STEAMS and DMAIC Curriculum for Data Scientists Using JMP 16©

Mason Chen, Charles Chen,
Patrick Giuliano

Stanford Online High School
STEAMS Club

Overview

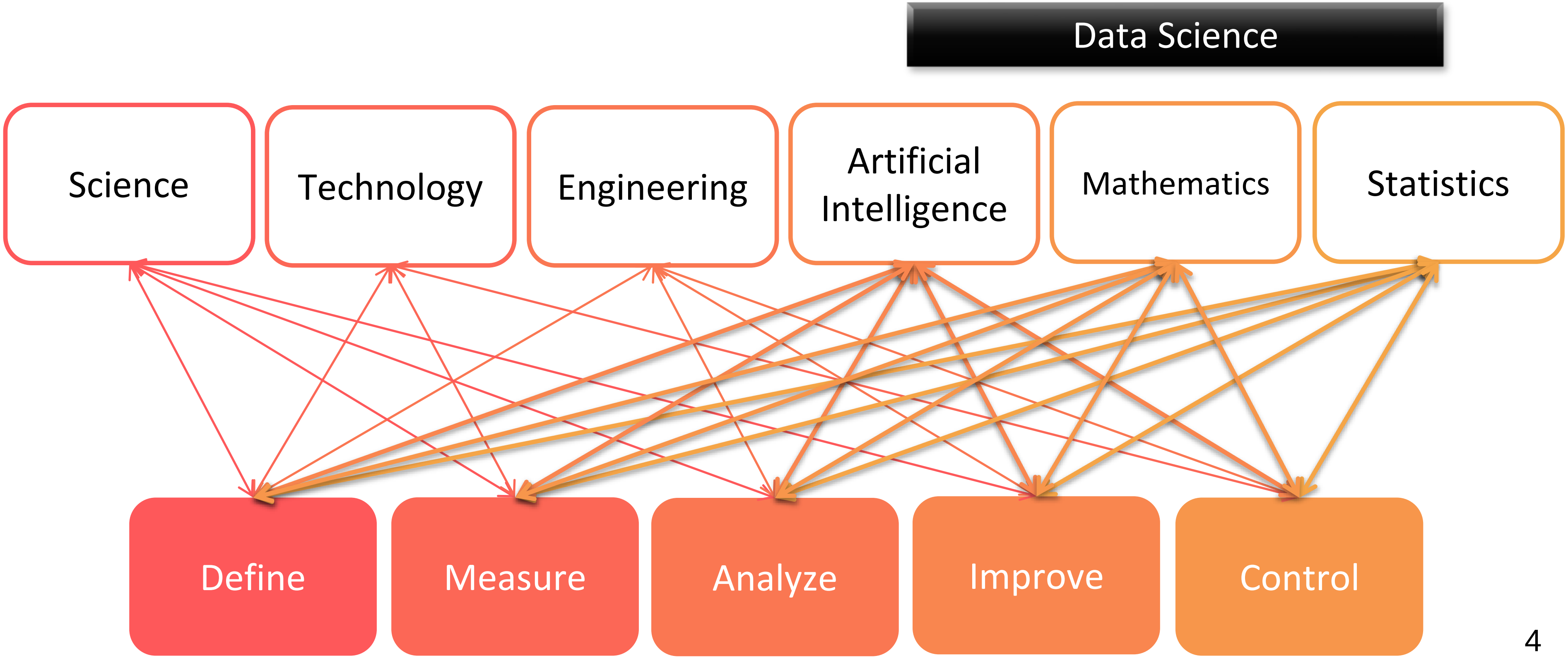
- **Opportunity statement** – the traditional Six Sigma DMAIC process combined with the interdisciplinary STEAMS methodology can help data scientists make greater contributions in the field of Big Data
- **Project objective** – develop a Six Sigma data science training curriculum for high schoolers to 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)

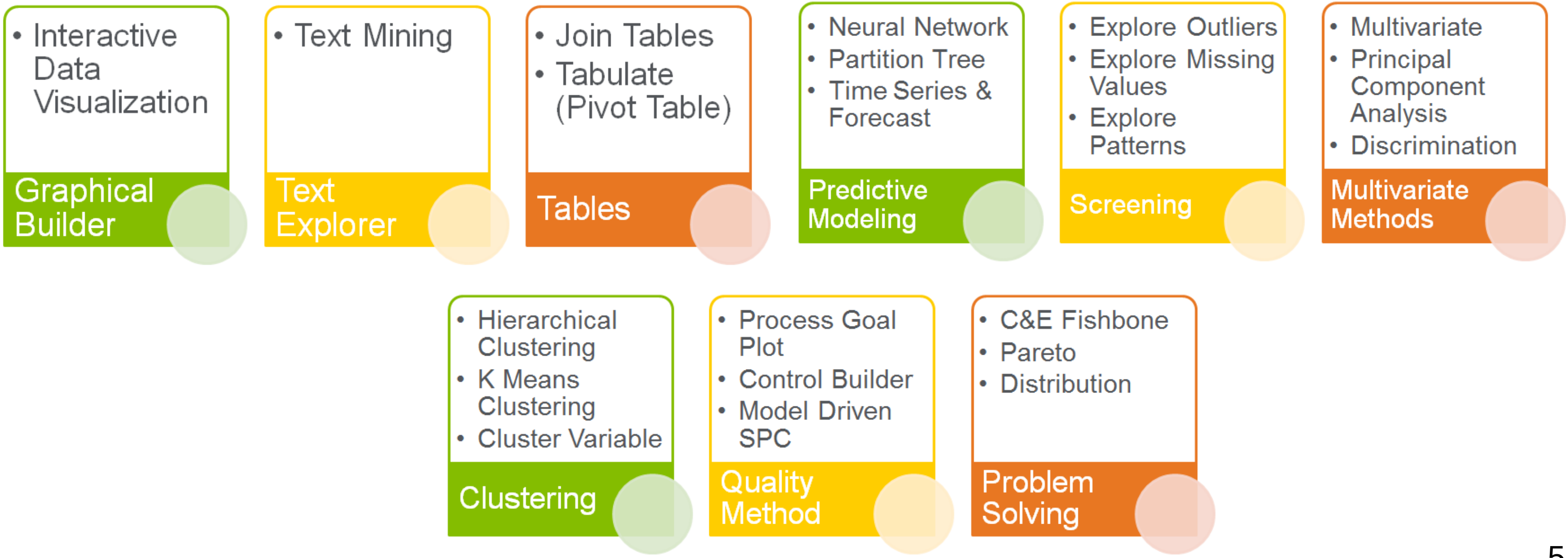
2019 ASA JSM Denver Conference at Rocky Mountain Summit (12,005 ft)

