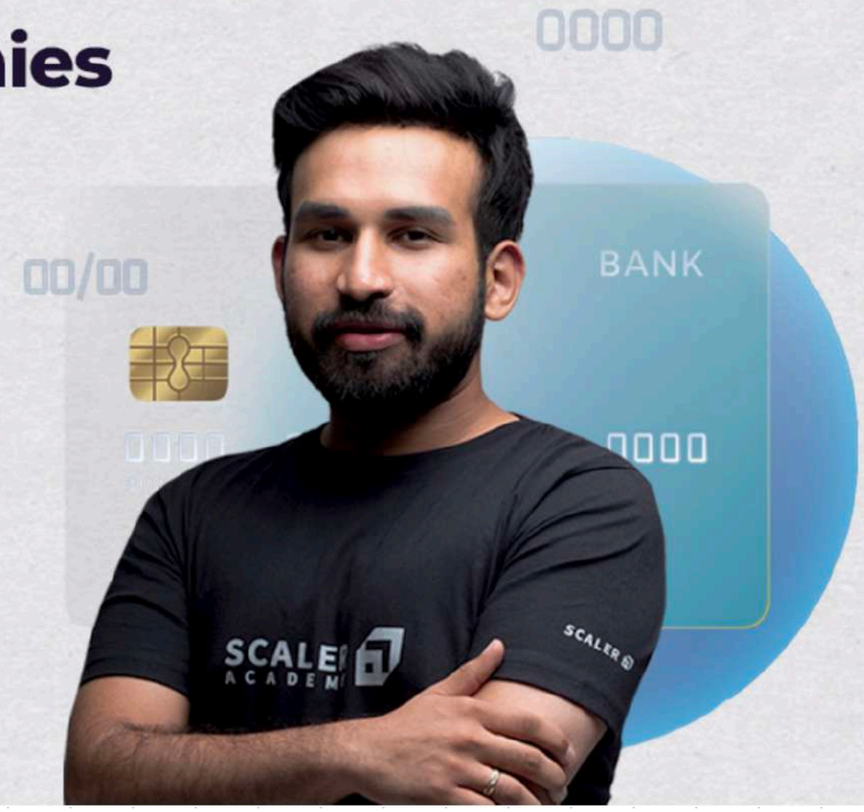
How Credit Card Companies use Machine Learning to prevent Fraud?

MasterCard VISA AMERICAN EXPRESS



FREE MASTERCLASS

By **Anant Mittal,**
Lead Data Scientist
& Instructor, Scaler

1st Dec, Thursday | 8 PM - 11 PM

Is this Masterclass for you?

→ Beginner Friend by (ML or Fraud Theory)

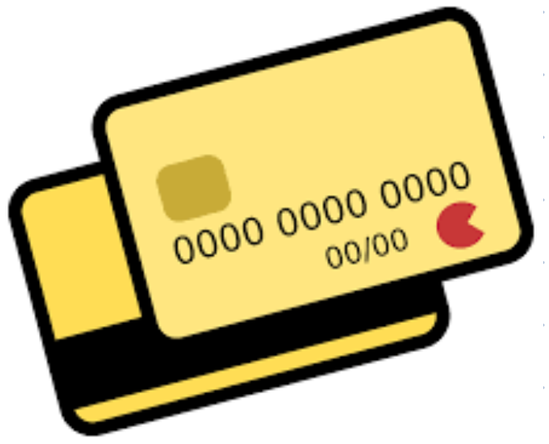
→ Overview (Fraud Detection + ML)

↓

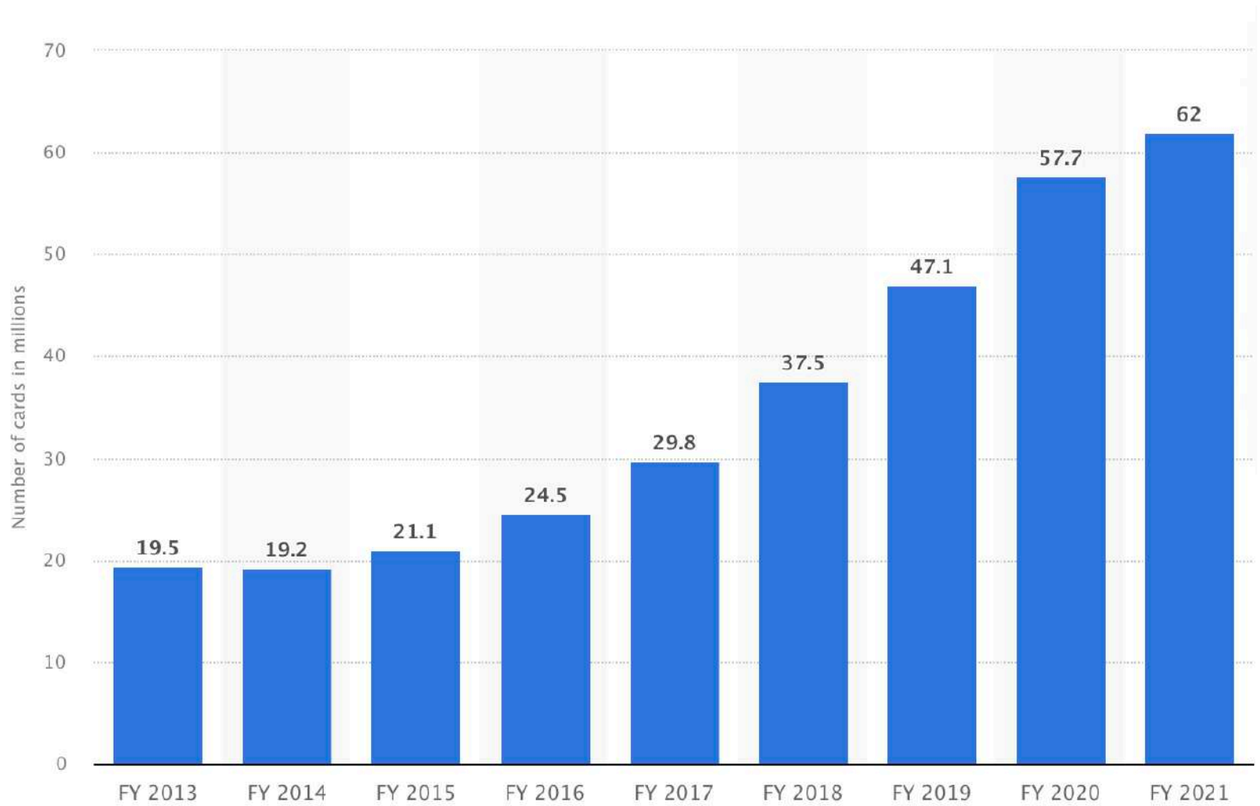
Diluted.

→ Notes LinkedIn + Discord

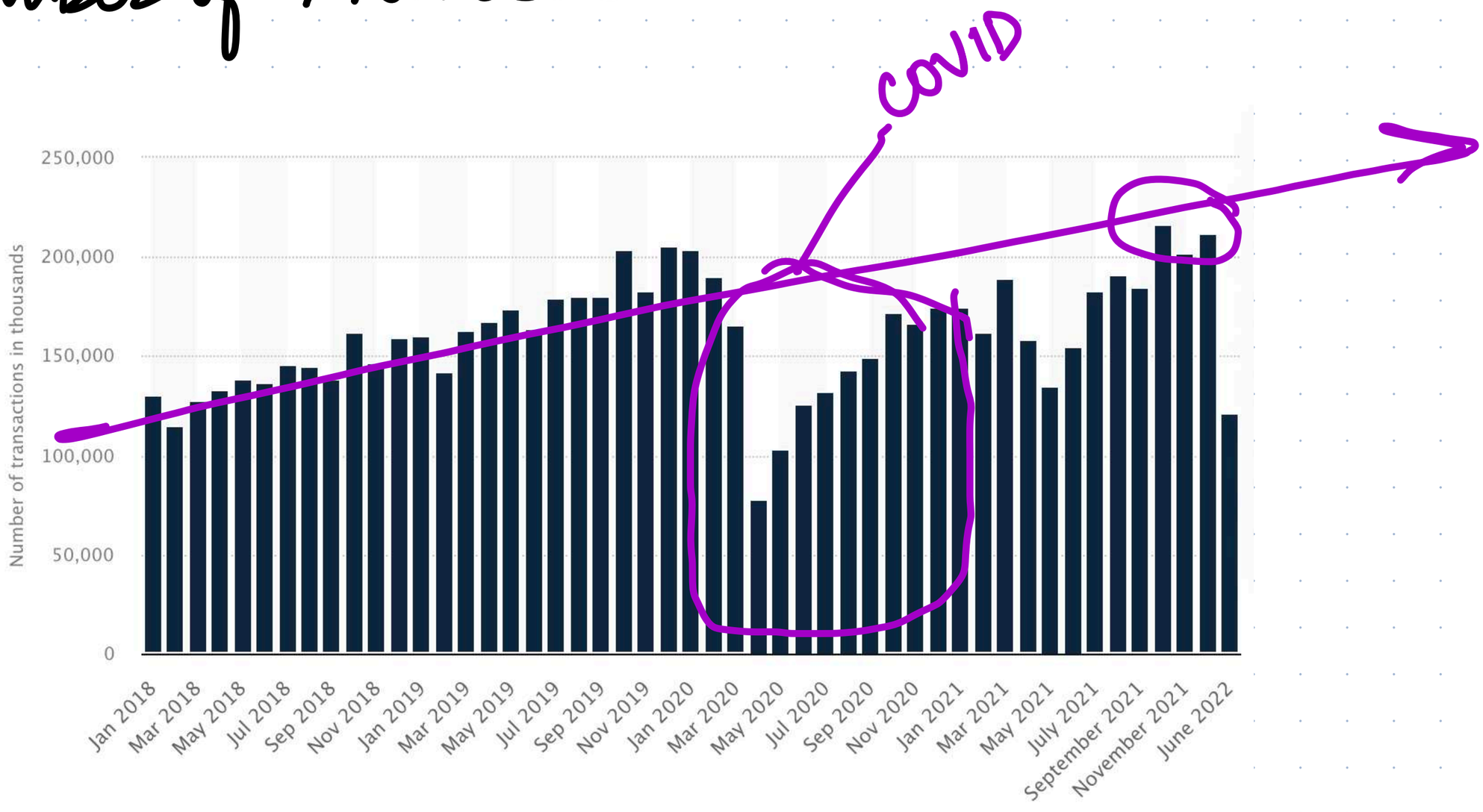
→ No coding


0000 0000 0000
00/00

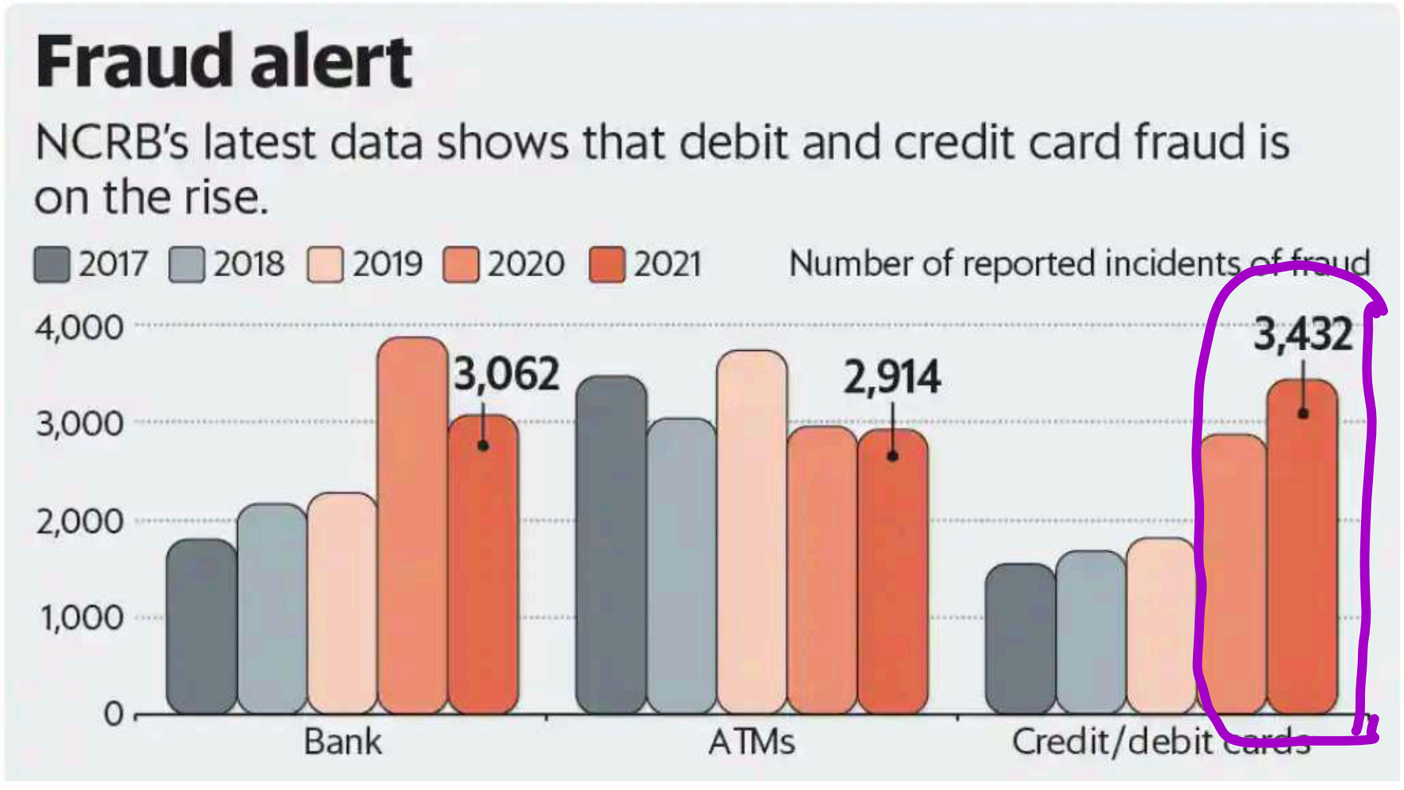
Credit card growth in India



Number of Transactions



Frauds



43%

Card, internet frauds increase to Rs 155 crore in 3,596 cases

Types of frauds

card not present

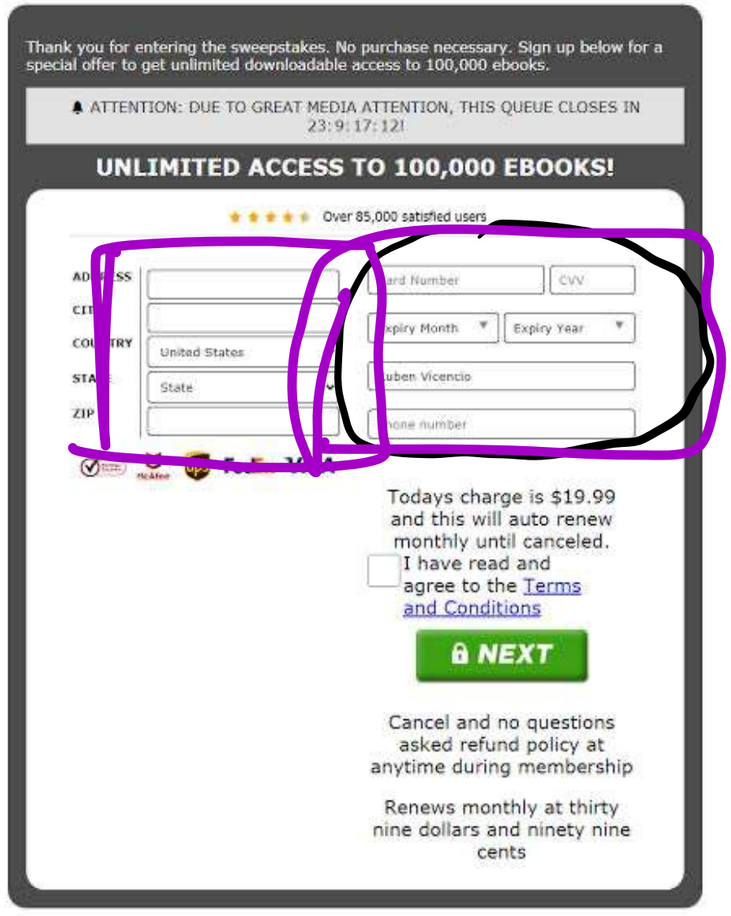
card present

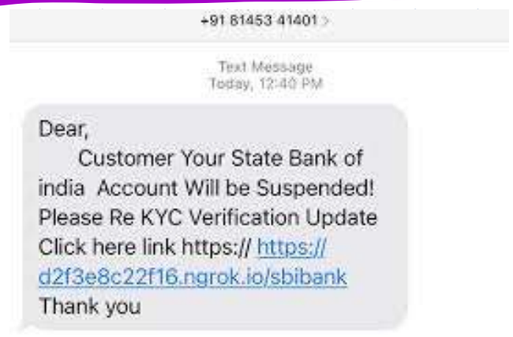
Card not present fraud

Digital or online transaction

Online Phishing

SMS





Email & SMS

Unauthorized Hyperlink

Card Not present fraud

OTP transaction.

Not frequent with gen age group

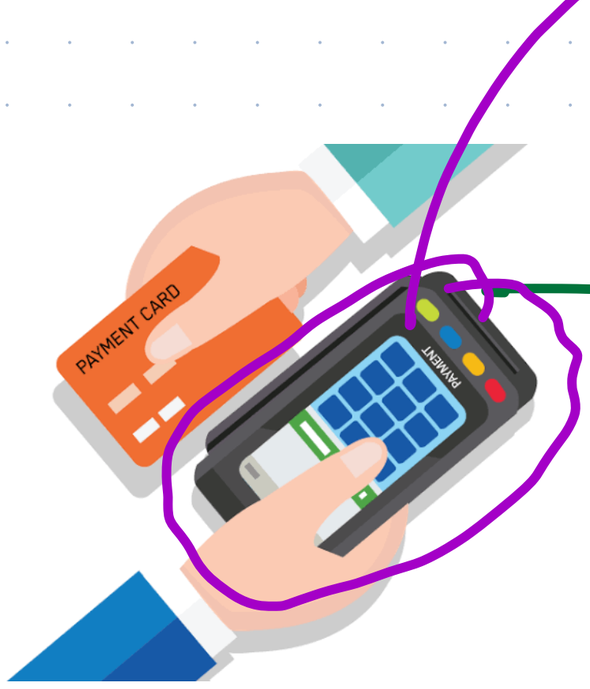
Old Age

Card Present Fraud

physically card is present during transaction

POS Fraud

tamper the POS machine



POS (Point-of-Sale)

card ~~Not~~ present fraud

ATM

skimming

- steal card info & make a duplicate card

Card lost/stolen frauds

Advert

Wireless Tap

Impact of Credit Card Fraud

- Customer
- Financial Institution (Banks/ Intermediary)
- Merchant

= Customer's negligence - Customer

↓

Bank

"Chargeback"

"FINANCIAL BODIES."

"Fraud Detection"

