Develop JMP 16_Based STEAMS and Six Sigma DMAIC Training Curriculum for Data Scientist

Mason Chen
Stanford Online High School
San Jose, CA 95131
Mason.chen.training@gmail.com

Charles Chen
Morrill Learning Center
San Jose, CA 95131
Charles.chen.training@gmail.com

Abstract

Traditional Six Sigma curriculums including DMAIC, DFSS, DFR, Lean are not developed specifically and effectively for today's AI Data Science fields. This project will demonstrate an innovative Six Sigma training curriculum for Data Scientists. There are several objectives in this modern Six Sigma training curriculum: (1) adopt modern Text Mining and Data Mining techniques on Root Cause and Problem Solving Analyses such as DFMECA, C&E, QFD, SIPOC, VSM, (2) integrate various JMP statistical platforms holistically to analyze pattern recognition and discover the insights, (3) enhance predictive modeling capability through Neural, Partition, Principal Component algorithms, (4) utilize modern Quality and Process platforms such as Goal Plot, Model-Driven Multivariate SPC, Process History Process and Screening to enhance production quality control, (5) map these modern JMP Platforms into the Six Sigma DMAIC/DFSS/Lean framework for facilitating the Six Sigma Project execution, and (6) integrate database through Query Builder to monitor the production status in real-time. This curriculum is not just designed for a Data Scientist but is also powerful for those working in process, quality, reliability, supply chin, business, statistics, and marketing fields who want to become a reliable decision maker and true project leader. This Six Sigma DMAIC for Data Scientist Approach can be commonly applied to other Quality Engineering, Project Management and Business Excellence methodology such as DFSS, DFR, Lean, 8D, PMP.

Keywords
Six Sigma, DMAIC, STEAMS, JMP, Data Science

1. Introduction
Traditional Six Sigma methodology and Interdisciplinary "STEM" (Science, Technology, Engineering, Math) or "STEAM" (Science, Technology, Engineering, Art, Math) are popular in School Education is a term used to group together these academic disciplines Colwell (2016).

1.1 Six Sigma DMAIC
DMAIC is the problem-solving approach that drives Lean Six Sigma. It's a five-phase method—Define, Measure, Analyze, Improve and Control—for improving existing process problems with unknown causes. DMAIC is based on the Scientific Method. DMAIC is a data-driven quality strategy used to improve processes. The letters in the acronym represent the five phases that make up the process, including the tools to use to complete those phases. It is an integral part of a Six Sigma initiative, but in general can be implemented as a standalone quality improvement procedure or as part of other process improvement initiatives such as lean.

1.2 STEM Methodology
This STEM term is typically used when addressing education policy and curriculum choices in schools to improve competitiveness in science and technology development. It has implications for workforce development, national security concerns and immigration policy. The acronym came into common use shortly after an interagency meeting

on science education held at the US National Science Foundation. In the early 1990's, a summer program called STEM Institute is arranged for talented under-represented students in the Washington, DC area. Based on the program's recognized success and expertise in STEM education NSF (2018) that NSF was first introduced to the acronym STEM.

1.3 Criticism of STEM Methodology

The focus on increasing participation in STEM fields has attracted many criticisms. The efforts of the U.S. government to increase the number of STEM graduates, the science and engineering occupations have been flat or slow-growing, and unemployment as high or higher than in many comparably-skilled occupations. The STEM Crisis Is a Myth": there was a "mismatch between earning a STEM degree and having a STEM job in the United States, with only around ¼ of STEM graduates working in STEM fields, while less than half of workers in STEM fields have a STEM degree Charette (2013). Based on the data, science should not be grouped with the other three STEM categories, because, while the other three generally result in high-paying jobs, "many sciences, particularly the life sciences, pay below the overall median for recent college graduates Casselman (2014). Efforts to remedy the perceived domination of STEM subjects has led to intense efforts to diversify the STEM workforce. Some critics feel that this practice in higher education, as opposed to a strict meritocracy, causes lower academic standards MacDonald (2018).

1.4 STEAM Methodology

STEAM fields are Science, Technology, Engineering, Art, and Math. STEAM is designed to integrate STEM subjects into various relevant education disciplines. These programs aim to teach students innovation, to think critically and use engineering or technology in imaginative designs or creative approaches to real-world problems while building on students' mathematics and science base Jolly (2016). STEAM programs add art to STEM curriculum by drawing on design principles and encouraging creative solutions Pomeroy (2016), Eger (2016).

