DATA MINING IS A HIGH ORDER CONCEPT

Creating -generating new ideas

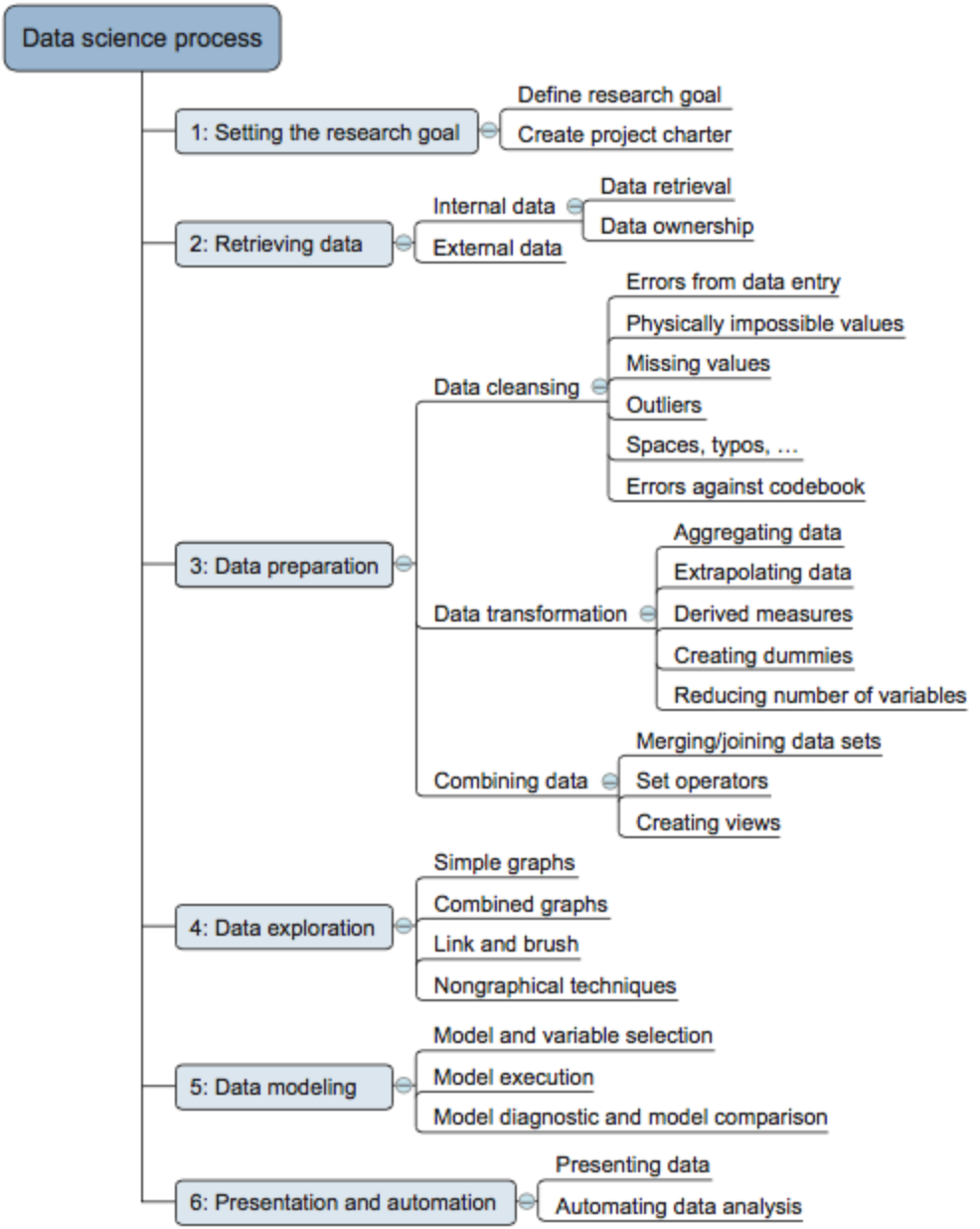
Evaluating - justifying a decision or choice

