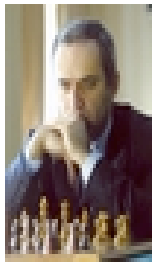


Essentials of Machine Learning (ML) for Business Leaders



“As one Google Translate engineer put it, “when you go from 10,000 training examples to 10 billion training examples, it all starts to work. Data trumps everything.”

– Garry Kasparov, Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins

Your Presenter - Prashanth Southekal

- Managing Principal of DBP-Institute
- Over 20 years of Information Management experience consulting for over 45 organizations including SAP AG, Shell, Apple, P&G, and General Electric.
- Author of the book - *Data for Business Performance*.
- Advisory board member at Grihasoft (India)
- Adjunct faculty of Data Analytics at University of Calgary (Canada) & IE Business School (Spain).



House Keeping

- This is a ~3 Hour Training
- My XL sheet is needed!
- Small changes to the slides
- Cell Phones/Laptops - Use your judgement.
- Questions/Comments - Any time!
- Pictures for LinkedIn/Twitter
- After Hours - Available after this training to discuss your unique questions.
- Door Prize - My Book; Signed by me (SAS Sponsored)

Course/Learning Objectives

1. Demystifying ML Solution Landscape
2. Strategies to acquire Quality Data
3. Understanding the 4 Key ML Algorithms
4. Implementing your ML Solution

Guest Presenter
Matt Joyce, Pre-Sales Lead, SAS-Canada

Designed for Senior IT and Business Professionals

Introduction: Definition of Analytics

What is your definition of Analytics?

Introduction: Definition of Analytics

Analytics is using **data** by asking relevant **questions** to gain **insights** for **decision making**

Table of Contents

1

- Introduction

2

- Quality Data

3

- Algorithms

4

- Embedded Analytics

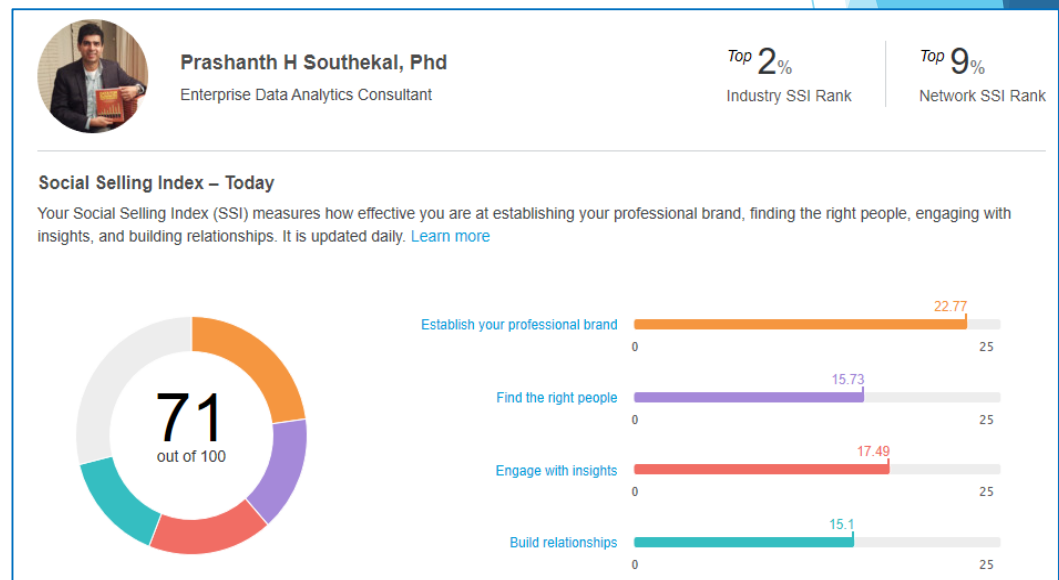
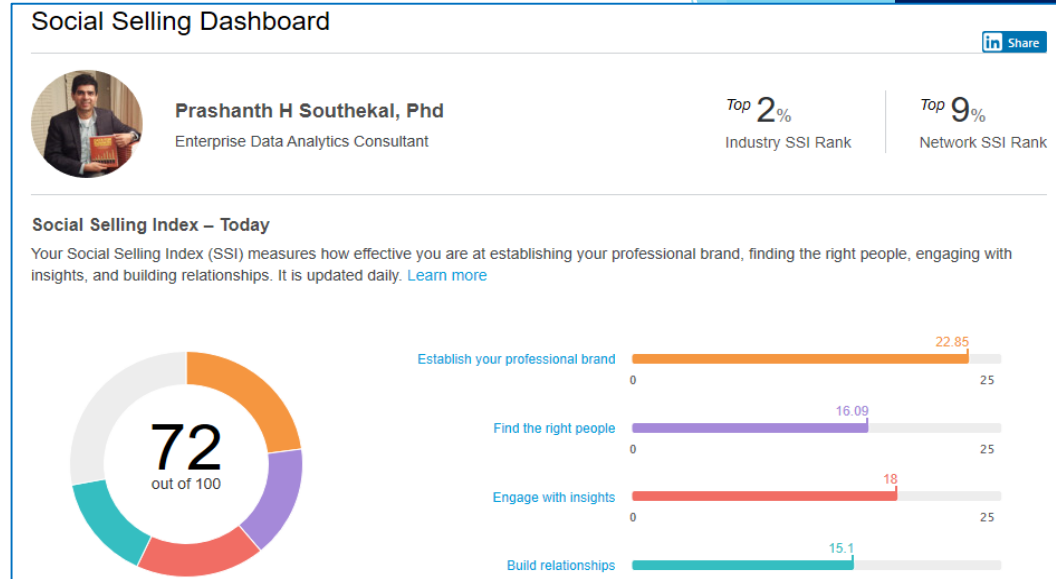
What is Machine Learning?



Machine learning is using **large amount of (quality) data** to implement **analytics insights** with **minimal human intervention**

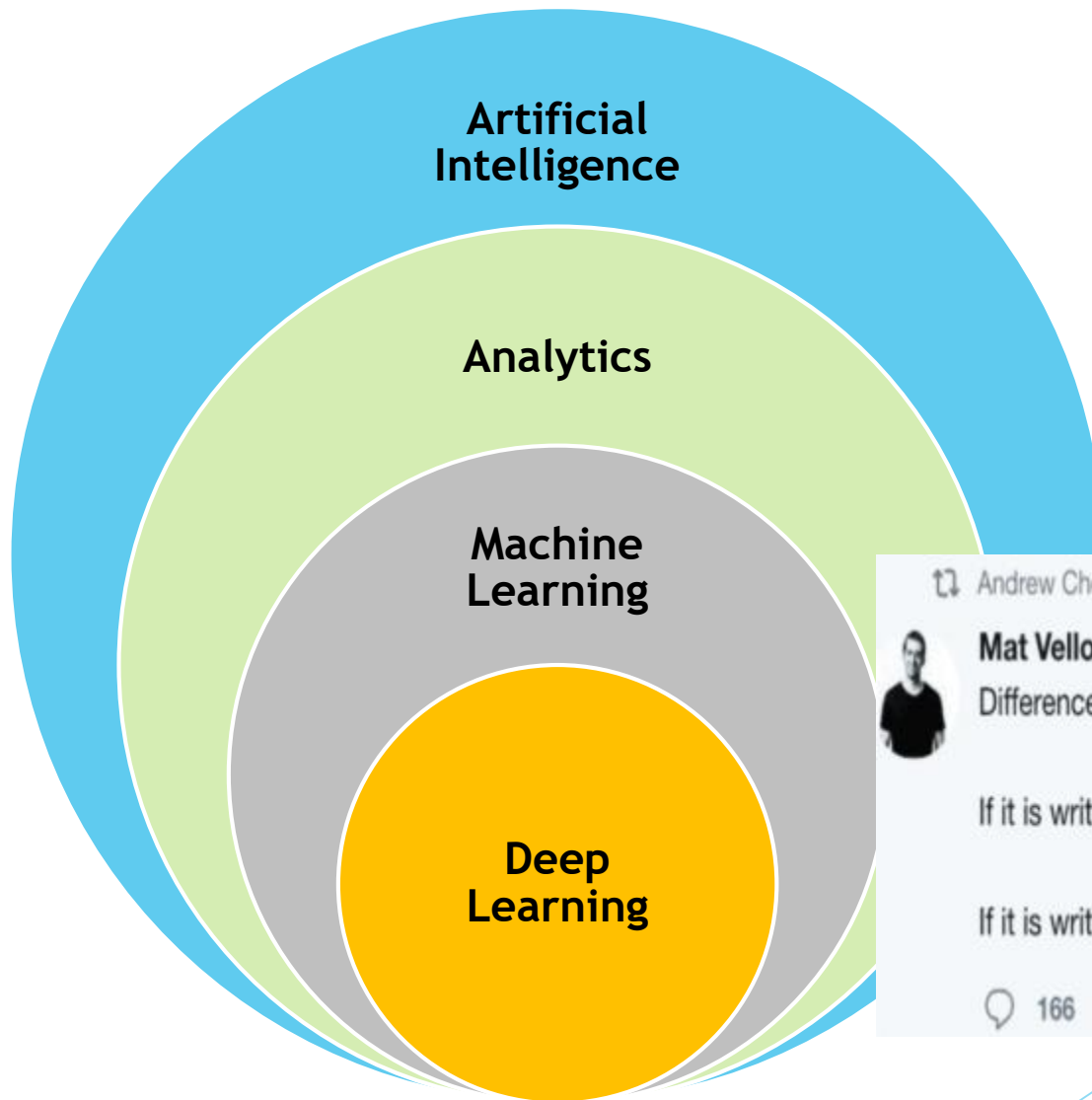
ML Examples

- Google Home & Amazon Alexa
- Self-driving Google car
- Recommendations from Amazon & Netflix
- Customer Churn in Retail
- Weather pattern prediction in Agriculture
- LinkedIn Posts Traction & SSI



Where does ML fit?

Layers in Predictive Analytics



ML and Predictive Analytics

- Machine learning (ML) works out **predictions** and recalibrates models in real-time **automatically**.
- Predictive analytics works strictly on “cause” data and must be **refreshed with “changed” data**.
- Unlike ML, predictive analytics still relies on **human experts** to work out and test causes and outcomes.

- ML is Dynamic; Predictive Analytics is Static
- ML is Data Driven; Predictive Analytics is Analyst Driven



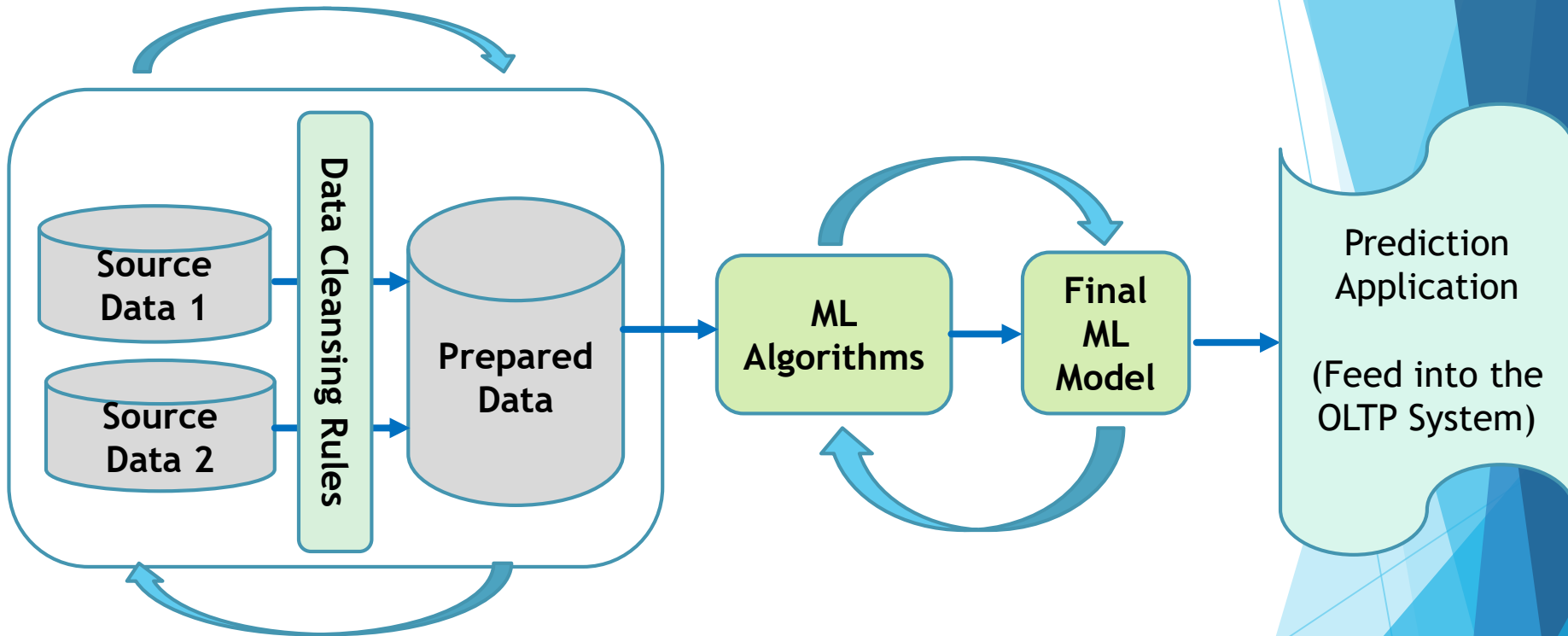
<https://www.youtube.com/watch?v=XH1wQEgROg4>

ML Process

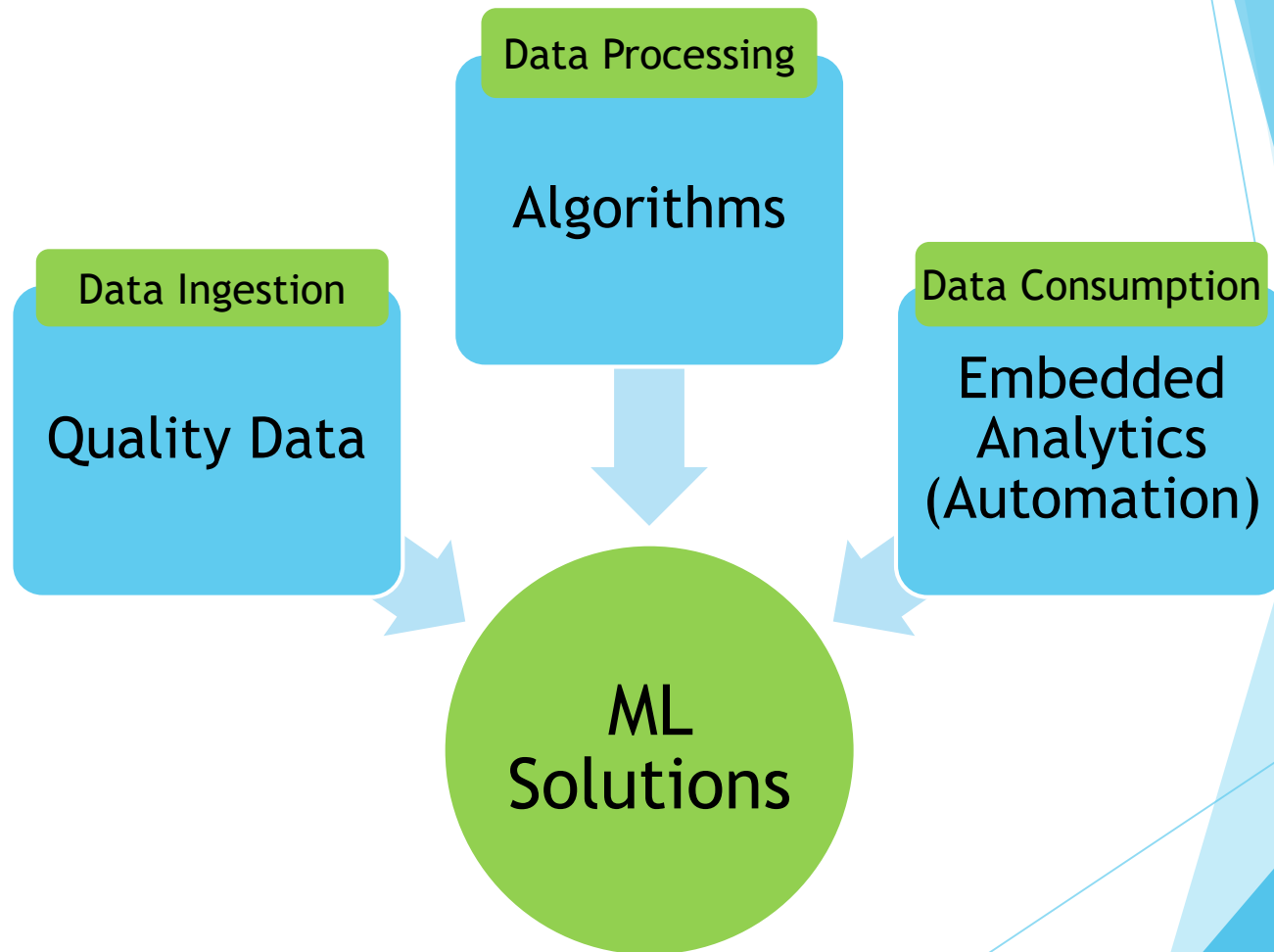
Data Ingestion

Data Processing

Data Consumption



ML Solutions



ML V/s Non-ML Solutions

A solution is deemed to be a ML solution if it meets 4 key criteria:

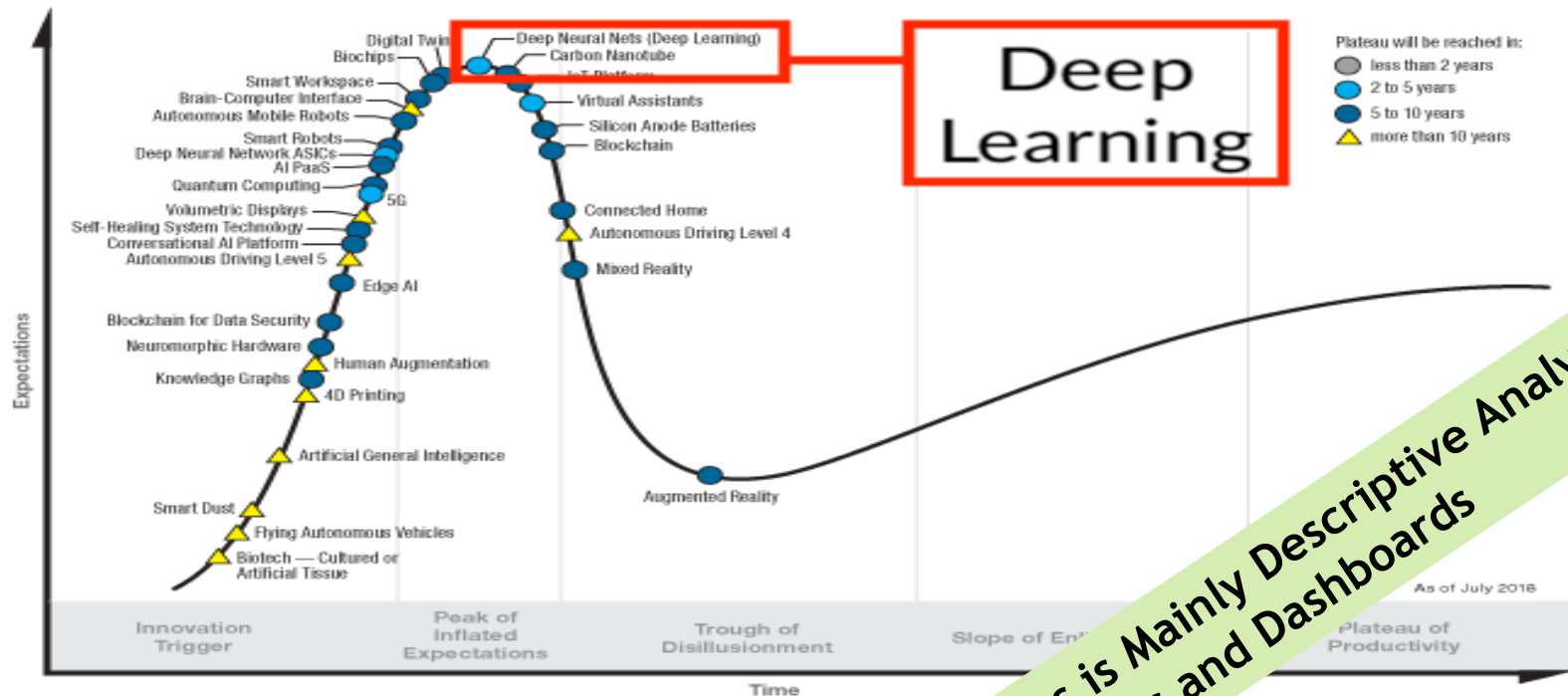
1. The output is **REFINED CONTINUOUSLY** i.e. data ingestion into the ML algorithm is continuous
2. There is **MINIMAL (OR EVEN ZERO) HUMAN INTERVENTION** in deriving and applying the output
3. The output is **PROBABLISTIC** as the solution is geared toward the FUTURE STATE.
4. The output provides answers to questions on mainly **EVENTS or TRANSACTIONS** (over entities or categories)

ML Implementation - Important Platforms



Current State of ML/DL

Hype Cycle for Emerging Technologies, 2018



gartner.com/SmarterWithGartner

Source: Gartner (August 2018)

© 2018 Gartner, Inc. and/or its affiliates. All rights reserved.

Gartner.

**~85% of Analytics is Mainly Descriptive Analytics
- Reports and Dashboards**

Table of Contents

1

- Introduction

2

- Quality Data

3

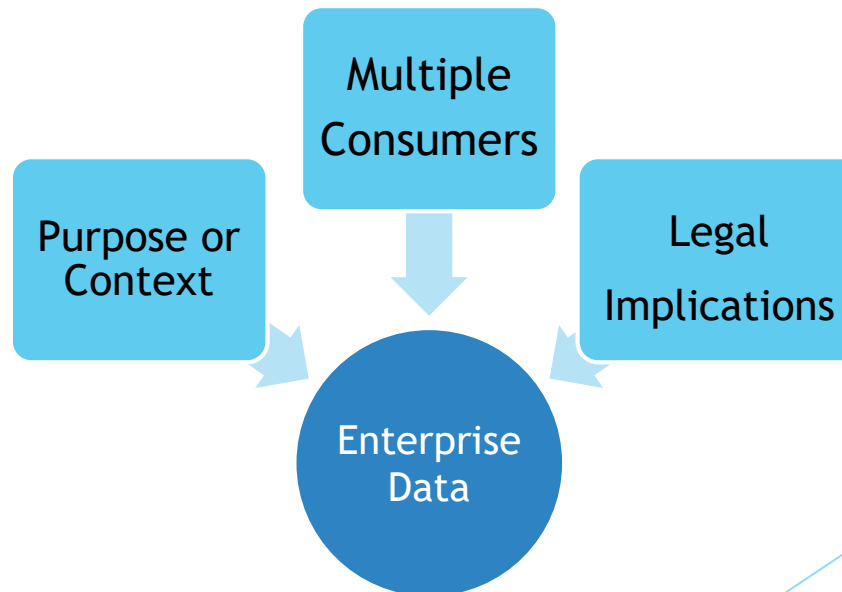
- Algorithms

4

- Embedded Analytics

What is Business Data?

Enterprise or Business Data is the data used in leading, planning, controlling and operating a business organization that provides goods and/or services to its customers



Business Data Characteristics

1. Multiple Stakeholders/Consumers

- Enterprise data is shared and reused
- The same Enterprise data can be viewed & consumed differently; the value varies based on the stakeholder type

2. Purpose/Context Driven

- Enterprise data is tied to the business process
- Enterprise data is heterogenous
 - Diverse sources - Internal and External
 - Data Type and Format; ~80% is unstructured data
 - Diverse technologies

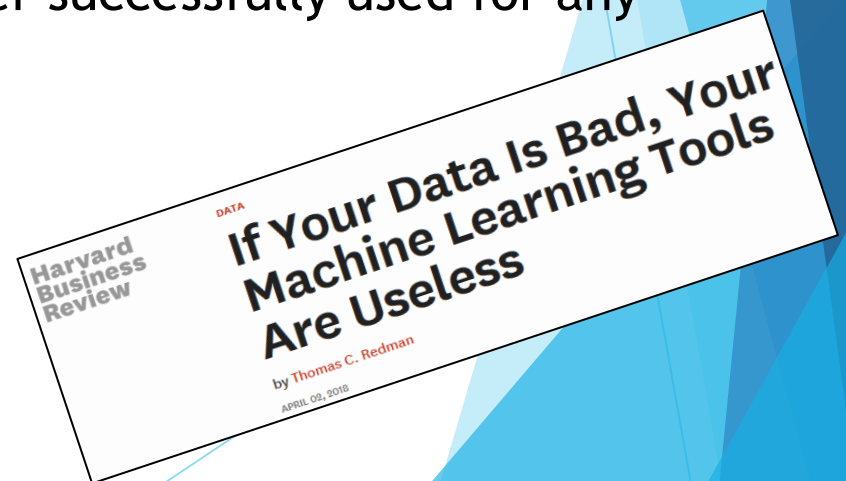
Business Data Characteristics

3. Legal Implications

- Enterprise data is secure & controlled
- Even though Enterprise data grows exponentially, it is rarely purged/archived
- Enterprise data follows compliance mandates
 - Privacy laws (PIPEDA, GDPR etc..)
 - Regulations (SOX, GAPP/IFRS....)
 - Standards (UNSPSC, Internal policies, UPC/EAN,PIDX, NACS, etc.)

The Challenge of Quality Data in Business

- An average user spends 2 hours a day looking for the right data (Mckinsey)
- Just 3% of the data in a business enterprise meets quality standards (HBR)
- Bad data costs 12% of the company's revenue (Experian Data Quality)
- 27% of data in the world's top companies is flawed (Gartner)
- 73% of data in an organization is never successfully used for any strategic purpose (Forrester)



What is Quality Data ?

*“The biggest problem in the analytics is having no idea what you are looking for in the **data**” -*

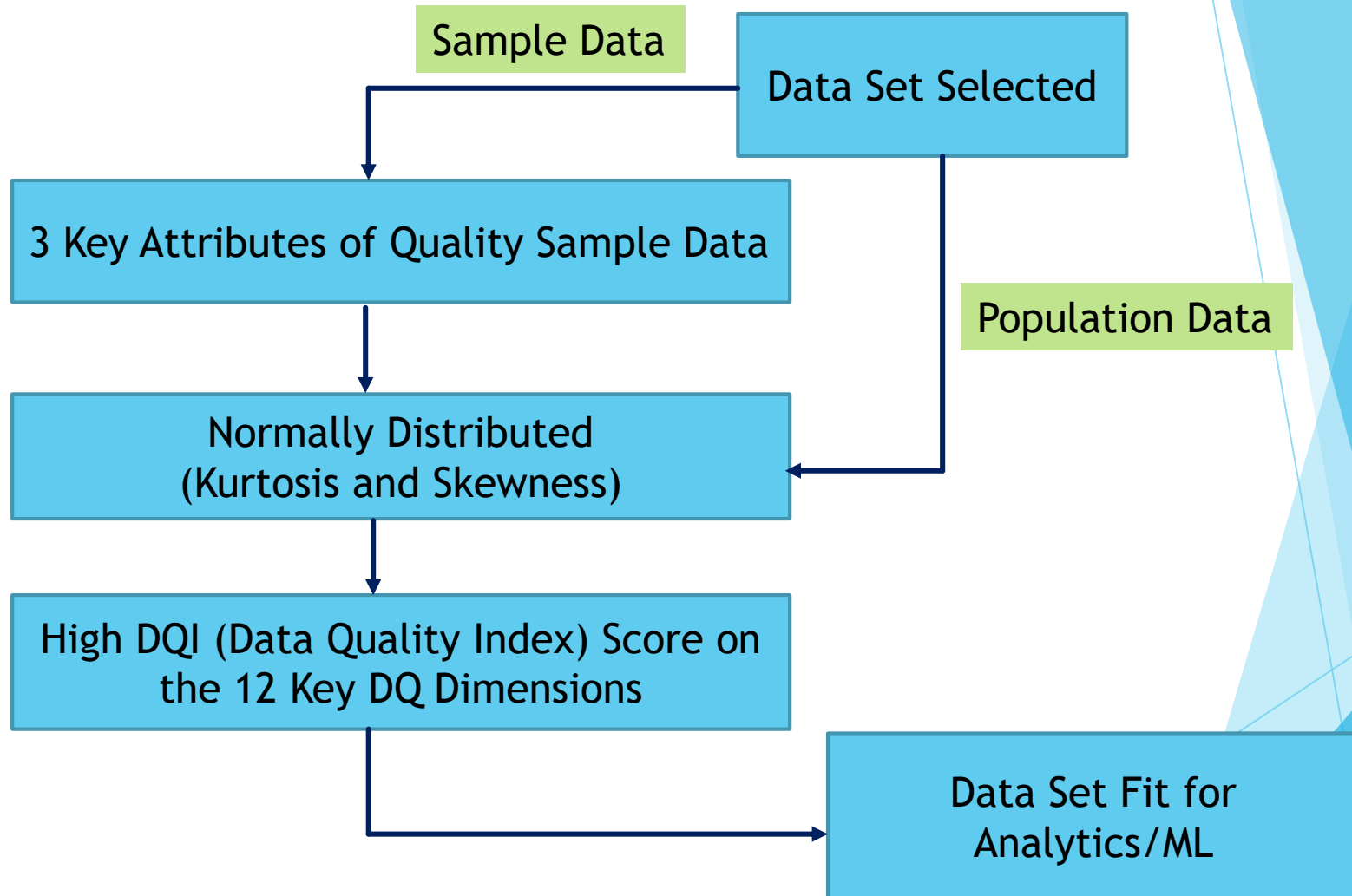
Tom Davenport

Definition of Quality Data ?

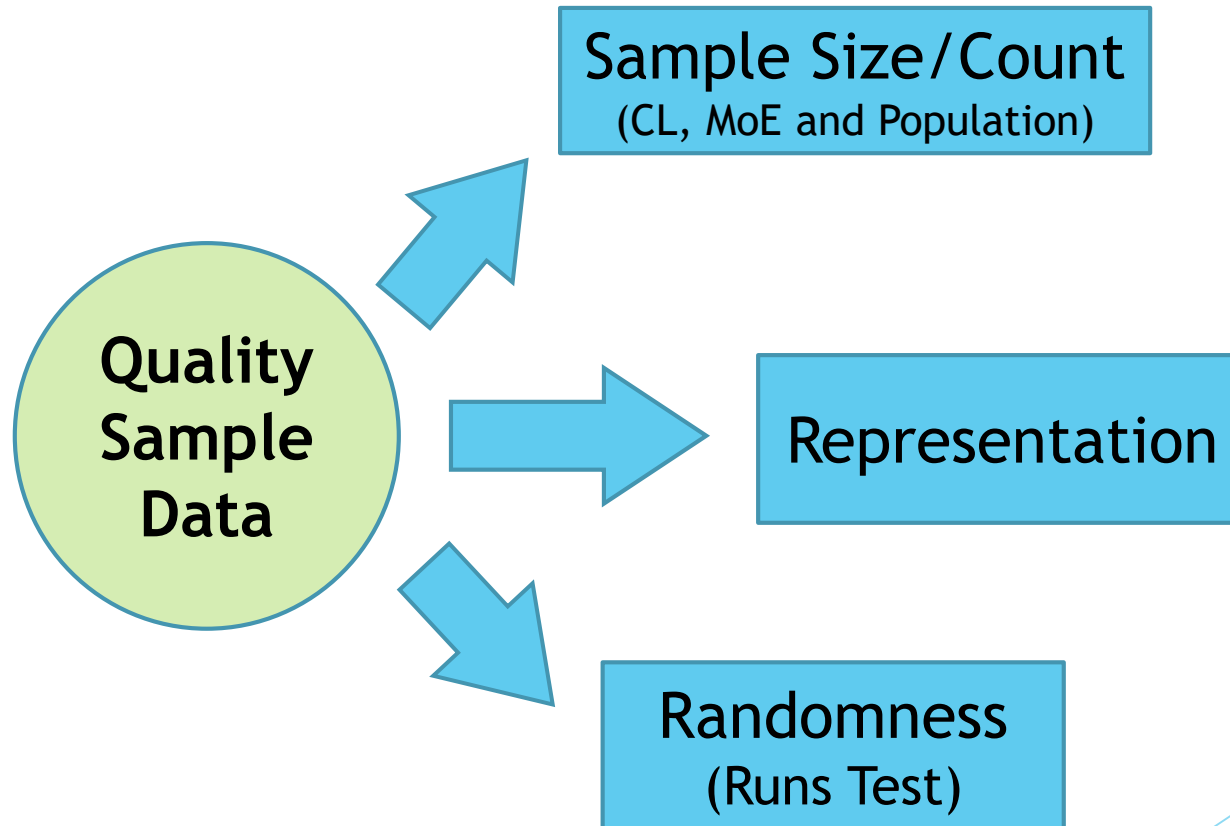
Data Quality is the assessment of data's **fitness** to serve its **purpose** in a given **context**.

