connections to the projectors). The faculty body, which exceeded 110 active members during any school year, and which numbered over 130 individuals including retirees and departed faculty members over the three-year period, were required to learn and use many new technologies. The faculty body comprised of a wide range of teachers, of different ages, and from diverse cultures and backgrounds. All faculty members had bachelor's degrees, and a majority of teachers had advanced degrees. Over a period of three years, the Director of Educational Technology and the site Dean of Faculty had anecdotally observed that the level of technological skill and usage varied widely by teacher ("C.L.", Dean of Faculty, personal communication, August 9, 2016).

While the school's administration had invested significant financial and time resources into preparing teachers to use and leverage the new technologies throughout the period that binds the case study, inevitably there were hundreds of requests for technological help submitted to the Information Technology helpdesk. The case study attempted to glean areas of technical

weakness of the faculty body by mining the archive of HD support requests submitted by teachers during the period from September 2013 to September 2016. Requests were placed by faculty members to IT support agents regularly as part of standard business operations, and the agents inputted these requests as data into the HDDB. (It was an assumption of this study that teachers indeed used the services of the IT helpdesk as expected). Over time, the HDDB grew into a unique, organic archive that the researcher believed might include valuable information about teachers' technology usage and skills. The researcher anticipated that themes extracted from the HDDB mining process could be used to provide focused PD to address the identified weaknesses.

When the study took place, the HDDB at the case study site was not leveraged or mined in any substantive way. The helpdesk platform periodically issued reminders to technicians to complete or close support requests, and the Director of Information Technology had the capability to review the durations of cases. The information contained within the HDDB, however, was not used to understand and improve areas of weakness of those submitting the support requests. This trend mirrors similar trends observed in the scholarly research on helpdesk data mining, which focused almost exclusively on improving IT operations (Jha & Hui, 2000; Blaaffladt, Johansen, Eide, & Sandnes, 2004; Forman, Kirshenbaum, & Suermondt, 2006; Marom & Zukerman, 2009; Joshi, Joshi, & Yesha, 2011; Andrews & Lucente, 2014). As a result, this case study attempted to inform decision makers at the site and inspire researcher-practitioners at other sites to determine the potential for HDDB mining at their locations.

As the Director of Educational Technology at the case study site, the researcher had access to the HDDB required for mining. TrackIt, the software that powered the on-site helpdesk, stored the data in an SQL-based database. To acquire the data for mining, a database

manager from the site would extract the SQL-based data and provide it to the researcher as a Microsoft Excel spreadsheet file. This Excel file would then have all columns with participant-identifying information deleted directly within Excel or in an enterprise-grade data mining tool such as RapidMiner Studio. However, straightforward access to the database did not necessarily or immediately grant the Director the right to use the data in a research study. Before acquiring the data, the researcher undertook a lengthy and detailed ethical review to ensure the privacy, confidentiality, and safety of the school and all participants. The Ethical Concerns and Participant Rights section of this chapter details these efforts.

Sampling and Participants

The pool of potential participants included all teachers employed at the site at the time when the case study was conducted, as well as teachers employed at the school during the period binding the case, September 2013 through September 2016. While any currently or recently employed teacher might have been a participant, it was possible that a potential participant might not have submitted an IT support request and thus were excluded from the study. As the goal of the case study was to use DM techniques to identify areas of technological skill set weakness amongst a faculty, the overall criteria for participation were straightforward. Participants must have been employed as a teacher during the study timeframe and must have submitted at least one technology help request during that time.

Owing to the archival nature of the data used in the case study, the undertaking did not require participant sampling methods. However, the study mimicked a sampling process during the data preparation phase. A form of purposeful sampling, similar to maximum variation sampling, was applied to determine potential participants within the study. After receiving approval to study the school-owned data, the HDDB was pared down to exclude tickets

submitted by students, administrators, and all non-faculty members. The researcher did not remove any faculty member support requests submitted within the study timeframe. In essence, this was a manifestation of maximum variation participant selection, as the study attempted to select the maximal number of participants, with the goal of representing the widest possible array of characteristics (Bloomberg & Volpe, 2012). This approach ensured that the helpdesk tickets selected were varied and diverse. Data mining of a subset of the HDDB based on this sampling method was anticipated to provide a comprehensive list of technological weaknesses.

It is important to note that the participants identified by maximum variation sampling might have been very similar to a participant list generated by convenience sampling. The latter sampling method would have relied on the availability and accessibility of partakers as a primary characteristic for inclusion, with less regard to the value of the information their participation might have provided (Bloomberg & Volpe, 2012). As this case study aimed to determine the value of data mining of the HDDB dataset across a faculty body, the largest body of participants was desired. Thus, in this case, maximum variation and convenience sampling would have yielded nearly identical participant groups. As helpdesk agents had already collected the data during standard business practices, the study did not seek additional participants, nor did the study require or include any interviews. Thus, the term *participants* was not wholly accurate as a descriptor of the individuals who provided data that helped build the database. As these individuals were more data contributors than participants, the term *data contributor* was used to refer to a person who had submitted IT help requests and whose requests were documented in the database and used in the study.

Since employees had generated the to-be-mined archival data during normal business practices by employees, the institution owned the data (“C.L.”, Dean of Faculty, personal

communication, October 3, 2016). Further, the database's non-private, non-confidential, archival nature allowed for use, study, and review by authorized employees of the school. The researcher sought and received consent from appropriate school administrators to use an anonymized version of the archived data containing information collected over time by IT support agents from data contributors.

Overview of Data Collection, Preprocessing, and Analysis

The data collection phase of this case study was unique compared to many qualitative case studies in that the archived data was available to the researcher and did not require collection. The data collection method was not unique, however, when compared to other data mining case studies. Indeed, a researcher conducting an HDDB data mining undertaking need not focus on data collection, as the purpose of the project was to review data that employees had collected *in situ*. Many system-generated datasets were leveraged in previous research case studies, including web-based tutoring data by Merceron and Yacef (2005), and helpdesk data by both Forman, Kirshenbaum, and Suermondt (2006) and Andrews, Beaver, and Lucente (2016). Learning management system data was used by Romero, Ventura, and Garcia (2008), and text chat interactions by both Abdous and He (2009) and He (2013). The study at hand leveraged data collected by the IT department and stored in the HDDB. Data collection was an ongoing process, as IT department employees recorded information about new helpdesk support requests on a regular basis.

Regarding the collection and preparation of data within data mining case studies, educational data mining researchers typically emphasized the data preprocessing steps as opposed to the collection methods since the data under analysis was previously collected and archived. The data analysis phase of this case study represented more than the means by which

the researcher might base conclusions or attempt to answer questions in that the nature of this case study was such that the data analysis processes were not only a means to an end but were the goal of the research as well. The study's data analysis phase employed the application of data mining procedures to the HDDB. The text mining and topic detection techniques leveraged in the study were clustering (wherein helpdesk records were automatically grouped by theme), and classification (wherein helpdesk records were assigned a label based on a customized algorithm). Mined data was distilled for human judgment and interpretation.

A professional data miner was employed to conduct the data mining using RapidMiner Studio version 7.3. At the time the study was conducted, this software system was an advanced data mining tool that automated the mining process. Many of the software's key algorithms and functions represented significant improvements to and automation of complicated and time-consuming mining tasks. Even with dedicated data mining software, the data mining steps were technologically complex. Many professional data miners have completed rigorous training, have years of experience, and have received certifications from data mining software providers. For this case study, the researcher leveraged the skills of Brian Tvenstrup, a RapidMiner-certified, professional data miner and president of the data mining service provider Lindon Ventures.

The anonymous dataset, once prepared for mining, was conveyed to the service provider via encrypted file transfer. In close collaboration with the researcher, the data miner performed the advanced data mining procedures in RapidMiner. In principle, the hiring of a third party for data mining procedures was akin to soliciting the skills of a transcription or coding service in a typical qualitative study. However, in this case study, the third party involved in data mining would never have access to subjects' private, confidential, or identifying information.

The case study required the miner to execute numerous data mining procedures. The researcher and data miner designed these analyses, discussing them in detail via phone calls and emails. The miner performed the analyses and returned the exported results in the form of Microsoft Excel and RapidMiner process files. The first data mining procedures created a wordlist, which contained all the unique words or tokens that appeared in the processed dataset. This file underwent extensive analysis in Microsoft Excel. The second data mining analysis used the clustering technique to highlight and identify clusters of similarly themed records within the HDDB. This data clustering process was performed iteratively and refined until the groupings highlighted important information about the subject and content of each cluster. As with the wordlist, Microsoft Excel was used for further analysis of this file. The third data mining procedure used a classification algorithm to assign a label to all helpdesk records, based on a manually-labeled subset of records. Once a subset of the technical support help request dataset was labeled, the researcher and the data miner undertook the classification procedure. This complex process generated a labeling model that was intended to learn from the manually-labeled cases and then automatically determine a label for all records in the dataset. The model was used to apply labels to the complete dataset, and the results of this labeling process were exported and analyzed. Once data mining procedures were completed, the wordlist, clustering, and classification results were exported from RapidMiner and analyzed in Microsoft Excel.

Detailed Methodology

Data mining procedures were applied to the HDDB data using the specialized, commercially available data mining software RapidMiner Studio version 7.3. A professional, certified data miner, under the oversight of the principal researcher, performed the data mining steps in RapidMiner. As the goal of this undertaking was not to pioneer data mining processes,

but, rather, to apply existing data mining process to the HDDB, the processes and methods used in RapidMiner were characteristic or standard text data mining techniques. All work in RapidMiner was comprised of creating a workflow, from data input to output, using procedures from the software's extensive, built-in library of processes. Many RapidMiner procedures have settings that a user can adjust before the execution of the workflow. The researcher sought the data miner's input regarding which standard methods and settings should be applied at each point in the mining procedure. The mining actions led to the creation of three outputted data files, by which the researcher could analyze, summarize, and generate conclusions about the value of data mining practices applied to an HDDB. The three outputs included a list of the words that appeared most often in the data sample, the result of an automated data-grouping process called clustering, and the results of a human-assisted data-grouping process called classification. The methods that led to the creation of these three RapidMiner outputs will be detailed after an examination of the processes of preparing the data for inputting into RapidMiner and structuring the data within the powerful data mining software.

Data Preprocessing Prior to Input into RapidMiner

The RapidMiner workflow was configured to accept text data as an input, arrange the data for analysis, and then run processes upon the data. This section focuses upon the preprocessing of the HDDB data, which included data preparation and cleansing in Microsoft Excel before and after importing the data into RapidMiner, and outlines the analyses performed on the dataset in RapidMiner. As the data preprocessing took several days to complete, intermediary data files were stored securely on a password-protected computer and backups were stored in an encrypted and password-protected cloud storage service.

Data File Preparation. An information technology agent at the research site exported the raw, unedited HDDB data from TrackIT, the IT support database system, to a Microsoft Excel spreadsheet file. The raw data file contained 35,623 records, encompassing a history of IT help requests from May 2005 through December 2016. Each record or row on the spreadsheet represented a single help request or support ticket. The raw data file contained 47 columns, each containing particular information about each support request. The columns included information such as the date of the help request, the name of the support requester, the support requester's department at the research site, and a detailed description of the problem requiring support.

The researcher deleted all records dated before September 1, 2013, and after September 30, 2016, from the dataset. The remaining dataset included 8,299 rows. Next, the researcher removed all records ascribed to a member of a non-academic department. These departments included the Admissions and Advancement offices, the Plant and Property and Power Plant departments, and Dining Services, for example. Further, the researcher deleted all student-submitted help requests as well as any blank or incomplete submissions. The researcher kept all submissions from members of academic departments, including the Arts, English, History and Social Sciences, Math, Science, and World Languages departments. At this point, 1,884 records remained. This was the final tally of HDDB records to be studied in the mining processes.

Having reduced the number of records to include only submissions from teachers, the intended groups of data contributors, the researcher next focused on eliminating columns from the dataset. The researcher removed columns for a variety of reasons. The researcher deleted any columns that contained possible identifiers. These included, for example, the name and contact information of the support requester, the name and contact information of the IT agent(s) who created and worked on the support request, the submitter's department, the location of the

incident, etc. The researcher removed other columns simply because they did not contribute to the study at hand. Many of these columns contained missing or irrelevant information. These columns included the date when an IT agent closed the ticket, the amount time an agent spent working on the case, a list of attachments included with the case (if the case was submitted via email), the support agent's qualification level, and a work order status for tracking the progress of the case.

Each record in the dataset contained a unique work order number (an integer), stored in a work order number column. The researcher deemed that this column was necessary for inclusion in the dataset. It is common practice in database analysis to ensure that all database records contain a unique identifier to avoid duplication or accidental data merging, but also for purposeful merging when combining multiple datasets (B. Tvenstrup, personal communication, December 23, 2016). However, the work order number could theoretically be used to identify the ticket and thus potentially reveal private information about the data contributor. Each work order number was increased by the same randomly generated 4-digit integer. Thus, each record in the database included a unique identifier that could not be traced back to the raw dataset.

After removal of columns, the dataset now contained 1,884 records and nine columns. These columns, or database fields, contained the unique (adjusted) work order number, support request date, and task (a support ticket title). The records also contained a request type (a very general ticket category), the priority assigned to the ticket, and a detailed description of the problem, including communication between the support requester and the IT agent(s). The description field contained the largest volume of text in each record and was deemed especially important for the impending data mining. The remaining columns were IT agent notes and two category fields for the type of work required to complete the project. These nine columns were

expected to be useful and relevant once the data miner began the mining processes. Before the data mining began, however, the dataset needed additional preprocessing, including data cleansing for rendering the data anonymous.

Data File Cleansing. Upon completion of the file preparation, the researcher set upon the task of rendering the data anonymous. This process was lengthy and required strict attention to detail, as maintaining the anonymity of data contributors and IT agents was of paramount importance. The researcher created Visual Basic programs (integrated into Excel) to automate the anonymising processes and also used Excel formulas and functions to support the procedure. All steps detailed in the ensuing steps were performed for any instances of potentially identifying information, regardless of case (upper or lower or a combination of both). For example, any code or procedure intended to remove the name “Fred” would also remove “fred,” “FRED,” and “fRed.” Further, the researcher created these programs to search and modify every spreadsheet record and column.

The researcher obtained a list of all IT support agents employed at the site during the period under review. A program was created and executed to replace all instances of IT agent first and last names with “ITAGENT.” A similar program was written to remove any and all instances of the site name, including abbreviations or common names. The program replaced these terms with the placeholder text “SITENAME.” Additionally, a program was executed to remove all instances of mailing addresses, and phone numbers were randomized so as to render them useless.

The researcher undertook two additional data cleansing steps, intended to format the data file for easier analysis in RM. The researcher replaced all instances of the text “Eastern Standard Time” throughout the dataset with “EST.” This small change made the dataset significantly

smaller in file size, and also reduced the overall word count. Similarly, the researcher replaced all line breaks (carriage returns) in the description column with a pipe character: “|.” This replacement significantly shortened the length of each support request description. Further, the pipe character is a commonly used text delimiter and, if necessary, could be easily replaced with another character (or even a line break) if the need arose.

Finally, the researcher obtained an extensive list of first and last names and executed a Visual Basic program to remove any and all instances of each name on the list anywhere it might appear in the dataset. The program replaced first and last names with “NAME.” For example, the name of a data contributor named “Mary Jones” was replaced by the alias “NAME NAME” in a record.

Upon completion of the anonymizing of the data, the researcher undertook a manual process of checking records to confirm anonymity. The researcher manually reviewed the content of all nine columns for 190 randomly selected records (just over 10% of the dataset) to confirm that no identifying information was present. Once the researcher determined that the program had successfully purged all identifying and private information from the data file, the researcher copied the entire dataset and pasted it into a new Excel file. This procedure ensured that no steps could be undone, nor could any versions be reverted to earlier instances. Except for the final version of the file, all versions of the data file (including those backed up) were securely and permanently deleted. The data file, transferred securely from the researcher to the data miner via encrypted email, was now ready for import into RapidMiner, where the data miner performed additional preprocessing steps.

Data Preparation in RapidMiner

While RapidMiner workflows are capable of performing advanced data analysis in an incredibly short operating time, any inputted data must be correctly configured for use in RapidMiner before a user can conduct such analyses. The data miner converted the prepared Excel file into a comma separated values (.csv) file, a standard and universal format for text-based data, and the file was imported into RapidMiner and stored as a RapidMiner data repository. The processes that would eventually be used to mine the HDDB data required a particular type and configuration of data input for proper execution. Thus, RapidMiner would need to reformat the data from rows and columns containing paragraphs of text to a specially formatted wordlist. The steps for converting text-based data to a wordlist are consistent across the data mining industry.

Creating the Wordlist in RapidMiner. To maximize the potentially useful words in the wordlist, the researcher and the professional data miner decided to combine the text of the task, description, type, and notes fields for each record on the data file. While the description field contained by far the most text within each record, the researcher believed that adding the text from the other fields might aid in text mining without introducing any potential problems. Once the data miner combined the four database fields into one in RapidMiner, the data miner initialized the processes in RapidMiner to create the wordlist.

The first and most simple process in creating the wordlist was case transformation. This step ensured that every word in the list was stored in lowercase. By default, data mining platforms would interpret identical words with different capitalization as separate entities, while a human would understand that the words were the same. Thus, case transformation ensured case consistency for all words on the list.

After RapidMiner completed the case conversion process, the next step was the tokenization of the words on the list. Tokenization was a typical but critical part of the preprocessing workflow. The tokenization process divided the large body of text (in this case, the combined text of four columns per record, for every record) into individual words. This was a principal step in the creation of the token list or wordlist.

On the advice of the data miner, the researcher agreed to remove tokens of 3 letters in length or shorter from the wordlist. This removal was, once again, considered standard practice, as these short words appeared frequently but provided little meaning or relevance from a data-mining topic discovery perspective. Similarly, the list was filtered to remove basic English stopwords, or common words that do not carry much meaning. RapidMiner contained a built-in dictionary of these terms, which included words such as *that*, *which*, *because*, *some*, *from*, and *hers*. Tokenization, removal of short tokens, and stopword filtering yielded a list of approximately 8,270 unique words in the wordlist.

At this point in the process, the data miner exported the existing wordlist to a human-readable comma-separated values file and conveyed it to the researcher via encrypted email. The researcher examined the file using Excel. The file contained a lengthy list of words and the count of appearances of each word throughout the dataset. The researcher sorted the list by the count, descending, which placed the most commonly appearing words at the top of the list. From this list, the researcher created a custom list of stopwords, similar, in principle, to the built-in list of stopwords in RapidMiner. The custom stopword list included words that frequently appeared in the dataset but which the researcher did not believe would contribute to the analysis or the data. The RapidMiner text mining processes might incorrectly evaluate a commonly appearing but relatively meaningless term. The custom stopwords list consisted of 68 words that the data miner

removed from the wordlist. The custom stopwords list included, for example, *SITENAME*, *NAME*, and *ITAGENT* (the placeholder terms applied during the anonymizing process). Additional custom stopwords included *thank*, *thanks*, *help*, *please*, *hello*, *sincerely*, and *greetings*, as well as the names and abbreviations for the days of the week. The process of filtering out and removing custom stopwords removed extremely frequent but hollow words that appeared at the top of the wordlist sorted by descending frequency of occurrence.

Pursuant to tokenization and stopword filtering, the data miner applied a stemming process to the wordlist in RapidMiner. Stemming was yet another commonly used filtering technique for text mining, in which related words that share a common root or stem were combined, with the intention of reducing the count of words in the wordlist. As with stopwords, RapidMiner contained a built-in dictionary of common stem words. The stemming process may have combined, for example, *company* and *companies* into the stem word *compan*, for example. As an additional example, the RapidMiner stemming process may have combined the words *come* and *coming* into *com*. (The word *com*, as in *amazon.com*, was removed during earlier filtering and thus *com* was a viable and satisfactory element of the wordlist). Stemming reduced the instances of individual words in the wordlist by removing multiple terms in favor of their shared stem. If the miner activated stemming, the system would search for and replace stemmed words with the stem.

The researcher determined several potential risks in using stemming. The primary risk was that certain words might have a similar stem and thus be akin linguistically, but have different meanings. In an HDDB, this problem could arise if, for example, the words *space*, *spacing*, and *spaced* were stemmed together into *spac*. The researcher considered the following example during decision-making regarding the use of stemming: a user might have written “I

spaced out and deleted my folder” in a help request, while another user might have written, “My space bar is broken.” A third user may have asked, “How do I set Google Docs to double space?” In these theoretical examples, the stem words are similar but have different meanings. During discussions regarding the use of stemming, the professional data miner noted that data mining was not an exact science and that robust data mining intended to deliver the best possible results rather than perfect results (B. Tvenstrup, personal communication, December 28, 2016). This small but important distinction, made clear during the data preprocessing stage, would resonate throughout the data mining phase of the project. The researcher decided to apply stemming techniques to the dataset, as the practice is common in text data mining, and potential gains were believed to outweigh the possible risks of falsely combining terms. Stemming the existing wordlist reduced the total count of words to 6,124.

Upon completion of the stemming process in RapidMiner, the researcher and the data miner discussed the use and creation of n -grams. Unlike stemming, which is intended to reduce the number of words or tokens in the wordlist, n -grams can (dramatically) increase the count of words. n -grams were computer-generated tokens of n length, where n consecutive tokens were combined to create a new token. For example, *google* and *drive* might be joined to generate the token *google_drive*, which in an n -gram of n value 2. The creation by RapidMiner of n -grams of length 2 would add many new tokens to the list, and n -grams of length 3 would add many more. The researcher and the data miner determined that n -grams of length 3 did not contribute enough unique and often-occurring tokens to warrant their creation and the corresponding inflation in total word count. Indeed, while the researcher authorized the creation of n -grams of length 2, which increased the size of the wordlist, the final data preprocessing step reduced the total number of tokens, including n -gram = 2 tokens.

The last data preprocessing step in RapidMiner was frequency pruning. This process, once again a typical process in a text data mining preprocessing workflow, calculated a frequency of occurrence for every word or token and then deleted all tokens beneath a user-determined threshold. Similar in principle to custom stopwords filtering, which deleted the few, common but meaningless words at the top of a wordlist sorted by descending frequency, pruning eliminated a large number of rarely-occurring words at the bottom of the sorted wordlist. The researcher determined that the most useful occurrence metric was the count of records, or helpdesk requests, which contained a given word. Recalling that the total number of records was 1,884, the researcher noted that a term that appeared in, for example, 1,600 records would have a very high frequency ($1,600 / 1,884$ or 84.93% of all tickets) and thus be an important word to include in the wordlist. A word that appeared in only 100 tickets (out of 1,884, or 5.31%) would be comparatively less important to include in the wordlist. The goal of pruning was to reduce the number of tokens dramatically, and thus the researcher was required to determine the threshold beneath which words were eliminated.

With the addition of the n -grams of length 2, the wordlist had grown to a large list of words or tokens with many occurring infrequently. Many of these tokens would not contribute to the forthcoming RapidMiner text data mining, but would slow the processes down, and, worse increase the likelihood of inaccurate results. The researcher and the data miner considered potential thresholds such as removing all tokens that only appeared in a certain number of cases, or that collectively represented the bottom 10% of all tokens. After an extensive review of the wordlist, the researcher and the data miner concluded that only tokens that appeared in 38 or more helpdesk requests were kept on the wordlist. The researcher noted that 38 out of 1,884 records equaled approximately 2%. Thus, for a term to stay on the wordlist, it must have

appeared in at least 2% (or 38) of all the helpdesk request submissions. With the pruning threshold set, the RapidMiner process removed words from the wordlist that did not meet this threshold.