Connecting STEAMS and DMAIC



Data Science JMP 16 Platforms

- Map JMP Platforms to Data Science Certification Program
- Based on Big Data 3Vs: Volume, Variety, and Velocity.



Graphical Builder
- Interactive Data Visualization

Text Explorer
- Text Mining

Tables
- Join Tables
- Tabulate (Pivot Table)

Predictive Modeling
- Neural Network
- Partition Tree
- Time Series & Forecast

Screening
- Explore Outliers
- Explore Missing Values
- Explore Patterns

Multivariate Methods
- Multivariate
- Principal Component Analysis
- Discrimination

Clustering
- Hierarchical Clustering
- K Means Clustering
- Cluster Variable

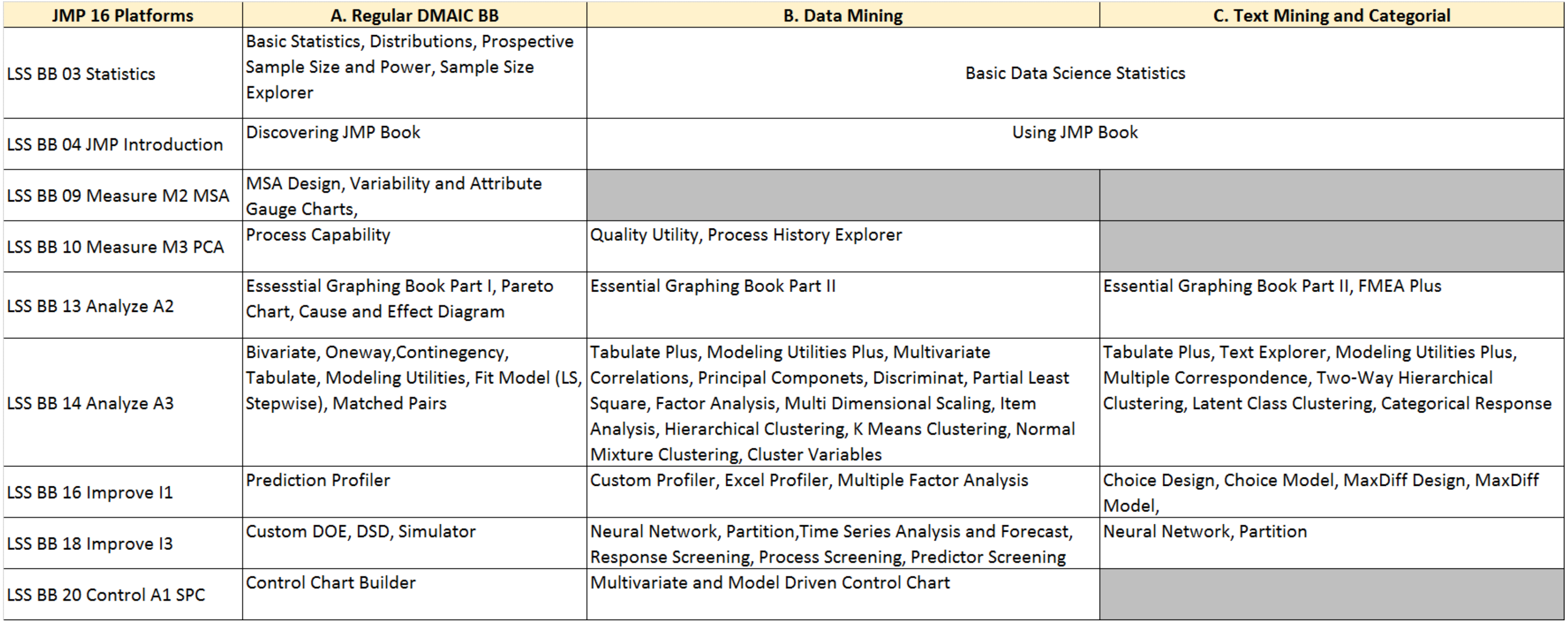
Quality Method
- Process Goal Plot
- Control Builder
- Model Driven SPC

