END-USER NEEDS OF FRAGMENTED DATABASES IN HIGHER EDUCATION

DATA ANALYSIS AND DECISION MAKING

Amanda Briggs

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Master of Science
in Human-Computer Interaction,
Indiana University

Accepted by the Graduate Faculty of Indiana University, in partial fulfillment of the requirements for the degree of Master of Science.

Thesis Committee

_____

Francesco Cafaro, PhD, MS, LS, L, Chair

_____

Lynn Dombrowski, PhD, MS, BS

_____

Khairi Reda, PhD, MS, BS

DEDICATION

For my husband, who listened to me complain and still told me I was amazing.

For my parents, who still think I can do anything.

For my son, please don't arrive until I'm finished editing.

ACKNOWLEDGEMENT

I would like to thank my thesis advisor, Dr. Francesco Cafaro, for his patience and support while keeping me on track. His guidance, motivation and wealth of knowledge were invaluable to me. I could not imagine a better advisor or mentor for this journey.

I would also like to thank my thesis committee, Dr. Lynn Dombrowski and Dr. Khairi Reda. Their insight and guidance helped propel my research into a more nuanced and interesting version of what was just a glimmer of an idea.

Amanda Briggs

END-USER NEEDS OF FRAGMENTED DATABASES IN HIGHER EDUCATION DATA ANALYSIS AND DECISION MAKING

In higher education, a wealth of data is available to advisors, recruiters, marketers, and program directors. However, data sources can be accessed in a variety of ways and often do not seem to represent the same data set, presenting users with the confounding notion that data sources are in conflict with one another. As users are identifying new ways of accessing and analyzing this data, they are modifying existing work practices and sometimes creating their own databases. To understand how users are navigating these databases, the researchers employed a mixed methods research design including a survey and interview to understand the needs to end users who are accessing these seemingly fragmented databases. The study resulted in a three overarching categories – access, understandability, and use – that affect work practices for end users. The researchers used these themes to develop a set of broadly applicable design recommendations as well as six sets of sketches for implementation – development of a data gateway, training, collaboration, tracking, definitions and roadblocks, and time management.

TABLE OF CONTENTS

List of Charts 11
List of Images 12
Chapter One: Introduction 13
Chapter Two: Literature Review 16
2.1 Framing the Data Problem 16
2.2 Big Data Decision Making 18
2.3 Infrastructure and Work Habits 20
2.4 Information Design and Data Visualization 24
2.5 Conclusion 25
Chapter Three: Methodology 27
3.1 Context for the Study 27
3.2 Participants 28
3.3 Research Design 31
3.3.1 Survey 31
3.3.2 Interviews 32
Chapter Four: Results 34
4.1 Survey Data Analysis 34
4.2 Interview Coding 41
4.2.1 Access 41
4.2.2 Understandability 43
4.2.3 Use 47
4.3 Thematic Analysis with Affinity Diagrams 51
4.3.1 Access 51
4.3.2 Understandability 54
4.3.3 Use 61
Chapter Five: Discussion 69
5.1 Design Recommendations 69
5.1.1 Access 69
5.1.2 Understandability 70
5.1.3 Use 71
5.2 Sketches 71
5.2.1 Gateway 72
5.2.2 Training 76
5.2.3 Collaboration 78
5.2.4 Tracking 81
5.2.5 Definitions/Roadblocks 82
5.2.6 Time 84
Chapter Six: Conclusion 87
Appendices 91
Appendix 1: Recruitment Materials 91
1.1 Survey Request – Internal 91
1.2 Survey Request – External 92
1.3 Interview Request from Survey 93
1.4 Interview Request – Database Administrator 94
Appendix 2: Informed Consent 95

2.1 Informed Consent – Survey 95
2.2 Informed Consent – Interview 100
Appendix 3: Affinity Diagrams 105
3.1 Data source 105
3.2 Talking to other systems 105
3.3 Data definitions/coding 106
3.4 Asking the right question 106
3.5 Experience 107
3.6 Relationship building 107
3.7 Reporting/analysis 108
3.8 Decision making 108
3.9 Time 109
References 110

LIST OF CHARTS

Chart 1: Education level of survey participants 30
Chart 2: Education level of interview participants 31
Chart 3: Identifying trends 36
Chart 4: Time spent monthly analyzing and reporting on data 38
Chart 5: Number of datasets accessed to create reports 39
Chart 6: Amount that users collaborate with other users in their unit 40

LIST OF IMAGES

Image 1: Proposed gateway solution for end users 74
Image 2: Proposed training solution 77
Image 3: Profile 80
Image 4: Data Definitions/Roadblocks 84
Image 5: Data Definitions/Roadblocks 84
Image 6: Workflows, Time, and the Data Ecosystem 86
Image 7: Workflows, Time, and the Data Ecosystem 86

ABSTRACT

In higher education, a wealth of data is available to advisors, recruiters, marketers, and program directors. However, data sources can be accessed in a variety of ways and often do not seem to represent the same data set, presenting users with the confounding notion that data sources are in conflict with one another. As users are identifying new ways of accessing and analyzing this data, they are modifying existing work practices and sometimes creating their own databases. To understand how users are navigating these databases, the researchers employed a mixed methods research design including a survey and interview to understand the needs to end users who are accessing these seemingly fragmented databases. The study resulted in a three overarching categories – access, understandability, and use – that affect work practices for end users. The researchers used these themes to develop a set of broadly applicable design recommendations as well as six sets of sketches for implementation – development of a data gateway, training, collaboration, tracking, definitions and roadblocks, and time management.

1. INTRODUCTION

Higher education encounters a wealth of data. From recruitment and enrollment data to current student and alumni data, accessing and managing data has become incorporated into everyday work practices. For the end users who work with this data regularly, the magnitude of available data available can be overwhelming. Currently, at a large, midwestern university there exist multiple ways for accessing and analyzing this data. These data sources can often appear to be fragmented from each other, each containing disparate data that does not communicate with other systems.

As we collect more data, we are better equipped to store data than we are to analyze it. [30] For end users, who access data regularly for their work, navigating these seemingly fragmented databases and analyzing the data they find can take an excessive amount of time, especially without proper training. Additionally, college and university leaders do not believe their institutions are making good use of the data available. [9]

End users, whose roles are typically advisors, recruiters, marketers, and program directors, are increasingly tasked with working as data analysts for decisions makers within their school or unit on their campus. While these users receive no formal training, they are asked to consistently access and analyze data for decision makers within their units. End users create reports that affect programs, recruitment, enrollment, marketing, student experience, and curriculum decisions. They are expected to produce

"high-confidence results" that further the credibility of their unit and overall their campus and institution. [32]

While the demand for data is increasing, end users are still navigating databases with minimal training and modifying the data retrieved or supplementing it with their own "homebrew databases." [24] To understand how users are accessing and navigating institutional databases, the researchers created a mixed-method research design including a survey and interview. Both the survey and interviews were designed to determine how users were accessing and analyzing data, the effect this has on work practices, and the demands of gathering and reporting on data. While the primary focus of this study was the end user needs while interacting with fragmented databases in higher education, the study design also included interviews with database administrators. The main goal of this study was to identify a set of themes that include the challenges and opportunities that occur in typical usage scenarios and craft a set of design recommendations that can be implemented.

The goal of this research was to identify a set of themes that include the challenges and opportunities that occur in typical usage scenarios and craft a set of design recommendations that can be implemented. The research was organized into three aims which included the following:

- Aim One: Identify data querying and visualization best practices and common operations focusing on incongruent and messy data sets.

- Aim Two: Identify all data sources used and data-driven decisions by marketing, recruitment and student support professionals at IUPUI, and other schools and universities.
- Aim Three: Develop design recommendations for a system that combines data and data visualization recommendations for higher education data systems.

The results of this study highlighted how end users and database administrators approach the following categories: relationship building, talking to other systems, data definitions/coding, asking the right question, data source, experience, reporting/analysis, decision making, and time. From these, findings, the researchers developed six design recommendations around the development of a data gateway, training, collaboration, tracking, definitions and roadblocks, and time.

The next chapters are organized as follows: In chapter two, a literature review is performed; chapter three details the mixed-method design of the study; chapter four highlights the results of the study, including the survey and interview results; chapter five uses the results to make a series of design recommendations for the current system; and chapter six summarizes the results and discussion in the conclusion, including the future work that could be accomplished.

2. LITERATURE REVIEW

Many organizations work with large, messy data sets and higher education is no exception. Bringing order to large datasets has been a subject of study in recent years as an ever-growing wealth of data is available to users. Researchers have sought to determine best practices for wrangling [20] these large data sets into manageable systems. As best practices have emerged across businesses, these are being adapted and implemented by higher education for the purposes of recruitment and marketing. However, as higher education continues to gather and collect increasingly wider swaths of data, it has become necessary to implement more refined data models to allow everyday users to keep up with all of the data available to them. "Our ability to collect and store data is growing faster than our ability to analyze it." [30] These broad data sets also create an impression of fragmented databases that don't work together creating additional work for users who must maintain their own "homebrew databases" [24] through a combination of existing databases and spreadsheets maintained by individual users.

2.1 Framing the Data Problem

As a society, we collect increasing amounts of data with each digital interaction. However, the accumulation of this data offers more questions for those who have to work with it daily – what does this data mean? What sort of analyses should be carried out on this data? [4] These questions are not always simple to answer. Large data sets pose questions about "the constitution of knowledge, the processes of research, [and] how we should engage with information." [4]

As organizations around the world continue to collect more data, industry leaders are questioning how to better utilize the knowledge that can be gleaned from this data. When top-performing companies "use analytics five times more than low performers" [13] it becomes imperative that organizations learn to excel at leveraging the data they can already access. And these companies are also using a centralized unit as the primary source of data and analytics in the organization. By creating a principal source for data and analytics, companies are becoming more effective in how they reach their audience.

Higher education is faced with the need to keep up with organizational demands [13] and yet has "applications, platforms and databases that do not 'talk' with one another and are difficult to integrate." [9] As higher education continues to transition to a recruitment-focused mindset [16], the need for skilled professionals who understand complex data sets and also databases that are designed to "talk" to one another is imperative. However, as advisors, recruiters and marketers spend increasing amounts of time "wrangling data" they are taken away from the primary duties of their job. The term "data wrangling," coined by Sean Kandel, represents the amount of the workday spent determining what data is available, organizing data, and understanding how to combine multiple sources of data. [20]

Since 1978, researchers have been discussing how to make databases and information systems easier to use for non-technically trained users [22], like the advisors, recruiters and marketers who are wrangling massive data sets today. Because of the volume of data

student success-focused researchers often wade through, the question of why students are choosing an institution is often lost [5] in the cacophony of data sources. On top of this, many institutional higher education leaders do not believe their college or university is making good use of the data it has available. [9]

Through data mining and predictive modeling, universities can discover the "hidden trends and patterns and mak[e] accuracy based predictions through higher levels of analytical sophistication." [16] These sorts of predictive analytics involve an extensive knowledge of mathematics and statistics that are typically employed by institutional researchers rather than school- or unit-level professionals tasked with identifying trends through detailed data analysis. Additionally, the task of wading through large datasets to begin creating meaningful visualizations and reports creates additional work for users. "The visual representation of data reduces the cognitive work needed to perform certain tasks." [31]

2.2 Big Data Decision Making

Big data can be defined as extremely large data sets that require computational power to analyze and identify trends and associations. As big data becomes more ubiquitous, it offers the opportunity to understand the world in a different way and to enable us to make more nuanced decisions. [15] In higher education this is more important than ever as universities and colleges are using data-driven decision making in marketing and recruitment. As the focus on attrition, graduation rates and interventions increases, the data used need to be accurate and timely [17]. Additionally, the database needs to be able

"talk" to other databases [9] without creating unnecessary additional work for the users analyzing the database.

Examining the use of big data in decision making, it is necessary to take an expanded look at big data. Traditionally, big data has been categorized with "three v's" – volume, velocity and variety. This can be expanded to include veracity, verification, value, validity and volatility [5]. Veracity is the trust analysts and decision makers have in their data. Verification includes corroboration and security. Value refers to the insights the data can provide. Validity is the accuracy of the data. And volatility refers to the longevity of the data. While it is important to focus on targeted interventions for students, these principles can be applied to both market and recruitment research in higher education.

Specifically examining recruitment market data and analysis across industries reveals that most analysis occurs by domain-specific experts rather than relying on large-scale models that can delineate "fine-grain trends" [26]. In higher education, at the school and unit level this sort of analysis is performed by a domain expert who examines multiple databases, rather than employing a data analysis model created for the specific marketing and recruitment needs of each school or unit.

In higher education and job recruitment, data analysis models have been created that bring large amounts of data into alignment to allow for the easy identification of trends and insights on specific populations. eduMRS is a framework developed by Qiu, et al.,

that captures real-time public opinion of higher education institutions by "retrieving, aggregating, quantifying and visualizing public opinions." [19] While this system is focused on the rankings of institutions, it represents a large-scale effort to bring data in alignment. In a similar recruitment domain, Zhu, et al., developed a novel data analysis model that can dynamically generate recruitment topics and discover recruitment market trends for job recruiters and job seekers without the need of a domain expert [26]. Efforts like these free up analysts' time and create databases that are able to "talk" to one another without a domain expert taking the time to bring the data into alignment.

2.3 Infrastructure and Work Habits

As everyday data users in schools and units are tasked with analyzing increasingly larger sets of data and access multiple databases, they are asked to play the role of data analyst for decision makers in their units. This role requires work that is exploratory, but also demand-driven in accessing trends and assessing audiences. These ad hoc data analysts are trying to communicate clear needs and goals for the organization while producing "high-confidence results" that further the credibility of the unit and the institution. Additionally, "the work creates a strong need to preserve institutional memory, both by tracking the origins of past decisions and by allowing repeatability across analyses." [32]

In non-profits like higher education institutions, users are often employing a mix of institutional databases and homebrew databases. "Homebrew databases," a term coined by Amy Voida, are data repositories that are often created in nonprofits and contain a diversity of information needs, stakeholders, and work contexts and often have

constraints of time, funding and expertise. [24] Because databases and data systems are often fragmented from each other in nonprofit, low-budget organizations, data administrators are forced to develop ad hoc homebrew databases. These databases are often an amalgamation of results from other databases, paper forms and event data. Database administrators have to enter and analyze information across multiple information systems that provide different functions to the administrator and the organization.

Bergman, et al., found that users consolidate project information when a system encourages it and "store and retrieve project-related information items in different folder hierarchies (documents, emails and favorites) when the design encourages such fragmentation." [2] Users relate to project information in a specific way. The creation of homebrew databases in higher education suggests that users are taking cues from the database design and creating work arounds for the fragmentation that exists across the multiple databases that they access regularly.

Data fragmentation can lead to database users feeling disempowered by the data they have access to. When users feel disempowered by data, as found by Bopp, et al., they experience an erosion of autonomy, data drift and data fragmentation [3]. An erosion of autonomy for data administrators is caused by a lack of consensus around metrics, a lack of understanding about the types of data to use for different purposes and prioritizing staff time for data work. Data drift occurs when "organizations change the kind of metrics that they collect and manage over time as their organizational identity, as

represented by data, evolves." [3] Finally, data fragmentation happens when organizations have a wealth of data that are not "connected to each other in any systematic way." [3]

When users encounter new data, they struggle to find a place to update it within existing systems that aren't designed for expanding with new data points. In 2015, Pine and Mazmanian assessed the data being used by healthcare providers. [18] The researchers found that it was cumbersome to update records. As new data demands were identified, the users had to create solutions around the newfound data. "Data must be located, quality of data assessed, idiosyncrasies identified, and inevitably a range of nearly intractable problems are discovered in the 'real messy world of data.'" [18] This experience can be translated to anyone working with fragmented databases that do not yet talk to each other. Another encumbrance the researchers found was the tension between quality and availability of data. While ground truth data was preferred, the time and effort associated with accessing the ground truth data was a hindrance to organizational effectiveness. Workarounds were developed to create another kind of access for analysts.

While the data systems and roadblocks encountered by data analysts have the potential to leave users feeling disempowered, they also have the potential to create collaborative work. Dantec and Edwards examine the notion of scale in cooperative systems, specifically looking at how hierarchies influence systems with a large number of users. [14] Often, these cooperative systems cross the boundaries of local, regional and national contexts, echoing the breadth of communities encompassed by large-scale higher

education institutions – units and schools, campus administrators and university administrators.

In 2016, Amy Voida found that interviewees in middle- and upper-management positions had similar experiences of quantitative data being desired for decision making while qualitative data "helped people 'connect emotionally' with the organization." [23] The data collected in these organizations was found to drive decision making, providing "credibility to organizational actions." [23] However, Voida posits that this is one of the myths of big data – hard numbers provide greater clarity and accuracy. Without the human stories behind the data, organizations can experience a shift in their identity, based on data collected and monitored over time – data drift.

For nonprofits, the solution for systemic data fragmentation does not reside with the individual organizations, but in the policy fields surrounding the data itself. [1] Researchers made the recommendation of banding together with other organizations to address data needs, rather than approaching them as individual organizations. The researchers felt that this would lead to systemic changes in data organization rather than individual organizations continuing to create workarounds suited to their particular needs. As education becomes increasingly reliant on digital data, it uses data to understand and predict human behavior, particularly potential student behavior. [25] These data technologies are becoming a part of university policy, creating a culture of data-based decision making in education.

2.4 Information Design and Data Visualization

As universities adopt a culture of data-based decision making, developing data analysis reports becomes a far more complicated process than just downloading data and organizing it into a report. Even the steps of preparing data for visualization involves a multi-step process that includes discovery, wrangling, profiling, modeling and reporting. [11] The most complicated part of this process is often the discovery and wrangling. "Visualizing the raw data is unfeasible and rarely reveals any insight. Therefore, the data is first analysed." [30] Even after preparing data to create visualizations, "reconstructing a repeatable workflow is difficult without a coherent linear history of the operations performed." [11] As analysts struggle to create meaningful, replicable visualizations, they also have to make sure they are communicating the patterns found in data. Meaningful information design is critical as "expert" analysts transfer information to decision makers via reports and dashboards. [6]

Reports and dashboards hold the ability to communicate powerful truths through visualizations for decision makers. However, the remarkable amount of information we are accruing often fails to be translated into meaningful visualizations. [7] While it seems obvious, data visualizations that achieve specific objectives are of most use to decision makers. All of the bells and whistles associated with data visualization software are often more distracting than insightful.

