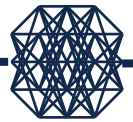
3. Machine Learning Basics

Prof. Jimeng Sun

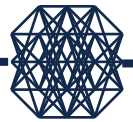
ILLINOIS

Unsupervised learning



- Dimensionality Reduction
- Clustering methods

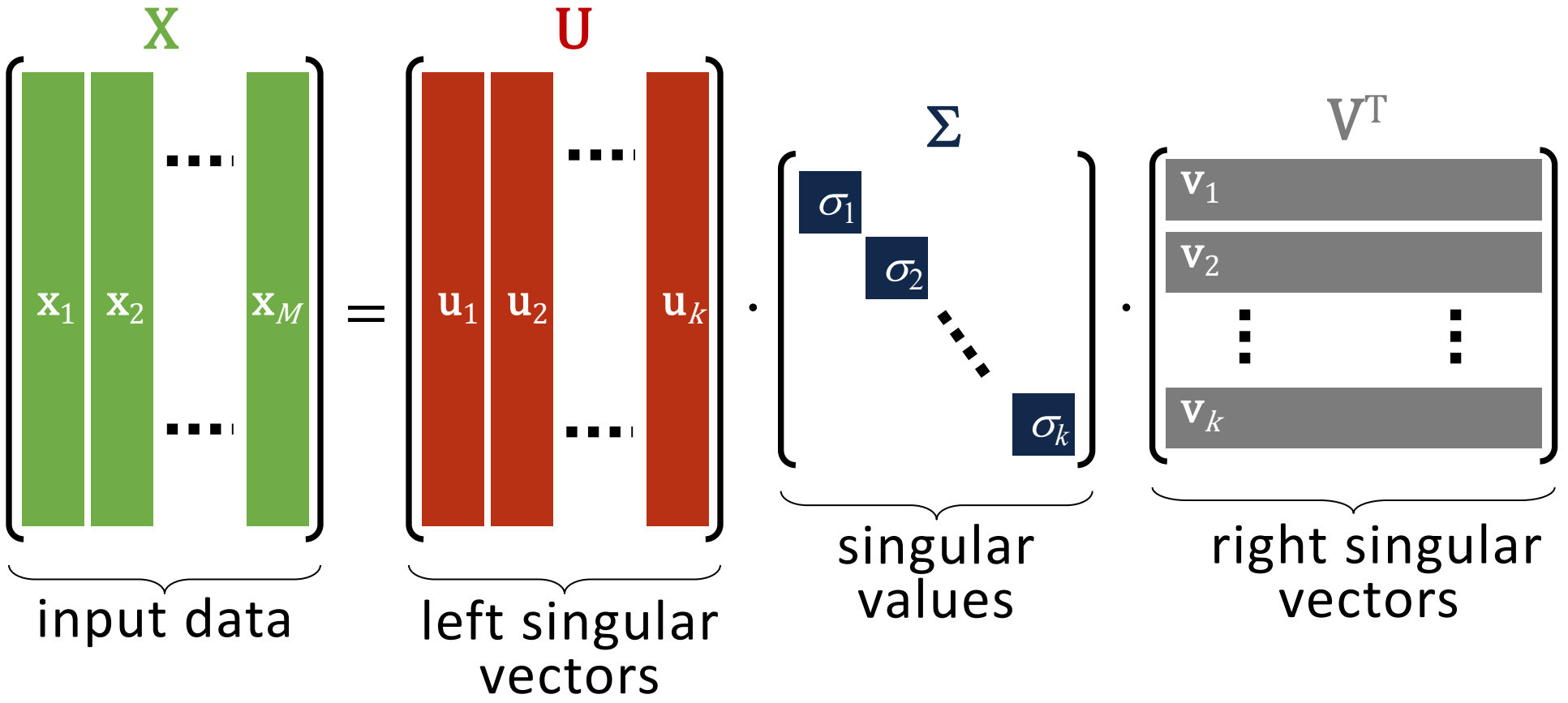
Dimensionality Reduction



- Singular Value Decomposition (SVD)
- Principal Component Analysis (PCA)

Singular Value Decomposition (SVD)

