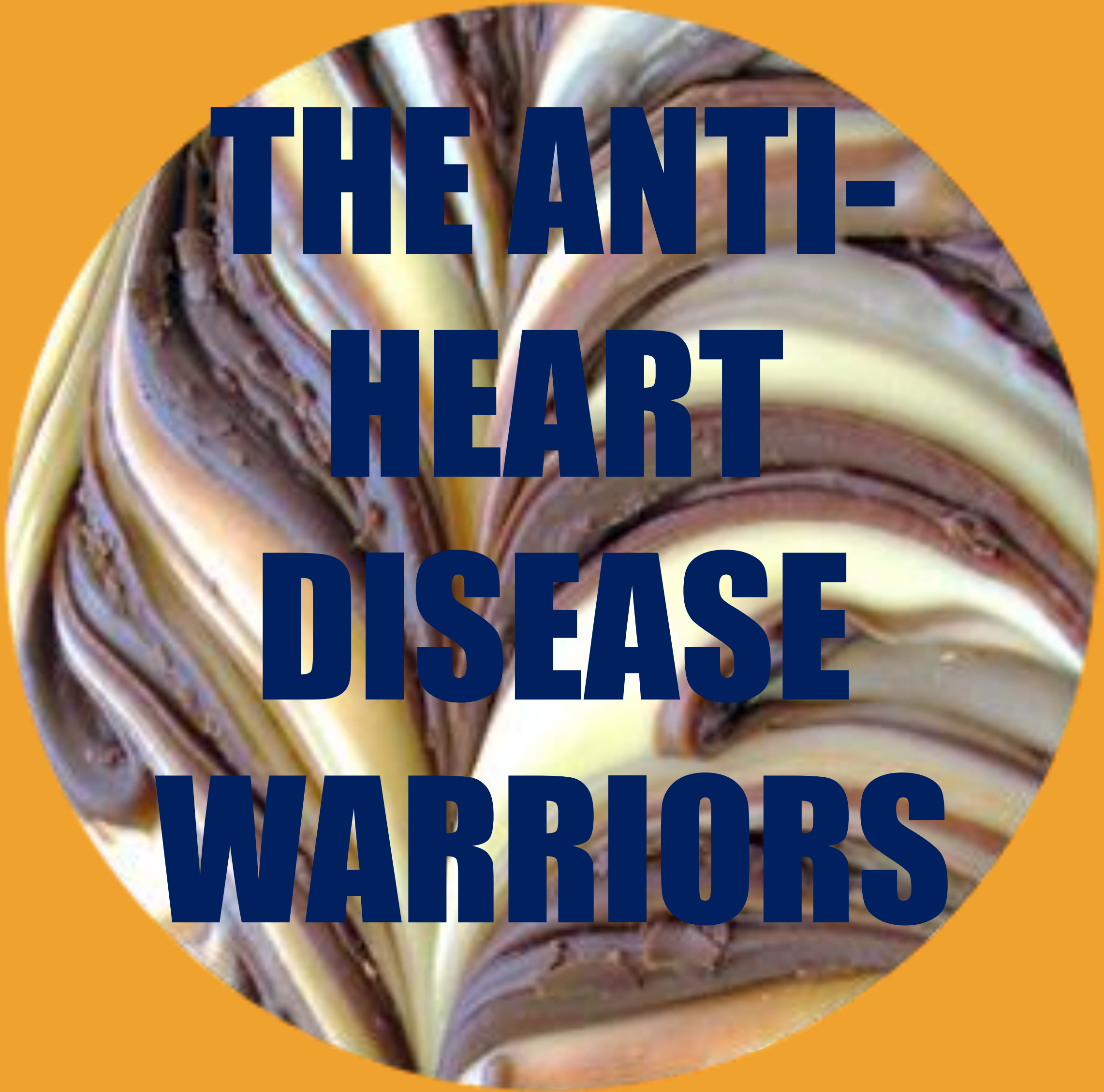
THE ANTI-HEART DISEASE WARRIORS



Mason Chen Black Belt, Stanford OHS

1st Place Best Contributed Paper, 2018 JMP Discovery Summit, Cary NC

Project Scope and Presentation Flow

S T E A M S

- Many people like chocolate, but have some concerns that chocolate is unhealthy.
- Some people who have **heart diseases** might need to eat chocolate, but do not know which one to eat.

1. Anti-Oxidant Science Literature Research

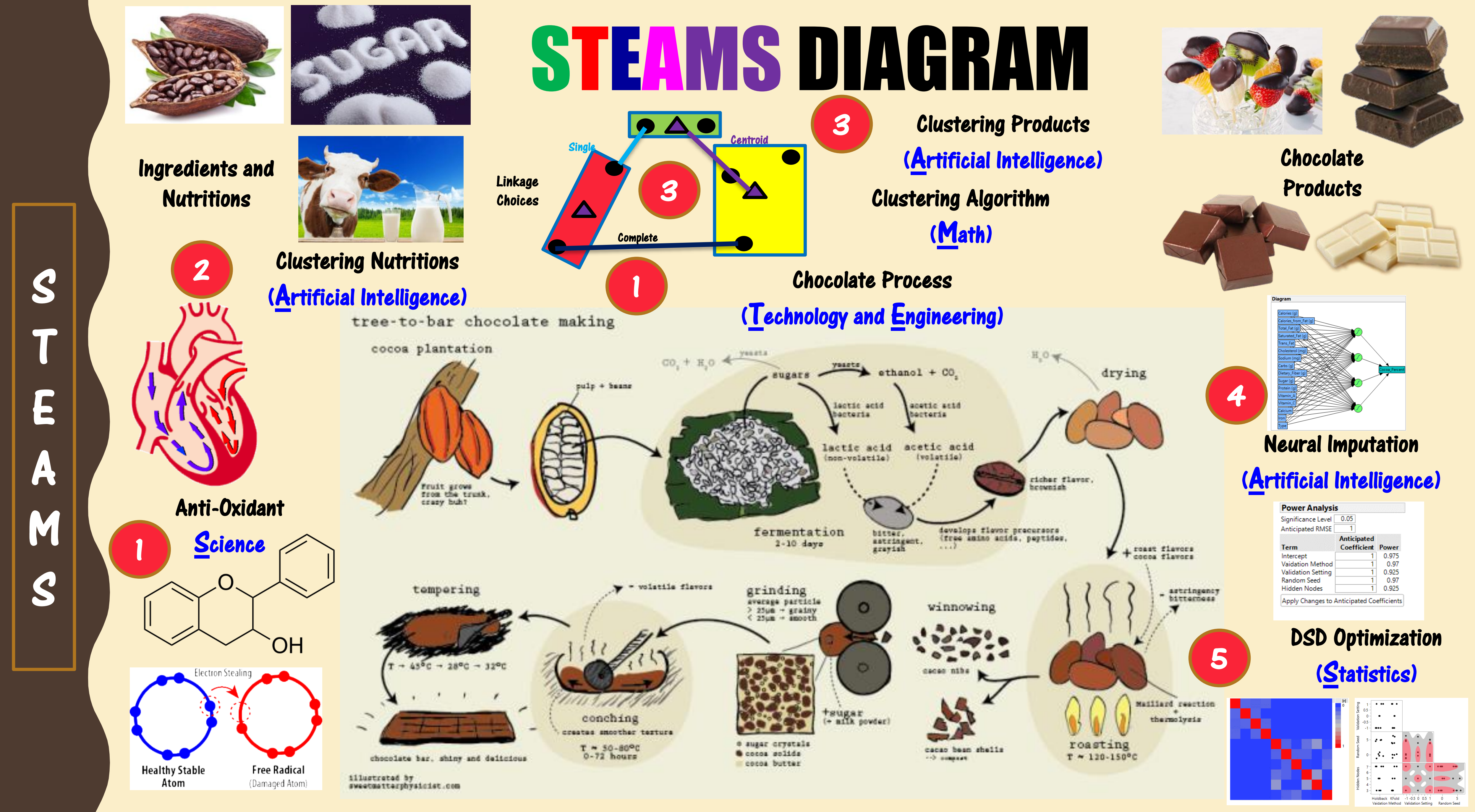
2. Clustering Nutritions & Science

3. Clustering Chocolate Types

4. Missing Value Neural Imputation of Cocoa%

5. DSD Optimization of Neural Setting

STEAMS DIAGRAM



Mason C., "STEAMS" Methodology of Conducting Chocolate Science Research", submitted to NSTA STEM Expo

2015-2018 "STEAMS" Journey

S T E M → T E A M S → S T E A M S

S T E A M S

Develop Math & Science Foundation
(Stanford OHS)
Math, Physics, Biology, Chemistry, JAVA, Statistics Literature Research/Writing

Certify 6 Professional Certificates:
IBM SPSS Statistics
IBM Modeler DA/DM (IBM000129876)
IASSC YB/GB/BB (GR764000541MC)
JMP Statistical Thinking (2018 Goal)
JMP DOE (2019 Goal)
JMP Script Specialist (2019 Goal)

Enhance STEAMS Skills
JMP/Pro, Python
Latex Paper Proceedings
Oral/Poster Presentations
Team Building

Hands-On

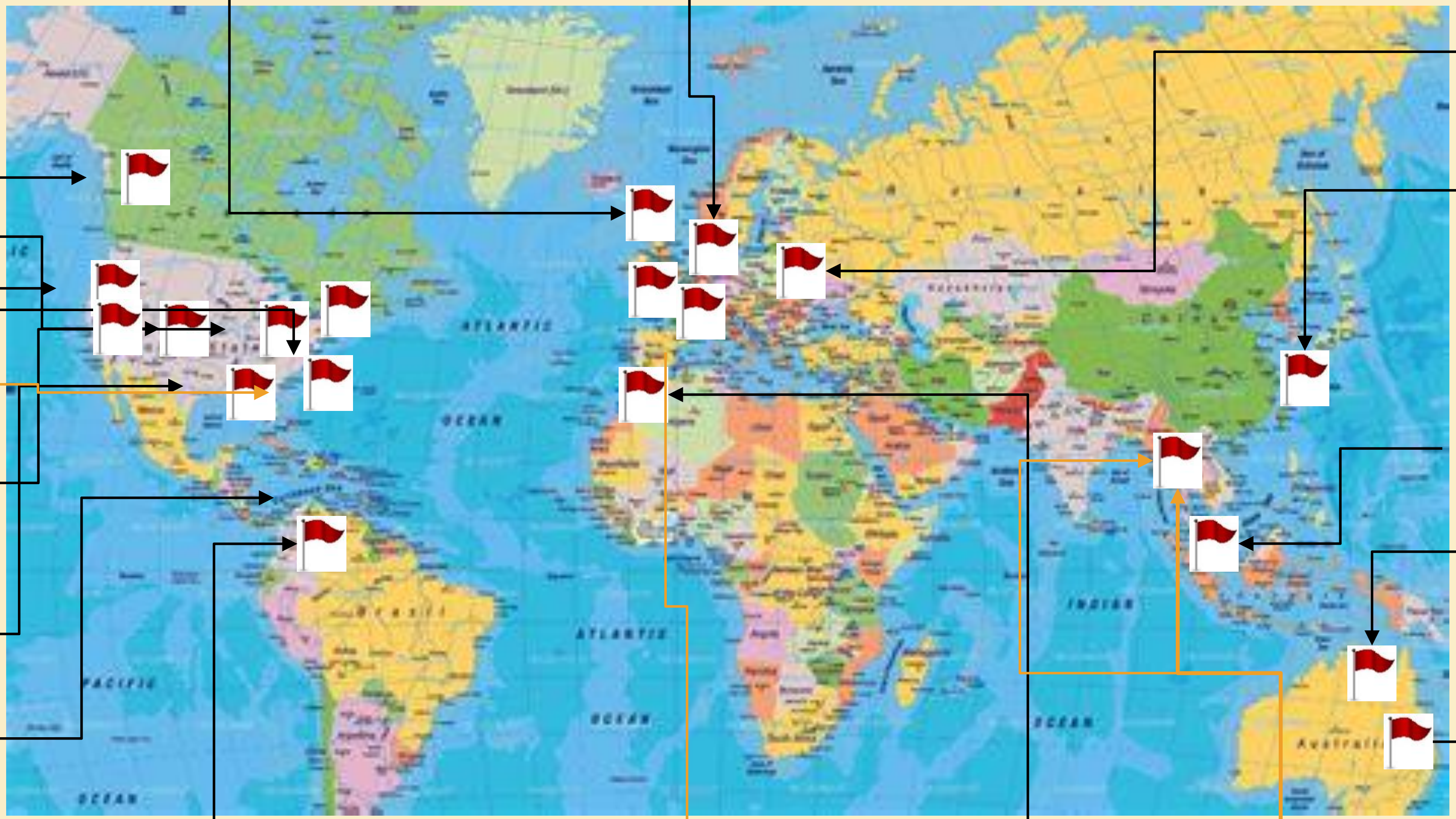
Fun, Real

Learning "STEAMS" techniques help motivate school learning on project-based and practical way

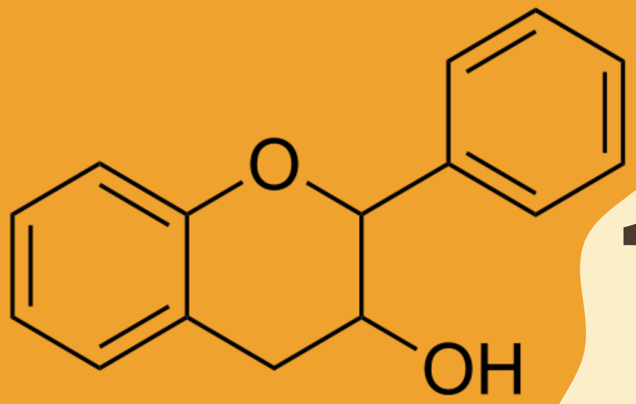
Global Vision Leadership: 2017-2018 Conferences

S T E A M S

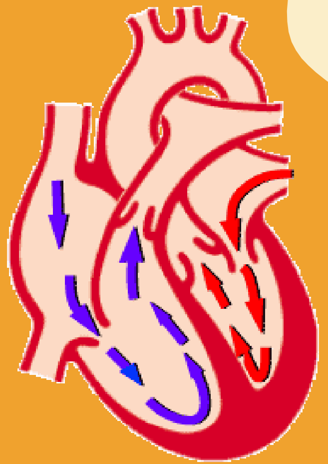
Culture-Shocking

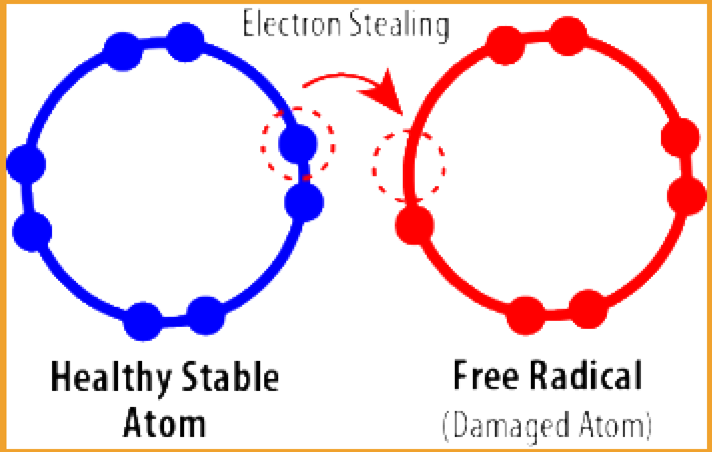


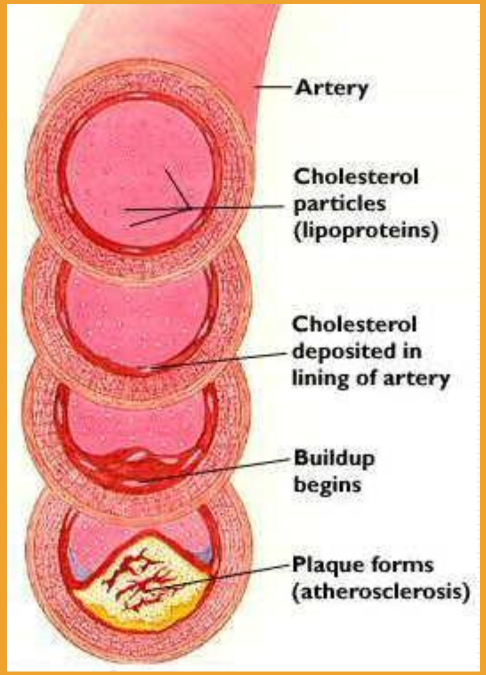
JMP13 >> Analyze >> Fit Y by X >> Nonpar Density



