02-08-2025

Agenda:

→ Bayes theorem
→ PDF, PMF, CDF
→ Skewness
→ Correlation and covariance

A = patient has the disease

B = patient test positive

Traditional prob:

Q: what is the prob of getting a positive test if a person has the disease?

$P(B|A)$

→ 99% of people with the disease test positive. (data - ground truth)

$P(B|A) = 0.99$ (if someone has disease, the test will say positive 99% of time)

→ Bayesian view: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A') \cdot P(A')}$$

→ 1% of population has the disease → $P(A) = 0.01$

→ $P(B|A) = 0.99$ ( test detect disease 99% of time ) ] TP

→ $P(B|A') = 0.01$ ( 1% false positive rate ) ] FP

Q. The patient tested positive, what is the prob that they actually have the disease?

test outcome → disease likelihood (Bayes)

disease likelihood → test positive (traditional)

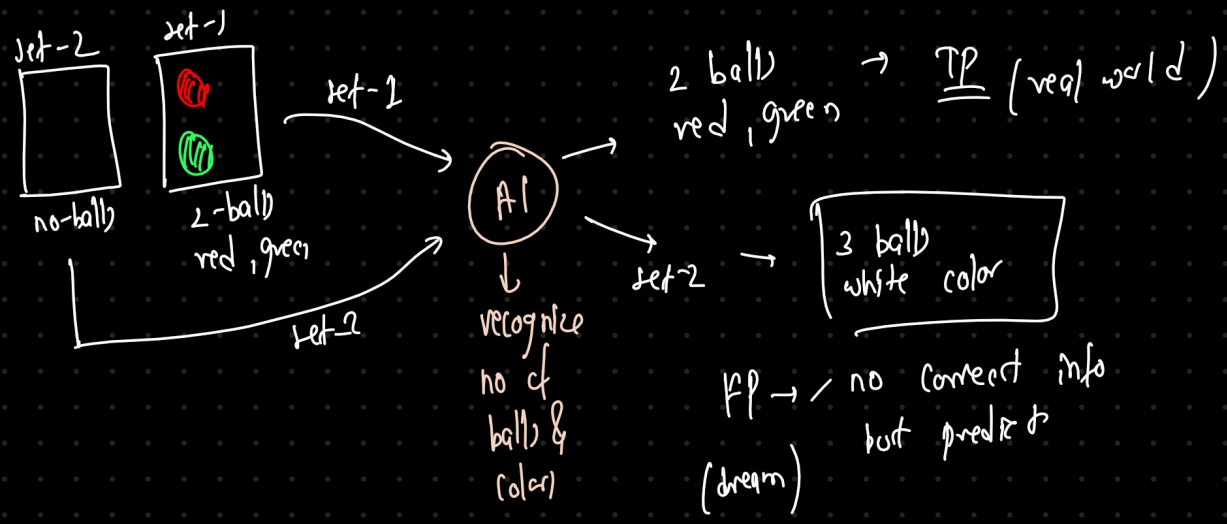
$$P(A|B) = \frac{0.99 \cdot 0.01}{0.0198} = \frac{0.0099}{0.0198} = 0.5$$

Even the test 99%, there only 50% chance the person actually has the disease. even after testing positive.

Recap:

Tradition :— what is the chance of positive test if person is sick?

Bayes :— Given a positive test, how likely is the person sick?

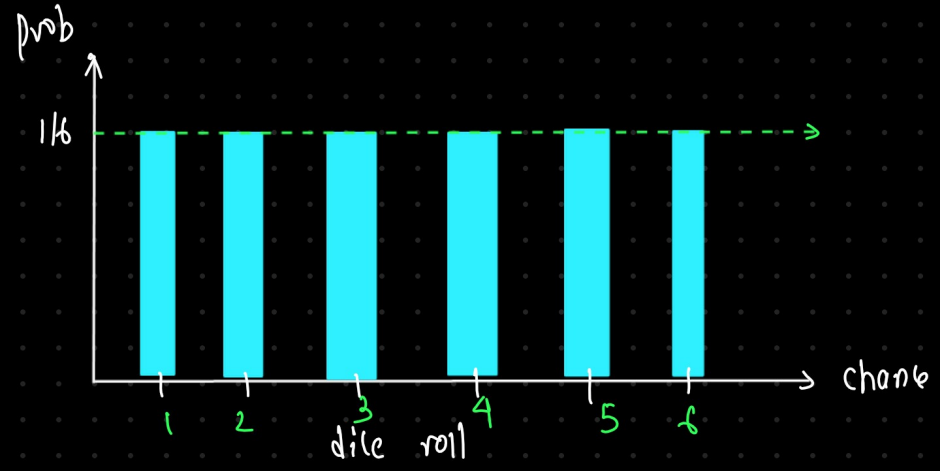



---

# PMF ( Probability Mass function )

→ It is for discrete data ( like dice rolls, coin flips, counts )

→ gives the prob. of each possible value in a dataset.