The final count of tokens on the wordlist was 469, which was deemed by the experienced data miner as a satisfactory number of tokens for the intended data mining processes applied to a text dataset of 1,884 rows and nine columns (B. Tvenstrup, personal communication, January 2, 2017). The wordlist, which included the occurrence frequency, was the first of three data mining deliverables from RapidMiner that the researcher used for determining results and drawing conclusions about the value of data mining processes applied to an HDDB. The wordlist was expected to be a useful resource for the researcher in the extraction and summarization of potential topics of interest from the dataset. However, prior to leveraging the wordlist for results generation, the list was exported from RapidMiner and analyzed further in Microsoft Excel.

Wordlist Analysis Methodology in Microsoft Excel. The creation of the wordlist was a by-product of the data mining preprocessing steps. While the wordlist was a modular element of a larger workflow, the wordlist allowed for analysis that eventually provided insight towards the case study's research questions and goals. Further, a thorough review of the wordlist was essential for illuminating the results of the clustering and classification analyses, as these procedures leveraged data contained within the wordlist. The wordlist, once exported from RapidMiner, was analyzed in the spreadsheet software, Microsoft Excel.

The wordlist was a simple document, consisting of three columns and 469 rows (one row per unique word in the dataset after preprocessing). The first column listed each word, or token, while the second column contained the total number of unique help requests that contained that token. Thus, no value in the second column could exceed 1,884, the total number of unique

support requests. The third column listed the total number of appearances of each token throughout the entire dataset. For example, if a support request contained the text “I thought my password was abc123, but that password isn’t working, could you please reset my password?” the total number of instances of the token *password* in that record would be three. In this manner, the value in the third column of the wordlist could exceed 1,884. An additional column was added to the wordlist, which calculated the percentage of unique HD requests that contained an individual token. This figure was calculated by dividing the total number of individual help requests that included a particular token by the total number of tokens, 1,884. An example of twenty tokens, the ten most frequently occurring and the ten least frequency occurring, ranked by percentage of appearances in unique help requests, is shown in Table 1.

Table 1

The Most and Least Frequently Occurring Tokens by Count of Unique HD Request Appearances

Token	Unique HD Request Appearances	Total Instances	% of In Unique HD Requests
us	676	1,225	35.88%
comput	536	1,226	28.45%
work	528	1,055	28.03%
set	477	826	25.32%
subject	469	836	24.89%
googl	414	1,363	21.97%
issu	414	668	21.97%
school	383	595	20.33%
know	375	620	19.90%
receiv	365	482	19.37%
		--	
account_creat	38	48	2.02%
campu	38	51	2.02%
displai	38	58	2.02%
exampl	38	39	2.02%
hear	38	43	2.02%
minut	38	44	2.02%
rememb	38	43	2.02%
sent_firstclass	38	50	2.02%
termin	38	60	2.02%
usernam	38	62	2.02%

The sample of the wordlist provided in Table 1 highlighted numerous elements of the data mining preprocessing steps that created it. The results of stemming were evident, as tokens such as *us* and *issu* represent stemmed versions of *use*, *used*, *user*, *uses*, and *issue*, *issued*, and *issues* respectively. Similarly, the tokens *account_creat* and *sent_firstclass* highlighted the creation of *n*-grams or new tokens created by combining individual tokens. Finally, the results of pruning, which eliminated all words that appeared in fewer than 2% of all help requests, were evidenced by the least frequently appearing tokens indicating an appearance rate of just over 2% of all support requests. Evidently, the preprocessing steps had a notable impact on the composition of the final wordlist.

A cursory review of the top and bottom ten tokens in the wordlist revealed a widely-varied rate of token appearances in individual helpdesk records. To better understand the variation in appearance rate, each token was assigned a range based on its appearance frequency in unique records, and the wordlist was summarized by this range. This summary is presented below.

Table 2

Number of Tokens in Percentage Ranges of Appearances in Unique HD Requests

Range of % of In Unique HD Requests	Number of Tokens
30% - 36%	1
24% - 30%	4
18% - 24%	6
12% - 18%	21
6% - 12%	111
2% - 6%	326

It quickly became evident that the vast majority of tokens appeared in relatively few individual help requests. Indeed, only 32 tokens appeared in 12% or more of all support requests. To better understand the tokens that appeared most frequently, the researcher generated a list of all words that appeared in 10% or more of all records. This list, which contained 59 tokens, is presented in Appendix A. A review of this list produced valuable information that would inform and complement ensuing analyses.

The subset of the wordlist that included the tokens that appeared in 10% or more unique helpdesk requests contained many words that on their own did not contain contextually relevant meaning. For example, the most frequently occurring word, *us*, (the stem word for *use*, *using*, *used*, etc.), while understandable to a human reader, did not in itself provide sufficient semantic meaning and context for relevance in a study intended to determine technology-related professional development topics from HDDB data. Similarly, many verbs appeared in the list of

frequently occurring tokens. These included, for example, *receiv*, *instal*, *change*, and *try*. These words may have a clear meaning to an English speaker, but do not, on their own, provide insight into the details of a helpdesk database. Other tokens, such as *school*, *click*, *http*, and *subject* are understandably present in an academic institution's HDDB but do not provide meaning or insight into the content of that database. Thus, many frequently occurring tokens did not provide clear insight into areas of technological weakness amongst the faculty and were excluded from future wordlist analyses.

To extract value from the wordlist, the researcher executed a crucial categorization analysis. After preprocessing, the HDDB wordlist was distilled to 469 unique terms, and the researcher rigorously reviewed each term to determine which were independently contextually useful and relevant to the study. This step was aided by the researcher's in-depth knowledge of the site's information technology practices. Had this phase been automated, and the wordlist was compared to an unaffiliated, pre-existing list of common technology terms, many valuable tokens would likely have been mislabeled as either useful or not useful, as many of the words had a meaning that was particular to the research site. The primary criteria by which the researcher gave a token a designation of *Meaningful* or *Not Meaningful* was that the token, taken on its own, was sufficiently semantically and contextually definitive to identify a specific hardware or software, or some other clear issue. In essence, a token was designated as *Meaningful* if the token clearly presented a topic that was relevant to the study at hand. Each meaningful term was also assigned a category, for grouping purposes. The outcomes of these designation and categorization processes are discussed in detail in the Results: Wordlist section of Chapter 4. The *Meaningful* designation assignment was the last analysis performed on the wordlist prior to results generation.

Creating the Word Vector in RapidMiner. Upon completion of the customization and operation of the RapidMiner processes that generated the wordlist, the final step before data mining was the conversion of the wordlist to a word vector. Sophisticated analyses such as clustering and classification required a properly formatted word vector, which was a standard element of text data mining procedures and an essential input into RapidMiner data mining processes. The word vector was the result of a procedure for quantifying the wordlist. The word vector was, essentially, a massive matrix that included one row for each help request (1,884 rows) and one column for every token on the wordlist (469 tokens). The intersection of each row and column was the frequency of occurrences of that term (a column) in that help request (a row). Thus, if a token (Token X) appeared 10 times in a help request (Request A) that contained 100 words, the term frequency for that token in that help request was 0.1 or 10%. If the same word (Token X) appeared 30 times in a different help request (Request B) that was 150 words long, the term frequency for that word in that request was 0.2 or 20%. Finally, if the same word (Token X) appeared 30 times in a help request (Request C) that was 400 words long, the term frequency for that word in that request was 0.075 or 7.5%. By using term frequency as the value at the intersection of a row and a column, as opposed to other frequency or summative options provided by RapidMiner, the frequency of appearance of any word in any help request record was normalized and thus comparable to all other values.

If absolute occurrences, for example, were used in RapidMiner in place of term frequency, the value at the intersection of a row and a column would simply have been the total number of times a word appeared in a support request. In the example above, the absolute occurrence value for Token X in Request A would have been 10, in Request B, 30, and in Request C, 30 again. These absolute numbers would not have provided adequate and valuable

insight into the relative importance of a word or token. The creation of the word vector, using term frequency as the value in the matrix where any token-column intersected a support request-row, allowed for the determination of the relative importance of a word or token.

The size and complexity of the word vector rendered it indecipherable to a human. The word vector, at 1,884 rows by 469 columns, contained 883,596 term frequency values, (including row and column titles). This normalized, quantified version of the wordlist was, however, the ideal input for RapidMiner's text data mining processes. The data miner designed two distinct data mining RapidMiner workflows that accepted the word vector as input. One workflow performed automated clustering analyses, in which the system automatically grouped helpdesk tickets together by identifying potentially related content amongst the tickets. The other workflow was designed to perform classification analyses, in which the researcher manually applied labels to a subset of the helpdesk tickets, and RapidMiner attempted to understand the labeling criteria and apply the same set of labels to the unlabeled support requests. Clustering was intended to create a small number of clusters from the word vector and the algorithm assigned each helpdesk request to a cluster. The researcher then reviewed and summarized the contents of each cluster for drawing results and conclusions. The methods for these processes follow.

Clustering Methodology in RapidMiner

While the wordlist presented opportunities to discover topics or trends within the dataset, additional data mining practices were undertaken to go beyond a summary of terms and frequencies. The first RapidMiner workflow to leverage the word vector was a clustering workflow. An important characteristic of this automated clustering process was that the researcher did not provide any information into RapidMiner regarding how support tickets might

be divided or grouped. The RapidMiner clustering process leveraged advanced data mining and statistical practices to create groups of related helpdesk tickets automatically. The clustering approach employed statistics to separate one group of documents from other groups based on the characteristics of words associated with each document. The researcher could then review the groups for analysis and to draw conclusions.

Clustering is a common approach to text-based data mining. The algorithms included in RapidMiner were intended to receive input from the user and output data that indicated which records were assigned to each cluster. The state-of-the-art algorithm functioned by determining a conceptual distance between all elements in the word vector. With theoretical distances determined, the algorithm then attempted to create clusters that simultaneously minimized the distances between data points within a cluster while maximizing the distances of data points between clusters (B. Tvenstrup, personal communication, January 5, 2017). The actual process of clustering is extremely complex, and for this reason, professional data miners rely on powerful and precise tools such as RapidMiner algorithms to perform the work and output the results. The outputted results included, amongst other data, the unique identifier for each help request and a code that indicated the cluster to which that request was assigned. This file was conveyed to the researcher via encrypted email. Using Excel, the researcher then merged each of these outputs with the cleaned, prepared, processed, and anonymized data file to create a list of detailed helpdesk requests that included the cluster assignments. The researcher could then review this file for findings and conclusions. However, before this important merging step, the researcher and the data miner configured the RapidMiner clustering algorithms to produce usable results.

The researcher and the data miner undertook two types of clustering processes. The first method was *k*-means clustering, in which the researcher and the data miner inputted a desired

number of clusters into the RapidMiner algorithm. The process of determining the correct number of desired clusters was complicated, as too many clusters would not yield results that were easily reviewable, and too few clusters would not provide adequate summarization. The researcher began by having the data miner run k -means clustering with $k = 2$. The algorithm would attempt to review all data points, defining 2 clusters with a conceptual line drawn between them. The process was run again, with $k = 3$, in which the dataset was divided intelligently into 3 clusters. The data miner outputted the result to a text file and conveyed it securely via encrypted email to the researcher for analysis. While this approach yielded useful results, k -means clustering required the researcher to determine the intended number of clusters, and the algorithm could not provide support as to the accuracy of that cluster count.

The second clustering process, x -means clustering, added a supplementary level of automated intelligence into the RapidMiner workflow. x -means clustering leveraged advanced machine learning technologies to attempt to determine the ideal number of clusters for a given dataset, and then to divide the data into those clusters. In principle, this clustering method determined the optimal number of clusters by processing the data with an increasing number of clusters in a set, from 2 through 50 or more, and calculating a score for each cluster set (B. Tvenstrup, personal communication, January 5, 2017). The algorithm measured the distinctness of each cluster created within each cluster set, but also penalized each cluster set for each additional cluster after the initial two. This complex process ran iteratively, creating and comparing scores for different counts of clusters. As with k -means clustering, the complex mathematics behind the x -means algorithm were kept largely out of sight from the RapidMiner user, and the power of the algorithm was in its ability to determine the ideal number of clusters quickly.

The researcher and data miner leveraged *x*-means clustering to calculate the ideal number of clusters given the word vector generated from the HDDB data. The algorithm determined that the dataset was comprised of information that could accurately be grouped into four clusters. While the *x*-means algorithm assigned each help request to one of the four clusters, the data miner advised that *k*-means clustering typically yielded more accurate cluster assignments (B. Tvenstrup, personal communication, January 5, 2017). Thus, having learned of the ideal number of clusters from *x*-means clustering, the data miner executed *k*-means clustering once again, with $k = 4$. This process generated a file containing, amongst other data, the unique identifier (the adjusted work order number) of each help request, and a cluster assignment for each support request, from Cluster 1 to Cluster 4. This file was then merged in Microsoft Excel with the original data file to create the final results of the clustering analysis: a list containing the details of each help request and a cluster assignment number. This information was then analyzed and prepared for finding and conclusion generation.

Clustering Analysis Methodology in Microsoft Excel. The second data mining procedure, in which a clustering algorithm was built and then applied to the HDDB word vector, yielded four distinct groups of helpdesk submissions. The details of these clusters were exported from RapidMiner and imported into Microsoft Excel where they were analyzed via conventional spreadsheet operations. While the clustering algorithm assigned each help request exclusively to one cluster, it was possible that all 469 tokens appeared in each cluster, as the tokens were not assigned exclusively to a group.

The initial analysis addressed the frequency of each token within each cluster. For each cluster, all tokens were assigned an average word frequency. RapidMiner determined this value by calculating the term frequency for each token within each helpdesk record assigned to a given

cluster and dividing that count by the total number of words in that record. These values were determined for each token within a given cluster, and the average term frequency was calculated for that token within that cluster. This process was repeated within RapidMiner for each token within each cluster, yielding an average term frequency for each token in each group. It was possible that a given cluster contained one or more tokens which had a term frequency of 0.00; this would indicate that that word did not appear in any help requests assigned to that cluster. To understand the contents of each group, the researcher analyzed the average term frequency by cluster in Microsoft Excel.

The goal of the clustering analysis was to separate different helpdesk requests from each other while grouping similar, thematically linked, requests together. To determine the theme for each cluster, the tokens assigned to each group were sorted in Excel by descending average term frequency, and the top- and bottom-most values were evaluated and reviewed. The fifty tokens with the highest average term frequency for each of the four clusters are included in Appendix B. Similarly, the fifty tokens with the lowest average term frequency for each of the clusters are included in Appendix C. By reviewing the most and least prominent tokens, the researcher determined the themes of each cluster. It was important to note that, due to the inexact nature of data mining as a computer-assisted practice, not all tokens, or helpdesk records, would align with the cluster themes. The purpose of this analysis was to glean general trends throughout each group, not to determine a link for all tokens in a cluster unequivocally. The outcomes of the application of human interpretation to the RapidMiner-generated cluster lists are addressed in the Results: Clustering Analysis section of Chapter 4. Continuing the investigation, the researcher next reviewed the content of the helpdesk tickets assigned to each cluster.

Merging Help Requests with the Cluster List in Microsoft Excel. In order to review the results of the clustering analysis for each helpdesk record, a list of all helpdesk tickets that included the detailed description field and the cluster assigned by the RapidMiner algorithm was created. To assemble this list, the researcher used Microsoft Excel to merge the RapidMiner output, which included the unique work order number and the cluster assignment, with the list of helpdesk tickets, which also included the work order number. Thus, since the work order number field was present on both files, a straightforward merge procedure allowed for the combination of the two files, keeping the essential columns from each document. The researcher then separated the combined file into four new spreadsheet pages, each containing the records assigned to one of the four clusters. The information contained within these files will be addressed further in the Results: HD Records with Cluster Number section of Chapter 4.

Merging the Cluster List with the Meaningful Words List in Microsoft Excel. In a process similar to the merge of the help requests and the cluster list, the researcher combined the list of tokens and assigned cluster with the list of 69 *Meaningful* words created during analysis of the wordlist. After merging the two files in Excel, the researcher created, for each cluster, a list that contained the token, its average term frequency within that cluster, and the category and subcategory of that token as assigned in the *Meaningful* words analysis. These lists were sorted by descending average term frequency so that the most frequently occurring words within each cluster were at the top of the list. As each cluster list contained all 469 tokens, the tokens deemed *Meaningful* were scattered throughout the sorted list of tokens. For any given group, words considered *Meaningful* might have had a high ranking average term frequency, or a middle- or low-ranked frequency. To consider only the most prevalent meaningful tokens, the researcher pruned the lists to include only *Meaningful* tokens that appeared in the top 100 ranked words,

while hiding any non-meaningful tokens in the top 100. This procedure generated four lists, one per cluster, which contained the *Meaningful* words that appeared in the top 100 words in each cluster when the lists were sorted by descending average term frequency. These lists provided valuable insight into the meaning of prevalent tokens in each group, and the lists and technology themes gleaned from them are presented in the Results: Clusters with *Meaningful* Tokens section of Chapter 4.

Having determined the top ranking *Meaningful* tokens in each cluster, the researcher combined all *Meaningful* tokens from all clusters to create a list of all the meaningful tokens that appeared in the top 100 tokens for any cluster. It was anticipated that important trends might be revealed when considering the most frequently occurring tokens across all clusters. In total, 54 unique, *Meaningful* words had appeared in the top 100 tokens for at least one cluster. This list was summarized and sorted in Excel, providing the count of clusters in which a token appeared amongst the top 100. The maximum number of clusters in which a token could have appeared was four, but no *Meaningful* word appeared in the top 100 tokens for all four groups, indicating that no *Meaningful* tokens were prevalent in all four clusters. The results from this analysis are addressed in the Results: Summary of *Meaningful* Tokens in all Clusters section of Chapter 4, and the list of these 54 tokens, sorted by descending number of clusters in which each token appeared, is included in Appendix D. This analysis was the last one performed on the clustering data mining data prior to results generation.

Classification Methodology in RapidMiner and Excel

In addition to the clustering workflow undertaken in RapidMiner, the researcher engaged in an alternate text data mining technique. As with the clustering processes, much of the complex statistical and mathematical work was performed by algorithms built into RapidMiner. The

classification process, however, required human intervention for a vital element of the procedure. Classification involved an attempt to train RapidMiner to assign a label to each record in the database. For RapidMiner to perform this labeling task, the data miner was required to input a list of pre-labeled helpdesk request records. The RapidMiner algorithm then attempted to learn from the examples provided and applied labels as accurately as possible to unlabeled records. Regarding the accuracy, the researcher and the data miner understood that perfect accuracy was impossible. Indeed, with two labels, a classification algorithm that could accurately assign labels with more than 50% accuracy (the accuracy of random assignments) was deemed satisfactory. An optimal accuracy range for two-label prediction and classification was 70-80% (B. Tvenstrup, personal communication, January 9, 2017). As the number of labels increased, the optimal accuracy levels decreased dramatically. For example, a classification system with three labels would be correct in 33% of random assignments. Thus, an acceptable accuracy level for computer-bolstered intelligent assignment to one of three labels was 33-50% (B. Tvenstrup, personal communication, January 9, 2017).

To maximize the accuracy of the classification model, and to support the project's goals, the researcher decided to use two labels. Based on the size of the dataset, the data miner recommended manually labeling at least 100 records with each label, or 200 records in total (B. Tvenstrup, personal communication, January 9, 2017). This value represented an absolute number rather than a relative number. While 200 labels represented approximately 10.6% of the entire 1,884-record dataset, a larger dataset - while possibly requiring more than 200 records for manual labeling - would not necessarily require labeling for 10% of all records.

Initially, the researcher used ATF (for actual technology failure) and PTF (for perceived technology failure) as labels to categorize the help requests. However, after the researcher began

the manual process of reading the details of helpdesk support request records from the HDDB and assigning the ATF and PTF labels, the researcher noted that these labels did not sufficiently describe the data. After reading nearly 200 helpdesk records, the researcher concluded that the ATF and PTF labels, as a dichotomic labeling system, were not adequately applicable to the data at hand. Upon thorough assessment of support requests, the researcher determined that many of the help requests were not failures, and couldn't be classified as such. For example, the researcher had originally anticipated a plethora of tickets in the style of "my computer crashed," "my Wi-Fi connection stopped," or "the light in my classroom projector is dim." With an expectation of these types of tickets, labels that referred to the ticket's underlying type of failure seemed appropriate, and the researcher would use human judgment to determine if the failure was avoidable had the submitter possessed increased technology savvy, or if the problem was truly unavoidable.

Upon a conducting a detailed review of the content of a random selection of the help requests, the researcher noted that support requests typically included a user asking the IT department for help with a problem. The researcher determined that some requests required an agent's participation to address the issue, while a technology-savvy user could have solved other requests. Revisiting the examples above, both a crashed computer and a malfunctioning Wi-Fi connection were potentially solvable by a support requester without the aid of an agent, (possibly by undertaking simple troubleshooting practices such as rebooting the computer or turning the Wi-Fi off and on). A projector bulb-related ticket couldn't be solved without a support agent, as only the IT members possessed replacement bulbs. The need to consider help requests not as failures but as problems was determined only after a thorough review of helpdesk records in Excel. The new scheme applied REQ-AGENT as the label for helpdesk requests that could not

be resolved without an agent's efforts, and NOREQ-AGENT as the label for requests that did not necessarily require an IT agent to resolve the case.

The dataset, by default, was ordered historically from oldest to newest. The researcher was concerned that the labeling of sequenced records would skew the data, as cases regarding particular technologies might be grouped together as a result of the technology implementation phases. Thus, while the original sequence of the dataset was preserved on the file, the dataset was sorted randomly, and the researcher marked individual records with one of the two labels after careful review and consideration.

By reading the details of each support request, the researcher was able to determine the most accurate label. For example, the researcher labeled support messages about Google Groups, which at the study site were created and maintained by the IT department, as REQ-AGENT. Similarly, any records that required an IT agent to drop off or pick up technology hardware or software were labeled by the researcher as REQ-AGENT. Further, support requests that required the IT department to create or modify user accounts were labeled as REQ-AGENT. Contrarily, support requests about activating a new device on the site's wireless network were labeled as NOREQ-AGENT, since the research site maintained published instructions for completing these steps. In addition, support requests related to printing and common printing errors were also labeled as NOREQ-AGENT, as configuration and troubleshooting information pertinent to printing was readily available at the site. Finally, support requests that asked for help with popular products such as Gmail and Google Drive were labeled as NOREQ-AGENT as users could easily search for solutions for these products on the Internet. An IT agent was not explicitly required to support problems with Gmail and Google Drive, and this requirement

characterized the label assignments. The researcher completed the labeling process in Microsoft Excel by labeling 641 records diligently and painstakingly.

Classification Analyses in Microsoft Excel and RapidMiner. After the manual labeling process was completed in Excel, the labels REQ-AGENT (requires an IT agent to resolve the case) and NOREQ-AGENT (does not necessarily require an agent to resolve the case) were assigned to a 641-record subset of the full 1,884 record dataset, and classification analyses proceeded using these records to build the classification model. Of the 641 labeled cases, 182 were labeled as NOREQ-AGENT and 459 were labeled as REQ-AGENT.