The fitness, purpose and context are determined by different dimensions.

Getting Quality Data



Getting Quality Sample Data



Sample Size (SS) for the “Right” Sample

$$\text{Sample size} = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \left(\frac{z^2 \times p(1-p)}{e^2 N} \right)}$$

- N = Population size
- e = Margin of error (MoE) (percentage in decimal form)
- z = z-score. The z-score for 95% confidence level is 1.96

If I need to know the opinion of citizen of Brussels on the 2026 Winter Olympics, the SS of the people to be interviewed are:

Population size ?	Confidence level (%) ?	Margin of error (%) ?
1200000	95	5
Sample size		
385		

Reference - <https://www.surveymonkey.com/mp/sample-size-calculator/>

Representation for the “Right” Sample

Bill Smith leads Nenshi by 17 points in latest poll



Pollster says 'late mistakes' could harm Smith but 'certainty' of his victory is becoming clearer

CBC News · Posted: Oct 07, 2017 6:00 AM MT | Last Updated: October 7, 2017



Mainstreet admits 'big polling failures' after predicting Nenshi would lose Calgary election

'Wonky sample'

Maggi believes his company's methodology isn't fundamentally flawed, but said it's investigating whether it failed to connect with some younger voters and those who don't speak English as their first language.

The company is also looking into the possibility that some people were simply hanging up on the automated calls that Mainstreet relies on to reach voters and gather their opinions.

In particular, Maggi admitted the second poll his company did, which had Smith ahead by 17 points, was based on a "wonky sample."

Quality Sample = Sample Count, Randomness, and Representative Sample

Representation (Runs Test) for the “Right” Sample

- Runs-test is a statistical test that examines whether the data set is occurring randomly.
- The runs test analyzes the occurrence of similar events that are separated by events that are different.

Getting Good Sample Data - Example 1

Do People of Hillsboro, Oregon Love
Shopping in Walmart?

I need a random good sample that
represents the population of Hillsboro,
Oregon for my Survey

A sample data that represents the population is
Representative Data

Getting Random Data Records

Population size ?

Confidence level (%) ?

Margin of error (%) ?

Sample size

68

RANDOM.ORG

Generate random integers (maximum 10,000).

Each integer should have a value between and

**Sample
Count**

**Population
Count**

**Random Data
Records
Identifier**

24378	25736	6769	28237	11286
5990	14816	22596	20084	24858
18066	28703	44	4226	22898
17288	29807	20188	12945	24438
2649	26291	23134	29385	20185
23934	605	12473	3472	7748
20600	23980	23639	21613	22282
11295	24479	704	12361	4742
16534	19782	271	5607	26896
27461	21824	7175	19047	2582
12967	17945	23609	13174	10036
3251	13505	11907	15761	4592
27606	25128	28840	25339	21142
22632	27273	3323		

Is this Representative of the Population?

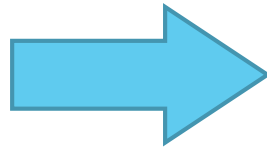
Age	Income	Ethnicity
0-20= 15%	0-40K = 20%	Asian = 35%
21-60 = 65%	40K-80K = 23%	South Asian = 24%
60+ = 20%	80K - 150K = 40%	White= 32%
	150K+ = 17%	Black = 5%
		Others = 4%

**Population
(of 30,000)**

Runs Test on Age (3 Levels)	Runs Test on Income (4 Levels)	Runs Test on Ethnicity (5 Levels)
P-value = 0.032. This means the sample data is random	P-value = 0.041. This means the sample data is random	P-value = 0.019. This means the sample data is random

**Sample
(of 68)**

Getting Quality Data - Population and Sample



1. Entirety
2. Consistency
3. Conformity/Validity
4. Uniqueness/Cardinality
5. Accuracy
6. Correctness
7. Accessibility
8. Security
9. Timeliness
10. Redundancy
11. Coverage
12. Integrity

Applying Data Quality: Case Study

#	Data Quality KPI	UoM (Reference Data)	Product (Master Data)
1	Conformance (Validity) to Standards	61%	0%
2	Completeness (Against mandatory fields)	88%	80%
3	Cardinality (Uniqueness)	100%	98%
4	Currency/Recency	100%	10%
5	Coverage (across 5 key systems)	50%	40%
6	Accuracy (for Data Consumers)	83%	55%
	Data Quality Index (DQI) (Composite Metric)	83%	44%

Do I have Quality Data?

Question	Tool
How do I get an average of a data set?	
Do I have a good data set (i.e. normal distributed data set) for my business analytics?	
Is data set(s) is more accurate over another data set?	
Is data set(s) is more precise over another data set?	
How do I determine Outliers ?	
How do I give good estimates ?	
Does the sample data encompass the 3 key features?	

Do I have Quality Data?

Question	Tool
How do I get an average of a data set?	Mean + Mode + Median
Do I have a good data set (i.e. normal distributed data set) for my business analytics?	Skewness and Kurtosis
Is data set(s) is more accurate over another data set?	Standard Deviation
Is data set(s) is more precise over another data set?	Standard Error
How do I determine Outliers ?	Z-score
How do I give good estimates ?	CI + CL
Does the sample data encompass the 3 key features?	Count + Randomness + Representativeness



Table of Contents

1

- Introduction

2

- Quality Data

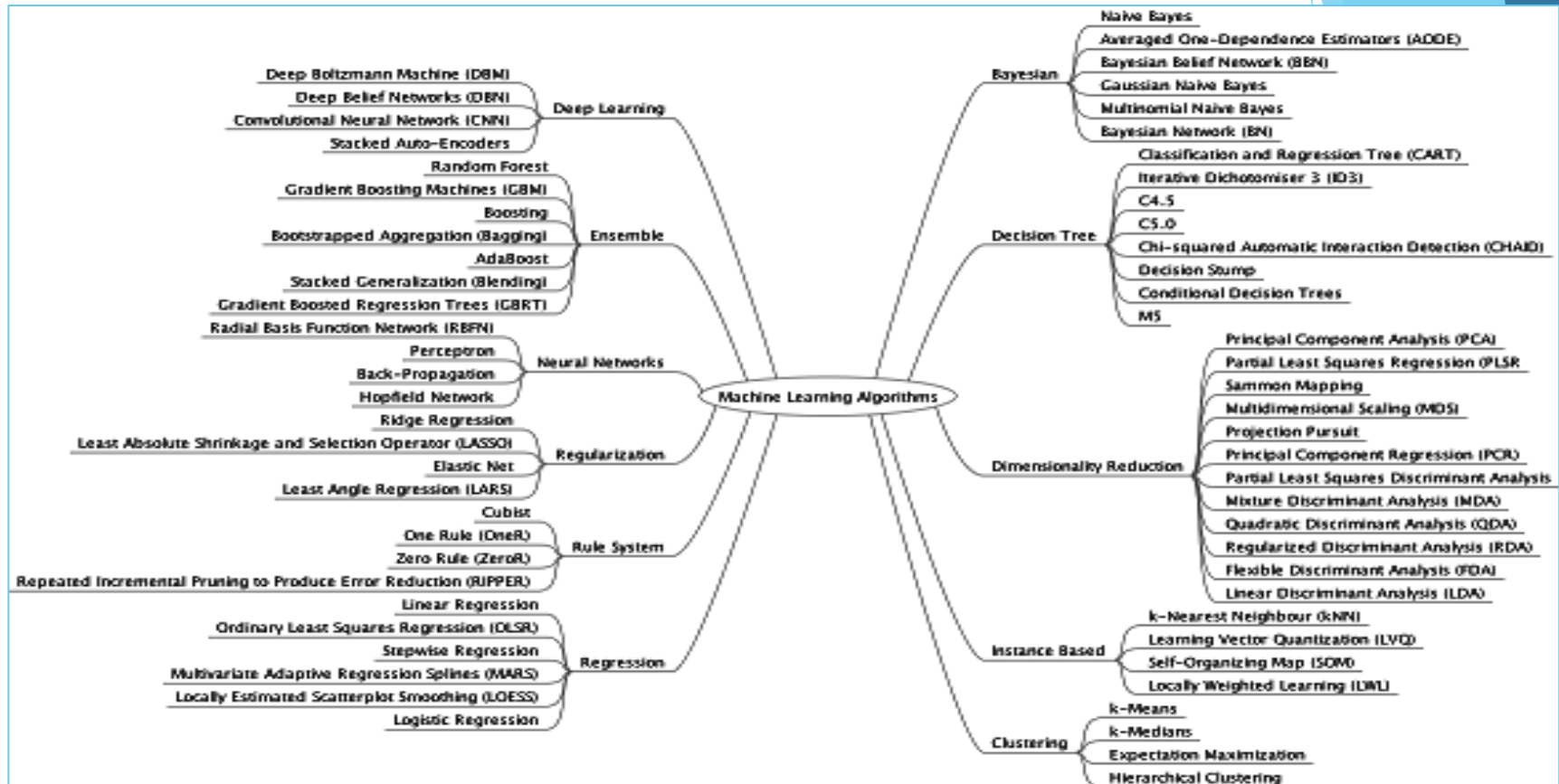
3

- Algorithms

4

- Embedded Analytics

ML Algorithms



“ Google’s self-driving cars and robots get a lot of press, but the company’s real future is in machine learning, the technology that enables computers to get smarter and more personal.

– Eric Schmidt (Google Chairman)

Which Algorithms Matter in Business?

The Four Functions of Management

Leading

Planning

Organizing

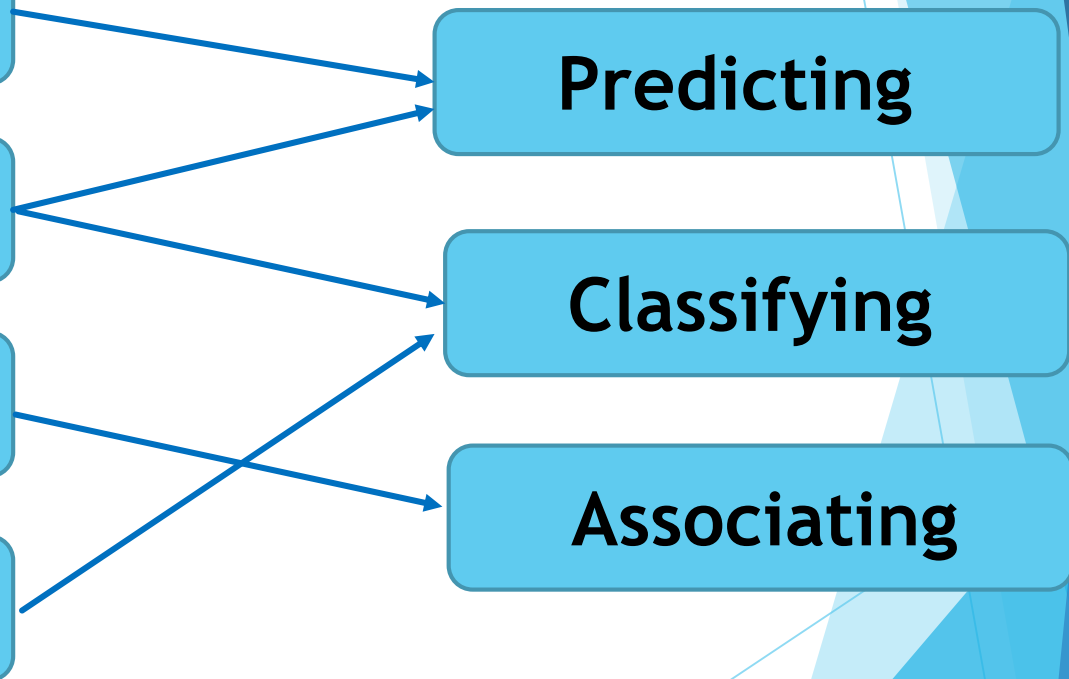
Controlling

Key Activities in Management

Predicting

Classifying

Associating



Types of ML Algorithms

Machine Learning

Supervised Learning

Build predictive models based on input and output data

Unsupervised Learning

Group data based on only input data

Regression

Focus

Classification

Clustering (Segmentation)

Association

Overview

Regression ML Algorithms

Regression

Regression Analysis is used to:

- Analyze the relationship among quantitative variables (and determine whether or not a relationship exists)
- Predict the value of one variable (the dependent variable, y) based on the value of another variable (independent variables x_i)

In Regression Analysis, you always predict one variable

Correlation V/s Regression

Basis for Comparison	Correlation	Regression
Purpose	Determines the relationship between 2 variables	Determines the relationship between 1 dependent variable & many independent variables
Usage	Represents linear relationship between two variables	Represents “ best fit line ” between the dependent variable & independent variables

Important Types of Regression Analysis

1. Simple Linear Regression (SLR)

- ▶ One independent variable X and One dependent variable Y

2. Multiple Linear Regression (MLR)

- ▶ Two or more independent variables
- ▶ One dependent variable Y

3. Regression Based on Dummy Variables

- ▶ two or more independent variables (*Continuous + Categorical*)
- ▶ One dependent variable Y (*continuous*)

4. Logistics Regression

- ▶ Two or more independent variables (*Continuous + Categorical*)
- ▶ One dependent variable Y (*Categorical*)

Key Assumptions in SLR and MLR

- ▶ Data of independent variables must be normally distributed.
- ▶ A linear relationship is assumed between the dependent variable and the independent variables.
- ▶ Absence of multicollinearity is assumed in the model, meaning that the independent variables are not highly correlated with each other (applicable only for MLR).
- ▶ The residuals are homoscedastic i.e. fall within the standard error (SE) limits.