PMF plot

PMF is a function, Not just a plot

sum of prob. = 1

why: what's the chance of x happening exactly?

↓

1, 2, 3, 4 → discrete data

1-2, 2-4 → cannot be range

→ PMF is not a plot, it's a function.

→ PMF function is the math rule behind the plot, the plot just visualize it.

↓

maths of cls → card

↓

dice

Concept → PMF

CLT →

Team → data

churn

→ probability with their age

↓

{ 40: 20%
60: 40
80: 30% }

## CDF (Cumulative Distribution function)

→ works for both discrete & continous data,

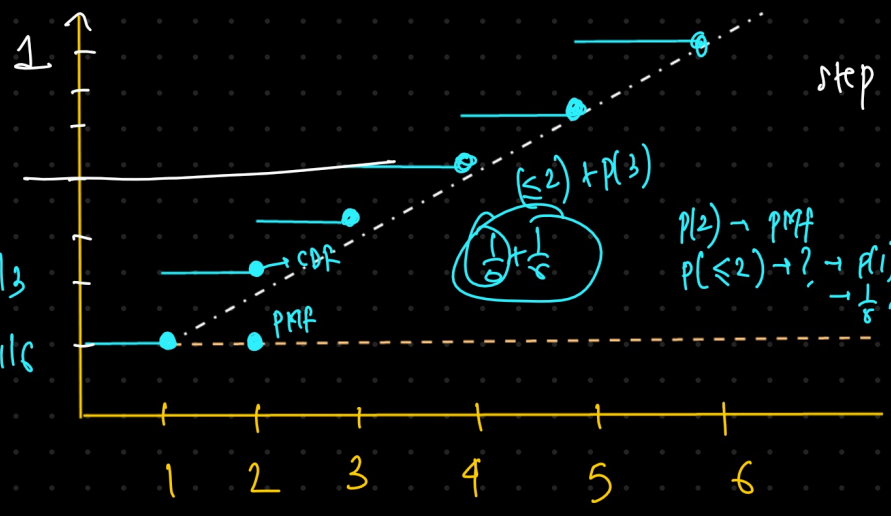
→ It gives the probability of $X \leq$ a value.

what is the chance of X being less than or equal to 4,

→ $P(\leq 4)$ on a dice!

↓

$P(1) + P(2) + P(3) + P(4)$




step → CDF for discrete counts

$P(\leq 2) + P(3)$

$P(2)$ → PMF

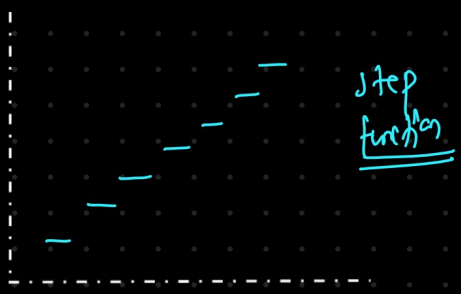
$P(\leq 2)$ → ? → $P(1) + P(2)$

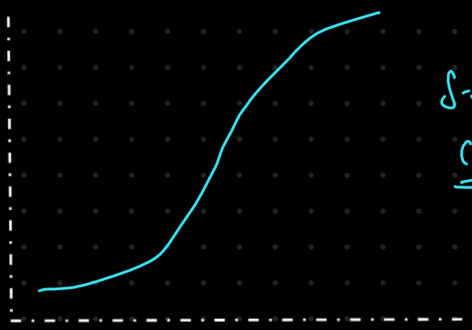
→ $\frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$

0 : 10
1 : 20
3 : 10
4 : 20
5 : 10

0: 10 + Nothing → 10
1: 20 + 10 → 30
3: 10 + 20 + 10 → 40
4: 20 + 10 + 20 + 10 → 60
5: 60 + 10 → 70

CDF always start with 0 and ends at 1




---

PMF → probability of outcome.
CDF → Cumulative probabilited

## PDF (probability Density function)

- PDF is for continous data
- shows the relative likelihood of a range of value (not exact point) → why?

why?