$$X = U\Sigma V^T \quad \text{where} \quad U^TU = V^TV = I$$



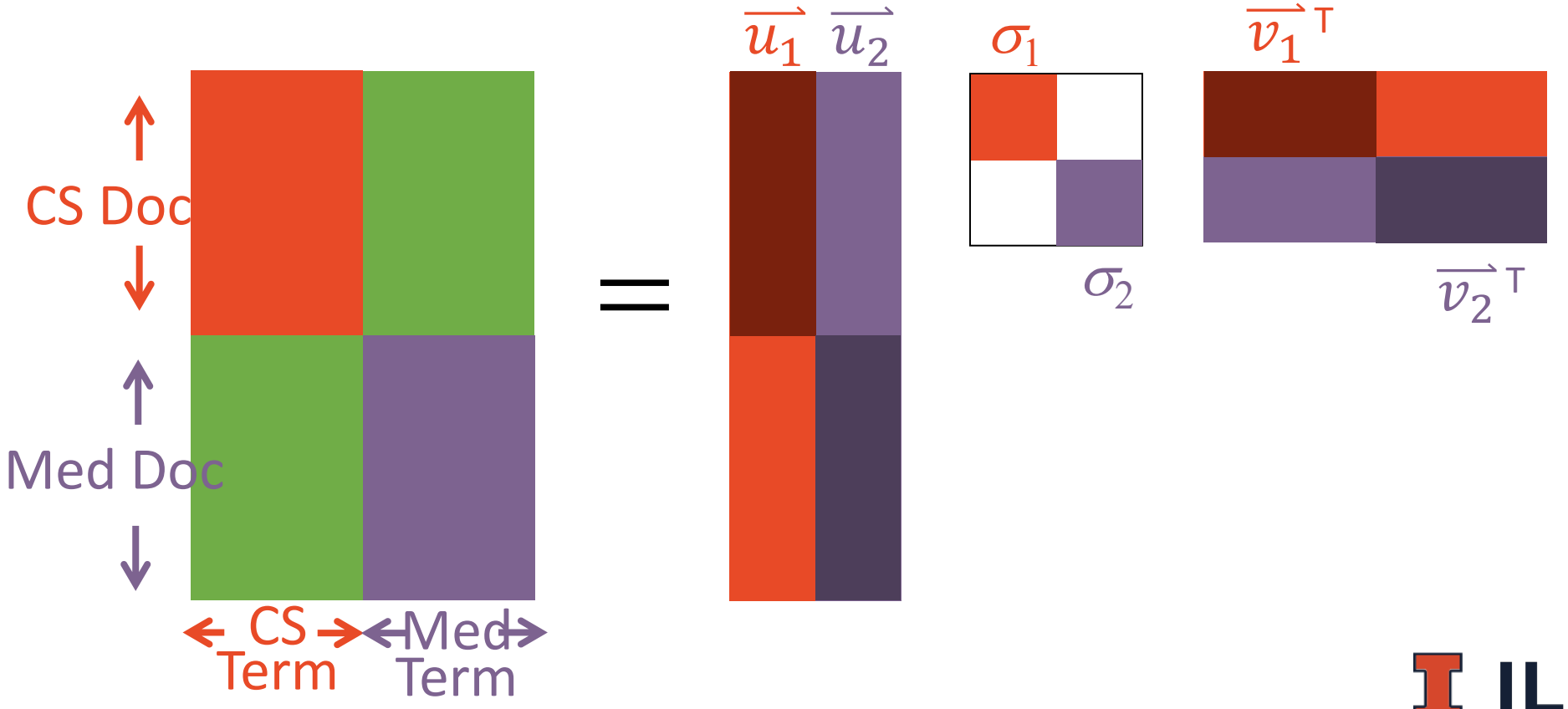
ILLINOIS

SVD Example

$$X = U\Sigma V^T = \sigma_1 \overrightarrow{u_1}\overrightarrow{v_1}^T + \sigma_2 \overrightarrow{u_2}\overrightarrow{v_2}^T + \ldots$$

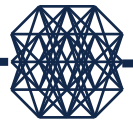
matrix view

spectral view



ILLINOIS

Quiz: SVD - Interpretation



Given a document-by-term matrix **A**, what is $\mathbf{A}^T\mathbf{A}$?

- ☐ document-to-document similarity matrix
- ☑ term-to-term similarity matrix
- ☐ term-to-document similarity matrix

