(Q1)

(a) K-means : optimization problem (course)

↳ euclidean-dist

$$\min_{M_1, M_2 \dots M_k} \left[\sum_{i=1}^{K} \sum_{x_j \in S_i} \| x_j - M_i \|^2 \right]$$

$M_1, M_2 \dots M_k$: d-dim (k x d)

M_i : centroid of i^{th} cluster

euclid. dist ✓

S_i : set of x_j's in cluster-i

[$x_j \in$ one cluster S_i $\forall j^{th}$ mathematically ✓]

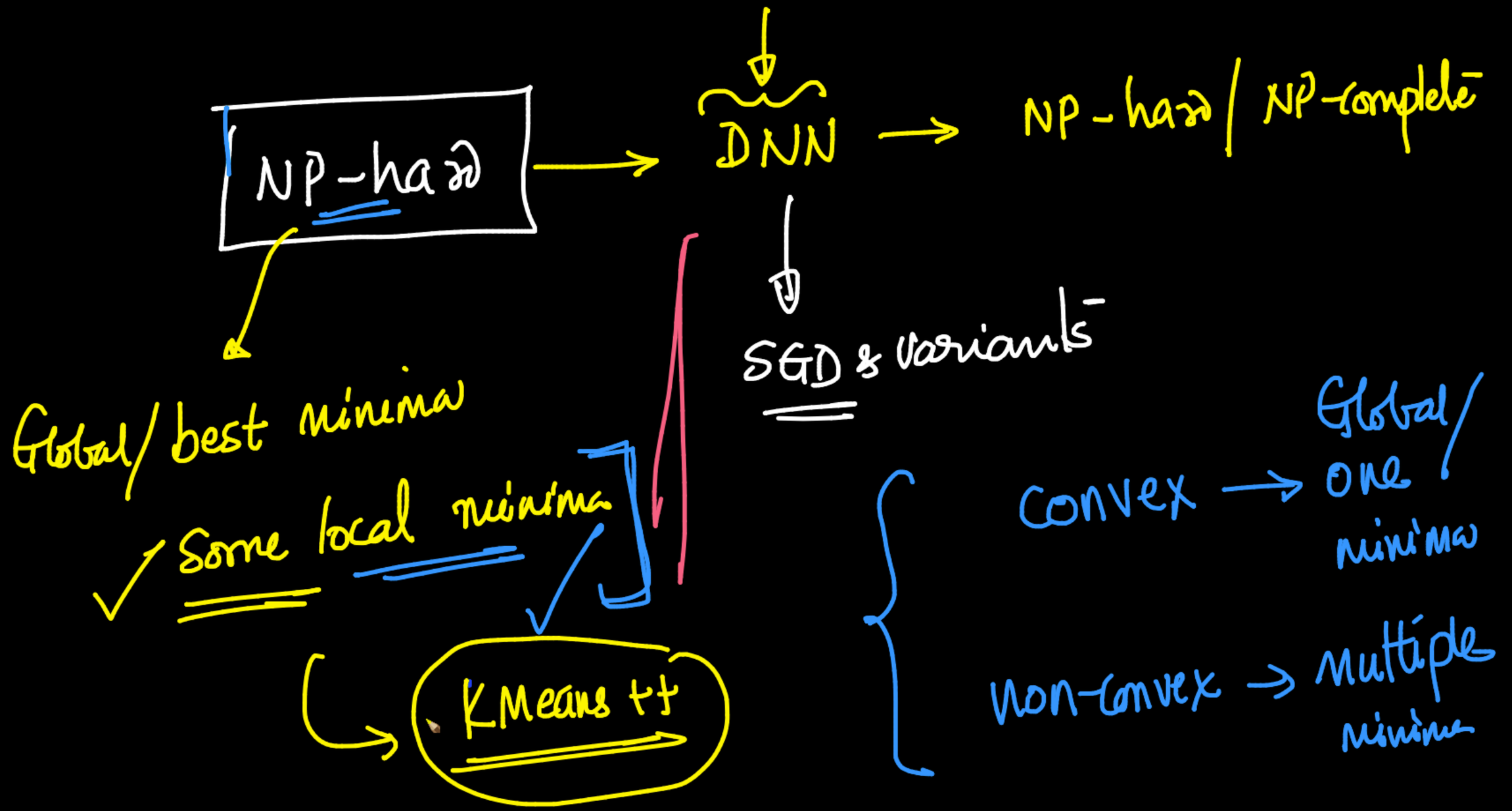
(*) how to identify clusters ✓

(b) Why can't solve it using GD methods ?