Simple Linear Regression (SLR) Exercise

Monthly Sales in C\$ V/s StoreManager's Experience in Months

Monthly Sales in C\$	StoreManager's Experience in Months
25000	150
24500	160
24000	155
26500	180
23000	140
24900	135
22000	145
28500	170
22000	130
21500	125
21000	175
29500	120
30000	145
28000	170
23000	130
22000	145
28500	170
22000	130
21500	125
31000	175

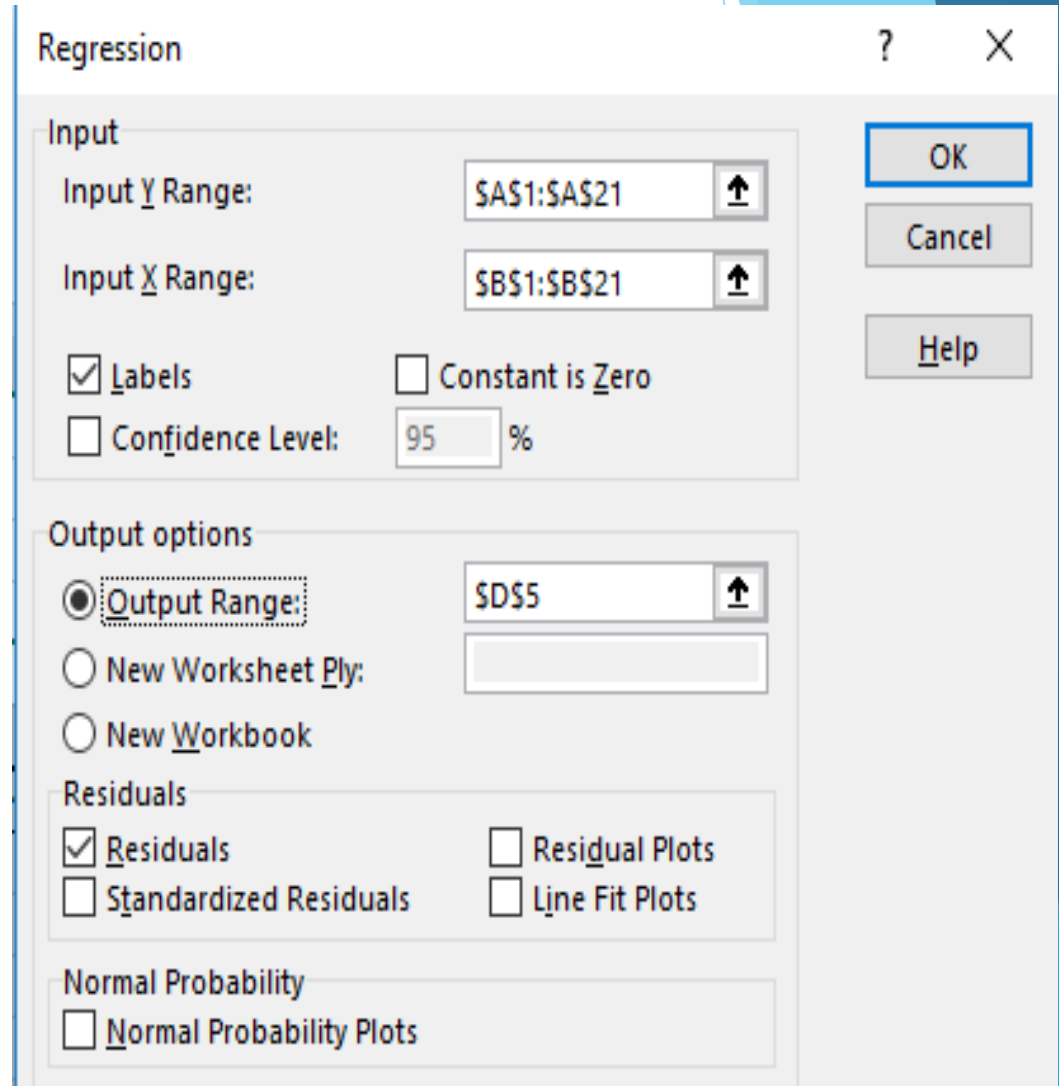
**Dependent
Variable i.e. Y**

**Independent
Variable i.e. X**

Simple Linear Regression With Excel

Data > Data Analysis > Regression

- *Input Y Range* (with header)
- *Input X Range* (with header)
- Leave CI at 95%
- Check *Labels*
- *Select Residuals*



The image shows the 'Regression' dialog box in Microsoft Excel. The 'Input' section has 'Input Y Range' set to '\$A\$1:\$A\$21' and 'Input X Range' set to '\$B\$1:\$B\$21'. Both ranges have selection icons to the right. The 'Labels' checkbox is checked, and 'Confidence Level' is set to '95 %'. The 'Constant is Zero' checkbox is unchecked. The 'Output options' section has 'Output Range' selected with a value of '\$D\$5'. 'New Worksheet Ply' and 'New Workbook' are unselected. The 'Residuals' section has 'Residuals' checked, while 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' are unchecked. The 'Normal Probability' section has 'Normal Probability Plots' unchecked. On the right side of the dialog are 'OK', 'Cancel', and 'Help' buttons.

SLR Output

$$\text{Sales (Y)} = 2119.22 + 150.93 * \text{Experience in Months (X)}$$

Multiple R	0.887474295					
R Square	0.787610625					
Adjusted R Square	0.775811215					
Standard Error	1566.846765					
Observations	20					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	163871841.9	163871841.9	66.75000224	1.81395E-07	
Residual	18	44190158.12	2455008.785			
Total	19	208062000				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2119.218071	2770.176063	0.765012051	0.454178364	-3700.705874	7939.142016
StoreManager's Experience	150.9296264	18.47348596	8.170067456	1.81395E-07	112.1182726	189.7409802

What do all these numbers mean?

5 Key Elements for SLR Evaluation

1. Coefficient of determination (R^2)
2. Standard Error
3. Co-efficients
4. Confidence Interval
5. Significance F (and p-value) of the Model

1. Coefficient of Determination - R^2

- The **coefficient of determination** denoted as **R^2** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- **R^2** gives the strength of one variable with regards to the other variable.
- If **R^2** is greater than 0.5 (here it is 0.787 or 78.7%), then there is a strong dependency between dependent and independent variables

2. Standard Error (SE)

- The standard error (of 1566.85) of the estimate is a measure of the precision of the predictions.
- SE represents the average distance that the observed values fall from the regression line. Conveniently, it tells you how “wrong” the regression model is on average using the units of the response variable
- Approximately 95% of the residuals (Data Points) should fall **within plus/minus 3*Standard Error** i.e. within 4700.6 of the independent variable i.e. Y (Homoscedasticity)
- In this example, look at the Residual Outputs and all 20 observations are within 4700.6. So this model is looks strong

3. Coefficients

	<i>Coefficients</i>
Intercept	2119.218071
StoreManager's Experience	150.9296264

For every additional **1 month of StoreManager's experience**, the store sales will increase by an average of **\$150.93**

4. Confidence Interval for Variable

Standard Error	1566.846765					
Observations	20					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	163871841.9	163871841.9	66.75000224	1.81395E-07	
Residual	18	44190158.12	2455008.785			
Total	19	208062000				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	2119.218071	2770.176063	0.765012051	0.454178364	-3700.705874	7939.142016
StoreManager's Experience	150.9296264	18.47348596	8.170067456	1.81395E-07	112.1182726	189.7409802

INTERPRETATION: We are 95% confident that the relationship between Store Sales and Experience is in a range such that One additional month of experience will have a Sales between \$112.12 and \$189.74.

5. Significance F (and p-value) of the Model

- In this case, the F Significance value is very small i.e. $1.81395127181234E-07$ (compared to 0.05) i.e. the F Significance value is statistically significant.
- So we **reject the null Hypothesis H_0** and accept the alternate hypothesis H_a .
- This means the Regression Model i.e. **$Y = 2119.22 + 150.93X$** is strong

Using the Model for Prediction

How much Store Sales (Y) would a Store manager with 190 months (X) generate?

$$\begin{aligned} Y &= 2119.22 + 150.93X \\ &= 2119.22 + 150.93 * 190 \\ &= \$ 30,795.92 \end{aligned}$$

How Confident are you on this figure? (R2)

I am 78.7% confident of this prediction.

SLR Comparison V1 v/s V2

	Iteration 1	Iteration 2
# of Records	20	22
R2	0.787	0.804
Standard Error	1566.9	1590.8
F-Value	1.814E-07	1.55E-08
Prediction Model	$Y = 2119.22 + 150.93X$	$Y = 739.85 + 160.77X$
Prediction Value (X = 190 months)	$Y = \$ 30,795.92$	$Y = \$ 31,286.15$

Multiple Linear Regression (MLR)

- MLR is used to explain the relationship between **one** continuous dependent variable (predicted variable) and **two or more** independent variables (predictor variables) .
- When “Store Sales”, is a function of Store Manager Experience, Areas of the Store, Community Population etc. you use MLR

Multiple Linear Regression (MLR)

Monthly Sales in C\$	StoreManager's Experience in Months	Store SFT	Community Population
25000	150	205	55000
24500	160	200	58000
24000	155	180	52000
26500	180	220	50000
23000	140	185	55000
24900	135	205	55000
22000	145	145	51000
28500	170	165	59500
20000	130	150	55000
25000	125	160	48000
29000	175	240	58000
26000	120	110	45000
23000	145	145	52000
28000	170	165	56000
23000	130	150	54000
22000	145	145	51000
28500	170	165	59500
22000	130	150	50000
21500	125	160	46000
31000	175	240	60500

**Dependent
Variable
i.e. Y**

**Independent
Variables i.e.
X1, X2, & X3**

Class_Store_Data_Descriptive_Predictive_IE.xlsx

6 Key Elements in The MLR Output

1. Multicollinearity
2. Adjusted R²
3. Standard Error
4. Co-efficients
5. Confidence Interval
6. Significance F (and p-value)

To Be
Discussed here

Already
Discussed in SLR

Multicollinearity in MLR

- ▶ Multicollinearity is a state of very **high inter-correlations** among the independent variables.
- ▶ Multicollinearity can result in several problems:
 - ▶ Multicollinearity results in a **change in the signs and magnitudes** of the regression coefficients.
 - ▶ Multicollinearity makes it **tedious to assess the relative importance of the independent variables**.

Avoid Multicollinearity as much as possible ($R^2 > 0.75$)

Multicollinearity (Correlation)

If there are any pairs of independent variables with high correlations (R2) i.e. above 0.75, then there is multicollinearity

	StoreManager's Experience in Months	Store SFT	Community Population
StoreManager's Experience in Months	1		
Store SFT	0.655885628	1	
Community Population	0.677225244	0.560537001	1

The Output of Excel here is R and NOT R2

In this example, there is NO multicollinearity

MLR Output

$$\text{Sales (Y)} = -5304.98 + 89.83 * \text{Experience in Months} + 20.18 * \text{StoreSFT} + 0.243 * \text{Population}$$

Regression Statistics						
Multiple R	0.941603467					
R Square	0.88661709					
Adjusted R Square	0.865357794					
Standard Error	1214.255819					
Observations	20					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	184471324.9	61490441.64	41.70491359	8.64673E-08	
Residual	16	23590675.09	1474417.193			
Total	19	208062000				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-5304.985256	3405.222564	-1.557896776	0.138815652	-12523.73461	1913.764103
StoreManager's Experience in Months	89.83825261	21.8305277	4.115257948	0.000810216	43.55960126	136.116904
Store SFT	20.17736964	10.95175067	1.842387601	0.08402828	-3.039304641	43.39404391
Community Population	0.242795675	0.086896615	2.794075197	0.012999244	0.058583082	0.427008269

What do all these numbers mean?

6 Key Elements in the MLR Output

1. Multicollinearity

- ▶ No Multicollinearity as the correlation between independent variables is less than 0.75

2. Adjusted R2

- ▶ Good Adjusted R2 at 0.86

3. Standard Error(Precision/Homoscedasticity)

- ▶ ~95% of the observations should fall within $\pm 3 \times \text{Standard Error}$ i.e. within $3 \times 1214.26 = 3642.8$ of the independent variable i.e. Y.
In this example, after look at the Residual Outputs and all 20 observations are within 3642.8. So this model is looks strong

6 Key Elements in the MLR Output

4. Co-efficients

	<i>Coefficients</i>
Intercept	-5304.985256
StoreManager's Experience in Months	89.83825261
Store SFT	20.17736964
Community Population	0.242795675

- For every additional 1 month of Store Manager experience, the store sales will increase by an average of \$89.83
- For every additional 1 SFT increase, the store sales will increase by an average of \$20.18
- For every additional 1 person increase in population, the store sales will increase by an average of \$0.24

Which Independent Variables are Important?

With 1 unit increase
or decrease in Store
Manager experience



The store sales will
increase or decrease
by \$89.83

With 1 unit increase
or decrease in the
SFT



The store sales will
increase or decrease
by \$20.18

With 1 unit increase
or decrease in the
population



The store sales will
increase or decrease
by \$0.24

Assuming the
**cost of
acquiring 1
unit of Store
Manager
experience,
Store SFT or 1
Person
population is
the same, the
maximum
return in on
Store Manager
experience**

Also look for the Sign of the co-efficients

6 Key Elements in the MLR Output

5. Confidence Interval

We are 95% confident that the relationship between Store Sales and:

- ▶ Store Manager Experience is in a range such that One additional month of experience will have a Sales between \$43.6 and \$136.1
- ▶ Store SFT is in a range such that One additional SFT will have a Sales between -\$3.1 and \$43.4
- ▶ Community Population is in a range such that One additional person will have a Sales between \$0.1 and \$0.43

6 Key Elements in the MLR Output

6. Significance F (and p-value)

- ▶ In this case, the F-value is very small i.e. $8.6467262003813E-08$ (compared to 0.05) i.e. the F Significance value is statistically significant.
- ▶ So we **reject the null Hypothesis H_0** and accept the alternate hypothesis H_a .
- ▶ This means the Regression Model is strong

Using the Model for Prediction

How much Store Sales (Y) would a Store manager with 190 months (X) generate, 225 SFT and 53,500 population?

$$\begin{aligned}\text{Sales (Y)} &= -5304.98 + 89.83 * \text{Experience in Months} + \\ &\quad 20.18 * \text{StoreSFT} + 0.243 * \text{Population} \\ &= -5304.98 + 89.83 * 190 + 20.18 * 225 + 0.243 * 53500 \\ &= -5,304.98 + 17,068 + 4,545 + 13,000 \\ &= \$29,308\end{aligned}$$

How Confident are you on this figure? (R2)

I am 86% confident of this prediction.

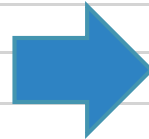
Multiple Regression Model with Dummy Variables

Dummy Variables in Regression

- Dummy Variables help to analyze regression equations when one or more independent variables are categorical.
- A dummy variable is a numeric variable that represents categorical data (reference data), such as gender, season, plants, currency, etc.
- Technically, dummy variables are dichotomous are limited to two specific values - 1 or 0.
 - 1 represents the presence of a qualitative attribute
 - 0 represents the absence of a qualitative attribute

HR Analytics

Employee	Performance Score	Age	Gender		Employee	Performance Score	Age	Gender
1	93	42	Male		1	92	42	1
2	86	40	Female		2	86	40	0
3	84	32	Male		3	84	32	1
4	77	30	Female		4	80	30	0
5	73	28	Male		5	73	28	1
6	74	27	Female		6	74	27	0
7	96	38	Male		7	96	38	1
8	81	37	Female		8	81	37	0
9	92	35	Male		9	92	35	1
10	75	33	Female		10	75	33	0
11	84	32	Male		11	84	32	1
12	77	30	Female		12	77	30	0
13	73	28	Male		13	73	28	1
14	76	27	Female		14	76	27	0
15	94	38	Male		15	94	38	1
16	81	37	Female		16	81	37	0
17	90	34	Male		17	90	34	1
18	75	33	Female		18	75	33	0
19	82	32	Male		19	82	32	1



MLR-Dummy Variable Output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.865936913					
R Square	0.749846737					
Adjusted R Square	0.718577579					
Standard Error	3.986391249					
Observations	19					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	762.1600096	381.080005	23.9803943	1.53338E-05	
Residual	16	254.261043	15.8913152			
Total	18	1016.421053				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	39.97862639	7.077590525	5.64862099	3.6283E-05	24.97480473	54.982448
Age	1.174123682	0.21280814	5.51728745	4.6845E-05	0.722990577	1.62525679
Gender	6.218580792	1.850329762	3.36079596	0.00397574	2.296056924	10.1411047

$$\text{Performance Score} = 39.98 + 1.174 * \text{Age} + 6.21 * \text{Gender}$$

Predicting the Employee Performance

$$\text{Performance Score} = 39.98 + 1.17 * \text{Age} + 6.21 * \text{Gender}$$

What is the Employee Performance Score for a
Female Employee who is 40 Years old?

$$\begin{aligned}\text{Performance Score} &= 39.98 + 1.17 * 40 + 6.21 * 0 \\ &= 86.78\end{aligned}$$

Introduction to Logistics Regression

- Logistic regression (Logit model) is a type of regression analysis when the dependent variable is typically a “dichotomous” (binary i.e. 1 / 0, Yes / No, True / False).
- Logistic regression is used to explain the relationship between one dependent binary variable and one or more nominal, ordinal, continuous level independent variables.
- In a Linear regression, we are predicting the **value**. But in Logistic regression, we are going to predict the **probability**
%

Logistic Regression Case Study

**Actionable Insights for Improved
Retailer Retention**

The Problem

At your convenience? Store owners working too hard

PUBLISHED THU, SEP 10 2015 - 8:38 AM EDT | UPDATED THU, SEP 10 2015 - 12:01 PM EDT



Alexandra Gibbs
@ALEXGIBBSY

SHARE f t in e ...

While working for yourself and the enjoyment of helping out people with their lives may hold a lot of appeal for independent retailers; rising property prices, the battle with online shopping and long, tough hours are taking their toll on convenience store owners, research shows.

Can't be
Wasabi's
cheaper
6x faster

According to a 2018 industry report, 153,237 convenience stores are operating in the U.S. These stores generated \$616.3 billion in sales for an average of nearly \$4 million per store.

