Practical Big Data Analytics

* Michael.Burgess(@qa.com)

~~1.30 — 7.~~30

The Skills to Succe

(4- 4.30)

9.30 — 4.30

Traing _ AP

c.12

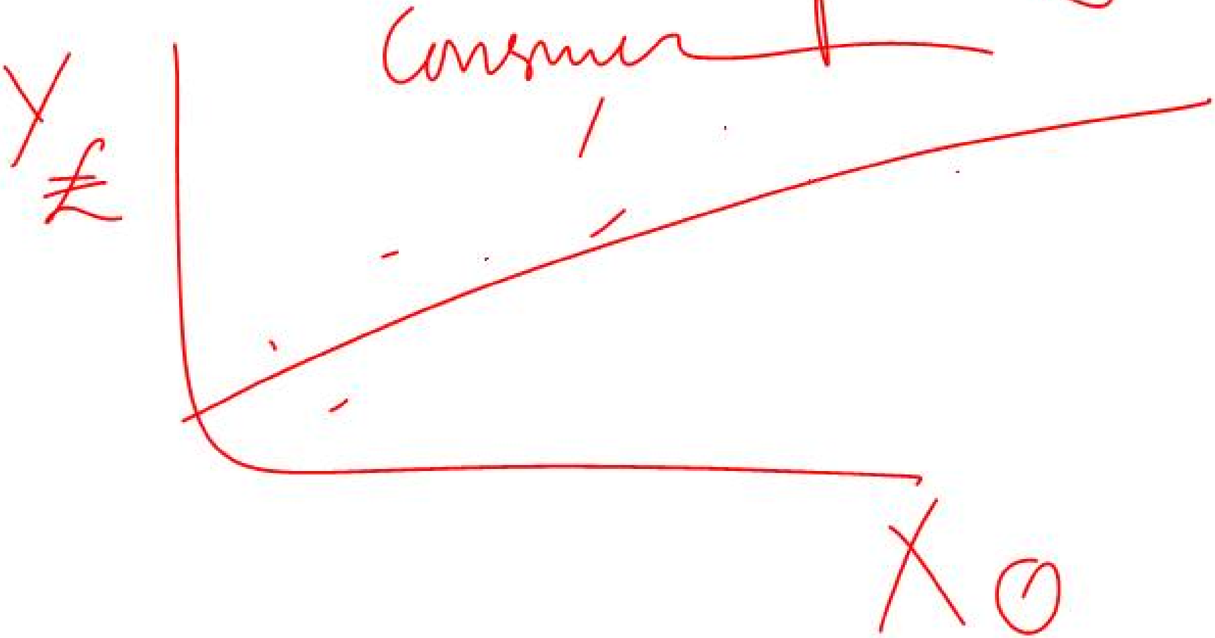
1pm

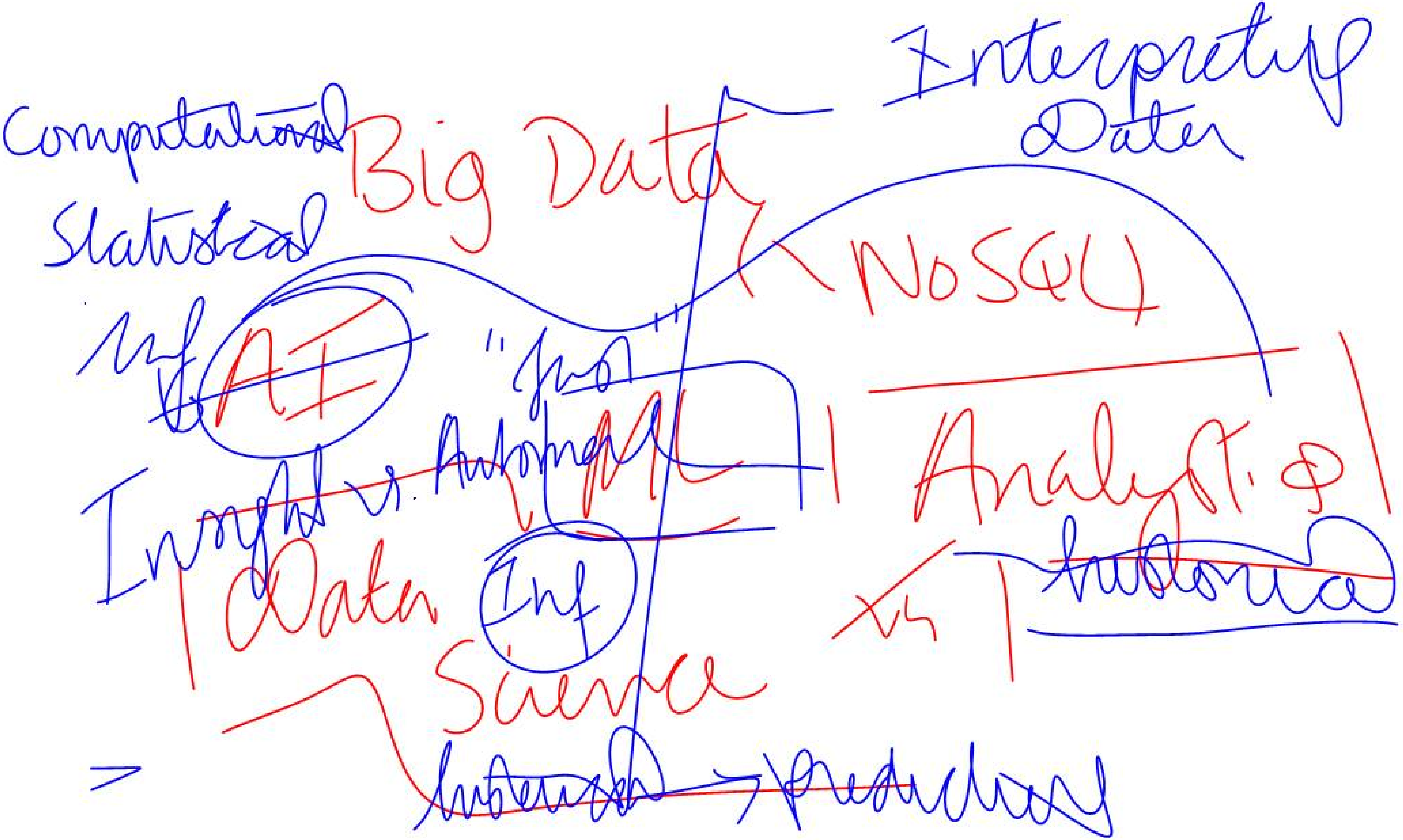
c

11am

c — 2 30pm

~~Computers~~ Data ↑




Computational
Statistical
Inf
AI
Big Data
Interpreting Data
NoSQL
"Info"
Insight vs Automation
ML
Analytics
Data Inf Science
vs
~~historical~~
>
~~inference~~ → prediction

~~V^3~~ ~~V^5~~

amount
structure

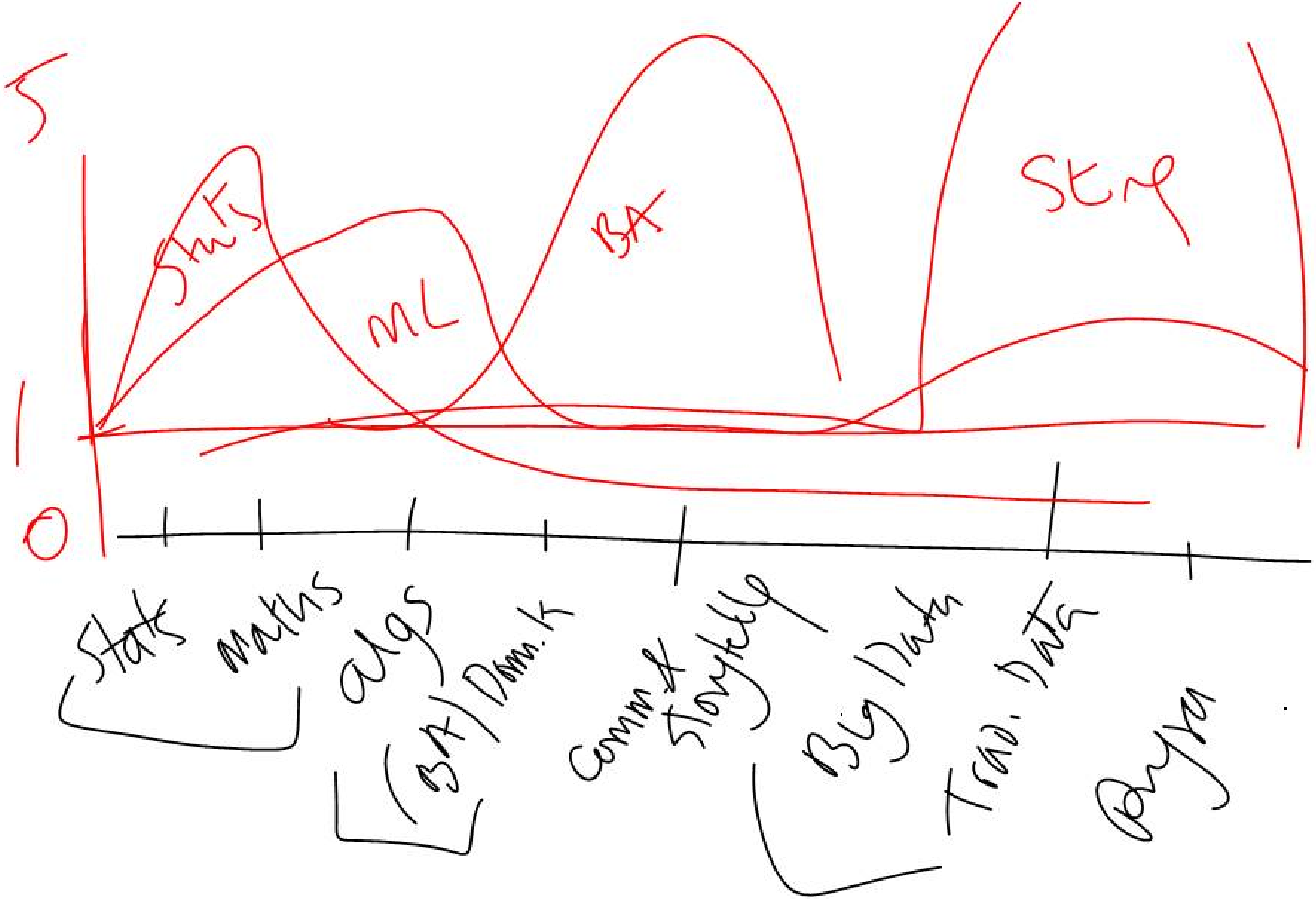
⊛ Vol.
Variety
Velocity

⊛ Skills &
tools

[Flexibility] BD = Non-trad method

Eng

Analytica


5
1
0
Stats
ML
BA
Strg
Stats
maths
algs
(BA) Dom. K
Comm & Storytelly
Big Data
Trad. Data
Python

7pm

May need QAPYTH3, or eq.

T Quick Review

- Wrap up python
  - fns
  - import (#2)
- Courseware Review

→ tinyurl.com/bda-10-02-20

✓[✓>5]

in python.

dir(x) ← lists all operations on x

Eg dir(dataframe)

⋮

read_csv()

⋮

to_csv()

⋮

@qa.com

m

michael.burgess@qa.com

burgess@qa

Machine learning → Self D

→ AI

> Maths

Rec

Alexa

Data → training

alg/model

X ← knew

Y ← ?

hist(X, Y)

$$\boxed{\hat{Y}(X) = \underset{\uparrow}{3}X + \underset{\uparrow}{5}}$$ model

Music Ca

Computational Stat-
inference

Parameters

Propositions

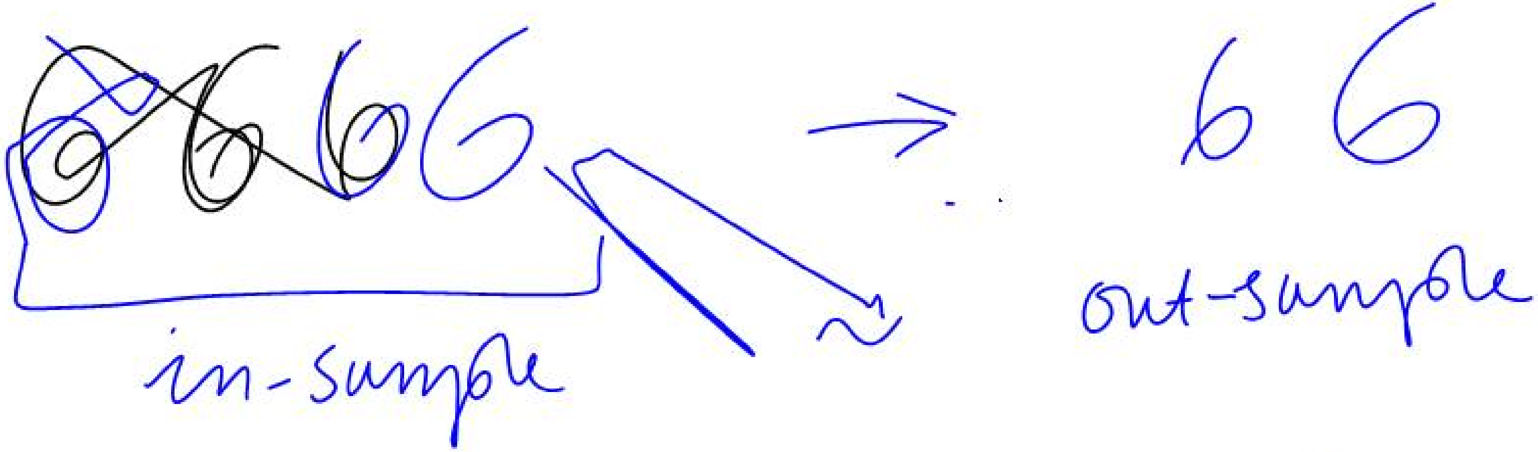
AI?

— 40s

↓ 50s

00s

```
if (a > 18):
    open()
else
    close() x
```



£UK ≠ €FR

France

anti

① in ~ out

② in ≠ out ← unsafe

ML

Regression ← $\hat{Y} \in \mathbb{R}$

1) { classification $\in \{-1, +1\}$ Binary Class

sup. learning $\in \{0, 1, \dots 5 \dots\}$

historical $(X, Y) \xrightarrow[\text{training}]{\text{learning}} \hat{Y}(X)$ min $\text{Loss}(\hat{Y}, Y)$

2) Unsup.