Log-reg:- $\min$ log-loss $+ \lambda \|w\|^2$

- ✓ MF:- $\min_{A_i, B_j} \|C_{n \times m} - A^T_{n \times d} B_{d \times m}\|^2$ no y_i's
- ✓ PCA:- optimization problem no y_i's
- ✓ SVM:-

$$\|x_j - M_i\|^2$$



$$\min_{m_i} \sum_{i=1}^{K} \sum_{x_j \in S_i} \|x_j - m_i\|^2 + \lambda_1 \square + \lambda_2 \square$$

non-differentiable

s.t

$$\begin{cases} x_j \in \text{one } S_i \\ S_i \cap S_j = \phi \quad i \neq j \end{cases} \rightarrow \text{set of pesali}$$

{ (c) Can you modify KMeans-optimization problem
s.t we can use GD methods: [V.H]

$$\min_{m_i, w_{ji}} \sum_{i=1}^{K} \sum_{j=1}^{n} \|x_j - m_i\|^2 \cdot w_{ji} + \lambda_1 \boxed{\checkmark} + \lambda_2 \boxed{\checkmark}$$

→ #clusters

→ #pts

assignment - matrix

$W_{n \times K}$

	1	2	...	i	...	K
1	0	0	0	...	1	... 0
2						
j				w_{ji}		
n						

s.t $x_j \in$ one s_i only

$s_i \cap s_j = \phi$ if $i \neq j$

$$\sum_{i=1}^{K} w_{ji} = 1$$ and $w_{ji} = 0$ or 1

GD

tricky

$w_{ji} \in \{0, 1\}$

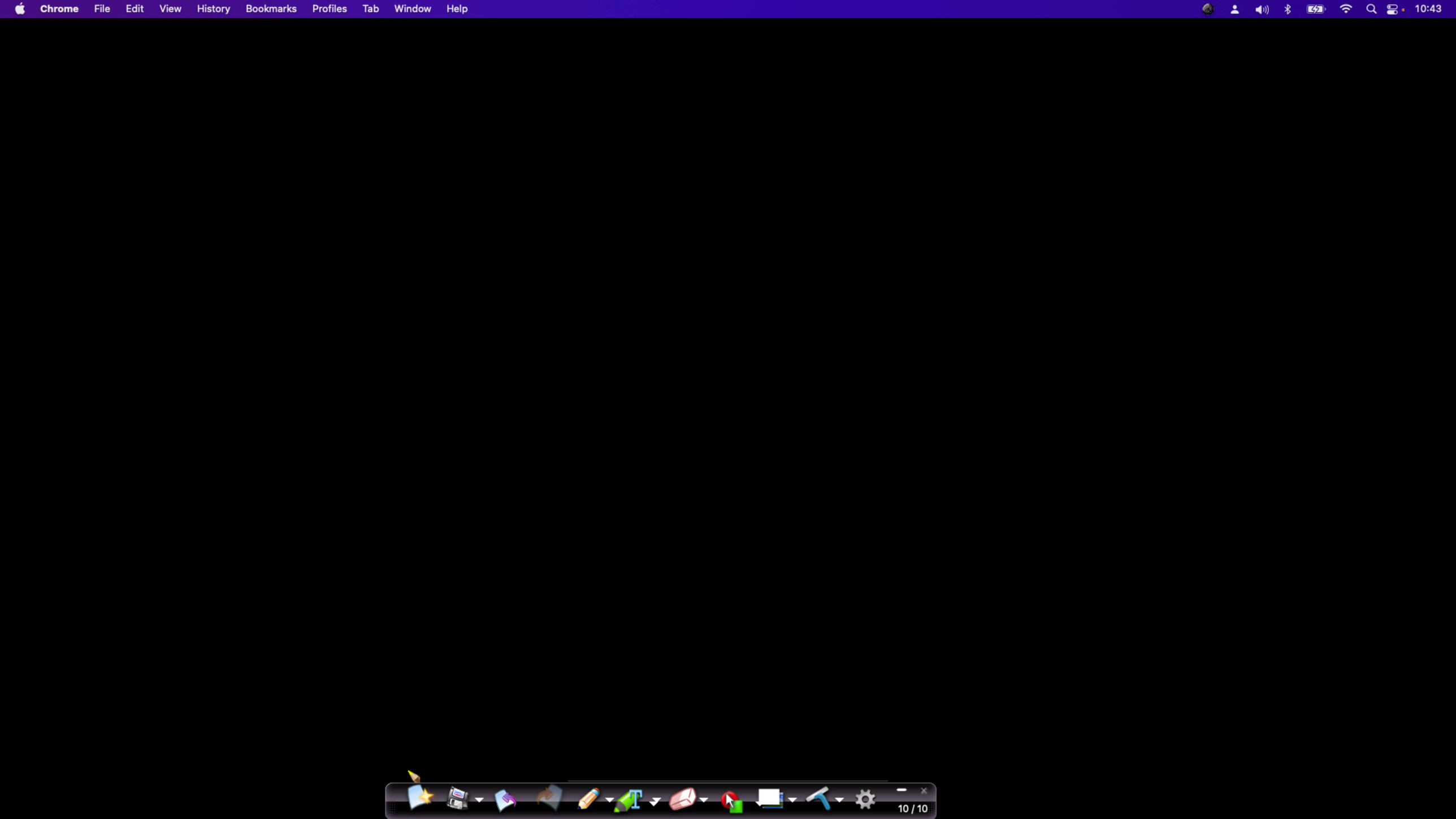
$w_{ij} \in \{0, 1\} \quad \forall i, j$