Problem Solving
- C&E Fishbone
- Pareto
- Distribution

JMP 16 Data Science Statistics

- **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



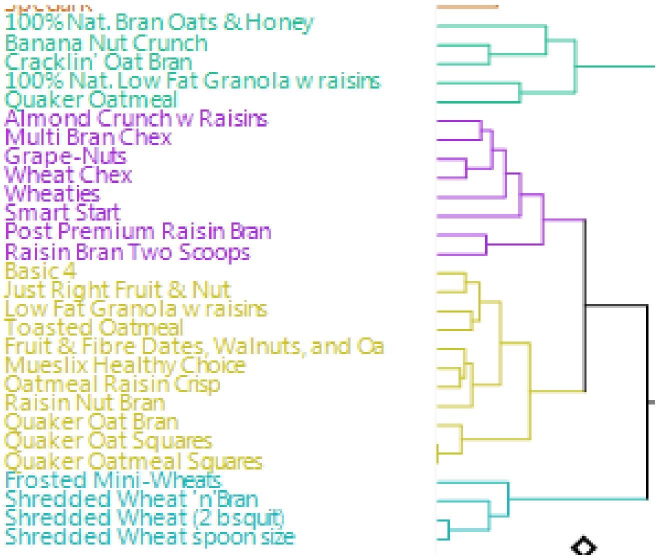
JMP 16 Platforms	A. Regular DMAIC BB	B. Data Mining	C. Text Mining and Categorial
LSS BB 03 Statistics	Basic Statistics, Distributions, Prospective Sample Size and Power, Sample Size Explorer	Basic Data Science Statistics	
LSS BB 04 JMP Introduction	Discovering JMP Book	Using JMP Book	
LSS BB 09 Measure 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	

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

Define Phase

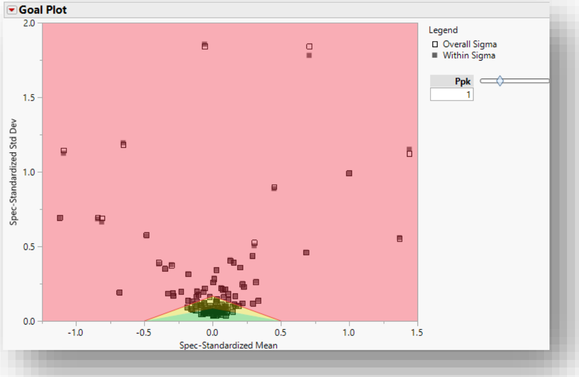
- **Overview**
  - Problem statement (voice of the customer, voice of business)
  - Project goal and objective (critical to quality)
  - Success criteria (specification limits)
  - Team building (forming, storming, norming, performing)

- **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



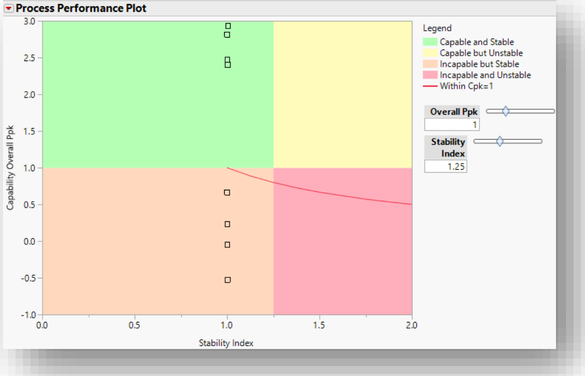
Measure Phase

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



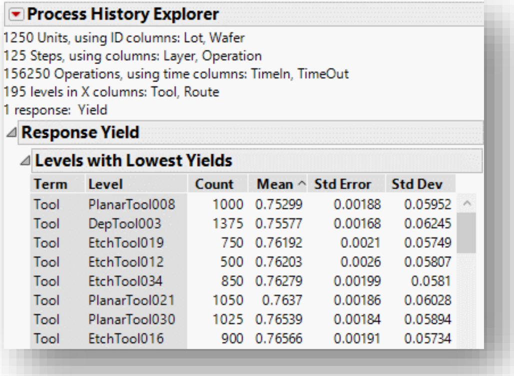
Goal Plot

How well variables are conforming to specification limits



Process Performance Plot

Divided based on process capability and stability



Process History Explorer

1250 Units, using ID columns: Lot, Wafer
125 Steps, using columns: Layer, Operation
156250 Operations, using time columns: TimeIn, TimeOut
195 levels in X columns: Tool, Route
1 response: Yield

Response Yield

Levels with Lowest Yields

Term	Level	Count	Mean ^	Std Error	Std Dev
Tool	PlanarTool008	1000	0.75299	0.00188	0.05952
Tool	DepTool003	1375	0.75577	0.00168	0.06245
Tool	EtchTool019	750	0.76192	0.0021	0.05749
Tool	EtchTool012	500	0.76203	0.0026	0.05807
Tool	EtchTool034	850	0.76279	0.00199	0.0581
Tool	PlanarTool021	1050	0.7637	0.00186	0.06028
Tool	PlanarTool030	1025	0.76539	0.00184	0.05894
Tool	EtchTool016	900	0.76566	0.00191	0.05734

