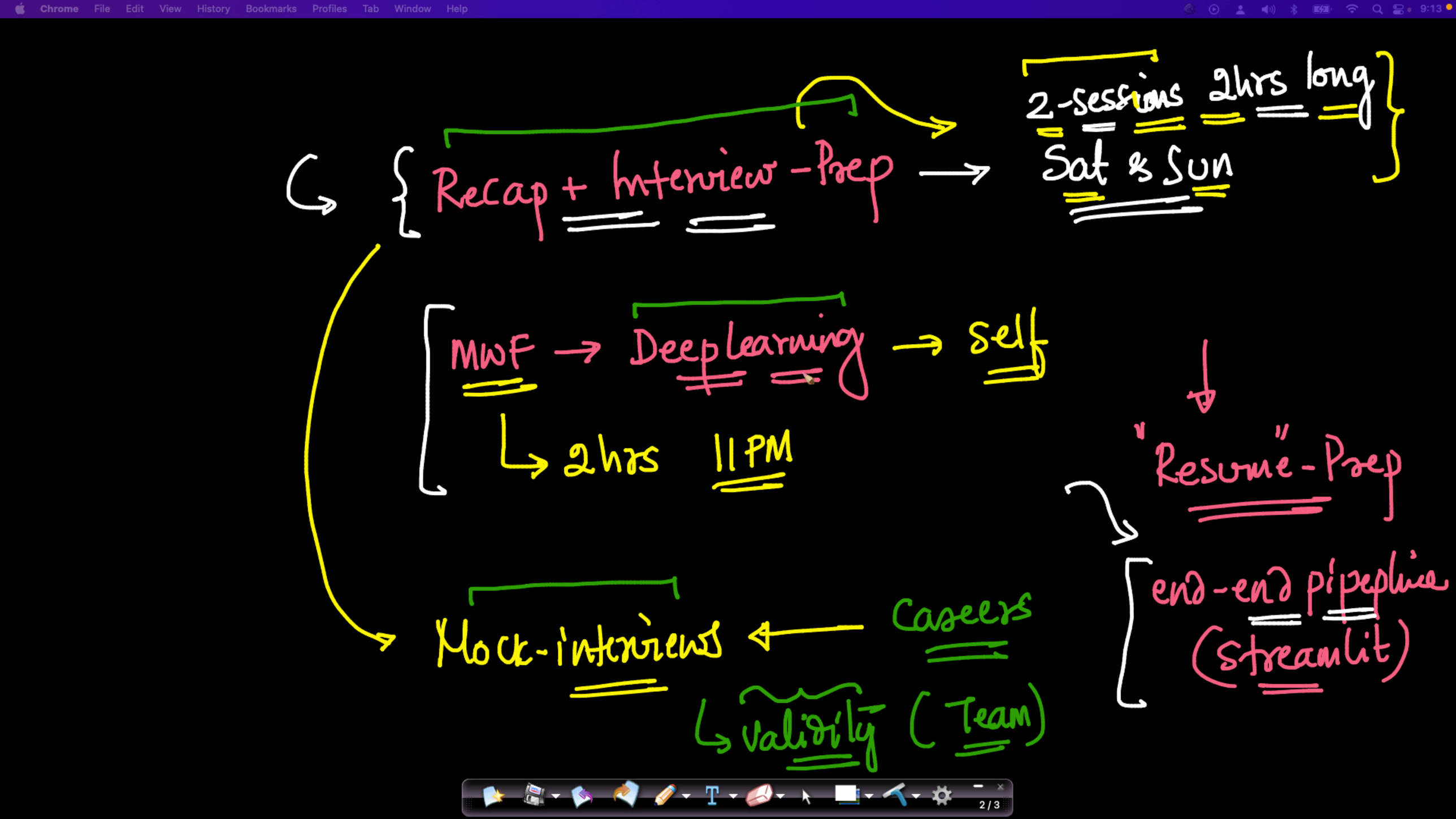
Obs: forgetten

{ ML-2

- Unsupervised Learning
- High-dim Data viz
- RecSys
- Time Series Analysis & Forecasting

Agenda: scenarios ✓; recap ✓; few interview ques (why?)


Recap + Interview-Prep
2-sessions 2hrs long
Sat & Sun
MWF → Deep learning → self
2hrs 11 PM
"Resume"-Prep
end-end pipeline (streamlit)
Mock-interviews
Careers
validity (Team)

Recap/IP

1. "How" it works internally → depth → Math → Programing: Coding
2. "Why" → harder →
3. Pros & Cons → when to use & "why"

t-SNE

K-means

↳ Stackoverflow / Quora / Wikipedia

Clustering:

- K-means, Kmeans++
  Lloyd's
- hierarchical → Linkages
- DBSCAN
- GMM

- Metrics
- DUNN Index

High-dim
Data viz

- UMAP, tSNE
- PCA,

RecSys:

- Content based: U-U; I-I; U-I
- Coll. filt → MF
  → MF + μ + b_u + b_i
- Hybrid-model & cold-start
- MF → Clustering
  → PCA ; NMF
  → SVD
- Matrix Completion

TS Analysis & Forecasting

- ACF, PACF; decomposition → (+, *)
- Imputation; TB-splitting; stationarity
- AR, MA, ARIMA; SARIMA, SARIMA
- FB Prophet
- SMA, EWMA; DES; TES
- C.I on TS
- Change detection

Anamoly detn /outlier

- Isolation forest
- DBSCAN; LOF
- elliptical envelope
- one-class SVM
- RANSAC
- { Isotonic-reg (Misc..)
  Platt scaling; elastic Net

Q1

n = 10M products

ecommerce

Cluster

(open-ended)

d = 40 (let)