$w_{ij} = 0 \text{ or } 1 \quad \forall i, j$

Mathematically as a differentiable fn...

$\checkmark\ w_{ij} = 0$

$\checkmark\ w_{ij} = 1$


Chrome File Edit View History Bookmarks Profiles Tab Window Help
10:43
10 / 10

d-dim

$$\mathcal{L} = \min_{m_i, w_{ji}} \sum_{i=1}^{K} \sum_{j=1}^{n} w_{ji} \, \| x_j - m_i \|^2 + \lambda_1 [\quad] + \lambda_2 [\quad] + \lambda_3 [\quad]$$

s.t

$$\sum_{i=1}^{K} w_{ji} = 1 \quad \forall j = 1 \to n$$

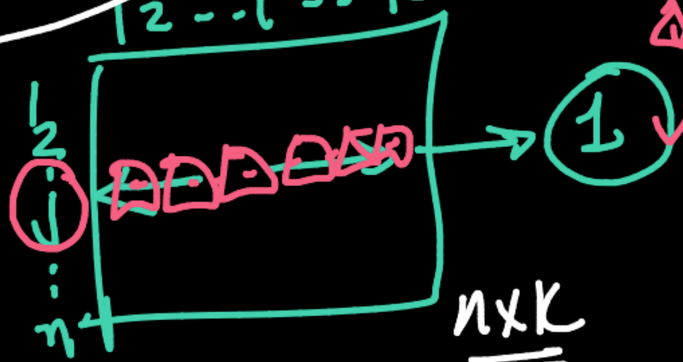
Problem ?