1.5 Artificial Intelligence and Digital Art

In the modern Big Data Society, Artificial intelligence (AI) is becoming a new dominant Data Science. AI sometimes called machine intelligence or machine learning, is intelligence demonstrated by machines, in contrast to the natural intelligence displayed by humans Schank (1991). In computer science AI research is defined as the study of "intelligent agents": any device that perceives its environment and takes actions that maximize its chance of successfully achieving its goals Paul (2006). Artificial intelligence was founded as an academic discipline in 1956, and in the years since has experienced several waves of optimism. Digital art is an artistic work or practice that uses digital technology as an essential part of the creative or presentation process. Since the 1970s, various names have been used to describe the process, including computer art and multimedia art Lieser (2009), Paul (2006). Recently, a lot of articles are about AI in art and design: a lot of the feature imagery, unsurprising as they're largely created by people from a science rather than arts background. And their research is often presented in the ultra-detailed format of scientific papers - heavy on words and, strangely to us. This digital art application of AI is being referred to as 'generative'. Generative AI will drive the next generation of apps for auto-programming, content development, visual arts, and other creative, design, and engineering activities. By 2019, most leading AI providers will offer tools and libraries for building AI-powered natural-language generation, image manipulation, and other generative use cases. Already, generative AI has proven itself - in both research and in commercial applications. In Generative graphics: AI can abstract visual patterns from artwork and then apply those patterns in the fanciful re-rendering of photographic images with the hallmark features of that artwork Kobielus (2017).

1.6 STEAMS Methodology

Instead of using classical STEM or STEAM, here, a new holistic “STEAMS” methodology is introduced. There are several novel concepts embedded in this new “STEAMS” methodology: (1) replace “Art” with “Artificial Intelligence”, (2) separate “Statistics” from “Math”, and (3) integrate all six “STEAMS” elements. “STEAMS” (Science, Technology, Engineering, Artificial Intelligence, Math, Statistics) methodology will be demonstrated. Authors have conducted STEAMS Chocolate project. The chemistry “Science” studied was cocoa bean nutrition, flavonoids, flavanols, and antioxidants. “Technology” is the chocolate manufacturing process to produce the commercial chocolate products from coca beans. Systematic “Engineering” problem solving techniques such as 5 whys and SIPOCs were deployed to understand the root cause analysis. “Artificial Intelligence” clustering algorithms were utilized to recognize the clustering patterns hidden among Chocolate nutritions and products. Clustering Distance “Math” was utilized to explain the clustering algorithm. Distribution “Statistics” was conducted to visualize the effect of using “Artificial Intelligence” and “Math” algorithms. All 6 STEAMS elements are critical to making

this Chocolate research project successful. The following sections will provide more details for each "STEAMS" element.

2. Develop Six Sigma Data Science Curriculum

In Section 2, the concept of developing a special Six Sigma Data Science curriculum was introduced based on STEAMS methodology, DMAIC phases and JMP 16 platforms.

2.1 STEAMS, DMAIC, Data Science Diagram

Modern Data Science (Artificial Intelligence, Mathematics, and Statistics) can help both STEAMS and DMAIC methodologies and their connections as shown in Figure 1. Modern Data Science techniques could be closely related to Artificial Intelligence, Mathematics, and Statistics in STEAMS methodology and mapped to Six Sigma DMAIC phases. There is a huge opportunity and need to innovate Six Sigma DMAIC curriculum through incorporating the modern Data Science elements.

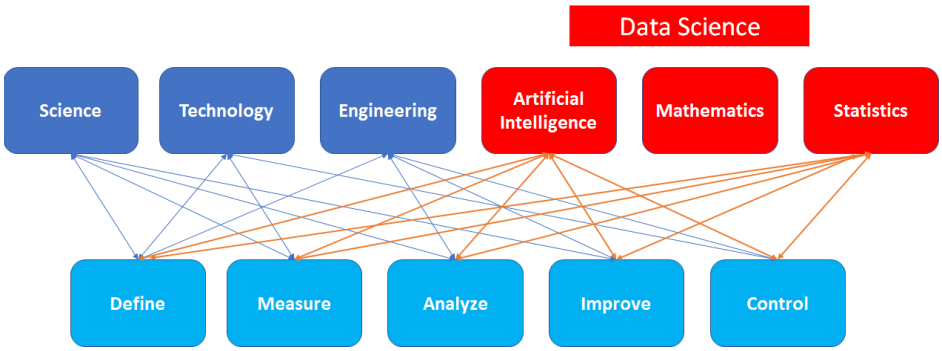


Figure 1. STEAMS, DMAIC, Data Science Diagram

2.2 JMP 16 Data Science Platforms

Identify the powerful Data Science platforms available in JMP 16 software as the one of the critical steps to develop an innovative Six Sigma Data Science curriculum as shown in Figure 2. The JMP platform selection criteria for Data Science curriculum is based on Big Data 3Vs: Volume, Variety, and Velocity.