Profit margins, however, are typically thin in the food industry, and convenience stores are no exception. Although in-store

The Impact - Store Performance Volatility (SPV)

1. Stores Closing
2. Ownership Changing
3. Structure Changing (DODO to CORO and Vice-versa)

SPV is ~5% per year from the 600+ PKI stores. This means there is a risk of about **C\$ 400 million/year** at stake for PKI

Solution - Support for the Retailer...But How?

Store Performance Data



Store Risk Profile Formulation



Store Categorization

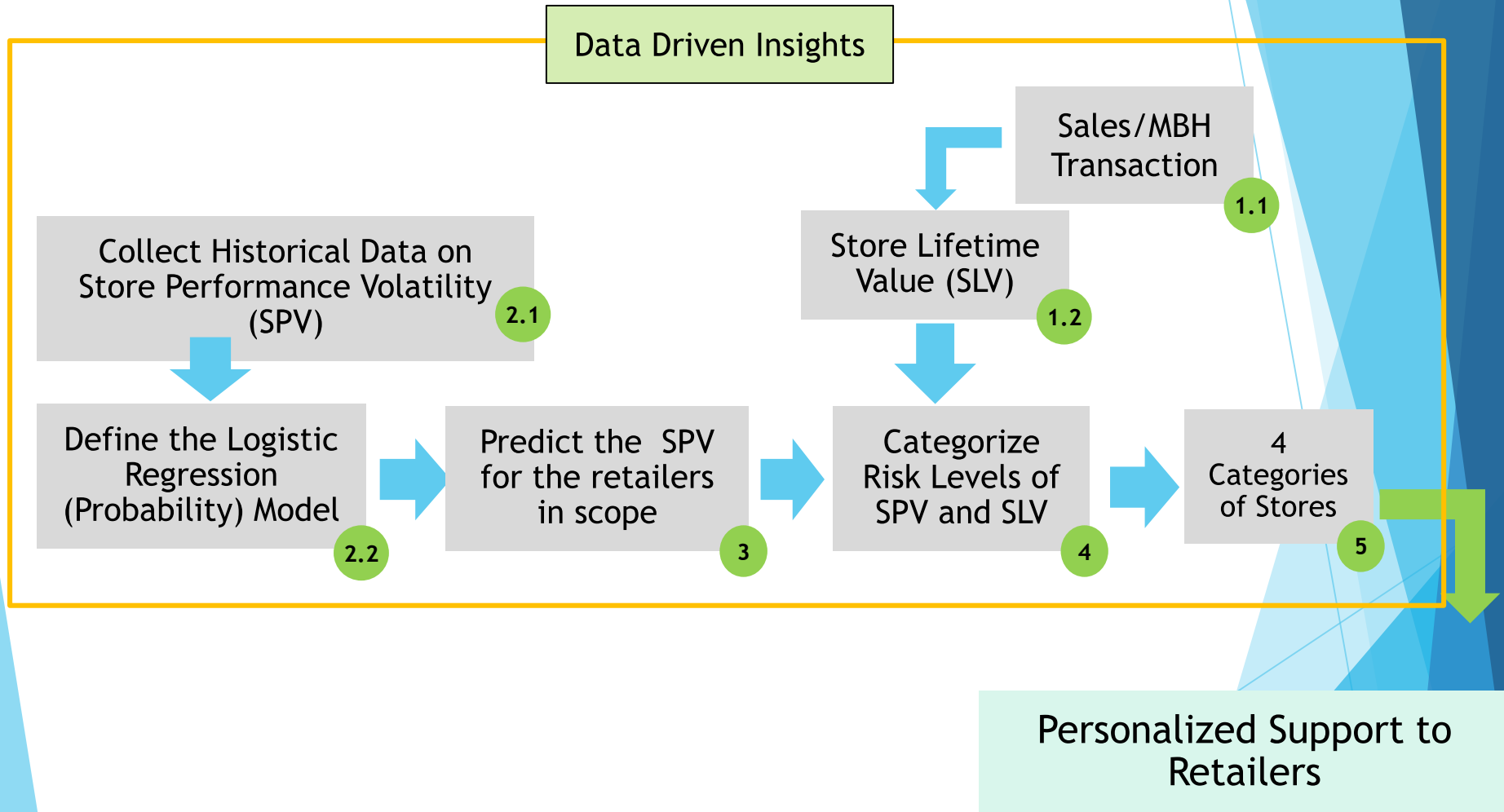


Personalized Support



Give the Market Dynamics and the size of our Retailer Network, how can we work on **optimally** use our **scarce resources** to support the **right** retailers?

The Insight Derivation Model



1. Sales/MBH Transaction + SLV

1.1 Sales/MBH Transaction

An analysis of 300 CORO stores in April of 2019, showed that the

- Average sales/day is C\$ 45,964 (Median is C\$ 39,019) (Fuel + Store Sales)
- Average # of transactions/day is 2,662

Transaction #	Product_Key	Quantity_Sold	Total_Price	Unit_Price	Store_ID	Item_Desc
205002995	1280	1	2.99	2.99	43105	COKE ZERO 1L
205002995	25038	1	1.89	1.89	43105	BG GO MUFFIN BANANA CHOC
205003006	1413	1	3.79	3.79	43105	MONSTER 473ML
205003006	4864	1	0.05	0.05	43105	Deposit .05
205003019	18618	1	70	70	43105	Rounding Variance
205003028	34378	1	0	0	43105	CIBC DCCA COUPON
205003036	48	10.644	20	1.879	43105	Sup Plus
205003038	3319	1	14.69	14.69	43105	BELMONT BLUE KS 10X20S
205002999	2544	1	11.99	11.99	43105	NEXT BLUE KS 10X20S
205002999	2544	1	11.99	11.99	43105	NEXT BLUE KS 10X20S
205002999	20789	2	7.98	3.99	43105	CORE POWER CHOCOLATE PET 340ML
205003001	8135	1	2.99	2.99	43105	POWERADE MIXED BERRY 710ML

Pivot Table

1.2 Store Lifetime Value (SLV)

- SLV = Average sales/day is C\$ 45,964 * Working Days/Year
- SLV = C\$ 45,964 * 350 days = C\$ 16 Million/Year

Store ID	Count of Transaction #	Sum of Total_Price
40001	3212	41448.12
40002	2927	38298.65
40003	1697	36708.54
40004	2640	33027.58
40005	1595	34193.64
40006	1983	29587.05
40007	1596	26472.4
40008	4018	52715.85
40009	2517	26785.45

2. Store Performance Volatility (SPV)

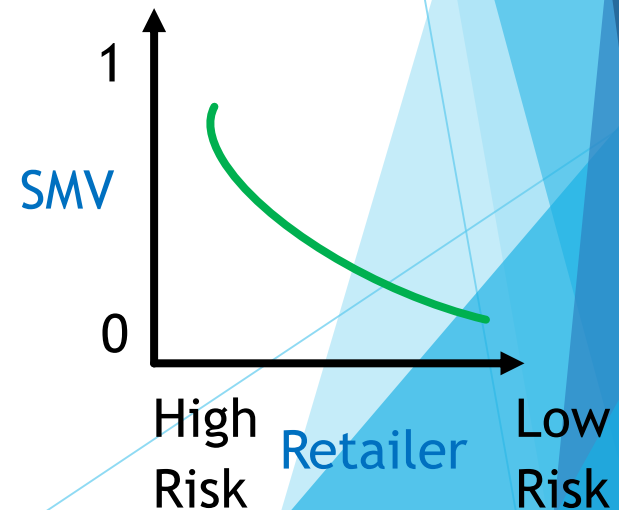
- Store Performance Volatility (SPV) is a Probability Value (Between 0 and 1)
- SMV is a Logistics Regression (Logit) Function of 3 independent variables
 - Average of IAuditor and Mystery (AIM) Shopper Score during the Change
 - Gas Buddy/Google Rating (GBR) during the Change
 - Provincial Economy Status (PES) during the Change
- Using data from 300+ Stores including data on AIM, GBR and PES we derived the Logit Model

$$SPV = -7.278 + 0.045*AIM + 1.253*GBR - 0.378*PES$$

The **hypothesis** is high SPV probability value is correlated with low AIM, GBS and PES scores. In other words, if there is high AIM, GBS and PES scores, then the SMV value will be low. And if I have low AIM, GBS and PES scores, the SMV value will be high.

3. Predict the SPV for the Retailers in scope

- Using data from 400 Stores including data on AIM, GBR and PES, the below Logit Model is applied.
 - $SPV = -7.278 + 0.045*AIM + 1.253*GBR - 0.378*PES$
- Low Risk (0 to 0.4) = 128 Stores
- Medium Risk (0.41 to 0.7) = 246 Stores
- High Risk (0.71 to 1.0) = 26 Stores



4. Categorize Risk Levels of SMV and SLV

SPV Risk Levels

- Low Risk = 0 to 0.4
- Medium Risk = 0.41 to 0.7
- High Risk = 0.71 to 1.0

SLV Risk Levels

- Low Value = 0 to \$20K
- Medium Value = \$20.1K to \$40K
- High Risk Value = \$40.1K Onwards

		SLV		
		Low	Medium	High
SMV	Low	Low	Low	Medium
	Medium	Low	Medium	High
	High	Medium	High	Very High



4 Retailor Categories

5.1 Integrate SPV and SLV

1	Store ID	IAuditor & Mystry Shopper Score	Gas Buddy Ranking	Provincial Economy Score	Predicted SPV Score	\$ Value/Day (SLV)
2	40001	86.0	4.1	3	0.5953	20724.06
3	40002	84.0	3.7	3	0.0041	19149.325
4	40003	100.0	3.8	3	0.8494	18354.27
5	40004	92.0	4.2	3	0.9906	16513.79
6	40005	89.0	3.9	4	0.1017	17096.82
7	40006	96.0	3.0	2	0.045	14793.525
8	40007	78.0	4.1	3	0.2353	13236.2
9	40008	86.0	4.1	3	0.5953	26357.925
10	40010	87.0	4.2	4	0.3876	18392.725
11	40011	87.5	3.9	2	0.7902	26772.945
12	40012	96.0	4.0	4	0.542	16845.205
13	40013	86.0	4.1	3	0.5953	14522.59
14	40014	95.0	3.4	3	0.1232	30105.145
15	40015	86.0	4.1	3	0.5953	29539.785
16	40016	87.5	3.8	3	0.2869	18018.805
17	40017	76.0	4.5	3	0.6465	19865.775
18	40018	97.5	3.9	4	0.4842	10153.635
19	40019	86.0	4.1	3	0.5953	13994.195
20	40020	98.0	3.5	3	0.3835	21985.6
21	40021	67.5	3.8	1	0.1429	25278.085
22	40024	82.0	4.2	3	0.5406	23591.78
23	40025	82.5	3.6	2	0.1893	25090.235
24	40026	86.0	4.1	3	0.5953	18035.835

5.2 Assign Stores into 4 Categories Based on SPV & SLV

		SLV		
		Low 0 to \$20K	Medium \$20.1K to \$40K	High \$40.1K Onwards
SPV	Low 0 to 0.4	68 Stores	53 Stores	7 Stores
	Medium 0.41 to 0.7	115 Stores	99 Stores	32 Stores
	High 0.71 to 1.0	15 Stores	9 Stores	2 Stores

	Cat 1		Cat 2		Cat 3		Cat 4
--	-------	--	-------	--	-------	--	-------

The Insights To Action

Personalized Support to Retailers

Store 1 (43005)

- Gas Buddy Rating (GBR) lower by 0.1 Points WoW (Week Over Week)
- Increase in the sale of baked goods by \$700 WoW
- 14 Weeks of low AIM scores on Store Cleanliness.
- 2.8 cms of rain last week in the Store Area

Store 2 (43123)

- Tobacco sales down by 3.7% WoW
- Carwash - Premium sales up by 7% WoW
- Carwash - Regular sales up by 3.7% WoW

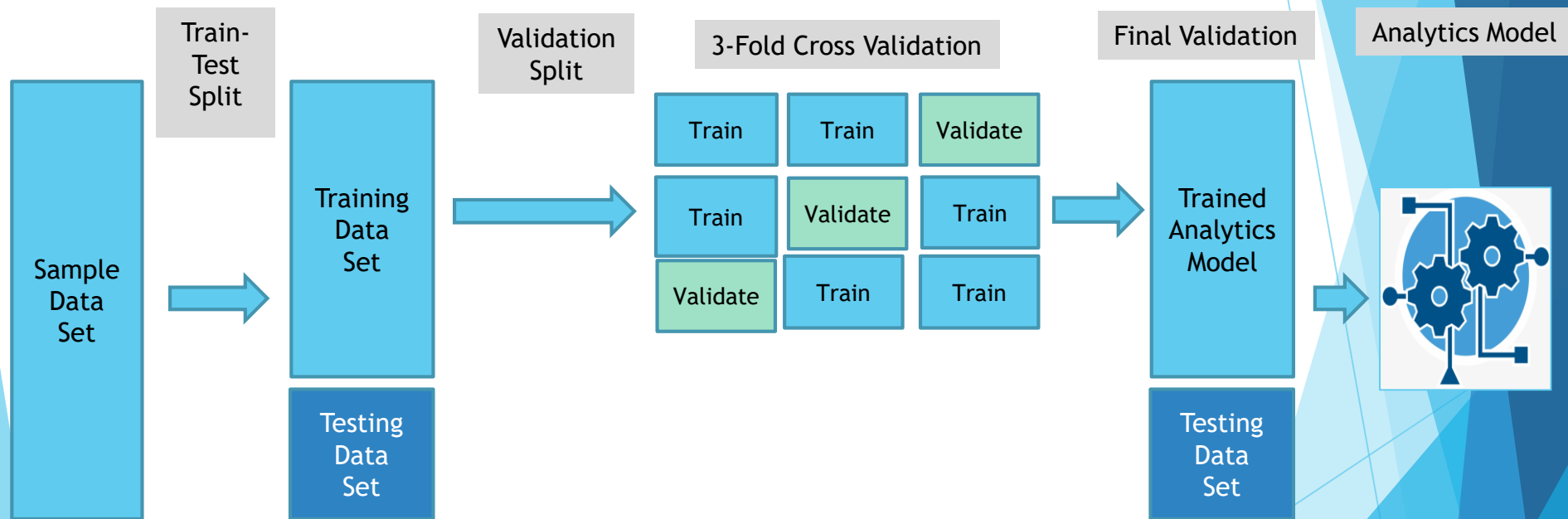
**Talking Points
with Retailers**

**(Input for
Embedded
Analytics)**

Validating the Predictive Analytics Model

When there is a prediction model, apart from R-square, residuals etc. what are the different ways of validating the prediction model? Basically, what are the different options to validate the Analytics prediction model given that the real/actual values will be available only in future?

Option 1: Data Splitting & Cross Validation



Option 2: Confirm O/P from Multiple Algorithms

Evaluate the output with multiple prediction model algorithms. Some examples are:

- Moving Average
- Logistic regression
- Time series models
- Classification and regression trees (CART)
- Neural networks
- Support vector machines (SVM)
- K-nearest neighbours
- And many more...

Break

Guest Presenter

Matt Joyce

Pre-Sales Lead

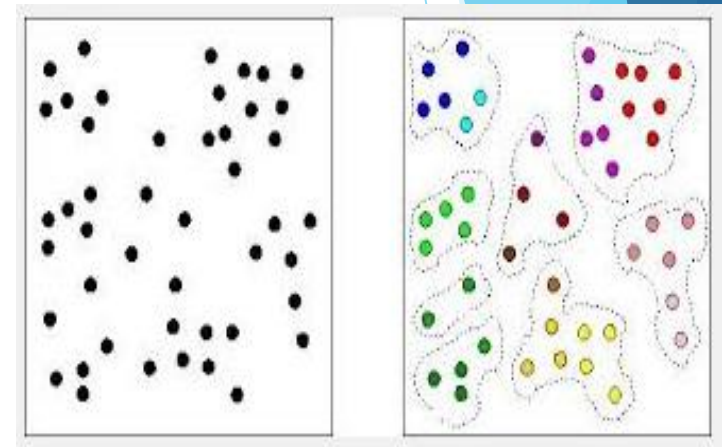
SAS-Canada



Classification ML Algorithms

Classification Algorithm

- The Classification algorithm takes the input data to classify the observation into appropriate class/group.
- This data set may be used for
 - **Bi-class** (like identifying whether the person is male or female or that the mail is spam or non-spam)
 - **Multi-class** too (like document classification or bio metric)



NOTE: Classification algorithms are used for assignment to classes; not to create classes. To create classes you use clustering.

