DAY 1 – 15-NOV-2025 (SAT)

- TRAINER INTRODUCTION
- SPECIFIC EXPECTATIONS
- GROUND RULES
- POLL

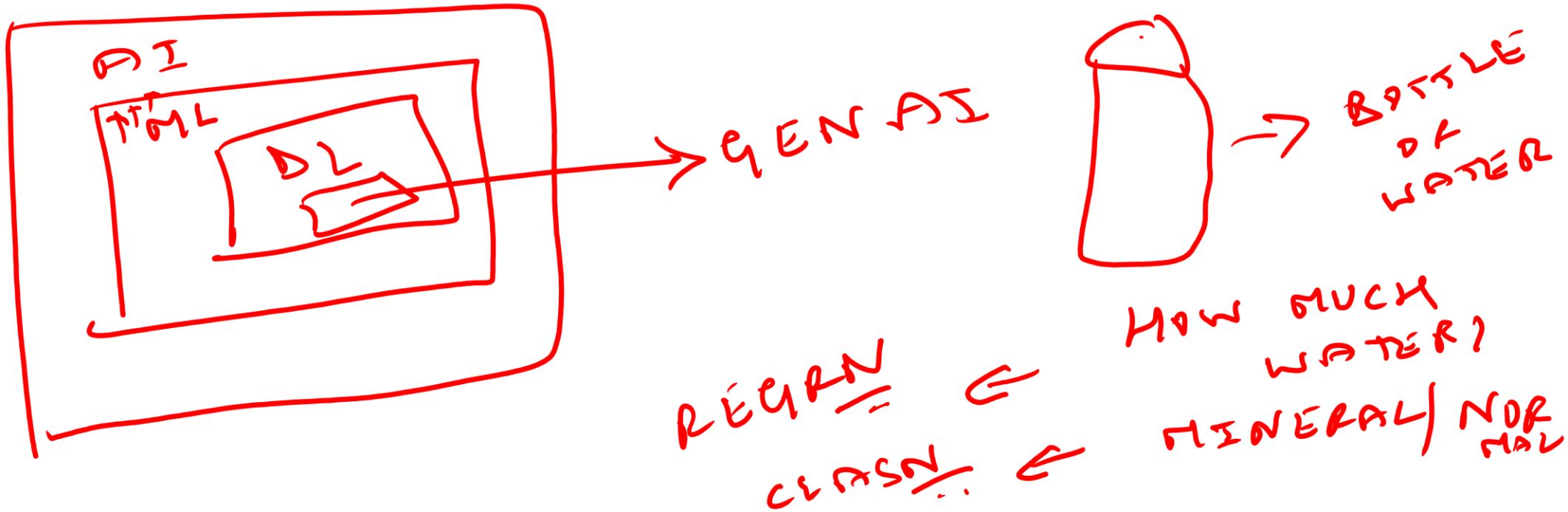
AGENDA – DAY 1 – 15-NOV-2025 (SAT)

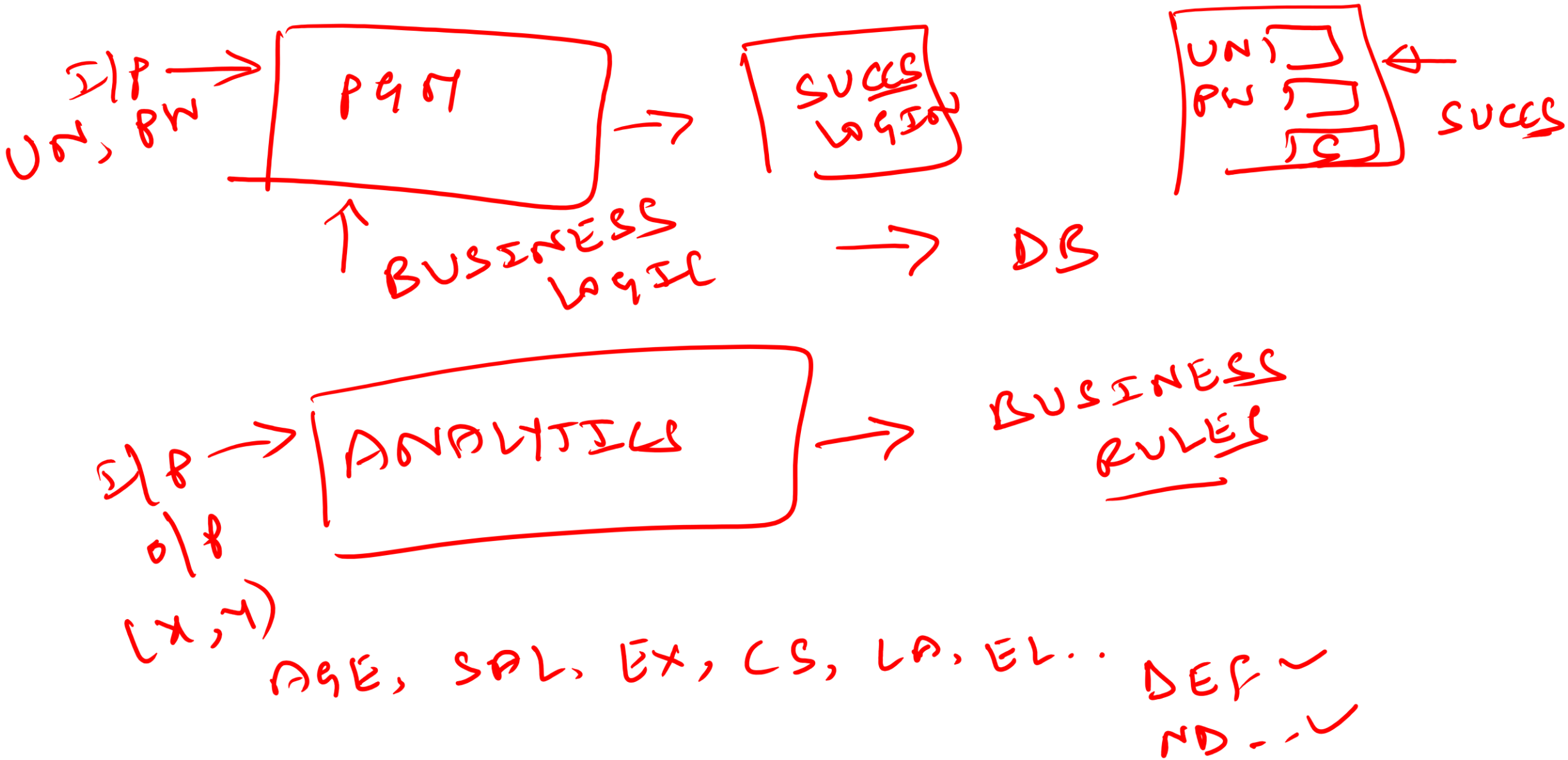
- Supervised Learning
- Introduction to Supervised Learning
  - What Is Supervised Learning?
  - Supervised Learning Categories
  - Applications of SL
- Linear Regression
  - Types – SLR, MLR
- Overfitting & Underfitting
- Non-linear Regression
  - Polynomial Regression
  - Model Performance Metrics
- Regularisation
  - Lasso, Ridge, Elastic-Net
- Model Optimisation
- Hands-On Demo
- Q & A
- SUMMARY, HEADS-UP FOR DAY 2 & CLOSURE