The 641 labeled cases were inputted into RapidMiner, and the classification algorithm was programmed to determine distinctions between the labels. A weighting formula was applied during the creation of the classification model to account for the discrepancy in the number of manually-labeled support requests for each label. In principle, the algorithm functioned by building a predictive model based on a subset of the labeled data, then testing the model on a different subset of the labeled data (B. Tvenstrup, personal communication, January 13, 2017). This process, called cross-validation, was performed 10-fold (with ten 64-case subsets of the 641 labeled cases) and iteratively grew more accurate. The algorithm also employed a Support Vector Machine model to learn from the labeled cases and determine a method for classifying unlabeled cases with the highest obtainable accuracy. The model, relying upon the cross-validation procedure, categorized the 641 manually-labeled cases and determined a confidence value for each prediction. The confidence value was a system-calculated figure between 0 and 1 that indicated the likelihood of an accurate prediction. Higher confidence values for a label assigned to a record indicated a greater likelihood of that label accurately reflecting the content of the record. The algorithm would compare confidence values to a threshold to determine the

most accurate label. The complex Support Vector Machine model iterated through many confidence thresholds, applied to each helpdesk request prediction, to determine the threshold that maximized the entire model's accuracy. The data miner programmed the model to accept a label prediction when the confidence value for that prediction exceeded a threshold, but the Support Vector Machine would be used to determine the threshold.

The algorithm might consider a baseline confidence threshold of 0.5 or 50% because this was the same confidence a two-label system would have generated in a prediction model based on random guessing. A confidence threshold that was too high, such as 0.8 or 80%, would only apply a label to a helpdesk record when the model was 80% confident in the accuracy of that label. A confidence level this high might leave many records unlabeled, or, in a binary label system where the algorithm must consider one label first, the algorithm might inaccurately assign the second label even if the confidence level for the first label is high (over 0.5), but under the desired threshold.

As had been the trend, the complicated functionality of the classification processes occurred transparently to the researcher and the data miner, and the system determined that the highest overall model accuracy was achieved with a confidence threshold of 0.6 or 60%. The algorithm compared confidence values to the 0.6 threshold to determine the appropriate label. For each record, if the RapidMiner-generated confidence for the REQ-AGENT assignment was greater than 0.6, the system assigned the REQ-AGENT label. If the confidence value for REQ-AGENT was below 0.6, the system assigned a label of NOREQ-AGENT.

The classification algorithm assigned a label that matched the manually-determined label with a cross-validation accuracy of 72.76%, per the RapidMiner logs. Thus, of the 641 manually-labeled cases, the classification model accurately classified approximately 465 records. The

researcher and the professional data miner agreed that the model, with a nearly 73% accuracy level (within the desired accuracy range of 70%-80%) had attained an acceptable and effective level of performance. While a random guessing algorithm might only accurately predict labels for one out of every two records, the classification model designed for this study accurately predicted nearly three out of every four records in the manually-labeled data subset.

Once satisfied with the functionality of the classification model, the researcher instructed the data miner to apply the labeling algorithm to the complete dataset. The miner applied the classification algorithm to the full dataset, and the process assigned each record a label of REQ-AGENT if the confidence value for the REQ-AGENT label for that record was 0.6 or greater. If the confidence level for the REQ-AGENT label for that record was less than 0.6, the model assigned the NOREQ-AGENT label to the record. In this manner, the algorithm classified each record. The results of the classification process were exported from RapidMiner for further analysis using Microsoft Excel. The exported file included information about each support request such as the work order number, the RapidMiner-determined confidence values for each label, the manually-assigned label (only for the 641 manually-labeled records), and the classification label from the algorithm. A sample of this data file is presented in Appendix E.

Merging Help Request Data and Classification Labels in Microsoft Excel. The researcher created a list of all helpdesk tickets (including the comprehensive description field) which also contained the classification label assigned by the RapidMiner algorithm. To create this list, the researcher used Excel to merge the RapidMiner output, which included the unique work order number and the classification label, with the list of helpdesk tickets, which also included the work order number. Since the work order number field was present on both files, a straightforward merge procedure allowed the combination of the two files, keeping the essential

columns from each document. The researcher then separated the combined file into two subsets of the data for closer analysis. One subset file included all the cases labeled with REQ-AGENT, or cases that required an agent for resolution, and the other included all the cases labeled with NOREQ-AGENT, or support requests that did not require an agent for resolution. The results of the analyses conducted upon these files are outlined in the Results: Help Request Data and Classification Labels section of Chapter 4.

Merging Clustering and Classification Results in Microsoft Excel. The final analysis undertaken in this study utilized the results of two other analyses. By combining the list of helpdesk tickets with a cluster number assigned, and the list of helpdesk tickets with a classification label designation, the researcher created a list of all helpdesk tickets, including the detailed description field, the cluster, and the label assigned to each support request. The researcher used a spreadsheet tool to merge the two spreadsheet files, thus assembling a master list that contained information from two of the three data mining procedures. This list was summarized to combine the results of the procedures, to determine the number of support tickets within each cluster that were resolvable with and without a support agent. The results of this summary are included in the Results: Clustering and Classification Combined section of Chapter 4. The creation of this merged file represented the final step in the classification procedure, itself the last step of the data mining phase.

Methodology Conclusion

The detailed methodology highlighted the procedures executed during this case study. From extracting the data to exporting resultant files, each step of the lengthy and complex procedure was provided to inform and inspire other researchers. The results of this multifaceted data-driven endeavor were manifested by three distinct output files from RapidMiner, the

wordlist, the clustering results, and the classification results. These files were summarized or combined with other data files to glean more information from them. Armed with the full mining results, as well as the summaries and merged files, the researcher was positioned to undertake extensive analysis and review of the mined data, from which conclusions were drawn. Microsoft Excel was used to summarize and examine the data mining outcomes to highlight trends and themes of technological weaknesses amongst the faculty body. Further, additional findings, including topics for professional development and information that might inform the planning of professional development, were gleaned from the data mining results. These findings are presented in Chapter 4.

Ethical Concerns and Participant Rights

The researcher conducting this case study was committed to protecting data contributors' rights and maintaining anonymity, respecting privacy, and ensuring confidentiality. The primary method by which the study accomplished these goals was by removing all identifiers from the HDDB data before analysis began.

The researcher identified two risks as potential unintended consequence of the case study. If the study did not maintain anonymity or confidentiality, it was possible that the study could accidentally identify one or more faculty members as lacking technological skills. If such an event were to occur, it was possible that the person(s) identified might be concerned about maintaining their employment. In this scenario, the individual's participation in the study would cause them harm.

The second risk potentially affected the faculty body and the institution as a whole. Even with anonymous data, if the researcher published a report that included mention of the technological weaknesses of a faculty body, it was possible that this information might be

misconstrued or potentially damaging to the image and reputation of the school. In theory, this information might affect hiring or enrollment.

The researcher reduced the risk to individuals to a reasonably low level by ensuring the anonymity of the data used in the study. Although the data, including identifiers, existed in an accessible form before the undertaking of this study and would continue to be generated after the researcher completed the project, the researcher stripped the data of identifying information before any analysis. Further, the researcher cleansed the data file of identifying information and then resaved, thus eliminating any chance of recovering the identifying information from the file. All data used in the study permanently lacked identifying information. The file itself was stored on the researcher's password-protected computer and backed up online in a secure, encrypted, password-protected cloud storage service. It is important to note that the study looked for trends in the data, and thus individual records were not at the forefront of the research, nor was attention ever called to a particular record.

Regarding potential harm to the school as a result of publishing or sharing the study's findings, several steps were taken to reduce the risk of harm to a reasonably low level. Firstly, the focus of the study was the capacity of data mining methods to identify the trends within an HDDB; the trends themselves were of secondary importance. The shared results for public consumption focused on the data mining topic discovery in the text-based data rather than on the particular trends, though the study included the trends in the findings. Any trends identified were not reviewed or measured for their impact on the school's operation, and are highlighted only to exemplify the usefulness of the data mining

At the time of the study, many independent boarding high schools employed an Information Technology helpdesk to support teachers in their use of technology. The researcher

reasonably assumed that all institutions' faculty bodies experienced trends in technology failures and weaknesses and that a similar study at any given location could have revealed such trends. It was the purpose of this case study to present the application of data mining techniques to HDDBs as a valid method for determining technological skill set weaknesses amongst a faculty body. If any given institution undertook and published a similar study, each school's technological weakness trend information would become available. Indeed, a primary goal of this undertaking was to inspire other researchers and institutions to do just that.

Consent

The study's leveraging of archival data, generated daily during normal business practices, placed the research study within the realm of studies with exempt research status. As the institution owned the data and authorized the researcher to study it, and the researcher had daily access to it by virtue of his professional role, the researcher was exempt from obtaining consent to view the data. Data contributors, who have helped create the data by submitting IT support requests during the three-year period under review, could reasonably have expected an authorized school employee at the Director level to examine and study the data.

The Director of Educational Technology and conductor of this case study intended to do no harm to data contributors or the institution. The researcher obtained a signed consent form from the Dean of Faculty, a member of the school's senior staff, indicating that the school had authorized the researcher to use anonymized helpdesk data for research. As the school owned the data, the Dean of Faculty was empowered to authorize the researcher to use HD submissions from teachers, including those no longer employed at the school. The Dean of Faculty's signed consent form fully authorized the researcher to perform an ethical study of the school's HDDB.

Data contributors in the case study based upon anonymous technology support tickets could reasonably expect no harm to come as a result of their involvement.

Conflict of Interest

The preservation of rigorous scientific inquiry and the protection of the participants were of the utmost concern in this study. As the study participants were employees at the same institution where the researcher was employed, and providing technological professional development to said employees was amongst the researcher's primary responsibilities, the researcher was mindful of any potential conflicts of interest. The study might have revealed areas of technological weakness for faculty members, and the researcher's employment responsibilities included training faculty members on technological topics. This was not perceived as a conflict of interest, as the data contributors and the researcher stood to benefit from new insight into faculty members' technological weaknesses. Any trends in technological weaknesses identified by the study might be addressed in future professional development sessions, thus benefitting the employees and contributing to the researcher's work role. The school administration had entrusted the researcher, in his role as Director of Educational Technology, to identify areas of weakness and provide professional development if possible. The Director of Educational Technology reported to the Dean of Faculty and did not have any authoritative control or power over, nor responsibility for, the members of the faculty body. Thus, the interests of the study were in line with the professional interests of the administration, the site, the data contributors, and the researcher.

Assumptions, Limitations, and Scope

Several concepts were held to be true in the undertaking of this study. As a research project, the identification of assumptions was both inevitable and required as these assumptions

would bolster and contribute to both the expectations and conclusions of the study. Though the study anticipated reliability for these assumptions as the study progressed, the study might have revealed that certain assumptions were inaccurate or unwarranted.

The study assumed that the HDDB contained a sufficient amount of valid and minable information for topic detection. Similarly, the issues discovered in the HDDB were assumed to be an accurate depiction of a faculty body's technology skill set deficiencies. Teachers at the study site were presumed to use the services of the helpdesk as expected and that helpdesk agents adequately performed the task of storing detailed information about each help request. A final assumption was that data mining methods could effectively be applied to analyze and summarize the data as desired. These assumptions were believed to be true as of the initiation of this research undertaking.

Limitations

A number of limitations were in place that could have restricted the scope of this case study. An important limitation pertained to the employees who have submitted helpdesk requests. While the HDDB at the research site contained requests from teachers, staff, employees, administrators, and students, the study only considered requests submitted by teachers. This limitation was a result of the restrictions placed upon the researcher in his role as the Director of Educational Technology. The Director's primary group of colleagues and constituents was the faculty, and the Director was empowered to provide professional development to the faculty body. Thus, while a similar study might have revealed useful information about students' and employees' technological skill sets, this case study focused on the group with which the Director had the most contact and administrative influence.

An additional limitation of this study was concerned with the transferability and reproducibility of the research. Nearly all HDDBs at any organization store similar data in different structures. While the data mining practices leveraged in this study might have worked in principle at other sites and in other studies, in practice the specific investigative procedures and conclusions drawn as a result of data mining were limited to this site and study. The study did not intend to recommend specific analytical approaches nor generate results that were wide-ranging representations or generalizations.

As the study leveraged archived data, any findings might have represented the state of the faculty's technological weaknesses during the timeframe of the study. The use of archival data as a sole data source in any study might yield information about how a situation *was* during the study timeframe, rather than precisely how a situation was at the current moment (Abowitz & Toole, 2009). As the study's timeframe ended in September 2016, and the completed analysis, findings, and results were not anticipated until spring 2017, it was possible, though unlikely, that the trends identified in the study were no longer accurate. However, as the study's goal was to determine the value of data mining techniques applied to the HDDB, the content of the trends was not as important to the study's findings as was the capability of revealing those trends in as much detail as possible.

Delimitations

While the HDDB contained many years of support request data, the study only leveraged data collected within the aforementioned three-year period in this study. This limitation was a result of the application of data distillation for human judgment as a data mining method. As the study required human judgment for topic detection, it was important that the human judge was familiar with the concepts contained within the data. The researcher was employed at the site for

the three years identified in the study and could thus competently distil data collected during this timeframe into logical and accurate groupings, trends, or topics.

Scope

The scope of this study was deliberately narrowed to ensure feasibility. The case study focused on the technological support requests placed by teachers at an independent boarding high school within the aforementioned three year period. The data examined in the study was mined using the commercial data mining tool RapidMiner Studio by a certified, professional data miner. The scope of the analysis was limited to topic detection within the HDDB.

Conclusion

This single-site instrumental case study mined archival helpdesk data to determine the value of data mining practices on an HDDB as a method to isolate and identify technological weakness trends within a faculty body, whose support requests had built the HDDB. The site's Dean of Faculty granted administrative approval for the study, and the researcher undertook all necessary precautions to ensure data contributor privacy. While the purpose of this study was to evaluate the usefulness of DM practices applied to the HDDB, the hope of the case study's researcher was to inspire other researchers to conduct similar studies at their sites. The methods employed in this study might not reveal generalizable results, but the methodology, buoyed by the findings as discussed in the ensuing section, could be leveraged by scholar-practitioners in future undertakings.

CHAPTER 4

ANALYSIS, RESULTS, AND FINDINGS

The data mining procedures executed in the methods chapter produced data that could be analyzed and summarized in attempt to glean findings, results, and conclusions from the HDDB. The study's topic, purpose, and research questions were leveraged throughout the results generation phase, as these provided valuable guidance, direction, and perspective while the data was reviewed and studied. The purpose of this case study was to employ data mining methods to an underused data source, the HDDB, to identify and examine areas of improvement for the technology skills of faculty members at an independent boarding high school in the United States in the recent past. The study's overarching goal was to determine the potential and value inherent in the application of data mining techniques to an HDDB. The desired outcome of the case study was to determine if data mining procedures, applied to a specific subset of a school's complete HDDB (for example, help requests from faculty members) could determine areas of weakness (and, potentially, improvement) for the very people whose technology help requests had created the content within the HDDB. Thus, the analysis and results generation phase of the case study approached the data with this outcome in mind.

As illustrated in the Methods chapter, the data mining work was divided into three distinct analyses. These were the creation of the wordlist, a clustering analysis to create groups of related helpdesk support request tickets, and a classification analysis to automatically and intelligently apply one of two labels to all help requests. Chapter 3 highlighted the planning and execution of these procedures, beginning with inputting processed data files into RapidMiner and finishing with outputted summary files further processed in Excel and made suitable for analysis.

This chapter will focus on the analyses of the mined outputted data and present the results of the analyses.

Analysis

While certain data files exported from RapidMiner and prepared for analysis in Microsoft Excel were ready for results generation, other data files required additional analysis prior to determining findings. Both the wordlist and clustering outputs required supplementary scrutiny through the lens of a human interpreter to glean useful conclusions that were relevant to topic determination. These additional processes are outlined below.

Additional Wordlist Analysis

The complete wordlist was reviewed in Microsoft Excel. From the list of unique tokens that appeared in the HDDB, the researcher designated as many terms as possible as being contextually *Meaningful*. Of the 469 unique tokens, the researcher found that 69 could be satisfactorily labeled as being *Meaningful*, while the remaining 400 were not designated as such. The 69 *Meaningful* tokens, along with their occurrences in unique HD records and manually assigned category and subcategory, are included in Appendix F. The 54 *Meaningful* tokens that appeared in the top 100 tokens for at least one cluster are included in Appendix D. Of the 54 most prevalent tokens, nine words appeared in more than 10% of all helpdesk records, as shown in Table 3.

Table 3

Meaningful Tokens that Appeared in 10% or More of All Helpdesk Records

<i>Meaningful</i> Token	In Unique HD Records	Total Instances	% of Unique HD Records
phone	342	620	18.15%
network	334	672	17.73%
password	283	663	15.02%
group	253	1,918	13.43%
classroom	218	327	11.57%
equip	204	228	10.83%
print	200	605	10.62%
vdt	196	371	10.40%
visual	196	212	10.40%

Table 3 provided exemplars of *Meaningful* tokens that appeared in many helpdesk records. The token *phone* was determined to refer to some aspect of the IT-managed phone system at the site. The words *network*, *password*, and *group* suggested issues with the wired or wireless data network, password resets or password-related login problems, and Google Groups, respectively. Google Groups was an important component of the Google Apps Suite for Education, which was launched at the site during the second year of the study timeframe and was used heavily since. Indeed, the Google Groups platform, primarily a tool for communicating digitally with large groups of users, contributed many tokens to the list of meaningful words. The tokens *classroom*, *equip*, and *visual* were deemed to refer to hardware (including audio/visual) issues in campus classrooms, while *print* referred to problems with the campus-wide printing system. The *vdt* token referred to a custom technology tool used at the site called Virtual Desktop, which was used by faculty prominently during the first year of the study timeframe. By reviewing their contextual connotation, all 69 tokens identified as *Meaningful* were labeled.

Just as the researcher determined semantically and contextually relevant terms from the wordlist, the 69 *Meaningful* tokens were divided into categories and subcategories. The

researcher believed that these categories represented overarching themes or areas of weakness within the HDDB. (The categories and subcategories assigned to each of the 69 *Meaningful* tokens are included in Appendix F.) Categories and subcategories were devised to provide both distinction between themes and descriptive detail within themes. If a token was sufficiently meaningful and distinct from other tokens in a category, the token was assigned to a subcategory. For example, the hardware category contained tokens that were distinct enough to warrant subcategories, such as classroom audio/visual hardware and Apple Mac computers. Some categories did not contain sufficiently distinctive tokens to warrant subcategories. For example, while the Virtual Desktop category contained three tokens (*vdt*, *virtual*, and *virtual_desktop*), the tokens lacked sufficient differentiation to justify the creation of subcategories. The tokens were summarized by category and subcategory, as shown in Table 4.

Table 4

Summary of Meaningful Tokens by Category and Subcategory

Category and Subcategory	Count of Appearances in HD Records	Count of <i>Meaningful</i> Tokens
Accounts - Creation	130	3
Accounts - Okta (SSO)	83	1
Accounts - Password Reset	336	2
Gmail	39	1
Google Drive	83	2
Google Groups	2,977	21
Hardware	314	2
Hardware - Classroom A/V	1,276	10
Hardware - Copiers	66	1
Hardware - iOS	79	2
Hardware - Mac	246	3
Network	743	5
Phone	521	3
Printing	301	2
Software - Academic Web	123	2
Software - FA Web	62	1
Software - FirstClass	258	3
Software - Google Chrome	82	1
Software - Web Browser	51	1
Virtual Desktop	491	3

The category summary analysis revealed valuable evidence that described the content of the HDDB. Eighteen *Meaningful* tokens were assigned to the Hardware category, with sixteen of those further classified amongst the classroom audio/visual equipment, photocopier, Apple iPhone and iPad, and Apple Mac computer subcategories. Seven *Meaningful* tokens were assigned to the Software category, divided into subcategories of commonly used software at the site. These included an intranet page called Academic Web, a grade book tool called FA Web, the Google Chrome web browser and another web browser, and FirstClass (also known internally as SWIS), an email system used at the site during the first year of the study timeframe. Finally, six *Meaningful* tokens were assigned to the Accounts category, into three subcategories

to distinguish between types of account-related issues. The subcategories represented problems with account creation, password resets, and problems with the site's single sign-on tool, Okta.

Additionally, several top-level categories without subcategories were identified. These included items from the Google Apps Suite for Education. The Google Groups category contained 21 tokens, Google Drive (web-based document storage, creation, and collaboration) contained two tokens, and Gmail (web-based email) had one *Meaningful* token. Finally, other standalone categories included network issues (five *Meaningful* tokens), phone issues (three tokens), issues with the Virtual Desktop platform (three tokens), and printing issues (two tokens). Thus, while only a small number of tokens were semantically or contextually relevant to warrant labeling, the labeled tokens provided insight into some of the content contained with the HDDB. The wordlist, now filtered, summarized, and organized through the lens of human judgment was ready for results generation. These results are outlined in the Results: Wordlist section of this chapter.

Additional Clustering Analysis

Just as the wordlist required additional manipulation before the determination of findings relevant to the study's goals, the clustering analysis output was further analyzed before results were pursued. To better understand the contents of the clustering analysis, the researcher carefully reviewed each cluster. RapidMiner created four clusters of tokens, each containing all 469 tokens, and the data mining software assigned an average term frequency for each term in each cluster. The cluster lists, once sorted by the average term frequency of each token in each group, could be manually reviewed to glean the topics or themes prevalent in each cluster. The top 50 tokens for each cluster, sorted by average term frequency, are included in Appendix B.

The first cluster, referred to as Cluster 1, was replete with words that were directly related to technological hardware. These tokens included, for example, *comput*, *projector*, *classroom*, *equip*, *audio*, *visual*, *room*, *classroom_equip*, *laptop*, *sound*, *mba*, *mac*, *pick*, *plug*, and *cabl*. These terms, among others in Cluster 1, were in some way related to or referencing hardware. Some tokens were self-explanatory, such as *projector* or *classroom*, and others were discernable stem words of technology tools, such as *comput* for *computer* and *cabl* for *cable*. Other tokens in this cluster required specific knowledge of the research site's technology vernacular to understand their meaning in context. For example, *mba* was an internal term used to refer to a teacher's MacBook Air computer. Similarly, in IT support requests, *pick* typically referred to an employee seeking to retrieve hardware from the IT office or to have an IT agent pick up hardware from a location on campus. By reviewing the token list sorted by average term frequency, the researcher began to understand the themes, topics, or trends that might unite many of the help requests assigned to Cluster 1.