: item-vec

Hierarchical → Comp. exp → 12 hrs

↳ natural

→ K Means →

Shirts

TSh

FSs

Hs

FS

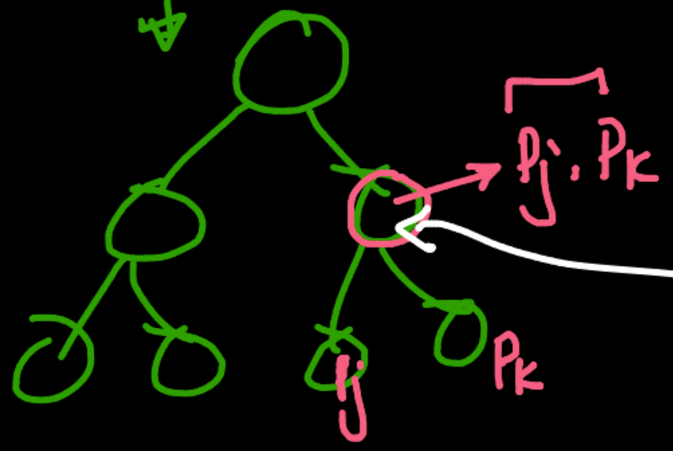
Speed-up (ex SDE)

→ dim-red (why?) → may not Speed up as much

→ Sampling

n' products → 500K

n' << n

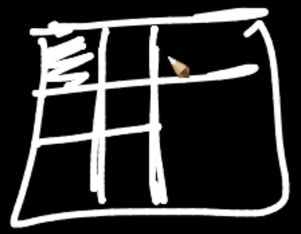


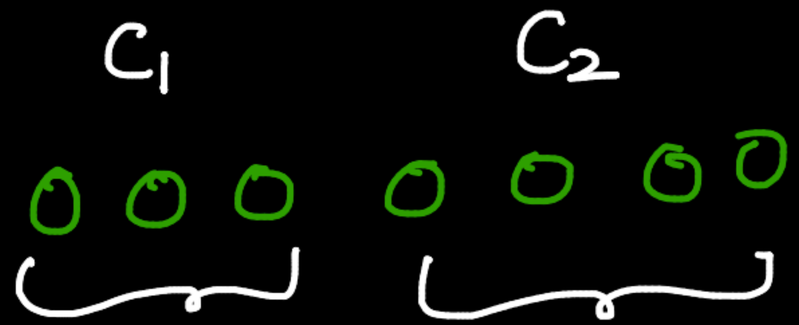
n-n' products

Pi → Pj

Fast Approx NN

O(1)




C1
C2

Q2

D

n-pls

{ Fast K-NN

clustering (?)

FAISS, SCANN

LSH (random plane)

Pre-cluster

1M

100K

900K

log n

x_1 x_{10} x_n

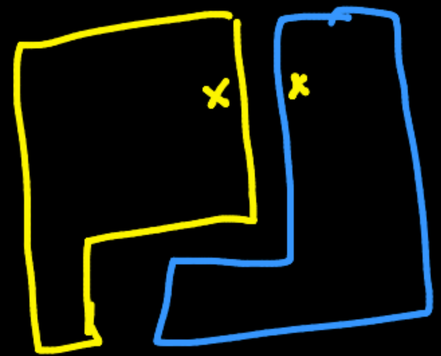
x_{q1} ✓ → 3-NN

✓

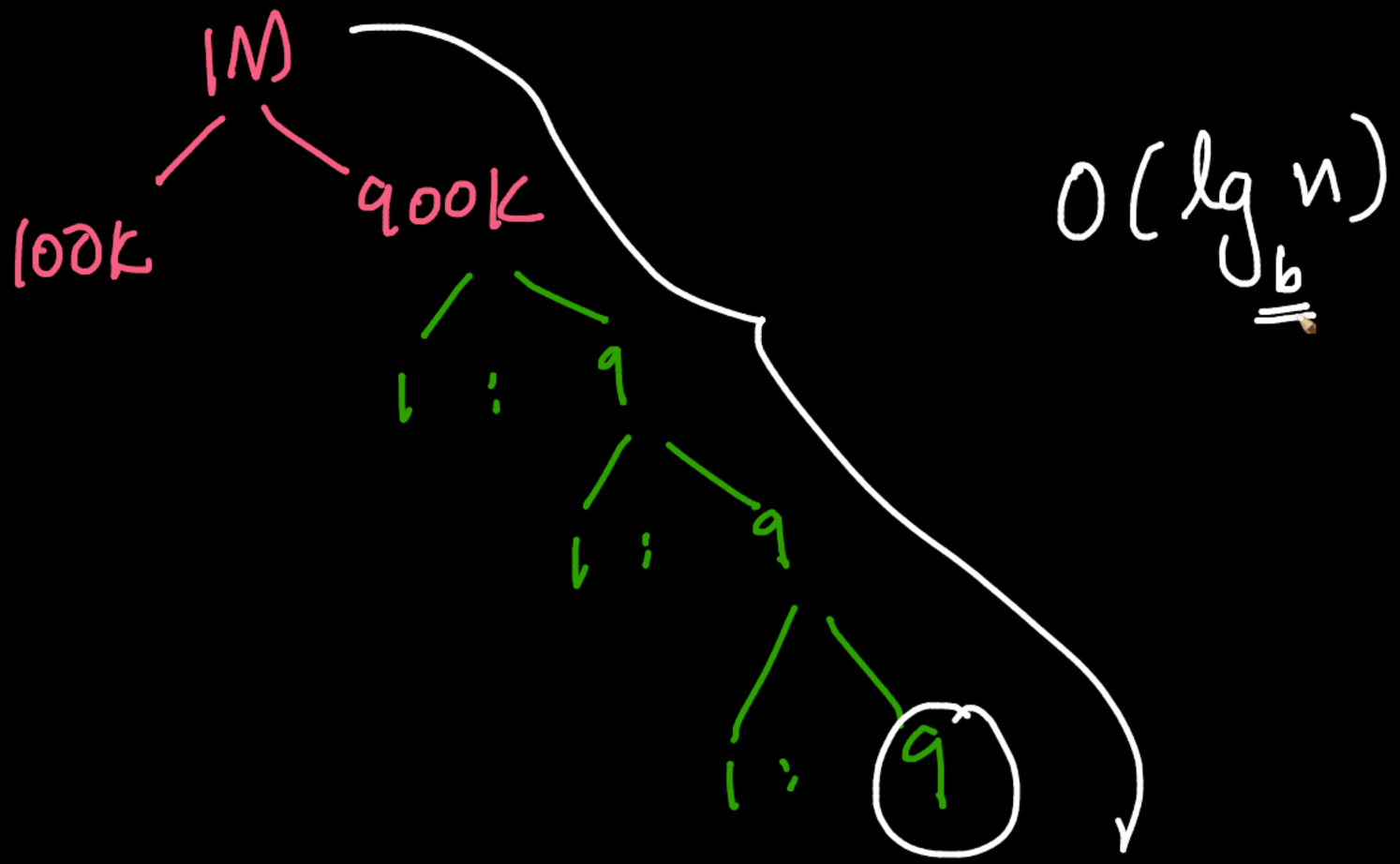
× [DBSCAN → R^* tree]

