IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

Course Name:

Supervised Machine Learning

By Prof. Vala A. Rohani
vala.ali.rohani@estsetubal.ips.pt

Session outline

1. Introducing myself
2. Course Topics
3. Introduction to Data Analytics
4. Installing R and R Studio

Introducing myself

- Post doctorate in Data Science
- PhD in Software Engineering (Recommender Systems)
- University Lecturer for more than 10 years
- Having more than 20 published papers in index conferences and journals

Some of my Professional Certificates:

- Hadoop Platform and Application Framework from University of California
- Practical Machine Learning from Johns Hopkins University
- R Programming from Johns Hopkins University
- MongoDB for DBAs from MongoDB
- Social Network Analysis from University of Michigan
- Mining Massive Datasets from Stanford University
- Pattern Discovery in Data Mining from Illinois University
- Process Mining from Eindhoven University of Technology

Vala Ali Rohani
Invited Assistant Professor
Founder and Chief Data Scientist at Data Corner
vala@data-corner.com

Course Topics

Course Topics

Docente	Sen	2ªF	3ªF	4ªF	5ªF	6ªF	Remoto/	Milestone		
VR	1 (4h)	08-Jan	09-Jan	10-Jan	11-Jan	12-Jan	P	Lab1	Presentation of course syllabus and evaluation schema. Introdcution to Data Science and Machine Learning Reviewing the essential topics in R Programming Exploratory Data Analysis in R	**Lab1** - Setting up the R and R Studio Reviewing the essential topics in R Programming Exploratory Data Analysis in R
VR	2 (8h)	15-Jan	16-Jan	17-Jan	18-Jan	19-Jan	R	Lab2	Preparation and Data Cleaning for ML (Machine Learning) in R Preparing Training & Test datasets Regression Linear Models (1)	**Lab2** - Data Cleaning in R Preparing Training & Test datasets Building Regression linear Models in R (1)
VR	3 (12h)	22-Jan	23-Jan	24-Jan	25-Jan	26-Jan	R	Lab3	Regression linear Models (2) Evaluating the Performance of Regression Models	**Lab3** - Building Regression linear Models in R (2) Applications
VR	4 (16h)	29-Jan	30-Jan	31-Jan	01-Feb	02-Feb	R	Lab4	Classification Models (1) Logistic Regression	Lab4 - Building Logistic Regression Applications
VR	5 (20h)	05-Feb	06-Feb	07-Feb	08-Feb	09-Feb	P	Lab5	Classification Models (2) Linear Discriminant Analysis	Lab5 - Building Classification Models in R (2) Applications
HP	6 (24h)			07-Feb			R	Lab6	Classification Models (3) Quadratic Discriminant Analysis Naive Bayes Comparison of Classification Methods	**Lab6** - Building Classification Models in R (3) Applications
HP	7 (28h)	12-Feb	13-Feb	14-Feb	15-Feb	16-Feb	R	Lab7	Tree-Based Methods (1) Regression Trees	**Lab7** -Building Tree-Based Methods (1) Applications
HP	8 (32h)			14-Feb			P	Lab8	Tree-Based Methods (2) Classification Trees	**Lab8** -Building Tree-Based Methods (2) Applications
HP	9 (36h)	19-Feb	20-Feb	21-Feb	22-Feb	23-Feb	R	Lab9	Tree-Based Methods (3) Bagging, Random Forests, Boosting, and Bayesian Additive Regression Trees	**Lab9** -Building Tree-Based Methods (3) Applications
HP	10 (40h)			21-Feb			R	Lab10	Final Project Presentations	
HP	11 (44h)	26-Feb	27-Feb	28-Feb	29/fev	01-Mar	P	Test	Written Test	
VR		04-Mar	05-Mar	06-Mar	07-Mar	08-Mar	P	Exam	Written Exam	

Marking Scheme

Marking Scheme

Marking (Continues)		Labs	20%	
		project (grp)	30%	
		Written test	50%	min 9
Exams (I, II [III])		Practice	50%	
		Written	50%	min 9

Introduction to Data Analytics

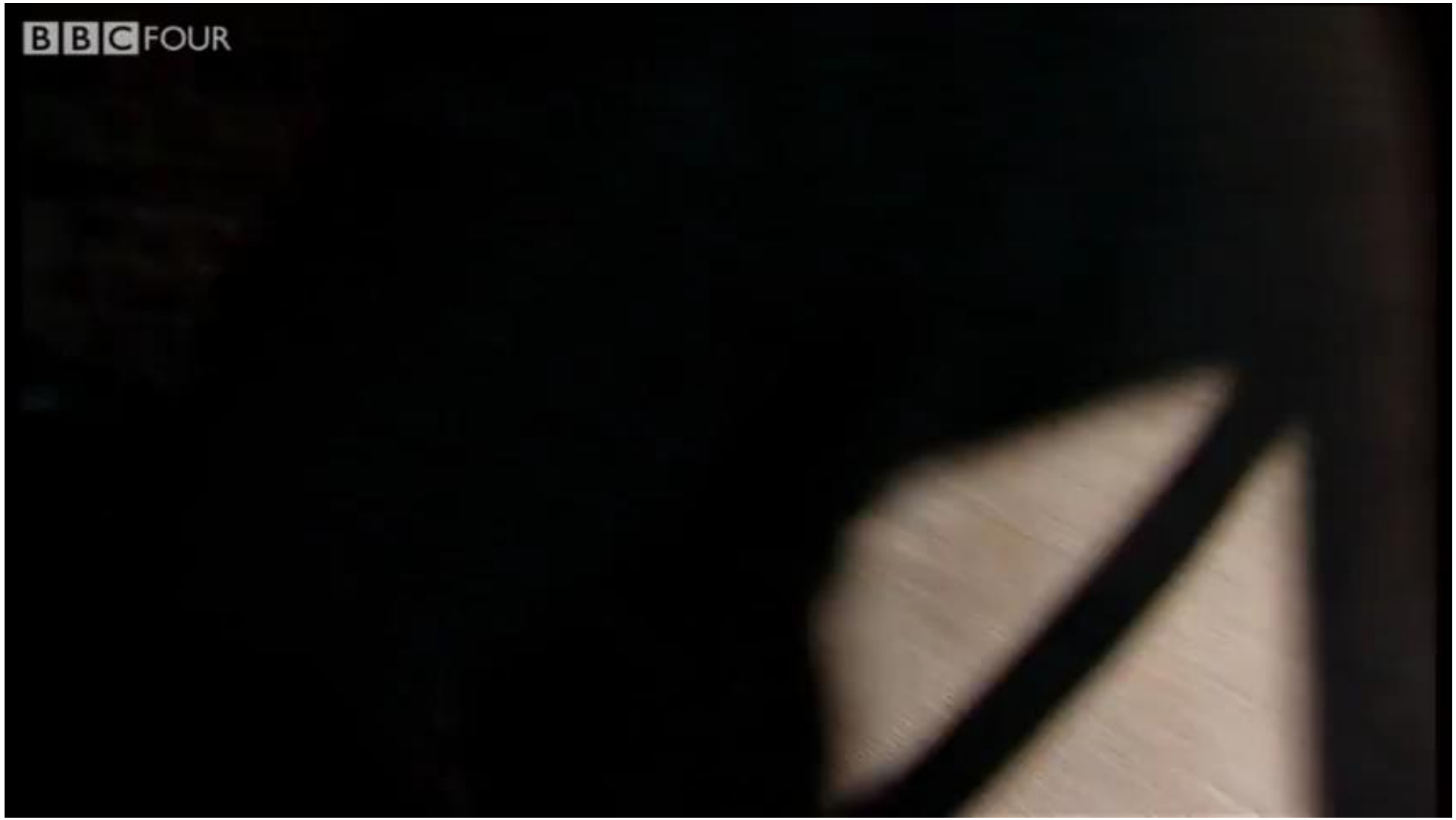
INFORMATION IS THE OIL AND ANALYTICS IS THE COMBUSTION ENGINE

What is Data Analytics?

Data Analytics is the discovery, interpretation, and communication of meaningful patterns in data.

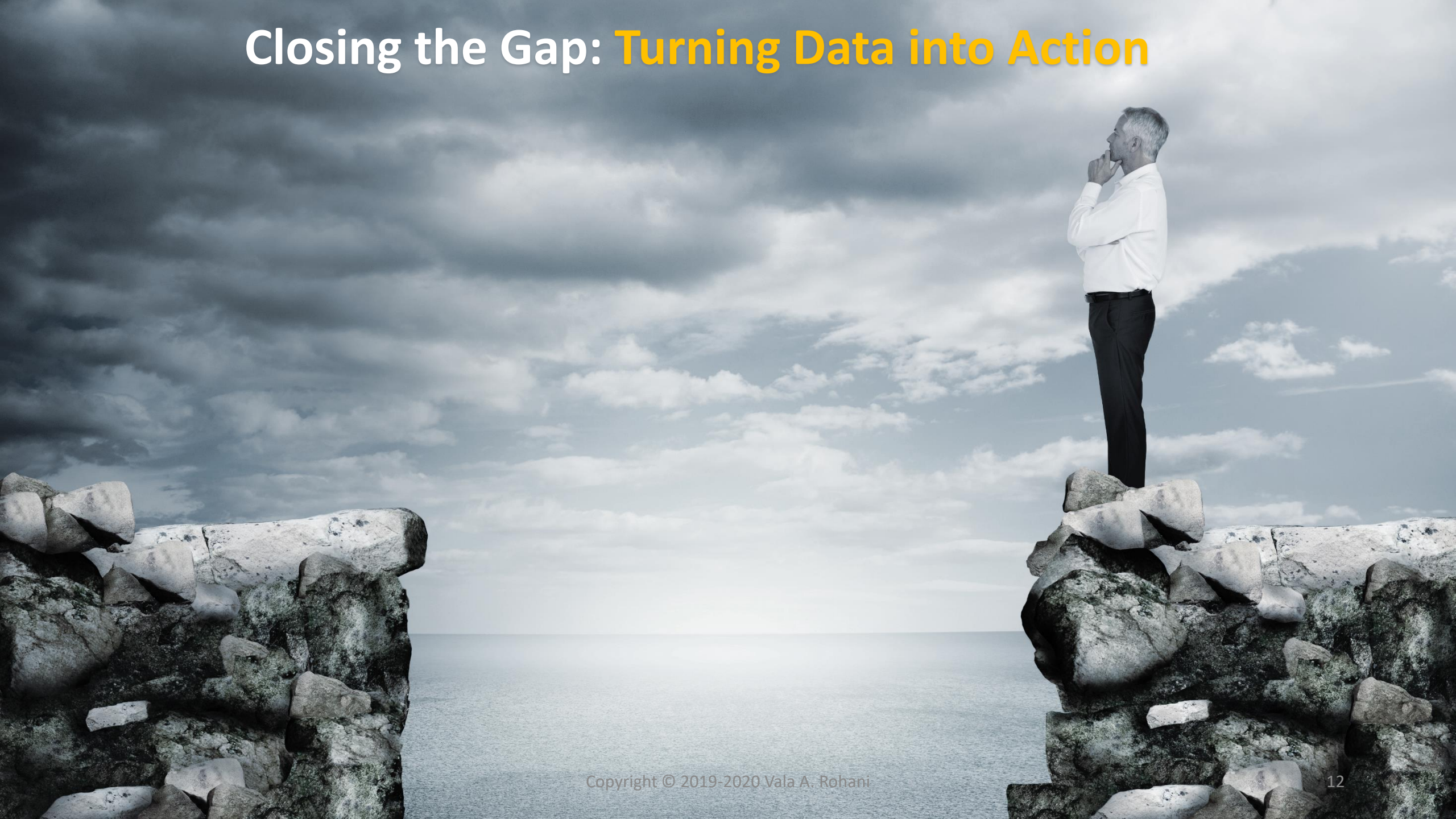
Analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance and often favors data visualization to communicate insight.

Let’s watch an inspired video on data visualization



https://www.youtube.com/watch?v=jbkSRLYSojo

Closing the Gap: Turning Data into Action



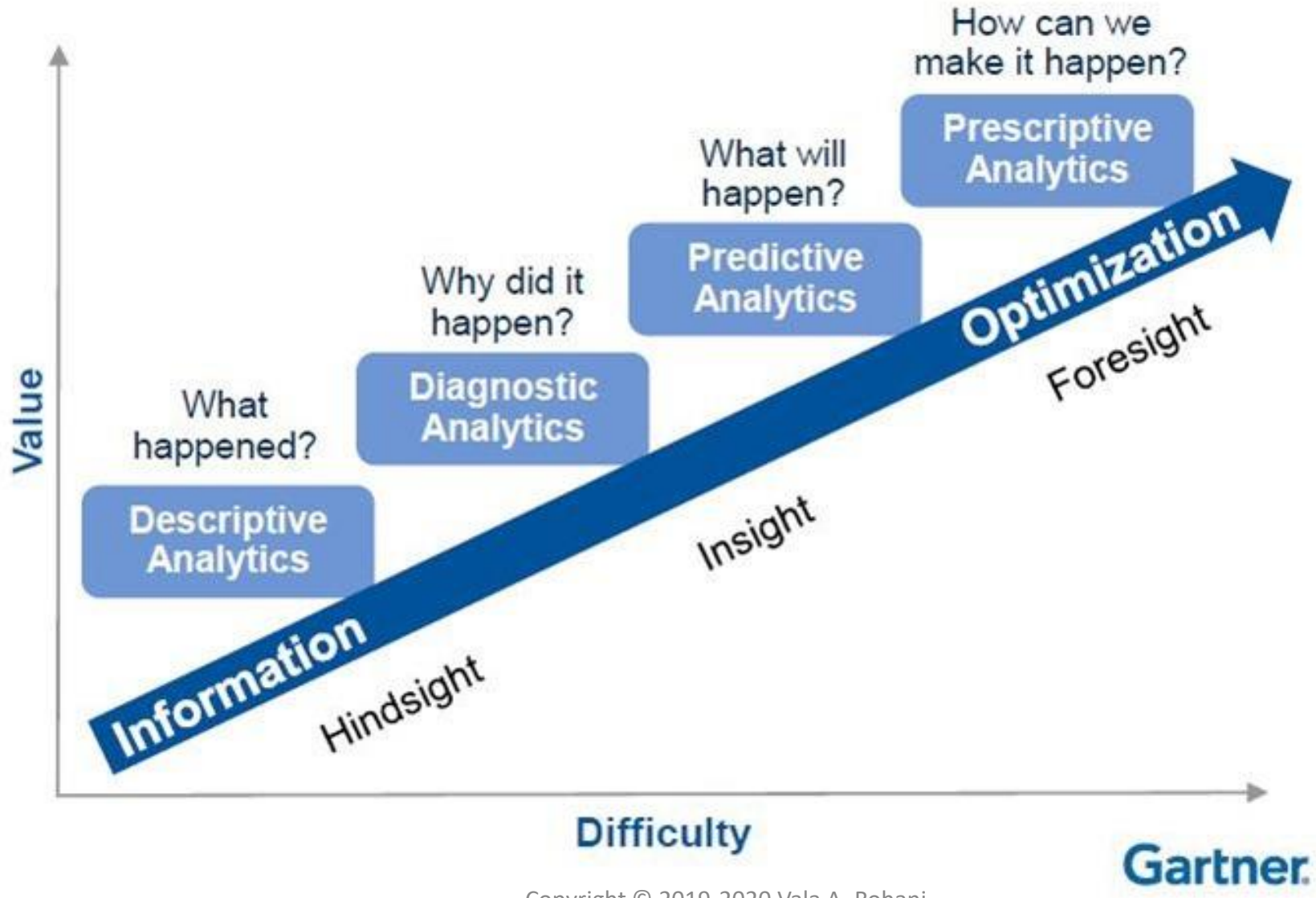
Rise over the noise …

Four key analytics types

Analytics solutions are of four principal types:

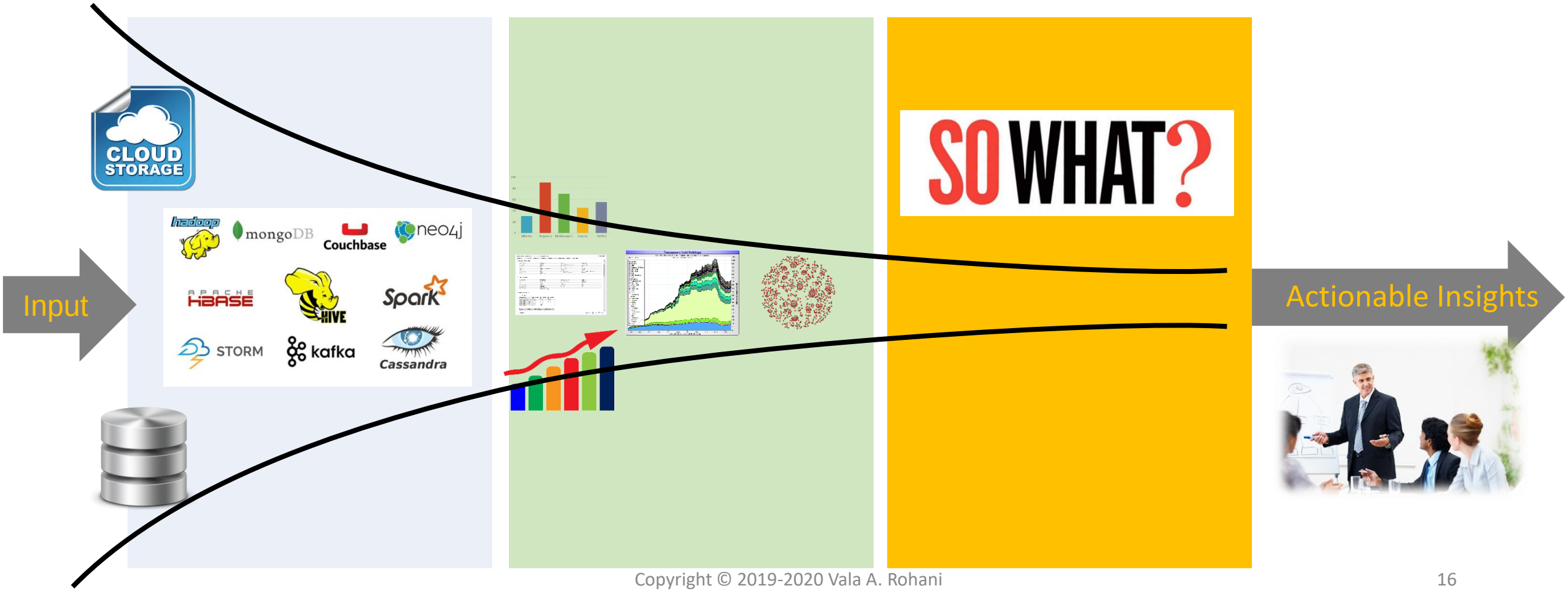
- **Descriptive:** which uses business intelligence and data mining to answer: "What has happened?"
- **Diagnostic:** which uses root cause analytics techniques to answer: "Why did it happen?"
- **Predictive:** which uses statistical models and forecasts to answer: "What could happen?"
- **Prescriptive:** which uses optimization and simulation to answer: "What should we do?"

The analytics maturity model – Where are we in the journey ?

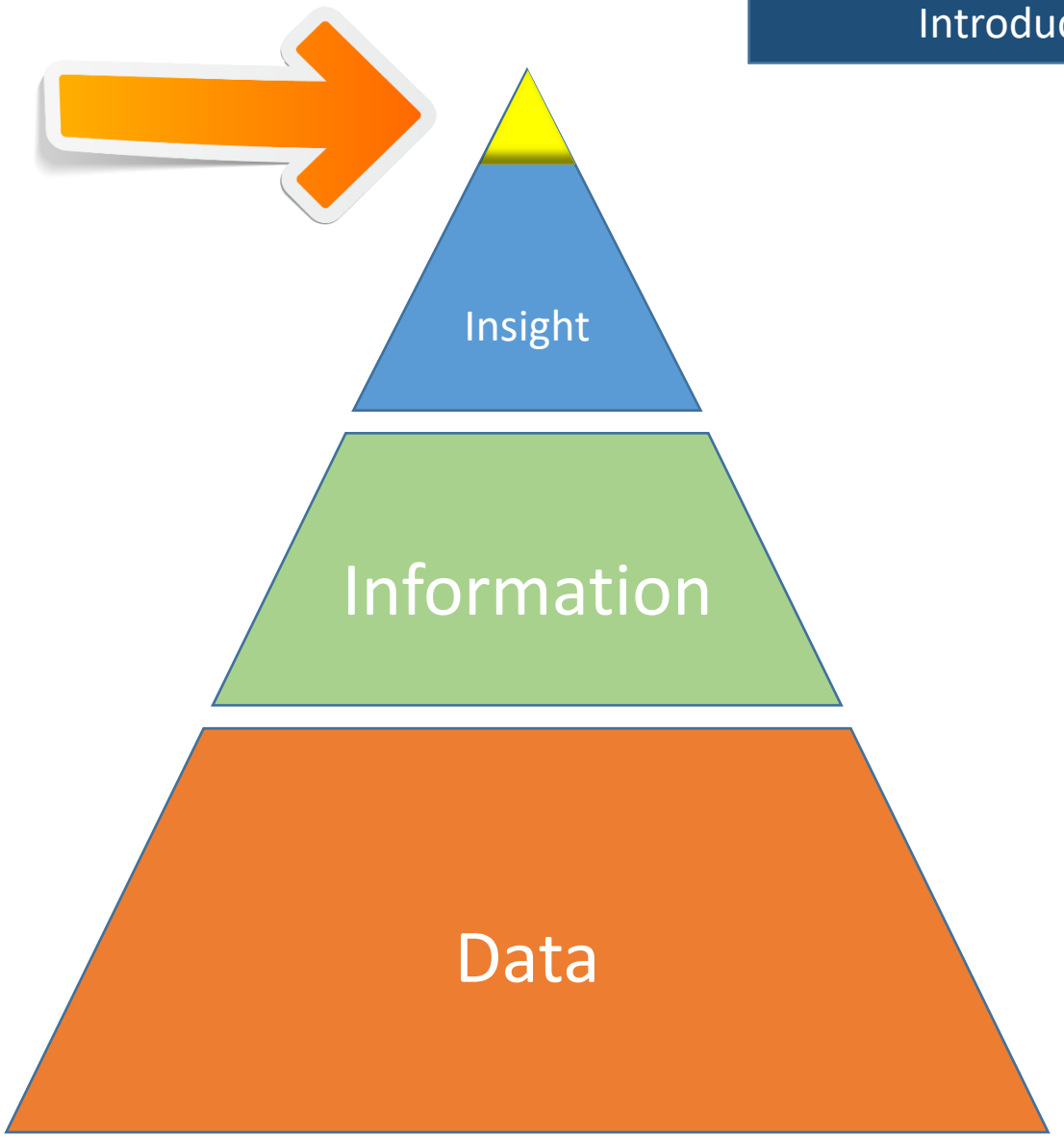


The journey from Data to Actionable Insights

Data Information Insights



Actionable Insights



Do not get trapped in your Data and Reports

Be an Analytics Ninja because …

Actions drives business forward not data or reports!

4 Angles of Actionable Insights

1) **What** specific action to perform?

2) **How** will the specific action improve the business

3) **Who** is responsible for the action?

4) By **when** should the action be performed?

How to generate Actionable Insights?

Data Democratization

Course: Data Analytics

Introduction

Data Democratization

Problem Statement:

- In the past, data has been primarily entrusted to only two privileged groups : **Executives** and **Data Specialists**
- But After decentralizing the decision-making and increasing responsiveness, Organizations are seeking to empower more workers by granting them the access to related Data
- However, managers have some fear of how people will use and interpret the data.
- As a result, more business questions end up funneling through the analytics team for answers.
- **Unfortunately, analytics teams often can't scale sufficiently to handle the increasing volume of data-related questions**

So what's an aspiring data-driven organization to do in this situation?

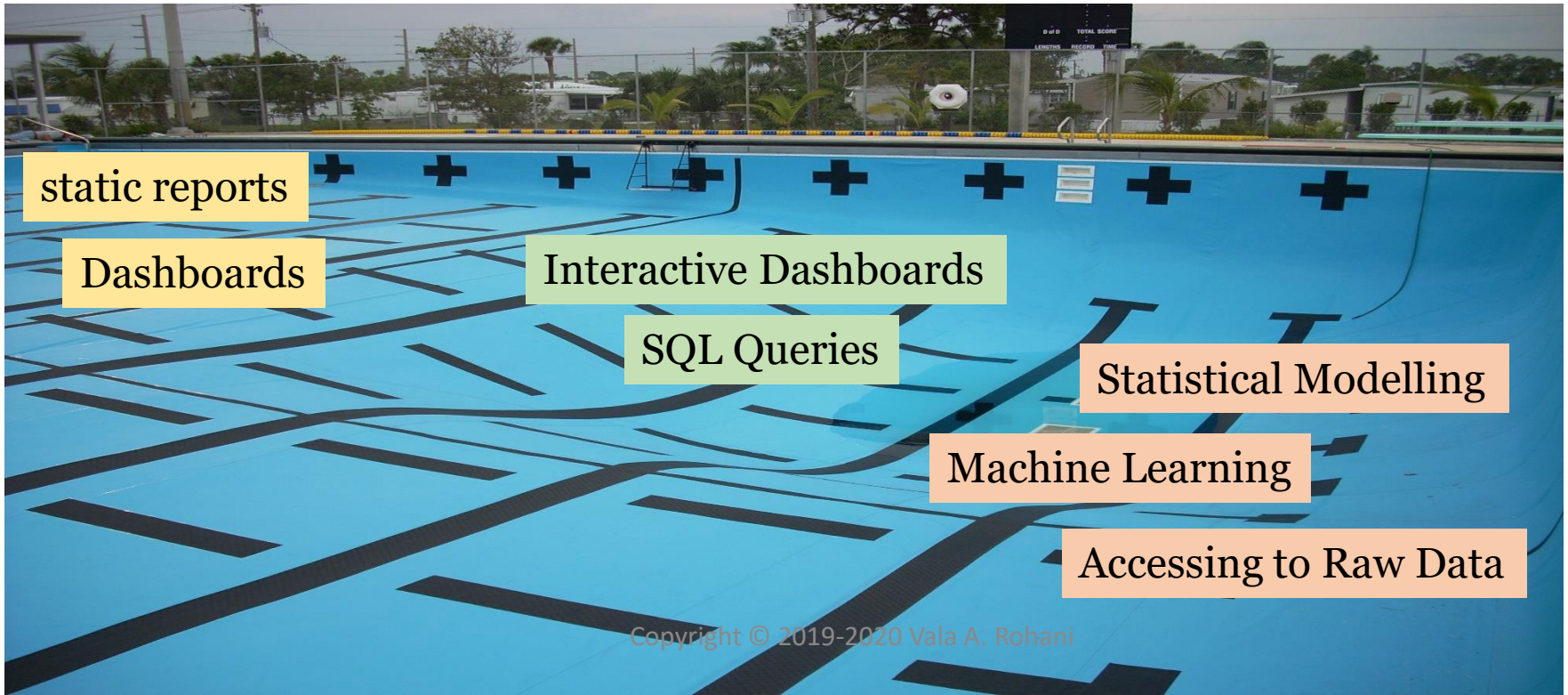
> You want your employees to be able to access the data they need, but you also don't want them to harm themselves with it.