w_{ji} can take any value

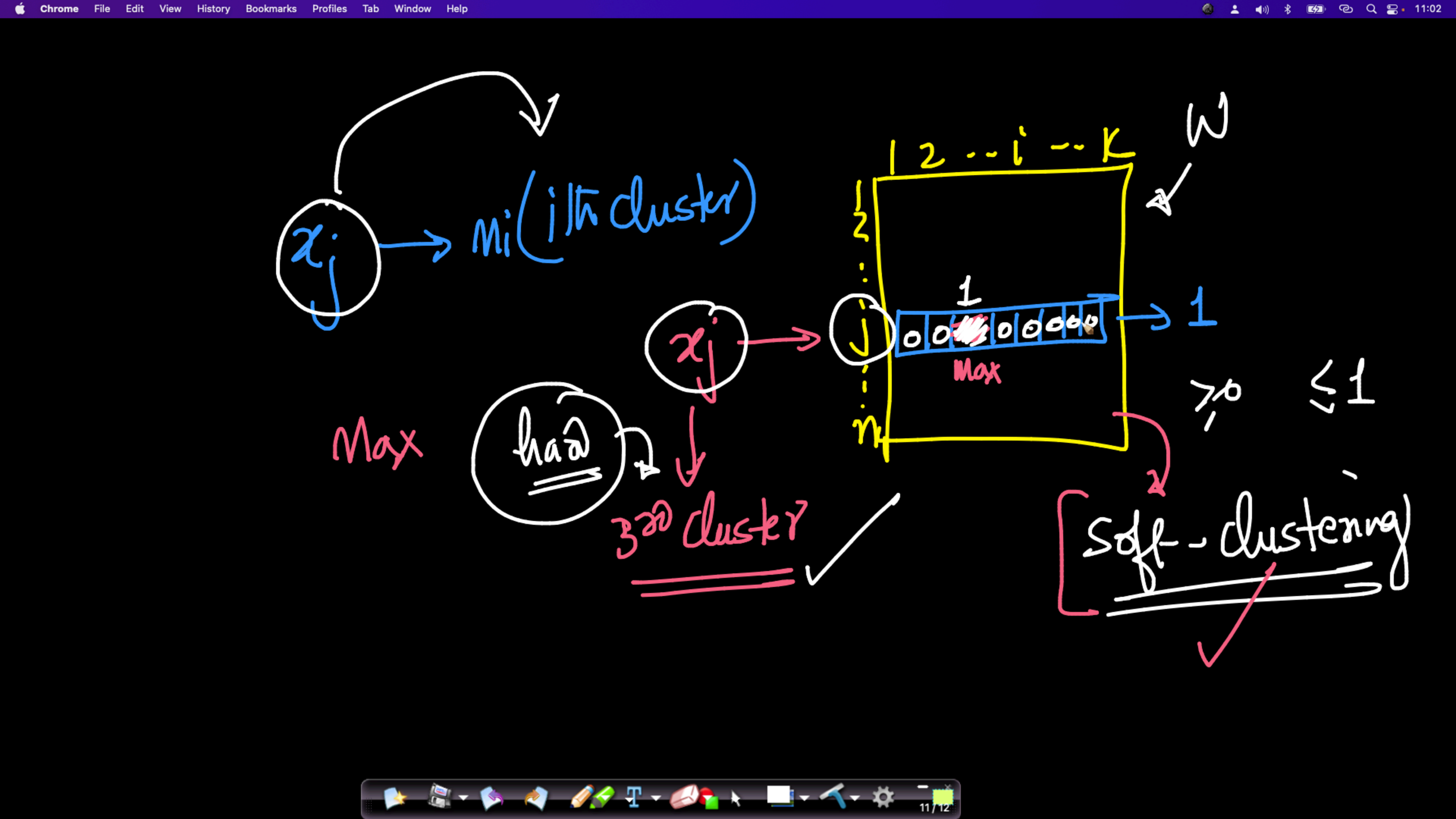
$w_{ji} = P(j^{th} \text{ pt} \in i^{th} \text{ cluster})$

$w_{ji} \geq 0$

$w_{ji} \leq 1$



$n \times K$


Chrome File Edit View History Bookmarks Profiles Tab Window Help
11:02
x_i
M_i (i th cluster)
1 2 .. i .. K
W
1
2
j
n
1
Max
1
x_j
Max
had
3rd cluster
≥ 0
≤ 1
soft - clustering
11 / 12

(Q2)

