MACHINE LEARNING

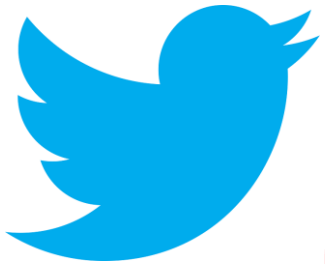
BY,
SANDESH G

CONTENTS

Vs of BIG DATA

VOLUME

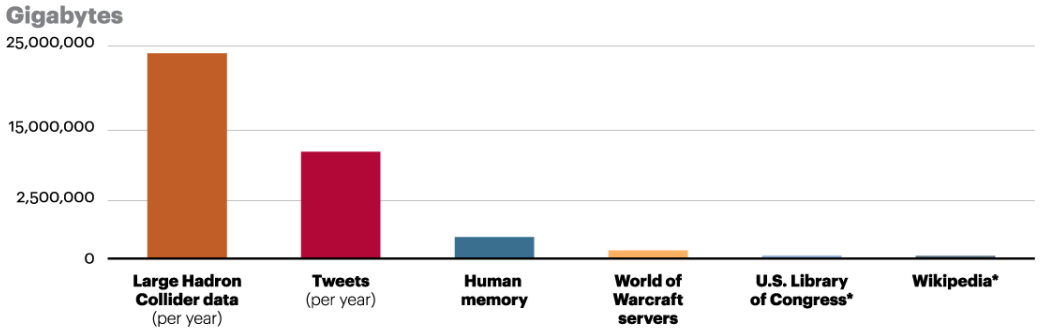
VARIETY



VELOCITY

Figure

The LHC collects about 25 million gigabytes of data per year



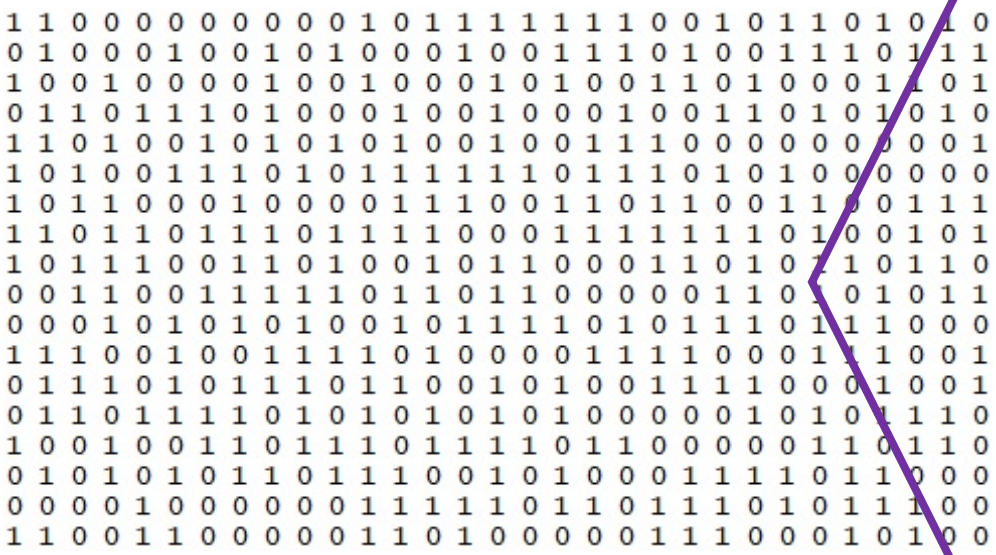
*Binary data

Note: All numbers are approximate.

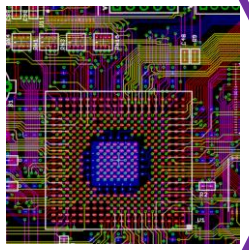
Source: "Particle Physics Tames Big Data," Leah Hesla, *Symmetry*, 1 August 2012