Kandel, et al., argue that the improvement of data visualization systems can help data analysts more effectively wrangle data. These systems would include a combination of

data verification, transformation, and visualization. [27] An improved visualization system could help domain experts spend less time wrangling data and more time working on their domain specialty. “Visual data exploration is especially useful when little is known about the data and the exploration goals are vague. Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals might be done automatically through the interactive interface of the visualization software.” [31] This could prove especially useful for novice users of data visualizations who struggle to identify appropriate views, execute appropriate interactions, interpreting visualizations and matching their expectations of the system with the reality of the visualizations. [29]

Considering the design process from initial data discovery to downstream data presentation and reporting can prove a useful solution for those in the position of reporting on data. For those in higher education, especially advisors, recruiters and marketers, working to discover the real-world problems they try to solve with data visualizations will improve the work environment for these ad hoc analysts. However, proving the profitability of a given data system remains the key driver to system-wide integration. [28]

2.5 Conclusion

Higher education gathers a wealth of data from potential students, applicants, enrolled students, and eventually alumni. As colleges and universities continue to gather data, the data that can be collected and stored is outpacing the ability of those who work there to sort through and analyze the data. Wrangling this abundance of data often falls on

domain experts, rather than those trained in data analysis. Often these domain experts are required to bring multiple datasets into alignment, which requires an investment of time and adapting work habits to acquire and analyze data.

While many leaders of higher education institutions do not believes their college or university is making good use of the data is has available, there does not seem to be a widely accepted and adopted way of accessing and analyzing data that encompasses both institution and unit-level solutions. Reports and dashboards have the ability to communicate data truths for decisions makers, but are not widely available across all data systems utilized by unit-level data analysts. The amount of information accrued by colleges and universities fails to be translated in meaningful visualizations causing analysts to create “homebrew databases” that satisfy needs within their unit.

The depth of study around big data and data visualization can be applied to higher education, however, a lack of across-the-board standards and adoption of said standards leaves a gap in the research around the use of data, how it affects work practices and the user’s experience around data access in higher education.

3. METHODOLOGY

The focus of this study was to explore the end-user needs of higher education database analysis and decision making by identifying a set of themes that include the challenges and opportunities that occur in typical usage scenarios. A mixed methods approach was used to investigate and determine the needs of student support and marketing professionals at colleges and universities who access multiple databases.

The mixed-method design primarily relied on qualitative interview data supplemented with quantitative and qualitative survey data to identify all data sources used and data-driven decisions by marketing, recruitment and student support professionals at Midwestern colleges and universities. Participants were initially recruited through the survey which was distributed by email based on publicly available directory information at their college or university of employment. Survey participants indicated if they were willing to be contacted for a follow-up interview. Additionally, database administrators were contacted to participate in an interview to understand their role and the future of these database systems.

3.1 Context for the Study

This study specifically examined the end-user needs of higher-education databases. The research methodology was designed to understand how users regularly access and utilize these databases, including work practices, the demands of data gathering, and the internal demands placed on users within their unit of employment within the college or university. While end users were the primary focus of this study, semi-structured qualitative

interviews were conducted with database administrators to understand the creation and refinement of university-based databases (often known as institutional research), the use of these databases, and the potential future for these databases.

3.2 Participants

Participants consisted of database users from five midwestern colleges and universities – Franklin College, Indiana State University, Indiana University, Purdue University, and Rose-Hulman Institute of Technology. Participants were recruited by email based on publicly available directories at their respective colleges or universities and the potential relationship of their position (advisor, recruiter, data coordinator, enrollment manager, marketing roles) with database use. *(Please find all recruitment materials in Appendix 1.)*

A total of 18 users participated in the survey, and ranged in age from 18 to 55 and consisted of five men and 13 women. All participants held an associates degree or higher and worked full-time at their college or university (see Chart 1). Participants self identified their roles which included: recruitment, advising, marketing, program coordination (recruitment, advising, student success, etc.), graduate admissions operations & advising, student and administrative services, marketing and recruitment, programming, and data and IT.

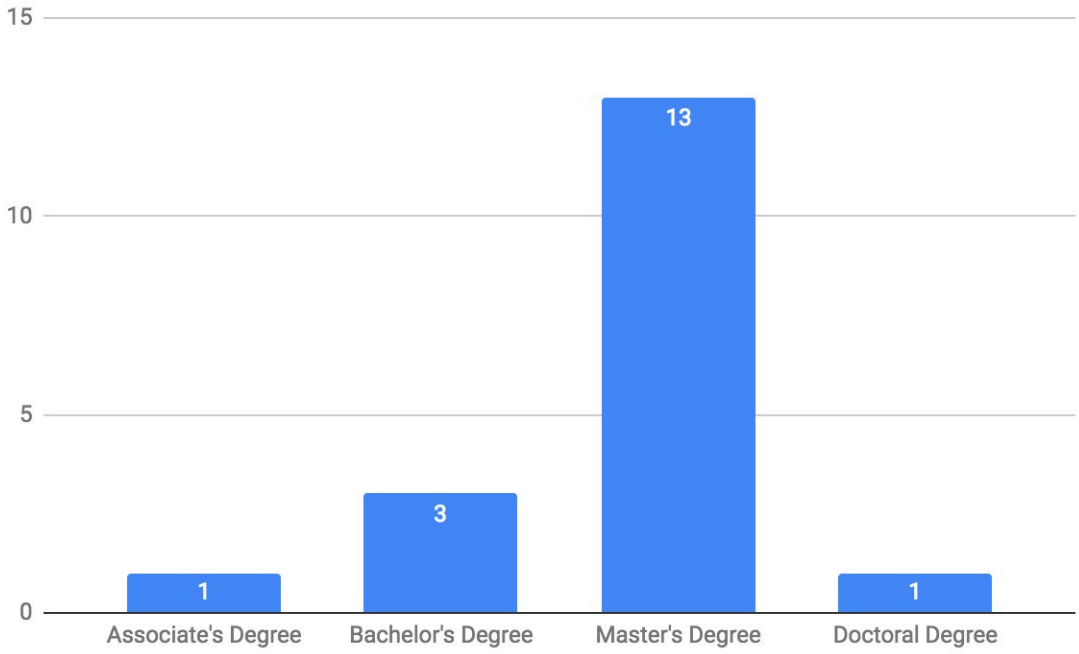


Chart 1: Education level of survey participants.

Of the 18 survey participants, the highest level of educational attainment for all participants was a degree. The majority of participants hold a master's degree and all participants work full-time in their role.

The 10 interview participants ranged in age from 26 to 55 and consisted of five end users and five database administrators. Interviews were conducted both in-person and by telephone. End users were recruited by volunteering to be contacted for a follow-up interview after completing the survey. The average age of end users was 37.2 years old with three interview participants identifying as female and two identifying as male. Their employment roles included academic advisors, program directors and admissions managers. Database administrators were recruited by email. Interview participants included three women and two men, with an average age of 46.8 years old and all participants held a bachelor's degree or higher, with three of the five participants holding

a doctoral degree. Database administrators employment roles included two directors, an assistant director, an associate director and a departmental vice chancellor.

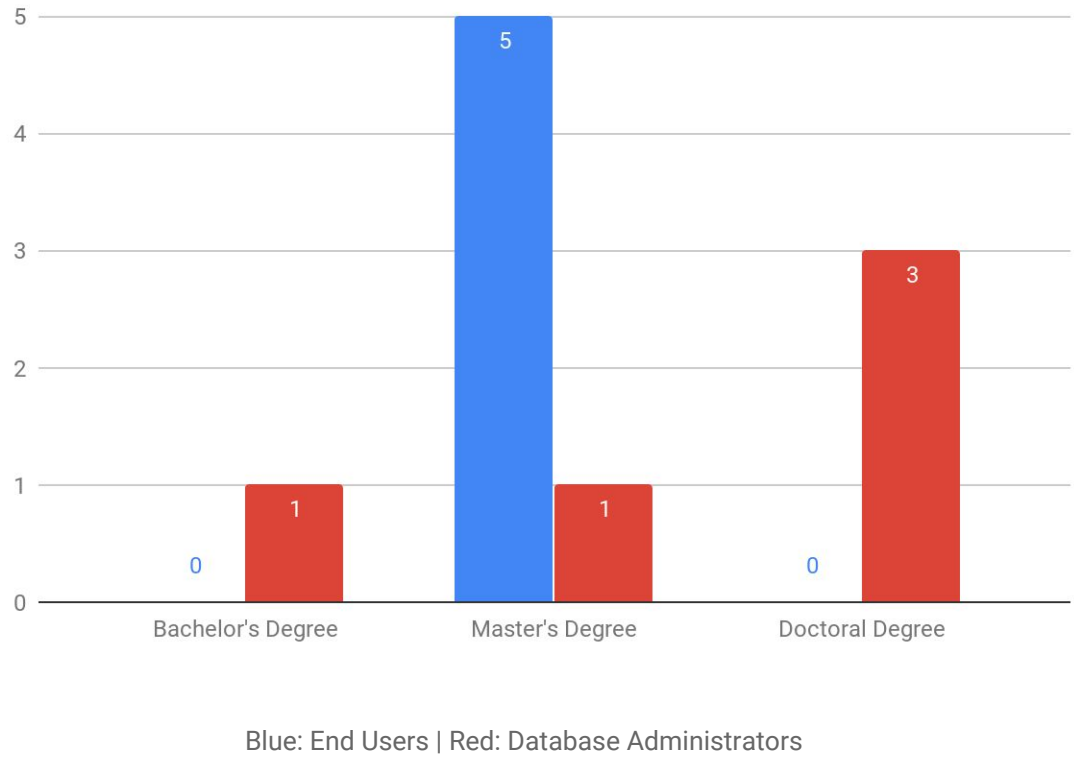


Chart 2: Education level of interview participants.

In the blue, on the left, end users identified their highest level of educational attainment. The red bar on the right represents the highest level of educational attainment of database administrators. All end users interviewed hold a master's degree, while three of the five database administrators interviewed held a doctoral degree.

3.3 Research Design

3.3.1 Survey

A survey was distributed to database end users. Participants included end users at five midwestern colleges and universities. The participants were briefed about the nature of the study in a recruitment email found in Appendix 1. Before beginning the survey, the participants read and agreed to an informed consent (Appendix 2) before being allowed to proceed with the survey. Participants then completed an online Google Forms survey including questions about the participants' use of databases at their respective institutions, data analysis and reporting, and the convenience of the data source, as well as demographic questions. At the end of the survey, participants were asked if they were willing to participate in a follow-up interview.

Database use questions included quantitative questions about the types of database systems used and how users identify trends in data. Following these questions, users were asked a qualitative question: "If you manipulate the data from any of the systems listed above, why do you do it?" Following this question, users were asked to identify what form their data analysis took post manipulation: "What format does the data take when you are finished? For example, is it a series of charts or a report with charts where you provide written explanation." Further, users were asked about the time they spend analyzing and manipulating data, the ease of finding all of the data they need in one place, the data sets needed to report on data, and how the data they report on is used. Users were also queried about the challenges they face when collaborating: "What are the challenges you generally encounter when collaborating with others outside of your unit?"

Finally, users were asked how they would improve functionality in the systems they use: "If you could improve any function in these systems, what would it be and which system(s) would it apply to?"

3.3.2 Interviews

Ten interviews were conducted with end users and database administrators at a large midwestern university. Interviews were semi-structured and took about 20 to 45 minutes to complete either in person or by phone. The semi-structured interview style created a consistency in the questions asked, but allowed the interviewer to follow up with clarifying questions when necessary. All interviews started with an overview of the study and an assurance of anonymity and confidentiality. A few interviews lasted longer than expected, but interviewees were comfortable continuing the interview. End users were recruited for interviews by opting in be contacted for a follow up interview at the end of the survey. Database administrators were recruited by email (found in Appendix 1).

End users and database administrators participated in semi-structured interviews with predetermined questions. End users interview questions included a series of questions to understand which database systems users engage with and what sort of information they provide for the user. Asking what users do with the data once they procure it was essential for understanding the work practices of users: "Do you use the data as it is provided? Or do you manipulate or combine the data?" and "How often do you perform analyses on the data you are accessing? What sort of analyses do you perform?" Following this inquiry users were asked about how they report on the data they access,

who they provide the reports to, how the reports are used, and the types of data used to create these reports. Additionally, to understand the full scope of report creation, users were asked: "Can you find all of the data you need for these reports in one place? How many data sets do you need to access on average? How much time does this take on average?" The interviews ended with a series of questions about what users like and dislike about the systems they use, what they would change if they could, and what they would like if they could have anything to make their work easier. Additionally, users were asked about their experience in data analysis and assessment.

Database administrators' interviews began with their history of working with the database system and what purpose the system serves. After asking about where the data for the system originates and the primary use of the system, database administrators were asked about the intended and current audience for the system: "Who was your intended audience for this system? Has that changed since the system was launched?" The next questions involved training – how often training is offered for users, how many people are trained to use the system, and what the training entails. Additionally, database administrators were asked about how often they see the system being used. Following this, participants were asked about what they feel the system does well and what could be improved. To understand how user testing and feedback is incorporated, administrators were asked: "When you create improvements for the system what sort of user testing/user feedback do you seek out?" Following this question, participants were asked if there were any plans to enable their system to "talk" to other systems employed by end users.

4. RESULTS

The research design was created to study to end-user needs of fragmented databases in higher education analysis and decision making. The mixed methods approach employed a survey consisting of quantitative and qualitative questions successfully completed by 18 respondents at five midwestern colleges and universities coupled with a semi-structured interview with ten participants consisting of five end users and five database administrators.

The survey consisted of end users who were asked about the databases they use regularly and their experience with accessing, acquiring and analyzing data and collaborating with other end users and decision makers within their unit and at the administrative level. Five interviews were conducted with end users to better understand their needs when it comes to data access and analysis, as well as decision making. An additional five interviews were conducted with database administrators to understand the decisions made around data access and visualizations including how their relationships with end users and university administration decision makers affects database use and modifications. The analysis of our results provided three categories of tasks and related challenges that end users and administrators encounter regularly when accessing data, collaborating with others, and presenting data for decision making purposes.

4.1 Survey Data Analysis

The survey consisted of 18 questions and eight demographic questions. Participants identified the database systems they used to most frequently access and analyze data.

When identifying trends in student data 66.66% of survey respondents create reports on trends by creating their own reports, using a combination of reports from university-created data sources, and reports from their own internal databases.

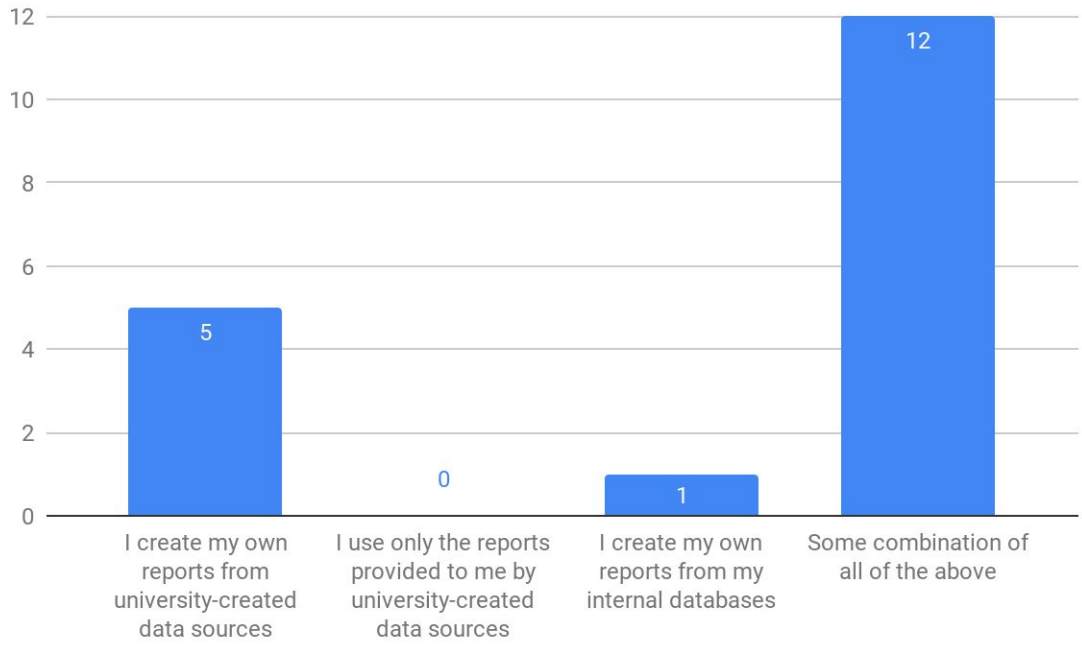


Chart 3: Identifying trends.

Survey respondents identified how they compile data into reports for decision making purposes in their unit. The majority of respondents identified that they create use a combination of data sources and reports to compile trend reports within their units. All respondents to the survey were end users at five midwestern colleges and universities.

When asked in the survey if participants manipulate data from the sources they listed, and why, survey responses included answering specific questions, to refine data for reporting, merging different data sets to tell one story, and to understand trends while there is still time to make a difference in recruitment and enrollment.

[S1] "To get specific answers to questions regarding the program I oversee or to build visuals/graphs to support trend analysis."

[S2] "To adjust data. Example for admissions there may be a student who deferred that was not included in the report that was initial ran therefore I would have to add the additional student into head count therefore requiring a manipulation of data."

[S3] "Yes, to assure I'm getting a diverse set of students in regards to degree, ethnicity and gender."

[S6] "Some is a merging of the different data elements/parameters that are not available in just one system and organizing to tell the story of our findings."

[S12] "Because I can't get the numbers I am looking for from a university created report until it's too late to move the needle."