K-means ✓

n = 1B-points

d = 20

K = 10

(EASY) ✓

practical

[fast-clustering →

- Single-box
- < 30 min

sdn-1

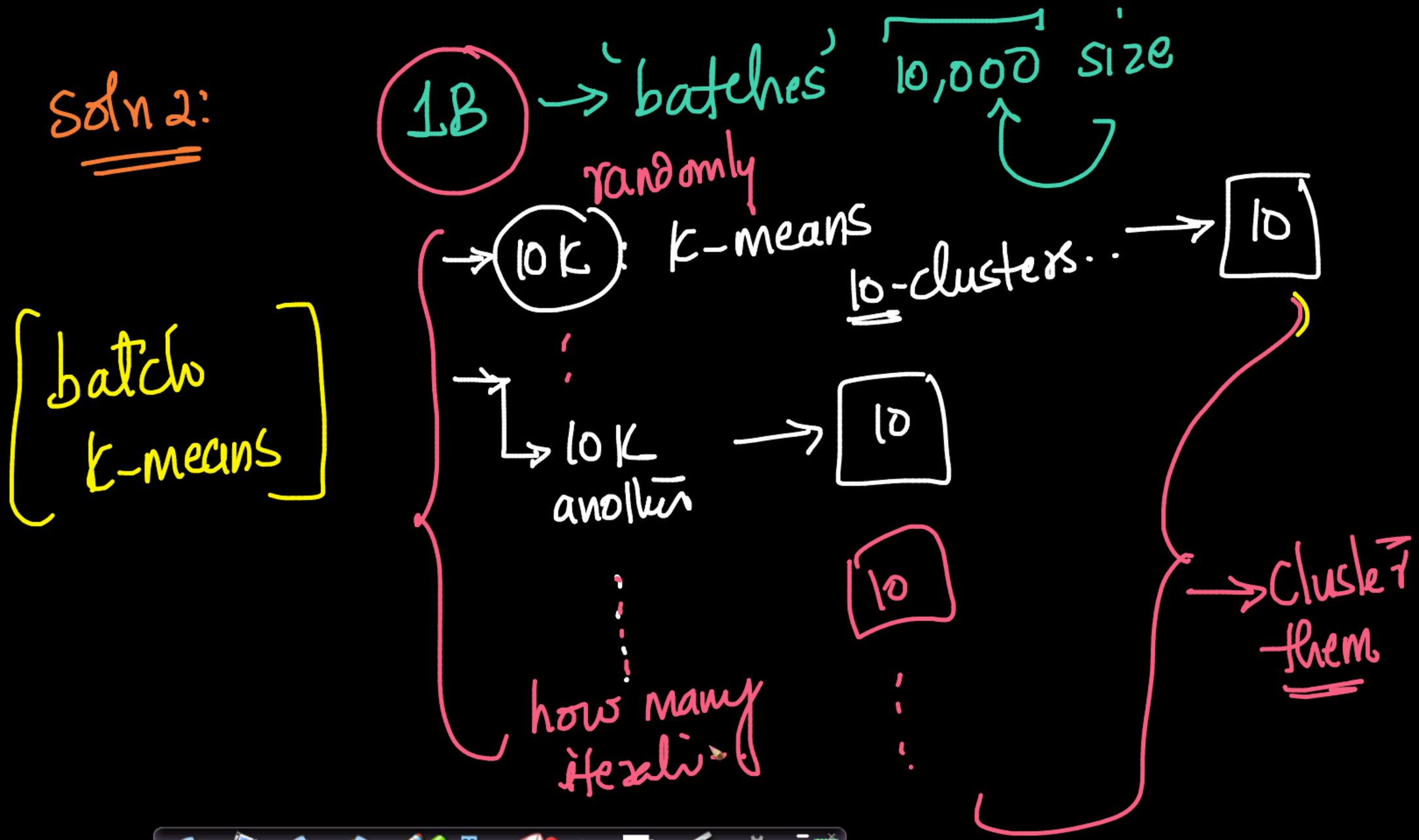
randomly sample 1M points ✓

→ generalizes well

→ KNN like to assign ...

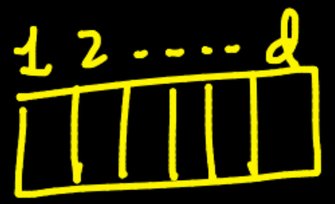
Chizag

Soln 2:



Q3

Text docs :- TF-IDF (fixed)



Practical

(a) K-means ? → Can be done

↓

d is high → Computation (ignore)

[euclidean dist]

→ Curse-of-dimensionality

✓ Alternatives: reduce d → removing some low utility words

→ PCA, SVD, NMF

→ Autoencoders..