1. Chocolate Anti-Oxidant Science





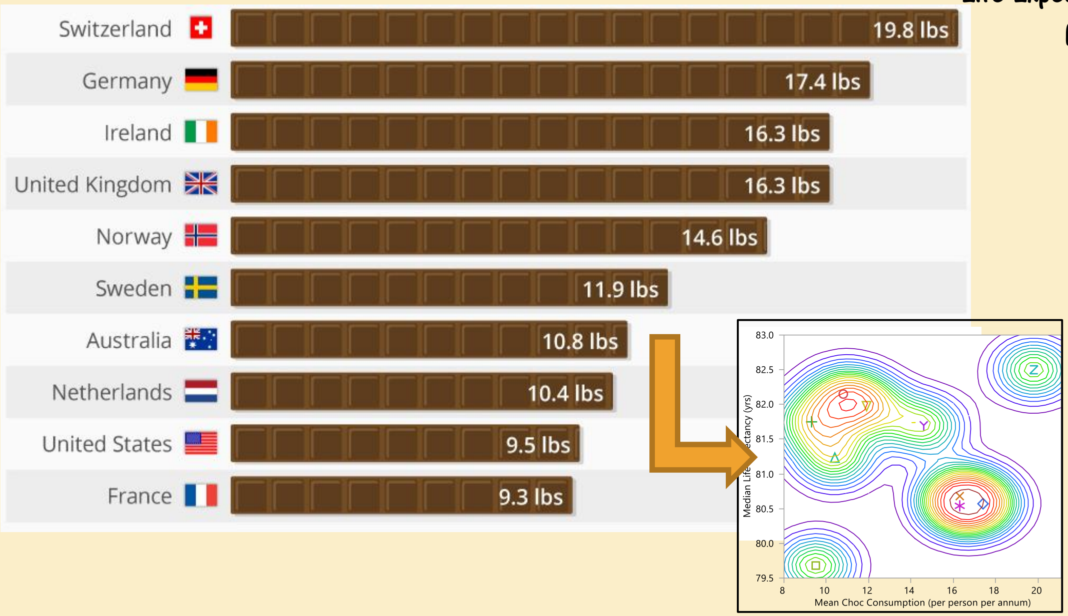


Mason C., (2018 July) "Multivariate Statistics of Antioxidant Chocolate", SMS IWSM Bristol Proceedings, Vol 2 37-40

ENGINEERING

IS EATING CHOCOLATE UNHEALTHY?

Chocolate has not been proven harmful.



Life Expectancy[8] (2015 Estimate)

Median: 74.75

#9: 82.50

#32: 80.57

#31: 80.68

#33: 80.54

#20: 81.70

#15: 81.98

#13: 82.15

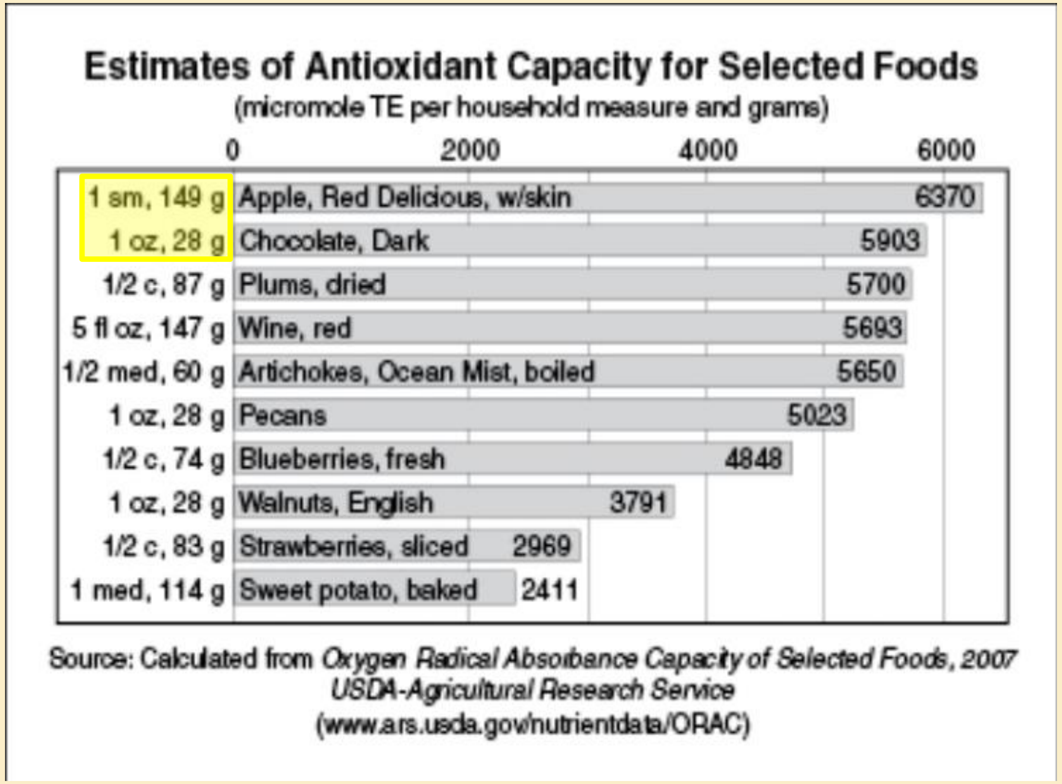
#24: 81.23

#43: 79.68

#19: 81.75

JMP13 >> Analyze >> Fit Y by X >> Nonpar Density

Anti-Oxidant Capacity/gram



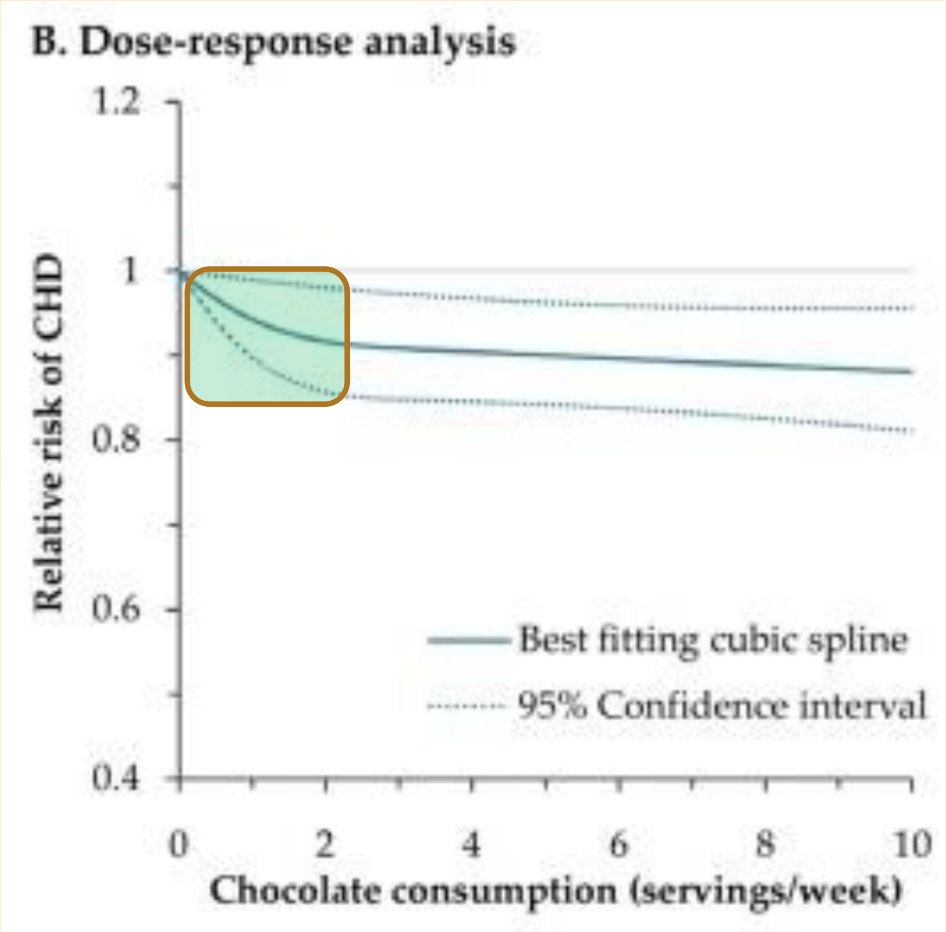
Dark chocolate is a powerful source of antioxidant. If chocolate's serving size is equal to that of an apple, it has the highest amount of antioxidant.

CHOCOLATE & ATRIAL FIBRILLATION (AF)

TECHNOLOGY

Lower Cardiovascular Heart Disease (CHD) risk if taking 2 Chocolate servings per week (1 serving = 30 g)

- Chocolate may be inversely associated with AF
- Dark chocolate may be a healthy snacking option
- AF = Atrial Fibrillation (a cardiovascular disease)
- Next, how Chocolate can reduce CHD risk and AF associated cardiovascular disease



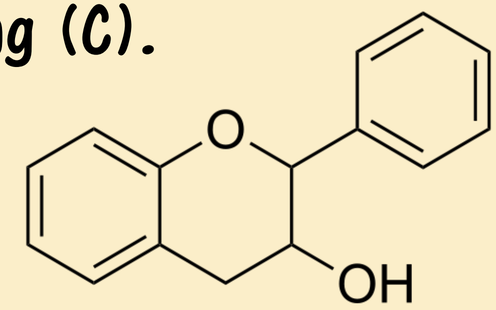
https://heart.bmj.com/content/103/15/1163

https://www.bmj.com/content/343/bmj.d4488

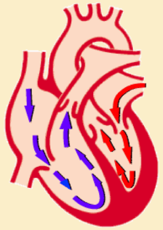
FLAVONOIDS SCIENCE & STRUCTURE

SCIENCE

- Flavonoids are the most abundant polyphenols in human diet that have antioxidant properties.
- Flavonoids have the general structure of a 15- carbon skeleton C6-C3-C6.
  - Consists of two phenyl rings (A and B) and a heterocyclic ring (C).



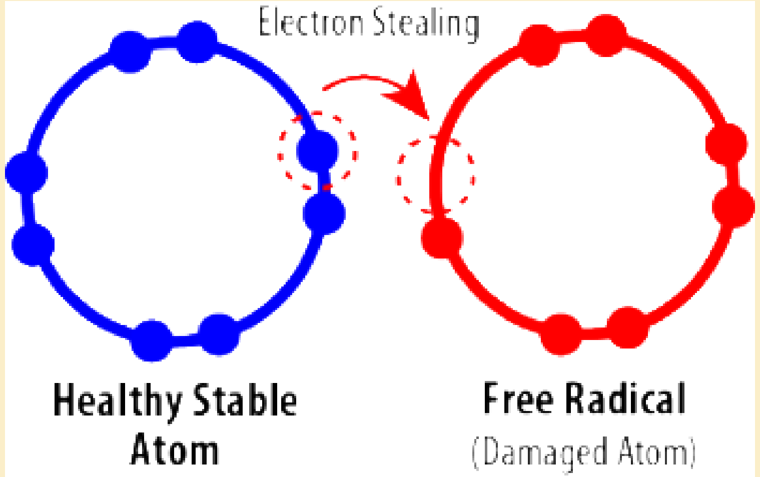
- There are seven different types of flavonoids based on its chemical structure:
  - Flavones, flavanol, flavanones, isoflavones, anthocyanidins, chalcones, catechins
- Chocolate flavonoids are flavanols which can promote healthy blood flow from head to toe.

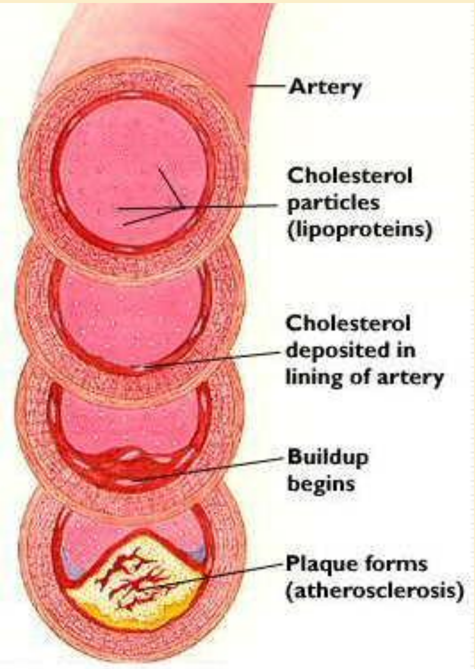


FREE RADICALS AND ANTIOXIDANTS

SCIENCE

- Free radicals are atoms with odd number of electrons
  - Antioxidants reduce free radical formation
  - Reactive free radicals causes cells mal-function
  - Excess free radicals damages blood vessel
- After the oxidation of free radicals, LDL (Low-density Lipoprotein) can cause CVD (Cardiovascular Disease)
  - The oxidized components attract macrophages which absorb & deposit Cholesterol





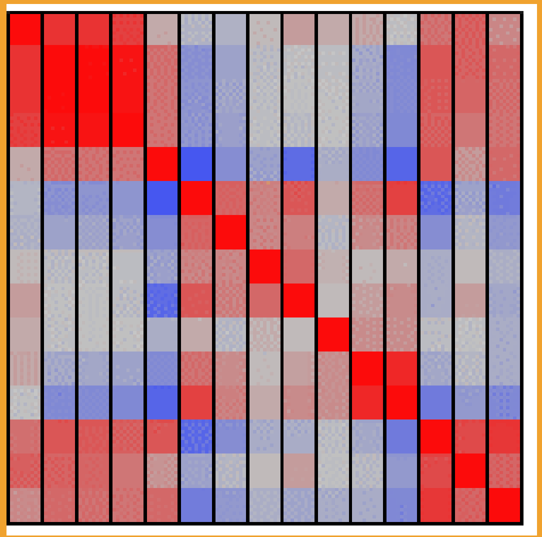
T
E
C
H
N
O
L
O
G
Y

DARK CHOCOLATE LITERATURE RESEARCH

- Benefits:
  - A lot of soluble fiber
  - A lot of minerals: iron, magnesium, copper, manganese, potassium, phosphorus, zinc, selenium
  - Powerful source of antioxidant
  - Improve blood flow and lower blood pressure
  - Increases HDL (good cholesterol) and decreases LDL (bad cholesterol)
  - Lower risk of cardiovascular disease (CVD)
  - Improve brain function[1]
- Concerns:
  - Causes migraines
  - Increases chance of kidney stones
  - Side effects from caffeine such as irregular heartbeat