↪ ε, minpls

↳ density-based



log (n)



$O(\log_b n)$

Q3

look resumes → BoW → dim : large

↳ cluster → hierarchical (cosine)

(+) Topic-modeling (LDA)

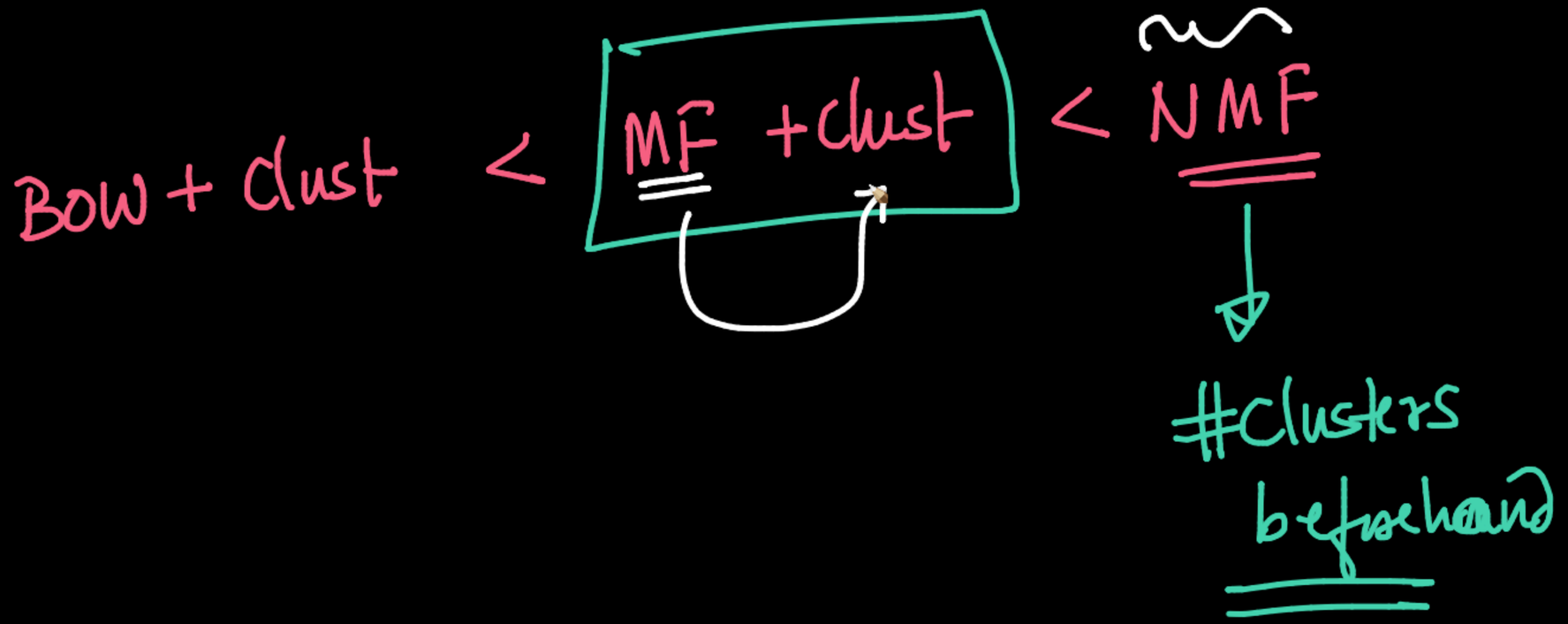
↳ MF: word-doc (matrix)

Very rich rep for each doc

+ hierarchical

XMF → soft clustering

≈



Q4

$n = 1MM$

$d = 2$

lat & long

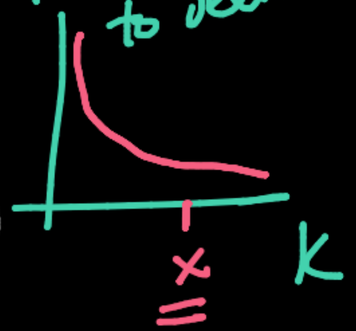
Src, dest

Deliveries (Zepto)

dark Stores

Cluster

median time to deliver



DBSCAN (ε, minpts)