More than half of the survey participants reported spending between one to five hours each month analyzing and reporting on data. Five of the 18 participants reported that they spent less than one hour analyzing and reporting on data; 10 of the 18 participants spend one to five hours each month; three participants spend between six to 10 hours on data each month; and no users report spending more than 10 hours a month on data analysis and reporting.

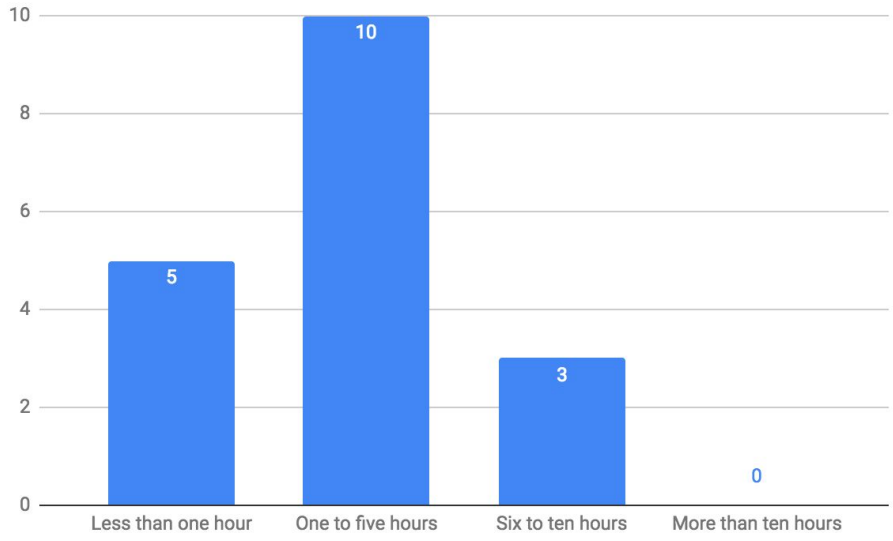


Chart 4: Time spent monthly analyzing and reporting on data.

Survey respondents identified the amount of time they spend monthly analyzing and reporting on data. The majority of respondents, all end users from five midwestern colleges and universities, indicated that they spend one to five hours each month analyzing and reporting on data.

When the survey participants were asked if they reported on the data they accessed for decision-making purposes within their unit or the university 14 of the 18 (77.77%) respondents answered yes. Half of users (nine of 18) access two datasets to create reports for decision making, with 14 of the 18 respondents accessing two or more datasets.

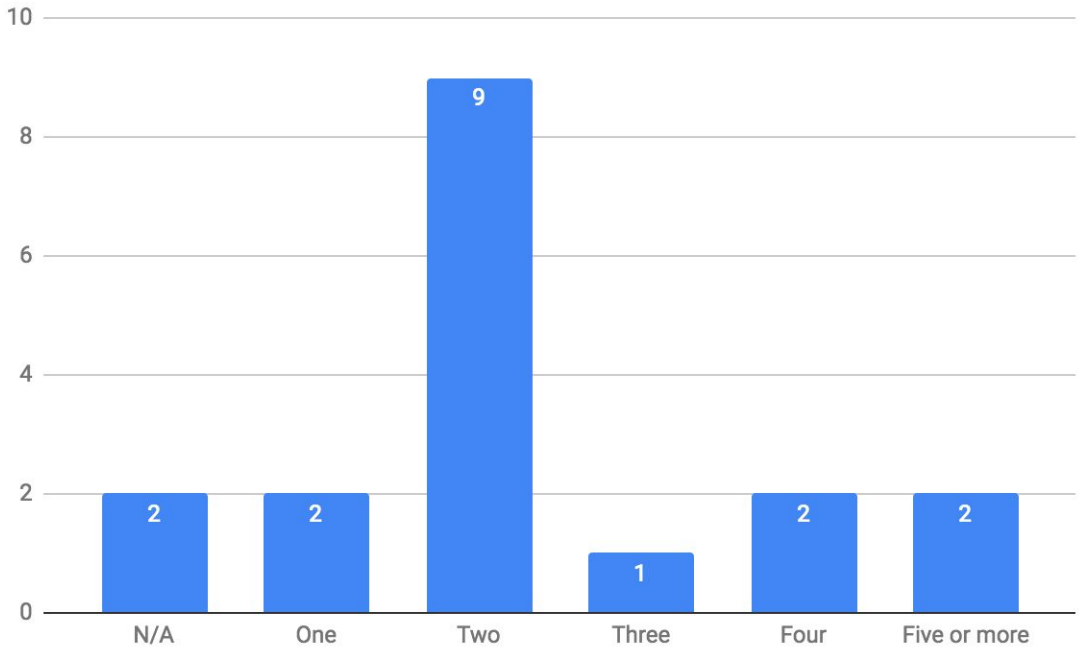


Chart 5: Number of datasets accessed to create reports.

Survey respondents indicated that they access at least two datasets to create reports within their unit.

When users were asked how often they collaborated with other users inside of their units, 12 of the 18 respondents indicated that they collaborated with other users within their unit at least once a month.

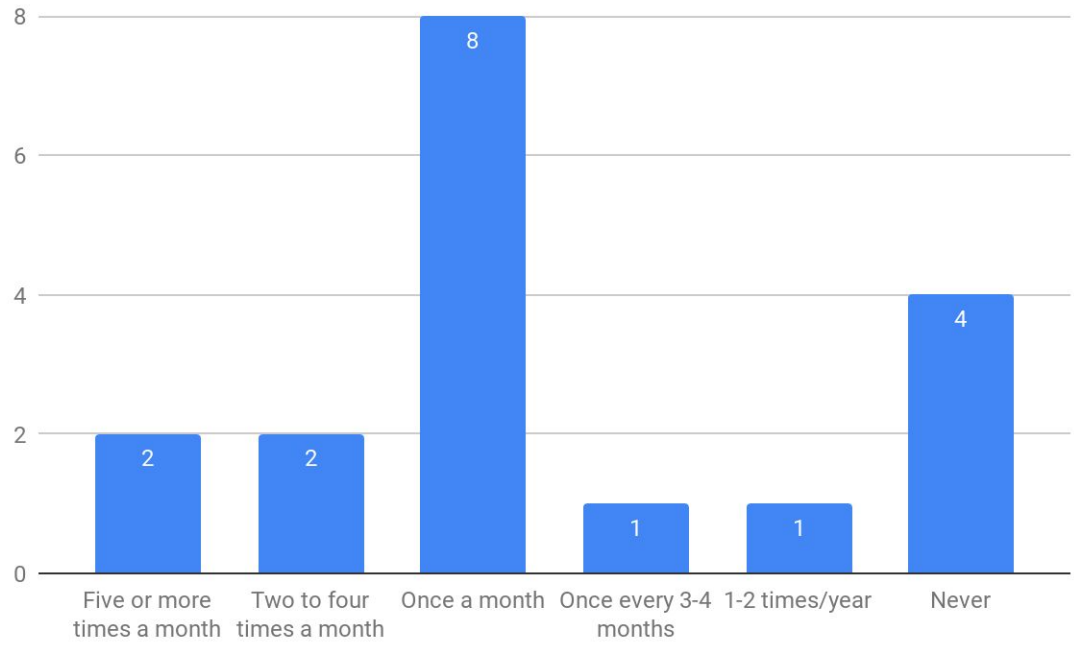


Chart 6: Amount that users collaborate with other users in their unit.

Survey respondents indicated that they collaborate with other users in their unit to access and report on data. The largest number of users (eight) indicated that they collaborated once a month with other users in their unit, four survey respondents indicated that they never collaborated for acquiring and analyzing data.

Survey respondents indicated that there were various challenges they encountered when collaborating with other users that they work with regularly. Included in the responses were a lack of consistent training, clarification on data sources, a lack of task clarity from supervisors, access, and familiarity with the systems.

> *[S1] "Some offices rely on internal databases (Excel) and have not trained on the campus-wide resources. That could mean we see wildly different information. It's better when all parties pull from the consistent place."*
>
> *[S2] "At times trying to get clarification as to what type of information is needed in order to decide which system to use to retrieve data."*

[S6] "Significant lack of knowledge about how/where to access data, lack of understanding what parameters to pull and analyze, and general inability manipulate data sets to get the desired output"

[S7] "Clarity in my data needs - the ability to state what I need clearly enough so that the data returned is what I need."

[S13] "Not everyone has the same access to and understanding of the reports."

[S16] "Familiarity with the system and functionality. Most users only use the very basic functions, leaving the more in depth work or reporting to directors."

Survey respondents indicated a variety of challenges when collaborating with others outside of their unit on accessing and analyzing data. Some of the challenges included ownership of content, knowing who to contact for data, scheduling conflicts, data not matching, unique needs, and access.

[S1] "Different units 'owning' content."

[S2] "Getting hold of someone to learn how to retrieve data that our department may need."

[S6] "Most generally it is scheduling conflicts as their primary responsibility is to their unit."

[S9] "Data does not always match."

[S13] "Each academic unit has unique variables that cannot be accounted for in one central system."

[S16] "Access is generally the largest challenge. Salesforce (TargetX) is not a campus wide program, it's just used in Admissions, so if we are working with someone in another department the terminology, and reporting procedures can cause confusion."

4.2 Interview Coding

The researchers used thematic analysis in psychology to evaluate the data from the interviews. After completing the interviews, the researcher transcribed all of the interviews for analysis. Following transcription, the researcher coded the interviews and identified three overarching themes: access, understandability, and use.

4.2.1 Access

A user's ability to access data, or their perception of their access to data, is essential to their ability to engage in use of the data systems available to them. When a user does not trust a data source or understand how a data source relates to the data within the ecosystem of the university as a whole, this leads the user to feel less credibility in the data provided as a whole. In these campus- and university-wide systems, there is a concern that data systems are not communicating with each other, which causes an additional sense of limited access to data sources. Within this overarching theme, two sub-themes were identified: data source and talking to other systems.

4.2.1.1 Data source

Standardization and consistency were common themes that emerged among comments on data sources. End users worried about credibility when pulling from different sources or finding conflicting answers with someone in their unit.

> *[P9] "There doesn't seem to be an agreed upon consistency of where the data should come from when we're making consistent decisions."*

Database administrators and end users both commented on the "shadow databases" that exist across units. These "shadow databases" are colloquially defined as the spreadsheets that a single user or single department will maintain. These databases are not shared widely and are used to understand data trends or datasets that cannot be accessed otherwise. Both administrators and users understood the need for these databases, with one end user mentioning that sometimes these databases will meet unit needs better than institutional resources.

> *[P10] "It seems like a lot of schools on campus have a 'shadow database.' A lot of it may be replicated in places like AdRx, but these databases match their needs a little better than what the university provides."*

A database administrator reiterated the occurrence of these "shadow databases," adding that some end users do not see the benefit of sharing this data or collaborating with institutional resources.

> *[P1] "There are a lot of folks where the 'spreadsheet on my hard drive' method has been working for 20 years, and they don't necessarily see the benefit of, you know, being able to share that data."*

4.2.1.2 Talking to other systems

A common refrain heard from end users and even some database administrators was the lack of communication between multiple campus- and university-wide systems, so that users felt they were not seeing consistency in the data. An end user mentioned that current platforms do not interface well with each other.

> *[P10] "I know the university wants us to use Salesforce, but it doesn't communicate with other things. Sometimes it's quicker to say this is going into Excel."*

A database administrator reiterated this sentiment, mentioning that in their role they could improve on utilizing and combining all the data sources available on campus.

> *[P5] "We could improve upon leveraging all the different data sources that are available on campus, like incorporating things coming from vendors like Academic Analytics and EAB."*

4.2.2 Understandability

Often users are working from a viewpoint that is influenced by their understanding of data definitions and their level of university work experience. Database administrators mentioned that some users operate under colloquial definitions specific to their experience or their unit as to what a coding term means, rather than using the university-specific definition. This often leads users trying to ask specific questions of the data that need to be refined by database administrators. Users with a higher level of

experience tailor their work practices around data based on the question they are trying to answer. This analysis led to three sub-themes: data definitions/coding, asking the right question, and experience.

4.2.2.1 Data definitions/coding

Database administrators often spoke about the need for common definitions among datasets. For example, one administrator mentioned that end users often use colloquial definitions of different common terms, not always abiding by the institutional definition, which can cause confusion.

> *[P4] "People can sort of operate from a vernacular definition of what 'enrolled' means, for example, or what 'first generation' means. But without being able to ask specifically what it means for the data owner, or the steward, can influence how it is interpreted at the back end when it is queried."*

Another administrator mentioned that she wanted consistent definitions:

> *[P3] "a consistency of definitions across campus."*

End users mentioned the consistency in data did not occur because different data sets were wrong, but that most end users operate with varying definitions of what data to be pulling, which could be caused by a lack of institutional knowledge.

> *[P9] "So there's a lot of institutional knowledge that if you don't have, you're totally at a disadvantage when you try to begin pulling reports."*

The same end user elaborated on wanting the university to address data definitions by building in more roadblocks – or internal checkpoints sanctioned by the university – to ensure everyone was at the same level when reporting data:

> *[P9] "I think building in a few more roadblocks to help with the validity of what we're using these datasets for, that would really take it to the next level."*

A few end users mentioned that different interpretations of data could lead to credibility issues within their unit. A database administrator mentioned that this happens at both the unit and institutional level.

> *[P1] "You'll be able to pull one number out of enrollment in a certain program, and then somebody else will have their number of enrollment in a certain program. And if it's even, like two or three off, you start fighting about whose number is right."*

Subsequently, an end user verified this experience saying that the inconsistencies are frustrating, even creating unnecessary credibility issues.

> *[P10] "Inconsistencies in the numbers are frustrating. It's off by just a handful. In academia, that creates a credibility issue."*

Another end user commented that these small inconsistencies cause units to question the validity of the data itself, when really it is a lack of institutional understanding.

> *[P9] "You sometimes have wildly different datasets being exchanged, which then causes everyone to question the value of the data itself, preventing better decisions from being made with the data."*

4.2.2.2 Asking the right question

Database administrators spoke often about their love to help others solve problems, mentioning how helping can be an especially enjoyable part of their job.

> *[P5] "When someone comes to us with a new question, or something that we can provide help with, we often get a lot of enjoyment out of doing a report that people are using."*

But sometimes database administrators have to help users ask the right question before they can begin to assist the user.

> *[P2] "Sometimes you have to ask questions to really dig in and try to figure out what they're trying to get at to determine whether we can even get at it."*

End users with a greater level of experience pulling data also mentioned that the way they work, analyze, and report on data is framed by the question they are trying to answer.

> *[P9] "I think it's a question of what are you really looking for? And are you appropriately querying or filtering? And again, that takes a lot of time and practice and training."*

4.2.2.3 Experience

Between database administrators and end users the level of formal training was vastly different. Database administrators had a high-level of training and experience with degrees in some field related to data or coding. For end users, everything is learned on the job or from colleagues with more experience.

[P9] "It's just users using the systems and becoming experts over time because they've used the systems. And not because they're brought in and given a clean training."

Another end user said that he has worked in academic units on campus and received training in the way of tutorials from others who were trained.

[P10] "a lot of tutorials from others trained in this."

4.2.3 Use

Through university data systems, users are building reports that aid with decision making in their units. Users were split between providing reports that were generated by the data systems and manipulating and combining the data in some way to create reports for their decision makers. There was a level of decision making associated with the use of these reports within units that included the allocation of budgetary resources, student experience planning, and campus- and university-level reporting. Relationship building emerged as a necessity for both end users and database administrators to understand both the data available within the system and the data needs of units. This larger theme generated four sub-themes: relationship building, reporting/analysis, decision making, and time.

4.2.3.1 Relationship building

As the interviews progressed, it became clear that to understand data needs, both end users and database administrators need to build relationships with the people providing or

requesting the data. For example, a database administrator commented that what facilitates job completion are the relationships they have developed across the university.

[P3] "What allows us to do our job is developing those relationships."

Another database administrator elaborated on this idea, commenting that understanding the context of the request and building credibility with others helps them to do their job well.

[P5] "I think that building the relationship helps us understand context more, and it also helps us in terms of building kind of that, for lack of a better term, kind of credibility to which we're all in this journey together."

For end users, relationship building created an additional benefit from sharing data with colleagues.

[P9] "Sometimes I share the reports that I've built with my colleagues from other departments, if they're wondering, you know, how to look at something."

4.2.3.2 Reporting/analysis

The level of analysis and reporting employed and to whom the reports were delivered differed in each unit for end users. Some end users provided reports based on what was available to them while others "usually manipulated and combined" the data. The frequency of the reporting varied from weekly to monthly for end users. In particular, one end user commented on weekly reporting to keep up with trends, then digging deeper each month.

[P6] "I do a point in cycle report every week or attempt to. And then the deeper dives in to this probably, maybe twice a month or monthly."

This same end user also mentioned that there is a limitation with finding reports where they do not have to manipulate data or write the reports themselves.

> *[P6] "I think sometimes finding some of the reports that I might want to have as just a straight written report that I don't have to write myself or manipulate the parameters. I think sometimes that's probably a limitation there."*

On the other hand, there was consistency in the level of analysis and reporting for database administrators. One administrator mentioned that he is pulling data and creating reports "every day, basically."

> *[P1] "I am performing analysis every day. I'm bringing data in and then merging it with other data sources. I'm, you know, pulling together and doing regressions and looking at the effectiveness of certain programs and stuff like that. That part I am doing every single day."*

4.2.3.3 Decision making

The reporting and analysis created by both end users and database administrators was used for decision-making purposes. Some of these reports were used to determine how resources were allocated, informing departmental budget conversations.

> *[P9] "That will usually inform some of the conversations that we have about where we want to spend some of our time and resources within our department."*

Other end users mentioned that the level of reporting they were performing helped to shape the student experience in their unit.

[P6] "It really helps us identify how we can shape our program and also meet the students' needs as we see what classes they're taking."

One database administrator primarily reported that they provided reporting for decision-making purposes at the campus and university level.

[P1] "That's literally why our office exists, for decision making."

Additionally, this administrator mentioned that students were also influential in this process.