CONCEPT

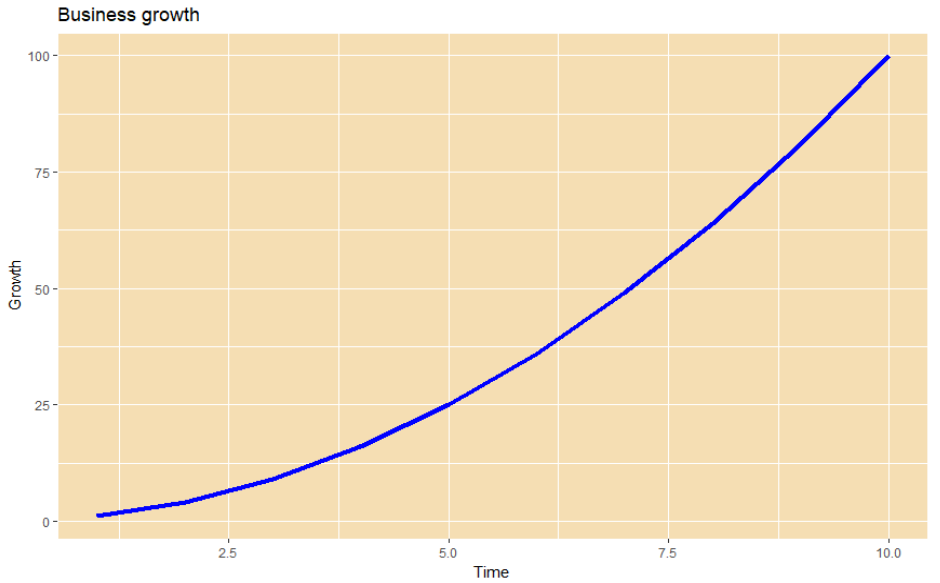
DATA



MACHINE



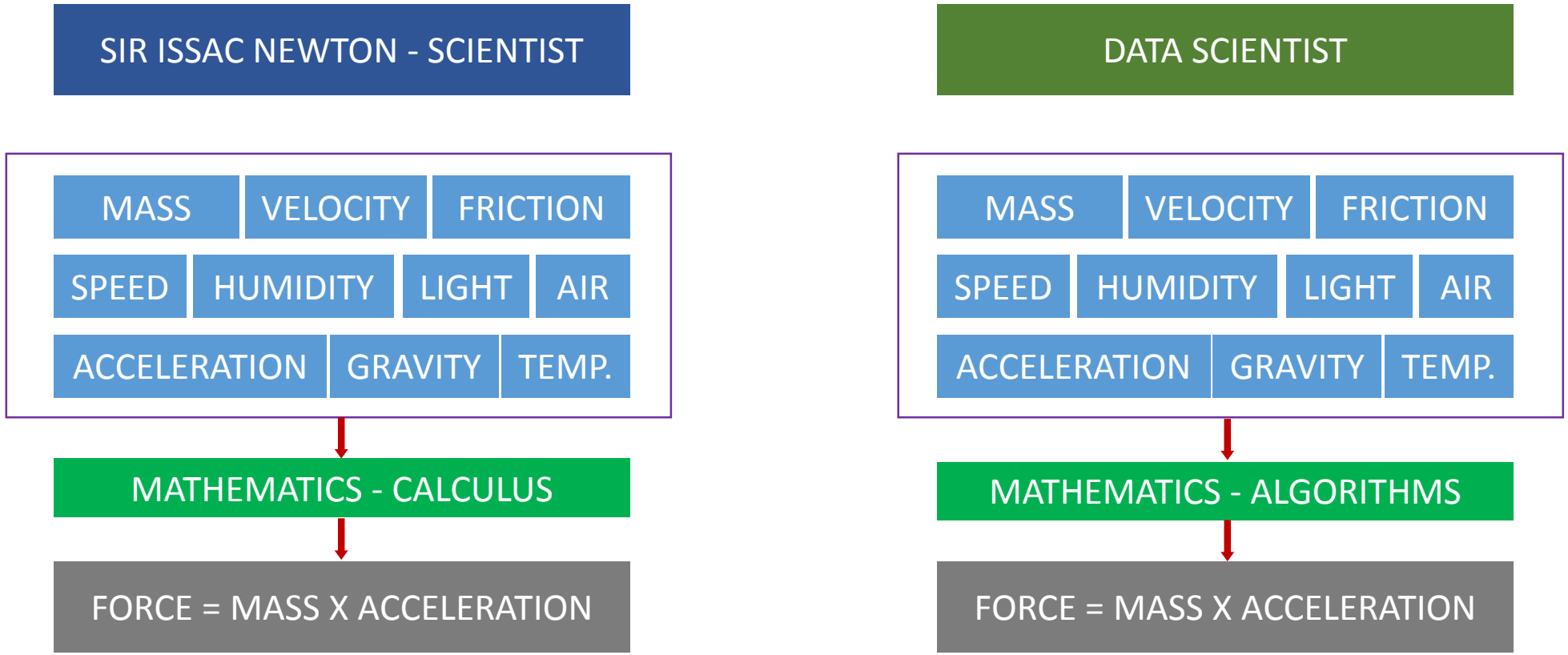
BUSINESS



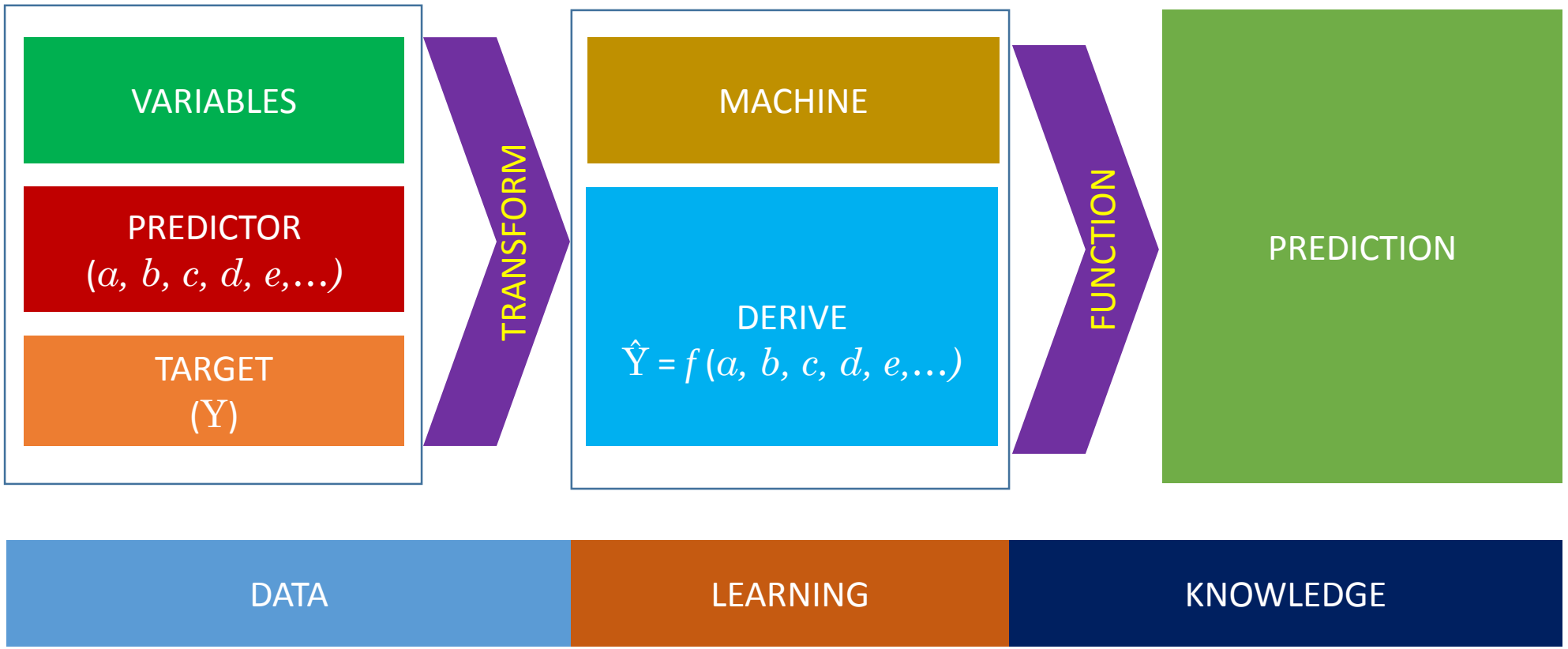
DATA → KNOWLEDGE

WIKIPEDIA: Machine learning is a field of computer science that gives computer systems the ability to "learn" (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed.