Your corporate data is like a giant swimming pool

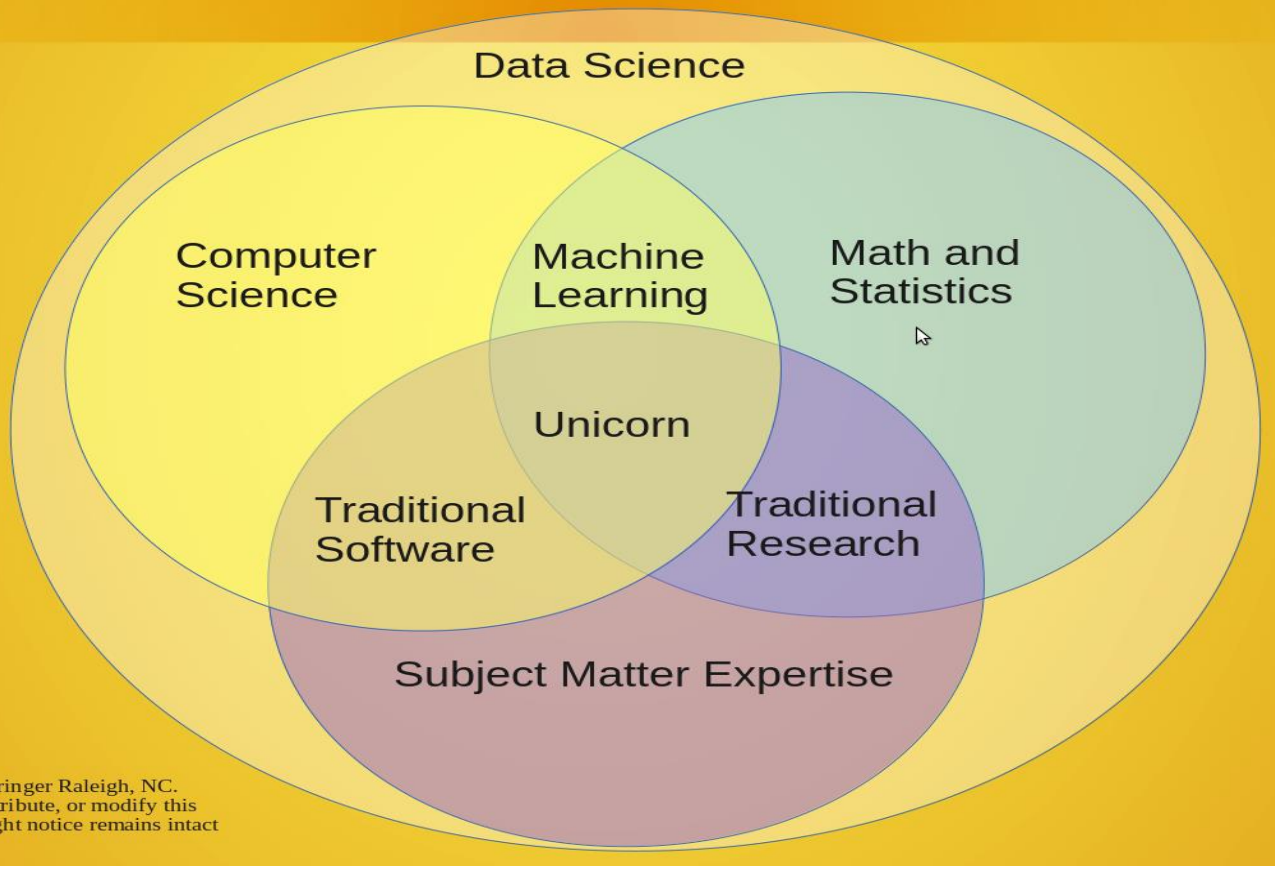
Data Democratization

As the Solution:

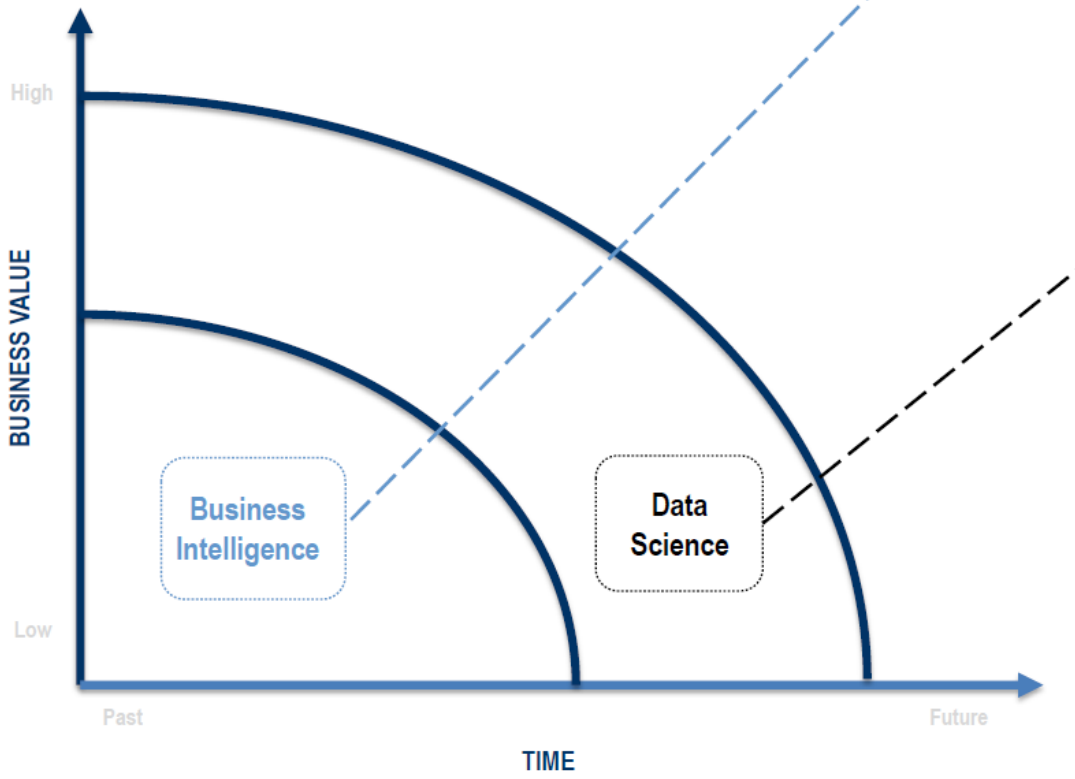
A multi-tiered data analytics approach is recommended, so various users can dive into the right depth of data based on their needs and analytical skills.



Data Science Venn Diagram v2.0



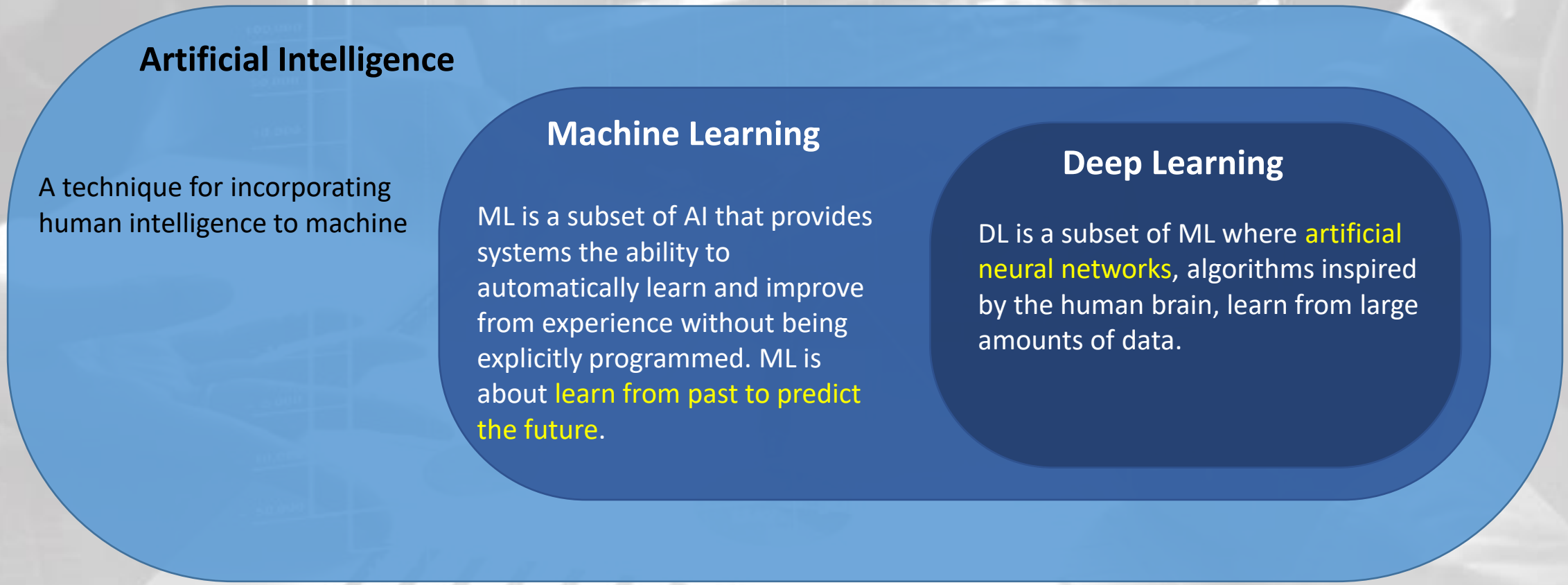
Business Intelligence vs Data Science



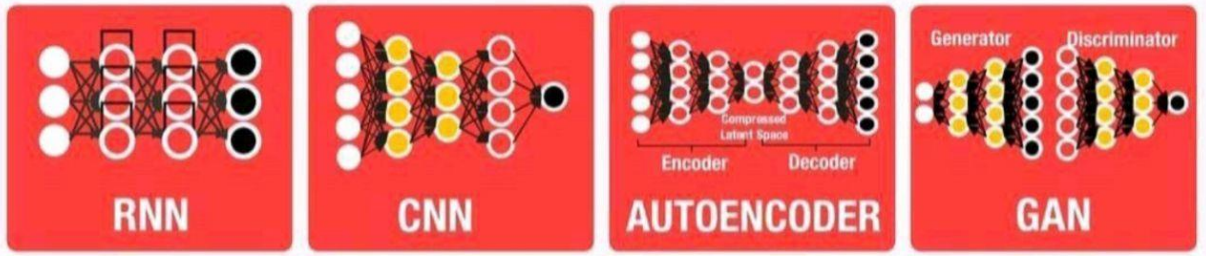
Typical Techniques	• Standard and ad hoc reporting • Dashboards, alerts, queries, details on demand
Data Types	• Structured data • Traditional sources • Manageable data sets
Common Questions	• What happened last quarter? • How many did we sell? • Where is the problem? In which situations? • How many? When? Where?

Typical Techniques	• Optimization, predictive modeling, forecasting, statistical analysis, Scenario Simulation etc.
Data Types	• Structured, Unstructured, Semi-structured data • Many Disparate data Sources • Very large data sets
Common Questions	• What if … • What is the best that can happen? • What will happen next? What if these trends continue? Why is this happening?

AI vs ML vs DL



DEEP LEARNING USE CASES



RNN	CNN	AUTOENCODER	GAN
Stock Price Prediction	Face Recognition	Dimensionality Reduction	Generating Image Data
Chatbots	Medical Image Detection	Image Compression	Generating Art
Voice Assistant	Object Detection	Feature Extraction	Image-to-Emoji Conversion
Weather Forecasting	Character Recognition	Anomaly Detection	Face Attribute Manipulation
Music Generation	Document Classification	Fraud Detection	3D Object Generation

DDO?

Data Driven Organization

What is NOT DDO?

Having Lots of reports does not make you data driven!

What is NOT DDO?

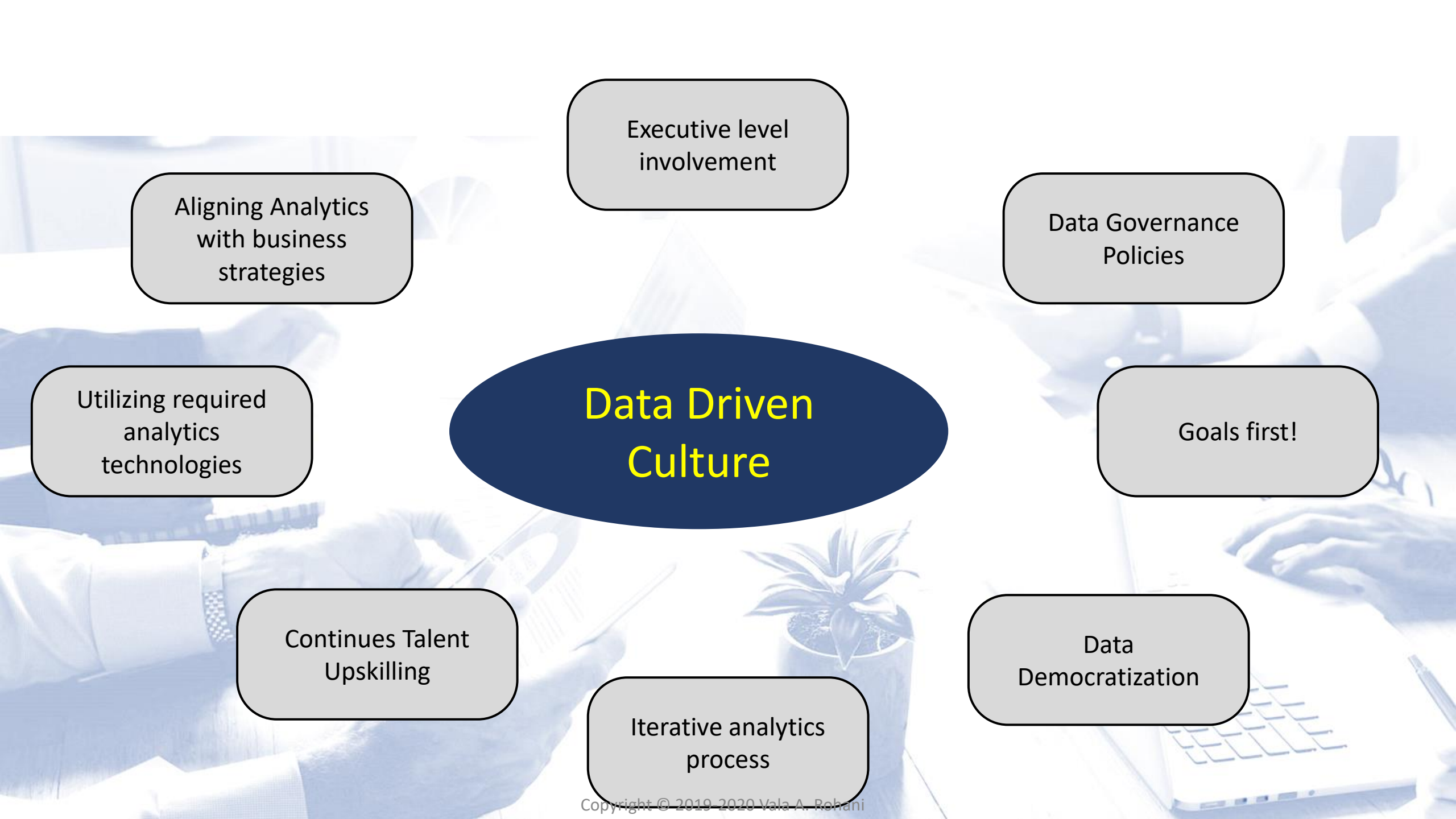
Having Lots of dashboards does not make you data driven!