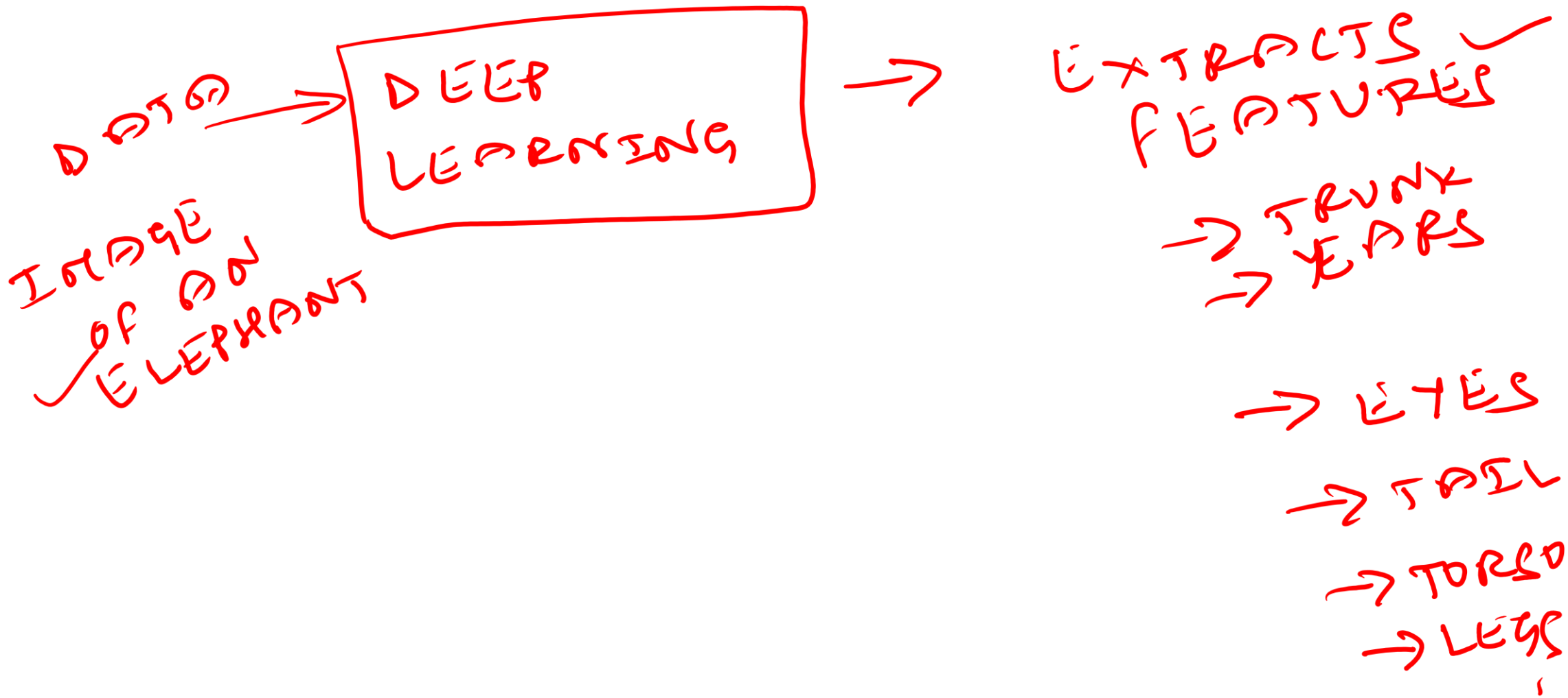
QUESTIONS DATA SCIENCE ANSWER?

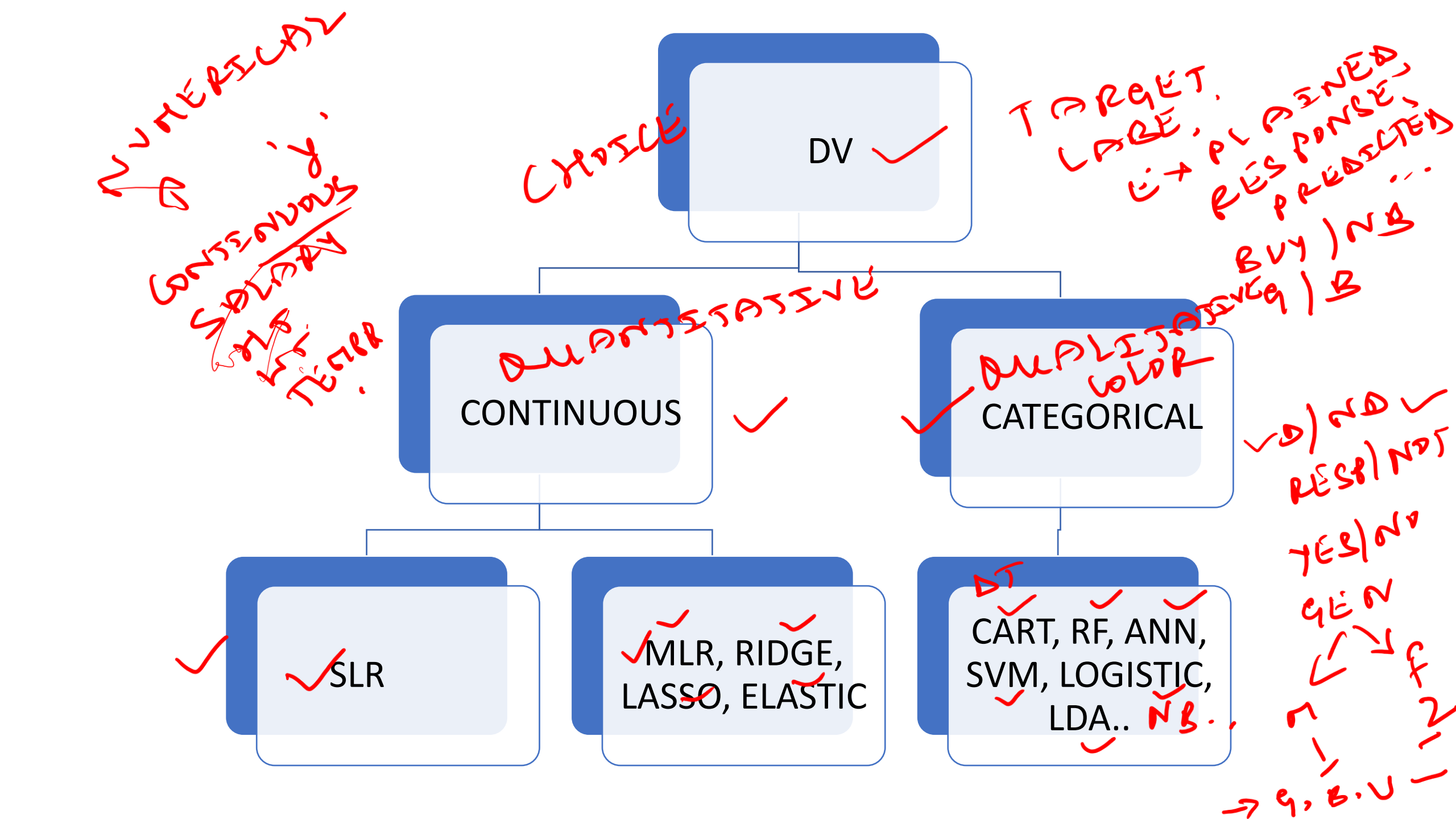
QUESTIONS	ALGORITHMS/MODEL
IS THIS A OR B (GOOF OR BAD)?	CLASSIFFICATION
HOW THINGS ARE ARRANGED?	CLUSTERING
HOW WEIRD OR STRANGE THIS IS?	ANAMOLY DETECTION
HOW MANY OR HOW MUCH	REGRESSION
WHAT SHOULD I DO NOW?	REINFORCEMENT NOW. PRESCRITIVE ANALYTICS