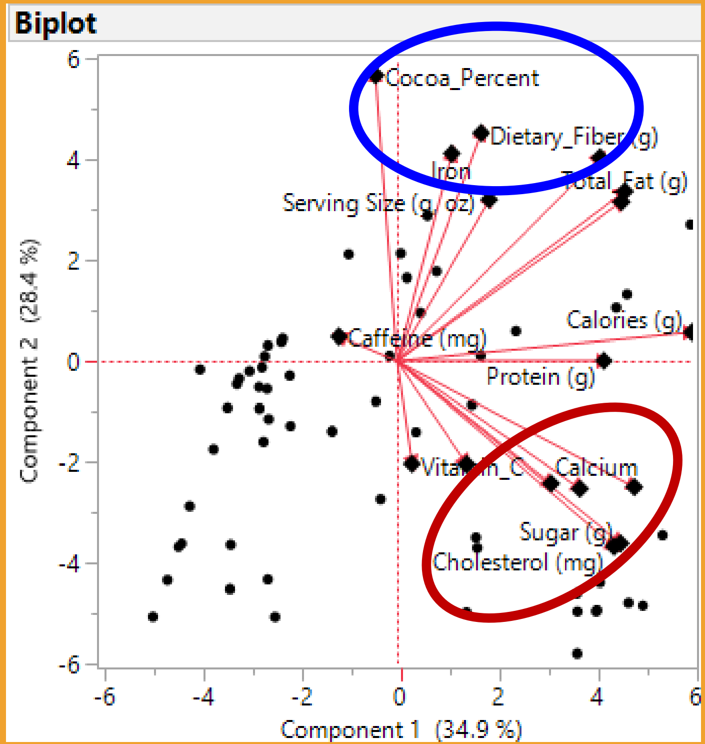
2. Clustering Nutritions & Science

(1) JMP 13 >> Analyze >> Distribution



(2) JMP 13 >> Analyze >> Multivariate Methods >> Multivariate

(3) JMP 13 >> Analyze >> Clustering >> Cluster Variables



(4) JMP 13 >> Analyze >> Multivariate Methods >> Principle Components >> Bi-Plot

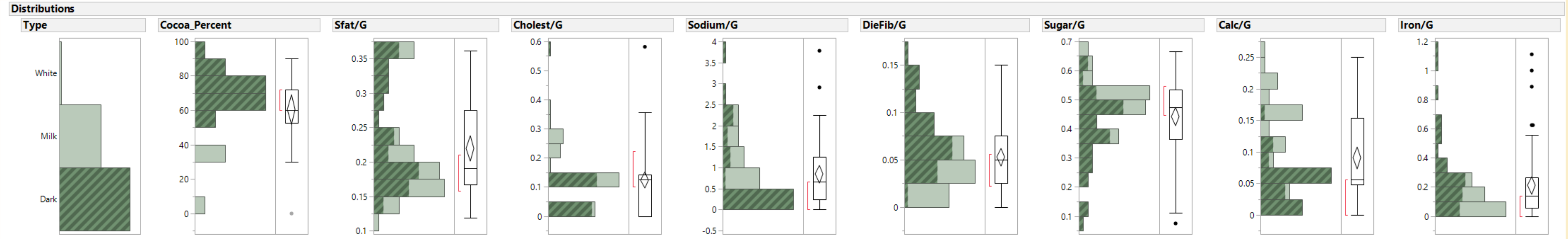
- Mason C., (2018 July), "Choose Healthy Chocolate", IEOM Europe Paris Proceedings, 434-441

STATISTICS

(1) CHOCOLATE NUTRITION DISTRIBUTION

JMP 13 >> Analyze >> Distribution

- 60+ Chocolate nutrition data collected from "Target" store.
- The quantities of the eight most critical ingredients were analyzed.



Most Dark Chocolate (**Qualitative** Clustering Criteria) has:

- 1st Cluster: higher Cocoa percent, Dietary Fiber and Iron
- 2nd Cluster: lower Cholesterol, Calcium, and Sugar

Chocolate Product Nutrition data has indicated that Dark Chocolate is healthier than the Milk and White Chocolate

(2) DARK CHOCOLATE CORRELATION

SC
I
EN
CE

- 1st Cluster: Sugar and Cocoa_Percent have a negative correlation of -0.9162.
- 2nd Cluster: Dietary Fiber and Iron have a positive correlation of 0.7722.

Any other better way to cluster nutritions?

A 100 gram bar of dark chocolate with 70-85% cocoa contains (1):

- 11 grams of fiber.
- 67% of the RDA for Iron.
- 58% of the RDA for Magnesium.
- 89% of the RDA for Copper.
- 98% of the RDA for Manganese.
- It also has plenty of potassium, phosphorus, zinc and selenium.

Correlations Pair-Wise Pearson Correlation

	Cocoa_Percent	Sfat/G_1	Cholest/G_1	Sodium/G_1	DieFib/G_1	Sugar/G_1	Calc/G_1	Iron/G_1
Cocoa_Percent	1.0000	0.5291	-0.3114	-0.0583	0.5482	-0.9162	0.2625	0.4597
Sfat/G_1	0.5291	1.0000	-0.1980	0.0184	0.0341	-0.7068	0.4161	0.0687
Cholest/G_1	-0.3114	-0.1980	1.0000	0.0302	-0.3666	0.3333	0.1732	-0.3304
Sodium/G_1	-0.0583	0.0184	0.0302	1.0000	-0.1344	0.0462	0.1667	-0.1862
DieFib/G_1	0.5482	0.0341	-0.3666	-0.1344	1.0000	-0.5804	-0.0207	0.7722
Sugar/G_1	-0.9162	-0.7068	0.3333	0.0462	-0.5804	1.0000	-0.3696	-0.4669
Calc/G_1	0.2625	0.4161	0.1732	0.1667	-0.0207	-0.3696	1.0000	-0.1037
Iron/G_1	0.4597	0.0687	-0.3304	-0.1862	0.7722	-0.4669	-0.1037	1.0000

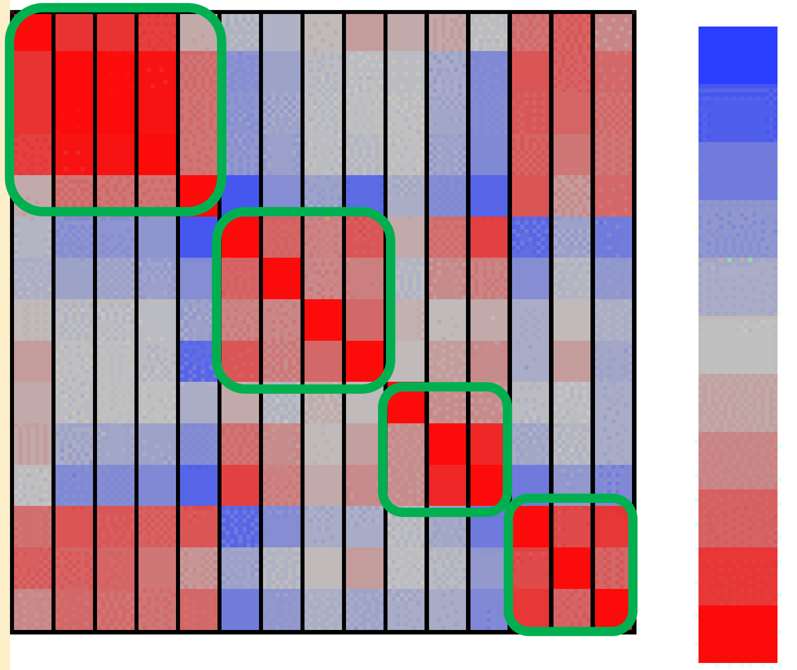
JMP 13 >> Analyze >> Multivariate Methods >> Multivariate

SCIENCE & AI

(3) VARIABLE CLUSTERING

JMP 13 >> Analyze >> Clustering >> Cluster Variables

Color Map on Correlations



Cluster Members

		Signal	Noise	S-N Ratio
Cluster	Members	RSquare with Own Cluster	RSquare with Next Closest	1-RSquare Ratio
1	Calories (g)	0.789	0.314	0.308
1	Calories_from_Fat (g)	0.976	0.456	0.044
1	Total_Fat (g)	0.977	0.426	0.04
1	Saturated_Fat (g)	0.935	0.361	0.101
2	Cocoa_Percent	0.742	0.366	0.406
2	Cholesterol (mg)	0.811	0.387	0.309
2	Vitamin_A	0.505	0.126	0.566
2	Vitamin_C	0.412	0.016	0.598
2	Calcium	0.726	0.079	0.297
3	Sodium (mg)	0.345	0.013	0.664
3	Carbs (g)	0.876	0.185	0.152
3	Sugar (g)	0.874	0.416	0.216
4	Dietary_Fiber (g)	0.888	0.403	0.187
4	Protein (g)	0.73	0.358	0.421
4	Iron	0.803	0.269	0.269

Clustering Nutritions can interpret the relevant Chocolate Science insight well:

Cluster 1: the higher the saturated fat, the higher the total fat, and the higher the calories.

Cluster 2: Calcium/Cholesterol, and Cocoa percent have a negative correlation.

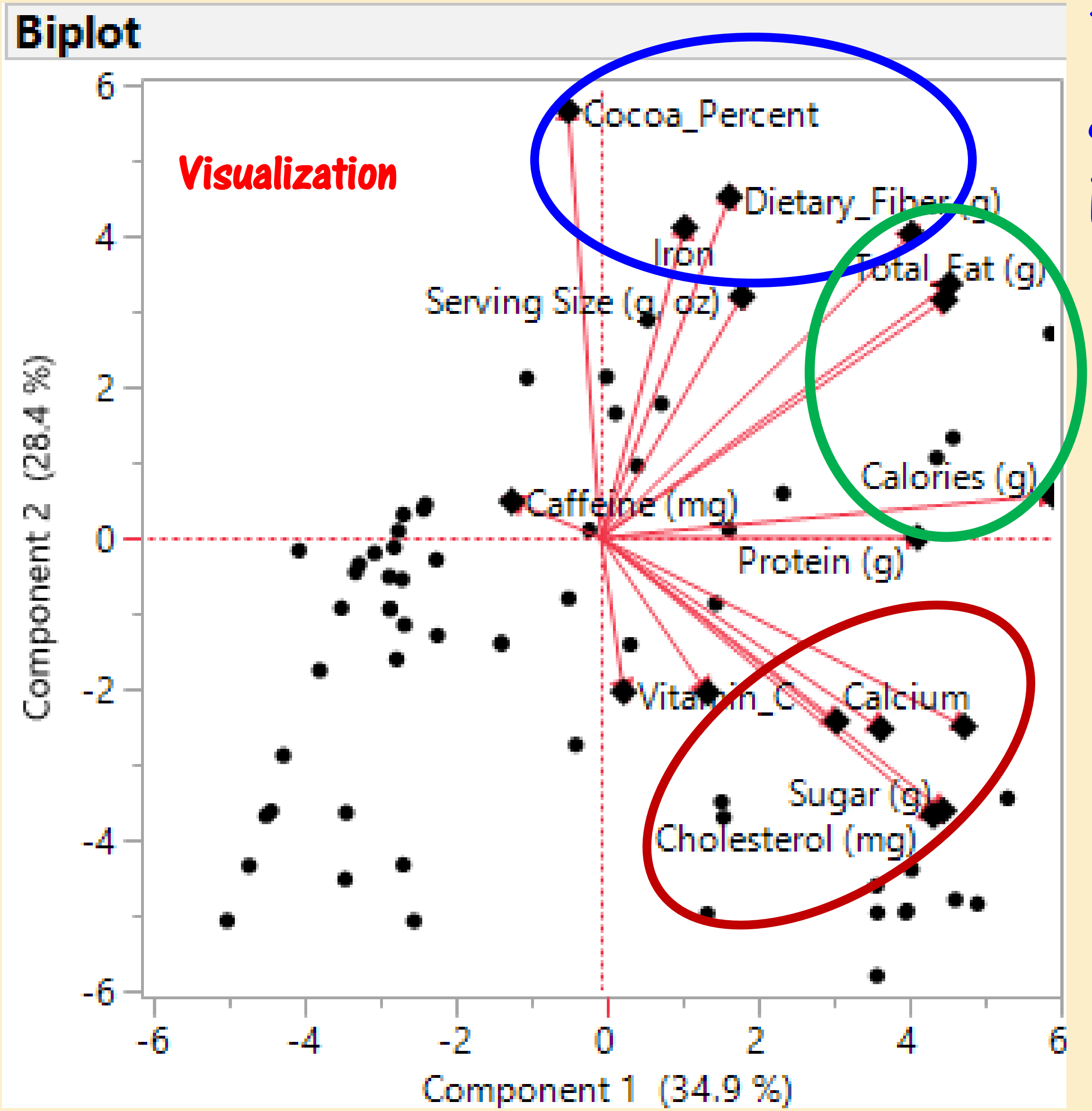
Cluster 3: the higher the sugar, the higher the carbohydrates.

Cluster 4: Iron and dietary fiber are positively correlated.

Common Sense

MATH & AI

(4) Principle Component Bi-Plot



1st Cluster: Cocoa Percent, Dietary Fiber, and Iron are near each other (Higher for Dark Chocolate)

2nd Cluster: Total Fat, Saturated Fat, and Calories

3rd Cluster: Calcium, Sugar, and Cholesterol are near each other (Higher for Milk/White Chocolate)

JMP 13 >> Analyze >> Multivariate Methods >> Principle Components >> Bi-Plot

Comparing Four Clustering Methods

E
N
GI
N
E
E
RI
N
G

Platform	Criteria	1st Cluster	2nd Cluster	3rd Cluster	4th Cluster
Interactive Distribution	**Qualitative**	Cocoa%, Dietary Fiber, Iron	Cholestrol, Sugar		
Multivariate Correlation	**Quantitative**	Dietary Fiber and Iron	Cocoa% and Sugar		
Clustering Variables	**Quantitative**	Saturated Fat, Total Fat, Calories	Cholesterol, Calcium, Cocoa%	Sugar, Carbohydrates	Iron, Dietary Fiber
Principle Component Bi-Plot	**Quantitative**	Cocoa %, Dietary Fiber, Iron	Saturated Fat, Total Fat, Calories	Calcium, Sugar, Cholesterol	

- *Four different clustering methods show similar clustering patterns*
- *Clustering "Statistics and Engineering" results match Chocolate "Science and Technology" Literature Research well (STEAMS).*

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster
JMP 13 >> Analyze >> Distribution

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster >> Clustering Distance Method

3. Clustering Chocolate Types

JMP 13 >> Analyze >> Distribution

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster >> Constellation Plot

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster >> Column Summary

JMP 13 >> Analyze >> Multivariate Methods >> Principle Components >> Eigenvalues

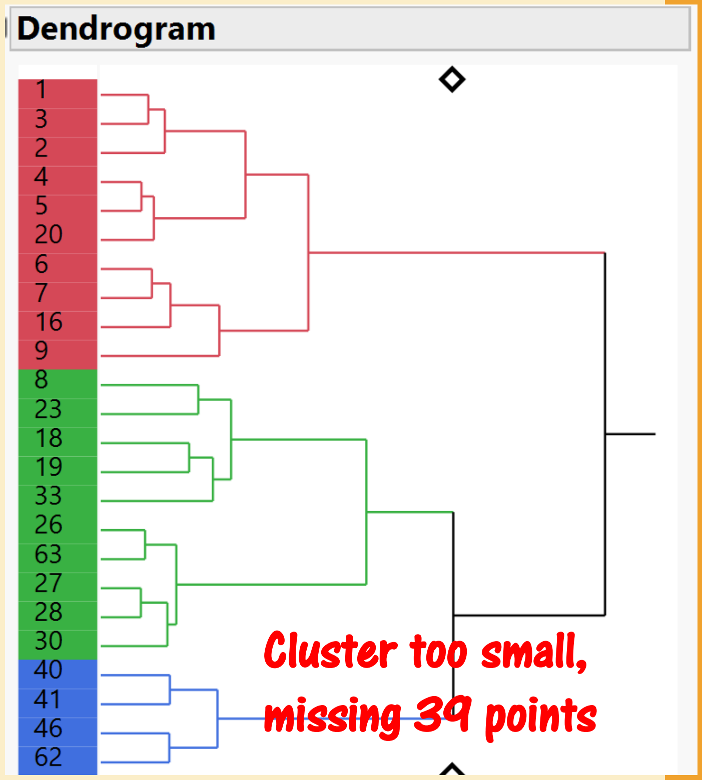
- Mason C., (2018 Dec.), "Statistics Application on the Study of Chocolate Science with Heart Disease", ASA SDSS Proceeding

S T A TI S TI C S AI

CLUSTERING PRODUCTS

Objective: find a way to identify healthy chocolate products for Heart Disease patients.

