Abstract

Estimation of software sustainment costs can consume from 60-90% [1] of the total ownership cost of a program, yet the software industry continues to struggle with the best way to predict these costs. Traditional cost drivers used for acquisition estimates do not necessarily apply to the sustainment portion of a project, particularly to some of the government costs during sustainment. This paper discusses an ongoing research project applying data mining techniques for collecting and analyzing actual cost, effort, programmatic and technical data from evolving software systems. The end goal of this data mining is to determine the best sustainment cost drivers, sustainment cost and schedule estimation relationships (CERs, SERs), and/or rules of thumb for estimating software sustainment activities.

Introduction

Estimating of software sustainment costs continues to be an issue for government agencies and their contractors. Software is not like hardware in that it does not wear out after a certain number of uses. Software is a much more malleable 'thing' than hardware; software developers are often asked to stretch and mold software solutions to make accommodations for limitations in the hardware or other software applications of a system.

For the purposes of this research, software sustainment costs cover all of the cost associated with keeping a software application up, running and meeting all functional and non-functional requirements of the system. The clock on software sustainment starts when the software is first delivered into production and continues until the software application has gone out of service. These costs cover a myriad of activities including adaptation, correction, minor enhancements, field support, certification and accreditation, addressing technical debt, etc. Some of these activities can be estimated using traditional software metrics, CERs and rules of thumb; some cannot.

Towards an identification of proper and comprehensive sustainment cost drivers along with cost and schedule estimating relationships, PRICE is involved in an on-going effort to collect and analyze software sustainment data through a data mining process. This data collection includes effort and cost data, as well as technical factors associated with the applications being studied. As is often the case with complex data collection projects involving multiple stakeholders, this research effort is not as far along as the author had hoped at this point in time. However, this is not a report on a failed project, but rather a report on progress towards success.

The first section of the paper discusses in more detail what software sustainment is and defines the activities associated with this phase of the software lifecycle. The second section contains a discussion of the process of data mining applied in this project. Following this, the paper walks through the data collection journey to date, discussing the pitfalls, challenges and lessons learned. This discussion includes details on the automation successes achieved in the process to date. The final section wraps up the discussion with presentation of lessons learned and next steps.

Software Sustainment

More and more systems are reliant on software for successful operations. There are many reasons for this. First and foremost is the need to keep up with ever improving technology options in both the

hardware and the software world. Due to budget constraints and the availability of money for research and development efforts, less new software is being developed while legacy software applications are being enhanced, adapted and modernized in an effort to meet new threats, mission requirements, coalition configurations, etc.[1] Software is often modified to accommodate changing requirements because it is easier to deploy than hardware. Often the most prudent solution to new and burgeoning requirements is to address issues with software rather than hardware. Due to this increased reliance on software and the need to make it last longer, software sustainment is a significant concern to all involved in fielding software intensive systems, cosuming up to 60-90% of total costs (effort) for many programs.

According to the Software Engineering Institute at Carnegie Mellon University (SEI/CMU) [2]:

> *"Software sustainment involves orchestrating the processes, practices, technical resources, information and workforce competencies for systems and software engineering, to enable system to continue mission operations and also to be able to be enhanced to meet evolving threat and capability needs."*

According to the Institute of Electrical and Electronics Engineering's (IEEE) Standard 12207 [2]:

> *Software maintenance is "the process of modifying a software system after delivery to correct faults, improve performance and adapt to changing environments"*

The terms software sustainment and software maintenance are sometimes used interchangeably. Depending on who you are and why you are talking about software maintenance (or sustainment) this might be acceptable. However, for many developers and consumers of software intensive systems, software maintenance is merely a subset of software sustainment. This is certainly true for the US Department of Defense (DoD) and its contractors. For them software sustainment includes everything associated with keeping fielded software operational, valuable, useful and easy to use. The activities associated with software sustainment include:

- Software changes – this includes the activities associated with requirements, design, implementation, integration and testing of software corrections, enhancements, etc. associated with an updated release of operational software. This activity includes changes associated with correcting bugs and addressing IAVA's (Information Assurance Vulnerability Alerts) or other security issues.
- Project and Technical Management – this includes the activities associated with planning, execution, configuration management, release management, measurement, contracting and other oversight activities associated with update releases of operational software
- Software Licenses – this includes both cost and effort associated with maintaining all the licenses necessary to maintain and support all third party and open source software that is part of the operational software system.
- Certification and Accreditations – this includes the activities associated with ensuring that the operational software continues to meet performance criteria associated with security, airworthiness, net-worthiness, IAVA's etc.

- Facilities – this includes cost and effort associated with creating, operating and sustaining facilities necessary to create an environment equipped to creating and sustaining operational software capability for the total operational life of the software
- Sustaining Engineering – this includes cost and effort associated with support necessary to sustain successful operation of the software throughout the lifecycle (investigations, test support, training, help desk, release delivery, etc.)
- Field Software Engineering – this includes cost and effort associated with on-site support of the software application in its operational environment (tech support, troubleshooting, installation assistance, on-site training etc.)
- Operational Management – this includes a portion of the overall operational management for non-system specific resources allocated to sustain the operation of a particular software system (operations, personnel management, financial management, change management, information management, etc.)

Data Collection and Analysis – Data Mining

Doing data collection right is not easy. Regardless of the problem you are trying to solve, there are many important steps that need to be taken in order to make sure that the data collection is structured to efficiently collect the best set of data to answer your question(s). According to Wikipedia "Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems." [3] Applying appropriate data mining techniques seems to be the proper path on our quest for better predictions of sustainment costs.

In 1999 a group of businesses got together and created the Cross Industry Standard Method for Data Mining (CRISP-DM). [4] This is a methodology that can be employed to apply structure to any data mining projects and acts as a sensible roadmap to help keep data junkies on track and focused. The CRISP-DM defines the six phases of the data mining process, though it is important to bear in mind that these steps are generally not entirely sequential. Most interesting data projects are very iterative in nature, with each iteration benefiting from and advancing lessons learned in previous iterations. The six phases are:

- Business Understanding

  This phase represents a very significant part of the project. Data mining must have a purpose and the purpose needs to be understood and accepted by all the project stakeholders. To put this more simply, the first step in a data mining exercise is to understand what question (or questions) the business needs an answer to. One would not start building software without first asking what requirement this software must fulfill (though this does occasionally happen – it never ends well). Similarly one should not start collecting and analyzing data without asking what problem they hope the data will help solve.

- Data understanding

  Once the question to be answered has been identified, the next obvious step requires investigation into where and how that data might be made available. Organizations tend to collect tons of data though that data is often stored and maintained in many separate silos throughout the organization. Chances are good that the data required to answer a business related question will require harvesting data from multiple groups within the organization. During

this phase, the data mining team begins to determine the data items that are likely to be required to answer the question as well as the places within the organization these data items can be acquired. Not only is it important to understand what data to collect and from where it can be obtained, it is equally important to determine the circumstances of data collection in each instance in order to make it possible to create a common ground for analysis

- Data Preparation

  Data is often ugly! Often, when one thinks of data analysis, the image is entirely of spreadsheets with endless rows and columns rich with numerical information of interest to the organization and related to the specific question being addressed. And in a perfect world maybe this is reality. More likely, particularly for those of us in the cost estimating community, this is not the case. Data is presented in many forms. While there are many numerical values for analysis, there is often also important non-numerical context data that has as much value as the numerical data in helping understand the answer to the question being posed. There are also many cases where some data is incomplete, missing or suspicious in nature. Data sets also may require filtering to remove pieces of data that are uninteresting or irrelevant to answering a specific question.

