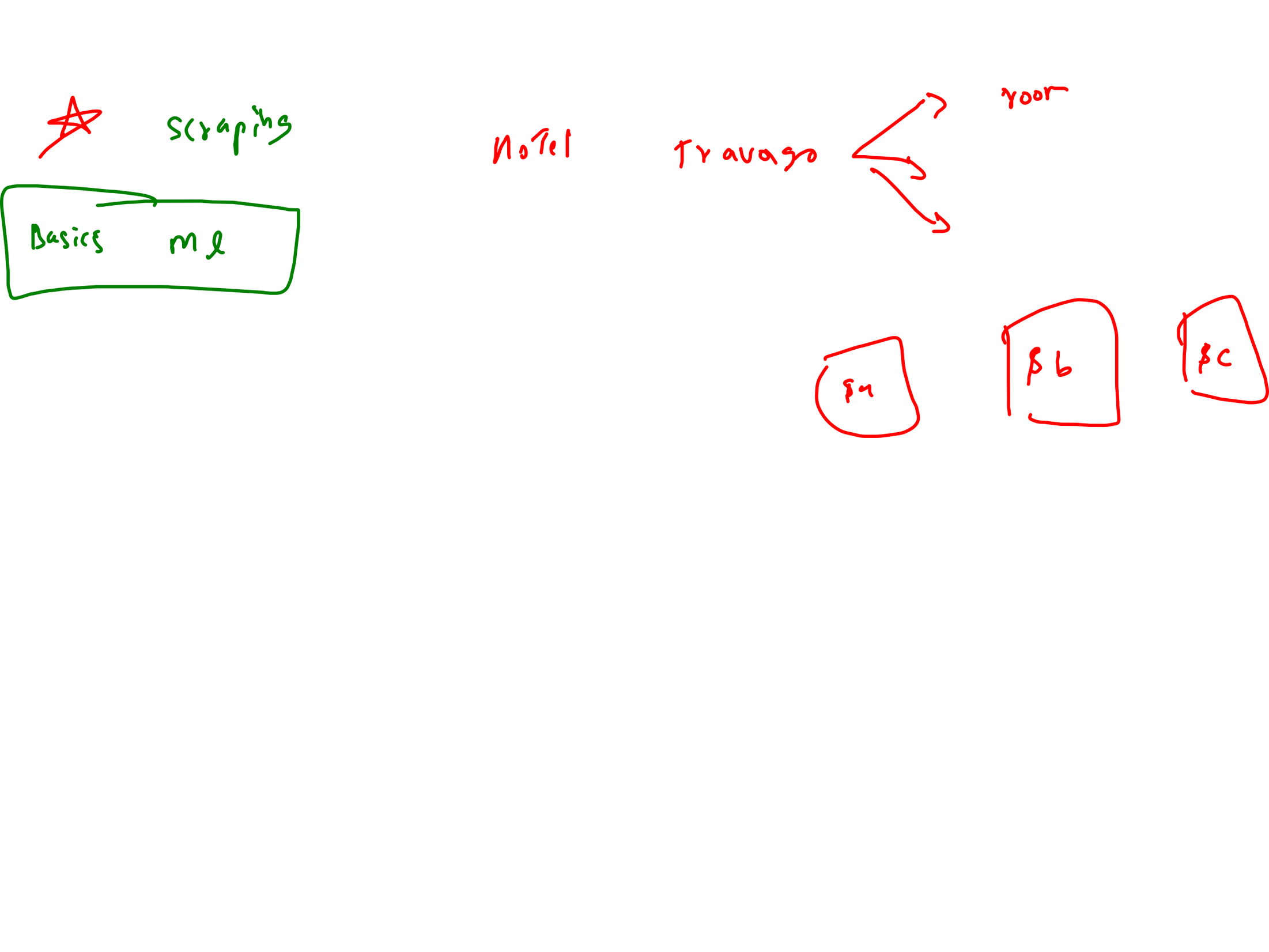

Scraping
Basics
ML
Hotel
Travago
Yoor
$m
$b
$c

}

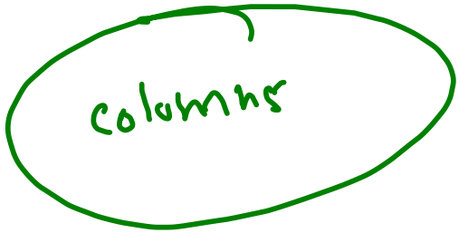
<a>

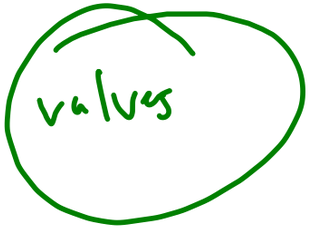
bs.findAll('a')

bs.find(div).findAll('a')

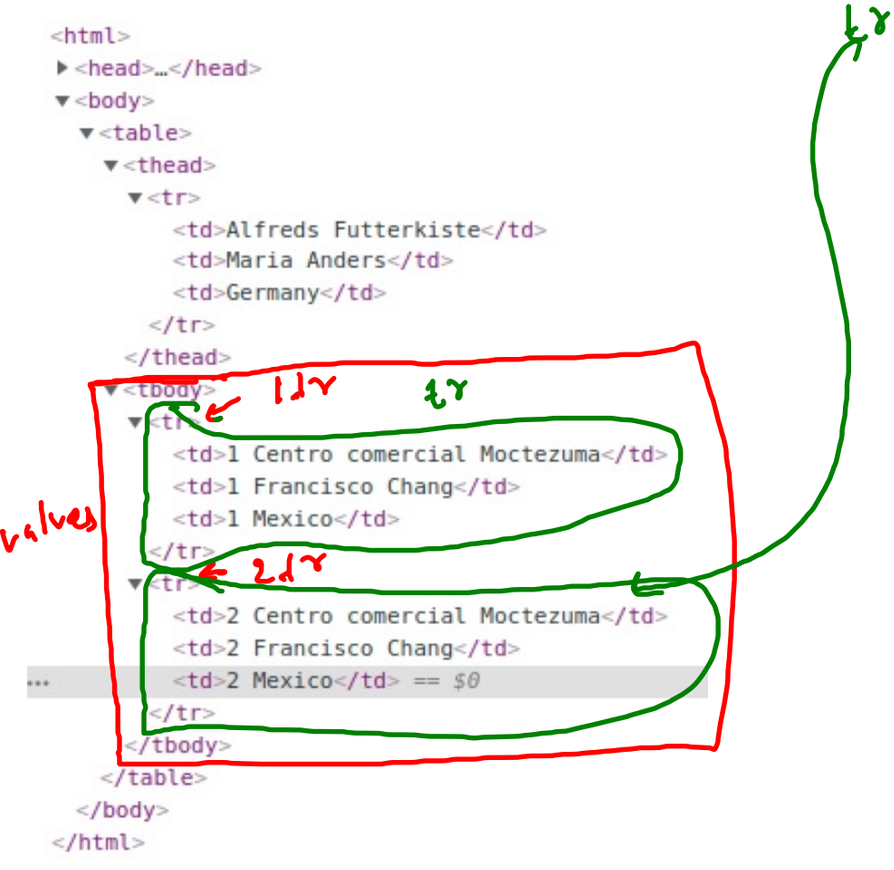
attrs = {'attr': value}
name

Data frame






```
<html>
▶ <head>…</head>
▼ <body>
  ▼ <table>
    ▼ <thead>
      ▼ <tr>
          <td>Alfreds Futterkiste</td>
          <td>Maria Anders</td>
          <td>Germany</td>
        </tr>
      </thead>
    ▼ <tbody>
      ▼ <tr>
          <td>1 Centro comercial Moctezuma</td>
          <td>1 Francisco Chang</td>
          <td>1 Mexico</td>
        </tr>
      ▼ <tr>
          <td>2 Centro comercial Moctezuma</td>
          <td>2 Francisco Chang</td>
          <td>2 Mexico</td> == $0
        </tr>
      </tbody>
    </table>
  </body>
</html>
```

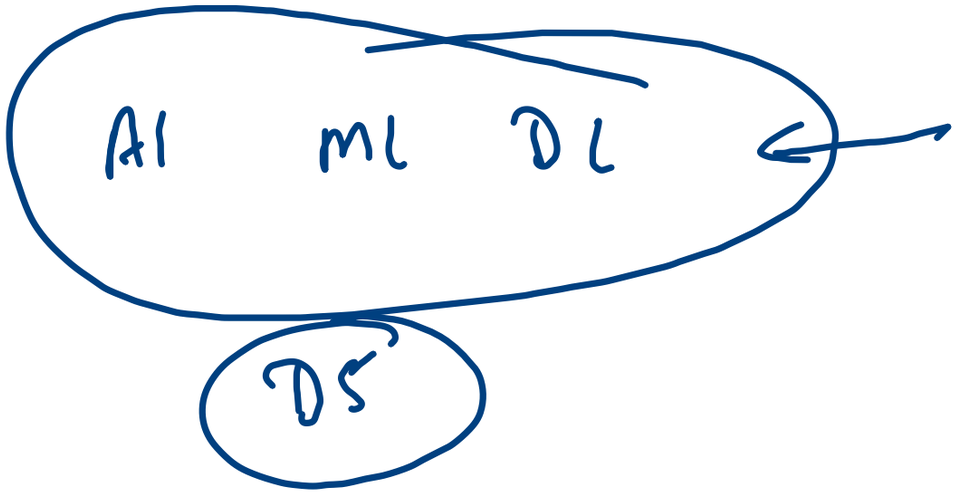



```
values = bs.tbody
values
```

```
pd_values = []

for tr in values.findAll('tr'):
    row = []
    for td in tr.findAll('td'):
      row.append(td.text)

    pd_values.append(row)

pd_values
```

```
[['1 Centro comercial Moctezuma', '1 Francisco Chang', '1 Mexico'],
 ['2 Centro comercial Moctezuma', '2 Francisco Chang', '2 Mexico']]
```