CONCEPT

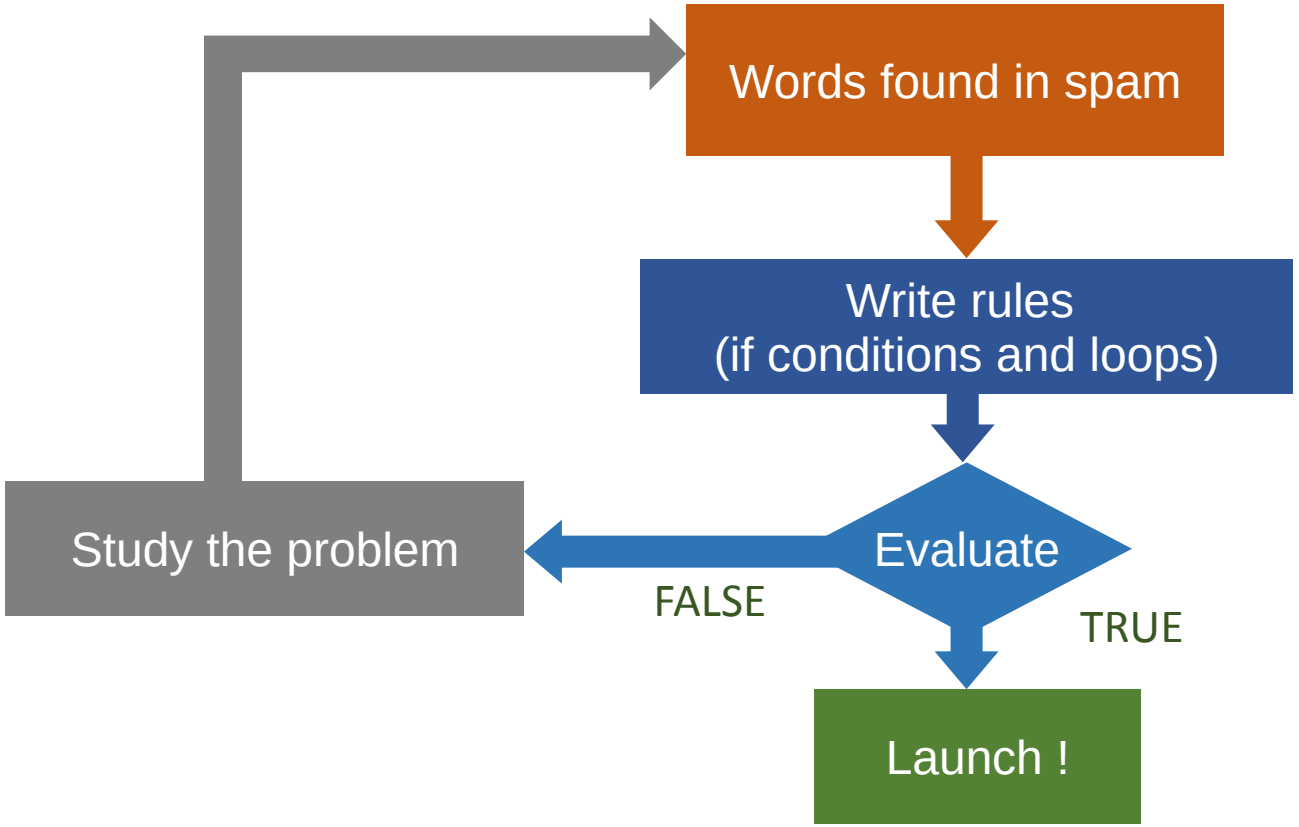

SIR ISSAC NEWTON - SCIENTIST
MASS
VELOCITY
FRICTION
SPEED
HUMIDITY
LIGHT
AIR
ACCELERATION
GRAVITY
TEMP.
MATHEMATICS - CALCULUS
FORCE = MASS X ACCELERATION
DATA SCIENTIST
MASS
VELOCITY
FRICTION
SPEED
HUMIDITY
LIGHT
AIR
ACCELERATION
GRAVITY
TEMP.
MATHEMATICS - ALGORITHMS
FORCE = MASS X ACCELERATION

CONCEPT


VARIABLES
PREDICTOR
(a, b, c, d, e,…)
TARGET
(Y)
TRANSFORM
MACHINE
DERIVE
Ŷ = f (a, b, c, d, e,…)
FUNCTION
PREDICTION
DATA
LEARNING
KNOWLEDGE