[P1] "Students are decision makers, too. So we might be able to give them access to information as well."

4.2.3.4 Time

The amount of time end users and database administrators are spending creating analysis and reports for decision making. A database administrator mentioned how much he would love to provide reports immediately because end users need data quickly.

[P1] "People want things now, you know, and that's not always possible, give how complicated data is."

End users mentioned that what often seems like a simple question can turn into complicated analyses of data.

[P9] "What should be like, 'Oh yeah, I can just get you that number,' becomes like, well, maybe in a day or two, after I stare at things for hours, I can get that answer to you."

4.3 Thematic Analysis with Affinity Diagrams

The survey addressed the research aim of identifying data querying and visualization operations among end users. The interviews addressed the research aim of understanding and analyzing data querying and data-driven decisions by professionals at schools and universities. The initial thematic coding resulted in three overall themes: access, understandability, and use. The researcher further analyzed each theme to identify common topics in the data. The data was then analyzed by both the researcher and a group of human-computer interaction (HCI) graduate students working in the IES lab. The data was organized into affinity diagrams by thematic analysis by both the researcher and the students. The affinity diagram helped to cluster ideas and large amounts of information to understand the relationships between the themed information. The titles of the following sections are the same designations used earlier in this chapter. *(For full affinity diagrams, please see Appendix 3.)*

4.3.1 Access

4.3.1.1 Data source

For end users, there is sometimes a perception that data is not standardized across different systems, creating conflicting data results within the same unit. Across this code, data generation, data discrepancies and inconsistencies, reliability and functionality and usage were common themes in the analysis.

There is some confusion for end users around the source of the data. Some end users have the perception that some systems contain wrong data.

> *[P9] "I think there's just all this confusion about where it's all coming from. And there's not, I think there's a misconception that the data is wrong in one system versus the other. And I don't think that's true."*

For end users, there is a lack of understanding as to where the data comes from for each system. The results are different in each system, but all of the information is all supposed to be coming from the same source.

> *[P9] "I don't fully grasp like our IUIE versus SIS versus AdRx, the information is all theoretically the same information. So understanding what system feeds what."*

As users are asked to create reports and provide information for decision making from the breadth of systems available to them, they are testing the accuracy of the systems. If the new data systems cannot provide accurate data, the users cannot trust them moving forward as their personal credibility relies on accurate data reporting.

> *[P6] "As we move forward with Salesforce, and there's some data retrieval options in AdRx, those are things I've been measuring their accuracy, just to gauge how good the data is. And right now I'm not confident in their data."*
>
> *[P10] "At times those two numbers [from different systems] do not match. In academia, that creates a credibility issue."*

To improve functionality and usage, database administrators work to add new and unique data for end users. In institutional research, administrators will create surveys to gather new data that helps unit, campus and university level decision makers. Often this new data will be integrated with existing data to create a broader and sometimes deeper understanding of various populations.

> *[P1] "A lot of the information that we get isn't in any kind of database at all to speak of. We have a number of different surveys that we administer to faculty, staff and students and people like that. We will take that information and integrate it with information that we extract from SIS."*

4.3.1.2 Talking to other systems

Database administrators and end users cited the need for data systems to be better integrated with each other to improve usability, data collaboration and overall integration and accessibility of the data between systems. Database administrators were keen to include recruitment and advising data in to their systems as a way to assist in both the enrollment and retention processes that could improve overall usability for end users.

> *[P1] " Advising data lives outside of our data system. So any kind of information that we would be working with, if we are working with say, health and life science advising, for example, we would need to work with somebody who has information with AdRx that we can then download and share the data file with us to be able to integrate that with our systems. Um, we're trying to understand that better Salesforce and some of those things that admissions is now using for*

recruitment of students and trying to see what kind of data that they have that might be able to, you know, assist with our enrollment projection processes and things like that."

This sort of data collaboration would make reporting easier for end users, rather than having to compare data across multiple systems.

[P9] "So then you have to open two systems and compare across."

[P6] "If there were a way to have that merger of data [be] a little easier."

But some of the limitations of data integration resulted in users not engaging with specific systems.

[P6] "Maybe Salesforce after we integrated a little more and they start getting more access to our information. But at this time, especially for the graduate population, because we don't have applications that feed into it we have more limited data, I feel."

4.3.2 Understandability

4.3.2.1 Asking the right question

For database administrators, a key to helping end users find the data they need is coaching them through how to ask the right question. However, most end users feel that either a lack of experience or training hinders their knowledge, causing them to feel that they are not appropriately using or querying the data system. The thematic analysis for this coding resulted in the following themes: location and time, granularity and

complexity, data collection, obtaining the "right" information and data and how to answer questions.

Database administrators try to be responsive to campus needs by being both proactive and reactive. They will often create reports to showcase new data and ensure it is being used, but also want to be able to answer specific questions that users bring to them.

> *[P5] "We do a combination of being proactive and reactive. We can be nimble to various questions that are posed. But I think we're contacted probably on a daily basis with new questions."*

Additionally, the complexity of questions that pass through a database administrator's office can be difficult to predict, making it essential that they have the ability to respond, even if they are working on school, department or college level projects.

> *[P5] "I would never have anticipated that the dean of business was asking me if our students are transferring to another institution to take accounting courses. That's just not something I would ever dreamed up in my wildest dream, you know what I mean? But we were able to address that question."*

Being able to identify trends like accounting students taking courses at another institution helps users formulate questions that enable database administrators research the answer. One database administrator said that examining the trends enables her to be able to formulate questions.

> *[P3] "It's still to answer questions, but it's also to formulate questions. In other words, what are the trends telling us, so what can we see?"*

For end users, obtaining the "right" data can feel like an exercise in trial and error. Not knowing where to obtain the specific data or report that they need for decision making can be frustrating.

> *[P6] "And you don't know if it's going to be the right report when you get it, you know. There's some questionable stuff, it's just a trial in there."*

Database administrators see themselves as serving as translators between the person who is making a data request. However, a knowledge of the university or campus culture is important in being able to meet users' needs.

> *[P4] "I think that's one of the real roles of folks who are in analyst positions or data manager positions is really helping to serve as the translator between the person who's maybe making a request or has a specific question."*
>
> *[P1] "You have to understand people's questions in order to know what data to use and when to use it and things like that. Well, part of that is kind of understanding the zeitgeist of the university."*

4.3.2.2 Data definitions/coding

Throughout interviews with administrators and users, a common theme emerged that there is a lack of consistency in how to define the terminology surrounding commonly used data. The common themes that emerged from this thematic analysis were data discrepancies, lack of documentation, improvements, performance, consistent data and unique data.

4.3.2.2.1 Data discrepancies

For database administrators, many of the discrepancies in the data can be attributed to an inconsistency in the data definitions among end users. One database administrator mentioned that end users will use a different definition of a term, which means they will never be able to talk about the same thing.

> *[P3] "You're using a different definition and we're never going to match."*

However, for end users, they are often trying to determine where the data that meets their needs is, without having to sift through the discrepancies.

> *[P6] "You know, finding those discrepancies and trying to determine which ones have the most accurate data."*

4.3.2.2.2 Lack of documentation

One reason that database administrators cited for the discrepancies in the data was that there is a lack of documentation for both administrators and end users.

> *[P1] "If your next question is what I don't like – what I don't like is that there's not as much documentation out there."*

4.3.2.2.3 Improvements

Database administrators spend time working with units and end users to understand their process and share with them how to use the system to their benefit.

> *[P1] "We do work with a lot of units to understand the process and kind of share with them also the benefits of being able to do it."*

One administrator mentioned that an information technology group is creating a cookbooks with data definitions to create a trail documentation, but the process has been slow.

> *[P2] "I think they're trying to do that with the data cookbook, but it is a very slow process."*

4.3.2.2.4 Consistent data

While end users perceive inconsistencies in the data, database administrators feel that there are not consistent definitions across their campus or university to create a common vocabulary for the users of the data in various units. When discussing some of the misconceptions of data definitions and what could help alleviate this problem, a database administrator said:

> *[P3] "Consistent definitions across the campus and across units."*

Because the definitions do not seem to consistent among end users, creating a perceived lack of consistency in the data, administrators collaborate with units as a solution to this problem.

> *[P3] "We can't be the answer to everyone. So we work with other offices for consistency in definitions to help them learn where the best tables are to pull data."*

4.3.2.2.5 Unique data

As administrators encounter unique data, they are working to code the data for users. But this unique coding is dependent on the users and the units they work in to help administrators understand not only the requirements of the coding, but the necessity of the coding.

[P1] "We've been working with schools to try to understand, okay you have academic departments, which of these plan codes is associated with these departments."

[P1] "There will be a group code for students who studied abroad, for example. If we have those group codes integrated in SIS, we can go ahead and pull that information directly and share around. But there's only a handful of group codes out there."

4.3.2.3 Experience

Analysis affirmed that many users begin their roles with a lack of training for analyzing and reporting on data. As users engage in self learning, they seek out more training. While this may not consist of formal training, like an advanced degree, the longer the time invested in practicing data analysis and reporting, users will begin to obtain expert status.

End users worry that their lack of experience leads to concerns about the validity of their analysis and reporting. Because there is no formal training available or required for users who's role requires them report on data, there is a constant sense of on the job learning.

[P9] "It's just become, out of necessity, to be able to answer questions that are given to me constantly, you just learn. And there's definitely error that comes with learning. That's the danger of not having that formal option available. And I get that the training is difficult, because everyone wants to do different things with the

data. But there's the validity questions that come up when we're all just kind of wild westing this experience of pulling information."

End users are learning from colleagues, asking questions of administrators and taking as many trainings as they can to become proficient in pulling data. However, most of them have no formal training in data analysis.

[P6] "When I was in financial aid, I had taken a few of the free entry-level workshops from UITS like SQL, data retrieval, stuff like that. So, no formal [training] outside of a couple of workshops."

While there may not be any formal training options available, both administrators and users invest time in self learning to better do their jobs. This helps them to stay on top of the latest trends and technologies and speak knowledgeably to their colleagues and supervisors about their work.

[P1] "Being able to stay on top of the latest trends and latest technologies and the latest, you know, resources to be able to do your job, I think is important."

The level of training administrators and users needed for their respective roles varied greatly. Most administrators were trained extensively in data retrieval and analysis with all interviewed administrators having some level of formal training ranging from a bachelor's degree to a doctoral degree. End users often have little to no formal training when they started their roles. While administrators are performing more detailed analysis for large-scale, campus and university decision making, this lack of training can be viewed as contributing to the varying levels of understanding that end users have regarding data sources, analysis and how to ask the right questions for better data

retrieval. Administrators acknowledge that becoming an expert in navigating these data systems is often about who you know that can help to explain the systems and sources.

> *[P1] "A lot of what you need to know to navigate these systems is dependent on whom you know. … And it's very easy for people to assume they know what they're doing when they don't actually. This is one of the reasons that numbers sometimes get mixed up and people are pulling wrong information. Well-meaning people who go in who look at some code and oh, yeah, that's what you need. And it's not what you need."*

4.3.3 Use

4.3.3.1 Relationship building

Relationship building was a unique byproduct of these data systems. While end users access data from their workstations across campus, both database administrators and end users discussed how this work built relationships with other administrators and users across their unit, campus and the university.

4.3.3.1.1 Accessibility/access to data

University- and campus-wide databases can serve as a means of access for end users and oftentimes are used by end users as a way communicate with students regarding opportunities available to them.

> *[P8] "... primarily just as a means of communication to students about scholarship opportunities or hey, you haven't signed up for classes yet. Or*

hey, you've got a registration hold because that's another data set I can pull in."

Additionally, when end users lack access to what they need to complete their job, they do not let that create a barrier to continuing to build relationships with students. They use additional notes and spreadsheets to facilitate relationships internally with colleagues and in furthering the student experience.

[P7] "If somebody were to ask me, 'Hey, how many prospectives contacted you this month?' I'm able to go back and look at my notes. … I can see how many students applied and I can see maybe the interactions I've had, too."

4.3.3.1.2 Operational benefits

Relationship building created obvious operational benefits for the unit, campus and university where both database administrators and end users worked. These operational benefits include collaboration, facilitating work through relationships, management and maintenance. One database administrator mentioned how important building relationships with end users and other administrators was for creating not only that sense of collaboration but also for facilitating the work that needs to be done.

[P4] "I think those relationships come back and are important, too. Because it allows sort of the the ability to work through a process or to talk through what a procedure looks like, and ultimately arrive at how we can help move the institution forward by really being willing to work together and understand what folks are trying to achieve."

Another database administrator commented on the management of relationships and how important that was for her to do her job well.

> *[P3] "The reason for me being so hard to schedule right now is, especially at this time of year, I'm doing a lot of relationship building. I sit in a lot of meetings and listen."*

For end users, some of the systems they use regularly have helped to build relationships that make the day-to-day maintenance of their jobs a little easier.

> *[P9] "I think the dawn of AdRx has been phenomenal for anyone who works with more of the academic day-to-day data, just because it's both the messaging maintenance system, as well as all of their stuff in one setting."*

4.3.3.1.3 Skill building

Relationship building is a necessary skill development for database administrators. It is a crucial way of understanding how end users are accessing and inputting data.

> *[P4] "That's something I think that been not just in, you know, my work here. But in my IR [institutional research] work or work with data across institutions is that the relationship piece, especially because there are so many different points of input into systems, that having a relationship with those various points of contact, I think, is really crucial to be able to understand what they are intending to use."*

4.3.3.1.4 Customers

For database administrators, working with end users to understand their needs was essential and it helps the administrators to provide their product with better service and turnaround to one set of their customers – end users.

> *[P1] "It's one less step for you to give that to us, it's one less step for us, we're able to turn around and generate this data a little bit more quickly for you. So I think once we talked to them and work with users a little bit, it kind of helps them to understand the process."*

4.3.3.1.5 Expert consultation

Database administrators felt that it was especially important to consult with the data content experts and the decision makers when understanding how to provide and present data.

> *[P5] "And truly building those kinds of relationships with people, where they feel comfortable coming to us, so we're not, we're not an afterthought data office, we're at the table when decisions making is made."*

4.3.3.2 Reporting/analysis

End users create reports for decision making purposes within their department or unit. The analysis that goes in to creating these reports involves understanding the availability of data then modifying the data in some way before analysis can begin. When users create reports they typically involve some sort of reports on customers (students) and either existing data visualizations or self-created visualizations for their supervisors or unit-level leadership.

Users at the unit level typically tend to modify data, or blend and combine data, to create a richer understanding of what is happening with student-level trends. Users who regularly report on data customize the reports they are providing.

> *[P10] "I customize my reports, make notes to them. I share these on a weekly basis, which include enrollment data, to school leadership."*

These reports help users to understand if they are on track within their unit and allow their leadership the opportunity to understand the trends in admission and enrollment data. It also allows users to understand the status of programs within their unit.

> *[P6] "Mostly to track our admissions and our applicants, make sure we're on par for at least trying to come close to or beat the last few years of admissions and headcount. … After we get some of the admits and metrics to see where those programs are."*

While users are often pulling together trend reports for their leadership, database administrators are trying to understand the system-level trends that are occurring and reporting out with white papers and research briefs to support decision makers.

> *[P1] "We're sharing what we're doing every day out there. We have all kinds of stuff on our website, you know, all kinds of, you know, white papers and research briefs and things like that."*

Administrators are also creating more data visualizations to aid users in reporting. Some users are combining these visualizations with their own internal data. The availability of these visualizations for end users has continued to increase through the use of Tableau reports.

[P1] "There were only a handful of people using Tableau for visualizations. And now, I mean, we have an entire section of the website that's moving to Tableau reports."

4.3.3.3 Decision making

Administrators want end users to be informed decision makers when accessing data, and simultaneously end users want to be able to provide comprehensive and accurate data to help their leadership be informed decision makers. For end users, they are helping their unit leadership determine which decisions are worth investing resources.

[P9] "We're usually looking at trend data in terms of, is it worth it?"

The data that users are accessing is also helping to determine how programs are structured and can result in changes to academic programs. However, some end users find that the qualitative information, or student stories, makes a bigger impact with decision makers than the quantitative data.

[P6] "We use that data when we speak with the faculty directors and co-directors about how the program is shaped, what we're seeing as far as trends and enrollment. … But, you know, sometimes I think it's probably more of the independent meetings with the qualitative information that gets through easier than the quantitative information."

End users create reports for decisions making with a frequency as high as once a week for their decision makers. While they may not be the final decision makers, they want to

provide as much information as possible for decision making about admissions, courses and program design.

> *[P10] "The weekly reports are used for decision making. Also, course sizes."*
>
> *[P7] "I'm not the one that makes the final decisions. But if I can take reports like this to people who do make the final decisions, I think it would be great to just even think about programs for the future."*

4.3.3.4 Time

For end users, the time it takes to find data can be frustrating. Most users are accessing data on a weekly basis which creates multiple hours each week spent retrieving and analyzing data for reporting. Sometimes what are perceived as simple requests within a unit turn into a full day of work for users who have to access different data sets and create comparisons.

> *[P9] "Something that I think most people think is going to be simple, you'll go into a meeting and someone will say, well, do we know how many pre-nursing majors came into our major last year? So, yeah, that sounds really easy. That is something that, realistically, because of how the data is set up, I won't even be able to find any clean answers to that. It's going to take pulling from different datasets, and eventually cross comparing the two, to answer that question. And that, in and of itself, is at least a full day realistically."*

These time consuming activities feel frustrating to users who want data sets to be simpler to access without as much cross comparison. While one user says that all data can be

technically accessed in one place, there are multiple data sets that need to be pulled to understand the trends and to perform analysis.

> *[P9] "The amount of time it takes is a little frustrating."*

Database administrators also find this sort of data querying to be time consuming and often feel like they need to continue challenging themselves to learn and improve their skills. One database administrator mentioned that you have to be proactive in staying on top of learning new things.