→ what is the chance of X falling between a & b?
→ what is the prob. of a student height in between 160cm & 170cm?

why PDF use range (Not exact value)

→ continous data has infinite precision.

→ 160.00 cm ? prob

160.0001 160.00002 160.000000009 160.X

↓ ∞ possibilities

$$P(X) = \frac{\text{No. of favourable outcome}}{\text{Total possible outcomes}} = \frac{1}{\infty} = 0$$

range → for continous data

continous prob are only meaniful over interval (a, b)

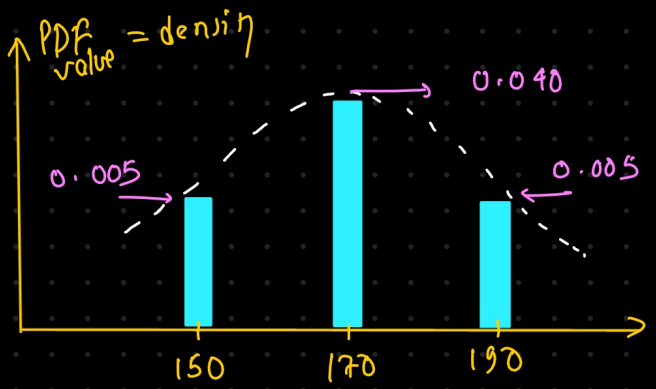
↓

160cm, 170cm

40%.

prob. of an exact point is 0 because there's no "width" the interval.

y → PMF → prob
↳ CDF → prob
↳ PDF → Density




Higher PDF value = More likely to find value nearby

Step 1: collect data

Height (in cm) of 20 people:

[162, 168, 171, 154, 175, 178, 165, 169, 172, 160, 167, 173, 170, 176, 164, 179, 166, 170, 174, 168] → len() → 20

Step 2: create a histogram.

bin-width = 5

| Height (in cm) | Frequency |
|---|---|
| 150 - 155 | 1 |
| 155 - 160 | 0 |
| 160 - 165 | 3 |
| 165 - 170 | 6 |
| 170 - 175 | 6 |
| 175 - 180 | 4 |

$$PDF(\text{Density}) = \frac{\text{Frequency of bin} \rightarrow \text{prob}}{\text{Total observation (all)} \rightarrow \text{prob}} \Big/ \text{Bin width} \rightarrow \text{bin width}$$

→ unit: (prob per unit height)

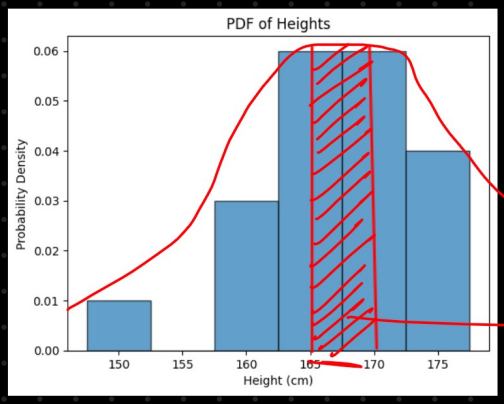
## Step3 : Normalize the Histogram

| Height (in cm) | frequency | PDF |
|---|---|---|
| 150 - 155 | 1 | (1/20x5) = 0.01 |
| 155 - 160 | 0 | 0 |
| 160 - 165 | 3 | 0.03 |
| 165 - 170 | 6 | 0.06 |
| 170 - 175 | 6 | (6/20x5) = 0.06 |
| 175 - 180 | 4 | 0.04 |

$$PDF = \frac{\frac{\text{Freqn. of bin}}{\text{Total observation}}}{\text{Bin width}} \rightarrow \frac{\text{Freq. of bin}}{\text{Total observ} \times \text{Bin width}}$$

→ Dividing a number is equivalent to multiplying by its reciprocal.

Density = probability per unit height




→ plot

0.06 × 5 = 0.30

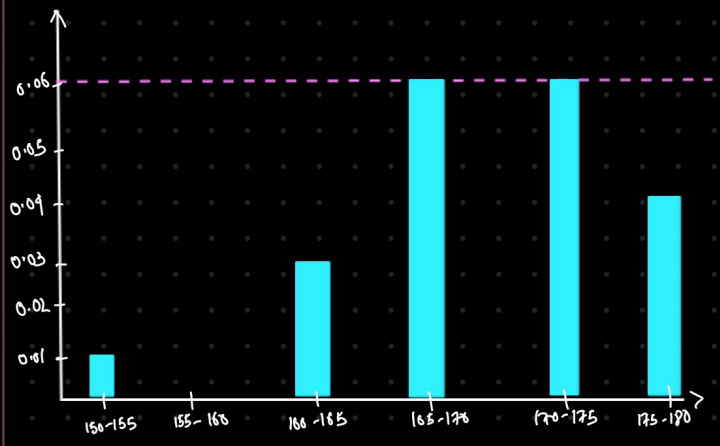
Step 4: Calculate area under PDF curve: → height ≥ 165 & ≤ 170

e.g. Find ( P(165 ≤ height ≤ 170 )

