Data Collection and Analysis Supporting Defendable Cost Estimates

Abstract

Cost modeling and estimation has a long and interesting history in the Aerospace and Defense since World War II. All sorts of mathematical and experiential models have been proposed and used over the years to help with bidding, planning, proposing and executing contracts. While general purpose models are useful, particularly in Rough Order of Magnitude and other early stage estimating needs, more and more industry and government professionals are asking for models built or tuned with data that is very specific to their industry and their organization. Unfortunately, many organizations do not have the infrastructure, processes or tools for collecting project data efficiently. And among those who do, some still struggle to find the best way to use their data effectively.

Over the last few years, PRICE Systems has been involved in several large scale pilots to help organizations develop processes for data collection, harvesting, and analysis to support more defendable estimates. This paper discusses the evolution of cost estimation as a practice, going from World War II to today. Following this, the two data collection and analysis pilots will be briefly describe and lessons learned will be presented.

An historical look at cost estimation

As World War II was winding down, a realization was dawning within the Aerospace and Defense community. General H. H. “Hap” Arnold summed this in the following quote: “During the war the Army, Army Air Forces, and the Navy have made unprecedented use of scientific and industrial resources. The conclusion is inescapable that we have not yet established the balance necessary to insure the continuation of teamwork among the military, other government agencies, industry and the universities. Scientific planning must be years in advance of the actual research and development work.” [1] Leaders within the community began believing that for the US to remain strong and safe, time and effort needed to be devoted to ensuring that we establish and maintain state of the art weapon systems. This marks the beginning of a period of time during which the Department of Defense has consistently been plagued with issues surrounding the costs of developing, producing and deploying weapons systems.

Project RAND was set up in October of 1945 to act as a think tank for Department of Defense (DOD) research and development efforts and in 1948 RAND separated from Douglas Aircraft Company to be its own independent entity. This represents the acknowledgement of the industry that Operations Research is an important part of any

good acquisition strategy. RAND was not exclusively focused on cost but addressing cost issues around DOD acquisitions is a recurring area of research for RAND to this day.

As time progressed things got worse, not better for DOD acquisitions. The 60's and 70's represented severe DOD cost overruns. A 1970 Government Accounting Office (GAO) study of 57 major DOD systems found 38 with at least 30 percent cost increase from point of contract award.[2] Projects understood that estimation was important but there was no consistent standards or practices for successful estimation across the industry. In the early 70's the notion of parametric estimation was introduced internally to RCA by Frank Freiman and was later introduced to the community as a whole as PRICE H and PRICE S for hardware and software estimating respectively. Parametric estimation is based on cost estimating relationships (CERs) that are derived from actual historical data. Other parametric cost estimating models followed and many program offices and contractors continue to use these models today for various aspects of the estimating process. There is a bit of a dark cloud over the use of general purpose parametric models; many consider them a 'black box' because there is little traceability to the actual data behind them. There is a drive across the industry for organizations to use historical data that is well known to the estimator to drive and support cost estimates for future systems.

Data driven estimation seems to be the flavor of the day for our industry. Not to say that the use of data driven estimation is mutually exclusive with the use of parametric models. In fact the use of parametric models is most successful when the approach is data driven. If an organization calibrates the model using their historical project data, then estimates going forward are data driven. If an organization uses historical project information to determine values for input variables to the parametric model, then that estimate is certainly data driven. Similarly if an organization uses analogies from historical data to validate a parametric estimate, that too is a data driven approach. But we're getting a bit ahead of ourselves since in order for any application of data to support an estimate, that data needs to be collected,analyzed and put into a useful form. The rest of this paper discusses two efforts to accomplish just that.

Software Project Data Collection Pilot

Motivation

The first pilot to be discussed focuses on simulation software being developed and integrated by two contractors for the US Army's Program Executive Office for Simulation, Training and Instrumentation (PEO STRI) to provide simulation, training and testing capabilities necessary to ensure the nation's security. One of the contractors

does the bulk of the development work while the other contractor is primarily responsible for the integration and test of the entire application and all of its branches.

