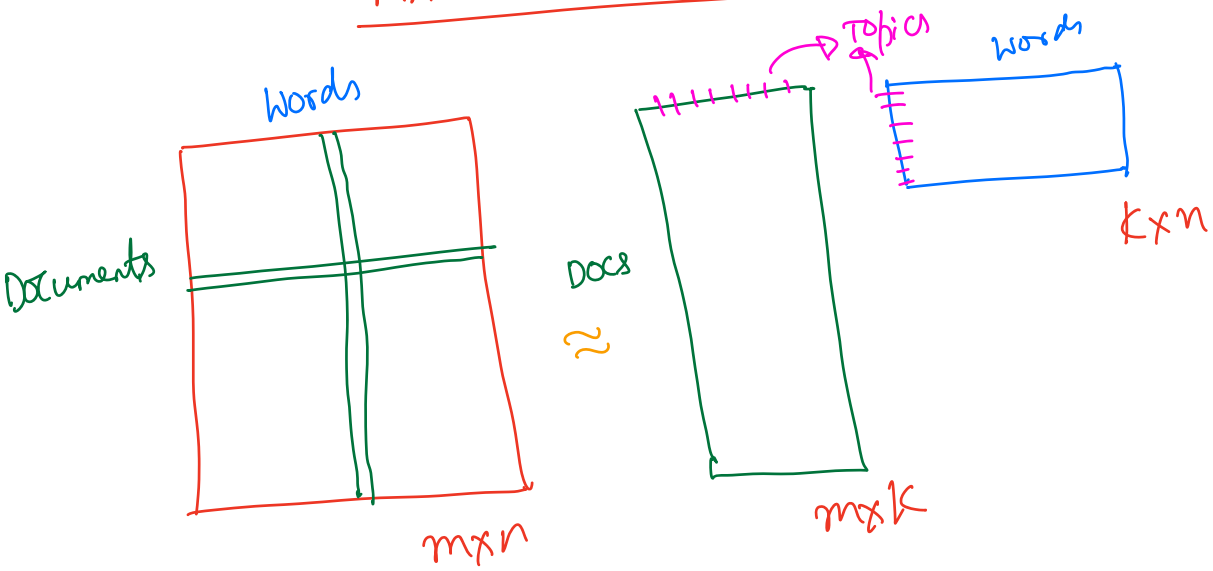
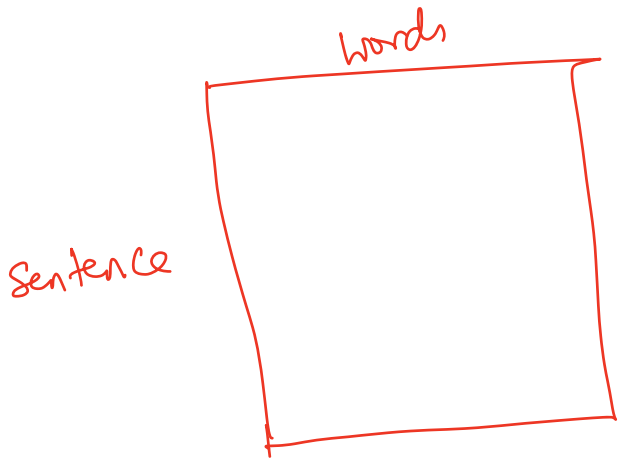
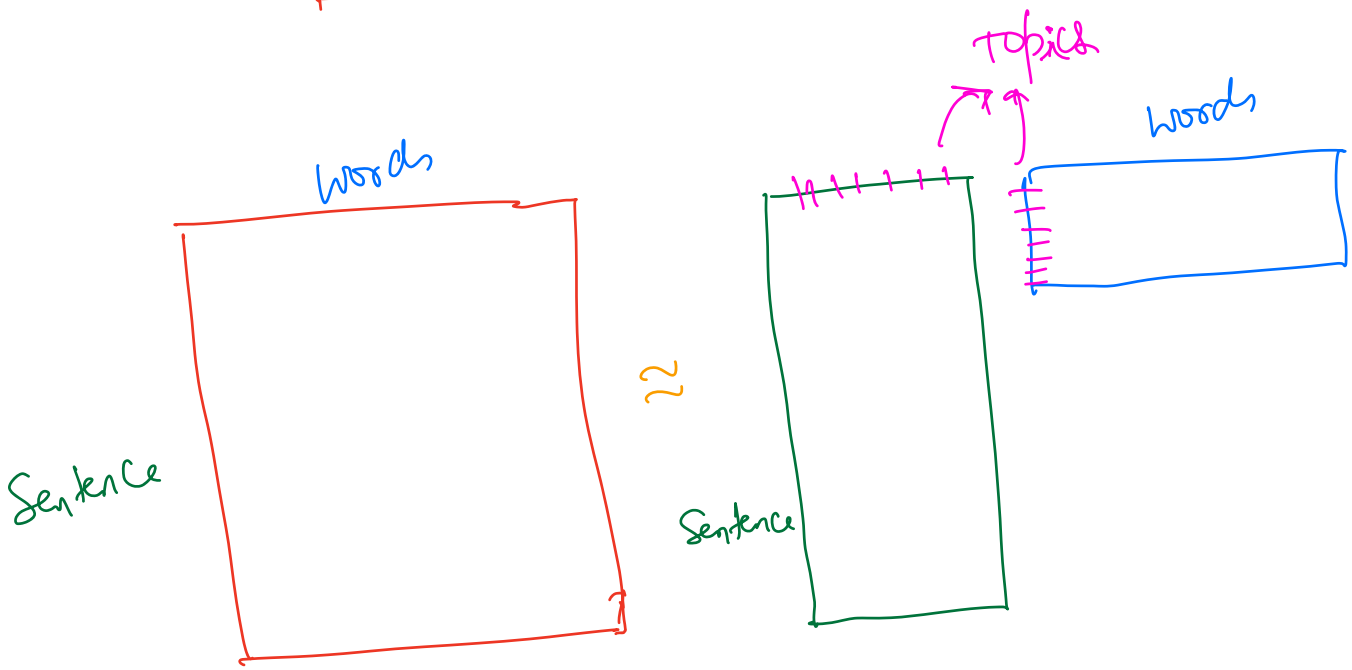
# LECTURE 12