Conversely, many of the tokens in a given cluster appeared with a very low, or zero, average term frequency, and thus did not often appear in help requests assigned to that cluster. These words did not contribute to the theme of the cluster but did highlight the topics that were irrelevant to the cluster, thus providing further clarification into the cluster's thematic nature. The bottom 50 tokens, by average term frequency, for each cluster are included in Appendix C. In Cluster 1, these tokens did not refer to technology hardware. These terms included, for example, *us_direct*, *set_manag*, *profil_match*, *owner*, *manag_group*, *exampl*, and *employee_number*. Once again, the researcher's familiarity with the site provided valuable input into deciphering the contextual meaning of these tokens. *set_manag* and *manag_group* (and many other words near the bottom of the list for Cluster 1) referred to Google Groups related

issues. Similarly, *profil_match* referred to a particular type of help request wherein a manager was soliciting the creation of an account for a new employee, where the account configuration was set to match an existing employee's configuration. The topics at the bottom of the list were useful for the researcher to determine the types of help requests that may not appear in Cluster 1. This information complemented the identification of tokens that highlighted Cluster 1 themes to paint a clearer picture of the content of Cluster 1. By leveraging the distilled data created by the data mining process, the researcher determined that help requests assigned to Cluster 1 typically focused on hardware-related technology issues. Of the 1,884 support requests in the database, the RapidMiner clustering algorithm determined that 643, or approximately 34%, might be related to hardware problems.

In a similar manner method used to determine the unifying topic of Cluster 1, the researcher identified themes for the remaining three clusters. Cluster 2 contained terms that were related to software, accounts, or technology services. These tokens included, for example, *print*, *password*, *reset*, *swi*, *okta*, *vdt*, *virtual_desktop*, *printer*, *access*, *account*, and *chrome*. Once again, many of these terms were self-explanatory, while others, such as *swi*, *okta*, *vdt*, and *virtual_desktop* required the researcher's site knowledge to understand that these tokens referred to specific software or technology services in use at the site. *swi* (stemmed for SWIS, also known as FirstClass) was the email platform used for the first year of the study timeframe prior to the implementation of Gmail. Similarly, *okta* (the single-sign-on tool in use at the site), *vdt* (an acronym for Virtual Desktop) and *virtual_desktop* (a custom technology used at the site) all represented software, accounts, and technology services provided by the IT department.

Tokens that did not appear in Cluster 2, or appeared very infrequently, included *wall*, *video*, *us_direct*, *unplug*, *task*, *speaker*, *set_manag*, *projector*, and *member_add*. While some of

these tokens (such as *wall*, *video*, *unplug*, and *projector*) referred to hardware and thus were better suited in Cluster 1, the researcher noted that certain tokens, such as *set_manag* and *member_add*, were likely related to Google Groups. The researcher was initially concerned about this find since Google Groups-related help requests might have been well-situated in a cluster of requests thematically linked by software, accounts, or technology services. Nevertheless, the results from the clustering analysis were accepted, and the algorithm had assigned 343 help requests, or approximately 18% of all help requests, to Cluster 2.

The RapidMiner clustering algorithm assigned the largest number of records, 701, or approximately 37%, to Cluster 3. The variety of support requests occurring in Cluster 3 indicated that topics were not as easily discernable as they were for Clusters 1 and 2. The most frequently occurring tokens in Cluster 3 were diverse. For example, amongst the top fifty tokens were *phone*, *account*, *us* (stemword for *use*, *using*, *used*, etc.), *network*, *access*, *messag*, *subject*, *telephon*, *issu*, *work*, *student*, *receiv*, *email*, *googl*, *instal*, *chang*, and *school*. Similar discord was found during a review of the least frequently appearing tokens in Cluster 2, which included hardware terms such as *hdmi* (a common audio/visual connection and wire), *speaker*, and *projector*, Google Groups terms such as *manag_web* and *manag_group*, and contextually unintelligible words such as *basic*, *hand*, *save*, and *descript*. Thus, owing to the wide variance in both frequent and rare tokens in Cluster 3, the researcher theorized that this cluster represented a miscellaneous, or catchall, cluster to which any record not allocated to a different group would be assigned. This concept was supported further after the researcher reviewed Cluster 4.

Cluster 4 unequivocally contained help requests that pertained to Google Groups. The most frequently occurring tokens included *group*, *googl*, *member*, *forum*, *add*, *group_googl*, *manag*, *http_group*, *add_member*, and *member_group*. Indeed, more than any other cluster, the

most frequently occurring tokens in Cluster 4 clearly referred to the common theme of the cluster. Similarly, many of the lowest frequency terms had no relation to Google Groups, further solidifying Google Groups as the topic of Cluster 4. Some of the lowest ranked tokens included *work_order*, *wireless*, *window*, *wall*, *virtual_desktop*, *turn*, and *supervisor*. The algorithm had assigned 197 help requests, or approximately 10% of the HDDB, to Cluster 4.

Both the wordlist and RapidMiner-generated clusters required significant human interpretation to prepare them for results generation. Upon completion of this interpretive task, the results of these analyses were reviewed independently, combined, and in conjunction with the results of the classification analysis to determine findings, results, and conclusions. Upon completion of all analyses, the process of generating results and conclusions began.

Findings and Results

Given the goals of the case study, results were expected to include valuable descriptions about potential professional development topics extricated from the HDDB. The data mining procedures applied to the text-based support request dataset were executed with an aim towards identifying themes or areas of weakness of the faculty body who had submitted the help requests, and each mining exercise provided valuable insight into these themes. Results from the wordlist analyses are presented first, followed by results from the clustering and classification analyses. Outcomes from combined or merged analyses are presented as well.

Results: Wordlist

The technology categories and subcategories derived from the *Meaningful* token list (and presented previously in Table 4) could be used as a starting point in the determination of professional development topics extracted from the HDDB. Included in the summary was the count of individual HD requests that included each token, by category. However, this value was

only relevant in comparisons between categories. The values themselves are unimportant, as they represent a summation of unique records for each token, and owing to the non-exclusivity of tokens in records, records were counted multiple times. If a support request included a token from multiple categories, that record would be counted twice, thus no longer rendering it unique. The Google Groups category had many tokens that appeared in a large number of helpdesk records. The hardware category also had many tokens that appeared in many help requests, as did the network category. The software, accounts, phone, and Virtual Desktop categories contained tokens that appeared in a similar number of requests, and the printing, Google Drive, and Gmail categories had a smaller number of tokens that appeared in help requests. Figure 1 presents these values visually:

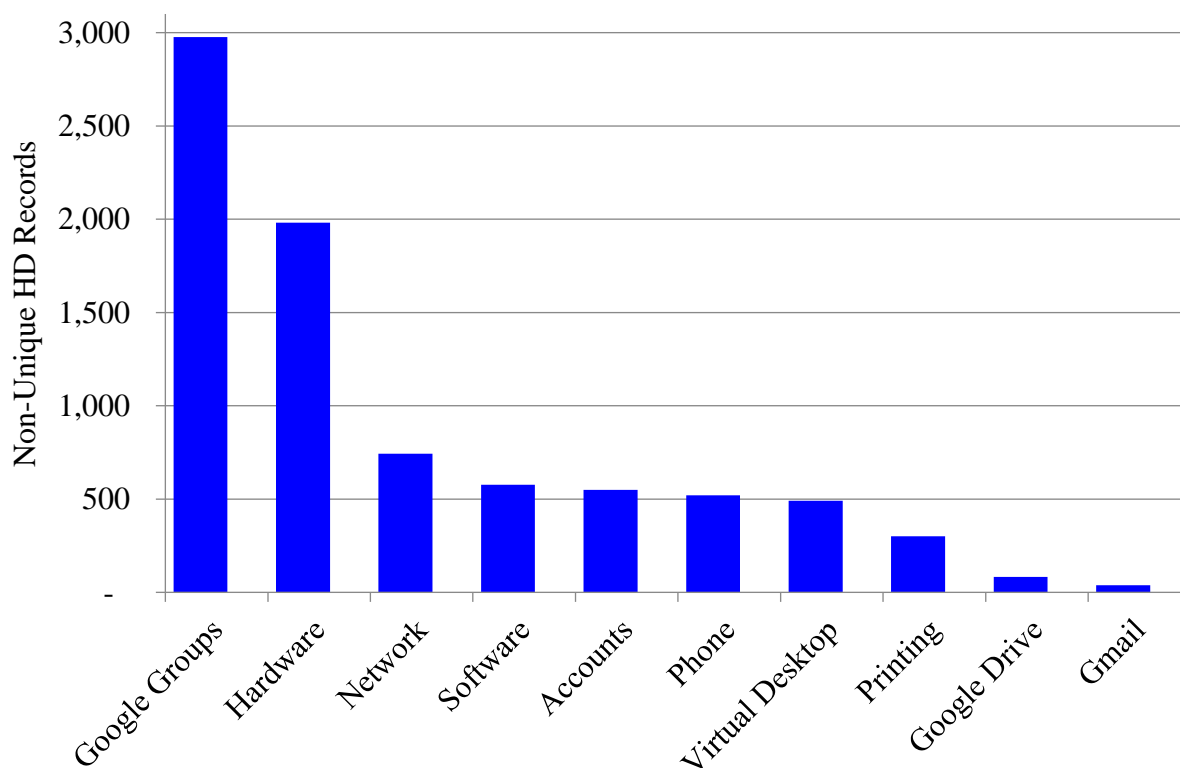


Figure 1. Number of appearances in HD records by category.

Professional development on any of these categories, or their subcategories, could potentially reduce the number of support requests related to these technologies. Indeed, from this analysis, it was determined that professional development on software systems such as Google Groups, Gmail, Google Drive, the Google Chrome web browser, the Academic Web intranet page and FA Web grade book, and account management tools such as the single-sign-on tool and the user-initiated password reset tool, could contribute to a reduction in help requests and empower teachers to work more confidently with these technologies. Similarly, professional development pertaining to hardware such as the classroom audio/visual tools, the enterprise-grade photocopiers and phones, handheld Apple products, and the school-issued MacBook Air computers might bolster and improve the use of these tools by the teachers while reducing help requests to the IT department. The identification of these topics would have been far more complicated without the application of text data mining techniques.

The categories of technologies for which users might benefit from professional development that were identified in this analysis almost certainly did not represent the entire range of topics contained within the HDDB. However, other themes could not be isolated from the wordlist simply because many tokens did not, on their own, offer semantic or contextual meaning. The wordlist analysis, which isolated, characterized, and summarized meaningful tokens, succeeded in highlighting trends of some problematic technologies mentioned in the HDDB that were hitherto unknown. This illumination continued during the review of other, more complex data mining analyses.

Results: Clustering Analysis

The clustering analysis in RapidMiner, followed by human judgment applied by the researcher, yielded four distinct clusters or groups of help requests within the HDDB. These

included hardware (Cluster 1), software, accounts, and technology services (Cluster 2), Google Groups (Cluster 4), and a catchall group (Cluster 3) whose thematic link amongst help tickets was a lack of relevance to any other cluster. The algorithm had assigned approximately 63% of all tickets in the HDDB to a meaningful cluster, leaving 37% allocated to the miscellaneous group. Thus, the clustering analysis provided not only major themes that indicated the types of incoming help requests, but also an approximate count of each type of ticket. Hardware support requests accounted for approximately 34% of incoming tickets, software, accounts, and technology services accounted for approximately 18%, and Google Groups-related tickets accounted for approximately 10% of incoming tickets.

Armed with this information, the institution might decide to provide professional development on the hardware technology employed at the site, as professional development on this topic might yield the greatest benefits to teachers and IT services alike. If teachers received training on the hardware they used on a daily basis, such as classroom audio/visual technology and MacBook Air computers, they might become more comfortable with the equipment and experience fewer hardware-related work stoppages, or they might be better prepared to handle issues that arose. Similarly, if the teachers received training on the hardware systems in use at the site, the IT department may receive fewer help requests on these technologies, allowing them to assign resources to other projects.

Further, the clustering analysis provided a potential professional development long-term plan. For example, professional development in the fall of a school year might be dedicated to the most problematic category, onsite hardware, while training in the winter might address software, accounts, and technology services. Finally, springtime professional development could address Google Groups problems, the category that generated the fewest support requests.

Additionally, once administrators became aware of the three foremost categories of identifiable helpdesk requests, the school leaders could decide on a different professional development plan, informed, but not necessarily dictated by, the anticipated volume of requests in each category. For example, while Google Groups tickets accounted for the fewest of the identified help requests, the administrators at the site might decide to train on this topic first as the researcher had anecdotally observed that the creation of Google Groups happened predominantly at the start of a school year. (This and other time-based trends could be supported by additional data mining analyses, if such analyses supported the goals of a project).

The results of the clustering text mining procedure provided invaluable insight into professional development topics at the site. In addition to the areas of focus (hardware, Google Groups, and software, accounts, and technology services), this analysis ascribed importance, or, potentially, urgency levels for addressing each area. This information, which could undoubtedly be used to develop professional development, would not have been available without the application of data mining techniques and human judgment applied to the mining outcomes. An additional analysis that followed this model was a detailed review of the full-length helpdesk requests assigned to each cluster. The researcher merged the cluster list with the full helpdesk dataset to create this file for analysis.

Results: HD Records with Cluster Number

The original intent of analyzing the detailed HDDB records with assigned cluster was to observe the accuracy of the RapidMiner clustering process, by spot-checking 10% (approximately 188) of the help requests to manually determine if the cluster assigned to a support request was accurate. This process was laborious, and the researcher was concerned with the scalability of this step in future studies that might contain a larger dataset. However, upon

review of the help requests assigned to each cluster, the researcher made an important discovery derived from the prevalence of particular tickets in Cluster 4, Google Groups.

Cluster 4 was identified as containing tickets that pertained to Google Groups. When the researcher spot-checked help requests in this group, it became apparent that many of the records in this cluster contained nearly identical text, which described a request by a user to have a Google Group created. At the research site, all employees used Google Groups, but only IT agents could create the Groups. This arrangement was a decision made several years prior by the Director of Technology and the Director of Educational Technology with the intention of reducing the total number of Google Groups, standardizing the naming and email addressing conventions for Groups, and ensuring that Groups were configured correctly for usage. Any time a user at the site wanted to create a Google Group, the user would fill out an online form that would automatically send an email to the helpdesk. This email contained nearly identical wording, the only exceptions being the desired name and email address of the Google Group, the group's owners and members, and a small number of settings. Each time the IT department received a request to create a new Google Group, the IT agent on hand had created a helpdesk ticket to manage the support request. The RapidMiner algorithm had assigned all requests to Cluster 4.

An Unexpected Finding. The finding that these nearly identical support requests were grouped together was a valuable discovery related to the application of data mining methods applied to an HDDB. The algorithm had created a cluster specifically for these types of tickets, effectively isolating them from other tickets, and keeping the topic of Google Group creation and management from attenuating the themes of other clusters. Further, the isolation of these tickets from all others provided a user-friendly indication that the complex clustering algorithm was

functioning correctly. These tickets behaved as a control group scattered throughout the dataset. While a human could not easily categorize many of the 1,884 tickets, a person could have easily categorized the Google Groups creation tickets, if necessary. That the algorithm classified the cases in this control group in the same manner as a human transcriber provided reassurance that the data mining clustering procedure was operating accurately. Cluster 4 thus provided not only the tickets related to Google Groups but also evidence of the clustering algorithm's adequate operation.

While Cluster 4 contained all of the Google Group creation requests, this cluster also contained other helpdesk requests related to Google Groups. These requests included, for example, a request to change ownership of a Group, a request to close an unused Group, or a request from a Group manager for assistance in some aspect of Group management. As this cluster accounted for just over 10% of all support requests, the researcher determined that Google Groups might be an important professional development topic. If users were educated upon proper creation and maintenance of Google Groups, then the responsibility for creating the groups could shift from the IT department to the users. This would empower the users, reduce the wait time between requesting the creation of a Google Group and accessing the Group, and lessen the load on the helpdesk. The prevalence of support request tickets related to Google Groups, as unearthed by the clustering data mining procedure, clearly highlighted a professional development opportunity that could benefit numerous stakeholders at the site.

Results: Clusters with Meaningful Tokens

While the cluster lists and *Meaningful* token list provided insight into the value of HDDB data mining, analysis of the merged files also proved very fruitful. By combining the cluster list with the *Meaningful* words list, the researcher generated four lists, one per cluster. These lists

contained the *Meaningful* words, and the manually determined and assigned categories and subcategories to which they belonged. Further, to identify the most prevalent topics, the lists were pared down to include only the top 100 *Meaningful* words in each cluster, when sorted by descending average term frequency. Non-*Meaningful* words that appeared in the top 100 words were removed from this investigation. Analysis for each of the four lists provided greater insight into each cluster, and this insight could be used to identify potential professional development topics in great detail.

Just as the clustering analysis yielded three distinct domains for professional development (Google Groups, hardware, and software, accounts, and technology services), analysis of the *Meaningful* words in the top 100 of each cluster provided superior insight into the suggested themes of each group. As shown in Table 5, Cluster 1, which contained help requests for hardware-related issues, contained frequently occurring tokens that were spread amongst several categories. The most commonly occurring category of *Meaningful* words in the top 100 tokens in Cluster 1 was the Hardware - Classroom A/V category. Indeed, ten of the nineteen *Meaningful* words in the list referred specifically to classroom audio/visual equipment. This was further proof that providing professional development on the usage of this hardware might prove extremely beneficial for the teachers, who often relied on these technologies to support lessons. Similarly, professional development that specifically addressed the audio/visual hardware in the classrooms might reduce the number of help requests on this subject sent to the IT office.

Table 5

Meaningful Tokens in the Top 100 of Cluster 1 – Hardware

Token	Avg. Term Freq.	Category - Subcategory	Rank in Cluster
projector	0.067062	Hardware - Classroom A/V	4
classroom	0.064270	Hardware - Classroom A/V	5
equip	0.056721	Hardware	7
audio	0.056271	Hardware - Classroom A/V	8
visual	0.052076	Hardware - Classroom A/V	9
audio_visual	0.047076	Hardware - Classroom A/V	11
classroom_equip	0.042807	Hardware - Classroom A/V	13
laptop	0.032673	Hardware	19
sound	0.030773	Hardware - Classroom A/V	21
mba	0.030565	Hardware - Mac	22
mac	0.029398	Hardware - Mac	24
network	0.021846	Network	36
speaker	0.019626	Hardware - Classroom A/V	44
monitor	0.014880	Hardware - Classroom A/V	61
phone	0.013409	Phone	69
video	0.012015	Hardware - Classroom A/V	79
chrome	0.011457	Software - Google Chrome	84
vdt	0.011322	Virtual Desktop	86
wifi	0.009873	Network	96

Numerous conclusions were drawn after analysis of the frequent *Meaningful* token list for Cluster 1. As tokens related to the MacBook Air computers (*mba* and *mac*) issued to all faculty were prevalent, professional development on the operation and troubleshooting of these devices could potentially benefit the teachers who used the devices, and the IT department who supported teachers in their usage. Similarly, the hardware that powered the phone system became apparent as a potential area of weakness and thus improvement. Further, proper usage and troubleshooting of the wired and wireless networks used in conjunction with both the classroom audio/visual systems and the MacBook Air computers seemed to warrant professional

development attention. A merge of the researcher-interpreted *Meaningful* words list, derived from RapidMiner's wordlist, and the RapidMiner-generated cluster list, clearly yielded detailed descriptions of hardware technologies that commonly caused teachers to reach out to the helpdesk. Professional development on these topics may contribute to improved use of these tools and a reduction in hardware-related support requests.

A similar analysis of the most prevalent meaningful tokens in Cluster 2, software, accounts, and technology services, yielded valuable and detailed information on potential professional development topics within this cluster. While both FirstClass and Virtual Desktop appeared prominently on the list, as shown in Table 6, the researcher, also the Director of Educational Technology at the site, knew that professional development on these topics was not a priority. FirstClass was replaced by Gmail during the study timeframe and was no longer in use, and teachers' usage of the Virtual Desktop platform was reduced dramatically since the introduction of Google Apps for Education two years prior. Thus, it appeared that the software, accounts, and technology services that most warranted attention through professional development were printing, password resets and usage of Okta (the single-sign-on tool), the Google Chrome web browser, and the Academic Web intranet page.

Table 6

Meaningful Tokens in the Top 100 of Cluster 2 – Software, Accounts, and Technology Services

Token	Avg. Term Freq.	Category - Subcategory	Rank in Cluster
print	0.204907	Printing	1
password	0.185354	Accounts - Password Reset	2
swi	0.099644	Software - FirstClass	4
okta	0.093202	Accounts - Okta (SSO)	5
virtual	0.086080	Virtual Desktop	6
vdt	0.082749	Virtual Desktop	8
virtual_desktop	0.074596	Virtual Desktop	9
printer	0.069921	Printing	10
mba	0.024048	Hardware - Mac	19
chrome	0.020737	Software - Google Chrome	24
phone	0.019354	Phone	26
copier	0.012361	Hardware - Copiers	48
network	0.010002	Network	58
telephon	0.009125	Phone	66
mac	0.006782	Hardware - Mac	89
academ_web	0.006486	Software - Academic Web	96
http_academ	0.006300	Software - Academic Web	100

During the study timeframe, a new, network-based printing platform was implemented, and the analysis of the prevalent, meaningful tokens in Cluster 2 indicated that teachers would likely benefit from additional training on the use and troubleshooting of this system. Similarly, Okta, the single-sign-on system intended to reduce the need for multiple passwords for various technical services in favor of using a single password was problematic for teachers. Thus, these elements of software, accounts, and technology services became a potential topic for professional development. Finally, the Google Chrome web browser was a prevalent token in Cluster 2, which was not surprising as the site required the use of this browser for nearly all web-based

activities. Proper use and troubleshooting of this critical software thus presented itself as a potentially valuable training topic. Once again, the frequently occurring contextually-relevant tokens from Cluster 2 provided great depth of content beyond the general theme of the group.

The prevalent, meaningful tokens in Cluster 3 were, as a result of the catchall nature of the cluster, less practical for highlighting detailed areas of professional development. However, these tokens, as shown in Table 7, still provided useful insight. Tokens related to the phone system were featured commonly, echoing their presence on the Cluster 1 and Cluster 2 lists. Similarly, network-related tokens were common, bolstering their potential importance. Finally, accounts and technology services terms were prevalent as well. Indeed, the tokens in Cluster 3 served to provide additional evidence that bolstered the need for professional development on topics devised from Clusters 1 and 2.

Table 7

Meaningful Tokens in the Top 100 of Cluster 3 – Miscellaneous

Token	Avg. Term Freq.	Category - Subcategory	Rank in Cluster
phone	0.089210	Phone	1
network	0.072082	Network	4
telephon	0.051603	Phone	8
swi	0.037487	Software - FirstClass	18
us_network	0.031049	Network	25
wifi	0.022663	Network	45
password	0.019207	Accounts - Password Reset	58
vdt	0.019030	Virtual Desktop	60
jack	0.017266	Network	66
wireless	0.014457	Network	84
copier	0.013423	Hardware - Copiers	95

Finally, the list of prevalent meaningful tokens in Cluster 4 was analyzed, and this analysis was straightforward. Nearly all contextually relevant tokens in this list, shown in Table 8, were directly related to the use, creation, and management of Google Groups.