How should we build a

Strong/Robust

FRAUD DETECTION SYSTEM

→ Classical Approach

→ Modern Approach

Classical Approach

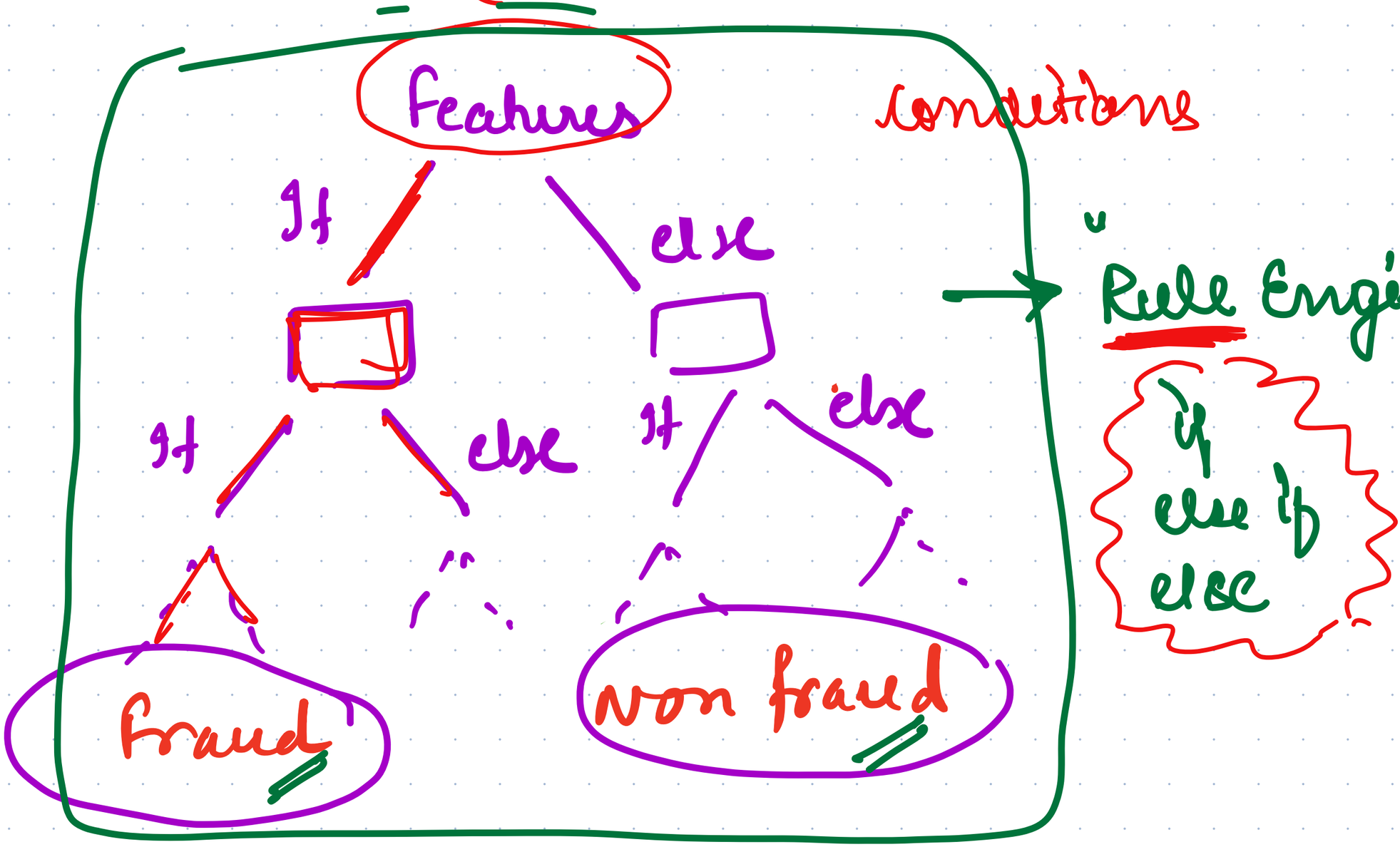
1) Hire a Business Analyst

↓

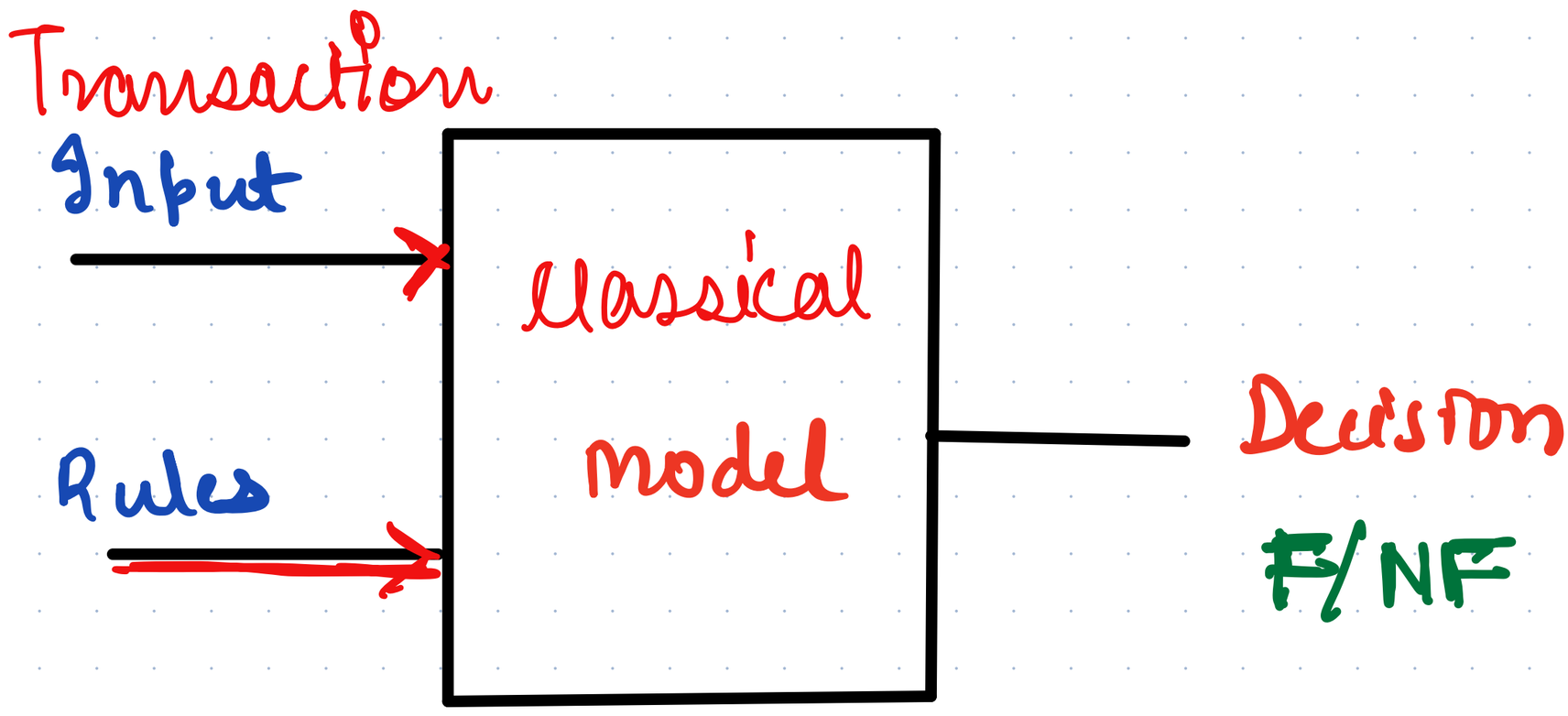
"Patterns" of Fraud
Past transaction.

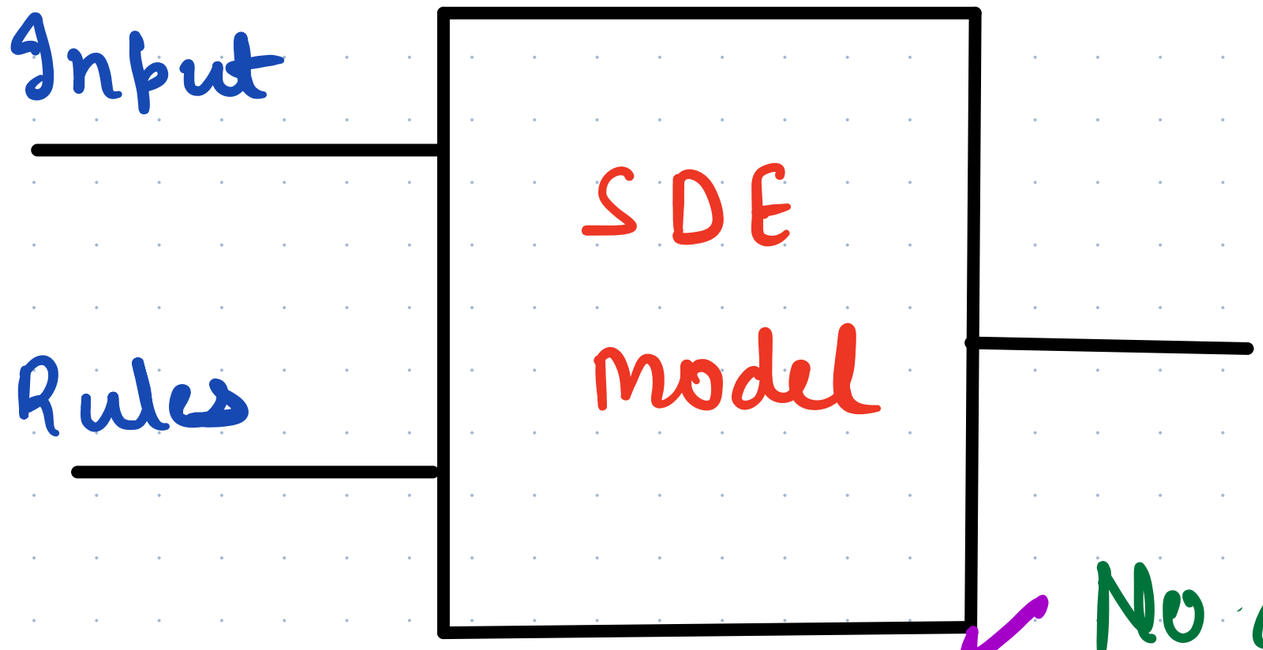
→ location
→ Past transaction
→ Time lapse (Interval)

2) Build a long if-elseif-else structure



Classical Rule based approach





- No adaptibility
- Tedious and Tough.
- Hard coded

Drawbacks?

What strategy?

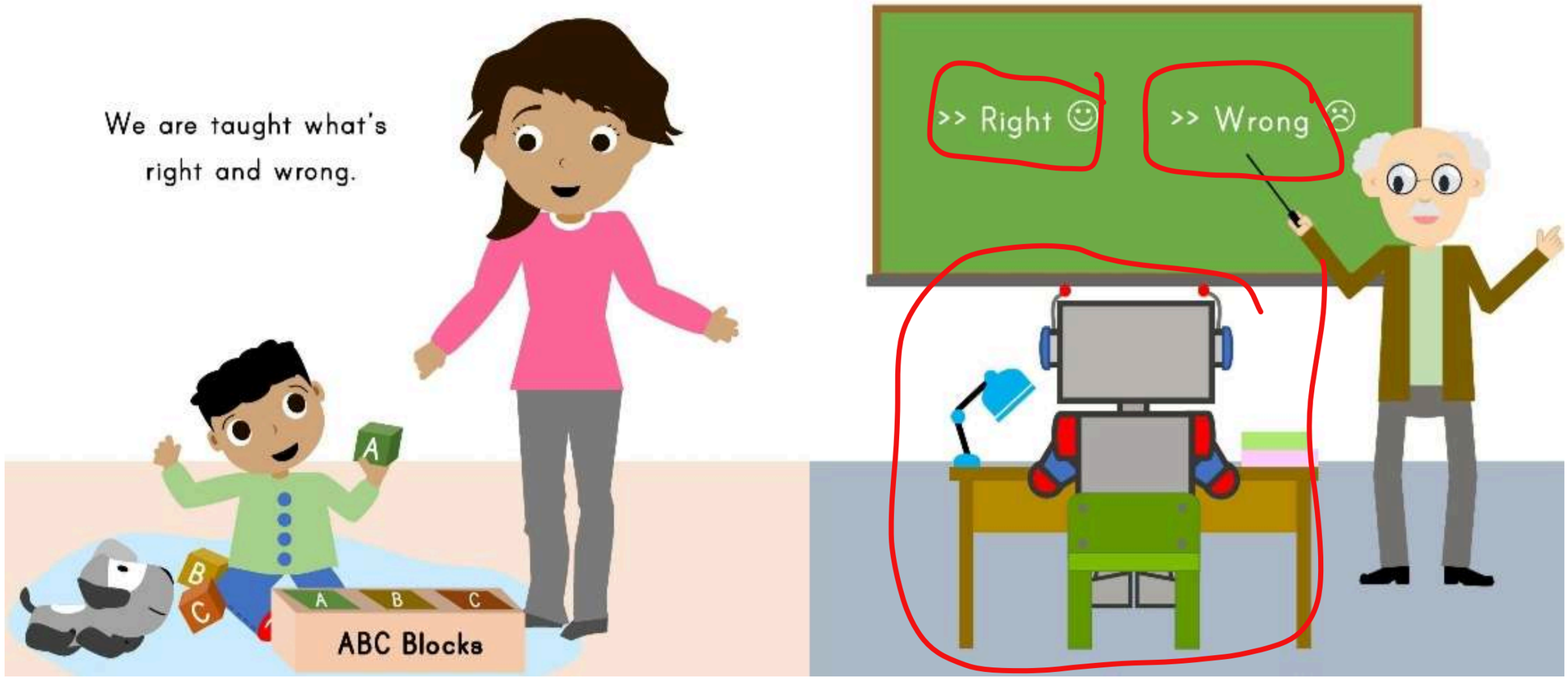
Idea behind ML (Predictive Analytics)

Outsource job of figuring patterns ~~human~~ computer.

Process of training these dumb machines to detect the pattern

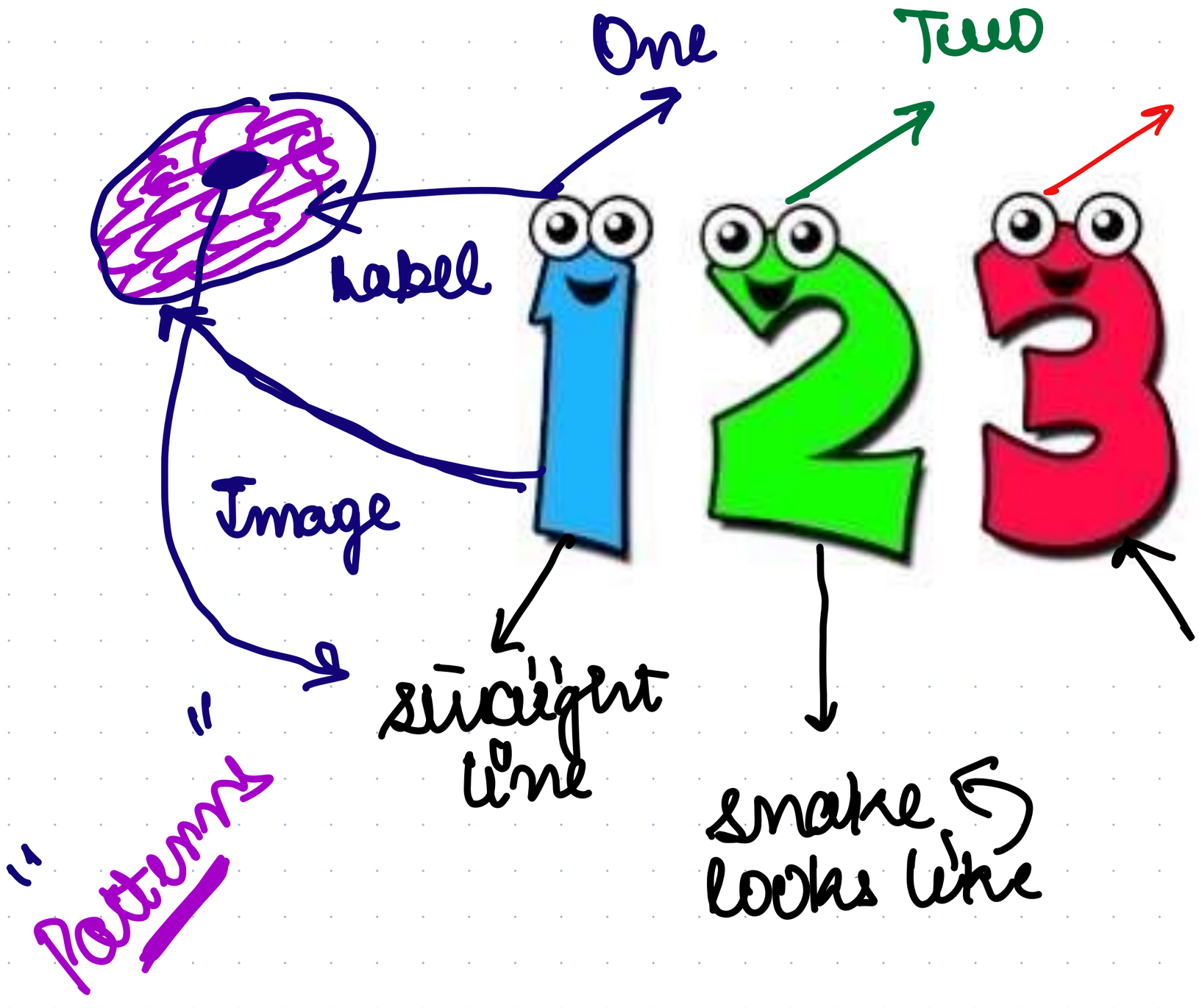
ML

Big Question : How do we train the machine to recognise patterns?

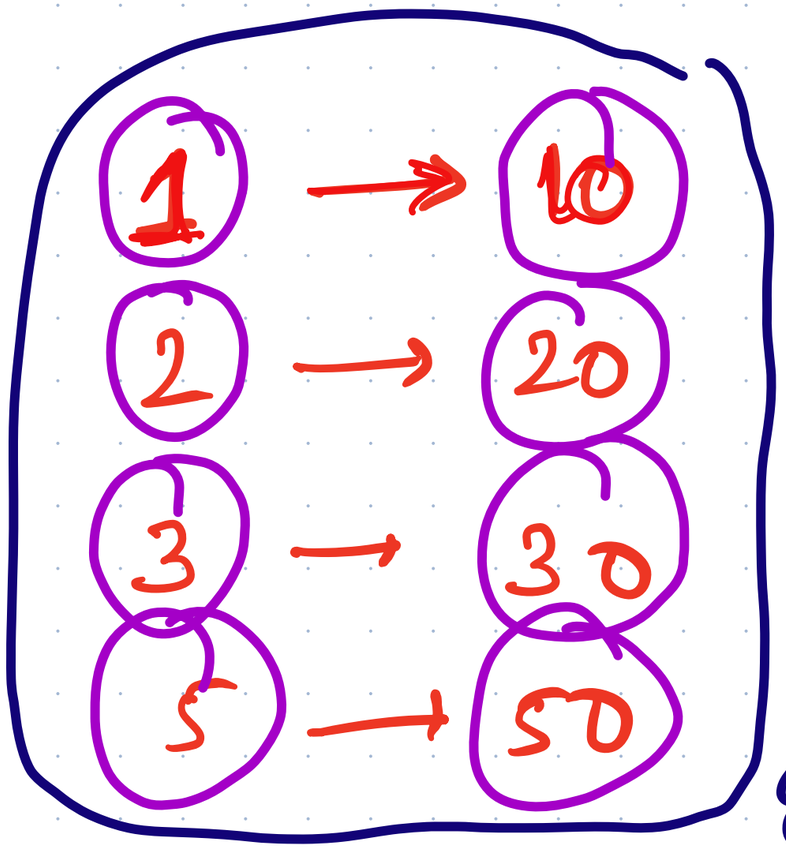


First day at your school

Numbers

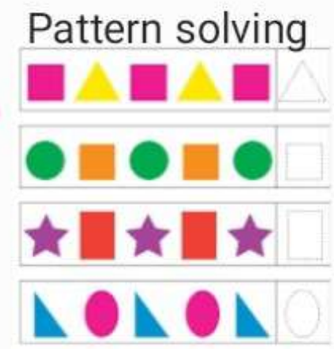

One
Two
Three
Label
123
Image
"Pattern"
straight line
snake looks like
Grandma's eye glass
DIET

1 → One.



Train my machine using some "training" data

Expert



Machine figure out some patterns.

$y = 10x$

$4 \rightarrow ?$ Guess the output?

40

Transaction

Data →

Rules →

Classical Approach

→ Output

F/NF

Data

Rules

F/NF

Transaction

Data

(Input) →

Labels →

N|F

Training

ML

→ Rules

ML Trading Pipeline

ML Modelling Pipeline

"Exploratory" Data Analysis Feature Engineering

lot of data → ~~Direct Modelling~~

Right Data

Finance Expert

Insights:

- Sanity checks (outliers/missing data)
- Determing certain pattern.

FE

→ 1 user OR Multiple users.,? #

→ Variation in the users - generalic model.

→ Months (Every) and for some years.

Machine learning

Modelling - train the ML Model

Training Data

Traditional 80% 20%

Million 99% 1%

Data & splitting Train Test

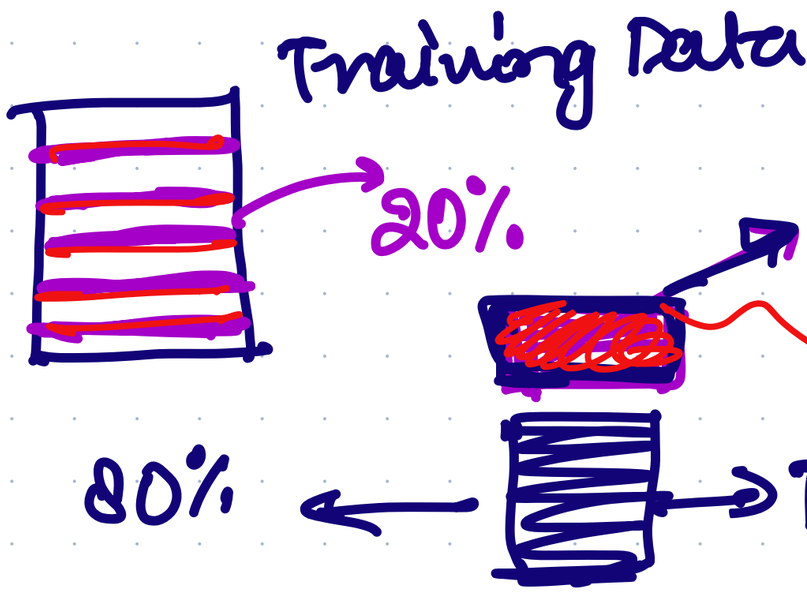
Validation

Trained Model

Deploy

Test

Can I use a proxy for evaluation data?



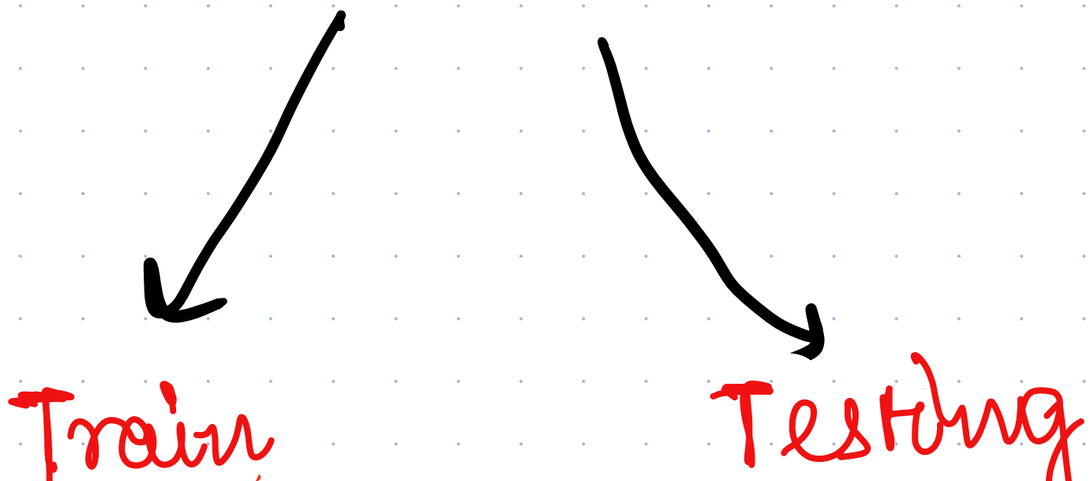
New data (labelled) F/NF

unseen during training process

Testing.

Evaluation

Split the data



This part will go into ML model

We can use this to test model

Feature Engineering ***

Buy a flat. cost of the flat.

- # bedrooms
- carpet area
- locality
- views
- floor num

features

Predict the price of your flat.

Q: What features do banks collect "when you make" credit card transaction?

Features.

Raw
- → Card Number
- → Name on the card
- → Transaction Amt
- → Transaction ID
- → Time
- → Geolocation.
- → Merchant Code.
- → Type of merchant
- → Transaction limit
- → Currency
- → CVV
- → Pin

Raw Features

enough? → NF/E

↓

NO

→ History
→ Usage Behavior.
→ Credit score.
→ Debts.
→ Payment method

User Behaviour/ Characteristics/ lifestyle

Why user's behaviors are important ?

Smart

User

does → Transaction (9-5)

Behavior.

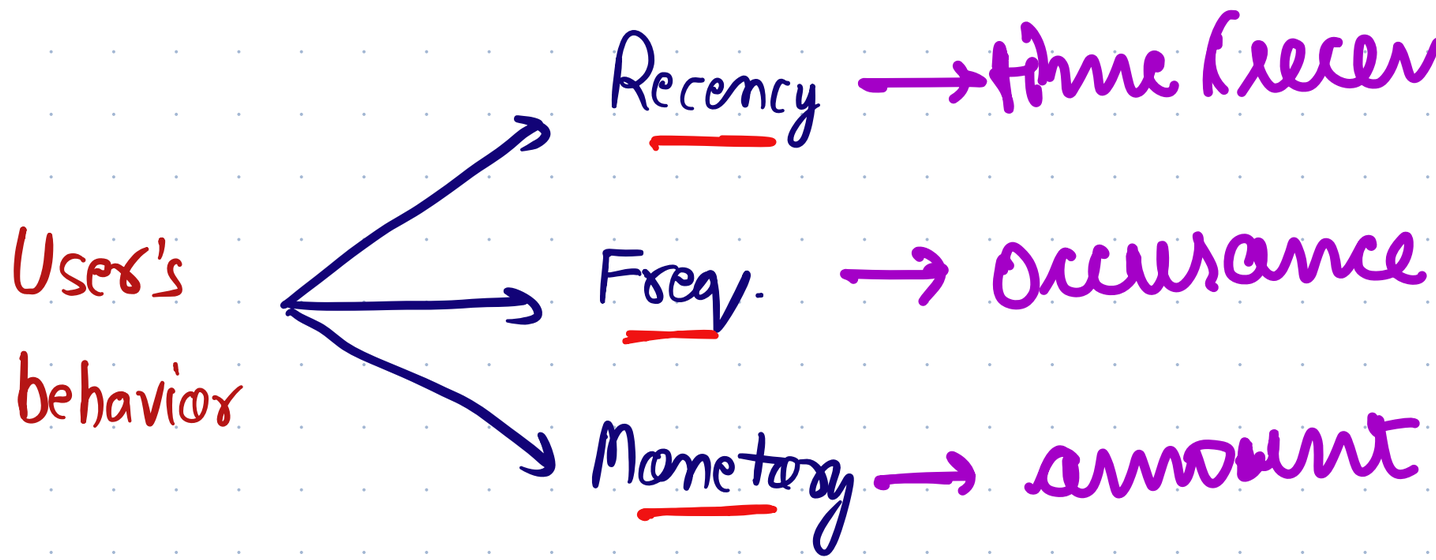
Transaction-N (midnight) → Fraud.