- Modelling

  Models can be used for classification or prediction, depending on the question to be answered. If one is trying to determine the best audience to target advertising for a new vehicle, they may want to create a model that classifies car owners based on previous purchases. If one is trying to predict the costs of developing or sustaining software, they will want to create a predictive model to accomplish this based on outcomes of similar software projects. Having said this, the cost model builders still may want to do some classification modeling before addressing the predictive problem, because as noted earlier there are often context details that drive stratifications in data sets.

- Evaluation

  Clearly, once a model has been developed, it is important to make sure that the model makes sense. This evaluation involves testing the model developed against a set of data with known outcomes and ensuring that it behaves properly. It is important to review via statistical tests, the 'goodness' of any models developed in order to ensure credibility and to provide context as to when it is, and is not, appropriate to use these models. It is desirable to hold back a part of the data set, when possible, to use as a test of the model developed to ensure that the model adequately responds to as many possible variations as possible. Equally important, the data mining team should always remember that sometimes the data is going to tell you something that is completely implausible; common sense should be an important tool in any data mining evaluation test kit.

- Deployment

  Once the data mining team is happy with the model created, the next big step is to introduce it to the stakeholders who need answers to the question originally posed. This phase may involve creating a physical implementations that makes the model easy for the end user to apply, documenting this implementation and training the end users on the proper ways to use it. End

users must fully understand the limitations of the models use as well; a model developed to predict the costs for aircraft that was developed using only commercial aircraft data would not be suitable for predicting the costs of a military air fighter. Deployment may also involve some evangelization on the part of the data mining team. Not everyone is going to believe a model is good just because its creator says so. There will be skeptics who need to be won over through successful presentation of evaluation results and some proof of concept exercises.

Data Collection Journey to Date

In the context of the phases outlined above, this software sustainment data mining projects is really just starting to get interesting. While there have been several iterations of 'test case' modellings and evaluations, the quantity and expansiveness of the data collected so far is not adequate to support good quality models at this point. The expectation is that additional data, soon to be delivered, will resolve that issue. Having said that, much progress has been made on the first three phases of the data mining project with an eye toward being ready for action when the team receives the next wave of data.

The question to be answered by this data mining exercise, simply put, is 'How can the organization do a better job of predicting software sustainment costs throughout its portfolio?' This question is pervasive in many organizations, especially since sustainment is often handled as a level of effort where there is a budget for sustainment and whatever highest priorities are identified during the budget cycle, those are the items that will be addressed. And to be fair, in almost every organization, regardless of what the sustainment priorities are – if a problem arises that is likely to create serious customer loss, business failure or tragic accidents – these will be addressed regardless of the sustainment budget plan. This fact does not, however, excuse an organization from having a sustainment plan for their portfolio of software and appropriately assigning budget to those sustainment projects with the highest business value. So in light of all of this, the question to be answered with this project can be more specifically refined to 'How can the organization do a better job of predicting sustainment costs across their portfolio in order to achieve maximum value for dollars spent?'

Understanding that data collection was time consuming and presented an expense to the contractors performing many of the sustainment activities, it was important that there were contractual incentives for the contractors to participate. This area was addressed by the organization in conjunction with the data mining team.

Armed with this question, the next phase requires gaining an understanding of the data. This process has taken some time and is still evolving. The organization is large and widely dispersed geographically. While there have been efforts to begin to institutionalize data collection, these efforts are in their infancy and lack rigorous enforcement in many groups within the organization. The challenge for the data mining team was to develop a data collection process that was reasonable with respect to the effort required to complete, comprehensive enough to fully answer the question, and aligned sensibly with data collection processes already in place within the organization.