↳ density difference

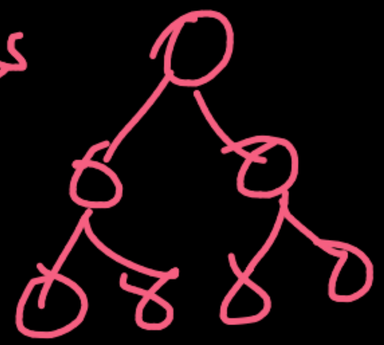
+

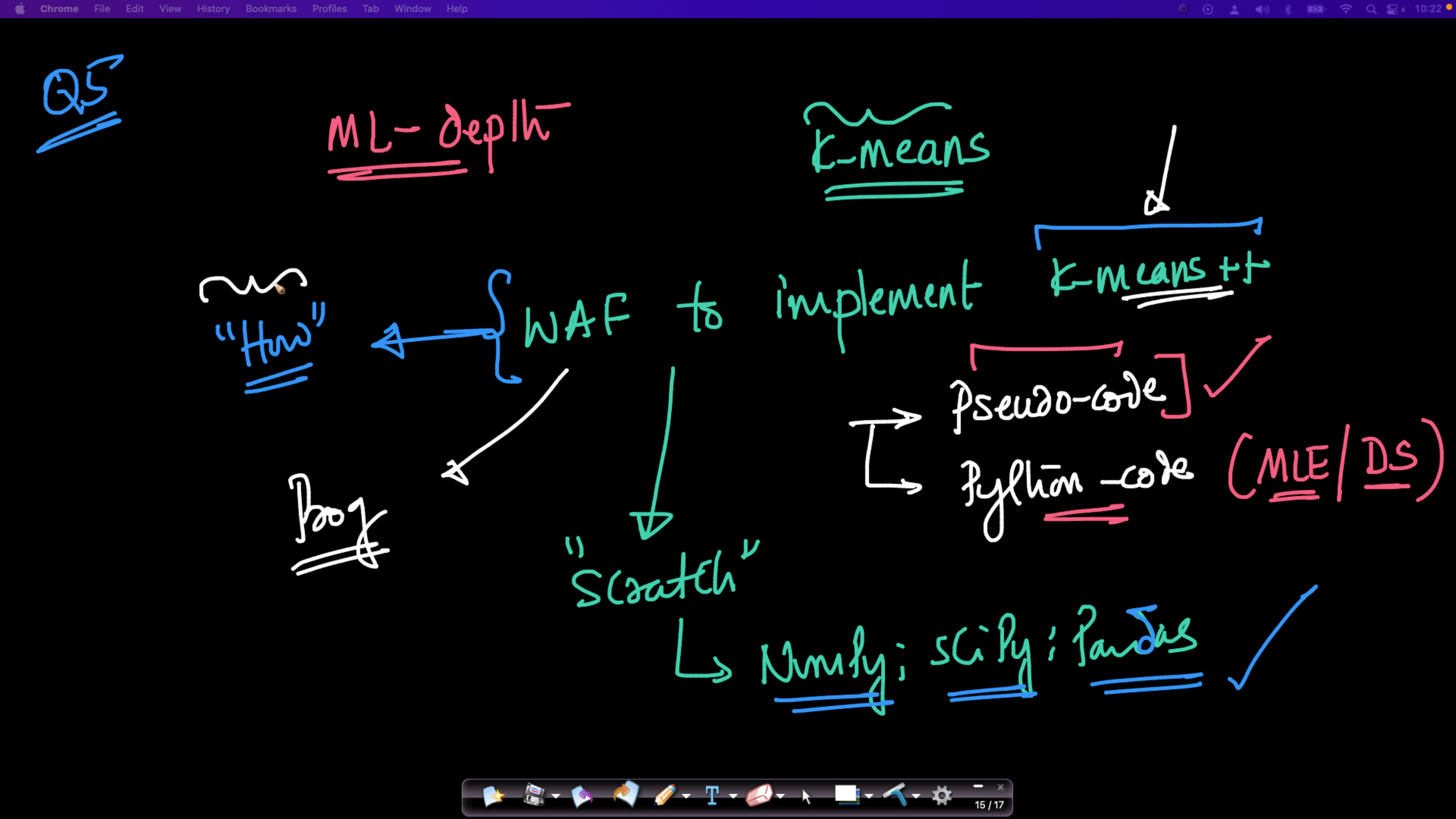
K-means

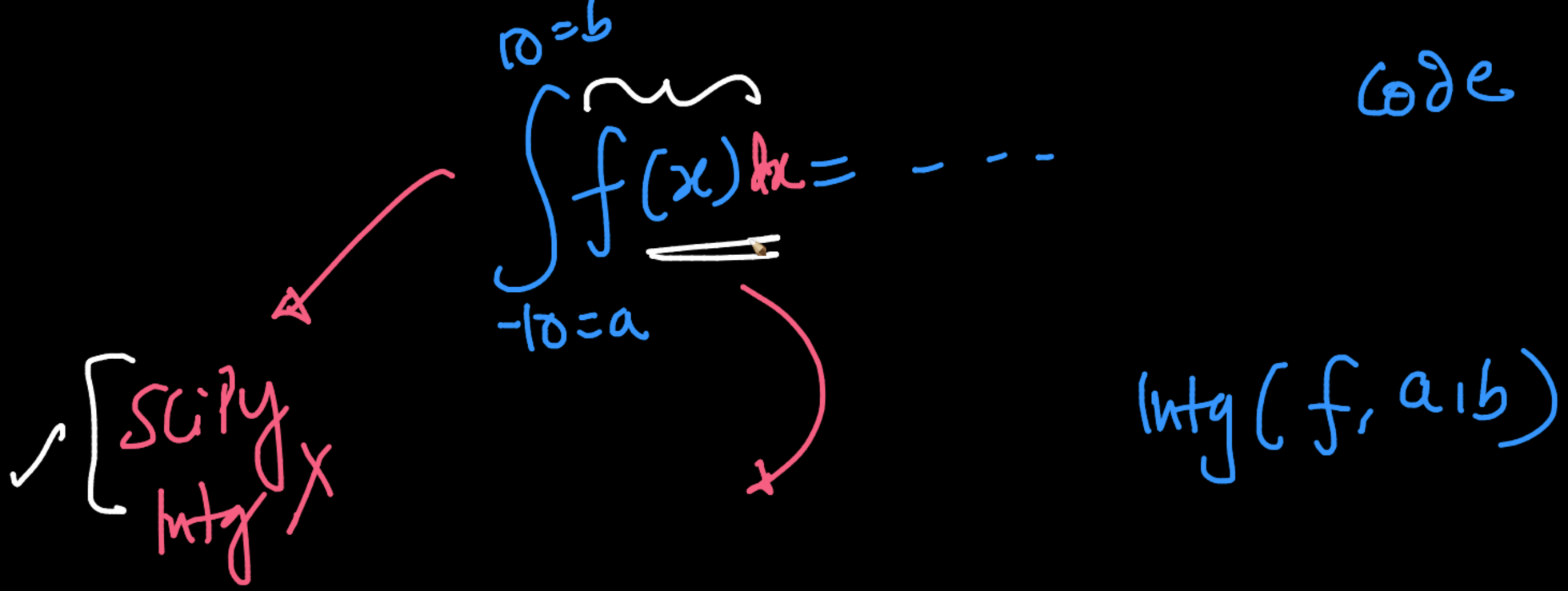
≈ equal sized Clusters

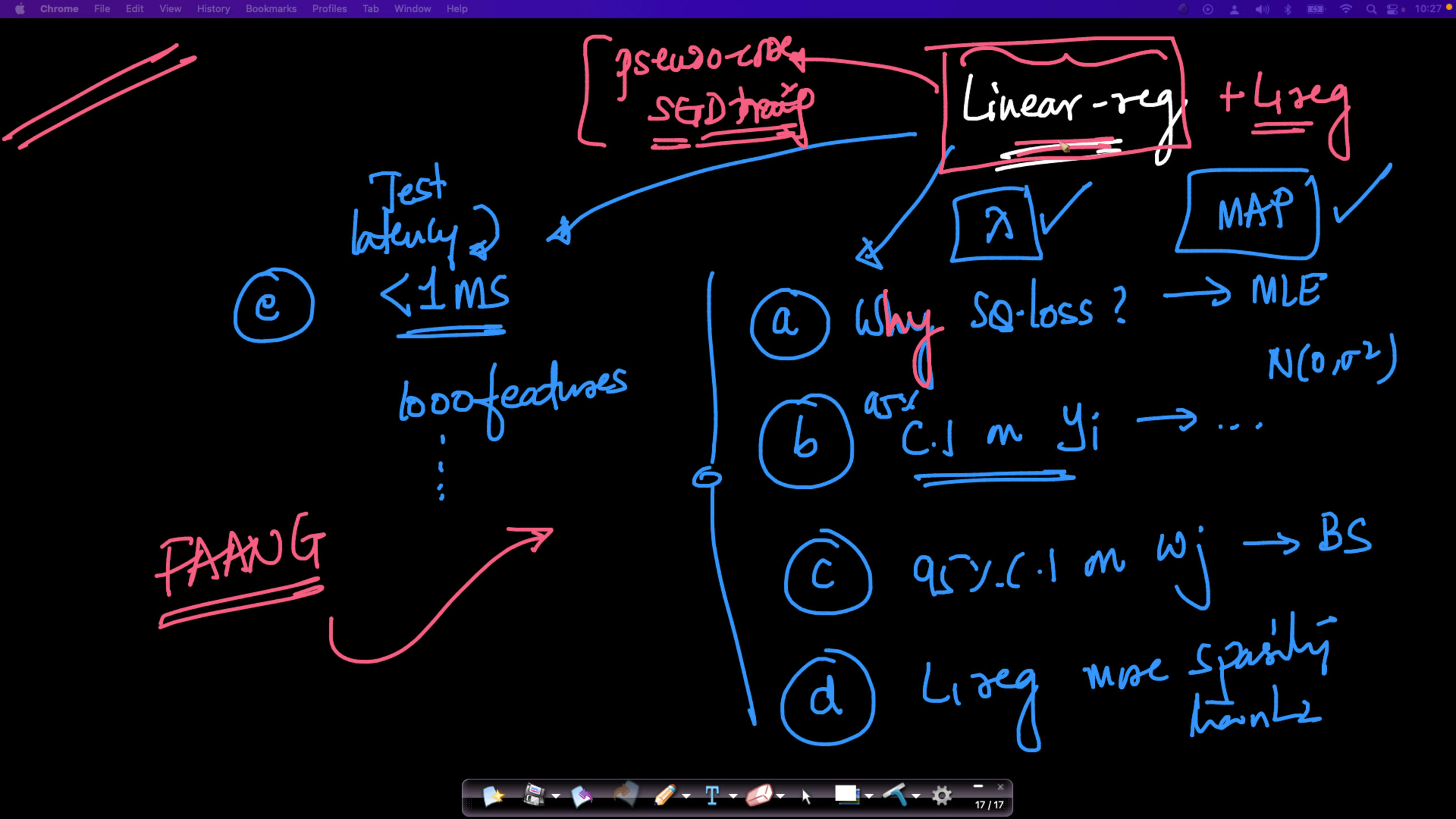
adv & disadv of each technique + "How"

Scenarios → Constraints = Properties




Chrome File Edit View History Bookmarks Profiles Tab Window Help
10:22
Q5
ML- Depth
K-means
K-means++
"How"
WAF to implement
Boost
"Scratch"
Pseudo-code
Python code
(MLE/DS)
Numpy; SciPy; Pandas
15 / 17


10=b
$\int f(x)\,dx =$ - - -
-10=a
[scipy
Integrate
Code