Analyzing - breaking into component parts

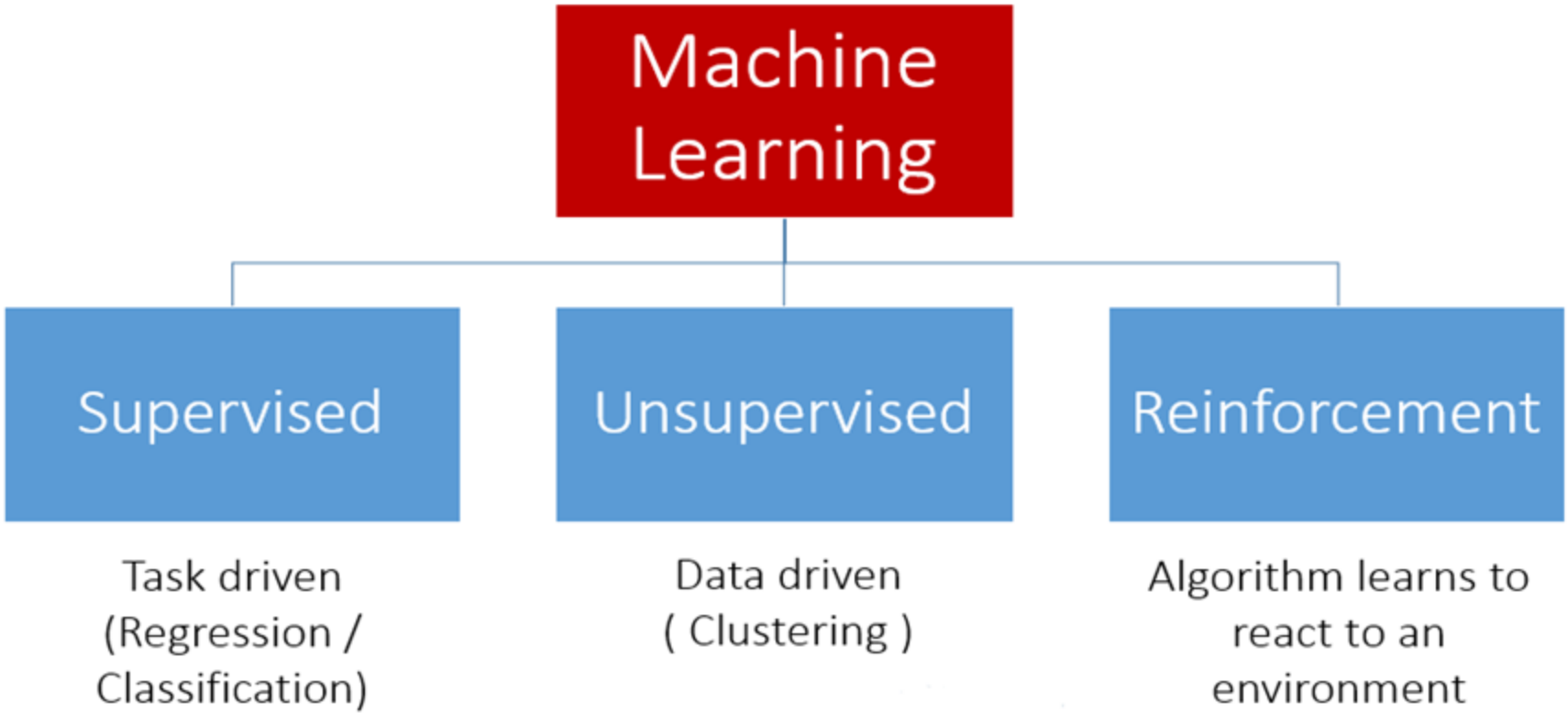
Applying - using information in a new setting