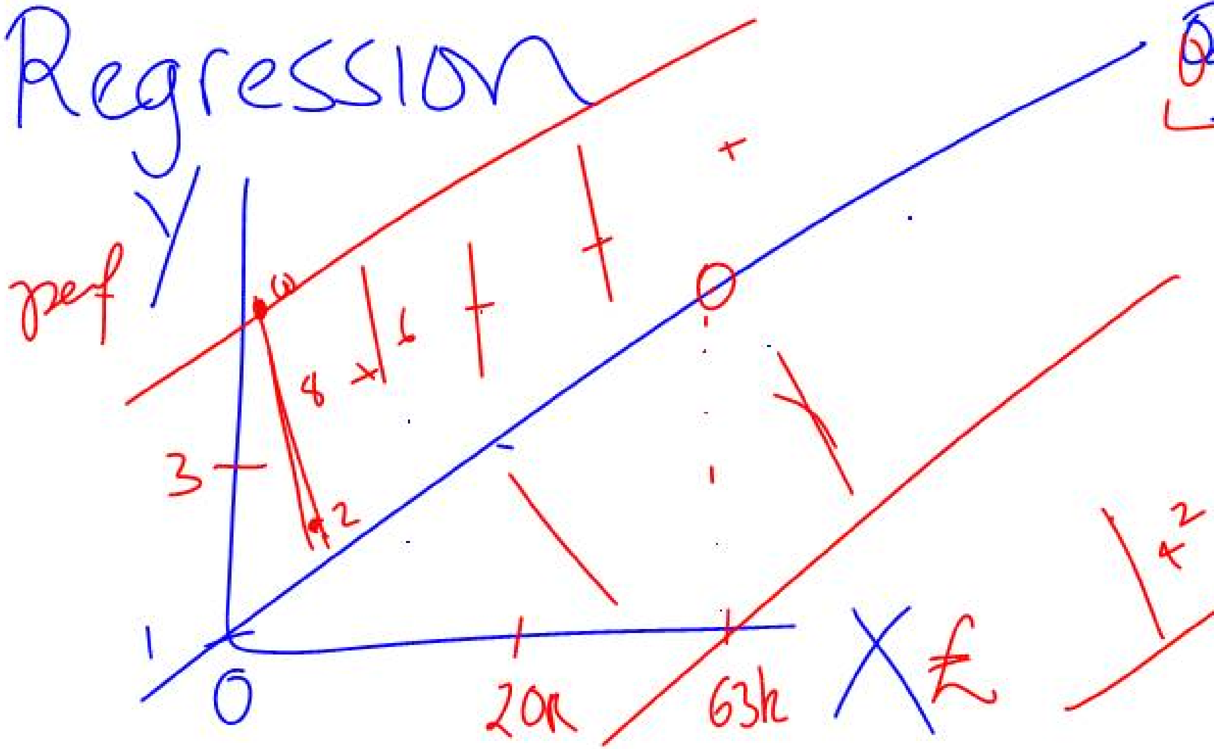
$(\overset{z}{X_1}, \overset{a_0}{X_2}, \overset{t}{X_3})$

① Simplify (Dim. Reduction)

② Common Occurrence

$P(\cancel{X}, X, X)$

Regression



$0.1x + 1$

$\frac{1}{10}$

(x, y)

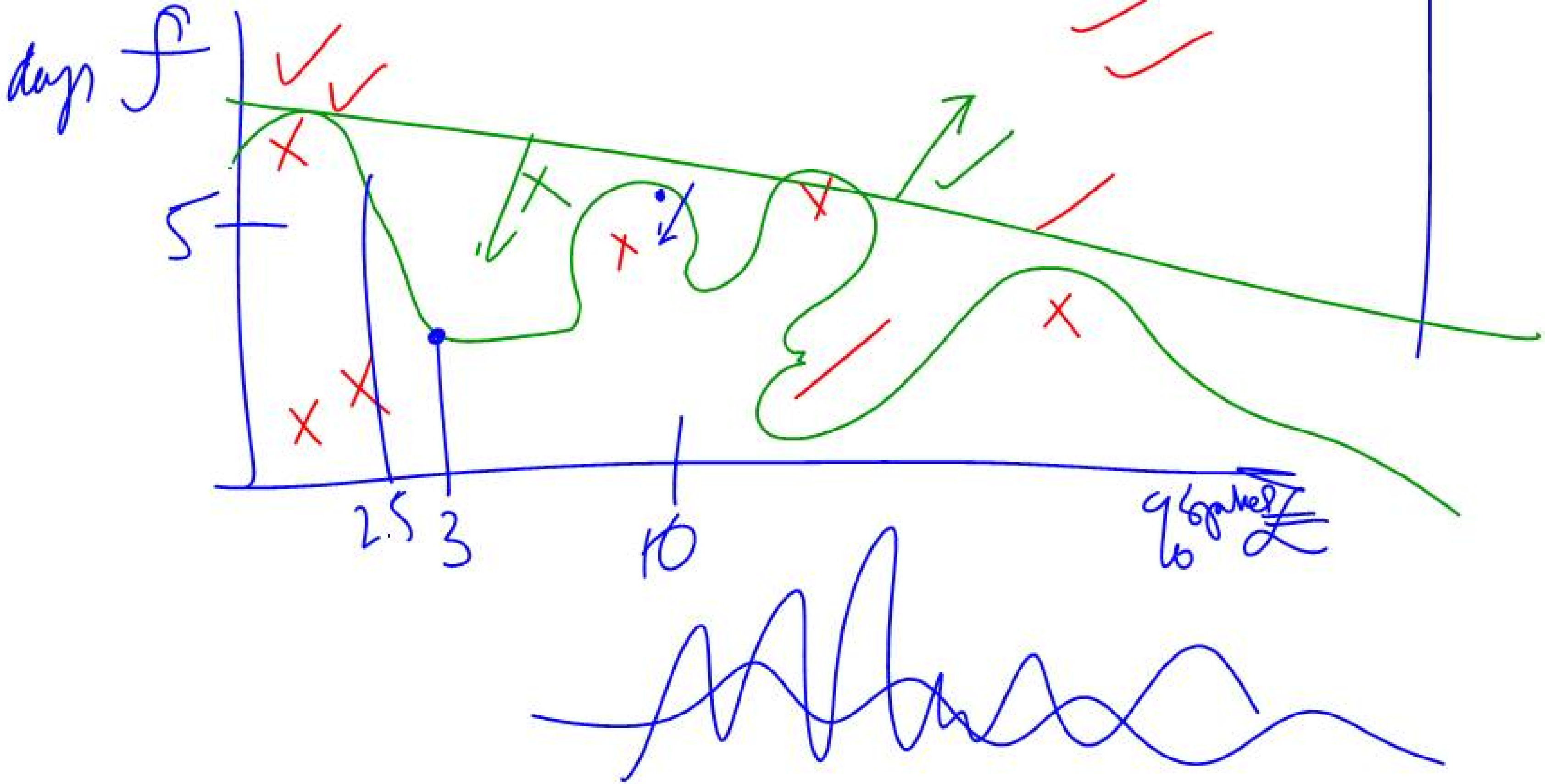
$(0.1, 1)$

$\text{loss}(\hat{y}, y)$

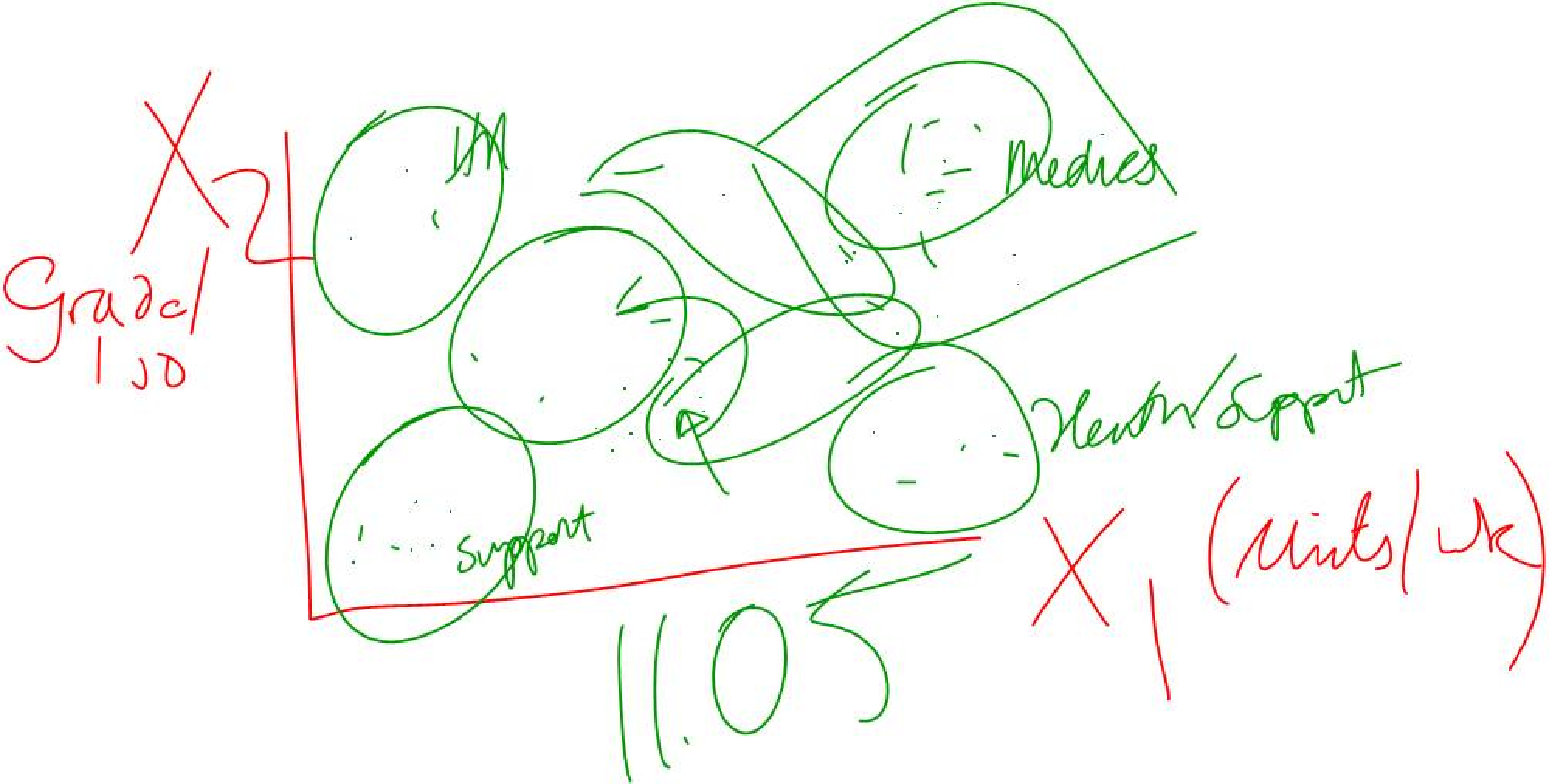
$|4^2 = 16$

$|-4^2 = 16$

Classification



X1	X2	X
100	10	✓
		✗
		✓
		✗


X2
Grade/
1-10
IM
Medics
Health/support
support
X1 (units/wk)
11.05

7pm

(x^i, y^i)

data model

stat model

assoc

Expl

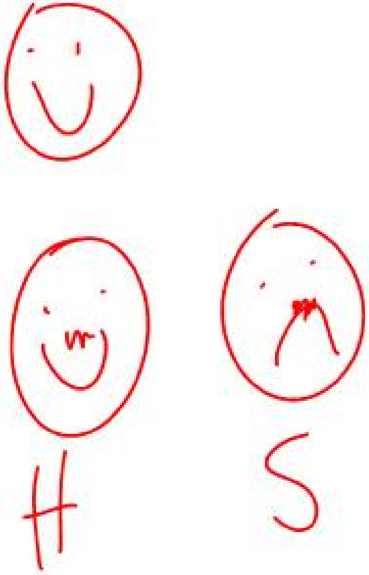
$$\hat{y}(x_i\ w) = w_1 x_1 + w_2 x_2 + \underline{w_3}$$