OUT OF SCOPE

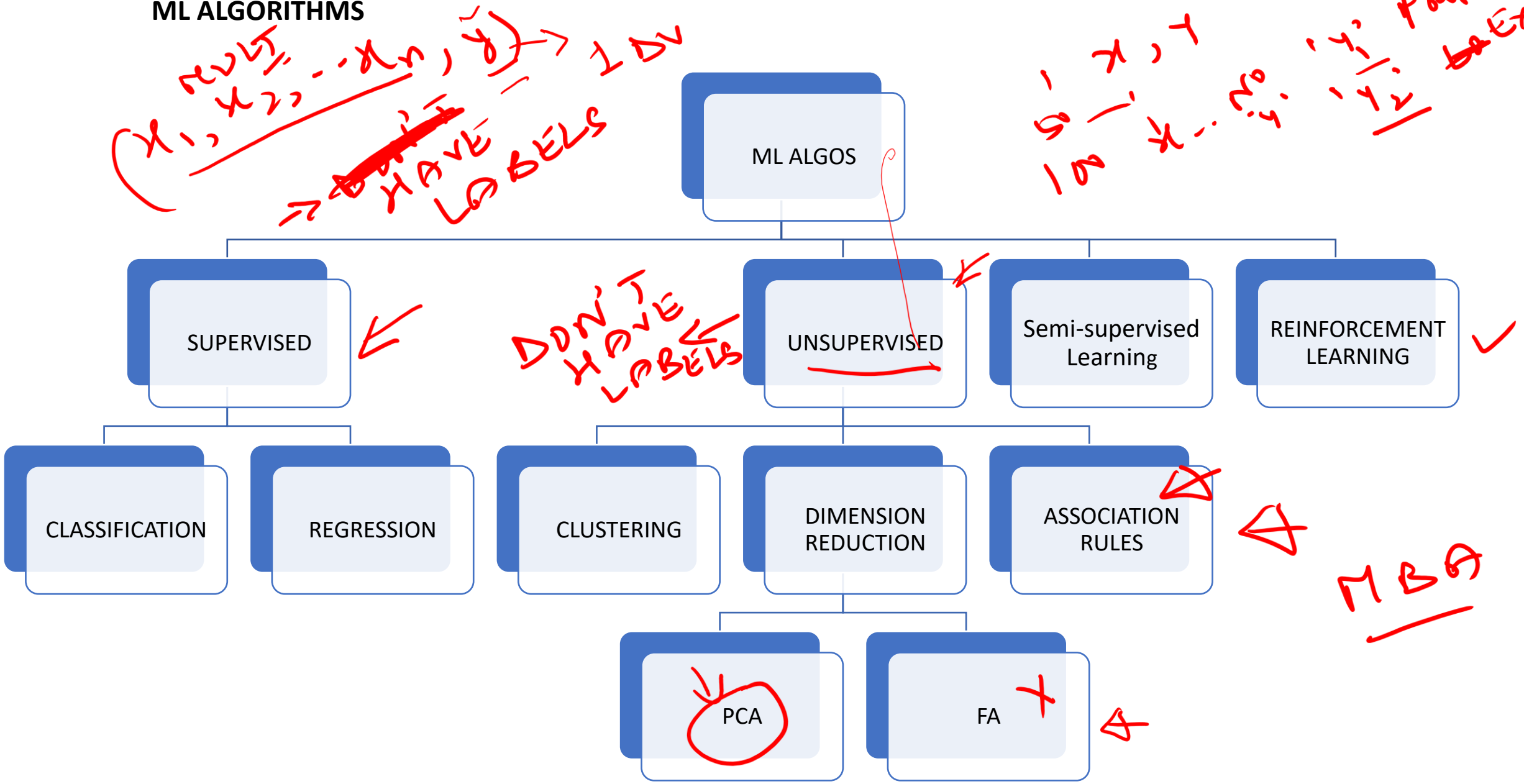



I/P
UN, PW
PGM
SUCCS LOGIC
UN
PW
IC
SUCCS
BUSINESS LOGIC
DB
I/P
O/P
(X,Y)
ANALYTICS
BUSINESS RULES
AGE, SAL, EX, CS, LA, EL..
DEF
ND..