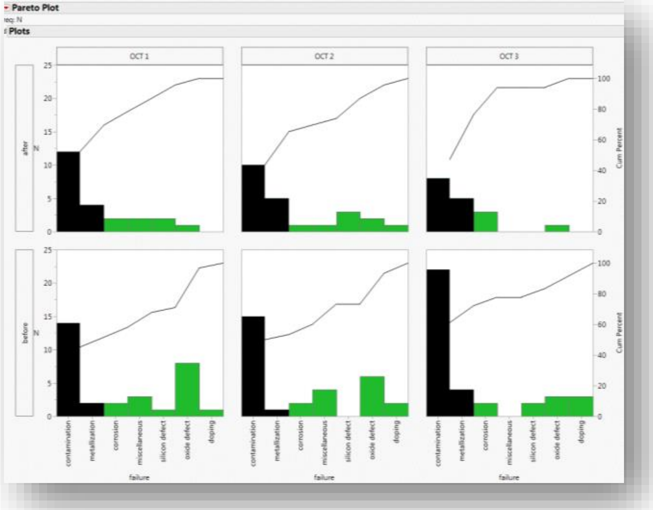
Process History Explorer

Helps identify factors associated with poor yield

Analyze Phase

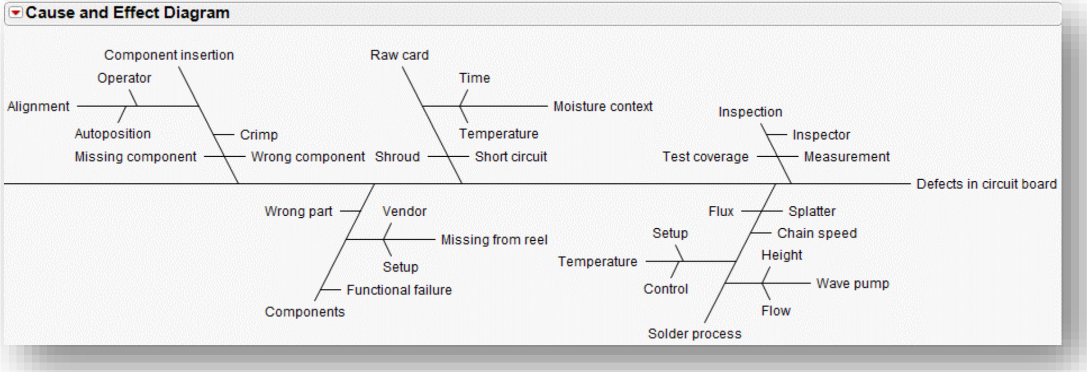
- **Overview**
  - Root cause analysis
  - Summarize complex datasets
  - Visualize and discover patterns and insights
  - Isolate and screen for important factors
- **JMP 16© platforms**
  - Root cause analysis – fishbone diagram
  - Table summary – tabulate
  - Text mining – text explorer
  - Multivariate 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

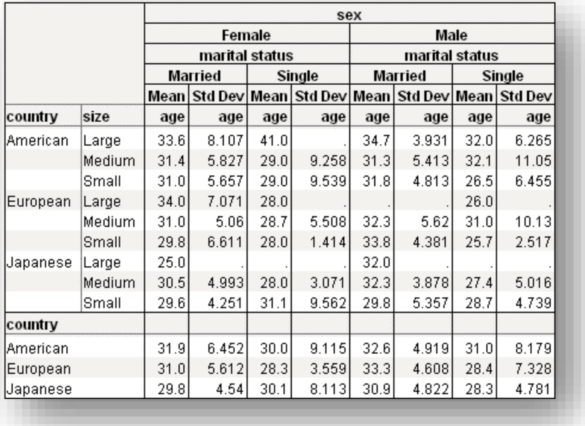
Highlights the severity of different problems



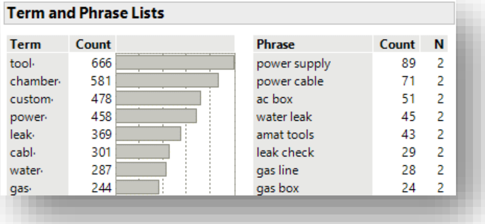
Fishbone diagram

Brainstorm and organize sources of the problem

Analyze: Data Summarization



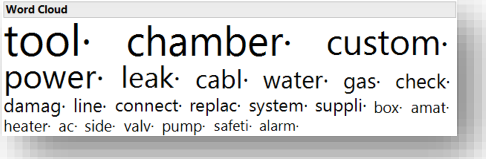
		sex							
		Female				Male			
		marital status				marital status			
		Married		Single		Married		Single	
		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
country	**size**	**age**	**age**	**age**	**age**	**age**	**age**	**age**	**age**
American	Large	33.6	8.107	41.0	.	34.7	3.931	32.0	6.265
	Medium	31.4	5.827	29.0	9.258	31.3	5.413	32.1	11.05
	Small	31.0	5.657	29.0	9.539	31.8	4.813	26.5	6.455
European	Large	34.0	7.071	28.0	.	.	.	26.0	.
	Medium	31.0	5.06	28.7	5.508	32.3	5.62	31.0	10.13
	Small	29.8	6.611	28.0	1.414	33.8	4.381	25.7	2.517
Japanese	Large	25.0	.	.	.	32.0	.	.	.
	Medium	30.5	4.993	28.0	3.071	32.3	3.878	27.4	5.016
	Small	29.6	4.251	31.1	9.562	29.8	5.357	28.7	4.739
country									
American		31.9	6.452	30.0	9.115	32.6	4.919	31.0	8.179
European		31.0	5.612	28.3	3.559	33.3	4.608	28.4	7.328
Japanese		29.8	4.54	30.1	8.113	30.9	4.822	28.3	4.781



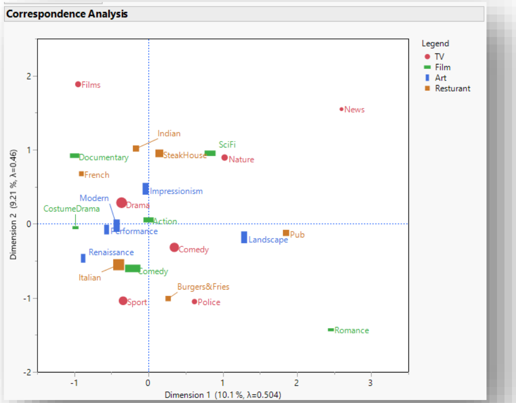
Term and Phrase Lists

Term	Count
tool·	666
chamber·	581
custom·	478
power·	458
leak·	369
cabl·	301
water·	287
gas·	244