var

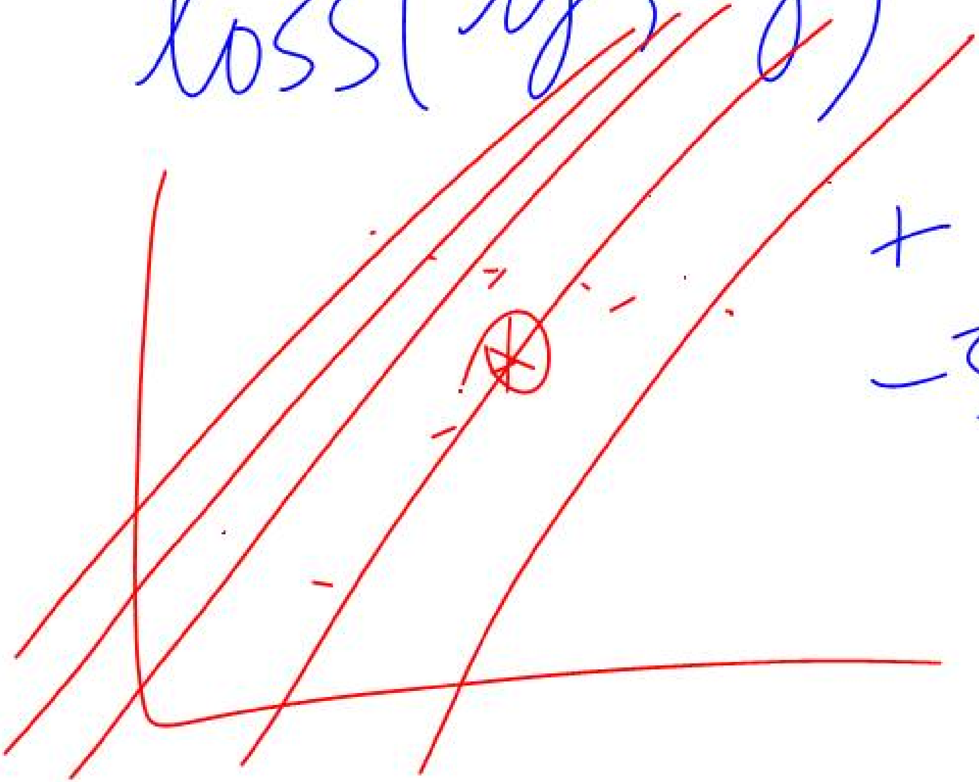
parameters

a single number

Causal


H
S

$$\text{loss}(\hat{y}, y) = (\hat{y} - y)^{-}$$



+5

−5

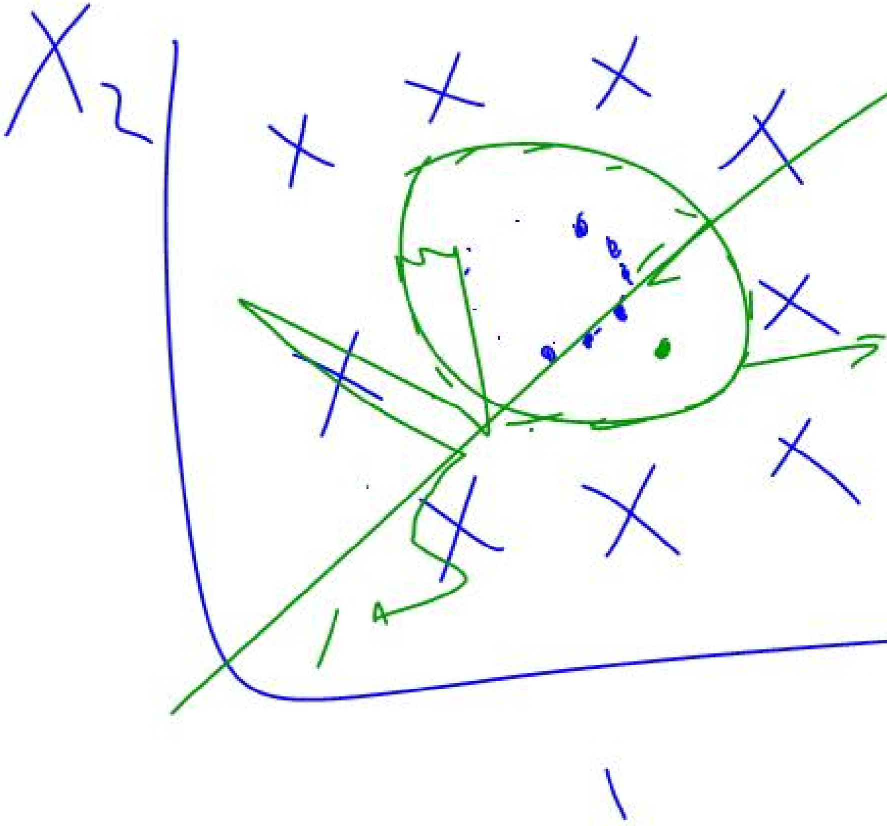
$y \in \mathbb{R}$

$x_1, x_2 \ldots \in \mathbb{R}$

<u>find</u> $\hat{y}(x\,.\,,w\,.\,.)$

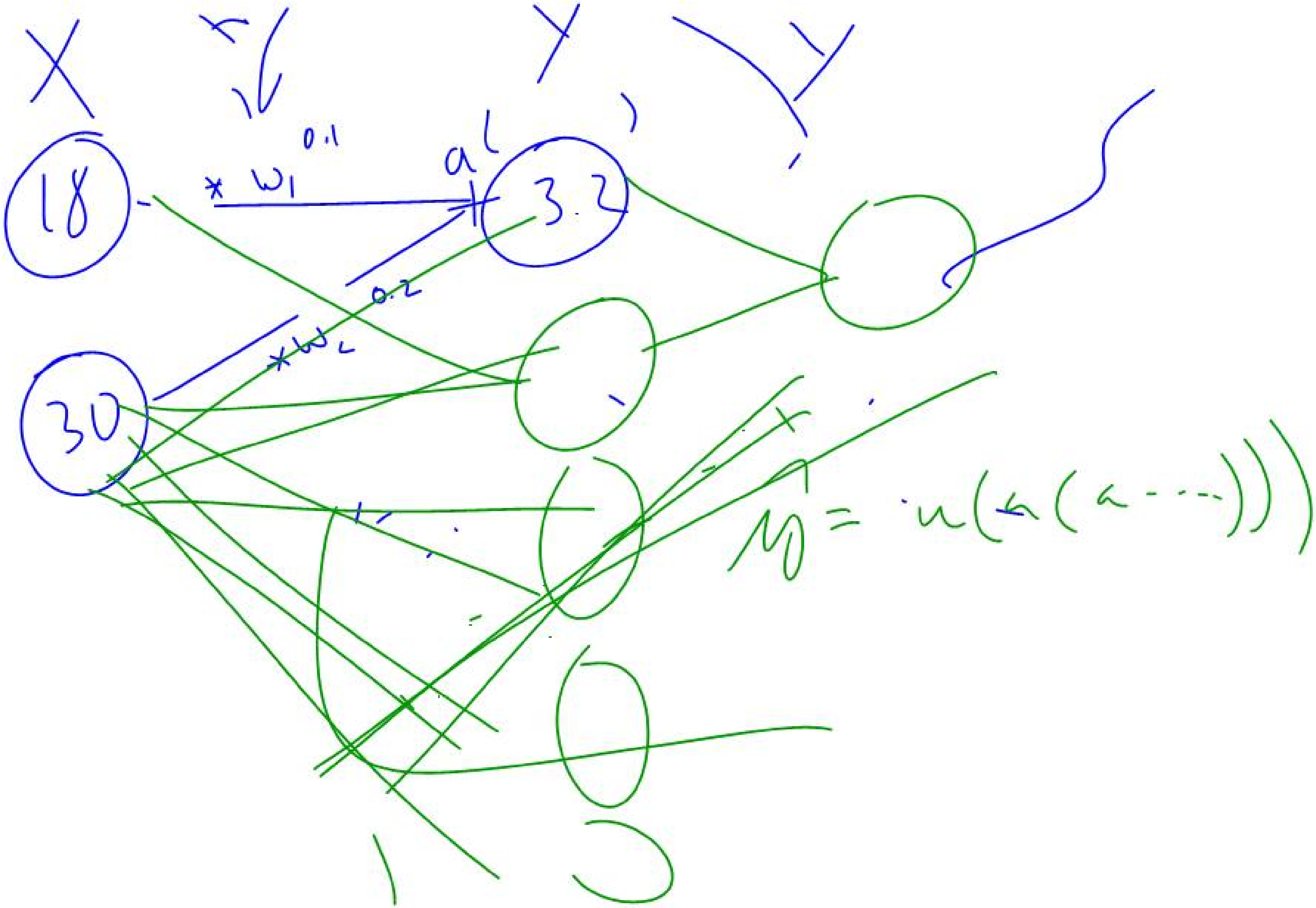
by $\boxed{\min}$ $\text{loss}(\hat{y}, y)$

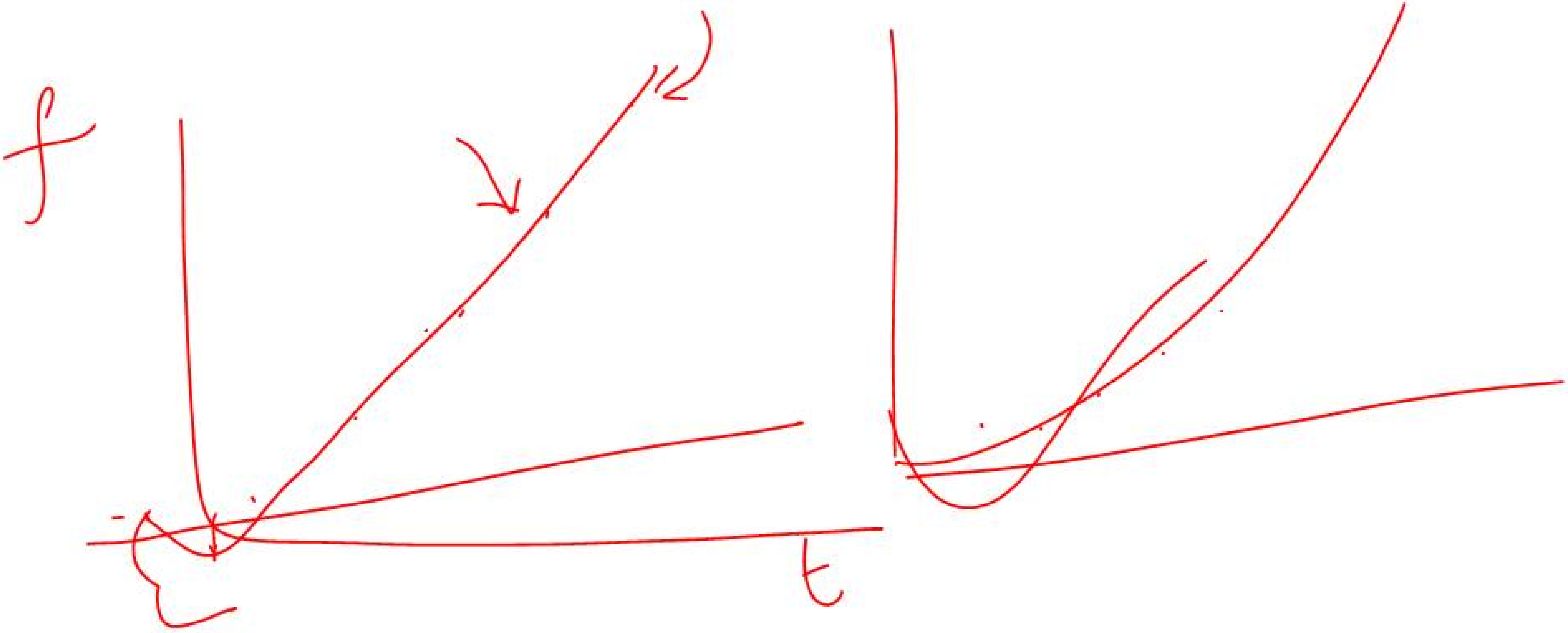
How? <u>Alg</u> — Grad Des.

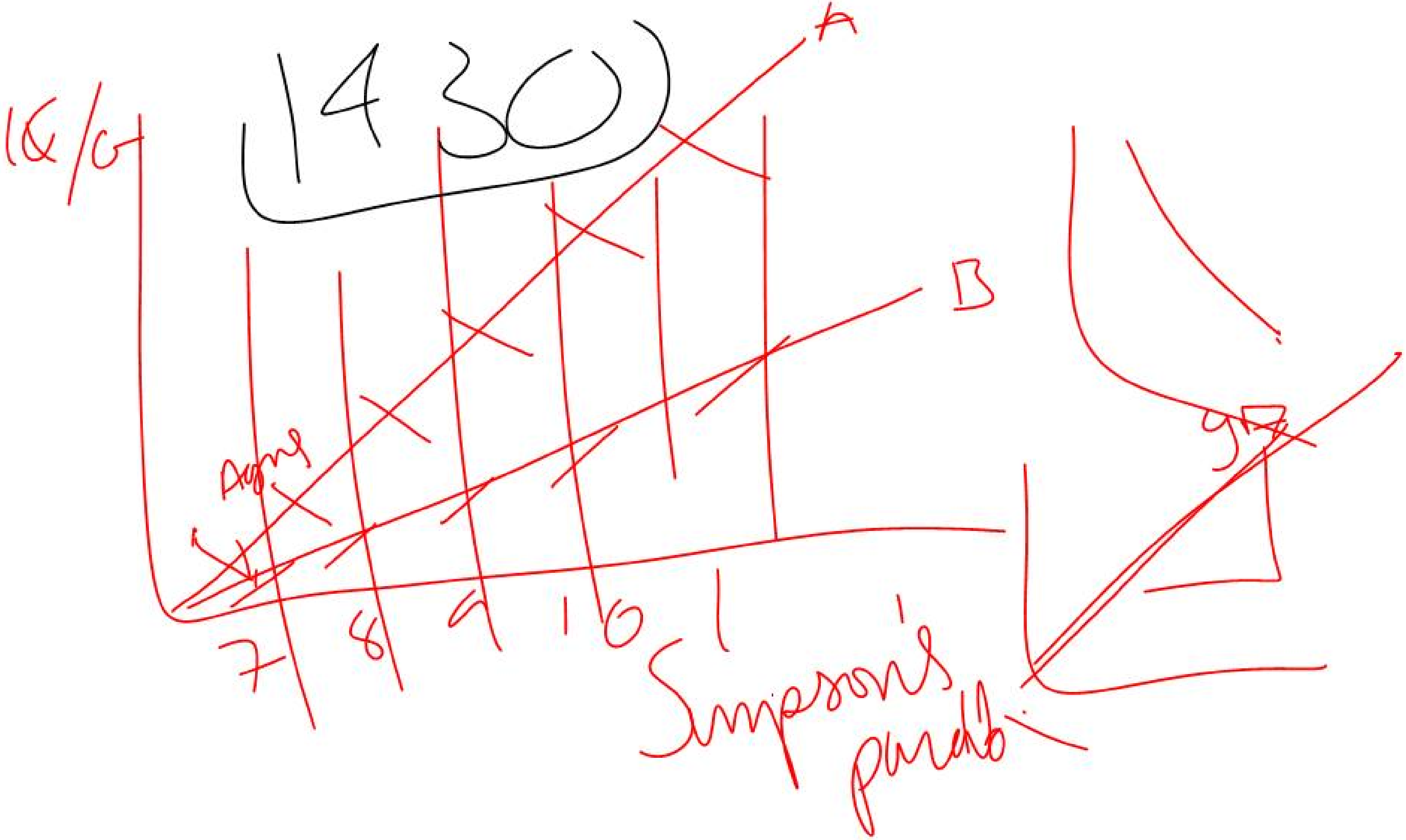


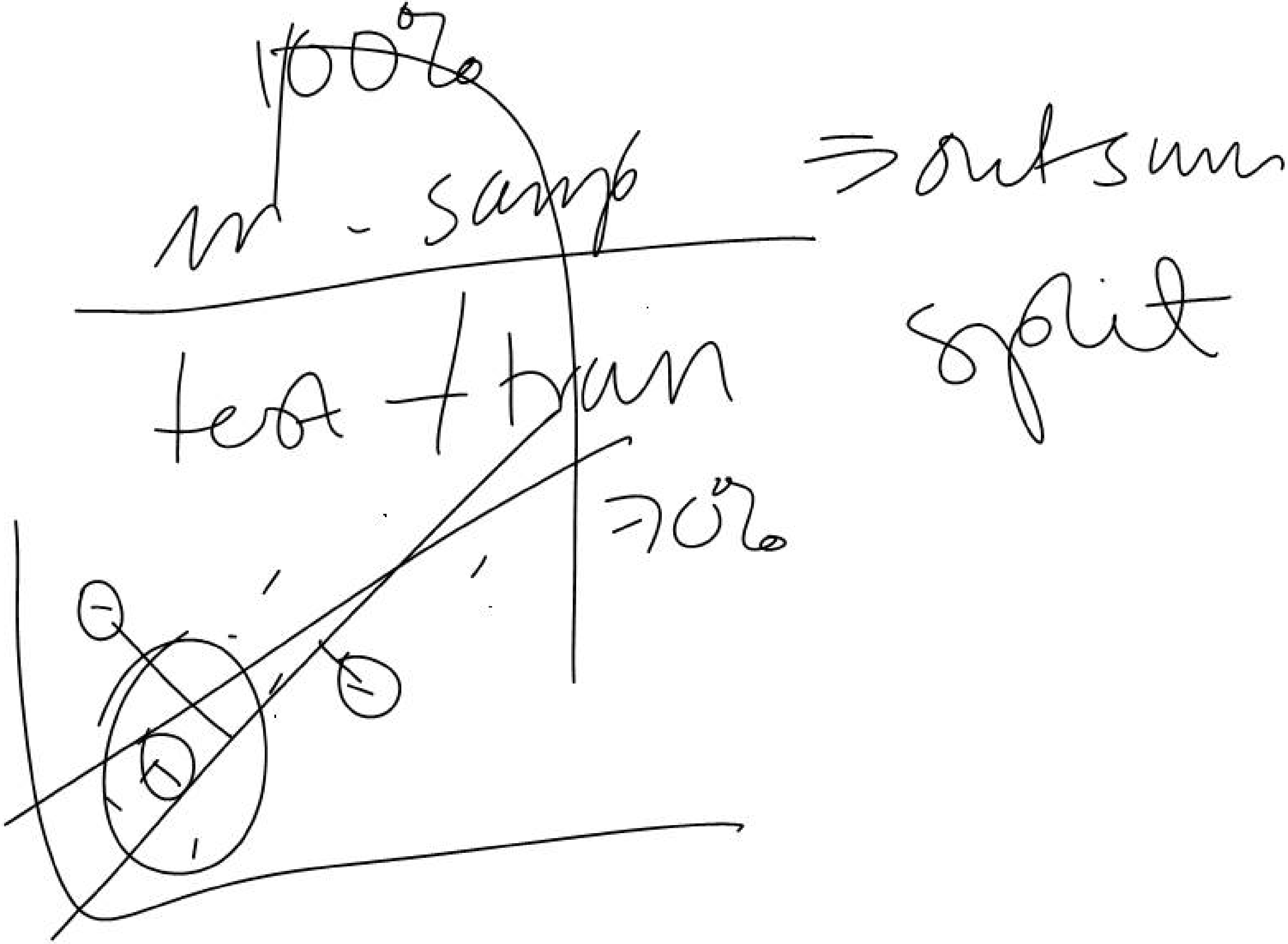
$$\hat{y}(x;w) = \boxed{w_1x_1 + w_2x_2 + \dots}$$

