BEYOND TECHNOLOGY: DESIGN A VALUE-DRIVEN INTEGRATIVE PROCESS MODEL FOR DATA ANALYTICS

by

CHAOJIE WANG

MBA in Finance, Loyola University Maryland, 2013

MS in Statistics, The University of Toledo, 1993

MA in Economics, The University of Toledo, 1993

BE in Management Information System, Tsinghua University, 1986

Submitted to the Graduate Faculty of the

School of Communications & Information Systems

in partial fulfillment of the requirements for the degree of

DOCTOR OF SCIENCE

IN

INFORMATION SYSTEMS AND COMMUNICATIONS

Advisor: Paul J. Kovacs, Ph.D.

Wenli Wang, Ph.D.

David F. Wood, Ph.D.

Robert Morris University

ABSTRACT

Data analytics is a key to unlock the untapped power of knowledge hidden in the trenches of big data in this Information Age. However, not all analytics efforts can achieve the desired outcomes and impacts. A successfully executed data analytics project relies on the application of an effective process model. Despite the abundance of existing analytics process models, many analytics professionals choose to use their own, or not to use any at all. This low adoption rate of process models and the lack of a universal model inhibit the maturity and growth of the analytics profession in satisfying the increasing demand for data analytics. This study takes the design science approach and designs a process model, the A2E Model, that is grounded in theory and unifies the best practices in industry. The A2E is an acronym for the ABCDE, the five steps in data analytics process: Assess Needs, Blend Data, Create Analytics, Discover Insights, and Explore Ideas. The utility of the model was illustrated by applying it to a real-world healthcare analytics effort aiming to improve dialysis care quality by reducing unplanned hospital readmissions. In addition, expert reviews were conducted to assess the relevancy and quality of the process model. This study contributes to the understanding and knowledge of data analytics process and helps both business and technical professionals achieve higher quality and deliver greater impacts for their data analytics efforts.

ACKNOWLEDGMENTS

I am grateful for the generous financial support from my employer. Its culture of multidisciplinary collaboration and system thinking helped shape this dissertation.

I am grateful for the doctoral faculties and administrative staff of the Doctor of Science in Information Systems & Communications (DISC) program at Robert Morris University. The unique cohort-based and in-person executive format provided an engaging and supportive environment and made this journey of learning both fun-filled and fulfilling.

I am grateful for my advisor Dr. Paul Kovacs and committee members Dr. Wenli Wang and Dr. David Wood. Their guidance and encouragement made the dissertation project less a burden and more an adventure. Dr. Kovacs's course on Data Warehouse, Business Intelligence, and Data Science helped set the foundation for this dissertation. His advice of "Keep it simple. Don't overthink." reflects the essence of the user-centered design and is applicable not only to doctoral projects but also to our daily livings. Dr. Wang's Advanced Quantitative Methodology course went beyond mechanics and techniques of inferential statistics and emphasized the importance of model development and theory building. Dr. Wood's Economics of Information Systems course provided a multidisciplinary and holistic view of information systems weaving together subject matters such as computer science, information systems, statistics, economics, and business strategy. Information is way beyond bits and bytes and that is part of the story this dissertation attempts to tell.

I am grateful for the fellow students of the Cohort 18. They came from all walks of lives with diverse cultural and intellectual backgrounds. What I learned from them enriched both my life experience and my intellectual capacity. Fred Hoffman, Alvi Lim, and Jin Kwon deserve special recognition. It was a blessing to be part of the Gang of Four, or shall I say, the Gang of

Food. Our countless dinner outings in Pittsburgh, Pennsylvania, and Ellicott City, Maryland, provided nutrition, nurture, and naughtiness much needed for the long and arduous journey.

I am grateful for my wife Libin Zhong. She endured many lonely nights when I was away from home at the doctoral residencies, and she ensured I was away from the TV set and stayed focused on my dissertation when I was at home. Our boys, Jayson and James are the fruits of love and the sources of joy. Keeping them in my thought is my secret sauce for completing this challenging journey.

I am grateful for my grandparents, who passed away many years ago, my parents, and my brothers and sister. I may be physically separated from this extended loving family, but I am deeply rooted in and shaped by them. There is no existence of an individual personhood outside a family. A family is the only tribe that you are part of forever and wherever.

I am grateful for Miss Caroline Von Stein. She helped review my manuscript, provided valuable editorial comments and grammatical corrections, and made my Chinglish read a lot better than it used to be.

Finally, I want to give myself a pat on the back. After completing two full marathons in my 40s, I did it again completing a different but equally challenging marathon – a doctoral degree in my 50s. Mental and physical health are the Yin and Yang of life and happiness; Having both is a blessing and keeping both is a lifelong pursuit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Over the years, similar terms have been used to describe the process and activities that uncover patterns and discover knowledge from data to enable evidence-based decision-making and problem-solving. Business analytics, business intelligence (BI), data analytics, knowledge discovery in databases (KDD), data mining, machine learning, and data science are among the most popular of them.

Of these terms, business analytics and BI appear to put more weight on the business side, while data mining, machine learning, and data science focus on the technology side. KDD is better known in academia than in industry and it is also the name for one of the three leading data analytics process models (Piatetsky-Shapiro, 2000). Data analytics appears to have a more generic, neutral, and balanced appeal and has gained favorable acceptance as a universally accepted term. It covers the whole spectrum between business analytics and data science.

For simplicity and consistency, this dissertation uses *data analytics* as a generic umbrella term to describe the process and activities that use statistical models, machine learning algorithms, and software tools to bring out hidden patterns, discover knowledge, inform decision-making, and support problem-solving through systematic collection, acquisition, preparation, analysis, presentation, and communication of data and information.

1.1 Background

This section provides a historic account of data analytics in the backdrop of rapid social and technological changes brought about by digital computers and the Internet. The opportunities brought by digitalization and globalization have been met with challenges of information overload, a term popularized by the best-selling book *Future Shock* (Toffler, 1970). Data analytics plays a key role in solving this dilemma by augmenting human intelligence with

artificial intelligence leading to the discovery of hidden knowledge in the tsunami of data to support informed decision-making and effective problem-solving. However, the data analytics community of practice has its own challenges in achieving impactful outcomes and delivering valuable benefits. This study is conceived and carried out from this context in response to this problem.

1.1.1 Opportunities and Challenges of the Information Age

With the invention of digital computers in the late 1940s, a new information technology industry was born, which in turn led to the proliferation of electronic data and propelled the society from the Industrial Age into the Information Age (Toffler, 1980). The advent of the now-ubiquitous Internet in the late 1990s, and the subsequent innovations in mobile computing and social media networks, have unleashed the power of information and communications technology (ICT) and created both unprecedented opportunities and challenges for individuals, communities, businesses, and governments around the globe.

As more and more data are generated from daily personal interactions, business transactions, and political discourses via the globally connected digital social media networks enabled by rapidly expanding ICT, organizations and individuals alike for the first time have the tools and the data to discover insightful knowledge, make impactful decisions in real or near-real time, and improve the profitability of businesses, the conditions of the society, and the quality of individual lives.

At the same time, society and individuals face the challenges of information overload (Toffler, 1970) and struggle to keep up with the exponential growth of data in volume, velocity, and variety, commonly referred to as “Big Data.” The maturing of cloud computing, coupled with the emerging Internet of Things (IoT), such as ubiquitous sensors and intelligent devices,

will rapidly generate even more data adding fuel to the fire.

1.1.2 Benefits & Barriers of Data Analytics

Human beings have practiced the discovery of useful knowledge from observations since the very beginning of civilization. However, the modern discipline of data analytics grew out of the opportunities and challenges of the Information Age. In the past three decades or so, data analytics in the form of business intelligence, knowledge discovery, and data mining has helped businesses, governments, and individuals deal with the challenges of information overload and create value from the vast amount of data in order to benefit society.

While data analytics has great potential to help organizations and individuals with decision making and problem solving, it is not without challenges, especially when facing the increasing complexity of the social, business, and technological environment exacerbated by the rapid growth of data.

Despite the leaps and bounds in the analytics platforms, software tools, algorithms, and techniques, many of analytics projects fail to achieve the desired outcomes.

From a project management perspective, according to Gartner (Gartner Inc., 2013), more than half of all analytics projects fail to complete within budget or on schedule or deliver the features and benefits that were optimistically anticipated at their outset. From a value creation perspective, PricewaterhouseCoopers and Iron Mountain surveyed 1,650 European and North American businesses and found that “only 4% of businesses can extract full value from the information they hold”, “43% obtain little tangible benefit from their information” and “23% derive no benefit whatsoever” (Reid, 2015).

Dun & Bradstreet and Forbes surveyed 300 business executives having various responsibilities (including Chief Executive Officer, Chief Financial Officer, Chief Information

Officer, Chief Analytics Officer, Chief Marketing Officer, and Chief Procurement Officer) from a broad range of industries (including financial, insurance, energy, retail, manufacturing, technology, and government) in both Europe and North America. While the findings revealed encouraging trends of businesses embracing analytics and benefiting from their investments in analytics, there were also indications that businesses were slow in adopting advanced analytics and still heavily relied on traditional spreadsheets and static reports. The survey identified a dozen or so challenges businesses face in unlocking the power of knowledge in data and using analytics to inform decisions. These challenges range from budget constraints to technology adoption, from data management and governance to skills gaps, from business and technology alignment to demonstration of return on investment (Dun and Bradstreet & Forbes, 2017).

1.2 Statement of the Problem

Having an established standard process model is critical to the success of any profession. As stated in the book *Data Science for Business: what you need to know about data science and data-analytic thinking*, data analytics "involves the application of a substantial amount of science and technology, but the proper application still involves art as well. But as with many mature crafts, there is a well-understood process that places a structure on the problem, allowing reasonable consistency, repeatability, and objectiveness" (Provost & Fawcett, 2013).

To combat the complexity of data analytics projects and to ensure quality outcomes and returns on investments, the data analytics community of practice has developed various process models and best practices since the late 1980s. In 1989, Piateskey-Shapiro first coined the term "knowledge discovery in databases" (KDD), which later became the name for the KDD process model (Fayyad, Piatetsky-Shapiro, & Smyth, 1996a, 1996b; Piatetsky-Shapiro, 2000). Around the same timeframe, several prominent companies, including SPSS, championed the Cross-

Industry Standard Process for Data Mining (CRISP-DM) funded by the European Commission (Wirth & Hipp, 2000). Another well-known process model with strong statistical orientation is Sample-Explore-Modify-Model-Assess (SEMMA) from the SAS institute as part of its data mining software product SAS Enterprise Miner (SAS Institute Inc., 2017). CRISP-DM, KDD, and SEMMA have become the three leading process models since then. Over time, both industry and academia have also created various similar process models and best practices.

The benefits of applying the aforementioned process models in analytics projects have been well documented (Maaskant, 2016) (Pramanik, Lau, Yue, Ye, & Li, 2017). These process models have also been tailored to fit specific industries and problem domains (Ahangama & Poo, 2015; Miškuf, Michalik, & Zolotová, 2017; Pramanik et al., 2017; Tahmasebian et al., 2017) .

However, despite the successful applications of these process models, their limitations were also recognized (Azevedo & Santos, 2008; Shafique & Qaiser, 2014). According to surveys conducted in 2007 and 2014 by KDNuggests.com, a leading website for data analytics, only about 60% of data professionals surveyed used one of the three leading process models. A significant 40% of data professionals surveyed either did not use, or used ad hoc or proprietary process models in their data analytics projects (Piatetsky-Shapiro, 2014). Figure 1 summarizes the survey results in a bar chart. As the chart indicates, the distribution has not changed much over the seven-year span.

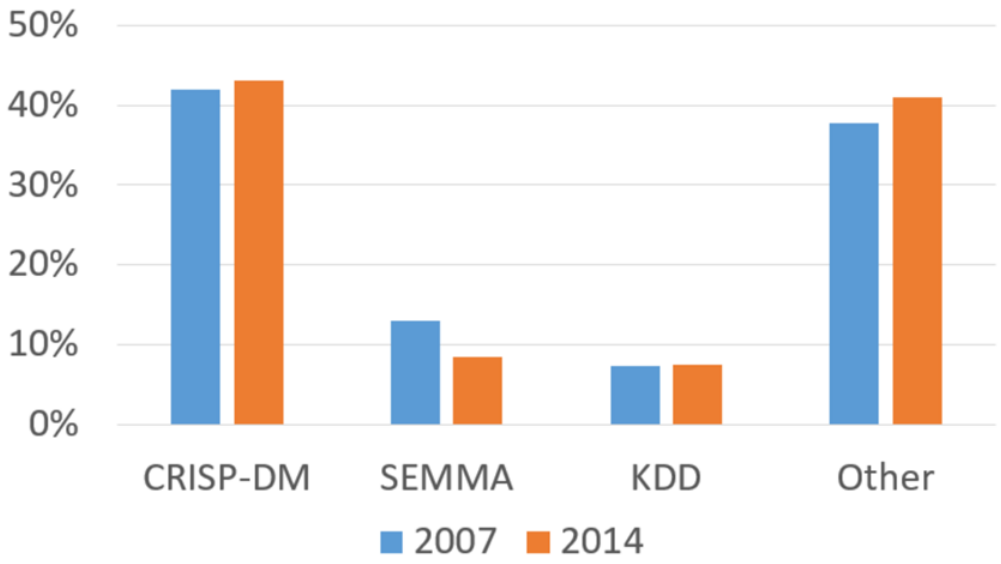


Figure 1. Adoption rate of data analytics process models

This lack of a universally accepted and widely adopted process model inhibits the application of systematic, consistent, repeatable, and transferable best practices for project success and increases the chances of project failures. This deficiency was also acknowledged as the number one of ten challenging problems in data analytics (Yang & Wu, 2006). The authors surveyed conference attendees and found that developing a unifying theory of data analytics was the number one concern and "the current state of the art of data mining research is too 'ad-hoc.' Many techniques are designed for individual problems" and "there is no unifying theory."

There have been various attempts made to create an improved process model that would unify the three leading ones and address their limitations (Mariscal, Marban, & Fernandez, 2010; Martins, Pesado, & García-Martínez, 2016). However, none of these attempts has materialized, and to date a universally accepted methodology is still lacking.

1.3 Purpose and Objectives of the Study

The purpose of this study is to apply the design science research methodology to develop a new and improved data analytics process model that is grounded in a strong theoretical

foundation and unifies the existing models by leveraging their strengths while addressing their limitations. The goal is to help data analytics professionals and business stakeholders gain a deeper and richer understanding of the data analytics process, achieve higher quality outcomes, and deliver greater impact. To achieve this goal, the study sets forth the following two research objectives:

Research Objective #1: Design a new and improved data analytics process model based on a strong theoretical foundation.

Research Objective #2: Evaluate the utility of the new model using an acceptable evaluation method.

1.4 Research Questions

To ground the design of the new model on the right foundation and path, this research starts by attempting to answer the following three questions:

Research Question #1: What are the limitations of the existing data analytics process models?

Research Question #2: What theories should be used as the foundation to design a new and improved process model?

Research Question #3: How well does the new and improved process model apply to a real-world situation?

Answers to the first two questions will help inform the design and development of the new model. Answer to the third question will help improve and evolve the new model.

1.5 Significance of the Study

The exponential growth of big data generated from various sources, including electronic transactions, social media networks, mobile devices, ubiquitous sensors, and the emerging IoT,

bring both challenges and opportunities to individuals, communities, businesses, and governments. Data analytics, enabled by increasing computing power, storage capacity, and advances in machine learning and artificial intelligence, plays a key role in helping society to gain knowledge from the vast amount of data and to make informed and timely decisions to solve real-world human problems. The potential power and benefits of data analytics can only be realized when the community of practice reaches common understanding of the data analytics process and adopts a standard process model which ensures the consistency, validity, repeatability, quality, and maturity of data analytics practices and outcomes.

This study aims to provide a well-designed and practical process model grounded in strong theoretical foundation and design principles and built on the strengths of the existing process models and best practices while addressing their limitations. The proposed process model will provide a conceptual framework and practical guidance to technical professionals, domain subject matter experts, business managers and executives, policymakers, and other stakeholders to achieve better outcomes in their data analytics and problem-solving endeavors.

According to Hevner, March, Park, and Ram (2004), " the key differentiator between routine design and design research is the clear identification of a contribution to the archival knowledge base of foundations and methodologies". This research, through the design of an improved process model and the empirical evaluation of its utility, contributes to the existing body of knowledge with improved methodology and improved understanding of the analytics process.

1.6 Boundaries/Delimitations of the Study

Utilizing a design science research methodology, this study is prescriptive in nature. It does not attempt to explain or predict the phenomenon of data analytics, which requires

empirical study of individual and organizational behaviors. However, this study leverages the results and contributions from empirical research efforts of the existing data analytics process models which provide motivations and urgency for designing a better process model. In addition, to evaluate the utility of the proposed process model, this study employs the traditional empirical methods by conducting an illustrative case study.

In addition to the foundational knowledge, the researcher also draws upon his own rich professional multidisciplinary knowledge and experiences in software engineering, systems engineering, project management, Agile development framework, business strategy, knowledge management, and organizational learning. While these experiences provide tremendous value to guide and enrich the design, they can potentially introduce personal styles and bias into the design.

Due to the time constraint, additional evaluation methods, such as action research and usability tests, have not been considered and can become potential research studies in the future.

1.7 Definition of Terms

The following definition of terms is adopted primarily from the fourth edition of a popular textbook with some minor edits (Sharda, Delen, & Turban, 2016):

Analytics. The science of analysis.

Big Data. Data that is characterized by its volume, velocity, and variety that exceed the reach of commonly used hardware environment and/or capability of software tools to process.

Big Data Analytics. Application of analytics methods and tools to big data.

Business Analytics. The application of analytics to business problems/data.

Business Intelligence. A conceptual framework for managerial decision support. It combines architecture, databases (data warehouses), analytical tools, and applications.

Cloud Computing. Information technology infrastructure (hardware, software, applications, and platform) that is available as a service, usually as virtualized resources.

Cross-Industry Standard Process for Data Mining (CRISP-DM). A sequence of six steps that starts with a good understanding of the business and the need for the data mining project (i.e., the application domain) and ends with the deployment of the solution that satisfied the specific business need.

Data. Raw facts that are meaningless by themselves (e.g., names, numbers).

Data Analytics Process Model. A standardized practice patterns such as steps, associated tasks and activities, principles, and guidelines for planning, organizing, and executing a data analytics effort or project.

Data Mining. A process that uses statistical, mathematical, artificial intelligence, machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases.

Data Science. An interdisciplinary field about scientific methods to extract knowledge from data. It is similar to data mining but stresses rigor, formality, and use of advanced machine learning and artificial intelligence.

Data Warehouse. A physical repository where relational data are specifically organized to provide enterprise-wide, cleansed data in a standard format.

Database. A collection of files that is viewed as a single storage concept. The data are then available to a wide range of users.

Datum. A piece of information/fact. Singular form of data.

Decision Support Systems (DSS). A conceptual framework for a process of supporting managerial decision making, usually by modelling problems and employing quantitative models for solution analysis.

Internet of Things (IoT). The technological phenomenon of connecting a variety of devices in the physical world to each other and to the computing systems via the Internet.

Knowledge Discovery in Databases (KDD). A term coined by Piatetsky-Shapiro in 1989. A process of discovering useful knowledge from a collection of data proposed by Fayyad, Piatetsky-Shapiro, and Smyth in 1996.

Process Model. A standardized practice pattern such as steps, associated tasks and activities, principles, and guidelines for planning, organizing, and executing a human endeavor. In modern society, human endeavors typically involve the use of technologies and the collaboration of people.

Sample-Explore-Modify-Model-Assess (SEMMA). A process for data mining projects developed by the SAS Institute and is incorporated in its SAS Enterprise Miner software product.

1.8 Structure of this Document

Chapter two starts with a comprehensive review of literature related to data analytics process models. The goal is to discover, acknowledge, critique, and leverage the scholarly and practical work already done in the field of data analytics process model. Gaps in the existing body of knowledge will be identified along the way, which will provide impetus and motivations for this study. This chapter also reviews literature related to the concepts and theories of information and communications, which provide theoretical basis to inform the design of a new model.

Chapter three introduces and justifies the use of the design science research and lays out a detailed plan for conducting this study, including the design and evaluation of the new model.

Chapter four describes the new process model in detail.

Chapter five documents the process and results of the evaluation of the model using an illustrative scenario.

Chapter six describes the researcher self-reflection of the effort in evaluating the new model using illustrative scenario, documents the results of the reviews of the model by two experts, and provides a summary of the research.

CHAPTER 2: REVIEW OF THE LITERATURE

This chapter presents a comprehensive review of the literature related to the study. Section 2.1 surveys the leading data analytics process models and assesses their strengths and limitations; Section 2.2 reviews the efforts made by fellow researchers aiming to improve the process models and to create better ones. The strengths and limitations of these efforts are also analyzed. Section 2.3 reviews foundational theories that will inform the design of the new process model. This review of the literature provides the motivations and guidance to the study of this dissertation.

2.1 Leading Data Analytics Process Models

As was mentioned in the chapter one, three leading data analytics process models take about 60% of share based on surveys of professionals (Piatetsky-Shapiro, 2014). **Table 1** provides a summary of them followed by a brief introduction and analysis on each of the three process models.

Table 1

Leading Data Analytics Process Models

Acronym	Name	Year Created	Creator
CRISP-DM	Cross-Industry Standard Process for Data Mining	1999	The consortium of SPSS (now part of IBM), Teradata, Daimler AG, NCR, and OHRA funded by European Commission (Wirth & Hipp, 2000). IBM's data mining software SPSS Modeler provides built-in support for the model (IBM, 2017).
SEMMA	Sample-Explore-Modify-Model-Assess	1997	SAS Institute. SAS's data mining software SAS Enterprise Miner provides built-in support for this model (SAS Institute Inc., 2017).
KDD	Knowledge Discovery in Databases	1996	Developed by academia (Fayyad et al., 1996b).

2.1.1 CRISP-DM

The Cross-Industry Standard Process for Data mining (CRISP-DM) is the de facto standard and the most adopted data analytics process model with 43% of share (Piatetsky-Shapiro, 2014). It was developed by the European consortium of several technology companies (Wirth & Hipp, 2000) and is currently incorporated in the IBM SPSS Modeler software product (IBM, 2017). In this process model, a data analytics lifecycle consists of six phases:

1) Business understanding
2) Data understanding
3) Data preparation
4) Modeling
5) Evaluation

6) Deployment

The model also defines various tasks to be performed within each phase. The six phases are both sequential and interrelated reflecting the iterative nature of the data analytics process as shown in **Figure 2**.

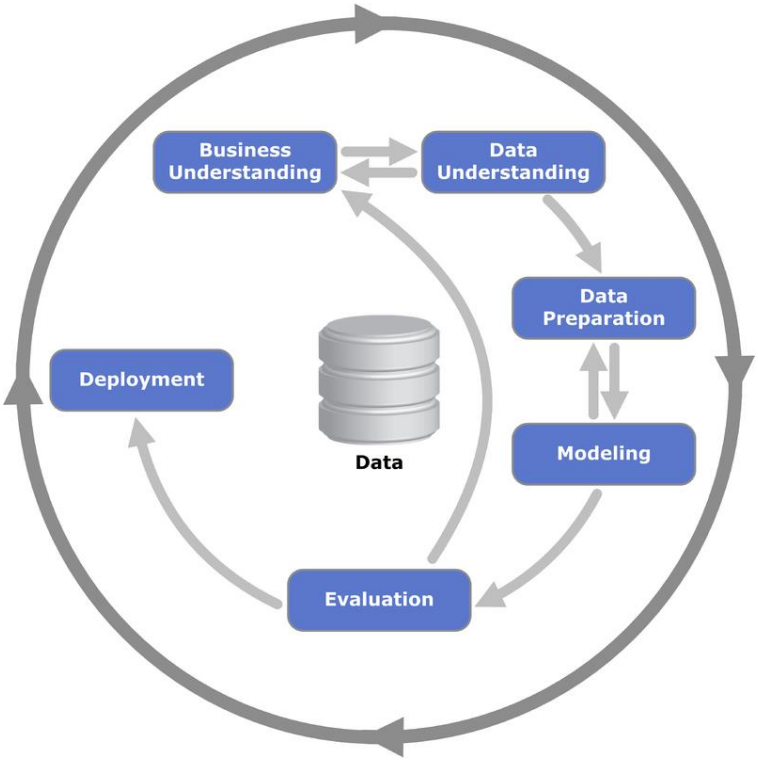


Figure 2. The six phases of the CRISP-DM process model

The strength of this model lies in its first step of "Business Understanding". Starting out with understanding of business problems, needs, goals, and objectives will align data analytics effort with the business strategy and ensure the outcomes add values and have impact on business.

However, this process model has several limitations. First, it treats data analytics as a pure technical effort by technical professionals and does not emphasize on the close collaboration between both technical and business professionals; Secondly, it does not incorporate the process of knowledge discovery, creation and application involving human-

machine symbiosis and human collaboration. Lastly, this model has not been updated to the changes in technology and business environment since its inception in the late 1990s. The consortium published a user guide providing detailed explanations of the methodology, process models, and activities in 1999 (Chapman et al., 2000). However, the model and its guide stayed as version 1.0 and has not been updated since then. There was a plan to update the model to version 2.0 but it failed to materialize. The consortium has been inactive, and its website has been unavailable for many years. This process model remains the most favorable by the industry despite the need to be updated and to stay current with the rapid changes in both social and technical landscape (Piatetsky-Shapiro, 2014).

2.1.2 SEMMA

The Sample-Explore-Modify-Model-Assess (SEMMA) process model was developed by SAS Institute, a leading statistical and analytics software provider. SEMMA consists of five sequential steps in an iterative cycle as shown in **Figure 3**.

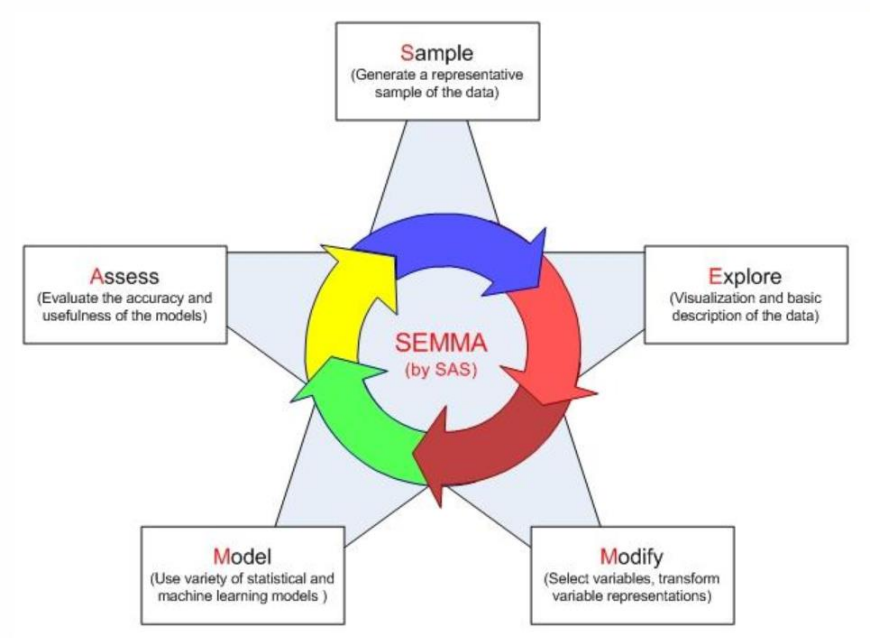


Figure 3. The SEMMA process model

The model is incorporated in SAS Enterprise Miner software product (SAS Institute Inc., 2017). SEMMA has a strong statistical orientation and starts straight from selecting the representative sample data. The understanding of business situation, needs, and goals along with data collection and understanding are assumed prior to the starting of the SEMMA process. This model is more useful for the users of the SAS data mining software and less useful to serve as a conceptual framework and guidance for general data analytics projects.

2.1.3 KDD

While CRISP-DM and SEMMA were developed by industry, Knowledge Discovery in Databases (KDD) process model was developed by academia (Fayyad et al., 1996b).

The KDD process model defines the data analytics lifecycle with nine steps:

1) Learning the application domain
2) Creating a target dataset
3) Data cleansing and preprocessing
4) Data reduction and projection
5) Choose the function of data mining
6) Choose the data mining algorithms
7) Data mining
8) Interpretation
9) Use the discovered knowledge

KDD strikes a balance between business and technology. It places the technical aspects of data analytics including data processing, algorithm selection, and data mining in the business context. It emphasizes the application of the discovered knowledge in decision-making and problem-solving. Another valuable contribution of this model is its coverage of knowledge

management and recognition of the limitations of technology. It was highlighted that "Researchers and practitioners should ensure that the potential contributions of KDD are not overstated and that users understand the true nature of those contributions along with their limitations" (Fayyad et al., 1996b).

Also notice that the KDD process model has a staircase process diagram that starts from data and ends at knowledge as shown in **Figure 4**. This diagram does not match exactly the nine-step outlined above. This mismatch is a deficiency of this process model from usability perspective. In addition, this process model does not stress the iterative nature of the data analytics process and does not have feedback loops specified in the model.

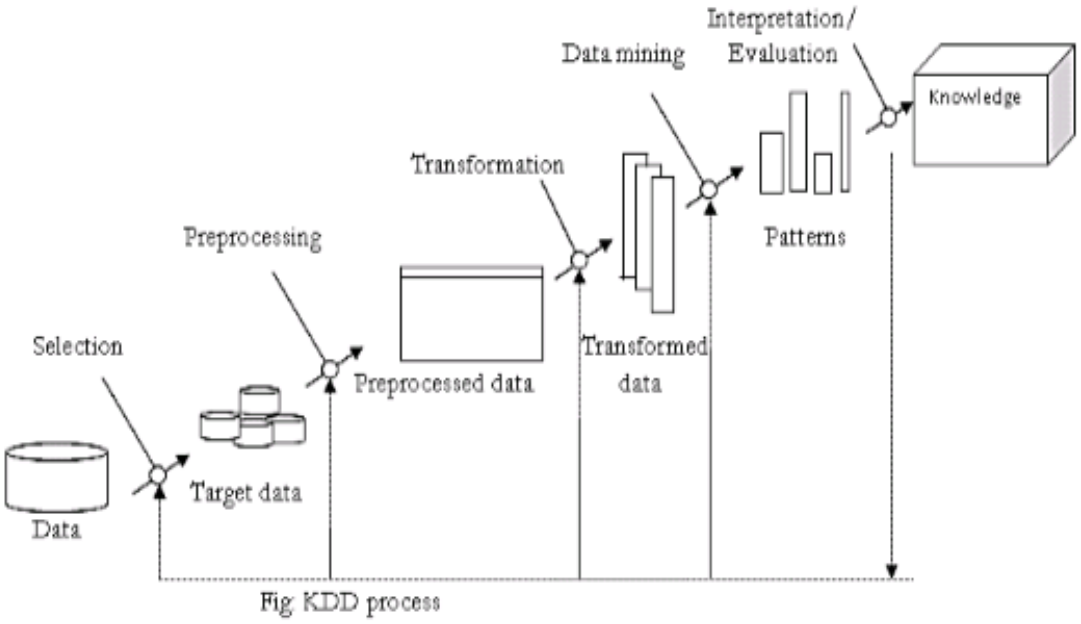


Figure 4. The Knowledge Discovery in Databases (KDD) process model

2.1.4 Summary

While the three leading data analytics process models are unique in their own way, they also share similarities. **Table 2** shows how the three models align with each other.

Table 2

Comparison of the Three Leading Data Analytics Process Models

CRISP-DM	SEMMA	KDD
1. Business understanding		1. Learning the application domain
2. Data understanding		
3. Data preparation	1. Sample	2. Creating a target dataset 3. Data cleansing and preprocessing 4. Data reduction and projection
4. Modeling	2. Explore 3. Modify 4. Model	5. Choose the function of data mining 6. Choose the data mining algorithms 7. Data mining
5. Evaluation	5. Assess	8. Interpretation
6. Deployment		9. Use the discovered knowledge

2.2 Efforts to Improve the Process Models

As CRISP-DM, KDD, and SEMMA were applied to various industry and academic projects and gained popularities, extensions have been developed to meet the needs of specific domains and attempts have been made to improve and unify them along the way. The section reviews these efforts.

2.2.1 Refined Data Mining Process

Mariscal et al. (2010) provided a comprehensive review of fourteen process models including the leading three and their extensions. These models have more commonalities than differences and each has its own strengths and limitations. To unify these similar but disparate models, the authors proposed "a new data mining and knowledge discovery process model named Refined Data Mining Process for developing any kind of data mining and knowledge discovery project" (p. 137). This generic new model defined three high-level processes: analysis, development, and maintenance. Each process is further divided into sub-processes. Overall,

seventeen sub-processes were defined drawing from steps and activities of the fourteen existing models. This represents the first attempt to develop a unifying approach to data analytics.