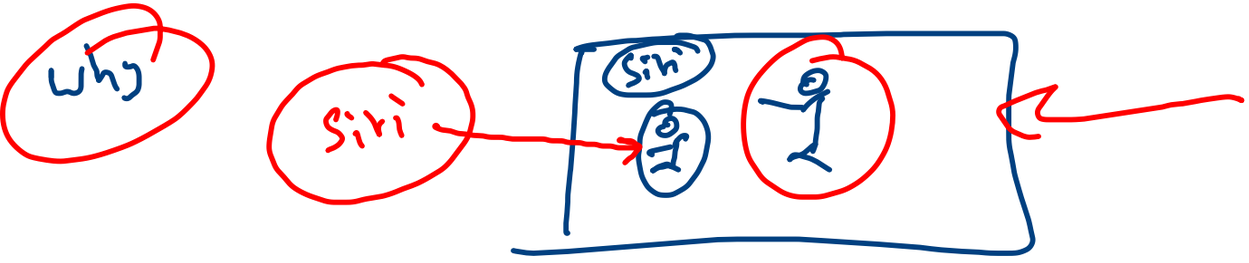
pd_values = [[1centro, 1franc., 1mexico], [2cent, 2franc., 2mexico]]



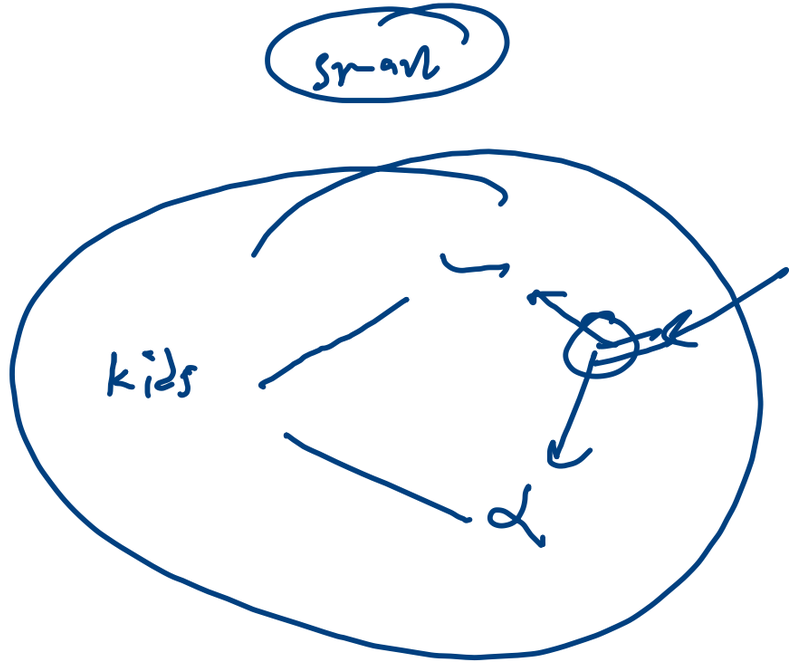
AI → Machine Intelligence
Artificial Intelligence

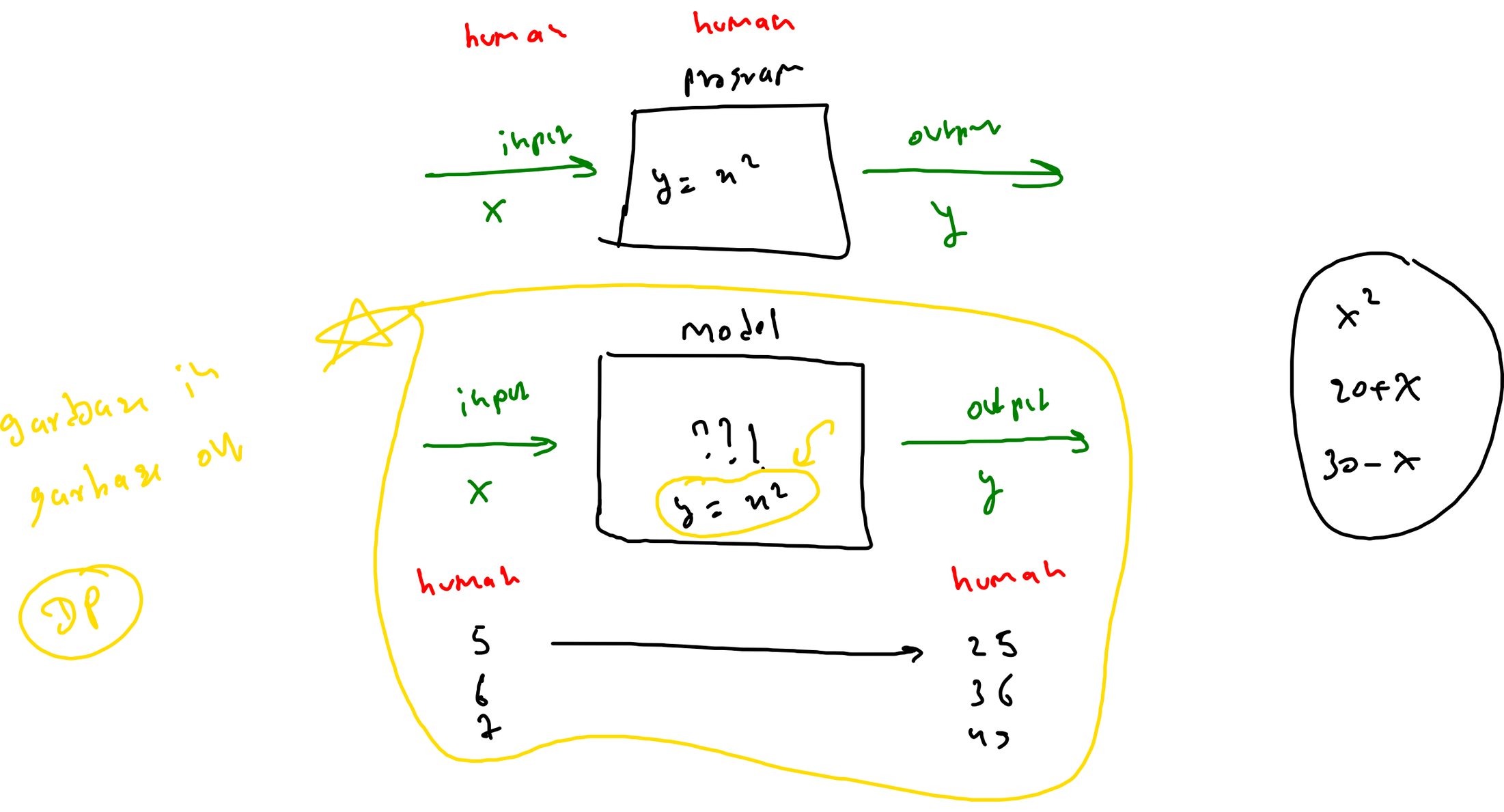
ML → ? stats

DL → neural networks


why
siri
siri


adoration


smart
kids


human
human
program
input
x
$y = x^2$
output
y
model
input
x
???
$y = x^2$
output
y
human
human
5
6
7
25
36
49
x^2
$20+x$
$30-x$
garbage in
garbage out
DP

ML Types

- Supervision
  - Supervised
  - Unsupervised
  - semi-supervised
  - Reinforcement Learning
- recurrent Learning
  - batch
  - online
- based on comparison
  - instance
  - model

Supervised

label

X → [ ] → Y

y

infinite set

finite set

Regression

[4, 5, 7, 8, 5.5]

5.1, 5.2, 5.3

[10k, 20k, 30k, 40k)