## MATRIX FACTORIZATION




k = # Topics




$K$ = #topics

→ Give us the weightage of "Latent topics" in a sentence

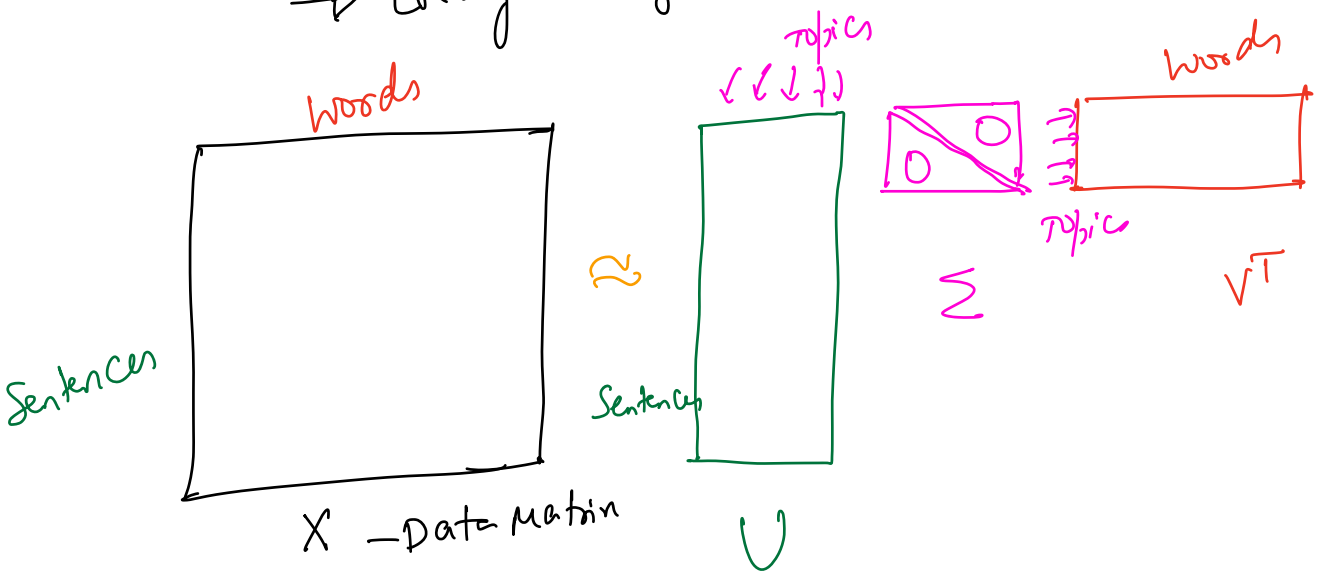
→ For a medical document, need to segment each sentence into a topic!

→ How do we assign "one topic" to a sentence?

# SVD

→ Singular value decomposition

→ Every single matrix has a SVD!




Σ - Digonal matrix

# ICE #0

Let $X$ be a data matrix and $X \approx U \Sigma V^T$ be the $k^{th}$ rank SVD of $X$.

i.e. $\Sigma$ has $k$ diagonal entries.