The authors provided high-level descriptions for the main processes and sub-processes without further elaboration and left the full development of the process model for future research. Nonetheless, this model's three overarching processes resembled the high-level systems and software development lifecycle of requirements analysis, design and development, and deployment and maintenance. This helps to align the data analytics process with the more mature systems and software engineering discipline.

One major drawback of this process model is its lack of visual representation. Its narrative-based description has many details and the 17 subprocesses are too many for people to remember. This drawback limits its usability and its practical adoption.

2.2.2 Integrated Knowledge Discovery & Data Mining

Sharma and Osei-Bryson (2010) believed that existing data analytics such as CRISP-DM stops at the descriptive level on the "What" and failed to provide the prescriptive guidance on the "How" to be applied effectively. They proposed an Integrated Knowledge Discovery & Data Mining (IKDDM) process model that is more granular. By making the dependences between various tasks explicit, they believe the integrated model will be able to support semi-automation of some of the tasks and hence improves efficiency and effectiveness.

However, this approach turns a process model into a detailed design flowchart and makes it much more complicated, harder to grasp, and less flexible to accommodate diverse analytics problems. **Figure 5** shows the complexity of the model which leads to its lack of usability.

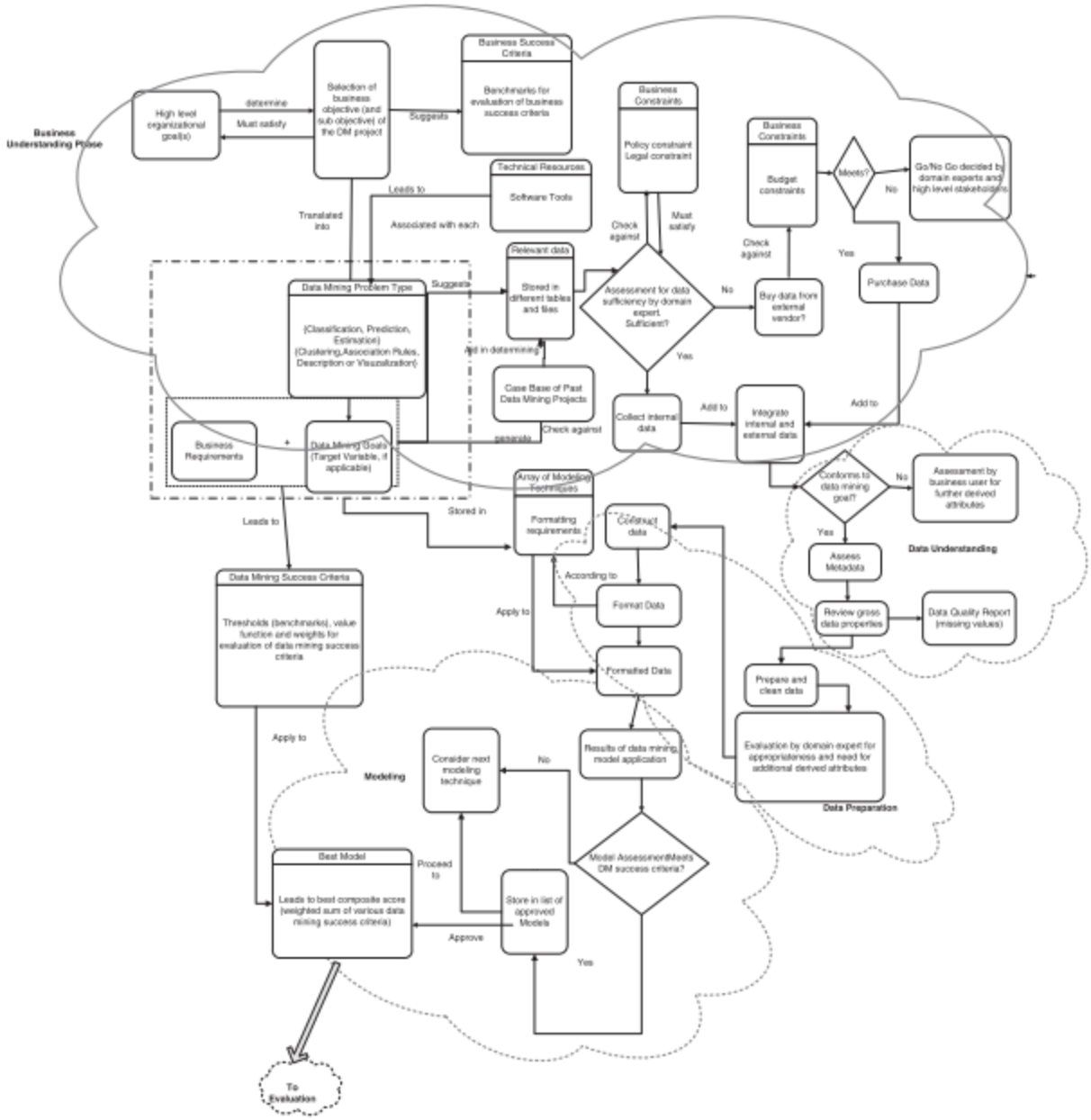


Figure 5. The IKDDM process model (Sharma & Osei-Bryson, 2010)

2.2.3 Unified Theoretical Framework for Data Mining

Khan, Mohamudally, and Babajee (2013) developed the unified theoretical framework for data mining by "formulating a unified data mining theory (UDMT) where the data mining processes; clustering, classification and visualization are unified by means of composition of functions". It is worth to note that the concept of unification in this model is much narrower than

that of the Refined Data Mining Process model described in the previous section. The authors further clarified that the foundation of the UDMT is that "without clustering there is no classification, without classification there is no visualization and hence without visualization there is no 'knowledge'". Based on that proposition, the authors defined four steps for knowledge extraction from a dataset:

1) Create (appropriate) partitions of the dataset
2) Create the clusters of each partition (clustering)
3) Construct decision rules of each cluster (classification)
4) Plot 2D or 3D graphs of each rule or classifier (visualization)

This model is narrowly focused on the technical aspects of data analytics and offers neither coverage nor guidance on the end-to-end data analytics process involving multidisciplinary collaboration of people in an organizational context.

2.2.4 Snail Shell Process Model

Li, Thomas, and Osei-Bryson (2016) developed the snail shell process model as an extension to the CRISP-DM model. The authors argued that traditional data analytics process models tend to target well-defined problems and offer little guidance on analyzing ill-structured complex problems. In addition, traditional models tend to deal with relatively smaller and static datasets and require no ongoing maintenance of statistical algorithms and computational models after they are deployed. To adapt to the increasing complexity of business situations and the increasing volume, variety, and velocity of big data, the snail shell model prefixes or front-loads the six-phase in the CRISP-DM model with a new phase called "Problem formulation" and appends or back-loads it with a new phase called "Maintenance" as shown in **Figure 6**.

Aside from the two additional phases, the snail shell model retains the core concepts of the original CRISP-DM model and does not address the its limitations outlined earlier such as lack of emphasis on human collaboration, knowledge discovery, creation, and application.

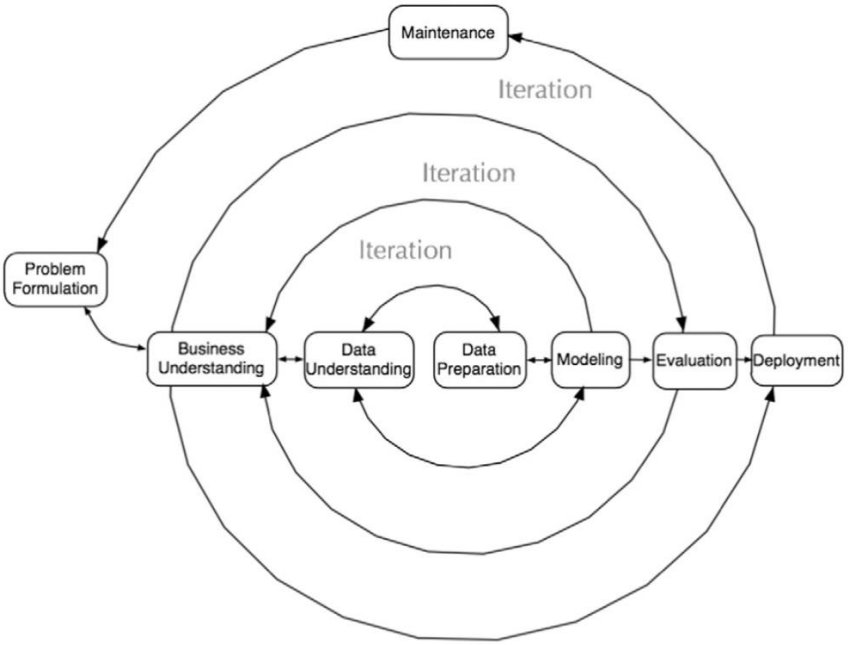


Figure 6. The Snail Shell Process Model adapted from (Li et al., 2016)

2.2.5 The Domain-Driven Data Mining

While the previous models extend or enhance the CRISP-DM model, the Domain-driven Data Mining model goes beyond the pattern discoveries using statistical models and computational algorithms and incorporates human factors, knowledge management, decision making, and business contexts into the process of data analytics. It is a transformation that moves from the traditional data-centric and technology-driven pattern discovery to a domain-driven and multi-perspective knowledge delivery (Cao, 2010). **Table 3** compares the two different paradigms.

Table 3

Data-centric vs Domain-driven Data Mining (Cao, 2010)

Aspects	Data-Driven	Domain-Driven
Rationale	Data tells the story	Data and ubiquitous intelligence disclose problem-solving solutions
Objective	Innovative and effective algorithms	Effective problem-solving
Data	Abstract, synthetic and refined data	Real-life data and surrounding information
Process	One-off	Multiple-step, iterative and interactive on demand
Mechanism	Automated	Human-centered or human-mining-cooperated
Infrastructure	Closed pattern mining systems	Closed-loop problem-solving systems in open environment
Usability	Predefined models and processes	Ad-hoc, dynamic and customizable models and processes
Deliverable	Patterns	Business-friendly decision-support actions
Deployment	Solid validation	Well-founded artwork in problem-solving
Evaluation	Technical metrics	Tradeoff between technical significance and business expectation

This model presented a valuable concept of ubiquitous intelligence which consists of the following five intelligences:

1) In-depth data intelligence
2) Domain intelligence
3) Network intelligence
4) Human intelligence
5) Social intelligence

The ubiquity of and the need to integrate the various intelligences provide impetus for incorporating human factors, human-computer interaction, and social and cultural context into data analytics process.

One noticeable limitation of this model is its excessive use of arcane mathematics symbols and equations in the elaboration and explanation of its design and techniques. This makes it inaccessible to non-technical people and runs counter to its domain-driven principle.

2.2.6 Business Analytics Methodology

The business analytics methodology (BAM) is by far the most recent and most business-friendly process model for data analytics. Recognizing the lack of attention given to business

contexts and business objectives by the KDD and SEMMA process model, Hindle and Vidgen (2017) started their research by asking the question "how can organizations align their business analytics development projects with their business goals?" and completed it with a four-stage model expanding on the business understanding of the CRISP-DM model as shown in **Figure 7**.

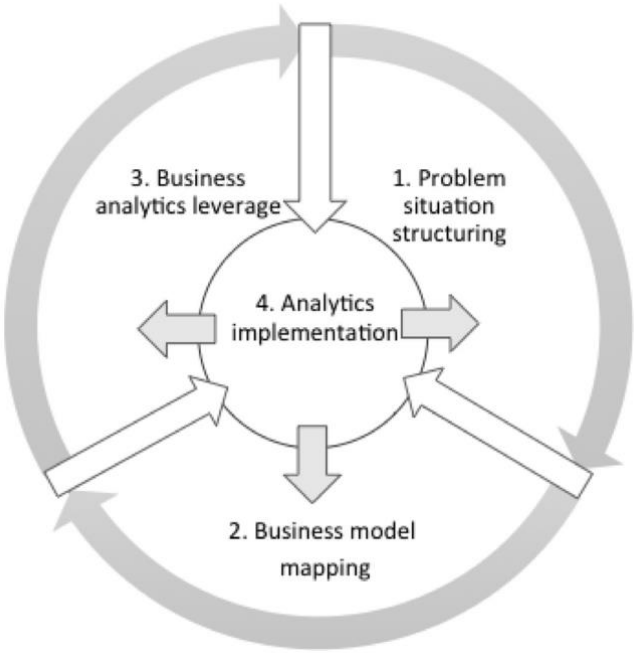


Figure 7. The Business Analytics Methodology (Hindle & Vidgen, 2017)

BAM drew upon knowledge from several areas including business modeling, systems thinking, and soft systems methodology to help analyze ill-defined and complex business problems, elaborate business needs, and articulate business objectives. This heavy front-load of business requirements analysis establishes a strong foundation to kick off the business analytics effort.

While this business-driven model clearly has advantages over the traditional technology-driven models, it does not account for the iterative nature of the data analytics process and could potentially suffer from the downside of the Waterfall methodology prevalent in the traditional software and systems engineering practices.

2.3 Theoretical Basis of the New Model

The previous section surveyed the three leading data analytics process models and identified their strengths and limitations. The efforts to extend and improve them were also evaluated and their strengths and limitations were analyzed. The results of the analysis provided the motivations for this research to design a new and improved process model that would address the limitations and leverage the strengths of the existing process models.

The section reviews relevant theories from the existing body of knowledge that provide a strong theoretical foundation for building this new and improved process model. **Table 4** summarizes these theories followed by detailed explanations.

Table 4

Theoretical Basis of the New Process Model

Theory	**Brief Description**	**Relevance to Data Analytics**
Wisdom Pyramid	Data, Information, Knowledge, and Wisdom (DIKW) represents the increasing level of human observation and understanding of the reality.	Data analytics process is to gain broader and deeper understanding of a phenomenon by extracting information from data, gaining insights from information and applying the discovered knowledge to inform decision-making.
Semiotic Ladder	The ladder has six steps. Between physical at the bottom and social world at the top there exist four escalating steps: empirics, syntactics, semantics, pragmatics like the Wisdom pyramid.	Data analytics process starts with observation and data collection at the physical level through sensors and devices and reach the final goal of solving social problems through evidence-based decision making and human collaborations.
Shannon-Weaver Model of Communication	There are seven components of communication: the sender, the encoder, the channel, the decoder, the receiver, the noise, and the feedback. Efficiency is affected by noise. Accuracy is improved via feedback.	Shannon-Weaver Model of Communication lays the foundation for computer science and information technology which provide the techniques and tools for data analytics.
Schramm Model of Communication	Interpretations of messages are influenced by people's field of experience. The message sender and the receiver may have different context to interpret the same message.	Schramm Model covers human communication and provides the theoretical basis for the social aspect of data analytics.
Human-Machine Symbiosis	Human develops technologies and tools to extend our physical and cognitive capability. Human beings and machines complement each other,	Data analytics is an exploratory and discovery process in which both human and artificial intelligence are applied. Machines perform tasks that

Theory	Brief Description	Relevance to Data Analytics
	and the partnership creates much more powerful capabilities.	require mass storage and speedy computation of big data and human beings perform tasks that require intuition, judgement, and purposes.
Human-Centered Design	The design of any system that is intended for human to use must put people first and accommodate people's needs, capabilities, and behaviors.	Data analytics process model is intended to provide the roadmap and guidance to practitioners and stakeholders in their analytics efforts. The process model must be designed with their needs in mind following the human-centered design principles.
Systems Engineering V Model	There are seven stages in the life cycle of a system from initial concept of operations to the final operations and maintenance. Validation effort is built in each step that provide feedback to ensure the previous steps were performed correctly and adequately.	Data analytics can be considered as a software engineering effort where technologies and tools are used to solve real-world problems. It requires a balance of rigor and agility.
Nine Laws of Data Mining	The Nine Laws address various aspects of data analytics. Business objectives, not technologies, are the drivers of data analytics. In addition, data analytics process must be agile to adapt to constant changes in business needs, technology innovations, and external environments.	Data analytics process model should embrace the principles articulated in the nine laws and incorporate the best practices they espouse. The process model must be driven by value instead of by technology and must be agile instead of rigid.

2.3.1 Wisdom Pyramid

The Wisdom Pyramid is also known as DIKW hierarchy where DIKW stands for Data, Information, Knowledge, and Wisdom.

Table 5 provides two complementary definitions of DIKW by two leading scholars, Zeleny (1987) and Ackoff (1989). The wisdom hierarchy provides a lens to look at data analytics process in an escalating order from deriving information from data, discovering knowledge from information, and combining knowledge, intuition, value, and belief to gain wisdom that leads to informed decisions and impactful actions.

Table 5

The Definitions of Data, Information, Knowledge, & Wisdom

Term	**Zeleny (1987)**	**Ackoff (1989)**
Data	Know nothing	Symbols
Information	Know what	Data that are processed to be useful; provides answers to who, what, where and when questions
knowledge	Know how	Application of data and information; answers how questions
Wisdom	Know why	Evaluated understanding

2.3.2 Semiotic Ladder

Semiotics is the study of signs, symbols, and their meanings. In his seminal work, Stamper (1993) created the semiotic ladder as part of the research on management and information systems in the purview of organizational semiotics as shown in **Figure 8**.

The semiotic ladder is like the wisdom pyramid but provide more granular layers starting from the physical world and ending at the social world. The physical world includes sensors, devices, and systems that generate data for the analytics to consume. The social world is where

people and organization apply intuition, judgement, value, and beliefs to the information and knowledge discovered from the data to make sound decisions and solve real-world problems.

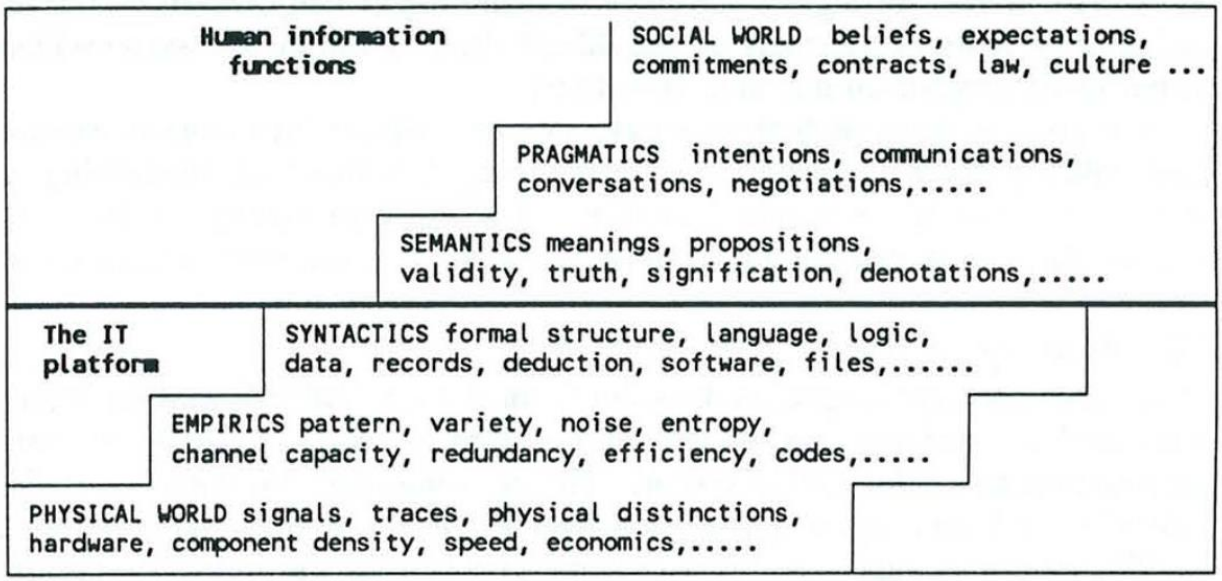


Figure 8. The semiotic ladder (Stamper, 1993)

2.3.3 Shannon-Weaver Model of Communication

Communication is at the heart of our personal life and social discourse. It is both a technical and a social matter. At the technical level, it concerns about the efficiency and accuracy of information transmission. At the social level, it concerns about the effectiveness of information exchange and understanding of meanings. In academia, there are many theories and models of communications. The most fundamental one is the Shannon-Weaver Information Theory (Shannon & Weaver, 1949), also known as Model of Communication, which lays the foundation for the modern computer science and information technology.

Shannon and Weaver were the first to make information measurable using binary digits (bits) (Stone, 2015). The Shannon-Weaver model represents the technical aspect of communication. It aims to minimize the impact of noise along the communication channel and to improve the efficiency of the information transmission through encoding and decoding. Shannon

and Weaver made the distinction between information and meaning and stated that "the word information, in this theory is used in a special sense that must not be confused with its ordinary usage. In particular, information must not be confused with meaning". As seen in **Figure 9**, this model has seven components:

- The sender
- The Encoder
- The channel
- The Decoder
- The Receiver
- The Noise
- The Feedback

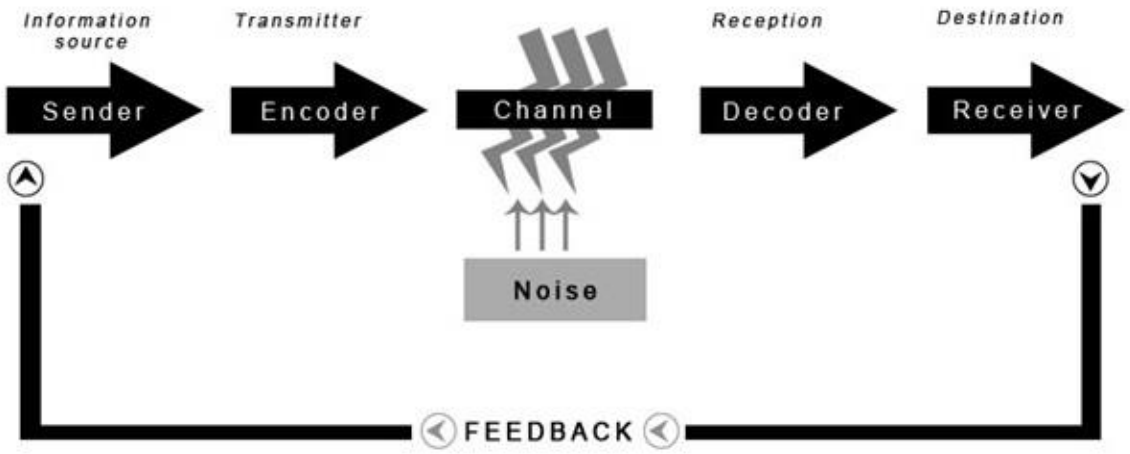


Figure 9. Shannon-Weaver model of communication

2.3.4 Schramm Model of Communication

Many scholars extend the Shannon-Weaver model when applying it to various domains. Notably among them is Schramm Model of Communication (Schramm, 1954) which has been widely used in social and political science.

Schramm (1954) introduced the concept of "field of experience" to highlight the differences in human experiences and the need to achieve mutual understanding through constant

and continuous feedback loops. In contrast to Shannon and Weaver model which focuses solely on the information itself and makes no reference to meaning, Schramm model focuses on the meaning of information and how the social context and individual experience influence the interpretation of information. The field of experience includes a person's cultural background, knowledge, and experiences which can influence his or her understanding and interpretation of the messages received and further influence his or her reactions and responses to the messages. Due to the influence of the field of experience on the communications, it is vital for constant and continuous feedback between the parties involved. **Figure 10** shows the overlap of the field of experiences from both the sender and receiver for both to reach shared understanding.

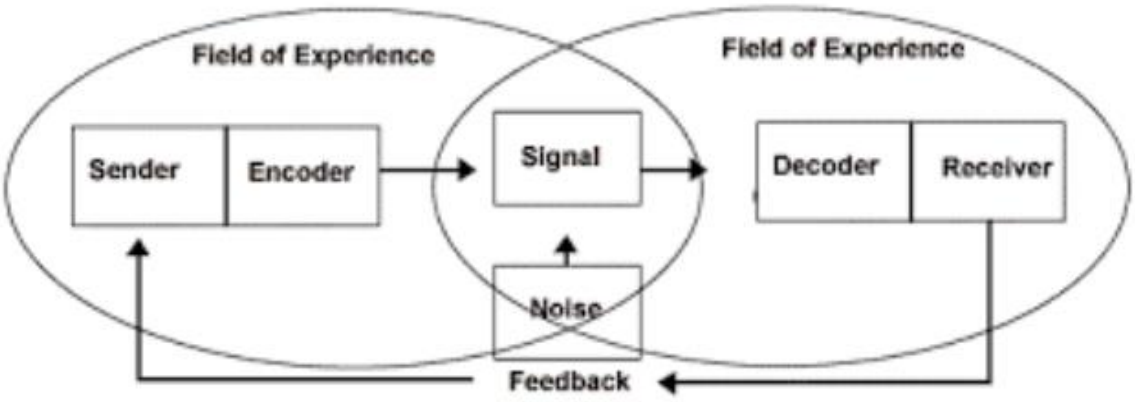


Figure 10. Schramm model of communication

2.3.5 Putting Pieces Together

Figure 11 illustrates the relationship between the four theories and the synergy among them.

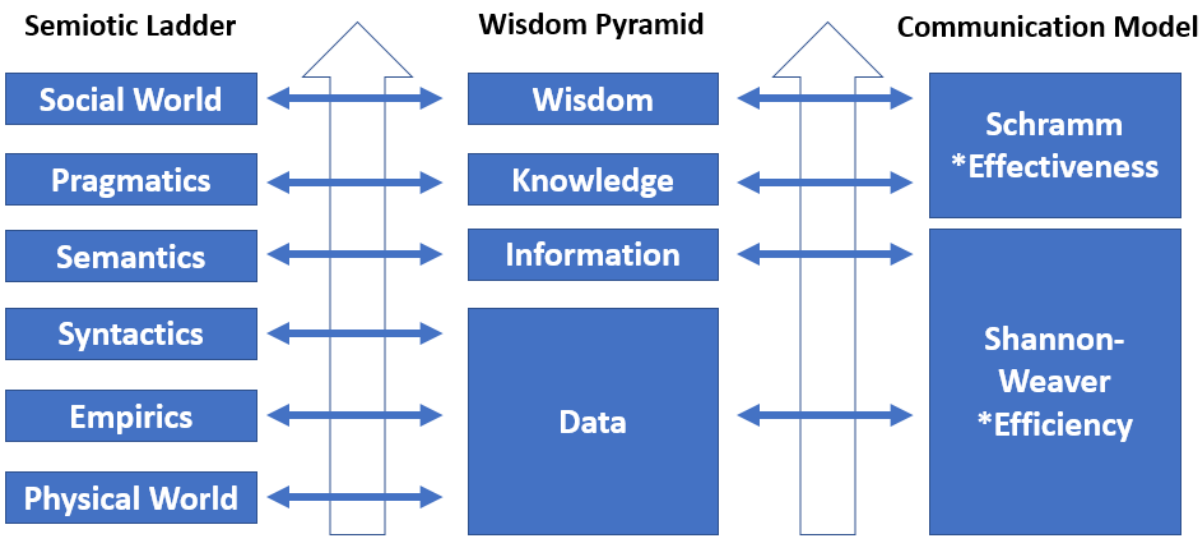


Figure 11. Putting pieces together – the fusion of four theories

At the center is the wisdom pyramid, which provides the theoretical constructs for the A2E Model. To the left is the Semiotic Ladder representing the increasing level of human understanding from the lowest level, our senses of the physical world, to the highest, our deep understanding of the social world. To the right are the two models of communication stacked one on top of the other. At the bottom is the Shannon-Weaver Model, which deals with data and information in the technical realm; on the top is the Schramm Model, which deals with knowledge and wisdom in the social realm.

The complexity increases as we move from physical realm to social realm. At the same time, the need increases to achieve deeper understanding of technology and its role in human activities. The data analytics process is like claiming the pyramid starting from the base with data and reaching the top with wisdom.

To summarize, data analytics activities should not be treated as purely technical tasks performed by only technical professionals using tools and techniques. Data analytics processes

must incorporate technology into the overall human experiences, organizational culture, business processes, and social value chain. **Table 6** summarizes the key takeaways.

Table 6

The Key Takeaways from the Four Theories

Theory	Key Takeaway
Wisdom Pyramid	Moving from senses to higher consciousness
Semiotics Ladder	Seeking meaning from signs and symbols
Shannon-Weaver's Model	Minimizing noise to achieve communication efficiency
Schramm's Model	Breaking human barriers to achieve common understanding and communication effectiveness

2.3.6 Human-Machine Symbiosis

More than half a century ago, American psychologist and computer scientist, J. C. R. Licklider (1960) envisioned how men and machines could potentially work together to perform complex tasks. He coined the term "Man-Computer Symbiosis". Licklider envisioned that:

> *In the anticipated symbiotic partnership, men will set the goals, formulate the hypotheses, determine the criteria, and perform the evaluations. Computing machines will do the routinizable work that must be done to prepare the way for insights and decisions in technical and scientific thinking (Licklider, 1960).*

Licklider hoped that "in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today."

Farooq and Grudin (2016) believed the traditional human-computer interaction (HCI) as part of the human-centered design (HCD) has now evolved into human-computer integration, a more

advanced partnership relationship beyond the simple request-response interactions, like the Licklider's man-computer symbiosis.

Data analytics relies on the close partnership between human intelligence and artificial intelligence to solve complex problems that are difficult to tackle by either men or machines alone.

Today, Licklider's hope has become a reality thanks to the advances in technology especially in big data, machine learning, natural language processing, and artificial intelligence. Data analytics, particularly big data analytics, has become the embodiment of Licklider's man-computer symbiosis.

The design of the new model will incorporate the man-computer symbiosis as a key concept.

2.3.7 Human-Centered Design (HCD)

In his popular book, *The Design of Everyday Things*, author Don Norman, an industrial psychologist and leading design expert, wrote eloquently, "Human-centered design (HCD), is an approach that puts human needs, capacities, and behavior first, then designs to accommodate those needs, capacities, and way of behaving" (Norman, 2013, p. 8). The goal of human-centered design is to enhance the "effectiveness and efficiency, improves human well-being, user satisfaction, accessibility and sustainability; and counteracts possible adverse effects of use on human health, safety and performance" (International Organization for Standardization, 2010).

One key to put people first is to keep the design simple and minimal to avoid the unnecessary complexity and information overload. According to the psychologist George Miller (1956), human brain can only deal with seven plus or minus two chunks of working memory effectively. Human will have trouble processing information beyond the maximum of nine

chunks. The less the number of the chunks, the easier for human to process information. Good design strives to reduce the memory load and avoid information overload for people.

The improved data analytics process model should have manageable number of steps and be easy to understand, remember, use, and repeat.

Hevner et al. (2004) believed designs should have styles and artifacts should be "aesthetically pleasing" echoing Simon (1996) and Norman (2013).

The new model will follow the HCD principles and be designed with simplicity and elegance in mind.

2.3.8 Systems Engineering V Model

The systems engineering V model (Forsberg, Mooz, & Cotterman, 2005) was developed to represent the systems development lifecycle and to guide the development process and project management.

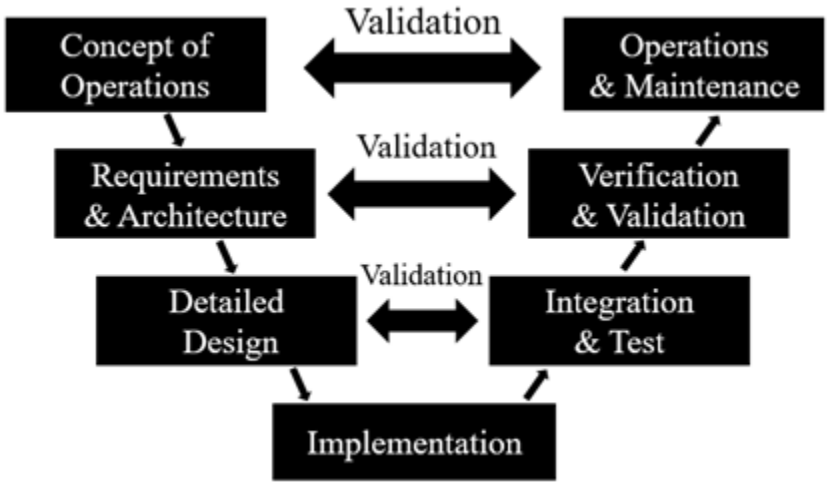


Figure 12. The systems engineering V model

As shown in the **Figure 12**, there are seven stages in the lifecycle of a system:

1. Concept of Operations
2. Requirement & Architecture

3. Detailed Design
4. Implementation
5. Integration, Testing, & Validation
6. System verification and validation
7. Operation & Maintenance

This model overcomes the limitation of the traditional Waterfall methodology by providing feedback loops to allow for testing and validation of the requirements, architecture, and design. While the development methodology has evolved over the years transiting from the Waterfall to the Agile methodology, the V model remains relevant and beneficial to business and IT professionals involved in the information systems development effort and continues to serve as a conceptual framework for system development and project management.