y y

finite set

classification

{ cat, dog, horse }

{ male, female }

{ 1, 2, 3, 4, 5, 6 }

X Iris

SL, SW, PL, PW

y

S
V
V

Un-Supervised

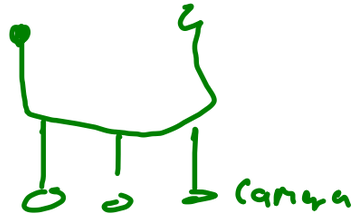
dog, cat, horse

input

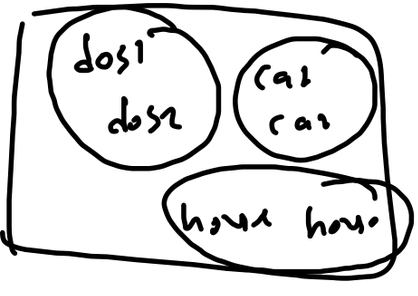
X

mnist

784



clustering
Grouping



Dimensional
deduction

784
⋮
2, 3

Anomaly
detection

outlier
inputs

Bank
Fraud

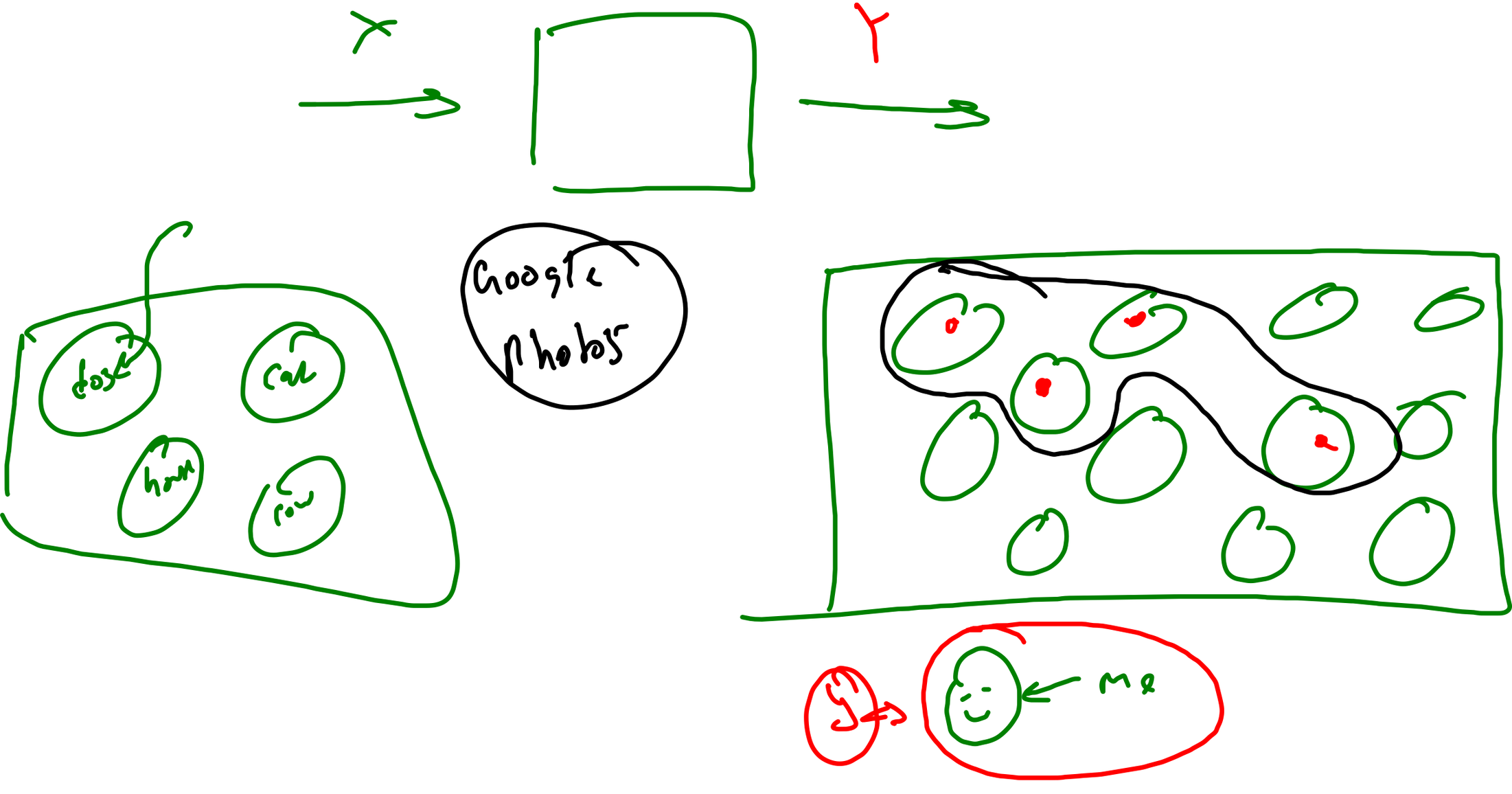
Association
rule learning

Walmart

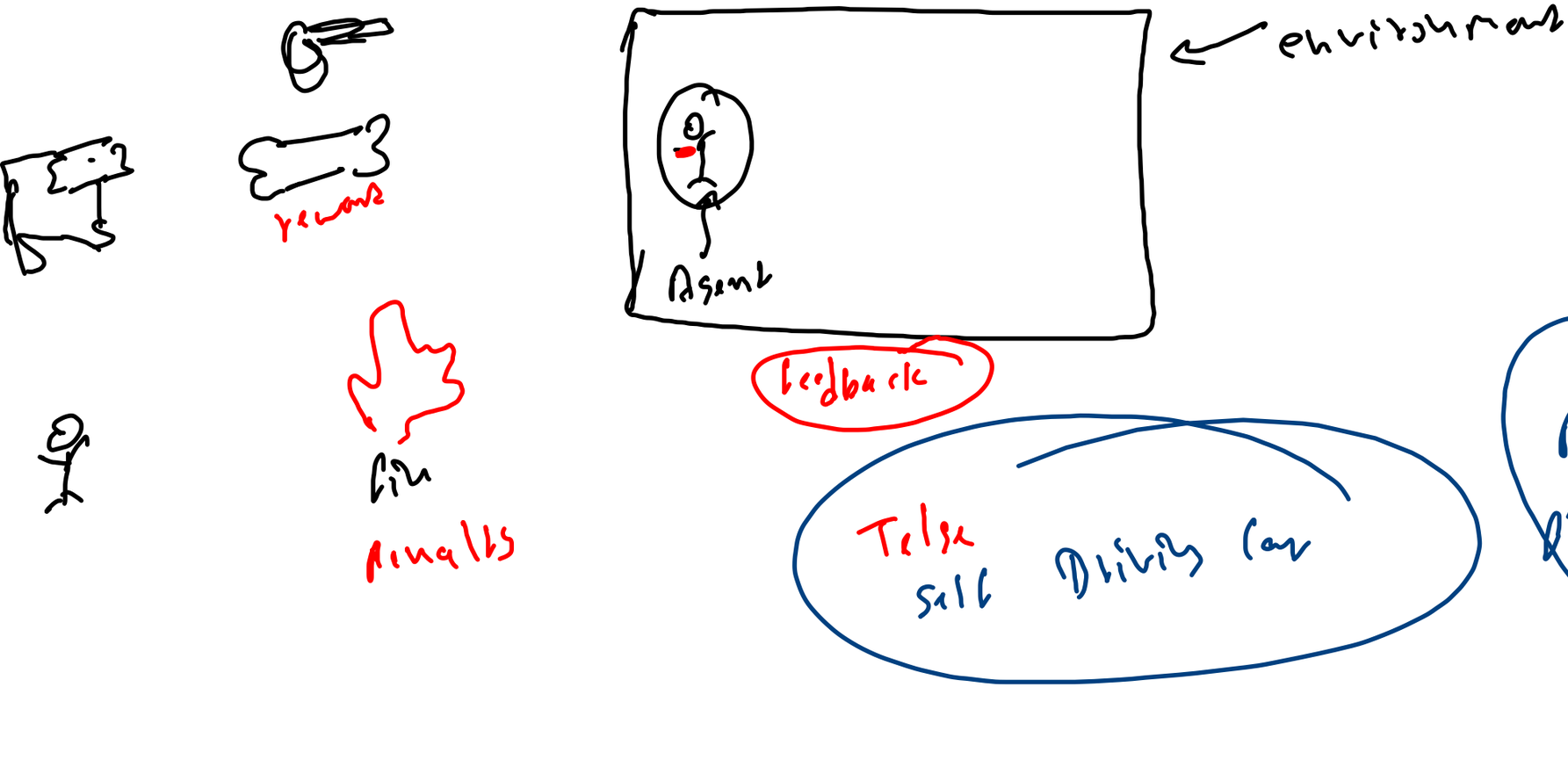
Beer - Diaper

24, tool, etc.
electronics

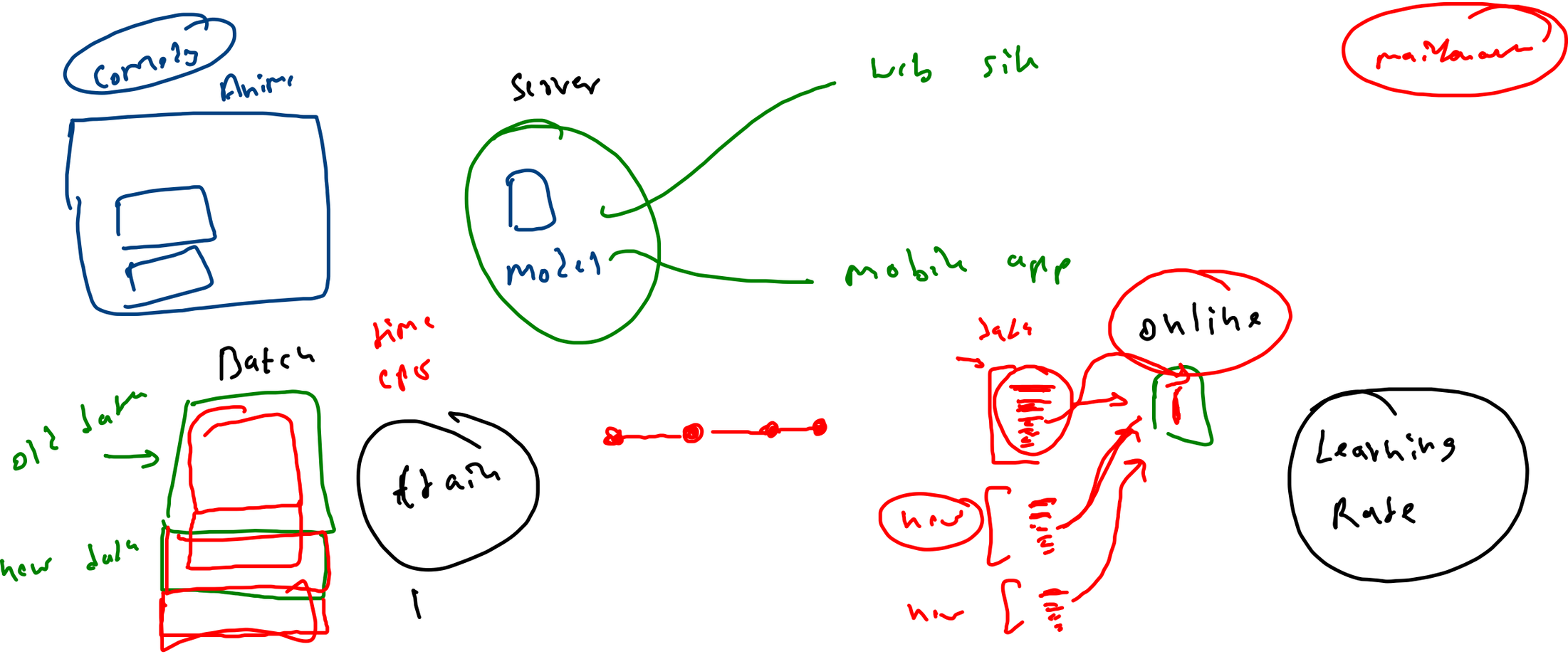
Semi- supervised



Reinforcement Learning