Personnel at PEO STRItake their role of delivering training to the warfighter seriously and intendto deliver the best of breed solutions to support their safety and wellbeing. They have been experiencing increasing pain due to struggles to estimate the costs of implementing capabilities, the effort to defend these costs to their customer, the Office of the Deputy Assistant Secretary of Defense for Cost and Economics (ODASA-CE) and their need to rebound appropriately when capabilities have to be cut because risk margins increase or budgets are slashed. When estimates are presented to ODASA the first question (and not an unreasonable one) is: “Where’s the data that supports this estimate?” Estimates not supported by actual data from similar programs are not deemed credible and often inspire significant risk margins. Additionally without insight into the projected costs for specific capabilities it is difficult to make the best trade-offs when capabilities must be cut.

This pilot is intended to be the first step toward rolling out a data collection process across all of PEO STRIs projects. The end goal is to establish a consistent, repeatable process to collect cost and technical date for each capability that is delivered in their programs that can be used to support estimates going forward. The pilot is a Six Sigma project with high level support within PEO STRI.

Implementation and Process

Getting the pilot into motion was no trivial task and required a great deal of collaboration between PEO STRI personnel, the contractors and the PRICE Team. Early on it was determined that it would be prudent to form an Integrated Product Team (IPT) with representation of all the stakeholders. The first job was to identify the target set of data for collection. Toward this end the team began work on a Data Item Dictionary (DID). The DID contained detailed definitions of what data was to be collected and how the data was to be collected. It also provided guidance on mapping data from contractor specific categories to more general categories. This was necessary to facilitate alignment of data collected by multiple contractors with different labor buckets and activities. It was important that all parties were involved in the creation of the DID to ensure that the right data was being collected and that data collection was aligned as much as possible with existing systems so as to minimize the burden of data collection. The DID developed was a negotiation between the stakeholders and continues to be a living document, changing as new situations are encountered.

The DID armed the contractors with a complete picture of what data was to be collected and how that data was to be collected. But data collection takes time and effort. Add to this the fact that, from a contractor’s perspective, there may be no apparent benefit to

the Army collecting productivity data; in fact it may seem like more of a risk than a benefit. It was important to motivate the contractor to participate in this effort and to create a comfort zone around the data collection effort to assuage concerns that data will be used for evil rather than good. Toward this end a data collection requirement around the DID was added to the Contract Deliverable Requirements List (CDRL). In parallel with this there was an education effort focused on making the contractor personnel comfortable that the goal was improved, defendable estimating based on data collected across many programs rather than an effort to identify productivities of specific contractors.

The DID requires a great deal of data to be collected. In order to make it possible for the contractor to collect this data consistently and efficiently it was necessary and prudent to introduce as much automation into the process as possible. Several tools were developed or extended to facilitate this automation. When collecting software project data, counting code is a significant chore. Originally it was thought that the code counter developed by the University of Southern California's Center for Software Excellence (USC CSE) would be adequate for this project. Due to the sheer volume of the data, the requirement to track data to capabilities and the fact that there were some languages not handled by the USC code counter, it was necessary to build a wrapper in Excel to facilitate consistent counting of new, modified, reused and deleted code. Additionally an Excel based Software Resource Report (SRR) form was developed to collect data. It guided data collection with appropriate drop downs, copy paste alignment with the output of the code counting tool and support for mapping activities and resources from contractor's designations to more generic categories. Figure1 is a snapshot of the code count by capability feature of the SRR.