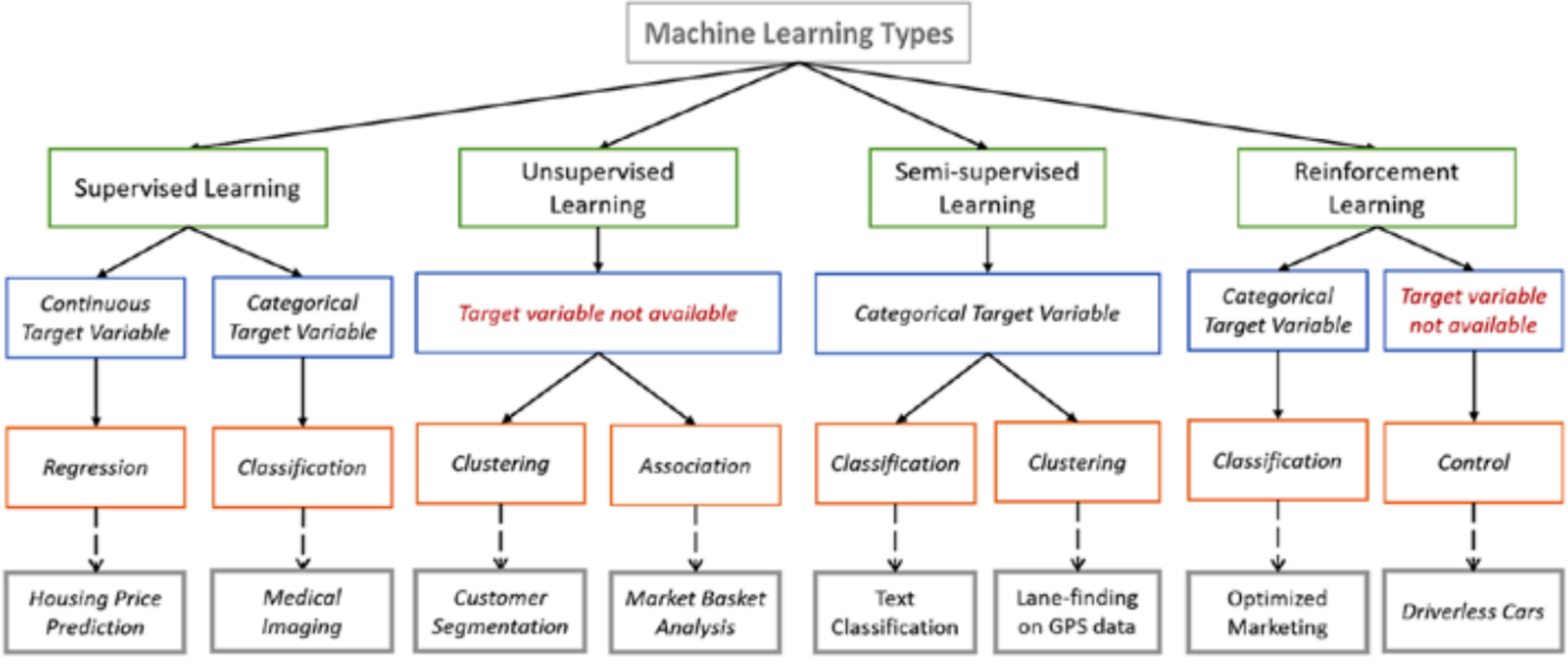
Understanding - explaining idea or concept