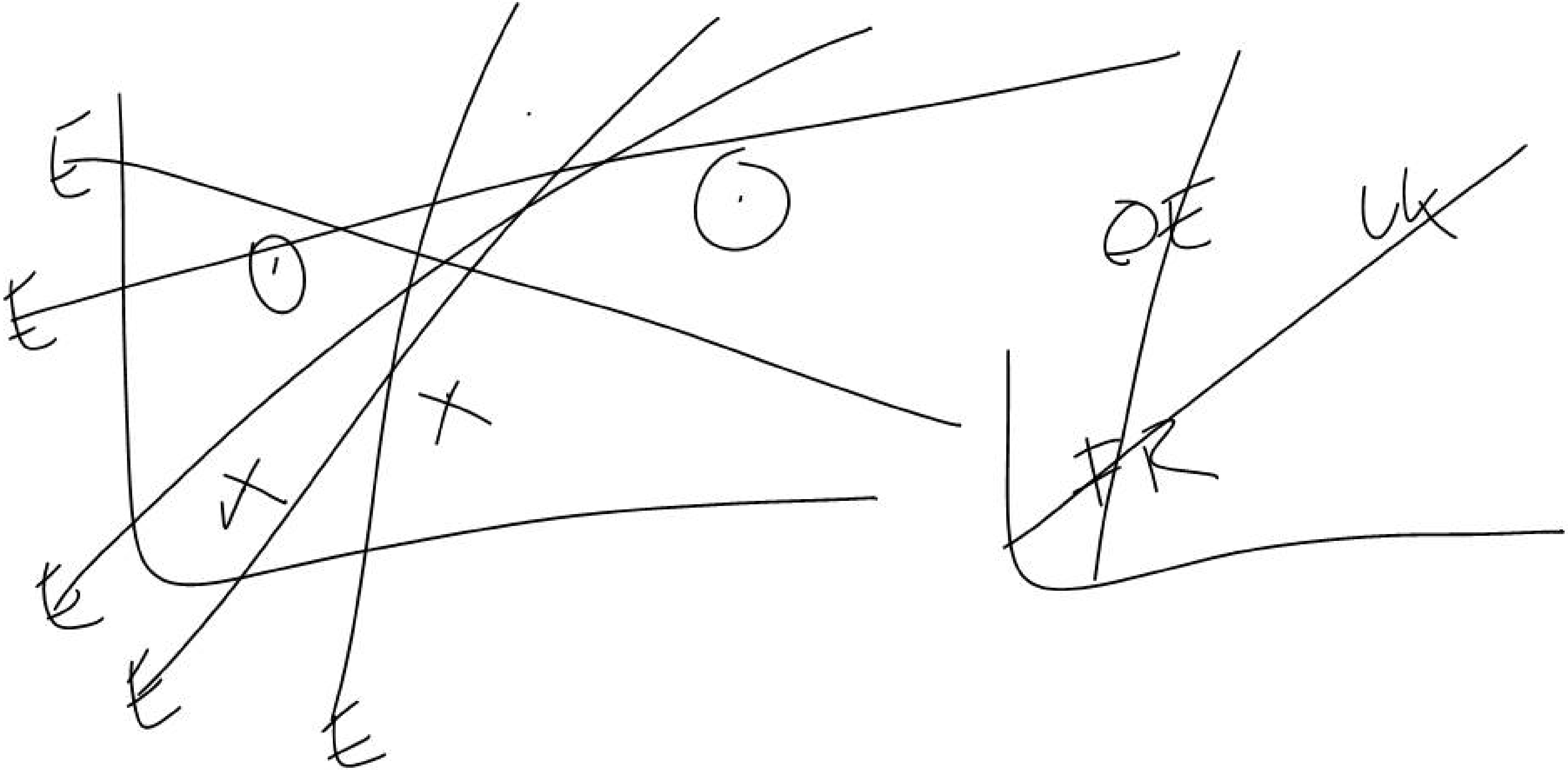
$$\sigma(a(a(wX)))$$

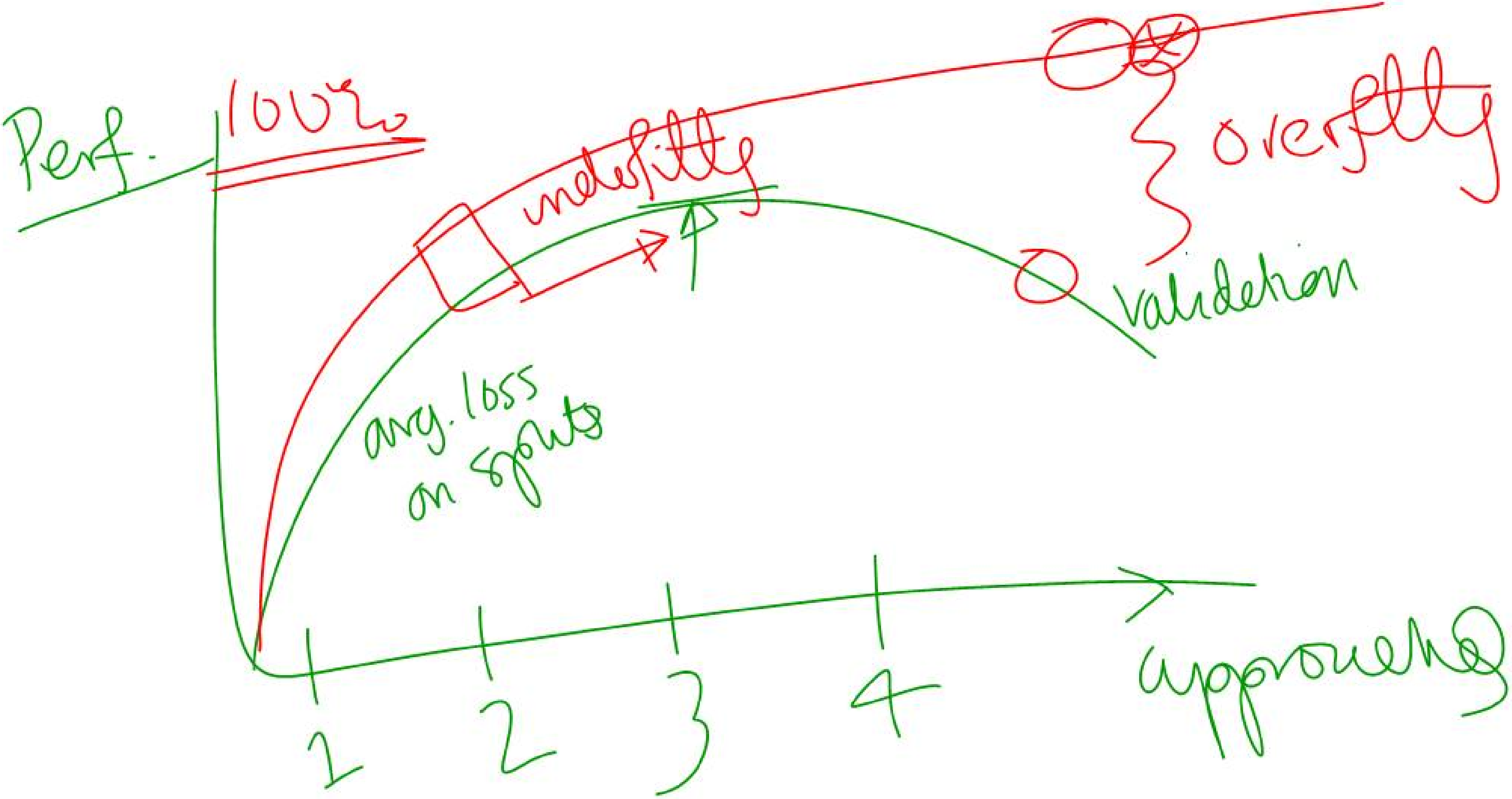

X
18
30
* w1
0.1
* w2
0.2
a(
3.2
y = a(a(a ....)))

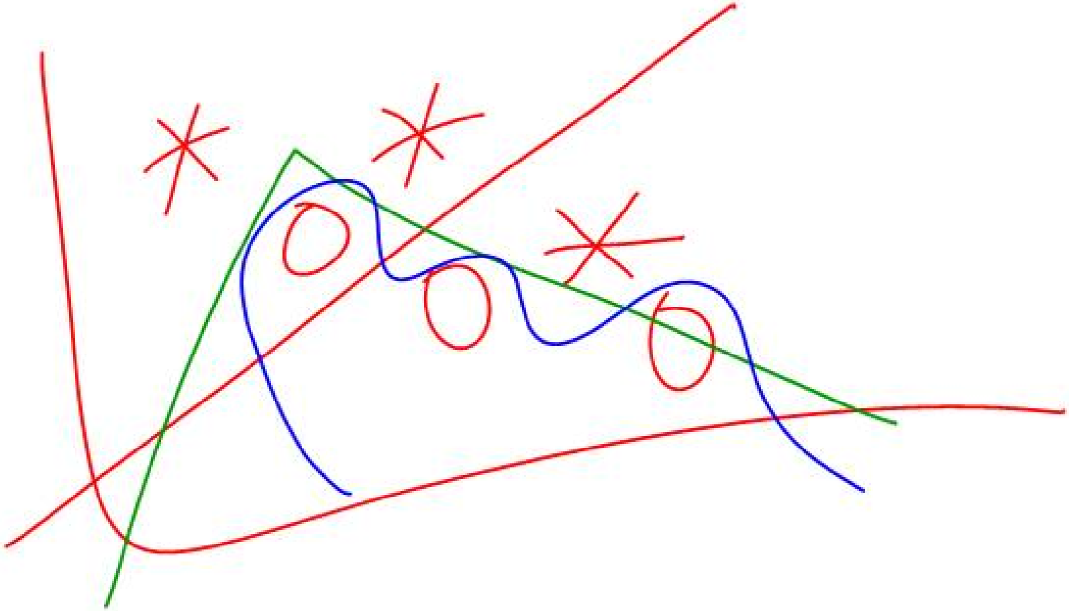

f
t


IQ/G
1430
A
B
Avg
7
8
9
10
1
Simpson's paradox


100%
in - samp
=> outsum
split
test + train
70%


E
E
E
E
E
OE
UK
FR


Perf.
100%
underfitting
overfitting
validation
avg. loss
on sports
approaches
1
2
3
4



1) ML workflow

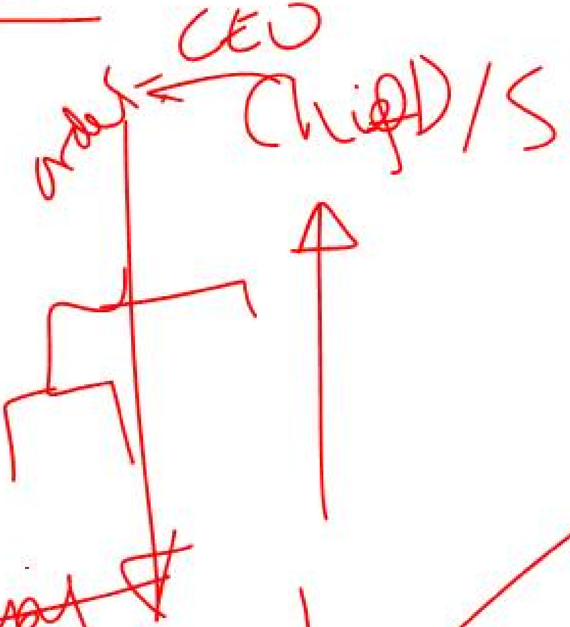
B3

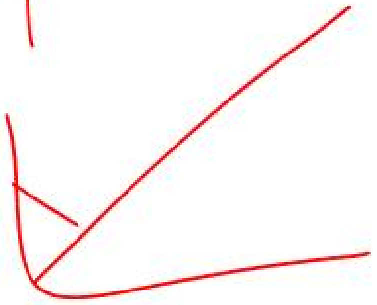
(Exploratory / (x, y))

⊗ Can we do you anything?

⊗ problems? ~~Gatekeepy~~

Opions — ~~HiPPO~~ Ia of sup.





1 Understand. Prob (BA)

2 Data Understand

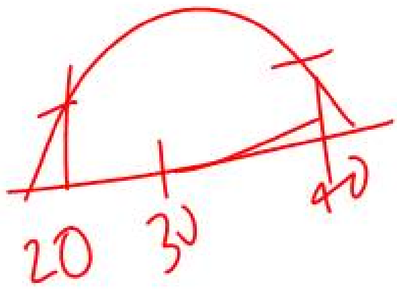
Clean

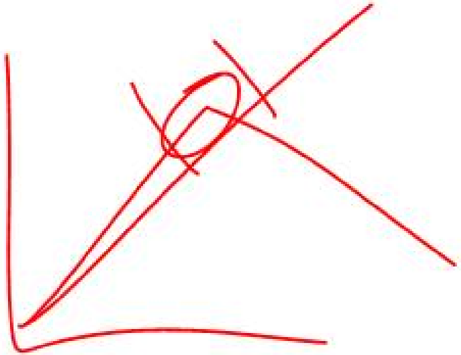
EX. D A

age

x_1 x_2 . . . .

1)





1.
N
N
:
y

N
N
N
N
N
N

1000

3

99276

③ Data Prep $\frac{kg}{m^2}$ $\frac{w}{h^2}$ w^{kg} h^{m} % Die

Feature Eng (x)

- Clean ← remove missing
- ETL ← Joining (Enriching)
- Transformation ← Changing repr
- XL ~~Encoder~~

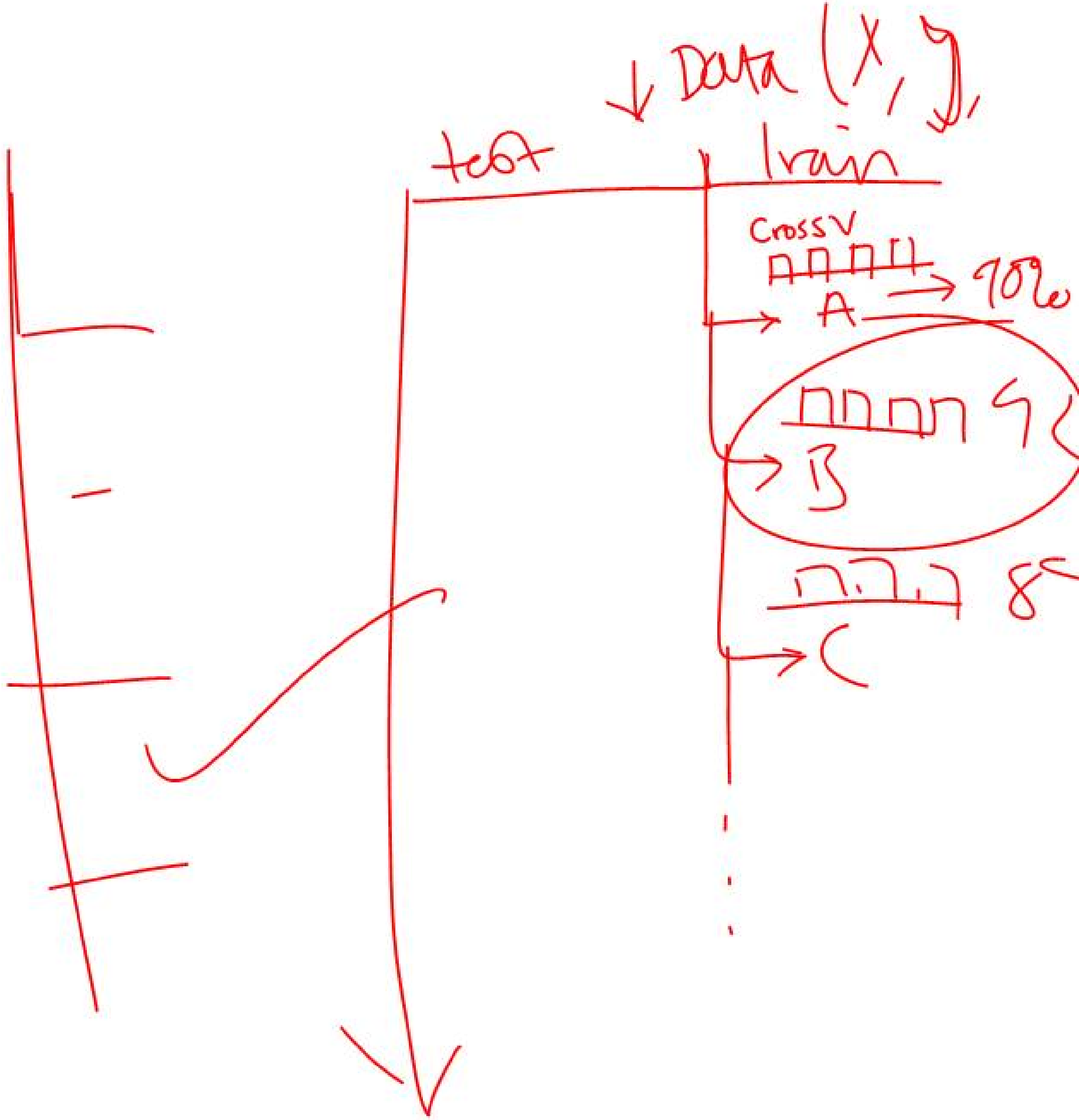
City
LOS → 1
MAN → 2 > ② 1
LON → 3

OHE →

isLON	isLOS	isMAN
1	0	0
0	1	0
0	0	1

④ Modelling

NN-1
-2
-3
LR-1
.
.
.



Evaluation

test

B

Eval

88%

100%

Retrain model in test + train data

Repeat

~~Deployment~~

58 to fly

55 [illegible]

① Exp] ② 1 mo +

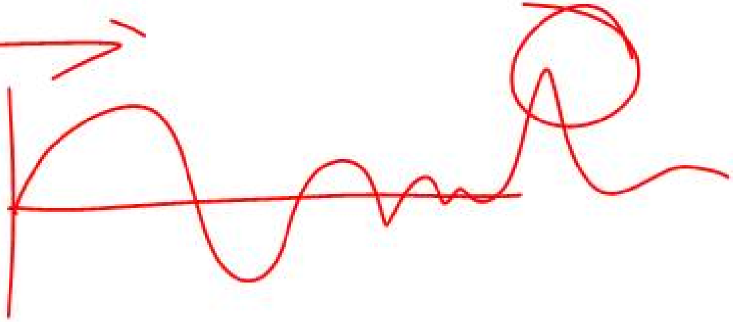
② Problem → $\frac{1}{2}$ mo +

③ Data 1 mo +

④ Modelly $\frac{1}{2}$ mo +

⑤ Eval $\frac{1}{2}$ mo +

⑥ Deployment $\frac{1}{2}$ mo (1 mo / 2 mo)



6mo —— years

Big Data

1000B
1TB

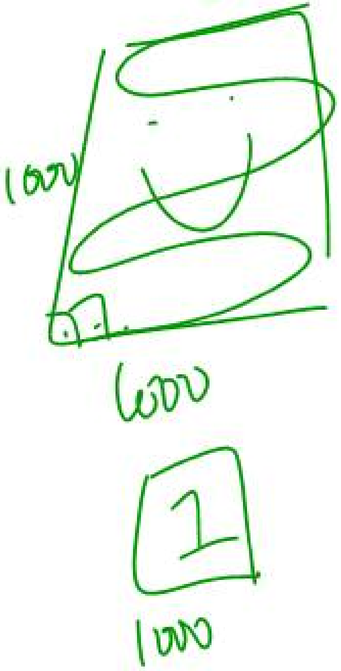
(non-tr.)