Parallels can be drawn between data analytics and system development since both use Information and Communications Technology (ICT) to help solve social and business problems. Both start with business requirements and require validation and verification before the outputs, whether in the form of an IT system or an analytics report or a dashboard, can be deployed and utilized in the field. Abbasi, Sarker, and Chiang (2016) compared CRISP-DM process model with systems development lifecycle (SDLC) and acknowledged the commonalities between the two.

2.3.9 Nine Laws of Data Mining

Khabaza (2010), one of the authors of the CRISP-DM model, created the nine laws of data mining from his many years of professional experiences as shown in **Table 7**.

Table 7

The Nine Laws of Data Mining (Khabaza, 2010)

Laws of Data Mining	**Description**
1. *Business Goals Law*	Business objectives are the origin of every data mining solution.
2. *Business Knowledge Law*	Business knowledge is central to every step of the data mining process.
3. *Data Preparation Law*	Data preparation is more than half of every data mining process.
4. *Not Free Lunch Law*	The right model for a given application can only be discovered by experiment or "There is No Free Lunch for the Data Miner"
5. *Watkins' Law*	There are always patterns.
6. *Insight Law*	Data mining amplifies perception in the business domain.
7. *Prediction Law*	Prediction increases information locally by generalization.
8. *Value Law*	The value of data mining results is not determined by the accuracy or stability of predictive models.
9. *Law of Change*	All patterns are subject to change.

These nine laws provide a comprehensive view of many issues arising in the data analytics process in an organizational context. They touch on both business and technology perspective and stress the importance of business goals (law #1), domain knowledge (law #2), Value proposition (law #8), the inherent nature of changes in everything from business needs, technology innovations, and social and market environments and the need for adaptation and agility (law #9).

It is worth noting that the number one law of data mining is the business goals law which states that "Business objectives are the origin of every data mining solution". The author also stated that "Data Mining is a Business Process" and further defined data mining as:

> data mining is concerned with solving business problems and achieving business goals. Data mining is not primarily a technology; it is a process, which has one or more business objectives at its heart. Without a business objective (whether or not this is articulated), there is no data mining. (Khabaza, 2010)

Overall these nine laws of data mining provide comprehensive lenses to evaluate the existing data analytics process models and guide the future effort in enhancing and improving them.

2.4 Summary

The above review of the literature revealed that there are many data analytics process models offered by industry and academia. Three of them stand out with CRISP-DM as the de facto industry standard followed by SEMMA and KDD. Many models were developed as extensions to CRISP-DM as it was adopted to various industries and domains. Efforts have been made to establish a unified model that incorporates and consolidates similar but disparate process models or create improved models that are novel and better than the existing ones. Despite these efforts, newer models fail to gain wide-spread acceptance and CRISP-DM remains the most favorable among the community of practice. The following summarizes the limitations of the currently available process models:

Most of the models are technology-driven and have been developed for the technical professionals with emphasis on models, algorithms, tools, and techniques. The insufficient incorporation of business knowledge and business value is not conducive to solving real-world social and business challenges. The Domain-Driven Data Mining process model touted to be business-driven as the name suggested but was filled with mathematical equations and jargons to be business user-friendly. The Business Analytics Methodology paid special attention to business

goals and objectives and applied theories and best practices from business modelling, systems thinking, and soft systems methodology to ensure thorough understanding and analysis of the complex business problems and situations. However, this heavy front-load of business requirement analysis has the disadvantages of a Waterfall-like software and systems engineering model and lacks the flexibility and adaptability of the proven and current Agile methodology.

One common issue with the existing process models is the lack of consideration for human-factors and user experience. These models target technical professionals and generally don't pay attention to simplicity and style which are the key elements of human-centered design.

The gaps in the existing data analytics process models remain and the need to develop a better one is urgent as data analytics moves to the front and center in the era of big data, IoT, AI, and the most pressing of all, the more complex and less safe world with too many wicked social and business problems to solve.

This chapter also reviewed various theories related to data, information, knowledge, communication, human-centered design, and man-computer symbiosis which provide theoretical foundation for the design of the new model.

CHAPTER 3: METHODOLOGY

3.1 Introduction

As stated in chapter one and further demonstrated in chapter two, there is currently not a data analytics process model that is business-technology balanced, universally-accepted and widely-adopted due to various limitations in the existing process models which are technology-centric and fail to incorporate human factors, knowledge management, and business values. This prohibits the growth of the data analytics practices, the maturity of the professional community, and runs the risk of a higher rate of project failures, lower outcomes, and lower returns on investments.

This study aims to contribute to the solution to this problem through achieving two related and sequential objectives:

Research Objective #1: Design a new and improved data analytics process model based on strong theoretical foundation.

Research Objective #2: Evaluate the utility of the new model using an acceptable evaluation method.

To achieve these two objectives, this study identifies three research questions:

Research Question #1: What are the limitations of the existing data analytics process models?

Research Question #2: What theories should be used as the foundation to design a new and improved process model?

Research Question #3: How well does the new and improved process model apply to the real-world situation?

Answers to the first two questions will help inform the design and development of the proposed process model. Answer to the third question will help improve and evolve the proposed process model.

The philosophical approach of this study is grounded upon the pragmatic worldview which emphasizes the practical application of knowledge (Creswell & Creswell, 2017). Hence, the desired goal of this study is to help data analytics professionals, domain subject matter experts, decision makers, and other stakeholders a gain deeper and richer understanding of the data analytics process by providing them with a conceptual framework and practical guidance and enabling them to achieve higher quality, better outcomes, and stronger impacts for their analytics efforts.

The objective of designing and evaluating a new data analytics process model to improve data analytics practices is well aligned with the frameworks and guidelines of design science research in information systems research (Gregor & Hevner, 2013; Hevner, 2007; Hevner & Chatterjee, 2010; Hevner et al., 2004; March & Smith, 1995; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007; Prat, Comyn-Wattiau, & Akoka, 2014; Wieringa, 2014). This project will adopt the Design Science as the research methodology.

3.2 Design Science Research

Different from traditional natural and behavioral science research which is descriptive, explanatory, and predictive in nature aiming to produce theories to help people gain a deeper understanding of the world in which we live in, design science research is prescriptive in nature aiming to produce tools and practice patterns to help people solve real-world problems. Design science dates to Nobel laureate Herbert Simon. In his seminal book, *the Sciences of the Artificial*, Simon (1996) stated that “the natural sciences are concerned with how things are” (p. 114) and

"Design, on the other hand is concerned with how things ought to be, with devising artifacts to attain goals" (p. 114). While natural sciences and behavioral sciences aim to describe, explain, and predict the natural and social phenomena, design science seeks to prescribe formulas and develop tools to enable people to take proactive actions in affecting these phenomena. Design science has been widely used in medicine, architecture, engineering, arts, and management.

Hevner et al. (2004) pointed out that design science research and traditional behavioral science research are not mutually exclusive. In fact, they complement each other. Each of the three cycles rely on or utilize empirical methods. Behavioral science not only provides prior knowledge required to inform the design of an artifact but also provides empirical methods such as case studies, expert reviews, and usability surveys to validate the utility of an artifact. In short, behavioral science provides the rigor of empirical methods to complement the creativity of design science. In return, design science takes the behavioral science to the next level where maximum benefits and impacts of knowledge can be achieved by enhancing human capability with tools and methods to solve social problems.

3.3 Design Science in Information Systems Research

Design science has been applied to information systems research over three decades. Hevner et al. (2004) were the first to propose an information systems research framework for conducting IS research combining behavioral science and design science paradigms as depicted in **Figure 13**. This framework lays out the key aspects of design science in information systems research and stresses the need to draw relevance from the problem situation and rigor from the existing body of knowledge. It prescribes the two-step process of design and evaluation resulting in the contribution of design artifacts to problem solution and the addition of new knowledge to the existing body of knowledge.

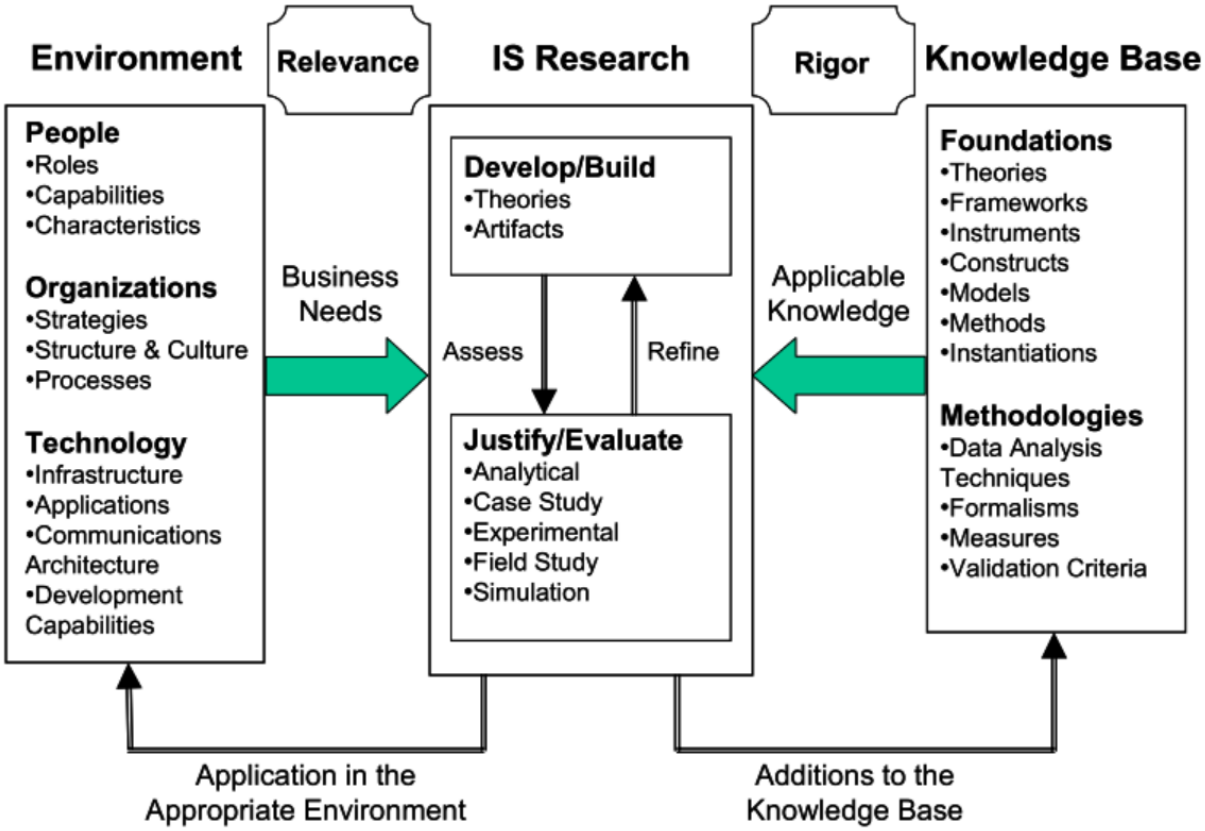


Figure 13. Information systems research framework (Hevner et al., 2004)

This seminal paper also provides seven guidelines for conducting information systems research using design science paradigm as shown in **Table 8**.

Table 8

Design Science Research Guidelines (Hevner et al., 2004)

Guideline	**Description**
1. Design as an artifact	Design-science research must produce a viable artifact in the form of a construct, a model, a method, or an instantiation.
2. Problem relevance	The objective of design-science research is to develop technology-based solutions to important and relevant business problems.
3. Design evaluation:	The utility, quality, and efficacy of a design artifact must be rigorously demonstrated via well-executed evaluation methods.
4. Research contributions	Effective design-science research must provide clear and verifiable contributions in the areas of the design artifact, design foundations, and/or design methodologies.
5. Research rigor	Design-science research relies upon the application of rigorous methods in both the construction and evaluation of the design artifact.
6. Design as a search process	The search for an effective artifact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
7. Communication of research	Design-science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

Hevner (2007) further refined the framework and rebranded it as three cycles of design science research as depicted in **Figure 14**.

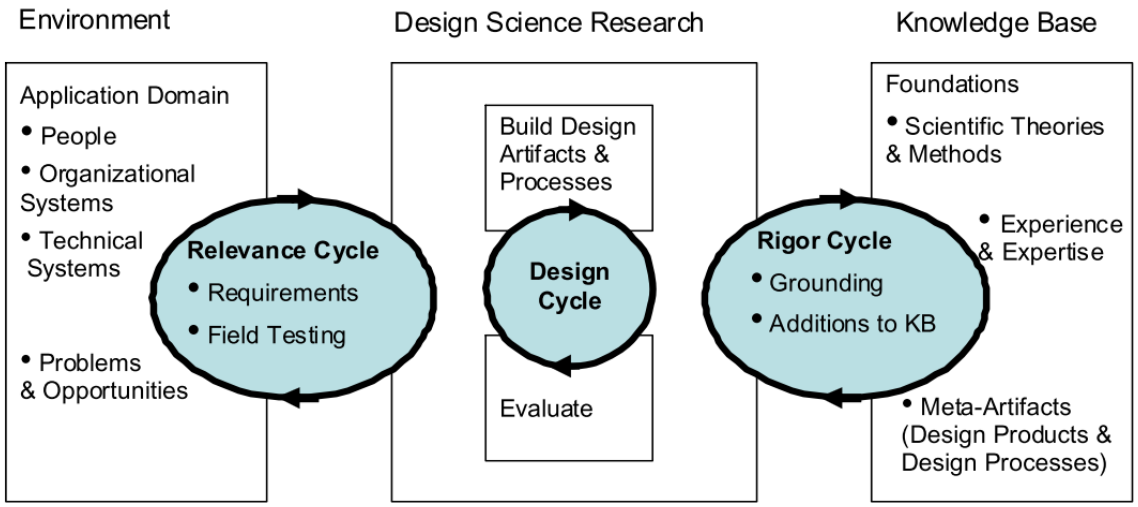


Figure 14. Three cycles of design science research (Hevner, 2007)

Design cycle is at the center of the process model and includes the design and evaluation of artifacts. The relevance cycle provides organizational context and business requirements as inputs to the design cycle and accepts the validated artifacts produced by the design cycle to help problem solving in the organizational setting. Rigor cycle provides existing knowledge to inform the design of artifacts and receives knowledge produced from the design cycle including the knowledge embedded in the artifacts and the knowledge obtained from the evaluation about the utility, quality, and efficacy of the design artifacts.

Peffers et al. (2007) proposed a similar but more elaborate six-step design science research process to further articulate and guide the process of conducting design science research in information systems as depicted **Figure 15**. Despite some minor differences Peffers six-step process model and Hevner three-cycle process model are well aligned.

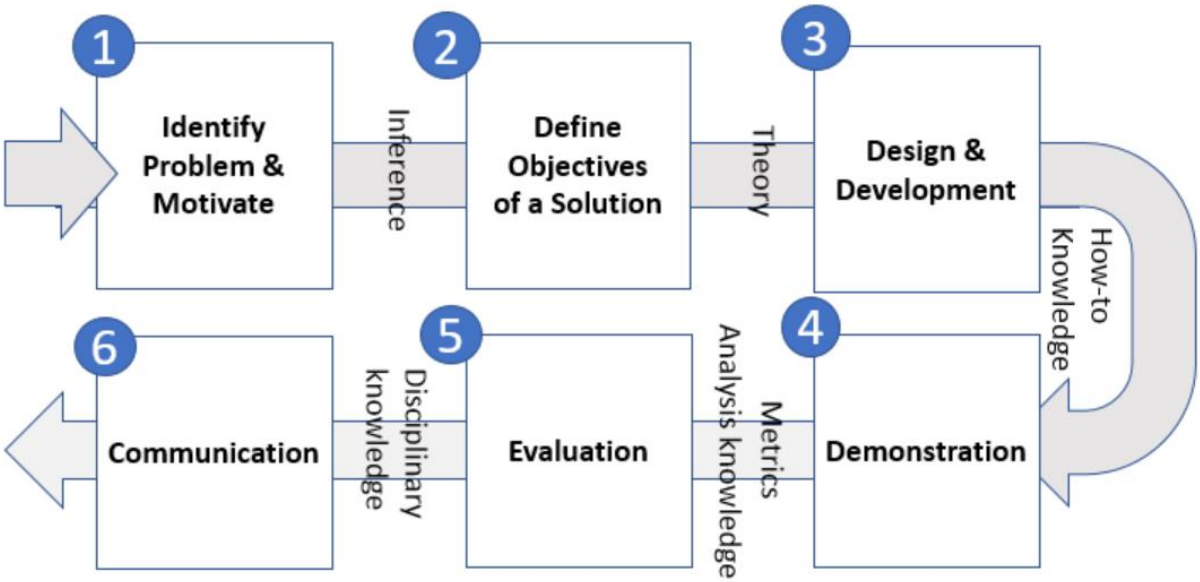


Figure 15. DSRM process model adapted from Peffers et al. (2007)

3.4 Evaluation of Design Artifacts

In design science research, designing and building an artifact is only half of the story. The evaluation of an artifact is equally important and ensures that the artifact is not only well designed with relevance to the problem situations and utilizing extant theories but also is efficient and effective for people to use in their problem-solving endeavor. To ensure the utility, quality, usability, and efficacy, an artifact must be evaluated through various methods, some of which are more rigorous than others, and in an iterative cycle.

Hevner et al. (2004) pointed out that "because design is inherently an iterative and incremental activity, the evaluation phase provides essential feedback to the construction phase as to the quality of the design process and the design product under development" (p. 85). This agile nature of design and the need for continuous feedback suggests that the evaluation process needs to be aligned and flow with the design process. Evaluation methods chosen should match the type and maturity of the design artifact and the stage of the design process. There is no one method that fits all and there is no one evaluation that does it all. If necessary, a design artifact

should be evaluated using multiple methods with multiple cycles to ensure its utility, quality, usability, and efficacy before it is finally deployed to the field. **Table 9** list literature related to design science artifacts evaluation methods.

Table 9

List of Literature on the Evaluation Methods of Design Science Research

Literature	**Summary**
Hevner et al. (2004)	Summarized the evaluation methods in five categories: 1. Observational; 2. Analytical; 3. Experimental; 4. Testing; 5. Descriptive
Peffers et al. (2012)	Established seven types of evaluation methods: 1. Logical Argument; 2. Expert Evaluation; 3. Technical Experiment; 4. Subject-based Experiment; 5. Action Research; 6. Prototype; 7. Case Study; 8. Illustrative Scenario
Venable, Pries-Heje, and Baskerville (2012)	Established a two-dimensional model where an evaluation method can be placed along the axis of artificial/natural and axis of ex ante/ex post.
Venable, Pries-Heje, and Baskerville (2016)	Revised the two-dimensional model by replacing ex ante/ex post with formative/summative. Most significantly, it introduced the concept of evaluation episodes, evaluation strategy, and a four-step evaluation process.
Wieringa (2014)	Presented a two-dimensional model where an evaluation method can be placed along the axis of robust mechanisms (idealized conditions, realistic conditions, and conditions of practice), and axis of scalable regularities (single case, sample, and population)
Prat, Comyn-Wattiau, and Akoka (2015)	Identified seven typical evaluation patterns: 1. demonstration; 2. simulation- and metric-based benchmarking of artifacts; 3. practice-based evaluation of effectiveness; 4. simulation- and metric-based absolute evaluation of artifacts; 5. practice-based evaluation of usefulness or ease of use; 6. laboratory, student-based evaluation of usefulness; 7. algorithmic complexity analysis.

Hevner et al. (2004) defined the five evaluation method categories along the spectrum of the artificial and natural divide. The observational category represents evaluation of artifacts in a

naturalistic setting and includes case study and field study. The last four categories can be viewed as non-observational, non-naturalistic or artificial.

Peffers et al. (2012) described seven types of artifact evaluation shown in **Table 10** which can be roughly separated into the artificial vs natural bucket with some ambiguity. While action study and case study are undoubtedly considered natural, the rest can be considered artificial except for illustrative scenarios. The authors defined illustrative scenario as "application of an artifact to a synthetic or real-world situation aimed at illustrating suitability or utility of the artifact" (p. 5). In the case of a synthetic situation, it would be considered artificial; In the case of real-world situation, it would be considered natural.

Table 10

The Design Science Research Evaluation Method Types by Peffers et al. (2012)

Evaluation Type	**Description**
Logical Argument	An argument with face validity.
Expert Evaluation	Assessment of an artifact by one or more experts (e.g., Delphi study).
Technical Experiment	A performance evaluation of an algorithm implementation using real-world data, synthetic data, or no data, designed to evaluate the technical performance, rather than its performance in relation to the real world.
Subject-based Experiment	A test involving subjects to evaluate whether an assertion is true.
Action Research	Use of an artifact in a real-world situation as part of a research intervention, evaluating its effect on the real-world situation.
Prototype	Implementation of an artifact aimed at demonstrating the utility or suitability of the artifact.
Case Study	Application of an artifact to a real-world situation, evaluating its effect on the real-world situation.
Illustrative Scenario	Application of an artifact to a synthetic or real-world situation aimed at illustrating suitability or utility of the artifact.

In addition to the artificial/natural dimension, Venable et al. (2012) added the concept of ex ante and ex post evaluation where ex ante represents the time before the design of an artifact is complete and ex post represents the time after the design of an artifact is completed. In a follow-on research, Venable et al. (2016) replaced the ex ante/ex post concept with the similar concept of formative/summative which better represents the iterative and incremental nature of the design and evaluation process. The formative evaluation tends to be process-oriented and is iterative and incremental in nature. A design artifact goes through multiple stages of design and evaluation until it becomes mature enough for the summative evaluation. Most significantly, the authors also introduced the concept of evaluation episodes, evaluation strategy, and a four-step evaluation process. A design artifact can undergo multiple evaluations during its design and evaluation lifecycle. Each evaluation is called an episode and a specific evaluation method that fits best can be utilized. A evaluation strategy is the sequence of multiple episodes with increasing rigor from being formative to being sumative and from being artificial to being natural.

3.5 Research Design of this Study

This study follows the three-cycle process by Hevner et al. (2004) and the six-step process by Peffers et al. (2007). These two approaches are compatible and complementary; together they provide a comprehensive framework and guidance for this research project.

Figure 16 shows the research plan of this study following Peffers six-step process (Peffers et al., 2007).

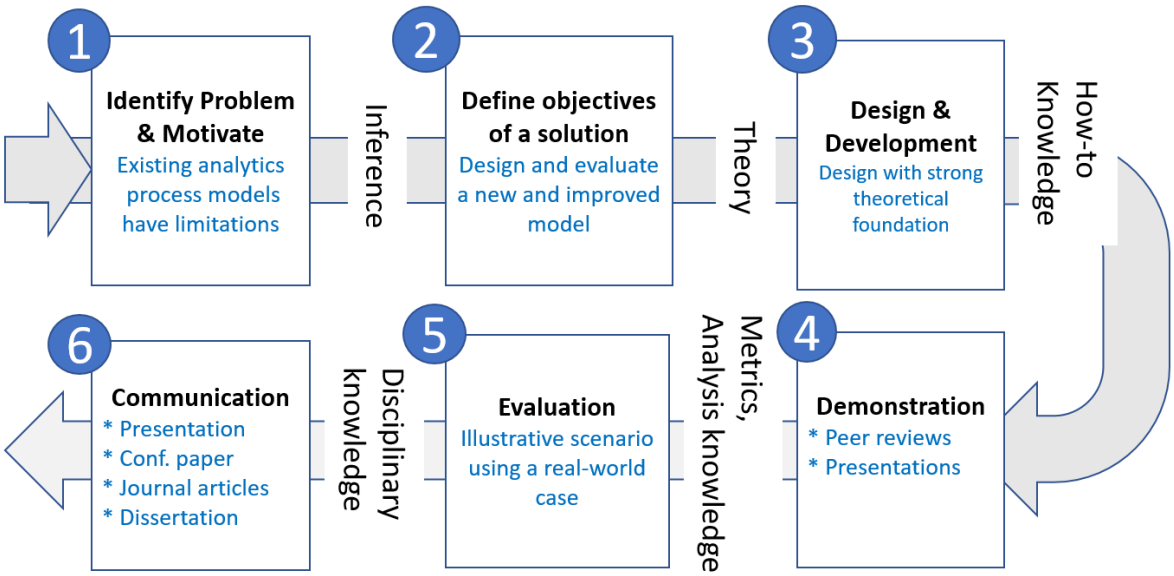


Figure 16. Research plan following Peffers six-step process

Step one (Identify Problem) and two (Define Objectives) have already been covered in Chapter one (Introduction) and Chapter two (Literature Review). Chapter one identifies the problem of insufficient adoption of data analytics process models and its impact on the effectiveness of data analytics effort. This prompts the need for improvement and leads to the objective of designing a new process model. Chapter two surveys the literature to gain a broader and deeper understanding of the strength and limitations of the existing process models. Step three (Design & Develop) is in part covered in Chapter two where existing theories in information systems, communications, and knowledge management are reviewed to provide theoretical foundation for the new process model. Chapter four (The Model) will describe in full detail the new process model. The new process model was first presented to the fellow doctoral students at Robert Morris University during a data analytics class and later presented to the colleagues of the researcher at the MITRE Corporation. Feedback was incorporated to improve the initial design. This covers step four (Demonstration). Step five (Evaluation) will be covered in Chapter five (Evaluation of the Model). Step six (Communication) will be forthcoming as the

researcher continues the journey with additional presentations and publications including this dissertation.

Figure 17 shows the research framework of this dissertation which ties research objectives, research questions, and research methods together in a holistic view.

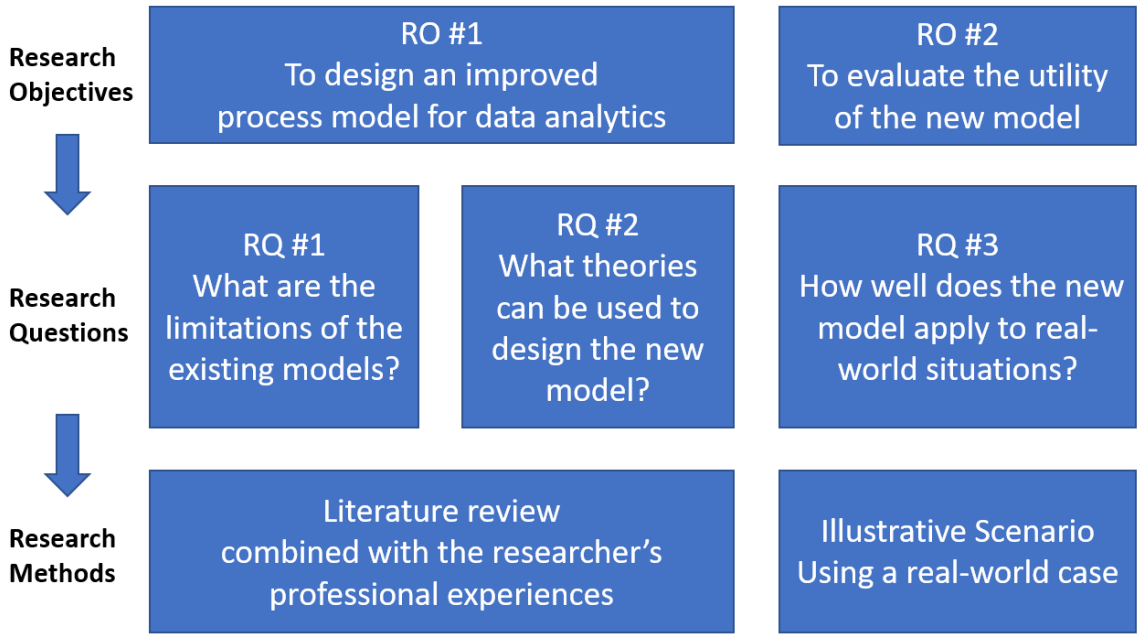


Figure 17. Research framework of this study

3.6 Evaluation Using Illustrative Scenario

Peffers et al. (2012) defined an illustrative scenario as the "application of an artifact to a synthetic or real-world situation aimed at illustrating suitability or utility of the artifact." The purpose of this evaluation is to illustrate the utility of the model by applying it in a real-world data analytics effort. While an illustrative scenario is typically a descriptive effort, the researcher went beyond and initiated a real-world data analytics project as a data scientist in collaboration with healthcare quality and policy experts. The evaluation performed in the realistic conditions with observational characteristics greatly improved the rigor of the illustrative scenario method.

Figure 18 shows where the illustrative scenario situates in the framework of scaling up to stable regularities and robust mechanisms (Wieringa, 2014).

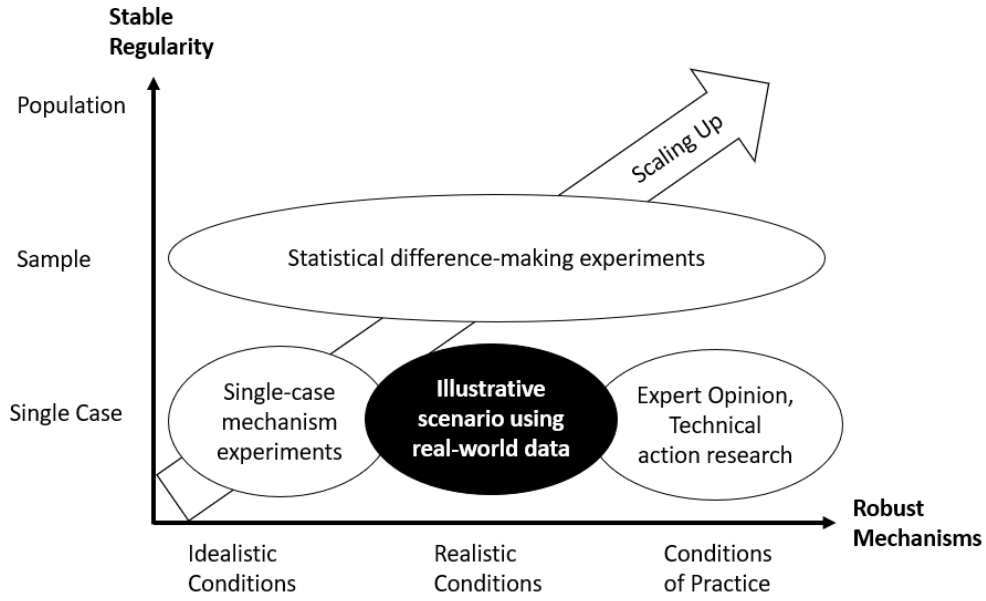


Figure 18. The illustrative scenario in the scalingup framework

For this study, a real-world problem will be used to illustrate how the A2E Model can be applied to a real-world situation. This scenario deals with the healthcare quality performance of U.S. dialysis facilities providing dialysis care for the Centers for Medicare and Medicaid (CMS) beneficiaries with End-Stage-Renal-Disease (ESRD). This illustrative scenario used both descriptive and predictive analytics to identify characteristics of dialysis facilities that may influence or contribute to the patients' hospital readmission. It also aims to explore ways of intervening to help reduce the hospital readmissions which put great financial burden on CMS and quality of life burden on the patients. Publicly available secondary data from several sources on dialysis care quality measures, dialysis facility organizational and geographical characteristics, and facility-level patient characteristics will be used as data inputs. For this study,

no individual patient personal and health data were used so there is no risk of information security and data privacy.

This illustrative scenario followed the five steps specified in the A2E Model and documented the effort and outcome of each steps. The goal is to illustrate how the A2E Model can be applied in a data analytics effort to address multiple perspectives of technology, people, and organization. This illustrative scenario also demonstrated the iterative and collaborative nature of the analytics process through continuous feedback loops and symbiosis between humans and machines.

There are limitations with this evaluation method. First, it only covers a single case in a single domain and its conclusions cannot be generalized to other cases and other domains. Future evaluations should include multiple cases in multiple domains for increased generalization. Secondly, this evaluation is illustrative in nature. Even though the data are from a real-world situation, the analytics effort only represents the application of the A2E Model under realistic condition and not the condition of practice as described by Wieringa (2014). Lastly, this illustrative scenario is only the first of the many more steps or iterations that are required to thoroughly validate an artifact before it is deployed to the field.

Additional evaluation methods and episodes should be considered to further validate and improve the A2E Model as potential future research projects. For example, future evaluation would select more sample cases from different industries and domains and adopt expert reviews, case study, or action research as the evaluation methods.

CHAPTER 4: THE A2E MODEL

4.1 Introduction

This chapter describes in detail the new process model for data analytics. It is worth to mention up front several prominent features of this model.

First, it is designed with strong theoretical foundation including information and communication theories as described in chapter two.

Secondly, it is value-driven as opposed to technology-driven. This model starts with business needs arising from social and organizational context and ends with ideas and recommendations which inform decision and actions to solve the problem or improve the situation.