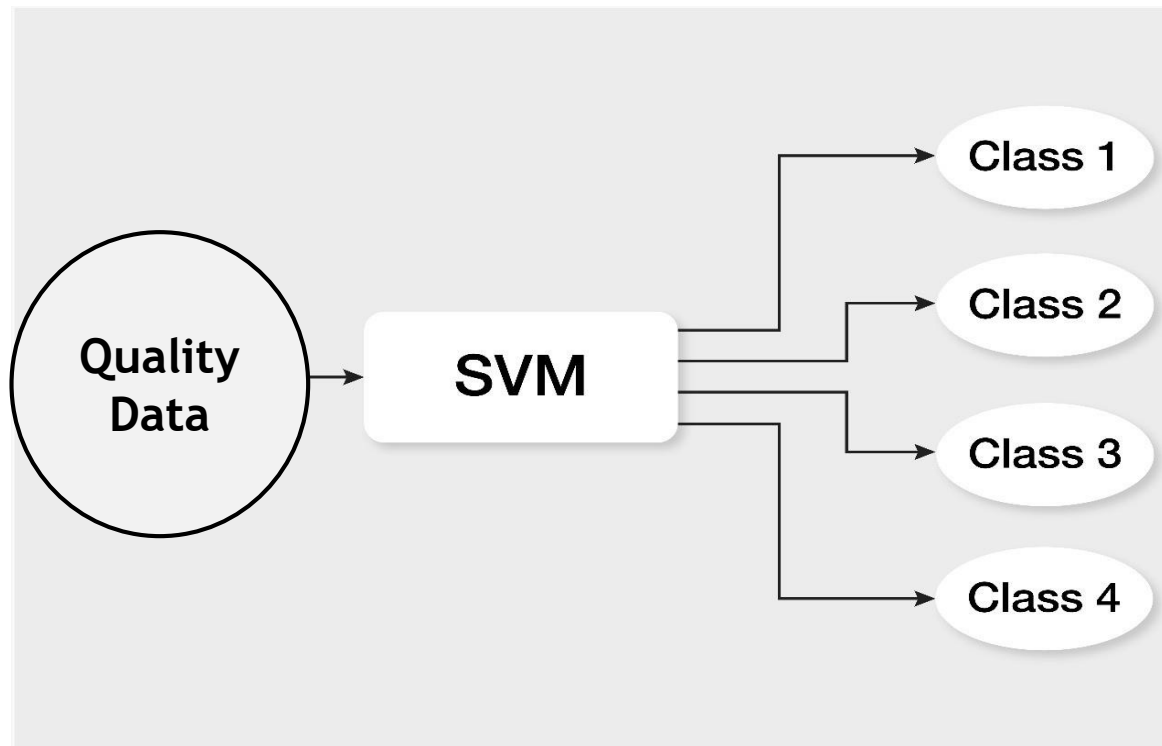
Types of Classification Algorithms

1. **Logistic Regression** - Two way classification/outcomes
2. **Support Vector Machines (SVM)** - Multiple classification/outcomes
3. **Decision Trees** - Break down the data set into smaller and smaller subsets.
4. **Nearest Neighbor (NN)** - Finding the point in a given set that is closest (or most similar) to a given point.



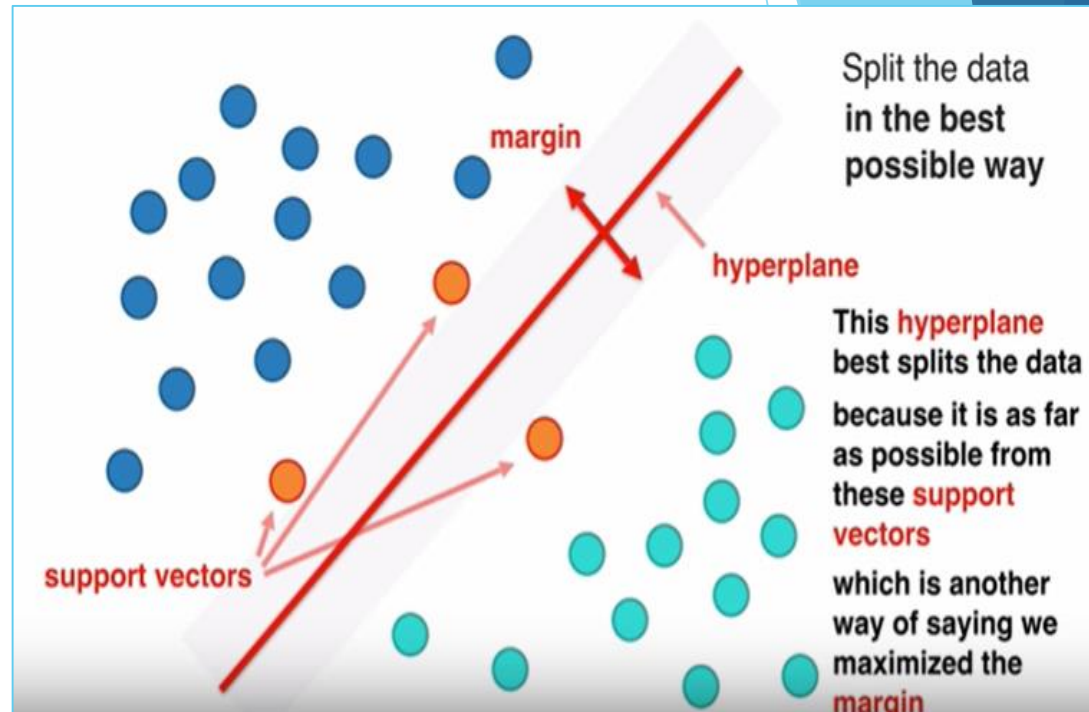
Our
Focus

Support Vector Machine (SVM)



Support Vector Machine (SVM)

- SVM a supervised classification algorithm
- It is based on determining the widest road (Margin) that separates the two groups.



SVM With a Real Example

You have 2 classes of Vendors - Good and Bad Vendors. They are classed based on delivery time and quality of goods supplied. I have a new vendor Precision Inc who has a delivery time score of 4.4/10 and quality of goods supplied score of 5.7/10.

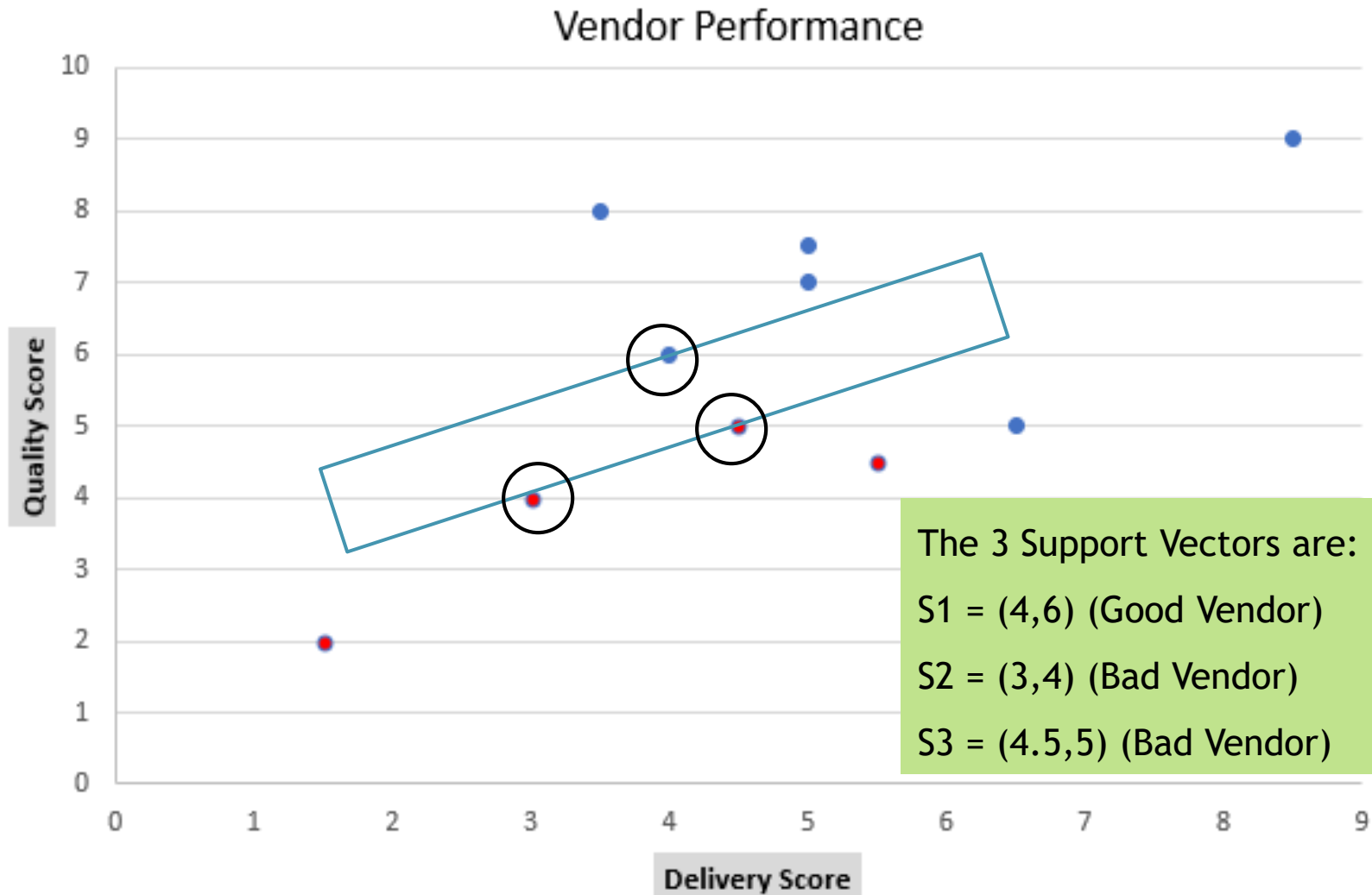
Is Precision Inc classified as a good vendor or bad vendor?

Step 1: Get the Data

Vendor ID	Delivery Score	Quality Score	Vendor Type
1	4	6	Good Vendor
2	1.5	2	Bad Vendor
3	3	4	Bad Vendor
4	5	7	Good Vendor
5	8.5	9	Good Vendor
6	4.5	5	Bad Vendor
7	5	7.5	Good Vendor
8	3.5	8	Good Vendor
9	6.5	5	Good Vendor
10	5.5	4.5	Bad Vendor

New vendor i.e. 11th vendor **Precision Inc** has **delivery time** score of 4.4/10 and **quality of goods** supplied score of 5.7/10.

Step 2: Visualize the Data & find 3 Support Vectors



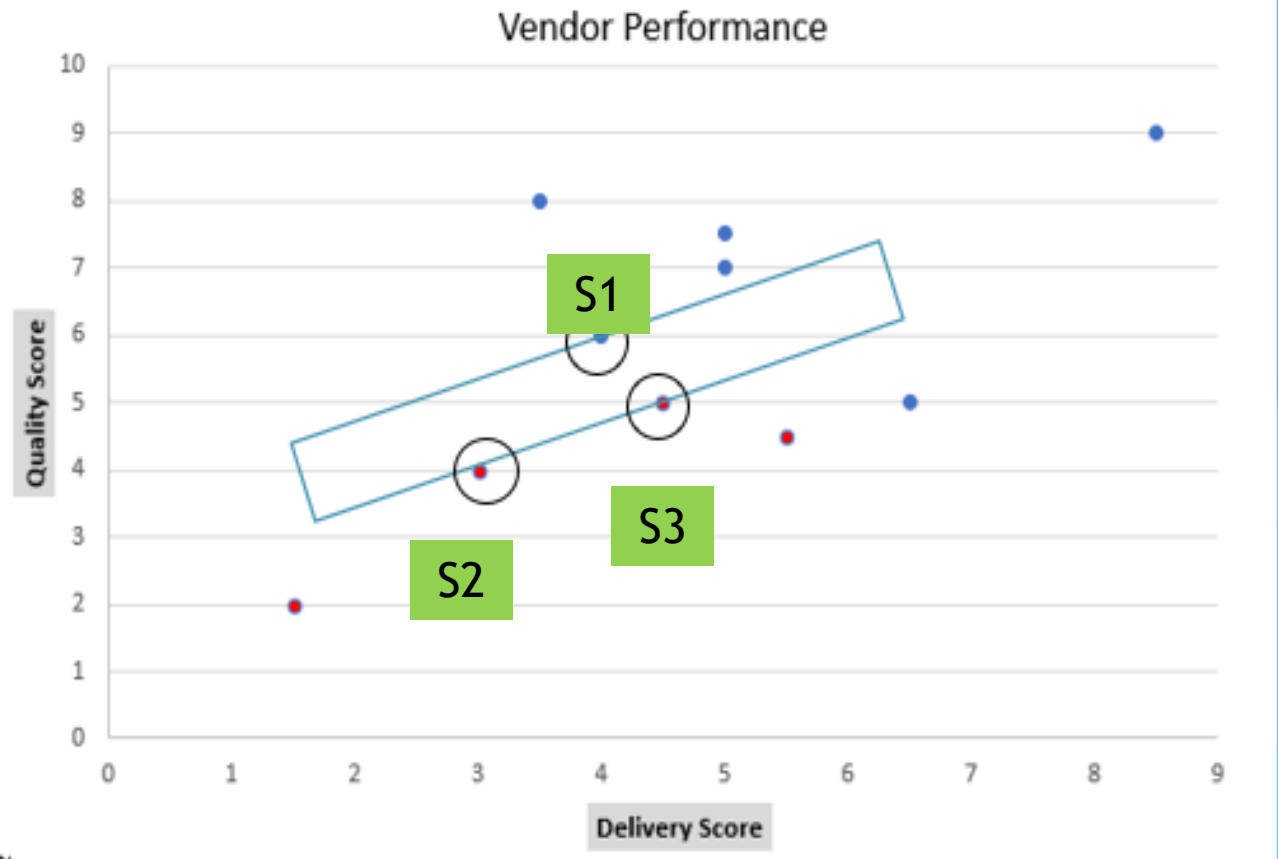
Step 3: Add bias input “1” to the 3 Support Vectors

The 3 Support Vectors now become:

$$S1 = (4, 6, 1)$$

$$S2 = (3, 4, 1)$$

$$S3 = (4.5, 5, 1)$$



Step 4: Formulate 3 Linear Equations

There are 2 bad vendors (represented by -1) and 1 good vendor (represented by +1) in the support vectors

$$a1.S1.S1 + a2.S2.S1 + a3.S3.S1 = +1 \text{ (for the 1st support vector i.e. good vendor)}$$

$$a1.S1.S2 + a2.S2.S2 + a3.S3.S2 = -1 \text{ (for the 2nd support vector i.e. bad vendor)}$$

$$a1.S1.S3 + a2.S2.S3 + a3.S3.S3 = -1 \text{ (for the 3rd support vector i.e. bad vendor)}$$

The 3 equations now are:

$$a1(4,6,1) (4,6,1) + a2(3,4,1)(4,6,1) + a3(4.5,5,1) (4,6,1) = +1$$

$$a1(4,6,1)(3,4,1) + a2 (3,4,1). (3,4,1) + a3(4.5,5,1) (3,4,1) = -1$$

$$a1(4,6,1)(4.5,5,1) + a2(3,4,1). (4.5,5,1) + a3 (4.5,5,1) (4.5,5,1) = -1$$

Step 5: Solve them

$$53a1 + 37a2 + 49a3 = +1$$

$$37a1 + 26a2 + 34.5a3 = -1$$

$$49a1 + 34.5a2 + 46.25a3 = -1$$

$a1 = 6.1, a2 = 14.8 \text{ \& } a3 = -17.8$
(the Hyperplane coefficients)

Step 6: Derive the Hyperplane

The hyperplane (classification pane) discriminates the positives class from the negative class.