Phrase	Count	N
power supply	89	2
power cable	71	2
ac box	51	2
water leak	45	2
amat tools	43	2
leak check	29	2
gas line	28	2
gas box	24	2



Word Cloud



Correspondence Analysis

Tabulate

Descriptive statistics and pivot tables

Text Explorer

Analyzes patterns between unstructured text

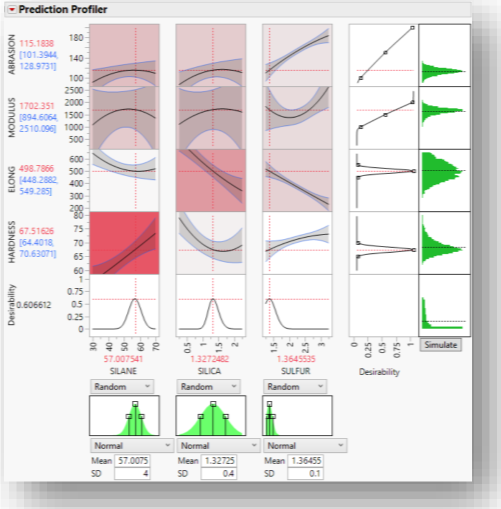
Multiple Correspondence Analysis

Associations between categorical levels

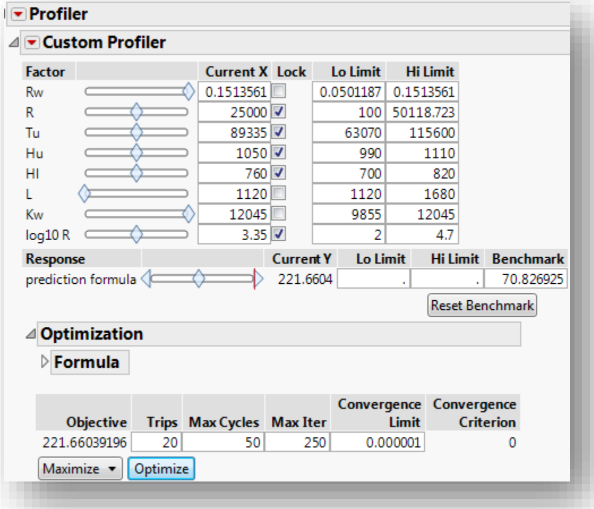
Improve Phase

- **Overview**
  - Build predictive models
  - Design new experiments
  - Improve production quality
- **JMP 16© 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



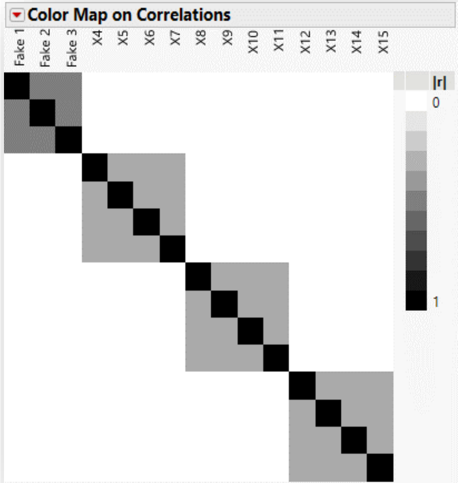
	SILANE	SILICA	SULFUR
	Random	Random	Random
	Normal	Normal	Normal
Mean	57.0075	1.32725	1.36455
SD	4	0.4	0.1



Factor	Current X	Lock	Lo Limit	Hi Limit
Rw	0.1513561	☐	0.0501187	0.1513561
R	25000	☑	100	50118.723
Tu	89335	☑	63070	115600
Hu	1050	☑	990	1110
Hl	760	☑	700	820
L	1120	☐	1120	1680
Kw	12045	☐	9855	12045
log10 R	3.35	☑	2	4.7

Response	Current Y	Lo Limit	Hi Limit	Benchmark
prediction formula	221.6604	.	.	70.826925

Objective	Trips	Max Cycles	Max Iter	Convergence Limit	Convergence Criterion
221.66039196	20	50	250	0.000001	0