Alternatives: cosine-sim / dist ✓

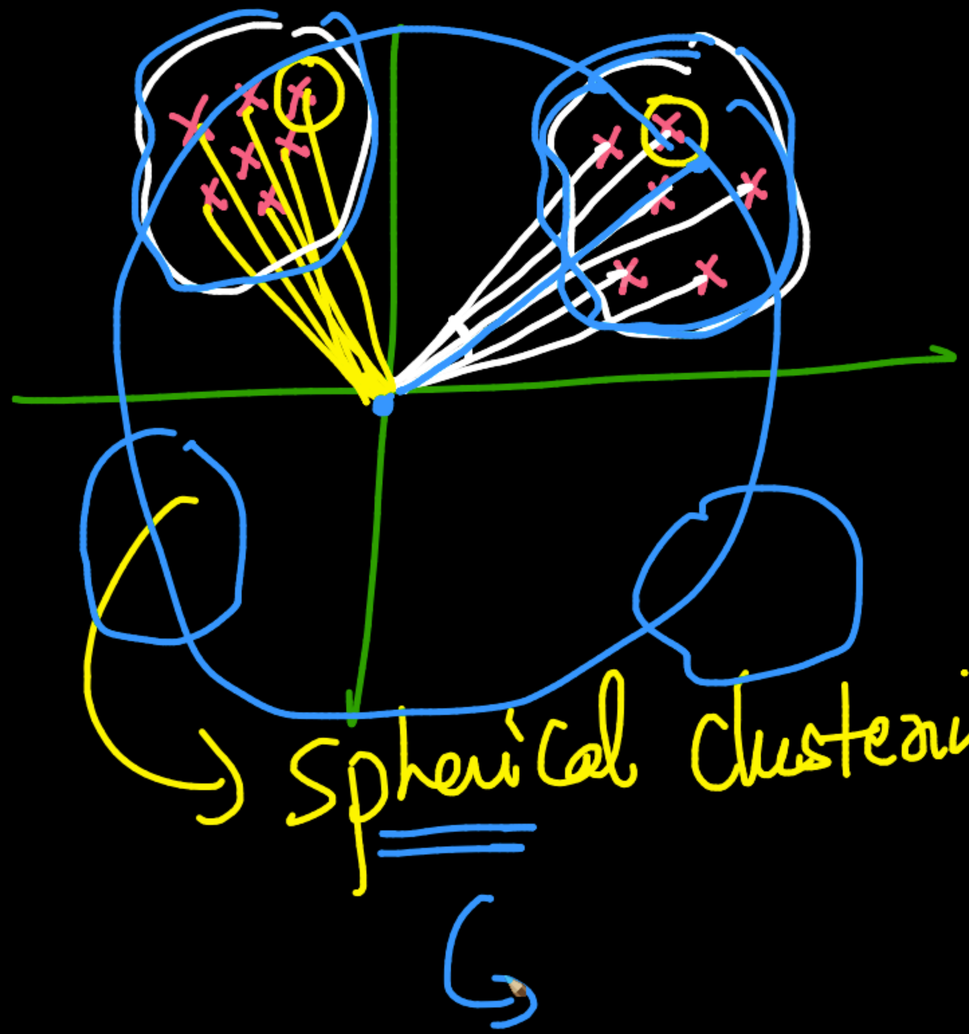
↓

how would you compute centroid

HINT: think geometrically...