The formula for the hyperplane “W” is:

$$W = \sum a_i S_i$$

$$W = (6.1)(4,6,1) + (14.8)(3,4,1) + (-17.8)(4.5,5,1)$$

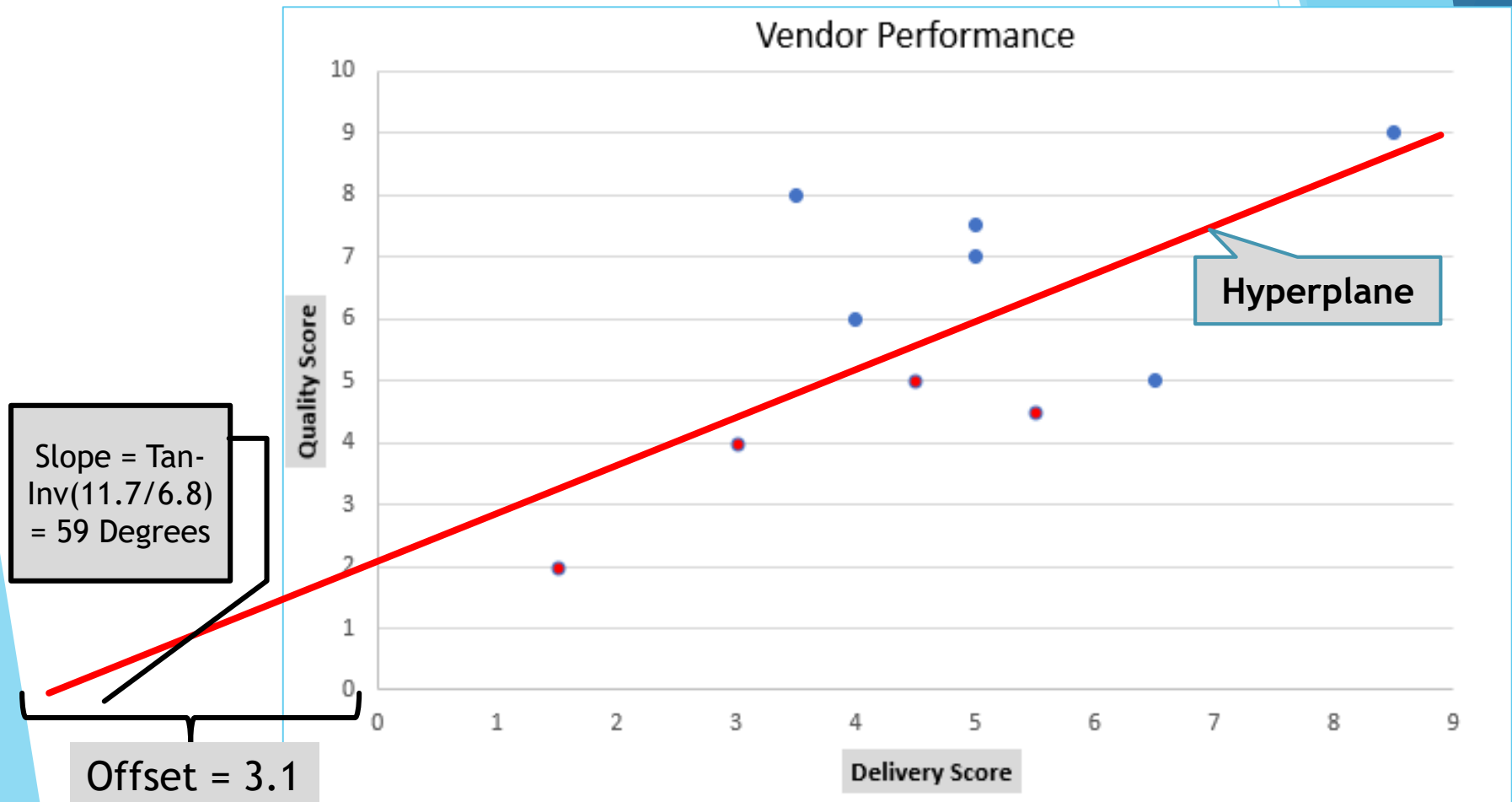
$$W = (-11.7, 6.8, 3.1)$$

The Equation of the “Biased” Hyperplane = (-11.7, 6.8, 3.1)

In terms of the linear equation $Y = Wx + B$, $W = (-11.7, 6.8)$ and offset $B = 3.1$

Step 7: Plot the Hyperplane Equation

In terms of the linear equation $Y = Wx + B$, $W = (-11.7, 6.8)$ and offset $B = 3.1$



Step 8: Classify the Vendor

Now we have a new vendor **Precision Inc** who has **delivery time** score of 4.4/10 and **quality of goods** supplied score of 5.7/10. Is this vendor classified as a Bad or Good Vendor?

Here the Point “V” = (4.4,5.7)

$W.V = (-11.7, 6.8) \cdot (4.4, 5.7) = -51.5 + 38.8 = -12.7 < -1.$

This means this vendor **Precision Inc** is a Bad Vendor.

Clustering ML Algorithms

ML Algorithms #3: Clustering

- Clustering is a method where we assign a set of observations into subsets known as clusters based on some similar conditions.
- Common clustering method are:
 1. **K-Means** used to cluster data in k-clusters (clusters are fixed beforehand) with **2 or more variables**
 2. **Sturges's rule** to classify observations based on $(1 + 3.3 \log n)$ clusters (where n is the number of observations) **with 1 variable**

Classification V/s Clustering

- In **classification**, you have a set of predefined classes/categories and want to know which to class the given (data) object belongs to.
- For example, in which of the 18 customer categories should I classify my new customer

- **Clustering** tries to group a set of (data) objects and find whether there is *some* relationship or categorization or grouping between the objects.
- For example, based on the 20000 invoices, is there a way to find some “homogeneous” or similar “K” groups
- The data elements are segmented based on the # of groups needed.

Clustering in Trump's Campaign

The Cambridge Analytica Case

87 Million American Facebook Users



5000 data points on American Voter

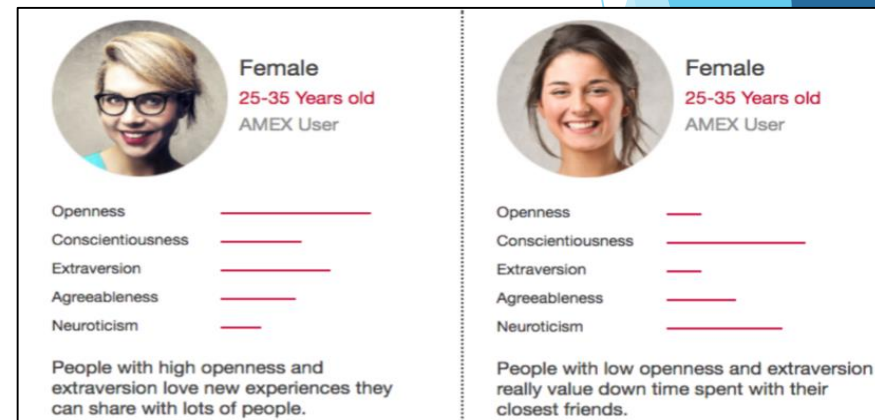


Psychrometric Clustering/Profiling



Targeted digital content

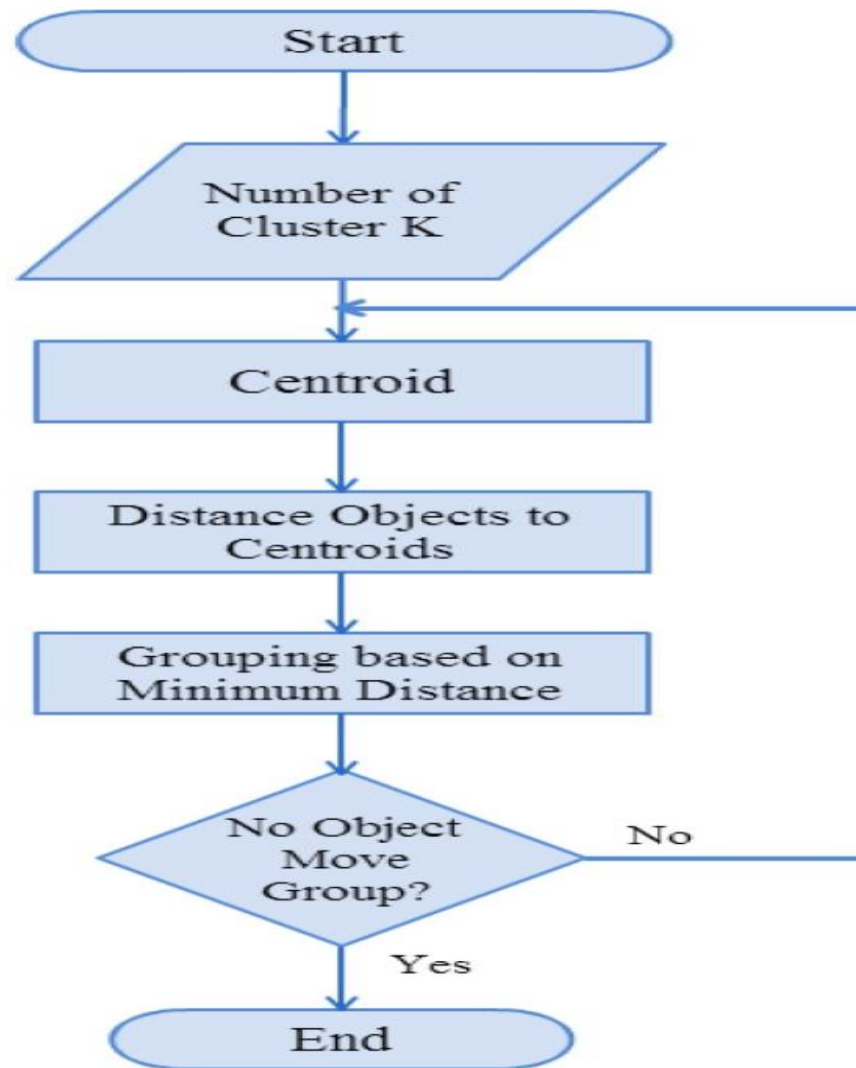
Trump, Views on President	Citizens United Supreme Court Decision
Border Wall With Mexico	Federal Educational Common Core Requirements
Momentum Voter	School Funding
Ban on Muslims Entering U.S.	Social Security Tax
Rideshare User	Union Support
Right To Work Laws	Military Intervention
Black Lives Matter Views	School Choice
Transgender Bathroom Use	Marijuana Legalization
Pathway to Citizenship for Undocumented Immigrants	Gay Marriage Support
Minimum Wage	Climate Change
Gun Laws	Fracking
International Humanitarian Intervention	Government Bailouts
Environment	Abortion



K-Means

(When the # of Clusters is determined)

K-Means Flowchart



Clustering Continuous Data using K-Means

Vendor ID	Delivery Score	Quality Score
1	4.0	6.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	8.5	9.0
6	4.5	5.0
7	5.0	7.5
8	3.5	8.0
9	6.5	5.0
10	5.5	4.5

The requirement is to group this multi-dimensional or vector data set (i.e. vendors) into **two clusters** i.e.

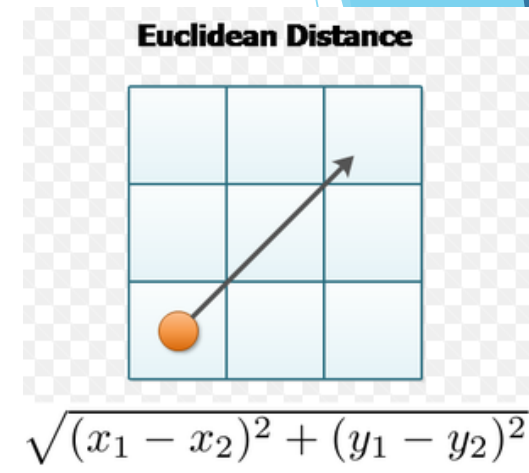
K=2

1. Bad Vendors
2. Good Vendors

1. Bad Vendors = Low Scores
2. Good Vendors = High Scores

Step 1: Calculate the Euclidean distance from Origin(0,0)

Vendor ID	Delivery Score	Quality Score	Euclidean Distance
1	4	6	7.21
2	1.5	2	2.50
3	3	4	5.00
4	5	7	8.60
5	8.5	9	12.38
6	4.5	5	6.73
7	5	7.5	9.01
8	3.5	8	8.73
9	6.5	5	8.20
10	5.5	4.5	10.98



Step 2: Find the two records that are furthest apart

In this case it is Vendor Id 2 and Vendor Id 5. This is the “centroid” of the 2 clusters

Step 3: Link Other Records to the 2 Centroids

Vendor ID	Delivery Score	Quality Score	Euclidean Distance	Distance to 2	Distance to 5	Closeness
1	6.5	7	9.55	7.05	2.83	Vendor ID 5
2	1.5	2	2.50			Vendor ID 2
3	3	4	5.00	2.50	7.38	Vendor ID 2
4	5	7	8.60	6.10	3.78	Vendor ID 5
5	8.5	9	12.38			Vendor ID 5
6	4.5	5	6.73	4.23	5.65	Vendor ID 2
7	5	7.5	9.01	6.51	3.37	Vendor ID 5
8	3.5	8	8.73	6.23	3.65	Vendor ID 5
9	6.5	5	8.20	5.70	4.18	Vendor ID 5
10	5.5	4.5	7.11	4.61	5.27	Vendor ID 2

Step 4: Have a Clean Cluster

Cluster	Vendors
Cluster 1	<ul style="list-style-type: none"> • Vendor ID 2 • Vendor ID 3 • Vendor ID 6 • Vendor ID 10
Cluster 2	<ul style="list-style-type: none"> • Vendor ID 1 • Vendor ID 4 • Vendor ID 5 • Vendor ID 7 • Vendor ID 8 • Vendor ID 9

Step 5: Start Iteration # 2 - Find the Centroid of the clusters

	Vendor ID	Delivery Score	Quality Score	Mean Vector Centroid of Cluster	Mean Vector Centroid of Cluster
Cluster 1	2	1.5	2	3.6	3.9
	3	3	4	3.6	3.9
	6	4.5	5	3.6	3.9
	10	5.5	4.5	3.6	3.9
Cluster 2	1	6.5	7	5.8	7.3
	4	5	7	5.8	7.3
	5	8.5	9	5.8	7.3
	7	5	7.5	5.8	7.3
	8	3.5	8	5.8	7.3
	9	6.5	5	5.8	7.3

Step 5: Start Iteration # 2

Compare each individual's distance to its own cluster mean and to that of the opposite cluster.

	Vendor ID	Delivery Score	Quality Score	Centroid of Cluster 1	Centroid of Cluster 2	Distance of each individual to Mean of cluster 1	Distance of each individual to Mean of cluster 2
Cluster 1	2	1.5	2	3.6	3.9	2.83	6.82
	3	3	4	3.6	3.9	0.61	4.33
	6	4.5	5	3.6	3.9	1.42	2.64
	10	5.5	4.5	3.6	3.9	1.99	2.82
Cluster 2	1	6.5	7	5.8	7.3	4.24	0.76
	4	5	7	5.8	7.3	3.40	0.85
	5	8.5	9	5.8	7.3	7.07	3.19
	7	5	7.5	5.8	7.3	3.86	0.82
	8	3.5	8	5.8	7.3	4.10	2.40
	9	6.5	5	5.8	7.3	3.10	2.40

Step 6: Validate the Clustering

Check if any records are nearer to the mean of the opposite cluster than its own

	Vendor ID	Distance of each individual to Mean of cluster 1	Distance of each individual to Mean of cluster 2
Cluster 1	2	2.83	6.82
	3	0.61	4.33
	6	1.42	2.64
	10	1.99	2.82
Cluster 2	1	4.24	0.76
	4	3.40	0.85
	5	7.07	3.19
	7	3.86	0.82
	8	4.10	2.40
	9	3.10	2.40

In this example there are NO data records that are nearer to the mean of the opposite cluster than its own.

NOTE: For example if vendor ID 3 had 5.61 instead of 0.61, then vendor ID 3 would have moved to cluster 2 from cluster 1

Step 7: Sort the final Clusters

The iteration will stop as there are no more relocations needed.

Here is the final 2 cluster and their associations

Cluster	Vendors
Cluster 1 (Bad Vendors)	<ul style="list-style-type: none">• Vendor ID 2• Vendor ID 3• Vendor ID 6• Vendor ID 10
Cluster 2 (Good Vendors)	<ul style="list-style-type: none">• Vendor ID 1• Vendor ID 4• Vendor ID 5• Vendor ID 7• Vendor ID 8• Vendor ID 9

Sturges's Rule

(When the # of Clusters is **NOT**
determined)

Clustering Continuous Data

Let's say, Vodafone Telecom gets billing amount data of 200 telecom customers in a specific geographic area. How can his data be categorized?