(1) Identify bin involved = 0.06 (Density)

Prob = Area = PDF value × Bin width = 0.06 × 5 = 0.30 (30% chance)

Verification = 165 - 170 → $\frac{6}{20}$ = 0.30



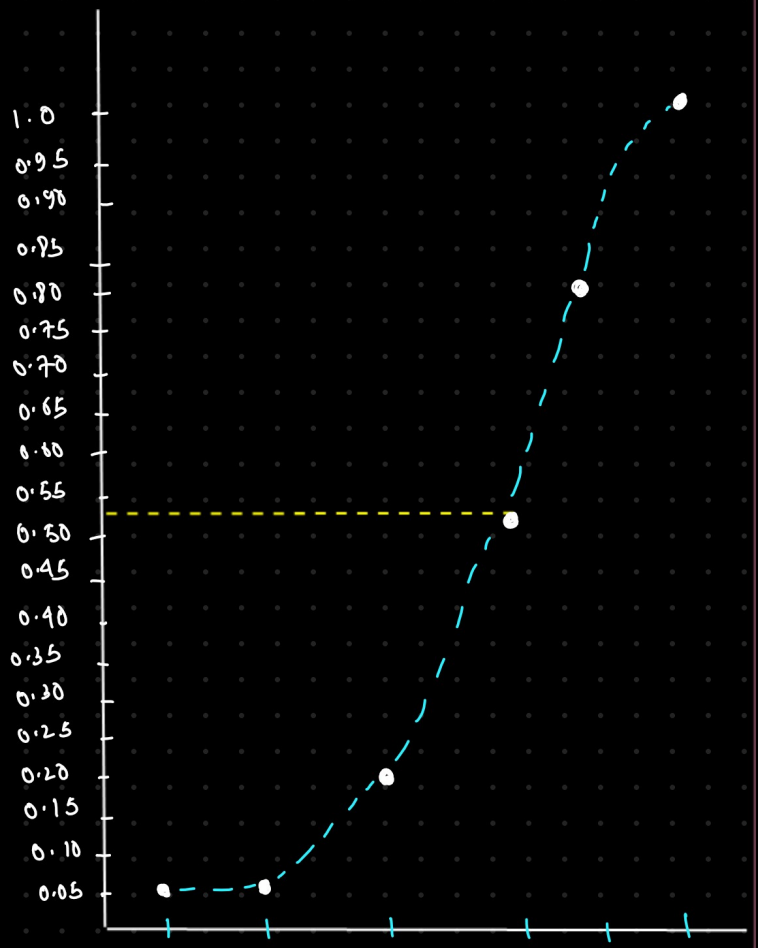

0.01 × 5 = 0.05 (Prob of bin)

