Using Systems Analysis and Design to Enhance the Business Understanding Stage in CRISP-DM

James J. Pomykalski
pomykalski@susqu.edu
Sigmund Weis School of Business
Susquehanna University
Selinsgrove, PA 17870, USA

Jan Buzydlowski
jbuzydlowski@holyfamily.edu
School of Business
Holy Family University
Philadelphia, PA

Abstract

The long history of Systems Analysis and Design (SA&D) has helped to develop a number of standard activities and models that assist in the development of information systems. This paper continues the work on the application of SA&D activities and models to the CRoss Industry Standard Process for Data Mining (CRISP-DM) process. The focus of this paper is the application of the activities and models from Systems Planning stage in SA&D to the Business Understanding stage in CRISP-DM.

Keywords: Analytics, CRISP-DM, Systems Analysis and Design, Planning

1. INTRODUCTION

Systems Analysis and Design (SA&D) is a development methodology (process), first described in the 1970's, that consists of narrative and graphical models used to plan, analyze and design information system solutions for the improvement of business processes (Whitten & Bentley, 2005; Kock, 2007). There are five major stages in the SA&D methodology planning, analysis, design, implementation, and maintenance (PADIM).

The CRoss-Industry Standard Process for Data Mining (CRISP-DM) methodology was originally developed in 1996 "based on the practical, real-word experience of how people conduct data-mining projects" (Chapman, et al., 2000, p. 3). The CRISP-DM methodology consists of six stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

The methodologies for both of these problem-solving approaches have been developed (independently) to assist traditional information systems development (Hoffer, George, & Valacich, 2013; Valacich & George, 2014; Whitten & Bentley, 2005) and data mining solution development (Chapman, et al., 2000; Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

This particular paper is the second in an on-going research effort to investigate the applicability of the narrative and the graphical models used in SA&D to aid the data-mining/analytics efforts now being used throughout many industries; the first paper is (Buzydlowski & Pomykalski, 2016). This paper focuses on the application of models used in the planning stage of the SA&D methodology to the business understanding phase of the CRISP-DM methodology. The goal is to demonstrate the applicability of SA&D models in the planning stage to the work undertaken in

the business understanding stage in the data mining development process (CRISP-DM).

This paper is organized as follows. Section two reviews the related research, including the competing models for developing data mining solutions. Section three applies the specific models of SA&D to the tasks in the business understanding stages of CRISP-DM. An on-going data analytics effort at Susquehanna University is used to illustrate the use/applicability of some of the models in the SA&D planning stage. The fourth section discusses the non-conforming aspects of CRISP-DM with SA&D. These aspects include culture, project teams, and agile project management methods. The final two sections look a future work including the impact on the education of future IS students, and the conclusions drawn from this work.

2. RELATED LITERATURE

The use of systems analysis and design methodologies, and their models, have been addressed in numerous texts and papers that describe the narrative and graphical models in the SA&D methodology (Dennis & Wixom, 2014; Whitten & Bentley, 2005; Hoffer, George, & Valacich, 2013; Valacich & George, 2014).

However, the work on process methodologies for performing data analytics are relatively young and not as thoroughly examined. For example, the CRISP-DM methodology was developed in 2000 and, at present, the most comprehensive work on this methodology is by Chapman et al. (2000). There are two other competing process methodologies for data analytics projects: Sample, Explore, Modify, Model, Assess (SEMMA) and Knowledge Discovery in Databases (KDD). Currently, only two short conference papers exist (Azevedo & Santos, 2008; Safique & Qaiser, 2014) that compare these process methodologies.

The KDD approach, which is a macro level view of the analytics project, was the first developed methodology, in 1996, to address "the need to scale up human analysis capabilities to handling the number of bytes [of data]" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) that is collected; as a macro level approach, the KDD methodology looks at analytics as a single stage of a larger knowledge discovery process and therefore does address the business problem in the initial stage.