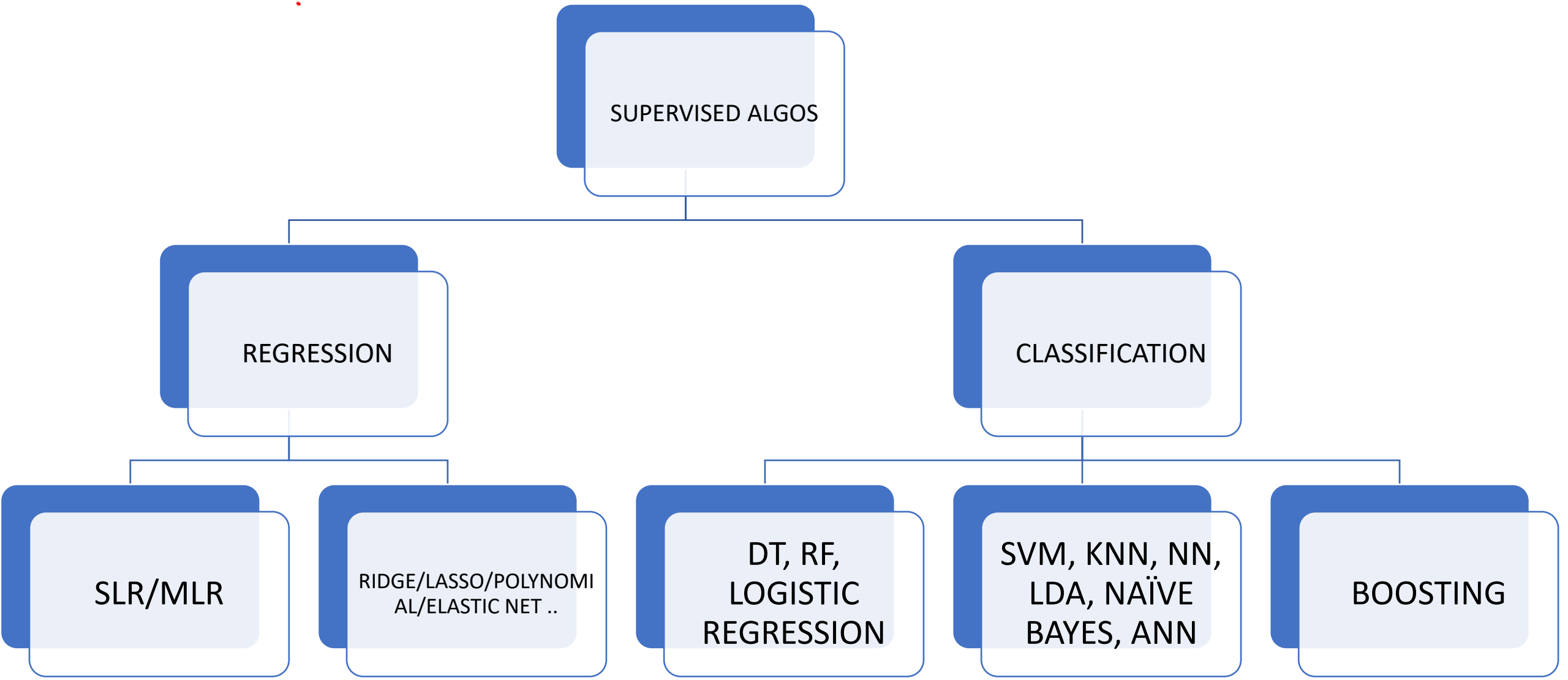

DATA
IMAGE OF AN ELEPHANT
DEEP LEARNING
EXTRACTS FEATURES
→ TRUNK
→ EARS
→ EYES
→ TAIL
→ TORSO
→ LEGS


DV
CONTINUOUS
CATEGORICAL
SLR
MLR, RIDGE, LASSO, ELASTIC
CART, RF, ANN, SVM, LOGISTIC, LDA..

ML ALGORITHMS



SUPERVISED LEARNING:



ALGORITHM

$$SI = \frac{P \times T \times R}{100}$$

P	T	R	SI

PROBLEM: SI

BUBBLE SORT

↳ 1. COMMON SENSE

↳ 2. FIRST PRINCIPLES

✓ ↳ 3. ALGORITHMIC APPROACH

↳ 4. DATA CENTRIC APPROACH.

2, 4, 6, 8, 10, ...