- Use hierarchical clustering to cluster chocolate products
- All Milk and white chocolate form the third cluster while dark chocolate split between the first and second cluster.
- Why are there two clusters for dark chocolate (why not one cluster for each Chocolate Type)?



Cluster too small, missing 39 points

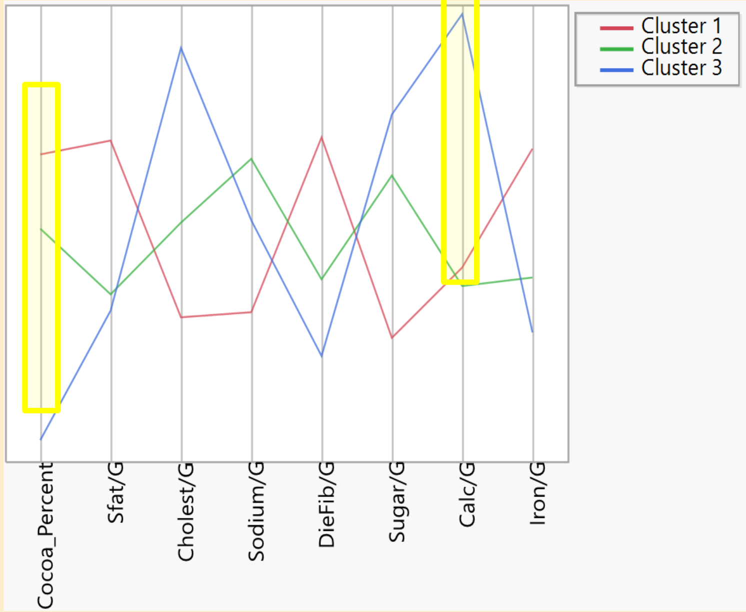
JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster

JMP 13 >> Analyze >> Distribution

AI & TECHNOLOGY

PRINCIPLE CLUSTERING DECIDING FACTORS

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster >> Column Summary



Column Summary

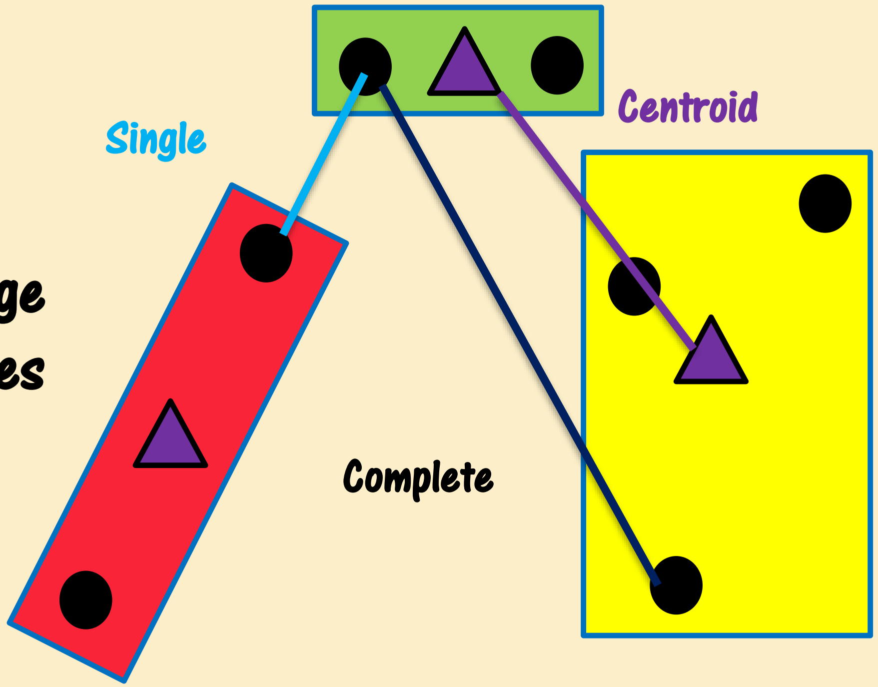
Column	RSquare	.2 .4 .6 .8
Cocoa_Percent	0.7788	
Sfat/G	0.4908	
Cholest/G	0.6978	
Sodium/G	0.3949	
DieFib/G	0.5788	
Sugar/G	0.6642	
Calc/G	0.7727	
Iron/G	0.4351	

- 1st Cluster: Dark Chocolate, High Cocoa%, and Low Calcium, Most Healthy?
- 2nd Cluster: Dark Chocolate, Medium Cocoa%, and Low Calcium.
- 3rd Cluster: Milk/White Chocolate, Low Cocoa%, and High Calcium

CLUSTERING DISTANCE METHODS

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster

M A T H



Average Linkage Distance for the average linkage cluster method is:

$$D_{KL} = \sum_{i \in C_K} \sum_{j \in C_L} \frac{d(x_i, x_j)}{N_K N_L}$$

Average

Centroid Method Distance for the centroid method of clustering is:

$$D_{KL} = \left\| \bar{x}_K - \bar{x}_L \right\|^2$$

Ward's Distance for Ward's method is:

Center-Center

$$D_{KL} = \frac{\left\| \bar{x}_K - \bar{x}_L \right\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

ANOVA (MS)

Single Linkage Distance for the single linkage cluster method is:

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i, x_j)$$

Min

Complete Linkage Distance for the Complete linkage cluster method is:

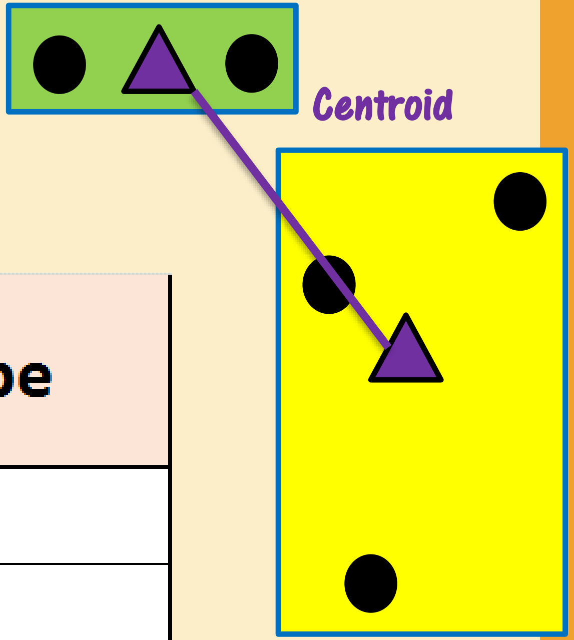
$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} d(x_i, x_j)$$

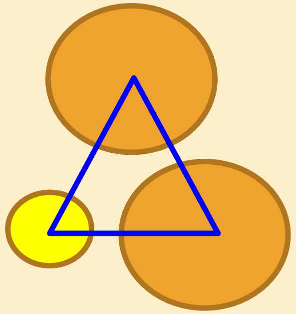
Maximum

Clustering patterns dependent on the cluster number observations, cluster variance, and outlier

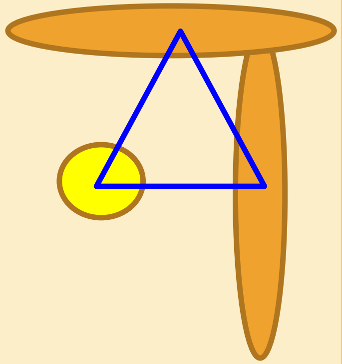
M
A
T
H

Selecting DISTANCE METHODS





	Size/ Variance	Outliers	Shape
Average	Smaller		
Centroid		Robust	
Ward	Smaller	Sensitive	
Single	Larger		Irregular/ Elongated
Complete	Smaller	Moderate	

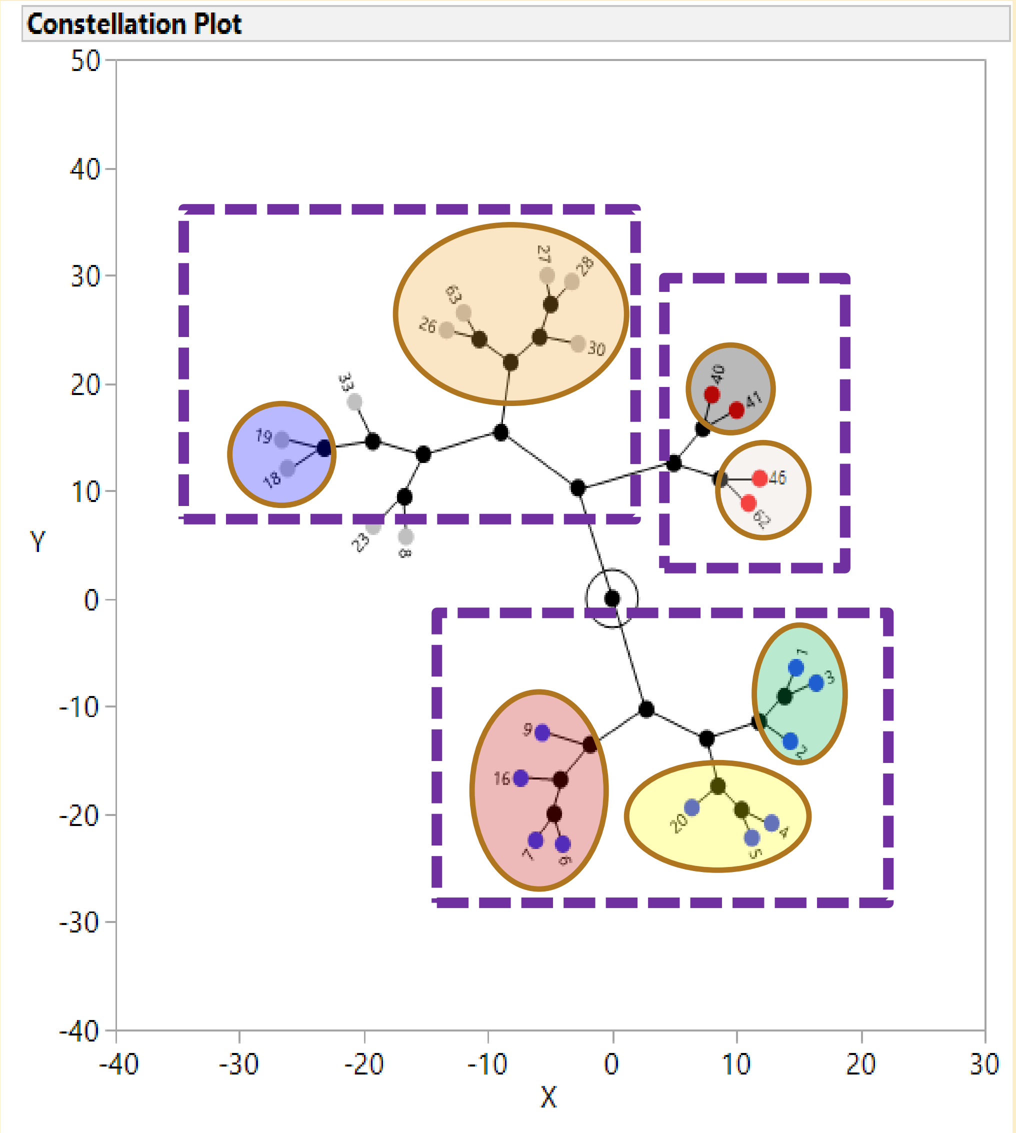


Depending on the data distribution, selecting an appropriate Clustering Distance algorithm is critical to Clustering Pattern Analysis

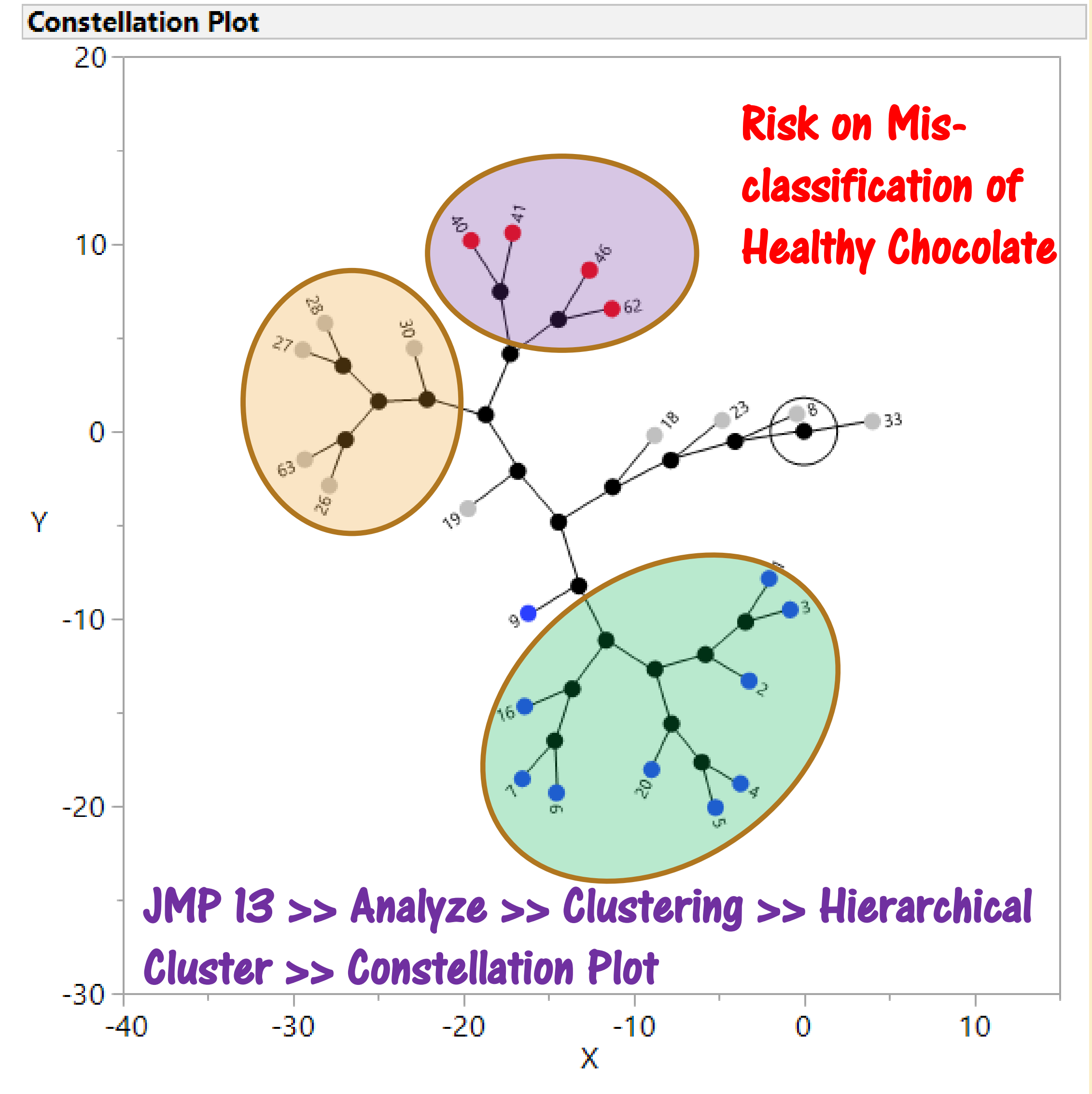
WARD VS SINGLE METHOD (10 Clusters)