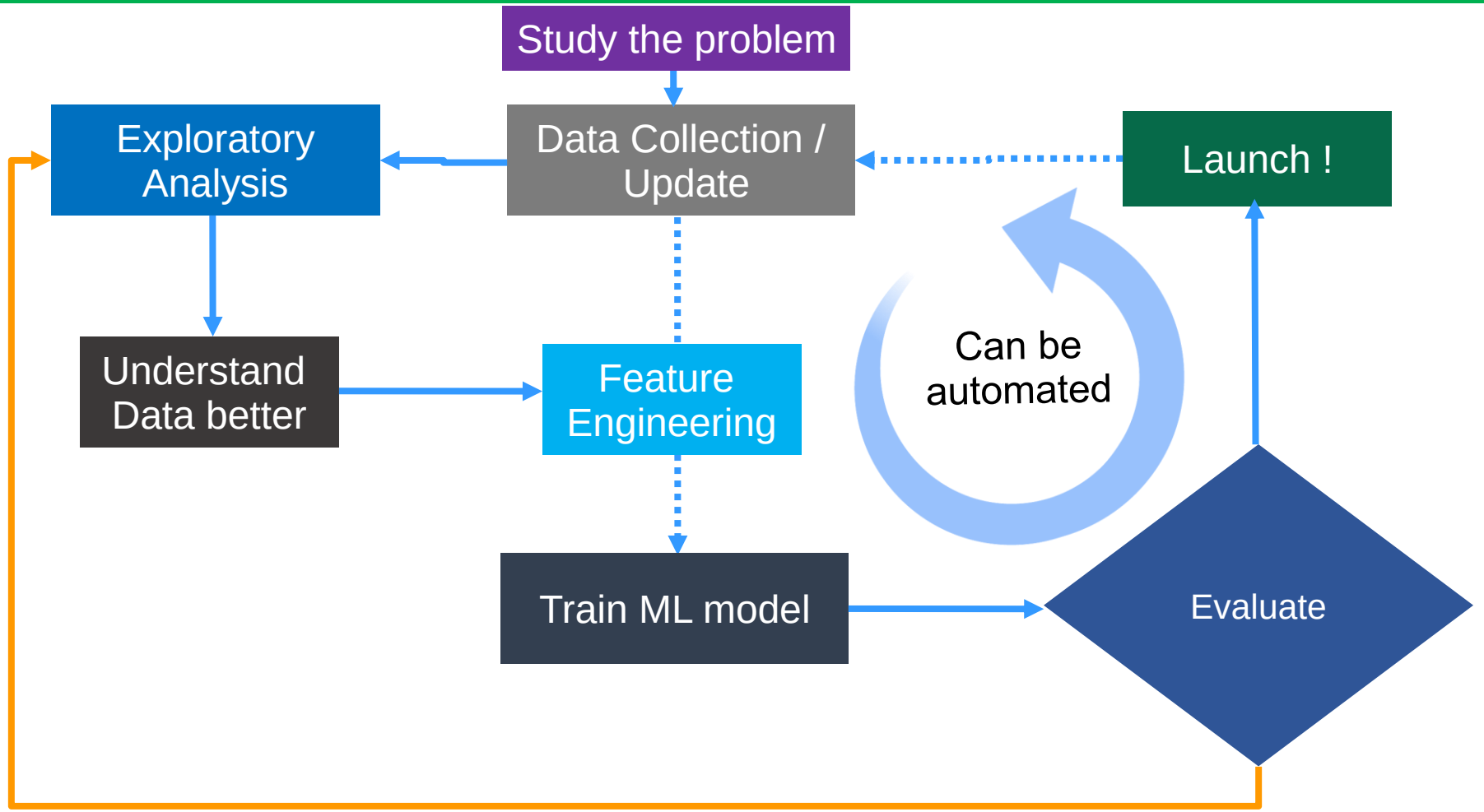
PROBLEM SOLVING APPROACHES

TRADITIONAL



PROBLEM SOLVING APPROACHES

MACHINE LEARNING



CLASSIFICATION

SUPERVISED	UNSUPERVISED	REINFORCED
REGRESSION	CLUSTERING	BRUTE FORCE
CLASSIFICATION	ASSOCIATION RULE MINING	MONTE CARLO METHOD
	DIMENSIONALITY REDUCTION	

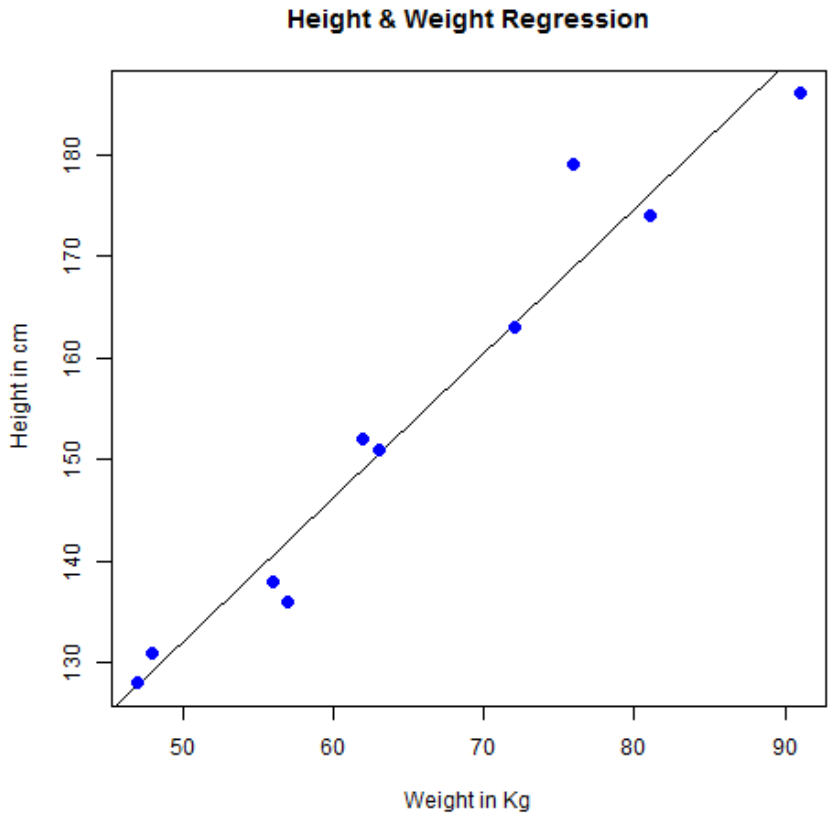
ALGORITHMS

REGRESSION	CLASSIFICATION	CLUSTERING
LINEAR REGRESSION	K NEAREST NEIGHBOURS	K MEANS CLUSTERING
NEURAL NETWORKS	NEURAL NETWORKS	HIERARCHICAL CLUSTERING
DECISION TREE	LOGISTIC REGRESSION	
RANDOM FOREST	SUPPORT VECTOR MACHINES	
	DECISION TREE	
	RANDOM FOREST	