0 × 5 = 0

0.03 × 5 = 0.15

0.06 × 5 = 0.3 , 0.06 × 5 = 0.3

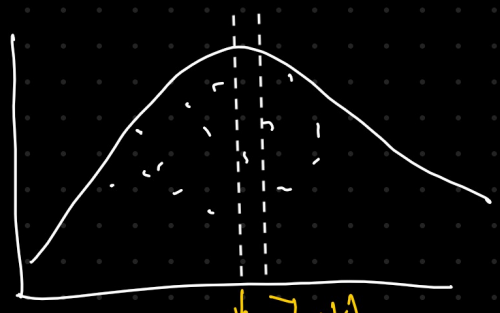
0.04 × 5 = 0.20



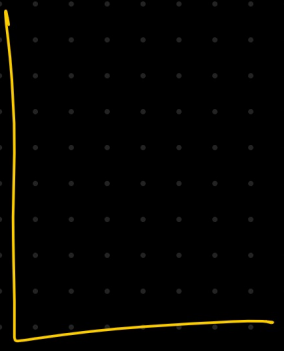
CDF → Area x width

CDF → Area

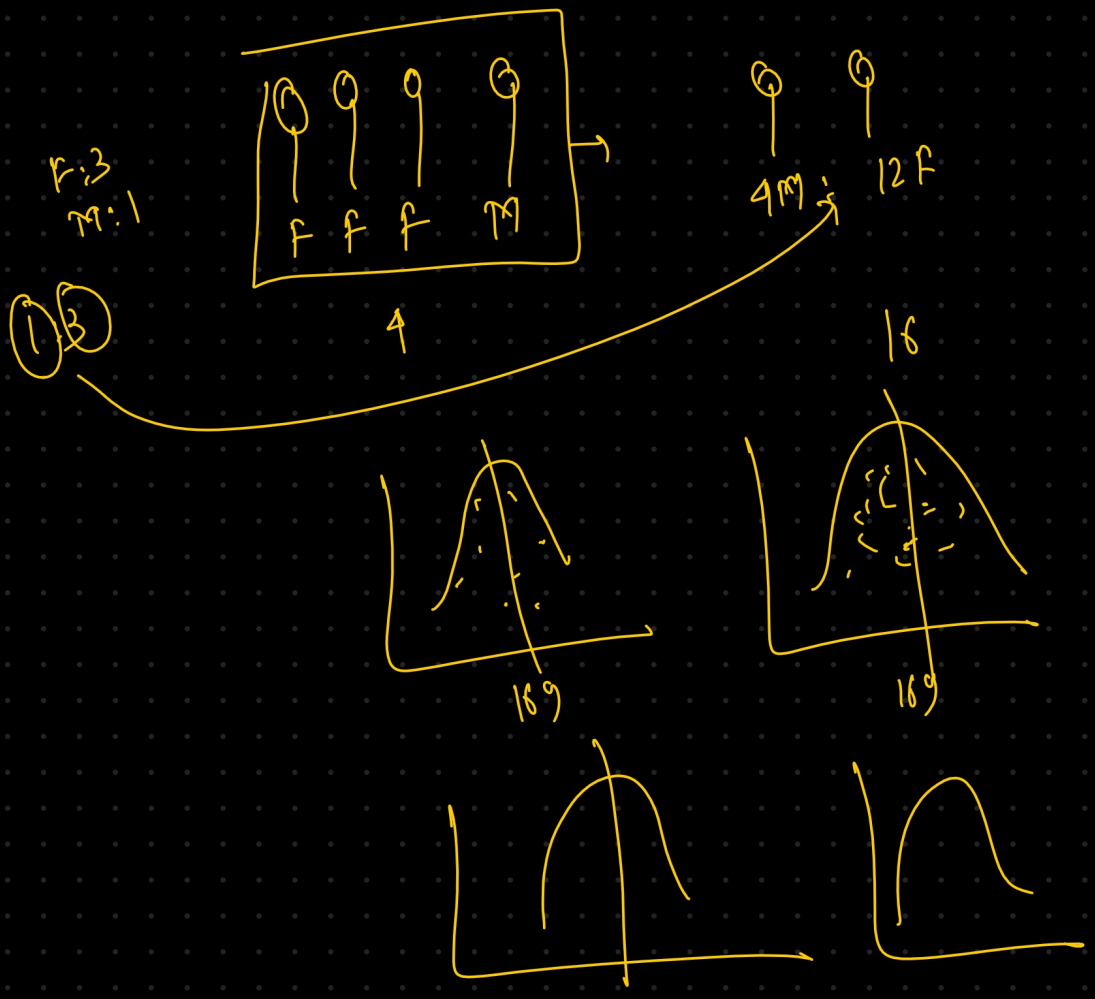
PDF gives Density

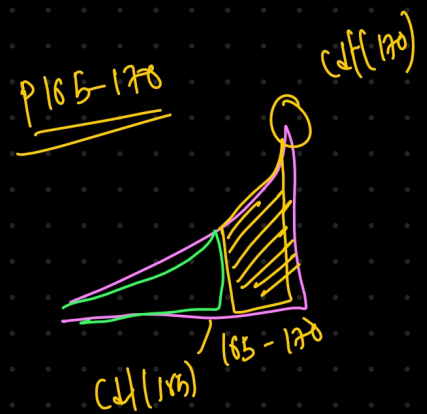


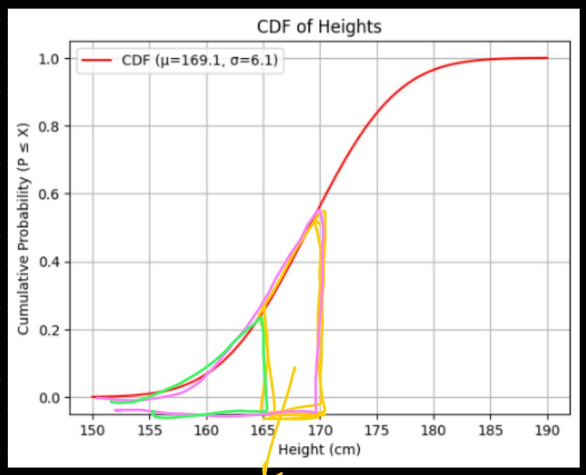

→ CDF =



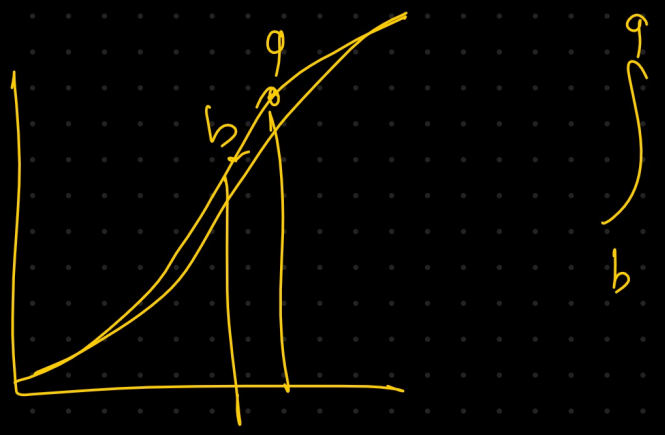
len()
20 → CD



F:3
M:1
F F F M
4
4M
12F
16
169
169


P165-170
CdF(170)
CdF(165)
165-170