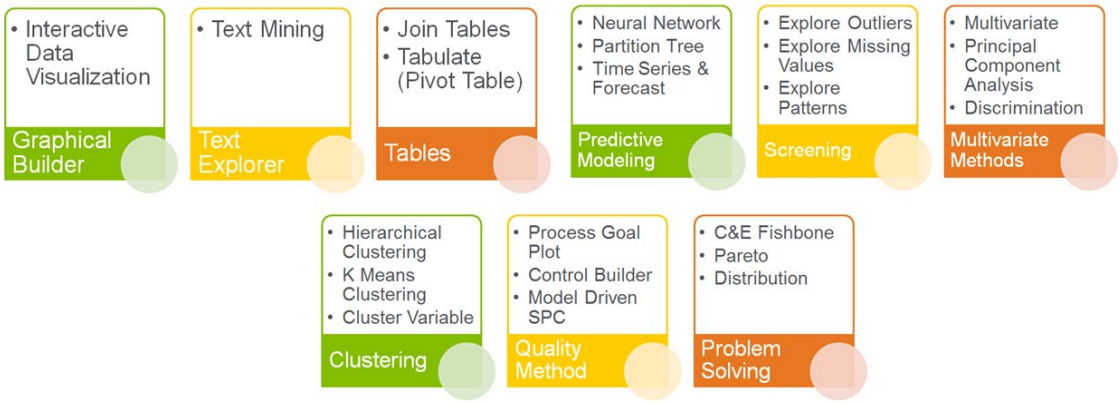


Figure 2. JMP 16 Data Science Platforms

2.3 Data Science Statistics

Learning data science statistics is very critical for building the foundation of developing an effective Six Sigma Data Science training program. In addition to traditional DMAIC Quality and Reliability Statistics and DFSS Design Modeling Statistics, Data Science curriculum should add the Linear Algebra Matrix Statistics, Data Mining Statistics, Time Series and Forecasting Statistics, Text Mining Analytics, Survey and Consumer Research Statistics as shown in Figure 3. These modern or more advanced statistics can help Data Scientist analyze complicated Big Data.

© IEOM Society International

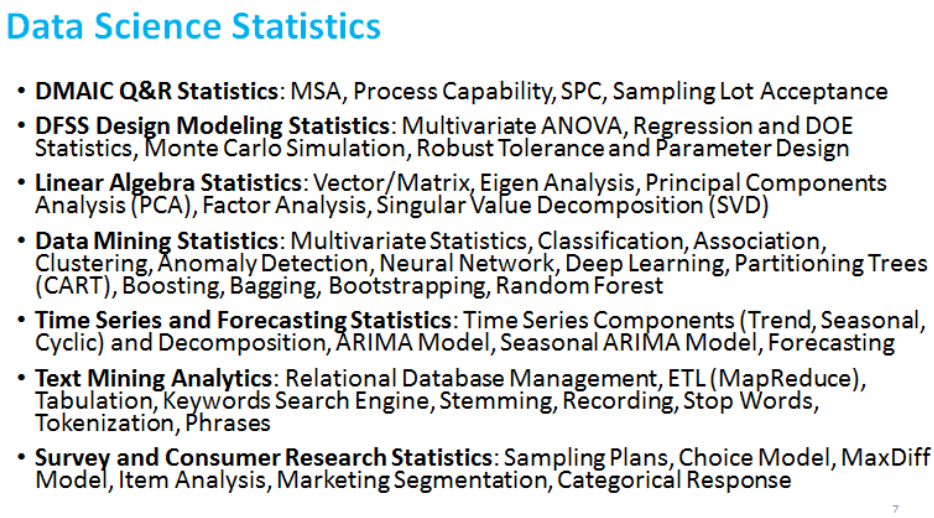


Figure 3. Data Science Statistics

2.4 Develop Six Sigma Data Science Curriculum

Six Sigma Data Science curriculum can be split into three stand alone programs: (A) Regular DMAIC Black Belt, (B) Data Mining, and (C) Text Mining and Categorical as shown in Figure 4. Three customized programs can help identify the trainee candidates and meet their training needs better. In the regular DMAIC BB program, trainees can learn more JMP Data Science Platforms related to Quality and Reliability functions. The B Data Mining program may fit better to data miner analyzing more continuous Database. The C Text Mining program may fit better to text miner or analyzing the categorical database.