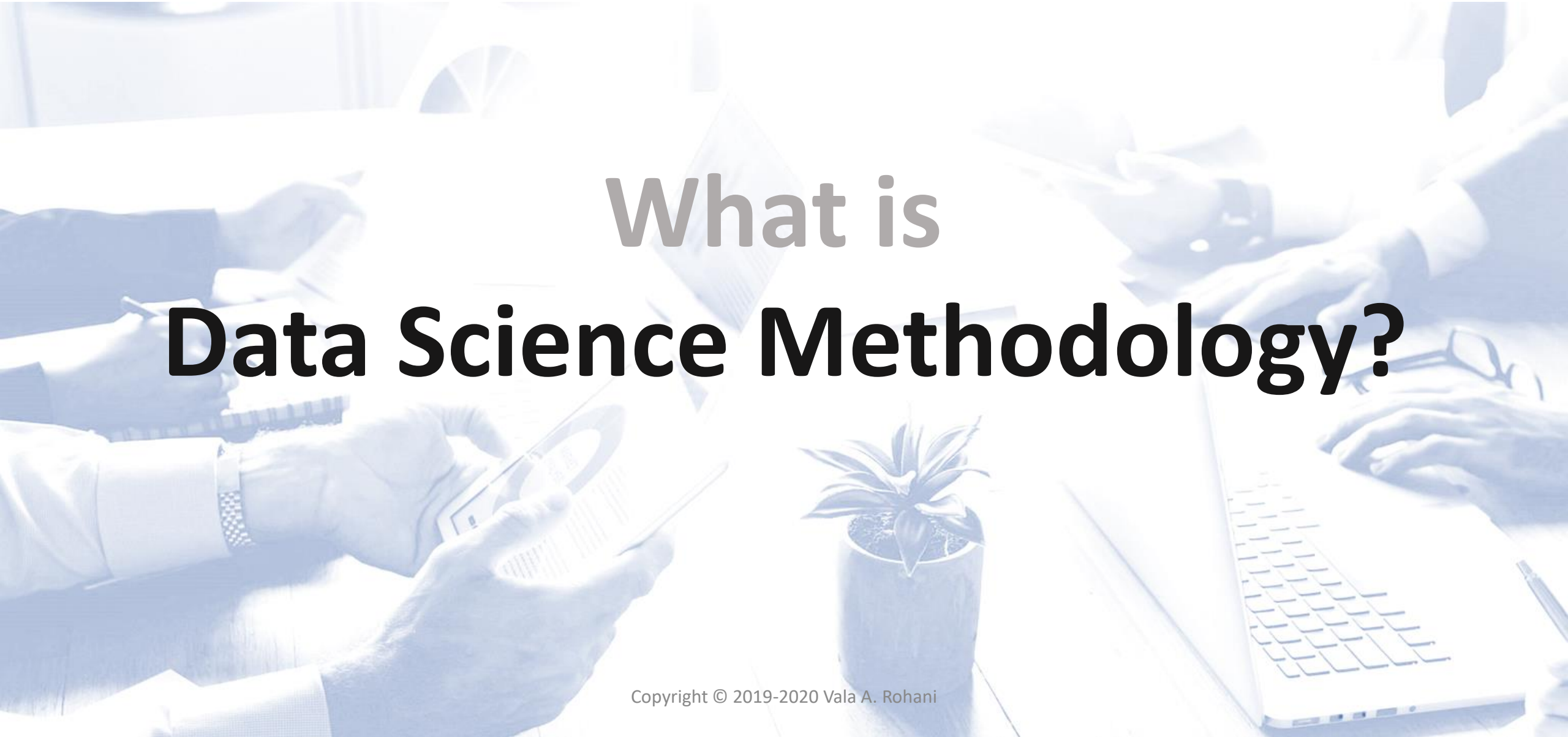
What is NOT DDO?

Having BIG Data clusters does not make you data driven!

What is DDO?

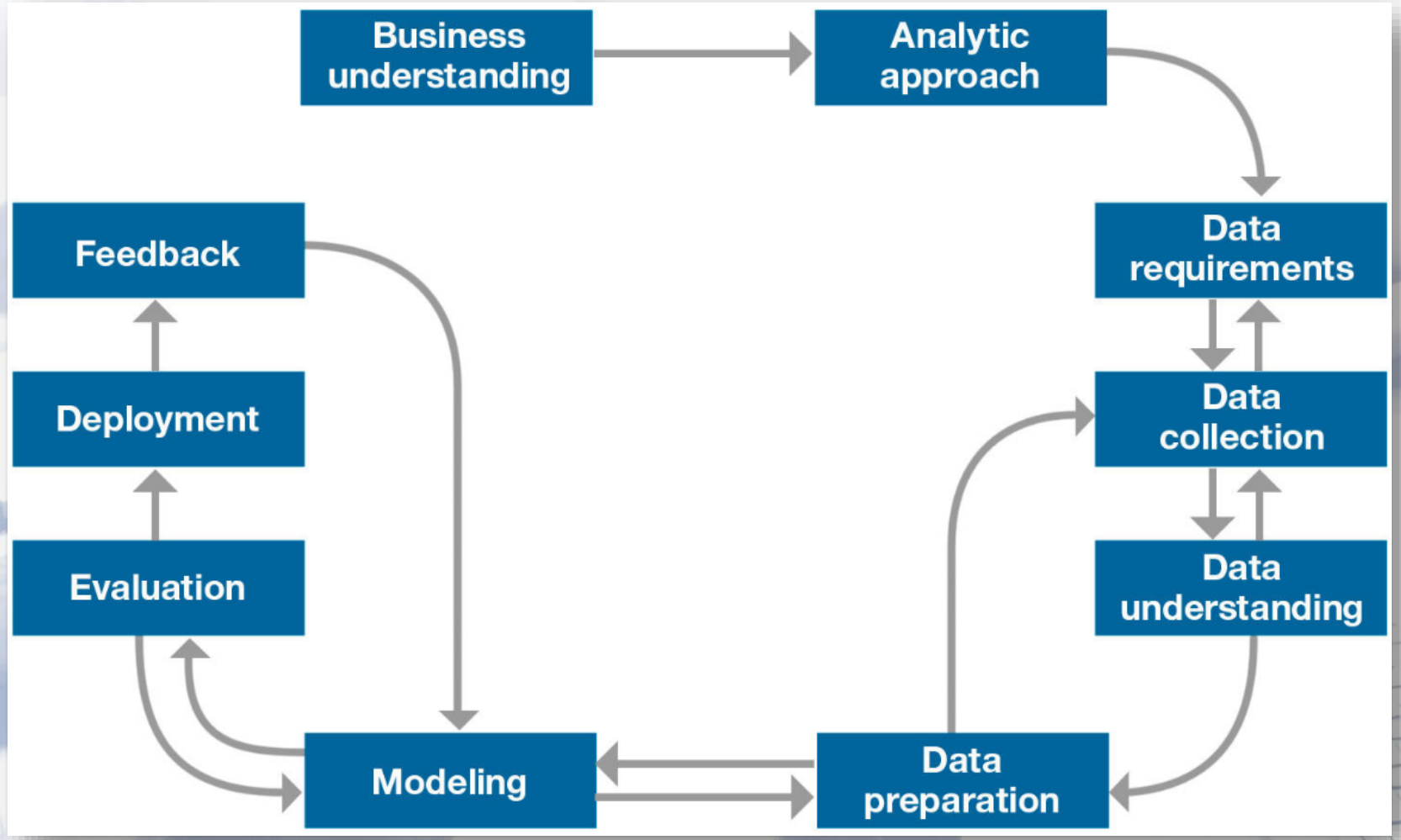
To become DDO, you need to create a **Data-Driven Culture** in your organization





What is
Data Science Methodology?

Data Science Methodology



https://www.ibmbigdatahub.com/blog/why-we-need-methodology-data-science

What skills you need to become Data Scientist?

DATA SCIENCE SKILL SET

BASIS 〉 CORE 〉 ENABLEMENT

TECHNICAL SKILLS	ANALYTICAL SKILLS	BUSINESS SKILLS
• CODING SKILLS	• ADVANCED STATISTICS & INFERENCE	• EVALUATION AND DEVELOPMENT OF BUSINESS CASES
• HANDLING DATA	• MODELLING & SIMULATION	• PROJECT MANAGEMENT
• COMPUTATIONAL TOOLS	• MACHINE LEARNING	• BUSINESS PROCESSES
• BASIC SOFTWARE DEVELOPMENT	• COMPUTER SCIENCE	• CHANGE MANAGEMENT
• BIG DATA	• ADVANCED MATH	• COMMUNICATION SKILLS
• HIGH PERFORMANCE COMPUTING	• DATA VISUALIZATION	• LEADERSHIP SKILLS
• PARALLEL COMPUTING	• EXPERIMENT DESIGN	
	• RESEARCH EXPERTISE	
TECH-SAVVY	***SCIENTIFIC PROBLEM SOLVING***	***APPETITE FOR BUSINESS PROBLEMS***

https://www.datasciencecentral.com/profiles/blogs/data-scientist-skill-set

Further details?!

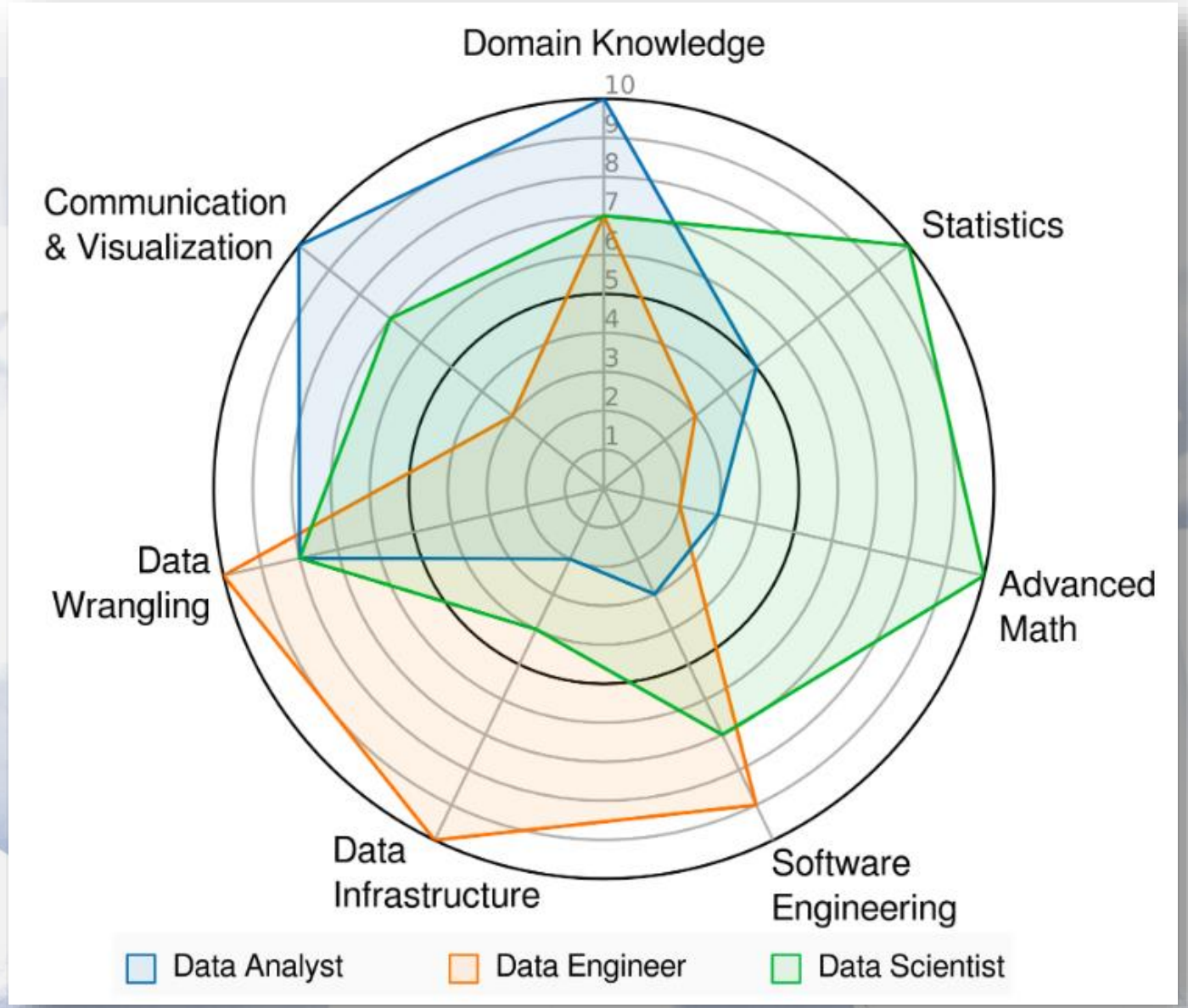

1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
15%
Factor Analysis
Install Pkgs
3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS
4. Machine Learning
Perceptron
Linear Regression
Ranking
Logistic Regression
Regression
K-Nearest Neighbor
Naïve Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
Classification
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
80%
Using ETL
How much Data?
Google OpenRefine
Data Survey
Sampling
Stratified Sampling
Principal Component Analysis
9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction
Transformation & Enrichment
8. Data Ingestion
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation
40%
Tableau
IBM ManyEyes
InfoVis
D3.js
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
7. Big Data
Job & Task Tracker
M/R Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
Author: Swami Chandrasekaran

Data Science: A mix of hard and soft skills

Data science is one area of the digital sector that is desperately in short of talent. According to EDM Monitoring Tool:

Europe will face a shortage of 925,000 data professionals by 2025.

The Data Science Team: Who Does What



How to accelerate the adoption of **data science** within an organization?

Prepared and Presented by Dr Vala Ali Rohani

How to accelerating the adoption of *data science* within an organization?

1. Access to Data, in the rawest form

2. Having strong data governance practices in place

3. Provide an analytic sandbox

4. Leverage data visualization tools

5. Keep your team structures flexible

6. Integrate data scientists with existing business units

7. Be sure your data science sponsors are 100% committed to supporting the team

http://www.informationweek.com/big-data/big-data-analytics/how-to-build-a-successful-data-science-team-/d/d-id/1113234

How to accelerating the adoption of *data science* within an organization?

8. Develop internal team skillsets

9. Iterate quickly

10. Be comfortable with data not being perfect

11. Plan now for scaling up

12. Think about producing Actionable Insights

How to manage a data science project

Prepared and Presented by Dr Vala Ali Rohani

What is your main responsibility as **Data Scientist?**

To use Data and Analytics to solve stakeholder business problems.