If $X$ is $m \times n$ matrix, what is the dimension of $U$?

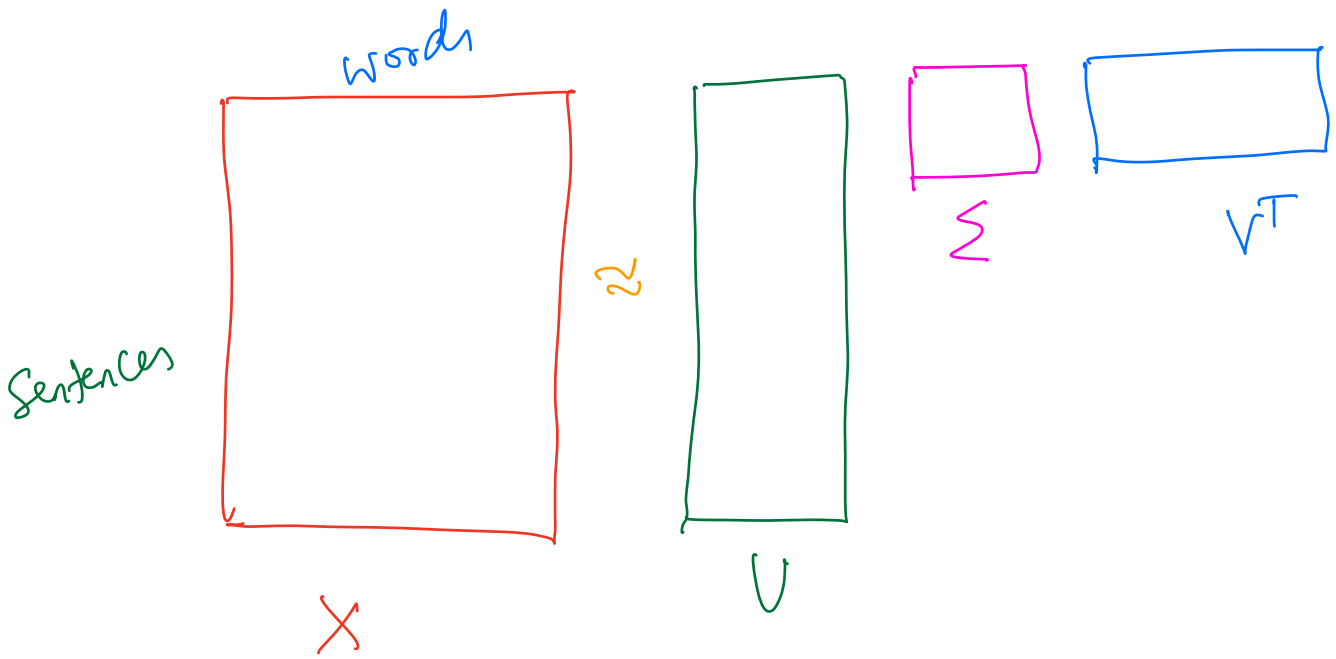
a) $k \times m$

b) $k \times k$

c) $m \times k$

d) $m \times n$

# SVD yields a trivial matrix factorization!




$$X \approx U\Sigma V^T = (U\Sigma)\, V^T$$

$U\Sigma$: factor 1; $V^T$: factor 2

How do we choose k, the number of latent topics?

→ SVD based matrix factorization doesn't guarantee that each sentence gets assigned a topic!

# Matrix Factorization

## – An optimization perspective!

$X$ – Data matrix – $m \times n$

$U$ – Left factor

$V^T$ – Right factor

**Objective:-** $\min_{U,V} \|X - U V^T\|_F^2$ (dimensions: $m \times n$, $m \times k$, $k \times n$)

without any constraints, $U$ & $V$ can be obtained from SVD as discussed!

# ICE #1

Consider the obj. fn. $\min_{U,V} ||X - UV^T||_F^2$.

Let's say $k = m$ & $X$ is of dim. $m \times n$.

What is the minimum value the loss fn. can take in this case?

(a) $-0.5$

(b) $0.5$

(c) $0$

(d) Some value $> 0$

Simply fix $U = I$ & $V = X^T$.