Q.: Features capturing user's behavior habits?

- → Frequency
- → Avg. amt of transaction
- → Time lapse.
- → Usual timing
- → Cyclic Trends
- → Merchant Type
- → How many attempts.
- → Preffered mode of payment

- Recency
- Frequency
- Monetary

capturing user behaviours

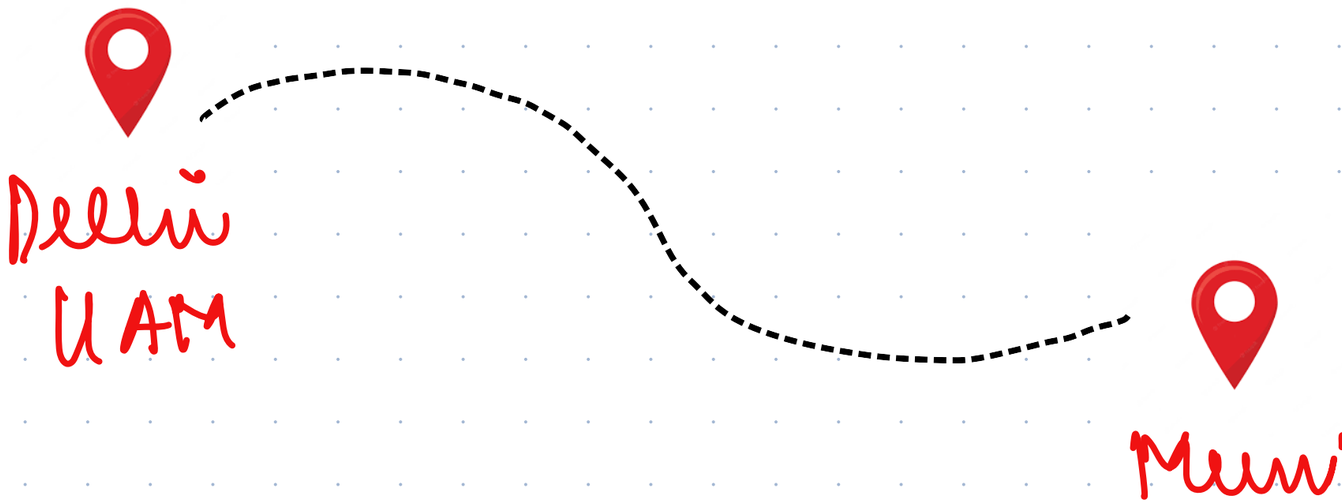


Q: Any suspicious transaction?

	Transac. Id	City	Time	Transac method	Amount
1	T420	Delhi	10:20	Card	100
2.	T013	Delhi	11:00	Card	40
3.	T169	Mumbai	11:20	Internet	200

Q: Feature that involves location & recency

Time



Engineer Feature

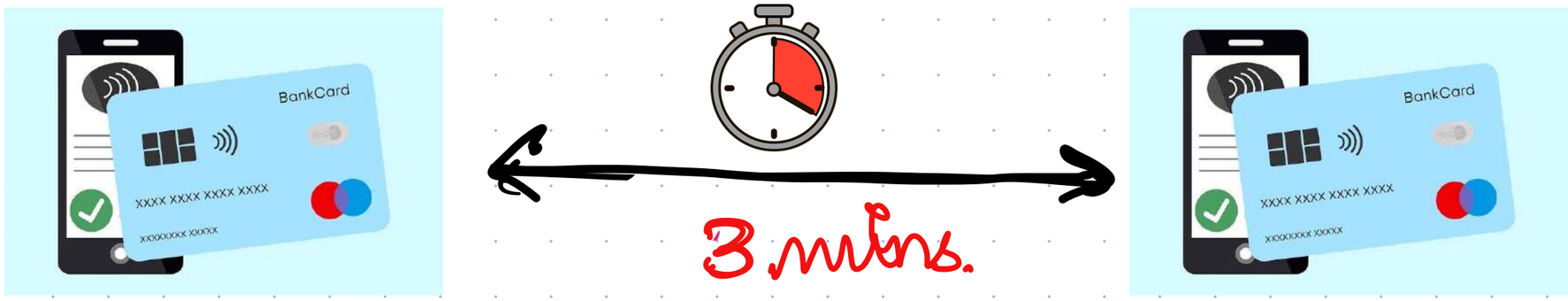
→ Velocity = $\frac{Lx - L(x-1)}{\text{time}}$ → $\frac{\uparrow}{\downarrow}$ = ↑↑↑

11:20AM

F

Recency

Gap between transaction



Transaction - 1

11:00 AM

large Amt.

Transaction - 2

11:03 AM

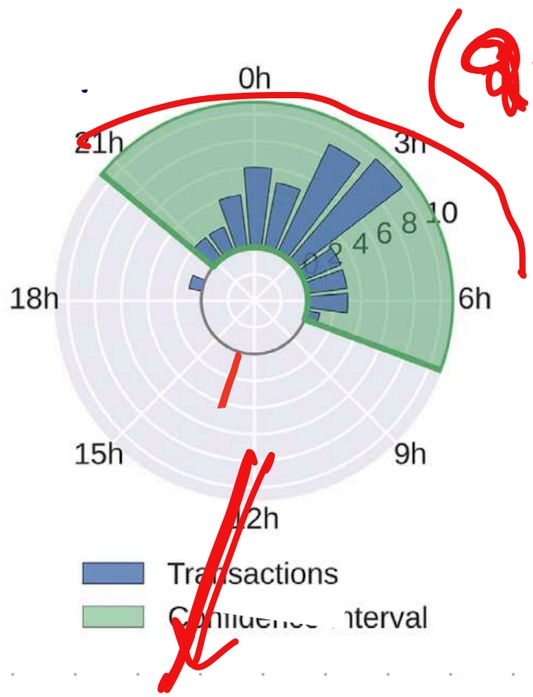
large Amt

→ Multiple large transaction in (less) time

I/O ⟸ F/NF

Unusual time of transaction

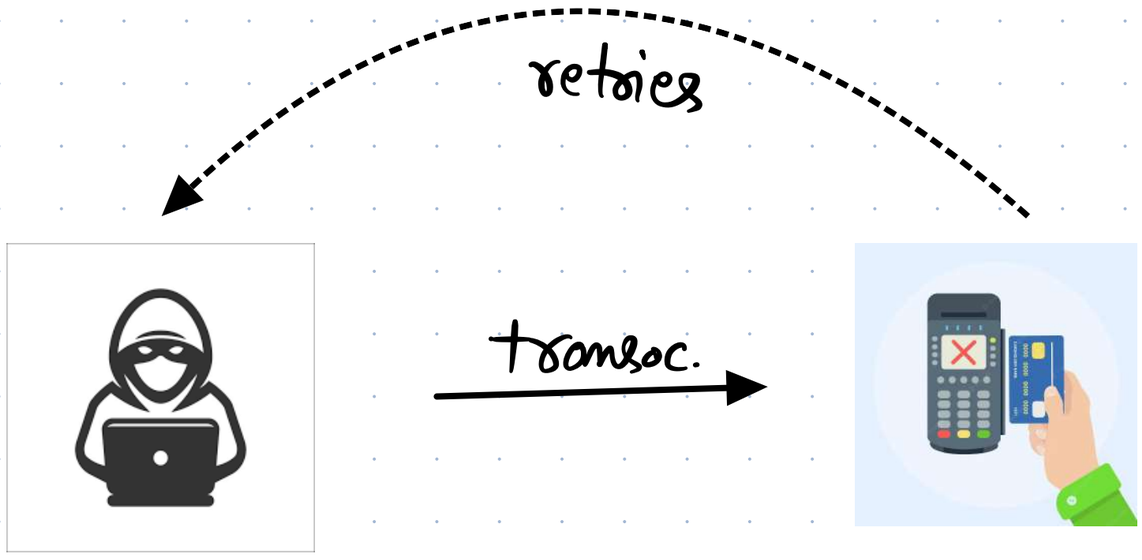
(q-5)



Frequency

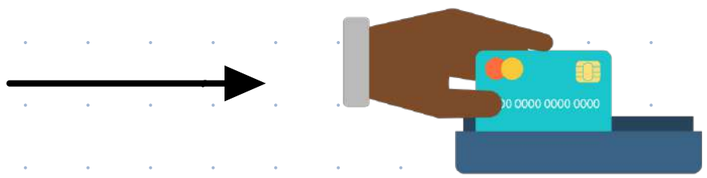
- Transaction features involving freq.!

failed transactions



of failed transaction to merchant

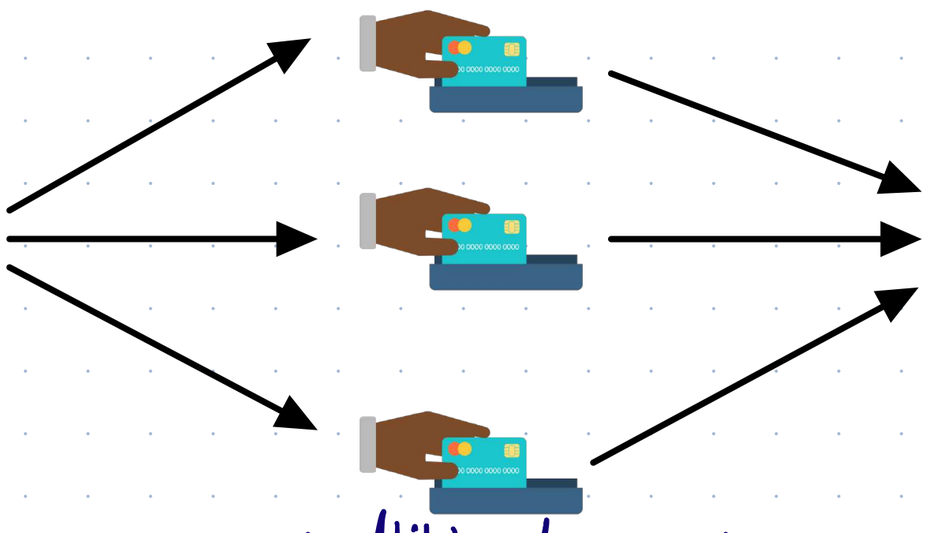
user



single transaction

merchant

fraudster



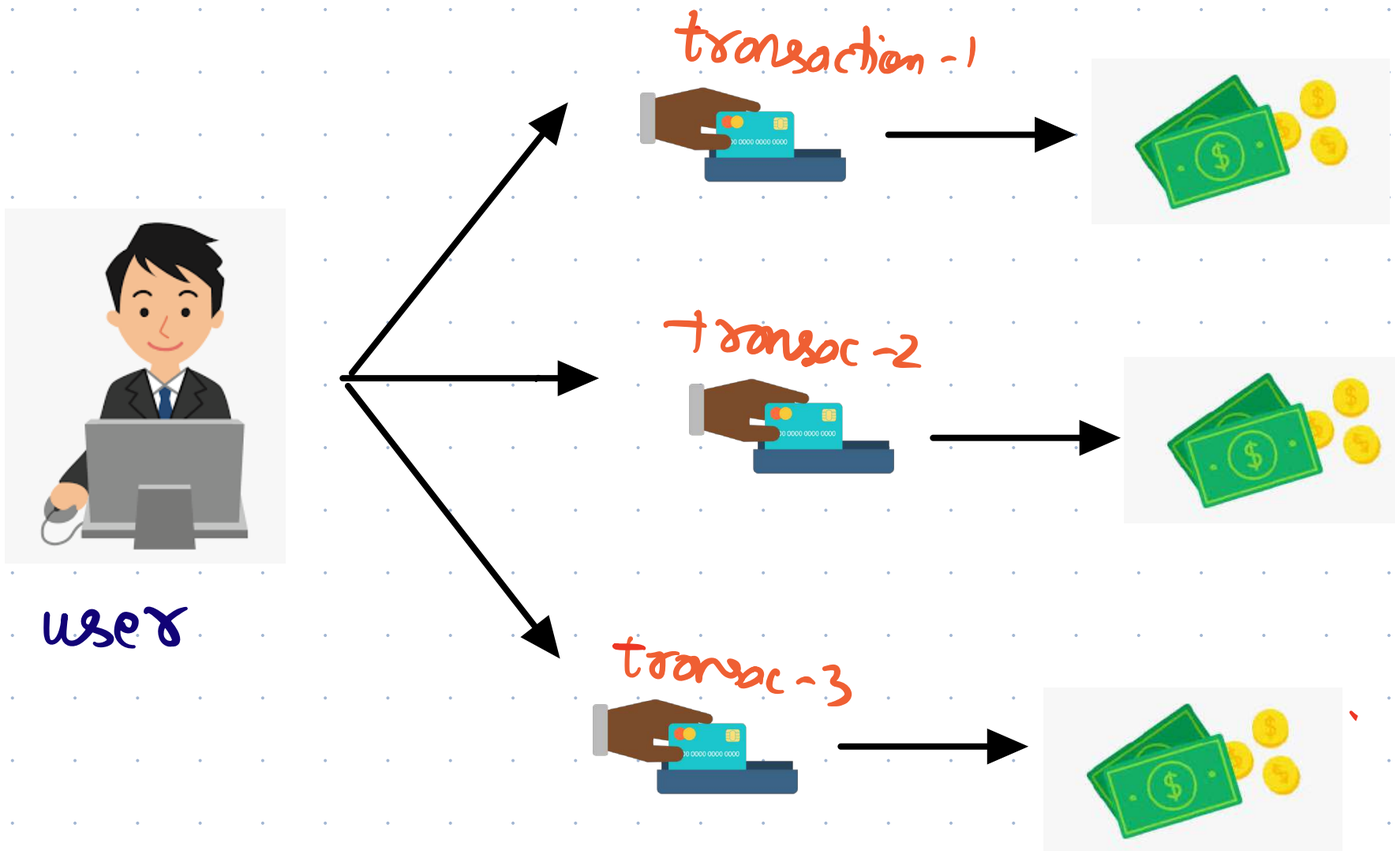
multiple transactions

merchant

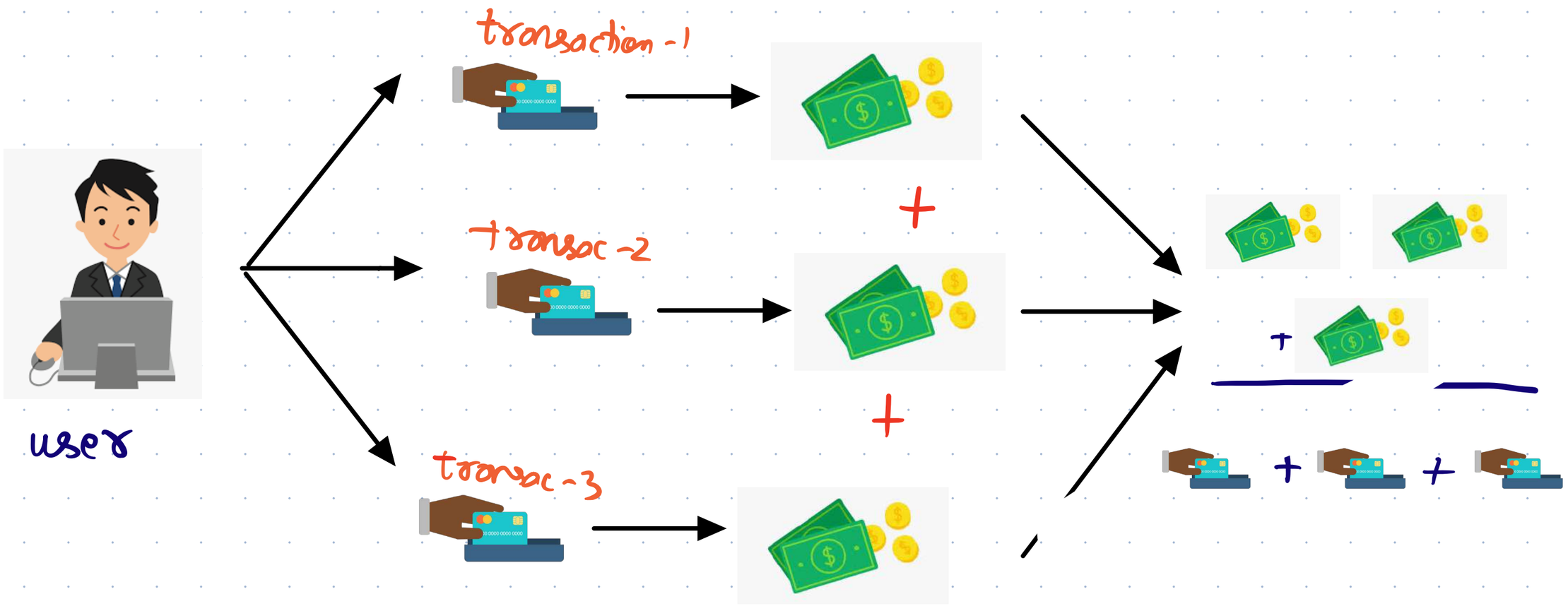
Monetary

Features involving transaction amount!

1.1 Cummulative transaction amount in a day



1.2 Average Amount spent in



Big one time transaction

elder

large transaction

targets

fraudster

3. Currency type of transaction

Indian user

usual transaction in Rs.

Suddenly,

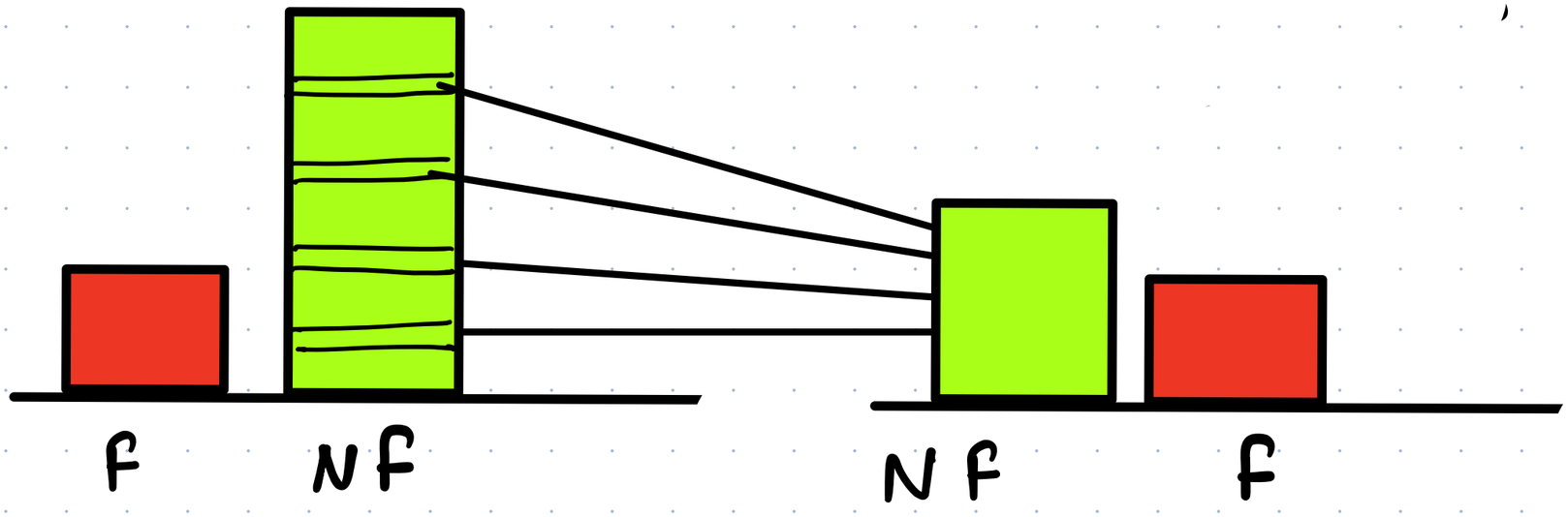
transaction in yuan

Is our data ready for modelling?