Step 1: Find the Range in the data set

Range = Max Value - Min Value = \$129.63 - \$10 = 119.63

Step 2: Apply Sturges's rule to determine the number of classes

of Classes = $1 + 3.3 (\log n)$; where n is the number of observations

of Classes = $1 + 3.3 (\log 200) = 1 + 3.3 \times 2.3 = 8.5 = 8$ groups or classes (You can also select 9 if you prefer)

Clustering Continuous Data

Step 3: Determine the Class Width

Class Width = Range/Number of Class = $119.63/8 = 14.95 = 15$
(rounded)

This means there will be 8 groups/classes which are separated by \$15.

1. Class 1 = \$10 to \$25 billing
2. Class 2 = \$26 to \$40 billing
3. Class 3 = \$41 to \$55 billing
4. Class 4 = \$56 to \$70 billing
5. Class 5 = \$71 to \$85 billing
6. Class 6 = \$86 to \$100 billing
7. Class 7 = \$101 to \$115 billing
8. Class 8 = \$116 to \$130 billing

Some Thoughts on Clustering

- Classification/clustering holds the key to good management.
- While you might be able to capture large amounts of time-series/continuous data, categorizing data is a fundamental building blocks for deriving insights and pursue appropriate actions.

Reference - <https://www.linkedin.com/pulse/transforming-interval-data-nominal-prashanth-h-southekeal-phd/>

Association ML Algorithms

ML Algorithms #4: Association

- Association ML algorithms uncover how items are associated to each other.
- There are 2 common algorithms to measure association
 1. Correlation
 2. Pure Association

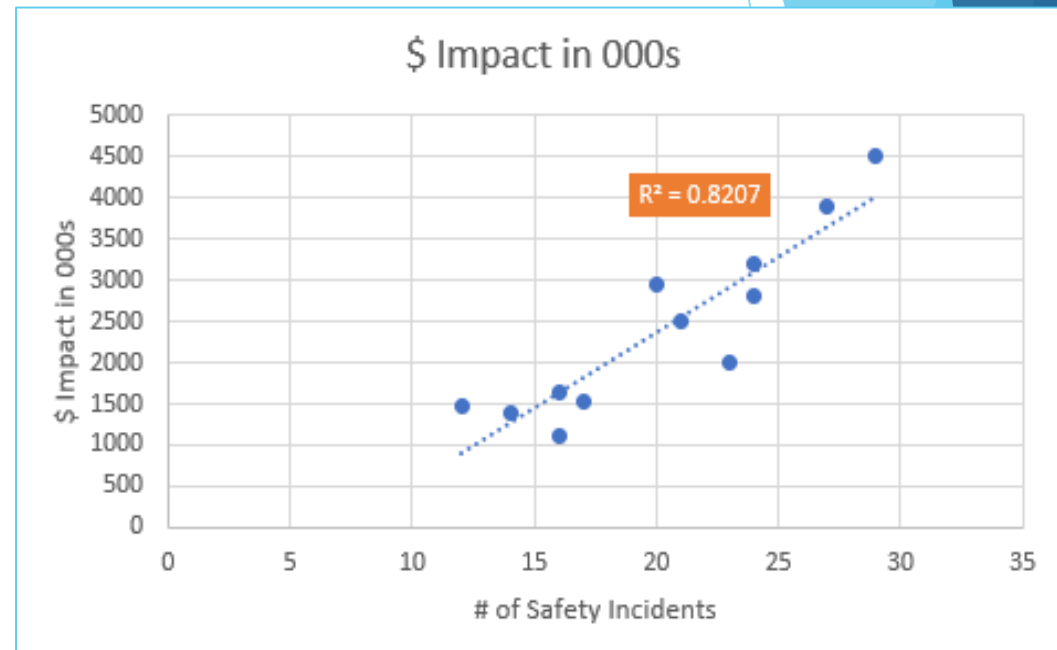
Association V/s Correlation

- Correlation is defined as the strength of the linear relationship of two variables.
- It measures whether, if we increase or decrease one variable by a certain factor, the other variable will also increase or decrease by that same factor or a factor that is somewhat close.

- Association is NOT strictly mathematical.
- It is often used to describe a group of people with a **common cause**, but it is also used to express any sort of **connection** between two things.

Correlation























- Correlation shown by the scatter plot displays the **relationship between the two continues variables**
- Scatter plots show how much one variable is affected by another and this relationship **strength** between two variables is called **correlation (R²)**
- A **trend line** within the scatter plot can be used to determine positive, negative or no correlation (**R**).



2 Pure Association ML Algorithms

1. Support Association ML Algorithm says how popular an item is, as measured by the proportion of transactions in which an item appears in a given set.
2. Lift Association ML Algorithm says how likely item Y is purchased when item X is purchased (while controlling for how popular item Y is).

Applying Association ML Algorithms

Transaction 1	   
Transaction 2	  
Transaction 3	 
Transaction 4	 
Transaction 5	   
Transaction 6	  
Transaction 7	 
Transaction 8	 

- The **support** of Apple is 4 out of 8 or 50%.
- The **lift** of purchasing Apple when Beer is purchased is 3 out of 6 or 50%.

Table of Contents

1

- Introduction

2

- Quality Data

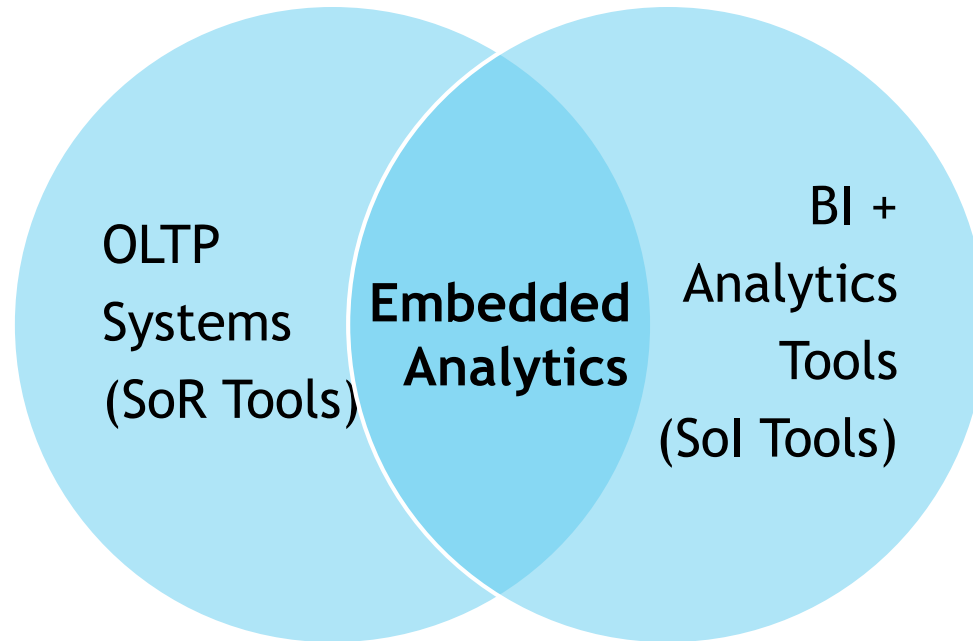
3

- Algorithms

4

- Embedded Analytics

Embedded Insights



Embedded Analytics = Workflow + API

Application of Embedded Insights

Power BI embedded analytics

Logistic
Regression
Models

This is a high risk
customer. Act now!
Offer her a free
perfume or 1000
Points!

SAP S/4HANA Embedded Analytics
The Comprehensive Guide

- Use embedded analytics for operational reporting and process analytics
- Configure SAP S/4HANA embedded analytics
- Integrate with other SAP tools for data warehousing, business intelligence, and predictive analytics

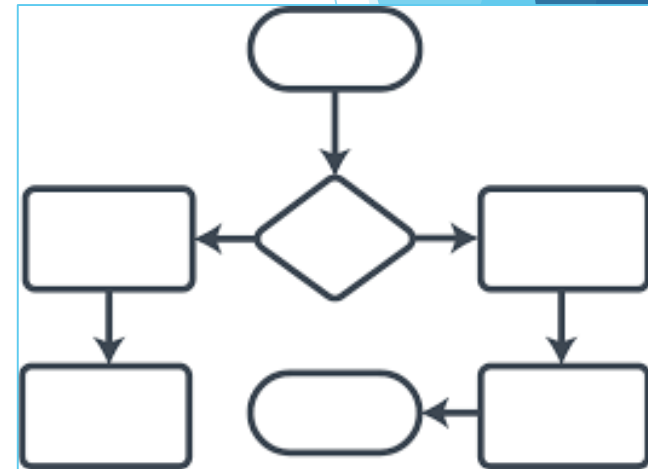


Embedded Analytics



Work Flow

- A workflow in the IT world is an orchestrated and repeatable pattern of activity, enabled by the systematic organization of resources, processes & information to achieve a result.
- Simply put a “workflow” is how the work is done by a person or group repeatedly and regularly with suitable approvals. Ex: PO Approval, Invoice payments, etc.
- The role of Workflow in Embedded Insights is to bring the insights from the Sol to SoR

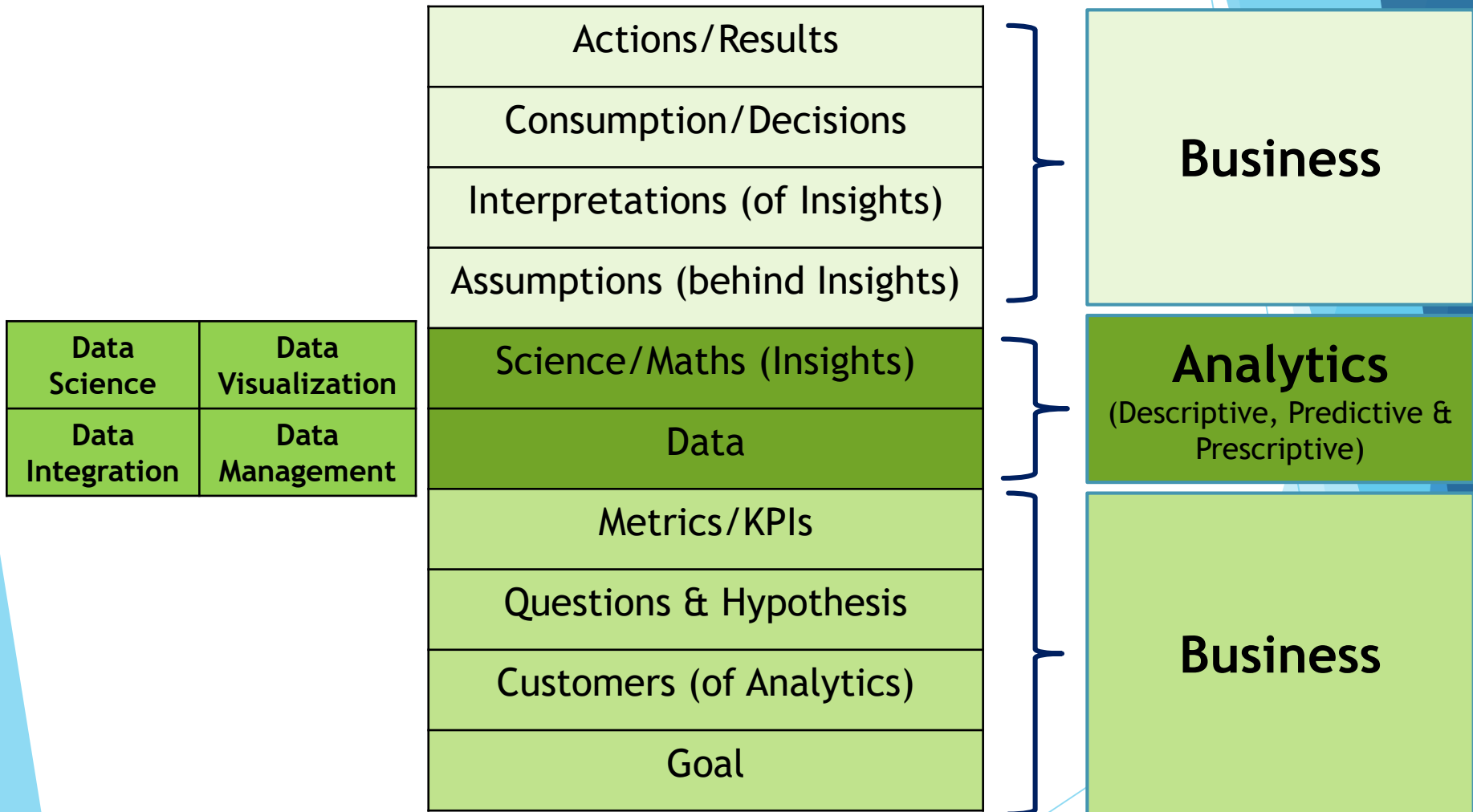


API - Glue in Sol & SoR Integration

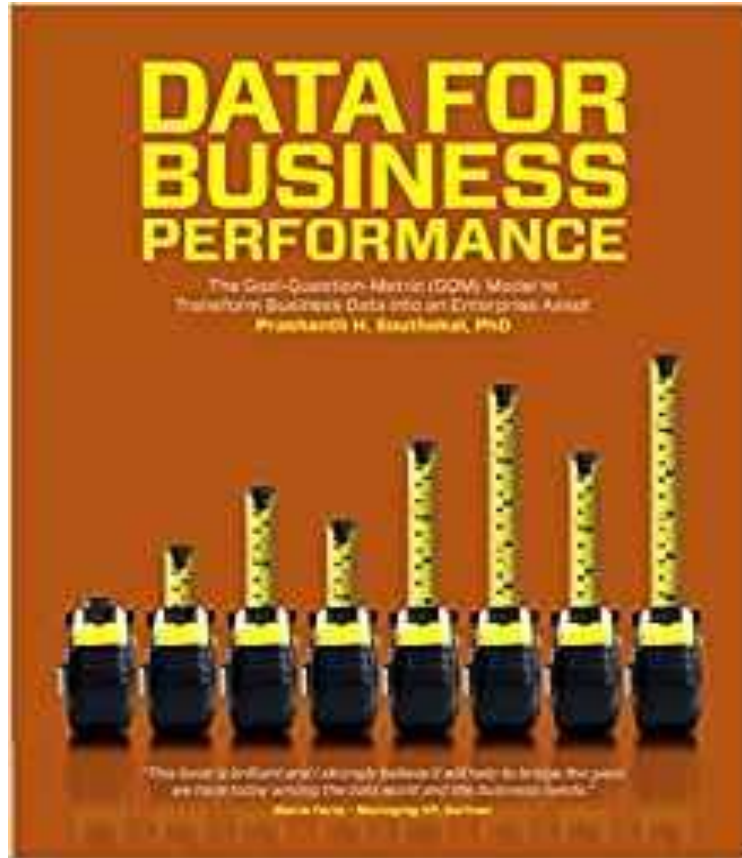
- API (Application Programming Interface) is a software intermediary that allows two systems to talk to each other.
- APIs can work with any common programming language; but the most popular approach is REST(Representational State Transfer) (SOAP is disappearing)
- API returns data in one of two possible formats:
 1. Extensible Markup Language (XML)
 2. JavaScript Object Notation (JSON)
- To control APIs from accessing data, the two primary mechanism are - Basic Auth & OAuth

Wrap-Up

Analytics Framework - Project Playbook



Winner of my Book



“For those interested in learning the basics of data quality, data standards, analytics and big data - in an interwoven way - need look no further than this book.”

Doug Laney,
Infonomics Author



Prashanth H Southeekal, Phd

Enterprise Data Analytics Consultant

2mo • Edited



Most of the Analytics projects these days have 2 main challenges . (1) There is no data at all (2) There is no quality data.

If you don't have data or quality data for validating your hypothesis, one option is to rework your hypothesis. If you are challenged with acquiring data internally, one approach is to get data from external sources. If you don't have good data for analytics, one strategy is to leverage sampling techniques and work with sample data. If you don't have precise/accurate data, one solution is to use ranges and confidence intervals.

Bottom line is Analytics is a probabilistic process and NOT a deterministic process. You can't expect a perfect situation in your Analytics projects. It simply doesn't exist. As they say, perfection is the opposite of getting things done!

#resultsmatter #analytics #dataquality #gettingthingsdone



Thank You

“An organization's ability to learn and translate learning/insights into action rapidly is the ultimate competitive advantage.”

Jack Welch