> *[P1] "You don't necessarily get the time to go and learn the latest and greatest new thing. You kind of have to be proactive and set aside that time for yourself."*

5. DISCUSSION

The results of the research created three thematic groups: access, understandability, and use. These overall themes were used to craft a set of recommendations for design. Within these themes emerged a set challenges and opportunities that allowed the researcher to create a series of broad sketches for possible implementation.

5.1 Design Recommendations

In an effort to improve the relevance of the data systems accessed by end users, the researcher has outlined a set of design recommendations that are based on user experience best practices. These recommendations are grouped by the overarching themes found in the results – access, understandability, and use – and are patterned to be broad recommendations that can be applied to other data systems in higher education and other industries regularly accessing large amounts of data.

5.1.1 Access

According to Hick's Law, the amount of time it takes a user to make a decision increases with the amount of choices available to the user. Users are accessing multiple data sources that sometimes contain seemingly different data, causing users to make a series of decisions as to which data to access and use. Additionally, these systems do not work together, creating confusion for users as to the source of the data. The following recommendations can be implemented to improve access for users:

- Simplify choices for users by placing all data systems in one place with a brief overview as to what each system does, which will avoid overwhelming users with too many options.
- Break down complex data retrieval steps into smaller steps for easier access to the data source.
- Make data retrieval choices quick and simple for users. The longer it takes users to interpret choices and make a decision creates more work for users that they do not want.

5.1.2 Understandability

Users spend most of their time on other sites, which means that it helps users to make your data retrieval system similar to other data retrieval systems. By creating a standardization of data definitions, users will be able to operate similarly as they do on other data systems. Creating a standardized structure of data retrieval for all systems will simplify the learning process for users and allow those with disparate levels of experience to work at similar levels. The researcher has developed recommendations to improve understandability for users:

- Make all data retrieval systems operate similarly to simplify the process for end users.
- Visually represent similar codes near each other. Users group elements together if they are sharing an area with a similar boundary. This will aid users in contextually understanding data definitions.

5.1.3 Use

As the research progressed, it became clear that users were engaged in both challenges and opportunities related to their regular usage of data systems. Included in this theme was the opportunity to build relationships to improve work practices, the reporting and analysis required within unit-level work, the influence these systems have on decision making, and the amount of time it takes to access these systems. The researcher developed a series of guidelines to address this theme:

- According to Parkinson's Law, all work will expand to fill the time allotted to it. Visualizing the amount of time a project takes and allowing other users to share their time will enable users to gain an understanding of how long a task should take.
- Creating savable, replicable, and shareable workflows will allow users to reduce the amount of time they are spending retrieving data and building reports and create a sense of credibility for decision makers that work is consistent.
- Apply principles of collaborative design to teamwork to improve relationship building. Users working together to solve problems creates a shared understanding of both the problem and the solution.

5.2 Sketches

Determining the availability of data as well as accessing and modifying data can consume a large portion of the workday for users, especially when they lack formalized training on these data systems. The lack of data definitions and roadblocks built in to a user's workflow also contributes to a sense of siloed data gathering that has no known

regulations or standardization. Compounding this sense of frustration is the amount of time required to retrieve, analyze and report on data for users. By creating a gateway for users that pulls together all data sources as well as recommended uses for each data set, users are provided with a standardized way of accessing and wrangling data. Including training within the gateway allows novice users to catch up to experienced users.

Siloed data gathering seems to occur regularly within units, to the point that it can create conflicting data presented by separate users to the same decision makers within a unit. This often calls into question the credibility of the data itself, rather than the individual work practices involved in gathering the data. Creating a way for users to collaborate without adding additional time burdens to their day is essential when creating a collaborative solution for users.

To improve collaboration and work practices, tracking data workflows is essential not only for users to replicate their work, but to facilitate collaboration and improve the unit-level understand of how data is accessed, analyzed and reported on. By creating replicable and accessible workflows, users can not only save time by replicating work, but share practices with other users and employ solutions developed by other users.

5.2.1 Gateway

"Data wrangling" – a term coined by Sean Kandel, constitutes the amount of time spent determining what data is available, organizing data, and understanding how to combine multiple sources of data [20] – remains a drawback for users. The amount of a workday

spent determining the availability of data coupled with organizing data from multiple campus and university sources is a source of frustration for end users. The researchers suggest creating a gateway for users to manage data sources, understand the varying types of data available and suggested uses for the available data.

During the survey and interviews, users mentioned that they sometimes struggle to identify the appropriate data source when beginning a project (4.2.1.1 data source). By alleviating this pain point on the front end of acquiring data, users will be given equal footing when compared with their more seasoned peers in understanding which data systems are best and what data is available in these systems (4.2.2.3 experience). Additionally, the brief explanation will provide with each data source will provide a university suggested use for the data, enabling users to understand where administration prefers specific data to originate (4.2.2.2 asking the right question), as well as serving as an institution-level sanction on the data (4.2.2.1 data definitions/coding).

> *[S6] "Significant lack of knowledge about how/where to access data, lack of understanding what parameters to pull and analyze, and general inability manipulate data sets to get the desired output" (4.1)*
>
> *[P9] "I don't fully grasp like our IUIE versus SIS versus AdRx, the information is all theoretically the same information. So understanding what system feeds what." (4.3.1.1)*

The gateway will consist of a visual interface accessible through a website that will allow users to discover and report on data [11]. This system should include a combination of

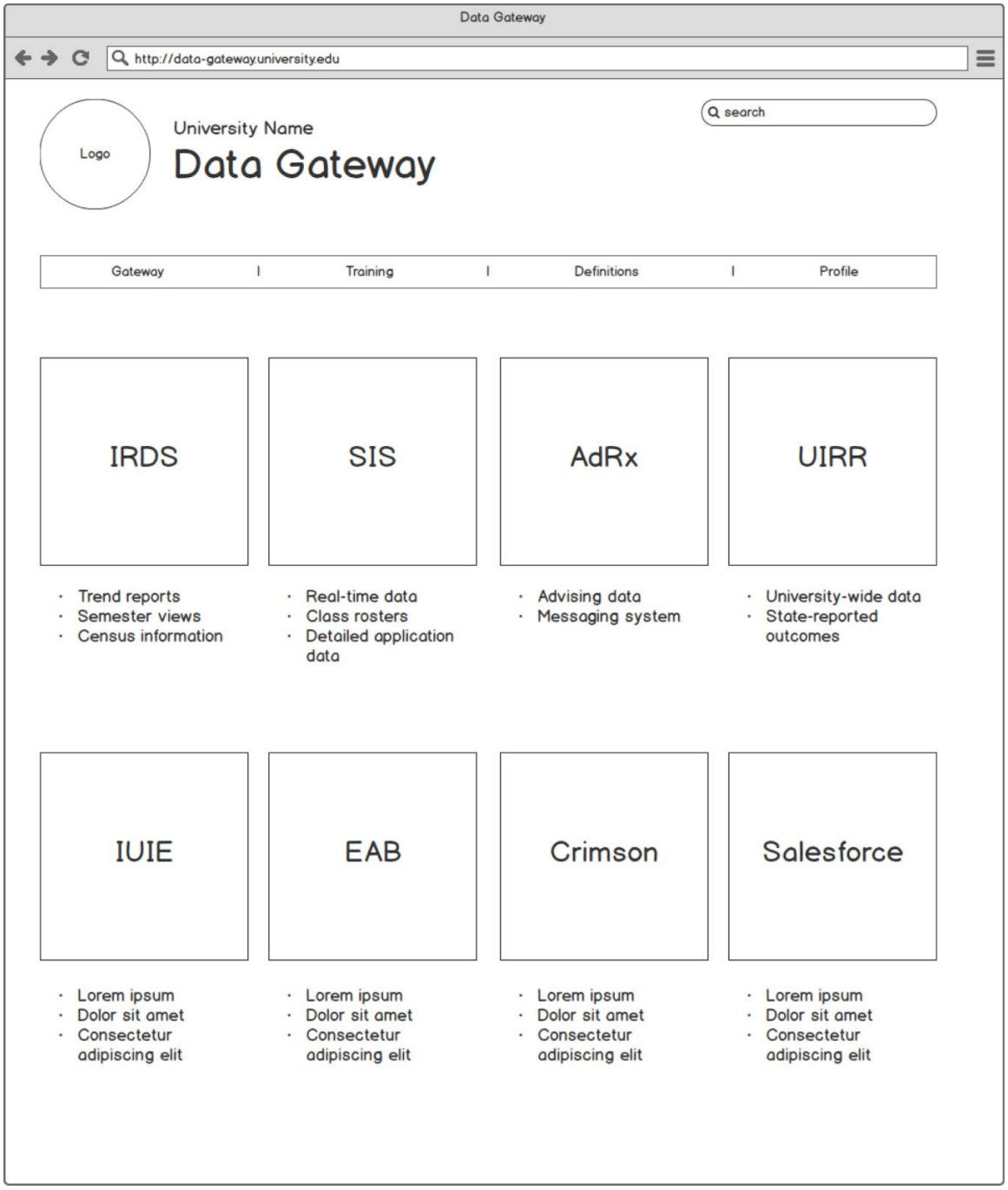


Image 1: Proposed gateway solution for end users

This solution creates a landing page for users that highlights each of the data sources and the main benefits of each data source, allowing users to quickly and easily understand which data source works best for their needs.

verification, transformation, and visualization as recommended by Kandel, et al. [28], and will be mainly accessed from a desktop interface as most users perform data analysis and transformation on their desktop.

The interface will allow for qualitative data to be imported to address the users' desire to tell a complete story for their decision makers, allowing them to "connect emotionally" with the customers they are trying to serve. [23] This emotional connection allows decision makers to remain rooted to the mission of the organization. It also allows users to tell the story of customers that backs the data discovered and aids in decision making.

This interface will allow novice users to access data in a controlled manner that will create the roadblocks desired by database administrators.

> *[P9] "So there's a lot of institutional knowledge that if you don't have, you're totally at a disadvantage when you try to begin pulling reports." (4.2.2.1)*

It will also help to guide users into asking the right question of the data through a series of checks and balances designed in the system to promote greater accountability. Additionally, the gateway will eliminate the frustration expressed by users over credibility (or lack thereof) that occurs with slight variances in data.

> *[P10] "Inconsistencies in the numbers are frustrating. It's off by just a handful. In academia, that creates a credibility issue." (4.2.2.1)*

Users will be able to point back to a replicable pathway as to how data was obtained, transformed and visualized. Users will also be able to footnote and source data as well, creating another layer of accountability for decision making and collaboration.

5.2.2 Training

Within the gateway, users will find a series of training modules that will guide them through how to use the systems available. The training will be a university-sanctioned and approved to assist users, who are domain-specific experts rather than data experts, to wrangle, analyze and visualize data to elaborate on the "fine-grain trends" [27] found in the school- and unit-level data (4.2.2.2 asking the right question). Users who are new to the university will be able to operate in a similar capacity to those with years of data experience (4.2.2.3 experience). As mentioned by one of the interviewees, there seems to be a standard where each data user is learning on their own without a formal training that is sanctioned by the university.

> *[P9]* "*And I get that the training is difficult, because everyone wants to do different things with the data. But there's the validity questions that come up when we're all just kind of wild westing this experience of pulling information." (4.3.2.3)*

At the end of the training, users will be able to use the interface within the gateway to verify the data they are using (4.2.1.1 data source), transform the data as needed and create meaningful visualizations for leadership within their organizations. Verification will occur by users engaging in a series of brief questions meant to delineate the data they are seeking to access. For instance, a user performing market research on fall semester applicants will require a different dataset than someone trying to understand the trends among students already on campus who change their major from other units.

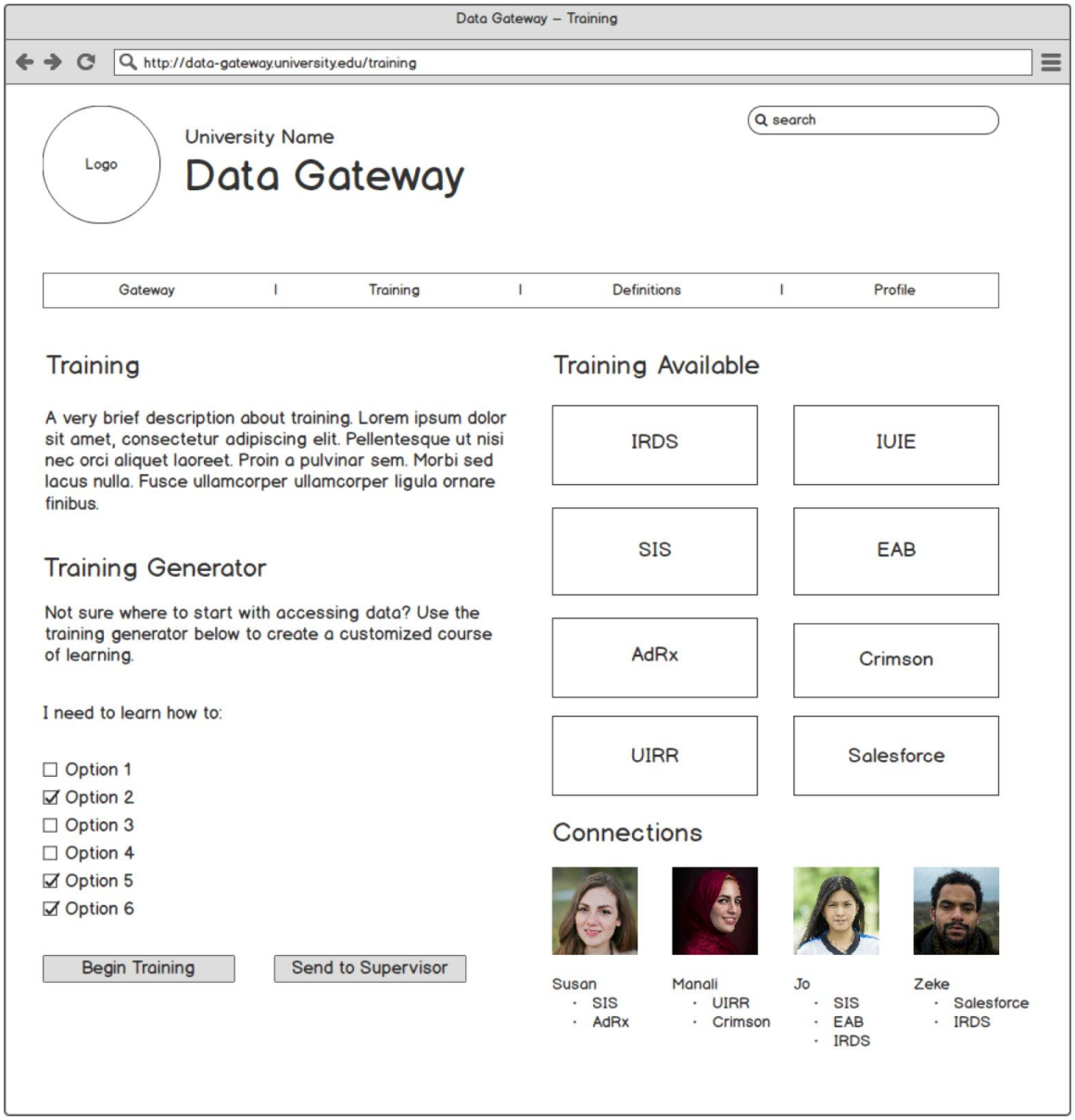


Image 2: Proposed training solution

This solution creates university-sanctioned training that allows end users to develop a custom-training solution for individual users, connect with other users, and have training approved by supervisors.

5.2.3 Collaboration

While a gateway and training will be beneficial to each individual user, one of the concerns expressed by interview participants was the difference in data pulled within units. To facilitate groups of users working together (within a unit, campus or university), the researchers recommend implementing a system that tracks and collects reports pulled by users for access to other users. Users will have a profile where they can access previous reports they have accessed as well the steps they used to access those reports, but these will be available to other users within their unit for ease of replication (4.2.3.2 reporting/analysis; 4.2.3.4 time). According to Bopp, et al., users can experience an erosion of autonomy when there is a lack of consensus around metrics (4.2.2.1 data definitions/coding). Creating a system that allows users to collaborate and develop shared metrics will alleviate the problem created by fragmentation in organizations when the data does not seem to be "connected to each other in any systematic way." [4] (4.2.1.1 data source)

> *[S13] "Not everyone has the same access to and understanding of the reports." (4.1)*

In addition to creating a lateral sense of collaboration among data users, the researchers recommend allowing both decision makers and database administrators access to collaborate as well (4.2.3.1 relationship building). Decision makers can alert users as to what metrics they want reported. Database administrators will have access to users' work practices to understand how they are accessing and sharing data. Administrators will also

be able to provide insight to users when needed or requested by users (4.2.2.2 asking the right question).

> *[S16] "Familiarity with the system and functionality. Most users only use the very basic functions, leaving the more in depth work or reporting to directors." (4.1)*

An additional consideration for creating a more collaborative environment is adapting smart boards and screens in conference rooms to be used to access data as a group or in meetings. These systems are already designed for participatory use and could be utilized as decision makers and analysts discuss reporting outcomes. The main consideration for use with large screens is ensuring that data either remains anonymized in these settings or additional steps are taken to ensure privacy.