JMP 16 Platforms	A. Regular DMAIC BB	B. Data Mining	C. Text Mining and Categorial
LSS BB 03 Statistics	Basic Statistics, Distributions, Prospective Sample Size and Power, Sample Size Explorer	Basic Data Science Statistics	
LSS BB 04 JMP Introduction	Discovering JMP Book	Using JMP Book	
LSS BB 09 Measure M2 MSA	MSA Design, Variability and Attribute Gauge Charts,		
LSS BB 10 Measure M3 PCA	Process Capability	Quality Utility, Process History Explorer	
LSS BB 13 Analyze A2	Essesstial Graphing Book Part I, Pareto Chart, Cause and Effect Diagram	Essential Graphing Book Part II	Essential Graphing Book Part II, FMEA Plus
LSS BB 14 Analyze A3	Bivariate, Oneway,Continegency, Tabulate, Modeling Utilities, Fit Model (LS, Stepwise), Matched Pairs	Tabulate Plus, Modeling Utilities Plus, Multivariate Correlations, Principal Componets, Discriminat, Partial Least Square, Factor Analysis, Multi Dimensional Scaling, Item Analysis, Hierarchical Clustering, K Means Clustering, Normal Mixture Clustering, Cluster Variables	Tabulate Plus, Text Explorer, Modeling Utilities Plus, Multiple Correspondence, Two-Way Hierarchical Clustering, Latent Class Clustering, Categorical Response
LSS BB 16 Improve I1	Prediction Profiler	Custom Profiler, Excel Profiler, Multiple Factor Analysis	Choice Design, Choice Model, MaxDiff Design, MaxDiff Model,
LSS BB 18 Improve I3	Custom DOE, DSD, Simulator	Neural Network, Partition,Time Series Analysis and Forecast, Response Screening, Process Screening, Predictor Screening	Neural Network, Partition
LSS BB 20 Control A1 SPC	Control Chart Builder	Multivariate and Model Driven Control Chart	

Figure 4. Six Sigma Data Science Training Programs

3. DMAIC Data Science JMP Platforms

In Section 3, authors would emphasize the Data Science concept and application in each DMAIC phases through utilizing the JMP platforms as the critical part of Six Sigma training curriculum.

3.1 Define Phase

The main objectives of Define Phase is to derive the CTQ (critical to quality) through the problem statement, voice of the customers (VOC), and voice of the business (VOB) impact. CTQ must be specific and measurable with a well-defined success criteria. Data Science techniques can significantly help the Define Phase with data-driven approach.

For example, JMP data science such as Consumer Research, Clustering can help identify the marketing segment and translate the VOC and VOB to CTQ as shown in Figure 5. Multi-dimensional Visualization platforms such as Graphical Builder, 2-Dimensional Pareto Chart, Bubble Plot, Variability Plot can help visualize the comprehensive data patterns. Certain Data Mining Hand (2001), Hastie (2009) and Text Mining can help discover the insights of the database. Consumer Choice and Maxdiff Design platforms can help survey analysis to understand more about the VOC and VOB objectively. These JMP Data Science platforms can help Define Phase on defining the Project Scope and Objective through defining CTQs well.