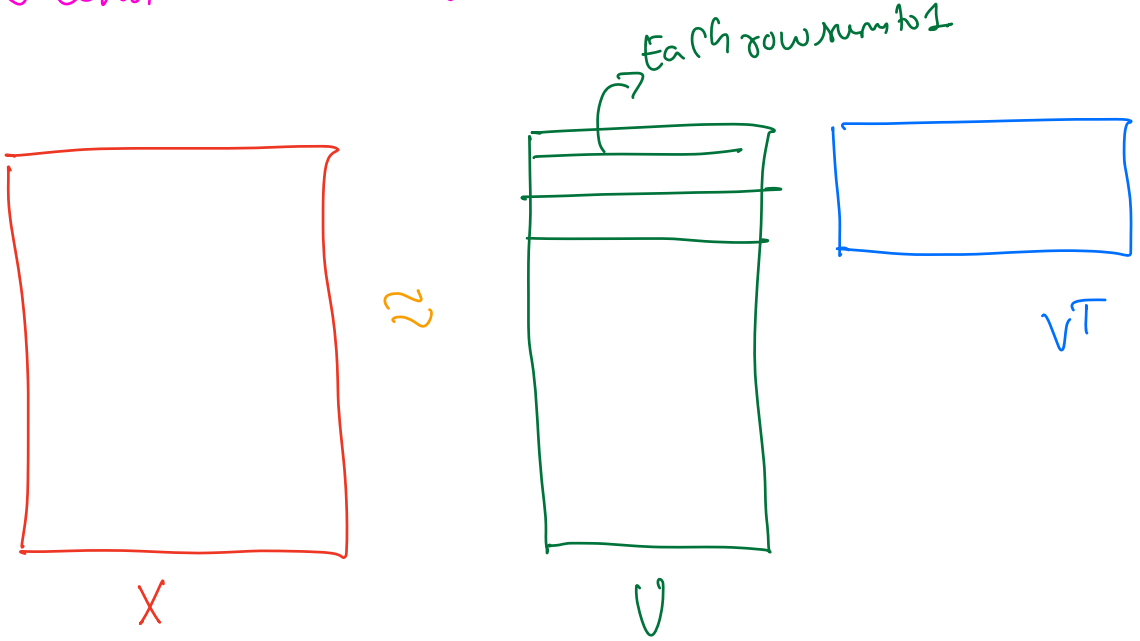
What's the loss fn. value now?

Segmenting each sentence into a "Latent Topic"
(or
– Unsupervised Topic Segmentation")

~ SVD or simple matrix factorization doesn't assign a single topic to a sentence

~ work around!

~ Constrain the left factor to be a probability matrix




$U1 = 1$ (Each row of U is a probability vector!)

$$\min \|X - UV^T\|_F^2$$

$U1 = 1,$

$U \geq 0, V$

This turns out to be exactly what we think it is!!

U – Assignment matrix

V – Centroid matrix

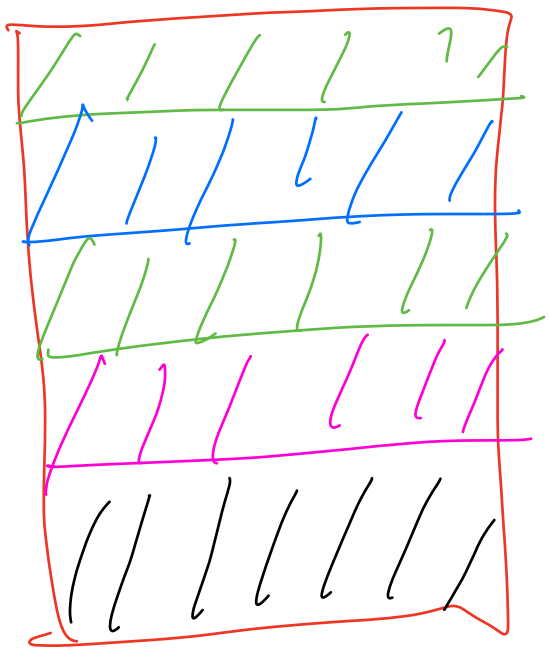
Algorithm:- (K-means)

1. Initialize V to be k random sentences from X
2. Iterate:-
   a) For fixed V, optimize for U
      – Assign sentences to a representative centroid
   b) For fixed U, optimize for V
      – Average sentence vectors in a cluster

k-means = Algorithm that optimizes a
"Matrix Factorization Lon
with Constraints!"

Probabilistic/Soft k-means
→ Instead of hard assignment
of a sentence to a topic,
soft assign!

# Topic Segmentation



# Two Problems in Health Care

1) Digital scribe notes to EHR

2) Curate EHR's → Info. retrieval!

↓

E.g.
1) Get all patient records with a particular straining disease?

2) Get % of patients that got treatment y

# Challenges to Info. Retrieval

1) Subject Identification

E.S. "Patients father had diabetes"

2) Temporality

E.S. "patient had diabetes until last year."

NLP can help: 1) Graph of entities & relationships

2) query graph

# Explainability of AI in health care.

Why?

- Informed Consent
- Liability Issues!
- Decision based on meta data
  - E.g. Huskies vs Wolves!
- Avoid medical malpractice Lawsuits (Legal issue)

Note:- Next Lecture!