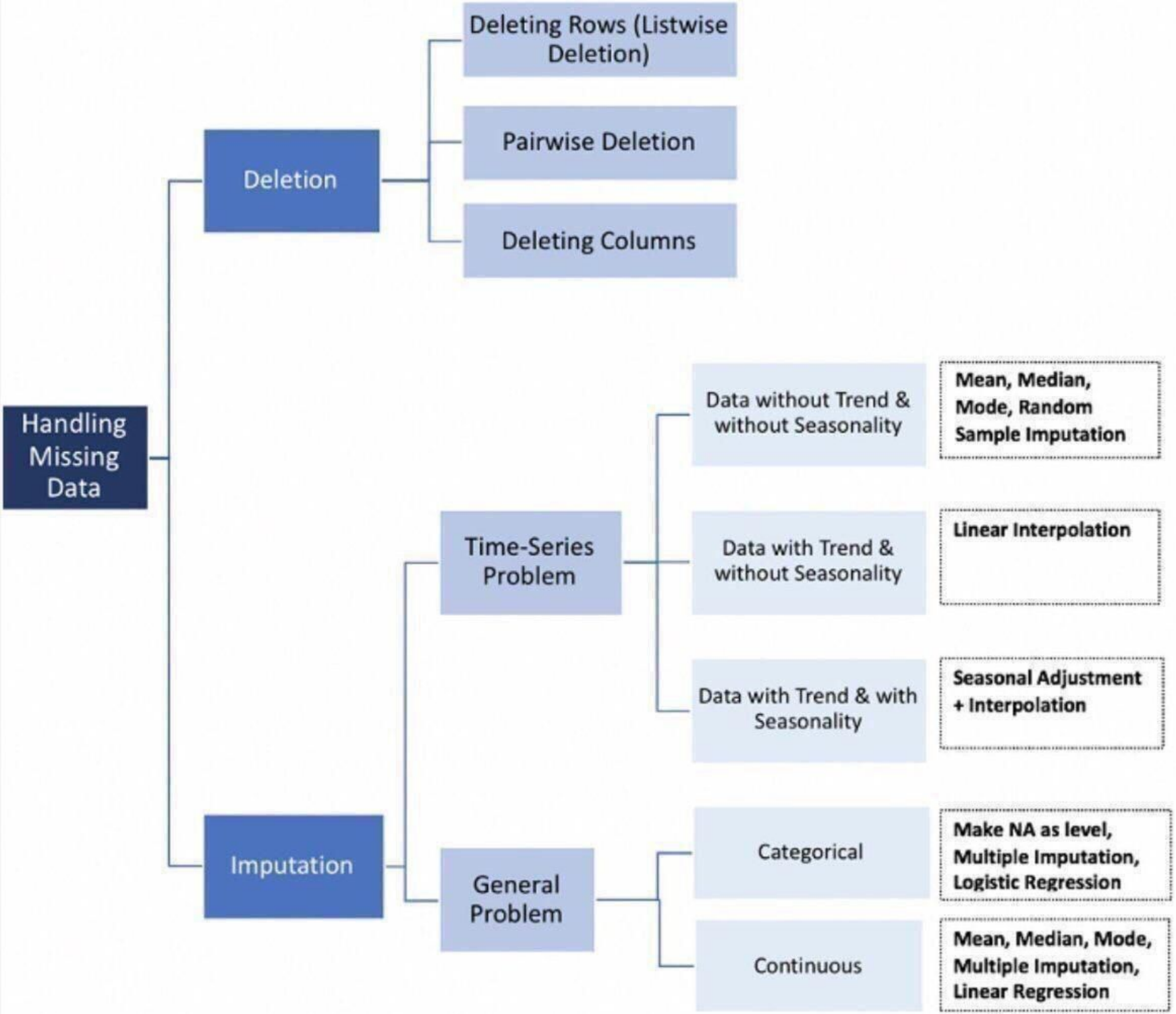
Remembering - recalling information

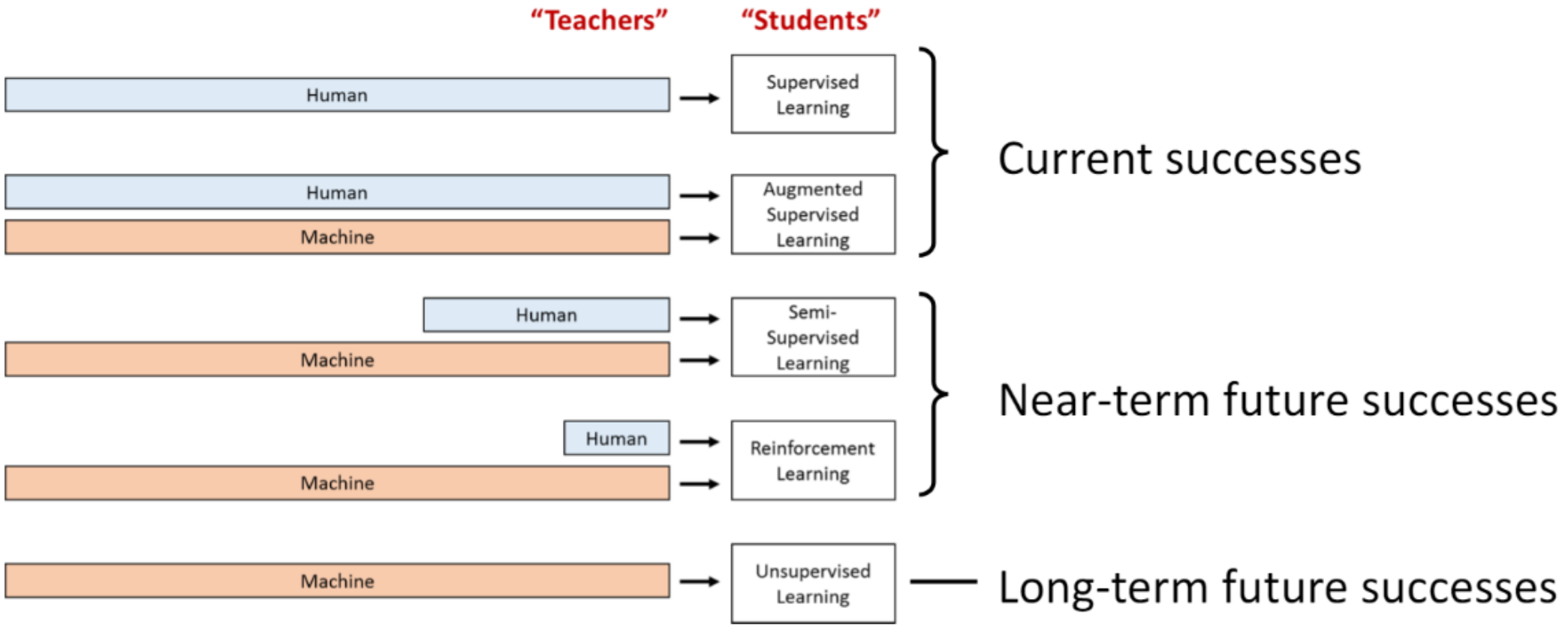

Data science process
1: Setting the research goal
Define research goal
Create project charter
2: Retrieving data
Internal data
Data retrieval
Data ownership
External data
3: Data preparation
Data cleansing
Errors from data entry
Physically impossible values
Missing values
Outliers
Spaces, typos, …
Errors against codebook
Data transformation
Aggregating data
Extrapolating data
Derived measures
Creating dummies
Reducing number of variables
Combining data
Merging/joining data sets
Set operators
Creating views
4: Data exploration
Simple graphs
Combined graphs
Link and brush
Nongraphical techniques
5: Data modeling
Model and variable selection
Model execution
Model diagnostic and model comparison
6: Presentation and automation
Presenting data
Automating data analysis