The SEMMA method, developed by SAS Institute to support model development in the SAS Enterprise Modeler software, primarily focuses exclusively on the analytic process and ignores many of the business aspects involved in most analytics projects.

The primary works on the CRISP-DM methodology were produced in 2000. Chapman, et al. (2000) describes the four levels of the CRISP-DM methodology (1) phases, (2) generic tasks that occur in most data mining situations, (3) specialized tasks that describe how the generic task are carried out in specific situations, and (4) process instances that record the actions, decisions, and results of the data mining effort. A second paper (Wirth & Hipp, 2000) describes the use of generic CRISP-DM process model in the planning, communication within and outside the project team, and in documentation. The paper reports on the experience with the model in practice.

The three competing methodologies were compared in previous papers (Safique & Qaiser, 2014) (Azevedo & Santos, 2008). Buzydlowski and Pomykalski (2016) extended this work and broadly applied the activities of the five systems analysis stages to the stages of these three data analytic methodologies. The primary outcome of that paper was the mapping of the five stages of SA&D onto the six stages of the CRISP-DM methodology. This mapping of the activities and models used in the initial of stages of both methodologies is the focus of this paper.

3. APPLICATION AND USE OF SA&D MODELS

In the systems planning stage, the most relevant models to the CRISP-DM business understanding stage are the Systems Service Request (SSR), Baseline Project Plan (BPP), Business Case, and the Statement of Work (SoW) (Whitten & Bentley, 2005; Valacich & George, 2014). To illustrate this applicability a student retention study, currently underway at Susquehanna University (SU), will be used.

The problem that SU sought to address was a decline in the retention rates of first year students in the period from 2006 to the present. Prior to 2006 SU enjoyed a retention rate of nearly 90% of first year students to the second year, however, since 2006 this rate has been declining and reached a low point of 82% in 2010. In 2015 SU began a concerted, structured effort to find the root causes of this drop in retention and to begin to develop new initiatives to raise the retention of the first year students.

While the four SA&D models are outcomes of the SA&D process, their development adds to the project understanding by both the user

community and the development teams both in terms of clarity of project objectives and the potential use of various analytical models. This point has been examined by Gass (1984) when discussing the need for documenting modeling efforts and again, by Gass (1993) when discussing the accreditation and validation of models. The simulation field has also dealt with issues of model understanding for years (Robinson, 2004).

The four major tasks that are undertaken in the CRISP-DM business understanding phase (Chapman, et al., 2000) are: determining the business objectives, assessing the situation, describing the data mining goals, and producing the project plan (see Appendix 1).

BASELINE PROJECT PLAN REPORT

1.0 Introduction
- A. Project Overview—Provides an executive summary that specifies the project's scope, feasibility, justification, resource requirements, and schedules. Additionally, a brief statement of the problem, the environment in which the system is to be implemented, and constraints that affect the project are provided.
- B. Recommendation—Provides a summary of important findings from the planning process and recommendations for subsequent activities.

2.0 System Description
- A. Alternatives—Provides a brief presentation of alternative system configurations.
- B. System Description—Provides a description of the selected configuration and a narrative of input information, tasks performed, and resultant information.

3.0 Feasibility Assessment
- A. Economic Analysis—Provides an economic justification for the system using cost-benefit analysis.
- B. Technical Analysis—Provides a discussion of relevant technical risk factors and an overall risk rating of the project.
- C. Operational Analysis—Provides an analysis of how the proposed system solves business problems or takes advantage of business opportunities in addition to an assessment of how current day-to-day activities will be changed by the system.
- D. Legal and Contractual Analysis—Provides a description of any legal or contractual risks related to the project (e.g., copyright or nondisclosure issues, data capture or transferring, and so on).
- E. Political Analysis—Provides a description of how key stakeholders within the organization view the proposed system.
- F. Schedules, Time Line, and Resource Analysis—Provides a description of potential time frame and completion date scenarios using various resource allocation schemes.