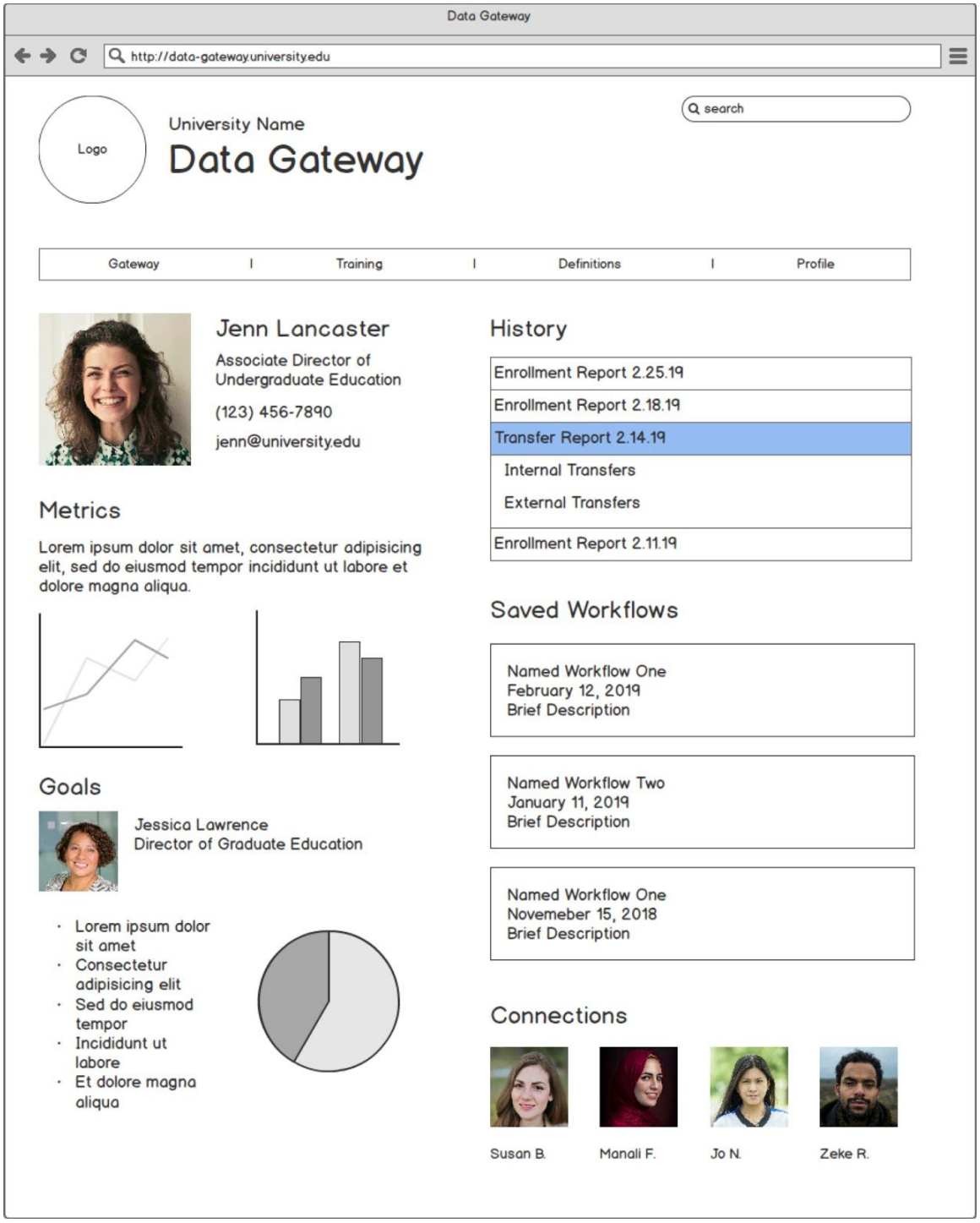


Image 3: Profile

The proposed data gateway will include profiles that will include the user's history for accessing later, metrics set by the user for quick access, saved workflows that the user would like to access again, goals set by the user's supervisor, and connections with other users.

5.2.4 Tracking

In addition to creating a tool to facilitate collaboration, users and database administrators should have a way of tracking data procurement and use. By enabling a ledger system, users and administrators can track the modifications and transformations that occur with data. Tracking the data origination and modifications will help to build trust among users (4.2.3.1 relationship building). A private, permission-only blockchain that users and database administrators can access will create a clear history of the data and how it has changed hands. What the blockchain will offer that a traditional spreadsheet or tracking software cannot is that it will create a single, accessible, incorruptible tracking system that can provide immediate results on data history, rather than hunting down the history of a data set (4.2.1.1 data source). It will provide a fool-proof way of understanding and visualizing the data source and its transformations while building credibility for those who access and report on data (4.2.2.1 data definitions/coding). Additionally, database administrators will save time hunting down data in the pipeline from origination to when it is input in the database (4.2.3.4 time).

> *[P4] "People can sort of operate from a vernacular definition of what 'enrolled' means, for example, or what 'first generation' means. But without being able to ask specifically what it means for the data owner, or the steward, can influence how it is interpreted at the back end when it is queried." (4.2.2.1)*
>
> *[P9] "I think building in a few more roadblocks to help with the validity of what we're using these datasets for, that would really take it to the next level." (4.2.2.1)*

The digital ledger can be included in the gateway interface as users access the history of the data from its origination to the eventual database they are pulling it from, as well as any use by other users in their unit. The ledger will enable users to build a sense of trust based on concrete assignations. This will serve as a system of checks and balances within units as data is accessed and the data access history is clear for all users (4.2.1.2 talking to other systems). Additionally, database administrators will be able to follow the breadcrumbs associated with data questions from end users.

5.2.5 Definitions/Roadblocks

Necessary for both database administrators and end users is a way to include definitions and roadblocks for all users. A shared vocabulary was considered an essential foundation for all who access data (4.2.2.1 data definitions/coding). While there is some work happening on creating a "cookbook" of data definitions for administrators and users, this system needs to be incorporated fully into the data access systems.

> *[P1] "If your next question is what I don't like – what I don't like is that there's not as much documentation out there." (4.3.2.2.2)*

As higher education integrates more data technologies, working to predict potential student behavior [26], it is necessary to build in roadblocks now so that all users are acquainted with the university's standard of data acquisition and reporting (4.2.3.2 reporting/analysis). As the demands for data continue to grow, it is essential that users have a sense of standardized procedures built into the expectations for their work practices.

> *[S1] "Some offices rely on internal databases (Excel) and have not trained on the campus-wide resources. That could mean we see wildly different information. It's better when all parties pull from the consistent place." (4.1)*

While roadblocks are often negatively viewed in user experience design, taking away from a possibly otherwise seamless experience, these sorts of roadblocks will ensure that users are all following a standardized method of acquiring and interpreting data that has been laid out and approved by database administrators and university decision makers (4.2.2.2 asking the right question; 4.2.3.3 decision making). While the roadblocks should be obvious enough to be read and incorporated into data acquisition work practices, they should not be so cumbersome as to prevent users from accessing the data systems.

The researchers propose building into the gateway a feed of recent activity on the landing page. Within this feed, the system will send notifications to users based on recent history with suggestions for better ways of pulling data or quicker ways to find data. As users progress through to find the data they are seeking, the gateway will provide advice on the data they are accessing. For instance, a user wants to develop a trend analysis of GPAs in specific general education courses in their major over the course of multiple semesters. Users will be able to select the parameters they want to access with explanations of how to use the parameters. Additionally, database administrators will be able to provide a "most used" feature that will either point to or provide access to the most used data requested by users and units.

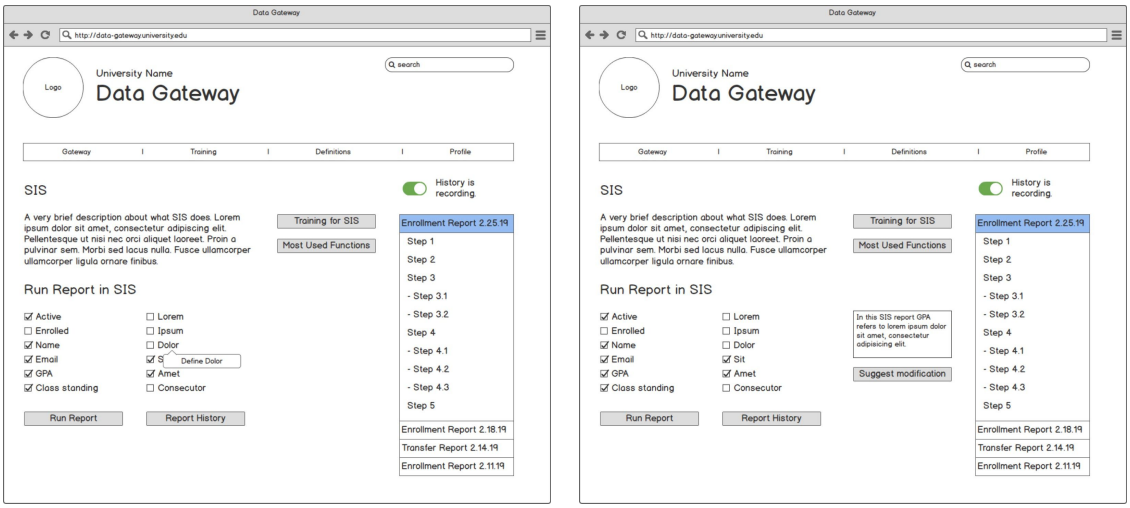


Image 4 and 5: Data Definitions/Roadblocks

Users can hover over terms and the system will define the term for users. Users will also have the option of suggesting modifications to the definitions for increased clarity.

5.2.6 Time

One of the most common frustrations of users was the amount of time it takes to access and analyze data (4.2.3.4 time). While some tasks can currently be automated, others that require combining data sources or accessing multiple data systems require users to spend time pulling together data that could otherwise be used for domain-specific work (4.2.3.2 reporting/analysis). Even preparing data to create visualizations, "reconstructing a repeatable workflow is difficult without a coherent linear history of the operations performed." [11]

> *[P9] "What should be like, 'Oh yeah, I can just get you that number,' becomes like, well, maybe in a day or two, after I stare at things for hours, I can get that answer to you." (4.2.3.4)*

The researchers propose incorporating a way to save workflow task actions that can be referenced through a visualization of actions and steps. Users will be able to access these saved workflows, share them with colleagues and also place them in the larger ecosystem of the university data retrieval.

Sharing best practices or found shortcuts will facilitate a sense of collaboration that is currently missing from the current system. By being able to replicate a colleague's work or understand how data was acquired will help individual users in units understand how other users acquire and use data as well as creating replicable workflows that can reduce the time spent on a given data-related task (4.2.2.3 experience).

This system will also create a way for users to track the time spent retrieving data. Visualizing the time spent on a task can be educational for users' supervisors and create a greater sense of the time spent on a given task. This particular visualization can also inform database administrators where the most time is being spent on data retrieval and the types of data being retrieved to aid in optimizing the system for end users (4.2.3.4 time).

> *[P1] "People want things now, you know, and that's not always possible, give how complicated data is." (4.2.3.4)*

For end users, placing the workflow visualization in the larger ecosystem of data use at the university will highlight the role they are playing in data storytelling and provide users a context for the work that they do (4.2.2.2 asking the right question). This

visualization will link a user's actions to other user's similar actions as well as the workflow that lead to users being able to access the data they are using, for instance, data is input, retrieved then input into systems, accessed by users, transformed, then reported to decision makers. Each step of this process will be visualized and available for users to see the full data universe. This context will also help users to build consensus among end users about the types of data being retrieved and the time users spend gathering data. [4]

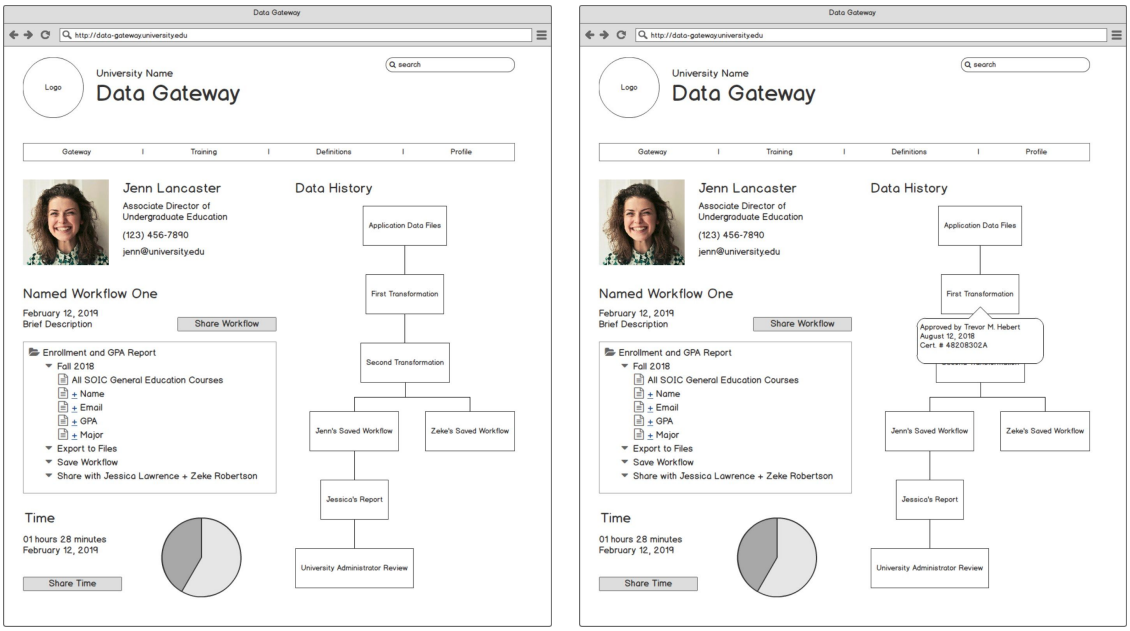


Image 6 and 7: Workflows, Time, and the Data Ecosystem

Users will be able to access their saved workflows, to access how the previously performed a task as well as automating that task for future use. Users will be able to see the time spent on a particular workflow to understand how to fit it within their daily work practices. Additionally, users will be able to see how a piece of data fits within the ecosystem of the university and how the data has been accessed and transformed throughout its lifespan.

6. CONCLUSION

At a large, midwestern university there currently exist varied methods for accessing and analyzing student data. On the surface, these data sources seem to be in conflict with one another, creating the impression of fragmented databases that do not "talk" to each other. Additionally, the wealth of data available in higher education only seems to be increasing creating a wealth of data that is available to users. As users maneuver through these virtual data spaces, they are modifying existing work practices and creating new ways of handling the large amounts of data they encounter regularly.

The goal of this research was to identify a set of themes that include the challenges and opportunities that occur in typical usage scenarios and craft a set of design recommendations that can be implemented. The research was organized into three aims including the following:

- Aim One: Identify data querying and visualization best practices and common operations focusing on incongruent and messy data sets.
- Aim Two: Identify all data sources used and data-driven decisions by marketing, recruitment and student support professionals at IUPUI, and other schools and universities.
- Aim Three: Develop design recommendations for a system that combines data and data visualization recommendations for higher education data systems.

The researcher performed a literature review to address best practices and common operations that users encounter in data retrieval and analysis. To understand how users are navigating these databases as well as how the databases are organized, the researchers created a mixed-method design with a survey and interview. In total, 18 end users from five midwestern colleges and universities completed the survey and 10 interviews with database administrators and end users were performed. The goal of the research methodology was to examine the end-user needs of higher-education databases. The survey and interviews were designed to understand how users access and utilize databases, including work practices, the demands of data gathering, and the internal demands placed on users. The primary focus of this study was the needs of end users, however, database administrators were interviewed to understand the creation of university-based databases and the current use of and future plans for these databases.

After analysis, the interview and survey resulted in three overarching themes: access, understandability, and use. These themes were used to further identify common topics in the data. The researcher worked with a group of HCI students to organize the data into affinity diagrams, clustering ideas and large amounts of information to understand the relationships between themed information.

A set of design recommendations were developed to address the three overarching themes of access, understandability, and use. These recommendations are patterned to be broad recommendations that can be applied to other data systems in higher education and other industries regularly accessing large amounts of data. Additionally, the researcher

developed a set of sketches that address the challenges and opportunities that influenced the user experience when accessing and working with data for end users in higher education. The sketches fell into the following categories: development of a data gateway, training, collaboration, tracking, definitions and roadblocks, and time. These categories provide a solution for the most compelling challenges identified in the challenges and opportunities surrounding the main themes.

These design recommendations can be incorporated into other data gathering work practices – that include discovery, acquisition, retrieval, transformation, and reporting – in higher education and extending into wider business practices, particularly those that collect an abundance of data, but lack the resources to consistently access, analyze, and report on the data. In particular, these sorts of industries can include health, construction, transportation, finance, government, and retail. Within health, the data created not only from patient records but also the Internet of Healthy Things, can be overwhelming to sort through and categorize as the industry continues to grow. Construction and transportation use logs that could benefit from not only a blockchain-style distributed ledger, but also a standardized training system and a way to collaborate both internally and externally. Both the finance industry and many state and local governments are already incorporating smart contracts, but both could profit from creating a tracking system that places the data in the larger ecosystem that each industry resides within. As higher education seeks to create a better customer experience for students, retail is focused on the same goal for both in-person and e-commerce shopping. Within these different realms, it is imperative

to harness the wealth of data accessed by end users whose work is focused on improving the experience for students, patients, clients, and the general public.

In pursuing further study on database usage in higher education, it will be necessary to understand user reaction to the design suggestions included in this study. The researchers would like to iteratively test these ideas and implementations with users to create improvements that meet end users needs while incorporating standardization created by database administrators. During testing with users, the researchers would like to develop methods of collaboration into the training portion of the website. While this study included research on collaboration while performing data work, it did not include any research on the benefits of collaboration during training, which could facilitate a greater understanding of data and better relationship building overall. The researchers would like to explore creating a user experience-focused solution for allowing disparate databases to "talk" to each other. While this solution will require a technical solution, users should feel that the solution is easy to use and does not create additional hardship for their everyday work. This will require further examination from the researchers. Finally, the researchers would like to perform a systematic review of all databases end users access and their incorporation into regular work practices. This review will enable the researchers to better address the end users' needs in the creation and execution of the gateway.

APPENDIX 1– RECRUITMENT MATERIALS

1.1 Survey Request – Internal

Good afternoon,

My name is Amanda Briggs and I am a graduate student in the School of Informatics and Computing at IUPUI. I also work at IUPUI in the Fairbanks School of Public Health as the director of marketing and communication.

I am working on my thesis research and have chosen to study the end-user needs of higher education database analysis and decision making (IRB #1808111834).

As a database user at IUPUI, I am asking for 15-20 minutes of your time to complete a survey regarding the data systems at your university. All answers to the survey will be anonymized and your identity will be protected if/when this research is published.

This survey will close on DATE.

If you have any questions please do not hesitate to respond to this email or my thesis advisor Dr. Francesco Cafaro at fcafaro@iu.edu.