Section 2.5 PRODUCT SIZE REPORTING								(Column will not print)	
SECTION 2.5.1 Capability Requirement Name	SECTION 2.5.2 Standardized Capability Name	SECTION 2.5.3 Language	SECTION 2.5.4 % Auto Generated	SECTION 2.5.5 New Code	SECTION 2.5.6 Deleted Code	SECTION 2.5.7 Modified Code	SECTION 2.5.8 Unmodified Code	SECTION 2.5.9 Application Type	SECTION 2.5.10 Estimated % of Total Effort (Build)
CR 65774 Reduced System Footprint - SCAMT Multiple Applications on Single Node	Complex Capability	Java		1,891	3,740	628	2,492,491	Assessment/Analysis Functions - Simulation [Constructive]	12.00%
CR 65774 Reduced System Footprint - SCAMT Multiple Applications on Single Node	Complex Capability	XML		63	36	36	8,242,067	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67136 Crowd Formation	Moderate Capability	Java		978	-	-	2,499,444	Assessment/Analysis Functions - Simulation [Constructive]	3.00%
CR 67136 Crowd Formation	Moderate Capability	XML		779	-	-	8,208,573	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67136 Crowd Formation	Moderate Capability	XLS File		121	-	-	815,757	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67136 Crowd Formation	Moderate Capability	CSV file		7	-	-	276,976	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67154 Dynamic UHRB Aperture Consistent Ordering	Complex Capability	C++		12	41	34	673,807	Assessment/Analysis Functions - Simulation [Constructive]	2.00%
CR 67154 Dynamic UHRB Aperture Consistent Ordering	Complex Capability	Java		238	87	98	2,496,600	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67156 Effects Guidance Matrices	Moderate Capability	Java		1,791	10	15	2,497,419	Assessment/Analysis Functions - Simulation [Constructive]	35.00%
CR 67156 Effects Guidance Matrices	Moderate Capability	XML		7	-	3	8,209,143	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67156 Effects Guidance Matrices	Moderate Capability	XLS File		14,066	-	-	813,585	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67156 Effects Guidance Matrices	Moderate Capability	CSV file		24	-	-	276,972	Assessment/Analysis Functions - Simulation [Constructive]	
CR 67164 Raytrace Feature Composition	Complex Capability	C++		982	4	35	676,299	Assessment/Analysis Functions - Simulation [Constructive]	13.00%

Figure 1: Snapshot of the SRR

Progress to Date

After several test runs in the lab, data collection began around the middle of 2012 and occurs every ten weeks. The first real iteration of data was thrown away because of automation failures in the contractor configurations and misunderstandings around some of the data definitions that came to light in practice. Additional meetings of the IPT, both in person and remotely, were required to iron out questions and uncertainties. To date three successful iterations have occurred, each resulting in slight refinements to the process, DID and/or SRR. Practice exercises using the data to calibrate TruePlanning, a commercial software estimating model, have led to additional changes to the data collection process and DID. Currently the IPT team is focused on determining the best long term strategy for configuration and storage of this data in a form that makes cost and technical information accessible on a capability basis.

Lessons Learned

A data collection effort for a complex program, involving cooperation across multiple organizations, spanning several years is not going to go off without a hitch. Issues such as technology glitches, competing agendas, office politics, and personalities all create barriers. Turning these barriers into opportunities is an on-going adventure. It is really important that the entire team understand the motivation for data collection. Not only does this alleviate concerns about being 'measured' it also helps facilitate discussions about what data to collect and how best to collect it.

Completely key to the success of this pilot was the ability to inject automation into processes that would otherwise have been manual and tedious. Not only does automation create efficiencies but it also ensures consistency in a process that, if manual, would be fraught with error, especially when there is schedule pressure (and when isn't there). It is important to note that in this pilot the data collection did not cost the program anything – there was no uptick on the contract. So creating automation to ease the contractor's burden was essential.

Communication and team work were paramount to success. Bringing everyone to the table and forming an IPT made it possible to iron out misunderstandings, disagreements and disputes. All kinds of issues arose throughout this pilot such as: whether we should count physical or logical lines of code, does every line of code in the base need to be included in the count, should we count XLS and CSV Files (a great deal of the simulation data is loaded into the system in this format), should we include auto-

generated code in our counts, etc. Having representation from all the stakeholders at the IPT made it possible to resolved these issues thoughtfully and practically.

It's also important to maintain an element of flexibility because not everything is possible or practical. In this pilot it was possible for us to do code counts by capability but the contractor's time keeping system did not facilitate tracking hours to capabilities. It was determined that expert judgment from the contractor personnel would allocate hours to specific capabilities at each ten week period.

Automotive Data Collection Pilot

Motivation

The motivation for this pilot is a complete turnaround from the software pilot discussed above. In this case we have an auto manufacturing that makes low volume luxury vehicles and collects vast quantities of cost and technical data at the part level for each automobile produced. The data is collected by several disparate systems and is stored in various files and formats that have little or no connections to one another.

Despite the plethora of data, the auto manufacturer is in a position where they are unable to predict the cost of a new or modified vehicle line until about two months prior to the day the first one rolls out of the plant. They would like to be able to predict at (or near) concept what the cost of the new or modified vehicle will be using a data driven question based estimating approach. The goal of the pilot was to focus on one specific aspect of a vehicle to determine whether the data supports the development of a parametric estimating model, along the way identifying what processes and tools need to be added or changed to make it feasible to harvest, mine and analyze the large, unwieldy set of data they currently collect and store.