Not only was it necessary to determine what data needed to be collected, it was also important to establish the periodicity for data collection. With a software development project there are generally discrete points at which data collection occurs; whether that be against delivery milestones within a

traditional waterfall project or within iteration and releases within an agile approach, the periodicity of data collection is relatively well defined and understood throughout the software development community. Sustainment projects tend to be different because releases can come at regularly planned intervals or at points in the lifecycle when a release is needed (to address a critical bug or a security vulnerability). Most sustainment efforts, by necessity, are a hybrid of these two models. For this project it was determined that the data collection should be done at two levels:

- Monthly data collection to include:
  - Program level data to include:
    - Field Support – Hours (by specific support activity and labor category) and Cost by Contractor by System
    - Program Support – Hours (by specific support activity and labor category) and Cost by Contractor by System
    - Infrastructure Support - Hours (by specific support activity and labor category) and Cost by Contractor by System
    - Other Direct Costs (ODC) – Costs by Contractor by Vendor by System
    - Travel – Costs by System by Trip
  - Engineering level data to include Hours (by specific activity and labor category) and Cost by Contractor by Requirements ID Number
- Release Data collection to include:
  - System level context Data such as Domain, Operating Environment, CMMI Level, Development Process, Schedule information
  - Software Size information by Requirement ID Number such as New Size, Deleted Size, Reused Size, Modified Size, Function Point Count, Functionality, Development Technology

The above list is a generalization of the data collection requirements for this project. It was developed by visiting several groups within the organization to learn from the Subject Matter Experts (SMEs) in the field as to their software sustainment cost drivers and also to ascertain what types data could be easily culled from existing data collection efforts (time cards, EVM, etc.). The data collection criteria were used to create a series of spreadsheets with a very detailed data dictionary to ensure, as much as possible, ease of use and consistency of data collected within the organizations

Armed with extensive spreadsheets, the data team was ready to deploy the data collection tools to select sustainment efforts within the organization to pilot the data collection process. Data collection, particularly of this magnitude, is often met with significant resistance. And while there was an enforcement mechanism through contracting means, on-going efforts already on contract had little incentive to participate at this point in time. Data collection efforts were temporarily halted while the data mining team attempting to identify projects willing to participate.

Eventually several data sets were identified that, while not completely aligned with the original data collection plan, were aligned enough to give the team a starting place. An important part of the data understanding phase is the acknowledgement that not all of the team's data wishes are likely to be met, at least not immediately; flexibility and patience are important skill sets to bring to the data mining table. The first data set was delivered to the data mining team which got to work on data preparation.

The data set delivered was not of sufficient quality or quantity to begin serious modeling, but it was adequate for data preparation based on the assumption that subsequent data sets would follow the same form and contain the same data attributes (with some possible additions). The tool selected for data preparation and analysis was RapidMiner, an open source application that provides powerful data mining capabilities. [5] The first wave of data was consolidated into an MS Excel® format and with very little modification could be imported into the RapidMiner Application. Figure 1 gives an indication of what the data looked like upon import. (Note – the form of the data mimics the team's initial results but the numbers are not from the actual data)