2005 → 2015

Skills

Variety ← "Dimensionality"
→ vary indpt

image



Velocity

24h.s

10MB/s

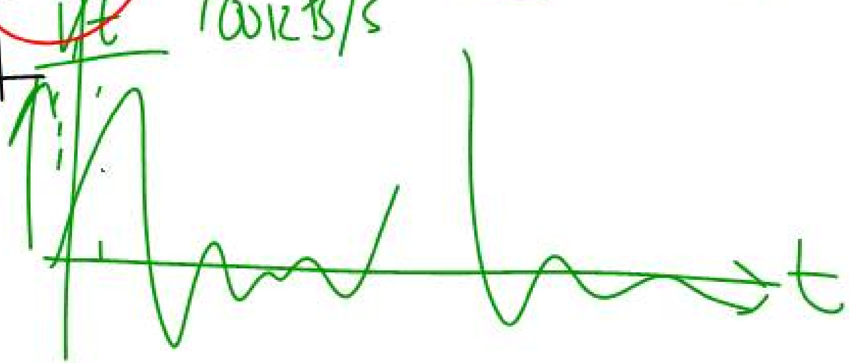
Volume

10TB

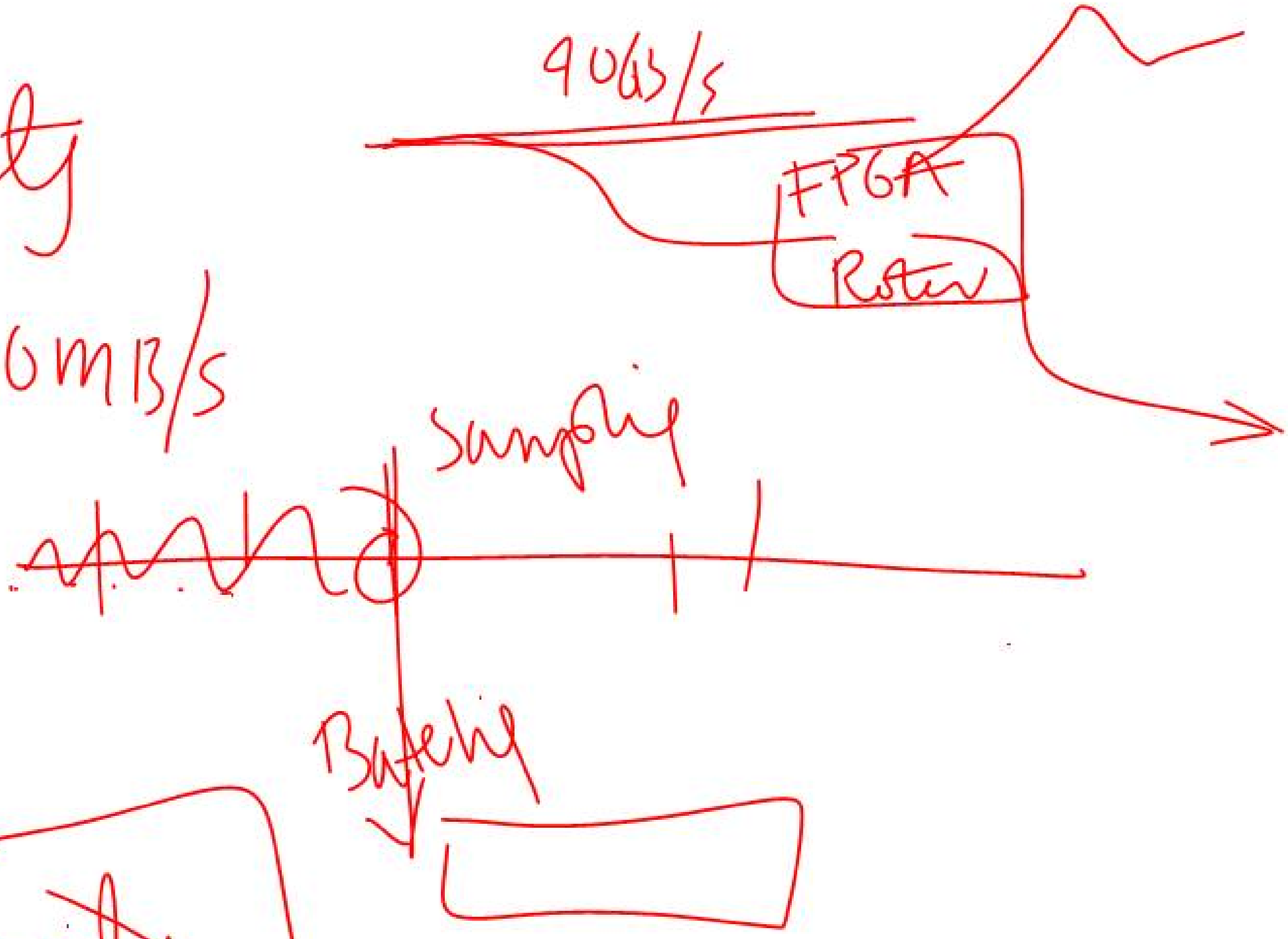
V / 1PB

V

[Json (K-V
F-C)] >



Velocity



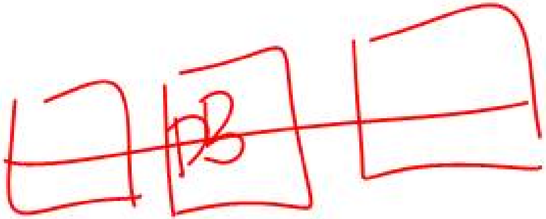
Variety

~~Value~~

Veracity

Non-traditional

- BI
- Relational — SQL →
- "One Machine"



Velocity

→ ingestion

Big Data

[illegible]

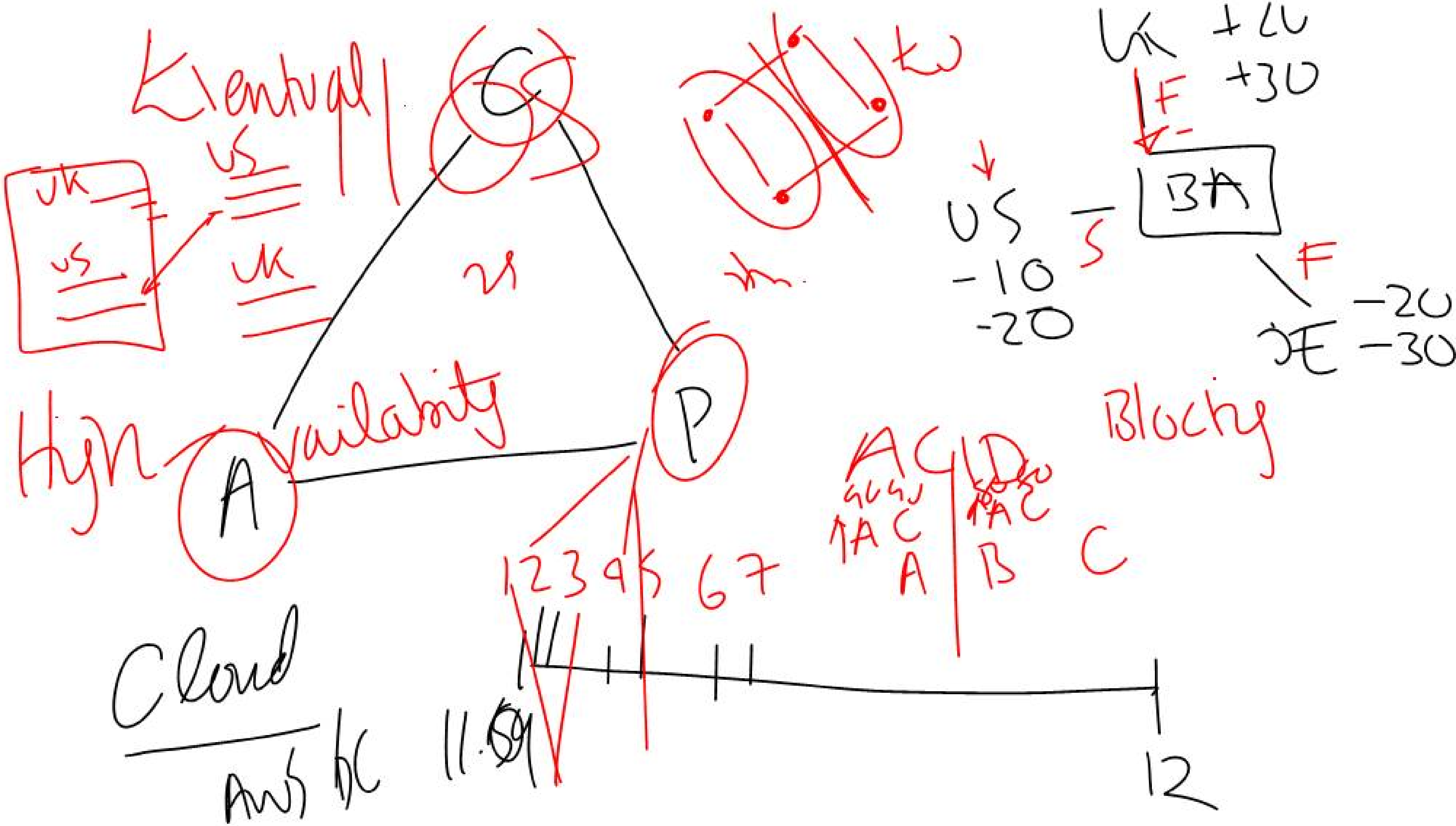
[illegible] Analysts [illegible] / Engineers

1) Data Structure → CAP

2) Query Structure → | Consistency
availability
partition tolerance

3) "Modelling Querying
~~"Representational~~ power"

"Readiness"


Eventual
C
A
P
21
tw
UK
US
US
UK
Hygh Availabity
Cloud
AWS DC
11.59
1 2 3 4 5 6 7
12
ACID
1A C
A
1A C
B
C
Blocking
UK +20
+30
F
BA
US
-10
-20
5
F
DE -20
-30