AI & MATH

Ward (Join Smaller Observations)



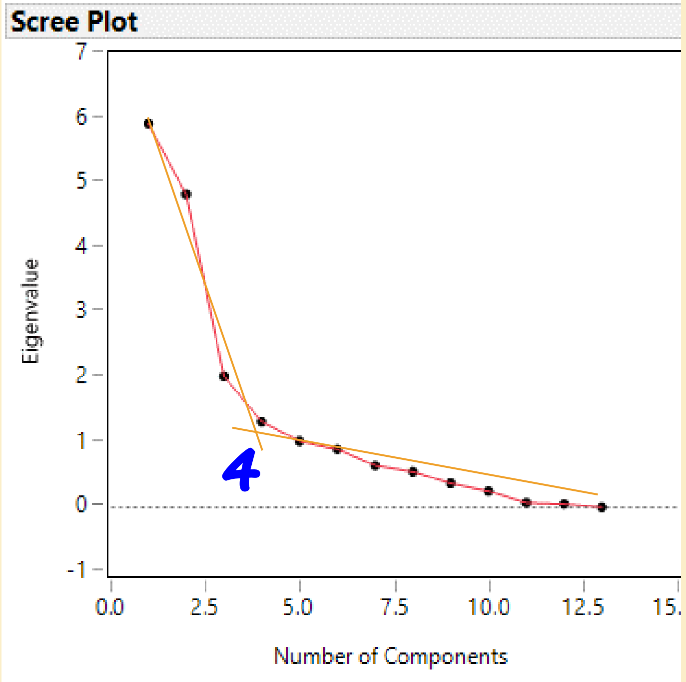
Single (Join Larger Variances)



M
A
T
H

DETERMINE NUMBER OF CLUSTERS

Clustering pattern result is highly dependent on the number of clusters



Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	5.9254	34.855		34.855
2	4.8336	28.433		63.288
3	2.0252	11.913		75.201
4	1.3256	7.797		82.998
5	1.0255	6.032		89.030
6	0.9010	5.300		94.330
7	0.6484	3.814		98.144

From both the scree plot and PCA eigenvalues (80% Pareto), we can pick **4** clusters

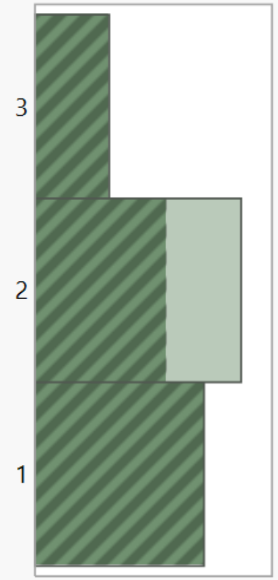
JMP 13 >> Analyze >> Multivariate Methods >> Principle Components >> Eigenvalues

S T A TI S TI C S

WARD VS SINGLE (3-5 CLUSTERS)

JMP 13 >> Analyze >> Distribution

Ward (3)



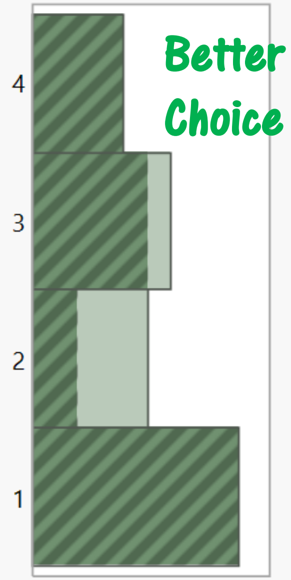
Frequencies

Level	Count	Prob
1	9	0.37500
2	11	0.45833
3	4	0.16667
Total	24	1.00000

N Missing 102

3 Levels

Ward (4)



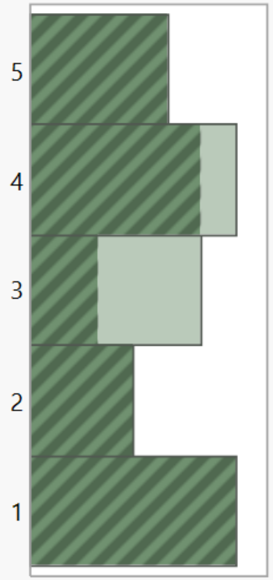
Frequencies

Level	Count	Prob
1	9	0.37500
2	5	0.20833
3	6	0.25000
4	4	0.16667
Total	24	1.00000

N Missing 102

4 Levels

Ward (5)



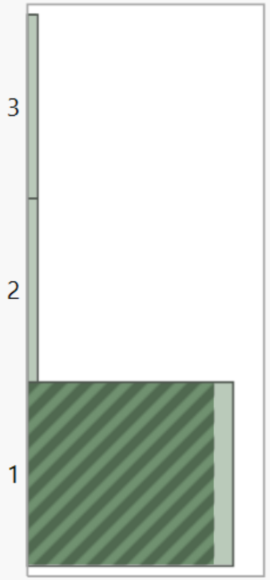
Frequencies

Level	Count	Prob
1	6	0.25000
2	3	0.12500
3	5	0.20833
4	6	0.25000
5	4	0.16667
Total	24	1.00000

N Missing 102

5 Levels

Single (3)



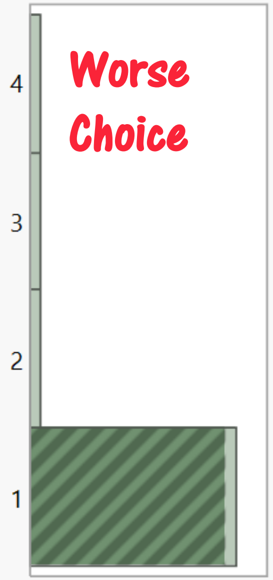
Frequencies

Level	Count	Prob
1	22	0.91667
2	1	0.04167
3	1	0.04167
Total	24	1.00000

N Missing 102

3 Levels

Single (4)



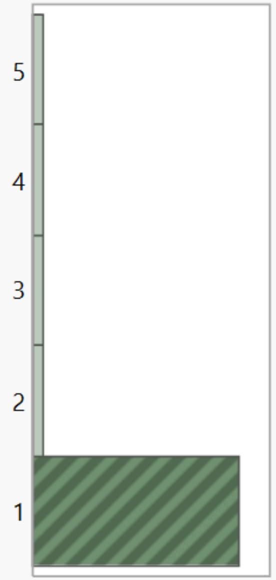
Frequencies

Level	Count	Prob
1	21	0.87500
2	1	0.04167
3	1	0.04167
4	1	0.04167
Total	24	1.00000

N Missing 102

4 Levels

Single (5)



Frequencies

Level	Count	Prob
1	20	0.83333
2	1	0.04167
3	1	0.04167
4	1	0.04167
5	1	0.04167
Total	24	1.00000

N Missing 102

5 Levels

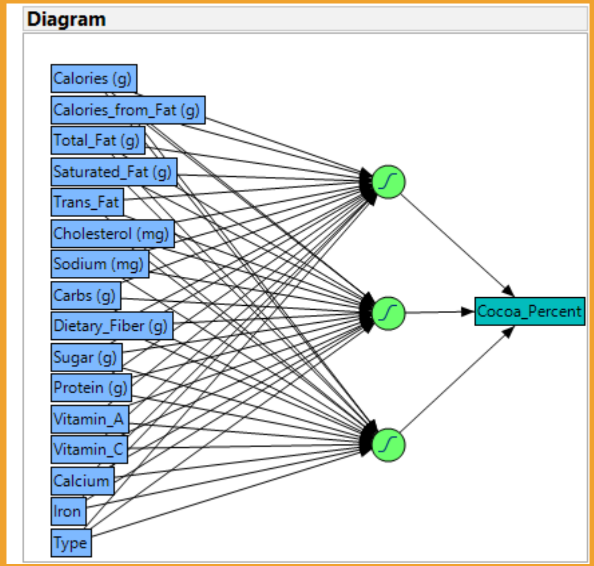
- Single does not show any significant difference between 3, 4, or 5 clusters
- Ward clusters become more similar in size with the higher the number of clusters

JMP 13 >> Analyze >> Clustering >> Hierarchical Cluster >> Missing Value Imputation

JMP 13 >> Analyze >> Predictive Modeling >> Neural

JMP 13 >> Analyze >> Screening >> Explore Missing Values

4. Missing Value Neural Imputation



JMP 13 >> Analyze >> Predictive Modeling >> Partition

JMP 13 >> Analyze >> Predictive Modeling >> Neural >> Diagram

- Mason C., "Neural Network Algorithm of Missing Value Imputation for Chocolate Science Research" submitted to SIAM SDM19

AI & MATH

Explore MISSING VALUES

JMP 13 >> Analyze >> Screening >> Explore Missing Values

- Objective: among 63 commercial chocolate products, 39 have missed the Cocoa % information **(most are Milk Chocolates)**

Missing Columns

☑ Show only columns with missing

Close

Column	Number Missing
Cocoa_Percent	39

Commands

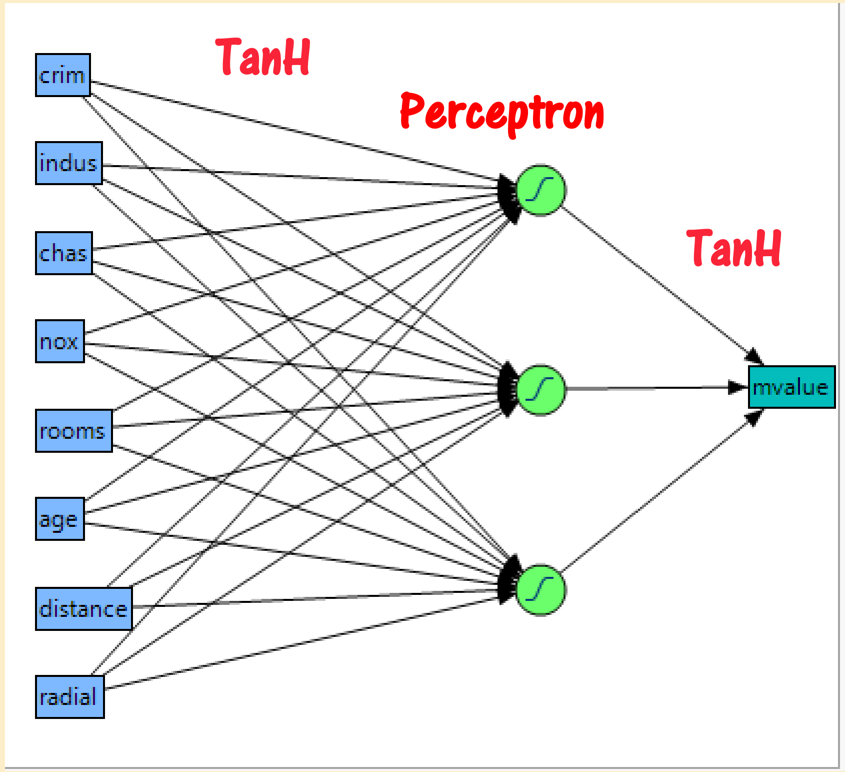
Missing Value Report	Number of missing values for each column
Missing Value Clustering	Hierarchical clustering of rows and columns missingness
Missing Value Snapshot	Patterns of missing values with graphical map
Multivariate Normal Imputation	Least squares prediction from the nonmissing variables in each row
Multivariate SVD Imputation	Imputation for wide problems using a singular value decomposition with the power-method adapted for missing values

- **Any other better imputation method?**

JMP Neural Network Platform

- Implements a fully connected **Perceptron (hidden nodes)** with one layer.

- **Main advantage:** can efficiently model different response surfaces given enough hidden nodes and layers.
- **Main disadvantage:** results are not easily interpretable, since there are intermediate layers (**Black Box**)



AI

Standard JMP Edition:
- **Only TanH activation function**
- **Can fit with one hidden layer.**

TanH The hyperbolic tangent function is a sigmoid function. TanH transforms values to be between -1 and 1, and is the centered and scaled version of the logistic function. The hyperbolic tangent function is:

$$\frac{e^{2x}-1}{e^{2x}+1}$$

More Powerful Exponential Transformation than PCA Linear Transformation

where x is a linear combination of the X variables.

AI & STATISTICS

MISSING VALUE - Neural Network

Training

Cocoa_Percent

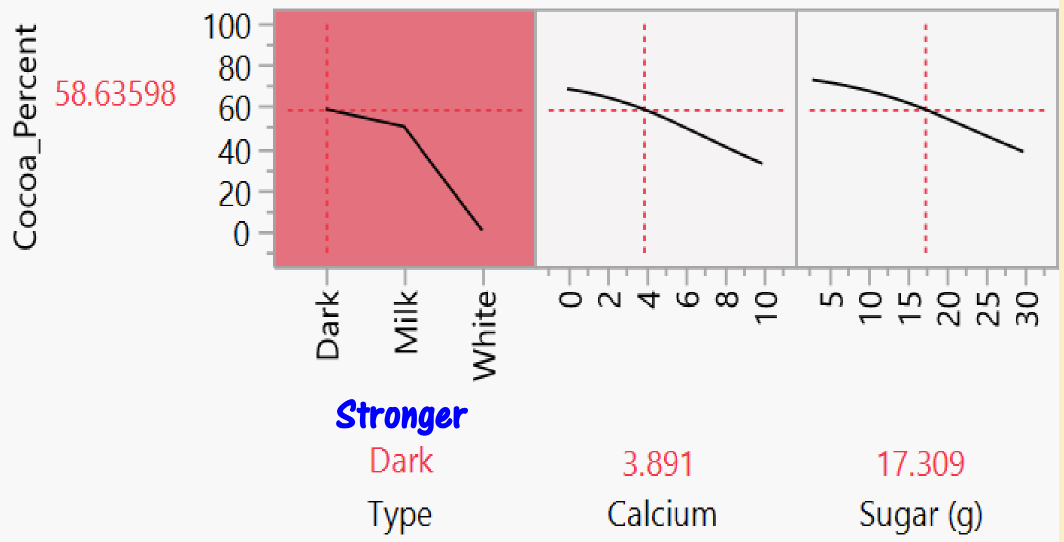
Measures	Value
RSquare	1
RMSE	1.5224e-7
Mean Abs Dev	1.1137e-7
-LogLikelihood	-214.1829
SSE	3.477e-13
Sum Freq	15

Validation

Cocoa_Percent

Measures	Value
RSquare	0.7898796
RMSE	5.9590559
Mean Abs Dev	4.5548399
-LogLikelihood	25.630805
SSE	284.08277
Sum Freq	8