Recurrent Learning



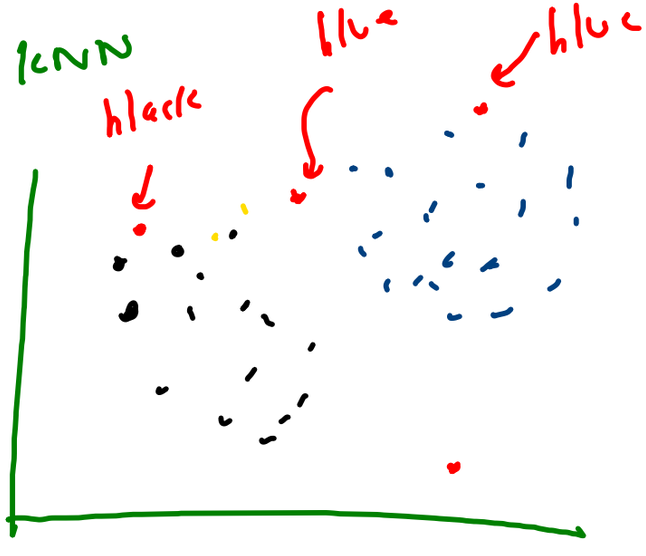
Based on Comparison

Instance

Model

KNN

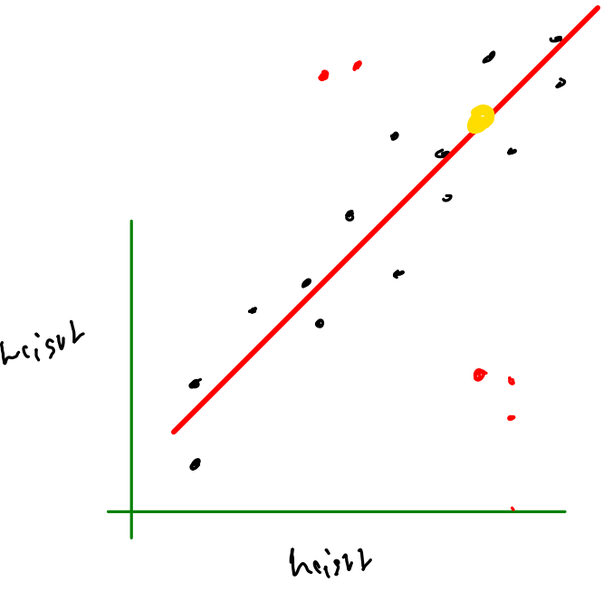
Rabbit



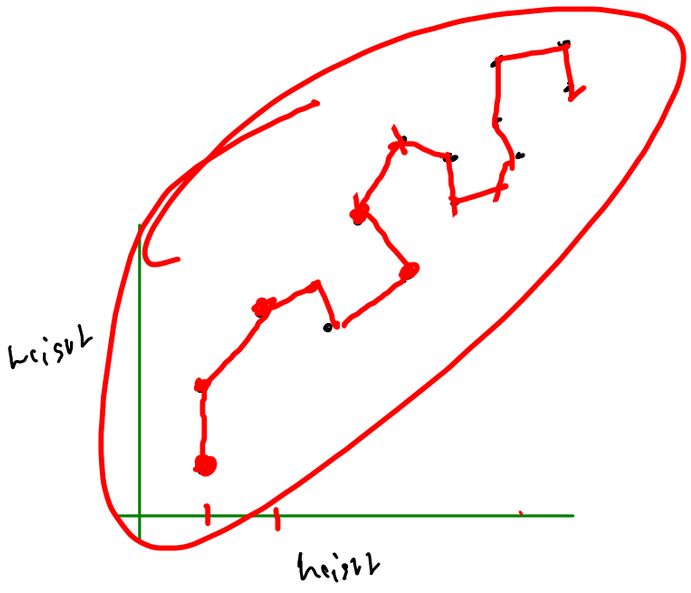
Logic, Generalize

challenges

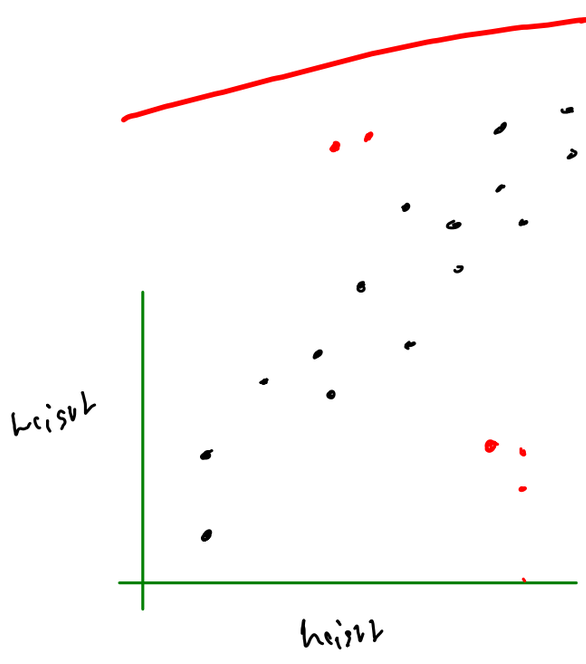
- insufficient data → [specific]
- how-representable data
- poor quality
- Irrelevant features
- Underfitting
- Overfitting



Correct



Overfitting

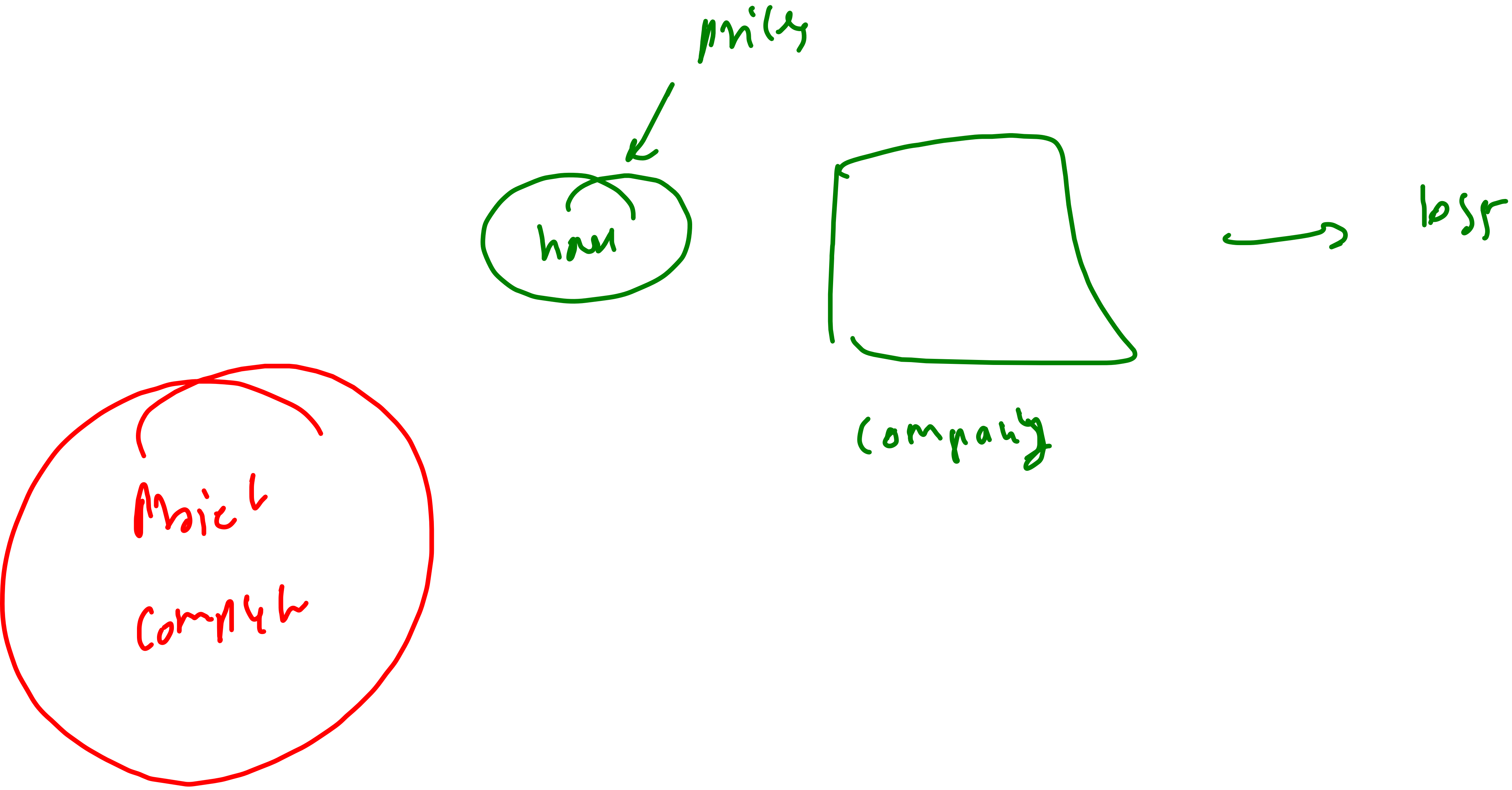


Underfitting

6.5

ML DLC

- frame problem
- Data gather
- preprocessing → human readable
- EDA
- feature engineering → model
- evaluation
- deployment
- maintenance