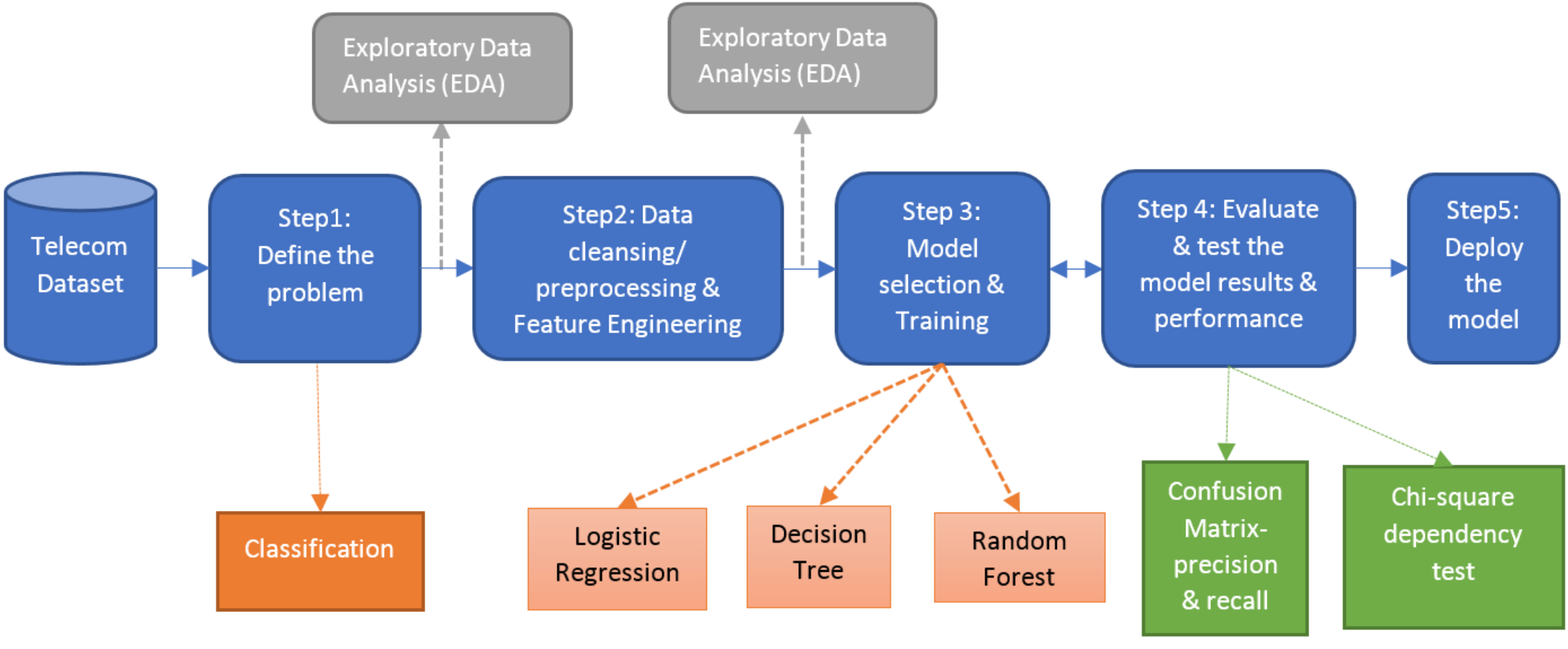
Types of Machine Learning




Machine Learning Types
Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target variable not available
Categorical Target Variable
Categorical Target Variable
Target variable not available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
Housing Price Prediction
Medical Imaging
Customer Segmentation
Market Basket Analysis
Text Classification
Lane-finding on GPS data
Optimized Marketing
Driverless Cars

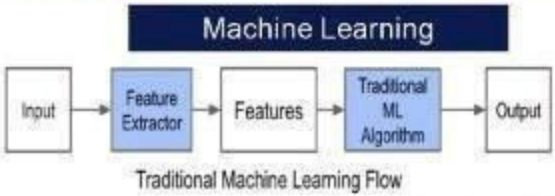

Handling Missing Data
Deletion
Deleting Rows (Listwise Deletion)
Pairwise Deletion
Deleting Columns
Imputation
Time-Series Problem
Data without Trend & without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend & without Seasonality
Linear Interpolation
Data with Trend & with Seasonality
Seasonal Adjustment + Interpolation
General Problem
Categorical
Make NA as level, Multiple Imputation, Logistic Regression
Continuous
Mean, Median, Mode, Multiple Imputation, Linear Regression

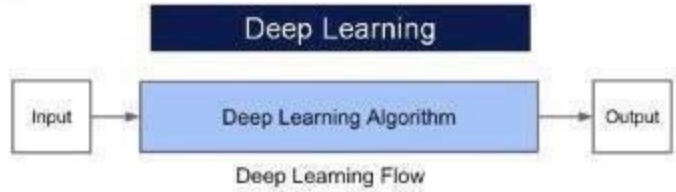

"Teachers"
"Students"
Human
Supervised Learning
Human
Machine
Augmented Supervised Learning
Current successes
Human
Machine
Semi-Supervised Learning
Human
Machine
Reinforcement Learning
Near-term future successes
Machine
Unsupervised Learning
Long-term future successes


Exploratory Data Analysis (EDA)
Exploratory Data Analysis (EDA)
Telecom Dataset
Step1: Define the problem
Step2: Data cleansing/ preprocessing & Feature Engineering
Step 3: Model selection & Training
Step 4: Evaluate & test the model results & performance
Step5: Deploy the model
Classification
Logistic Regression
Decision Tree
Random Forest
Confusion Matrix- precision & recall
Chi-square dependency test

Machine Learning vs Deep Learning

Process





Compute

Can work with commodity hardware (including consumer laptop)

Needs High end machines with GPU

Feature Extraction

Feature extraction with business context/domain has to be done

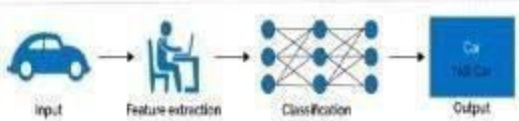
Automatically learns features layer by layer