Prediction Profiler

Studies response distribution and factor sensitivity

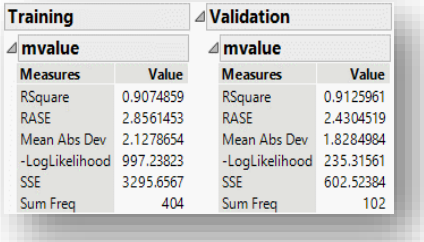
Custom Profiler

Finds optimal factor settings without graphs

Group Orthogonal Supersaturated

For designs with a greater number of factors than runs

Improve: Predictive Models



Training

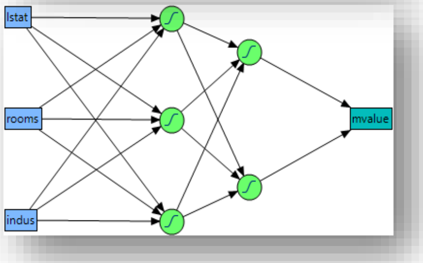
mvalue

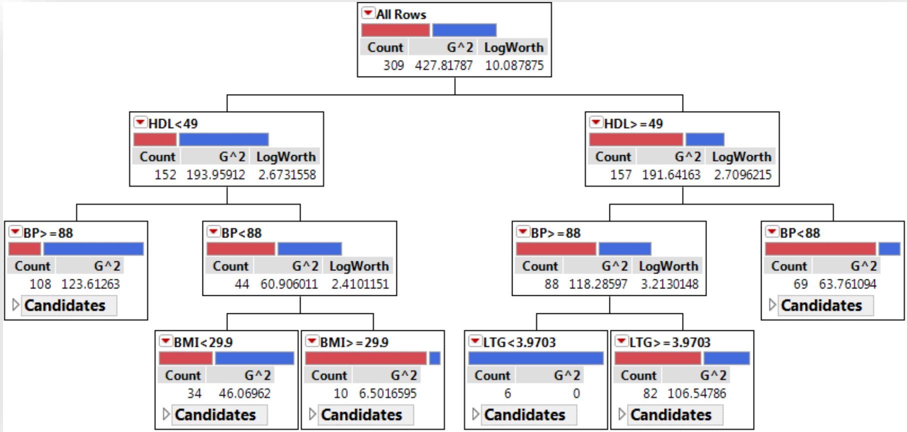
Measures	Value
RSquare	0.9074859
RASE	2.8561453
Mean Abs Dev	2.1278654
-LogLikelihood	997.23823
SSE	3295.6567
Sum Freq	404

Validation

mvalue

Measures	Value
RSquare	0.9125961
RASE	2.4304519
Mean Abs Dev	1.8284984
-LogLikelihood	235.31561
SSE	602.52384
Sum Freq	102





All Rows

Count	G^2	LogWorth
309	427.81787	10.087875

HDL<49

Count	G^2	LogWorth
152	193.95912	2.6731558

HDL>=49

Count	G^2	LogWorth
157	191.64163	2.7096215

BP>=88

Count	G^2
108	123.61263

Candidates

BP<88

Count	G^2	LogWorth
44	60.906011	2.4101151

BP>=88

Count	G^2	LogWorth
88	118.28597	3.2130148

BP<88

Count	G^2
69	63.761094

Candidates

BMI<29.9

Count	G^2
34	46.06962

Candidates

BMI>=29.9

Count	G^2
10	6.5016595

Candidates

LTG<3.9703

Count	G^2
6	0

Candidates

LTG>=3.9703

Count	G^2
82	106.54786

Candidates

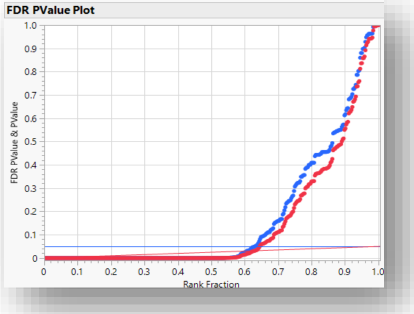
Neural network

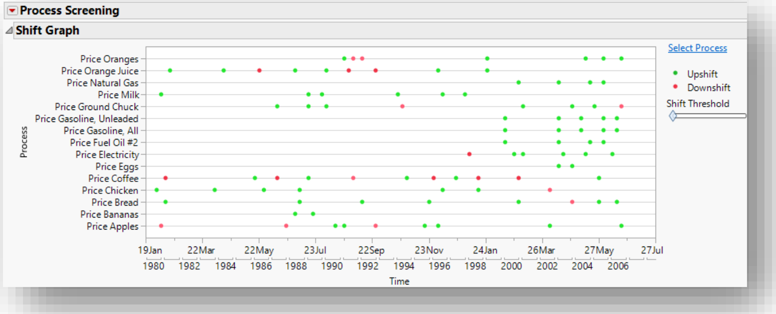
Uses a transfer function to predict response variables

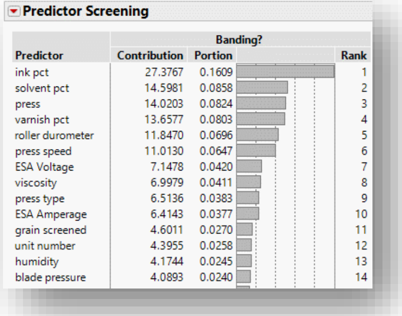
Partition

Creates a decision tree by recursively partitioning data

Improve: Design Optimization







Predictor Screening

Predictor	Banding?			
	Contribution	Portion		Rank
ink pct	27.3767	0.1609		1
solvent pct	14.5981	0.0858		2
press	14.0203	0.0824		3
varnish pct	13.6577	0.0803		4
roller durometer	11.8470	0.0696		5
press speed	11.0130	0.0647		6
ESA Voltage	7.1478	0.0420		7
viscosity	6.9979	0.0411		8
press type	6.5136	0.0383		9
ESA Amperage	6.4143	0.0377		10
grain screened	4.6011	0.0270		11
unit number	4.3955	0.0258		12
humidity	4.1744	0.0245		13
blade pressure	4.0893	0.0240		14