Integrate(f, a, b)


pseudo-code
SGD from scratch
Linear-reg +L2reg
Test latency
<1ms
λ
MAP
a Why SQ-loss ? → MLE
N(0,σ²)
b as c.1 m yi → ...
c as c.1 on wj → BS
d L1reg more sparsity than L2
1000features
FAANG

Q

Why tSNE not preserve "global" struc. → preserving neighbours

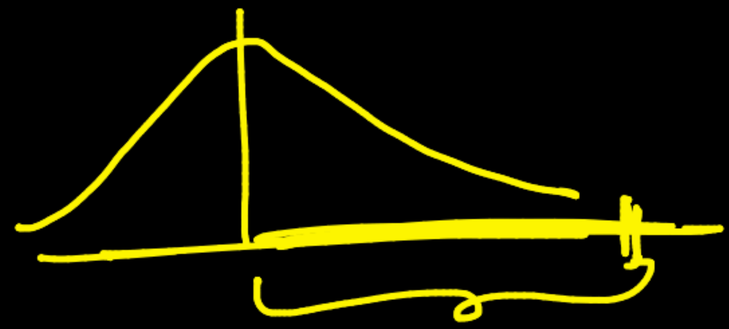
- crowding problem ✓
- Stochasticity → may not different solns for different runs
- HINT: obj. fn in optmzn-problem
  - min KL(P, Q)
    - d_ij in P_{ij} = Prob that x_i & x_j are neighbours → Gaussian
    - d'_ij in Q_{ij} = y_i & y_j ... → t-dist