Thirdly, it is integrative. It integrates multiple perspectives including technology, people, and organization; It integrates and aligns business and technology; It integrates and aligns human intelligence and artificial intelligence; It integrates multiple disciplines including knowledge management, organizational learning, human-centered design, and human-computer symbiosis.

Fourthly, it follows the Agile principles. Data analytics is an exploratory and learning process. It takes multiple iterations over time to gain insights from data. It also takes multiple feedback loops in the applications of the insights in decision making and problem solving.

Fifthly, the model is universal and generic and is technology and industry agnostic. It can be used by one person on a small effort, or many people on a large project. It is especially powerful and beneficial when a team of multidisciplinary professionals collaborate in a large data analytics project. For example, in a healthcare data analytics project, the team may include data engineers, data scientists, business analysts, physicians, nurses, and healthcare quality and policy experts. This model provides the basis for establishing a shared mental model for a team

and helps facilitate and enable effective communications, coordination, cooperation, and collaboration.

Finally, it is designed with simplicity and elegance in mind for usability and user experience. The model uses the A, B, C, D, E mnemonics to help people easily remember the five essential steps as shown in **Figure 19**. The model does not inundate the practitioners with complicated steps, tasks, and activities that are incoherent and hard to follow and remember in practice. For this reason, we conveniently name this innovative model the A2E Process Model for Data Analytics, or the *A2E Model* for short.

4.2 Five Themes

The A2E Model can be summarized with five themes as shown in **Figure 19** to articulate the various aspects of data analytics process. Through these five themes, the practitioners and stakeholders can gain a deep and holistic understanding of the data analytics process in an organizational context. The remaining sections provide details on each of these themes.

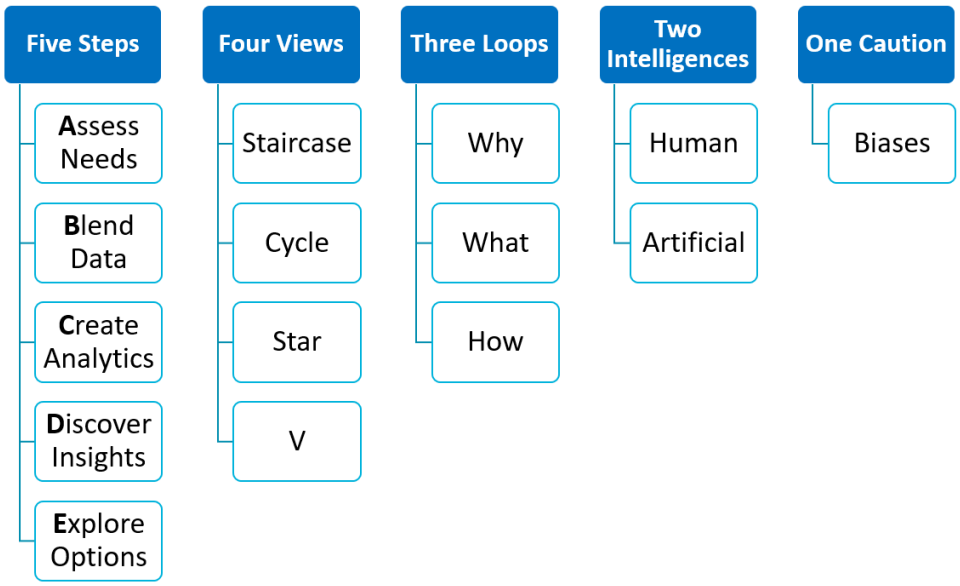


Figure 19. The five themes of the A2E data analytics process model

4.3 Theme One - Five Steps

The core concept of the A2E Model is the five-step process that represents the major stages or phases in a data analytics process:

1) **A**ssess Needs
2) **B**lend Data
3) **C**reate Analytics
4) **D**iscover Insights
5) **E**xplore Ideas

Figure 20 show the five steps and the exemplary activities that may be performed during each step. These activities are not meant to be exact and strictly followed but rather serve as examples.

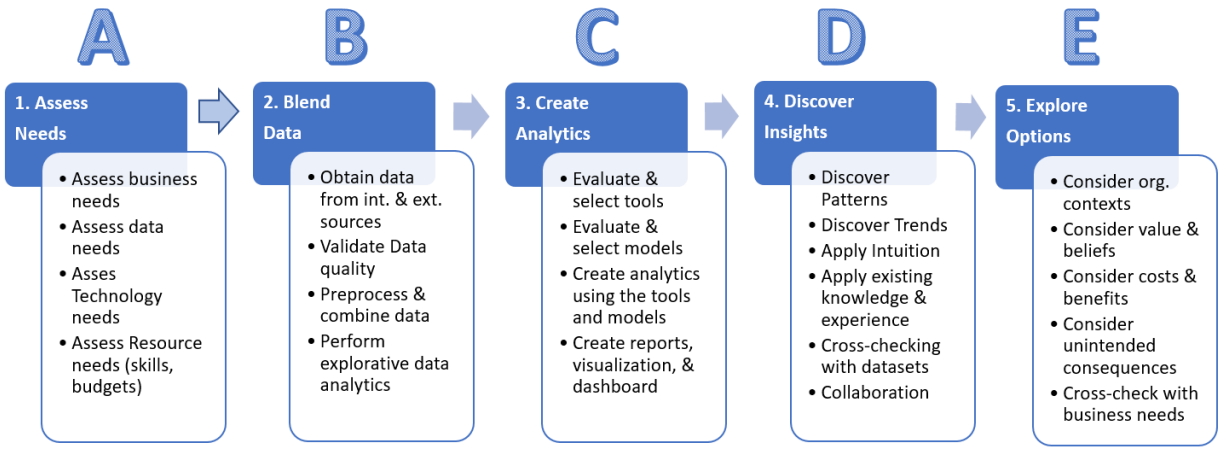


Figure 20. The five steps

Table 11 provides a general description for each step and how it maps to the Wisdom Hierarchy and the nature of the activities performed within it.

Table 11

The Five Steps of the A2E Model

Step	**Description**	**Wisdom Hierarchy**	**Nature of activities**
Assess Needs	Access business needs, data needs, technology needs, and resource needs.	All	All
Blend Data	data from various sources are acquired, cleansed, filtered, and merged to form coherent datasets. Data quality is inspected and ensured. Data privacy is evaluated and ethical concerns if any are addressed. Data schema, semantic definitions are documented in data dictionary. Exploratory and descriptive data analysis is performed to help gain understanding of the data.	Data	Transactional
Create Analytics	Use data visualization tools, statistical models, and machine learning algorithms to create analytics to support knowledge discovery.	Information	Analytical
Discover Insights	Use the analytics to discover insights and create knowledge that can be used to inform decision making. Business expertise and existing body of knowledge are utilized to interpret and validate the analytics.	Knowledge	Intellectual
Explore Ideas	Engage with all stakeholders and subject matter experts to share the insights and knowledge, leverage their experiences and expertise, accommodate their needs and interests, collaborate on solutions and actions, explore potential interventions, and execute on the plan to solve the business problem or improve social conditions.	Wisdom	Transformational

The A, B, C, D, E mnemonics are easy to remember. The name of each step is phrased using the construct of Verb + Object for increased impacts.

The five steps may appear to be sequential, however, they are iterative as well. This will be explained in later sections.

4.4 Theme Two - Four Views

There are four ways to look at the data analytics process.

4.4.1 The Staircase View

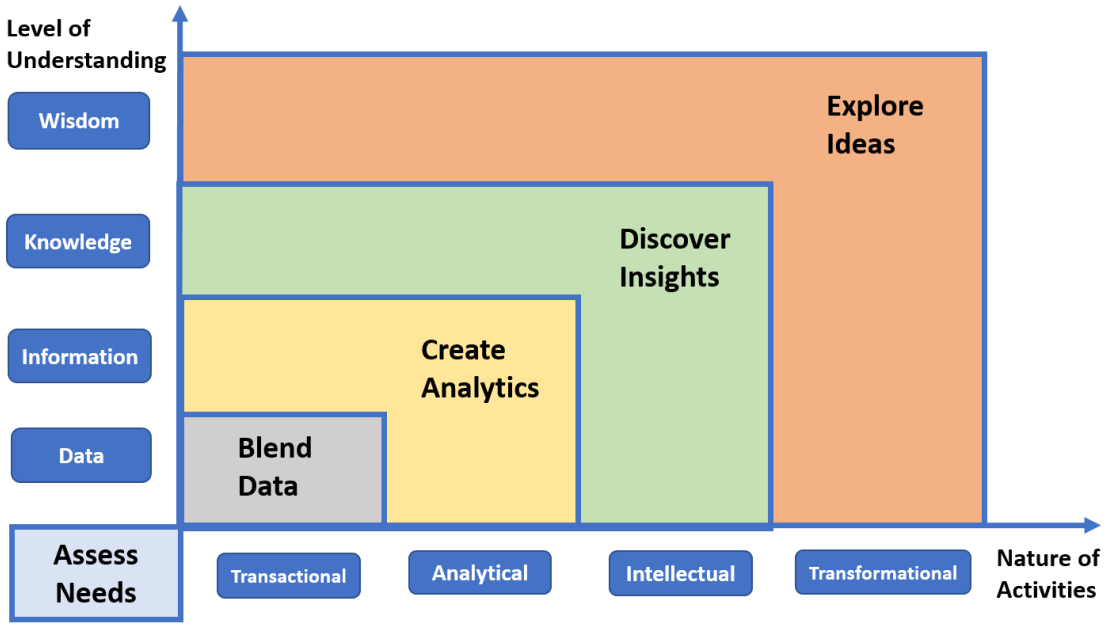


Figure 21. The staircase view

The staircase view as shown in **Figure 21** represents the essence of the A2E Model. This view follows two lines of reasoning. On the vertical line, the y-axis represents the wisdom hierarchy as mentioned in chapter two. Moving from data to information, from information to knowledge, and from knowledge to wisdom represent the increasing level of understanding of the problem or situation at hand.

On the horizontal line, the x-axis represents the nature of the activities as we move from transactional to analytical, from analytical to intellectual, and from intellectual to transformational.

The first step "Assess Needs" is outside of the normal range since it is the front and center of the process and must incorporate all aspects and elements from data to wisdom and from transactional to transformative. This will become more evident when we look at the star view.

4.4.2 The Cycle View

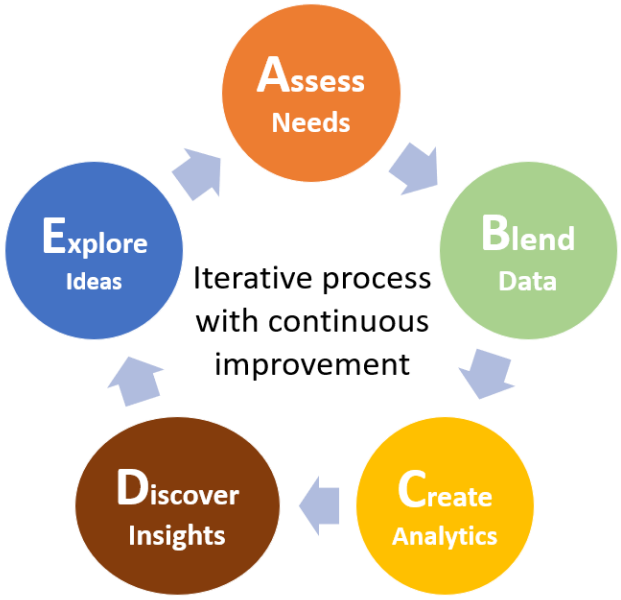


Figure 22. The cycle view

The purpose of the staircase view as shown in **Figure 21** is to provide a simple and elegant structure and step-by-step process guidance for planning and performing data analytics activities. The downside of this simplification is that it makes it appear to be sequential and waterfall-like. However, data analytics is an iterative and incremental process which requires constant and continuous feedbacks and validations within every step and between all steps. It should follow the Agile principles currently prevalent in the system development. The needs and requirements may not be crystal clear at the outset and will become more apparent and granular along the

various steps. The cycle view in **Figure 22** helps to alleviate this downside and instill agility into the process.

The cycle view follows the Deming Plan-Do-Check-Act (PDCA) cycle and stresses the importance of iterative and continuous improvements that are necessary for a successful data analytics project.

4.4.3 The Star View

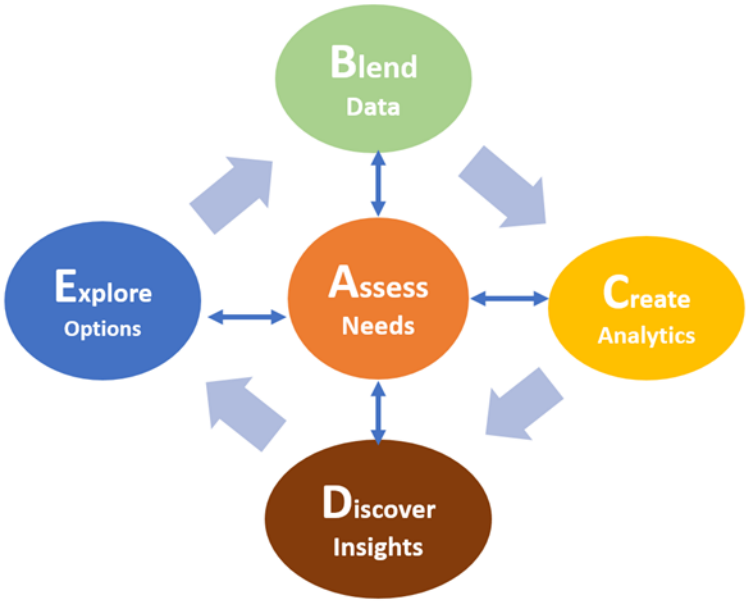


Figure 23. The star view

While each of the five steps is equally important, the first step "Assess Needs" has a special place in the process. We all know that technology is only a means to an end, it is not an end by and of itself. Technology must serve human needs. As was mentioned in chapter two, the mission of data analytics is to support evidence-based decision making to solve social problems and improve human conditions.

The star view as shown in **Figure 23** places the first step "Assess Needs" at the center as the central hub. At every step along the way, activities must be constantly and continuously validated against the needs. This is particularly critical when dealing with complex, ill-defined,

and wicked problems and situations. The business needs may not be clear upfront, and the data needed to support the business needs may not be well-defined or readily available early on.

To summarize, the model is value-driven as opposed to technology-driven and is human-centric as opposed to technology-centric.

4.4.4 The V View

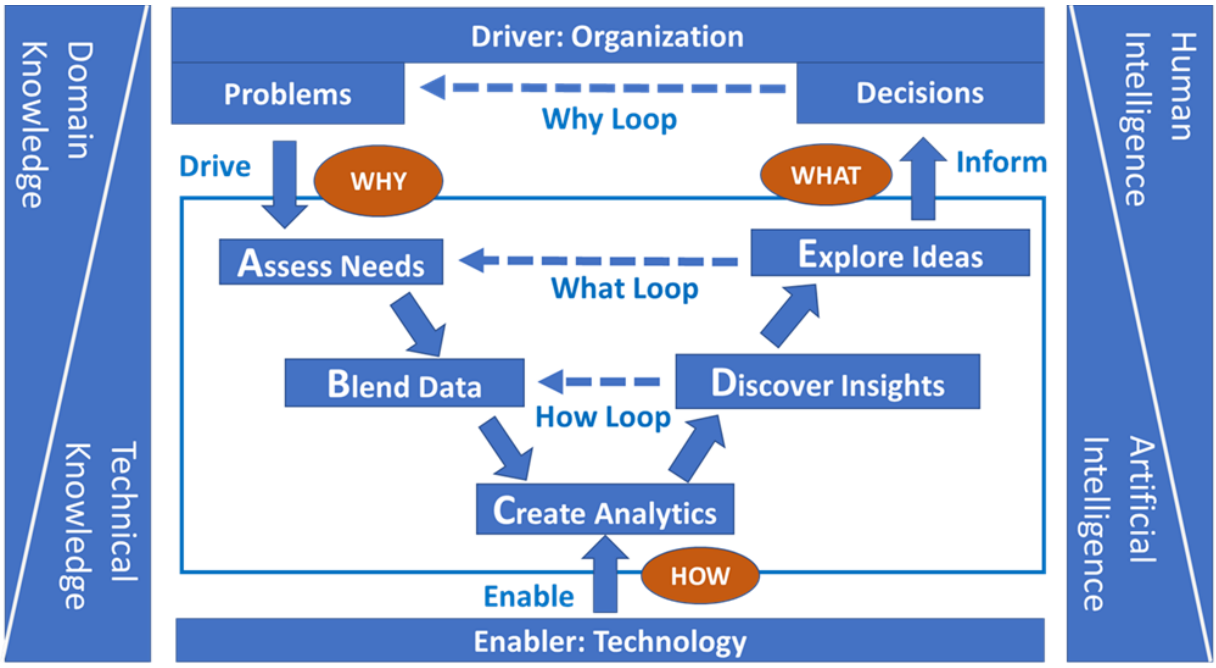


Figure 24. The V view

The V view as shown in **Figure 24** is more sophisticated and systemic compared to the previous three views. It encapsulates the fusion and synergy of multiple perspectives and multiple disciplines and hence deserves more elaboration and attention.

At a high level, data analytics share similarity with systems and software engineering in that they both use Information and Communications Technology (ICT) to process data and information and support decision making to solve social and business problems. The V view leverages the concepts of the V model from the systems and software engineering as described in chapter two. The five steps in the model fit well with V model as shown in **Table 12**:

Table 12

Mapping Between Steps in the Model and Phases in the V Model

Step in the ABCDE Model	Phase in the V Model
***A**ssess Needs*	Concepts of Operations Requirements & Architecture
***B**lend Data*	Detailed Design
***C**reate Analytics*	Implementation
***D**iscover Insights*	Integration, Test, and Verification
***E**xplore Ideas*	System Verification and Validation Operation and Maintenance

The V view places data analytics at the heart of the organizational context. The advances in machine learning and artificial intelligence have provided data analytics with better computational algorithms and software tools but technology is only the enabler not the driver. More effort should be focused and spent on understanding the mission, the value, the goals, the objectives, the problems, the needs, and the requirements. This requires the fusion and balance of domain knowledge and technical knowledge, human intelligence and artificial intelligence. Data analytics is both an art and a science.

In summary, data analytics must begin with asking the "Why" question before moving on to the "How" and should produce the "What" that informs decision making and support problem solving.

4.5 Theme Three - Three Loops

Previous section touched on the "Why", "How" and "What" of data analytics. In his 2011 book, *Start with Why: How Great Leaders Inspire Everyone to Take Action,* author Simon Sinek created the Golden Circle of Why, How, and What (Sinek, 2011). It provides a simple and yet

powerful compass to guide any human endeavor including the data analytics process as shown in **Figure 25**.

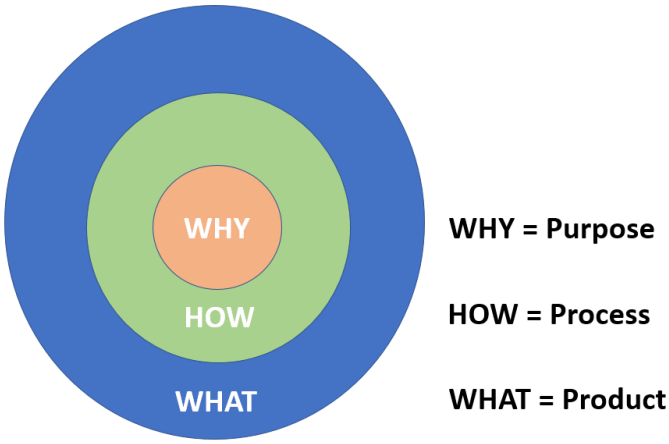


Figure 25. The golden circle (Sinek, 2011)

The V view mentioned in the previous section contains three loops that map to the golden circle as seen in **Figure 26**. The three loops represent three dimensions of data analytics and are good indicators of data analytics maturity level.

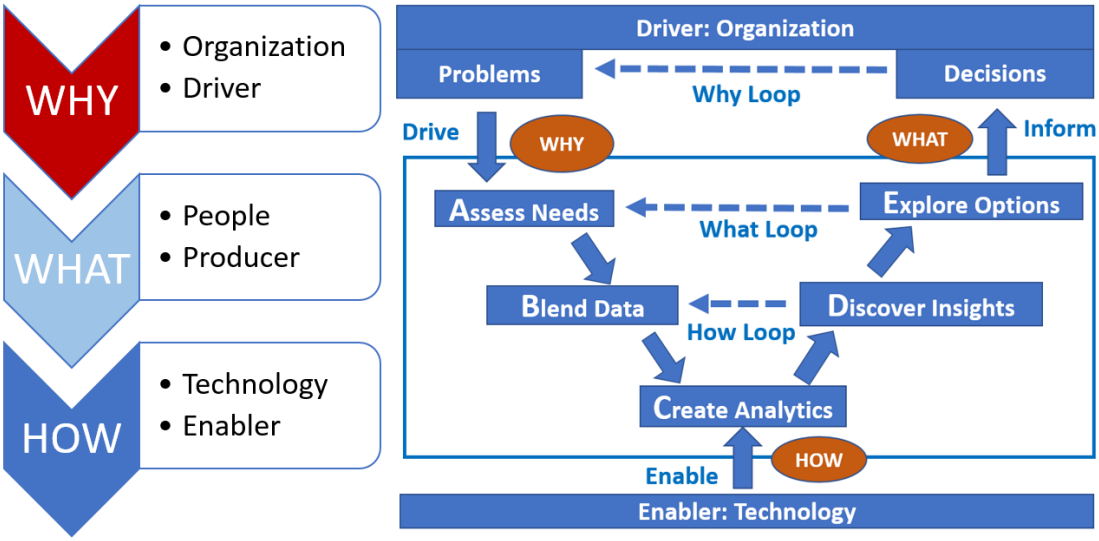


Figure 26. The three loops

4.5.1 The How Loop

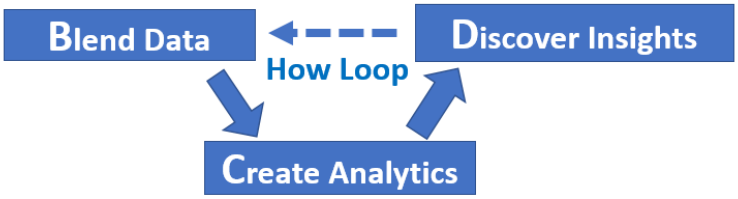


Figure 27. The How loop

The How loop as shown in **Figure 27** is at the bottom of the V view. It travels from step B (build datasets) to step C (create analytics) to step D (discover insights) and back to step B for feedback and validation. This loop focuses more on the application of technology and less emphasis is placed on understanding of the problem situation and incorporation of domain knowledge. It focuses on generating information and patterns and not on providing insights and solutions.

The How loop focuses on technology dimension where technology is an enabler for the analytics. However, technology is only a piece in a larger pie, it is only a mean to an end, not an end by and of itself. This loop represents a low maturity level of data analytics when an organization stops at this level and fails to move up to the What loop and the Why loop.

4.5.2 The What Loop

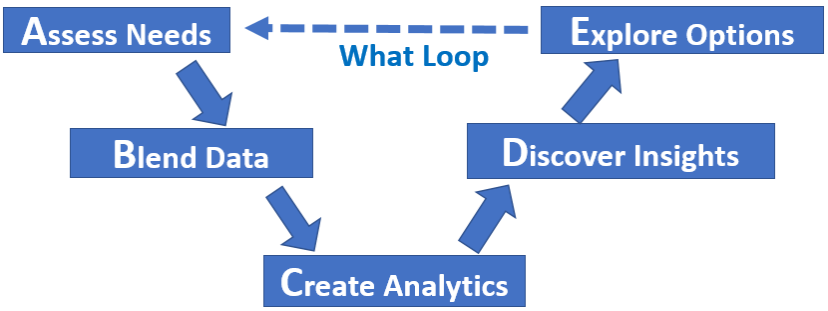


Figure 28. The What loop

The What loop as shown in **Figure 28** travers from A->B->C->D->E and back to A for feedback and validation. This is the typical scope of data analytics. It starts by assessing the business needs and end with the options that are evaluated and validated against the needs.

The What loop focuses on the people dimension where people are the key performer and producer responsible for the recommended decisions based on knowledge discovered. There is no silver bullet to solve human problem other than for human to engage in a collaborative, communal process. Even with the aid of technology and automation of manual tasks, people must make the ultimate judgement calls and make the final decisions. The Why loop focuses on generating insights and creating knowledge and represents a medium level of maturity. Organizations performing at this loop should strive to move up to the higher level by performing at the Why loop.

Davenport, Harris, and Morison (2010) pointed out that "moving from purely information-oriented questions to those involving insights is likely to give you a much better understanding of the dynamics of your business operations."

4.5.3 The Why Loop

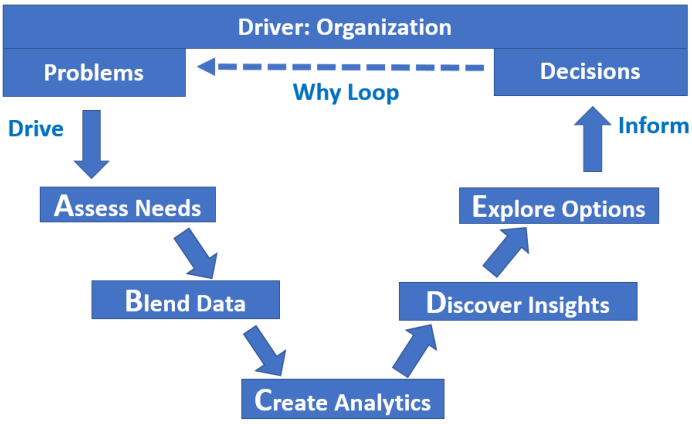


Figure 29. The Why loop

The Why loop shown in **Figure 29** starts from the problem that drives the whole data analytics process (A->B->C->D-E) which produces the knowledge, insights, and recommendations to inform decision making. The decision leads to actions that aim to solve the problem. This loop aligns the data analytics process with the organizational decision-making and problem-solving process.

Technology is only a means to an end and not an end in and of itself. Human creates technology to sever a purpose that is to extend human cognitive ability and to solve social problems and improve human conditions. As eloquently stated by the renowned futurist Gerd Leonhard in his book *Technology vs. Humanity: the coming clash between man and machine*, "in the end, technology is not what we seek, but how we seek" (Leonhard, 2016).

The Why loop represents the social, cultural, and organizational dimension where social needs, organizational missions, and business strategies drive the analytics activities and ultimately determine the success or failure of the analytics project. Organizations that regularly perform at this level of maturity have much better chance of success.

4.6 Theme Four - Two Types of Intelligence

Data analytics is a collaborative effort involving multiple disciplines from science, technology, engineering, and math (STEM) to management, sociology, psychology, and art. It encompasses multiple perspectives including technical, people and organizational perspective.

There are two intelligences involved in data analytics effort. One is human intelligence; the other is artificial intelligence. The former is the driver and the latter is the enabler as depicted in the V view of the model.

In many ways, human intelligence and artificial intelligence are complementary. On one hand, AI powered by computer processors, storages, and algorithms has the capability to process

large volume of data much faster than human beings. On the other hand, with more complex situations where social and cultural context and value judgement are required to make decisions, AI in its current stage will not be able to replace human beings. The human-machine symbiosis is about taking advantage of the best of both worlds.

The human-machine symbiosis is not the only type of symbiosis. There is also a need for machine-machine symbiosis where multiple algorithms and multiple tools are utilized to corroborate, cooperate, and complement each other to minimize machine biases and maximize the accuracy. This can be called machine-machine symbiosis.

Collaboration between diverse individuals and groups will help reduce cognitive bias and bring out innovative ideas. This can be called human-human symbiosis or simply teamwork. **Figure 30** shows the three types of symbiosis in a four-quadrant.

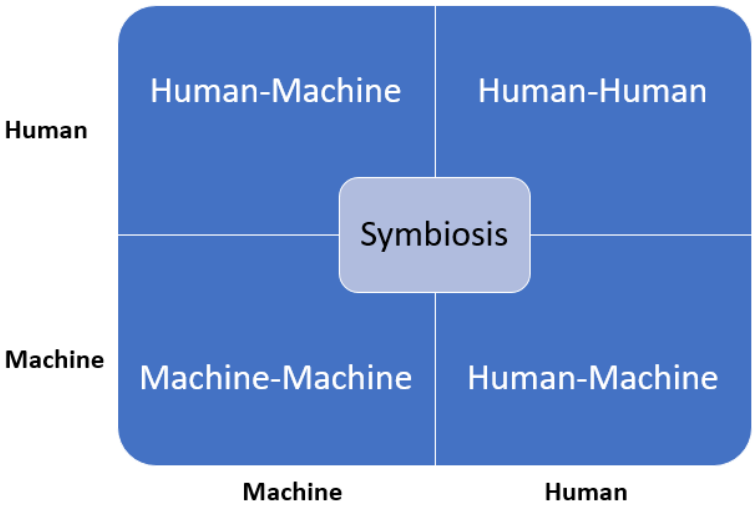


Figure 30. The three types of symbiosis

To extend this concept, we also recognize the symbiosis of the business professionals and technology professionals as well as the symbiosis of the domain knowledge and technical knowledge.

Striking the right balance between human intelligence and artificial intelligence and between business and technology is a critical success factor for data analytics and is the subtle message the A2E Model tries to deliver.

4.7 Theme Five - One Caution

Biases exist in data, algorithms, and human minds. Data analytics is inherently biased since it relies on data as inputs, algorithms as the enabler, and humans as the driver. Great care should be given to the accuracy, applicability, limitations, and consequences (intended or unintended, desired or undesired) of data analytics.

The reality is complex, and the true reality is unknowable as famously stated by Laozi in the 2500-year old Taoist text *Tao Te Ching*, "The tao that can be told is not the eternal Tao. The name that can be named is not the eternal Name. The unnamable is the eternally real."(Laozi, Mitchell, Roig, & Little, 1989). Echoing Laozi, statistician George Box (1976) was famously quoted "all models are wrong" because models are only approximations of the true reality. Statistical and computational models that underlie data analytics are no exceptions. They come naturally with their designers' cognitive bias (intentional, or unintentional).

It is also worth to point out that data, whether they are collected through human observations or sensory devices, are also approximate measures of the actual properties of the observed entities. Even correct or accurate data collected cannot escape from the bias of those who define the measurement, who measure and collect data or design data collection instruments. In addition, incomplete, inaccurate, and missing data are typical and can distort the outcome of the analytics - "garbage in garbage out".

King, Keohane, and Verba (1994) defined scientific research as "an ideal to which any actual quantitative or qualitative research, even the most careful, is only an approximation". The

authors further presented four characteristics of scientific research, one of which is "The conclusions are uncertain". They stated that:

> *"By definition, inference is an imperfect process. Its goal is to use quantitative or qualitative data to learn about the world that produced them. Reaching perfectly certain conclusions from uncertain data is obviously impossible. Indeed, uncertainty is a central aspect of all research and all knowledge about the world. Without a reasonable estimate of uncertainty, a description of the real world or an inference about a causal effect in the real world is uninterpretable" (King et al., 1994).*

We should also recognize that both men and machines have limitations. Even though the partnership between the two helps overcome the limitations to some extent, it will not eliminate the limitations. This leads to the realization that data analytics empowered by human intelligence and enabled by artificial intelligence are not panacea to all problems. Not all complex problems have clear-cut solutions. One quality that is important for all professional, business or technical, is humility. To be humble is to know:

- the complexity, uncertainty, and unknowability of the reality
- the limitation of human mind and machine power
- what to believe and what to doubt
- when to act and when not to act
- the why before performing the how and producing the what
- to anticipate the unintended consequences of actions
- to understand the unforeseen benefits of inactions
- to recognize an individual's limitation and the need for collaboration and teamwork
- the limitation of technology and the wisdom of humanity

CHAPTER 5: ILLUSTRATIVE SCENARIO OF THE A2E MODEL

5.1 Background

According to the 2018 U.S. Renal Data System (USRDS) Annual Data Report (Saran et al., 2018), in 2016 more than 700,000 Americans suffered from the End-Stage Renal Disease (ESRD), an irreversible kidney failure which requires either kidney transplant or renal replacement therapy also known as dialysis. In the same year, about 95,000 ESRD patients died and about 125,000 people were newly diagnosed with ESRD leading to the increase of the overall ESRD population over the prior year. Of all ESRD patients, only less than 3% received kidney transplants and the rest must rely on the ongoing maintenance dialysis care to sustain their lives. Because of the tremendous health, quality of life, and financial burden on the ESRD patients and their families, ESRD has been the only disease that enjoyed the universal healthcare insurance coverage by the federal government since 1972 under the social security legislation.