Non-fraud >>> #fraud (1%)

Re-Sambling

Under-Sampling (Reduce the majority)

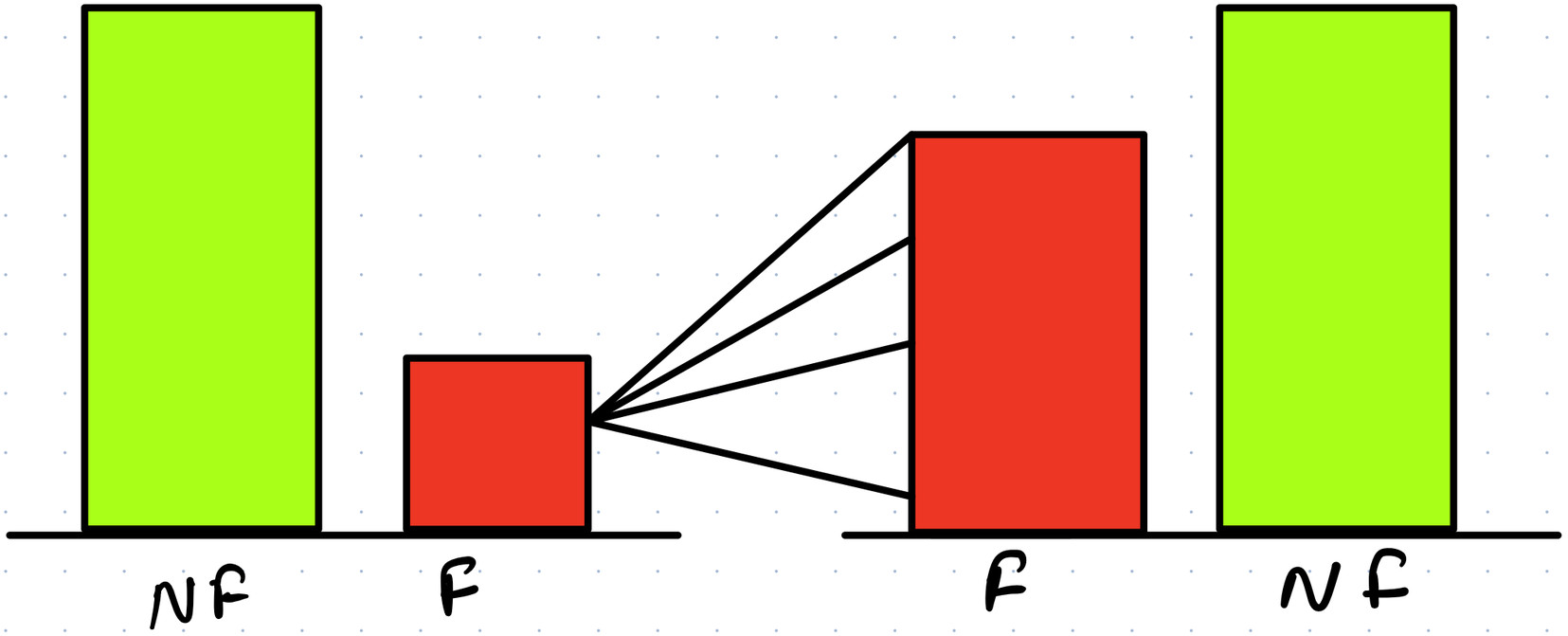


Original

Under-Sampled

% difference is less now

Over-Sampling (Increase the minority)

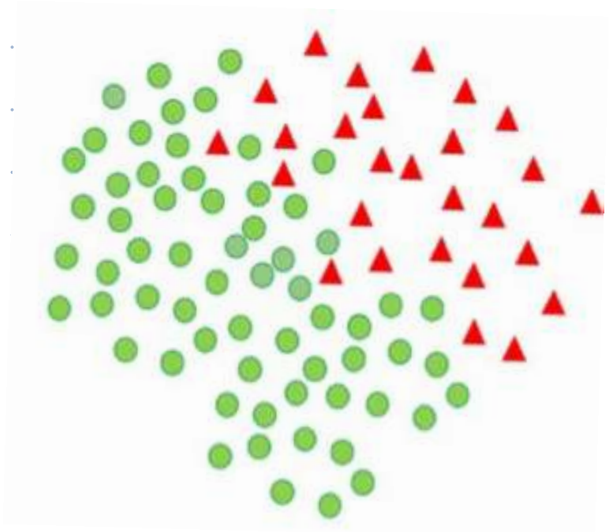


How to increase fraud transaction?

SMOTE

Δ → minority Class

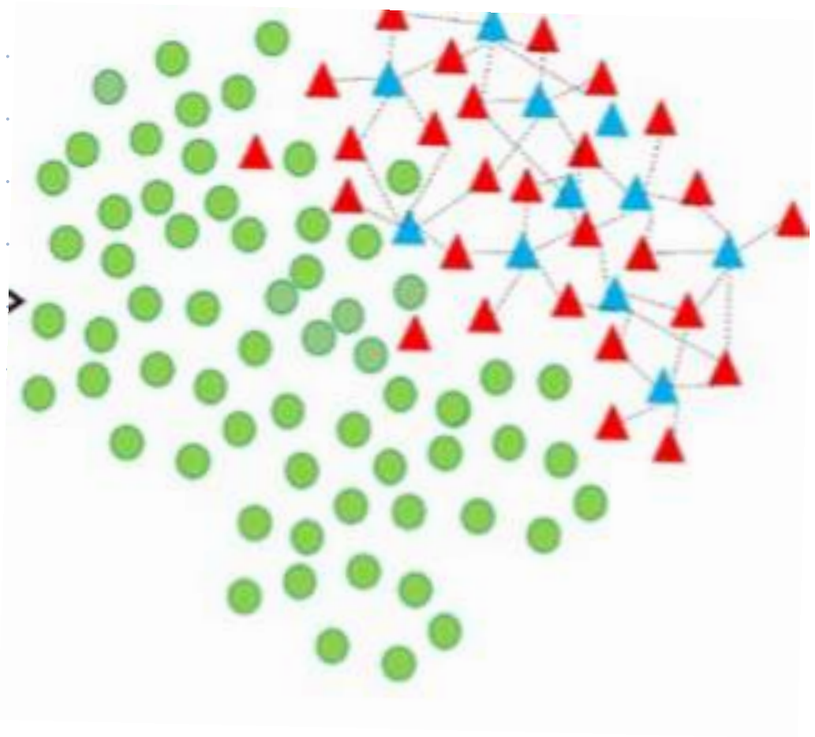
O → majority Class

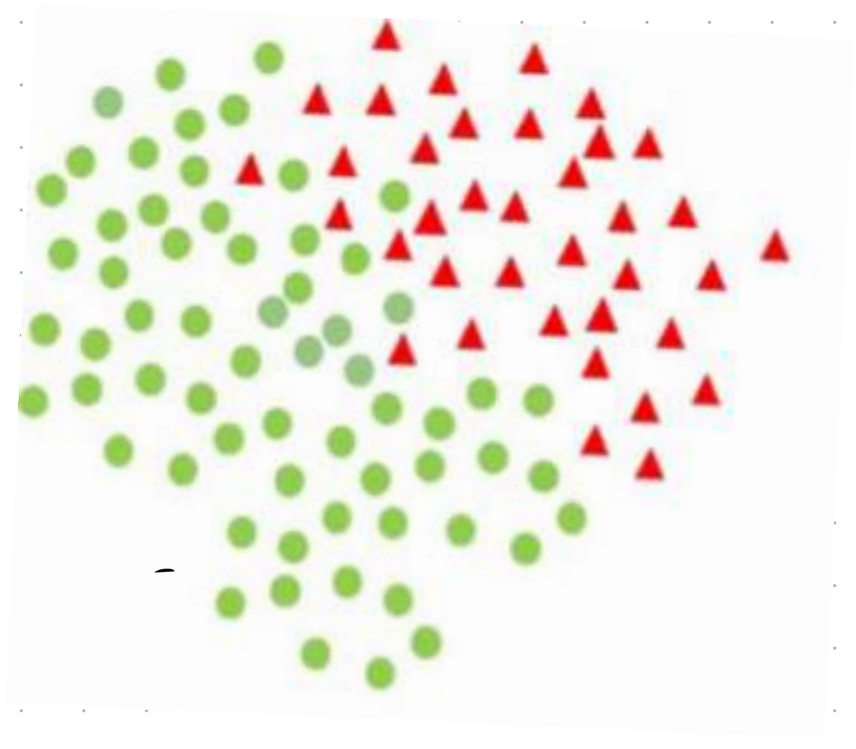


Imbalanced Dataset

How we can increase minority Class ?

▲ → synthetic minority class data points

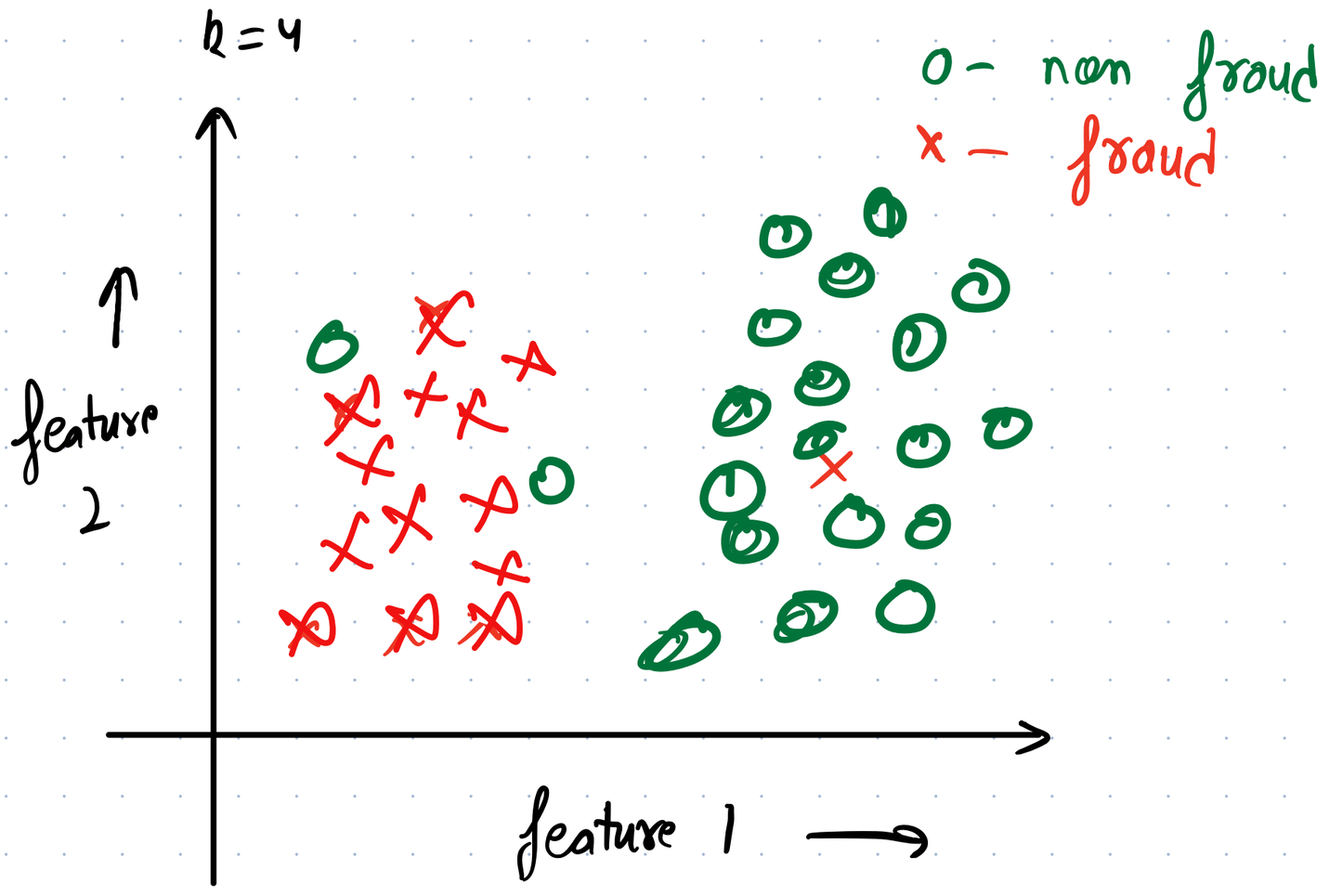




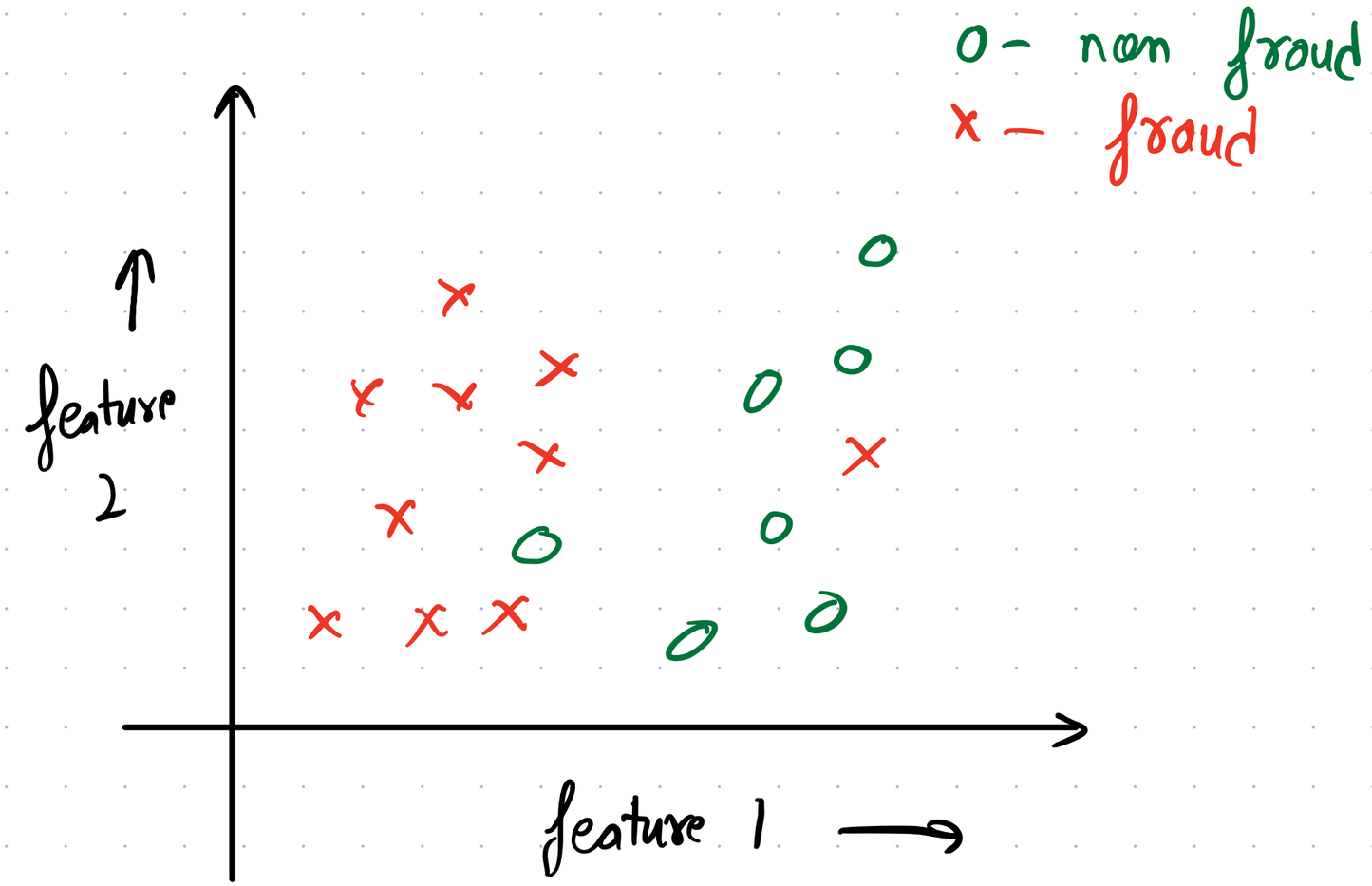
ML Modelling:

What do we want to predict?