4.0 Management Issues
- A. Team Configuration and Management—Provides a description of the team member roles and reporting relationships.
- B. Communication Plan—Provides a description of the communication procedures to be followed by management, team members, and the customer.
- C. Project Standards and Procedures—Provides a description of how deliverables will be evaluated and accepted by the customer.
- D. Other Project-Specific Topics—Provides a description of any other relevant issues related to the project uncovered during planning.

Figure 1: Outline of Baseline Project Plan (Hoffer, George, & Valacich, 2013, p. 132)

In business understanding, the first task is to thoroughly understand, from a business perspective, what the client wants to accomplish. In SA&D, this is accomplished in the development of the baseline project plan. The baseline project plan (BPP) consists of four major sections: introduction, the system description, feasibility, and management issues (see Figure 1). The system description is developed, in conjunction with the business stakeholders, based on the preliminary problem statement in the systems service request (SSR).

The baseline project plan could also be used to assess the situation (the second task in business understanding) which details the resources, constraints, and assumptions associated with the project.

A feasibility analysis, usually constructed with any SA&D project, can be used to assess the risks (technical feasibility analysis), costs and benefits (economic feasibility analysis) of the study. The results of the feasibility analysis are documented within the Baseline Project Plan.

SSR/BPP/Business Case

In the retention study, a systems service request (of some form) was used by the administration to try to describe the fundamental problem. Technical personnel, comprised of faculty in economics and social science research, began to reformulate the vague problem description in the SSR into a statement that could lead to more concerted efforts to collect data and model the problem.

While a "formal" Baseline Project Plan was not written, a Business Case was developed and shared with administrative personnel, the members of the Board of Trustees and other interested University groups to describe both the feasibility and the need for this analytical undertaking.

Pine Valley Furniture **Prepared: 9/20/2003**
Statement of Work

Project Name: Customer Tracking Systems
PVF Project Manager: Jim Woo

Customer: Marketing
Project Sponsor: Jackie Judson

Project Start/End (projected): 10/1/03–2/1/04

PVF Development Staff Estimates (labor-months):

Programmers:	2.0
Jr. Analysts:	1.5
Sr. Analysts:	0.3
Supervisors:	0.1
Consultants:	0.0
Librarian:	0.1
TOTAL:	**4.0**

Project Description

Goal
This project will implement a customer tracking system for the marketing department. The purpose of this system is to automate the . . . to save employee time, reduce errors, have more timely information, . . .

Objectives
- minimize data entry errors
- provide more timely information
- . . .

Phases of Work
The following tasks and deliverables reflect the current understanding of the project:
In Analysis, . . .
In Design, . . .
In Implementation, . . .

Figure 2: Sample Statement of Work (Hoffer, George, & Valacich, 2013, p. 132)

A project plan, and the corresponding Statement of Work (SoW), can be used in data analytic

projects as well. The Statement of Work uses the elements of the Baseline Project Plan to give the systems owners (business stakeholders) an overall plan for the development of the information system. The SoW, shown in Figure 2, defines the "what and when" of the project. The SoW acts a contract with business stakeholders—both the systems owners and systems users—to develop or enhance an information system; the SoW defines the vision, scope, constraints, high-level user requirements, schedule, budget, and most importantly, the deliverables of the project.

Project Plan/Statement of Work
This three to five year retention study is underway. Currently, data sources are being identified and analysis on these data sources is being carried out. In addition, the project plan calls for the identification and collection of additional data (both quantitive and qualitative) that will help to identify additional areas for further study.

Work on a formal Statement of Work is also currently underway. The results to date have given the administration and the Board of Trustees hope that this effort will yield positive results in the future.

4. NON-CONFORMITY WITH SA&D WORK

There are a number of tasks outlined in the CRISP-DM methodology that need to be handled differently than in traditional SA&D projects. The first of these deals with the organizational culture itself. Changing to an analytics culture is difficult and time consuming and must be done starting with the C-level officials (Davenport T. , 2006).