ALGORITHMS

REGRESSION

LINEAR REGRESSION



Y = AX + B

MEAN

HEIGHT = NUMBER X WEIGHT + BIAS

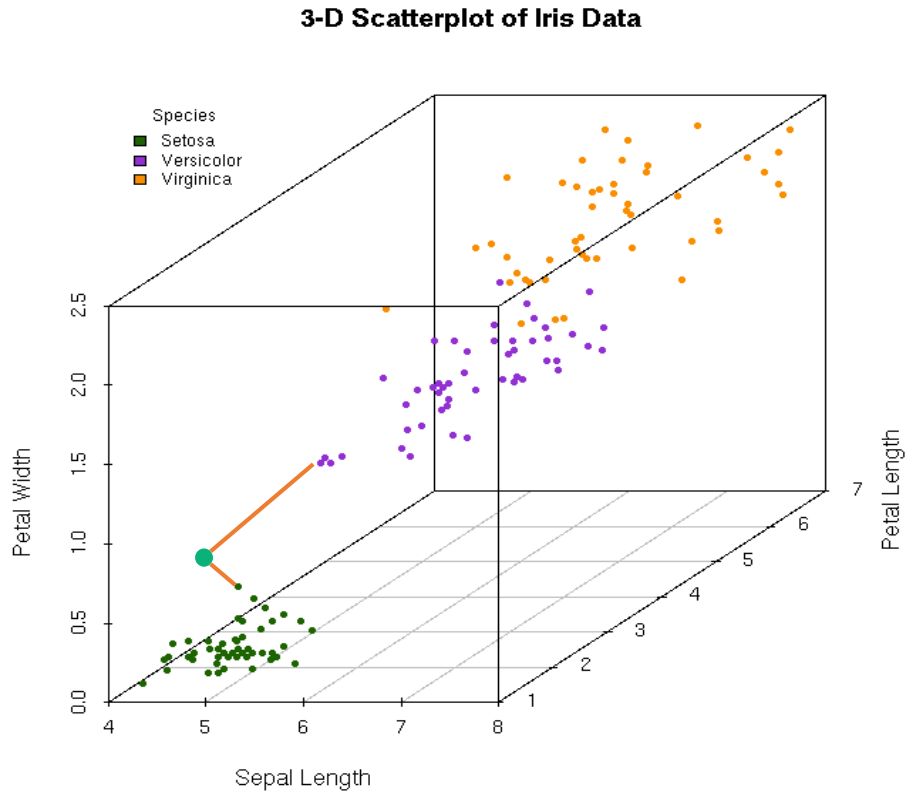
MEDIAN

```
x <- seq(-5,5,0.1)
y <- 2 * x + 3
plot(x,y,type = “l”, col = “blue”)
grid()
```

ALGORITHMS

CLASSIFICATION

K - NEAREST NEIGHBOURS



DISTANCE MEASURES

EUCLIDEAN

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

MANHATTAN

$$d_1(\mathbf{p},\mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^{n} |p_i - q_i|$$