Best regards,

Amanda Briggs

1.2 Survey Request – External

Good afternoon,

My name is Amanda Briggs and I am a graduate student in the School of Informatics and Computing at IUPUI. I also work at IUPUI in the Fairbanks School of Public Health as the director of marketing and communication.

I am working on my thesis research and have chosen to study the end-user needs of higher education database analysis and decision making (IRB #1808111834).

As a database user at your university, I am asking for 15-20 minutes of your time to complete a survey regarding the data systems at your university. All answers to the survey will be anonymized and your identity will be protected if/when this research is published.

This survey will close on DATE.

If you have any questions please do not hesitate to respond to this email or my thesis advisor Dr. Francesco Cafaro at fcafaro@iu.edu.

Best regards,

Amanda Briggs

1.3 Interview Request from Survey

Good afternoon,

Thank you for participating in my research survey "Fragmented Databases in Higher Education." You chose to be contacted for a follow-up interview.

I would like to schedule a time for an in-person, web-video, or phone interview. The interview will last approximately 30-45 minutes and will only ask about your work practices.

If you have any questions please do not hesitate to respond to this email.

Best regards,

Amanda Briggs

1.4 Interview Request – Database Administrator

Good afternoon,

My name is Amanda Briggs and I am a graduate student in the School of Informatics and Computing. I also happen to work at IUPUI in the Fairbanks School of Public Health.

I am working on my thesis research and have chosen to study the end-user needs of higher education database analysis and decision making (IRB #1808111834).

As a database administrator/creator, I would like to interview you to understand your role in the creation of this system and the future you see for these types of databases at IUPUI and other universities.

I would like to schedule a half hour of your time for an interview. We can conduct this interview in-person or via web-based video. I expect this interview will take 20-30 minutes to complete and will only ask you about your work practices.

If you have any questions please do not hesitate to respond to this email or my thesis advisor Dr. Francesco Cafaro at fcafaro@iu.edu.

I look forward to hearing back from you.

Best regards, Amanda Briggs

APPENDIX 2 – INFORMED CONSENT

2.1 Informed Consent – Survey

IRB STUDY # 1808111834

INDIANA UNIVERSITY INFORMED CONSENT STATEMENT FOR

Understanding End-User Needs of Databases in Higher Education Data Analysis and Decision Making

You are invited to participate in a research study to examine the needs of student support database users at IUPUI. You were selected as a possible subject because you work at IUPUI and access institutional databases as part of your daily work. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

The study is being conducted by: Amanda Briggs and Dr. Francesco Cafaro. This study is not funded by scholarships or federal programs.

STUDY PURPOSE

The purpose of this study is to explore the different work practices and needs of institutional database users.

NUMBER OF PEOPLE TAKING PART IN THE STUDY

If you agree to participate, you will be one of 300 participants who will be participating in this research.

PROCEDURES FOR THE STUDY

If you are eligible to participate, you will be asked to take a 20-minute survey that details your use of institutional databases at IUPUI. You will be asked a series of closed- and open-ended questions.

The answers to the survey will be stored on a password-protected computer so that the researchers can refer back to the answers for further examination.

Following the survey, you will be asked if you are willing to participate in an interview. This interview will take 30 minutes to complete and take place at the location of your choosing.

RISKS OF TAKING PART IN THE STUDY

While on the study, the risks are:

- While we will take every effort to protect your data and ensure your confidentiality as a part of this study, there is still a potential risk for a breach in confidentiality.
- There is the potential for physical fatigue while participating in the survey or interview.

The researchers will try to minimize these risks by:

- All data from the surveys will be stored in an encrypted location and will be promptly destroyed after they have been transcribed and the study is complete.
- Non-ephemeral records of contact (i.e. email messages) will be deleted after the study is completed.
- Research personnel will de-identify the transcriptions of the recordings. All of the de-identification process will be performed on a password-protected laptop. After this process, all identifiable data will be immediately deleted from this laptop and all original recordings will be destroyed.
- Transcription and coding of the data will be performed on a password-protected computer.
- A portion of the de-identified data from the user studies will be used for the purpose of dissemination at academic conferences. No data records will make use of personally-identifiable information (randomly assigned unique codes will be used to distinguish between participants).

BENEFITS OF TAKING PART IN THE STUDY

You will personally contribute to system-wide database recommendations at a university.

ALTERNATIVES TO TAKING PART IN THE STUDY

Instead of being in the study, you may choose to not participate.

CONFIDENTIALITY

Efforts will be made to keep your personal information confidential. We cannot

guarantee absolute confidentiality. Your personal information may be disclosed if required by law. Your identity will be held in confidence in reports in which the study may be published and databases in which results may be stored. Only the researchers of this study will have access to the audio recordings of the interviews, and it will be stored in an encrypted location. Once the study is over, the audio recordings will all be destroyed. Information collected from you for this study may be used for future research studies or shared with other researchers for future research. No information will be shared that can identify you.

Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the study investigator and his research associates, the Indiana University Institutional Review Board (IRB) or its designees, and (as allowed by law) state or federal agencies, specifically the Office for Human Research Protections (OHRP).

PAYMENT

There will be no payment for participating in this study.

CONTACTS FOR QUESTIONS OR PROBLEMS

For questions about the study, contact the researchers:

Amanda Briggs

amanbrig@iu.edu

Dr. Francesco Cafaro

fcafaro@iu.edu

For questions about your rights as a research participant or to discuss problems, complaints or concerns about a research study, or to obtain information, or offer input, contact the IU Human Subjects Office at (317) 278-3458 for Indianapolis campus or (812) 856-4242 for Bloomington campus.

VOLUNTARY NATURE OF THIS STUDY

Taking part in this study is voluntary. You may choose not to take part or may leave the study at any time. Leaving the study will not result in any penalty or loss of benefits to which you are entitled. Your decision whether or not to participate in this study will not affect your current or future relations with the IUPUI, Indiana University (IU), or the School of Informatics and Technology.

2.2 Informed Consent – Interview

IRB STUDY # 1808111834

INDIANA UNIVERSITY INFORMED CONSENT STATEMENT FOR

Understanding End-User Needs of Databases in Higher Education Data Analysis and Decision Making

You are invited to participate in a research study to examine the needs of student support database users at IUPUI. You were selected as a possible subject because you work at IUPUI and access institutional databases as part of your daily work. We as that you read this form and ask any questions you may have before agreeing to be in the study.

The study is being conducted by: Amanda Briggs and Dr. Francesco Cafaro. This study is not funded by scholarships or federal programs.

STUDY PURPOSE

The purpose of this study is to explore the different work practices and needs of institutional database users.

NUMBER OF PEOPLE TAKING PART IN THE STUDY

If you agree to participate, you will be one of 300 participants who will be participating in this research.

PROCEDURES FOR THE STUDY

If you are eligible to participate, you will be asked to participate in a 30-minute interview that details your use of institutional databases at IUPUI. You will be asked a series of closed- and open-ended questions. The interviews will be audio recorded and some photos may be taken of your work environment. If faces are captured, these will be anonymized by blurring prior to publication of the study.

Your answers and all audio recordings and photos will be stored on a password-protected computer so that the researchers can refer back to the answers for further examination.

RISKS OF TAKING PART IN THE STUDY

While on the study, the risks are:

- While we will take every effort to protect your data and ensure your confidentiality as a part of this study, there is still a potential risk for a breach in confidentiality.
- There is the potential for physical fatigue while participating in the survey or interview.

The researchers will try to minimize these risks by:

- All data from the surveys and recordings of the interviews will be stored in an encrypted location and will be promptly destroyed after they have been transcribed and the study is complete.

- Non-ephemeral records of contact (i.e. email messages) will be deleted after the study is completed.
- Research personnel will de-identify the transcriptions of the recordings. All of the de-identification process will be performed on a password-protected laptop. After this process, all identifiable data will be immediately deleted from this laptop and all original recordings will be destroyed.
- Transcription and coding of the data will be performed on a password-protected computer.
- A portion of the de-identified data from the user studies will be used for the purpose of dissemination at academic conferences. No data records will make use of personally-identifiable information (randomly assigned unique codes will be used to distinguish between participants).

BENEFITS OF TAKING PART IN THE STUDY

There is no direct benefit to participation. You will personally contribute to system-wide database recommendations at a university.

CONFIDENTIALITY

Efforts will be made to keep your personal information confidential. We cannot guarantee absolute confidentiality. Your personal information may be disclosed if required by law. Your identity will be held in confidence in reports in which the study may be published and databases in which results may be stored. Only the researchers of

this study will have access the audio recordings of the interviews, and it will be stored in an encrypted location. Once the study is over, the audio recordings will all be destroyed. Information collected from you for this study may be used for future research studies or shared with other researchers for future research. No information will be shared that can identify you.

Organizations that may inspect and/or copy your research records for quality assurance and data analysis include groups such as the study investigator and his research associates, the Indiana University Institutional Review Board (IRB) or its designees, and (as allowed by law) state or federal agencies, specifically the Office for Human Research Protections (OHRP).

PAYMENT

There will be no payment for participating in this study.

CONTACTS FOR QUESTIONS OR PROBLEMS

For questions about the study, contact the researchers:

Amanda Briggs

amanbrig@iu.edu

Dr. Francesco Cafaro

fcafaro@iu.edu

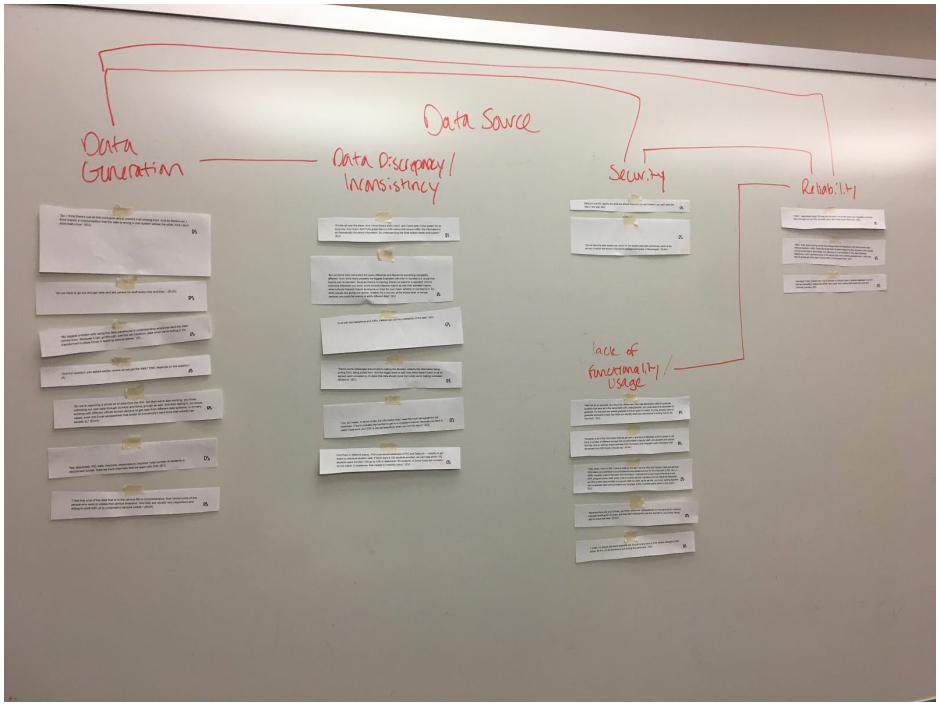
For questions about your rights as a research participant or to discuss problems, complaints or concerns about a research study, or to obtain information, or offer input, contact the IU Human Subjects Office at (317) 278-3458 for Indianapolis campus or (812) 856-4242 for Bloomington campus. You may also email irb@iu.edu.

VOLUNTARY NATURE OF THIS STUDY

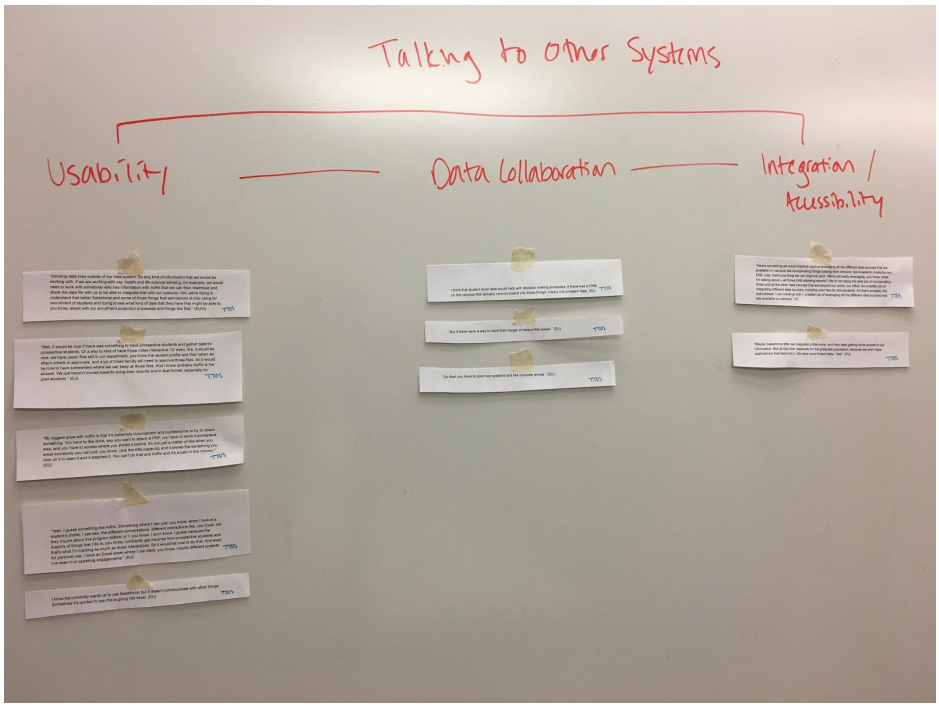
Taking part in this study is voluntary. You may choose not to take part or may leave the study at any time. Leaving the study will not result in any penalty or loss of benefits to which you are entitled. Your decision whether or not to participate in this study will not affect your current or future relations with the IUPUI, Indiana University (IU), or the School of Informatics and Technology.