Development time

Relatively less training time

Long training time as DL learns all the diverse set of parameters

Schematic Example



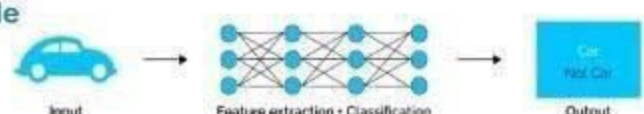
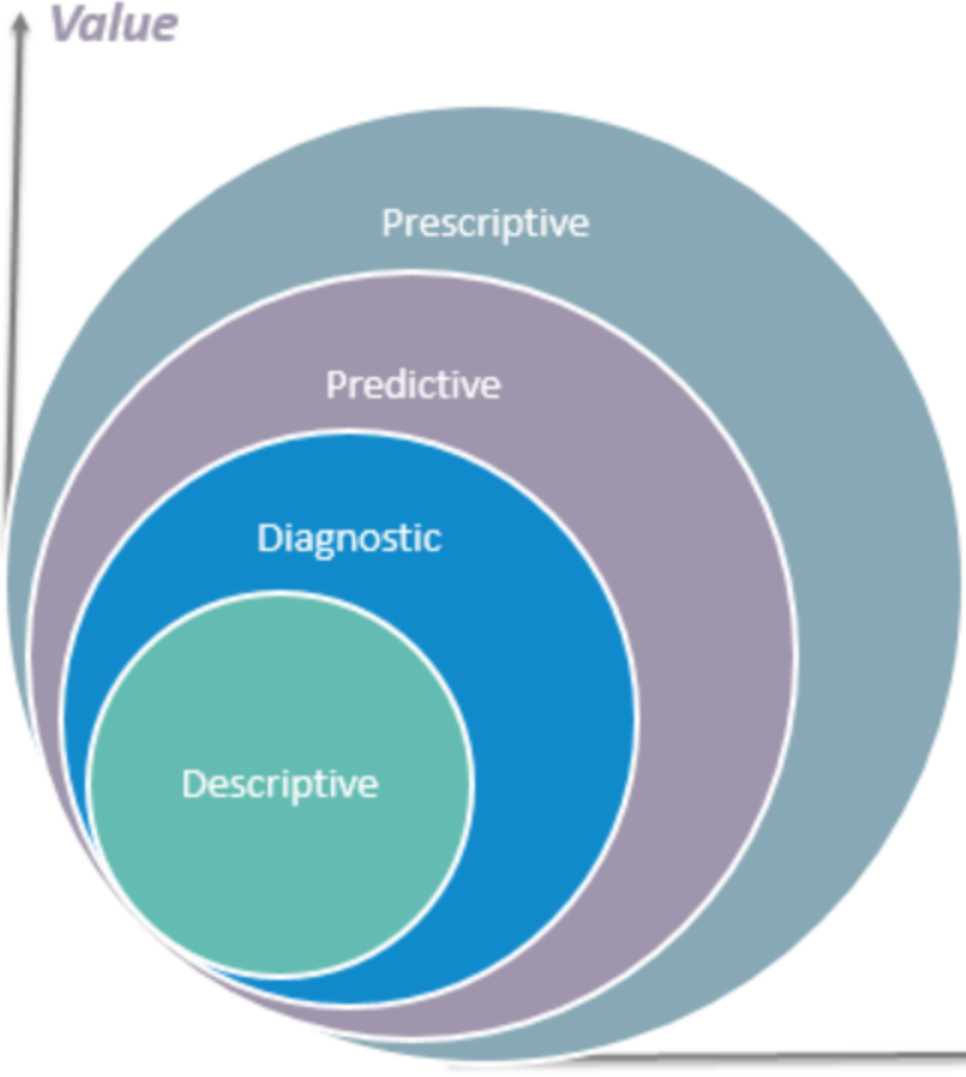


Table 1: The three components of learning algorithms.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

4 types of Data Analytics



What is the data telling you?

Descriptive: *What's happening in my business?*

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: *Why is it happening?*

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: *What's likely to happen?*

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: *What do I need to do?*

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa

www.principa.co.za

10mo

6 Steps Approach to **#DataScience** for **#businesses** & practitioners

1. Set Research Goal - what are you going to research, how does it benefit the business, define a problem statement, what kind of data & resources are needed, a timeline of deliverables.

2. Data Collection - how to retrieve the data, check the quality & accessibility, data comes from many sources and formats, can be **#Excel** spreadsheets, **#MySQL** or other databases, **#Bigdata**

3. Prepare the Data - most crucial step in Data Science and **#MachineLearning** because data collection is prone to errors, it needs enhancement.