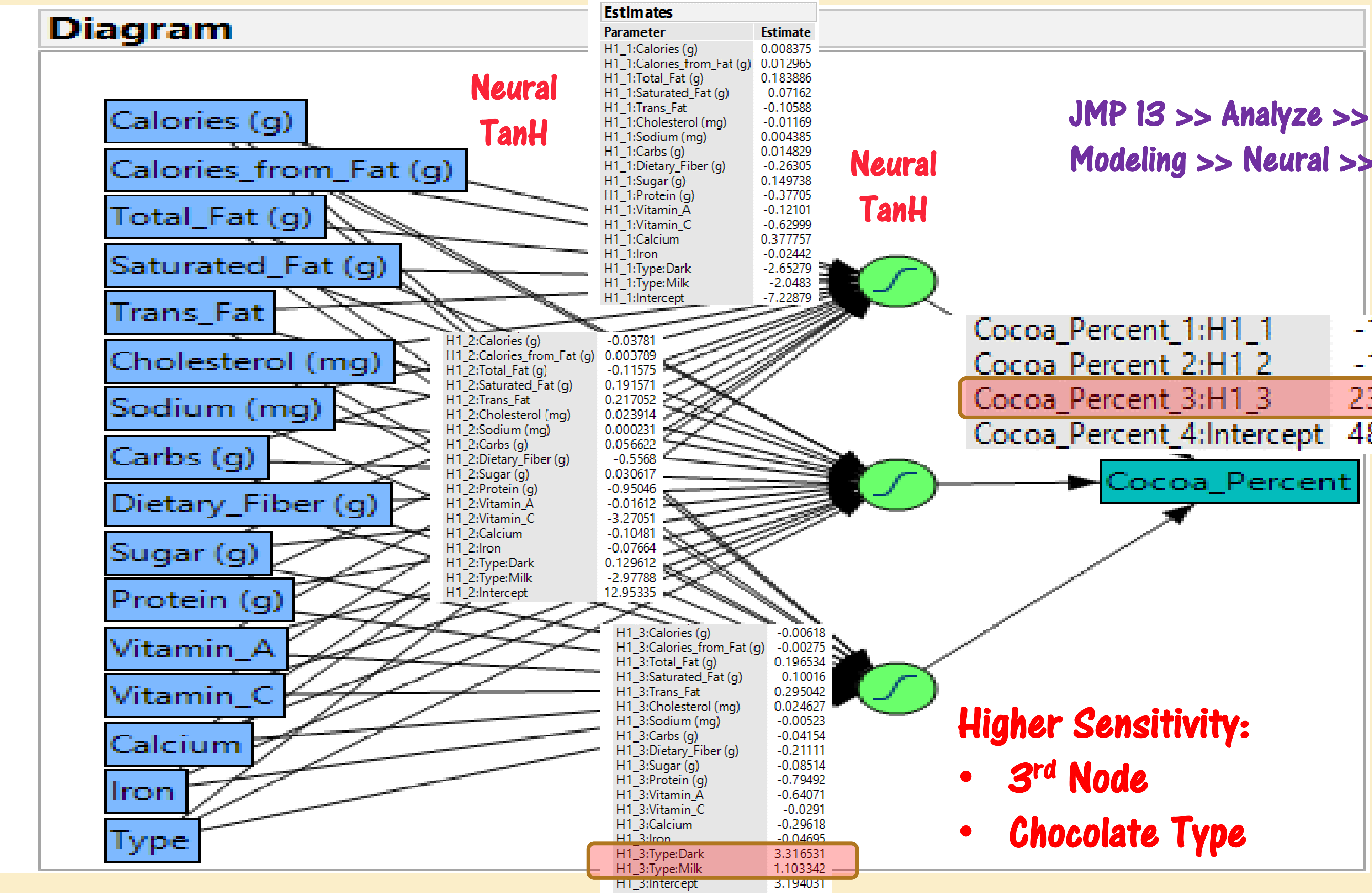
Prediction Profiler



JMP 13 >> Analyze >> Predictive Modeling >> Neural

- The R-square of both Training and Validation are above 0.7.
- Though validation portion is weaker (typical Over-fit concern for Neural).
- Chocolate Type, Calcium, and Sugar as top three predictors for predicting the Cocoa%

Neural Network- Estimates (JMP Default Setting)



Estimates

Parameter	Estimate
H1_1:Calories (g)	0.008375
H1_1:Calories_from_Fat (g)	0.012965
H1_1:Total_Fat (g)	0.183886
H1_1:Saturated_Fat (g)	0.07162
H1_1:Trans_Fat	-0.10588
H1_1:Cholesterol (mg)	-0.01169
H1_1:Sodium (mg)	0.004385
H1_1:Carbs (g)	0.014829
H1_1:Dietary_Fiber (g)	-0.26305
H1_1:Sugar (g)	0.149738
H1_1:Protein (g)	-0.37705
H1_1:Vitamin_A	-0.12101
H1_1:Vitamin_C	-0.62999
H1_1:Calcium	0.377757
H1_1:Iron	-0.02442
H1_1:Type:Dark	-2.65279
H1_1:Type:Milk	-2.0483
H1_1:Intercept	-7.22879

H1_2:Calories (g)	-0.03781
H1_2:Calories_from_Fat (g)	0.003789
H1_2:Total_Fat (g)	-0.11575
H1_2:Saturated_Fat (g)	0.191571
H1_2:Trans_Fat	0.217052
H1_2:Cholesterol (mg)	0.023914
H1_2:Sodium (mg)	0.000231
H1_2:Carbs (g)	0.056622
H1_2:Dietary_Fiber (g)	-0.5568
H1_2:Sugar (g)	0.030617
H1_2:Protein (g)	-0.95046
H1_2:Vitamin_A	-0.01612
H1_2:Vitamin_C	-3.27051
H1_2:Calcium	-0.10481
H1_2:Iron	-0.07664
H1_2:Type:Dark	0.129612
H1_2:Type:Milk	-2.97788
H1_2:Intercept	12.95335

H1_3:Calories (g)	-0.00618
H1_3:Calories_from_Fat (g)	-0.00275
H1_3:Total_Fat (g)	0.196534
H1_3:Saturated_Fat (g)	0.10016
H1_3:Trans_Fat	0.295042
H1_3:Cholesterol (mg)	0.024627
H1_3:Sodium (mg)	-0.00523
H1_3:Carbs (g)	-0.04154
H1_3:Dietary_Fiber (g)	-0.21111
H1_3:Sugar (g)	-0.08514
H1_3:Protein (g)	-0.79492
H1_3:Vitamin_A	-0.64071
H1_3:Vitamin_C	-0.0291
H1_3:Calcium	-0.29618
H1_3:Iron	-0.04695
H1_3:Type:Dark	3.316531
H1_3:Type:Milk	1.103342
H1_3:Intercept	3.194031

Cocoa_Percent_1:H1_1	-12.6766
Cocoa_Percent_2:H1_2	-12.9672
Cocoa_Percent_3:H1_3	23.00243
Cocoa_Percent_4:Intercept	48.51717

JMP 13 >> Analyze >> Predictive Modeling >> Neural >> Diagram

Higher Sensitivity:

- 3rd Node
- Chocolate Type

5. DSD of Neural Setting

JMP 13 >> DOE >> Definitive Screening Design

JMP 13 >> Analyze >> Fit Y by X

JMP 13 >> DOE >> Design Diagnostic >> Evaluate Design

JMP 13 >> Analyze >> Distribution

JMP 13 >> Save Script >> To Data Table or To Script Window >> Edit/Save/Run Script

JMP 13 >> Analyze >> Fit Model

- Mason C., "Optimize Neural Network Algorithm of Missing Value Imputation", submitted to 2019 ASA ENAR Spring Meeting

ENGINEERING

Resolve Neural Over-Fit Concern

JMP 13 >> DOE >> Definitive Screening Design

Objective: optimize Neural settings to resolve over-fit by improving R-Square of both Training and Validation for **Cocoa Missing Imputation**

JMP Neural Validation Methods:

- **Holdback:** randomly divides the original data into the training and validation (holdback portion) sets.
- **Kfold:** divides the data into K subsets. **Each K** set used to validate the model fit on the rest of the data, fitting a total of **K models**. Chose model giving the **best validation statistic**. Best for small data sets (makes efficient use of limited data)

Four DOE Input Variables:

- Validation Method (Categorical)
- Validation Setting (Continuous) "**Nested**" under Validation Method
- Random Seed (Categorical)
- Number of Hidden Nodes (Continuous)

Two DOE Output Responses:

- R-Square of Training Set
- R-Square of Validation Set (**More Important-Neural Over-fit)**

E N GI N EE RI N G

Evaluate DSD of Optimizing Neural Settings

JMP 13 >> DOE >> Design Diagnostic >> Evaluate Design

Power Test of Sign (> 90%)

Correlation of Confounding (<0.3)

Uniformity of Prediction Power

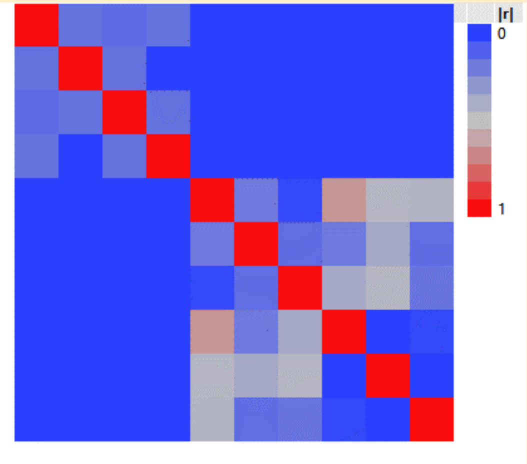
14 DSD Runs

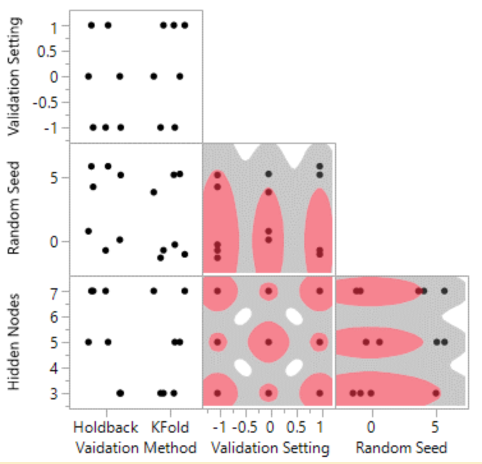
Power Analysis

Significance Level	0.05
Anticipated RMSE	1

Term	Anticipated Coefficient	Power
Intercept	1	0.913
Vaidation Method	1	0.89
Validation Setting	1	0.776
Random Seed	1	0.89
Hidden Nodes	1	0.783

Apply Changes to Anticipated Coefficients





Add Four Random Corner Points

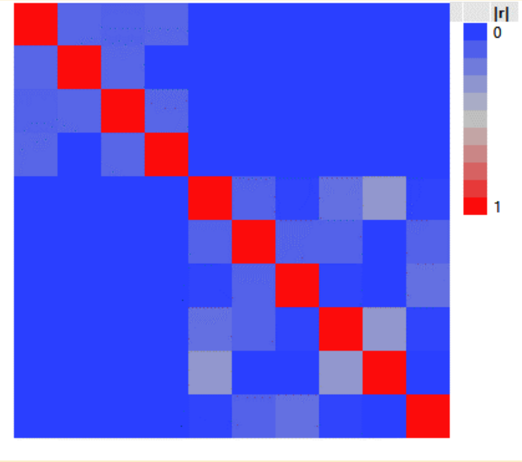
✓ 18 DSD Runs is safer on Power

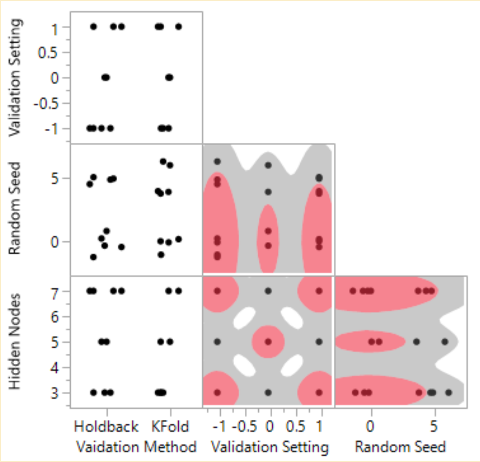
Power Analysis

Significance Level	0.05
Anticipated RMSE	1

Term	Anticipated Coefficient	Power
Intercept	1	0.975
Vaidation Method	1	0.97
Validation Setting	1	0.925
Random Seed	1	0.97
Hidden Nodes	1	0.925

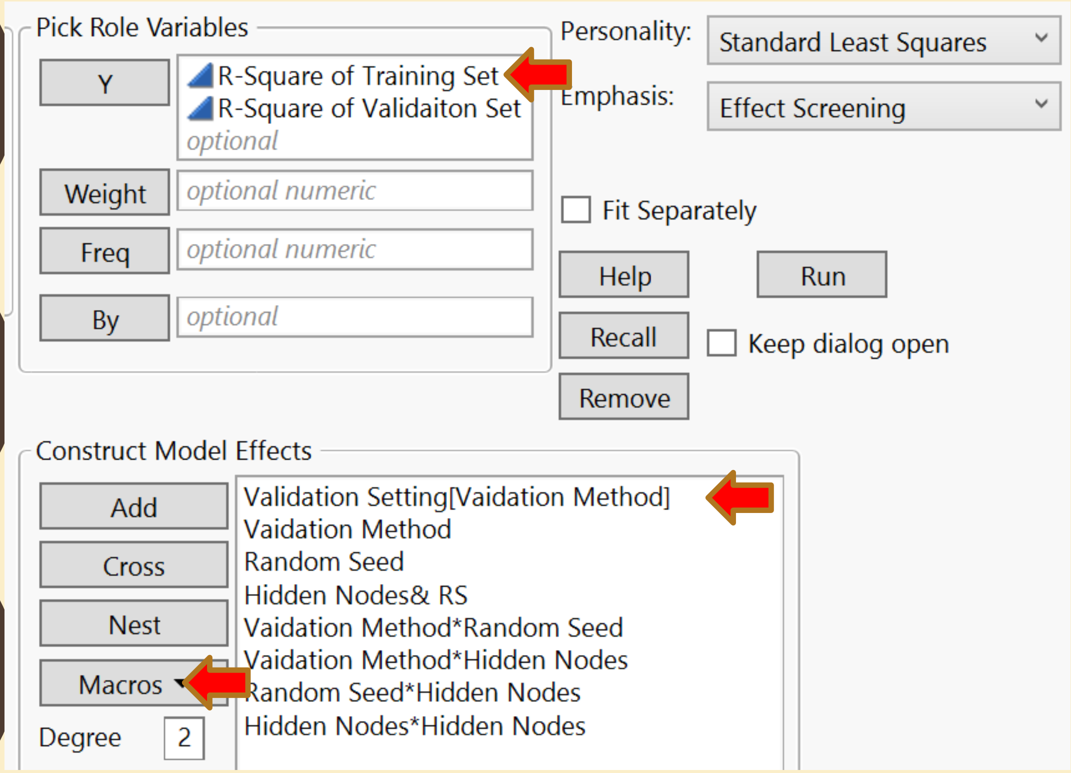
Apply Changes to Anticipated Coefficients