Organizational Culture
"Analytics competitors must instill a companywide respect for measuring, testing, and evaluating quantitative evidence" (Davenport T. , 2006, p. 6). This change in organizational culture is reflected in adjustments to activities, resources, and personnel. In particular, the changes deal with project management and appreciating working with the analytical outcomes (KDnuggets, 2017).

Project Teams
In his recent book, *Keeping up With the Quants* (Davenport & Kim , 2013), Tom Davenport and Jinho Kim make the case that the development of analytical solutions would take a team of people in order to maximize the value of the work. In fact, the first two questions of Davenport and Kim's Stakeholder Analysis Worksheet (Davenport & Kim , 2013, p. 26) asks if all the potential executives (business stakeholders) have been included and briefed on the data analytics project.

In another work, Pomykalski (Pomykalski, 2015) outlined the roles of both the business stakeholders and the "quants" (quantitative professional) over the entire CRISP-DM lifecycle. This work proposed the roles of each group and the leadership responsibilities in order for the work to gain maximum value.

In the business understanding stage, the business stakeholder's (users and owners of the business problem) primary responsibility is to formulate and state the business problem and the specific business objectives that are to be met in this work. The quant's role is to reformulate the business problem into the data mining goals discussed below (Davenport & Kim , 2013).

In addition, the quant must be able to design the preliminary project plan and the success metrics to assess the effort, while the business stakeholder needs to see that the chosen metrics properly measure the success of the project in terms of the business objectives (Pomykalski, 2015).

Project Selection and Initiation
Given the change in the organizational culture there should naturally begin to develop new criteria for project selection. In traditional SA&D projects the end-user (business stakeholder) drafts a systems service request (SSR) that includes a vague problem description faced by the business unit. This SSR is used to assess, by some type of project selection committee, the project in terms of its current status. This selecton committee must be made aware of the differences that exist in data analytic projects.

With data analytic projects the data governance structures need to be set up to allow more members of the organization access to data sources (Harris, 2015). Data governance policies are necessary for enhancing the collaboration between the "business problem analytics is trying to solve, data quality experts, data modelers, technical architects managing the analytics infrastructure, and, of course, those lauded data scientists applying their insight-generating statistical" knowledge (Harris, 2015).

Data Mining Goals
One of the key tasks in the business understanding stage of the CRISP-DM

methodology is the development of goals of the data mining activities. "A data mining goal states project objectives in technical terms" (Chapman, et al., 2000, p. 9) such as what in particular is going to "predicted". In a typical SA&D development effort, the goals of the information system are not specifically called out and documented to this level.

These technical goals could be—and should be—added to the Statement of Work to show a complete picture of the work to be done to the clients and the technical stakeholders.

Agile Project Management

Traditional project management approaches to analytics projects have been found to be limiting. More success has been found using agile project management techniques (Larson & Chang, 2016).

The particular needs of analytics projects, "the continuous delivery of business value throughout the development lifecycle" (Castro & Jain, 2016, p. 2), require a different approach to project management. Agile development speed the timeliness of analytic solutions for quicker delivery of value to business stakeholders. Castro and Jain (2016) outline the effectiveness of agile project management to analytics projects.

5. FUTURE WORK

This paper is the second in a planned series of papers to show how systems analysis and design methods and models can be applied to the CRISP-DM method for analytics projects. This paper focuses on the overlay of the planning stage of SA&D to the business understanding stage of the CRISP-DM process.

Work also needs to be done to begin to educate future IS professionals (students) on the benefits and activities in agile project management. The work of Landry and McDaniel (2015), in including agile development methods with a traditional project management course, seemed to show some promise. Further work in moving project management to a blend between the traditional and agile methodologies needs to be continued.