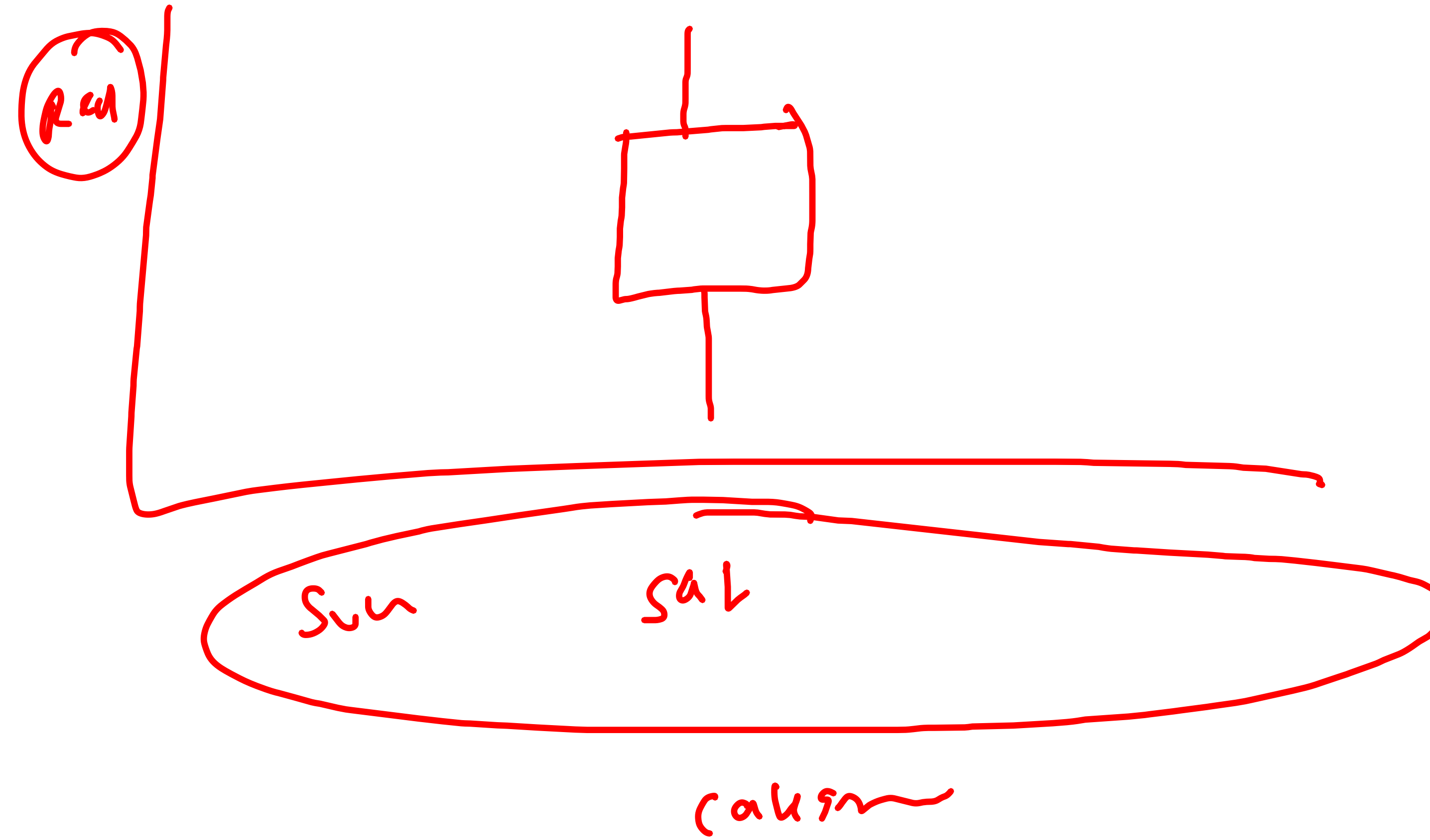
policy
home
company
loss
Public
Company

S

Sun

hips

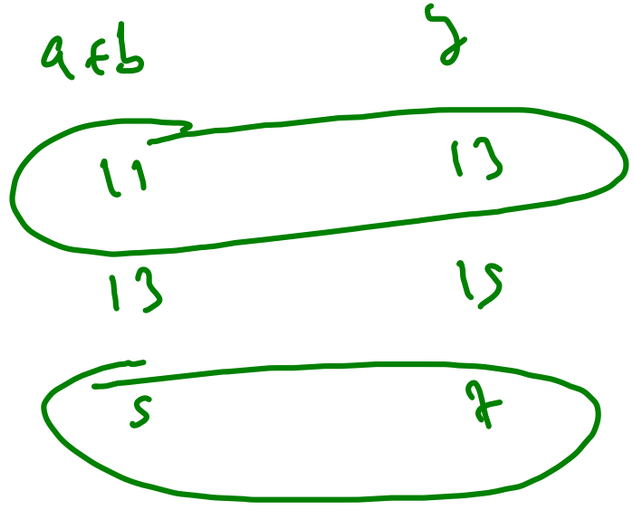
fir

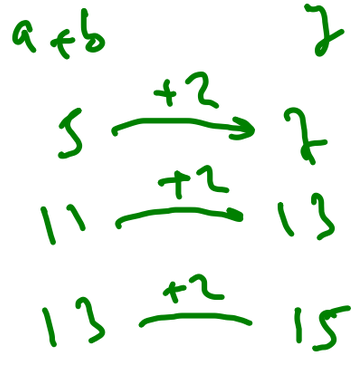


humby

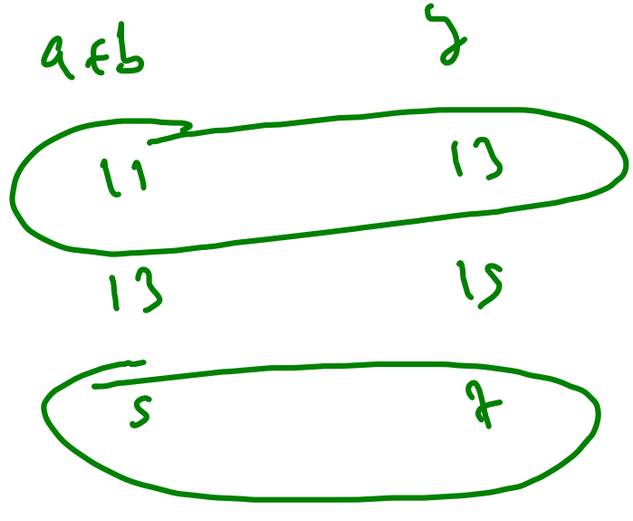
bar

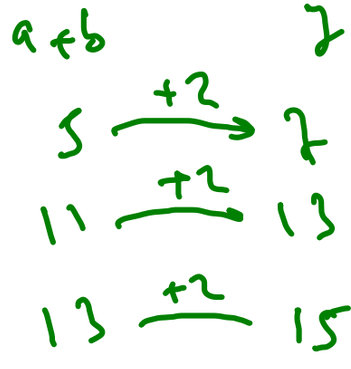
a	b	a+b	y
1	10	11	13
6	7	13	15
9	-4	5	7

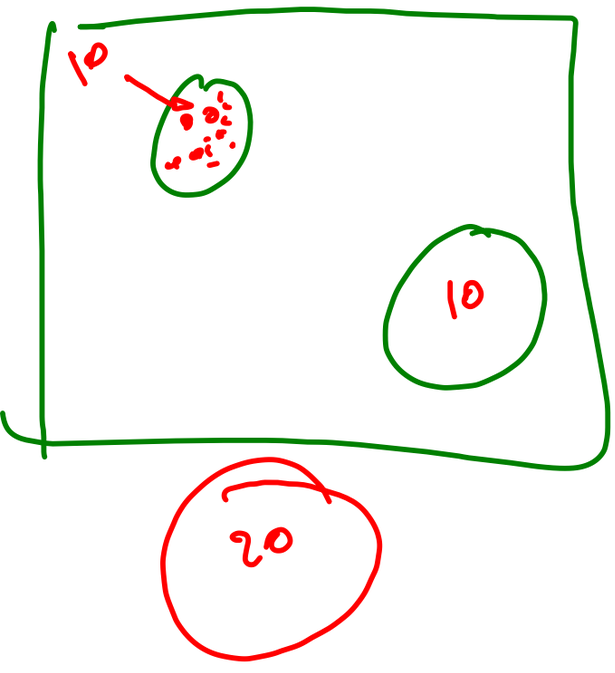