ALGORITHMS

CLUSTERING

K - MEANS

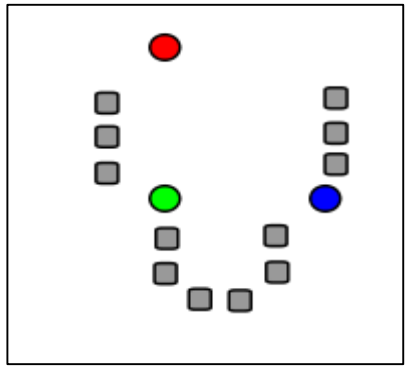
RANDOM INITIALIZATION OF CENTROIDS

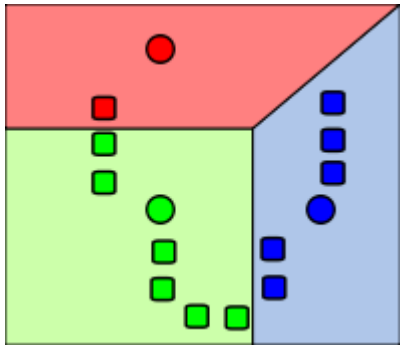
CALCULATE DISTANCES TO NEAREST POINTS

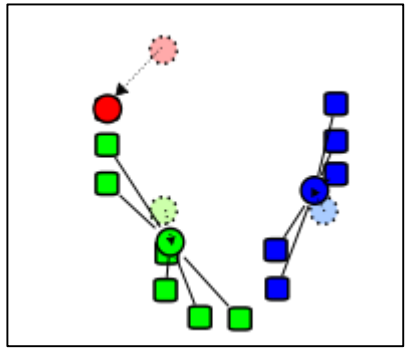
CALCULATE MEAN OF DISTANCES

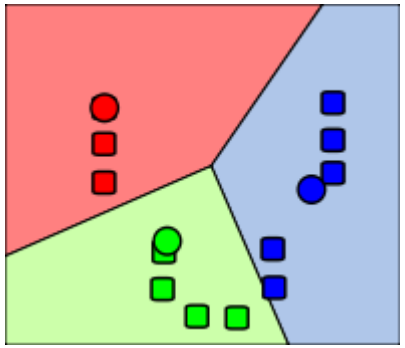
SHIFT TO AVERAGE

BACK TO STEP 2





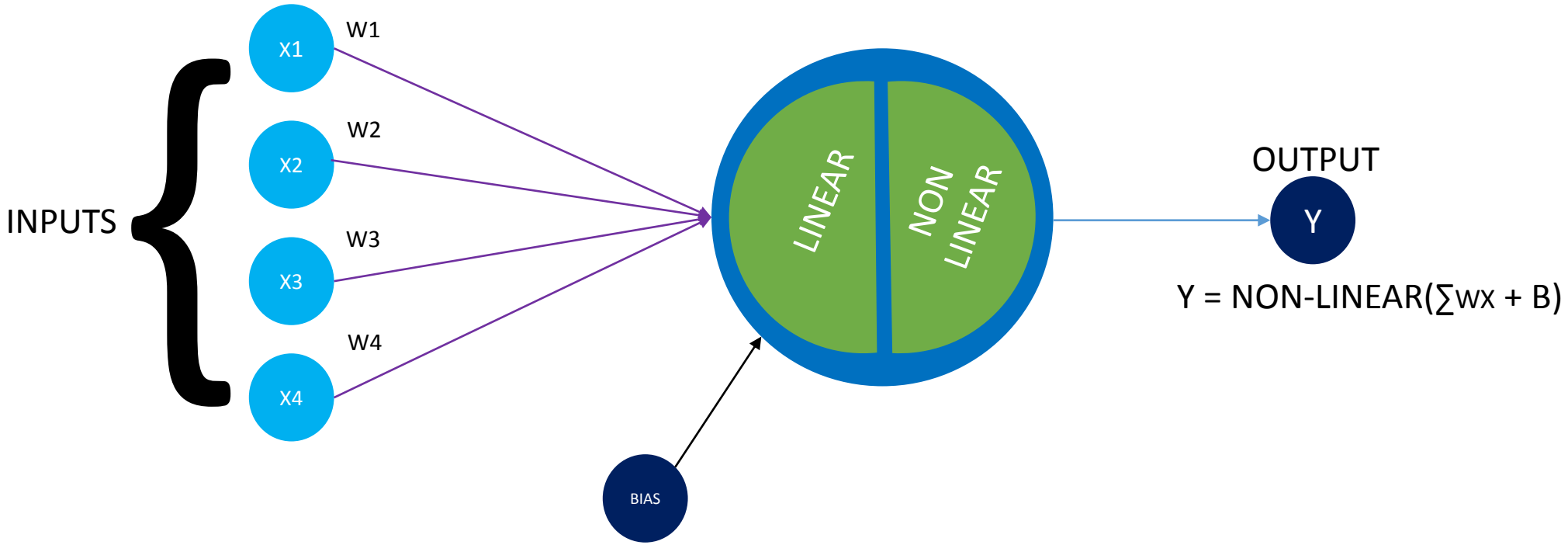




ALGORITHMS

NEURAL NETWORKS