Data 11.05

Data Modle

Columnar

MySQL — Table (Relation

Neo4j — Graph (Network

Spark

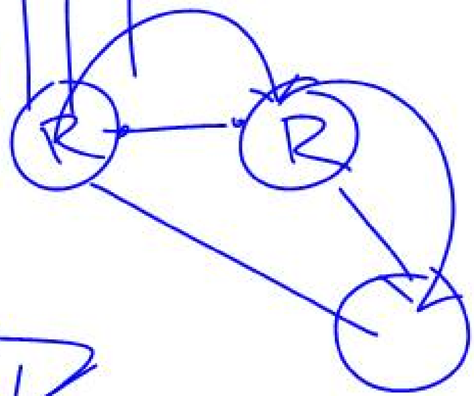
~~hdfs~~

Unsr. /schemaless {"u.p.a" [ ] }

{"tag": 3}

Redis — Key Value

*Mongo — Documents



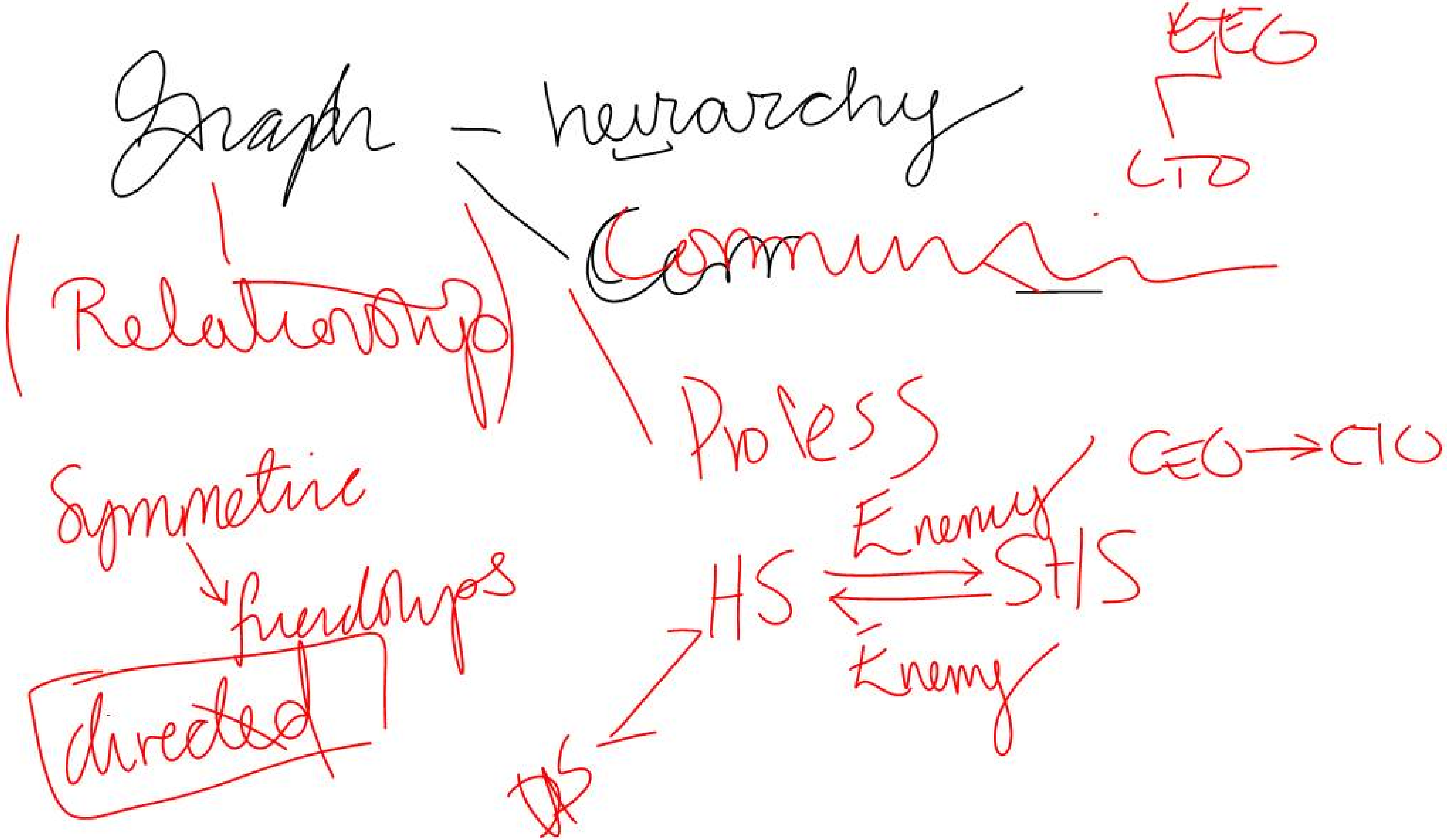
{name
loans: {1: { }
2: { }}
}

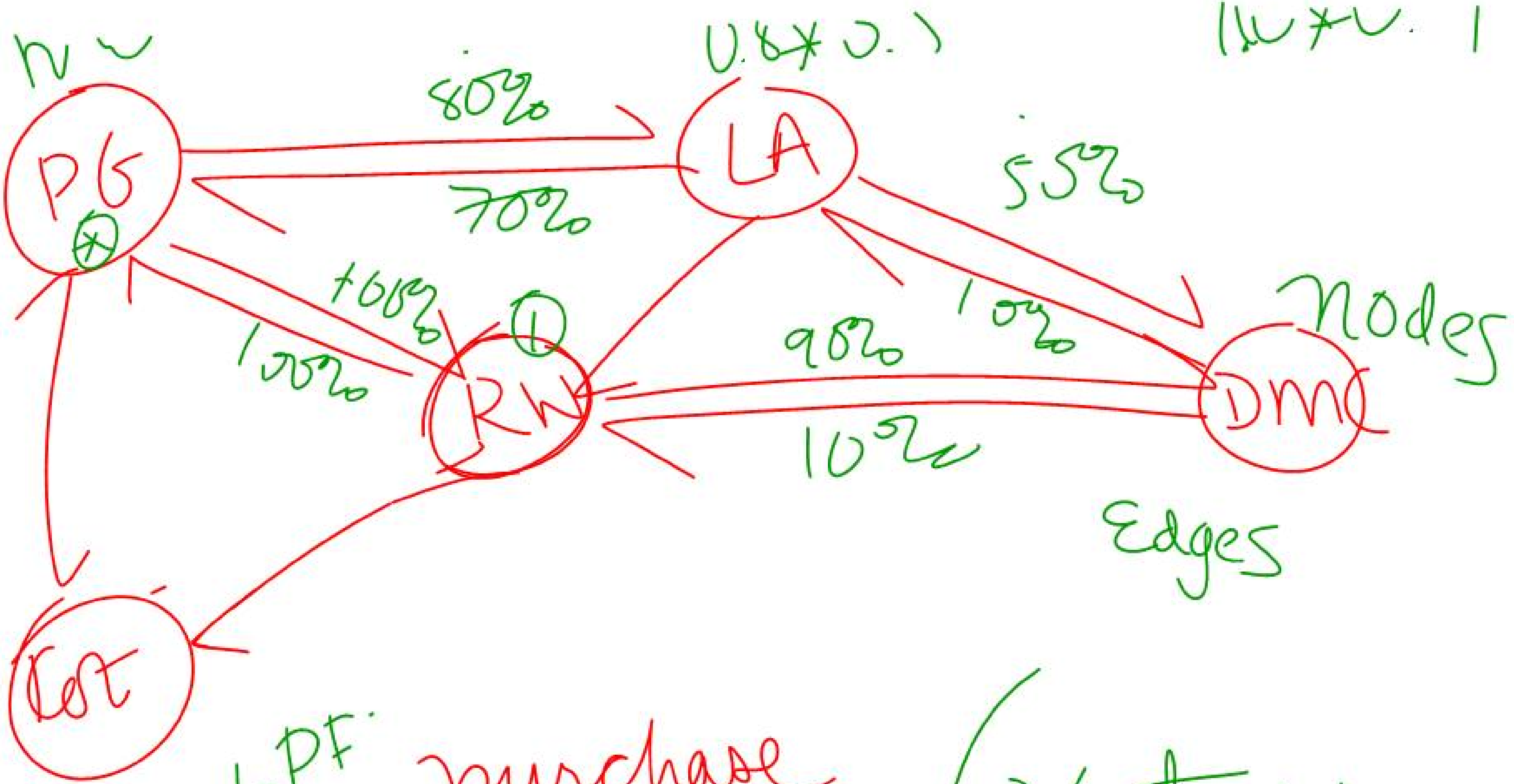
1pm

Graphs & Graph Analytics

[(u, v, w]

"M" "A" 10


Graph — heirarchy
CEO
CTO
Communi
Relationships
Process
CEO → CTO
Symmetric
friendships
directed
Enemy
HS
SHS
Enemy
DS



Graph + PF: purchase
BF
Traversal: weighted

(Vertex
≡
Node)

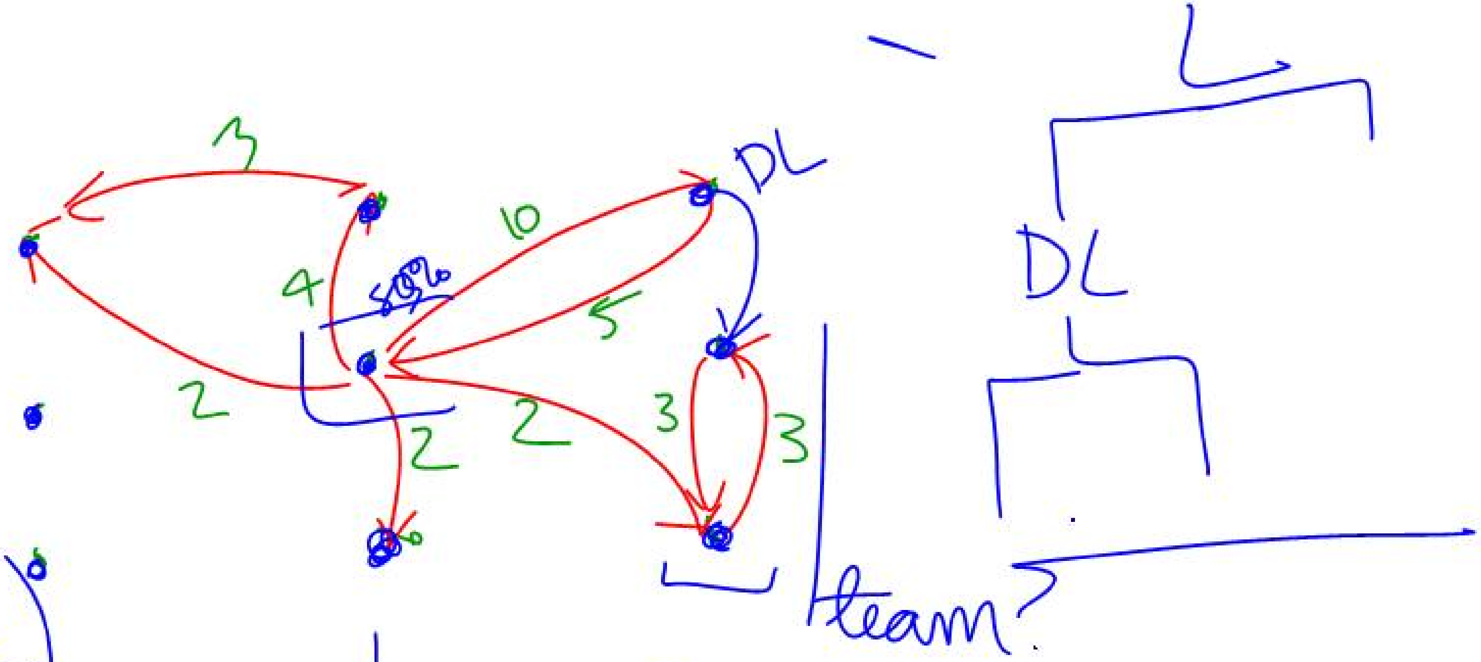
Queen Bee

↓ neigh.

"Centrality"

(importance)

| + neighbors weight



middle

out

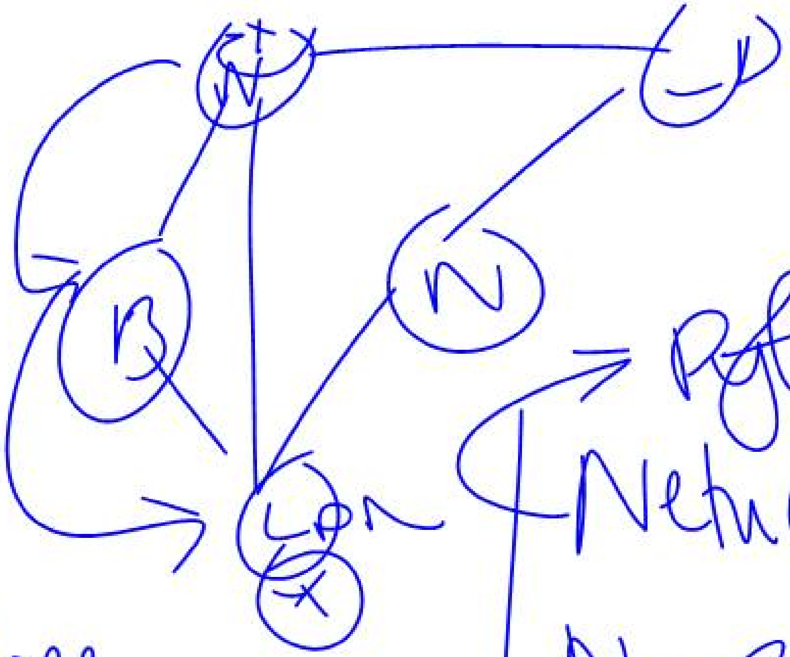
terrorist friendship group

Travel, Map

Subgraph

- Min. Span Tree

: Shortest path "

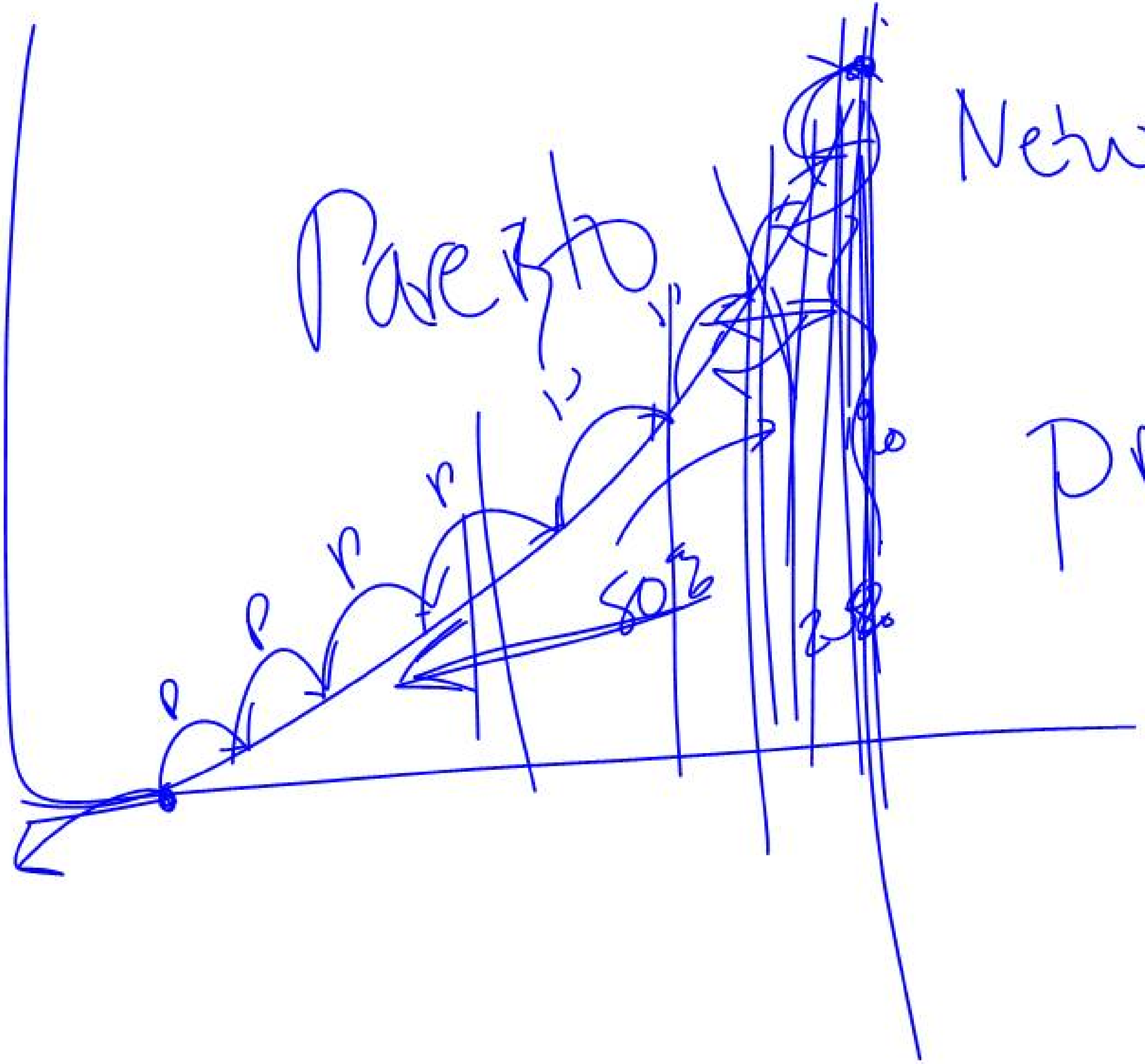


Python

Networkx

Neo4j

DB



Pareto

Networks

Pref. Attach.

T:0/2/3

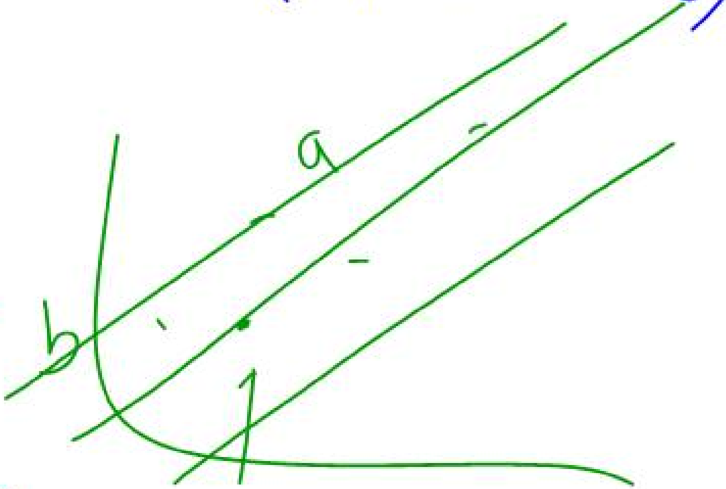
5

$$\overset{\vee}{\longleftarrow} \qquad \overset{\smile}{\longrightarrow}$$

$$\hat{y}(x; a, b) = ax + b \qquad (x, y)$$

$$l(\hat{y}, y) = (\hat{y} - y)^2 = l(a, b; x, y)$$

$$\min_{a,b} l(a, b; x, y)$$



$a := \text{Rand}()$

$b := \text{Rand}()$

~~[illegible]~~

$$a_{NEW} := a_{OLD} - \overset{\text{Hiding}}{\overline{\text{change}(l, a)_{OLD \to NEW}}}$$

i = [30 0 .. 10 . . 20]

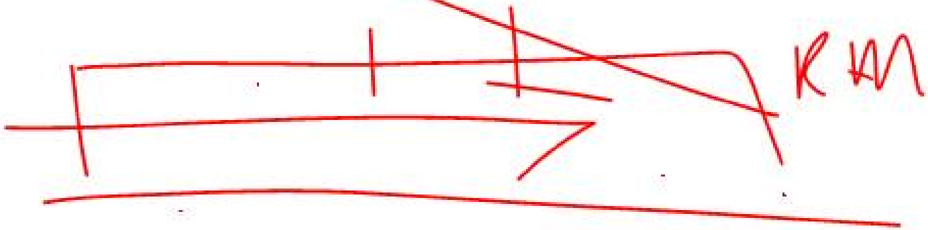
0 1 2 3 4

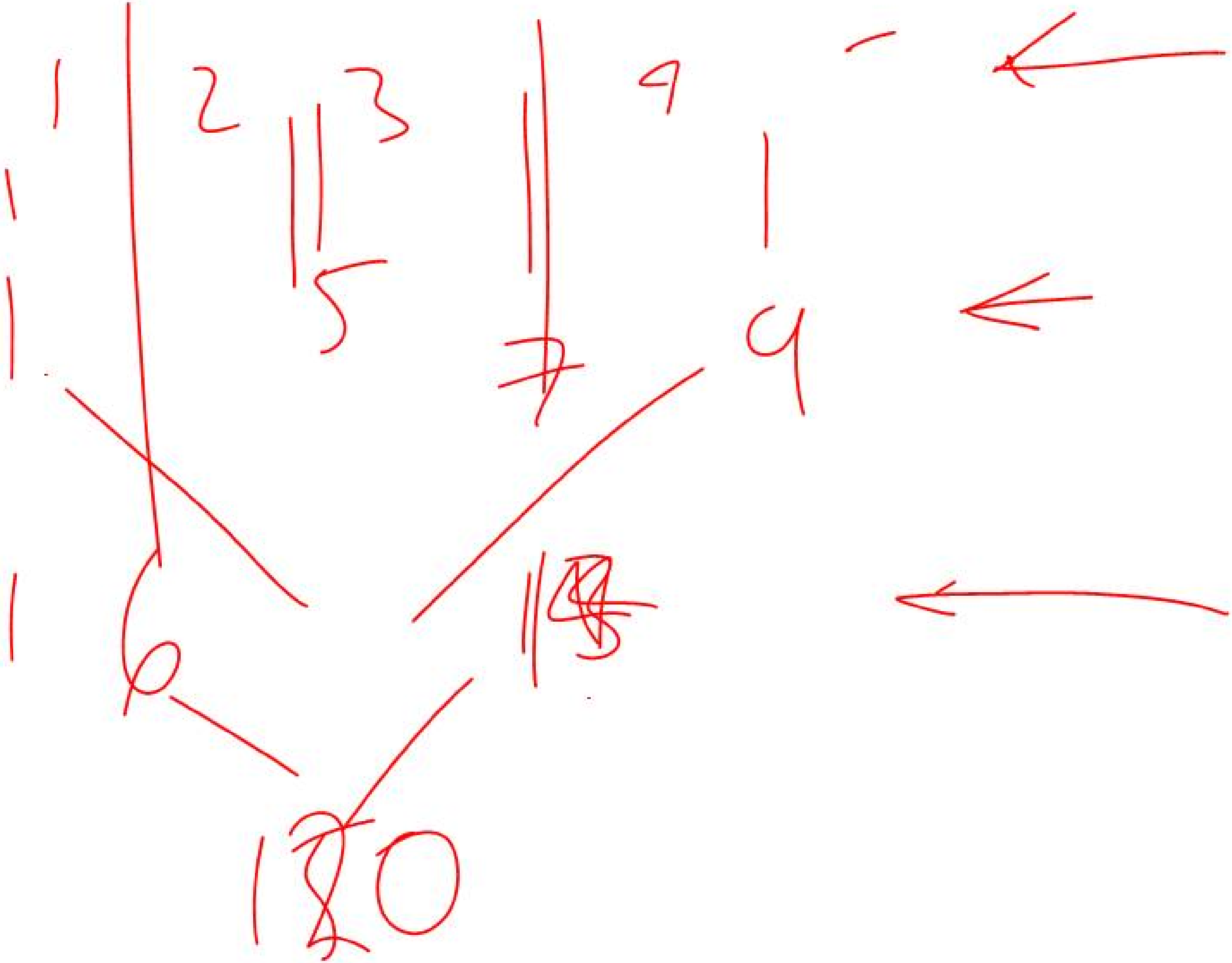
Sum(images)

