# Beyond Technology:
## An Integrative Process Model for Data Analytics

**Chaojie Wang**
https://orcid.org/0000-0001-8521-9420
*The MITRE Corporation, USA*

## INTRODUCTION

Over the years, many different terms have been used to describe the process and activities that extract information and discover knowledge from data to enable evidence-based decision making and problem solving. For example, data mining, knowledge discovery in databases, business intelligence, business analytics, data analytics, big data analytics, and data science are among the most popular ones. For simplicity and consistency, in this chapter we use data analytics as a generic umbrella term to cover a myriad of activities that use statistical models, machine learning algorithms, and software platforms and tools to uncover patterns, discover knowledge, and inform decision making through systematic acquisition, preparation, analysis, and presentation of data and information.

It is well understood that software engineering efforts require the adoption of a development process model, such as the Waterfall, the Agile, or a hybrid of both. Since the core of data analytics is software, it can greatly benefit from the adoption of a process model to improve its efficiency and effectiveness and achieve meaningful and impactful outcomes. A process model serves as a shared mental model for a team and improves the communication and collaboration and increases the chances for success. However, according to surveys conducted by KDNuggets.com about 40% of data scientists and business analysts don't use any process model even though many are available (Piatetsky-Shapiro, 2014). This low adoption rate can be attributed to two main limitations in the existing models. Firstly, they were derived mainly from practical experiences and lack strong theoretical basis. Secondly, they were developed by and for technical professionals and do not incorporate sufficient consideration for human factors and organizational contexts.

This chapter introduces an integrative data analytics process model grounded on human-centered design principles and industry best practices, the A2E Process Model for Data Analytics. It was created and evaluated using design science methodology as part of a doctoral dissertation project (Wang, 2019). The author reviewed the three leading process models for data analytics along with various efforts by prior researchers and practitioners to improve and extend them (Wang, 2019, p. 25 - 37). "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" (Wang, 2019, p. 52). Recognizing this gap, the author applied the human-centered design principles in the development of the A2E Model, A2E is an acronym for A, B, C, D, and E, representing five steps in data analytics process: Assess Needs, Blend Data, Create Analytics, Discover Insights, and Explore Ideas. This model was evaluated through a real-world case study in a healthcare quality improvement data analytics effort to demonstrate its utility and efficacy (Wang, 2019, p. 83 – 131). In addition, the model was reviewed by subject matter experts for

DOI: 10.4018/978-1-7998-9220-5.ch067

relevancy and quality (Wang, 2019, p.136 – 140). Both the case study and the expert review confirmed that the model is effective in helping data scientists and domain experts communicate and collaborate to achieve higher quality and deliver greater impacts for their data analytics efforts.

This chapter provides detailed description of the A2E Model including its theoretical foundation and step-by-step descriptions and guidelines. The goal of this chapter is to provide a concise, easy to follow guidance for data scientists and domain experts to apply this novel process model in their day-to-day analytics effort.

## BACKGROUND

Data analytics is a key to unlock the untapped power of knowledge hidden in the trenches of big data in the age of information and intelligence. 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 that facilitates multidisciplinary collaborations and balance technical efficiency with organizational effectiveness. Despite the abundance of existing analytics process models, many analytics professionals and project teams choose to use their own, or not to use any at all (Piatetsky-Shapiro, 2014). 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.

The abundance and rapid growth of digital data and the increasing demand for analytics bring forth opportunities as well as challenges. To ensure consistency, repeatability, quality, and maturity of data analytics, to reduce the risk of project failure, and to improve the outcomes and impacts, the data analytics community has used a variety of process models and best practices over the years. Leading among them are the Cross Industry Standard Process for Data Mining (CRISP-DM), the Knowledge Discovery in Databases (KDD), and the Sample-Explore-Modify-Model-Assess (SEMMA). **Table 1** summarizes these three leading data analytics process models.

*Table 1. The three leading data analytics process models*

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

According to surveys conducted in 2007 and 2014 by KDNuggets.com, a leading online community for data analytics professionals, about 60% of professionals surveyed used one of the above three methodologies and the remaining 40% used either a proprietary process model or did not use any model at all. There had been little change in the adoption rate over the seven-year span between 2007 and 2014 (Piatetsky-Shapiro, 2014).