According to the same report, in 2016, the federal government through the Medicare program spent $35.4 billion on healthcare for about half a million Medicare beneficiaries with ESRD. While the population of ESRD patients represents only less than 1% of total population of Medicare beneficiaries, the cost of caring for them represents 7.2% of total Medicare cost (Saran et al., 2018).

According to CMS Dialysis Facility Compare website, there are more than 7,000 dialysis facilities nationwide that provide dialysis care to all CMS beneficiaries with ESRD (Centers for Medicare and Medicaid Services, n.d.-a) Improving the quality and reducing the cost of dialysis care for the Medicare beneficiaries are of major concerns of the federal government. Over the years CMS has established multiple programs to ensure quality and transparency of dialysis care. These include:

- Dialysis Facility Compare (DFC) (Centers for Medicare and Medicaid Services, n.d.-a)
- Dialysis Facility Reports (DFR) (University of Michigan Kidney Epidemiology and Cost Center, n.d.)
- ESRD Quality Incentive Program (QIP) (Centers for Medicare and Medicaid Services, n.d.-b).

These programs collect administrative, clinical, and claim data on all dialysis facilities, calculate quality measures, and produce public reports to help ESRD patients, providers, policymaker, and the public to assess, compare, and improve the quality of dialysis care.

While the public reporting data files are useful, they are in the traditional file formats such as Excel spreadsheet, PDF, or CSV and are neither easy to read nor friendly to use. Furthermore, there are hidden patterns in the data that can be uncovered through modern data analytics to improve quality of dialysis care.

This research, serving as a method to evaluate the proposed A2E analytics process model, will perform data analytics using data from these public reporting files to identify factors that may influence the quality of care by dialysis facilities. The results will enable the dialysis facilities and CMS policymakers to make informed decisions and take effective actions to improve care quality for the hundreds of thousands of ESRD patients. This evaluation research, serving as an illustrative scenario, aim to demonstrate the relevant and effective application of the principles and guidelines prescribed by the A2E model.

This illustrative project will be carried out primarily by the researcher in collaboration with several healthcare quality experts. The researcher has many years of hands-on experience in software programing, database design, systems engineering, and project management. In the past

few years, the researcher has also received training on data visualization and machine learning and recently completed doctoral courses on statistics, data analytics, and research methodology. In addition, the researcher has a Bachelor of Engineering degree in Management Information Systems, a Master of Art degree in Economics, a Master of Science degree in Statistics, and a Master of Business Administration degree in Finance. The combination of strong IT experiences, formal education in Science-Technology-Engineering-Math (STEM), economics, and business, and training in the latest data analytics tools and techniques uniquely qualifies the researcher to take on this analytics effort.

Since 2017, the researcher has been on a project supporting CMS's ESRD QIP program as a business requirements and systems development subject matter expert (SME). This work assignment has provided opportunities for the researcher to learn and acquire valuable knowledge about healthcare quality management and dialysis care for ESRD patients.

5.2 Donabedian Quality Model

As a part of preparation for taking on this effort, the researcher spent time learning about healthcare quality models and their applications in population health. About a year ago, a healthcare quality expert recommended the Donabedian Structure-Process-Outcome model for healthcare quality Donabedian (1966) to the researcher as a tool to guide research in healthcare quality improvement. Donabedian model is a simple yet practical and adaptable framework for examining the quality of healthcare and driving improvement of care quality and remains one of the most adopted quality models for healthcare ever since it was introduced half a century ago. The model measures the healthcare quality from three interconnected dimensions as depicted in **Figure 31**:

- *Structure:* How the care is organized such as physical and organizational settings in which care is delivered
- *Process:* What is done to improve the health of patients such as treatments of conditions and care for patients.
- *Outcome:* What is the result of the care. The result can be either positive and adverse.

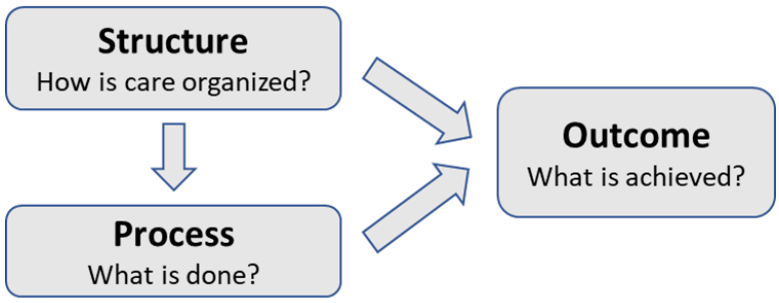


Figure 31. Donabedian healthcare quality model by Donabedian (1966)

The structural measures include both the physical and organizational settings in which care is organized and carried out. For example, these can include the facility, equipment, human resources, and financial resources. Structure can affect how the care is delivered and can also affect the outcomes of care as well.

The process measures include actions taken, care delivered, or any practice patterns that aim to improve patient health. These can include clinical measures such preventive care, diagnosis, treatment or non-clinical measures such as patient education and coordination of care.

The outcome measures are effects of healthcare on patients or population and can be either positive or negative such as survival or mortality. Adverse outcomes are the result of overuse, underuse, or misuse of care (Orszag, 2008). Outcome measures can be clinical such as the cure of a disease, the removal of a tumor, and the lowering of blood pressure which are easier to measure or non-clinical such as patient experience and satisfaction and patient quality of life which are much harder to measure.

Donabedian quality model was used as a framework to categorize quality measures and the associated data elements for this healthcare analytics effort. The remainder of this chapter will chronicle the application of the five-step A2E process model in this analytics effort as an evaluation method to illustrate how the A2E process model can be applied to data analytics practice.

5.3 Step One – Assess Needs

Activity #1. Determine the business need. The researcher presented the idea of performing data analytics on care quality of Medicare certified dialysis facilities using publicly available datasets to several healthcare experts and asked for advice and support. In addition, the researcher presented the initial idea of using Donabedian's Structure-Process-Outcome healthcare quality model to guide the selection of quality measures and asked for advice on what measures are important and should be included.

There are many measures for the outcome of dialysis care. The following list some of them:

- Mortality or survival
- Hospital admission
- Hospital readmission
- Patient experience or satisfaction
- Patient quality of life
- Cost of care

While each one of the outcome measures is important and deserves attention, it is impractical for the researcher to work on all of them due to the limited time and resources available. In addition, the goal of this research is to illustrate how the A2E analytics process

model can be applied to healthcare analytics. It would be prudent to keep the scope limited and manageable. One of the experts recommended focusing on the hospital readmissions among ESRD patients. He pointed out that hospital readmission not only is a huge financial burden on the federal government but also a disruption and burden to a patient's daily life and overall quality of life. If we could uncover insights from the data about factors that may influence the hospital readmission, it would help the dialysis facilities take steps and help the federal government adjust policies to reduce hospital readmissions. According to the U. S. Renal Data System (USRDS),

> *Hospital admissions, subsequent readmissions, and emergency department visits are a major burden for patients with ESRD. On average, patients with ESRD are admitted to the hospital more than once a year, and more than one in three hospital discharges are followed by a readmission within 30 days. Furthermore, inpatient treatment represents a significant societal and financial burden, accounting for approximately 33% of total Medicare expenditures for patients with ESRD. (Saran et al., 2018)*

The U.S. Renal Data System defines readmission as "a hospital admission occurring within 30 days of a hospital discharge, excluding emergency room visits and those intended for rehabilitation purposes". Readmissions have grave adverse effect and are associated with "increased morbidity and mortality and reduced quality of life" and "pose a significant societal and financial burden, particularly for ESRD patients." (Saran et al., 2018). The increased understanding of the importance of hospital readmission along with the recommendation from the healthcare expert led to the decision of focusing on the readmission outcome for this analytics effort.

Activity #2 Determine the data need. After the business need is identified, the researcher set out to identify data sources that would provide the data elements necessary for

measuring various quality dimensions related to dialysis care. To keep the scope of this analytics effort simple and manageable, the researcher decided to utilize secondary data sources that are available in the public domain. These data sources were published by the federal government and contain data related to dialysis facilities and do not contain data on individual patients. This eliminates the potential risks of information security and data privacy concerns such as exposure of Personally Identifiable Information (PII) and Protected Health Information (PHI).

The main data source is the Fiscal Year 2018 Dialysis Facility Report (DFR) which contains data for calendar year 2016. The yearly DFR is a flat file in Comma Separated Value (CSV) format and published by the CMS and available for public access on CMS website. The FY2018 DFR has 6,574 rows each of which represents a single dialysis facility and 3,610 columns representing various quality measures, organizational characteristics, and patient population characteristics.

There is a growing body of research linking demographic and socioeconomic factors such as race, ethnicity, poverty level, education level characteristics to health outcomes under the general scheme of social determinants of health (Barr, 2014; Marmot, 2005; McGovern, Miller, & Hughes-Cromwick, 2014). Research linked socioeconomic factors to the health outcome of chronic kidney disease (CKD) including End-Stage Renal Disease (ESRD) (Nicholas, Kalantar-Zadeh, & Norris, 2015). Research also found that the demographic and socioeconomic characteristics of the neighborhood where a dialysis facility is located are associated with a dialysis facility's care quality (Zhang, 2016).

Data on demographic and socioeconomic characteristics of the patient population of a facility is already included in the DFR file, however, the file does not contain data elements for the socioeconomic characteristics of the community where a facility is located. The community

characteristics can be obtained from the U.S. Census Bureau American Community Survey data based on zip code. Since the DFR dataset does not have data element for the zip code of a facility, two other data sources were used to obtain the zip code: The Dialysis Facility Compare (DFC) and the ESRD Quality Incentive Program (QIP). The U. S. Census region and division codes list were used to determine the geographical region and division of a facility and the Rural-Urban Commuting Area Codes (RUCA) were used to determine whether a facility is in an urban or a rural area.

Table 13 summarizes the data sources used in the analytics effort. The combination of these data sources will provide a comprehensive view of the potential factors influencing the hospital readmissions of ESRD patients under the care of dialysis facilities. This demonstrated the importance of utilizing and integrating data from multiple sources.

Table 13

The Data Sources

Data Source	Description
Dialysis Facility Report (DFR) Fiscal Year 2018	The annual DFR is provided by CMS to stimulate quality improvement efforts for all dialysis facilities. This report allows comparisons of the characteristics of a facility's patient populations, patterns of treatments, and outcomes such as hospitalization and mortality to other facilities and to the state and national averages.
Dialysis Facility Compare (DFC) Fiscal Year 2018	The DFC is a rating system developed by CMS for ranking dialysis facilities by comparing the quality and outcome of dialysis care provided by all facilities. Each facility is graded on nine different health statistics such as mortality, hospitalizations, and blood transfusions.
ESRD QIP Payment Year 2018 Performance Score Summary Report (PSSR)	CMS's ESRD Quality Incentive Program calculates a quality performance score for each facility and compares it to a national minimal standard. Facilities scoring below the standard will have its payment for dialysis reduced by up to 2%.
U.S. Census Bureau five-year American Community Survey (ACS) 2016	DP03 – Economic Characteristics (income, employment) DP02 - Social Characteristics (language, education) DP05 - Demographic and Housing Estimates (sex, age, race, ethnicity)
U. S. Census Bureau Region and Division	The Census Bureau divides the 50 states and the District of Columbia into four regions and nine divisions. Puerto Rico and the Island Areas are not part of any census region or census division.
Rural-Urban Commuting Area (RUCA) Codes	The rural-urban commuting area codes (RUCA) uses data from the decennial census on urbanization, population density, and daily commuting to classify geographical areas.
CMS Manual	This manual specifies a dialysis facility hospital affliction based on the range of the last four-digit of its six-digit CMS Certification Number (CCN).

After the data sources were identified and the datasets were retrieved and reviewed, the researcher identified data elements from those data sources that are relevant to dialysis care quality measures based on the Donabedian's model. **Table 14** lists those data elements categorized according to the Strucutre-Process-Outcome model:

Table 14

Quality measures categorized according to the Donabedian quality model

Donabedian's Dimension	**Dialysis Care Quality Measures**
Outcome	• Standardized Readmission Ratio (SRR)
Structural - Facility Geographical Characteristics	• Region • Division • Network • State • Urbanicity
Structural - Facility Organizational Characteristics	• Profit Status • Chain Ownership • Hospital Affiliation • Compliance Status • Staff Patient Ratio • Station Patient Ratio
Structural – Facility Patient Characteristics	• Age • Gender (% Female) • Race (% African American) • Ethnicity (% Hispanic) • Insurance Coverage (% Medicare Coverage)
Structural – Facility Community Characteristics	• Race (% African American) • Ethnicity (% Hispanic White) • Income (% Family with Income Below Federal Poverty Level) • Unemployment Rate • English proficiency (% People aged 5 and over that have a primary language other than English and speak English less than well)
Process	• Ultrafiltration Rate (UFR) • Kt/V Dialysis Adequacy • Hemoglobin Level • % Patients ESA Prescribed • % Patients Influenza Vaccinated • % Patients on Arteriovenous (AV) Fistula • % Patients on Catheter Only Over 90 Days • Dialyzer Reuse • Evening Shift

Activity #3. Determine the technology need. Although the DFR dataset is not considered big data, it is still a large dataset with about 200 megabytes in size and has 3,574 rows and 3,610 columns. In addition, the DFR dataset needs to be joined with data from other identified sources. Clearly, Microsoft Excel is not an ideal choice for preprocessing and blending multiple datasets and performing analytics. The researcher reviewed and evaluated programing language R and Python and decided to use Python for blending data for its powerful and flexible built-in data processing packages including Numpy and Pandas.

To make the analytical process transparent, reproducible, and sharable, Jupyter Notebook was used. Jupyter Notebook is an integrated development environment (IDE) for developing, documenting, and sharing code. It generates an interactive document that embeds software code and its results with comments and narratives so that the process can be easily shared with fellow researchers and practitioners for collaboration, repeatability, and transparency. There are several cloud-based platforms available for developing and hosting Jupyter Notebook. Kaggle Jupyter Notebook, Microsoft Azure Notebooks, and IBM Data Science Experience (DSX are among the most popular ones and are freely available. The researcher chose Kaggle as the development environment for data preprocessing and blending using Python. Kaggle provides free storage to host the data files used and the Jupyter notebooks developed in addition to providing the run-time development environment. Kaggle is a leading online community for data scientists and data analysts.

The researcher used Tableau Public Desktop to develop interactive data visualizations and used Tableau Public website to publish and share data visualizations. Tableau Public is a free service provided by Tableau and makes it easier for the research to develop and share the

visualizations with healthcare policy and quality experts. The data analytics tools selected are summarized in **Table 15**.

Table 15

The Data Analytics Tools

Tool	**Description**
Python (v3.7)	Python is the most popular program language used by data scientists due to its versatile capabilities and large number of packages for scientific computing, data processing, statistics, and machine learning. This research project takes advantage of the Panda package and its data frame object for blending data and the statsmodel and scikit-learn package for predictive analytics.
Jupyter Notebook	The Jupyter Notebook is a development environment that facilitates the creation and sharing of interactive documents that contain source code, the outputs including visualizations, and narrative text. It helps facilitate and promote collaborative, reproducible, sharable, and transparent programing, data analytics, and general research effort.
Kaggle	Kaggle provides cloud-based Jupyter Notebooks development environment. The source data files, intermediate files, final data file, and the Jupyter Notebooks that perform the data blending can be accessed from the following URL: **https://www.kaggle.com/wcj365/rmudsc** (requires Kaggle account login)
Tableau Public Desktop (v2018.3)	Tableau Public Desktop is the free version of the commercial Tableau data visualization tool that allows researchers and practitioners to develop interactive and visual analytics for publishing to the free Tableau Public website.
Tableau Public Website	Tableau Public is a free website that hosts data visualizations developed using Tableau Public Desktop. The visual and interactive analytics developed as a part of this research project can be accessed from the following URL: **https://public.tableau.com/profile/jaywang**

5.4 Step Two - Blend Data

The following activities were performed to preprocess and blend data from the identified sources to produce a final consolidated dataset for developing visual analytics:

Activity #1 - Import and consolidate DFR, DFC, and QIP datasets. The DFR dataset has the most comprehensive information about dialysis facilities, clinical quality measures, and patient population characteristics and is the main data source for this effort. Even though DFR dataset contains facility name, city, and state, unfortunately it does not contain the zip code of a facility's physical location. We need the zip code to get information about the demographic and socioeconomic characteristics of the community where a facility is located by using US Census Bureau American Community Survey data and to determine whether a facility is in a rural or urban area by using Rural Urban Commuting Areas (RUCA) data. These demographic and socioeconomic characteristics of a facility may influence the quality of care provided by the dialysis facilities. CMS publishes two additional datasets related to dialysis facility quality measures and they both contain zip code of a facility's physical location. One is dialysis facility compare (DFC) dataset. The other is ESRD Quality Incentive Program (QIP) performance score summary report dataset. All three datasets contain the CMS Certification Number (CCN), a unique identifier for a facility. CCN numbers were used to match facilities and merge zip codes from the DFC and QIP datasets into the DFR dataset.

Activity #2. Obtain the demographic and socioeconomic characteristics of the community where a facility is located. The five-year American Community Survey data for 2016 were used.

First, we import the ACS's DP05 dataset which has demographic information about gender, age, race, and ethnicity. Two data elements were obtained:

- HC03_VC50 - Percent of RACE - One race - Black or African American.
- HC03_VC88 - Percent of HISPANIC OR LATINO AND RACE - Total population - Hispanic or Latino (of any race)

Secondly, we import the ACS's DP03 dataset which has socioeconomic data elements. Two data elements were obtained:

- HC03_VC07 – Unemployment Rate
- HC03_VC161 - % Family with Income Below Federal Poverty Level

Thirdly, we import the ACS's DP02 dataset. Data element HC03_VC173 is the percentage of people aged 5 and over that have a primary language other than English and speak English less than well. This is a good measure of English proficiency of the community which may be indicator of difficulty in communication that can potentially lead to hospital readmission.

Activity #3. **Determine a facility's chain ownership.** A facility can be either independently owned or belongs to a chain organization. The largest two dialysis chains are DaVita and Fresenius Medical Care (FMC) and are commonly called large dialysis organizations (LDO). There are many other chains of varying size, we group them into a new category named "Other Chains". The independently owned and operated facilities are not part of any chain organization and are placed in the "Independent" category. After the data conversion, we end up with four categories. **Figure 32** shows the distribution of facilities among the four categories. About two-third of all facilities belong to one of the two largest chains; about one-fifth belongs to other chains; about one-tenth are independent facilities.

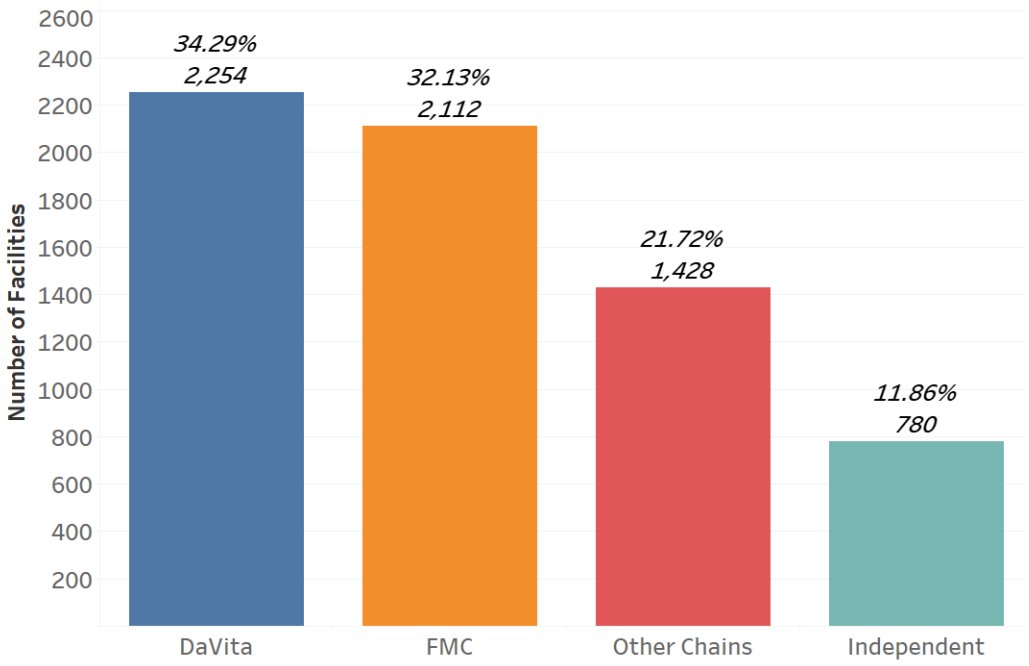


Figure 32. Dialysis facilities distribution based on chain ownership

Activity #4. Determine facilities' hospital affiliation. A dialysis facility can be either hospital affiliated or independent. Hospital affiliation may have influence on readmission since readmission may be attributed to the lack of or insufficient coordination between the dialysis facility where a patient receives dialysis care and the admitting hospital where a patient was initially admitted. While the data sources do not contain data element about a facility's hospital affiliation, the CMS operational manual provided information about the association of facility types and CCN number ranges. CCN is six-digit long and the ranges of the last four digits are assigned by CMS for specific type of facilities as indicated below (Centers for Medicare and Medicaid Services, n.d.-c):

- 0001-0879 Short Term (General and Specialty) Hospitals
- 2000-2299 Long Term Hospitals
- 2300-2499 Chronic Renal Disease Facilities (Hospital-Based)
- 2500-2899 Non-Hospital Renal Disease Treatment Centers

- 2900-2999 Independent Special Purpose Renal Disease Facilities
- 3300-3399 Children's Hospitals
- 3500-3699 Renal Disease Treatment Centers (Hospital Satellites)
- 3700-3799 Hospital-Based Special Purpose Renal Dialysis Facilities

This mapping was used to derive a facility's hospital affiliation based on its CCN number. **Figure 33** shows the results of the data transformation which indicates that most facilities are not affiliated with hospitals.

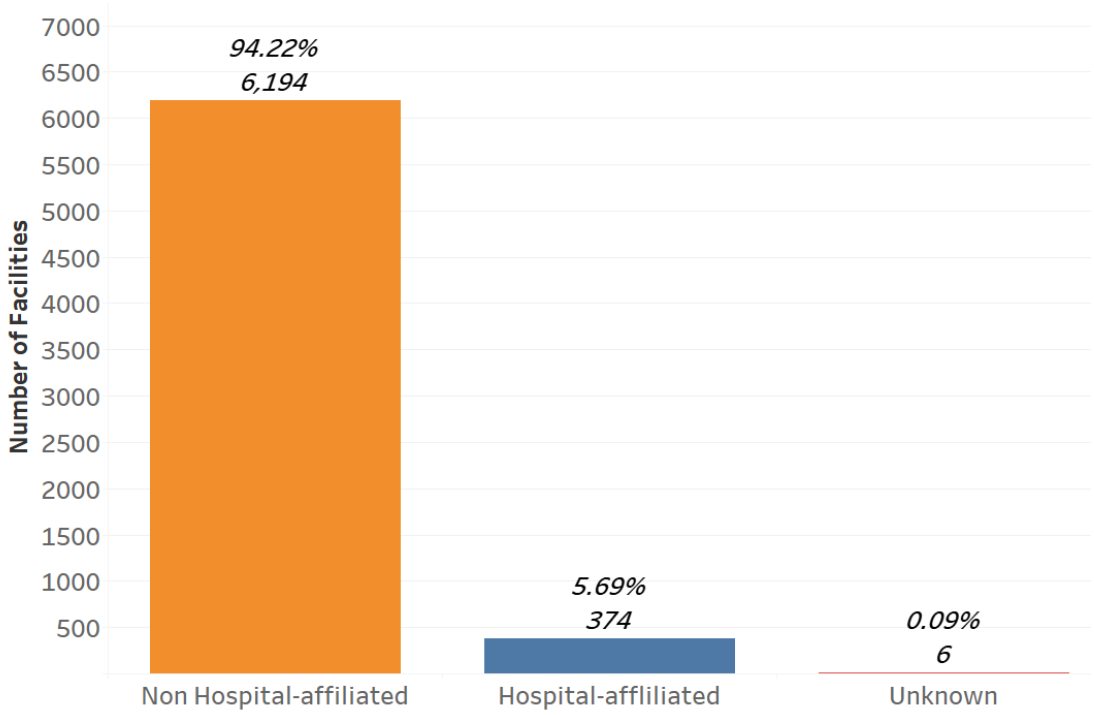


Figure 33. Dialysis facilities distribution based on hospital affiliation

Activity #5: Determine the geographical region. The U. S. Census Bureau's state, region and division mapping was used to determine the geographical areas of a facility based on the facility's state code. **Figure 34** show the distribution of facilities along the four regions. 57 facilities are US oversea territories. Most facilities are located in the South.

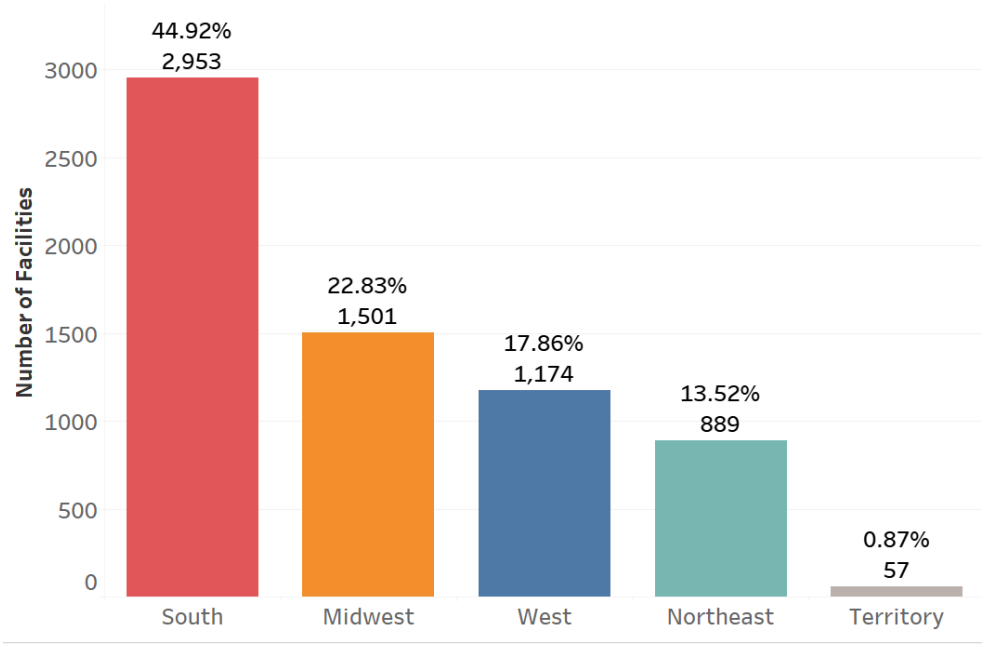


Figure 34. Dialysis facilities distribution based on census region

Activity #6. Determine a facility's urbanicity. The Rural Urban Commuting Area (RUCA) codes were used to determine the urbanicity of a facility based on the zip code of its physical location. The codes classify U.S. census tracts using measures of population density, urbanization, and daily commuting. The classification specifies metropolitan, micropolitan, small town, and rural commuting areas based on the size and direction of the primary (largest) commuting flows. Metropolitan areas are considered urban while the rest are considered rural. **Figure 35** shows the distribution of facilities based on urbanicity. Close to three-fourth of facilities are in the urban areas. Note that there are 417 facilities (over 6%) are not classified due to the lack of match in zip code.

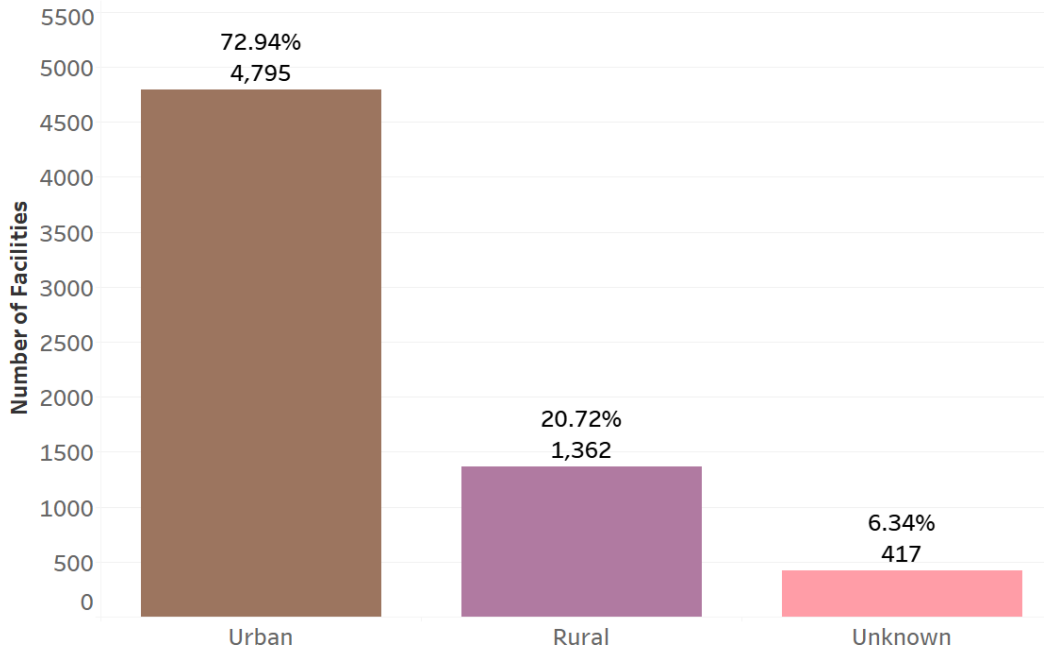


Figure 35. Dialysis facilities distribution based on urbanicity

Activity #7. Determine facilities' years in operation. The DFR dataset comes with a data element for the initial CMS certification date. This date was used to calculate the years since a facility was certified by CMS for operation.

Data blending is the most time-consuming phase of data analytics process and is also error-prone. The quality of data determines the quality of the analytics. As often said, "garbage in, garbage out". To ensure the quality of data and promote reproducibility and transparency of data blending, the researcher used Jupyter Notebook. Notebooks are "documents integrating prose, code and results—offer a way to publish a computational method which can be readily read and replicated" (Kluyver et al., 2016, p. 87). This provides visibility into how data are preprocessed and blended, what the rationale is and what the results are at every step.

Activity #8. Share Data and Data Blending Source Code. To be transparent and share knowledge, the researcher made public both the datasets used and source code files developed on

Kaggle. **Table 16** lists the URLs on Kaggle.com to access the source data files and the Jupyter notebooks.

Table 16

The Web Links to Access the Data Files and Jupyter Notebooks

Content	**URL**
Home page and data sources	http://www.kaggle.com/wcj365/rmudsc
Notebook for activity one	https://www.kaggle.com/wcj365/activity-one-blend-dfr-dfc-qip-data
Notebook for activity two	https://www.kaggle.com/wcj365/activity-two-blend-acs-data
Notebook for activity three through seven	https://www.kaggle.com/wcj365/activity-three-to-seven

Figure 36 shows the home page of the project sowing the data sources.