0	936.7699631...	2753.426	1292.251	2.694	Unknown	0	?	235238.3766...	179964.781	73645.037	19343.068	Best Guess ...	0
0	1188.319805...	1370.050	2726.292	3.530	Unknown	0	?	62169.01590...	500447.058	311519.950	18764.110	Best Guess ...	0
0	unknown	3000.607	7187.161	8.504	unknown	0	?	3156607.944...	2915061.812	2347668.718	478891.343	0	0
0	unknown	538.640	288.790	3.067	unknown	0	?	91498.20328...	283726.663	73219.090	165578.656	0	0
0	unknown	252.789	1878.768	3.278	unknown	0	?	104322.7288...	153358.465	250866.439	148956.025	0	0
0	unknown	831.374	207.950	2.336	unknown	0	?	156487.5488...	251084.829	279896.161	71922.747	0	0
0	unknown	88.458	1368.998	0.980	unknown	0	?	125624.9703...	268313.243	252964.188	279374.801	0	0
0	unknown	1043.323	1198.060	0.335	unknown	0	?	201109.7852...	103566.218	193148.377	46279.190	0	0
0	unknown	804.784	1696.621	2.410	unknown	0	?	35280.38317...	103222.625	278364.659	164119.731	0	0
0	unknown	47.999	1373.936	3.311	unknown	0	?	57919.68820...	103624.680	78403.754	134691.935	0	0
0	unknown	1490.551	104.470	2.457	unknown	0	?	255008.4125...	147105.600	98181.387	97088.984	0	0
0	unknown	1163.077	1008.643	1.269	unknown	0	?	20026.08152...	34150.866	138491.264	60034.751	0	0
0	unknown	1700.790	173.889	2.209	unknown	0	?	77994.10730...	159998.348	277856.959	297514.337	0	0
0	unknown	42.905	808.027	0.413	unknown	0	?	45787.28754...	178023.180	100765.756	118728.090	0	0
562.665	5463.260666...	8614.135	16806.488	3.750	112680.5109...	189386.502	70559.878	1939636.465...	2109652.616	1298956.112	629834.798	0	The contracto...
1008.876	28666.05533...	88409.375	38157.941	57.908	180566.3142...	28625.770	195631.697	58198.96055...	1348684.888	8976920.952	7275564.213	0	The contracto...
1349.389	38,400 Hr	6342.272	106023.142	54.964	903.0710353...	547511.227	310119.953	3940230.523...	931706.088	204712.865	10375340.846	0	The contracto...
514.237	24672 Hr	31987.020	80970.021	20.292	$581k	1425172.775	148957.064	$3326k	6416213.250	2928915.203	10096676.040	Data from PMR	C-SR Block 6...
7240.074	56160 Hr	103074.670	39708.390	66.082	$533k	424494.614	1290039.271	$6583k	1743393.946	7917160.681	19422575.036	Data from PMR	1) C- Sr. Savil...
5309.067	57600Hr	13264.226	100035.187	16.078	378K	283555.897	850787.347	7461K	9059909.464	13311249.472	12121450.023	From PMR D...	1. CIB mainte...
0	Unknown	0	0	0	Unknown	0	?	Unknown	0	?	?	0	N/A
0	Unknown	0	0	0	Unknown	0	?	Unknown	0	?	?	0	N/A
0	Unknown	0	0	0	Unknown	0	?	Unknown	0	?	?	0	N/A
0	Unknown	0	0	0	Unknown	0	?	Unknown	0	?	?	0	N/A
0	Unknown	0	0	0	Unknown	0	?	Unknown	0	?	?	0	N/A
0	Unknown	0	0	0	Unknown	0	?	Unknown	0	?	?	0	N/A

Figure 1 - Example of raw data upon import

Upon closer inspection, it was obvious that there were quite a few areas where this data required further analysis. If is important to note that this is merely a snapshot of the data – there are rows above and below and columns to the right and left. With more than 50 records each of which had over 100 columns of attributes, this analysis is likely to be tedious, time consuming and likely to be fraught with oversites or errors.

A very powerful feature of RapidMiner is its ability to present analysis of the meta-data of each of the data attributes in a format that presents a top level view of the situation and a clear roadmap of the work to be done. Figure 2 shows a snapshot of the Information that is available for each data attribute once the import process has been completed. (Once again this is a representation of the form but not the actual data). Note the information that is available for each data attribute:

- Data type – RapidMiner makes a guess on import as to the type of data of each data attribute - Note for Attribute 5 in Figure 2 that the value was assumed to be polynominal (meaning it has multiple discrete values) but it clearly has some numeric values - further investigation reveals that some entries contain text such as unknown or NA. This indicates that part of the preparation for this data set would be to address these text fields and change the attribute type to numeric
- Number of examples (data points/rows) for which the attribute is missing – Note for Attribute 7 there are missing values. This indicates that for this attribute part of the preparation should address the best way to deal with missing attribute values.

- The Statistics Section of the Meta Data includes information based on the data type:
  - Min, Max and Average for real or numeric data
  - Least, Most and Values for polynominal and binominal data
  - Earliest, Latest and Duration for Date Data
- For each Attribute the user has the option to dig a bit deeper in the Meta Data as shown with Attribute 8. Enabling this option provides a visualization of the distribution of a particular attribute.