To this end, we need to know …

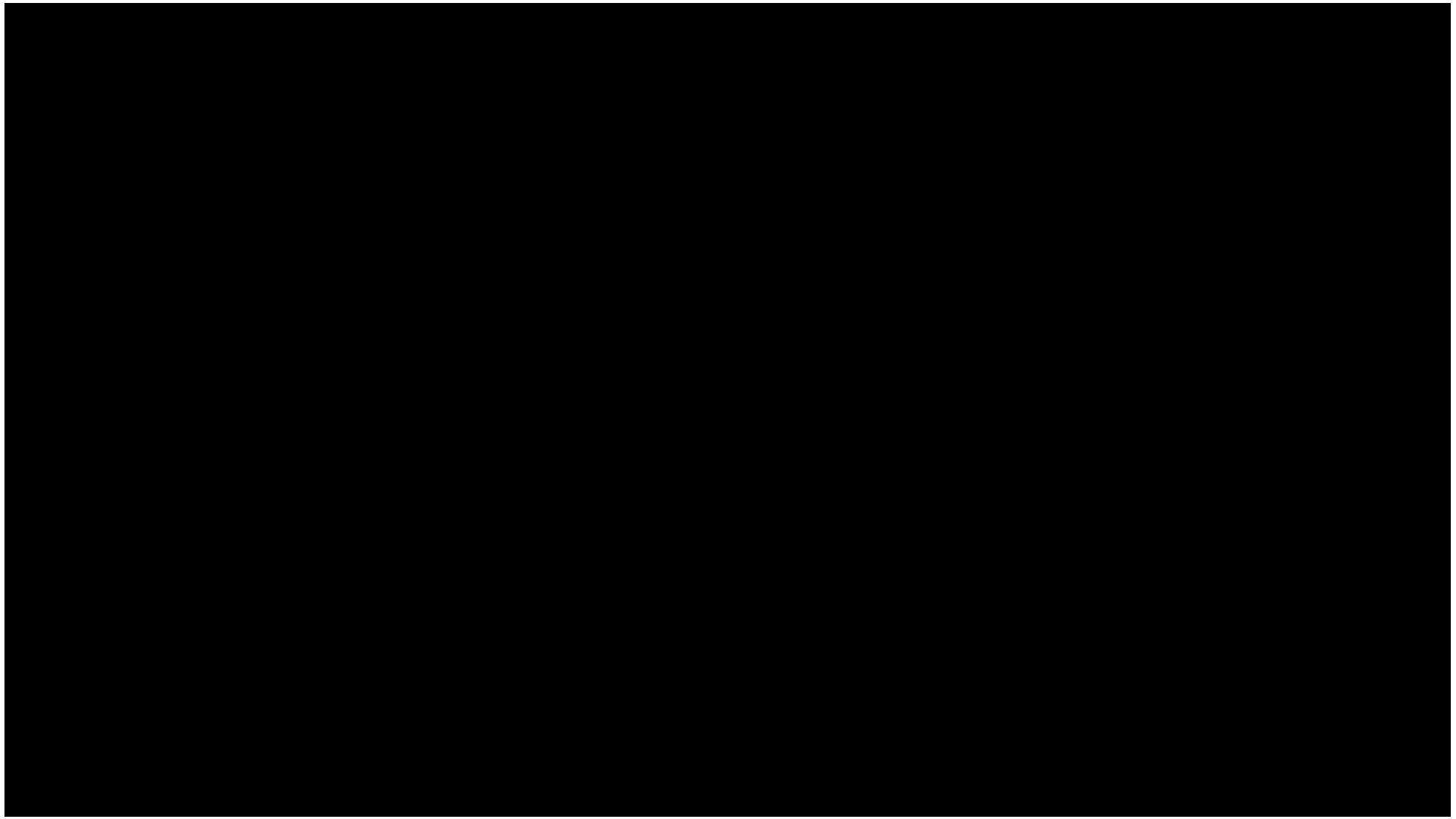
What the real business problem is?

Meet the primary people: ***Stakeholders***

And, ask questions …

- What the problem is in this business that you are hoping to solve using analytics?
- How this problem is affecting your business?
- What is your ideal outcome for this project?

Let’s watch a story about one of real challenges in Analytics Projects!



https://www.coursera.org/learn/analytics-tableau/lecture/ruazC/rock-projects

To ….

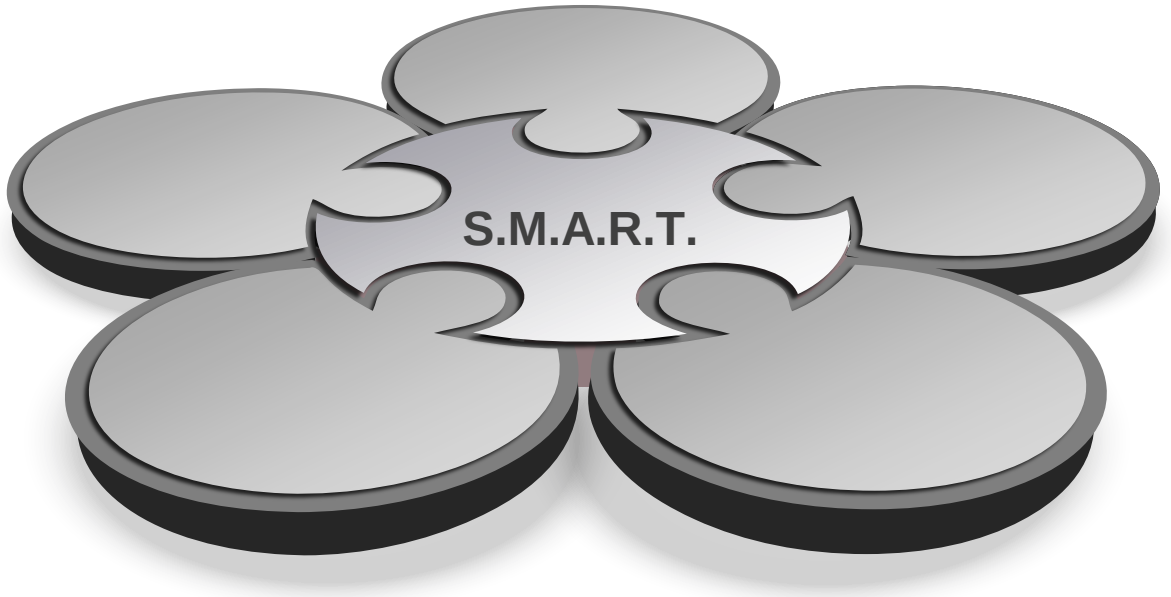
Identify effective ***objectives***

But How???

S.M.A.R.T. *Framework*

George Duran, 1981

The establishment of projects objectives is suggested be created using the **S.M.A.R.T.** philosophy.



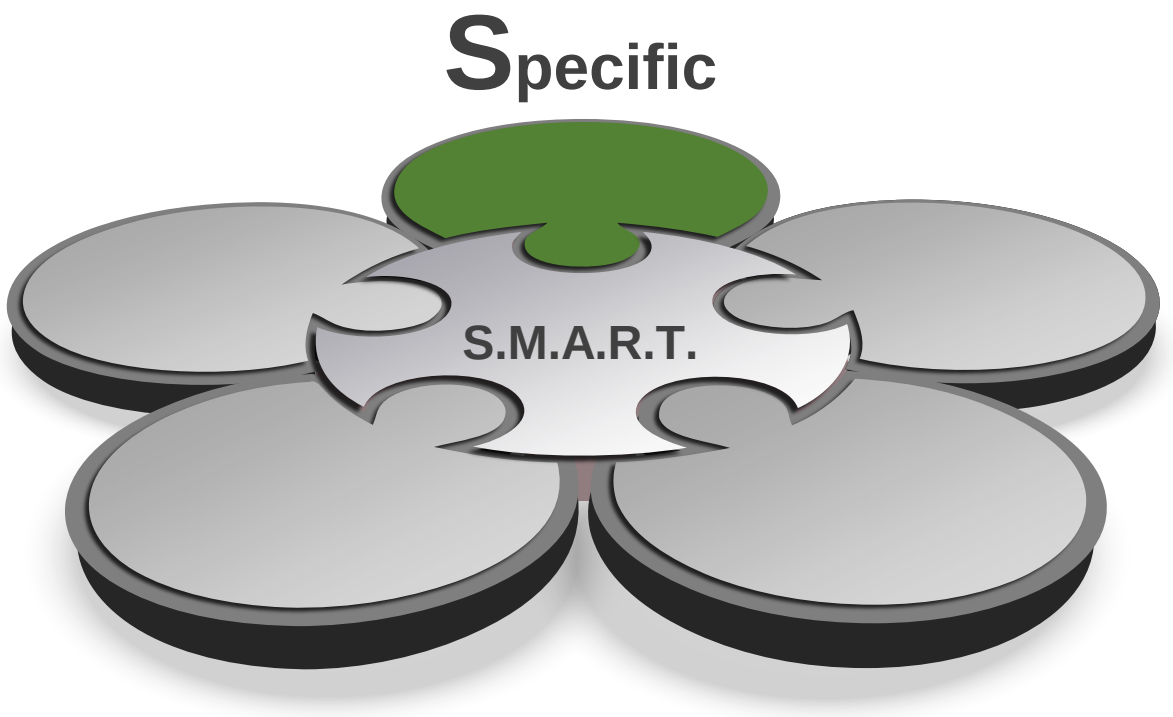
Let’s start with an example:

Imagine your client is an online retailer of the men’s clothing store.

Project manager describe the problem as follows:

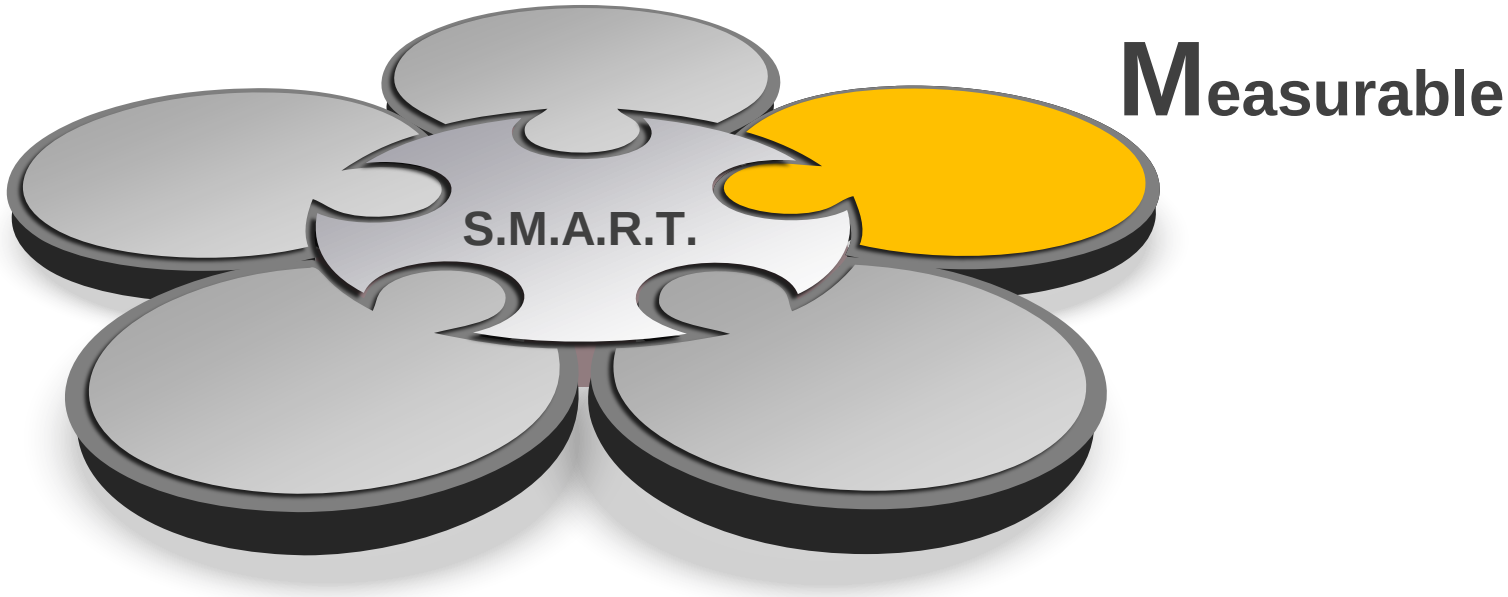
> People aren’t returning back to our website after they login for the first time. I need you to tell us how convert first time visitors to returning visitors.

Phrasing the Project Goal using **SMART** method:



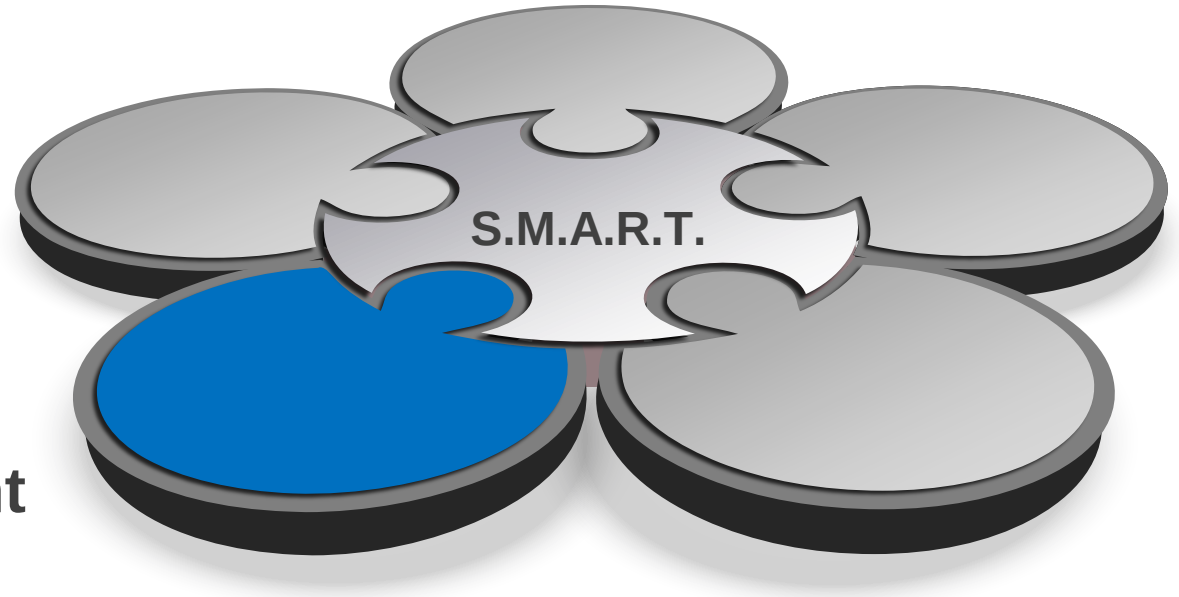
Increasing the number of returning visitors

Phrasing the Project Goal using **SMART** method:



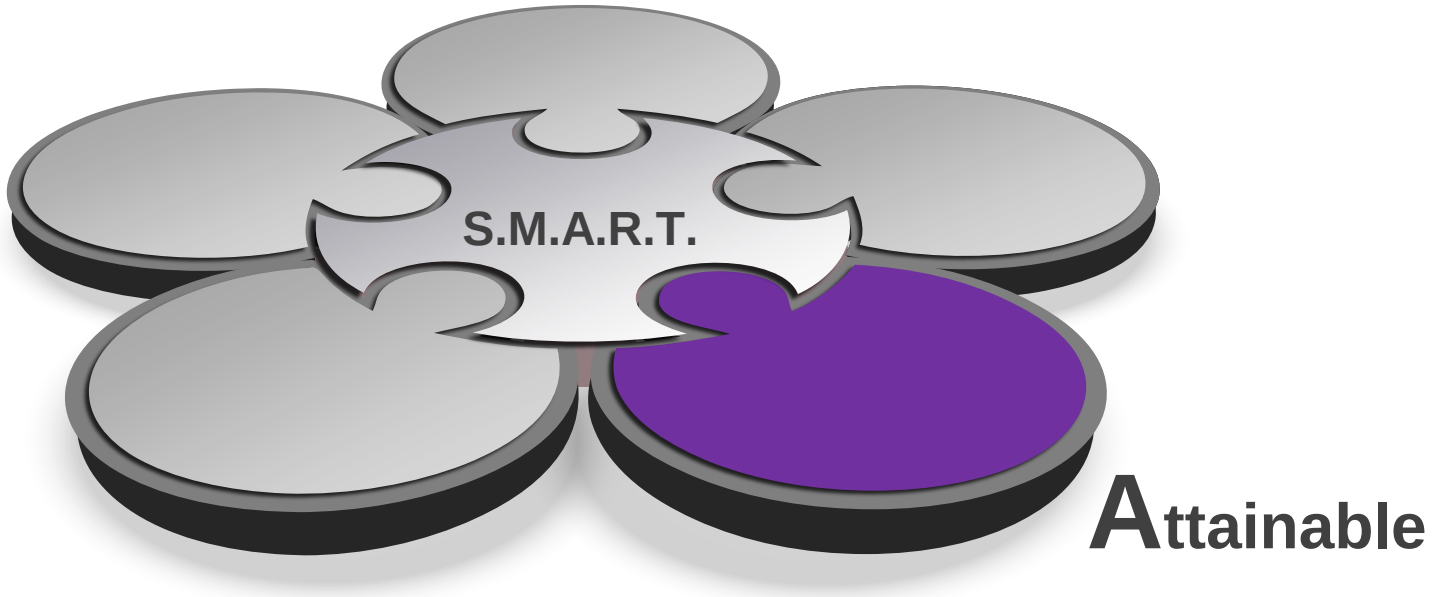
Increasing the number of returning visitors **on a month-by-month basis by at least 15% compare to the last month during last year.**

Phrasing the Project Goal using **SMART** method:



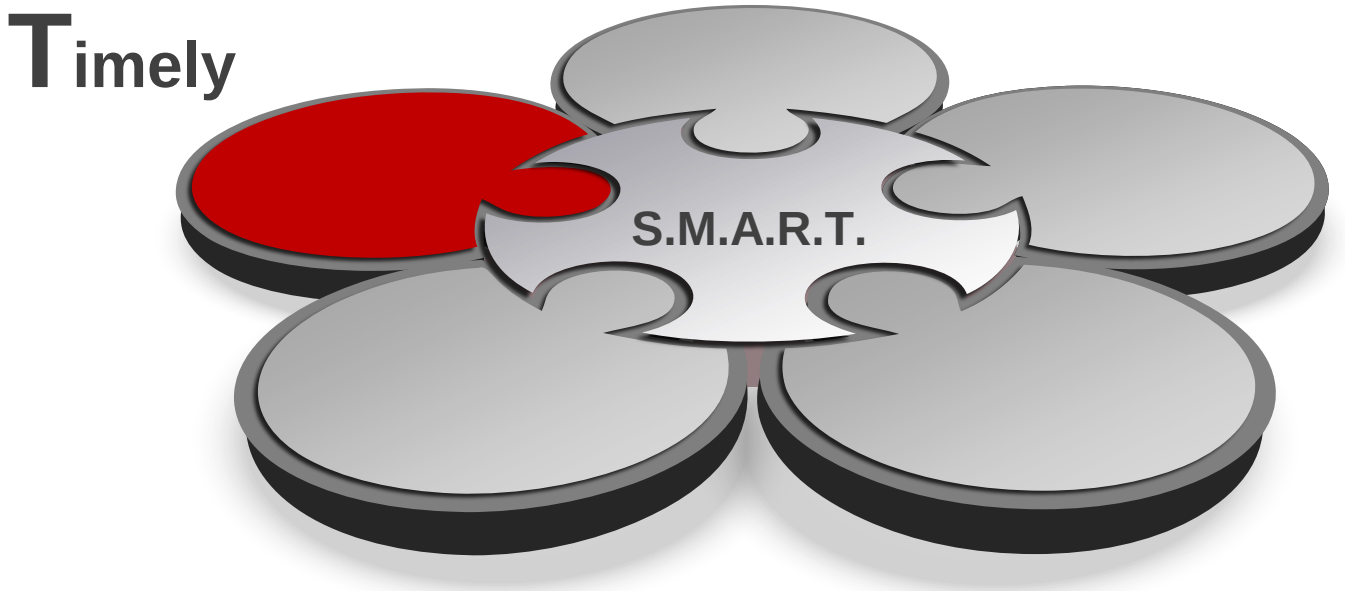
The goal of the project is **to determine the website changes that will most efficiently increase revenues** by 15% on a month-by-month basis compared to the same month last year.

Phrasing the Project Goal using **SMART** method:



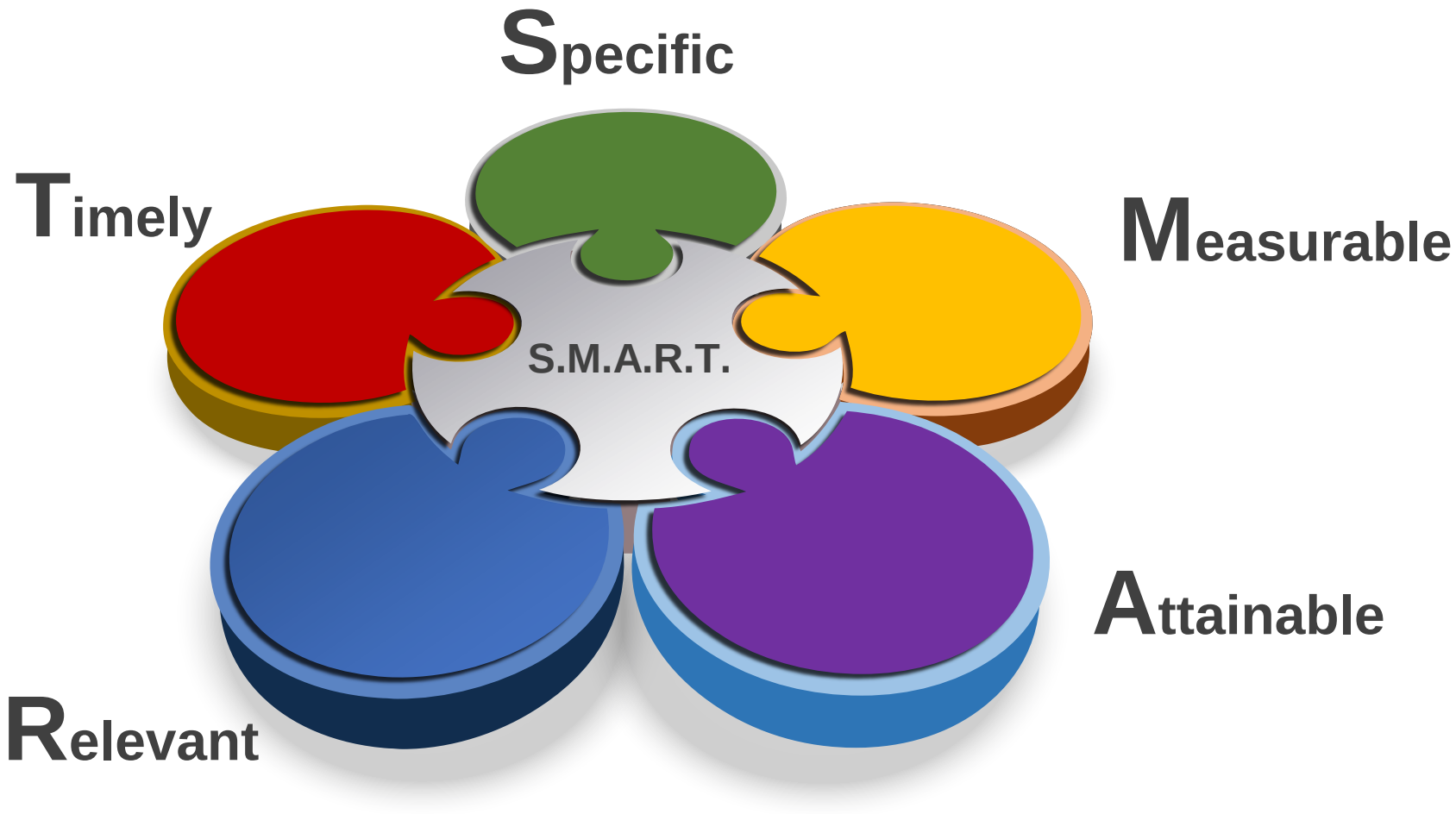
The goal of the project is **to analyze archived click-stream data** to determine the website changes that will most efficiently increase revenues by 15% on a month-by-month basis compared to the same month last year.

Phrasing the Project Goal using **SMART** method:



The goal of the project is to**, by the end of two months,** analyze archived click-stream data to determine the website changes that will most efficiently increase revenues by 15% on a month-by-month basis compared to the same month last year.

Phrasing the Project Goal using **SMART** method:



Phrasing the Project Goal using **SMART** method:

Here's a more complete version, that might have two parts:

First in 3 months, install a system that will collect and store click-stream data in a cloud-base relational database.

Then, the second part could be. By the end of 2 months, after the system has been installed, this data will be analyzed to determine the website changes that will most efficiently increase revenues by 15% on a month-by-month basis compared to the same month last year.

How to Design a Data Analytics Project?

Structured

Pyramid

Analysis

Plan

S
P
A
P

Question: What's a variable?

Answer: A variable is an object, event, idea, feeling, time period, or any other type of category you are trying to measure. There are two types of variables-independent and dependent.

S
P
A
P

Question: What's an independent variable?

Answer: It is a variable that stands alone and isn't changed by the other variables you are trying to measure.
For example, someone's age might be an independent variable.
In fact, when you are looking for some kind of relationship between variables you are trying to see if the independent variable causes some kind of change in the other variables, or dependent variables.

S
P
A
P

Question: What's a dependent variable?

Answer: It is something that depends on other factors. For example, a test score could be a dependent variable because it could change depending on several factors such as how much you studied, how much sleep you got the night before you took the test, or even how hungry you were when you took it.

Usually when you are looking for a relationship between two things you are trying to find out what makes the dependent variable change the way it does.

S
P
A
P

Some Examples:

1. You are interested in whether a higher minimum wage impacts employment rates.
2. A scientist studies the impact of a drug or age or smoking on cancer.
3. Educators are interested in whether participating in after-school math tutoring can increase scores on standardized math exams.

Dependent var
Independent Var

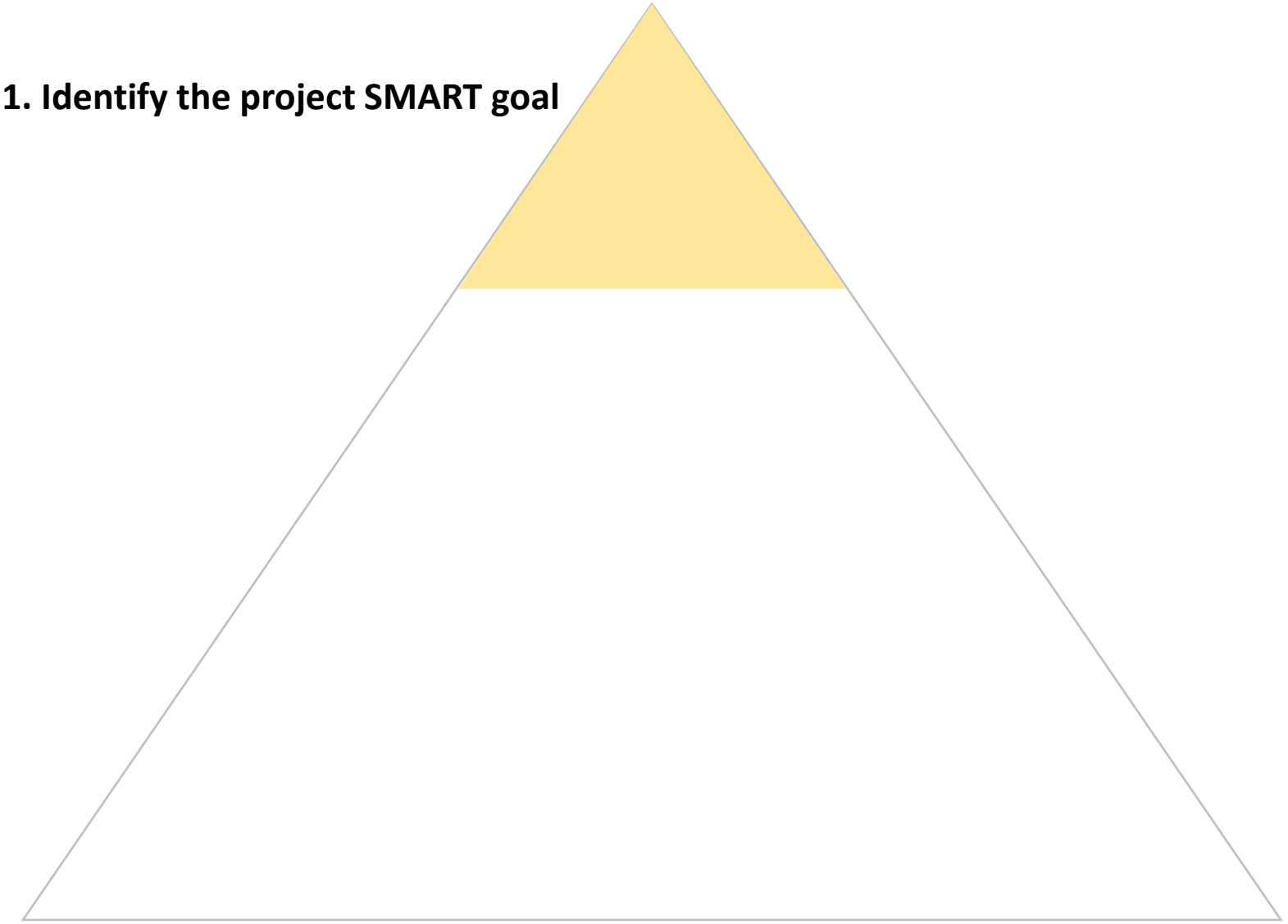
S

P

A

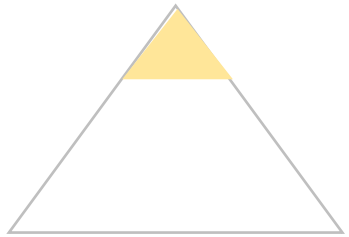
P

1. Identify the project SMART goal



S
P
A
P

1. Identify the project SMART goal



In 2 months, analyse archived click-stream data to determine the website changes that will most efficiently increase revenues by 15% on a month-to-month basis compared to the same month last year.