Table 8

Meaningful Tokens in the Top 100 of Cluster 4 - Google Groups

Token	Avg. Term Freq.	Category - Subcategory	Rank in Cluster
group	0.493158	Google Groups	1
member	0.179757	Google Groups	3
forum	0.159037	Google Groups	4
group_googl	0.125306	Google Groups	6
http_group	0.104855	Google Groups	12
add_member	0.104487	Google Groups	13
com_forum	0.104278	Google Groups	14
member_group	0.055603	Google Groups	22
forum_forum	0.054630	Google Groups	23
direct_add	0.054617	Google Groups	24
set_manag	0.051827	Google Groups	27
group_manag	0.051358	Google Groups	28
forum_managememb	0.049778	Google Groups	29
managememb	0.049778	Google Groups	30
member_activ	0.049778	Google Groups	31
group_web	0.049464	Google Groups	32
manag_web	0.048796	Google Groups	35
member_add	0.048755	Google Groups	39
manag_group	0.048419	Google Groups	40
group_group	0.041595	Google Groups	42
googl_group	0.031721	Google Groups	43
chrome	0.007960	Software - Google Chrome	76
password	0.006822	Accounts - Password Reset	82

This finding was not surprising, given the predictable nature of the Google Groups tickets as a control group of support requests. Interestingly, both the Google Chrome web browser and a

need for resetting passwords were included in the list. The browser was strongly recommended for using Google Groups, and a password was required for accessing the service. Thus, the prevalent, meaningful tokens in Cluster 4 supported the need for professional development regarding accessing and using Google Groups and highlighted the value of a discussion on shifting the management of Google Groups from the IT department to individual users.

The predominant, *Meaningful* tokens from each cluster highlighted the potential professional development topics within each cluster. To review professional development topics from a wider perspective, the researcher undertook an analysis to determine the most prevalent *Meaningful* words across all clusters. This analysis yielded the subjects of the utmost importance for professional development.

Results: Summary of Meaningful Tokens in all Clusters

By summarizing the list of meaningful tokens per cluster, the researcher had determined that 54 *Meaningful* tokens had appeared in the top 100 tokens for at least one cluster, (see Appendix D). Analysis and summary of this list yielded extraordinary results, as five *Meaningful* words appeared in the top 100 tokens in three out of the four clusters, and six *Meaningful* tokens appeared in the top 100 words in two of the four groups, as shown in Table 9.

Table 9

Meaningful Tokens that Appeared in the Top 100 Tokens for Two or More Clusters

<i>Meaningful Token</i>	Count of Clusters where Token Appears in top 100	Clusters
chrome	3	(1) hardware, (2) software, accounts, and technology services, (4) Google Groups
network	3	(1) hardware, (2) software, accounts, and technology services, (3) miscellaneous
password	3	(2) software, accounts, and technology services, (3) miscellaneous, (4) Google Groups
phone	3	(1) hardware, (2) software, accounts, and technology services, (3) miscellaneous
vdt	3	(1) hardware, (2) software, accounts, and technology services, (3) miscellaneous
copier	2	(2) software, accounts, and technology services, (3) miscellaneous
mac	2	(1) hardware, (2) software, accounts, and technology services
mba	2	(1) hardware, (2) software, accounts, and technology services
swi	2	(2) software, accounts, and technology services, (3) miscellaneous
telephon	2	(2) software, accounts, and technology services, (3) miscellaneous
wifi	2	(1) hardware, (3) miscellaneous

The results from this analysis provided remarkable input into the potential professional development topics that might empower teachers and reduce reliance on the IT support office. The Google Chrome web browser, network issues, password issues, problems with the phone system, and the Virtual Desktop platform appeared as prevalent tokens in three out of the four clusters. The overwhelming prevalence of these tokens, each of which ranked highly in average term frequency for nearly all clusters, provided robust substantiation that these specific topics were problematic. The researcher anticipated that teachers might grow more adept at using or

troubleshooting these problematic technologies if professional development that unambiguously addressed these technologies was provided.

Each of the technologies found in three of the four clusters represented a potential area for training and improvement. Google Chrome was a prominent technology tool at the site, thought to be used daily by every teacher, and yet it was a common source of difficulty. Similarly, teachers were reporting many issues with the school network, and these problems could be addressed with training that included providing teachers with standard troubleshooting steps. While an attempt was made to implement a single-sign-on tool and provide teachers with easier password management, evidently password problems still plagued teachers. Professional development on school password policy and, perhaps more importantly, the user-initiated password change tool built into Okta (the single-sign-on tool), might reduce the prevalence of password related help requests. In the same vein, professional development on proper usage of the phone system, including adjusting settings and resolving common issues, might help teachers gain an understanding of this important technology. Finally, though Virtual Desktop usage had diminished in recent years, it still featured prominently in three out of the four clusters, and thus training on this platform might be beneficial.

While the most prevalent tokens appeared in three out of the four clusters, six tokens appeared amongst the top 100 tokens in two out of the four groups. Two of these tokens, pertaining to the wireless Wi-Fi network and the telephone system echoed the importance of professional development on the network and phone system as determined from a review of the tokens that appeared in three out of the four clusters. Further, while a token pertaining to SWIS or FirstClass, the previous email system, was prevalent in two of the four clusters, this tool was no longer in use at the research site, and thus did not warrant training. (The administration at the

study site might be interested in the discrepancy between the number of FirstClass-related help requests, and the number of Gmail and Google Drive requests, as the two tools replaced much of the functionality of FirstClass). The remaining tokens that appeared in two out of the four clusters warranted in-depth consideration as professional development topics.

Clearly, certain hardware tools, beyond the phone system and the network, were problematic for the teachers. Two tokens relating to the school-issued MacBook Air laptop computers appeared in two clusters. The prevalence of these tokens in more than one group serves as an indication that teachers might benefit from additional training on the usage and troubleshooting of these computers. Teachers are expected to use their school-issued computers daily, and yet it was apparent that issues with these computers were causing teachers to reach out to the IT support team for help. Professional development on proper use and troubleshooting of these computers might empower teachers and reduce strain on the IT department. Similarly, the school-wide printer and photocopier system, which was also expected to be utilized heavily by teachers, was problematic. Once again, professional development on the use of these machines might greatly benefit the teachers.

The compelling results generated from the merge of the cluster list and the *Meaningful* words list epitomized the usefulness of data mining methods applied to the HDDB. The *Meaningful* words list was a by-product of the data mining process, then fine-tuned and interpreted by the researcher. The clusters were produced by a robust algorithm that grouped thematically-linked help requests while also calculating a prevalence value for each token within each cluster. By combining the two data mining output files, the researcher was positioned to identify numerous contextually-important, evidence-backed topics for professional development. Amongst other prevalent issues, it seemed that the teachers at the research site would benefit

from professional development on the Google Chrome web browser, the wireless and wired networks, the school-issued computers, the phone system, and the password reset tools available to teachers. Perhaps with professional development, the number of tickets that required an IT support agent would reduce, as teachers may begin to troubleshoot or avoid the issues they encountered with these technologies. Pursuant analyses similarly investigated the interplay between support requests that required or did not require an agent for resolution.

Results: Help Request Data and Classification Labels

To glean value from the classification analysis, the classification output was combined with the raw HDDB dataset. Once the classification model was applied to the complete dataset and merged with the full helpdesk dataset, the resultant file contained all helpdesk requests and each request's algorithm-assigned label, REQ-AGENT or NOREQ-AGENT. This data file consisted of all 1,884 records, though the researcher divided the file into two smaller files composed of all the support requests that were assigned the REQ-AGENT label and all records that were allocated the NOREQ-AGENT label. These files, which contained the full text of each help request, were laborious to review, and an evaluation of data files at the individual record level was unsustainable for similar studies conducted upon larger datasets. Thus, this file was summarized to determine the total number of help requests that received each label. Of the 1,884 records, 422 (or 22.40%) were assigned the REQ-AGENT label. The remaining cases, 1,462 records (or 77.60%), were classified with the NOREQ-AGENT label.

While this summary did not provide input into topics or themes for professional development offerings, this data mining result did offer insight into certain aspects of training planning. Specifically, the percentage of records classified into each of the REQ-AGENT and NOREQ-AGENT labels provided a snapshot of the then-current ratio of support requests that

required and didn't require an agent. These figures could function as a bar or benchmark that could be revisited after the implementation of a professional development plan. The data mining revealed that a majority of the support requests, nearly 78%, could potentially be resolved by a well-trained user without the input of an IT agent. This figure could be recalculated after professional development was conducted, and the new value might reveal a reduction in cases that did not explicitly require the work of a support agent. This summary analysis provided crucial information about the current state of technology support requests, highlighting a near 3.5:1 ratio of tickets that did not require an agent to tickets that needed an agent. This ratio and the overall percentages could be recalculated periodically to track the faculty body's progress and, potentially, to highlight a reduction in support requests that a properly trained teacher could handle. A decrease in the ratio might be leveraged as a motivator and tracking tool for the increasing technology savviness of the faculty body. Further, a reduction in requests that did not require an agent might represent both empowerment for the teachers and a reduced support load for the IT department. To further examine the informative value of the classification analysis upon professional development, the classification results and the clustering results were merged.

Results: Clustering and Classification Combined

The list of helpdesk records with clusters was merged with the classification list, generating a single file that combined the results of the classification and clustering procedures. This master file was synthesized with a goal of determining the number of support tickets within each cluster that were resolvable with and without a support agent. Further, the master list was divided into eight subsets, each capable of providing detailed insight into a particular combination of support requests. These combinations included:

- Cases labeled as REQ-AGENT and Cluster 1 - hardware
- Cases labeled as REQ-AGENT and Cluster 2 - software, accounts, and technology services
- Cases labeled as REQ-AGENT and Cluster 3 - miscellaneous
- Cases labeled as REQ-AGENT and Cluster 4 - Google Groups
- Cases labeled as NOREQ-AGENT and Cluster 1 - hardware
- Cases labeled as NOREQ-AGENT and Cluster 2 - software, accounts, and technology services
- Cases labeled as NOREQ-AGENT and Cluster 3 - miscellaneous
- Cases labeled as NOREQ-AGENT and Cluster 4 - Google Groups

These lists provided clear, detailed samples of records with each combination of label and cluster. The content of these lists could be called upon if, for example, the researcher or site administration sought exemplars of hardware-oriented help requests that did not require an agent's intervention to resolve. However, reviewing the actual details of each record was a laborious task, and the researcher did not consider such a review as scalable to similar studies that utilized larger datasets. Thus, analytical efforts were focused on reducing the master file.

The combined classification-clustering file was summarized to reveal the total number of support requests in each combination of classification label and cluster. These results are presented in Table 10.

Table 10

Summary of Classification Label Assignments by Cluster

	REQ-AGENT	NOREQ-AGENT	Total
Cluster 1 - Hardware	100	543	643
Cluster 1 - Label % of Cluster Total	15.55%	84.45%	
Cluster 2 - Software, Accounts, Tech. Services	15	328	343
Cluster 2 - Label % of Cluster Total	4.37%	95.63%	
Cluster 3 - Miscellaneous	163	538	701
Cluster 3 - Label % of Cluster Total	23.25%	76.75%	
Cluster 4 - Google Groups	144	53	197
Cluster 4 - Label % of Cluster Total	73.10%	26.90%	
Total	422	1,462	1,884

The summary by classification label and cluster provided valuable insight into the divisions of support tickets that required an agent or did not require an agent within each group. In Cluster 1, hardware, approximately 84% of the tickets were classified as not requiring an agent. In Cluster 2, software, accounts, and technology services, nearly 96% of requests were categorized as not requiring an agent. In Cluster 3, miscellaneous support requests, approximately 77% were classified as not requiring an agent. Cluster 4, Google Groups, yielded contrary results, with approximately 73% of support requests classified as requiring an agent.

These results provided valuable information that could contribute to the development of a professional development model. For example, nearly 96% of all Cluster 2 cases (software, accounts, and technology services such as printing, password resets, and the Google Chrome web browser) were potentially solvable by a properly trained teacher. Thus, a professional development model born out of data mining of this HDDB might dictate the emphasis of initial efforts on the technologies associated with Cluster 2-type issues. Cluster 2 issues that did not require an agent for resolution represented approximately 17.4% (or 328) of all 1,884 records. Thus, professional development on Cluster 2 technology problems might contribute a comparable reduction in support requests. Similarly, a professional development model might be developed to address the approximately 84% of Cluster 1-type issues that could be handled by an appropriately skilled teacher. Cluster 1 (hardware, including MacBook Air computers and classroom audio/visual systems) requests that did not require an agent for resolution represented 28.82% (or 543) of the 1,884 support requests. Professional development on Cluster 1 technologies could contribute an analogous reduction in support requests. Thus, while this summary did not provide additional understanding of specific professional development topics,

the summary provided useful insight into which professional development topics might have the biggest impact on reducing support requests that do not require an agent to resolve.

On the contrary, the 53 Cluster 4 support requests that did not require an agent represented a mere 2.81% of all support requests. Thus, if professional development on Google Groups, the theme of Cluster 4, was provided, it was likely that only a marginal reduction in support requests would occur. The high percentage of Cluster 4 cases that required an agent (approximately 73%) was not surprising to the researcher. Under the current IT helpdesk configuration, only IT agents are capable of creating and configuring Google Groups. Thus, support requesters often required an IT agent for many Google Groups-related tasks. However, nearly all Cluster 4 cases, 144 that needed an agent and 53 that did not, representing just over 10% of all support requests, might be removed from the helpdesk processes if the site were to change its policy on creating and managing Google Groups. If such a decision were made and professional development was offered to teachers for proper use of Google Groups, the site might experience a reduction of up to 10% of its support requests. This finding, an important topic of discussion for the site administrators and a potentially valuable area of professional development would not have become apparent without the combined outputs of the clustering and classification text data mining analyses.

Conclusion

Text data mining methods applied to teacher-submitted technology help requests successfully identified areas of technological weakness amongst the faculty members. Human interpretation of the text data mining outcomes provided additional depth and detail. The three data mining procedures (the development of the wordlist, and the creation and application of a clustering and classification algorithm) not only yielded revealing information as standalone

endeavors, but also complemented or supplemented each other's results. The results of all data mining efforts could collectively guide the creation of a professional development model.

Analysis of the wordlist revealed prominent problematic topics and topic categories identified throughout the dataset. The wordlist analysis also provided general insight into the relative magnitude of the topic categories. The clustering analysis identified four prevalent topic areas into which all helpdesk support requests were categorized and again offered an approach by which the categories could be compared or ranked. Further, the clustering analysis highlighted the potential advantages of having a control group of support request tickets, an unexpected finding. When the clustering results were merged with results from the wordlist analysis, many prominent problematic technologies within each cluster were identified. The merged clustering-wordlist investigation also yielded a list of the most prevalent identifiable problematic technologies across all teacher-submitted support requests. The classification procedure highlighted the relative occurrences of support requests that required an agent to resolve and support requests that did not necessarily require an agent to resolve. Finally, the merged classification-clustering results provided useful information regarding the division of tickets that required and didn't require support agents within each of the four cluster categories.

The results of the data mining procedures contributed indispensable evidence upon which a technological professional development plan for teachers could be constructed. The site's teachers would benefit from training on hardware, such as MacBook Air computers, the classroom audio/visual systems, and the wired and wireless networks. Hardware-related professional development might generate a decrease of nearly 29% of support requests. Further, training on software, accounts, and technology services such as printing, password resets, and usage of Okta (the single-sign-on tool), the Google Chrome web browser, and the Academic

Web intranet page might generate a decrease of up to 17% of help requests. Finally, the data mining of the helpdesk database confirmed that professional development on Google Groups, and a change in Google Groups-related policy, might yield a 10% decrease in support requests.

CHAPTER 5

CONCLUSIONS, RECOMMENDATIONS, AND IMPLICATIONS

This study generated two types of findings: those that were relevant to the research site, and those that were pertinent to scholar-practitioners beyond the site. While discoveries about technological weaknesses and potential professional development opportunities at the site represented prospects for site leaders, the capabilities of HDDB data mining represented the potential for powerful knowledge gathering and change endeavors at innumerable educational institutions. At the most fundamental level, the ultimate conclusions from this study were that HDDB data mining could inform training undertakings and successfully highlight trends in technology weaknesses for users who have submitted support requests. Similarly, the foremost recommendation was that other researchers or practitioners should attempt HDDB data mining studies at their sites. Finally, the principal implication of this study was the correlation between identifying technology skill weaknesses and improving educational technology integration. If technology weaknesses amongst teachers were impacting educational technology integration, (and research showed a relationship between skillset and integration efforts), and if HDDB data mining can highlight skillset weaknesses, professional development based on the results of HDDB data mining could lead to improved educational technology undertakings. This chapter will interpret the findings of the research, and present conclusions, recommendations, and implications of the case study.

Interpretation of Findings

This single-site instrumental case study was guided and informed by three research questions. The unifying theme of the research questions was an inquiry into the value of data mining methods applied to an independent boarding high school's helpdesk database. The

study's purpose was to determine the knowledge that this data mining might reveal with regards to teachers' technological skills or lack thereof. Professional development might be devised based on the identification of the skills that teachers lacked and technology problems that could have been a result of these shortcomings. These research questions provided direction in support of the purpose of the study:

1. In what ways could data mining be leveraged to best extract the desired information from the HDDB?
2. What does data mining of the HDDB reveal about gaps in teachers' technology skill sets?
3. How could data mining of the HDDB be used to determine topics and plan technology-related professional development for teachers?

To best interpret the findings of the study, these research questions will be reviewed in detail, supported and bolstered from the results introduced in Chapter 4.

Question 1

The first research question sought to determine the data mining methods by which teachers' technology weaknesses and potential professional development topics might be extricated from an HDDB. At the heart of this question was an investigation into data mining practices that could be applied to a dataset to reveal themes and trends of a faculty body's technology skill weaknesses. The three data mining procedures (the wordlist creation and analyses, and the creation and application of clustering and classification algorithms) yielded different contributions towards the goal of determining areas of professional development for teachers. Similarly, analyses that merged the results produced by the data mining methods provided valuable, supporting insight.

Creation and Analysis of the Wordlist. The wordlist analysis was proficient at highlighting general themes amongst the help requests submitted by teachers during the study's timeframe. The goal of the wordlist and wordlist analyses, aside from being an elemental component of other data mining analyses, was to minimize the total number of tokens considered in more sophisticated data mining procedures. In this vein, the wordlist creation and analyses were reductive in nature. The stemming (reducing words to a shared root word) and pruning (removing infrequently occurring words) preprocessing steps were intended to reduce the number of words leveraged in the analyses. The final wordlist contained 469 unique tokens. This list was eventually pared down to 69 *Meaningful* words, all of which provided insight into the topics and themes of the dataset. Only nine of the 69 *Meaningful* words appeared amongst the top 10% of tokens that occurred in distinct helpdesk records. The nine *Meaningful* tokens that appeared in 10% or more of all helpdesk records, shown in Table 3, provided the first firm indication of the areas of technological weaknesses of the teachers who had submitted support requests to the helpdesk.

The nine most frequently occurring *Meaningful* tokens highlighted numerous potential professional development topics. These topics included: training on the campus-wide phone system, network access and usage, password resets and management, printing, the custom Virtual Desktop platform, and the hardware that powered the classroom audio/visual systems. Numerous critical areas of teachers' technological weaknesses became apparent after completion of the first data mining procedure and straightforward supplemental work performed by the researcher as a human interpreter of the data. Had the wordlist analysis ceased at this point, the study would still have succeeded in highlighting potential areas for improvement amongst the faculty body that were hitherto unknown.

While analysis of the nine most frequently occurring tokens provided a watershed moment in leveraging data mining procedures for topic discovery within an HDDB, many important and revealing analyses followed. Continuing the trend of reducing the HDDB towards smaller subsets, a manual review of the 69 *Meaningful* tokens yielded twenty categories and subcategories into which the *Meaningful* tokens were distributed. The top-level categories of weakness as highlighted by the HDDB included account management, Gmail, Google Drive, Google Groups, hardware, the wired and wireless network, the phone and printing systems, various softwares, and the Virtual Desktop platform, (see Table 4). These categories provided valuable differentiation amongst professional development topics, presenting specific subject matter groupings for potential teacher training endeavors. These findings suggested that Google Groups and hardware issues might be the categories most in need of professional development attention.

Thus, the generation of the wordlist within RapidMiner and careful summarization by the researcher yielded professional development topic areas, as well as a potential indicator of importance. The contributions from the wordlist are particularly noteworthy since the wordlist was a precursor to other data mining analyses, not a terminal data mining objective. The wordlist would continue to contribute to the study's goals in later analyses when it was merged with results from the clustering procedure.

Creation and Application of a Clustering Model. The clustering procedure within the RapidMiner data mining software was developed to divide support tickets into distinct groups. After review of the four clusters, the themes of each cluster were identified. The clusters included tickets pertaining to (1) hardware, (2) software, accounts, and technology services, (3) Google Groups, and (4) miscellaneous requests that could not be categorized accurately. These

themes echoed the topics derived from the *Meaningful* tokens but provided valuable information regarding the division of all helpdesk tickets into each cluster. The hardware cluster contained approximately 34% of all support requests, while the software, accounts, and technology services cluster contained approximately 18% of all help requests. The Google Groups cluster comprised approximately 10% of all incoming help requests, and the catchall group included approximately 37% of all help requests. Thus, the clustering procedure provided clear, evidence-based indicators of which technologies were most problematic and where professional development attention should focus. For example, professional development that pertained to the hardware used by teachers at the research site might lead to a reduction in nearly a third of all help requests. From this data mining procedure, technology topics that contributed to approximately 63% of all support requests could be identified and addressed in professional development.

The data generated from the clustering analysis might help dictate the sequence in which topics were covered in professional development. Once aware of the distinct categories of technology problems faced by a site's users, and occurrence frequencies of each problem type, site administrators could determine when to address each category. For example, the hardware cluster, representing approximately 34% of all help requests, might warrant immediate attention, due to its large volume. However, Google Groups problems might warrant attention at the start of the school year, when these issues were (anecdotally) observed to occur most frequently. The categorical groupings derived from the clustering analysis provided insight into the content of the HDDB and assisted in creating training strategies.

The category assigned during the clustering procedure was merged back to the original helpdesk data. While the study proceeded in a direction away from reviewing individual records, spot-checking of the data yielded valuable and useful information. First and foremost, spot-

checking produced the concept of using a control group of tickets, to test for clustering algorithm accuracy and also to avoid unnecessarily reducing cluster accuracy. Secondly, the merged helpdesk dataset with cluster created new data that did not exist before the clustering data mining procedure. This new data was a detailed list of support requests with a problem category applied. If a researcher or practitioner were interested in reviewing detailed information about distinct groups of support tickets, that information was now easily and readily available to them. Such a task might be time-consuming, but if the chore were necessary for a site's or project's goals, the merging of the clusters onto the detailed help request list provided information which had not been previously available.

Continuing the trend that highlighted the usefulness of merging data mining results, the merged cluster list and *Meaningful* tokens list yielded results that were amongst the most compelling examples of HDDB topic detection from data mining methods. By merging these lists, the most common *Meaningful* tokens within each cluster became apparent. Thus, while the clustering analysis had unveiled the importance of addressing hardware issues, the merged cluster and *Meaningful* list highlighted many specific problematic hardware technologies and their rank within the group. For example, the merged list highlighted classroom technologies, such as the projector and audio/visual equipment, as amongst the most problematic hardware technologies in the cluster that generated the most problems, (see Tables 5, 6, 7, 8). Indeed, the merged cluster and *Meaningful* token list highlighted the minutia within each cluster, providing a level of detail and insight into the themes of the HDDB that had not been previously achievable.

