Estimating Software Sustainment Costs

Arlene Minkiewicz

ICEAA Conference, June 2018

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Agenda

- Introduction
- Software Sustainment
- Data Collection and Analysis – Data Mining
- Data Collection Journey to Date
- Lessons Learned and Next Steps
- Conclusions

Introduction

- Estimating software sustainment costs continues to be an issues for organizations that deploy software intensive systems and the contractors that support these systems
  - Software isn't like hardware - more 'malleable'
  - Software developers are often asked to stretch and mold to accommodate for limitations in hardware or other software in a system
- For the purpose of this research, software sustainment covers costs of all activities necessary to keep a system up, running, and meeting all functional and non-functional requirements
  - Some of these activities can be estimated with traditional software metrics
  - Some cannot
- This presentation discusses an on-going data mining projects intended to address better ways to estimate comprehensive software sustainment costs

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

About this data mining project….

- Collection from actual software sustainment efforts of…
  - Costs and Effort
  - Technical Data
  - Programmatic Data
- Progress on actual data collection has been slower than anticipated
  - Not surprising given the nature of data collection as we know it
- This is not a report on a failed data mining project….
- But rather a report on progress toward success
- In other words… we haven't found the Holy Grail but there is 'A path! A path! A path'

Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Software Sustainment

- More and more systems are reliant on software for successful operation
- Budget constraints and available money for Research and Development has led to …
  - Less new software being developed
  - Legacy applications being enhanced, adapted and modernized to meet new threats, mission requirements, coalition configurations, etc.
- Software changes are easier to deploy than hardware changes
- Software sustainment consumes 60-90% of program budget for many software intensive programs

Software Sustainment

- *"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 Software Engineering Institute – Carnegie Mellon University (SEI CMU)*
- *Software maintenance is "the process of modifying a software system after delivery to correct faults, improve performance and adapt to changing environments"*
  - According to the Institute of Electronics and Electrical Engineers (IEEE) Standard 12207

What's Software Sustainment Include?

- Software changes – software requirements, design, code and test for items such as…..
  - Bug fixes
  - Enhancements
  - Addressing IAVAs or other security issues
- Project and Technical Management – oversight activities for sustainment period such as…
  - Planning
  - Execution
  - Configuration Management
  - Release Management
  - Measurement
  - Contracting
- Software Licenses
- Certifications and Accreditations

What's Software Sustainment Include?

- Facilities
- Sustaining Engineering – including activities such as …
  - Investigations
  - Test Support
  - Training
  - Help Desk
- Field Support – including on-site activities such as …
  - Technical Support
  - Troubleshooting
  - Installation Support
  - On-site Training
- Operational Management for non-system related resources needed to sustain a particular system

Data Mining

- Doing data collection and analysis right is not easy
- "Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems."
  - According to Wikipedia
- In 1999, several large businesses collaborated on a proscribed methodology for successful data mining
  - Cross Industry Standard for Data Mining (CRISP-DM)
  - Applies structure to the data mining process
  - Sensible roadmap to help keep data junkies on track and focused

CRISP-DM

- This methodology divides the data mining space into six phases
  - Business Understanding – What's the question?
  - Data Understanding – What's the data, how are we going to get it, how is it going to be collected?
  - Data Preparation – How do we make the data useful for analysis?
  - Modelling – How do we figure out what the data tells us towards answering the question?
  - Evaluation – How well does our model work?
  - Deployment – How do we convince others to believe our model, how do we help others to be successful using our model?

Data Collection Journey

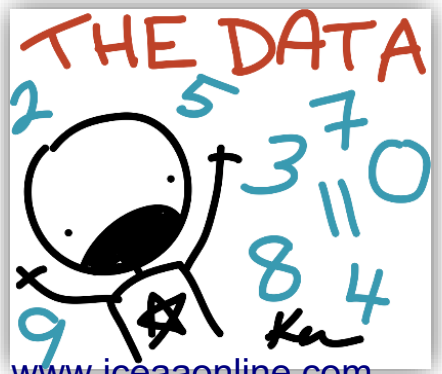
- In context of CRISP-DM methodology – this project is just starting to get interesting
- Several iterations of 'test case' modeling and evaluations, quantity and expansiveness of data collected so far is not adequate to support quality models
- Significant progress on the first three phases