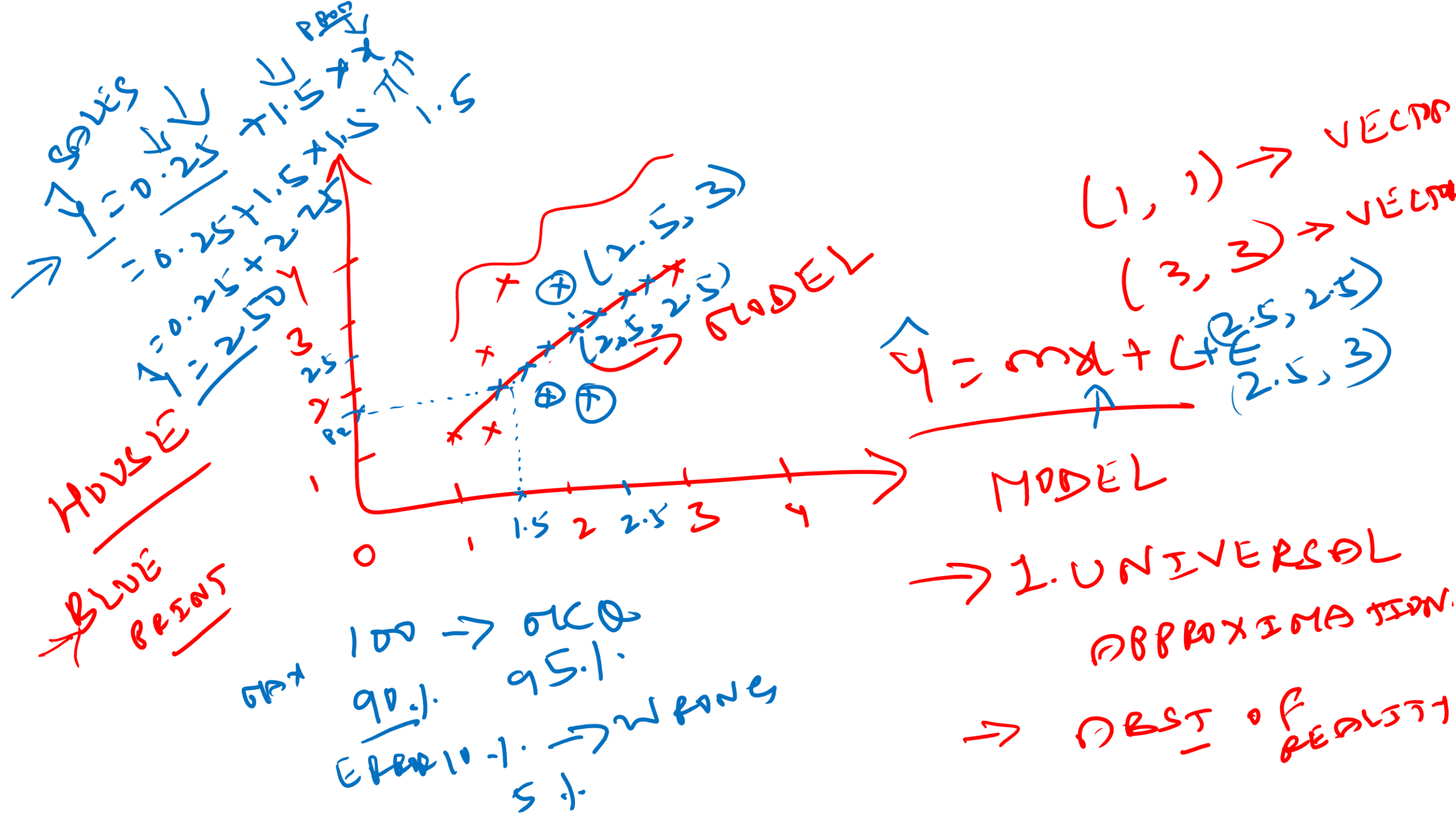
I/P

O/P : I/P + 2

= I/P × 2

→ 2, 4, SIN30, $\frac{1}{e^{-17267}}$, $\log\left(e^{+2025}\right)$ ×

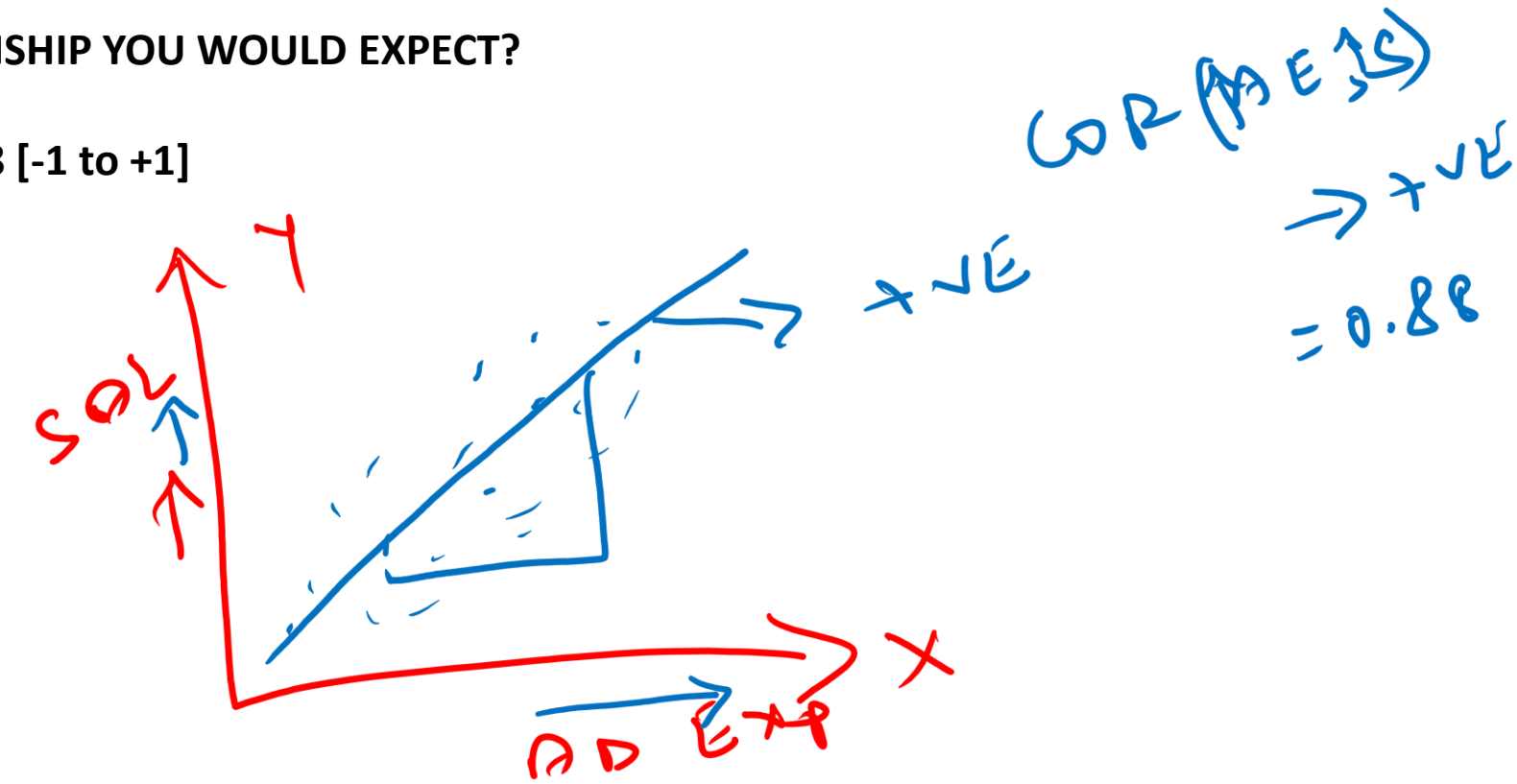
→ ML MODEL ?


HOUSE
BLUE POINTS
MODEL
1.5
2
2.5
3
4
0
1
(2.5, 3)
(2.5, 2.5)
$\hat{y} = mx + c$
MODEL
(1, 1) → VECTOR
(3, 3) → VECTOR
(2.5, 3)
1. UNIVERSAL APPROXIMATION.
→ BEST OF REALITY

NEED FOR REGRESSION?

SALES, AD EXP →WHT KIND OF RELATIONSHIP YOU WOULD EXPECT?

CORRELATION → POSITIVE → COR = 0.88 [-1 to +1]



QUESTION:

22% FLAT INCREASE IN AE WOULD RESULT IN HOW MUCH SALES?

Correlation will only tell us strength & direction, not how much sales increase

SICK → GO TO A DOCTOR.

PRESCRIPTIVE ANALYTICS

IF $T > 101°F$ MED 1

MED 2

DESCRIBE (DESC ANALYTICS) ✓

↳ HEADACHE

↳ FEVER

→ DIAGNOSIS ✓ (DIAGNOSTIC ANALYSIS)

→ VIRAL FEVER ✓

→ BLOOD TESTS → REPORTS

→ PREDICTIVE ANALYTICS ✓

→ VIRAL FEVER

WHY REGRESSION?

- COME UP WITH STATISTICALLY SIGNIFICANT OUTCOMES

- UNDERSTAND THE IMPACT OF MULTIPLE INDEPENDENT VARIBALES (FACTORS) ON TARGET VARIABLE
  - IMPACT OF AE & PROMOTION ON SALES

- KPIs – REVENUE, PROFITS, LOSS, COST, DEMAND, PRICE, MARKET SHARE
  - CUSTOMER SATISFACTION, EMPLOYEE ATRRITION etc etc

- "BUSINESS PLANNING" & FORECASTING / PREDICTING

- **ACCESS DATA**
  - **CROSS-SECTIONAL DATA**
  - **TIME SERIES DATA – ABC COMPANY STICK PRICES ON A PERIODICAL PRICES (DAILY)**
  - PANEL DATA → CS + TS DATA

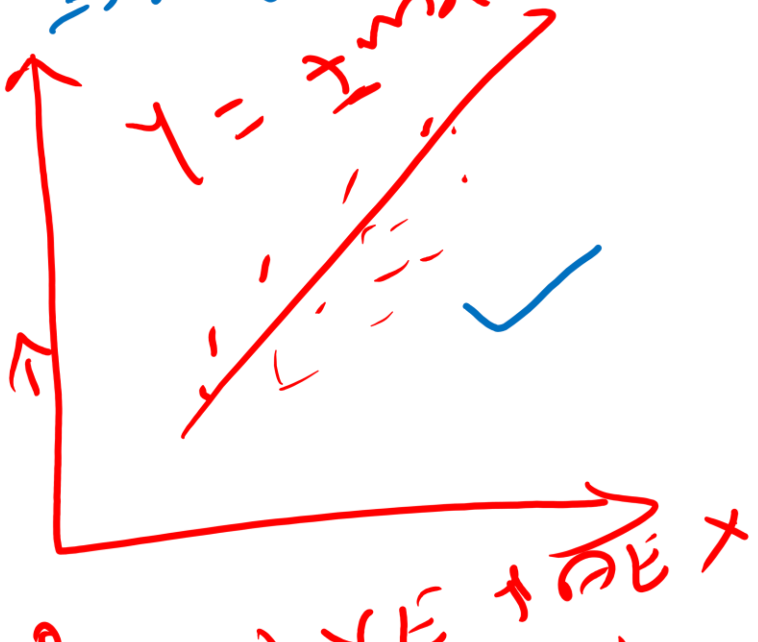
2 PTS → ST LINE
3 PTS → PLANE
→ HYPER PLANE