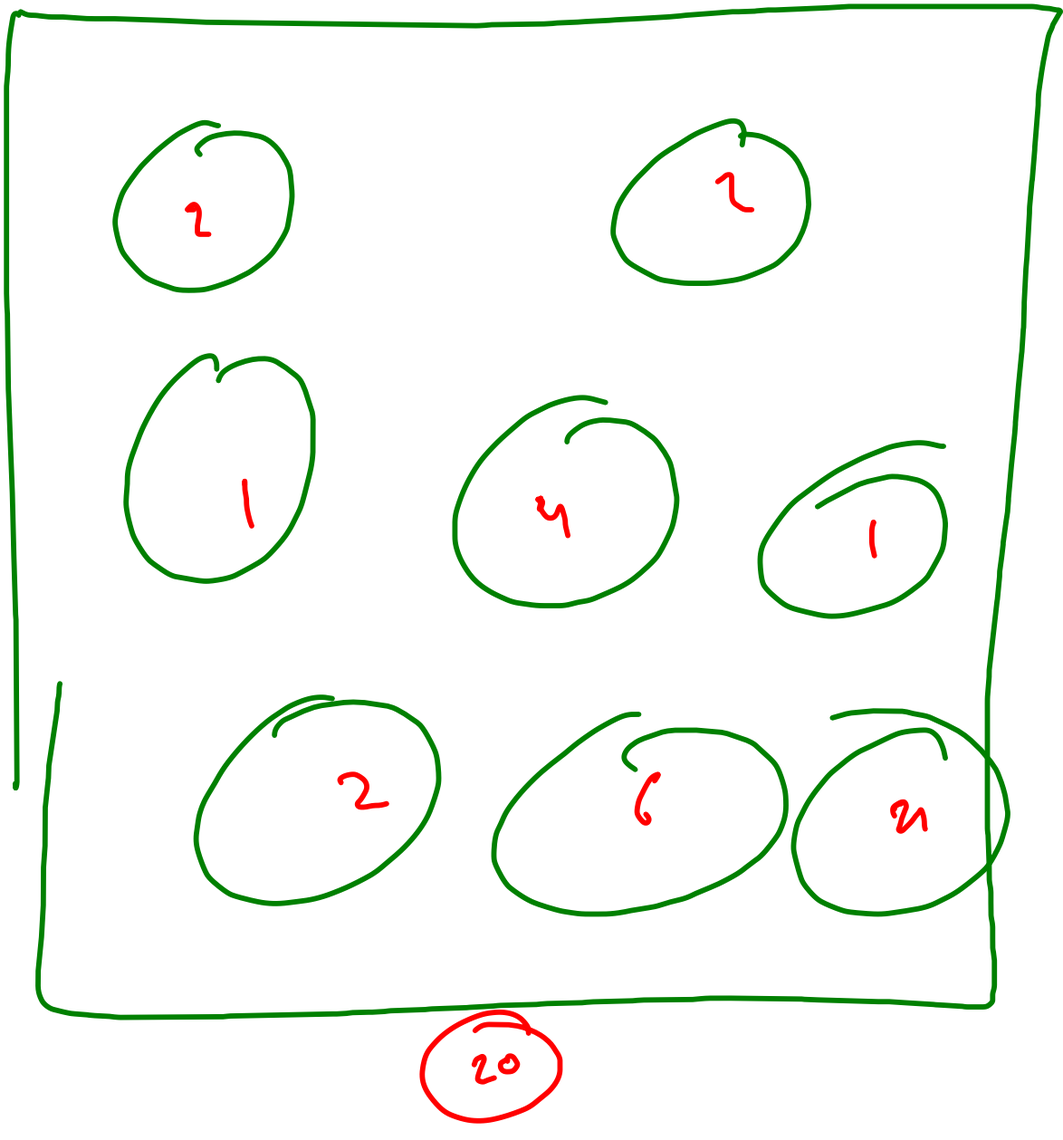


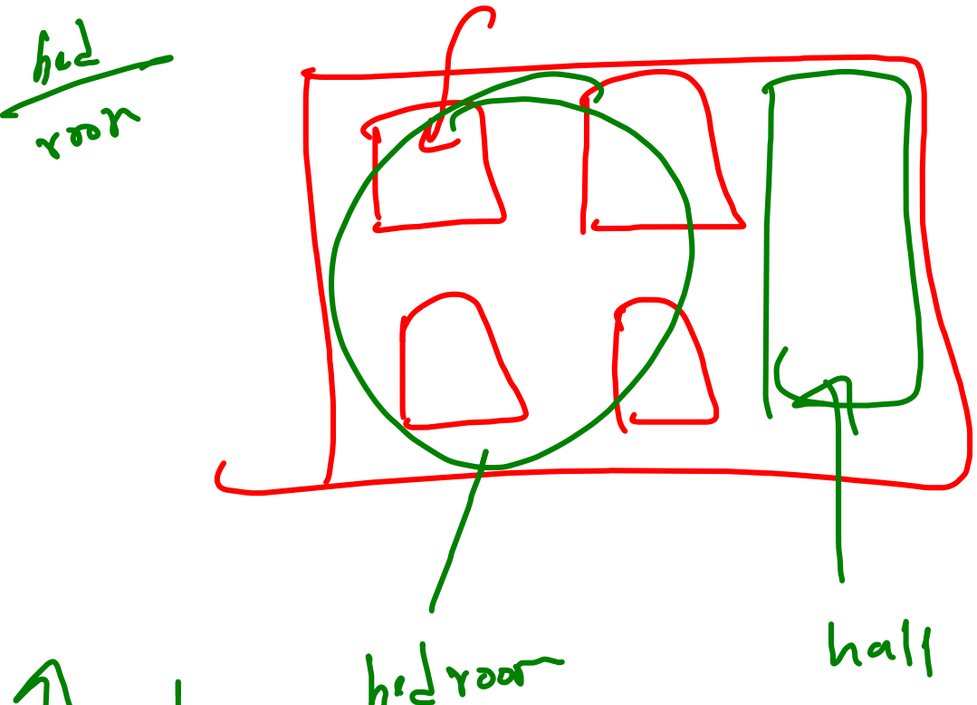
a	b	a+b	y
1	10	11	13
6	7	13	15
9	-4	5	7

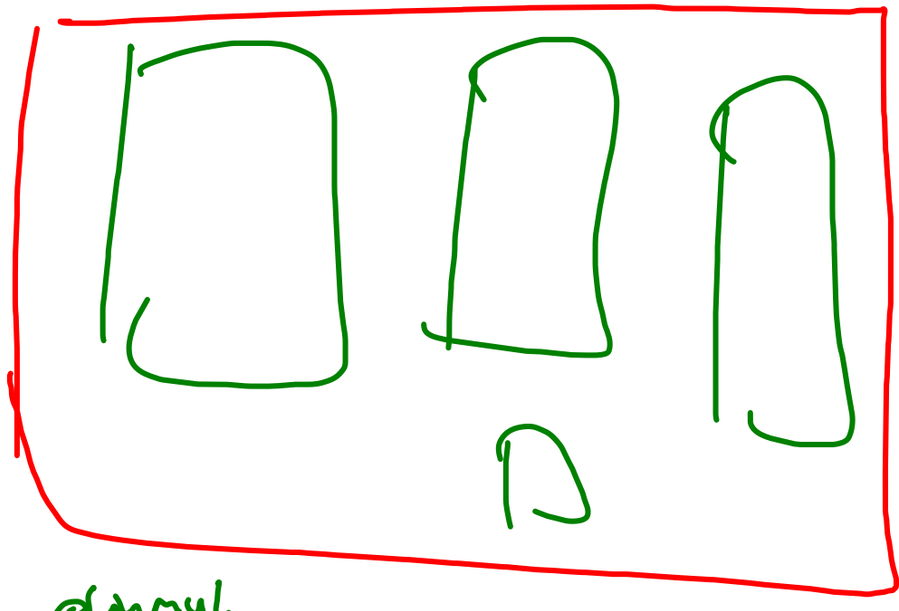


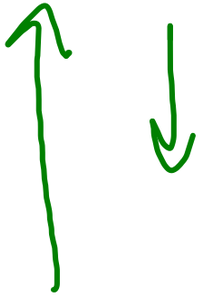


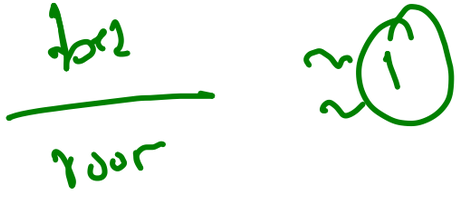

10
10
20

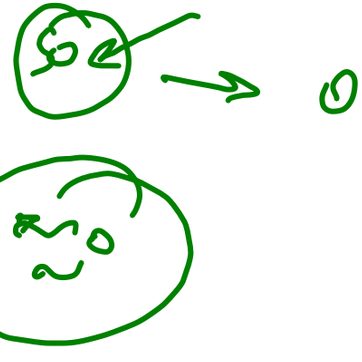

2
2
1
4
1
2
6
4
20