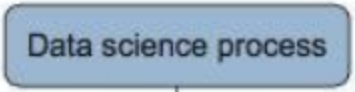
a) Data Cleansing - remove inconsistencies
b) Data Integration - enrich your data by combining sources
c) Data Transformation - ensure data is suitable for the model

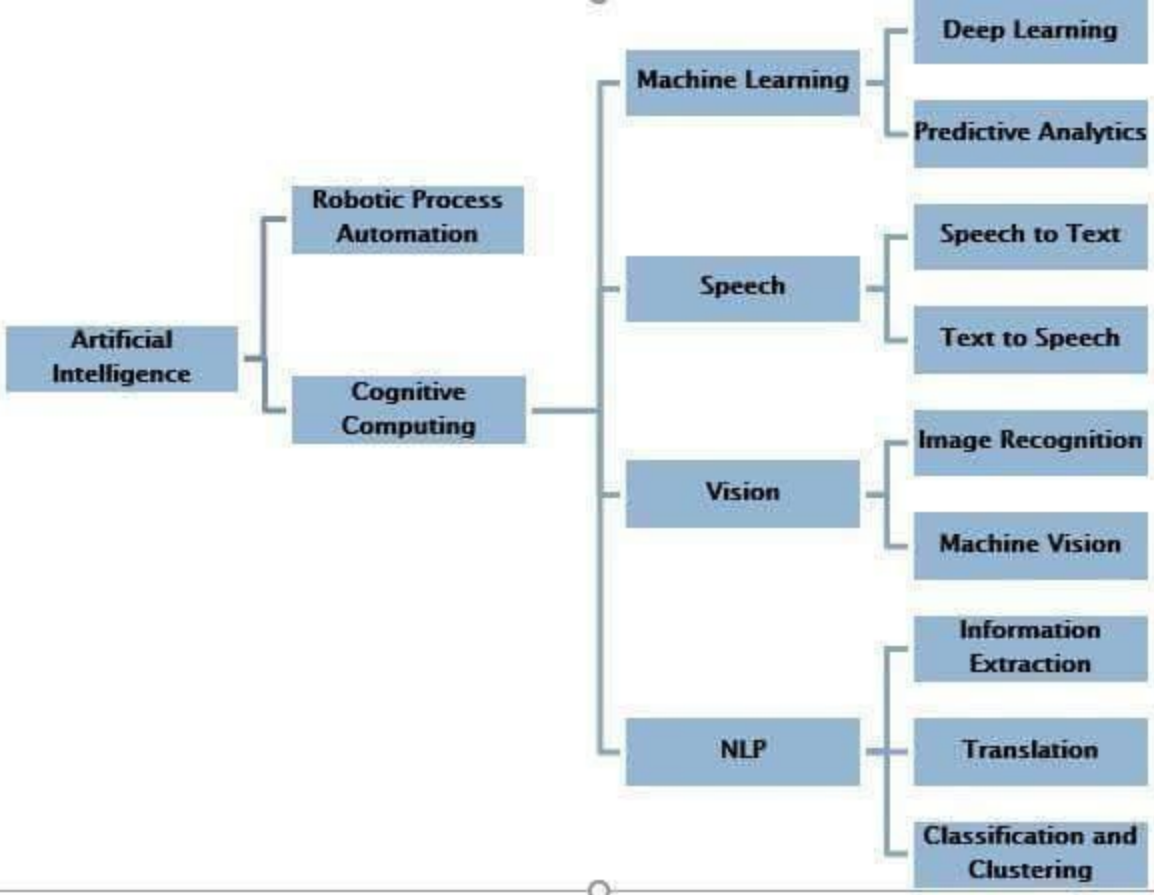
4. Data Exploration - how variables interact, how data is distributed, identify outliers, gain insights. Use **#statistics**, **#DataVisualization**

5. Data Modeling - build a model, use domain expertise, use insights to answer research questions. Use stats, ML or Operation Research. Re-iterate to find a model that fits the best.

6. Presentation & Automation - share end results as an **#analysis**, report or an automated model that can be applied to other business problems.

#Analytics #ArtificialIntelligence #Automation #IoT




Artificial Intelligence
Robotic Process Automation
Cognitive Computing
Machine Learning
Deep Learning
Predictive Analytics
Speech
Speech to Text
Text to Speech
Vision
Image Recognition
Machine Vision
NLP
Information Extraction
Translation
Classification and Clustering