CDF of Heights
CDF (μ=169.1, σ=6.1)
Cumulative Probability (P ≤ X)
Height (cm)
1.0
0.8
0.6
0.4
0.2
0.0
150
155
160
165
170
175
180
185
190

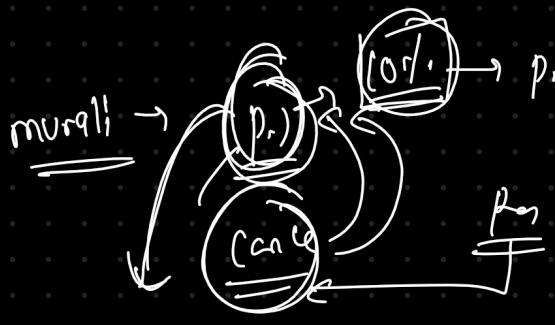
Area / pr




PMF → discrete prob

CDF → cum prob → discrete
↘ continuous

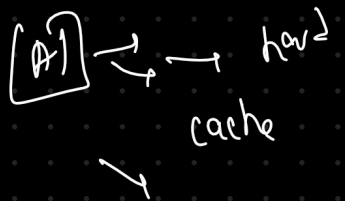
PDF → prob of range of value → CDF ( PDF )
↓
Area under curve

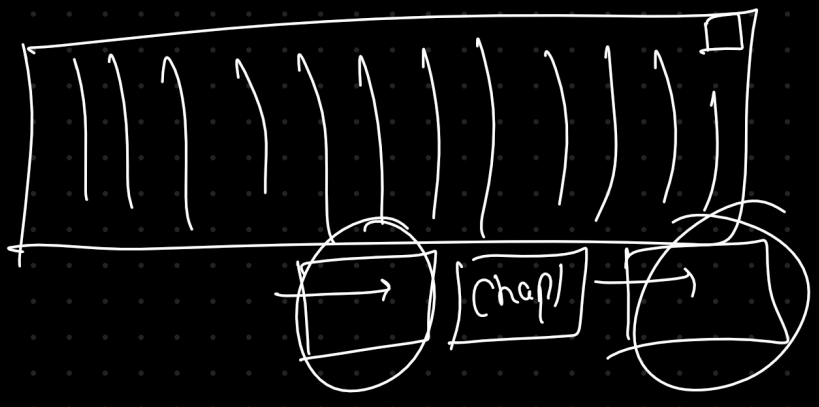




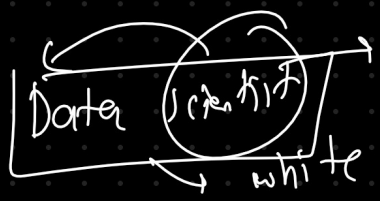
core concept

Branch



AI




↳ (ab)

↳ Dark