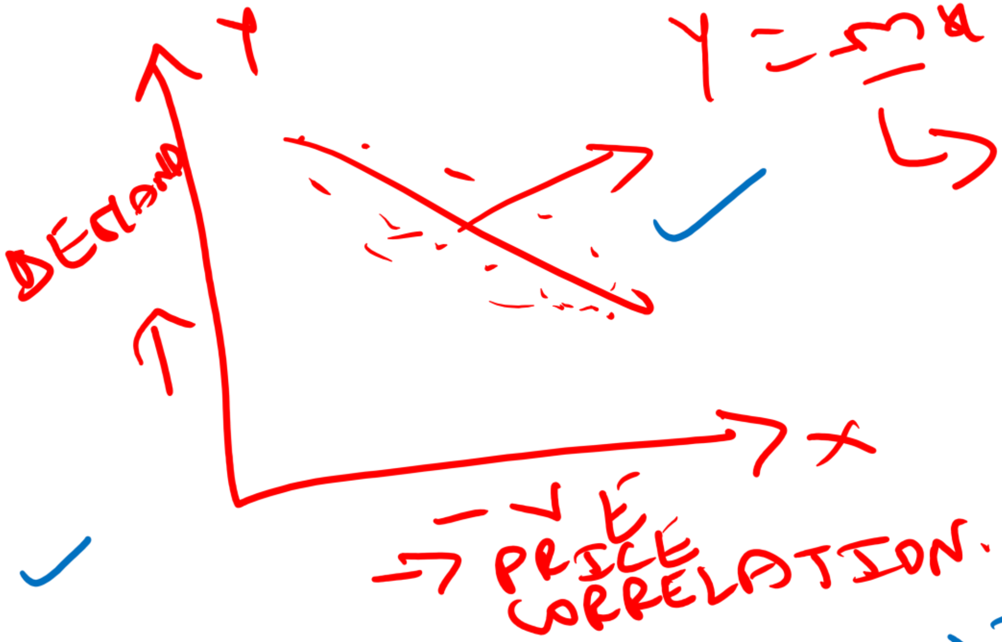
Y, $x_1, x_2, x_3, \ldots x_n$

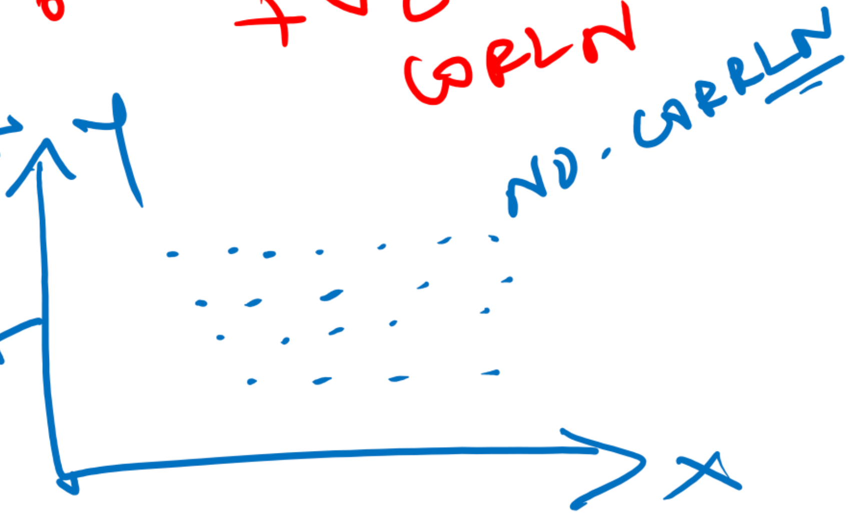
$y = f(x)$
↓
LINEAR

LINEAR REGRN

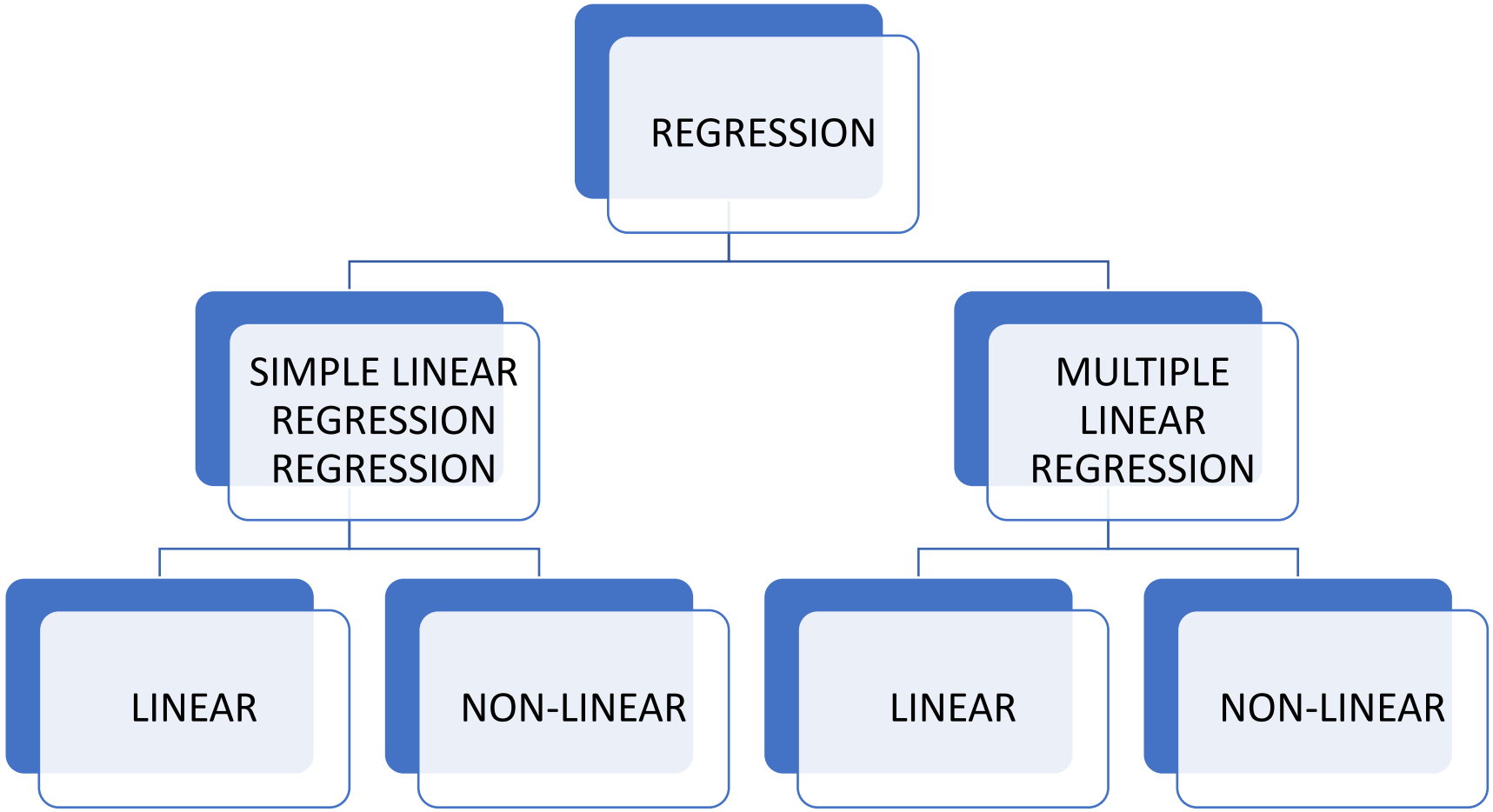
SLR
MLR







NO LINEAR RELATIONSHIP


REGRESSION
SIMPLE LINEAR REGRESSION REGRESSION
MULTIPLE LINEAR REGRESSION
LINEAR
NON-LINEAR
LINEAR
NON-LINEAR

REGRESSION – REPRESENTATION & NOMECLATURE

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$ → SLR ⇒ 1 Y, 1 X

Y = matrix

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$ → MLR

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

DETERMINISTIC

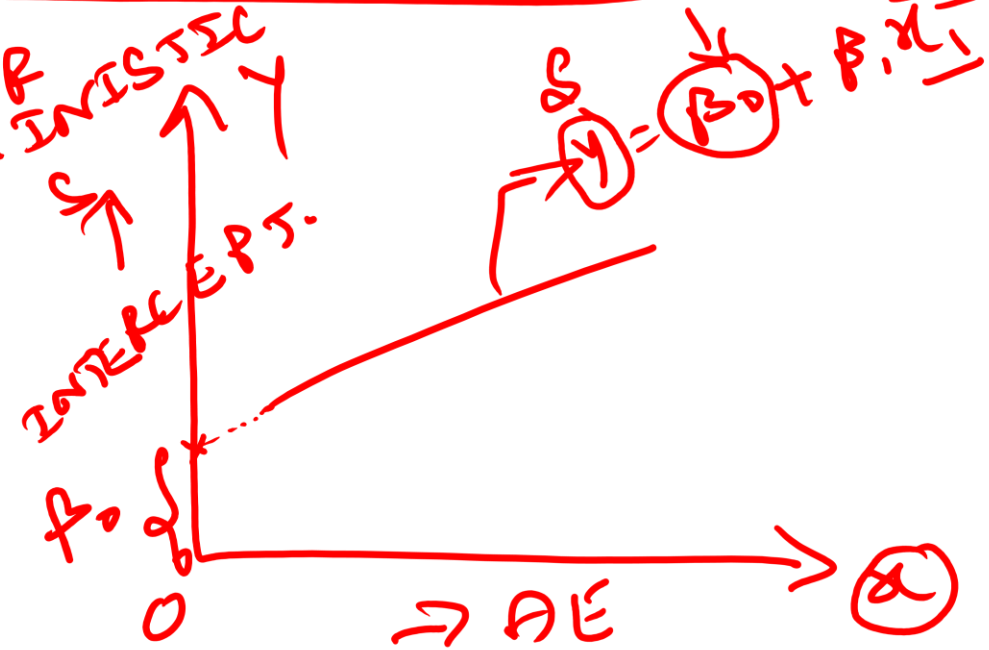