Business Understanding

- 'How can the organization do a better job of predicting software sustainment costs throughout their portfolio?'
  - Question that is pervasive in many organizations
- In many organizations sustainment costs are handled as a level of effort
  - Highest priority items that emerge throughout the budget cycle get the funding
  - To be fair – all organizations (whether they plan or not) will apply needed funds to divert from tragedy (software issues that will cause financial disaster, loss of life, loss of critical customers)
- Not an excuse for an organization to not have a sustainment plan that allows for funds to be allocated to the projects with the highest business value
- 'How can the organization do a better job of predicting sustainment costs across their portfolio in order to achieve maximum value for dollars spent?'
- Business and data mining team also addressed the necessity for the business to get buy in from contractors required to support data collection through contractual means

Data Understanding

- Important to determine data to collect and periodicity of data collection
  - Traditional software development projects generally have natural points for data collection
    - *Milestone reviews when an waterfall like approach is employed*
    - *Iteration, increments and release when an agile or incremental approach is employed*
  - Software sustainment projects generally make releases using a different scenario
    - *Regularly schedule releases with enhancements, bug corrections and adaptations*
    - *On demand releases to address serious defects, security issues or other show stopping issues*
    - *Most sustainment projects are a hybrid of the two scenarios listed above*
- Data collection targets were determined through interviews with Subject Matter Experts (SMEs) within the business and their contractor community
- Data collection for this project was done on two levels
  - Monthly data collections
  - Data collections aligned with each release



Data Understanding

- 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 Cost (ODC) - Costs by Contract 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 System
- Release data to include:
  - System Level context data such as Domain, Operating Environment, CMMI Level, Development Process, Schedule Information
  - Software Size information by Requirement such as New Size, Deleted Size, Modified Size, Reused Size, Functionality, Development Technology

Data Collection

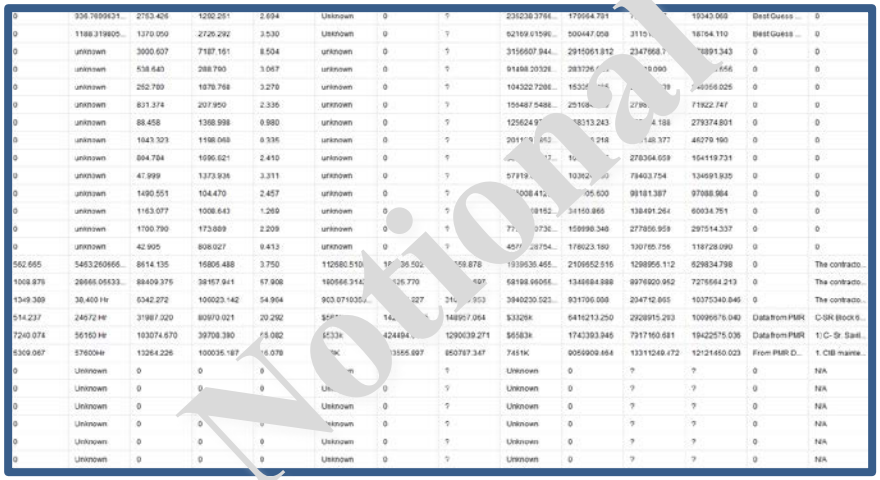
- Data collection was sluggish in the beginning
- Data collection is costly and time consuming and is often viewed with skepticism by those being 'measured'
- Measures to enforce participation through contract were thwarted by the fact that many on-going sustainment efforts were already on contract
- Data mining team was forced to start work with data from previous collection exercises
  - Subset of the data originally outlined
  - Seen as a starting place
- Patience and flexibility are important data collection tools

Data Preparation

- Data provided was of insufficient quantity and quality to support modeling and evaluation
- It was however sufficient to support data preparation
- Tool selected for automation of the data preparation processes was RapidMiner
  - Open source software (available for free from https://rapidminer.com/)
  - Powerful data mining capability
  - Easy to use drag and drop operations for building and maintaining data preparation and analysis processes

Data Preparation

- Data collected was consolidated into a single MS Excel®spread sheet
- Spread sheet was imported into RapidMiner
- Snap shot of what the imported data looked like (not the real numbers):