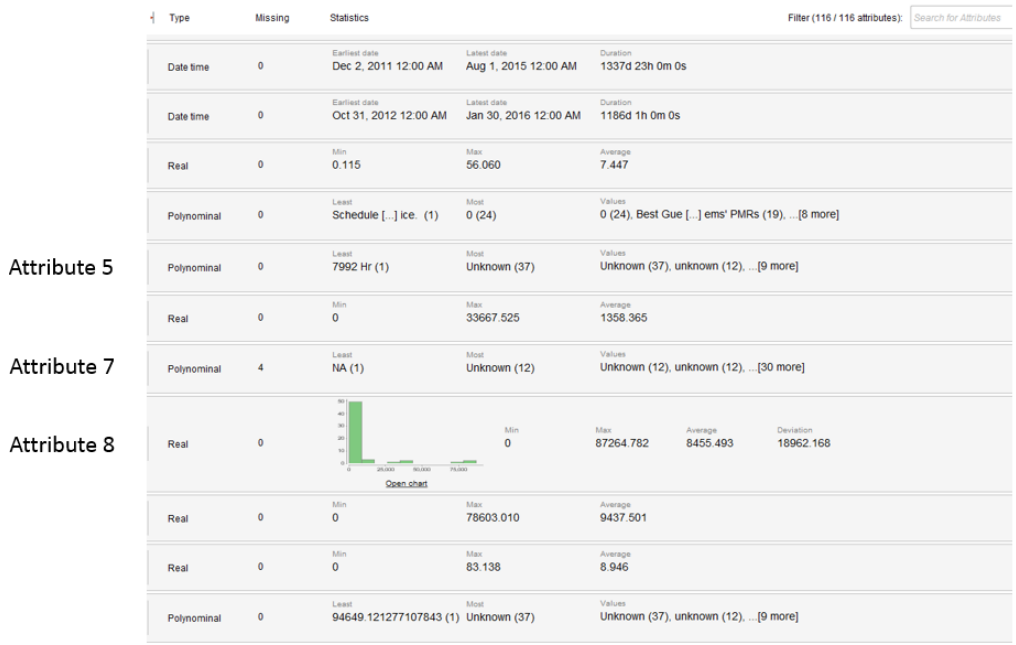

Figure 2 - RapidMiner Stats around each data attribute

Analysis of all of the meta-data for a data set provides a roadmap into the areas where further investigation is needed to complete the data preparation task.

RapidMiner provides the capability to create repeatable processes with an easy to use visual drag and drop interface. It comes with hundreds of operators that enable the end user to eliminate useless attributes, replace missing values with a more appropriate value (if one can be assumed or calculated), to filter out data attributes that are unnecessary to the current analysis, etc. Figure 3 shows some of the filtering and cleansing operators available through RapidMiner.

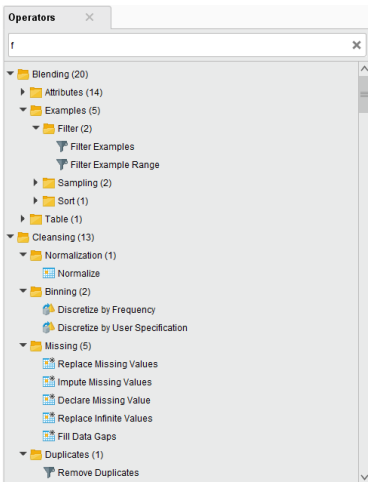

Figure 3 - Example of Operators for Filtering and Cleansing

Figure 4 shows a sample of such a process created by the data mining team for the sustainment data being collected