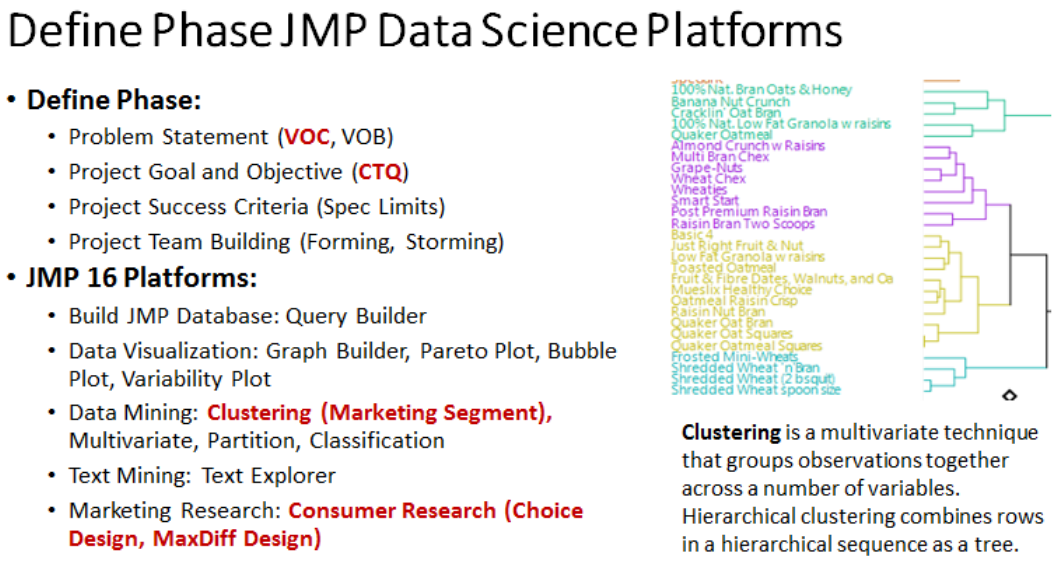


Figure 5. Define Phase JMP Data Science Platforms

3.2 Measure Phase JMP Data Science Platforms

The main objectives of the Measure Phase are: Measurement System Analysis (MSA) and Process Capability Analysis (PCA). There is little Data Science relevancy to MSA. Though, Data Science techniques can help Process Capability Analysis through several JMP platforms as shown in Figure 6. Process Capability Goal Plot is a powerful visualization tool to plot the Cpk and Ppk process capability index of large scale MFG Production data (different parameters or lots) Bissell (1990).. The goal plot can decompose process capability index to two components: (1) meeting the target (X-horizontal) and dispersion (Y-vertical). It's a great snap shot to visualize the production process capability. The middle Process Performance plot is another great process capability plot to address both Process Capability (Y-vertical) and Process Stability (X-horizontal) together. The left Process History Explorer can summarize the historical production process capability. All these three listed JMP platforms above can help measure the Production Data more effectively through JMP Data Science platforms.