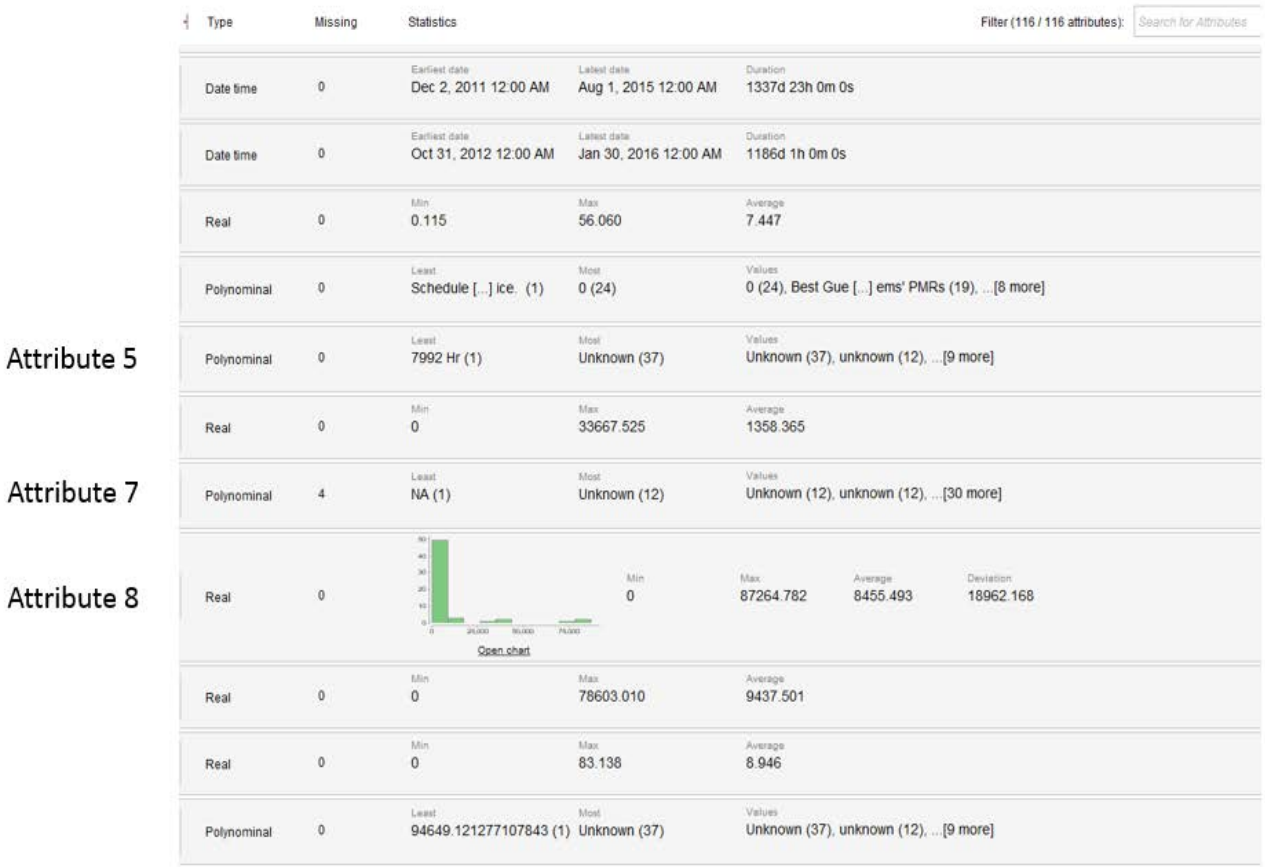
- Lots of missing data in many of the columns

Data Preparation

- RapidMiner offers a very powerful feature in that it prepares meta-data for each data attribute to provide an window into the strengths and weaknesses of a data set
- This meta-data include
  - Data Type – Polynominal, Binominal, Numeric, Date
  - Number of attributes for which this data attribute is missing
  - Statistics around the data attribute – min, max, average, least, most, value, earliest, latest, duration, etc.
  - Option to view visualization of each data attributes statistics

Data Preparation

- Meta-data provides a roadmap to areas where preparation should focus:



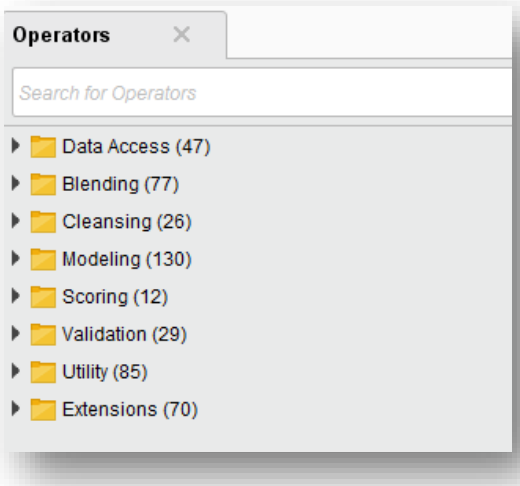
	Type	Missing	Statistics		
	Date time	0	Earliest date: Dec 2, 2011 12:00 AM	Latest date: Aug 1, 2015 12:00 AM	Duration: 1337d 23h 0m 0s
	Date time	0	Earliest date: Oct 31, 2012 12:00 AM	Latest date: Jan 30, 2016 12:00 AM	Duration: 1186d 1h 0m 0s
	Real	0	Min: 0.115	Max: 56.060	Average: 7.447
	Polynominal	0	Least: Schedule [...] ice. (1)	Most: 0 (24)	Values: 0 (24), Best Gue [...] ems' PMRs (19), ...[8 more]
Attribute 5	Polynominal	0	Least: 7992 Hr (1)	Most: Unknown (37)	Values: Unknown (37), unknown (12), ...[9 more]
	Real	0	Min: 0	Max: 33667.525	Average: 1358.365
Attribute 7	Polynominal	4	Least: NA (1)	Most: Unknown (12)	Values: Unknown (12), unknown (12), ...[30 more]
Attribute 8	Real	0	Open chart	Min: 0	Max: 87264.782 / Average: 8455.493 / Deviation: 18962.168
	Real	0	Min: 0	Max: 78603.010	Average: 9437.501
	Real	0	Min: 0	Max: 83.138	Average: 8.946
	Polynominal	0	Least: 94649.121277107843 (1)	Most: Unknown (37)	Values: Unknown (37), unknown (12), ...[9 more]

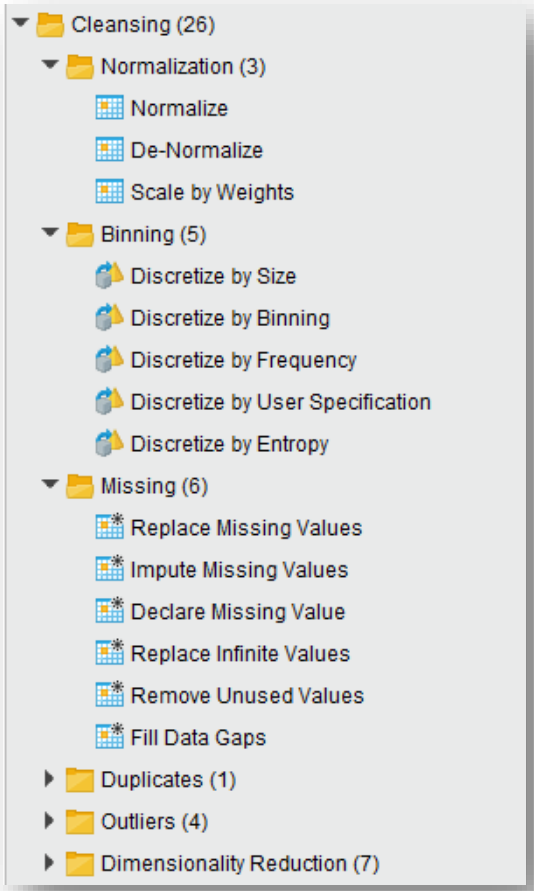
Filter (116 / 116 attributes): Search for Attributes

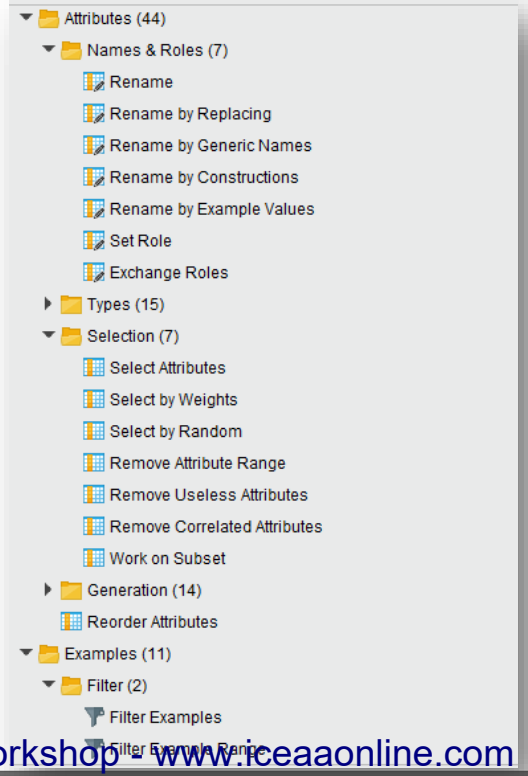
Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Data Preparation