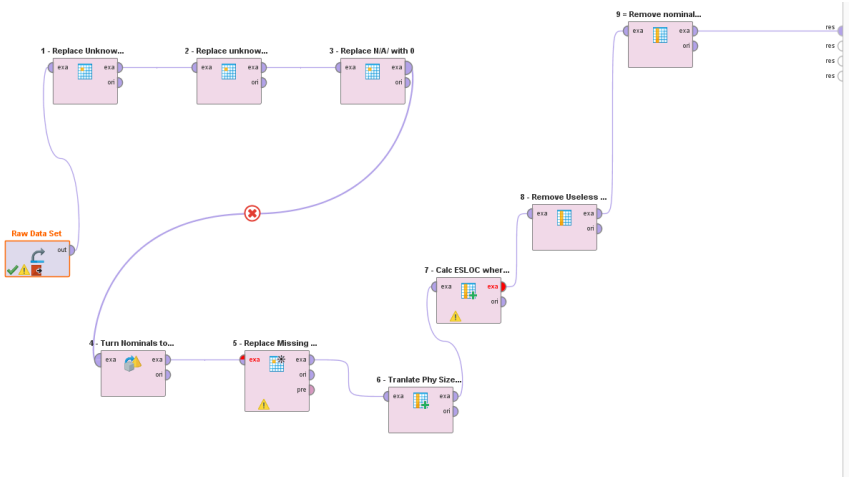

Figure 4- RapidMiner Process to Prepare Data

In Figure 4 the following things are being done:

Step 0 - The Raw Data Set is identified as the example set (example set is RapidMiner terminology for the data set to be prepared)
Step 1 – All instances of the term 'Unknown' for a selected subset of attributes are set to 0
Step 2 – All instances of the term 'unknown' for a selected subset of attributes are set to 0
Step 3 – All instances of the term 'N/A' for a selected subset of attributes are set to 0
Step 4 – Selected attributes of type nominal have their data types changed to numeric
Step 5 – Missing values are replaced with 0 for all size related inputs (since size can be new, modified, reused, deleted, etc. empty spaces are likely to indicate 0 for that category)
Step 6 – Software lines of code (LOC) that are entered in physical size units are translated to Logical lines of code via conversion factors collected in the data set

Step 7 – In cases where LOC>0 in one or more of the above categories, but no ESLOC is calculated, this calculation is performed for this data attribute.
Step 8 – Useless attributes are removed – RapidMiner removes attributes which meet certain user specified criteria (such as where all entry values are the same, or all or most values are missing, etc.)
Step 9 – Removes from the data set a selected subset of attributes that have nominal values (text), were used in calculations and thus are already represented, or have missing values

Figure 5 shows a snapshot of the data set after the data preparation phase (Once again this represents the form not the actual numbers). One can see that the question marks and zero slots have been replaced with values and this has resulted in a much smaller set of usable attributes. In fact further analysis indicates that until a larger and more complete data set is attained, further analysis of this data is likely to be unproductive.
Like Figure 1, this is also a snapshot of the data set but in this case while there are rows above and below the ones selected – all the columns are shown.

16.756	57600	62040	7466226.358	7844491.144
13.503	56160	62606.880	6654199.939	7192964.723
11.959	52241.600	66801.600	4384729.607	5086235.722
11.959	51763.200	63723.200	3325580.536	3974384.560
11.959	43142	44707.250	2847476.233	2930534.374
3.515	38400	41952	3871710.197	4075853.097
7.885	34560	35448	4046295.360	4151420.196
20.435	24672	32664	3414433.836	4010881.840
6.867	14373	14373	1214682.278	1214682.278
16.427	10272	12048	1484447.182	1597371.933
15.080	7430.530	7430.530	1196919.234	1196919.234
2.957	2870	2870	307754.978	307754.978
2.957	2870	2870	323597.058	323597.058
2.990	2870	2870	310217.152	310217.152
2.661	2870	2870	331431.282	331431.282
2.891	2870	2870	357392.915	357392.915
2.891	2662	2662	285017.689	285017.689
2.990	2444.962	2444.962	289617.912	289617.912
2.661	2357.442	2357.442	254428.384	254428.384
2.661	1639.024	1639.024	176892.752	176892.752

Figure 5 - Snapshot of example set after data preparation process

Although analysis is not yet practical, having a handle on data preparation for sets of data with this attribute set puts the data mining team in a good place for handling additional submissions of data. The basic data preparation process outlined above has been used as the basis for processes which do the following

- Perform Decision Tree Analysis for various context attributes
- Filter the set by Super Domain and perform correlation and regression analysis
- Filter the set by Operating Environment and perform correlation and regression analysis

- Prepare data from the Monthly support data reports and perform analysis on these example sets as well.