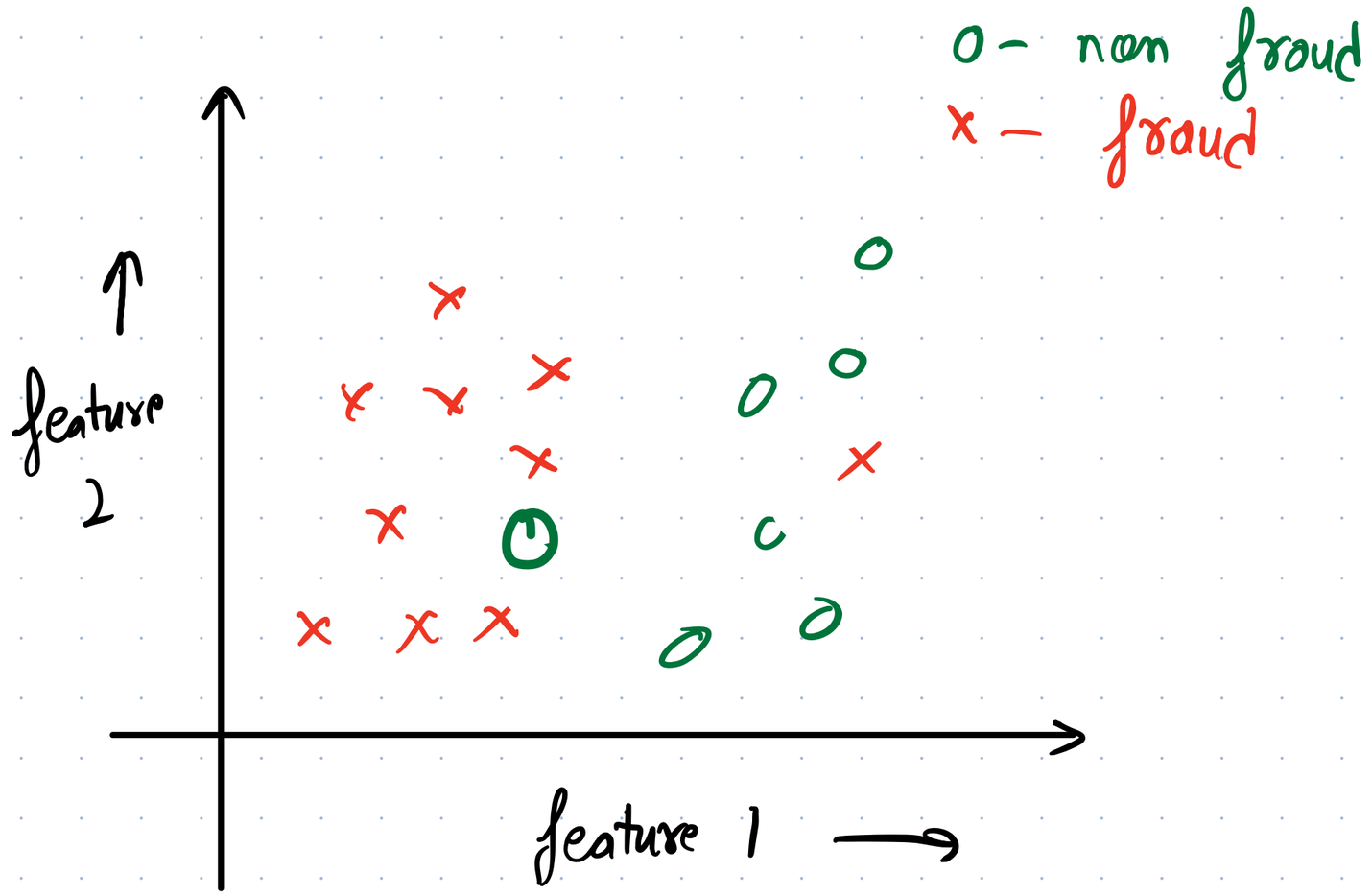
KNN

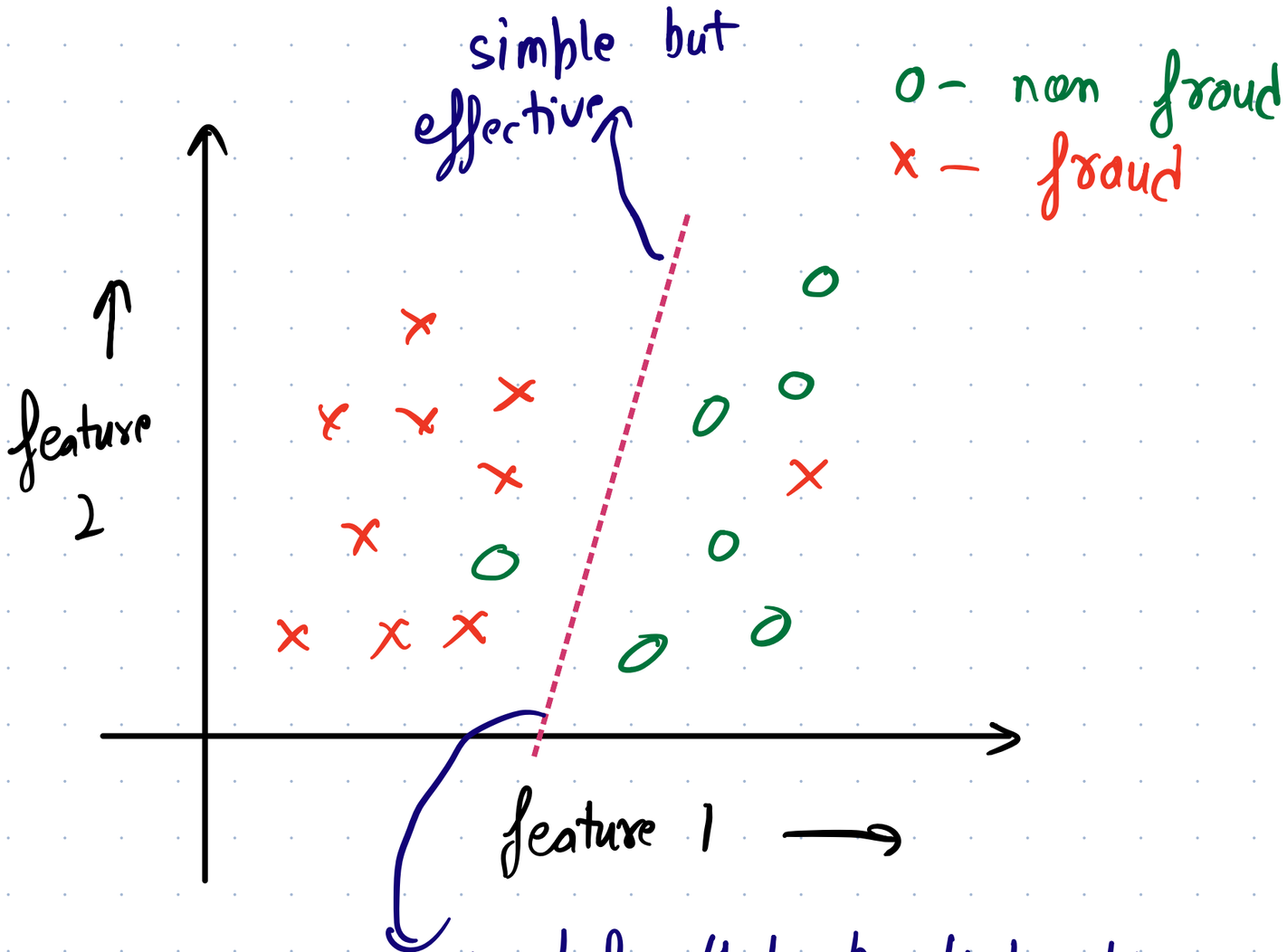


Logistic Regression



More complex boundaries

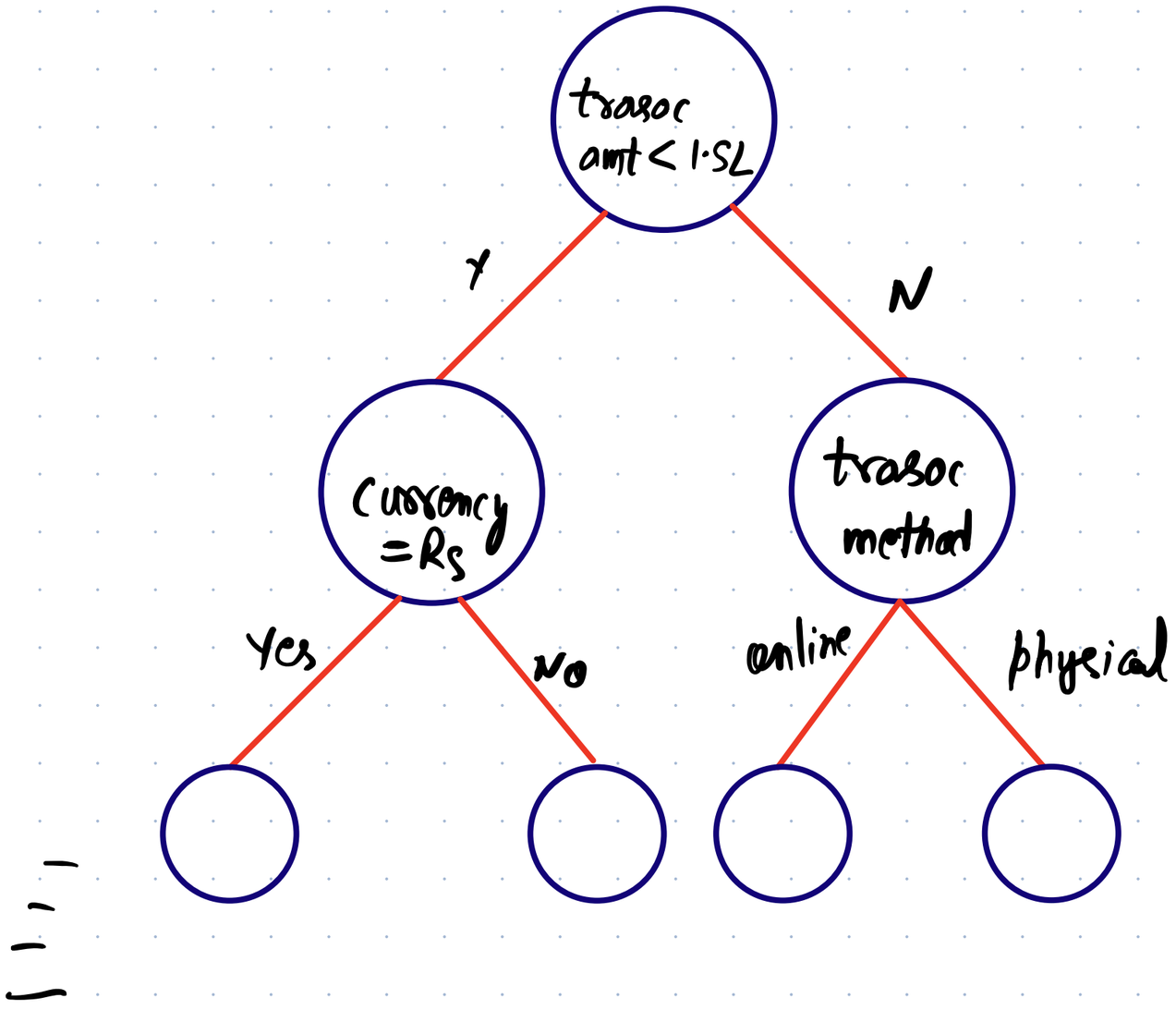




model that predicts linear boundary

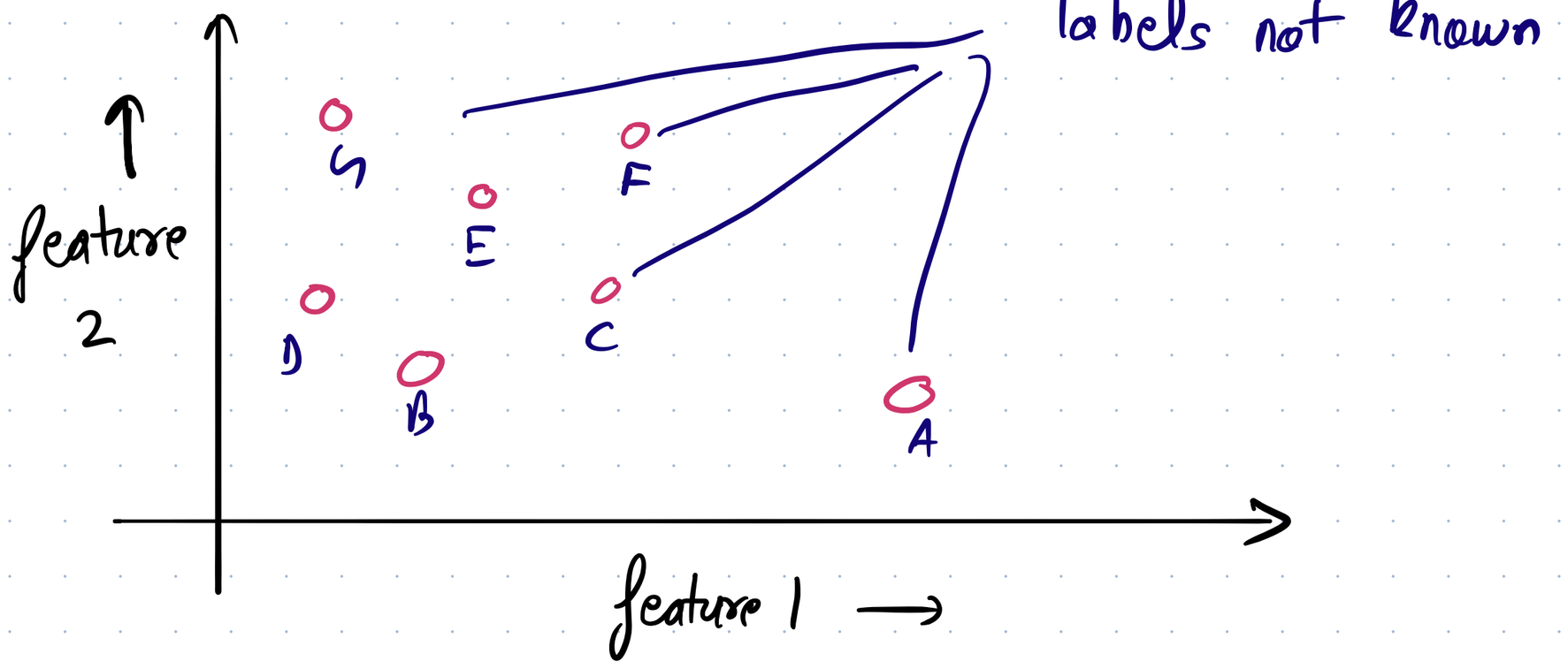
⇒ Linear regression

Decision Tree



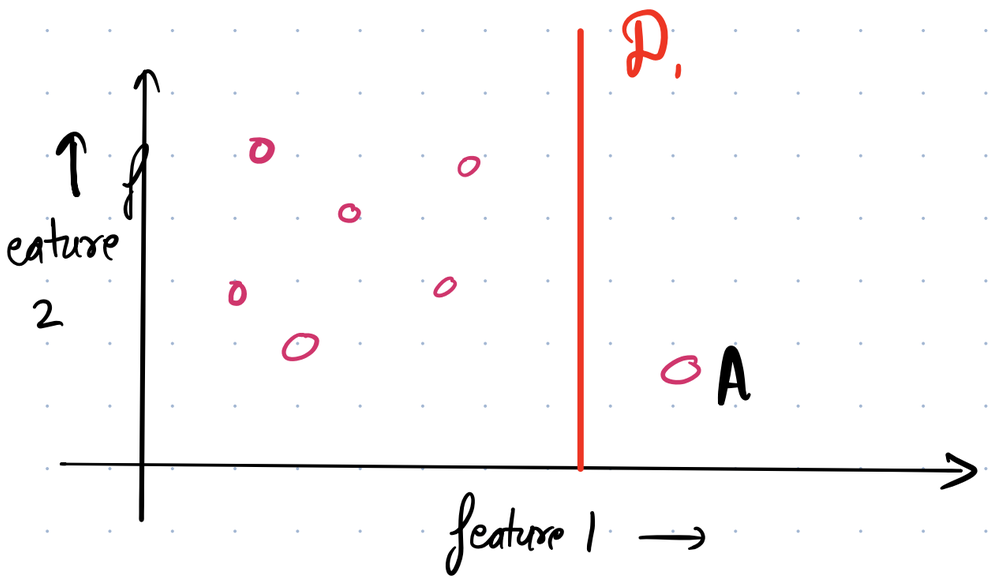
Isolation forest

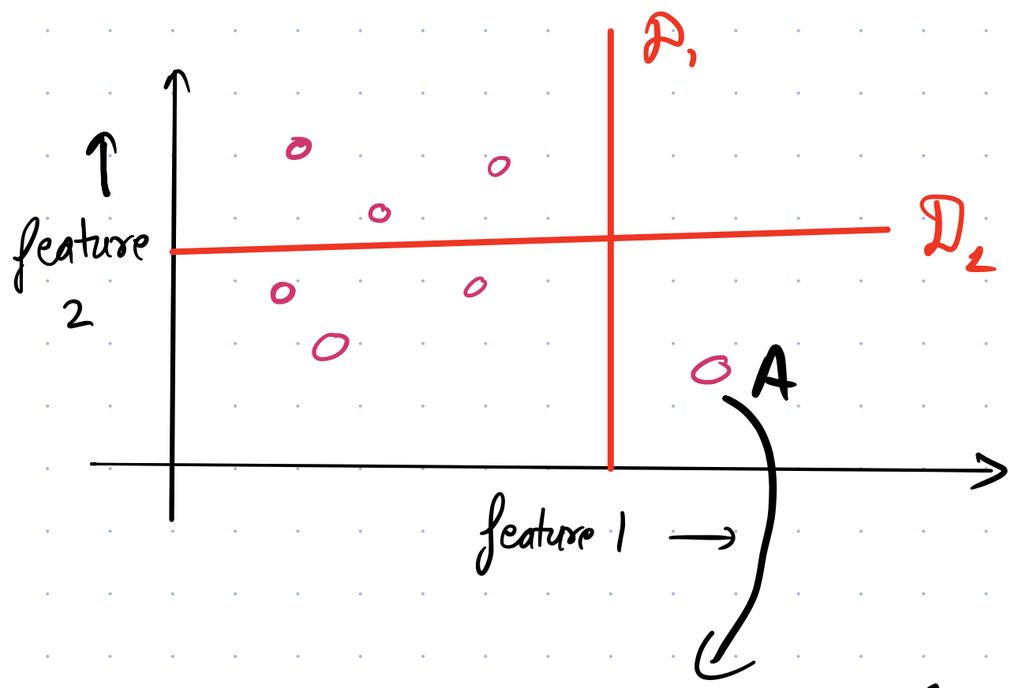
- find datapoints that are very different from rest of datapoints

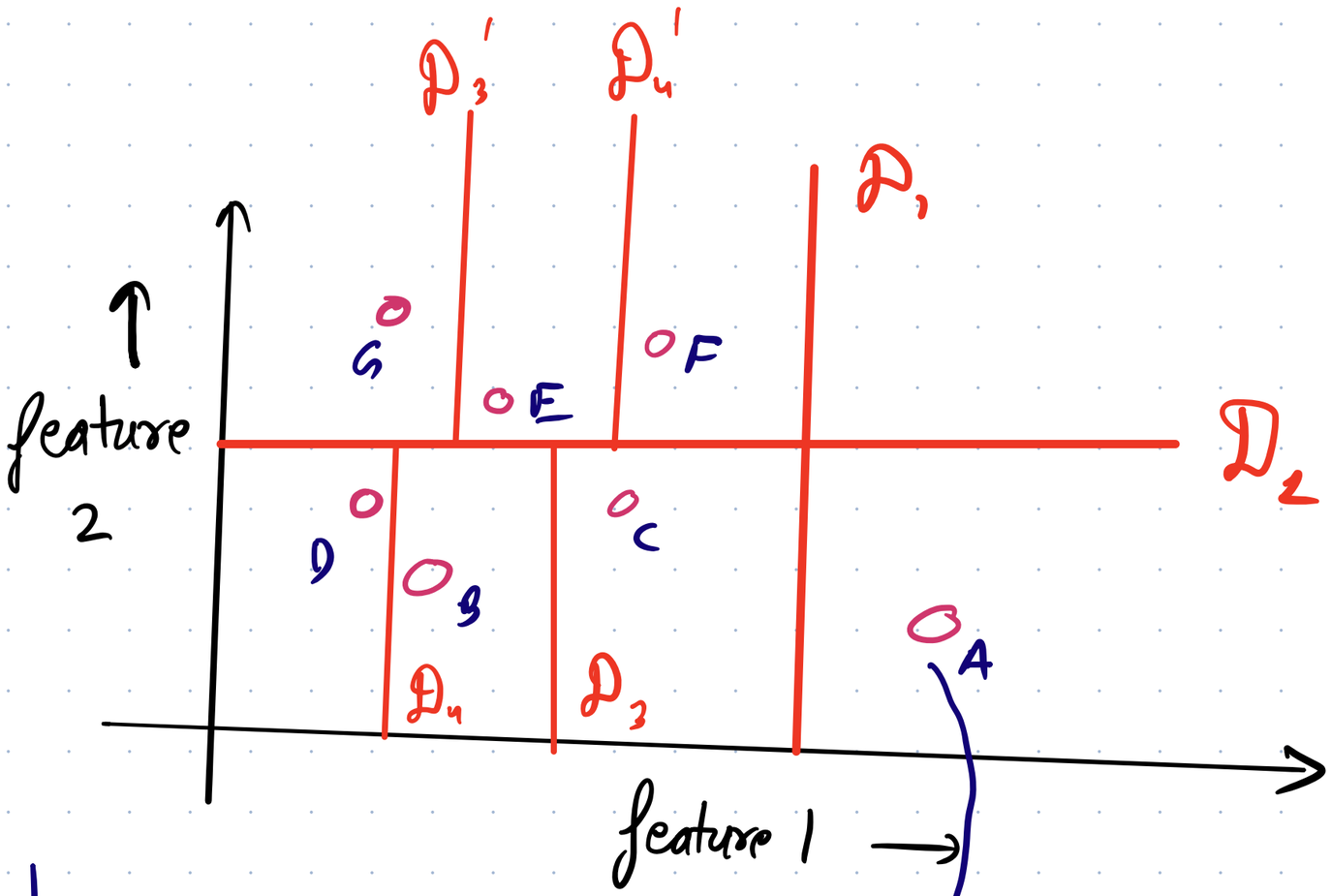


How do Isolation forest work?

— isolate each datapoint from every other point







fraud ← different from other points ← isolated earliest

Model Evaluation

How do we know model is good ? ⇒ Check Accuracy

Test data ⇒ Model ⇒ 99 % Accuracy

1 out of 100 prediction is wrong

↙

Is this a Good Model ?

Non Fraud >>>> # Fraud

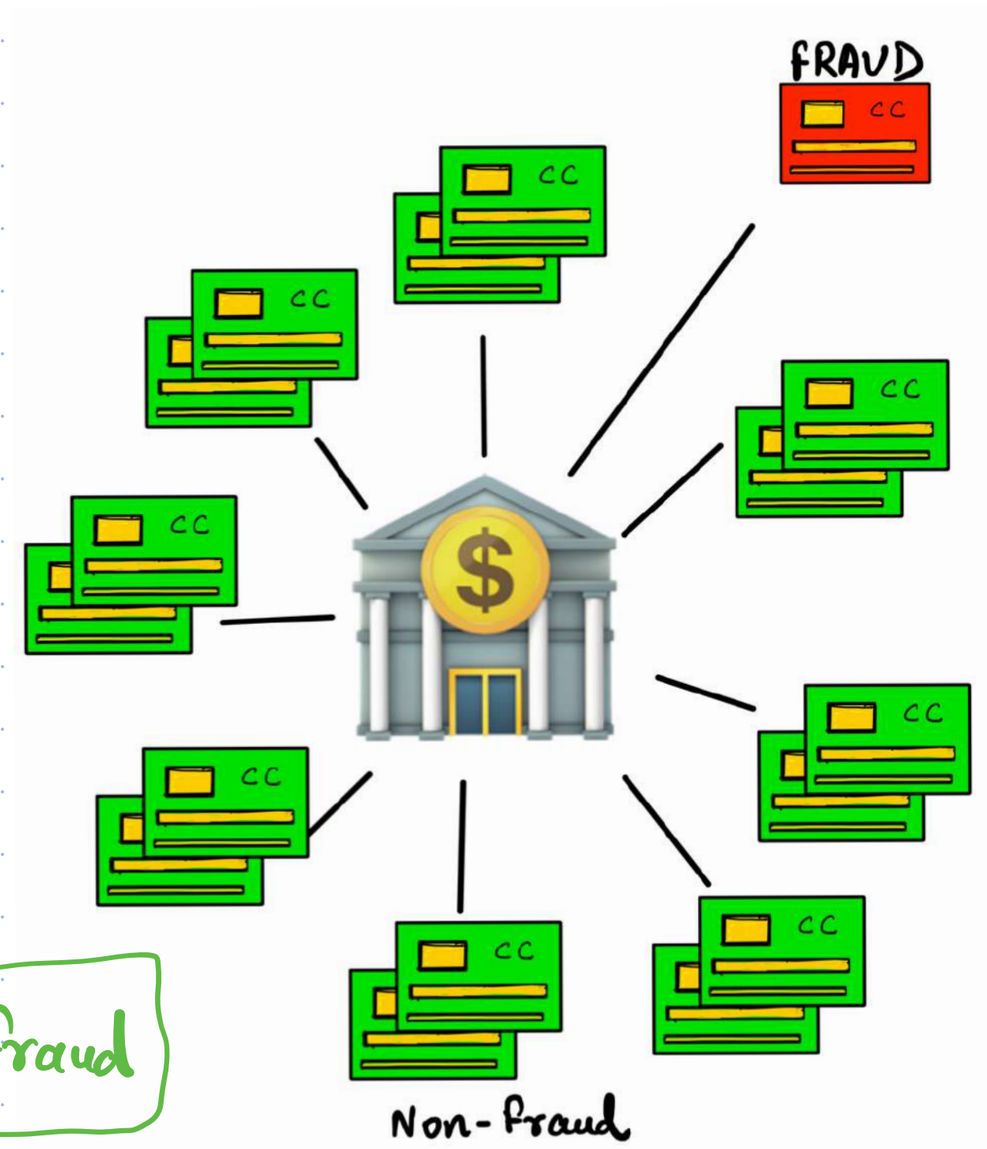
(99:1)

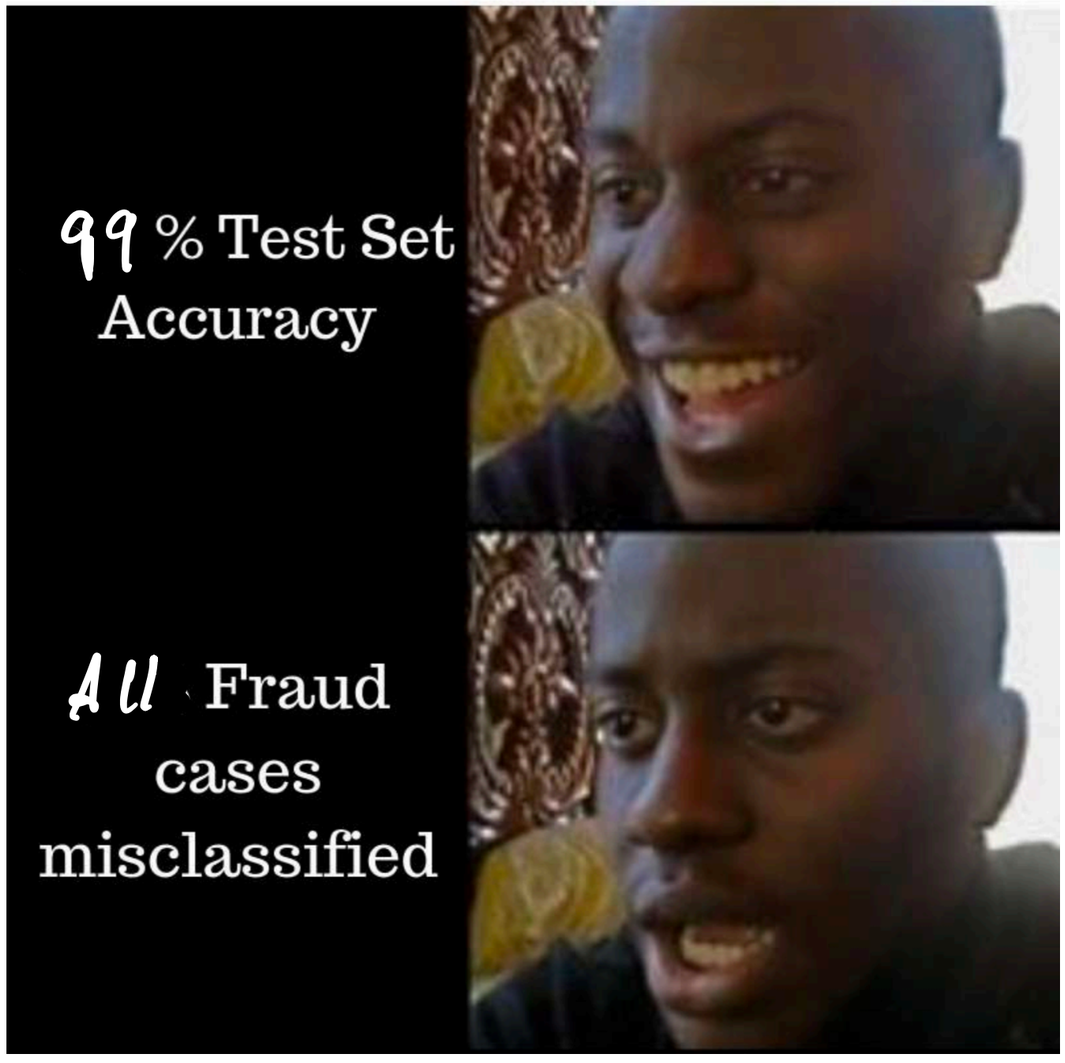
Suppose 100 Transaction

99 Non-fraud

1 Fraud

What If model predicts every Transaction as Non-fraud





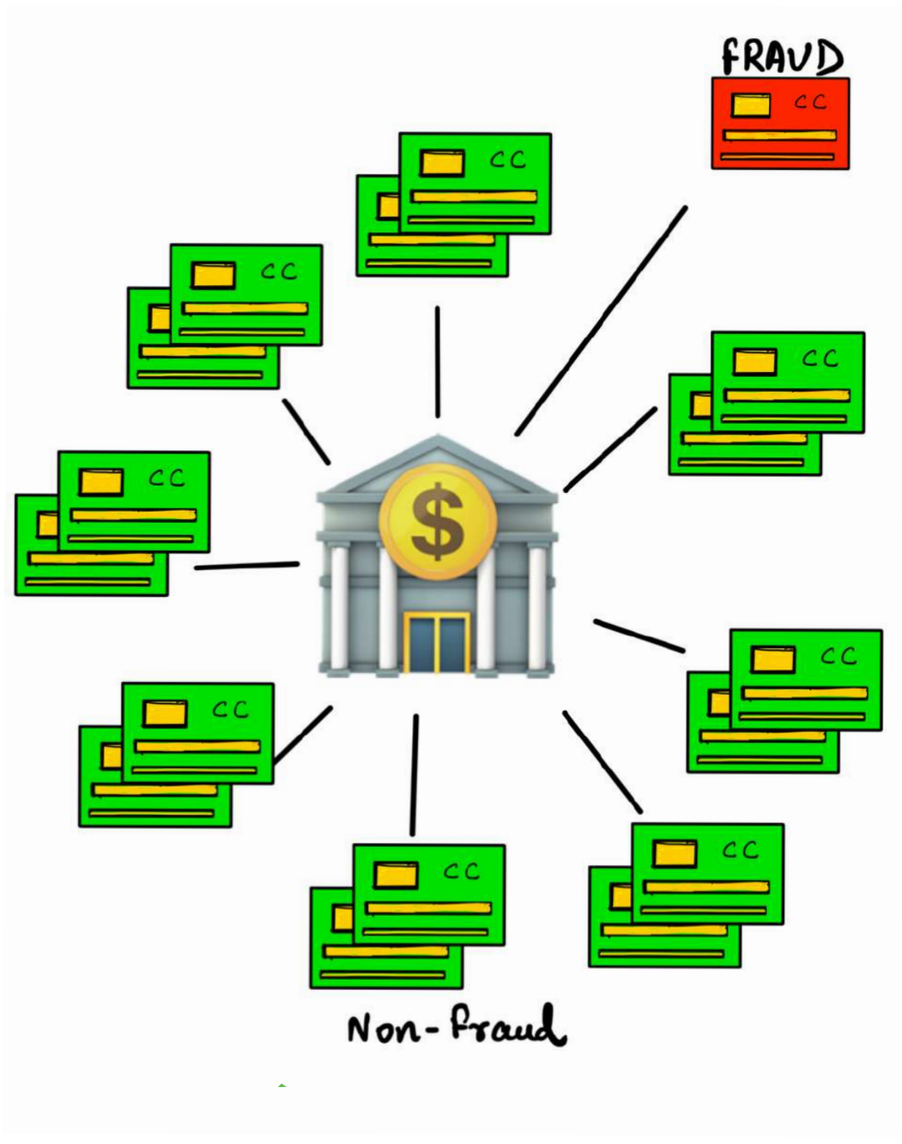
What is Accuracy ?

$$\Rightarrow \frac{99}{100} = 99\%$$

Is it Reliable ?