APPENDIX 3 – AFFINITY DIAGRAMS

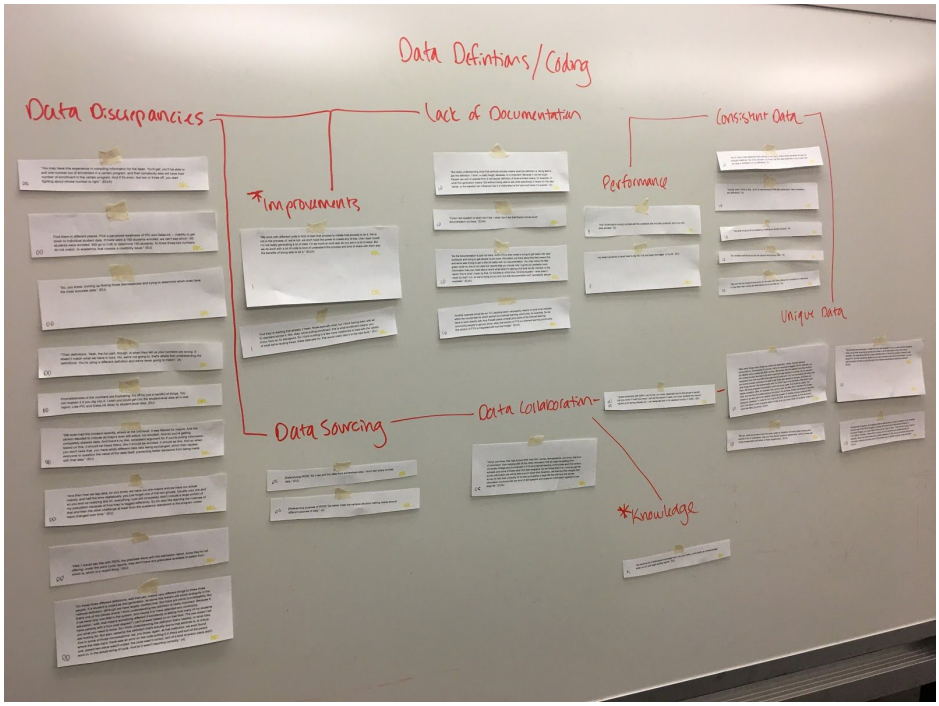
3.1 Data source



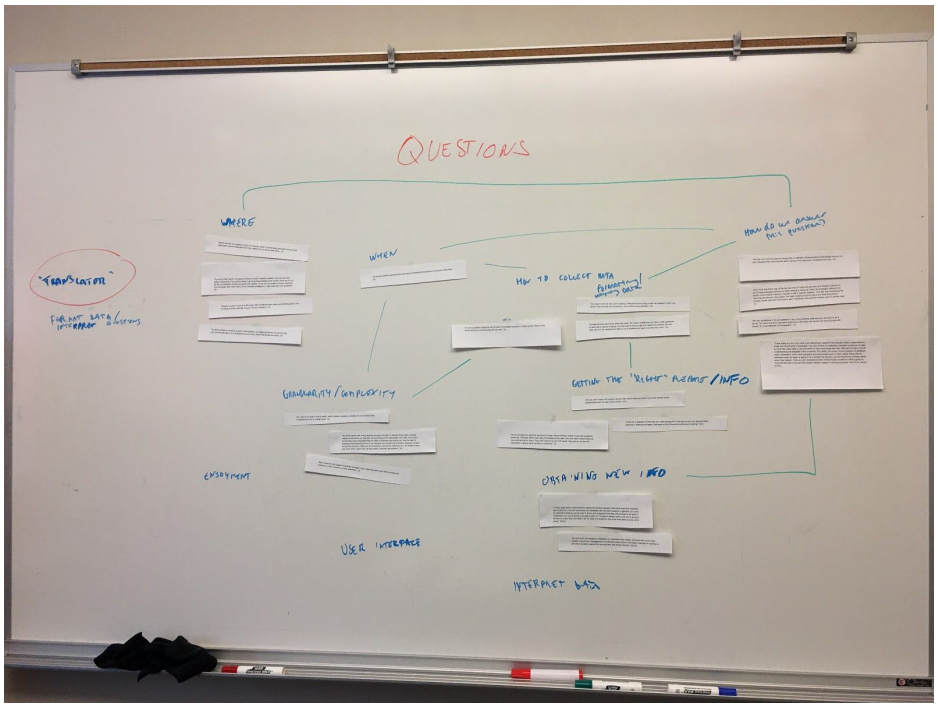
3.2 Talking to other systems



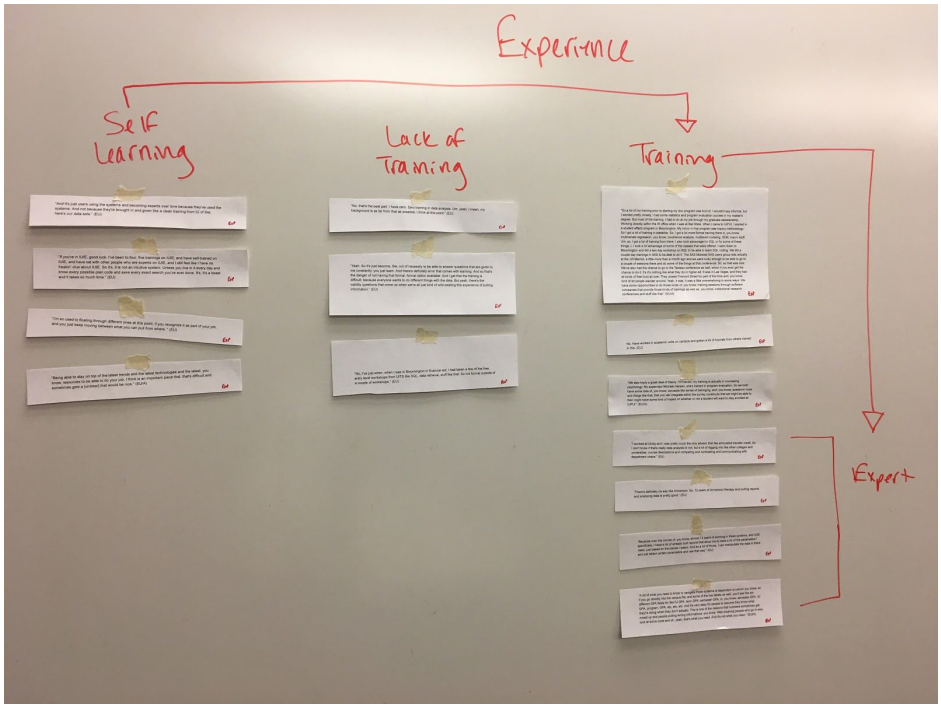
3.3 Data definitions/coding



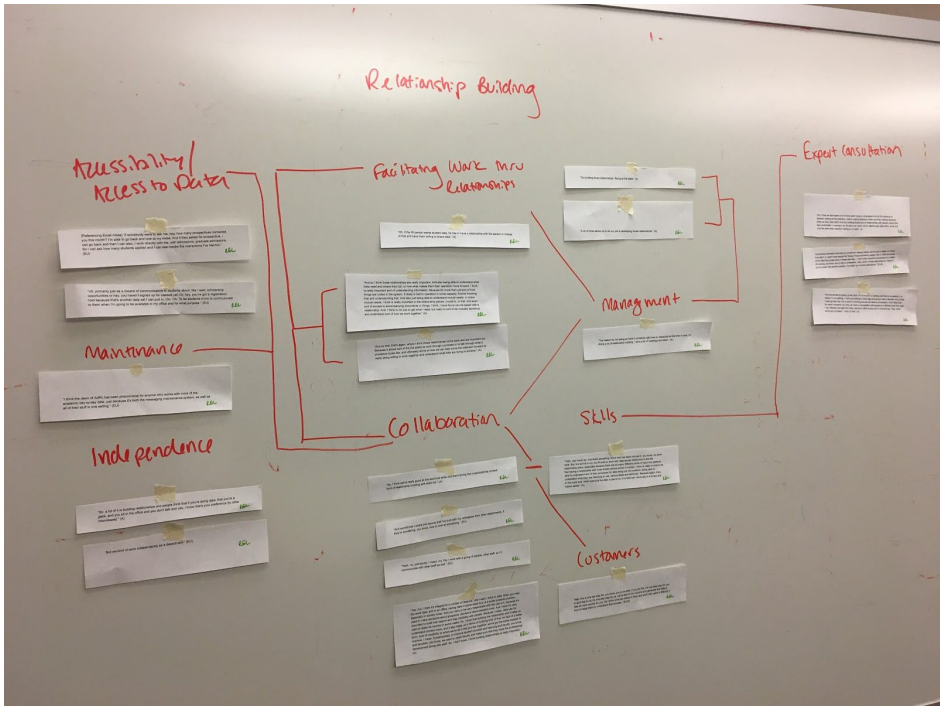
3.4 Asking the right question



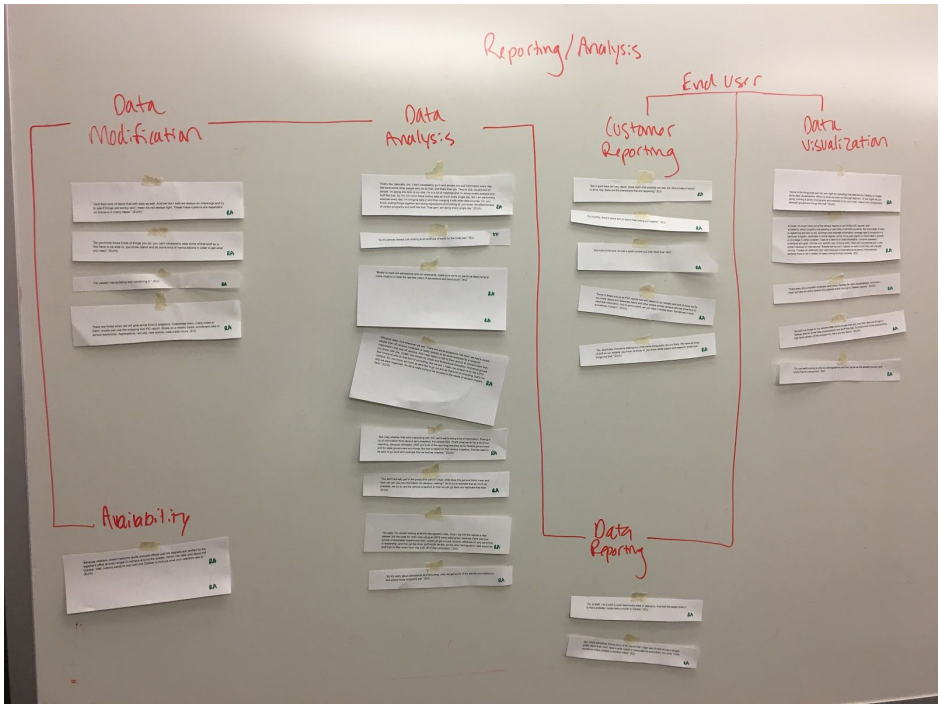
3.5 Experience



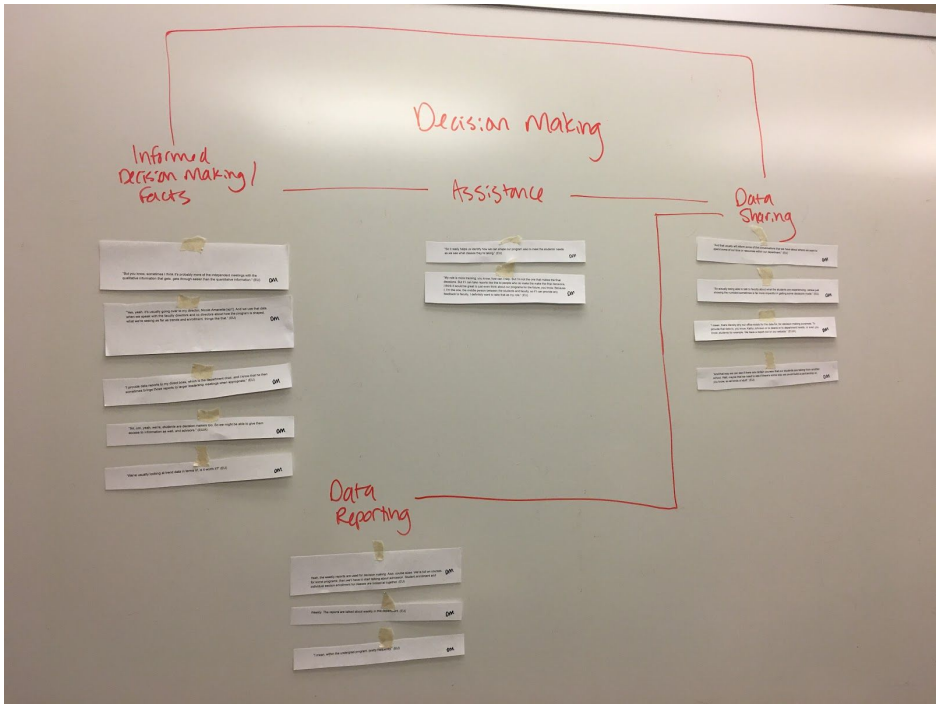
3.6 Relationship building



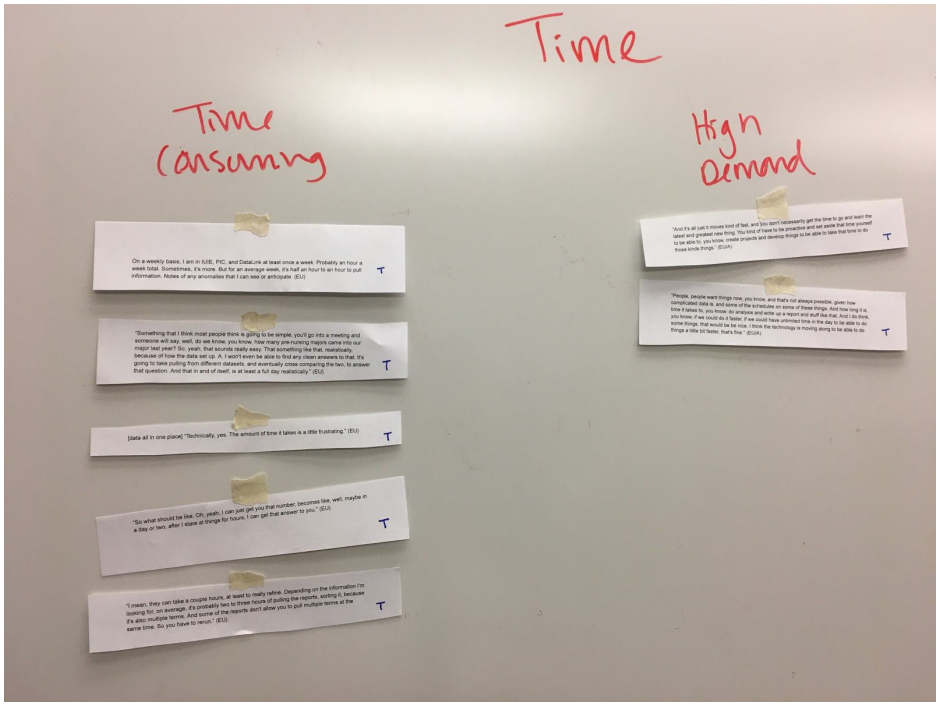
3.7 Reporting/analysis



3.8 Decision making



3.9 Time


Time
Time consuming
High demand
On a weekly basis, I am in IUIE, PIC, and DataLink at least once a week. Probably an hour a week total. Sometimes, it's more. But for an average week, it's half an hour to an hour to pull information. Notes of any anomalies that I can see or anticipate. (EU)
"Something that I think most people think is going to be simple, you'll go into a meeting and someone will say, well, do we know, you know, how many pre-nursing majors came into our major last year? So, yeah, that sounds really easy. That something like that, realistically, because of how the data set up. A, I won't even be able to find any clean answers to that. It's going to take pulling from different datasets, and eventually cross comparing the two, to answer that question. And that in and of itself, is at least a full day realistically." (EU)
[data all in one place] "Technically, yes. The amount of time it takes is a little frustrating." (EU)
"So what should be like. Oh, yeah, I can just get you that number, becomes like, well, maybe in a day or two, after I stare at things for hours. I can get that answer to you." (EU)
"I mean, they can take a couple hours, at least to really refine. Depending on the information I'm looking for, on average, it's probably two to three hours of pulling the reports, sorting it, because it's also multiple terms. And some of the reports don't allow you to pull multiple terms at the same time. So you have to rerun." (EU)
"And it's all just it moves kind of fast, and you don't necessarily get the time to go and learn the latest and greatest new thing. You kind of have to be proactive and set aside that time yourself to be able to, you know, create projects and develop things to be able to take that time to do those kinds things." (EUA)
"People, people want things now, you know, and that's not always possible, given how complicated data is, and some of the schedules on some of these things. And how long it is, time it takes to, you know, do analyses and write up a report and stuff like that. And I do think, you know, if we could do it faster, if we could have unlimited time in the day to be able to do some things, that would be be nice. I think the technology is moving along to be able to do things a little bit faster, that's fine." (EUA)

REFERENCES

[1] L. M. Benjamin, A. Voida, and C. Bopp, “Policy fields, data systems, and the performance of nonprofit human service organizations,” Human Service Organizations: Management, Leadership & Governance, vol. 42, no. 2, pp. 185-204, 2018.

[2] O. Bergman, R. Beyth-Marom, and R. Nachmias, “The project fragmentation problem in personal information management,” in Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06, 2006.

[3] C. Bopp, E. Harmon, and A. Voida, “Disempowered by Data,” in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17, 2017, pp. 3608-3619.

[4] D. Boyd, and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” Information, communication & society, vol. 15, no. 5, pp. 662-679, 2012.

[5] B. K. Daniel, "Big Data in Higher Education: The Big Picture," Big Data and Learning Analytics in Higher Education, pp. 19-28, 2017.

[6] B. I. U. Dur, “Data visualization and infographics in visual communication design education at the age of information,” Journal of Arts and Humanities, vol. 3, no. 5, pp. 39-50, 2014.

[7] S. Few, “Information dashboard design,” 2006.

[8] J. S. Gagliardi, and P. Wilkinson, “Big Data on Campus,” 2017.

[9] K. C. Green, “The Babel Problem with Big Data,” 2018.

[10] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, “Research directions in data wrangling: Visualizations and transformations for usable and credible data,” Information Visualization, vol. 10, no. 4, pp. 271-288, 2011.

[11] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, “Enterprise data analysis and visualization: An interview study,” IEEE Transactions on Visualization & Computer Graphics, no. 12, pp. 2917-2926, 2012.

[12] B. c. Kwon, B. Fisher, and J. S. Yi, “Visual analytic roadblocks for novice investigators,” in 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), 2011, pp. 3-11.

[13] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, “Big data, analytics and the path from insights to value,” MIT sloan management review, vol. 52, no. 2, pp. 21, 2011.

[14] C. A. Le Dantec, and W. K. Edwards, "Across boundaries of influence and accountability: The multiple scales of public sector information systems." pp. 113-122.

[15] S. Lohr, “The age of big data,” New York Times, vol. 11, no. 2012, 2012.

[16] J. Luan, “Data mining and its applications in higher education,” New directions for institutional research, vol. 2002, no. 113, pp. 17-36, 2002.

[17] A. G. Picciano, “The evolution of big data and learning analytics in American higher education,” Journal of Asynchronous Learning Networks, vol. 16, no. 3, pp. 9-20, 2012.

[18] K. Pine, and M. Mazmanian, “Emerging Insights on Building Infrastructure for Data-Driven Transparency and Accountability of Organizations,” iConference 2015 Proceedings, 2015.

[19] R. G. Qiu, R. R. Ravi, and L. L. Qiu, “Aggregating and visualizing public opinions and sentiment trends on the US higher education,” in Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services - iiWAS '15, 2015, pp. 1-5.

[20] T. Rattenbury, J. M. Hellerstein, J. Heer, S. Kandel, and C. Carreras, Principles of Data Wrangling: Practical Techniques for Data Preparation: " O'Reilly Media, Inc.", 2017.

[21] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz, "Evaluating information visualization in large companies: challenges, experiences and recommendations." pp. 79-86.

[22] B. Shneiderman, “Improving the human factors aspect of database interactions,” ACM Transactions on Database Systems (TODS), vol. 3, no. 4, pp. 417-439, 1978.

[23] N. Verma, and A. Voida, “On Being Actionable,” in Proceedings of the 19th International Conference on Supporting Group Work - GROUP '16, 2016, pp. 325-334.

[24] A. Voida, E. Harmon, and B. Al-Ani, "Homebrew databases: Complexities of everyday information management in nonprofit organizations." pp. 915-924.

[25] B. Williamson, “Digital education governance: data visualization, predictive analytics, and ‘real-time’ policy instruments,” Journal of Education Policy, vol. 31, no. 2, pp. 123-141, 2015.

[26] C. Zhu, H. Zhu, H. Xiong, P. Ding, and F. Xie, “Recruitment Market Trend Analysis with Sequential Latent Variable Models,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016, pp. 383-392.

[27] B. chul Kwon, B. Fisher, and J. S. Yi, "Visual analytic roadblocks for novice investigators." pp. 3-11.

[28] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono, “Research directions in data wrangling: Visualizations and transformations for usable and credible data,” Information Visualization, vol. 10, no. 4, pp. 271-288, 2011.

[29] M. Sedlmair, P. Isenberg, D. Baur, and A. Butz, "Evaluating information visualization in large companies: challenges, experiences and recommendations." pp. 79-86.

[30] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, “Interactions with big data analytics,” interactions, vol. 19, no. 3, pp. 50-59, 2012.

[31] D. A. Keim, “Visual exploration of large data sets,” Communications of the ACM, vol. 44, no. 8, pp. 38-44, 2001.

[32] D. A. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler, "Challenges in visual data analysis." pp. 9-16.