API

```
for (i=0; i<len; i++) {
    total += images[i];
```

[x+1 for x in I]

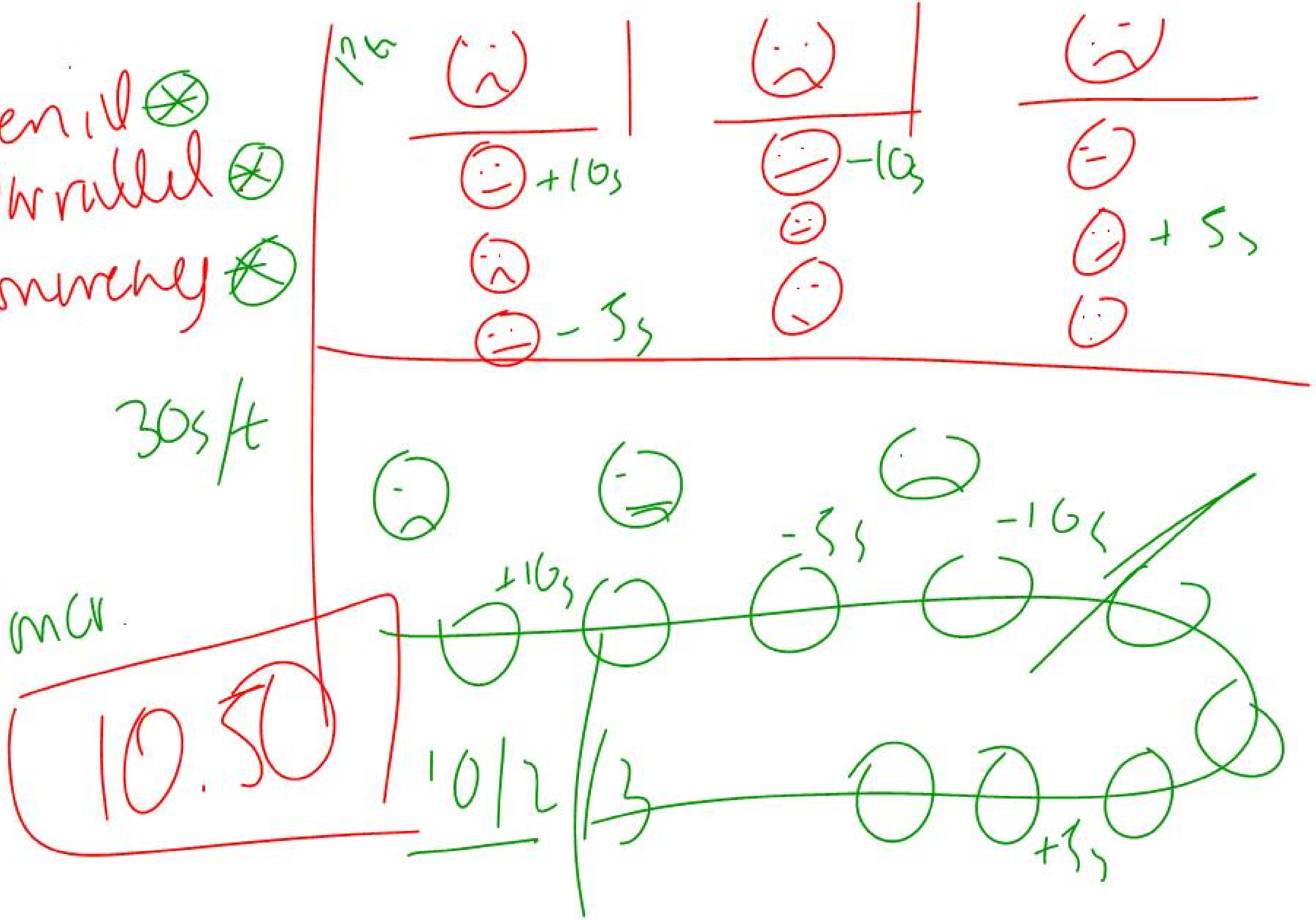



1
2
3
4
5
7
9
1
6
15
120

Serial

Parallel

Concurrency



⊗ Streaming (RAM)

HDD

1n

— constant memory

$[\quad\quad\quad\quad]_{1000}$

→ Streaming API

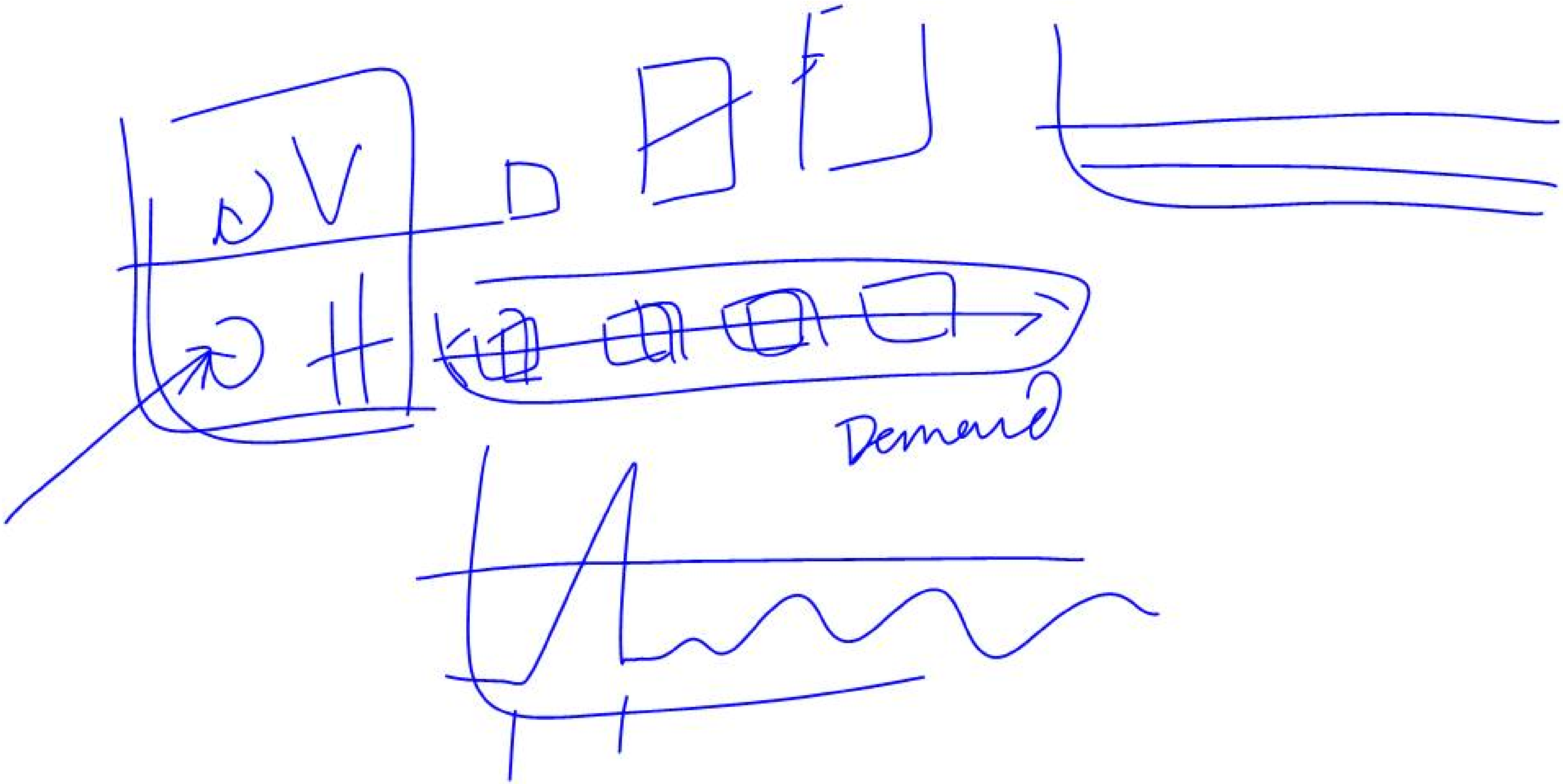
⊗ Parallelization

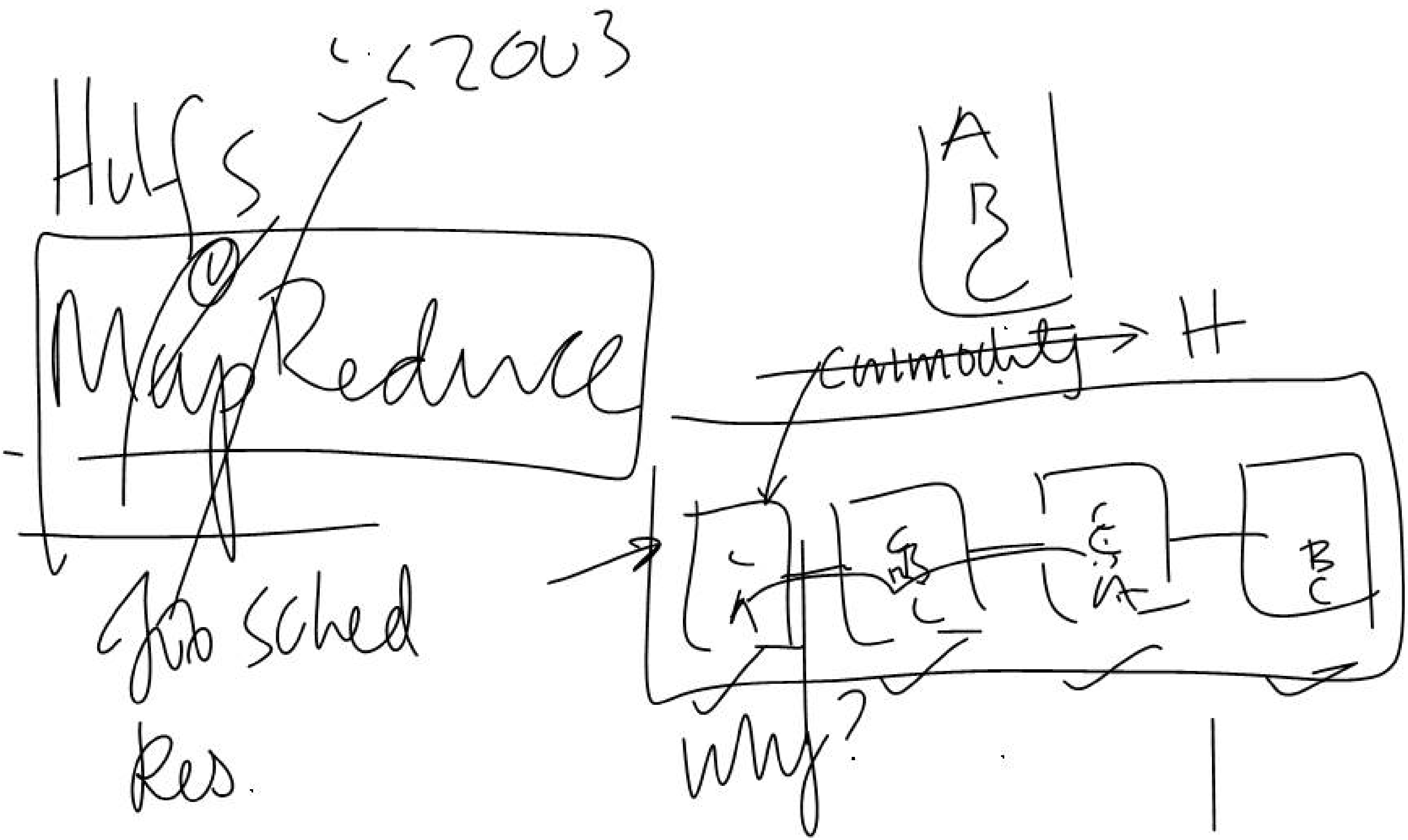
cpu

for

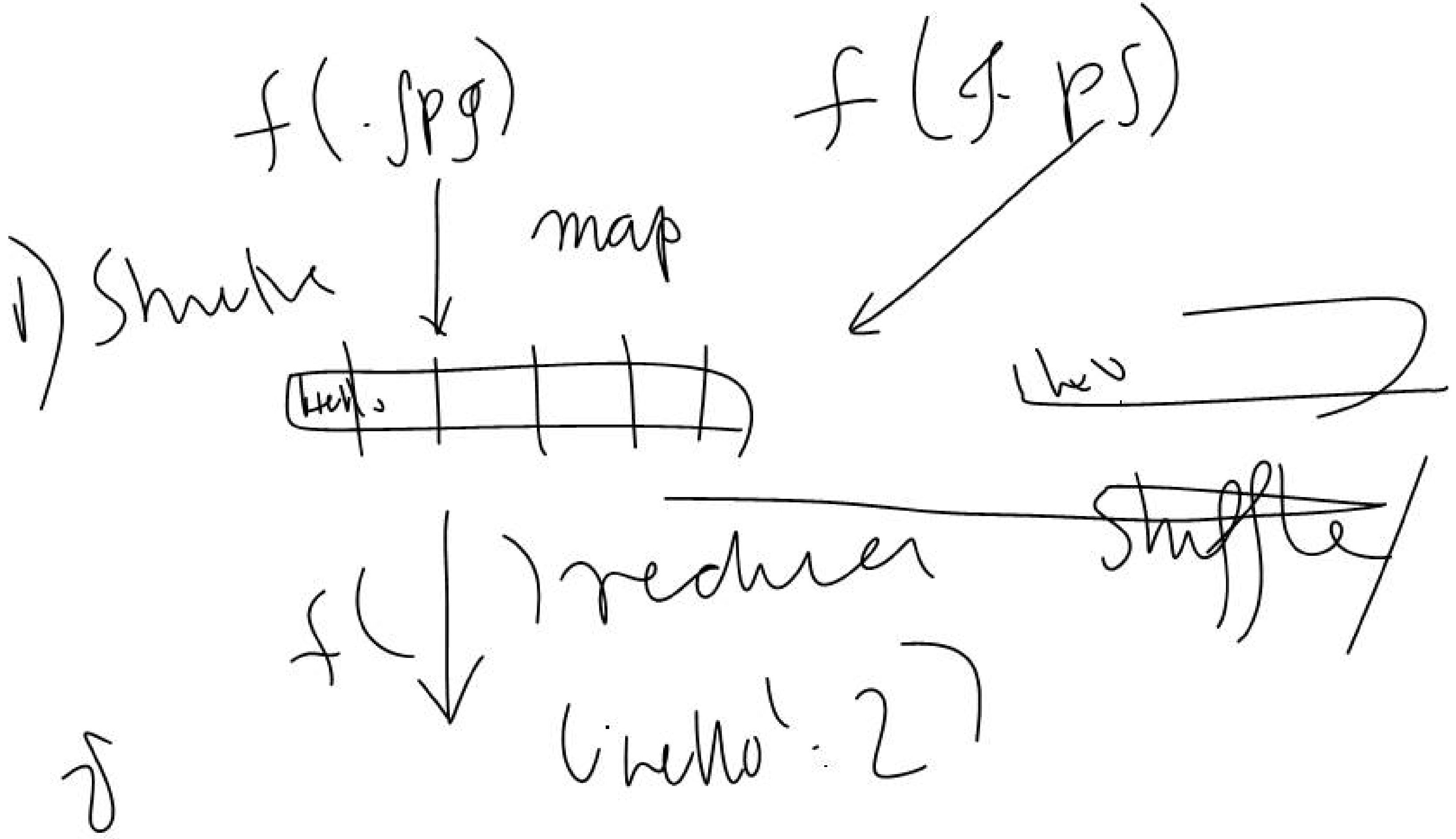
ds

ds[i]