NON-DETERMINISTIC

$x_1 = 0$

$Y = \beta_0 + \beta \times 0$

$Y = \beta_0 + 0$

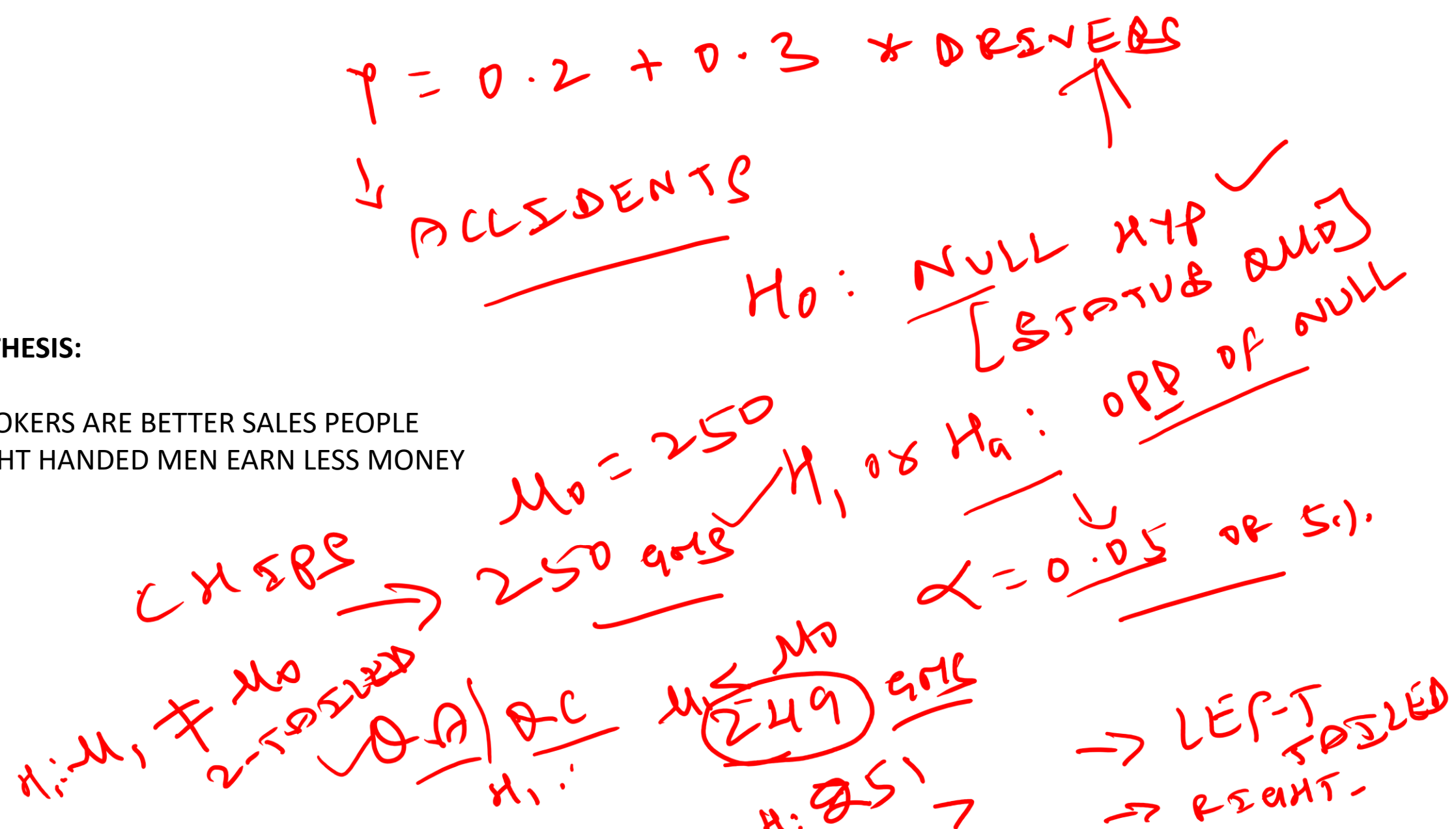
$Y = \beta_0$



$E(Y) = \beta_0 + \beta_1 x_1$ → DE

HYPOTHESIS:

- SMOKERS ARE BETTER SALES PEOPLE
- RIGHT HANDED MEN EARN LESS MONEY



$\alpha = 0.05$ or 5% → TYPE I ERROR

$\beta = 1 - \alpha = 1 - 0.05 = 95\%$

$\beta = 10\%$

P-VALUE = 0.06 < ALPHA 0.05 → CANNOT REJECT NULL

COMPUTE P-VALUE = $\frac{0.06}{0.044}$

IF CPV = 0.06

SMALL SAMP

SD UNKN

t-TEST

CPV < ALPHA

REJECT NULL

LARGE SAM

SD KNOWN

Z-TEST

ANOVA:

NATURE of $\vec{X}$ CAT, NATURE of Y $\rightarrow$ CONT

INSURANCE PROVIDER: IS THE AVG INS CLAIM ACROSS 4 REGIONS CAT IS SAME OR NOT!