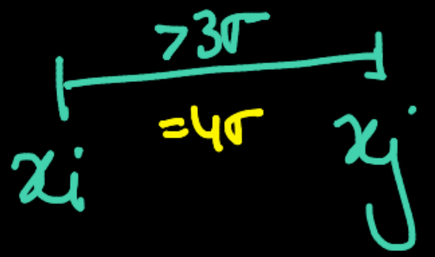
x_i's → y_i's

high low

V. small

Gaussi → P_{ij} $\approx$ Q_{ij} ← $\sim$ t-distb 1





$$\sum_{ij} P_{ij} \log \frac{P_{ij}}{Q_{ij}}$$

Q1

Amazon.in

orders per min

Forecasting

10-days

weekly

s = Monthly ...
s = 24 hrs
Trend

t

10days

~ 8-yrs

daily

$y_t \approx y_{t-1}$

exogenous - festivals

SARIMAX

Y

N

d=0

Multiple Seasonalities

FB-Prophet

Q

additionally, Quadratic features(s)

y_{t-4} y_{t-5}

y_{t-3} y_{t-4}

y_{t-2} y_{t-3}

t-1
t
t+1

easier

Statsmodels

SARIMAX → lib

p d q

ϵ_t

exogenous

$$y_{t+1} = \sum_{i=0}^{d} \alpha_i y_{t-d}$$

Introduce non-lr features into existing implm of SARIMAX

Q

Pseudo code on PACF ✓

ACF ✓

what

"Hw" ✓

y_t y_{t-1} y_{t-2}

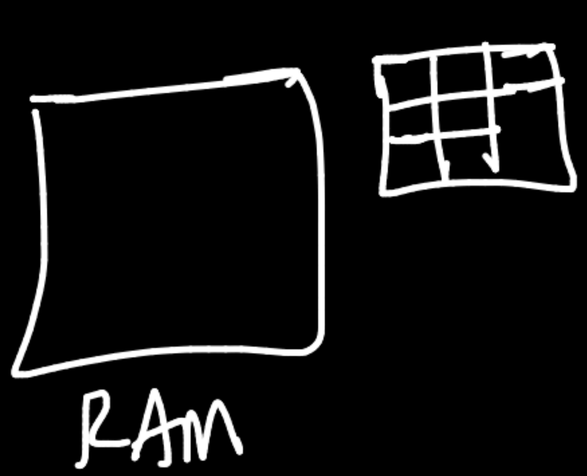
Cor · Coeff (y_{t-2}, δ_t)