Implementation and Process

This project also had its fits and starts. Initial meetings were held with the customer to better understand their mission, their business processes and their lexicon. Understanding their business processes helped facilitate the identification of sources of potential data. Cost and technical data is collected for every vehicle at a sub-system and part level. The original plan was to focus on a small subset of these parts from various subsystems to prove the concept. It was soon obvious that data at the part level was way too down in the weeds to support estimation at a concept level. The focus of the pilot shifted to a single sub-system, the seat assembly

Having identified the target for our pilot, the next step was to identify cost drivers. Perusal of the data and discussions with the customer led us to examine a set of feature codes associated with the individual vehicles that are related to the seat assembly. The feature codes did indeed seem to be the right path to pursue but presented completely new challenges. Every vehicle has its own unique set of feature codes. Feature codes are used to identify one unique characteristic of the seat such as the material for the seat covers or whether or not they are bench seats. As an example every vehicle will have one of three feature codes to indicate that the seat adjustment is manual, electric or electric with Memory. To create a question based estimation each related set of feature codes needed to be combined to create a single question with a set of potential answers. The effort was further complicated by the fact that collecting the necessary set of cost and feature code information for a single vehicle required visiting at least three separate files.

Clearly the project screamed for automation. After a first pass of analysis, the team sat down with a developer to outline the manual process for collecting the data and aligning the feature codes. An application was developed that made it possible to harvest cost and part information at a sub-assembly level and to align the costs with the feature codes for further analysis. A target for future automation would be the aggregation of related feature codes into a single question but this will require collaboration with the customer to understand the many aspects of the various codes. The automation tool also provided some visualization capability to view cost trends as demonstrated in Figure 2

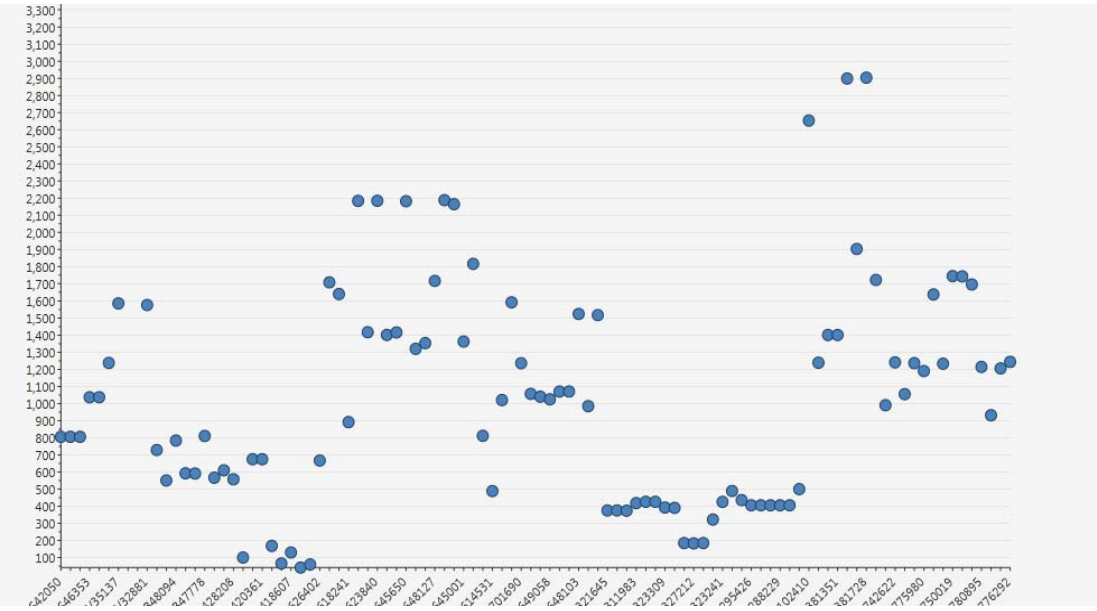
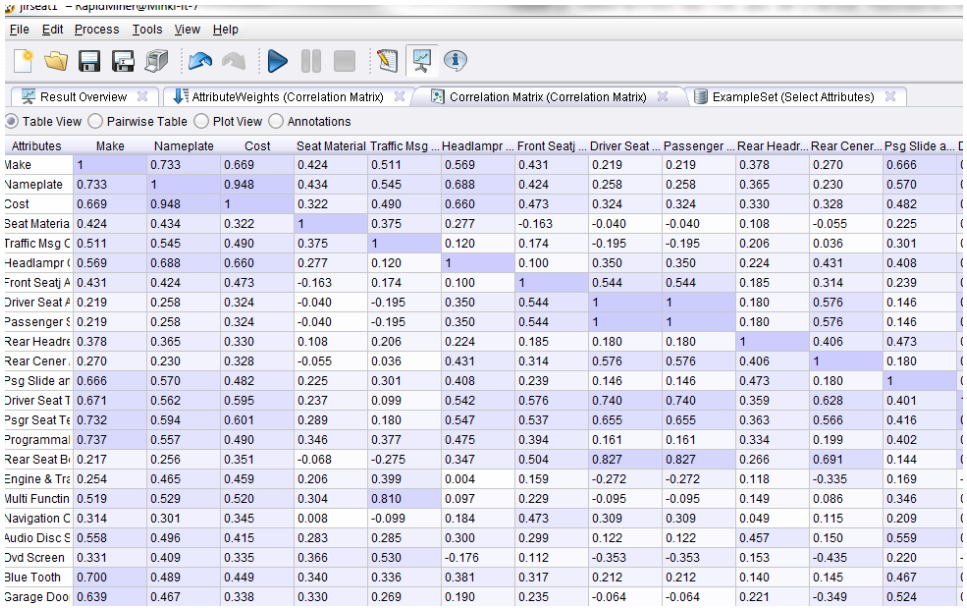