S

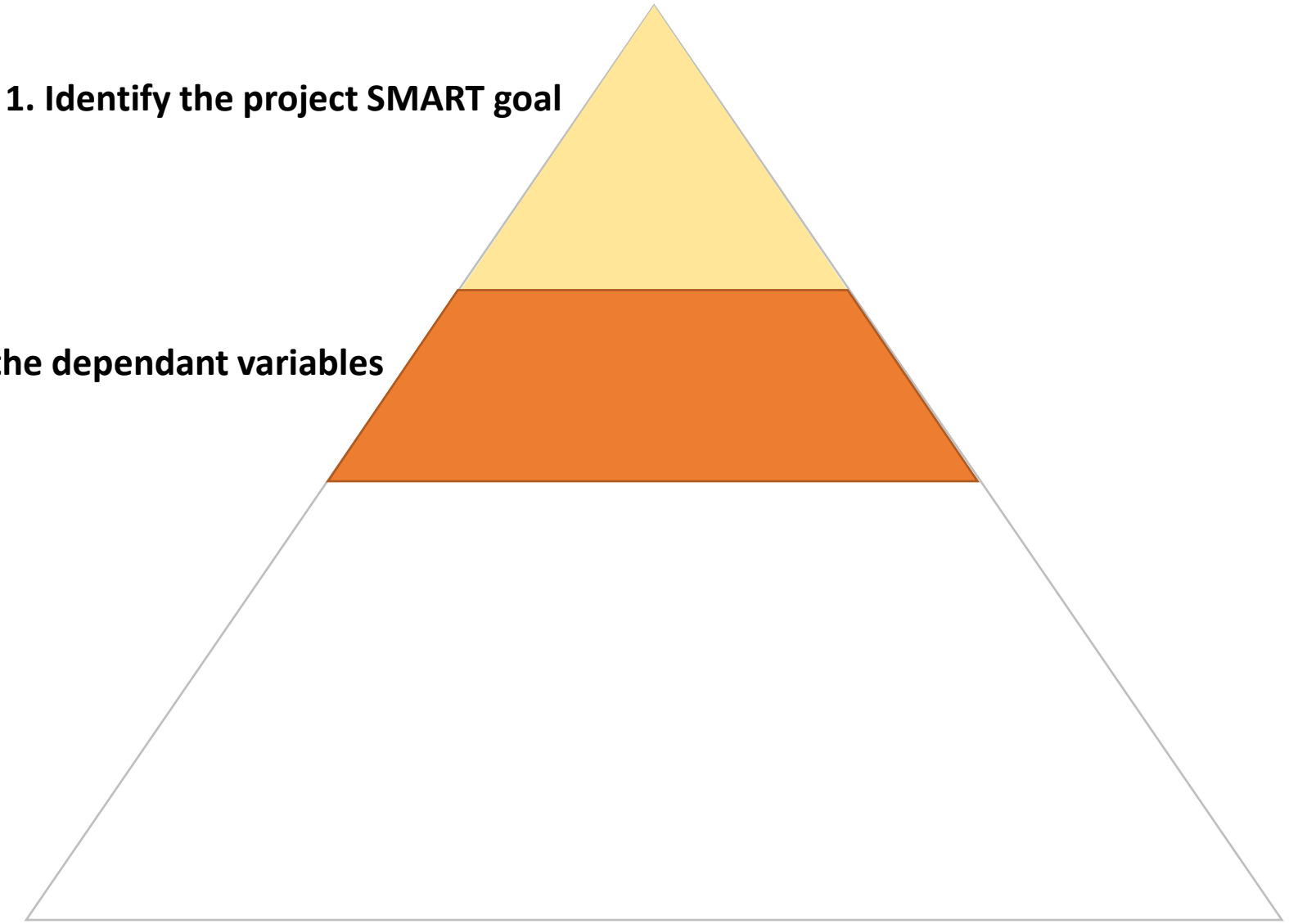
P

A

P

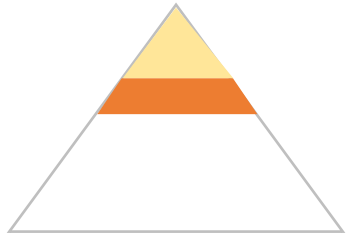
1. Identify the project SMART goal

2. Define the dependant variables



S
P
A
P

2. Define the dependant variables



DV1: Total $ spent per transaction

In clickstream database, “total spent” field aggregated by SUM over each transaction ID

DV2: Total $ spent per month

In clickstream database, “total spent” field aggregated by SUM over date (month)

DV3: Total $ spent per customer

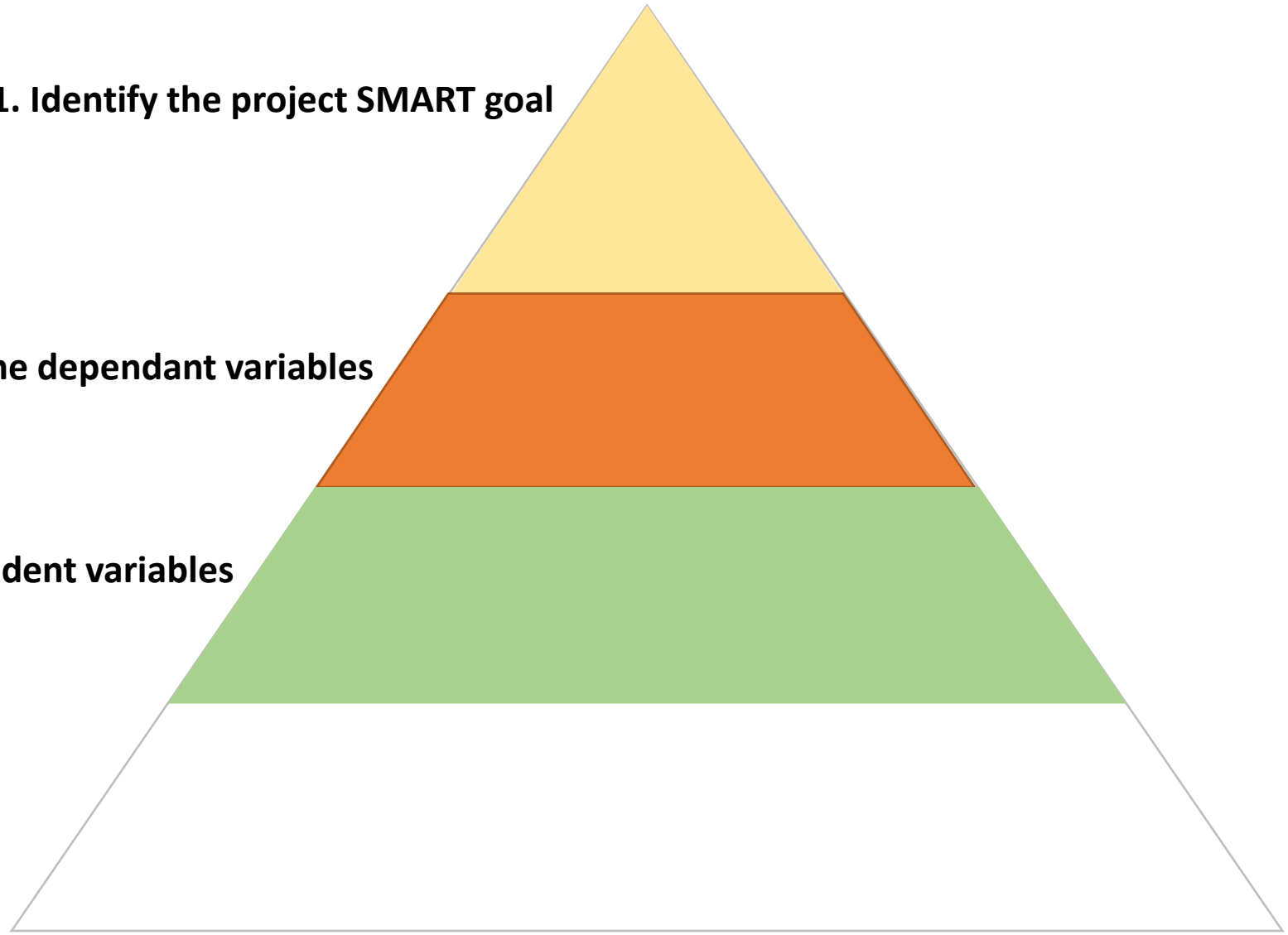
In clickstream database, “total spent” field aggregated by SUM over each customer ID

S
P
A
P

1. Identify the project SMART goal

2. Define the dependant variables

3. Define independent variables



S
P
A
P

3. Define independent variables

Do specific demographics contribute to revenue?

- Age?
- Gender?
- Income?

Do specific behaviors contribute to revenue?

- Longer time on site?
- More visitis to site?

Do specific marketing strategies contribute to revenue?

- Promotional emails
- Facebook ads?
- Tweets?

S
P
A
P

3. Define independent variables

Do specific demographics contribute to revenue?

- Age?
- Gender?
- Income?

Do specific marketing strategies contribute to revenue?

- Promotional emails
- Facebook ads?
- Tweets?

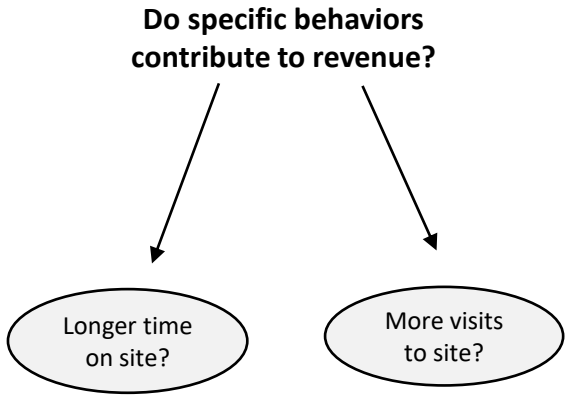
Do specific behaviours contribute to revenue?

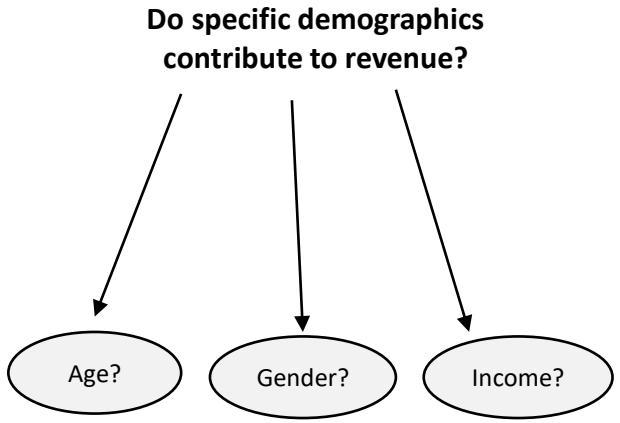
- Longer time on site?
- More visitis to site?

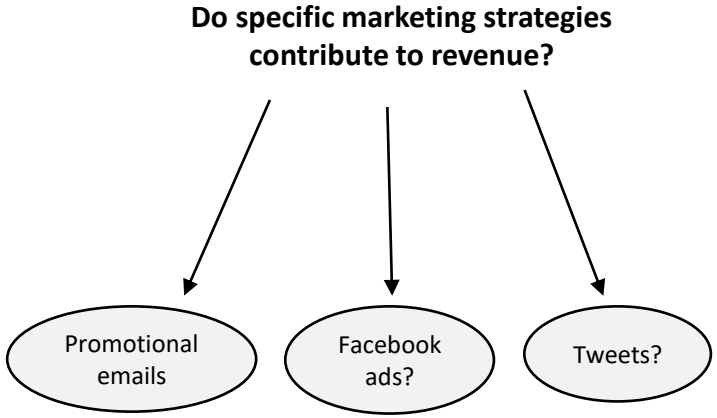
S
P
A
P

3. Define independent variables

Once you know what data variables you are going to use for each category or subcategory, the next step is to assign each category or subcategory a priority based on **who suggested them**, **how much impact you suspect they could have, and how feasible you suspect they will be to assess.**