mean → euclidean-dist

→ cosine-dist



Centroid : Unit vec

mean-angular-sense

↳ spherical clustering

↳

Q4

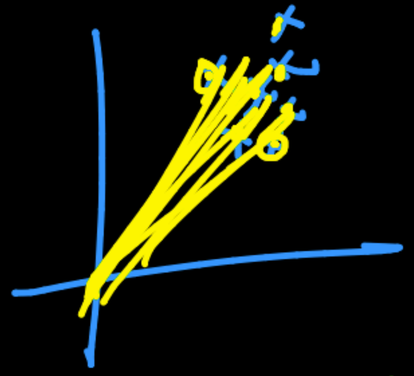
DBSCAN

precautions in practice

DICE

(Minpts*, ε*) best

elbow Method



Vary ε carefully

MinPts, ε : params

range-queries

core pts, border outliers

① → Small change in params ⇒ different clusters

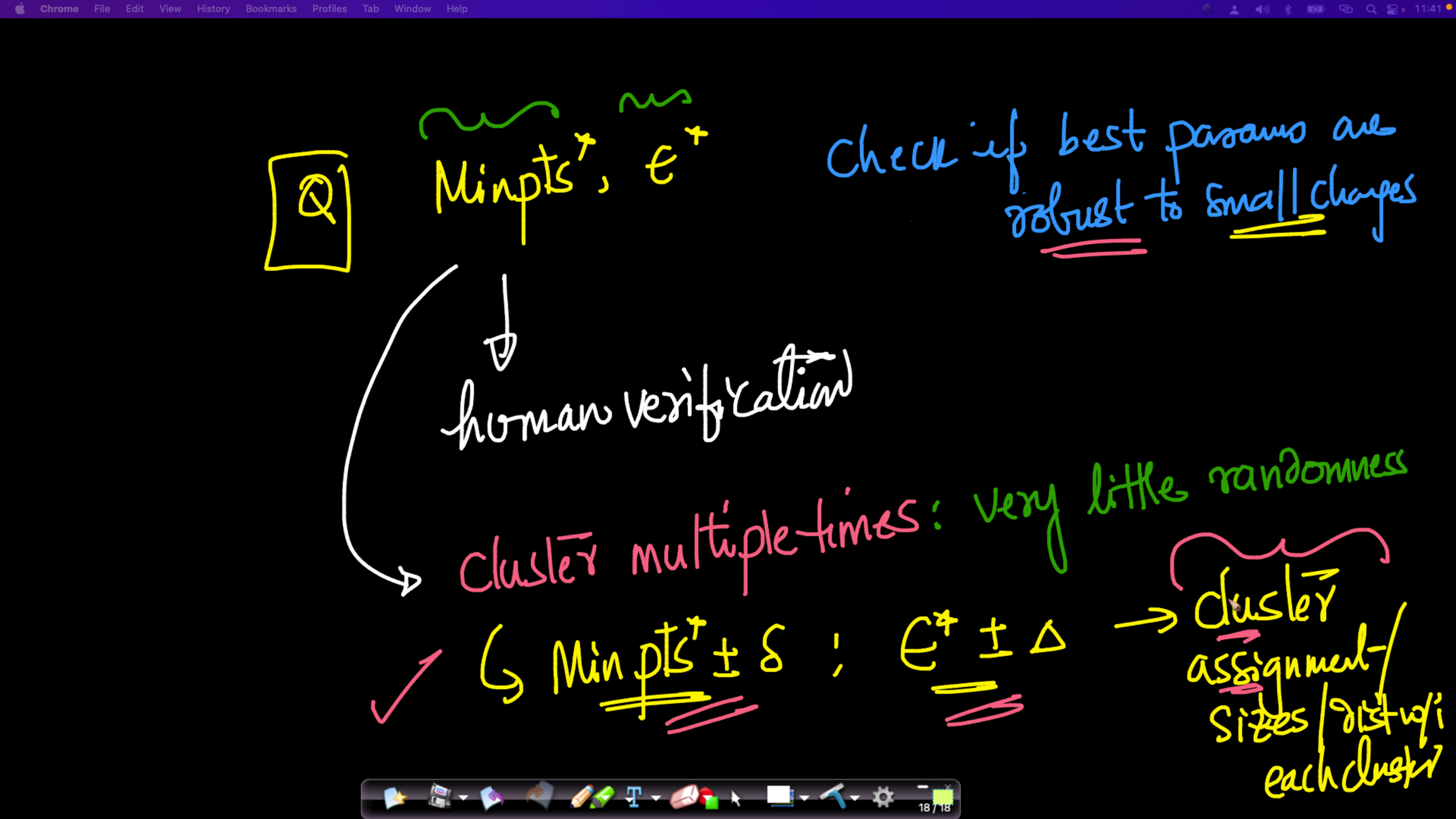
reduce dim

② → dim is large – curse of dim

d = 1000

Cost (one-hot encoded)

cosine sim


Q
MinPts*, ε*
Check if best params are robust to small changes
human verification
cluster multiple times: very little randomness
MinPts* ± δ ; ε* ± Δ
cluster assignment/ sizes / dist w/i each cluster

- Single-linkage: MIN-agglo..
- Complete " : MAX - "

↳ cases where they fail (Course)

Q

KMeans ++ initialization

IM

[how to optimize KMeans++ for speed

↳ parallelizes (not trivially)

1 randomly sample + ✓

10K ✓

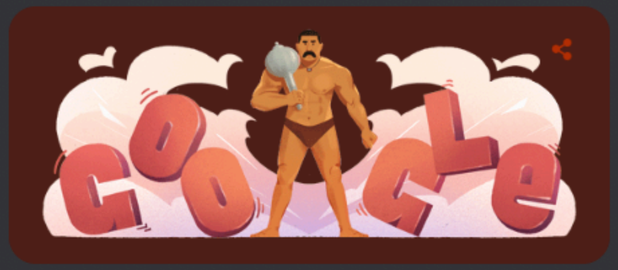
Chrome File Edit View History Bookmarks Profiles Tab Window Help 12:17

https://www.airmeet.c | Airmeet: Interview que | Applied AI Course(Scr | Unsupervised learning | Hierarchical clustering | DBSCAN.docx - Goog | Microsoft PowerPoint | Find Open Datasets a | New Tab

Search Google or type a URL

Apps | Google Bookmark | Google Scholar | Share to Google+ | Stumble! | Live Tv Channels... | User generated g... | User Generated C... | Best Implementati... | Mod (video gamin... | YouTube - Automa... | Kinect hack move... | Other Bookmarks

Gmail Images



Search Google or type a URL

Applied Roots | Airmeet | Google Drive | youtube.com | GATE CS App...