© IEOM Society International

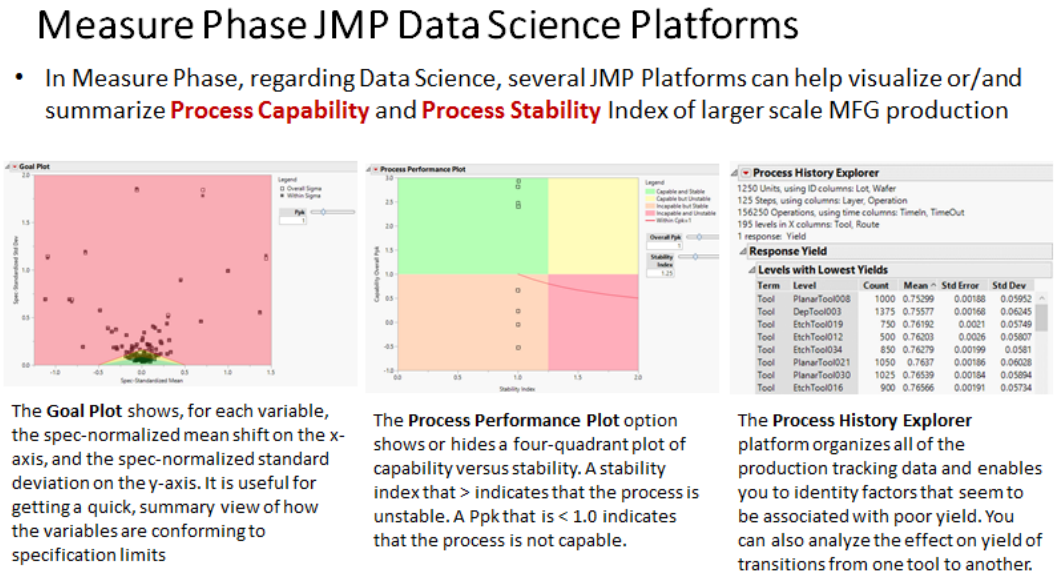


Figure 6. Measure Phase JMP Data Science Platforms

3.3 Analyze Phase JMP Data Science Platforms

The main objectives of traditional Analyze Phase are to identify the root causes and validate the root causes. In addition to C&E fishbone diagram, there are many powerful JMP Data Science Platforms can help the Root Cause Analysis and Problem Solving Process as shown in Figure 7. JMP Tabulate Platform, similar to Excel Pivot Table, can summarize the database through multidimensional and hierarchical layering table format with add-in statistics metric available. JMP Text Explorer platform can search the keywords and help identify the failure modes or root causes. JMP Multivariate methods and platforms can utilize Eigen analysis such as principal component analysis Candes (2009) and singular value decomposition Jackson (2003). to discover the hidden patterns Anderson (1958). Clustering platforms can split the data or variables into homogeneous clusters based on similarity or entropy to group the similar terms for further root cause analysis Hartigan (1981), Milligan (1980).. JMP Modeling Utilities can identify the multi-dimensional outliers through the Euclidean Distance algorithm Benford (1938), Penny (1996), Rousseeuw (1987). Consumer Categorical Response platform can help cluster the Categorical data. All these listed JMP Data Science platforms can help conduct the Analyze Phase to identify and validate the potential root causes and vital Few Xs.

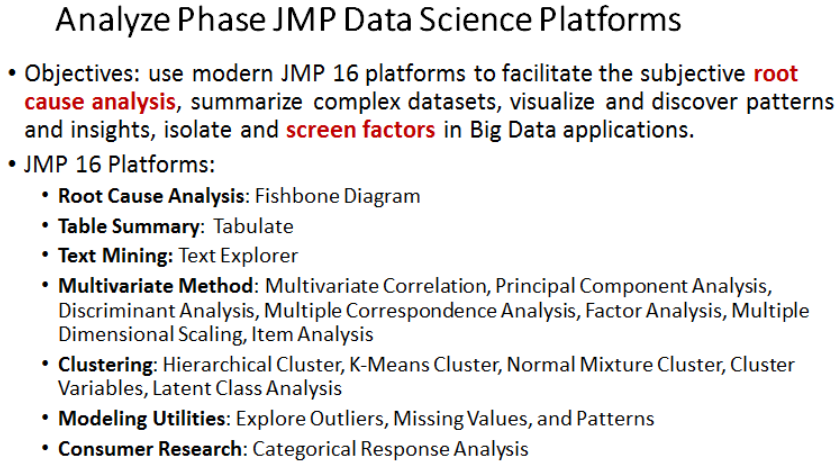


Figure 7. Analyze Phase JMP Data Science Platforms

3.4 Improve Phase Data Science Platforms

The objectives of the Improve Phase are to find and implement the optimal solutions to improve process or optimize the design. Traditional Regression Belsley (1980), Hocking (1985), and DOE modeling techniques are very powerful to analyze structured experimental data to avoid confounding risks. Though, in today's Big Data environment, modern Data Science techniques may provide different thinking for the Improve Phase as shown in Figure 8. Prediction Profiler or Custom Profiler can be powerful on building the model on the unstructured database. Parameter and Robust Design can simulate the integrated or accumulated system variance through Monte Carlo simulation. Consumer Research can consider the Survey DOE such as Choice Design and MaxDiff Design to provide the best product based on VOCs. Group Orthogonal Super Saturated Design can block the variables from upstream to downstream to reduce the experimental sample size while sustaining a good orthogonal structure. Specialized Data Mining modeling techniques such as Neural Network, Partitioning, Screening Platforms can help the Improve Phase to verify or/and validate process improvement or design optimization. All these modern JMP Data Science platforms can facilitate the Improve Phase efficiency.