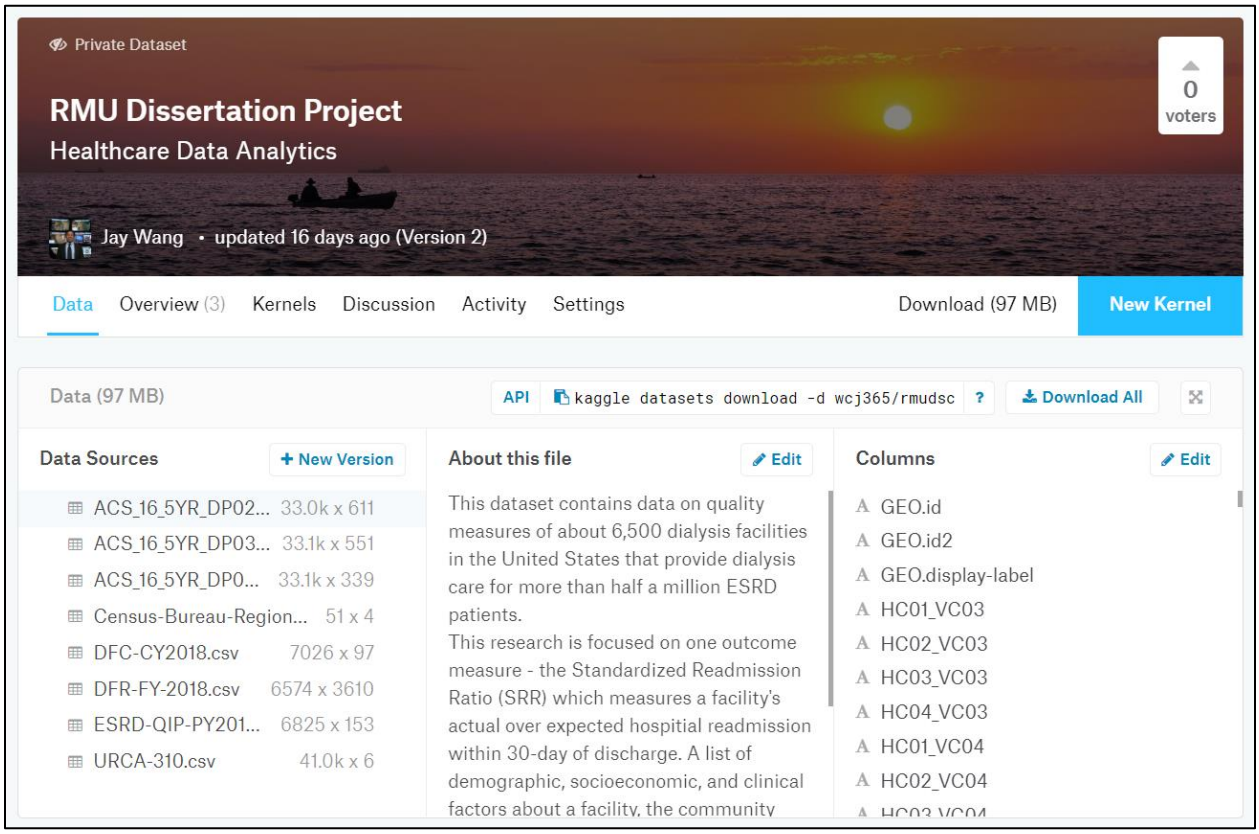


Figure 36. The home page of the project

Figure 37 shows a sample section of the Jupyter Notebook illustrating the mixture of narrative text, Python code, and the results.

Figure 37. A sample section of a Jupyter notebook

After the data blending effort is completed, we now have one single consolidated dataset that contains all the relevant data elements needed for creating analytics including descriptive statistics, visual analytics, and predictive analytics.

5.5 Step Three - Create Analytics

Analytics comes in many shapes and forms. It is usually categorized into three types: Descriptive, Predictive, and Prescriptive. Data visualization is an indispensable tool for all types of analytics and can be performed as an independent analytics effort or the stepping stone into the more advanced and sophisticated analytics. It is particularly powerful in discovering and communicating hidden patterns and trends without the employment of complex algorithms and

complicated interpretations. Visual analytics can show what has happened (descriptive), what may happen in future (predictive), and what actions may help to exert influence (prescriptive). We will use data visualization software to create interactive visual analytics first followed by predictive analytics using automatic machine learning.

5.5.1 Descriptive Analytics using Interative Data Visualization

The consolidated final dataset was imported into Tableau Public Desktop (version 2018.3), a leading data visualization and analytics tool. Tableau provides easy to use interface for data visualization and exploratory data analysis (EDA) and requires no hardcore programing skills. The data visualizations are interactive and allow for explorations by placing mouse over different spots and selecting single quality measure from a list. The visual analytics produced from Tableau are visually appealing and easy to understand. **Figure 38** shows an example of the Tableau Public Desktop design window.

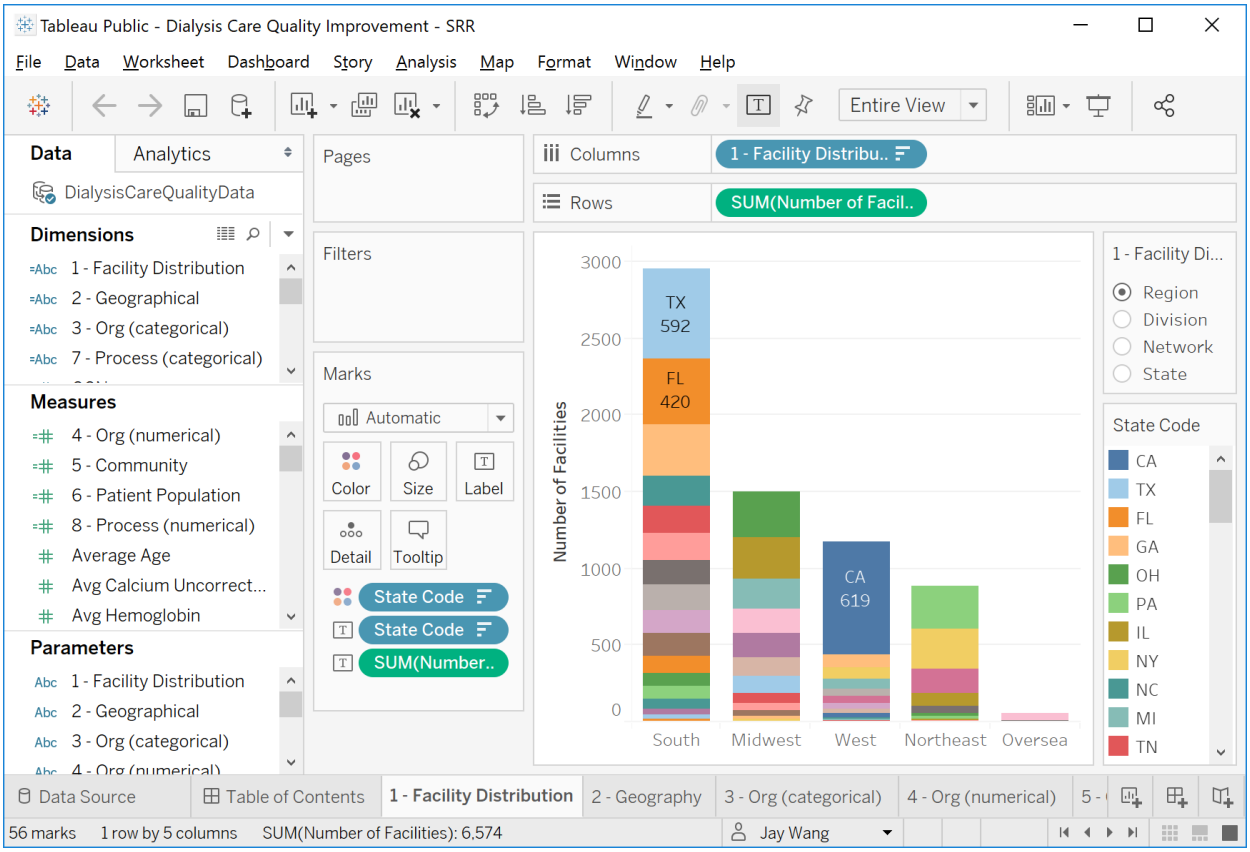


Figure 38. Tableau Public Desktop design window

Table 17 list the ten visualizations developed using Tableau Public Desktop:

Table 17

The Ten Visualizations

Visualization	**Description**
1. Facility Distribution	Explore how the 6,574 facilities are distributed across various regions, divisions, networks, and states.
2. Geographical	Explore how the Standard Readmission Ratio (SRR) varies among various geographies.
3. Organizational (categorical measures)	Explore how SRR varies among various categorical organizational factors.
4. Organizational (numerical measures)	Explore how SRR varies among various numerical organizational factors.

5. Community Characteristics	Explore how socioeconomic characteristics of the community where a facility is located may be associated with SRR.
6. Patient Population	Explore how the socioeconomic characteristics of a facility's patient population may be associated with SRR.
7. Process (categorical measure)	Explore how the categorical clinical process measures may be associated with SRR.
8. Process (numerical measure)	Explore how the numerical clinical process measures may be associated with SRR.
9. Profit status, Staffing level, and SRR	Explore the association between a facility's profit status, staffing level, and SRR.
10. Urbanicity, Staffing level, and SRR	Explore the association between a facility's urbanicity, staffing level, and SRR.

The visualizations were published to the Tableau Public website for sharing with healthcare experts as shown in **Figure 39**. The URL for visiting this web page is:

https://public.tableau.com/profile/jaywang

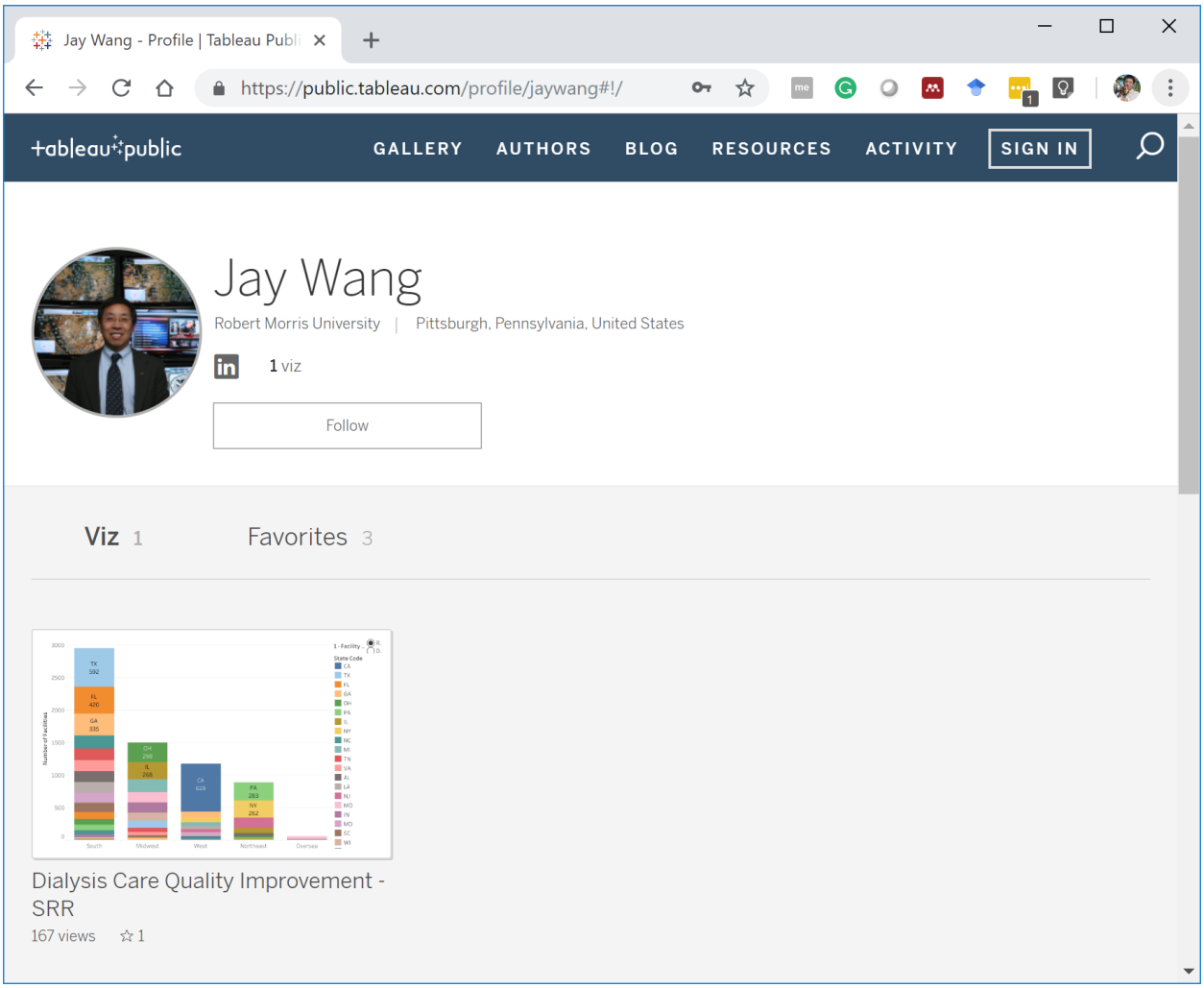


Figure 39. The home page for accessing the visual analytics

Visualization #1. This visualization was created per the suggestion of a healthcare expert. Unlike the rest of nine visualizations which focus on the relationship between the Standardized Readmission Ratio (SRR) outcome measure and various organizational, geographical, socioeconomic, and clinical factors, it explores how facilities are distributed along various geographies. This provides helpful background information.

Instead of creating one worksheet for each geographical category, a single Tableau worksheet was created with a list of ratio buttons for a user to pick and choose. A single category can be selected from the radio buttons to show how the facilities are distributed along that category. This greatly reduced the number of worksheets and made it much easier to explore and understand the descriptive statistics. The rest of the visualizations followed this same design pattern. **Figure 40** shows the screen shot.

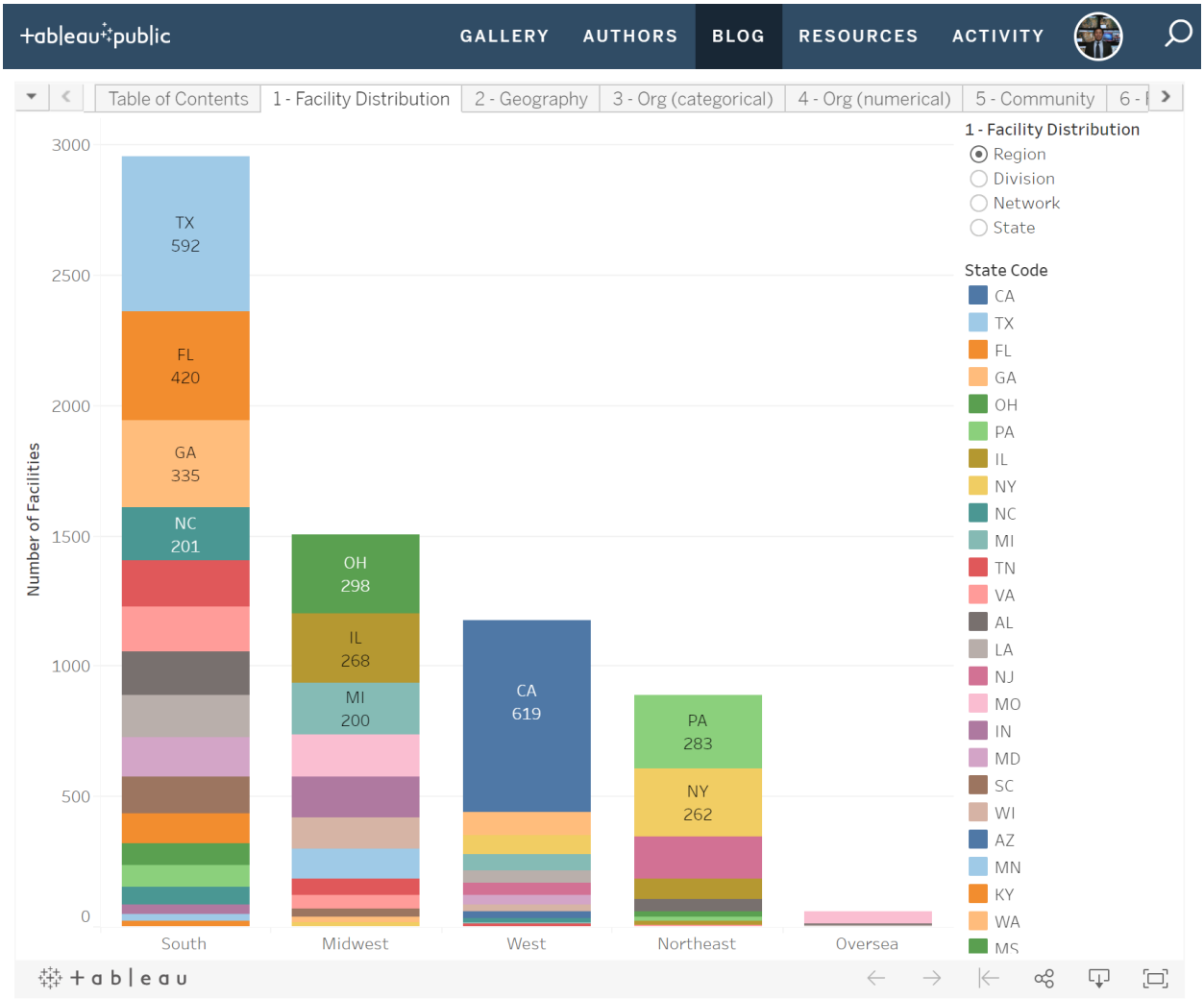


Figure 40. Facility distribution across various geographies

Visualization #2. This visualization explores the variations in SRR along various geographical areas including Region, Division, Network, State, and Urbanicity. CMS divides all facilities into eighteen networks. Each network covers facilities from one to several adjacent states. A network is an administrative entity contracted with CMS to help with quality assurance, education, and training.

The Y-axis represents the average SRR of the facilities; The X-axis represents the categorical variable selected from the list of radio buttons on the upper right corner. The visualization changes as a different category is selected.

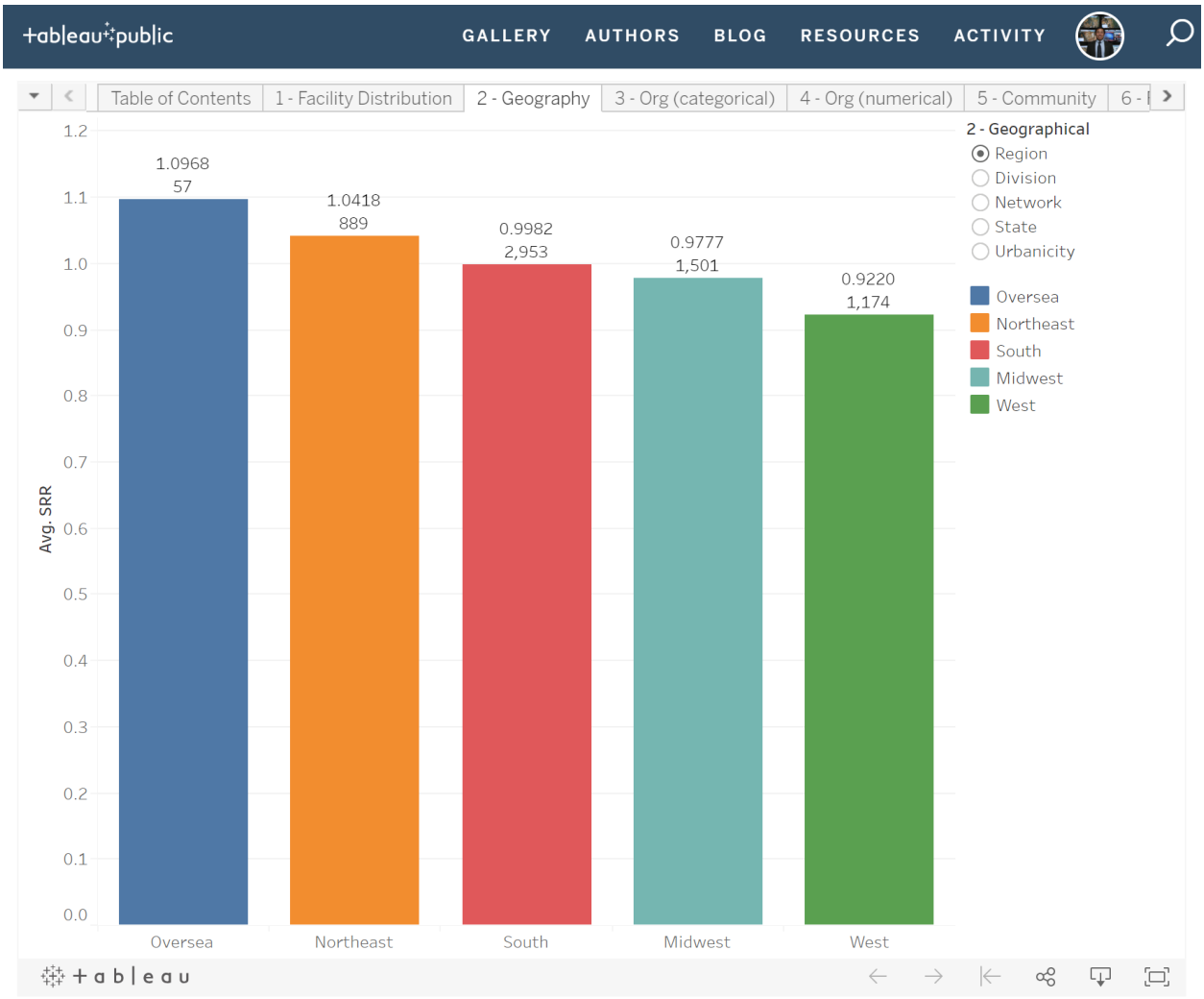


Figure 41. Average SRR across different geographies

Visualization #3. This visualization explores the variations in SRR along various categorical organizational factors including hospital affiliation, china ownership, profit status, and compliance status.

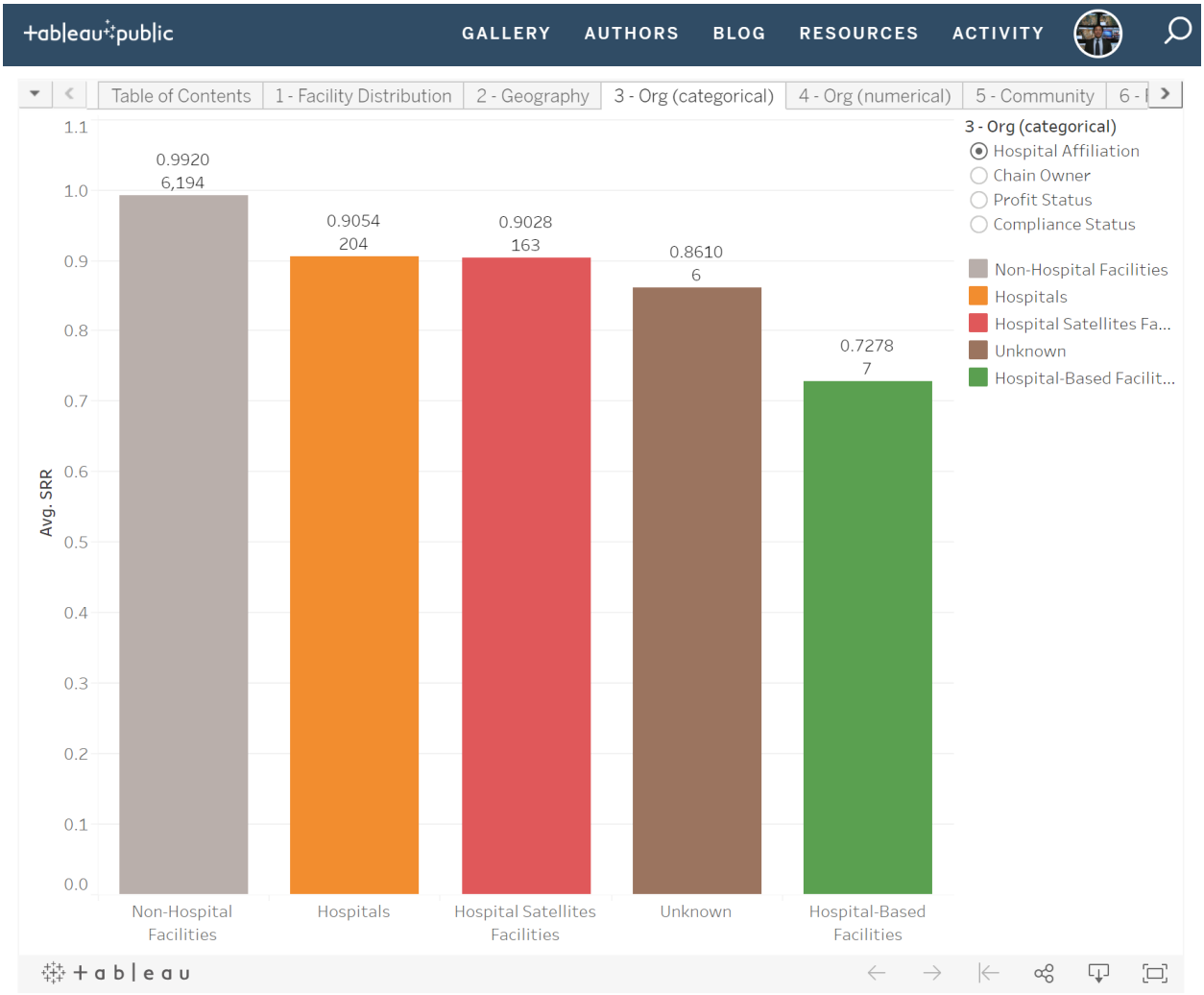


Figure 42. Average SRR across categorical organizational factors

Visualization #4. This visualization explores the variations in SRR along various numerical organizational factors including staff patient ratio, station patient ratio, and years in operation. Staff patient ratio is calculated by dividing total number of staffs by the total number of patients; Station patient ratio is calculated by dividing total number of dialysis stations by the total number of patients. Both ratios represent the capability of a facility to provide care for its patients.

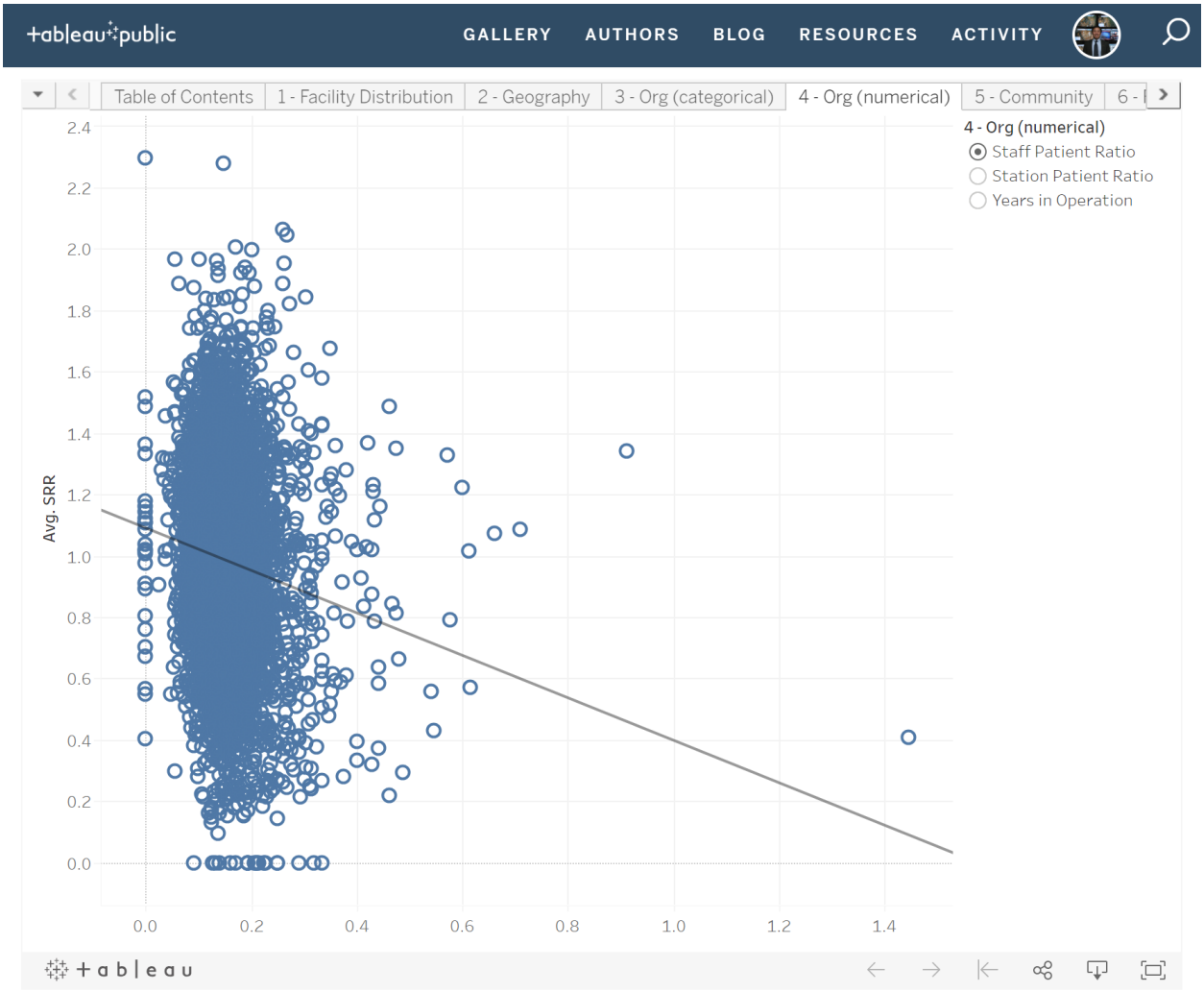


Figure 43. Average SRR across numerical organizational factors

Visualization #5. This visualization explores the variations in SRR along various socioeconomic factors of the community a facility is located. These factors include race (the percentage of African American), ethnicity (the percentage of Hispanic White), income (the percentage of families with income below the federal poverty level), English proficiency (the percentage of population aged 5 and over that have a primary language other than English and speak English less than well), and unemployment rate.

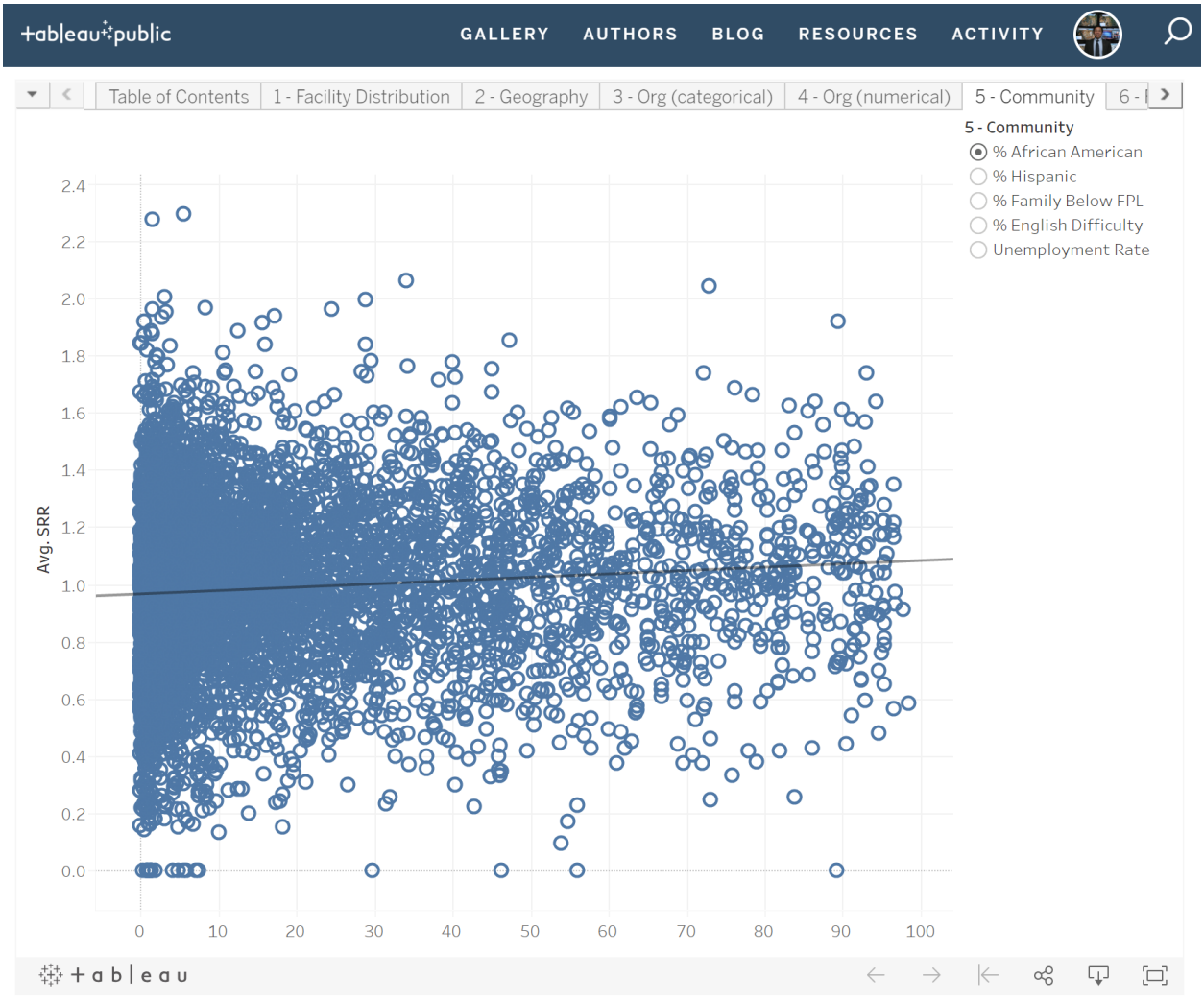


Figure 44. Average SRR across numerical socioeconomic factors (community)

Visualization #6. This visualization explores the variations in SRR along various socioeconomic characteristics of the patient populations a facility serves. These factors include race (the percentage of African American), ethnicity (the percentage of Hispanic White), gender (the percentage of female), age (average age), and insurance coverage (the percentage of patients with Medicare coverage).