Owing to the level of detail generated from the merge of the cluster and *Meaningful* token list, a leader creating a professional development plan would be armed with valuable information regarding the sources of technology problems. Hardware-related problems, which

accounted for approximately 34% of all help requests, were largely the result of issues with common technologies including audio/visual systems, the computers used by teachers, and the wired and wireless network. Because of the importance of these technologies in everyday teaching, and the potential disruptions to teaching when these technologies fail, the value of addressing these technologies via professional development was made clear.

A similar trend was observed when reviewing the details of the software, accounts, and technology services cluster. Many of the problematic technologies in this cluster were amongst the most fundamental technologies in use at the school. Common tasks such as printing and copying, changing and managing passwords, and using the custom Virtual Desktop tool and intranet site were found to be contributors to help requests within this category. As before, there appeared to be a trend linking the most common technologies with the largest number of support requests. This trend was also noted in cluster 4, Google Groups (another important and relied-upon tool at the site). Whether the link between common technologies and an increased number of support tickets was based on increased problems with the technologies, or just increased usage (and therefore more opportunities for problems) might be determined in a subsequent study, but the importance of professional development on these topics was irrefutable.

An Important Analysis and Finding. Perhaps the most noteworthy analysis born out of the data mining of the HDDB was the summary of clusters with *Meaningful* tokens. As shown in Table 9, five tokens appeared in the top 100 most frequently occurring tokens in three out of the four clusters. Similarly, six additional tokens appeared in two groups. Thus, these problematic technologies were identified as being the most troublesome technologies across the entire dataset. The most prevalent clearly identifiable technologies included the Google Chrome web browser, network and password issues, and the phone and Virtual Desktop systems. Nearly as

prevalent were problems with the photocopier system, teachers' MacBook Air computers, a legacy email platform, the telephone and wireless network systems. The clustering data mining procedure merged with the distilled and interpreted wordlist, thus precisely identified many of the technologies that were causing disruptions for teachers at the research site. In a case study that had topic identification as a central goal, the names and occurrences of specific problematic technologies was a spectacular find.

Creation and Application of a Classification Model. While the wordlist and clustering analyses yielded information that pertained to the topics hidden within the HDDB, the classification data mining procedure shed light on the nature of these topics. The primary purpose of the classification analysis was to determine the percentages or ratio of support requests that required an IT agent for resolution compared to those requests that were resolvable without an agent. Thus, the classification procedure, when viewed through a lens of topic discovery methods, offered less value than the clustering and wordlist investigations. The classification analysis indicated that nearly 78% of all support requests could be addressed by a properly trained teacher without the aid of an IT agent. This figure could be used as a benchmark to determine the impact of professional development.

Analysis of the merged classification and clustering results provided additional insight into areas of focus for professional development efforts. Per Table 10, approximately 84% of hardware-related support requests could potentially be addressed without an agent and thus would be valuable content for professional development. Similarly, over 95% of support requests regarding software, accounts, and technology services could be addressed through professional development. As these two categories of tickets represented the majority of requests within the identifiable clusters (non-miscellaneous) of support requests, this analysis yielded valuable

information on the potentially most impactful areas for professional development. This information, combined with the earlier identification of relevant topics to cover in training, might contribute to the creation of a professional development plan.

Clearly, data mining procedures including the development and analysis of a wordlist, and the creation and application of a clustering model could be leveraged to extract potential professional development topics from an HDDB. A classification procedure could complement and inform decision making as well. Combined analyses of all three data mining procedures can further highlight topics, trends, and areas of concern within the dataset.

Question 2

The application of data mining procedures to the support requests submitted by teachers at the research site from September 2013 through September 2016 revealed valuable information about the gaps in teachers' technology skill sets. Many of the topics and themes derived from data mining the HDDB pertained to technologies that the teachers were expected to use on a daily basis. The data mining highlighted that teachers experienced problems with technologies for which administrators had anticipated an adequate level of competency. While site administrators including the researcher in his role of Director of Educational Technology implemented and trained upon technologies such as new audio/visual hardware in classrooms, MacBook Air computers, and Google Groups for mass communication, teachers often lacked the necessary skills to use these technologies competently without encountering problems and work stoppages. Similarly, teachers may have lacked the necessary skills to troubleshoot problems with these technologies when issues arose.

An important gap in teachers' technology skills sets identified during this case study was insufficient expertise with the Google Chrome web browser. Chrome was an essential tool for

nearly all web-based tasks performed at the school and was the IT department's recommendation for browsing the Internet. Proper operation of the browser was vital for correct usage of web-based resources. Further, as Gmail and Google Drive are both web-based platforms, Google Chrome was the site's recommended conduit for accessing these email, communication, and document creation tools. Similarly, Chrome was the recommended web browser for accessing FA Web, the school's online gradebook, and the Academic Web intranet site, as well as the school's library databases and other resources. Indeed, of all technological tools in use at the research site, Google Chrome was amongst the most important ("C.L.", Dean of Faculty, personal communication, February 25, 2017).

Google Chrome appeared as a problematic technology in the top 100 tokens for three out of the four clusters and was identified as a serious gap in teacher skillset. Further, in the software, accounts, and technology services cluster, all but two of the tokens that appeared more frequently than the *chrome* token were software, accounts, and technology services that required or leveraged Google Chrome for proper use, (see Table 6). These included problematic technologies such as printing, password resets, and the single-sign-on tool, all of which were performed or accessed using Google Chrome. Also included was the Virtual Desktop platform, which was launched via Google Chrome and also contained a version of Chrome within the platform itself. Indeed, many of the problematic technologies in the software, accounts, and technology services cluster necessitated or relied upon Google Chrome and issues with Chrome might have been a contributor to issues with other technologies. Data mining of the HDDB thus identified a gap in teachers' technology skillsets by highlighting the frequency and breadth of issues experienced with a critical technology tool.

As was the case with Google Chrome, the data mining procedures and analyses revealed extensive usage issues with the school-wide wired and wireless networks. Indeed, the token *network* appeared in three of the four clusters, and the token *wifi* appeared in two of the three clusters, (see Table 9). Further, of all the most prevalent *Meaningful* words that appeared in two or more clusters, only *telephone* and *phone* did not rely upon or require proper use of the site's data network. Technologies such as browsing the web, printing, resetting passwords, and effectually using a MacBook Air computer all required a stable connection to the school's network. Other identified problematic technologies, such as the Virtual Desktop, classroom audio/visual hardware, and Google Groups also necessitated a connection to the school's network. The prevalence of network-related support requests highlighted a potential gap in teachers' skillsets with regards to using or troubleshooting this important technological component at the research site.

Similarly, the frequency of tokens related to the teachers' usage of their school-issued MacBook Air computers highlighted potential gaps in teachers' knowledge in the operation of these machines. Two distinct *Meaningful* tokens related to these computers appeared amongst the top 100 tokens in two clusters, (see Table 9). As with Google Chrome and the school networks, the MacBook Air computers were all but essential for the successful use of educational technology at the research site. These machines were school-owned and issued to members of the faculty as the standard work computer. Indeed, the audio/visual hardware configuration in each of the school's classrooms was designed, selected, and built specifically to ensure compatibility and full functionality with MacBook Air computers. The prevalence of support requests and tokens which insinuated that the MacBook Airs were problematic for the

faculty highlighted a potential gap in the skillset required to operate this technology on a regular basis adequately.

An additional gap in teachers' skillsets pertaining to technology hardware was revealed from the combined *Meaningful* token list and cluster merge. While MacBook Air related tokens appeared prominently in more clusters, classroom audio/visual hardware tokens were ranked higher than MacBook Air tokens in the hardware cluster, (See Table 5). Indeed, eight out of the top ten most frequently occurring *Meaningful* tokens in the hardware cluster, including the top seven, were directly related to the audio/visual systems in the classrooms. (The other two tokens in the top ten were related to the MacBook Air computers). Further, of the nineteen *Meaningful* tokens that appeared in top 100 words by average term frequency in the hardware cluster, ten pertained directly to audio/visual equipment, by far the largest category. While audio/visual tokens only appeared prominently in one cluster, (likely owing to their precise clustering into the hardware group), their thematic dominance in the most problematic non-miscellaneous cluster highlighted their overall importance. Clearly, teachers were encountering issues with the classroom audio/visual technology, and the HDDB data mining revealed this gap.

By considering the combined results of the data mining procedures applied to the support requests submitted by teachers and collected in the HDDB, an overarching gap within teachers' technological skillsets became apparent. The four prominent technologies for which skill gaps became apparent, Google Chrome, use of the network, the classroom audio/visual hardware, and the MacBook Air computers, were all required for technology-bolstered education at the site. The MacBook Air computers were the hardware that teachers used for technology-enhanced instruction, and the Google Chrome web browser was the most important software to be used with the computers. The audio/visual system was the means by which teachers presented digital

content from their computers to their students. The network provided the path by which the hardware-software combination accessed all onsite resources, such as the printers and audio/visual hardware, and off-site resources, including email and the Internet, and thus all four were interconnected and crucial. The four technologies supported and required one another, and thus the four as a collective represented the core of technological use by teachers at the site. Data mining revealed that these four technologies, individually and in concert, were problematic for teachers. However, given that many support requests related to these technologies were solvable without an IT agent, (nearly 85% for hardware including the audio/visual system, computers, and the network, and nearly 96% for software, accounts, and technology services), the mining procedures also revealed that teachers' skillsets were lacking for these essential technologies. Clearly, data mining of the HDDB had the potential to identify and diagnose gaps in teachers' technology skillsets.

Question 3

Data mining of an HDDB could be used not only to determine topics but also to plan technology-related professional development for teachers. As has been suggested, certain data mining procedures provided insight into the relative importance of topics gleaned from the HDDB, and thus provided a means for prioritizing topics during professional development plan. For example, the visualization in Figure 1 of the frequency of *Meaningful* words in help requests highlighted that the Google Groups and hardware categories had tokens that appeared very frequently in support requests and thus warranted professional development. Network, software, accounts, phone, and the Virtual Desktop platform also had reasonably high numbers of frequently appearing *Meaningful* words. From this analysis alone, a professional development

planner might determine that training efforts should focus on Google Groups and problematic hardware.

The data mining procedures and analyses, however, were most revealing when they were compared with one another rather than standing alone. The clustering analysis divided all support requests into four categories, in which Google Groups, hardware, and software issues were represented. However, this analysis highlighted that hardware issues (34% of all support requests) represented many more support requests than Google Groups (10%). The results of this analysis align with the previous analysis in stressing the importance of professional development on hardware issues while diminishing the necessity of Google Groups training. Further, while the *Meaningful* words analysis underscored the importance of hardware training as a whole, the clustering analysis provided valuable insight into which specific hardware technologies warranted training. The general hardware theme was thus clarified further, emphasizing the importance of training on the MacBook Air computers, the classroom audio/visual systems, and the wired and wireless networks. Similarly, the *Meaningful* words analysis called attention to the particular software systems that necessitated a high priority in professional development. These included the Google Chrome web browser, the password management and reset platform, and the printer/copier system. As the software, accounts, and technology services category accounted for more support requests (18%) than did Google Groups, once again a priority or sequence of training topics became apparent as a result of merging data mining results.

Continuing the trend of adding value to data mining procedures by merging, combining, and comparing results, summarizing the *Meaningful* words list and identifying the number of clusters in which each *Meaningful* word appeared within the top 100 tokens provided significant insight into professional development planning. Numerous text data mining procedures have thus

far highlighted the importance of hardware, software, and accounts as potential professional development topics. The count of clusters containing *Meaningful* words amongst each cluster's top 100, considered amongst the most important procedures, further supported the findings by noting the particular software (Google Chrome) and hardware (MacBook Air) computers that required the most attention.

Finally, the summaries derived by merging the clustering and classification analyses provided additional insight into the development of a professional development plan. The understandings gleaned from prior analyses were complemented with new information about the potential gains ascribed to each professional development topic. While Google Chrome was highlighted as a central element of a professional development plan and was deemed a prevalent token in three clusters, the software cluster overall generated far fewer support requests (18.2% of all requests) than did the hardware cluster (34.1%), (see Table 10). Further, many hardware-related tickets, an estimated 29% of all requests, could potentially be solved without the input of an IT agent. Numerous software requests, an estimated 17% of all requests, could be solved without an IT agent. An overall goal of the study was to use HDDB data mining to determine an effective professional development model, and this combined analysis highlighted that professional development efforts should be directed towards hardware problems, specifically on the MacBook Air computers, the data network, and in-classroom audio/visual technology. Google Chrome, as a prevalent theme within a less problematic cluster, warranted secondary professional development attention.

Thus, data mining of the HDDB was used to determine the topics for professional development as well as the relative importance of each topic. These characteristics may be essential elements of a well thought out professional development plan. Further, the development

of the training plan might also benefit from a perusal of the complete helpdesk dataset, merged to include both a cluster and a REQ-AGENT/NOREQ-AGENT label. While reviewing data at the individual record level is time-consuming, the task might be fruitful for developers of professional development. For example, by considering both the cluster and label of help requests, trainers can identify site-specific, original, and highly relevant examples to leverage during the presentation of professional development. These examples, pulled directly from the database of help requests submitted by the very people participating in professional development, might help to situate and frame the relevance of the professional development. Finally, if a third party provided the professional development, genuine examples of clustered and classified helpdesk records might assist the third-party trainers to customize their content.

Implications of the Study

The results of this case study have implications for many potential beneficiaries and stakeholders. The purpose of this instrumental, single-site case study was to determine the usefulness of applying text data mining procedures to an HDDB that contained technical support requests from teachers. The particular findings pertaining to problematic technologies at the research site were meaningful and relevant for the site's employees and administrators, but to interested parties beyond the study site, the most remarkable result of data mining applied to an HDDB was the efficacious capability of the methodology to highlight areas for professional development. This result underlines the case study's purpose as a proof of concept regarding the value of text data mining a technical support request dataset. Indeed, data mining of an institution's HDDB has the potential to inform change endeavors that can positively impact the lives of many people affiliated with an organization. From individuals to communities, HDDB data mining to determine professional development topics for teachers has far-flung implications.

The consequences of this case study were divided into two categories: Implications for Practice and Implications for Leadership.

Implications for Practice

Data mining of the HDDB was capable of highlighting areas of technological weakness for a faculty body, as well as potential professional development topics. This finding has implications for those involved in the design, use, and practice of educational technologies.

These beneficiaries include those responsible for technology integrations at schools, the teachers who use the technology and seek help when problems arose, the technology support team who handles the issues, and the institutions' administrations.

The researcher, in his role as Director of Educational Technology, was initially concerned that teachers at the research site lacked the technical skills to operate the technologies they were expected to use regularly. This concern prompted the investigation that was eventually manifested by the case study. As the Director of Educational Technology at the research site, and tasked with helping teachers to bolster their lessons with the use of technology, the researcher had hoped to confirm the notion that teachers lacked technological skills, and to highlight which technological skills, in particular, were problematic. The case study served to confirm the researcher's beliefs that teachers lacked necessary core competencies. Scholar-practitioners can use HDDB mining not only to identify topics for professional development but also to confirm or disprove their opinions or beliefs about the status of technology usage and problems at their institution. In addition to confirming beliefs about teachers' technology skillsets, the results of an HDDB data mining undertaking can potentially influence professional development models and existing practices of the determination of training topics.

Scholar-practitioners and on-site educational technology experts employed at educational institutions may leverage assessment instruments to identify the technological weakness of a group of teachers. These instruments were occasionally generated by combining other instruments (Hancock, Knezek, & Christensen, 2007) or developed from scratch (Conrad & Munro, 2008). Some instruments were designed specifically to assess technology undertakings at educational institutions (Davies, 2011), while others were general purpose evaluative tools that measured technology usage (Rosen et al., 2013). Unfortunately, technology assessment instruments were considered unreliable measurement tools or weakness detectors. Numerous scholars, including Kopcha and Sullivan (2007) and Reinhart, Thomas, and Toriskie (2011) determined that such assessment instruments were often inaccurate or unreliable. Further, Maderick (2013) and Maderick, Zhang, Hartley, and Marchand (2015) concluded that when assessments were the sole measure of technological capability, the assessment results were often imprecise.

As helpdesk database data is collected organically, *in situ*, and within the context of the site, the contents of the database represent genuine problems with the specific technologies in use at the site. This data is potentially free from biases that might affect assessment instruments and is a clear and accurate portrayal of technology problems experienced by the technology users at a particular site. A prominent implication of this case study is that data mining of the HDDB may be a new and accurate method to determine teachers' technological weaknesses. This new method may replace or complement existing assessment methods, and thus educational technology leaders informed by HDDB data mining evidence may be better positioned to provide relevant and valuable professional development for teachers.

Teachers at schools that undergo HDDDB data mining projects stand to benefit in their teaching and organization practices that leverage the use of technology. Numerous scholars highlighted a lack of technology skills as a barrier to educational technology integration (Inan & Lowther, 2010b; Ertmer et al., 2012; Pilgrim & Berry, 2014). Other researchers noted that a lack of technological skills might affect a teacher's confidence and self-efficacy and thus negatively impact their pedagogic use of technology (Conrad & Munro, 2008). Professional development on topics gleaned from the HDDDB, which itself was populated by the technology help requests from teachers, might be leveraged to address areas of weakness in teachers' technology skill sets. Thus, HDDDB data mining for professional development topic detection, following by properly managed and designed professional development, might very well address gaps in teachers' technology skillsets and improve their technology-supported instruction.

Improvements to teachers' technological skillsets might impact teachers' lives in realms beyond the classroom. Owing to the pervasiveness of technology in contemporary North American life, if teachers gain new skills and competencies with technology in their workplace, they can apply these new capabilities to projects and undertakings outside of their professional life. Indeed, as teachers learn about the same technology issues that have caused them enough trouble to warrant a call to the IT helpdesk, teachers may experience a sense of empowerment and newfound confidence when using technology. Thus, the identification of problematic technologies and professional development intended to address these technologies could affect teachers in many facets of their lives.

As teachers develop greater technology skillsets, equally applicable at work and home, the communities in which they reside may experience positive growth as well. This trend may be especially true of teachers at a boarding school, who share living, working, and recreational

spaces, but the potential for community improvement goes beyond boarding institutions.

Professional development may help teachers to master technologies or troubleshoot issues when they arise, and teachers can also help others learn and understand proper technology usage.

Technology-savvy teachers are capable of supporting each other with technology problems, benefitting both the individuals and their personal and professional affiliates.

Just as teachers might experience numerous benefits as a result of an HDDB data mining project at their school, the implications of this case study highlighted that institutions themselves stand to benefit from such an undertaking. Organizations, leaders, and policymakers typically invest significant time, energy, and financial resources into educational technology projects, such as classroom audio/visual systems and computers distributed to teachers. If usage of these technologies was problematic, knowledge of that fact might be critical to school administrators. Indeed, leaders also stand to benefit if they learn that the selected technologies aren't particularly problematic.

Data mining of an HDDB can reveal the current status of technology usage by a particular group of users within an organization, and this information may be paramount for decision makers. Aside from implementing a professional development platform to assist teachers in overcoming weaknesses, a data mining undertaking might inform operational decisions. These decisions may concern new hardware and software technology purchases, hiring practices in information technology and educational technology offices, and the efficacy of current institutional technology support systems.

Regarding institutional technology support, a successful HDDB data mining case study implies robust potential for improvement in information technology helpdesk offices. IT departments may see a reduction in support requests as teachers receive professional

development on the most problematic technologies. As teachers become empowered to avoid or troubleshoot technology problems, the IT support office may be positioned to allocate resources to other projects, as the load on the helpdesk decreases. The potential for time, energy, and financial savings is palpable, as are the opportunities to apply technology human resources to new undertakings.

Finally, the successful data mining of a helpdesk database for topic detection has important implications for the field of helpdesk data mining. While few contemporary educational data mining studies leveraged the HDDDB for analysis, this data source had a lengthy history of analysis within the corporate realm (Jha & Hui, 2000; Blaaffladt, Johansen, Eide, & Sandnes, 2004; Forman, Kirshenbaum, & Suermondt, 2006; Maron & Zukerman, 2009; Andrews & Lucente, 2014; Andrews, Beaver, & Lucente, 2016). Nearly all of the studies on helpdesk data in an enterprise environment attempted to improve information technology helpdesk practices and procedures. The success of this case study implies that data mining of the HDDDB could be used for purposes beyond investigating helpdesk operations. In the educational realm, the success of the case study denotes the value of such studies at academic institutions. In the corporate sphere, this case study highlights the potential to improve the technology skills of the employees requesting support from their internal helpdesks.

Knowledge gleaned from HDDDB mining can guide decision-making efforts that might affect many aspects of an academic institution's operations. Analysis and summarization of data mining outputs can inform discussions on a variety of topics, including professional development planning, technology policies, and hardware purchases. While this research suggested implications for changes in practice, the study, which championed data-informed transformation efforts, also offers unique implications for key decision makers.

Implications for Leadership

The success of this HDDB data mining study has implications beyond technology and professional development. Specifically, the methods employed in this case study could be used to provide valuable guidance for institutional leaders tasked with affecting positive change at their organization. Many of the case study's most valuable implications relate to leadership improvements as experienced through a lens of transformative leadership. Transformative leadership is a leadership theory that unequivocally unites institutional change undertakings with societal advancement (Shields, 2010).

Transformative leadership attempts to address issues of equality and inequality, at the individual, organizational, and societal levels (Shields, 2010). Indeed, the principal tenets of transformative leadership include addressing both critiques of the status quo and the promise or potential of a changed environment, undertaking efforts to affect profound and equitable change, emphasis on both individual achievement and the public good, and a voluntary acknowledgment of power (Shields). The case study has implications for all of these leadership concerns.

Shields (2010) described the process by which transformative leadership efforts might affect change. Transformative leadership, she wrote, "recognizes the need to begin with critical reflection and analysis and to move through enlightened understanding to action," (p. 572). These actions were intended to emphasize equality. Shields' brief summary of the mindset behind a transformative leadership project underlines the ideology behind an HDDB data mining undertaking. Data mining procedures, such as the creation of a wordlist, and the creation and application of clustering and classification models, provide an opportunity to reflect upon the status quo of a group of users' technological weaknesses. The data yielded from the mining endeavors can be analyzed and contextualized to provide greater insight and understanding of the

target group. With a greater comprehension of both the present status and advantages of potential changes, a leader equipped with data gleaned from an HDDB data mining project can lead professional development supported change that can potentially improve the working lives of all faculty members.

By data mining the HDDB with an eye on highlighting areas of weakness, a transformative leader can evaluate a group of users and determine a status quo of the current state of technology weaknesses amongst that group of users. This representation is neither critical nor discriminating; rather it is diagnostic in nature. The mined data will provide a hitherto unavailable objective picture of current technology competence amongst a group of users. Correspondingly, the data mining endeavor might help a transformative leader understand and convey to others how a changed status quo might look. The new status quo, wherein teachers are empowered to avoid or solve many technology problems, can become a shared vision or goal that will benefit all members of a community who embrace the idea. Numerous transformative leaders espoused the concept of designing and sharing a communal vision as an element of an effective change program (Bates, 1995; Weiner, 2003; Shields, 2010). Further, teachers can assist one another in achieving that goal by working together and sharing or spreading new skills.