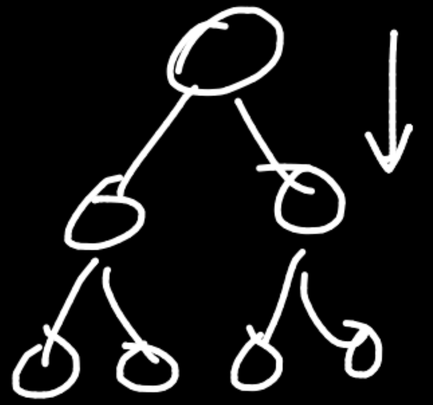
$$\delta_t = y_t - f_{AR}(y_{t-1})$$

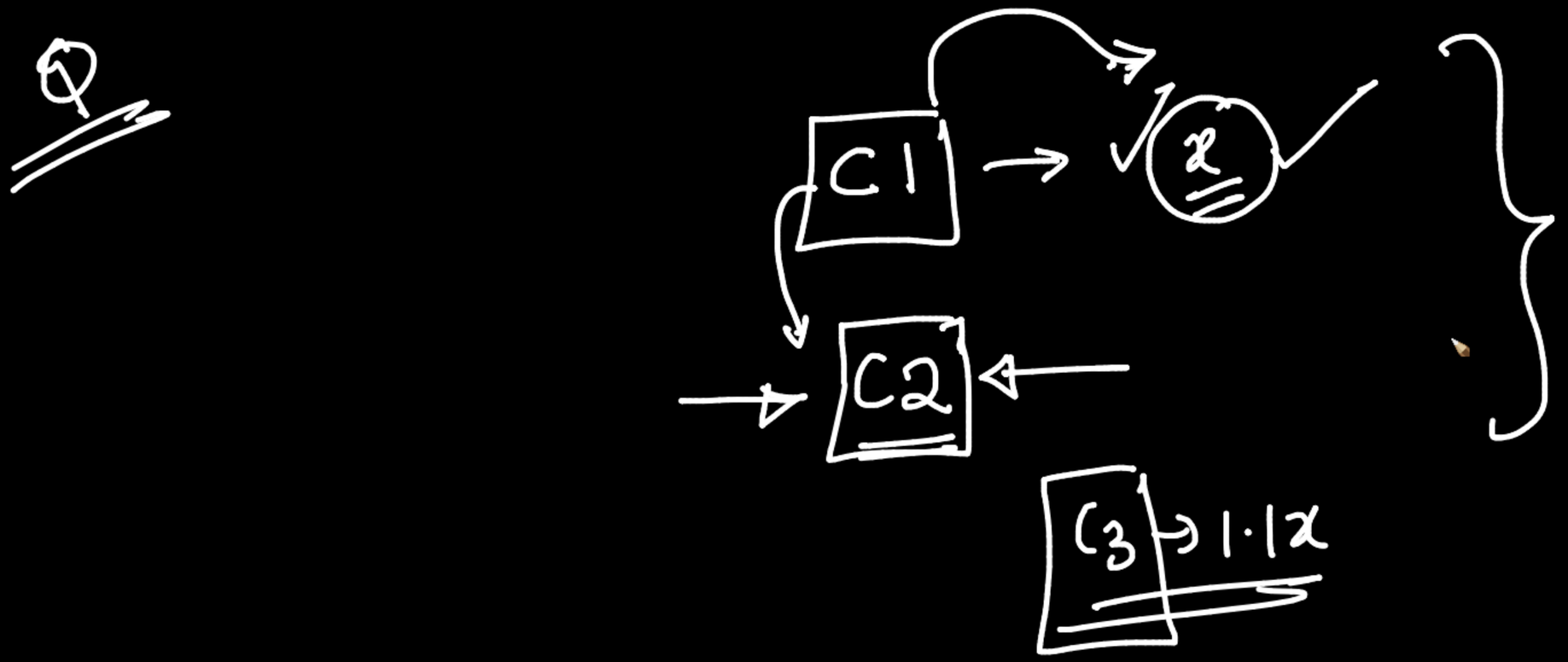
$$y_t \leftarrow f_{AR}(y_{t-1})$$

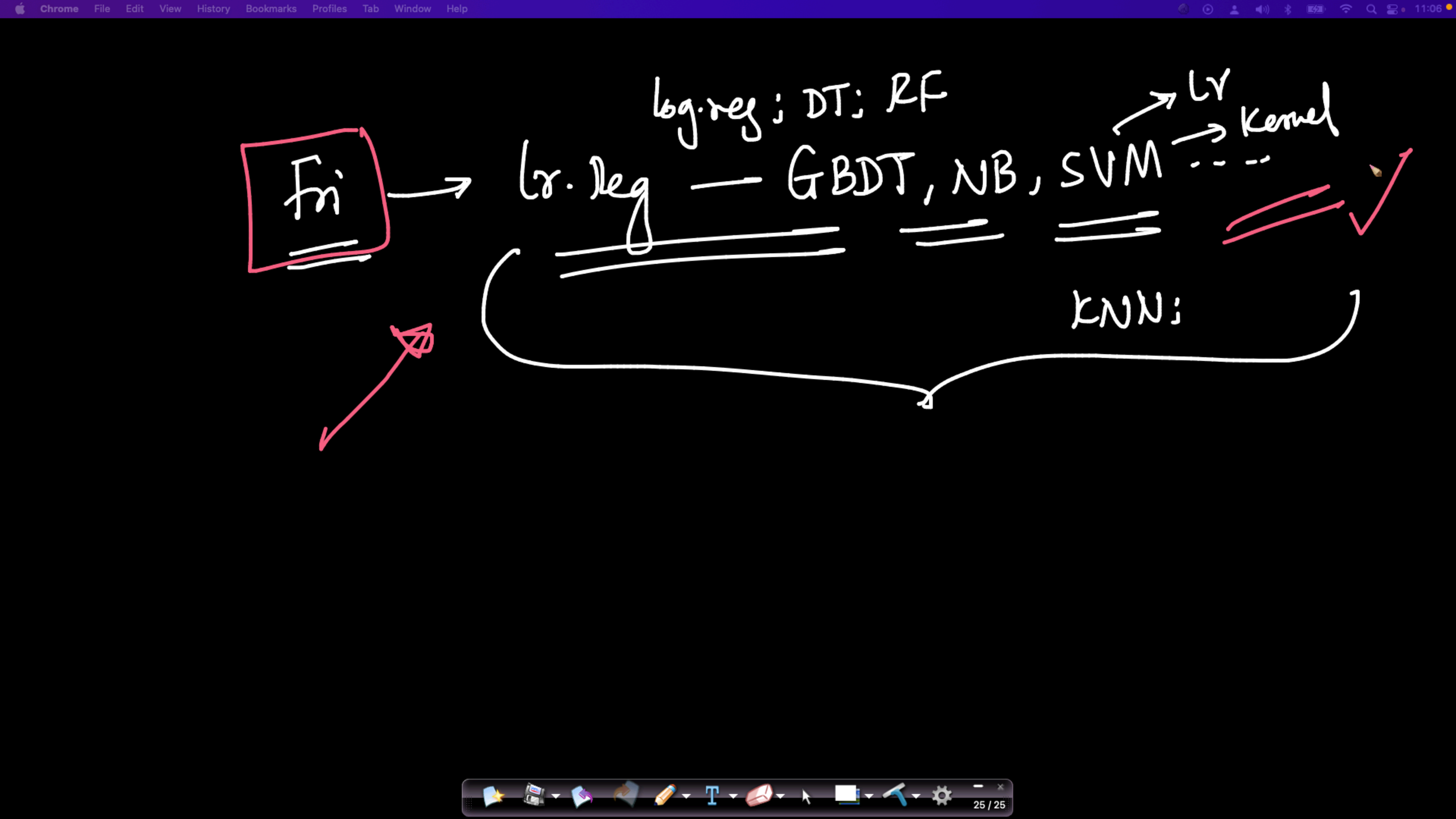
Q

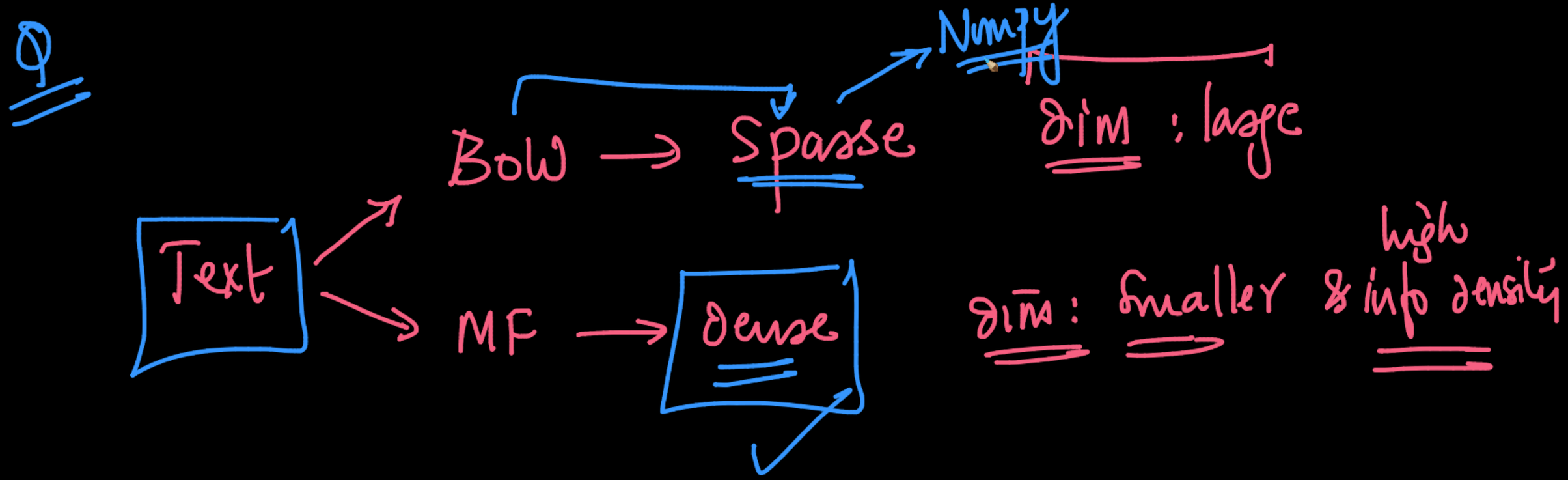
disb impltn of hier. clustering → divisive

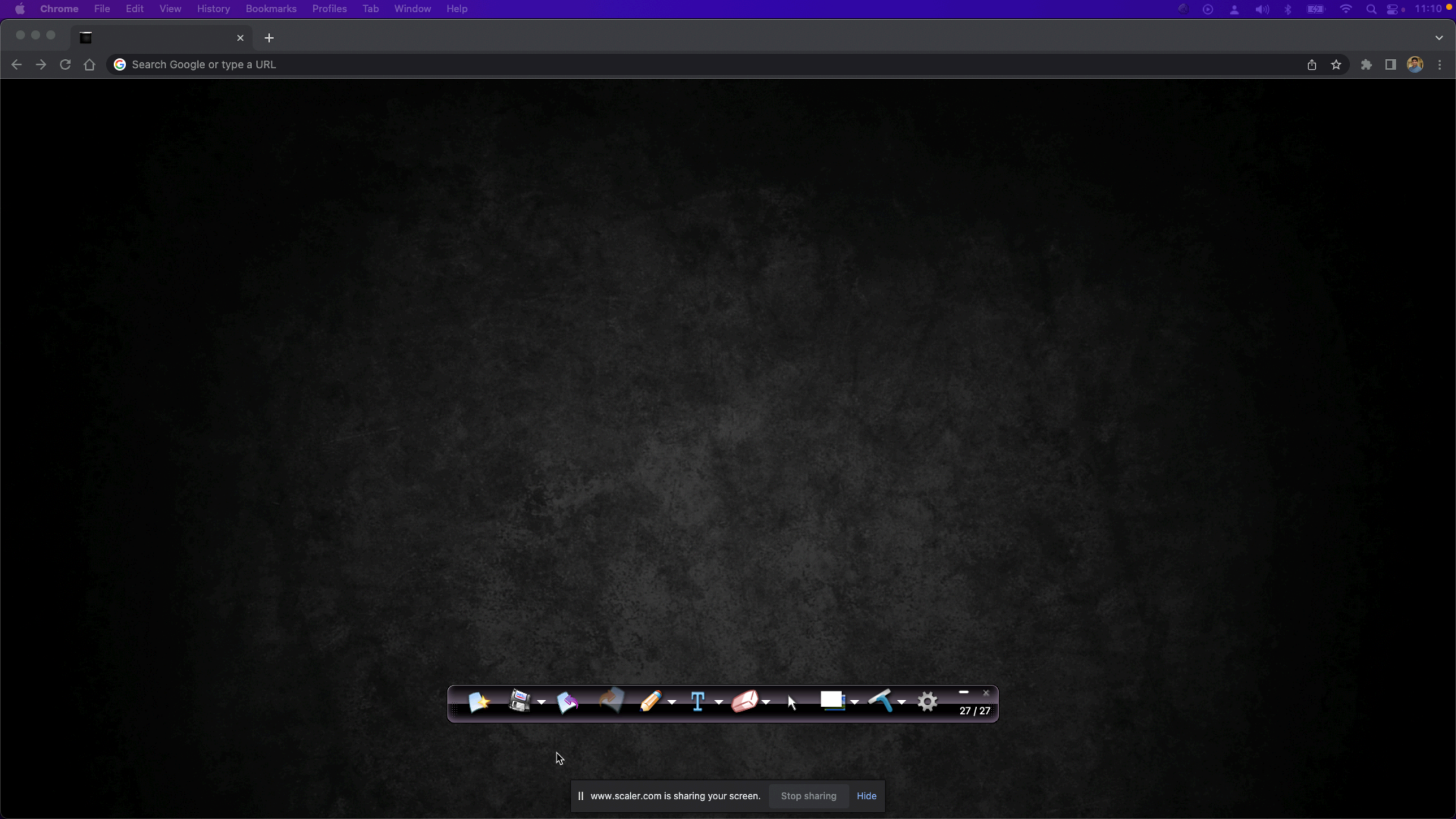





Q
C1
x
C2
C3 → 1·1x


log.reg ; DT ; RF
Fn
lr. Reg — GBDT, NB, SVM
LV
Kernel
KNN;


Q
BoW → Sparse
Numpy
dim : large
Text
MF → Dense
dim: smaller & high info density


Chrome File Edit View History Bookmarks Profiles Tab Window Help
Search Google or type a URL
27 / 27
www.scaler.com is sharing your screen.
Stop sharing
Hide