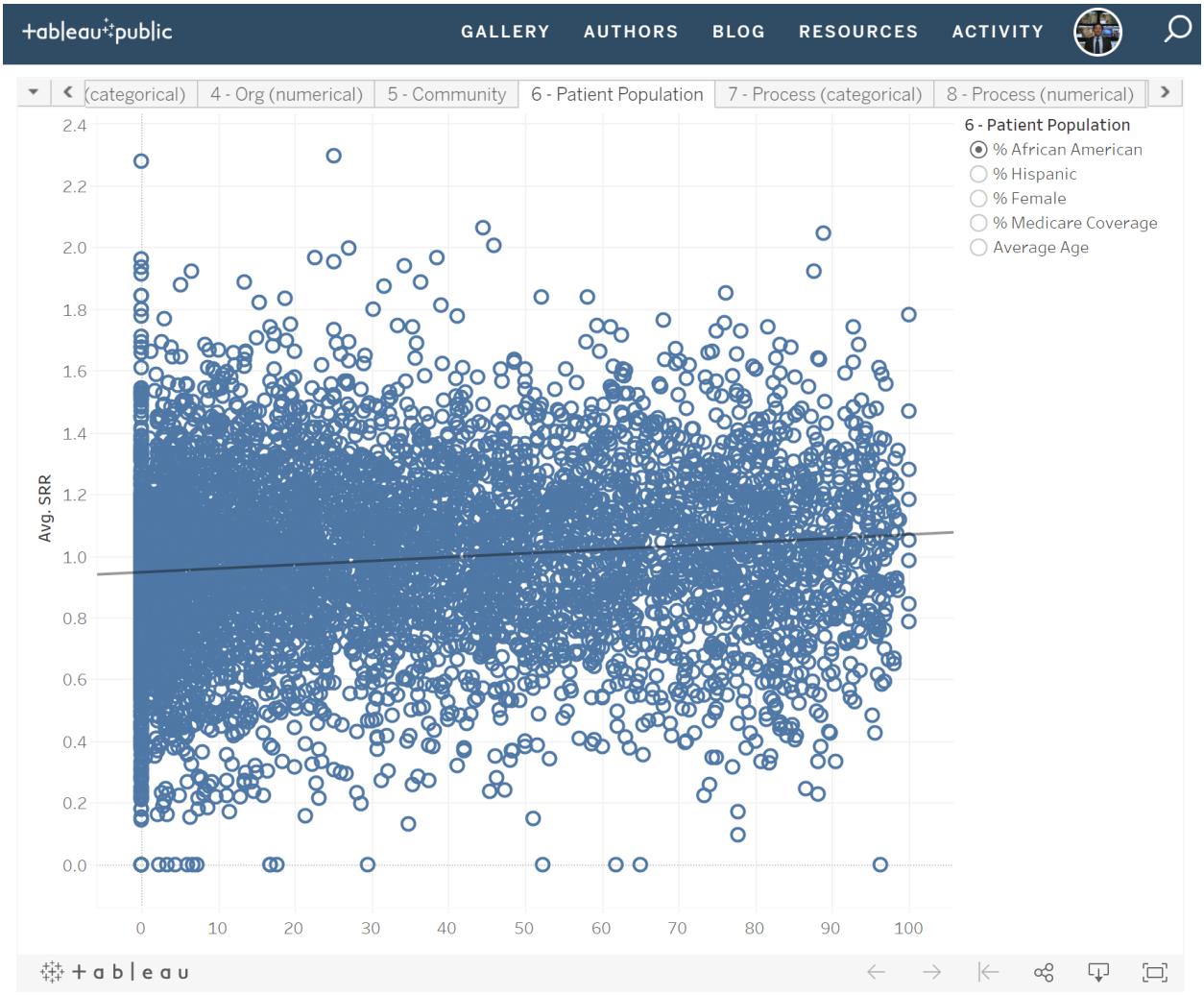


Figure 45. Average SRR across numerical socioeconomic factors (patient population)

Visualization #7. This visualization explores the variations in SRR along two categorical clinical process measures. The dialyzer reuse measure indicates whether a facility reuses dialyzer for its patients; A dialyzer, also known as “artificial kidney”, performs the function of a healthy kidney by removing the waste and fluid from the blood. Dialyzers can remain functional after more than one use. While patients don’t share dialyzers, they have the choice to whether to reuse their dialyzers or not. The evening shift measure indicates whether a facility provide dialysis care after the regular hour.

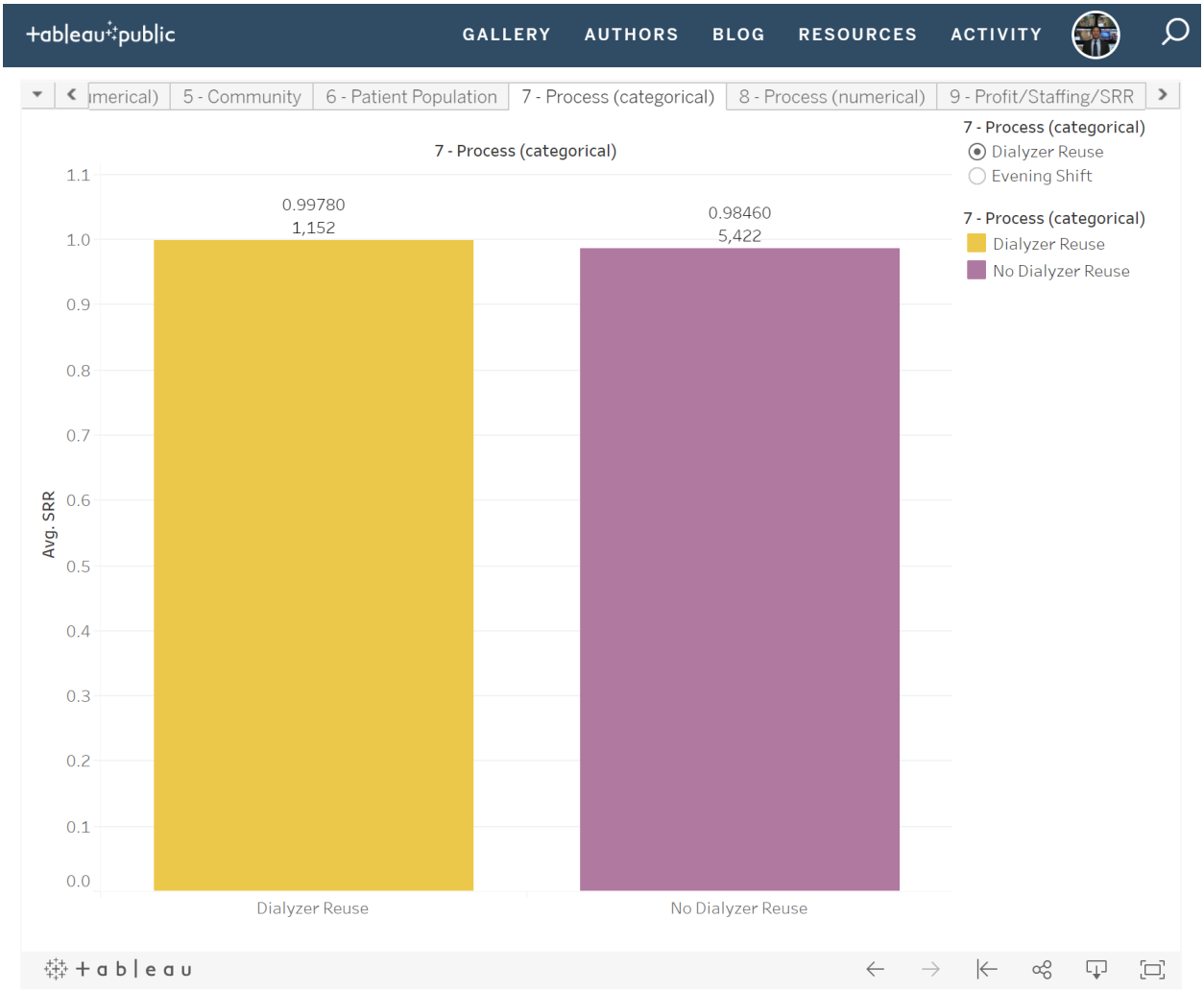


Figure 46. Average SRR across categorical clinical process factors

Visualization #8. This visualization explores the variations in SRR along various numerical clinical process measures.

Erythropoietin (EPO) is produced by the kidney and used to make red blood cells. ESRD patients typically suffers from anemia and are prescribed Erythropoietin-stimulating agents (ESA) to treat the deficiency in red blood cells.

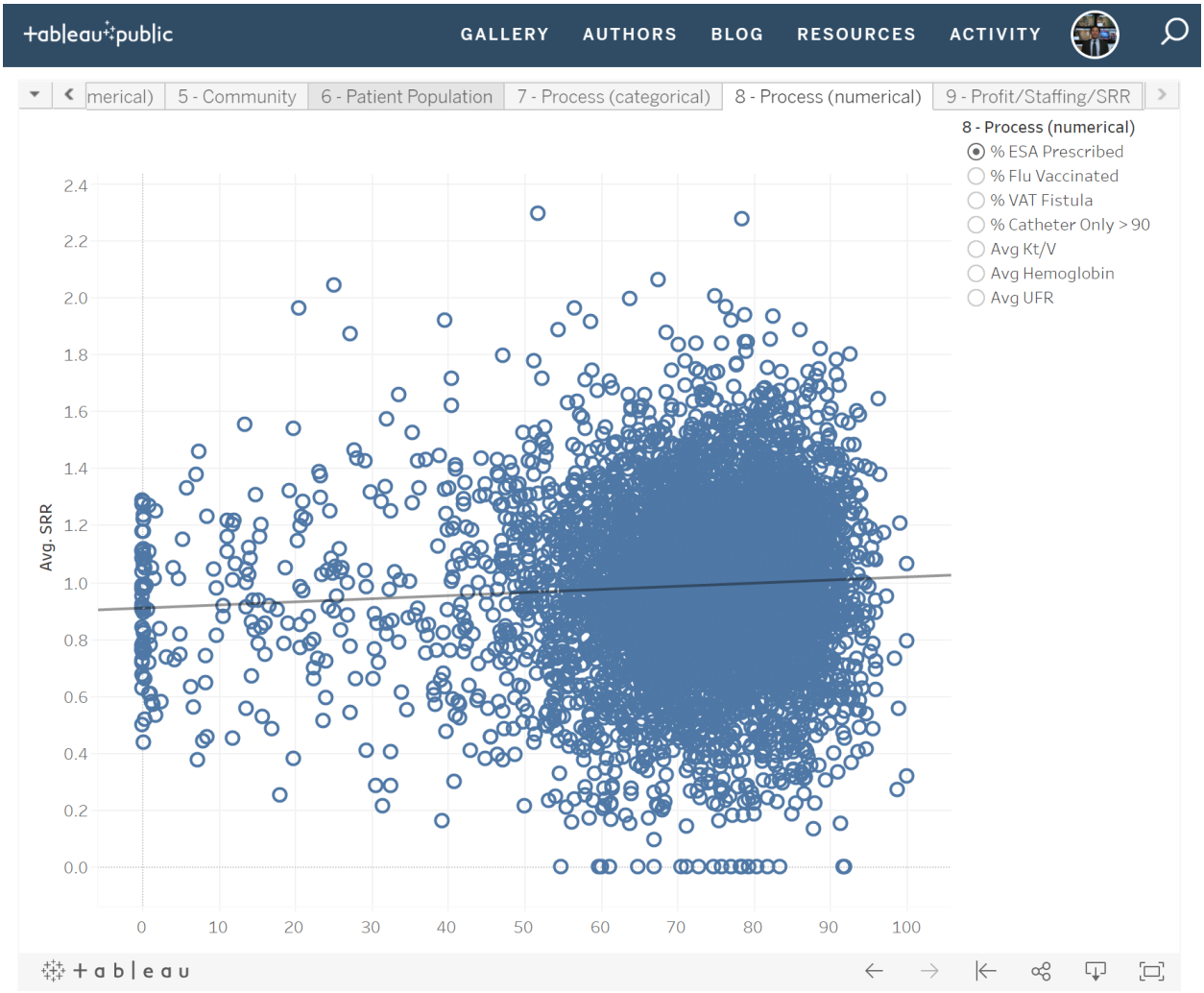


Figure 47. Average SRR across numerical clinical process factors

Visualization #9. This visualization explores the variations in SRR along both facility profit status and staffing level. This visualization was created after feedback was received from a healthcare quality expert. The question was raised why for-profit facilities have higher average SRR than not-for-profit facilities. It may be convenient to say it is because for-profit facilities care for profits more than patients. However, the expert likes to know the underlying mechanism at play beyond a simple answer. Since it was discovered that facilities with higher staff patient ratio have a lower SRR, it was reasonable to investigate the difference between the staffing levels of the for-profit and not-for-profit facilities. This visualization shows that for-profit facilities have low staffing level which may explain their higher SRR.

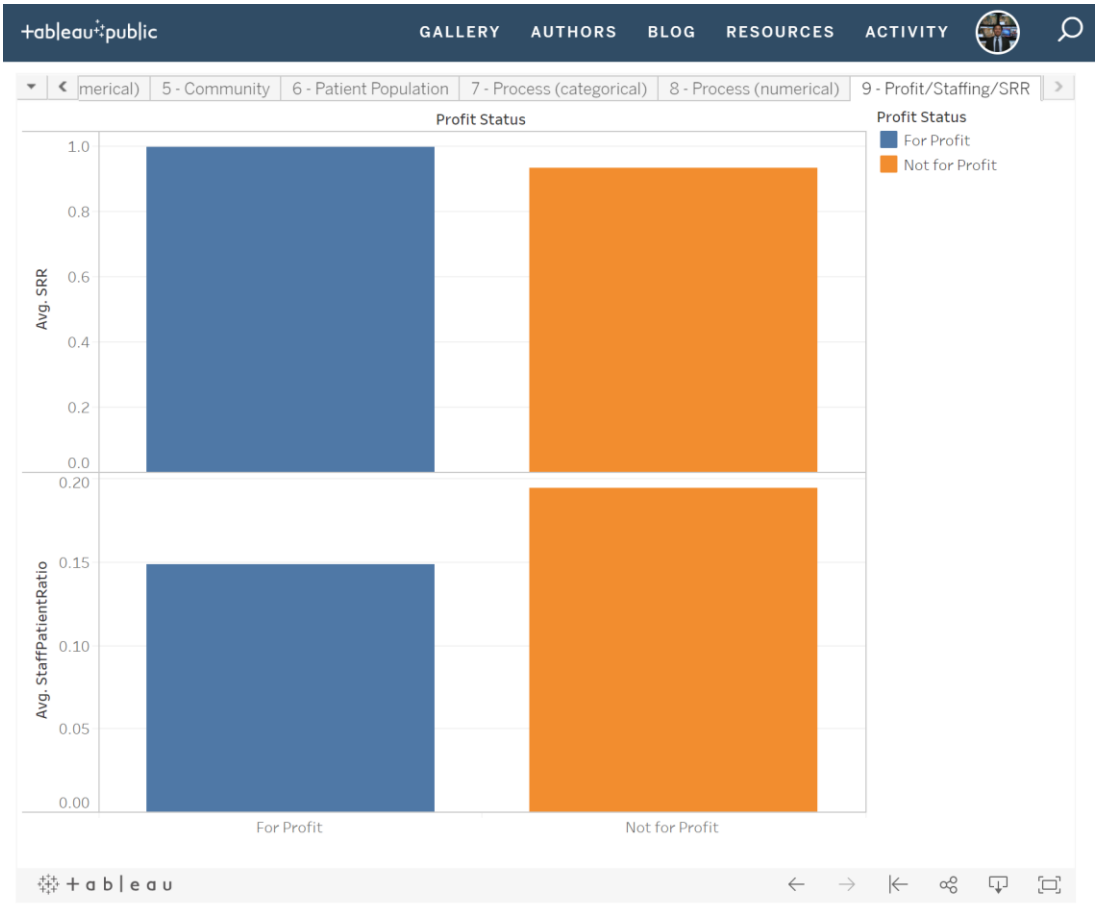


Figure 48. Associations of average SRR, profit-status, and staffing level

Visualization #10. This visualization explores the variations in SRR along both facility urbanicity and staffing level. This visualization was created after feedback was received from a healthcare quality expert. The question was raised why rural facilities have lower average SRR than urban facilities. The expert likes to know the underlying mechanism at work. Since it was discovered that facilities with higher staff patient ratio have a lower SRR, it was reasonable to investigate the difference between the staffing levels of the rural and urban facilities. This visualization shows that rural facilities have higher staffing level which may explain their lower SRR.

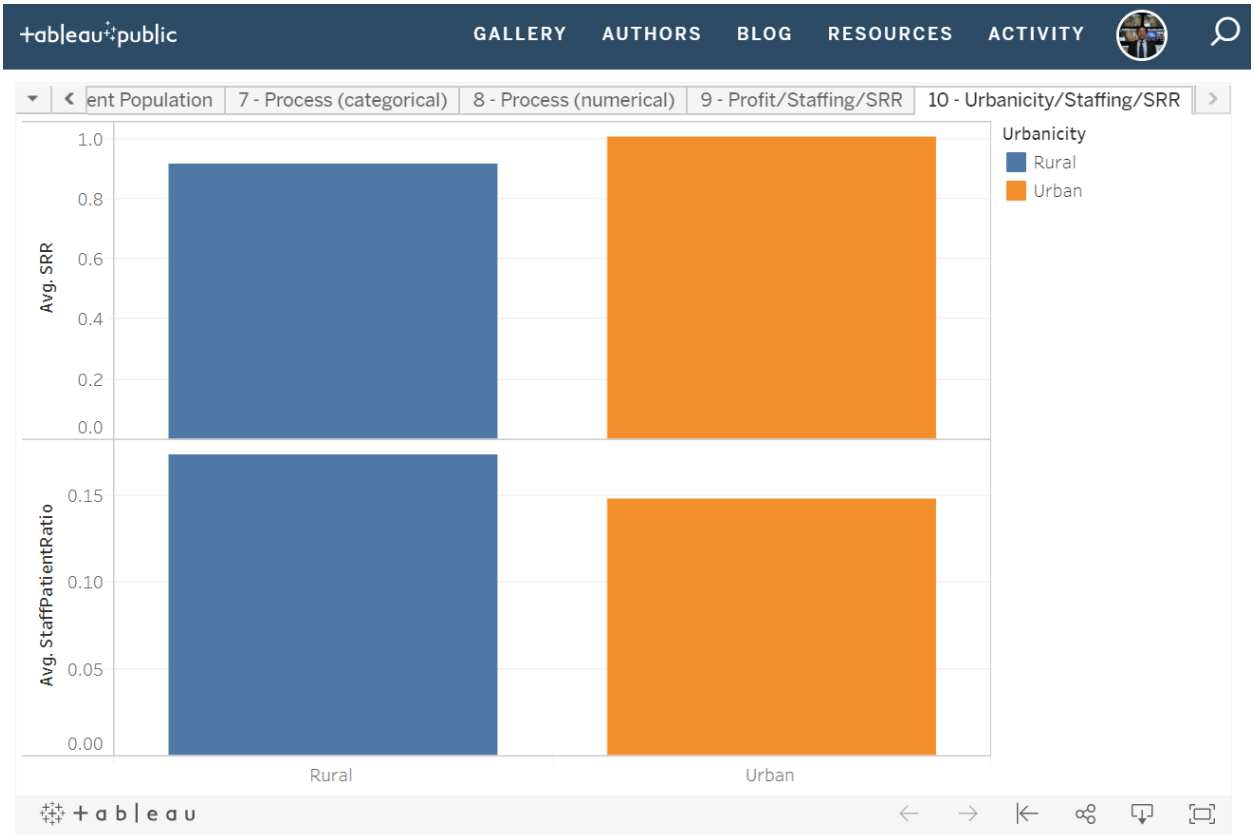


Figure 49. Associations of average SRR, urbanicity, and staffing level

Summary. Table 18 summarizes the association of the various structural and process factors with the outcome measure SRR. For each numerical measure, the scatter plot on the visualization includes a fitted line. When the mouse is placed over the line, it displays the regression equation with the intercept and coefficient, the p-value of the coefficient, and R-squared. The determination of association with SRR is based on the p-value. For categorical measures, the determination of association with SRR is subjective.

Table 18

Factors and Their Association with SRR

Quality Dimension	**Quality Measures**	**Association with SRR**
Structural - Facility Geographical Characteristics	Region	Yes
	Division	Yes
	Network	Yes
	State	Yes
	Urbanicity	Yes
Structural - Facility Organizational Characteristics	Profit Status	Yes
	Chain Ownership	Yes
	Hospital Affiliation	Yes
	Compliance Status	Yes
	Staff Patient Ratio	Positive, p-value < 0.0001, R-squared = 2.8%
	Station Patient Ratio	Positive, p-value < 0.001, R-squared = 0.7%
	Years in operation	No
Structural – Facility Patient Characteristics	% African American	Positive, p-value < 0.0001, R-squared = 1.5%
	% Hispanic	No
	% Female	No
	% Medicare Coverage	Negative, p-value < 0.0001, R-squared = 1.1%
	Average Age	No
Structural – Facility Community Characteristics	% African American)	Positive, p-value < 0.0001, R-squared = 0.8%
	% Hispanic White	No
	% Family with Low Income	Positive, p-value = 0.002, R-squared = 0.2%
	% English Difficulty	Positive, p-value < 0.0001, R-squared = 0.3%
	Unemployment Rate	Positive, p-value < 0.0001, R-squared = 0.4%
Process	% ESA Prescribed	Positive, p-value < 0.0001, R-squared = 0.3%
	% Influenza Vaccinated	Negative, p-value < 0.0001, R-squared = 0.8%
	Average Kt/V	Negative, p-value < 0.0001, R-squared = 1.0%
	% VAT Fistula	Negative, p-value < 0.0001, R-squared = 0.8%

Quality Dimension	Quality Measures	Association with SRR
	% Catheter Only > 90 Days	Positive, p-value < 0.0001, R-squared = 0.3%
	Average Hemoglobin	Negative, p-value < 0.0001, R-squared = 2.0%
	Average UFR	No
	Dialyzer Reuse	No
	Evening Shift	No

5.5.2 Predictive Analytics using Automated Machine Learning

Automated Machine Learning (AutoML) is an emerging technology that aims to address the issue of low supply of data science skills and high demand for data analytics. Data science is multidisciplinary and interdisciplinary in nature and requires the knowledge and skills in mathematics, statistics, computer science, software engineering, information technology, communications, business strategy, and the problem domain. It takes years of education, training, and hands-on experience to become a data scientist. In addition, the process of applying machine learning requires collecting, cleansing, and consolidating data from multiple sources and selecting, training, tuning, and comparing of many machine learning algorithms. This process is both human and computer resources intensive. It is also time consuming and requires many weeks if not months to complete. In the meantime, businesses are challenged by the increasing volume, variety, and velocity of big data and the critical need to use them to inform strategic decision making and solve pressing problems quickly.

AutoML emerged as a solution to the problem of the mismatch between the supply and demand for data science skills. By automating many of the mundane repeatable tasks, professionals with diverse background and varying degree of data science skills can use AutoML platform and software to rapidly apply machine learning, dramatically shorten the lifecycle, and reduce the resources need. This research used two AutoML products to perform predictive analytics: DataRobot and Rapidminer.

Predictive Analytics using Data Robot. DataRobot is one of the leading AutoML products in the market for predictive analytics. It is a cloud-based platform and can be accessed via a web browser. No desktop software is required. The AutoML process is straightforward and only requires a few mouse clicks. After the blended data file was uploaded, the researcher selected SRR as the target for predictive analytics and initiated the automatic machine learning process. The process took several minutes to complete. What was going on behind the scene is the automatic process of feature engineering, feature selection, algorithm selection, hyper parameter tuning, cross-validation, and finally the recommendation of a best performing model.

The results showed that a total of 36 models were run and the best performing model is the Ridge Regressor, a regularized liner regression model. **Figure 50** shows the DataRobot dashboard that list all models with the recommended model at the top.

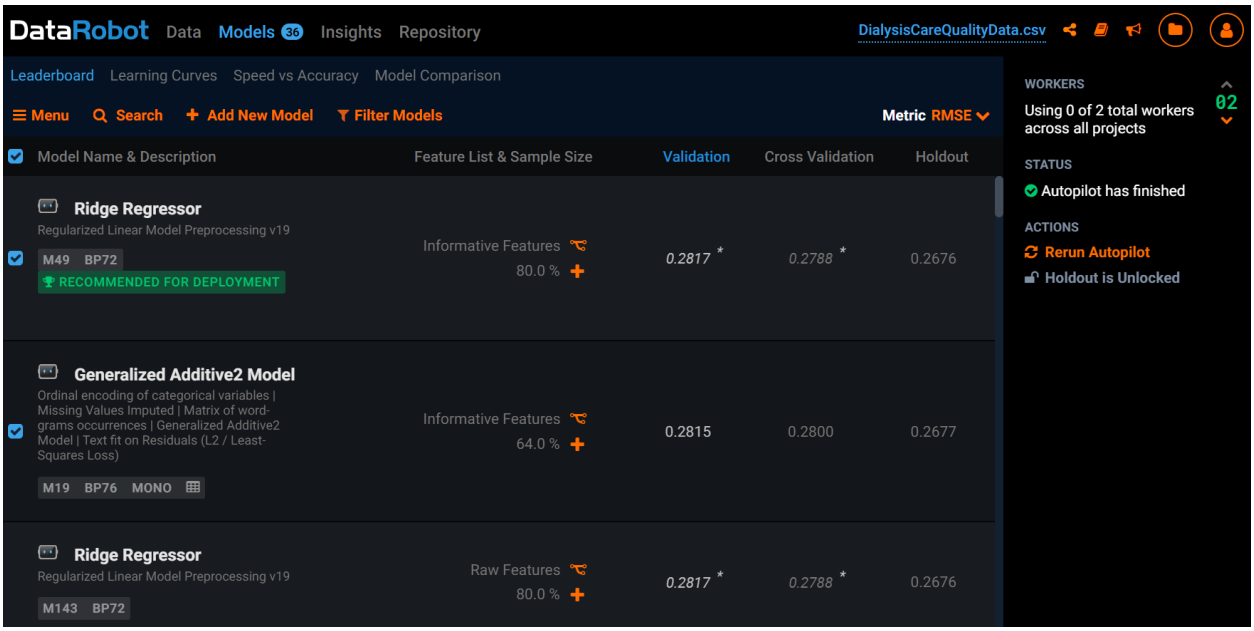


Figure 50. DataRobot dashboard shows the recommended model at top

DataRobot provides information about various aspects of the predictive analytics. **Figure 51** shows the ranked list of all features according to their impact on predicting the target. It shows that Average Kt/V and average hemoglobin have the most impact followed by staff

patient ratio and percentage of people with difficulty in English in the community where a facility is located. The plot in the middle of the figure provides a visual cue about the direction of the impact (positive vs negative) and the accuracy of prediction for a selected feature on the list to the left.

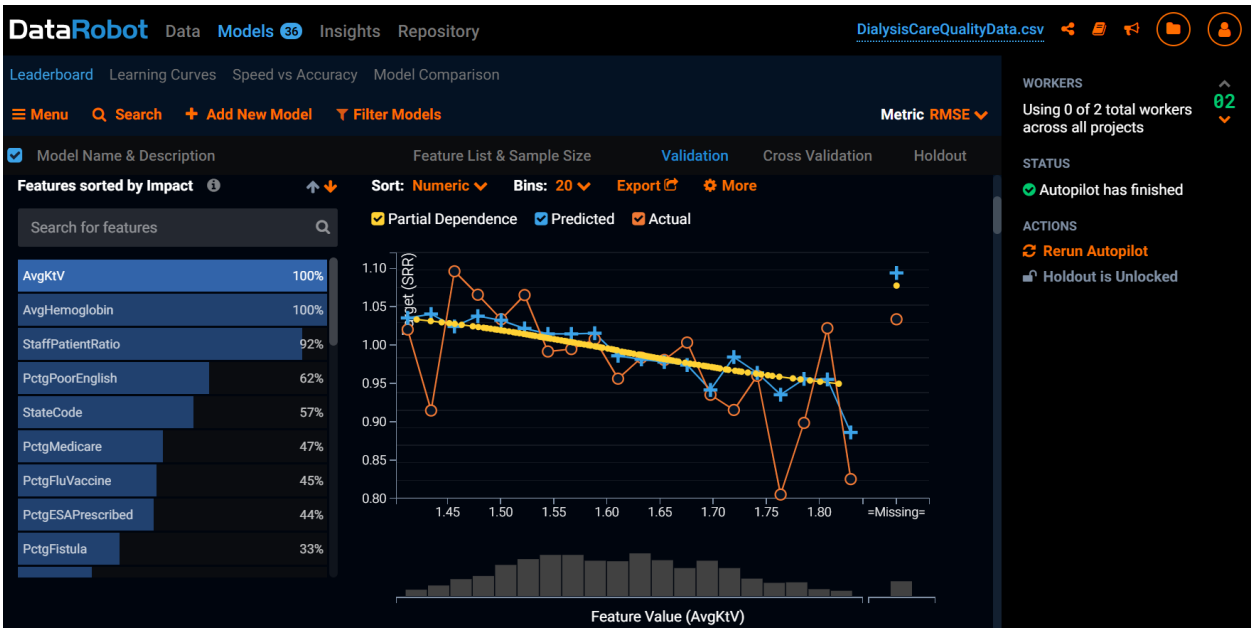


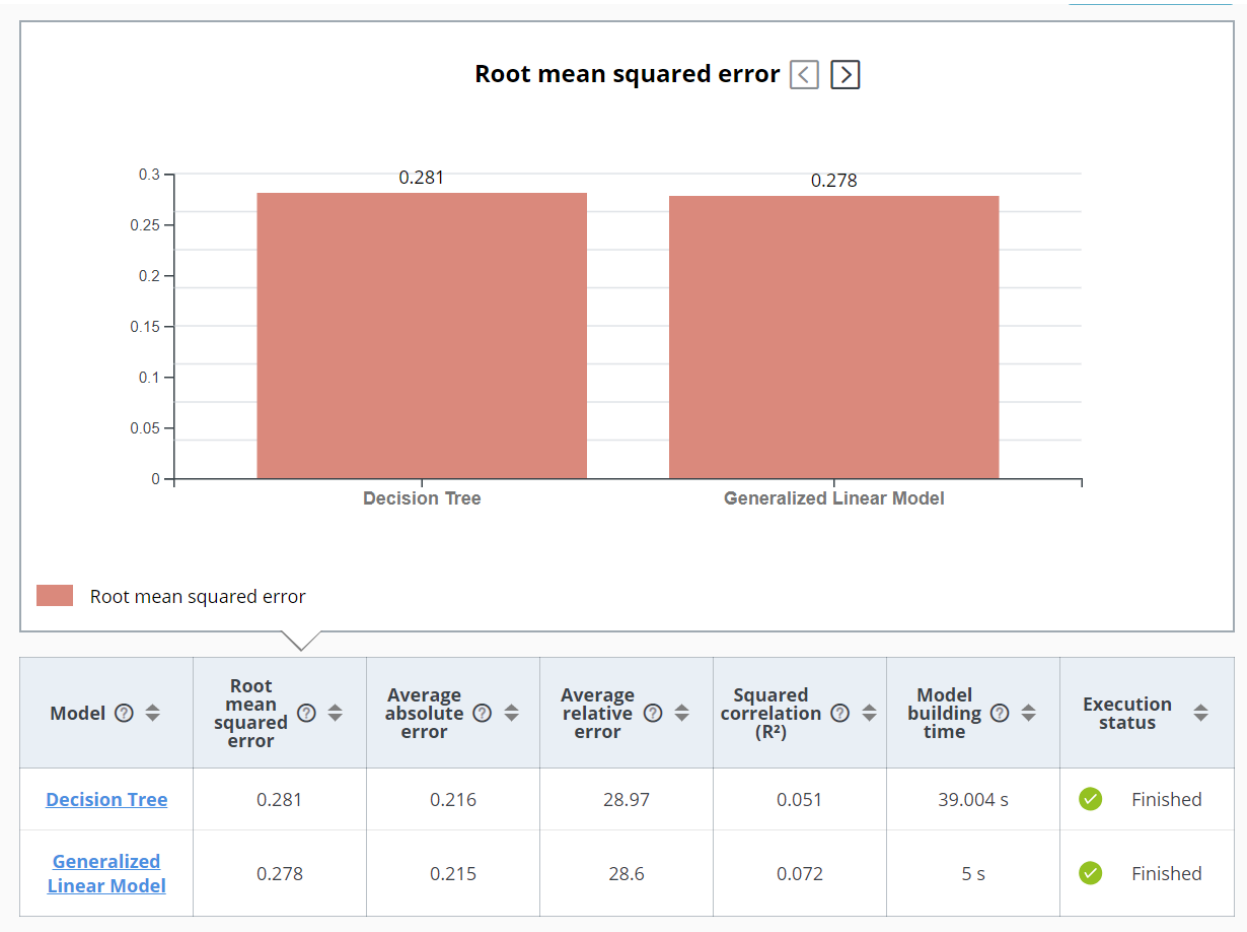
Figure 51. DataRobot dashboard shows the impact ranking of features

The importance ranking and visual analytics are available for all the models. The results vary from one model to another and can be compared to gain additional insights. Even though only one model was selected as the best performer, the performance difference among the top few models were small. In addition to showing the feature impact ranking and plot indicating the direction of impact of each feature, DataRobot also provides a list of features ranked according to its importance in tree-based models. **Figure 52** shows the ranked list of featured from the Random Forest Regressor model. It differs from the list from the Ridge Regressor model but

overall, they are similar.