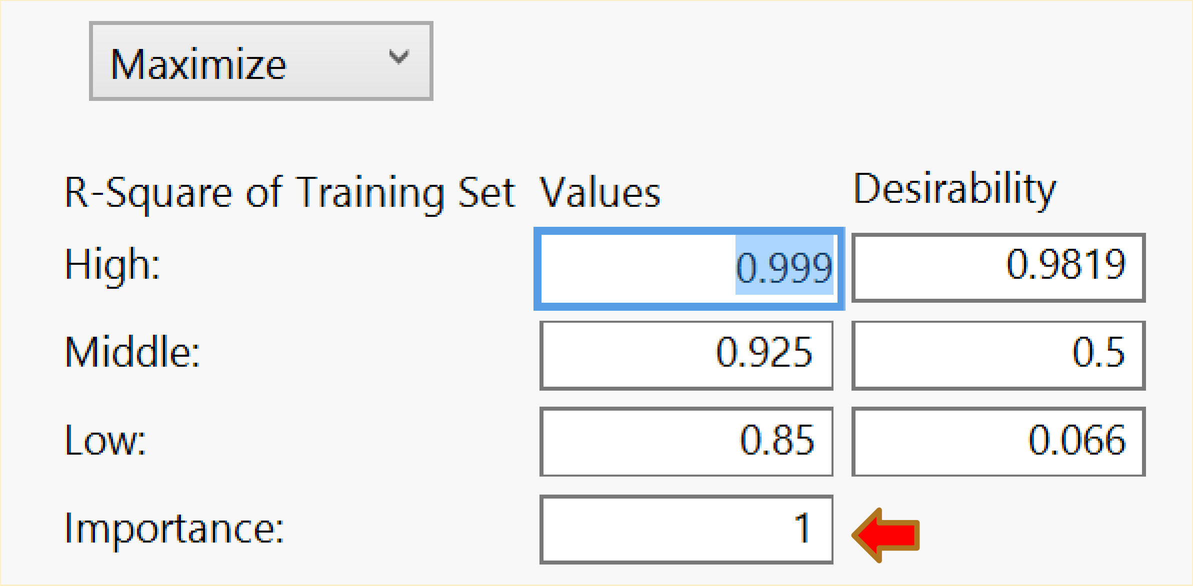


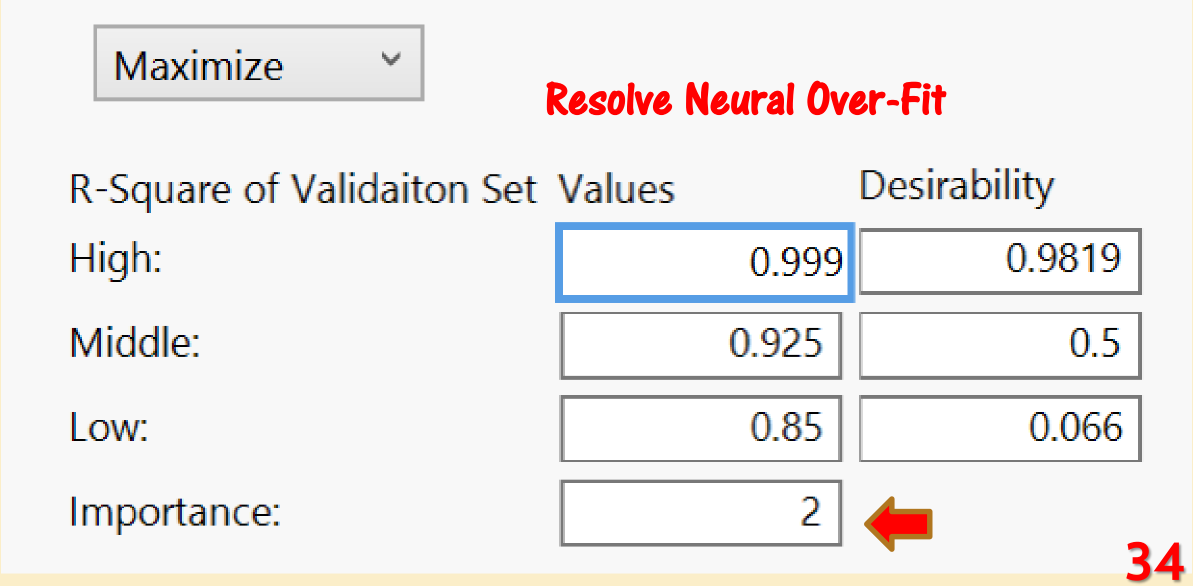


Fit Model (Nested) and Set Desirability

JMP 13 >> Analyze >> Fit Model







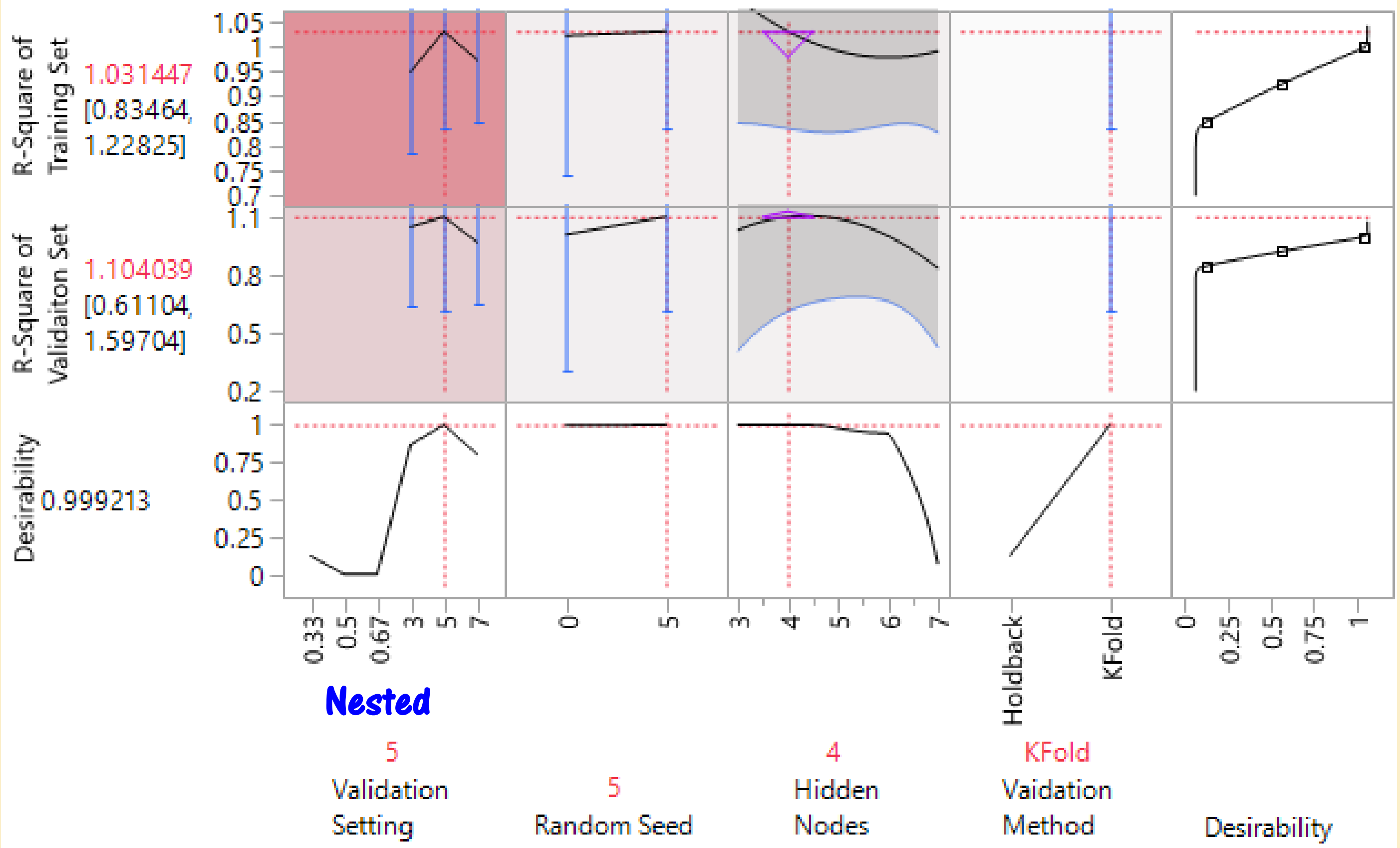
Construct Model Effects:

- Validation Setting is "Nested" under Validation Method
- Choose Response Surface (RS)

Optimal Neural Network Setting

JMP 13 >> Analyze >> Fit Y by X

S T A T I S T I C S



Optimal Neural Setting:

- **Kfold** is better than Holdback (small sample size and favor validation)
- **5 Kfold** numbers (24/5 ~ 5 data for validation set)
- Use Random **Seed= 5 to** improve reproducibility
- **4 Hidden Nodes** is best (Constrained by 15 input variables for one layer)
- Achieve **>99% R-Square** fit on predicting Cocoa%

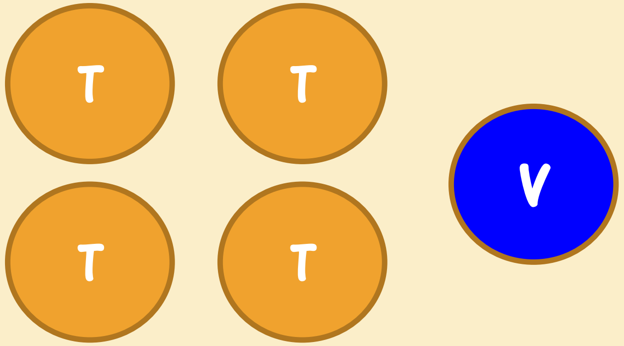
Future Work:

- R-Square > 100%, not following Normal Distribution?
- Wider Confidence Interval: Small Validation Dataset, or Outliers?

Understand Neural Optimization (Future Work)

Why Kfold over Holdback?

Holdback Portion = 0.2



Kfold K=5, Select the Best among 5 Models

Consider Neural Over-fit (lower Validation R-Square)

- If K is large, small size in each K cluster, making validation Over-Fit concern worse
- If K is small, losing advantage of using Kfold over Holdback
- When total sample size is smaller, may prefer Kfold method with smaller K number

Coincidence with Four Hidden Nodes?

- The optimal Neural suggests four hidden nodes of transforming the 15 Input Nutrition Variables
- Section 2 Clustering Variables also suggests four clusters
- Neural related to PCA Eigen algorithm (TanH ~ Linear)?

E N GI N E E RI N G

Neural Model Enhancement

JMP 13 >> Analyze >> Predictive Modeling >> Neural

Original Neural Setting

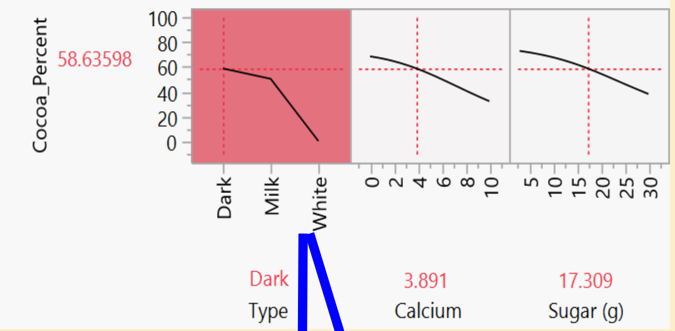
Training — Cocoa_Percent

Measures	Value
RSquare	1
RMSE	1.5224e-7
Mean Abs Dev	1.1137e-7
-LogLikelihood	-214.1829
SSE	3.477e-13
Sum Freq	15

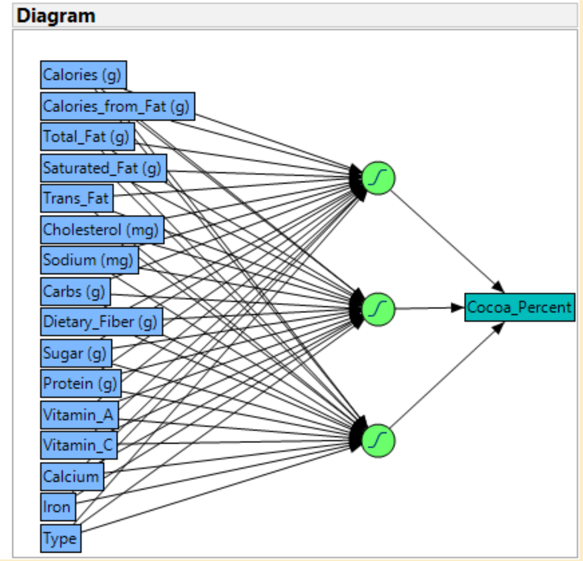
Validation — Cocoa_Percent

Measures	Value
RSquare	0.7898796
RMSE	5.9590559
Mean Abs Dev	4.5548399
-LogLikelihood	25.630805
SSE	284.08277
Sum Freq	8

Prediction Profiler



Diagram



Validation R-Square improved by **20%**

Indicating 4th Chocolate Type- Fruit Chocolate (Vitamin C)

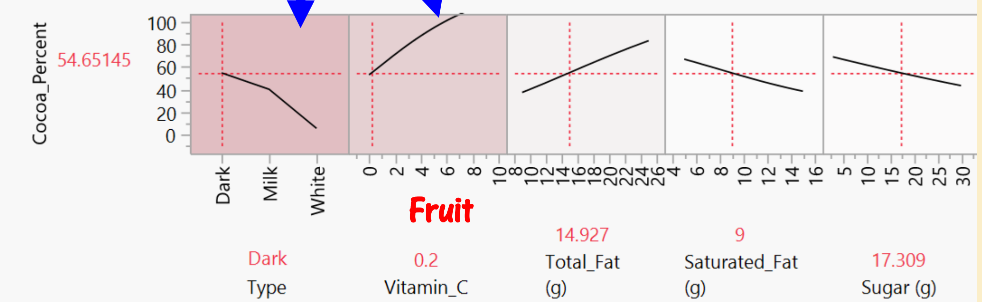
Training — Cocoa_Percent

Measures	Value
RSquare	0.9997931
RMSE	0.3087843
Mean Abs Dev	0.2368498
-LogLikelihood	4.6326999
SSE	1.8116074
Sum Freq	19

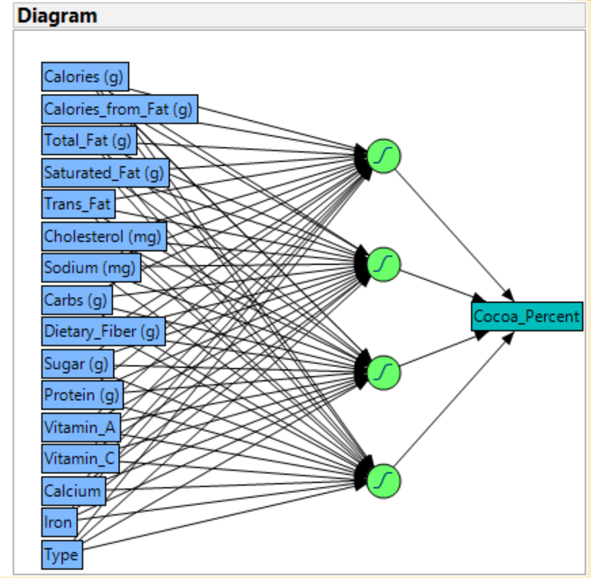
Validation — Cocoa_Percent

Measures	Value
RSquare	0.9982076
RMSE	0.7109563
Mean Abs Dev	0.5761011
-LogLikelihood	4.311177
SSE	2.0218356
Sum Freq	4