All fraud
misclassified

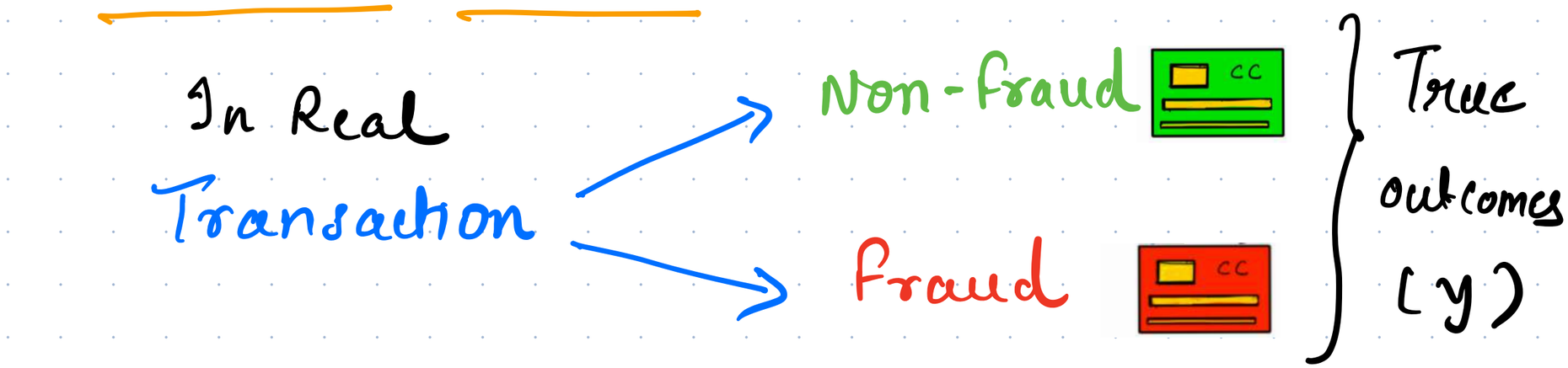
Imbalance Dataset

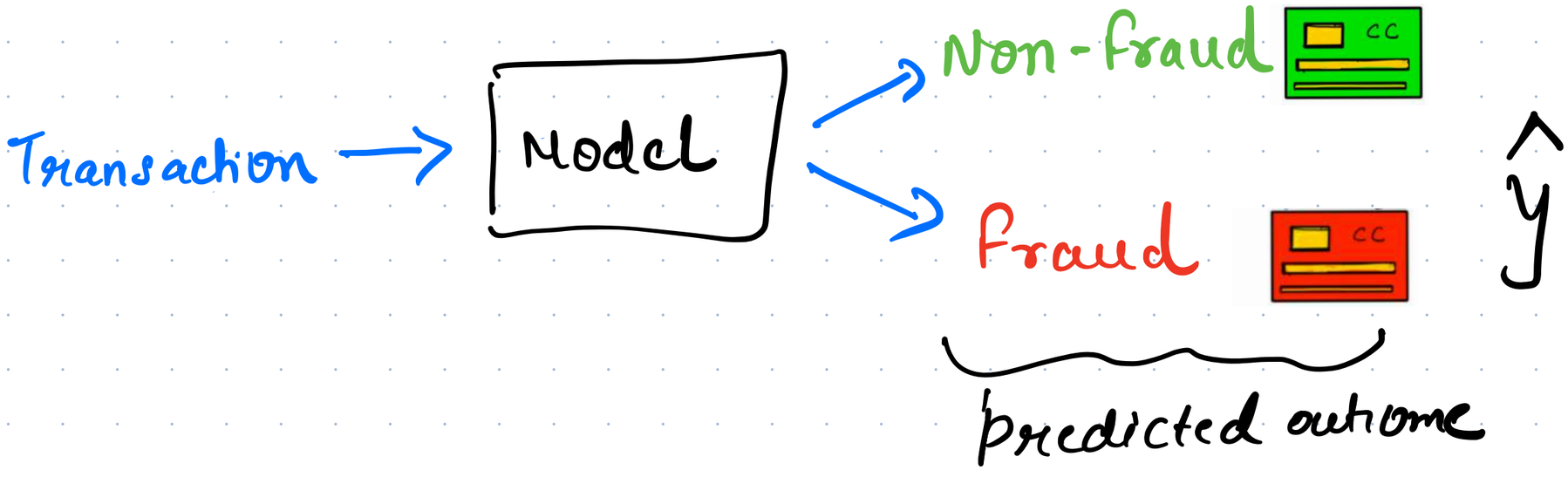


Majority data belongs to one class (Non-fraud)

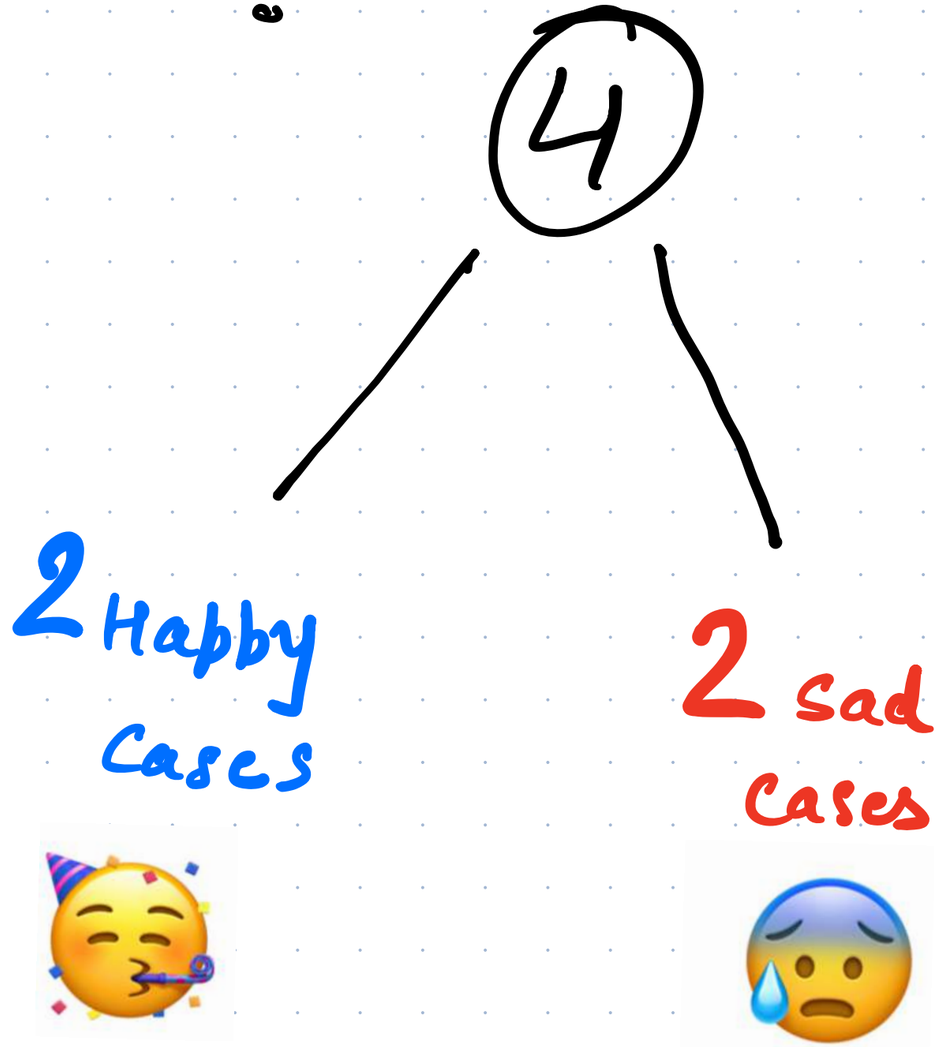
Any other alternative ?

Confusion matrix





How many possible combination of model predictions ?

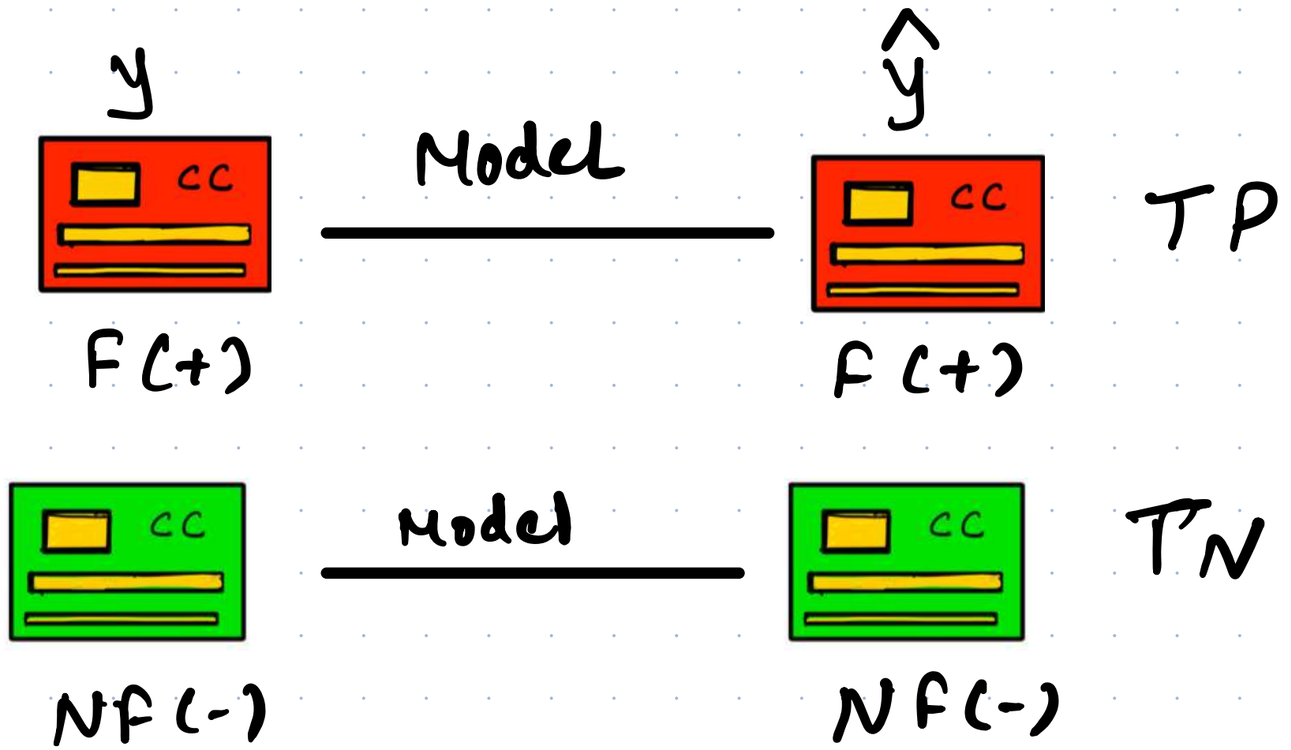


Happy Cases

⇒ Fraud is Caught, money saved

⇒ Profit to the Merchant

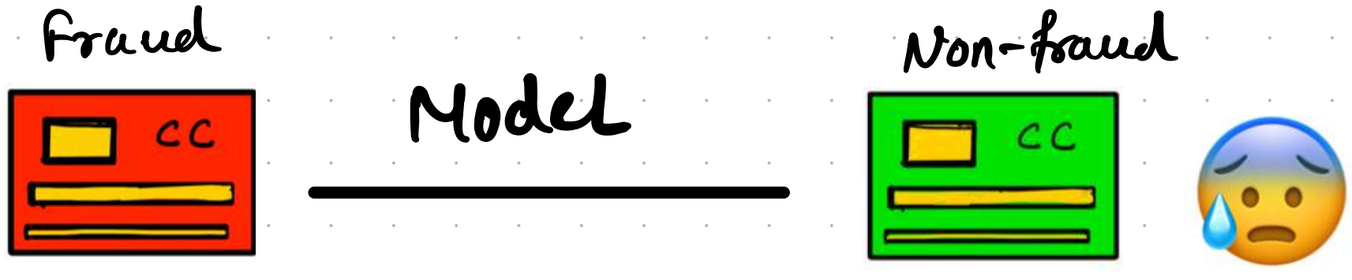
Considering fraud as tve class



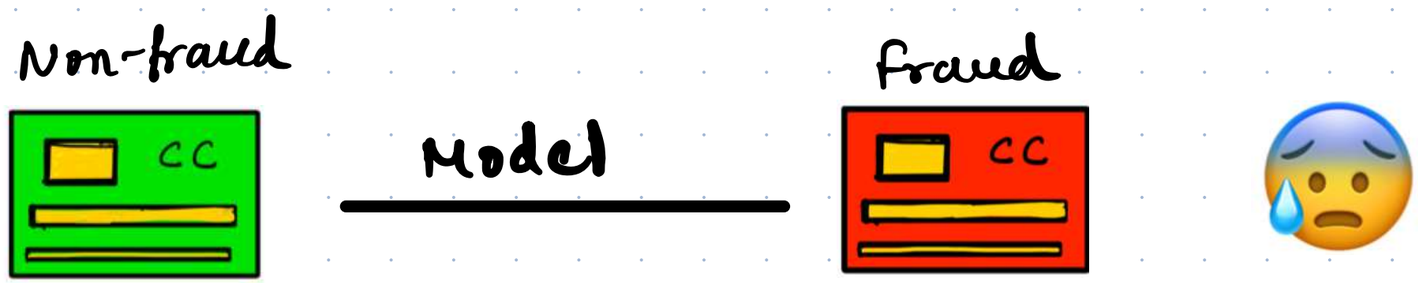
$y = \hat{y}$

So True Outcomes

Sad Cases



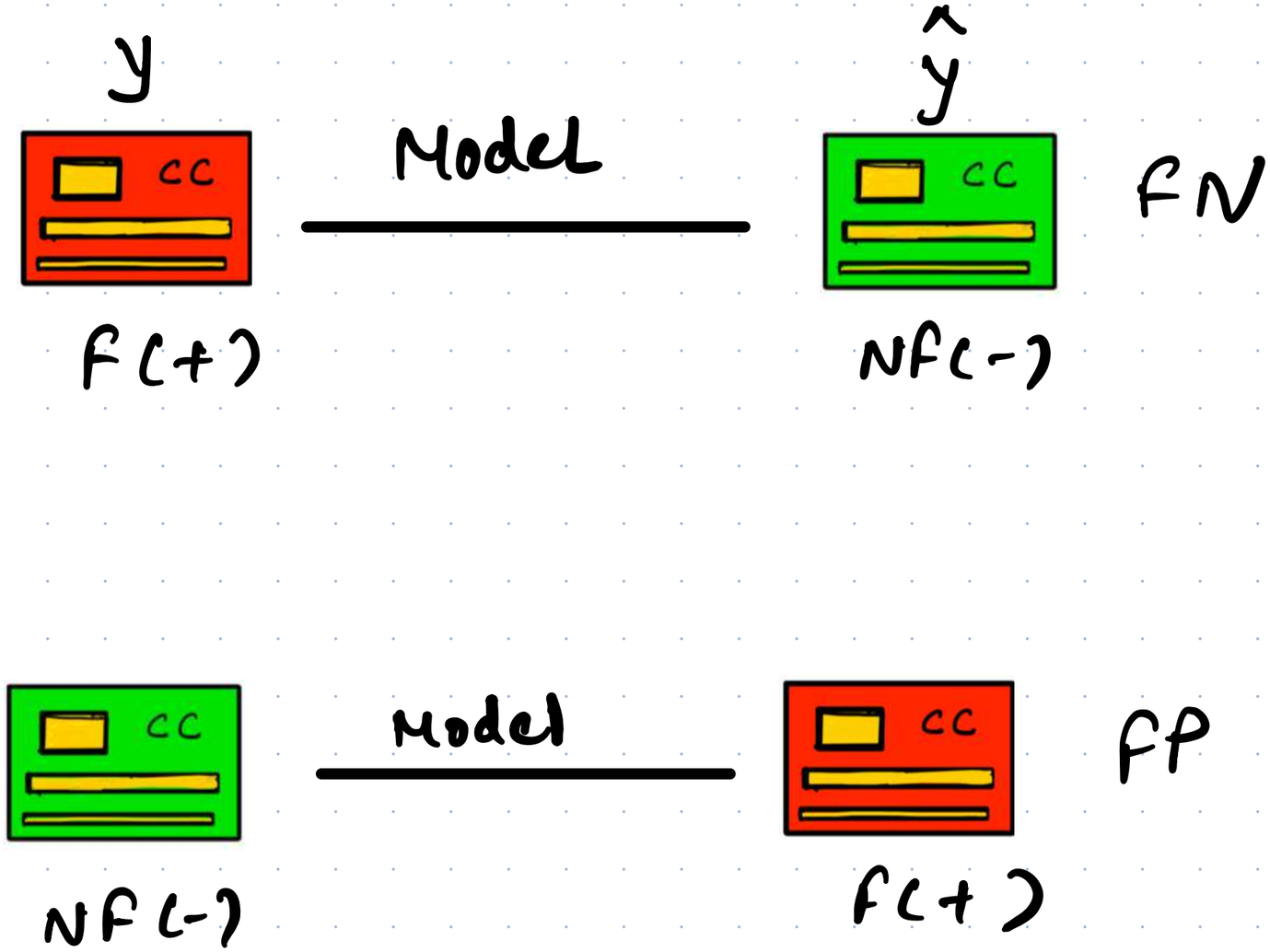
Model missed a fraud, loss to Merchant



Model blocked a fair transaction

Poor customer experience

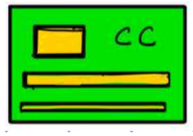
$y \neq \hat{y}$ so false Outcomes



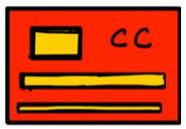
Confusion Matrix (2 x 2) grid

Model predictions ($\hat{y}$)


F

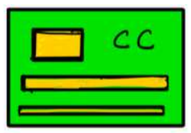

NF

Actual Values (y)


F


TP

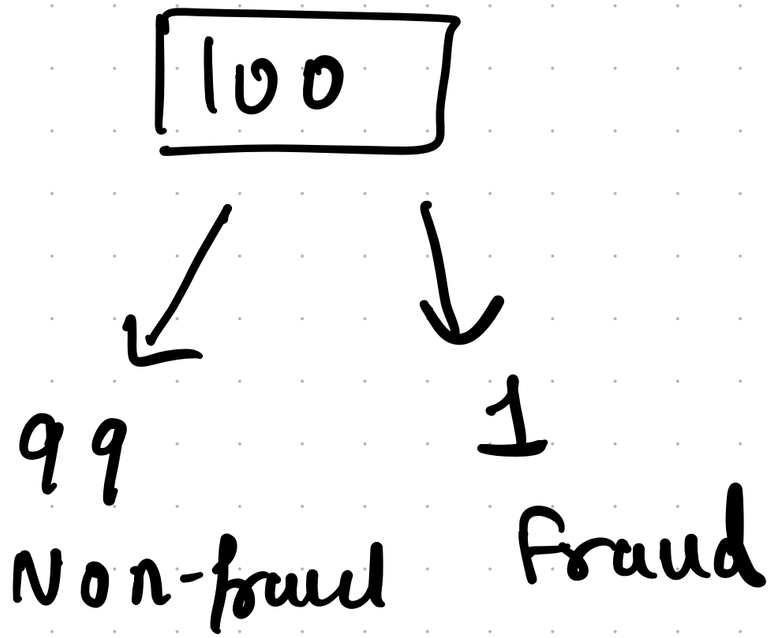

FN


NF


FP


TN

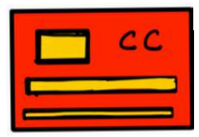
How will the CM for dumb model looks like ?