6. CONCLUSIONS/SUMMARY

Given the long history of the application of Systems Analysis and Design activities and models, their application to the analytics work, especially the CRISP-DM methodology, is important. This work focused on the application of the models and activities in the systems planning stage of SA&D to the business understanding stage in CRISP-DM.

Many of the models and activities will allow the CRISP-DM to be more thoughtful and improve the documentation of the activities. This improved documentation could lead to more standardization of analytic modeling projects which in turn could lead to better outcomes and value of data.

In this paper, key things like data governance, project management and project teamwork are also shown to be key elements.

7. REFERENCES

Arthur, C. (2014, June 30). Facebook emotion study breached ethical guidelines, researchers say. *The Guardian*. Retrieved September 11, 2016, from https://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA, and CRISP-DM: A Parallel Overview. *IADIS European Conference on Data Mining.* Amsterdam.

Buzydlowski, J., & Pomykalski, J. (2016). Comparing and Contrasting Systems Analysis Methodologies with Data Analytic Frameworks. *Northeastern Association of Business, Economics and Technology (NABET).* State College, PA.

Castro, L., & Jain, R. (2016). Achieving Effective Data Analytics:. *Conference on Information Systems Applied Research. 9*, pp. 1-6. Las Vegas, NV: ISCAP.

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.* Somers, NY: IBM. Retrieved June 2017, from http://www.crisp-dm.org/CRISPWP-0800.pdf

Davenport, T. (2006). Competing on Analytics. *Harvard Business Review*, 1-9.

Davenport, T., & Kim , J. (2013). *Keeping Up With the Quants: Your Guide to Understanding and Using Analytics.* Cambridge, MA: Harvard Business Review Press.

Dennis, A., & Wixom, B. (2014). *Systems Analysis and Design* (Sixth ed.). New York, NY: John Wiley & Sons.

EPS Cloud Fabric. (2012, August 24). The Importance of Integrity in Business. Retrieved June 27, 2016, from https://www.nimbo.com/blog/the-importance-of-integrity-in-business/

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3), pp. 37-54.

Gass, S. (1984). Documenting a Computer-Based Model. *Interfaces, 24*, 84-93.

Gass, S. (1993). Model accreditation: A rationale and process for determining a numerical rating. *European Journal of Operational Research, 66*(2), 250-258.

Harris, J. (2015, August 20). Data Governance and Analytics. *The Data Roundtable*. Retrieved from http://blogs.sas.com/content/datamanagement/2015/08/20/data-governance-analytics/

Hoffer, J., George, J., & Valacich, J. (2013). *Modern Systems Analysis and Design* (Seventh ed.). Pearson.

KDnuggets. (2017, July). How to Turn your Data Science Projects into a Success. Retrieved from http://www.kdnuggets.com/2017/07/olavlaudy-turn-data-science-projects-into-success.html#%2EWWkQNCn4tJ8%2Elinkedin

Kock, N. (2007). *Systems Analysis & Design Fundamentals: A Business Process Redesign Approach.* Thousand Oaks, CA: SAGE Publications.

Landry, J., & McDaniel, R. (2015). Agile Preparation Within a Traditional Project Management Course. *EDSIG Conference* (pp. 1-7). Wilmington, NC: ISCAP.

Larson, D., & Chang, V. (2016). A review and future direction of agile, business intelligence, analyticsand data science. *International Journal of Information Management, 36*, 700-710.

Miliard, M. (2014, October 6). Facebook would like your health data. *Healthcare IT News*. Retrieved September 11, 2016, from http://www.healthcareitnews.com/news/facebook-would-your-health-data

Pomykalski, J. J. (2015). The Analytics Team: Contributions of the Business and Technical Stakeholders. *Northeastern Association of Business, Economics and Technology (NABET).* State College, PA.

Robinson, S. (2004). *Simulation: The Practice of Model Development and Use.* Chichester, England: John Wiley & Sons.

Safique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, 217-222.

