STEAMS and DMAIC Curriculum for Data Scientists Using JMP 16©

Mason Chen, Charles Chen, Patrick Giuliano

Stanford Online High School STEAMS Club
Overview

- **Opportunity statement** – the traditional Six Sigma DMAIC process combined with the interdisciplinary STEAMS methodology can help data scientists make greater contributions in the field of Big Data.

- **Project objective** – develop a Six Sigma data science training curriculum for high schoolers to industry professionals by mapping JMP 16© platforms onto DMAIC phases.
Case Study: Mason Chen’s Learning Experience

- 2015 **Big Data Statistics** Summer Camp (10 years old)
- 2016 May IBM **SPSS** Statistics Certified (10 years old)
- 2016 August IASSC **Minitab DMAIC Black Belt** Certified (11 years old)
- 2016 August ASQ/ASA/JMP Joint Annual **STEAMS** Speaker (11 years old)
- 2016 October IBM **Modeler Data Mining** Certified (11 years old)
- 2017 April IEOM Rabat **DMAIC EV3** Robotics Best Paper Award (11 years old)
- 2017 April IEOM Rabat **Java** Best Paper Award (11 years old)
- 2017 August Found **STEAMS Organization** (12 years old)
- 2018 October **JMP** USA DS Best Contributed Paper Award (13 years old)
- 2019 Youngest **IEEE** Presenter, **JMP** Principal Component & Clustering (13 years old)
- 2020 **JMP STIPS** Certification- **Data Mining** (14 years old)
- 2020 Learning **JMP** DOE Cert Exam (15 years old)
- 2021 March **JMP** Europe Discovery Summit Best Student Poster Award (15 years old)
- 2021 Learning **JMP 16** Text Mining and Time Series Forecast (15 years old)
- 2021 June Stanford Summer Course: **Linear Algebra** (16 years old)
- 2021 August Stanford OHS **Data Science R** Course (16 years old)
- **2021 August JMP 16_Based Six Sigma Data Science Program (16 years old)**
- 2021 September **R-Based** Six Sigma DMAIC Statistics Curriculum (16 years old)
Connecting STEAMS and DMAIC

Science
Technology
Engineering
Artificial Intelligence
Mathematics
Statistics

Define
Measure
Analyze
Improve
Control

Data Science
Data Science JMP 16 Platforms

- Map JMP Platforms to Data Science Certification Program
- Based on Big Data 3Vs: Volume, Variety, and Velocity.
• DMAIC quality and reliability – measurement systems analysis (MSA), process capability, statistical process control (SPC), lot acceptance sampling
• DFSS design modeling – analysis of variance (ANOVA), regression, design of experiment (DOE), Monte Carlo simulation, robust tolerance
• Linear algebra – eigen analysis, principal component analysis (PCA), factor analysis, singular value decomposition (SVD)
• Data mining – classification, neural network, partition trees, random forest
• Time series and forecasting – time series decomposition, autoregressive integrated moving average (ARIMA) models, forecasting
• Text mining – stemming, recoding, tokenization, phrases
• Survey and consumer research – sampling plans, choice model, MaxDiff model, marketing segmentation
### Six Sigma DMAIC Data Science Curriculum

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

Three training programs – standard, data mining, and text mining
Define Phase

**Overview**
- Problem statement (voice of the customer, voice of business)
- Project goal and objective (critical to quality)
- Success criteria (specification limits)
- Team building (forming, storming, norming, performing)

**JMP 16© platforms**
- Build JMP database – query builder
- Data visualization – graph builder, Pareto plot, bubble plot, variability plot
- Data mining – clustering, multivariate, and partition methods
- Marketing Research: Consumer Research
Measure Phase

- Several JMP Platforms can help visualize or/and summarize **Process Capability** and **Process Stability** Index of larger scale MFG production
Analyze Phase

- **Overview**
  - Root cause analysis
  - Summarize complex datasets
  - Visualize and discover patterns and insights
  - Isolate and screen for important factors

- **JMP 16 platforms**
  - Root cause analysis – fishbone diagram
  - Table summary – tabulate
  - Text mining – text explorer
  - Multivariate methods – multivariate correlation, factor analysis
  - Clustering – hierarchical clustering, k-means clustering, cluster variables
  - Survey and consumer research – categorical response analysis
Analyze: Identifying the Root Cause

Pareto plot

Highlights the severity of different problems

Fishbone diagram

Brainstorm and organize sources of the problem
Analyze: Data Summarization

Tabulate
Descriptive statistics and pivot tables

Text Explorer
Analyzes patterns between unstructured text

Multiple Correspondence Analysis
Associations between categorical levels
Improve Phase

- **Overview**
  - Build predictive models
  - Design new experiments
  - Improve production quality

- **JMP 16e platforms**
  - Predictive modeling – prediction profiler, custom profiler
  - Design of experiment (DOE) – custom DOE, mixture DOE, group orthogonal supersaturated designs, augmentation
  - Specialized models – neural network, partition model, response screening, process screening, predictor screening
  - Survey and consumer research – choice model, MaxDiff design
Improve: Design Optimization

**Prediction Profiler**
Studies response distribution and factor sensitivity

**Custom Profiler**
Finds optimal factor settings without graphs

**Group Orthogonal Supersaturated**
For designs with a greater number of factors than runs
Improve: Predictive Models

Neural network
Uses a transfer function to predict response variables

Partition
Creates a decision tree by recursively partitioning data
Improve: Design Optimization

Response Screening
Features that aid the analysis of large datasets

Process Screening
Process capability and stability for many responses

Predictor Screening
Ranks predictors using bootstrap forest partitioning
Improve: Consumer Research

Choice Design
Used to find the best combination of features

MaxDiff Design
Only considers most and least preferred items
Control Phase

- **Overview**
  - Scale-Up Process Control
  - Sustain Improvement over long period
  - Upstream-Downstream Multivariate Process Control

- **JMP 16® platforms**
  - **Classical Control Charts**: Control Chart Builder
  - **Time Sensitive Control Charts**: CUSUM, EWMA Control Charts
  - **Multivariate Control Charts**: T2 Control Chart, Model Driven Multivariate Control Chart
  - **Consumer Research**: Multiple Factor Analysis
  - **Time Series Analysis**: Time Series Decomposition and Smoothing, ARIMA, Forecasting
Control: Multivariate Tools

Change Point Detection plot
Detects a shift in the mean by dividing the data

T Square chart
Uses principal components for process stability

Multiple factor analysis
Uses eigenvalue decomposition to compare items
Control: Time Series Techniques

Model diagnostics
Identify trend, seasonal, and cyclic components

ARIMA models
Fits data using seasonal or non-seasonal methods

Forecasting
Finds the optimal model to predict future points
Takeaways

• Traditional **Six Sigma DMAIC** and Interdisciplinary **STEAMS** methods can help develop Data Scientist on leadership and team building

• **Modern JMP 16** platforms are mapped to DMAIC Phases to help deploy 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
Thanks!