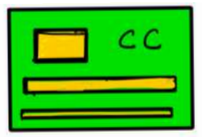
Model Predicts every transaction as Non-fraud

Dumb Model

Model predictions ($\hat{y}$)



F

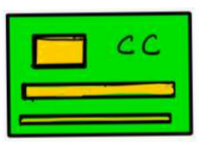


NF

Actual Values (y)

F

	0 TP	1 FN
	0 FP	99 TN



NF

Accuracy X

So what can be a good metric

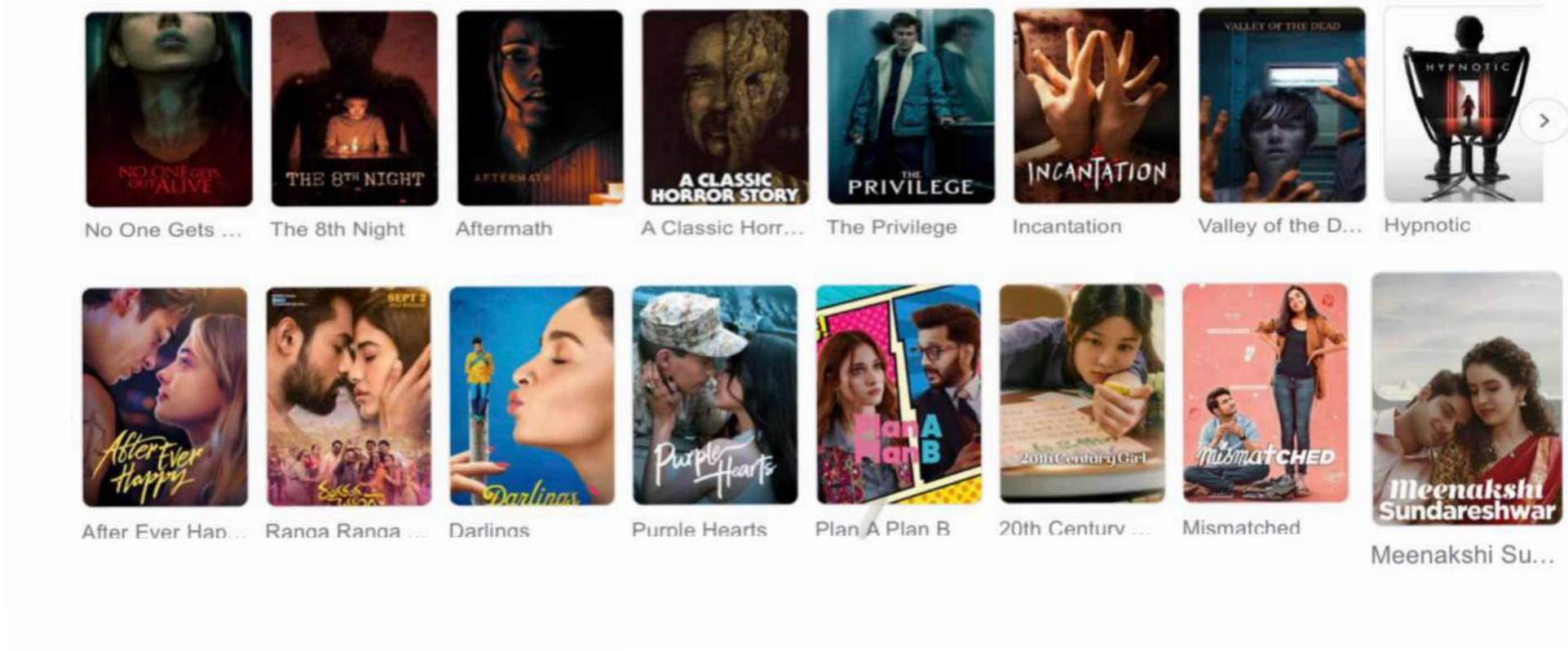
⇒ Precision & Recall

Netflix Recommendation

Horror movies only with her lover

Priya

She loves romantic movies



Total horror movies = 8

Total romantic movies = 8

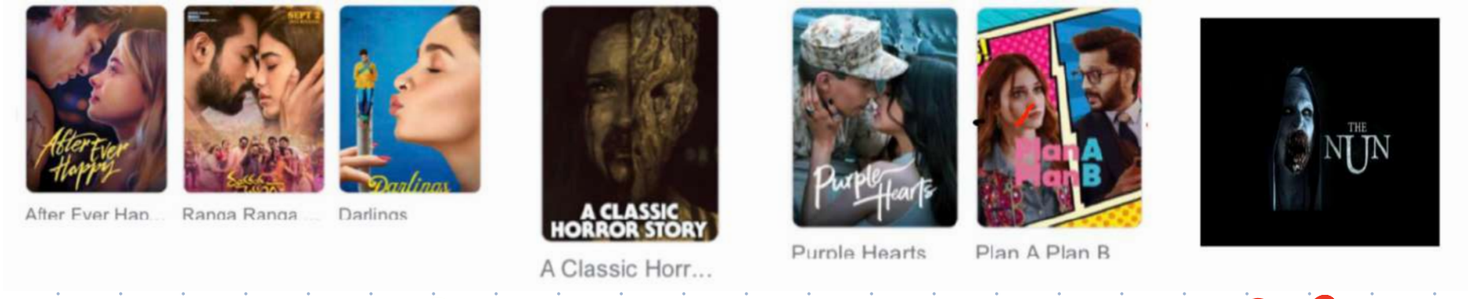
What should be the correction recommendation approach?

→ Precise about recommendations

→ No Horror movies

→ Suggest more & more Romantic movies (latest)

Romantic Recommendations



✔ ✔ ✔ ✘ ✔ ✔ ✘

How much Precise? (How many correct recommendations)

Precision = $\frac{5}{7} = 0.7$

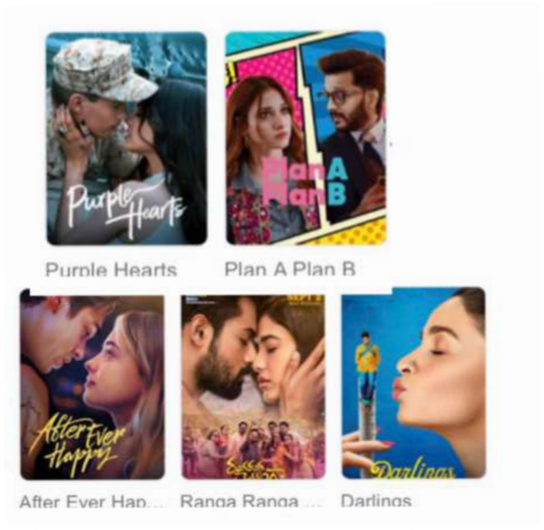
(For romantic)

All Romantic Movies

Recall ?
(for romantic)

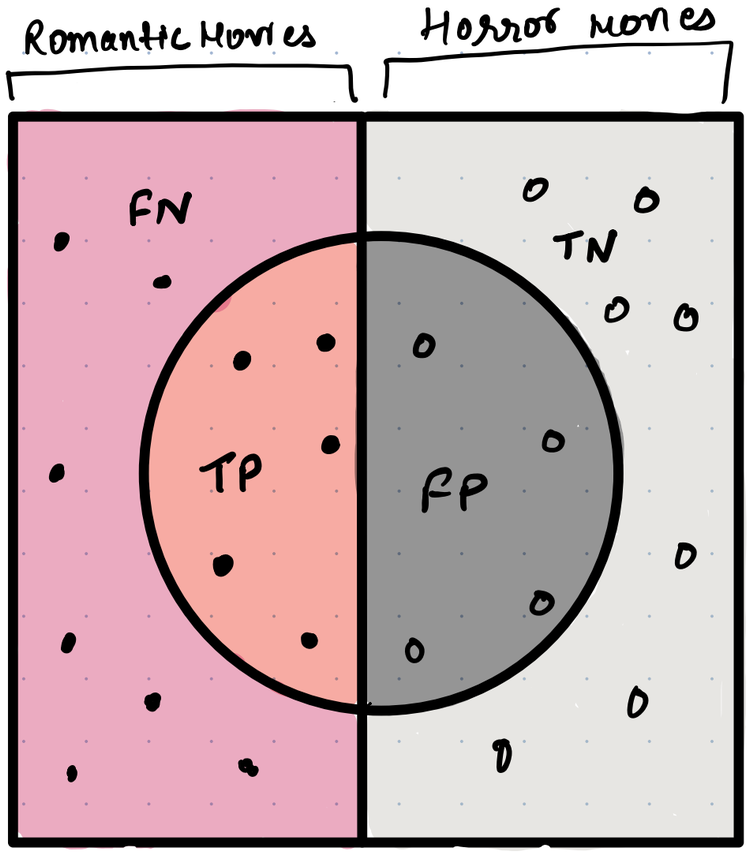
Out of all romantic how many model was able to recall ?

Recommended



Recall = $\frac{5}{8}$

Let's Summarise



○ Recommendations □ Whole Data

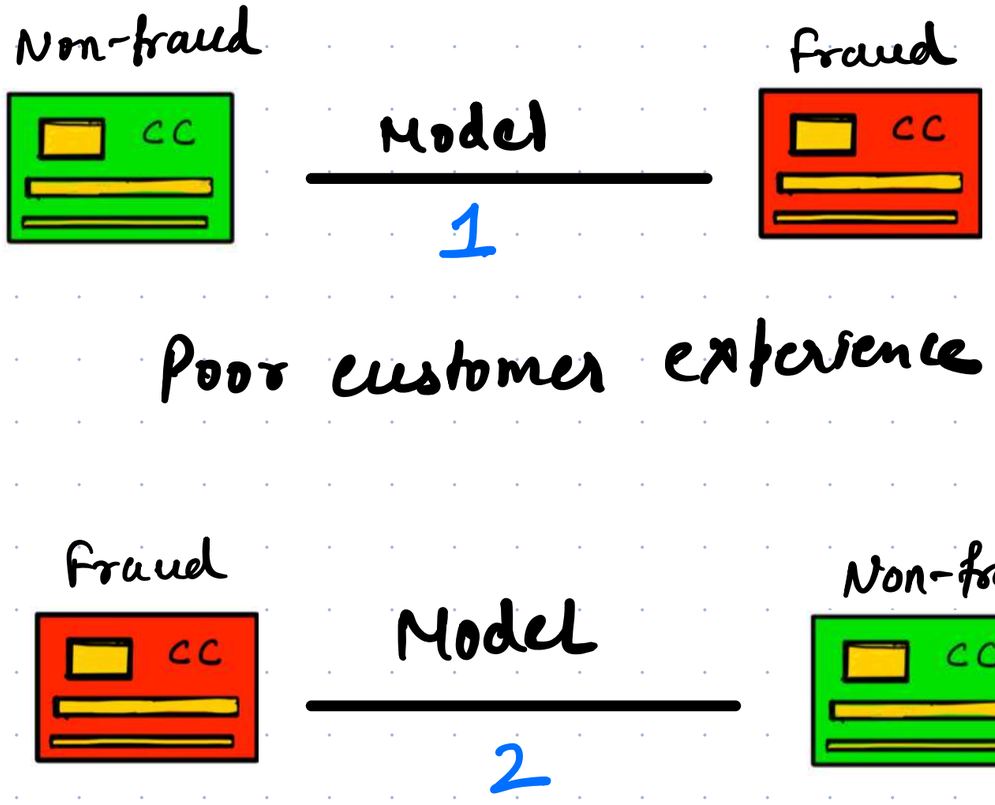
How many romantic in recommendations?

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

How many recommended out of all romantics

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Two Goals in CC Transactions



Remember we are calculating for fraud class

Which metric for Case-1 ?

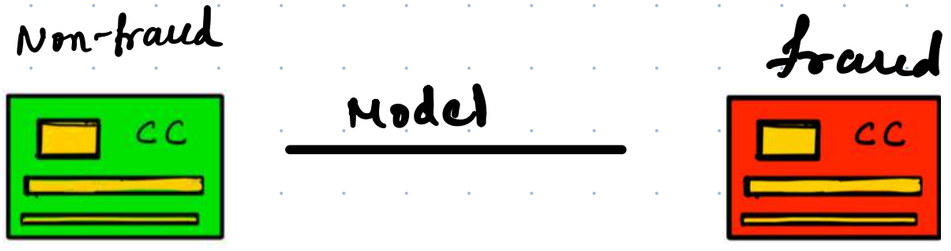
What is Model Prediction in Case-1 ? (fraud)

We need to be Precise

Precision Goal ?

=> How many correct fraud out of all !

=> Reduce misclassifying



Example

⇒ 10 Transaction

⇒ 4 are fraud

Case-1

Precision

⇒ model Predicted

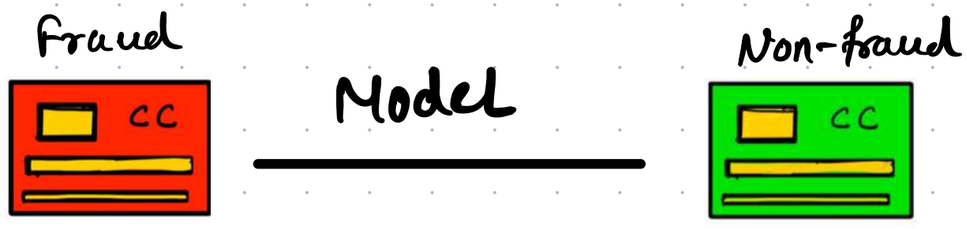
6 fraud, 4 Non-fraud

Precision ? $\frac{4}{6} = 0.66$

Which Metric for Case-2 ?

Model should **Recall** every fraud

So reduce misclassifying a



Example

⇒ 10 Transaction

⇒ 5 are fraud

⇒ model predicted 4 fraud, 6 Non-fraud

Case-2 ↓ Recall ↑

$$Recall = \frac{4}{5} = 0.8$

What mindset while evaluating ?

→ Model should not block a fair transaction
(High Precision)

→ Model should not allow a fraud transaction
(High Recall)

We need **High Precision, High Recall**

Precision & Recall

Ideal values ? 1.0

Precision = 1.0
Recall = 1.0

⟸ Possible ?

Look at this Example

Block if score ≥ 90 (High threshold)

Id	Score	Model outcome	True outcome
1	5	Not-fraud	Not fraud
2	10	Not fraud	Not fraud
3	15	Not fraud	Not fraud
4	30	Not fraud	Not fraud
5	40	Not fraud	fraud
6	90	Fraud	Fraud

Precision = $1/1$

= 1.0

Recall = $\frac{1}{2}$ = 0.5

Precision = 1.0 but low Recall

Let's try to increase Recall by reducing threshold

Why reducing ? low Threshold → More fraud detection → More Recall

Block if score ≥ 30 (low threshold)

Id	Score	Model outcome	True outcome
1	5	Not-fraud	Not fraud
2	10	Not fraud	Not fraud
3	15	Not fraud	Not fraud
4	30	fraud	Not fraud
5	40	fraud	fraud
6	90	fraud	fraud

Precision = $\frac{2}{3}$

$= 0.67$

Recall = $\frac{2}{2}$

$= 1.0$

Recall 1.0 but Precision fell

Trade-off

Recall ↑ → More fraud detection → More chances of them being wrong

↓

Precision ↓ decreased

So what other metric ?

⇒ Combine Precision & Recall

$$F_1 \text{ score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \cdot P \cdot R}{P + R}$$

Precision Recall

What if we want to give more Importance to One metric?

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{(\beta^2)(Precision + Recall)}$$

$\beta > 1$

$$F_2 = \frac{(1+2^2) \cdot P \cdot R}{P + R}$$

Recall more than Precision

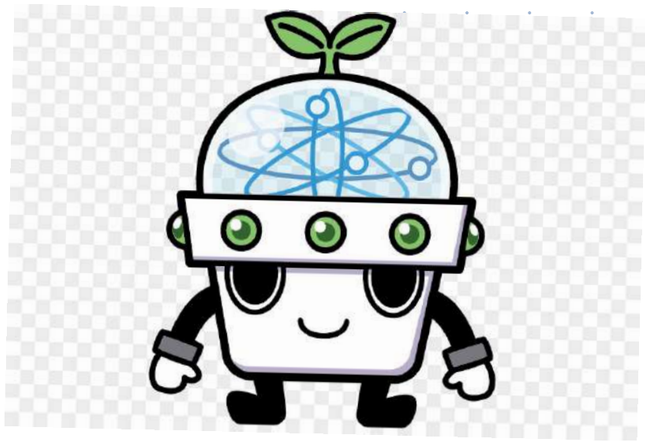
$\beta < 1$

$$F_{0.5} = \frac{(1+0.5^2) P \cdot R}{P + R}$$

Precision more than Recall

Which Model ?

Model-1



δ_1 / δ_2 / δ_3 \

M_1 M_2 M_3

Model-2

δ_1 / δ_2 | δ_3 \

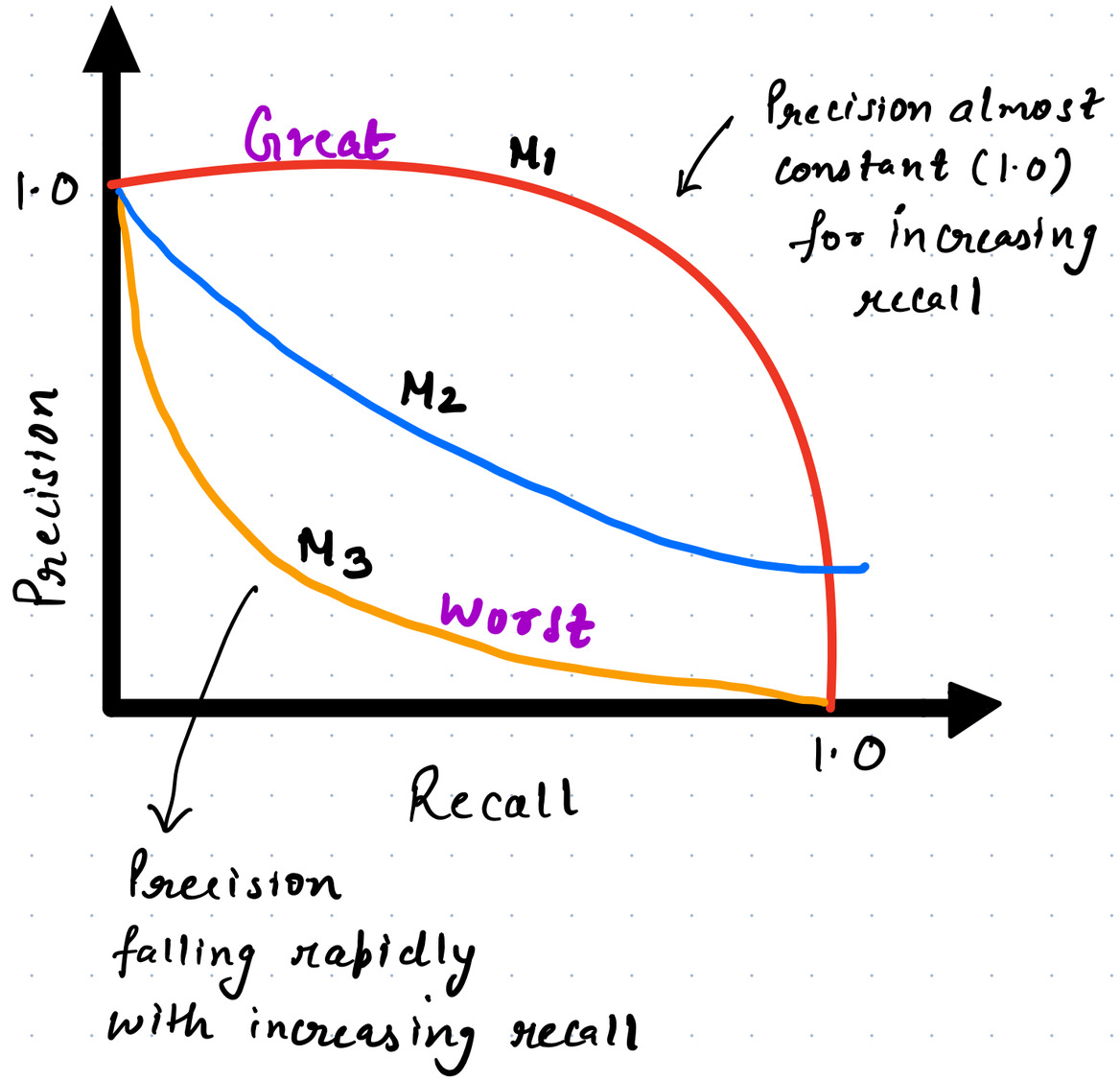
M_1' M_2' M_3'

$S \rightarrow$ Threshold

$M \rightarrow$ Metrics

(Precision, Recall, f score)

Precision - Recall Curve



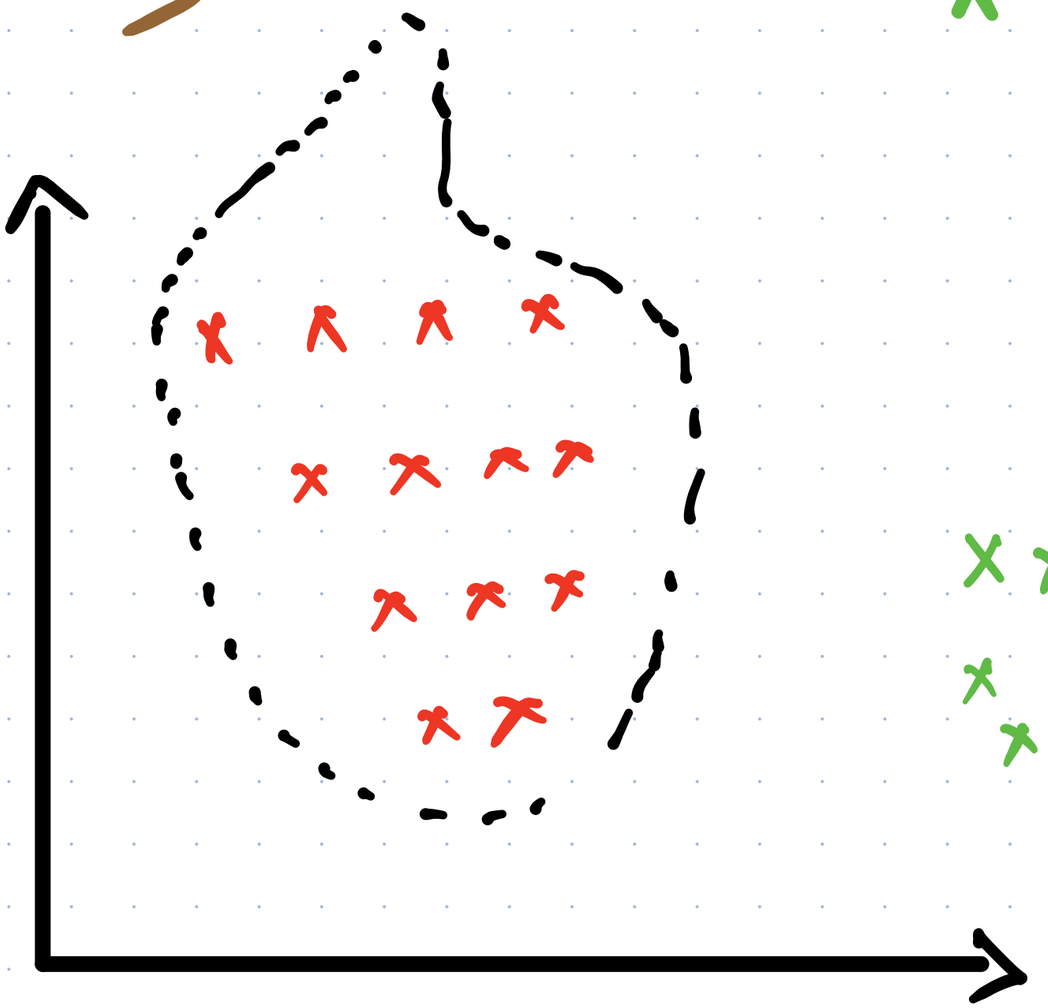
How this will affect

Our model ?

⇒ Is it going to degrade ?