An HDDB data mining project that contributes to professional development topic identification can be used by a transformative leader to affect broad and equitable change. Such an undertaking deliberately ignores the particular circumstances of any individual teacher and holds at its core the notions of equality and objectivity. Mining a helpdesk dataset containing support requests from teachers intends to look for trends amongst all teachers, with a goal of providing professional development back to all teachers. The content of the professional development is rooted deeply within the faculty body, and the potential for positive change is

essentially equal for all faculty members. Indeed, the goal of the study is to consider the competencies of a faculty body collectively and to attempt to address those skills so that all faculty members are capable of the same usage and troubleshooting steps. Equitable change is a core feature of the deliverables of a successful helpdesk data mining project.

Just as equitable change is a core characteristic of the outcome of an HDDDB mining project, data analysis lies at the heart of the mining work. Regarding equitable change and the importance of objectivity, Shields (2010) borrowed from other researchers to discuss objectivity in data. Citing Evers and Wu (2006), Shields highlighted the importance of considering and analyzing data with a conscientious and thoughtful approach, allowing for patterns to emerge naturally. These trends or patterns should be brought to light without predispositions. By highlighting that equitable change is achieved through careful and objective analysis, Shields may be unintentionally endorsing the use of data mining practices. The actions taken during text data mining procedures are principally automated, and the computer algorithm is designed to look at trends without prejudice. A clustering data mining algorithm, for example, assembles groups of support requests with no regard to the semantic meaning of the data content. The mining of an HDDDB honors a characteristic of transformative leadership regarding the objectivity of data analyses.

The leader of an HDDDB data mining project, typically an administrator such as a Director of Educational Technology, has a professional responsibility to help teachers learn to use, troubleshoot, and teach with technological tools. Indeed, that leader may be in a position of power to select topics and areas of focus for the faculty body and must make decisions that will impact their leadership. While many practitioners would leverage this power to make decisions that benefit the institutions, even these potentially benevolent decisions might be affected by the

leader's biases. For example, a technology leader might recommend to a faculty that professional development efforts focus on a particular technology with which the manager is familiar or fond, but which may not be the most suitable tool for the faculty. Even if professional development on the technology tool is worthwhile, the leader might use her power to focus training efforts on that technology when the status quo dictates that different technology should be prioritized. By leading based on informed perspectives gleaned from the HDDDB, the administrator reduces the role of positional power in their decision making. Professional development goals may no longer be dictated by a single individual or a group of leaders, but, rather, extracted from the support requests of the very people to undergo professional development. This shifts the power from an individual or group to a consensus. The transformative leadership tenet of acknowledgment of power and privilege (Shields, 2010) is pervasive throughout a duly organized and executed HDDDB data mining undertaking at an educational institution.

Another power shift may occur when faculty members are equipped to avoid or troubleshoot many of the technology problems they experience. By empowering teachers to address their technology problems, power is reduced from the information technology helpdesk and shifted to those who might ask the helpdesk for support. In this case, the power under consideration corresponds to knowledge, and a change of both authority and knowledge from the few to the many may be beneficial for all members of a community. Indeed, a shift in knowledge brought about by careful and objective study and analysis is a fundamental characteristic of transformative leadership (Shields, 2010).

Shields (2010) described an important obligation of the transformative leader, regarding their role as an agent of change. Shields wrote, "a fundamental task of the educational leader in this transformative tradition is to ask questions, for example, about [...] which ideas should be

taught,” (p. 570). This statement applies impeccably to the case study at hand. Using data mining, questions were asked of the HDDB, populated by teachers, and the answers to these questions highlighted the topics that would be of greatest value to the data contributors.

Recommendations for Action

This case study generated numerous recommendations for action. The principal recommendation was straightforward: practitioners, scholars, Directors of Technology and Educational Technology, administrators, and institutional leaders are encouraged to undertake an HDDB data mining study. By leveraging data mining procedures and analyses, with an emphasis on identifying the most problematic technologies that affect the users requesting support from the helpdesk, leaders can expose previously unknown trends of technology weaknesses amongst the technology users. These weaknesses can then be address through professional development or training. With the helpdesk data readily available, as it will be at a majority of North American educational institutions, only a data mining undertaking stands between school administrators and new, unique, and potentially transformative contextual insight. Data mining is a complex field that leverages sophisticated technologies, but many third-party data mining service providers are available. Simply put, this study yielded a recommendation for decision makers to conduct HDDB data mining studies, as valuable information might be generated with a relatively small investment of time and money.

While an HDDB data mining study can be conducted upon helpdesk data of virtually any quality or thoroughness, this case study unearthed particular recommendations that, if accepted and applied to helpdesk practices, might improve the quality of the database and any data mining studies that ensue. As the data at the heart of any HDDB mining study will come from the information technology helpdesk office, many of the principal recommendations pertain to

actions that this office can take to generate better data. If HDDDB management improves, the clarity and detail of mining findings might improve commensurately.

Recommendations for Action: Help Desks

IT supervisors are encouraged to use existing helpdesk database fields correctly, consistently, and thoroughly. In this case study, the vast majority of information about each support request was stored in a single field, the problem description. Potentially useful fields, such as the request category or the type of technology in use, were leveraged only sporadically and thus could not be relied upon during the mining procedures. If these fields were populated correctly and consistently by IT agents at the time when they created each ticket, database records would be categorized with content-relevant labels. Carefully labeled and categorized support requests would scaffold structured and hierarchal cataloguing, resulting in more sophisticated classification opportunities.

For example, the database used in this case study did not contain any database fields that would have supported advanced classification. To support automated labeling, a new field was added, with only two potential values (REQ-AGENT for cases that required an agent for resolution and NOREQ-AGENT for cases that could be resolved without an agent), and classification analyses considered these two labels. Had the case study's dataset included a correctly populated "Technology Type" field, which technical support agents used to indicate if a request was related to *hardware* or *software*, then the classification analysis could have been more elaborate. For each support request, the classification system could have yielded four categories (*hardware* - REQ-AGENT, *hardware* - NOREQ-AGENT, *software* - REQ-AGENT, *software* - NOREQ-AGENT) rather than two categories, providing far more insight into the types and volumes of support requests that required an agent for resolution.

Similarly, IT managers are recommended to mindfully limit the amount of data in the key database fields, such as the support request description. The support request description field in the case study's dataset often included entire email messages or email chains copied and pasted as a single block of text. As a result, the data preprocessing and data mining procedures were required to account for a large number of tokens, many of which were not contextually relevant to the support request. These words, such as *thanks*, *sincerely*, *Friday*, *table*, or *bag*, were not likely to provide semantic meaning for either the helpdesk agent or a data mining study. Some words that may have multiple meanings, such as *monitor* (a computer screen or the action of overseeing), *power* (as in electricity, referring to a button, or referring to the on/off state of a device), and *project* (the verb for casting video content to a screen, or the noun for an undertaking), should be used judiciously by helpdesk operators as they create each ticket to ensure clarity. Any actions that might decrease the number of tokens per support request while maintaining or increasing clarity might yield an improved dataset for mining. Helpdesk managers are advised to modify ticket creation practices with this goal in mind.

The procedures leveraged in creating support requests are not the only helpdesk practices that might improve as a result of this research. While this case study performed classification and clustering analyses upon a dataset collected over a period of three years, the benefits and information gleaned from these analyses suggest an important recommendation for action. Helpdesk managers and information technology leaders are encouraged to run classification and clustering analyses frequently, perhaps monthly, daily, or even in real time as support requests are created. While this recommendation may necessitate procedural changes in the technology office operations, the benefits may outweigh the challenges. A clustering procedure performed in real time as support requests are created could intelligently and automatically divide support

requests into, for example, hardware and software categories. These tickets could be assigned to the IT support agent best suited to handle these issues. This approach echoes Marom and Zukerman's (2009) methodology which used clustering and prediction models to divide support requests into categories by complexity, with different IT resources allocated to address each request.

Proper and immediate assignment of support requests to the most qualified agent may yield performance improvements in the IT department. Similarly, classification analyses run regularly may provide IT and professional development leaders with a list of support requests that require an IT agent for resolution and requests that could be handled by a well-trained user. An agent can deal with the former set of requests, while a professional development planner can determine the most immediate and pressing issues that face the user base from the latter list. Even if these analyses were not run in real time, helpdesk offices and school leaders could gain an understanding of the current problems experienced by, or areas of skill weakness of, the user group. Helpdesk offices may be particularly interested in trends that reveal possible hardware or software failures, while professional development decision makers may be interested in the technologies with which users struggle.

Other Recommendations for Action

This case study attempted to leverage standard data mining practices to glean information from the helpdesk database at an independent, boarding high school in North America. The study did not attempt to pioneer new data mining methods or procedures. Indeed, the study deliberately employed the same techniques that many other data mining studies had used but applied them to an underused dataset with a new goal. The results of the study provided valuable insight about the technological weakness in teachers' skillsets based on the support requests submitted by

teachers. The findings of this case study were not intended for generalization to other learning institutions. To interested parties beyond the study site, the primary value of the study's findings was that the methodology for HDDDB data mining for professional development topic detection successfully identified topics and themes that warranted professional development attention. While the study's most practical finding, (that the research site's faculty might benefit from training on technologies including the classroom audio/visual systems, MacBook Air computers, proper network usage and troubleshooting, and the Google Chrome web browser), isn't transferrable to other institutions, the project's methodology is recommended as a course of action.

If other institutions undertake an HDDDB data mining that leveraged this case study's methodology, the study may yield valuable information regarding the status of technology usage at the study site. Data mining methods such as the generation of a wordlist and the creation and application of clustering and classification models may produce beneficial insight. Leaders of HDDDB data mining undertakings are strongly encouraged to go beyond simple execution of these analyses. By applying human judgment and interpretation to the analysis results, the insights generated are likely to improve noticeably, providing not just more detail, but also contextual information about the use of technologies at the site and for the site's particular goals. Finally, practitioners are also encouraged to merge and combine the results of data mining procedures to glean even more information from their mining undertakings.

If practitioners and school leaders undertook HDDDB data mining studies as described and recommended within this case study, they might gain a deeper understanding of the areas of technological weaknesses for the group of data contributors whom the researchers have chosen to study. By highlighting problematic technologies, the practitioners may devise professional

development to address the identified skillset weaknesses. If this training is successful, teachers' technology capabilities and competencies may improve. Many researchers and scholarly studies have associated improved technology skills with greater educational technology integration. Improvements in technology skills were linked to augmented educational technology undertakings and technology-supported education (Inan & Lowther, 2010b; Ertmer et al., 2012; Pilgrim & Berry, 2014). Similarly, growth in teachers' technology competency may also contribute to the development of teachers' self-efficacy, and technology-related beliefs and attitudes (Conrad & Munro, 2008; Ertmer & Ottenbreit-Leftwich, 2010; Inan & Lowther, 2010a). Conrad and Munro linked advancement in these traits to improved educational technology efforts. Güneş, Gökçek, and Bacanak (2010) identified a direct correlation between technology competence and technology confidence and attitudes, with improvements in one generating progress in the other.

An HDDB data mining study, in the style of the case study under review, may be a powerful new method for identifying skillset weaknesses and areas of improvement. Regarding the importance of context-relevant technology skills, Pilgrim and Berry (2014) elucidated simply that, "in order to teach with technology, the teachers had to be familiar with the technology" they were expected to teach with (p. 137). An HDDB data mining study can highlight the specific technologies with which teachers lack familiarity and competency. HDDB-derived technology weaknesses present a new level of detail regarding professional development topics and are backed by firm evidence. This study highlighted a list of technologies, by name, which warranted professional development attention. In a complex educational environment, teachers might use many technologies as part of their organizational or pedagogic work. The determination of general categories of problematic technologies may not be enough to

understand precisely where teachers are struggling. Mining of a school's HDDB data is capable of presenting highly detailed insight into the technologies used by teachers that are most problematic. An HDDB data mining study is recommended for any site where this type of information may be useful.

Recommendations for Further Study

Scholars and scholar-practitioners are encouraged to undertake HDDB data mining studies with a goal of identifying areas of weakness or training topics for those who submitted the support requests. These studies may leverage in-house data miners, or as with this case study, may outsource the complex data mining tasks to a trained professional. Regardless of who performs the data mining procedures, institutions stand to benefit from conducting HDDB data mining studies for their helpdesk data, or a subset thereof. While new studies may be customized, designed, and modified by the research team undertaking the investigation, this case study has identified numerous opportunities and recommendations for further scholarly undertakings.

As this study may be amongst the first of its kind, future studies may attempt to replicate the methods described in this case study. A nearly identical study, conducted at a different site, may prove beneficial in supporting the case study's primary finding: that data mining methods applied to an HDDB can generate meaningful and relevant professional development topics. While the results from nearly any HDDB data mining study conducted at different research sites are likely non-generalizable and non-transferable, continued attempts at detailing and refining the methodology will serve to add to the growing body of knowledge regarding HDDB data mining for professional development topic discovery.

Scholars and scholar-practitioners are also encouraged to replicate the study while using a control group of support requests. In this case study, many nearly identical support requests regarding Google Groups were useful as a control group. The presence of these control group support requests helped to confirm the functionality of the clustering algorithm. Further, the algorithm successfully separated these cases from mingling with support requests in other clusters. This isolation might be very useful at other research sites where many similarly-worded support requests are anticipated. Researchers who undertake HDDDB data mining studies should consider including a control group of support requests to confirm the functionality of their clustering algorithm, and to ensure that predictably similar tickets do not get incorrectly bundled with other cases. A research study that compared the results of the same data mining methods applied to the same dataset with and without a control group of tickets would also add to the collective body of knowledge.

In addition to replication studies that attempt to mimic the data mining and analysis methodology outlined within this case study, additional studies with minor modifications to the inputted helpdesk data might benefit the particular site and the field of research. For example, future studies may attempt to apply the same data mining techniques to data extracted from the helpdesk for the same site but at different time periods. This method might yield valuable insight regarding technology weakness trends over time. Similarly, a study that opted to address various groups of data contributors might uncover valuable training information unique to each group. For example, the application of HDDDB data mining practices to help requests submitted by members of various academic departments or buildings within a single site may highlight differences or similarities in technology skillset themes by logical groups of users.

Further, while this case study analyzed the support requests from a single location, a future study might mine the HDDBs from separate but affiliated sites. For example, the same data mining methodology applied to helpdesk data from numerous schools within the same district might highlight technological skillset trends or differences between faculty bodies. These results may be particularly interesting if all schools within the district (or within the study set) are expected to use the same educational technologies. If the results for one or more schools yield noticeably different results from other schools, decision makers and district leaders might be positioned to leverage this information for district-wide improvements. Similarly, if trends of technology problems are widespread, professional development could be arranged, or new technologies considered. A single study might call for data mining practices applied to the helpdesks of different schools, or a meta-study may be conducted that considered the results of numerous data mining studies to search for large-scale trends.

Correspondingly, a far-reaching study or a survey of studies about a particular technology might generate valuable insight into the common problems with that technology. For example, many schools, both public and private, use Google Apps for Education as a communication and collaboration platform. A collection of individual HDDB data mining studies which themselves only focused on Google Apps-related support requests may reveal particular trends of problems across a broad user base. In this manner, educational technology hardware and software developers may become aware of the facets of their technologies that most commonly lead to user work stoppages.

While many recommendations presented for further study suggest replication, other scholars may be interested in modifying the data mining procedures employed in this study. This case study created n -gram tokens by combining up to n consecutive words to create a single

token. For example, the text *broken projector wire* would have generated the following n -grams of length two: *broken_projector* and *projector_wire*. This study did not create n -grams of non-consecutive words, though the technology to perform this sophisticated step existed at the time of the study (B. Tvenstrup, personal communication, December 28, 2016). In a future study, where non-consecutive words are joined into new tokens, the example phrase may have generated the following n -grams of length two: *broken_projector*, *projector_wire*, and *broken_wire*. From this example, it seems that n -grams of non-consecutive tokens might prove important and prevalent in an HDDB data mining study aimed at topic discovery. Researchers are encouraged to undertake new versions of this study that leverage different data preprocessing and mining decisions.

While this study leveraged clustering and classification procedures, other studies may employ different data mining steps. For example, sentiment analysis, in which text data is mined not only for content but for the emotions that underlay the content (Romero & Ventura, 2013; Povoda et al., 2015), might provide valuable information if leveraged during the mining of an HDDB. Sentiment analysis or other text data mining procedures might provide additional insight that complements clustering or classification procedures.

Conclusion

This case study yielded numerous recommendations for further study. These recommendations included modifying the data to be mined and changing the mining methods and tactics. New studies that leverage the methodology introduced in this case study but adjust the inputted data or mining procedures might make valuable contributions to this burgeoning field of research. However, as the goal of this case study was to unite HDDB data mining and technological professional development, the most important recommendation for further research

is straightforward. The researcher recommends that educational institution leaders perform HDDB data mining studies and provide professional development on topics gleaned during the mining. This mining-training process should be performed repeatedly and cyclically to monitor growth and progress in users' technology skillsets. The long-term value of mining helpdesk databases for professional development topic discovery will only be understood once a site leader mines its helpdesk database, gleans topics for professional development, and then provides training on those topics. In due time, another mining-training cycle can occur, with particular emphasis on the issues addressed during the professional development. Over time and across cycles of research, a reduction in helpdesk support requests pertaining to the subjects identified and trained upon in earlier cycles will validate the usefulness of HDDB mining for topic discovery.

Arguably the most important connotation from this case study is an imploration to educators, scholars, and practitioners to ask questions of data they already possess. This case study highlighted the value of searching for answers from an existing data resource, rather than collecting new information for analysis. Clearly, an institution may already possess a data source containing extraordinarily useful and context-relevant information in an accessible format. The data waits for a transformative leader to pose questions of it. In this instrumental, single-site case study, as in other studies it may inspire, the researcher's only action was to ask existing data to reveal its secrets, and the clear, objective truths revealed themselves.

REFERENCES

- Abdous, M. & He, W. (2009). Using text mining to uncover students' technology-related problems in live video streaming. *British Journal of Educational Technology*, 42(1), 40-49.
- Abowitz, D. A., & Toole, T. M. (2009). Mixed method research: Fundamental issues of design, validity, and reliability in construction research. *Journal of Construction Engineering and Management*, 136(1), 108-116.
- Andrews, A. A., Beaver, P., & Lucente, J. (2016). Towards better helpdesk planning: Predicting incidents and required effort. *Journal of Systems and Software*, 117, 426-449.
- Andrews, A., & Lucente, J. (2014, April). From Incident Reports to Improvement Recommendations: Analyzing IT helpdesk Data. In *2014 23rd Australian Software Engineering Conference* (pp. 94-103). IEEE.
- Baker, R. S. (2010). Data mining for education. *International encyclopedia of education*, 7, 112-118.
- Baker, R. S., & Siemens, G. (2014) Educational data mining and learning analytics. In Sawyer, K. (Ed.) *Cambridge Handbook of the Learning Sciences: 2nd Edition*, pp. 253-274.
- Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1), 3-17.
- Bates, R. (1995, July). *A socially critical perspective on educational leadership*. Paper presented at the Flinders University Conference on Educational Leadership, Adelaide, S. Australia.
- Bauer, J., & Kenton, J. (2005). Toward technology integration in the schools: Why it isn't happening. *Journal of Technology and Teacher Education*, 13(4), 519-546.
- Bickel, S., & Scheffer, T. (2004, September). Learning from message pairs for automatic email

- answering. In *European Conference on Machine Learning* (pp. 87-98). Springer Berlin Heidelberg.
- Blaafflad, A.N., Johansen, B.H., Eide, N.E., & Sandnes, F. E. (2004). A text-mining approach to helpdesk and e-mail support. In *Proceedings of the annual Norwegian Computer Science Conference*.
- Bloomberg, L.D. & Volpe, M. (2012). *Completing your qualitative dissertation: a road map from beginning to end*. Thousand Oaks, CA: Sage Publications.
- Buabeng-Andoh, C. (2012). Factors influencing teachers' adoption and integration of information and communication technology into teaching: A review of the literature. *International Journal of Education and Development using Information and Communication Technology*, 8(1), 136-155.
- Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In *Evolution of teaching and learning paradigms in intelligent environment* (pp. 183-221). Springer: Berlin, GE.
- Conrad, A. M., & Munro, D. (2008). Relationships between computer self-efficacy, technology, attitudes and anxiety: Development of the computer technology use scale (CTUS). *Journal of Educational Computing Research*, 39(1), 51-73.
- Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage Publications.
- Davies, R. S. (2011). Understanding technology literacy: A framework for evaluating educational technology integration. *TechTrends*, 55(5), 45-52.
- Ertmer, P. A., & Ottenbreit-Leftwich, A. T. (2010). Teacher technology change: How knowledge, confidence, beliefs, and culture intersect. *Journal of Research on Technology*

- in Education*, 42(3), 255-284.
- Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship. *Computers & Education*, 59(2), 423-435.
- Evers, C. W., & Wu, E. H. (2006). On generalising from single case studies: Epistemological reflections. *Journal of Philosophy of Education*, 40(4), 511-526.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37.
- Flowers, C.P., & Algozzine, R.F. (2000). Development and validation of scores on the basic technology competencies for educators inventory. *Educational and Psychological Measurement*, 60(3), 411-418.
- Foo, S., Hui, S. C., & Leong, P. C. (2002). Web-based intelligent helpdesk-support environment. *International Journal of Systems Science*, 33(6), 389-402.
- Forman, G., Kirshenbaum, E., & Suermondt, J. (2006, August). Pragmatic text mining: minimizing human effort to quantify many issues in call logs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 852-861). ACM.
- Groff, J., & Mouza, C. (2008). A framework for addressing challenges to classroom technology use. *AACE Journal*, 16(1), 21-46.
- Güneş, G., Gökçek, T., & Bacanak, A. (2010). How do teachers evaluate themselves in terms of technological competencies? *Procedia-Social and Behavioral Sciences*, 9, 1266-1271.
- Hancock, R., Knezek, G., & Christensen, R. (2007). Cross-validating measures of technology integration: A first step toward examining potential relationships between technology

- integration and student achievement. *Journal of Computing in Teacher Education*, 24(1), 15-21.
- Hanna, M. (2004). Data mining in the e-learning domain. *Campus-wide information systems*, 21(1), 29-34.
- He, W. (2013). Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*, 29(1), 90-102.
- Heathcote, E. & Dawson, S. (2005) Data Mining for Evaluation, Benchmarking and Reflective Practice in a LMS. In *Proceedings E-Learn 2005: World conference on E-learning in corporate, government, healthcare & higher education*, Vancouver, Canada.
- Hew, K. F., & Brush, T. (2007). Integrating technology into K-12 teaching and learning: Current knowledge gaps and recommendations for future research. *Educational Technology Research and Development*, 55(3), 223-252.
- Hooper, S., & Rieber, L.P. (1999). Teaching, instruction, and technology. In A.C. Ornstein & L.S. Behar-Horenstein (Eds.), *Contemporary issues in curriculum*. 252-264. Boston: Allyn and Bacon.
- Hui, S. C., & Jha, G. (2000). Data mining for customer service support. *Information & Management*, 38(1), 1-13.
- Inan, F. A., & Lowther, D. L. (2010a). Factors affecting technology integration in K-12 classrooms: A path model. *Educational Technology Research and Development*, 58(2), 137-154.
- Inan, F. A., & Lowther, D. L. (2010b). Laptops in the K-12 classrooms: Exploring factors impacting instructional use. *Computers & Education*, 55(3), 937-944.
- Joshi, K. P., Joshi, A., & Yesha, Y. (2011, March). Managing the quality of virtualized services.