Response Screening

Features that aid
the analysis of
large datasets

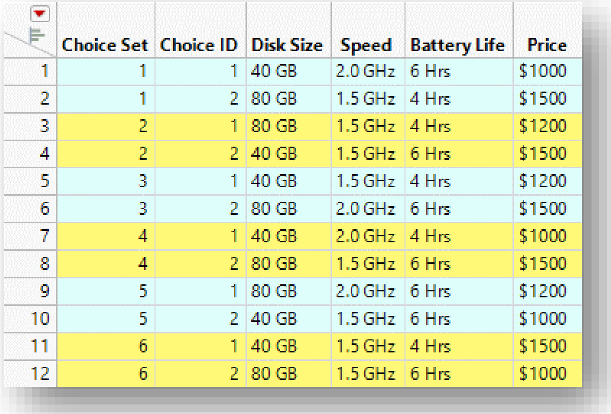
Process Screening

Process capability
and stability for
many responses

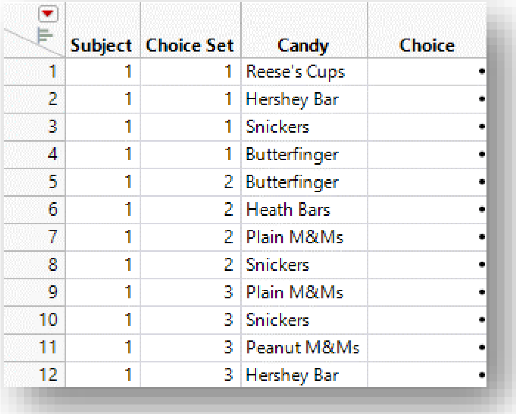
Predictor Screening

Ranks predictors
using bootstrap
forest partitioning

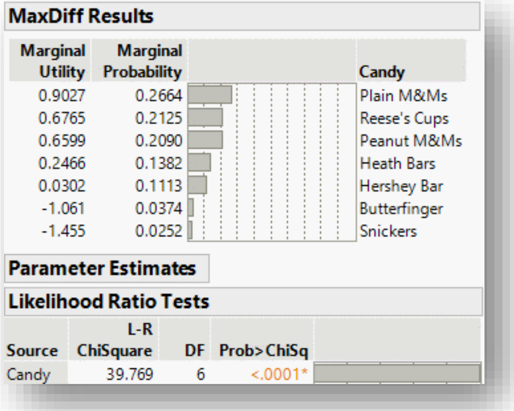
Improve: Consumer Research



	Choice Set	Choice ID	Disk Size	Speed	Battery Life	Price
1	1	1	40 GB	2.0 GHz	6 Hrs	$1000
2	1	2	80 GB	1.5 GHz	4 Hrs	$1500
3	2	1	80 GB	1.5 GHz	4 Hrs	$1200
4	2	2	40 GB	1.5 GHz	6 Hrs	$1500
5	3	1	40 GB	1.5 GHz	4 Hrs	$1200
6	3	2	80 GB	2.0 GHz	6 Hrs	$1500
7	4	1	40 GB	2.0 GHz	4 Hrs	$1000
8	4	2	80 GB	1.5 GHz	6 Hrs	$1500
9	5	1	80 GB	2.0 GHz	6 Hrs	$1200
10	5	2	40 GB	1.5 GHz	6 Hrs	$1000
11	6	1	40 GB	1.5 GHz	4 Hrs	$1500
12	6	2	80 GB	1.5 GHz	6 Hrs	$1000



	Subject	Choice Set	Candy	Choice
1	1	1	Reese's Cups	•
2	1	1	Hershey Bar	•
3	1	1	Snickers	•
4	1	1	Butterfinger	•
5	1	2	Butterfinger	•
6	1	2	Heath Bars	•
7	1	2	Plain M&Ms	•
8	1	2	Snickers	•
9	1	3	Plain M&Ms	•
10	1	3	Snickers	•
11	1	3	Peanut M&Ms	•
12	1	3	Hershey Bar	•



MaxDiff Results

Marginal Utility	Marginal Probability	Candy
0.9027	0.2664	Plain M&Ms
0.6765	0.2125	Reese's Cups
0.6599	0.2090	Peanut M&Ms
0.2466	0.1382	Heath Bars
0.0302	0.1113	Hershey Bar
-1.061	0.0374	Butterfinger
-1.455	0.0252	Snickers

Parameter Estimates

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Candy	39.769	6	<.0001*

Choice Design

Used to find the best combination of features

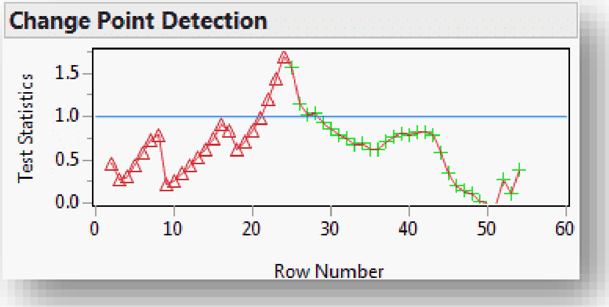
MaxDiff Design

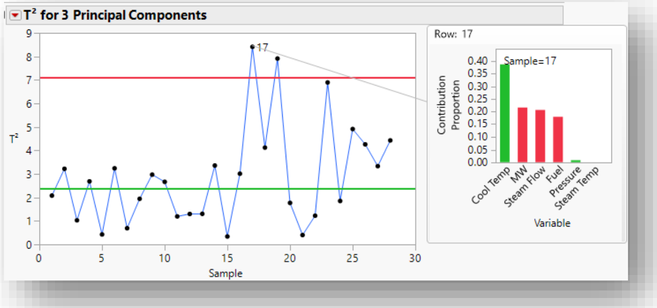
Only considers most and least preferred items

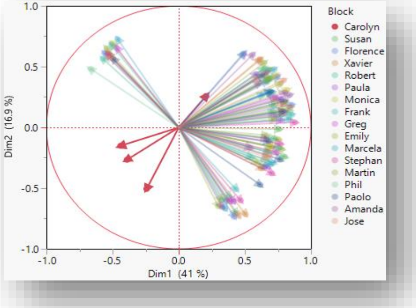
Control Phase

- **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 Ana**lysis: Time Series Decomposition and Smoothing, ARIMA, Forecasting