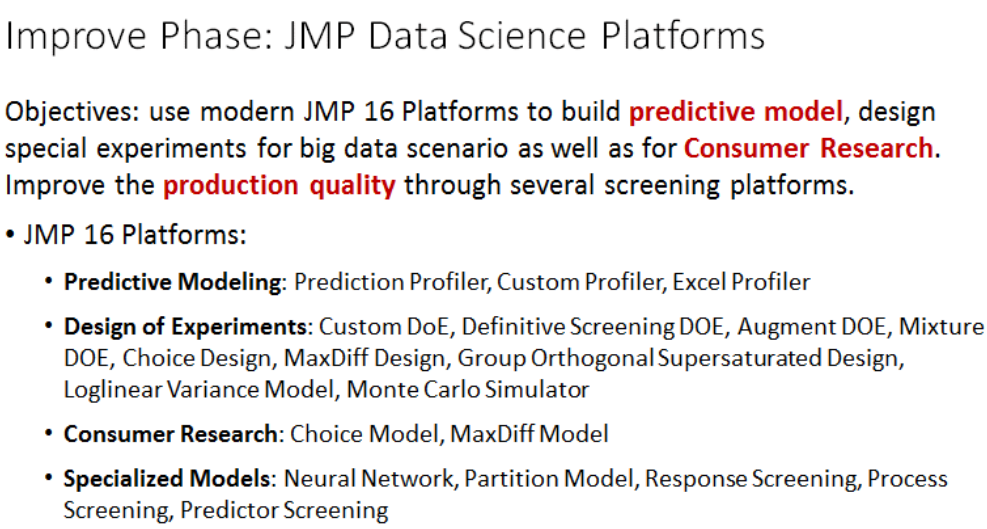


Figure 8. Improve Phase JMP Data Science Platforms

3.5 Control Phase Data Science Platforms

The main objectives of the Control Phase is to establish the control method or system to sustain the improvement for a ling term success. Traditional Univariate Statistical Process Control Chart (SPC) was very popular and powerful in traditional Six Sigma community AIAG (2005), Nelson (1984, 1985, 1996), Wheeler (2004). Certain JMP Data Science platforms may further enhance the Control Phase to deal with multivariate database. As shown in Figure 9, Multivariate Control Chart Mason (2002), Nomikos (1995), Tracy (1992), and Model Driven Multivariate Control Chart Kourti (1996) can combine many univariate control charts into one Multivariate T-Square Control Chart while decomposing the failure mode patterns simultaneously. Multiple Factor analysis can also help identify the critical factors to be considered in the Multivariate Control Chart to enhance statistical Signal-Noise ratio. Time Series and Forecasting Box (1994), Hyndman (2008, 2018) platforms can also handle the time series data which may be consisted of Trend and Season components Baillie (2002), Shiskin (1967). These JMP Data Science platforms can facilitate the Control Phase by monitoring the right factors and detecting the process stability through multivariate platforms.

© IEOM Society International

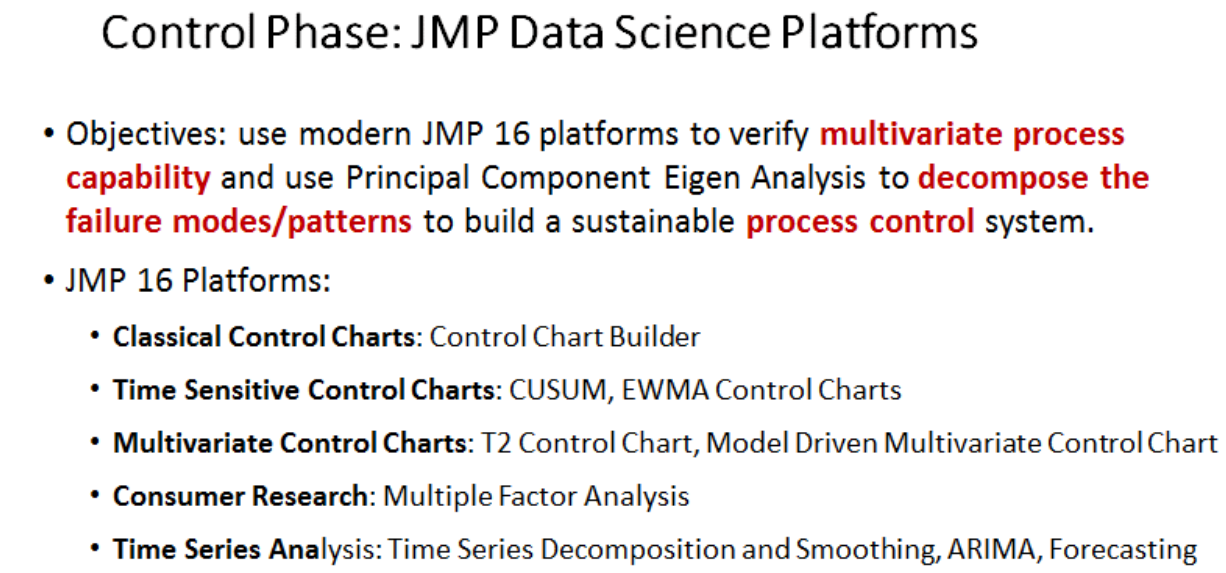


Figure 9. Control Phase JMP Data Science Platforms

4 Conclusions

Traditional Six Sigma DMAIC and Interdisciplinary STEAMS methods can help develop Data Scientist on leadership and team building. JMP 16 platforms are mapped to DMAIC Phases to help execute Six Sigma Projects in Data Science fields. Database Management, Applied Engineering Statistics, Data Mining and Text Mining are all critical to today's Data Scientific Analytics.