- RapidMiner has hundreds of operators to handle various kinds of filtering and cleansing necessary to prepare data for analysis

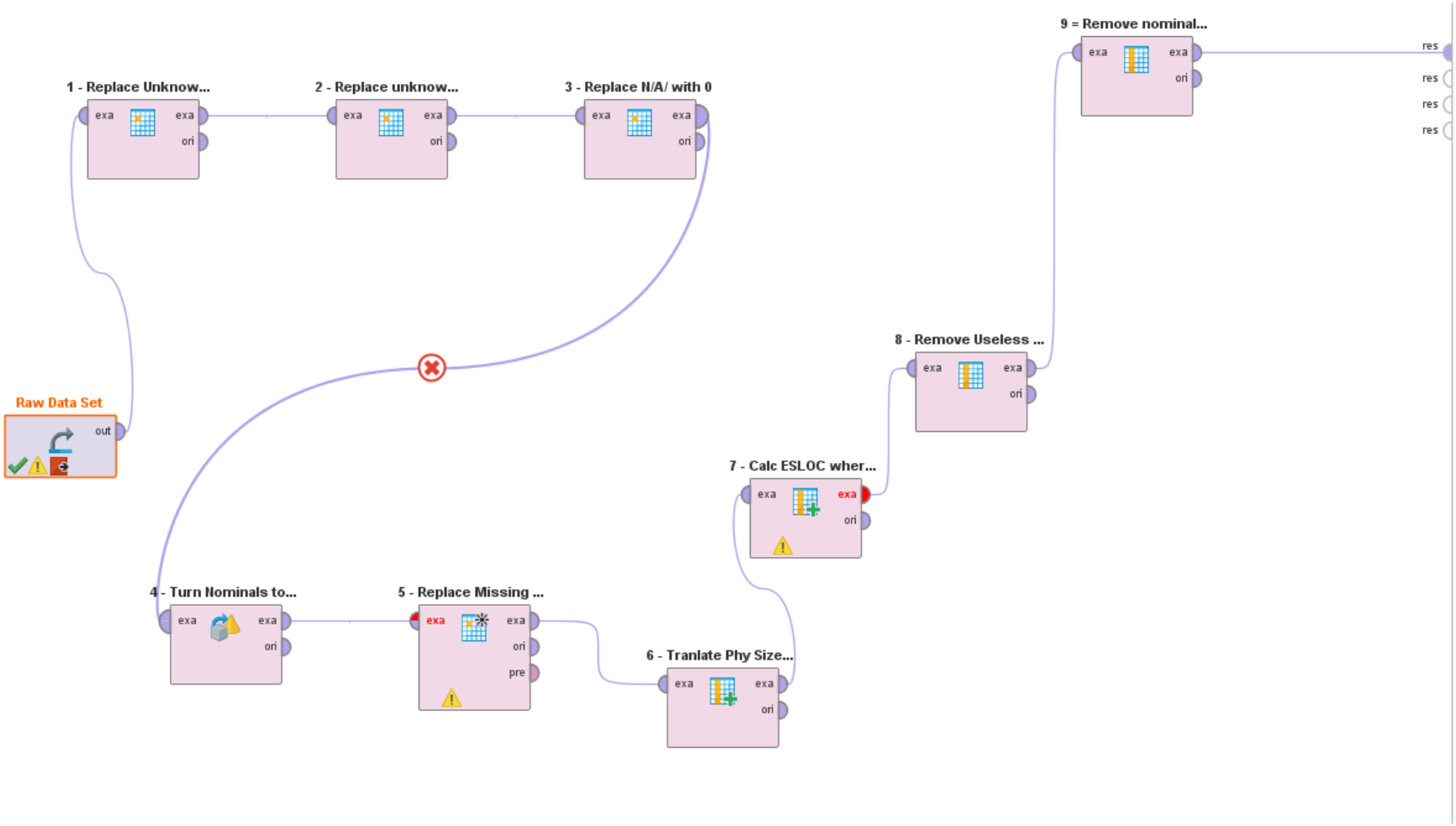






Data Preparation

- Using RapidMiner a basic data preparation process was developed

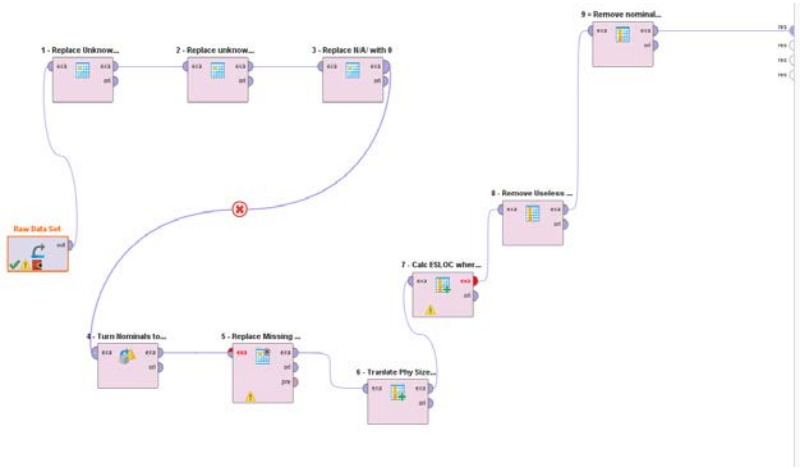


Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Data Preparation

- The steps in this process are….

- **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



Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

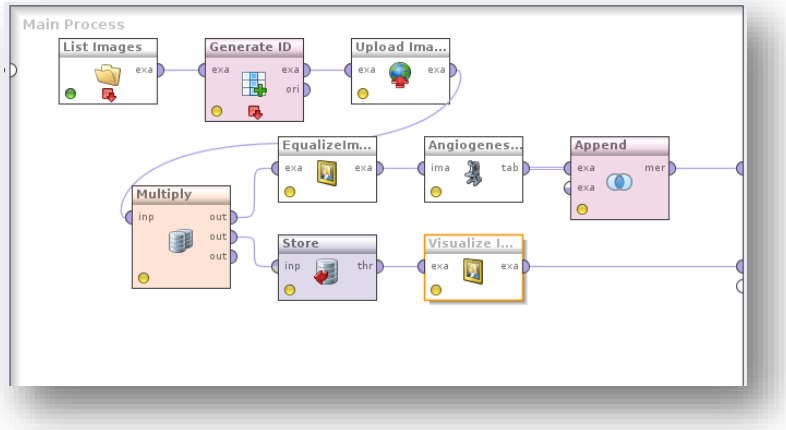
Data Preparation

- Snapshot of resulting data view (numbers not real)
- Many fewer data attributes but more complete data set with no missing or useless attributes

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	3325580.5	3974384.560
11.959	43142	44707.250	2847476.233	3930534.374
3.515	38400	41952	3871710.1	4[illegible]5853.097
7.885	34560	35448	[illegible]6295.[illegible]0	4151420.196
20.435	24672	32664	341[illegible]3.836	4010881.840
6.867	14373	14373	[illegible]682.278	1214682.278
16.427	10272	12048	1484447.182	1597371.933
15.080	7430.530	74[illegible]30	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	[illegible]	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

Data Preparation

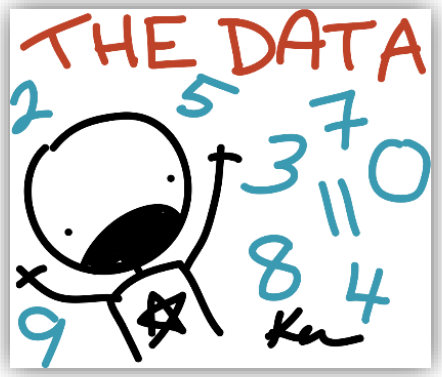
- Using this as a base, other processes were created to do such analysis as
  - Design tree analysis for context data
  - Filter the data set by super domain and perform correlation and regression analyses
  - Filter the data set by operating environment and perform correlation and regression analyses
  - Prepare data from the monthly support data reports and perform analysis on these example sets as well



Presented at the 2018 ICEAA Professional Development & Training Workshop - www.iceaaonline.com

Conclusions

- The focus of this project up to this point has involved …
  - Understanding the question to be answered and gaining consensus
  - Understanding the data needed to answer the question and the data available to answer the question
  - Creating processes to prepare and analyze the data
- The actual data collection part of the project has been disappointing – data miners need to be patient and flexible



Next steps

- Data will be run through the existing processes as it is received
- The processes will be refined as the team learns more about the data
- Additional avenues for data collection are being identified and will be pursued
- Data collection processes will be institutionalized as new contracts are issued which require data collection
- Data analysis processes will be institutionalized as best practices are spread throughout the business