S

P

A

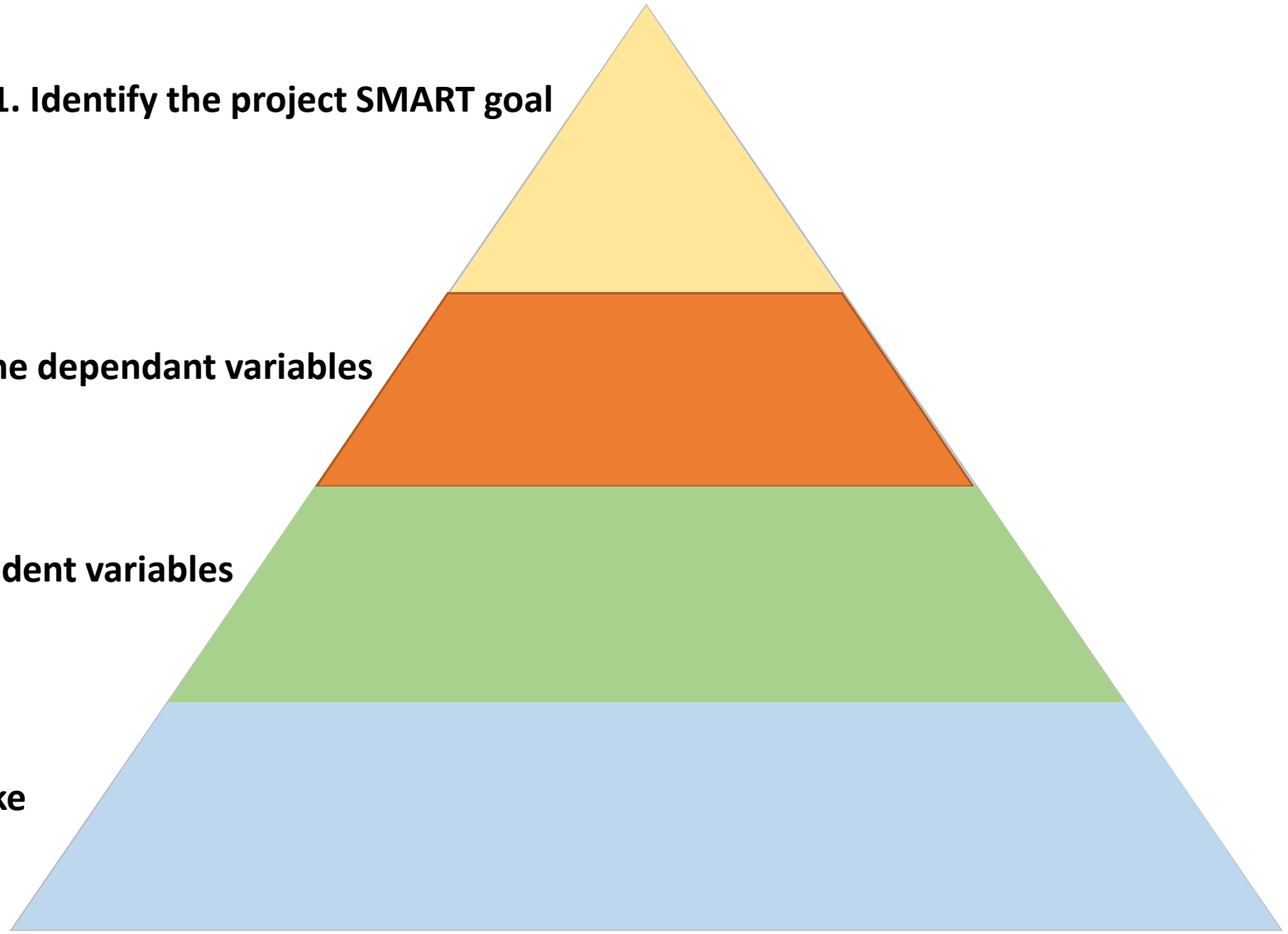
P

1. Identify the project SMART goal

2. Define the dependant variables

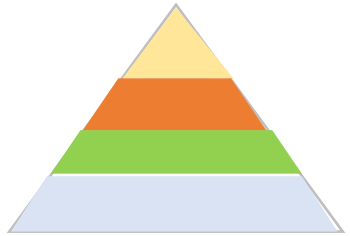
3. Define independent variables

4. Specify analyses to run & graphs to make



S
P
A
P

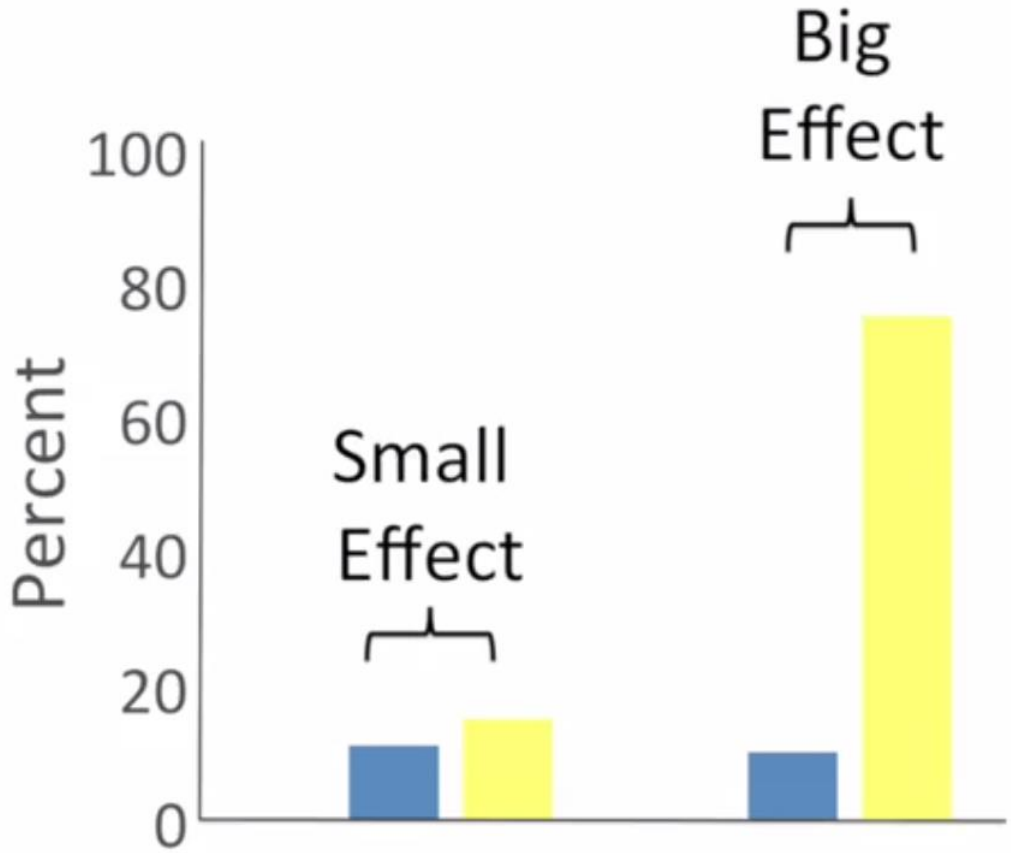
4. Specify analyses to run & graphs to make



The final step of an SPAP is to actually implement the plan and illustrate the most important issues or independent variables that will help you solve your business problem.

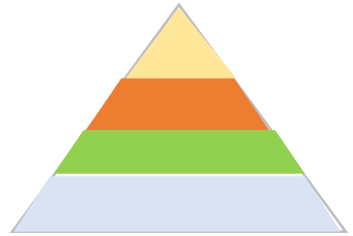
But How?

Using Data Visualizing Software



S
P
A
P

4. Specify analyses to run & graphs to make



Do specific behaviors contribute to revenue?

- Longer time on site?
- More visits to site?

Do specific demographics contribute to revenue?

- Age?
- Gender?
- Income?

Do specific marketing strategies contribute to revenue?

- Promotional emails
- Facebook ads?
- Tweets?

Line Chart:

Visit on x-axis
$ Spend on y-axis

Bar Chart:

Age on x-axis
$ Spend on y-axis

Bar Chart:

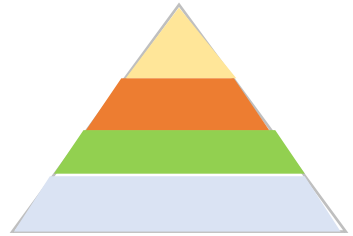
Income on x-axis
$ Spend on y-axis

Bar Chart:

Gender on x-axis
$ Spend on y-axis

S
P
A
P

4. Specify analyses to run & graphs to make



Do specific behaviors contribute to revenue?

- Longer time on site?
- More visitis to site?

Do specific demographics contribute to revenue?

- Age?
- Gender?
- Income?

Do specific marketing strategies contribute to revenue?

- Promotional emails
- Facebook ads?
- Tweets?

Line Chart:

Visit on x-axis
$ Spend on y-axis

Bar Chart:

Age on x-axis
$ Spend on y-axis

Bar C[illegible]:

Inc[illegible]e on x[illegible]xis
$ [illegible]end on y-a[illegible]is

Bar Chart:

Gender on x-axis
$ Spend on y-axis

Do any eye-catching effect you observe?

S

P

A

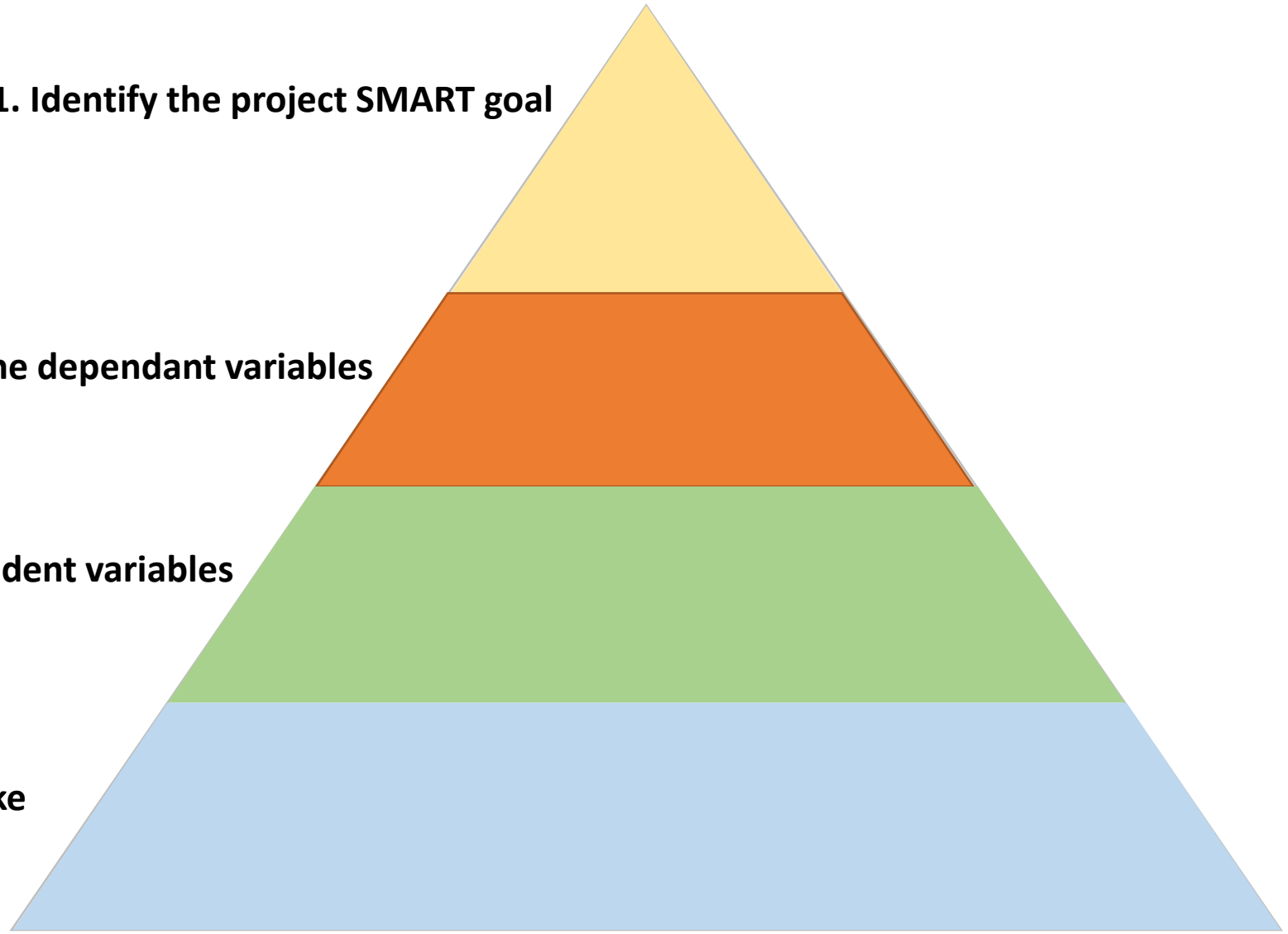
P

1. Identify the project SMART goal

2. Define the dependant variables

3. Define independent variables

4. Specify analyses to run & graphs to make



1.Installing R

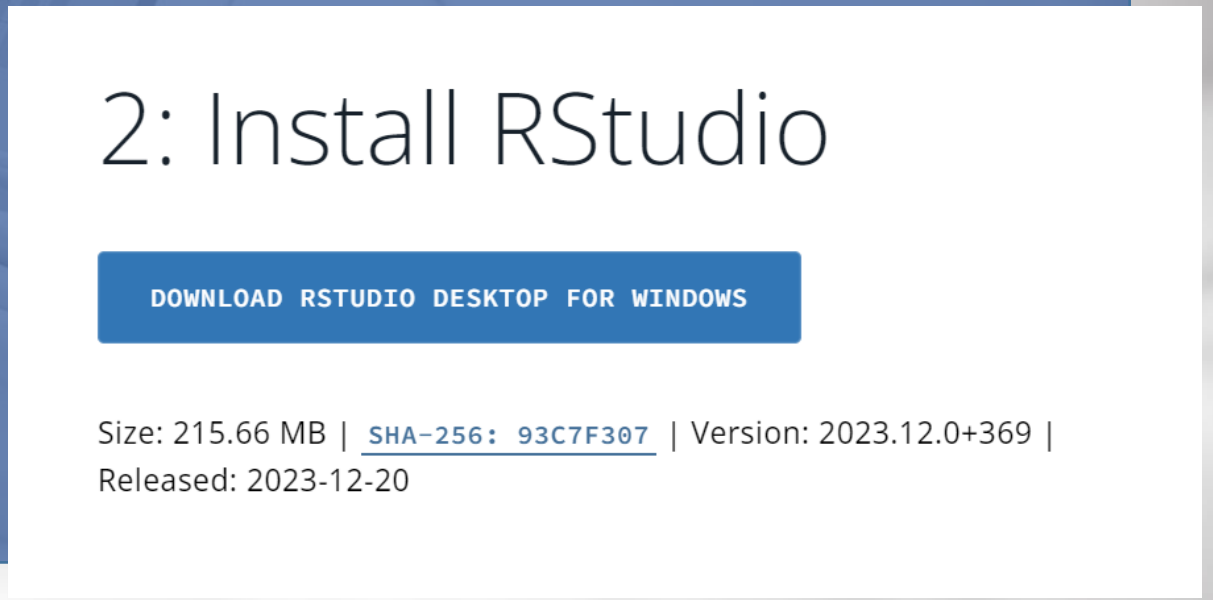
Latest version: 4.2.3 for windows

https://cran.r-project.org/

2.Installing R Studio

RStudio Desktop 2023.03.0+386

https://posit.co/downloads/



Assignment 1.1

1. Identify a DS project objective using the SMART methodology

2. Design your project using the SPAP methodology (Identify the DV and IVs)

Any questions ?

vala.ali.rohani@estsetubal.ips.pt