Future Work

Authors are continue current Antarctic Glacier project: Map JMP 16 Data Science Platforms to Six Sigma DFSS, Lean, DFR training curriculum.

Acknowledgements

Thanks to IEOM STEM Co-Chairs Dr. Ali and Dr. Reimer.

References

AIAG (2005). *Statistical Process Control.* 2nd ed. Troy, MI: Automotive Industry Action Group.
Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.
Baillie, RT, Chung, SK, 2002, "Modeling and forecasting from trend-stationary long memory models with applications to climatology", *International Journal of Forecasting*. Volume 18, Issue 2, Pages 215-226
Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.
Benford, F. (1938). "The law of anomalous numbers." *Proceedings of the American philosophical society*, 551–572.
Bissell, A. F. (1990). "How Reliable is Your Capability Index?" *Applied Statistics* 30:331–340.
Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
Candes, E. J., Li, X., Ma, Y., and Wright, J. (2009). "Robust Principal Component Analysis?" *Journal of the ACM*, 58:1–37.
Casselman, Ben (2014). "The Economic Guide to Picking A College Major". *FiveThirtyEight*.
Charette, Robert N. (2013). "The STEM Crisis Is a Myth". *IEEE Spectrum*.
Christiane Paul (2006). *Digital Art*, pp. 7–8. Thames & Hudson.
Colwell Rita, (2016) *Compare: "Guest commentary: A "STEM" in Collier County to reach their future".*
Eger, John. (2016) "National Science Foundation Slowly Turning STEM to STEAM". *Huffington Post*.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.
Hartigan, J. A. (1981). "Consistency of Single Linkage for High–Density Clusters." *Journal of the American Statistical Association* 76:388–394.
Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag.
Hocking, R. R. (1985). *The Analysis of Linear Models*. Monterey, CA: Brooks/Cole.
Hyndman, R.J., Athanasopoulos, G. 2018, *Forecasting: Principles and Practice*, 2nd
Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer-Verlag.
Jackson, J. E. (2003). *A User's Guide to Principal Components*. Hoboken, NJ: John Wiley & Sons
Jolly, Anne. (2016) "STEM vs. STEAM: Do the Arts Belong?". EdWeek.org. *Education Week: Teacher*.
Kobielus James (2017). *InfoWorld* | "How generative AI is changing art and design"
Kourti, T., and MacGregor, J. F. (1996). "Multivariate SPC Methods for Process and Product Monitoring." *Journal of Quality Technology* 28:409–428.
Lieser, Wolf. (2009) *Digital Art*. Langenscheidt: h.f. ullmann. pp. 13–15
MacDonald, Heather (2018). "How Identity Politics Is Harming the Sciences". *City Journal*. Manhattan Institute.
Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications.* Philadelphia: SIAM.
Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms." *Psychometrika* 45:325–342.
Nelson, L. (1984). "The Shewhart Control Chart—Tests for Special Causes." *Journal of Quality Technology* 15:237–239.
Nelson, L. (1985). "Interpreting Shewhart X Control Charts." *Journal of Quality Technology* 17:114–116.
Nelson, P. R. C., Taylor, P. A., and MacGregor, J. F. (1996). "Missing Data Methods in PCA and PLS: Score calculations with incomplete observations." *Chemometrics and Intelligent Laboratory Systems* 35:45–65.
Nomikos, P., and MacGregor, J. F., (1995). "Multivariate SPC Charts for Monitoring Bach Processes." *Technometrics* 37:41–59.
NSF (2018), President Bush Honors Excellence in Science, Mathematics and Engineering Mentoring | NSF - National Science Foundation".
Penny, K. I. (1996). "Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance." *Journal of the Royal Statistical Society*, Series C 45:73–81.
Pomeroy, Steven Ross. (2016) "From STEM to STEAM: Science and Art Go Hand-in-Hand". *Scientific American*.
Rousseeuw, P. J., and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.
Schank, Roger C. (1991). "Where's the AI". *AI magazine*. Vol. 12 no. 4. p. 38.
Shiskin, J., Young, A. H., and Musgrave, J. C. (1967). *The X-11 Variant of the Census Method II Seasonal Adjustment Program. Technical Report 15*, US Department of Commerce, Bureau of the Census.
Tracy, N. D., Young, J. C., and Mason, R. R. (1992). "Multivariate Control Charts for Individual Observations." *Journal of Quality Technology* 24:88–95.
Wheeler, D. J. (2004). *Advanced Topics in Statistical Process Control.* 2nd ed. Knoxville, TN: SPC Press.