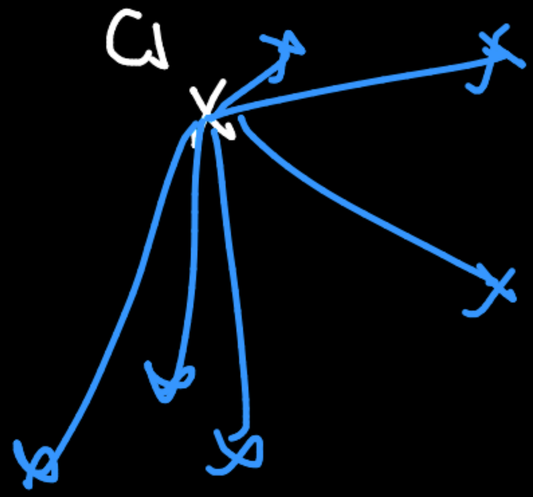
Google Drive | Applied Roots | Scaler by Inte... | Inbox (1,299) | Add shortcut

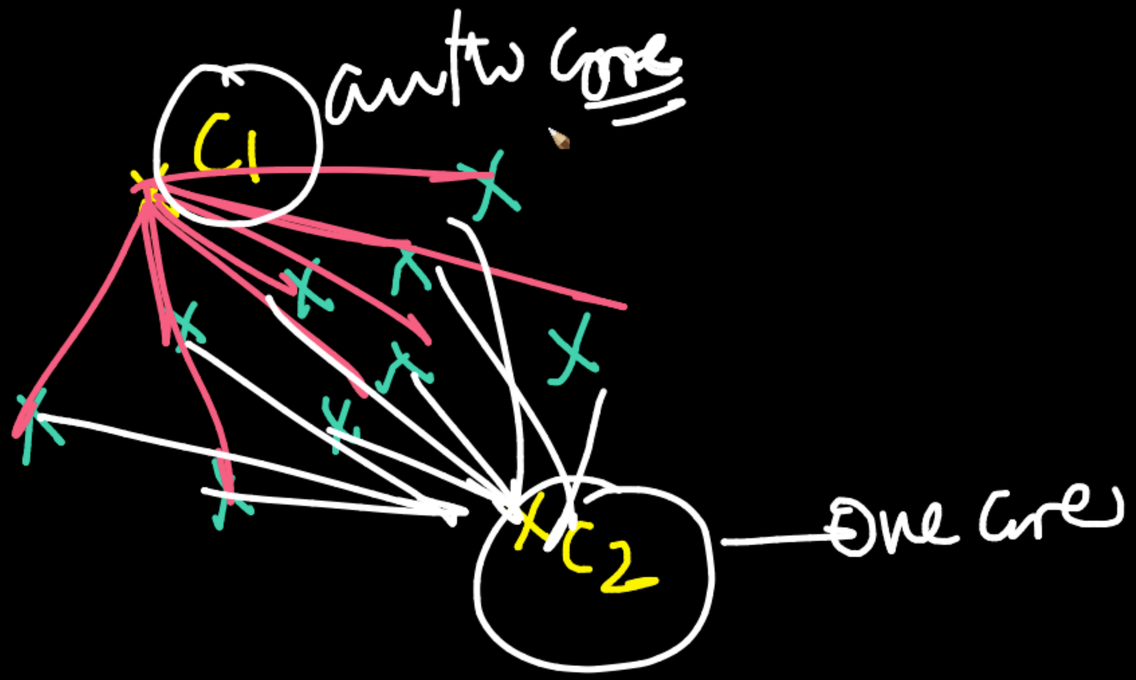
21 / 24

Customise Chrome

prob. fn → remove outliers...

pick farthest pt




C1
C2
One core