Figure 52. Tree-based variable importance from the Random Forest Regressor model

Predictive Analytics using RapidMiner. RapidMiner is another leading machine learning product. It is a desktop software that provides functions for a data scientist to define and perform machine learning process. An AutoML function, called “Auto Model”, was recently developed. It is available from both the desktop version and a cloud version. Unlike DataRobot which comes with tens of algorithms, Rapid Minder currently offers a few algorithms representing various classes of supervised and unsupervised machine learning to choose from. The researcher picked Decision Tree and Generalized Linear Model (GLM). **Figure 53** shows the performance comparison between the two models.



Model	Root mean squared error	Average absolute error	Average relative error	Squared correlation (R^2)	Model building time	Execution status
Decision Tree	0.281	0.216	28.97	0.051	39.004 s	Finished
Generalized Linear Model	0.278	0.215	28.6	0.072	5 s	Finished

Figure 53. RapidMiner comparison of two models

Based on the comparison, the GLM performs better than the Decision Tree model both in terms of accuracy and speed even though the difference in accuracy is minimal. In addition, The GLM provides the ranking of the importance of the features as shown in **Figure 54**. The ranking is based on the standardized coefficient. The larger the standardized coefficient is, the more important the feature is. The sign of the coefficient indicates the direction of the association of a feature with the target. The positive sign indicates a feature has a positive association with the target and the negative sign indicates a feature has a negative association with the target. The ranking shows that the Staff Patient Ratio has the highest negative impact on the hospital readmission while the percentage of African American patients has the highest positive impact

on the hospital readmission. The feature importance ranking produced by RapidMiner GLM and DataRobot Ridge Regressor model (**Figure 51**) are different.

Generalized Linear Model Visualization:

Influence of columns

Column	Coefficient	Std. coefficient
StaffPatientRatio	-0.507	-0.027
AvgHemoglobin	-0.099	-0.025
PctgMedicare	-0.001	-0.025
AvgKtV	-0.171	-0.023
Urbanicity2.Rural	-0.023	-0.023
PctgFluVaccine	-0.001	-0.014
PctgBlack	0	0.013
PctgFistula	-0.001	-0.012
PctgESAPrescribed	0.001	0.012
PctgPoorEnglish	0.001	0.007

< 1 2 3 4 5 ••• 47 >

Figure 54. RapidMiner ranking of feature importance

RapidMiner Auto Model is still at the early stage of development and is not as user-friendly and feature-rich as DataRobot. However, its simplicity makes it easier to understand. DataRobot's user interface is more sophisticated and takes a while for a user to comprehend.

5.6 Step Four - Discover Insights

After the various interactive visual analytics were created and published to Tableau Public website, the researcher shared the link with four healthcare domain experts (expert A, B, C, and D) and scheduled a walk-through session with each one of them separately. During the walkthrough, the expert asked questions, identified patterns, provided comments, and made suggestions. The following summarizes the results of the four walkthrough sessions in two categories: Suggestions for improving the analytics and discovery of insights.

5.6.1 Suggestions for Improvement

- Expert A asked for analytics that show the distribution of facilities across the various geographical areas including regions, divisions, networks, and states. This helps the stakeholders see the overall national statistics such as which geographical area have the greatest number of facilities. The researcher made a new visualization title “Facility Distribution” and made it the first visualization as seen in **Figure 40**.
- Both expert C and D indicated that it would be helpful to show analytics on the population size in addition to number of dialysis facilities in the various geographical areas. This would require obtaining additional data elements from the Census Bureau and could be a task for future research.
- Expert A suggested identifying top three factors that influence SRR the most. This will help guide practices and policymaking. Expert A also suggested clustering facilities based on similarity in their characteristics. To address the feedback, the researcher performed predictive analytics using Automatic Machine Learning (AutoML) platform DataRobot and RapidMiner. DataRobot’s Ridge Regressor model identified Average Kt/V, Average Hemoglobin, and Staff Patient Ratio as the top three features that have

most impact on the prediction of SRR while RapidMiner's Generalized Linear Model identified the Staff Patient Ratio, Average Hemoglobin, and Percentage African American Patients as the three leading factors.

5.6.2 Key Insights Discovered

- Expert D pointed out that the "Facility Distribution" visualization showed the South has the largest number of facilities: Close to three thousand facilities which is more than 50% of all facilities nationwide. The expert explained that this may be attributed to the South having low income level, high poverty level, high sugar consumption, and high diabetes prevalence. These factors lead to high prevalence of Chronic Kidney Disease (CKD) and End-Stage Renal Disease (ESRD), and hence the need for large numbers of dialysis facilities.
- All experts discovered that facility staffing level is associated with SRR: the lower the staffing level, the higher the SRR. Expert D pointed out that nurse fatigue has been an issue in healthcare and has negative impact on care quality. This discovery aligns with existing literature on the association of nurse staffing level and patient health outcome (Griffiths et al., 2016). In the hospital setting, researchers found that higher staffing level is associated with lower readmission rate (McHugh, Berez, & Small, 2013). In dialysis care, researchers placed great emphasis on care coordination which relies very much on the effort and dedication of all patient care staff as a priority to reduce readmission (Fishbane & Wish, 2015; Mathew, Strippoli, Ruospo, & Fishbane, 2015).
- All experts discovered that for-profit facilities have higher SRR than the not-for-profit facilities. However, the visual analytics does not tell the reason why for-profit facilities perform worse than not-for-profit facilities in this measure. One possible explanation is

that for-profit facilities tend to be more cost-conscious and may allocate less resources to patient care. From previous discovery on the negative relationship of staffing level and SRR, we hypothesized that this may be due to the differences in staffing level. The researcher created a new visualization that shows how staffing level varies across between for-profit and not-for-profit facilities. As shown in **Figure 48**, the for-profit facilities have lower staffing level compared to not-for-profit facilities which leads to the higher SRR.

- Expert D was particularly interested in the difference of SRR between rural and urban communities. The visualization showed that rural communities have a lower average SRR than that of urban communities. This is counter-intuitive since rural communities tend to have less resources and in general underperform in healthcare due to various socioeconomic disparity (Warshaw, 2017). Expert D reasoned that this may be due to the limited access to hospitals in rural areas which explains the lower number of hospital admissions and readmissions. This would be a further research project for future.
- The following socioeconomic factors of the community a facility is located are positively related to the SRR. These factors include:
    - The percentage of African American
    - The percentage of families with income below federal poverty level (FPL)
    - The percentage of people age 5 and above with English difficulty
    - The unemployment rates
- The following socioeconomic factors of the patient population served by a facility are related to the SRR. These factors include:
    - The percentage of African American (positive relationship)

  - The percentage of patients with Medicare coverage (negative relationship)

- The difference in the average SRR between facilities that reuse dialyzer and facilities that don't is minimal.
- The difference in the average SRR between facilities that offer evening hour services and facilities that don't is minimal.
- Most of the clinical measures are related to SRR.

5.7 Step Five - Explore Ideas

During the meetings with the experts, in addition to discover insights from the descriptive visual analytics and the predictive analytics, the researcher and the experts also discussed ideas to apply the insights to dialysis care quality improvement. **Table 19** summarizes the key ideas to reduce unplanned hospital readmissions categorized by the quality dimensions from the Donabedian quality model.

Table 19

Ideas to Reduce Unplanned Hospital Readmission

Quality Dimension	**Insights Discovered from Visual Analytics**	**Ideas for Quality Improvement**
Structure – *Geographical Characteristics of Dialysis Facilities*	Facilities located in rural areas have lower SRR than facilities located in urban areas. This may be partially due to the difference in staffing level.	This requires further research to find out underlying causes for this disparity. Potential questions to ask include: 1) What are the other factors than the staffing level that may contribute to the lower SRR in rural facilities? 2) Was the lower SRR in part due to the lower number of hospital admissions due to the limited access to hospitals in rural area?
Structure – *Organizational*	Lower staffing level is associated with higher SRR.	**Dialysis facilities:** Facilities with lower staff patient ratio may consider increasing the

Quality Dimension	Insights Discovered from Visual Analytics	Ideas for Quality Improvement
Characteristics of Dialysis Facilities	Facilities located in a metropolitan area and for-profit facilities tend to have lower staff patient ratio.	staffing level and provider the new staff members with adequate training. For-profit facilities and facilities located in metropolitan area need to pay special attention to the risk of understaffing. **CMS Policymakers:** Consider adopting staff patient ratio as a quality measure in both the Dialysis Facility Compare (DFC) and ESRD Quality Incentive Program (QIP) program. For DFC, the first step could be to add the staff patient ratio to the website as a measure for comparison of facilities. The next step could be to include the staff patient ratio in the five-star rating.
Structure – *Socioeconomic Characteristics of Patient populations and Communities*	Facilities serving higher percentage of socioeconomically disadvantaged patient populations and facilities located in communities with higher percentage of socioeconomically disadvantaged populations have higher SRR.	**Dialysis Facilities:** All dialysis facilities need to be keenly aware of social conditions of the patients they server and pay special attention to those patients who are socioeconomically disadvantaged and have higher risks of hospital readmission. Facilities serving higher proportion of socioeconomically disadvantaged patients and facilities located in a community with higher proportion of socioeconomically disadvantaged populations need to gain better understanding of the barriers and challenges these patients and the communities are facing and find opportunities and means to remove the barriers and confront the challenges. The barriers and challenges may be in various forms such as languages and communications, foods and nutrition, transportation, and social support.

Quality Dimension	Insights Discovered from Visual Analytics	Ideas for Quality Improvement
		CMS Policymakers: Consider adding socioeconomic factors in the risk adjustment of the hospital readmission measure so that the risk adjustment of the measure accounts for social determinants of health (SDoH) and is a fair measurement of care quality for providers.
Process	The difference in average SRR between facilities that reuse dialyzer and facilities that don't is minimal.	This is consistent with the current understanding that dialyzer reuse is safe and may not increase the risk of hospital readmission. Facilities and patients can benefit from the cost savings of dialyzer reuse.
	The difference in average SRR between facilities that offer evening hour services and facilities that don't is minimal.	Providing evening hour service may not increase the risk of hospital readmission but have the benefits of convenience for ESRD patients.
	Average Kt/V and average Hemoglobin have the most impact on SRR according to the results of the automated machine learning.	Facilities should continue to monitor these two measures to ensure patients get adequate dialysis care and their anemia is sufficiently managed.

CHAPTER 6: EVALUATION OF THE A2E MODEL

6.1 Reflection on the Illustrative Scenario

In chapter 5, the researcher used a real-world scenario to illustrate how the A2E Model can be applied to a data analytics effort. This had been a journey of learning, collaboration, and discovery. At the very early stage of this effort, the researcher started out with a very broad inquiry: Exploring the factors influencing the quality of dialysis facilities in the US using the publicly available datasets. The researcher faced with two questions. First, what are the measures of dialysis care quality? and secondly, what are the factors that influence dialysis care quality? Through consultation with a healthcare quality expert, the researcher was able to gain knowledge about a well-known Healthcare Quality Model by Donabedian (1966). The Structure-Process-Outcome construct of Donabedian Model provided a clear framework for selecting both the quality outcomes and influential factors. The researcher was able to use this model as a guide to examine the data elements in the various data sources and select only the ones that are relevant. The initial examination of the data sources revealed that there were many outcome measures such as mortality, hospital admission, unplanned hospital readmission (readmission within 30-day of discharge), patient satisfaction, etc. For an illustrative scenario, it would be too complex to deal with many outcome measures at once. A data analytics project typically focuses on one target or dependent variable at a time. The researcher consulted another healthcare expert and was advised to focus on the hospital readmission as the target measure since it is an important outcome measure that has both financial impact on the government and quality of life impact on the patients. This clearly demonstrated the importance of engaging domain experts at the beginning of data analytics in which both the business need and the scope of work can be discussed and specified. This also demonstrated that the need to perform a preliminary

examination of the data sources to make sure the data are there to support the effort. During this process, the researcher also spent time evaluating various programming languages and software tools including R, Python, Tableau, and Jupyter Notebooks. Various development platforms were also evaluated including Microsoft Azure Notebooks, IBM Data Science Experience, and Kaggle. The lesson learned is that the first step, "Assess Needs", is not only a key step and sets the stage for the future steps but is also a mini project in and of itself. The mini project include preliminary descriptive and visual analytics and requires a close collaboration of both technical professionals and domain experts.

The second step, "Blend Data", was the most time-consuming step requiring intensive effort in examining, cleansing, and merging data from different sources. This step involves many trials and errors and is iterative and incremental. The researcher has a strong background in software engineering and has benefited from the disciplined approach to data management and programing. The researcher set the goal of transparency and reproducibility at the very beginning to make sure the process of blending data is well documented and can be repeated by others. The researcher first used a local development environment on a laptop computer to develop the Python programs for blending the data sources and later migrated both the data sources and Python programs to Kaggle to make them live in the cloud for easy access and sharing with others. The researcher decided to use Jupyter Notebooks for Python so that the documentation, the source code, and the outputs including the textual, tabular, and visual outputs are in one single place for easier access, understanding, and replicated validation. During this step, the researcher consulted domain experts occasionally when questions arose about the meaning of a data element or the relationship among data elements. Other than occasional consultation with domain experts, this process had mostly been an independent effort. In a data analytics project,

this effort could be performed by a separate data management team. In a large organization with complex IT infrastructure such as multiple databases and data warehouses, the effort would also require the support from the IT department. In addition to coding in Python or other programing languages such as R or SQL for data blending, there are also software tools that offer drag and drop functionality. These tools are typically categorized as Extract-Transform-Load (ETL) or Data Preparation tools and could be evaluated as an aspect of future research related to the step two, “Blend Data”.

Compared to the second step, “Blend Data”, the third step, “Create Analytics” is more straightforward and less tedious and time-consuming. The researcher first tried to use various Python visualization libraries to get familiar with data visualization techniques. Later, the researcher tried Tableau, the leading data visualization tool, and decided to go with Tableau for its ease of use and the ability to publish the visual analytics in the Tableau Public website for sharing with domain experts. This proved to be a good decision since during the step four and five, the availability of these visualizations on the web made it very easy and convenient for sharing and discussing with healthcare experts. The original plan was to only perform visual analytics for the purpose of illustration. However, additional predictive analytics were conducted to make the illustration more comprehensive and stronger. DataRobot and RapidMiner, two of the leading cloud-based automatic machine learning (AutoML) platforms were used. AutoML is on the rise and has the potential to make machine learning more accessible and productive. The researcher uploaded the blended final dataset to DataRobot platform, it automatically performed feature engineering, algorithms selection, hyperparameters tuning, cross-validation, model performance comparison, and finally recommended the best model based on accuracy and speed. DataRobot automated the most challenging, yet mundane manual work typically performed by

well-trained data scientists. RapidMiner Auto Model works in a similar fashion. This further demonstrated the validity of the A2E Model that stresses on the people and organizational aspect of data analytics since technology is just an enabler but not a driver. As technology matures and becomes more automated, more effort will be shifted to the non-technical human and organizational aspects such as communications, collaborations, business strategies, and organizational decision making and performance evaluation.

After the visual analytics and predictive analytics were developed, the researcher shared the results with several healthcare experts as part of the step four, “Discover Insights”. The experts identified patterns from the visual analytics and attempted to make sense of them. They used their professional experience and knowledge to interpret the results. It is interesting to see how each expert brought different perspective. One expert has many years of experience as a nurse and knows a lot about the ESRD patients and connected the analytics with patients’ experience. Another expert has experience in rural area healthcare issues and policies and provided good insights on the disparity of healthcare between rural and urban communities. This speaks to the need for incorporating different perspectives and the importance of multidisciplinary collaboration.

The step five, “Explore Ideas”, was more of follow-up with the step four, “Discover Insights”. During the discussion with healthcare experts, insight discovery was typically followed by exploration of ideas about how to use the insights to inform decision. This seemed to be a natural progression. During the discussion, the experts also asked more questions some of which required future work to incorporate additional data elements and perform additional analytics. This demonstrated the iterative and incremental nature of the data analytics process as a continuous exploratory process. Given the amount of data and the complexity of the domain,

the data analytics process does not stop at any point in time. It is an ongoing process and can keep going as more and more data become available.

6.2 Expert Evaluation of the A2E Model

Peffers et al. (2012) defined expert evaluation as the "assessment of an artifact by one or more experts (e.g., Delphi study)" (p. 5). While a full-fledged expert evaluation such as Delphi study is not in scope for this research and could be a potential future research project, the researcher did request review and feedback from two experts. Expert A is the Research Director for a competitive intelligence consulting and software firm and a former human intelligence professional. Expert B is the Senior Manager of Market Research for a global biopharmaceutical company responsible for market research and data analysis to support decision making and drive business growth.

The feedback from both experts are positive and encouraging. Both thought the A2E Model is well designed, easy to follow, and reflects their experience in data analytics effort and their understanding of the data analytics process.

Expert A drew upon his experience in competitive intelligence, business intelligence, and knowledge management and believed that the A2E Model "would have utility for helping a variety of organizational shareholders to understand what needs to be done, why, by whom, and at what cost, as part of a data analytics undertaking". He further commented that the A2E Model is "simple without being simplistic", a highly valued trait of any good model.

Expert B commented that "In my real-world environment, we don't have a well specified process to stick to, but in general we follow the 5 steps in market research and analysis". Expert B further elaborated that "the model should be useful for analytical teams. It should especially be impactful in an environment that does not have clear process for analysis". Expert B liked the

way the model is specified using five themes and five steps. He commented that "theme 5 (one caution) is a critical point because very often more time and efforts are spent on data cleansing to ensure quality of the analysis".

Both expert A and B identified areas for improvement and provided valuable suggestions.

Expert A pointed out three areas that may need improvement. First, there is a need to emphasize on "information requirements assessment" and perhaps it should precede the first step "Assess Needs" because relevant information is the prerequisite for the assessment of business needs. However, he also recognized that first step "Assess Needs" encompasses the assessment of information requirements. Secondly, expert A suggested adding "Identify appropriate data sources" and "develop a data acquisition plan" as subtasks or activities under "Assess Data Needs". He further commented that:

> *It's not enough to know what data you need/desire; you also need to identify where that data resides (there may be multiple places), and then identify how that data will be acquired, by whom, and what the associated resource costs will be (resource, time, opportunity).*

Expert A's emphasis on information requirements and data acquisition is consistent with the experience of data analytics practitioners as they often report that most of their time is spent on data collection, understanding, and cleansing. Thirdly, expert A believed that the "cross-check with business needs" under step five "Explore Ideas" should be applied to other steps as well. He used navigating terrain using a map and a compass as a metaphor:

> *It is important to periodically stop and verify one's location to ensure one is not straying off course. Being off course by one degree matters little over 50 meters but matters progressively more the farther one travels. Therefore periodic "compass checks" are*

helpful, especially given that information requirements may sometimes change over time, and changes in the data acquisition/analysis efforts may need to change, too.

The Star View, one of the four views in the A2E Model (see Figure 23 on page 73) placed the first step "Assess Need" at the center with connections to the other four steps. This view is consistent with the expert A's opinion.

Expert B drew upon his experience in managing analytics projects to emphasize the importance of communication and felt that the role of communication could be elaborated more in the A2E Model. He commented:

"In my experience, one critical component for success is communication. Before data collection and analytics start, business questions should be clearly defined; during the project, regular touch points are needed to ensure data and analytics results make sense; in the end, results need to be clearly presented to ensure business question are addressed."

These suggestions are valuable and will be incorporated in the future revision of the A2E Model.

6.3 Summary

This research followed Design Science Research methodology to design a novel process model for data analytics to address the limitations of the existing models and to help increase the quality and effectiveness of data analytics efforts. A process model helps guide the planning and execution of any human endeavor. Much like a software development process model guiding the effort of developing a software system to meet its quality standards, schedule, and budgetary constraints, a data analytics process model aims to guide the effort of data analytics to ensure that it delivers quality outcomes and achieves valuable impacts within schedule and budget.

Design Science calls for both relevance and rigor. To address the aspect of relevance, the researcher was motivated by the problem of low adoption rate of existing data analytics process models due to their limitations and the poor outcome and low impact achieved from data

analytics efforts in the industry. As the volume, velocity, and variety of data increase exponentially due to the proliferation of emerging technologies including Internet of Things, mobile and cloud computing, and social media networks, the demand for extracting actionable insights from data to inform decision making is increasing exponentially as well. A better process model is much needed to help improve the efficiency and effectiveness of data analytics effort. To address the aspect of rigor, the researcher surveyed the landscape of existing data analytics process models from both industry and academia with focus on the three leading ones and their extensions. In addition, the new and improved process model for data analytics, the A2E Model, was designed based on strong theoretical foundation including the wisdom pyramid, the semiotic ladder, the information and communication theory, the systems engineering model, and the human-centered design principles.

Design an artifact is only half the story in Design Science Research. Evaluation of an artifact for its utility and efficacy is the other half and is equally important. While there are multiple evaluation methods, this research project adopted the illustrative scenario method using a real-world healthcare quality improvement scenario. The researcher illustrated how the five steps of the A2E Model were followed and described the activities and outcomes of each of the steps. The scenario demonstrated the importance of collaboration between the technical and business professionals. The scenario also demonstrated the concept of human-machine teaming and the machine-machine teaming. Both descriptive and predictive analytics were performed using data visualization and machine learning. In addition to the illustrative scenario, the researcher also solicited expert reviews of the A2E Model and received positive feedback along with suggestions for improvement.

The contributions of this research are three-fold. First, the A2E Model provides a theoretical framework to the understanding and knowledge of the data analytics process. In contrast to the existing process models which are technology-oriented and lack the emphasis on organizational context and human factors, the A2E Model integrates organization, people, and technology in a holistic and balanced view. Secondly, the A2E model provides a shared mental model to guide both technical and business professionals and their diverse stakeholders in the data analytics effort. Thirdly, the A2E Model provides a pedagogical framework to guide the curriculum design of balanced multidisciplinary data analytics programs.

REFERENCES

Abbasi, A., Sarker, S., & Chiang, R. H. (2016). Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems, 17*(2).

Ahangama, S., & Poo, D. C. C. (2015). *Designing a process model for health analytic projects.* Paper presented at the PACIS, Singapore, Singapore.

Azevedo, A., & Santos, M. F. (2008). *KDD, SEMMA and CRISP-DM: A parallel overview.* Paper presented at the IADIS European Conference Data Mining.

Barr, D. A. (2014). *Health disparities in the United States: Social class, race, ethnicity, and health*: JHU Press.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association, 71*(356), 791-799.

Cao, L. (2010). Domain-driven data mining: Challenges and prospects. *IEEE Transactions on knowledge and data engineering, 22*(6), 755-769.

Centers for Medicare and Medicaid Services. (n.d.-a). Dialysis Facility Compare. Retrieved from https://www.medicare.gov/dialysisfacilitycompare

Centers for Medicare and Medicaid Services. (n.d.-b). ESRD Quality Incentive Program. Retrieved from https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/ESRDQIP/index.html

Centers for Medicare and Medicaid Services. (n.d.-c). *Pub 100-07 State Operations Manual*. Retrieved from https://www.cms.gov/Regulations-and-Guidance/Guidance/Manuals/Downloads/som107c02.pdf.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide* Retrieved from https://www.the-modeling-agency.com/crisp-dm.pdf

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.): SAGE Publications, Inc.

Davenport, T. H., Harris, J. G., & Morison, R. (2010). *Analytics at work: Smarter decisions, better results*: Harvard Business Press.

Donabedian, A. (1966). Evaluating the Quality of Medical Care. *The Milbank Memorial Fund Quarterly, 44*(3), 166-206. doi:10.2307/3348969

Dun and Bradstreet, & Forbes. (2017). *Analytics accelerate into mainstream*. Retrieved from Dun & Bradstreet: http://www.dnb.com/perspectives/analytics/analytics-accelerates-into-the-mainstream-report.html

Farooq, U., & Grudin, J. (2016). Human-computer integration. *ACM Interactions, 23*(6), 26-32.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Fishbane, S., & Wish, J. B. (2015). Quality Measurement in Wonderland: The Curious Case of a Dialysis Readmissions Measure. *Clinical Journal of the American Society of Nephrology*, CJN. 02770315.

Forsberg, K., Mooz, H., & Cotterman, H. (2005). *Visualizing project management: Models and frameworks for mastering complex systems* (3rd ed.). Hoboken, N.J.: J. Wiley.

Gartner Inc. (2013). *Gartner predicts business intelligence and analytics will remain top focus for CIOs through 2017*. Retrieved from http://www.gartner.com/newsroom/id/2637615

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS quarterly, 37*(2).

Griffiths, P., Ball, J., Drennan, J., Dall'Ora, C., Jones, J., Maruotti, A., . . . Simon, M. (2016). Nurse staffing and patient outcomes: Strengths and limitations of the evidence to inform policy and practice. A review and discussion paper based on evidence reviewed for the National Institute for Health and Care Excellence Safe Staffing guideline development. *International Journal of Nursing Studies, 63*, 213-225. doi:https://doi.org/10.1016/j.ijnurstu.2016.03.012

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems, 19*(2), 4.

Hevner, A. R., & Chatterjee, S. (2010). *Design research in information systems: Theory and practice*. New York, NY: Springer.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly, 28*(1), 75-105.

Hindle, G. A., & Vidgen, R. (2017). Developing a business analytics methodology: A case study in the foodbank sector. *European Journal of Operational Research*. doi:http://dx.doi.org/10.1016/j.ejor.2017.06.031

IBM. (2017). *IBM SPSS Modeler CRISP-DM Guide* Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.1/en/ModelerCRISPDM.pdf

International Organization for Standardization. (2010). 9241-210 2010. Ergonomics of human system interaction Part 210: Human-centred design for interactive systems.

Khabaza, T. (2010). Nine laws of data mining. Retrieved from http://khabaza.codimension.net/index_files/9laws.htm

Khan, D. M., Mohamudally, N., & Babajee, D. K. R. (2013). A unified theoretical framework for data mining. *Procedia Computer Science, 17*(Supplement C), 104-113. doi:https://doi.org/10.1016/j.procs.2013.05.015

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*: Princeton university press.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., . . . Corlay, S. (2016). *Jupyter Notebooks-a publishing format for reproducible computational workflows.* Paper presented at the ELPUB.

Laozi, Mitchell, S., Roig, J. V., & Little, S. (1989). *Tao te ching*: Kyle Cathie.

Leonhard, G. (2016). *Technology vs humanity: The coming clash between man and machine*. Kent, UK: Fast Future Publishing

Li, Y., Thomas, M. A., & Osei-Bryson, K.-M. (2016). A snail shell process model for knowledge discovery via data analytics. *Decision Support Systems, 91*, 1-12.

Licklider, J. C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics*(1), 4-11.

Maaskant, M. (2016). *Data mining approaches for calculating the energy consumption of buildings.* (Master Thesis), Eindhoven University of Technology.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251-266.

Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review, 25*(2), 137-166.

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet, 365*(9464), 1099-1104.

Martins, S., Pesado, P., & García-Martínez, R. (2016). *Information mining projects management process.* Paper presented at the SEKE.

Mathew, A. T., Strippoli, G. F., Ruospo, M., & Fishbane, S. (2015). Reducing hospital readmissions in patients with end-stage kidney disease. *Kidney international, 88*(6), 1250-1260.

McGovern, L., Miller, G., & Hughes-Cromwick, P. (2014). *The relative contribution of multiple determinants to health outcomes*: Project HOPE.

McHugh, M. D., Berez, J., & Small, D. S. (2013). Hospitals with higher nurse staffing had lower odds of readmissions penalties than hospitals with lower staffing. *Health affairs, 32*(10), 1740-1747.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review, 63*(2), 81.

Miškuf, M., Michalik, P., & Zolotová, I. (2017). *Data mining in cloud usage data with Matlab's statistics and machine learning toolbox.* Paper presented at the IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), Herl'any, Slovakia.

Nicholas, S. B., Kalantar-Zadeh, K., & Norris, K. C. (2015). Socioeconomic Disparities in Chronic Kidney Disease. *Advances in chronic kidney disease, 22*(1), 6-15. doi:10.1053/j.ackd.2014.07.002

Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. New York, NY: Basic Books.

Orszag, P. R. (2008). *The overuse, underuse, and misuse of health care*.

Peffers, K., Rothenberger, M., Tuunanen, T., & Vaezi, R. (2012). *Design science research evaluation.* Paper presented at the Seventh International Conference on Design Science Research in Information Systems and Technology, Las Vegas, NV, USA.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45-77.

Piatetsky-Shapiro, G. (2000). Knowledge discovery in databases: 10 years after. *ACM SIGKDD Explorations Newsletter, 1*(2), 59-61.

Piatetsky-Shapiro, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved from http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pramanik, M., Lau, R. Y., Yue, W. T., Ye, Y., & Li, C. (2017). Big data analytics for security and criminal investigations. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7*(4).

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2014). *Artifact evaluation in information systems design science research - a holistic view.* Paper presented at the PACIS.

Prat, N., Comyn-Wattiau, I., & Akoka, J. (2015). A taxonomy of evaluation methods for information systems artifacts. *Journal of Management Information Systems, 32*(3), 229-267.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media, Inc.

Reid, C., et al. (2015). *Seizing the information advantage: How organizations can unlock value and insight from the information they hold*. Retrieved from http://www.ironmountain.com/Knowledge-Center/Reference-Library/View-by-Document-Type/Landing-Pages/P/PWC.aspx

Saran, R., Robinson, B., Abbott, K. C., Agodoa, L. Y., Albertus, P., Ayanian, J., . . . Chen, J. L. (2018). US renal data system 2018 annual data report: epidemiology of kidney disease in the United States. *American Journal of Kidney Diseases, 71*(3).

SAS Institute Inc. (2017). *Getting started with SAS®Enterprise Miner™ 14.3* Retrieved from http://documentation.sas.com/api/docsets/emgsj/14.3/content/emgsj.pdf?locale=en#nameddest=n0gurzayfolktun1e8vl6n7qsi88

Schramm, W. (1954). How communication works. *The process and effects of mass communication*, 3-26.

Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research, 12*(1), 217-222.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Sharda, R., Delen, D., & Turban, E. (2016). *Business intelligence, analytics, and data Science: A managerial perspective*. New York, NY: Pearson.

Sharma, S., & Osei-Bryson, K.-M. (2010). Toward an integrated knowledge discovery and data mining process model. *The Knowledge Engineering Review, 25*(1), 49-67.

Simon, H. A. (1996). *The sciences of the artificial*. Cambridge, MA: MIT press.

Sinek, S. (2011). *Start with why: How great leaders inspire everyone to take action*. New York, NY: Penguin Group.

Stamper, R. (1993). A semiotic theory of information and information systems.

Stone, J. V. (2015). *Information theory: A yutorial introduction*: Sebtel Press.

Tahmasebian, S., Ghazisaeedi, M., Langarizadeh, M., Mokhtaran, M., Mahdavi-Mazdeh, M., & Javadian, P. (2017). Applying data mining techniques to determine important parameters in chronic kidney disease and the relations of these parameters to each other. *Journal of renal injury prevention, 6*(2), 83.

Toffler, A. (1970). *Future shock*. New York: Bantam Books.

Toffler, A. (1980). *The third wave* (1st ed.). New York: William Morrow & Company, Inc.

University of Michigan Kidney Epidemiology and Cost Center. (n.d.). Dialysis Facility Report. Retrieved from https://www.dialysisdata.org/content/dialysis-facility-report-methodology

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). *A comprehensive framework for evaluation in design science research.* Paper presented at the International Conference on Design Science Research in Information Systems.

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77-89.

Warshaw, R. (2017). Health disparities affect millions in rural U.S. communities. Retrieved from https://news.aamc.org/patient-care/article/health-disparities-affect-millions-rural-us-commun/

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Heidelberg, NY: Springer.

Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining.* Paper presented at the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making, 5*(04), 597-604.

Zhang, Y. (2016). The association between dialysis facility quality and facility characteristics, neighborhood demographics, and region. *American Journal of Medical Quality, 31*(4), 358-363.