Control: Multivariate Tools







Change Point Detection plot

Detects a shift in the mean by dividing the data

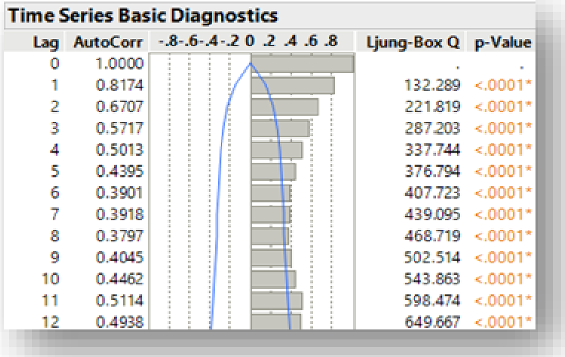
T Square chart

Uses principal components for process stability

Multiple factor analysis

Uses eigenvalue decomposition to compare items

Control: Time Series Techniques

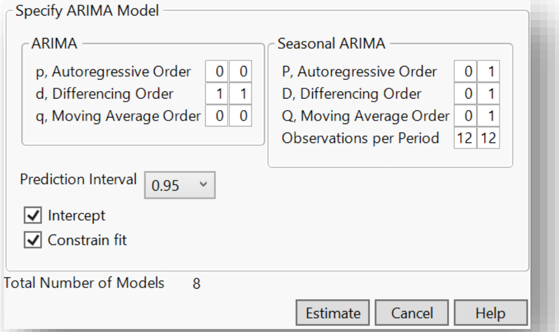


Time Series Basic Diagnostics

Lag	AutoCorr	-.8-.6-.4-.2 0 .2 .4 .6 .8	Ljung-Box Q	p-Value
0	1.0000		.	.
1	0.8174		132.289	<.0001*
2	0.6707		221.819	<.0001*
3	0.5717		287.203	<.0001*
4	0.5013		337.744	<.0001*
5	0.4395		376.794	<.0001*
6	0.3901		407.723	<.0001*
7	0.3918		439.095	<.0001*
8	0.3797		468.719	<.0001*
9	0.4045		502.514	<.0001*
10	0.4462		543.863	<.0001*
11	0.5114		598.474	<.0001*
12	0.4938		649.667	<.0001*

Model diagnostics

Identify trend, seasonal, and cyclic components



Specify ARIMA Model

ARIMA

p, Autoregressive Order	0	0
d, Differencing Order	1	1
q, Moving Average Order	0	0

Seasonal ARIMA

P, Autoregressive Order	0	1
D, Differencing Order	0	1
Q, Moving Average Order	0	1
Observations per Period	12	12

Prediction Interval 0.95

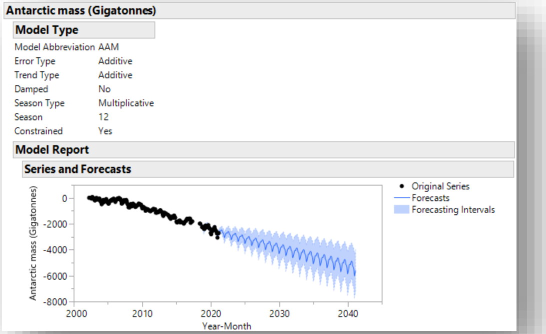
☑ Intercept

☑ Constrain fit

Total Number of Models 8

ARIMA models

Fits data using seasonal or non-seasonal methods



Antarctic mass (Gigatonnes)

Model Type

Model Abbreviation	AAM
Error Type	Additive
Trend Type	Additive
Damped	No
Season Type	Multiplicative
Season	12
Constrained	Yes

Model Report

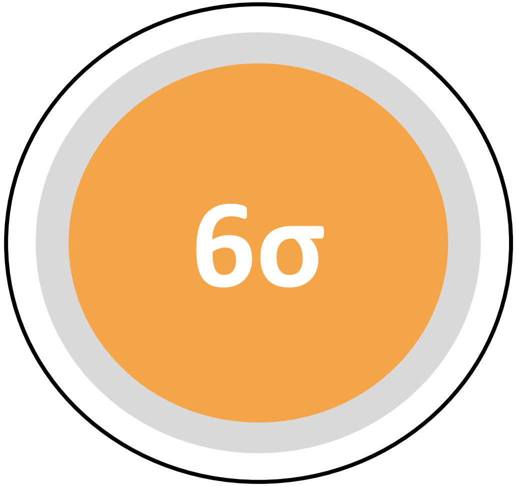
Series and Forecasts

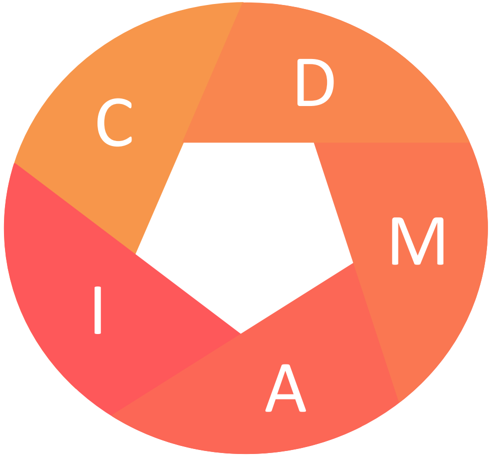
Forecasting

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!