Figure 2: Data visualization

Progress to Date

With automation simplifying data harvesting, actual analysis could begin. Data was collected for one hundred vehicles across various models and automation was applied to extract and align cost and feature information. At the end of this process was a spreadsheet with 100 rows and 44 potential cost drivers. Data mining techniques were applied to identify which features were most strongly correlated to cost. The Rapid Miner Data Mining application was chosen for this initial analysis. Figure 3 presents an example of the type of correlation matrix one might use to identify cost drivers in a large data set – color coding is used to highlight good correlations.


File Edit Process Tools View Help

Result Overview | AttributeWeights (Correlation Matrix) | Correlation Matrix (Correlation Matrix) | ExampleSet (Select Attributes)

Table View | Pairwise Table | Plot View | Annotations

Attributes	Make	Nameplate	Cost	Seat Material	Traffic Msg ...	Headlampr ...	Front Seatj ...	Driver Seat ...	Passenger ...	Rear Headr...	Rear Cener...	Psg Slide a...
Make	1	0.733	0.669	0.424	0.511	0.569	0.431	0.219	0.219	0.378	0.270	0.666
Nameplate	0.733	1	0.948	0.434	0.545	0.688	0.424	0.258	0.258	0.365	0.230	0.570
Cost	0.669	0.948	1	0.322	0.490	0.660	0.473	0.324	0.324	0.330	0.328	0.482
Seat Materia	0.424	0.434	0.322	1	0.375	0.277	-0.163	-0.040	-0.040	0.108	-0.055	0.225
Traffic Msg C	0.511	0.545	0.490	0.375	1	0.120	0.174	-0.195	-0.195	0.206	0.036	0.301
Headlampr (	0.569	0.688	0.660	0.277	0.120	1	0.100	0.350	0.350	0.224	0.431	0.408
Front Seatj A	0.431	0.424	0.473	-0.163	0.174	0.100	1	0.544	0.544	0.185	0.314	0.239
Driver Seat A	0.219	0.258	0.324	-0.040	-0.195	0.350	0.544	1	1	0.180	0.576	0.146
Passenger S	0.219	0.258	0.324	-0.040	-0.195	0.350	0.544	1	1	0.180	0.576	0.146
Rear Headre	0.378	0.365	0.330	0.108	0.206	0.224	0.185	0.180	0.180	1	0.406	0.473
Rear Cener.	0.270	0.230	0.328	-0.055	0.036	0.431	0.314	0.576	0.576	0.406	1	0.180
Psg Slide ar	0.666	0.570	0.482	0.225	0.301	0.408	0.239	0.146	0.146	0.473	0.180	1
Driver Seat T	0.671	0.562	0.595	0.237	0.099	0.542	0.576	0.740	0.740	0.359	0.628	0.401
Psgr Seat Te	0.732	0.594	0.601	0.289	0.180	0.547	0.537	0.655	0.655	0.363	0.566	0.416
Programmal	0.737	0.557	0.490	0.346	0.377	0.475	0.394	0.161	0.161	0.334	0.199	0.402
Rear Seat Bi	0.217	0.256	0.351	-0.068	-0.275	0.347	0.504	0.827	0.827	0.266	0.691	0.144
Engine & Tra	0.254	0.465	0.459	0.206	0.399	0.004	0.159	-0.272	-0.272	0.118	-0.335	0.169
Multi Functin	0.519	0.529	0.520	0.304	0.810	0.097	0.229	-0.095	-0.095	0.149	0.086	0.346
Navigation C	0.314	0.301	0.345	0.008	-0.099	0.184	0.473	0.309	0.309	0.049	0.115	0.209
Audio Disc S	0.558	0.496	0.415	0.283	0.285	0.300	0.299	0.122	0.122	0.457	0.150	0.559
Dvd Screen	0.331	0.409	0.335	0.366	0.530	-0.176	0.112	-0.353	-0.353	0.153	-0.435	0.220
Blue Tooth	0.700	0.489	0.449	0.340	0.336	0.381	0.317	0.212	0.212	0.140	0.145	0.467
Garage Doo	0.639	0.467	0.338	0.330	0.269	0.190	0.235	-0.064	-0.064	0.221	-0.349	0.524