Prediction Profiler



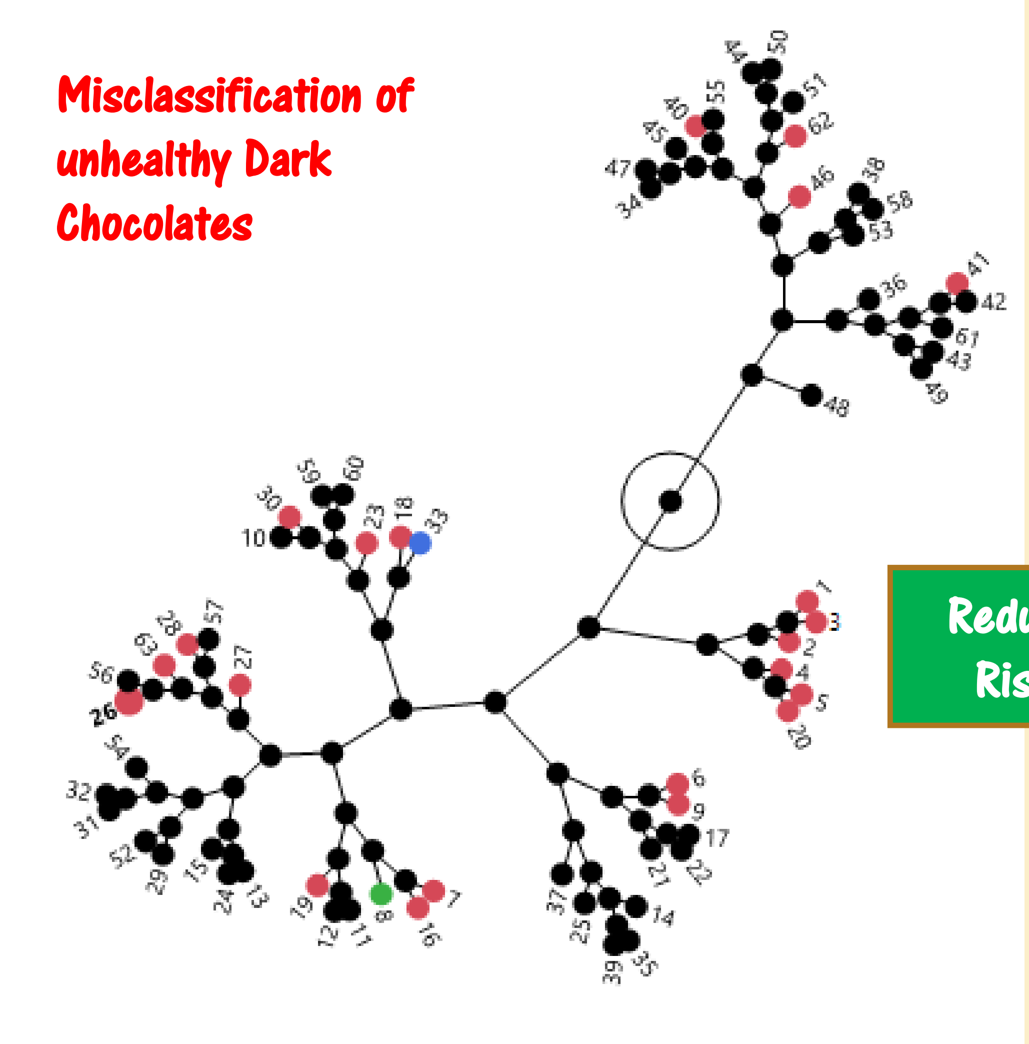
Diagram



Optimal DSD Neural Setting

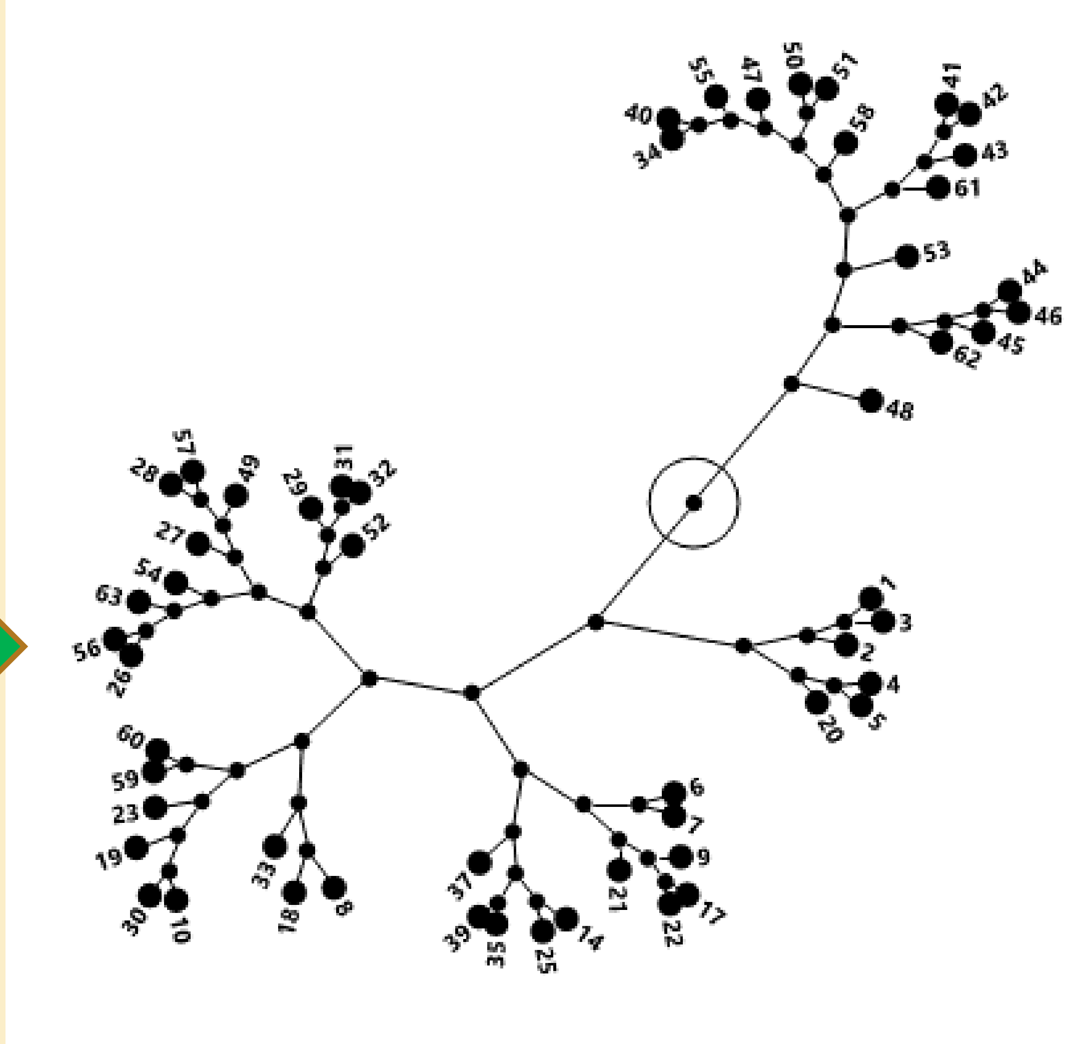
ENGINEERING

Default: Missing Value Imputation



Reduce Risk

Optimal Neural (Predicted Cocoa %)



Achievements AND FUTURE RESEARCH

Achievements:

- ✓ Adopted and Integrated "STEAMS" methodology successfully
- ✓ Learned Chocolate Products, Nutrition Anti-Oxidant Science
- ✓ Applied Multivariate Statistics, Clustering and Neural Algorithms
- ✓ Conducted DSD optimization on Resolving Neural Overfit

Future JMP Research:

- ❑ Investigate "Fruit" Chocolate Type, Outlier Effect
- ❑ JMP Pro Partition: Bootstrap Forrest, Boosted Tree, K-Nested, Naïve Bayes
- ❑ JMP Pro Neural: Deep Learning, Hidden Layer Structure, Fitting Options
- ❑ Certify JMP Script Specialist

JAVA/Latex Advisor: Dr. Ying Huang

Biology Advisor: Dr. I-Chen Chen

STEAMS Advisor: JMP Chuck Boiler

STEAMS Advisors: Stanford OHS

Statistics//MATH/AI Advisor: Dr. Charles Chen

Q&A

Thank YOU

Biology/Writing Advisor: CMQ/OE Patrick Giuliano

STEAMS Advisor: ASA Dr. Chris Barker

STEAMS Advisor:
IEOM Dr. Ali Ahad,
IEOM Dr. Don Reimer

STEAMS Advisor: ASQ Fellow Dr. John Flaig

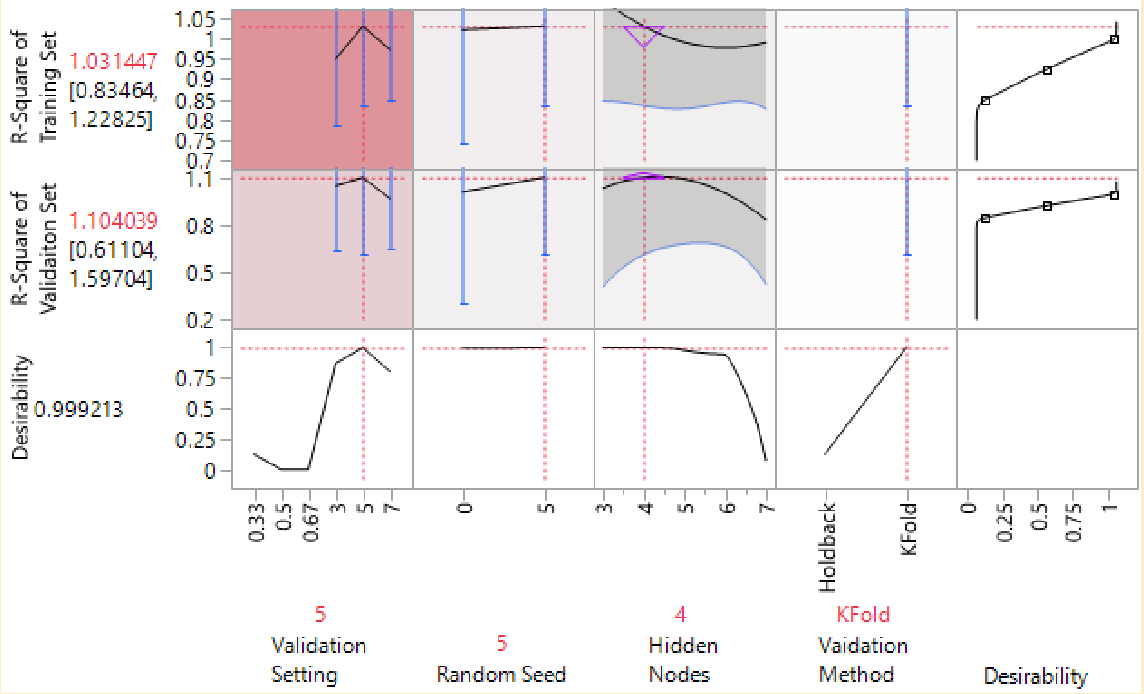
Robotics Advisor: CQE/CRE Roland Jones

Optimize Desirability Function Importance

ENGINEERING

Optimize the Number of Hidden Nodes:

- Higher R-Square of Training Set at Nodes=3
- Higher R-Square of Validation Set at Nodes=4
- Relative Importance can impact the Optimal Setting



Importance Setting		Optimization Result	Optimal Neural Settings			
Importance of Training Set	Importance of Validation Set	Overall Desirability	Validation Method	Validation Setting	Random Seed	Hidden Nodes
1	1	0.9997	KFold	5	5	3
1	1.5	0.9997	KFold	5	5	4
1	2	0.9997	KFold	5	5	4
1.5	1	0.9997	KFold	5	5	3
1.5	1.5	0.9997	KFold	5	5	3
1.5	2	0.9997	KFold	5	5	4
2	1	0.9997	KFold	5	5	3
2	1.5	0.9997	KFold	5	5	3
2	2	0.9997	KFold	5	5	3

Conduct a 2-Factor & 3-Level Full Factorial on comparing the relative importance in (1,2) range

- Set the Desirability Function Range of (0.999, 0.95, 0.9)
- In General, the optimal result shows the similar trend: 3 hidden nodes favor training set and 4 hidden nodes favor validation set

Little room for further improvement on setting the relative importance between Training Set and Validation Set

Document Key JMP Scripts

JMP 13 >> Save Script >> To Data Table or To Script Window >> Edit/Save/Run Script

AI

```
Partition(
Y( :Cocoa_Percent ),
X(
:Type,
:Name( "Calories (g)" ),
:Name( "Calories_from_Fat (g)" ),
:Name( "Total_Fat (g)" ),
:Name( "Saturated_Fat (g)" ),
:Trans_Fat,
:Name( "Cholesterol (mg)" ),
:Name( "Sodium (mg)" ),
:Name( "Carbs (g)" ),
:Name( "Dietary_Fiber (g)" ),
:Name( "Sugar (g)" ),
:Name( "Protein (g)" ),
:Vitamin_A,
:Vitamin_C,
:Calcium,
:Iron
),
Minimum Size Split( 3 ),
Validation Portion( 0.6 ),
Split History( 1 ),
Informative Missing( 1 ),
Column Contributions( 1 ),
Initial Splits( :Name( "Cholesterol
(mg)" ) >= 5 ),
SendToReport( Dispatch( {}, "Partition",
FrameBox, {Frame Size( 480, 56 )} ) )
);
```

```
Neural(
Y( :Cocoa_Percent ),
X(
:Name( "Calories (g)" ),
:Name( "Calories_from_Fat (g)" ),
:Name( "Total_Fat (g)" ),
:Name( "Saturated_Fat (g)" ),
:Trans_Fat,
:Name( "Cholesterol (mg)" ),
:Name( "Sodium (mg)" ),
:Name( "Carbs (g)" ),
:Name( "Dietary_Fiber (g)" ),
:Name( "Sugar (g)" ),
:Name( "Protein (g)" ),
:Vitamin_A,
:Vitamin_C,
:Calcium,
:Iron,
:Type
),
Informative Missing( 0 ),
Validation Method(
"KFold", 5 ),
Fit(
NTanH( 4 ),
Diagram( 1 )
),
```

```
Fit Model(
Y( :Name( "R-Square of Training Set" ),
:Name( "R-Square of Validaiton Set" ) ),
Effects(
:Validation Setting[:Vaidation Method],
:Vaidation Method,
:Random Seed,
:Hidden Nodes & RS,
:Vaidation Method * :Random Seed,
:Vaidation Method * :Hidden Nodes,
:Random Seed * :Hidden Nodes,
:Hidden Nodes * :Hidden Nodes
),
Personality( "Standard Least Squares" ),
Emphasis( "Effect Screening" ),

:Name( "R-Square of Training Set" ) << {Summary of Fit( 0 ),
Analysis of Variance( 0 ), Lack of Fit( 0 ), Sorted
Estimates( 0 ),
Plot Actual by Predicted( 1 ), Plot Regression( 0 ),
Plot Residual by Predicted( 1 ), Plot Studentized Residuals(
1 ),
Plot Effect Leverage( 0 ), Box Cox Y Transformation( 1 )},
:Name( "R-Square of Validaiton Set" ) << {Summary of Fit( 0
),
Analysis of Variance( 0 ), Lack of Fit( 0 ), Sorted
Estimates( 0 ),
Plot Actual by Predicted( 1 ), Plot Regression( 0 ),
Plot Residual by Predicted( 1 ), Plot Studentized Residuals(
1 ),
Plot Effect Leverage( 0 ), Box Cox Y Transformation( 1 )}
),
```