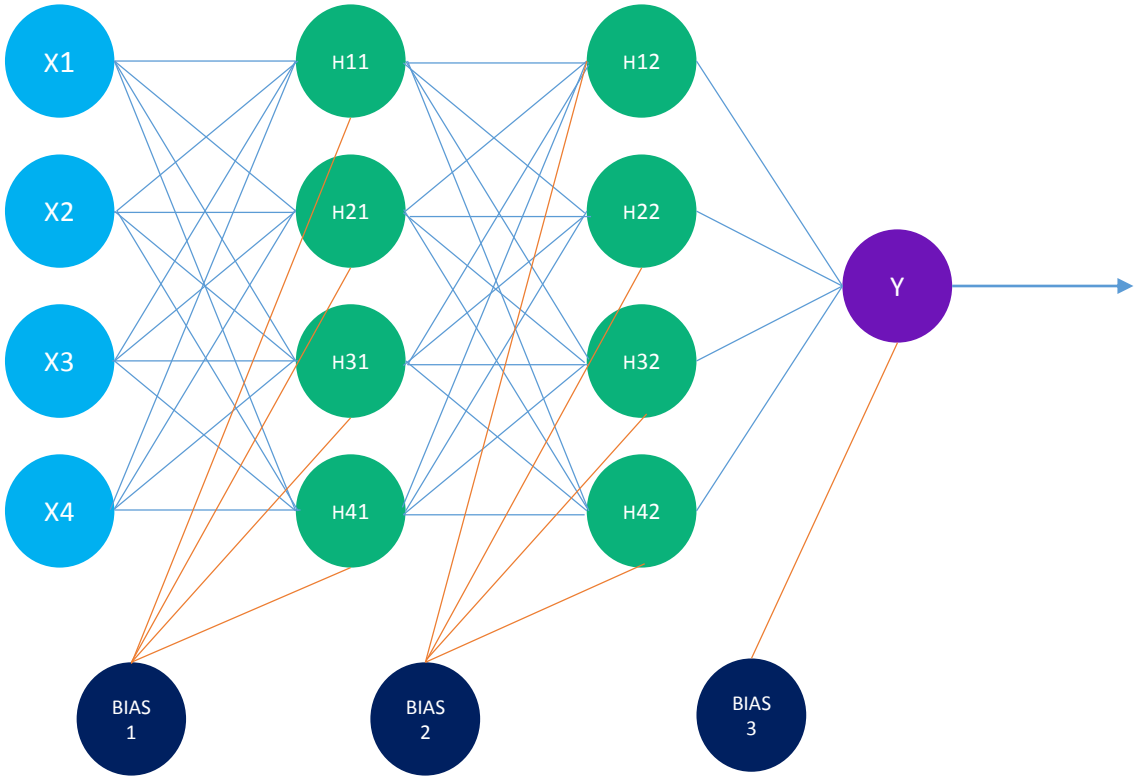
PERCEPTRON



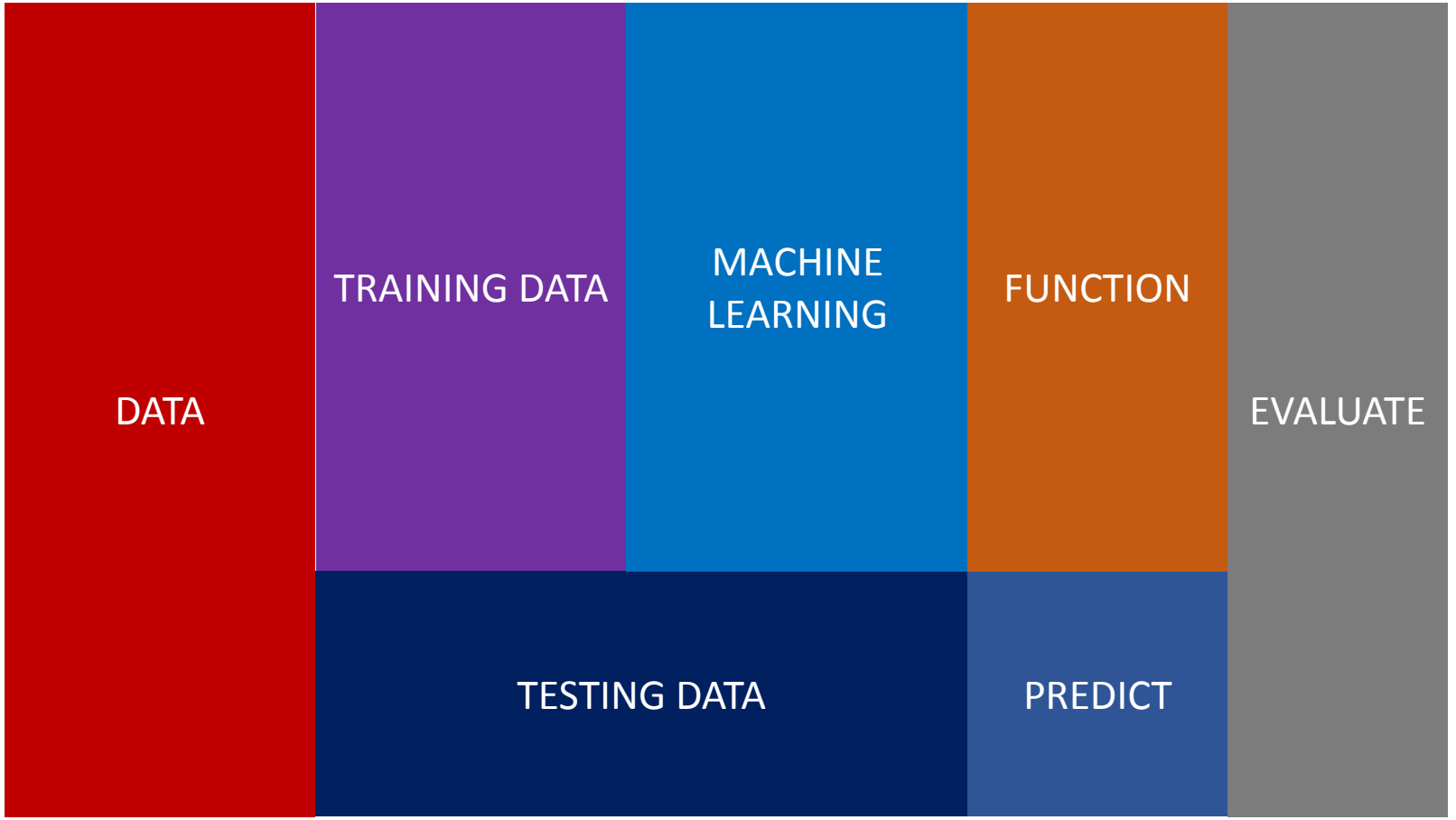
ALGORITHMS

NEURAL NETWORKS

MULTI LAYER PECEPTRON



MODEL EVALUATION



REGRESSION

R^2

Mean Squared Error

Root Mean Squared Error

MODEL EVALUATION

CLASSIFICATION

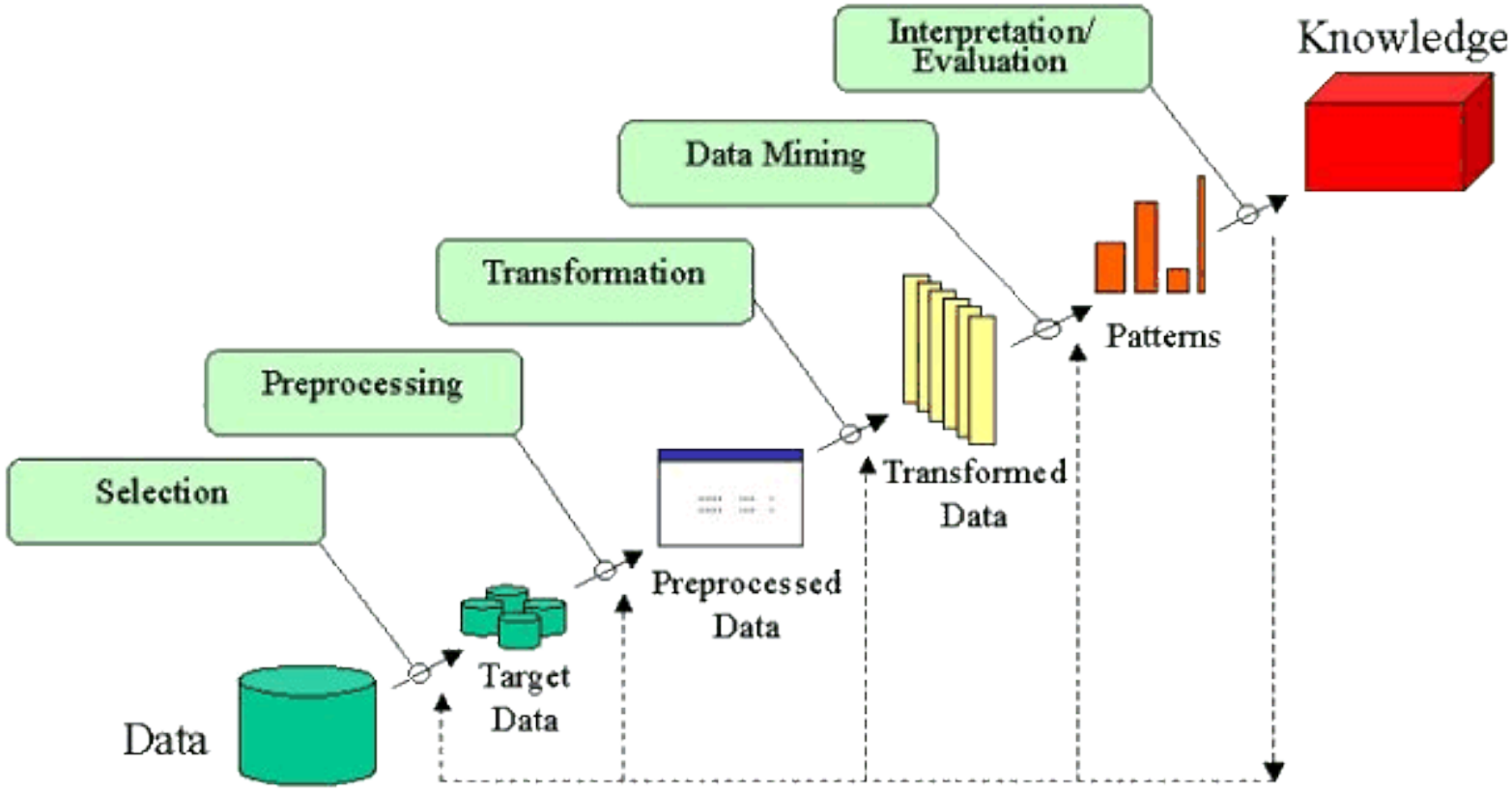
		ACTUAL 1	ACTUAL 0	
PREDICTED	1	TRUE POSITIVE	FALSE POSITIVE Type I Error	Precision = TP/ PTP