Demand



<2003
HDFS
MapReduce
Job sched
Res.
A
B
C
commodity
H
C
A
B
C
C
A
B
C
why?


f(. jpg)
f(J. ps)
map
1) Shuffle
hello
hello
~~shuffle~~
f( ) reducer
{'hello': 2}
2)

Spark

hdfs

mr

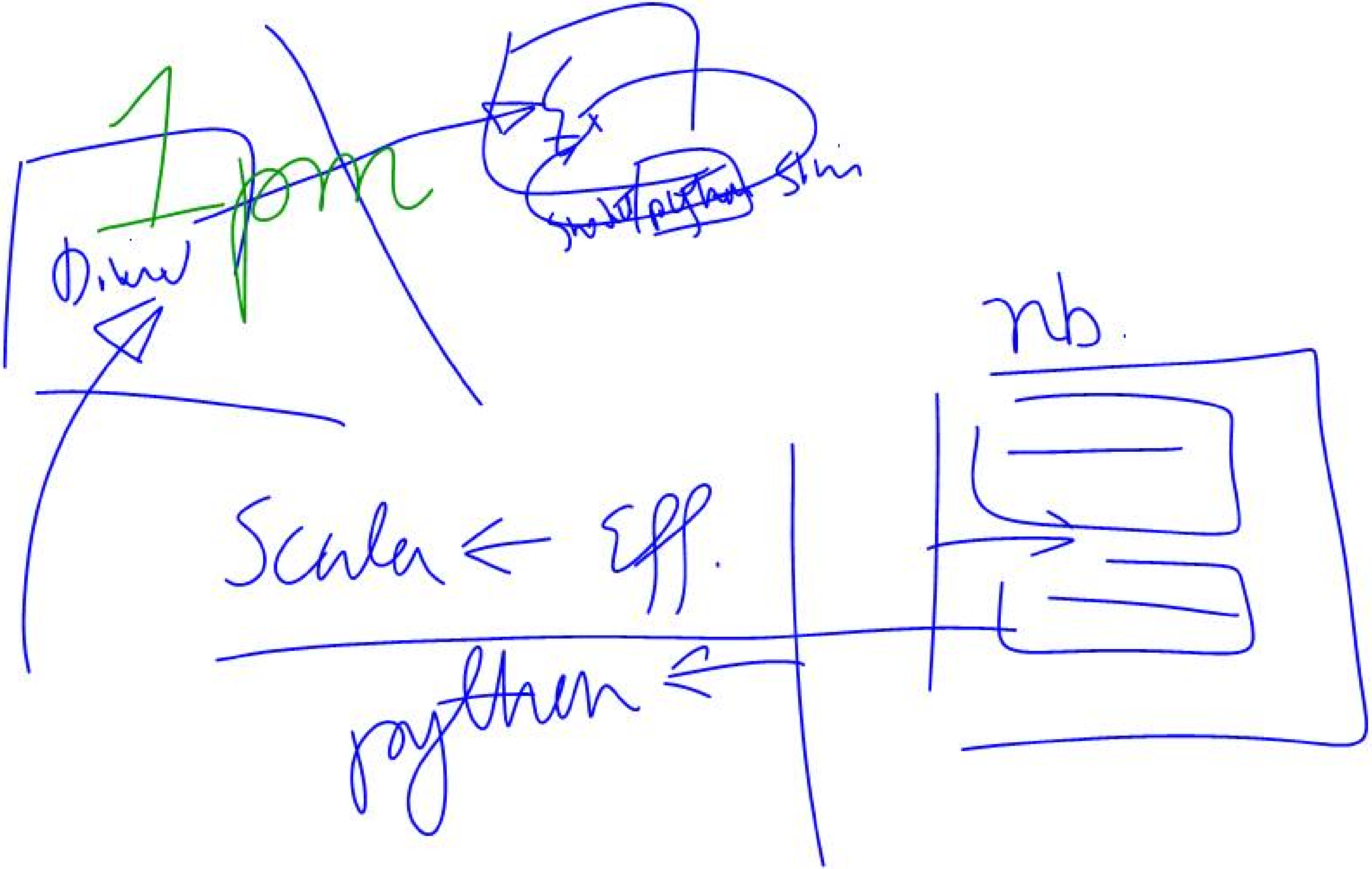
Big Data

RDD

Resilient

Distr.

Dataset


1 jvm
D.ww
stdout|python stdin
nb.
Scala ← Eff.
python ←

by " "

Kry/SD

“[ ]”

RDD

Resiliant

v1 FP + Dror.

v2 Tabular + Dataset

SQL

215

[illegible]

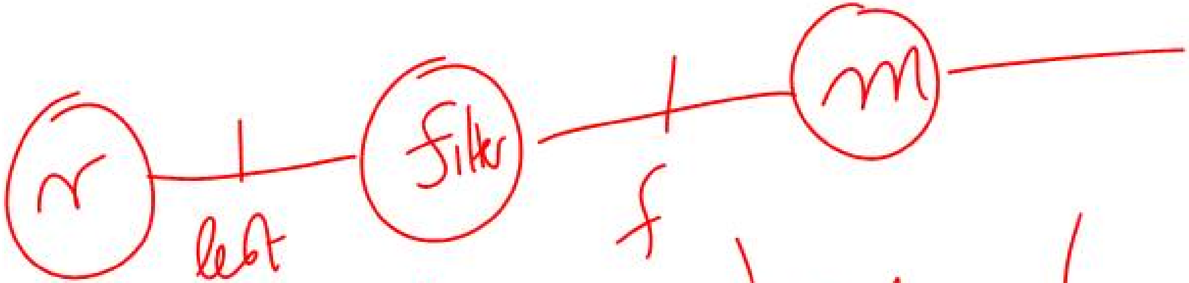
r.filter(test).map(f).collect()

transformation

a =

a =

- a =



a= (r.map(g)).filter(—)

filter (text)

flatMap (lists)

fold ← Reduce
aggregate

map ← 1) Projection / "Structure"
2) Calc

max()

collect()

top()

Questions

→ **Next Steps**

- Presentations / output ...
  - → psychology
  - → Evidence & argument

11am

15-20m Ex
5-10 Break

10.20

Projects / Pipelines

Problem / Solution

BDA

1 (- Data < Steaming / online ; Batch / DB / offline

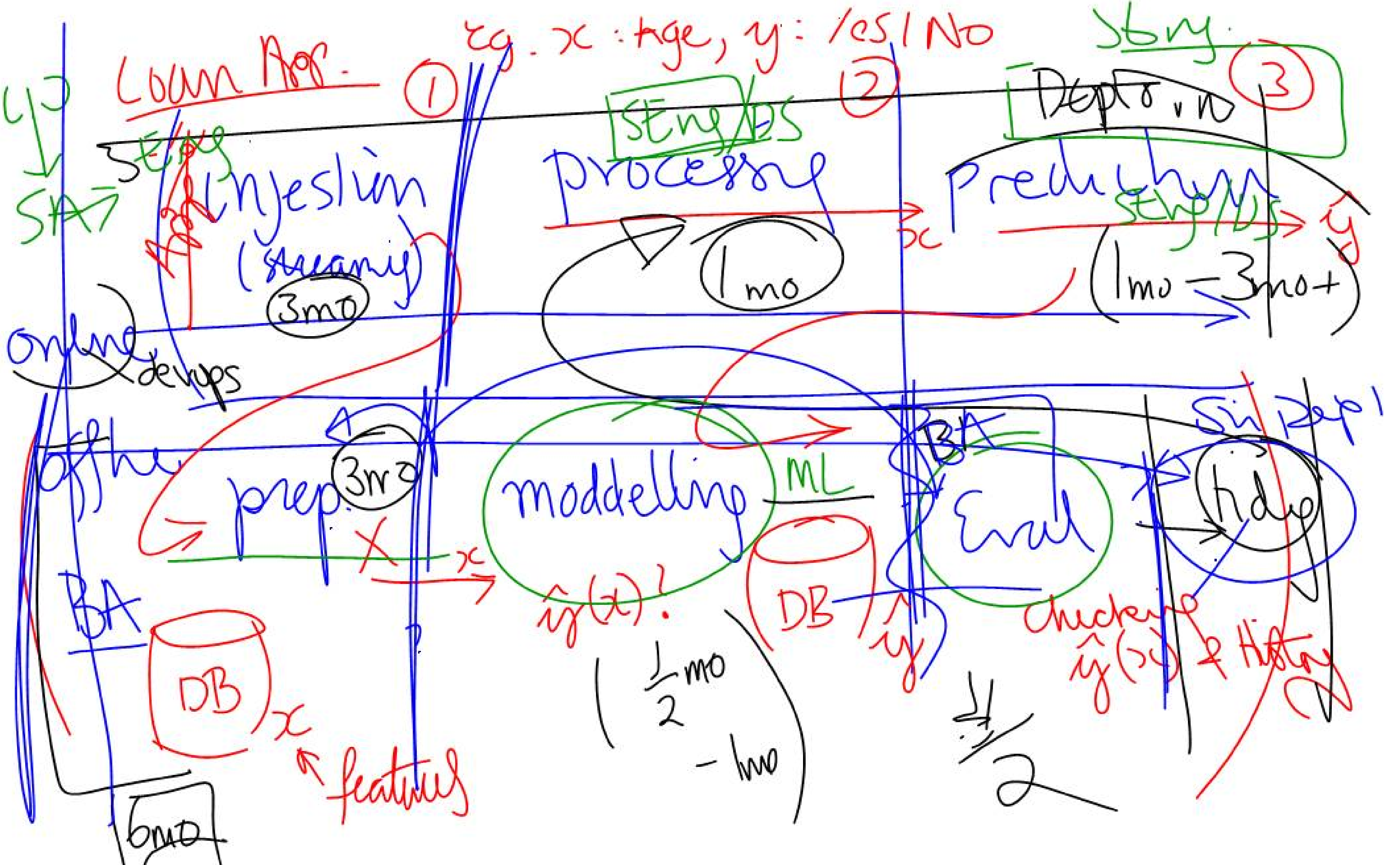
- ingestion

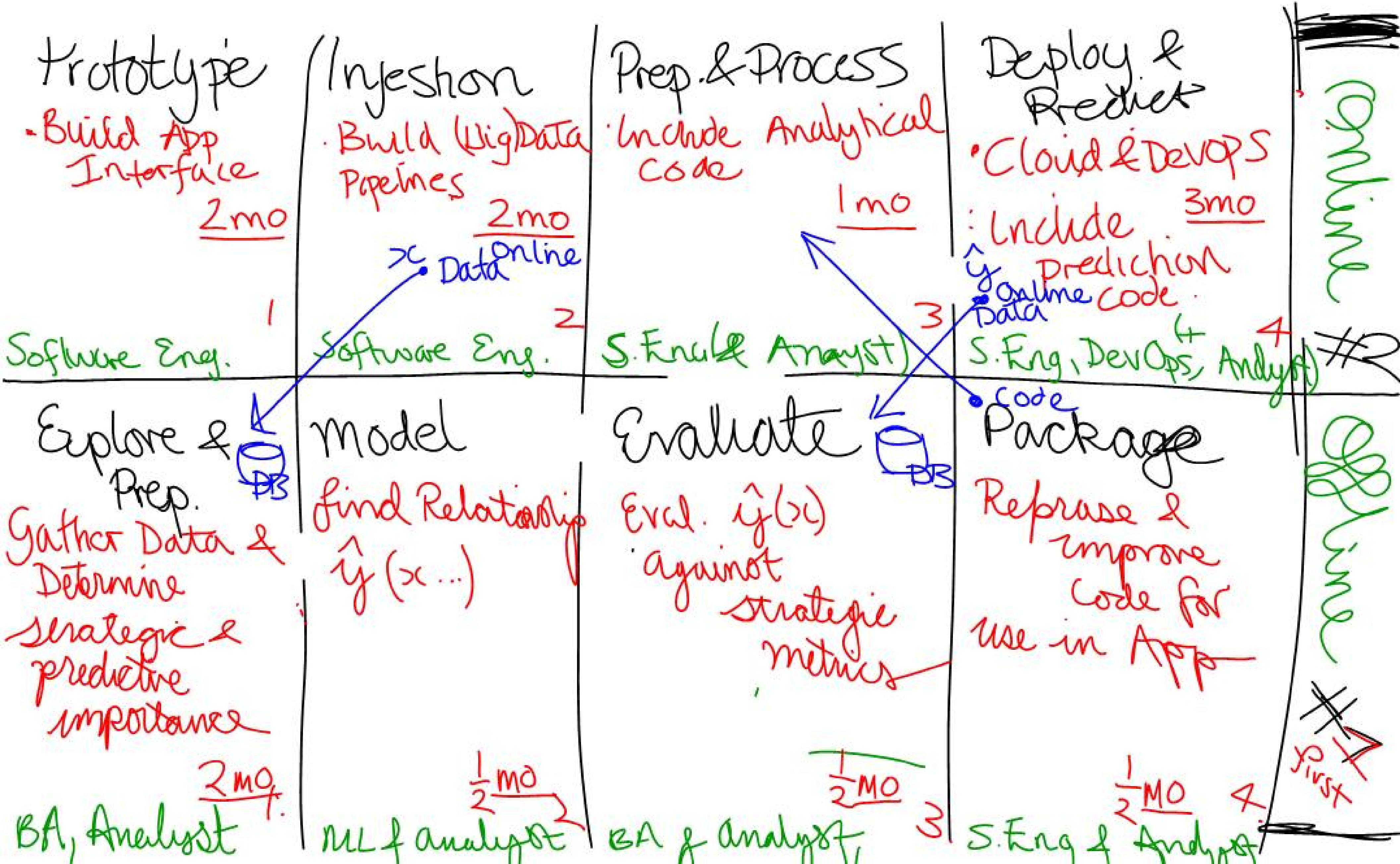
$\hat{y}(x_1, x_2 \ldots)$

Lr

Online	Offline
Processing	Ex. Dep. Analysis
Prediction	Analysis / Modelling - Modelling - Eval.

St

Loan App.
eg. x: Age, y: Yes/No
Stry
1
2
3
Dept'o.n
Steng/DS
Q/S
Steng
SA
Arch
Injestion
(streaming)
3mo
Processing
1mo
x
Prediction
Steng/DS
y
(1mo - 3mo+)
Online
devops
Offline
prep.
3mo
x
x
modelling
ML
ŷ(x)?
(1/2 mo - 1mo)
DB
y
BA
Eval
1/2
Sin Rep
hdp
checking
ŷ(x) & History
BA
DB
x
features
6mo



Prototype
- Build App Interface

2mo

1

Software Eng.

Injeshon
- Build (Big)Data Pipelines

2mo

2

Software Eng.

Prep.&Process
- Include Analytical Code

1mo

3

S.Eng(& Anaylst)

Deploy & Predict
- Cloud & DevOPS
- Include prediction code

3mo

4

S.Eng, DevOps, Andyst)

Explore & Prep.

Gather Data & Determine strategic & predictive importance

2mo

BA, Analyst

Model

find Relationship $\hat{y}(x\ldots)$

$\frac{1}{2}$mo

2

ML & analyst

Evaluate

Eval. $\hat{y}(x)$ against strategic metrics

$\frac{1}{2}$MO

3

BA & analyst

Package

Reprase & improve code for use in App

$\frac{1}{2}$MO

4

S.Eng & Andyst