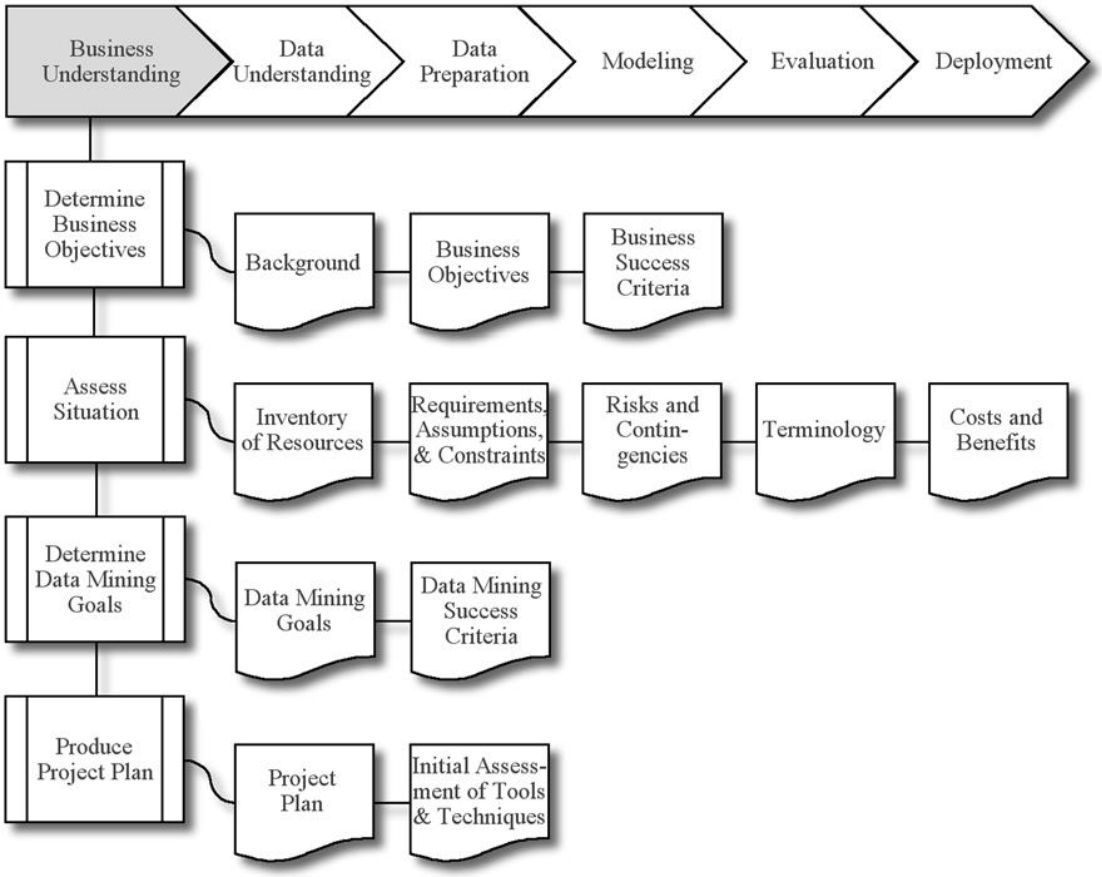
Steijn, W., & Vedder, A. (2015, July). Privacy under Construction: A Developmental Perspective on Privacy Perception. *Science, Technology, and Human Values*, pp. 615-637.

Valacich, J., & George, J. (2014). *Essentials of Systems Analysis and Design* (Sixth ed.). Pearson.

Whitten, J., & Bentley, L. (2005). *Systems Analysis and Design Methods* (Seventh ed.). McGraw-Hill/Irwin.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *4th international conference on the practical applications of knowledge discovery and data mining*, (pp. 29-39). Retrieved June 2017, from https://pdfs.semanticscholar.org/48b9/293cfd4297f855867ca278f7069abc6a9c24.pdf

Appendix 1



CRISP-DM Business Understanding Phase (Chapman, et al., 2000, p. 7)