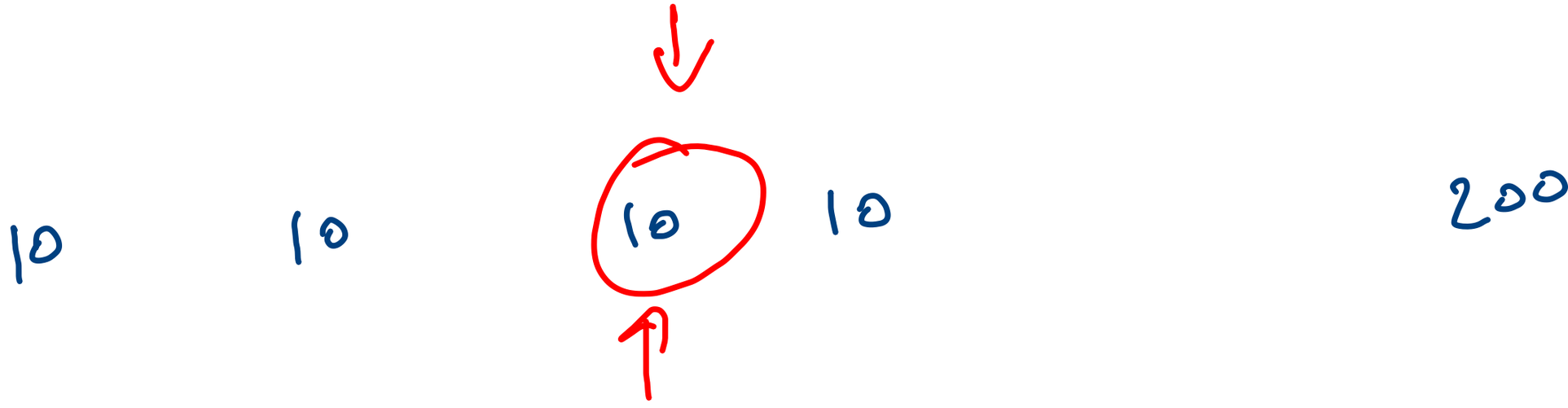

bed room
bedroom
hall



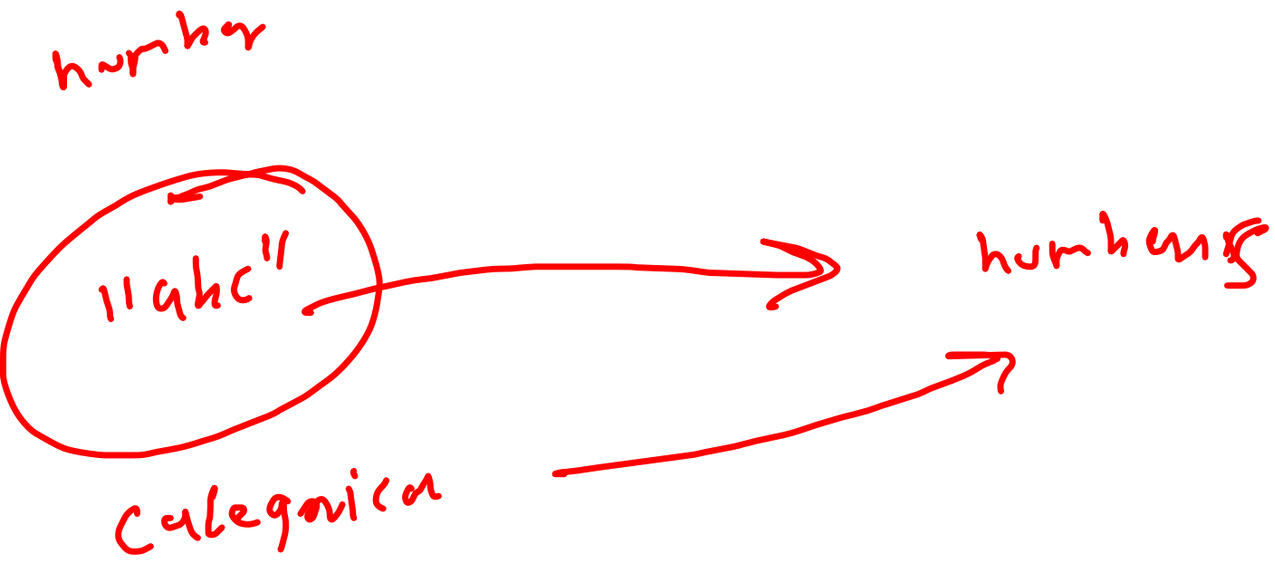


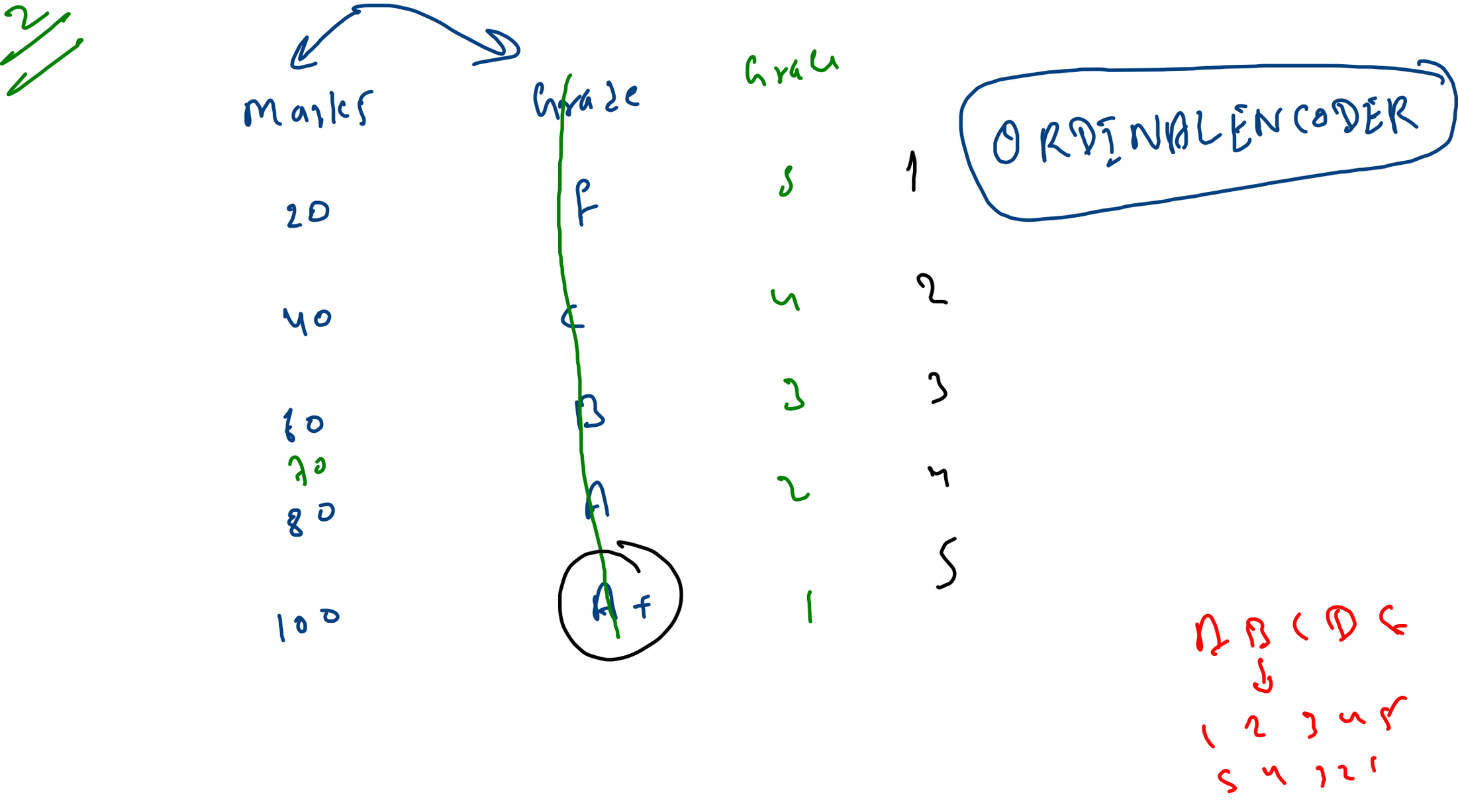

bed
room




10
10
10
10
200


240
5
=
48


number
"abc"
numbers
Categorical


Marks
Grade
Grade
F
E
B
A
A f
20
40
60
70
80
100
5
4
3
2
1
1
2
3
4
5
ORDINAL ENCODER
A B C D E
1 2 3 4 5
5 4 3 2 1