HR HEAD
AVG PERF

INDIA → E, W, N, S

1- WAY ANOVA
2- WAY ANOVA

Avg SP = f(REGION)

y → CONTINUOUS

y = f(x)

INS = f(R) → CAT

CONT

SLR

$Y = \beta_0 + \beta_1 x + \epsilon$

HYPOTHESIS

$\beta \in (\mu, \sigma)$ — GUESS SSD

MLR

LINEAR

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

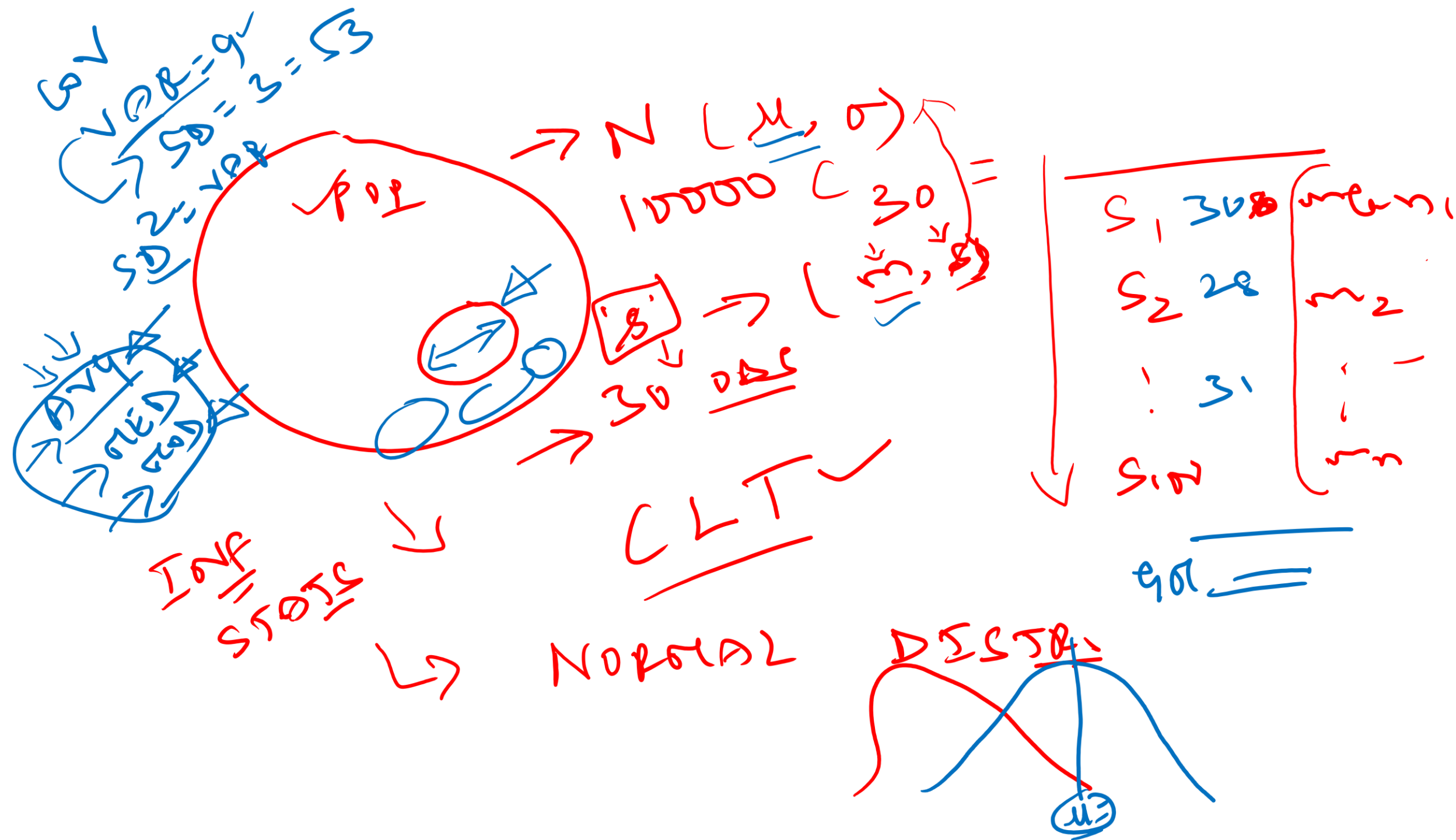
β_0 ✓ P — β_1 D — PRED

H_0: NO LINEAR REL^N B/W Y & X ✓

H_1 OR H_a: LINEAR RELATIONSHIP.

H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_k = 0$ ✓

H_1: $\beta_1 = \beta_2 = \beta_3 \neq \beta_4 = \cdots \beta_k$


pop
N (μ, σ)
10000
30
CLT
INF
NORMAL
DISTRI

SOURCE : INTERNET