- In *2011 Annual SRII Global Conference* (pp. 300-307). IEEE.
- Klösgen, W., & Zytkow, J. M. (2002). *Handbook of data mining and knowledge discovery*. Oxford University Press, Inc.
- Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A., & Rosé, C. P. (2015). Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(4), 333-353.
- Kopcha, T. J. (2012). Teachers' perceptions of the barriers to technology integration and practices with technology under situated professional development. *Computers & Education*, 59(4), 1109-1121.
- Kopcha, T. J., & Sullivan, H. (2007). Self-presentation bias in surveys of teachers' educational technology practices. *Educational Technology Research and Development*, 55(6), 627-646.
- Maderick, J.A. (2013). *Validity of subjective self-assessment of digital competence among undergraduate preservice teachers* (Doctoral dissertation). Retrieved from UNLV Theses/Dissertations/Professional Papers/Capstones. Paper 1941.
- Maderick, J.A., Zhang, S., Hartley, K., & Marchand, G. (2015). Preservice teachers and self – assessing digital competence. *Journal of Educational Computing Research*, 54(3), 326 – 351.
- Malik, R., Subramaniam, L. V., & Kaushik, S. (2007, January). *Automatically Selecting Answer Templates to Respond to Customer Emails*. In *IJCAI* (Vol. 7, pp. 1659-1664).
- Marom, Y., & Zukerman, I. (2009). An empirical study of corpus-based response automation methods for an e-mail-based help-desk domain. *Computational Linguistics*, 35(4), 597-635.
- Means, B. (1994). Introduction: Using technology to advance educational goals. *Technology and*

- education reform: The reality behind the promise*, 1-21. San Francisco, CA: Jossey-Bass.
- Merceron, A., & Yacef, K. (2005, May). Educational Data Mining: a Case Study. In *AIED* (pp. 467-474).
- Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. *Procedia-Social and Behavioral Sciences*, 97, 320-324.
- Monk, D. (2005). Using data mining for e-learning decision making. *The Electronic Journal of e-Learning*, 3(1), 41-54.
- Okojie, C., Olinzock, A., & Okojie-Boulder, T. C. (2006). The pedagogy of technology integration. *Journal of Technology Studies*, 32(2), 66-71.
- Palmieri, L., Semich, G., & Graham, J. (2009). A needs assessment of basic technology competencies with a suggested model of improving teacher preparation and technology integration at Robert Morris University. *Society for Information Technology & Teacher Education International Conference*, 2009(1). 3505-3508.
- Pilgrim, J., & Berry, J. (2014). Technology integration with teacher candidates in a summer-camp setting. *Journal of Digital Learning in Teacher Education*, 30(4), 131-138.
- Povoda, L., Arora, A., Singh, S., Burget, R., & Dutta, M. K. (2015, October). Emotion recognition from helpdesk messages. In *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2015 7th International Congress on* (pp. 310-313). IEEE.
- Reinhart, J. M., Thomas, E., & Toriskie, J. M. (2011). K-12 teachers: Technology use and the second level digital divide. *Journal of Instructional Psychology*, 38(3), 181-193.
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert*

- systems with applications*, 33(1), 135-146.
- Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(6), 601-618.
- Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1), 12-27.
- Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1), 368-384.
- Rosen, L. D., Whaling, K., Carrier, L. M., Cheever, N. A., & Rokkum, J. (2013). The media and technology usage and attitudes scale: An empirical investigation. *Computers in Human Behavior*, 29(6), 2501-2511.
- Schibeci, R., MacCallum, J., Cumming-Potvin, W., Durrant, C., Kissane, B., & Miller, E. J. (2008). Teachers' journeys towards critical use of ICT. *Learning, Media and Technology*, 33(4), 313-327.
- Shields, C. M. (2010). Transformative leadership: Working for equity in diverse contexts. *Educational Administration Quarterly*, 46(4), 558-589.
- Shum, S. B., Knight, S., & Littleton, K. (2012). Learning analytics. In *UNESCO Institute for Information Technologies in Education. Policy Brief*.
- Siemens, G., & Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.
- Tasir, Z., Abour, K. M. E. A., Halim, N. D. A., & Harun, J. (2012). Relationship between Teachers' ICT Competency, Confidence Level, and Satisfaction toward ICT Training

- Programmes: A Case Study among Postgraduate Students. *Turkish Online Journal of Educational Technology-TOJET*, 11(1), 138-144.
- Wartena, C., & Brussee, R. (2008, September). Topic detection by clustering keywords. In *2008 19th International Workshop on Database and Expert Systems Applications* (pp. 54-58). IEEE.
- Weiner, E. J. (2003). Secretary Paulo Freire and the democratization of power: Toward a theory of transformative leadership. *Educational Philosophy and theory*, 35(1), 89-106.
- Wenger, E., & Lave, J. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge, UK: Cambridge University Press.
- Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of information science*, 18(1), 45-55.
- Zhao, M., Leckie, C., & Rowles, C. (1996). An interactive fault diagnosis expert system for a helpdesk application. *Expert Systems*, 13(3), 203-217.
- Zhao, Y., & Bryant, F. L. (2006). Can teacher technology integration training alone lead to high levels of technology integration? A qualitative look at teachers' technology integration after state mandated technology training. *Electronic Journal for the Integration of Technology in Education*, 5(1), 53-62.
- Zhong, N., Matsunaga, T., & Liu, C. (2002, August). A text mining agents based architecture for personal e-mail filtering and management. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 329-336). Springer: Berlin, GE.

APPENDICES

Appendix A

Wordlist Tokens in 10% or More Unique Helpdesk Requests.

Token	In Unique HD Requests	Total Instances	% of Unique HD Requests
us	676	1,225	35.88%
comput	536	1,226	28.45%
work	528	1,055	28.03%
set	477	826	25.32%
subject	469	836	24.89%
googl	414	1,363	21.97%
issu	414	668	21.97%
school	383	595	20.33%
know	375	620	19.90%
receiv	365	482	19.37%
phone	342	620	18.15%
network	334	672	17.73%
messag	324	677	17.20%
http	320	948	16.99%
instal	320	512	16.99%
add	316	668	16.77%
com	315	1,065	16.72%
access	300	675	15.92%
password	283	663	15.02%
close	278	294	14.76%
chang	274	482	14.54%
connect	273	542	14.49%
system	269	355	14.28%
student	267	671	14.17%
email	264	621	14.01%
web	263	525	13.96%
group	253	1,918	13.43%
see	247	411	13.11%
problem	246	399	13.06%
account	237	725	12.58%
link	235	390	12.47%
try	235	352	12.47%
click	223	662	11.84%
look	223	307	11.84%
abl	222	313	11.78%
made	222	262	11.78%
check	221	311	11.73%
send	220	328	11.68%

Appendix A (continued)

Token	In Unique HD Requests	Total Instances	% of Unique HD Requests
classroom	218	327	11.57%
troubl	218	234	11.57%
inform	213	376	11.31%
desktop	211	349	11.20%
make	211	313	11.20%
take	211	278	11.20%
come	208	310	11.04%
equip	204	228	10.83%
wrote	202	312	10.72%
mail	200	375	10.62%
print	200	605	10.62%
page	199	506	10.56%
resolv	197	200	10.46%
test	197	349	10.46%
vdt	196	371	10.40%
visual	196	212	10.40%
continu	195	214	10.35%
room	193	392	10.24%
reset	191	253	10.14%
sent	191	269	10.14%
administr	189	223	10.03%

Appendix B

Top 50 Tokens by Average Term Frequency for Each Cluster

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.
comput	0.237	print	0.205	phone	0.089	group	0.493
work	0.099	password	0.185	account	0.080	googl	0.267
instal	0.068	reset	0.142	us	0.073	member	0.180
projector	0.067	swi	0.100	network	0.072	forum	0.159
classroom	0.064	okta	0.093	access	0.066	add	0.135
us	0.062	virtual	0.086	messag	0.053	group_googl	0.125
equip	0.057	desktop	0.086	subject	0.053	com	0.121
audio	0.056	vdt	0.083	telephon	0.052	http	0.116
visual	0.052	virtual_desktop	0.075	issu	0.050	googl_com	0.115
connect	0.048	printer	0.070	work	0.049	page	0.110
audio_visual	0.047	googl	0.055	student	0.046	manag	0.108
room	0.047	work	0.043	receiv	0.045	http_group	0.105
classroom_							
equip	0.043	log	0.033	email	0.041	add_member	0.104
appl	0.042	chang	0.030	googl	0.040	com_forum	0.104
subject	0.041	subject	0.029	instal	0.040	web	0.101
test	0.035	set	0.029	chang	0.039	web_page	0.098
know	0.034	issu	0.029	school	0.039	page_http	0.092
updat	0.033	access	0.025	swi	0.037	click	0.091
laptop	0.033	mba	0.024	connect	0.037	set	0.083
set	0.032	comput	0.024	mail	0.037	us	0.061
sound	0.031	account	0.023	set	0.036	direct	0.056
mba	0.031	save	0.023	know	0.034	member_group	0.056
issu	0.030	us	0.022	system	0.033	forum_forum	0.055
mac	0.029	chrome	0.021	offic	0.032	direct_add	0.055
replac	0.029	school	0.020	us_network	0.031	link	0.054
pick	0.028	phone	0.019	dorm	0.030	activ	0.053
plug	0.027	system	0.019	close	0.030	set_manag	0.052
come	0.027	abl	0.019	administr	0.029	group_manag	0.051
						forum_	
offic	0.027	delet	0.019	inform	0.028	managememb	0.050
cabl	0.024	messag	0.018	see	0.027	managememb	0.050
check	0.023	know	0.017	number	0.027	member_activ	0.050
school	0.023	clear	0.017	troubl	0.026	group_web	0.049
softwar	0.023	tri	0.016	add	0.026	made	0.049

Appendix B (continued)

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.
problem	0.022	email	0.016	com	0.026	us_direct	0.049
try	0.022	web	0.016	file	0.026	manag_web	0.049
network	0.022	secur	0.016	summer	0.025	activ_click	0.049
student	0.022	instal	0.016	folder	0.025	click_link	0.049
lab	0.022	try	0.015	problem	0.024	link_us	0.049
				administr_			
power	0.021	problem	0.015	system	0.024	member_add	0.049
dvd	0.021	document	0.014	creat	0.024	manag_group	0.048
make	0.021	http	0.014	resolv	0.024	made_set	0.043
look	0.021	com	0.014	list	0.023	group_group	0.042
class	0.020	login	0.014	contin	0.023	googl_group	0.032
speaker	0.020	sent	0.014	take	0.023	student	0.030
dorm	0.019	administr	0.013	wifi	0.023	address	0.025
		administr_					
meet	0.018	system	0.013	address	0.022	chang	0.025
desktop	0.018	test	0.012	issu_resolv	0.021	email	0.024
				contin_			
abl	0.018	copier	0.012	troubl	0.021	creat	0.023
show	0.018	sign	0.012	send	0.020	class	0.019
turn	0.018	user	0.011	resolv_close	0.020	request	0.014

Appendix C

Bottom 50 Tokens by Average Term Frequency for Each Cluster

Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.
instal_add	0.000	add_member	0.000	tag	0.002	further_word	0.000
form	0.000	audio	0.000	past	0.002	given	0.000
match	0.000	audio_visual	0.000	macbook	0.002	hdmi	0.000
		classroom_					
past	0.000	equip	0.000	action	0.002	head_employe	0.000
databas	0.000	com_forum	0.000	main	0.002	instal_add	0.000
googl_com	0.000	contact_safeti	0.000	owner	0.001	instruct	0.000
page_http	0.000	databas	0.000	termin	0.001	internet	0.000
access_top	0.000	direct_add	0.000	sent_firstclass	0.001	ipad	0.000
account_creat	0.000	displai	0.000	understand	0.001	iphon	0.000
activ_click	0.000	dorm_dorm	0.000	video	0.001	lab	0.000
add_member	0.000	drop	0.000	monitor	0.001	level_group	0.000
addit_comment	0.000	duti	0.000	exampl	0.001	line	0.000
com_forum	0.000	exampl	0.000	descript	0.001	locat	0.000
comment	0.000	forum	0.000	save	0.001	mac	0.000
contact_safeti	0.000	forum_forum	0.000	hand	0.001	machin	0.000
		forum_					
creation	0.000	managememb	0.000	googl_group	0.001	match_addit	0.000
depart_app	0.000	googl_group	0.000	plai	0.001	mayb	0.000
dept_head	0.000	group_googl	0.000	basic	0.001	namech	0.000
						network_	
direct_add	0.000	group_group	0.000	audio	0.001	account	0.000
						number_	
drive_access	0.000	group_manag	0.000	unplug	0.001	depart	0.000
employe_							
number	0.000	group_web	0.000	audio_visual	0.001	pdf	0.000
exampl	0.000	hdmi	0.000	web_page	0.001	phone_fax	0.000
				classroom_			
forum_forum	0.000	http_group	0.000	equip	0.001	plug	0.000
forum_manage							
memb	0.000	intern	0.000	hdmi	0.000	prefer_posit	0.000
googl_group	0.000	jack	0.000	add_member	0.000	printer	0.000
group_googl	0.000	languag	0.000	link_us	0.000	profil_match	0.000
group_manag	0.000	link_us	0.000	dvd	0.000	receiv_subject	0.000
group_web	0.000	made_set	0.000	member_group	0.000	replac	0.000

Appendix C (continued)

	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.	Token	Avg. Term Freq.	
head_employe	0.000	main	0.000	group_googl	0.000	safeti	0.000	
http_group	0.000	manag_group	0.000	us_direct	0.000	said	0.000	
level_group	0.000	manag_web	0.000	projector	0.000	save	0.000	
link_us	0.000	managememb	0.000	activ_click	0.000	select	0.000	
made_set	0.000	member_activ	0.000	com_forum	0.000	sent_firstclass	0.000	
manag_group	0.000	member_add	0.000	direct_add	0.000	softwar	0.000	
manag_web	0.000	member_group	0.000	forum	0.000	station	0.000	
managememb	0.000	network_account	0.000	forum_forum	0.000	submit_take	0.000	
match_addit	0.000	owner	0.000	managememb	0.000	supervisor	0.000	
						supervisor_		
member_activ	0.000	page_http	0.000	group_group	0.000	dept	0.000	
member_add	0.000	past	0.000	group_manag	0.000	tag	0.000	
member_group	0.000	plug	0.000	group_web	0.000	take_effect	0.000	
number_depart	0.000	projector	0.000	http_group	0.000	turn	0.000	
owner	0.000	run	0.000	made_set	0.000	us_network	0.000	
prefer_posit	0.000	safeti	0.000	manag_group	0.000	virtual	0.000	
profil_match	0.000	set_manag	0.000	manag_web	0.000	virtual_desktop	0.000	
set_manag	0.000	speaker	0.000	managememb	0.000	wall	0.000	
submit_take	0.000	task	0.000	member_activ	0.000	window	0.000	
supervisor_dept	0.000	unplug	0.000	member_add	0.000	wireless	0.000	
take_effect	0.000	us_direct	0.000	page_http	0.000	word_close	0.000	
top_level	0.000	video	0.000	set_manag	0.000	work_fine	0.000	
us_direct	0.000	wall	0.000	speaker	0.000	work_order	0.000	

Appendix D

Prevalent Meaningful Tokens with Count of Clusters

Meaningful Token	Count of Clusters where Token Appears in top 100	Meaningful Token	Count of Clusters where Token Appears in top 100
chrome	3	group_manag	1
network	3	group_web	1
password	3	http_academ	1
phone	3	http_group	1
vdt	3	jack	1
copier	2	laptop	1
mac	2	manag_group	1
mba	2	manag_web	1
swi	2	managememb	1
telephon	2	member	1
wifi	2	member_activ	1
academ_web	1	member_add	1
add_member	1	member_group	1
audio	1	monitor	1
audio_visual	1	okta	1
classroom	1	print	1
classroom_equip	1	printer	1
com_forum	1	projector	1
direct_add	1	set_manag	1
equip	1	sound	1
forum	1	speaker	1
forum_forum	1	us_network	1
forum_managememb	1	video	1
googl_group	1	virtual	1
group	1	virtual_desktop	1
group_googl	1	visual	1
group_group	1	wireless	1

Appendix E

Sample of Classification Label Predictions and Confidences

Unique Work Order Number	NOREQ-AGENT Confidence	REQ-AGENT Confidence	Manually-Assigned Label	Classification Label
31191	0.816381190	0.183618809	NOREQ-AGENT	NOREQ-AGENT
31196	0.533837216	0.466162783		NOREQ-AGENT
31201	0.515040211	0.484959788		NOREQ-AGENT
31205	0.599702397	0.400297602		NOREQ-AGENT
31213	0.545032577	0.454967422		NOREQ-AGENT
31230	0.558520161	0.441479838	REQ-AGENT	NOREQ-AGENT
31236	0.646361644	0.353638355		NOREQ-AGENT
31240	0.793575478	0.206424521		NOREQ-AGENT
31245	0.433053561	0.566946438		NOREQ-AGENT
31259	0.600634562	0.399365437		NOREQ-AGENT
31262	0.698552349	0.301447650		NOREQ-AGENT
31270	0.351998840	0.648001159		REQ-AGENT
31281	0.604642929	0.395357070		NOREQ-AGENT
31309	0.599795020	0.400204979	REQ-AGENT	NOREQ-AGENT
31310	0.460292633	0.539707366		NOREQ-AGENT
31313	0.416363009	0.583636990		NOREQ-AGENT
31314	0.579496592	0.420503407		NOREQ-AGENT
31315	0.387074519	0.612925480		REQ-AGENT
31322	0.381731910	0.618268089		REQ-AGENT
31329	0.473620282	0.526379717		NOREQ-AGENT
31342	0.510367067	0.489632932		NOREQ-AGENT

Note. Cases labeled with REQ-AGENT were predicted to require an agent for resolution. Cases labeled with NOREQ-AGENT were predicted to not require an agent for resolution.

Appendix F

Wordlist Tokens Deemed *Meaningful*, or Contextually Relevant

Token	In Unique HD Requests	Total Instances	% of Unique HD Requests	Category	Sub- Category	Category and SubCat Combined
phone	342	620	18.15%	Phone		Phone
network	334	672	17.73%	Network		Network
password	283	663	15.02%	Accounts	Password Reset	Accounts - Password Reset
group	253	1918	13.43%	Google Groups		Google Groups
classroom	218	327	11.57%	Hardware	Classroom A/V	Hardware - Classroom A/V
equip	204	228	10.83%	Hardware		Hardware
print	200	605	10.62%	Printing		Printing
vdt	196	371	10.40%	Virtual Desktop		Virtual Desktop
visual	196	212	10.40%	Hardware	Classroom A/V	Hardware - Classroom A/V
us_network	186	190	9.87%	Network		Network
audio	177	220	9.39%	Hardware	Classroom A/V	Hardware - Classroom A/V
swi	175	458	9.29%	Software	FirstClass	Software - FirstClass
member	170	751	9.02%	Google Groups		Google Groups
group_googl	162	411	8.60%	Google Groups		Google Groups
virtual	162	274	8.60%	Virtual Desktop		Virtual Desktop
classroom_ equip	159	159	8.44%	Hardware	Classroom A/V	Hardware - Classroom A/V
audio_visual	155	155	8.23%	Hardware	Classroom A/V	Hardware - Classroom A/V
forum	142	577	7.54%	Google Groups		Google Groups
http_group	142	374	7.54%	Google Groups		Google Groups
com_forum	141	372	7.48%	Google Groups		Google Groups
forum_forum	141	203	7.48%	Google Groups		Google Groups
group_manag	141	176	7.48%	Google Groups		Google Groups

Appendix F (continued)

Token	In Unique HD Requests	Total Instances	% of Unique HD Requests	Category	Sub- Category	Category and SubCat Combined
member_group	140	219	7.43%	Google Groups		Google Groups
add_member	138	394	7.32%	Google Groups		Google Groups
direct_add	138	199	7.32%	Google Groups		Google Groups
forum_managememb	138	170	7.32%	Google Groups		Google Groups
group_web	138	170	7.32%	Google Groups		Google Groups
managememb	138	170	7.32%	Google Groups		Google Groups
member_activ	138	170	7.32%	Google Groups		Google Groups
manag_group	136	166	7.22%	Google Groups		Google Groups
manag_web	136	167	7.22%	Google Groups		Google Groups
member_add	136	167	7.22%	Google Groups		Google Groups
set_manag	136	195	7.22%	Google Groups		Google Groups
virtual_desktop	133	195	7.06%	Virtual Desktop		Virtual Desktop
projector	131	363	6.95%	Hardware	Classroom A/V	Hardware - Classroom A/V
telephon	128	135	6.79%	Phone		Phone
mac	114	206	6.05%	Hardware	Mac	Hardware - Mac
laptop	110	239	5.84%	Hardware		Hardware
group_group	104	112	5.52%	Google Groups		Google Groups
printer	101	223	5.36%	Printing		Printing
sound	94	186	4.99%	Hardware	Classroom A/V	Hardware - Classroom A/V
mba	89	111	4.72%	Hardware	Mac	Hardware - Mac
okta	83	198	4.41%	Accounts	Okta (SSO)	Accounts - Okta (SSO)
chrome	82	177	4.35%	Software	Google Chrome	Software - Google Chrome
wifi	82	179	4.35%	Network		Network
jack	74	174	3.93%	Network		Network

Appendix F (continued)

Token	In Unique HD Requests	Total Instances	% of Unique HD Requests	Category	Sub- Category	Category and SubCat Combined
google_group	69	126	3.66%	Google Groups		Google Groups
wireless	67	172	3.56%	Network		Network
copier	66	114	3.50%	Hardware	Copiers	Hardware - Copiers
academ_web	63	81	3.34%	Software	Academic Web	Software - Academic Web
faweb	62	104	3.29%	Software	FA Web	Software - FA Web
http_academ	60	76	3.18%	Software	Academic Web	Software - Academic Web
credenti	53	62	2.81%	Accounts	Password Reset	Accounts - Password Reset
speaker	52	120	2.76%	Hardware	Classroom A/V	Hardware - Classroom A/V
browser	51	80	2.71%	Software	Web Browser	Software - Web Browser
fax	51	96	2.71%	Phone		Phone
profil_match	50	50	2.65%	Accounts	Creation	Accounts - Creation
video	48	84	2.55%	Hardware	Classroom A/V	Hardware - Classroom A/V
monitor	46	88	2.44%	Hardware	Classroom A/V	Hardware - Classroom A/V
firstclass	45	64	2.39%	Software	FirstClass	Software - FirstClass
drive_access	43	43	2.28%	Google Drive		Google Drive
macbook	43	53	2.28%	Hardware	Mac	Hardware - Mac
exist_profil	42	42	2.23%	Accounts	Creation	Accounts - Creation
googl_drive	40	76	2.12%	Google Drive		Google Drive
iphon	40	48	2.12%	Hardware	iOS	Hardware - iOS
gmail	39	83	2.07%	Gmail		Gmail
ipad	39	79	2.07%	Hardware	iOS	Hardware - iOS
account_creat	38	48	2.02%	Accounts	Creation	Accounts - Creation
sent_firstclass	38	50	2.02%	Software	FirstClass	Software - FirstClass