Let's see

But I remember this as fraud

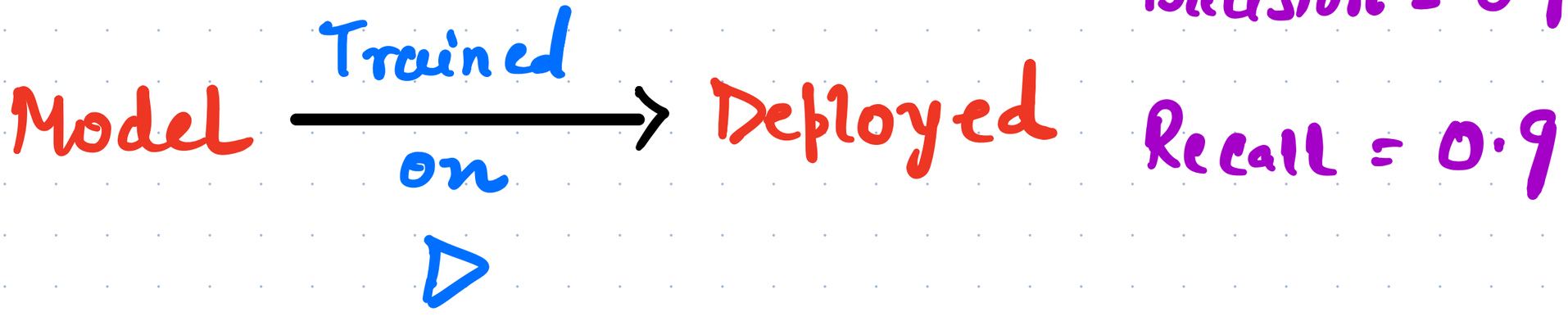


x → old data

x → new type of fraud

→ undetected

Let's see an example



Everything is Good

One month later

Transactions → Model → Detections

(some new types)

Precision = 0.9

Recall = 0.85

Blocks 90 % of fraud

Still good enough

6 months later

New Fraudster emerged ⇒ New Data ⇒ Model

Precision = 0.7

Recall = 0.75

Blocks 70% fraud

Poor performance

Model

(Blocks only 30.1 fraud)

Continously degrading

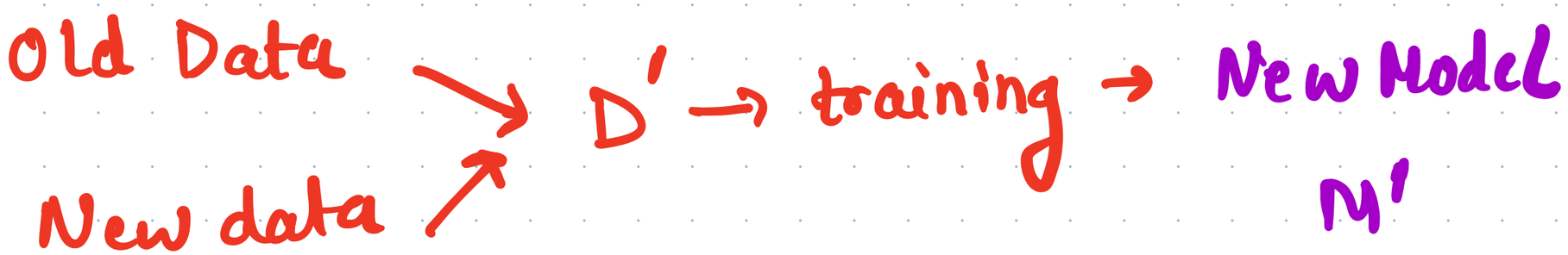
What should we DO ?

Retraining ?

Is it easy to Retrain?

⇒ what do we need for retraining?

1) Old data
2) new data



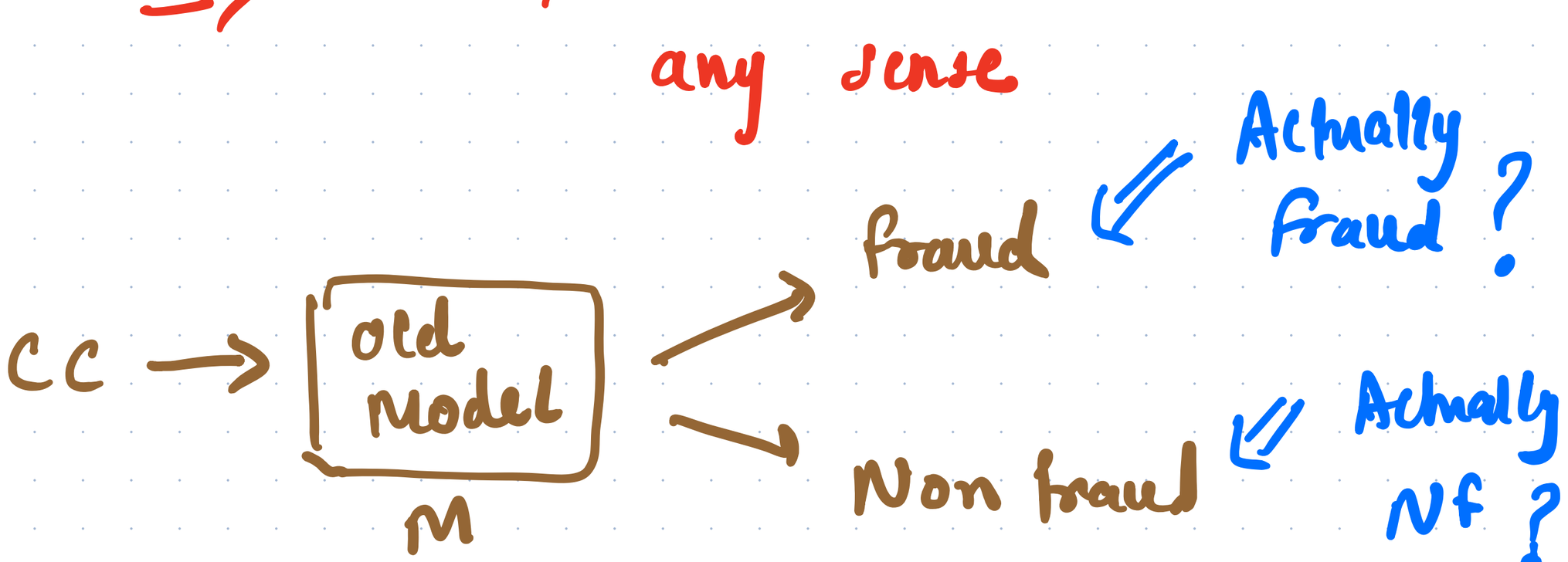
Easy Right?

But do we labels for new data?

we only have what is predicted by old model

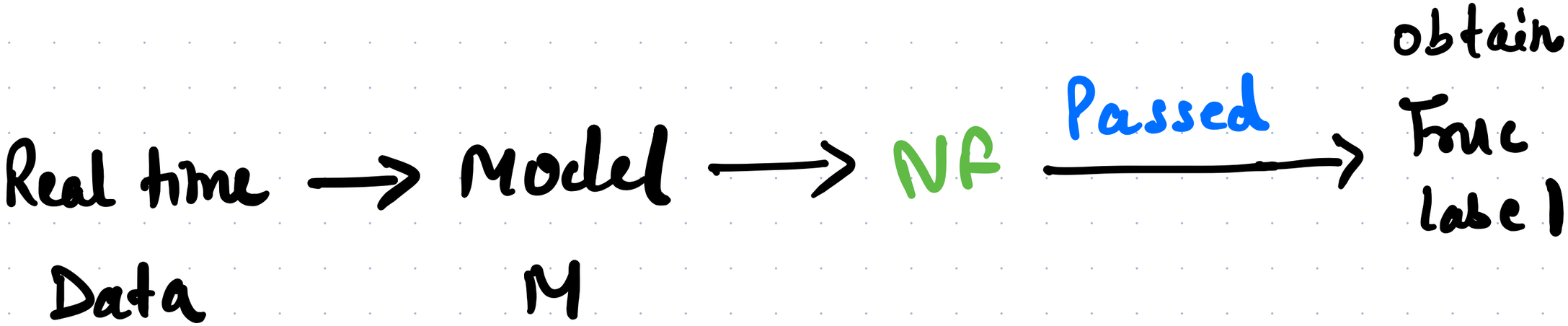
But should we trust old model predictions for retraining ?

⇒ No, this doesn't make any sense



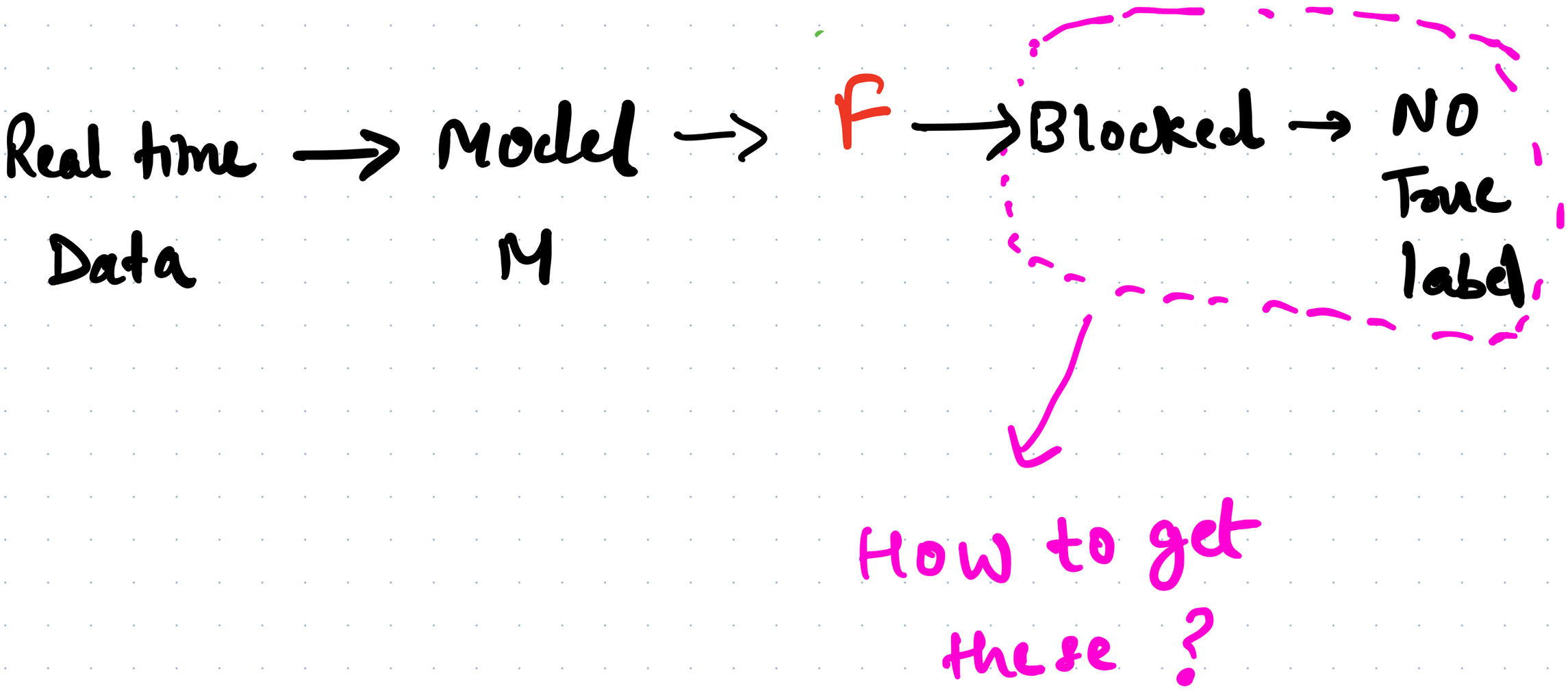
We need to know true labels

But How ?

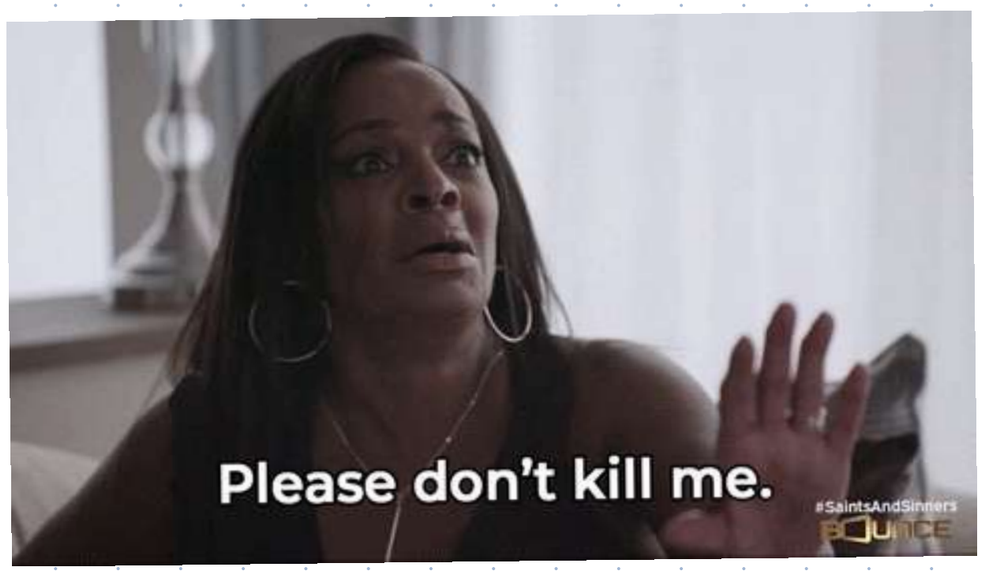


We need to know true labels

But How ?



What if we allow few fraud transactions to Pass !



⇒ To get their true labels obviously

But How many fraud should we Allow?

All (100%)

X

it will be a useless model

a few fraction say 10%

True labels obtained for these 10%

Let's see an example

Model's action ↑ | We chose ↗ | True label ↙

Id	Score	Default Action	Chosen Action	y
1	5	Allow	Allow	NF
2	30	Allow	Allow	F
3	80	Block	Allow	NF
4	95	Block	Allow	F

} ↗ few % we allowed to pass

For transaction - 1

Id	Score	Default Action	Chosen Action	y
1	5	Allow	Allow	NF

Both Model & we decided to Pass

True label ⇒ NF } By manual checking

For transaction - 2

Id	Score	Default Action	Chosen Action	y
2	30	Allow	Allow	F

Both Model & we decided to Pass

True label ⇒ Fraud

By manual checking

} Wrong detection
Include it in
new data

For transaction - 3

Id	Score	Default Action	Chosen Action	y
3	80	Block	Allow	NF

part of ./. we allowed to pass

Model achon = Block Our Achon = Pass

True label ⇒ NF

By manual checky

→ Include in New Data D'

as wrong Fraud detected

For transaction - 4

Id	Score	Default Action	Chosen Action	y
4	95	Block	Allow	F

part of 'I' we allowed to pass

Model action = Block

Our Achon = Pass

True label ⇒ F } By manual checking

* So 2 new data Points

Id	Score	Default Action	Chosen Action	y
1	5	Allow	Allow	NF
2	30	Allow	Allow	F
3	80	Block	Allow	NF
4	95	Block	Allow	F

✓
X
X
✓

} → Wrongly predicted by model in Production

Let's Evaluate model on new data

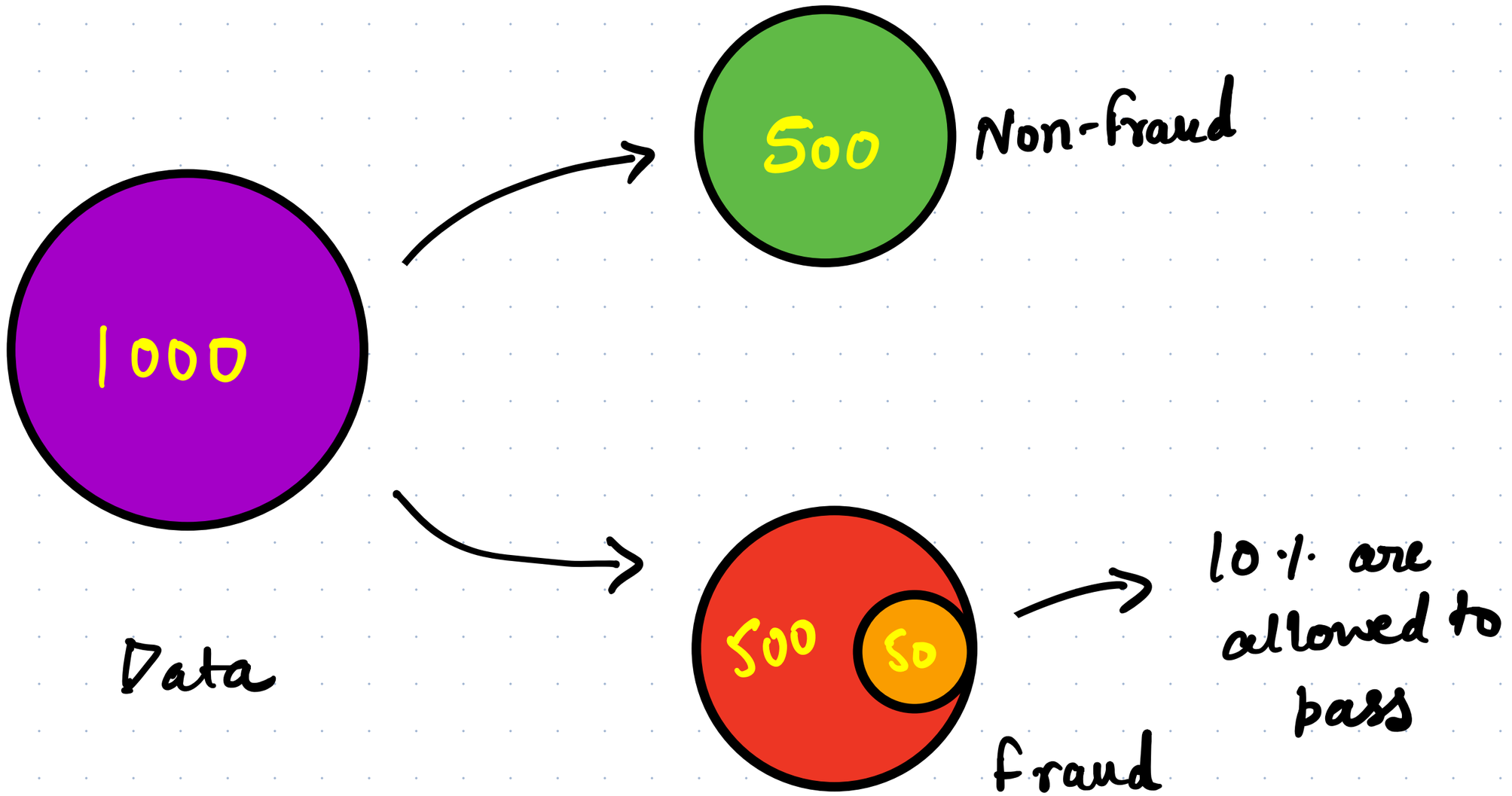
Id	Score	Default Action	Chosen Action	y	
1	5	Allow	Allow	NF	✓
2	30	Allow	Allow	F	✗
3	80	Block	Allow	NF	✗
4	95	Block	Allow	F	✓

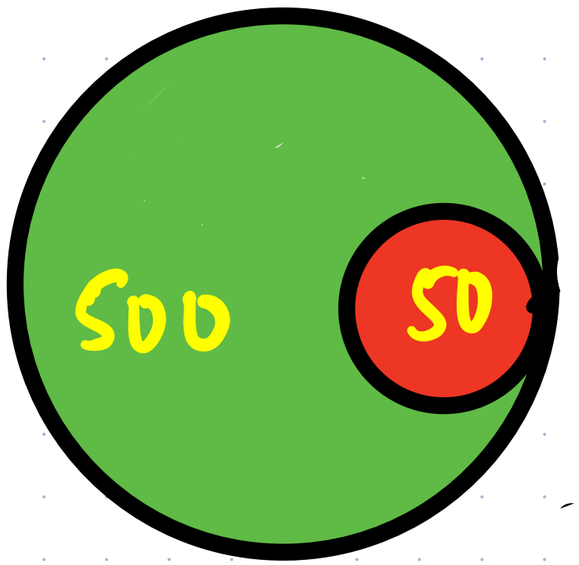
$$\text{Precision} = \frac{1}{2}$$

$$\text{Recall} = \frac{1}{2}$$

But is it a right way ??

Do you see the Problem ?

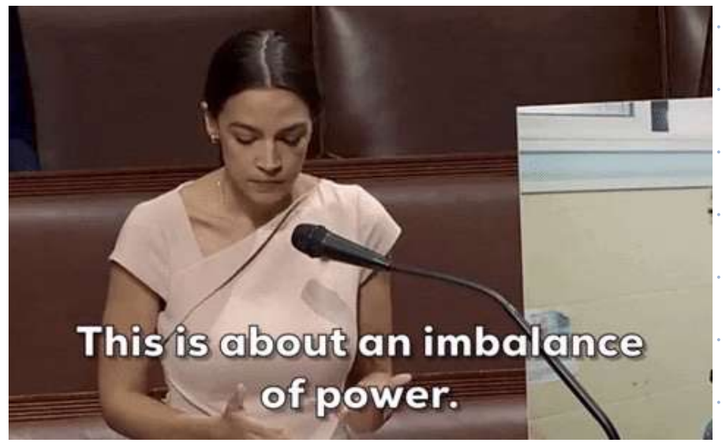




only 10% of 500 are representing New fraud fraud

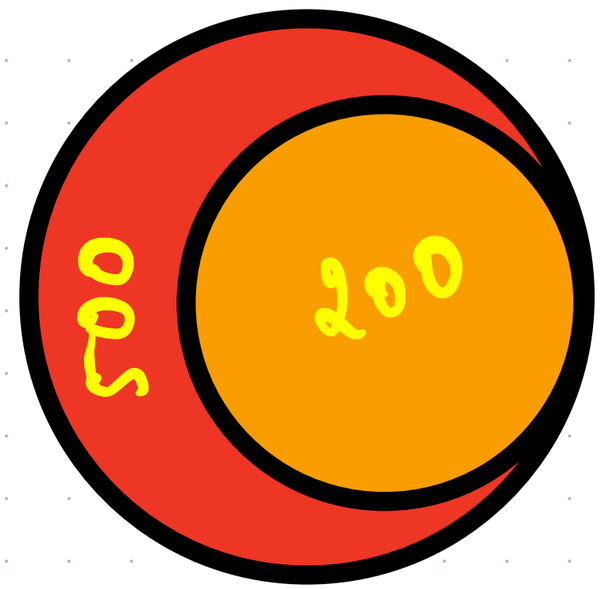
Loss of Data

How to handle this Imbalance while evaluating ?



What if we allow more fraud transaction to pass ?

✗ model will be useless



Then what to do ?

How about we bump up Weight of each sample ?

Consider

1 sample ≃ 10 samples


What's next?

Evaluate again

Precision
Recall

Precision
Recall

Re-train