Figure 3: Sample correlation matrix from Rapid Miner

From the correlation matrix, five top drivers were identified, their values were digitized and regression analysis provided a CER, which has been implemented in the TruePlanning framework for interface with the front end question based estimation portal developed for this pilot. There is general acknowledgement that the fidelity of the CER is in question because the sample set was so small, especially compared to the large set of potential cost drivers. The point of the pilot was not to develop a great CER but rather to determine whether it was worth the effort to proceed with the larger data set. Recommendations have been made for streamlining the data collection process and additional automation targets have been recommended. Once these issues have been addressed, large scale data harvesting and analysis will commence for the rest of the sub-systems in the vehicles.

Lessons Learned

At the risk of sounding like a broken record, automation was clearly a key factor for success in this pilot as well as the earlier described effort. The first pass at data

harvesting was a very manual exercise taking several weeks to accomplish. The addition of automation pared analysis time from several weeks to two and a half days. This was completely acceptable for the pilot where there were only one hundred data points and forty four potential cost drivers but moving on to larger data sets more automation is necessary to make similar analysis possible. Furthermore, analysis of large quantities of data requires visualization and data mining automation. Tools such as Rapid Miner, R and Excel are essential to help focus the analysis on the most likely cost drivers and weed out the noise/

Communication is an extremely important tool. In this case it was important not only to understand the customer's mission but also to understand their business and the lexicon they use when referring to their vehicles and features. Initial versions of the automation tool had erroneously highlighted several features completely unrelated to the seat as potential cost drivers. One needed to understand what the feature codes mean to recognize such an error. Automation and visualization are powerful tools but taken in a vacuum without thoughtful analysis and discussion, they can be dangerous.

Having lots of data is a problem we would all like to have. Having lots of data in lots of different places with different alignments and cryptic labeling is just a problem. The first step to successful use of data for cost estimation is to have it be accessible, understandable and possess a high degree of quality. Serious thought should go into existing and future data collection processes and policies with respect to how well they do or do not support data driven estimation.

Conclusions

The need for data driven estimation is driving many organizations in Aerospace and Defense, as well as industry in general to look for new ways to collect and use data. This is really nothing new, we have been using data (real or experiential) to justify our estimates for as long as we have been estimating. The new requirement is transparency and openness about the data and how it is being used.

To support this requirement, organizations may find the need to establish or improve formal processes for collecting, harvesting and analyzing their project data. This paper highlights two such organizations who have realized that they want their data to drive their estimates but recognized that processes and tools needed to be in place to make this happen. Both of these organizations are seeing some level of success but recognize there is a long road ahead. The pilot projects described have resulted in the development and deployment of a set of tools and processes they will need as they travel this road.

.

[1] http://www.rand.org/about/history.html

[2] https://www.acquisition.gov/comp/aap/documents/Chapter1.pdf