PREDICTED	0	FALSE NEGATIVE Type II Error	TRUE NEGATIVE	
		Recall = TP/ TTP		

$$\text{F1 MEASURE} = 2\frac{\text{PRECISION X RECALL}}{\text{PRECISION+RECALL}}$$

$$\text{ACCURACY} = \frac{\text{TP +TN}}{\text{ALL}}$$

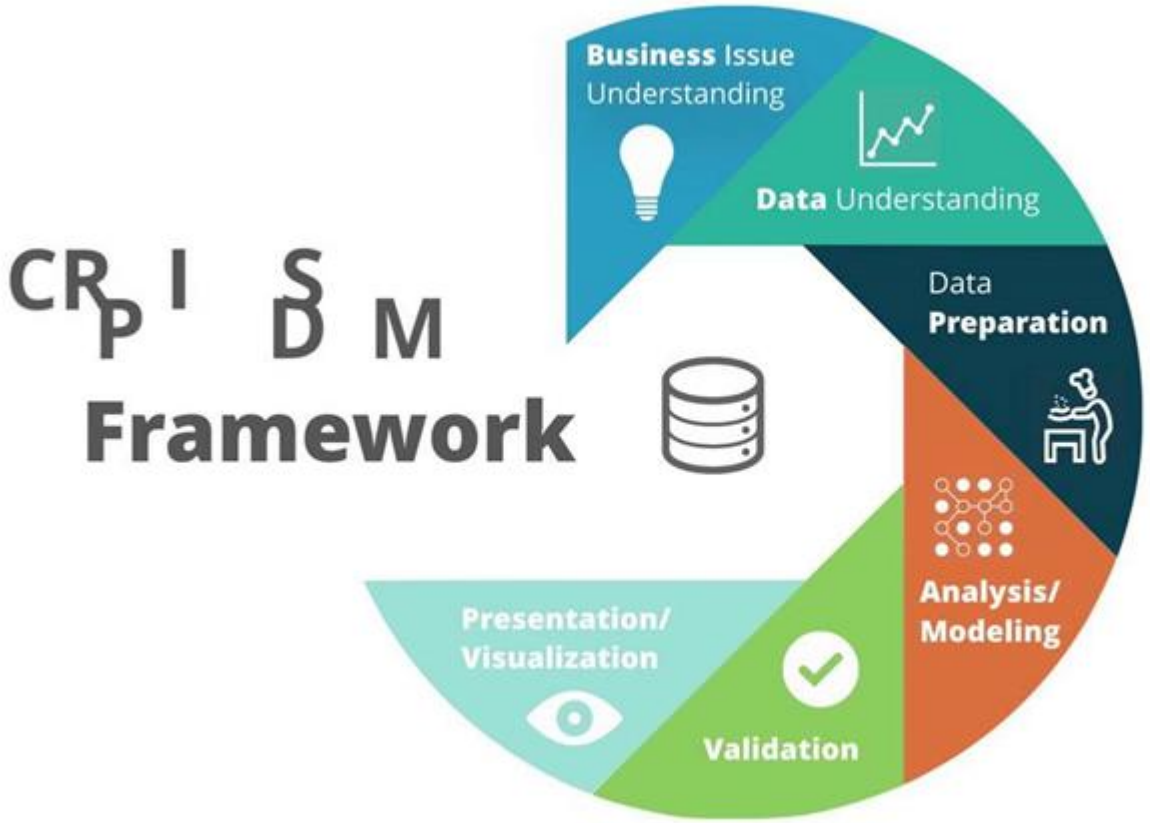
DATA SCIENCE METHODOLOGY

KNOWLEDGE DISCOVERY IN DATABASES



DATA SCIENCE METHODOLOGY

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING



Q & A

THANK YOU