These analysis models mentioned above have been executed and evaluated against the current data set and the results have been, not unexpectedly, disappointing based on the limited quantity of data and the overall availability of attribute data.

Lessons Learned and Next Steps

The focus of the project up to this point has been on business understanding, data understanding and data preparation. Business understanding has been achieved; the project team understands the question the business wants answered and has buy in from the stakeholders that this understanding is correct. The data mining team also has a commitment from the business to continue to support and enforce data collection requirements going forward. Both the business and the data mining team have learned that patience and flexibility are important skills to bring to the table.

Data understanding has been achieved through interviews with stakeholders and subject matter experts, followed by an iterative process of developing and refining data collection tools that serve to collect project and release level attributes from the contractors supporting the sustainment projects as well as internal resources supporting the sustainment efforts. Data preparation has been addressed through automated processes developed in the RapidMiner tool which can be used as is or easily adapted to incorporate changes in the data collection mechanism or changes in the data set.

Collection and Evaluation processes have been considered and prototyped but as yet have not been well vetted or exercised due to the sparsity and incompleteness of the example set of data provided so far. For this reason there are also no real findings to report at this point in the process – there is not enough quality data to support conclusions at this point in time

Going forward the data mining team will continue to apply the work done so far to subsequent submissions of example data. There is reason to believe that future submissions will alleviate many of the concerns raised by analyses focused on this initial data set. Though lessons learned to date (and through the history of cost research and analysis) have taught the team not to expect this to be completely true, nor to expect that there will not be additional concerns that arise. Once additional data has been processed and prepared, the next steps in this process will be to apply traditional and non-traditional modeling techniques to help answer the question posed during the business understanding phase of this project.

In conjunction with continued analysis of the proposed data sets (containing a subset of the original data items outlined), the team will continue to pursue, with the business and its contractors, additional avenues of collection more closely aligned with the original data collection requirements determined in the data understanding phase of the project. The goal for the business to eventually institutionalize data collection analysis process through automation of data collection, preparation and modeling throughout the entire portfolio.

As stated earlier, this project is not as far along at this point in time as the author had hoped when first proposing this paper. This in itself, is a lesson learned. Even when there is strong commitment within a business to improve processes, there are often many obstacles. Obstacles aside, the project is inching forward and while the primary question remains unanswered to date, the work done so far is a path to

success. As more data is received and as additional parts of the business are brought into the fold, the processes and methodologies put in place in these early phases should act as a springboard to a successful start to answer the question posed earlier - 'How can this organization do a better job of predicting sustainment costs across their portfolio in order to achieve maximum value for dollars spent?' Note however, that this is just the start. More important than an immediate answer to the question is the fact that there are processes in place which are being used, and improved with each use, to institutionalize the data mining processes for constantly improving the chances of getting better informed answers to this question as more data is fed into the system.

References

[1] McLendon, et. al., "Addressing Software Sustainment Costs for the DoD", Crosstalk Magazine, January/February 2014, available at (retrieved 3/2018): http://static1.1.sqspcdn.com/static/f/702523/24156564/1388991346767/201401-McLendon.pdf?token=2sB1RBJakWI3Azmvvrh9vNGWs%2BY%3D

[2] "DAU Acquisition Encyclopedia, Software Sustainment", available at https://www.dau.mil/acquipedia/Pages/ArticleDetails.aspx?aid=544442b8-70d3-4a5c-afd4-42ffab0a9f41, December, 2017, retrieved March 2018

[3] https://en.wikipedia.org/wiki/Data_mining#Situation_in_the_United_States

[4] https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

[5] https://rapidminer.com/