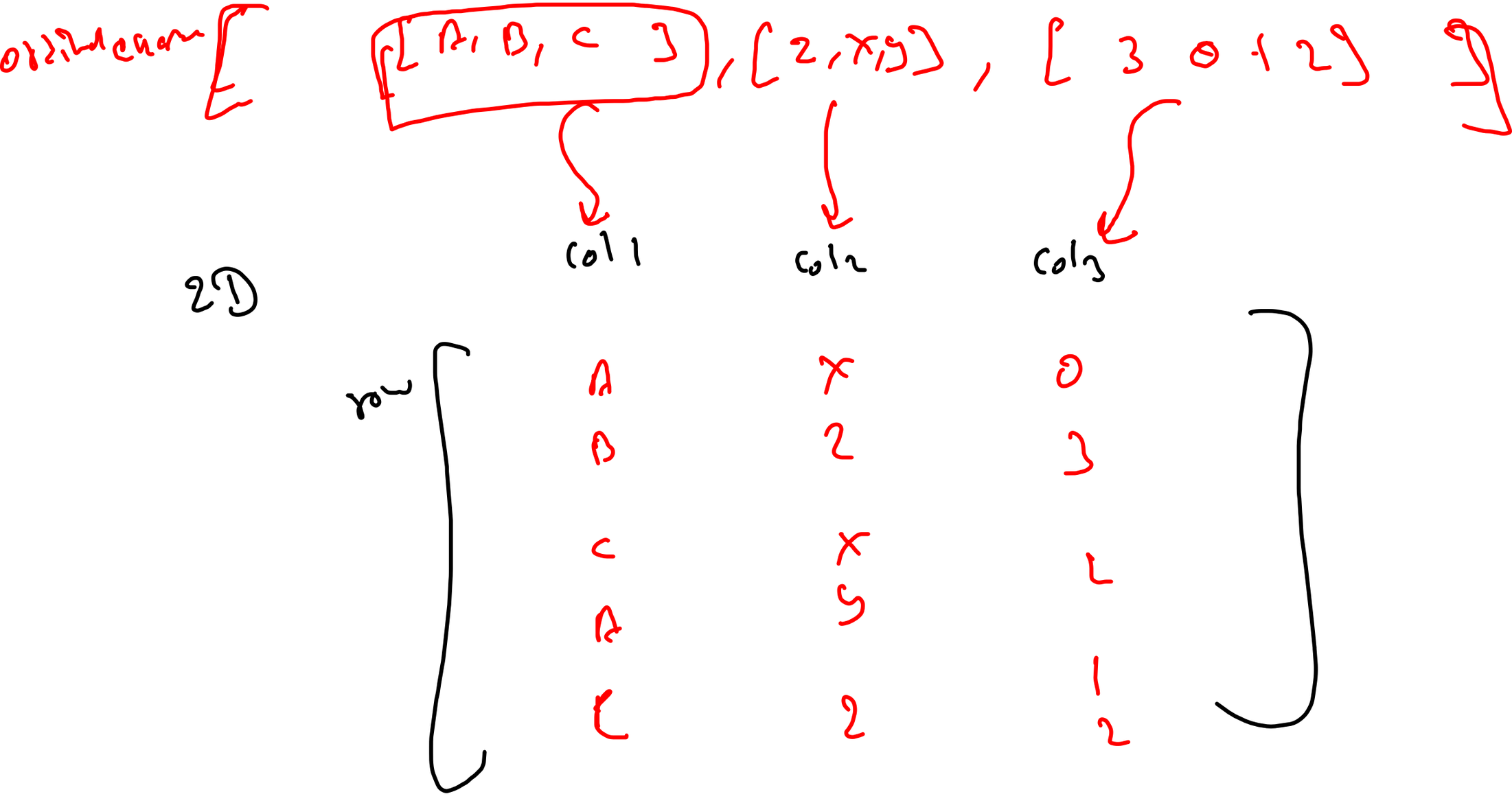
ONE HOT ENCODER

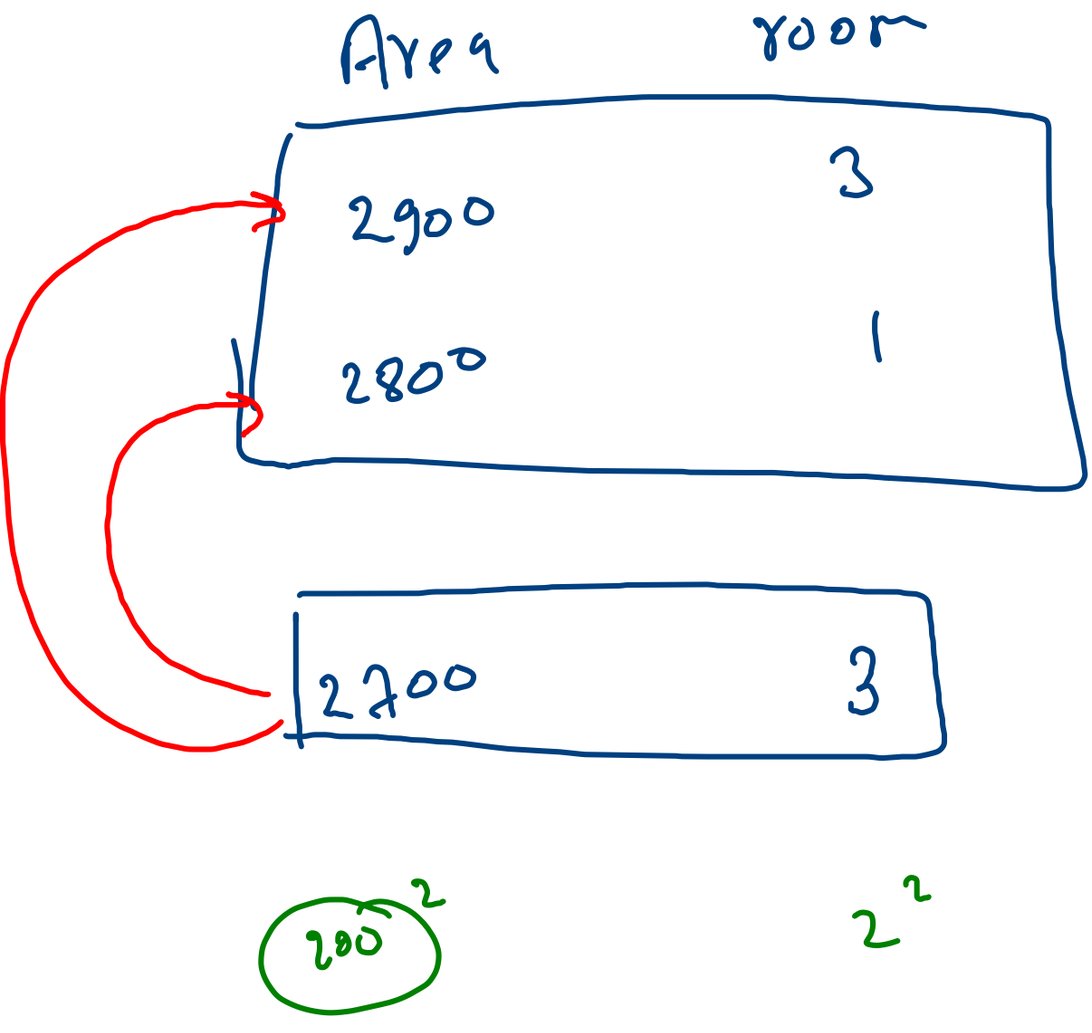
get_dummies

Marks
20
40
60
80
100

State	
Goa	1
Mumbai	2
Kolkata	3
Delhi	4
Delhi	5

Goa	Mumbai	Kolkata	De
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
0	0	0	1


[[A, B, C], [2, X, y], [3 0 1 2]]
col 1
col 2
col 3
2D
row
A X 0
B 2 3
C X 1
A y 1
C 2 2



200^2 + 0^2 — greater

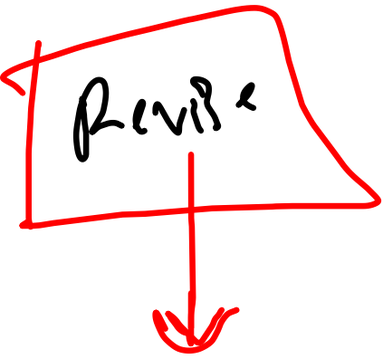
100^2 + 2^2 — smaller

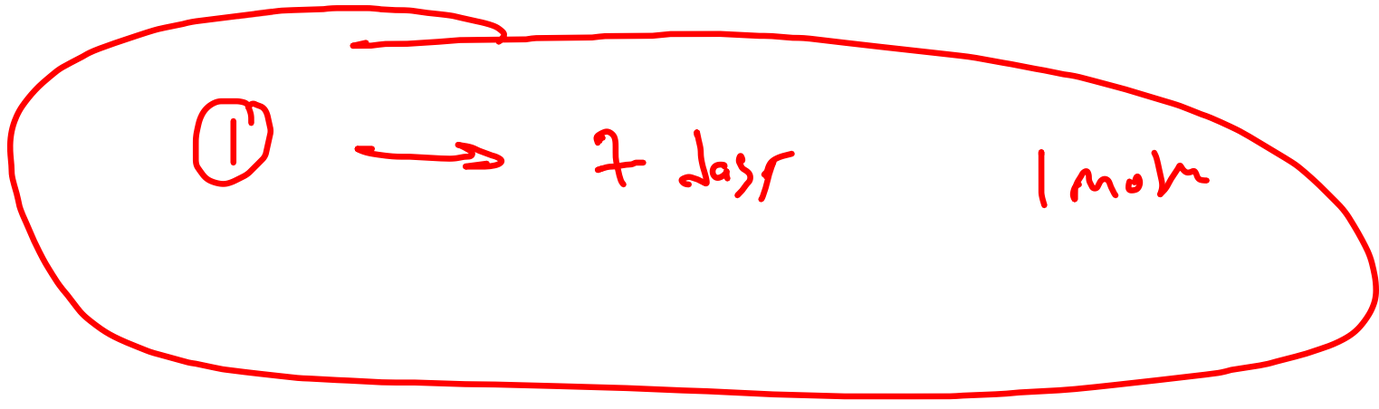
200^2 2^2

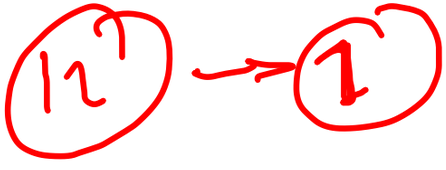
Standardization → μ σ

Normalization → $[x, y]$

- min-max
- Robust Scaler


Revise


1
7 days
1 mon


12
1