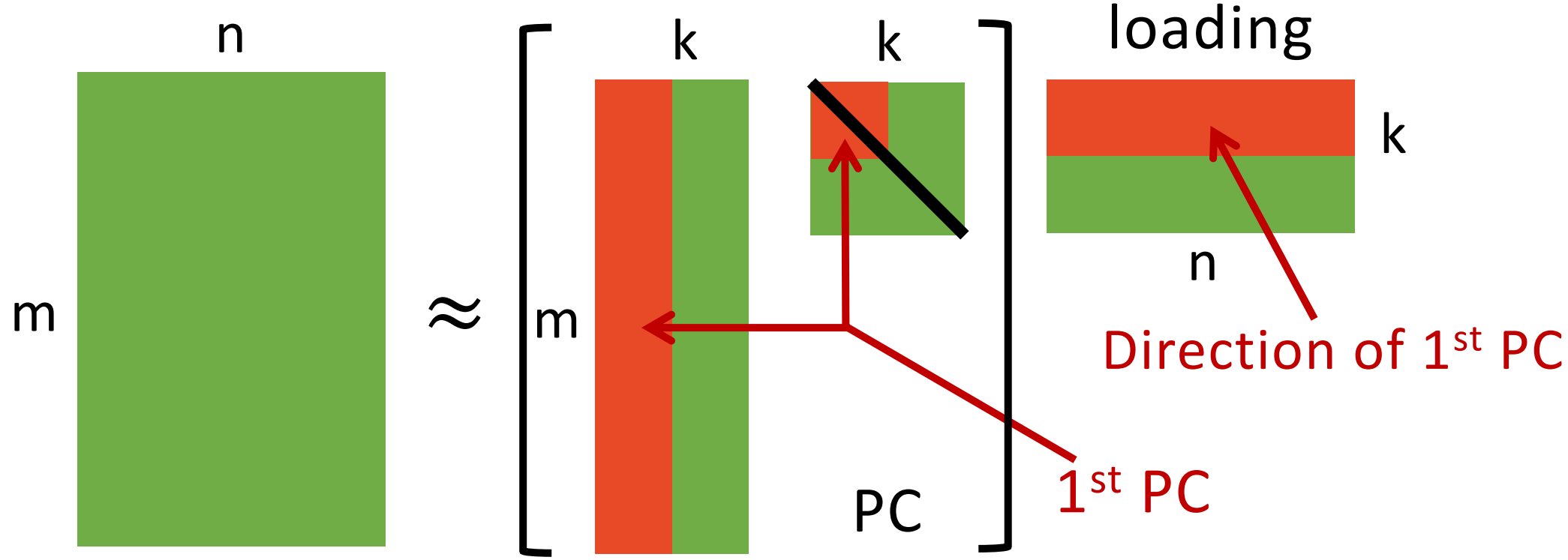
ILLINOIS

Principal Component Analysis (PCA)

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

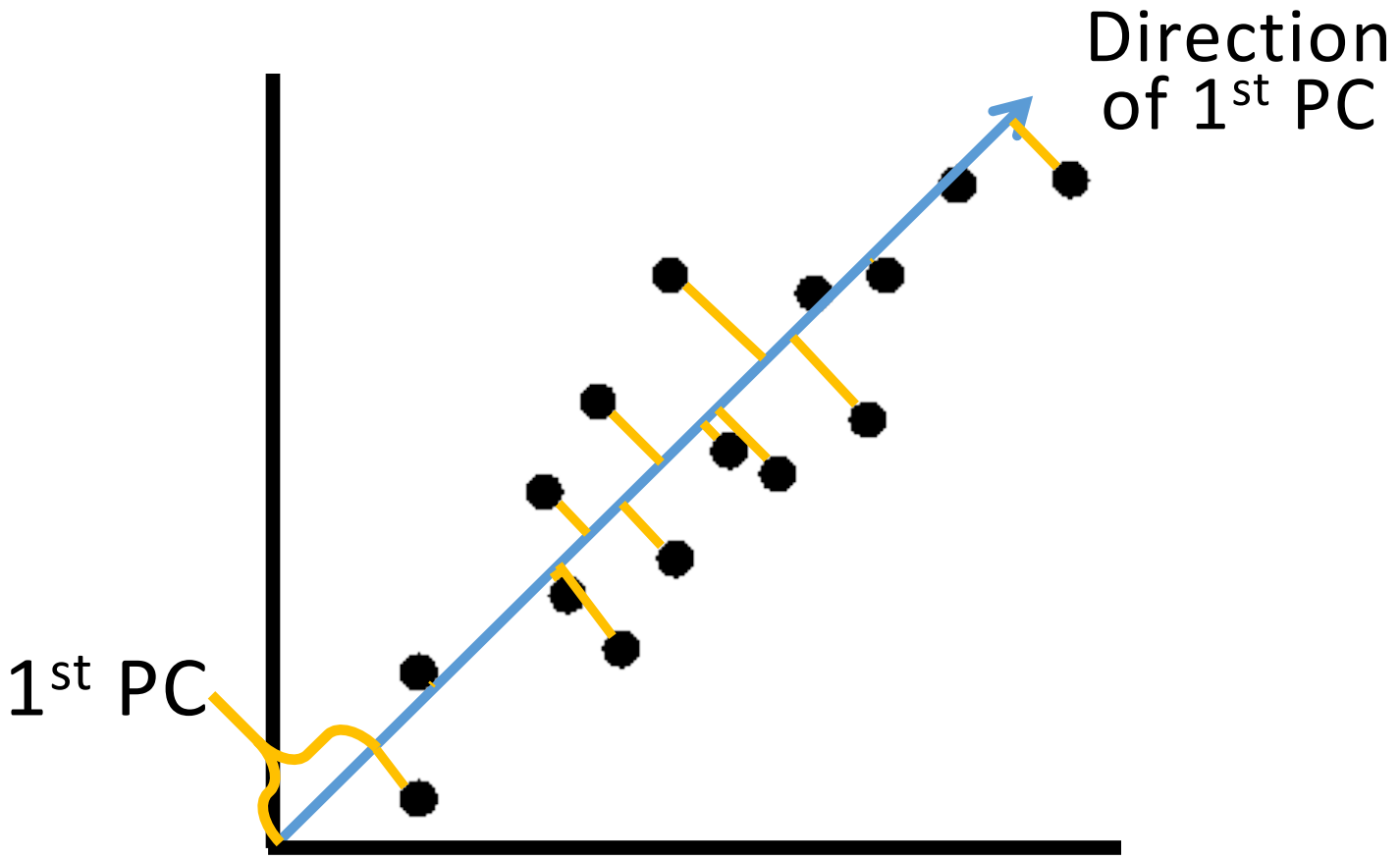
PC

loading



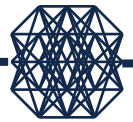
ILLINOIS

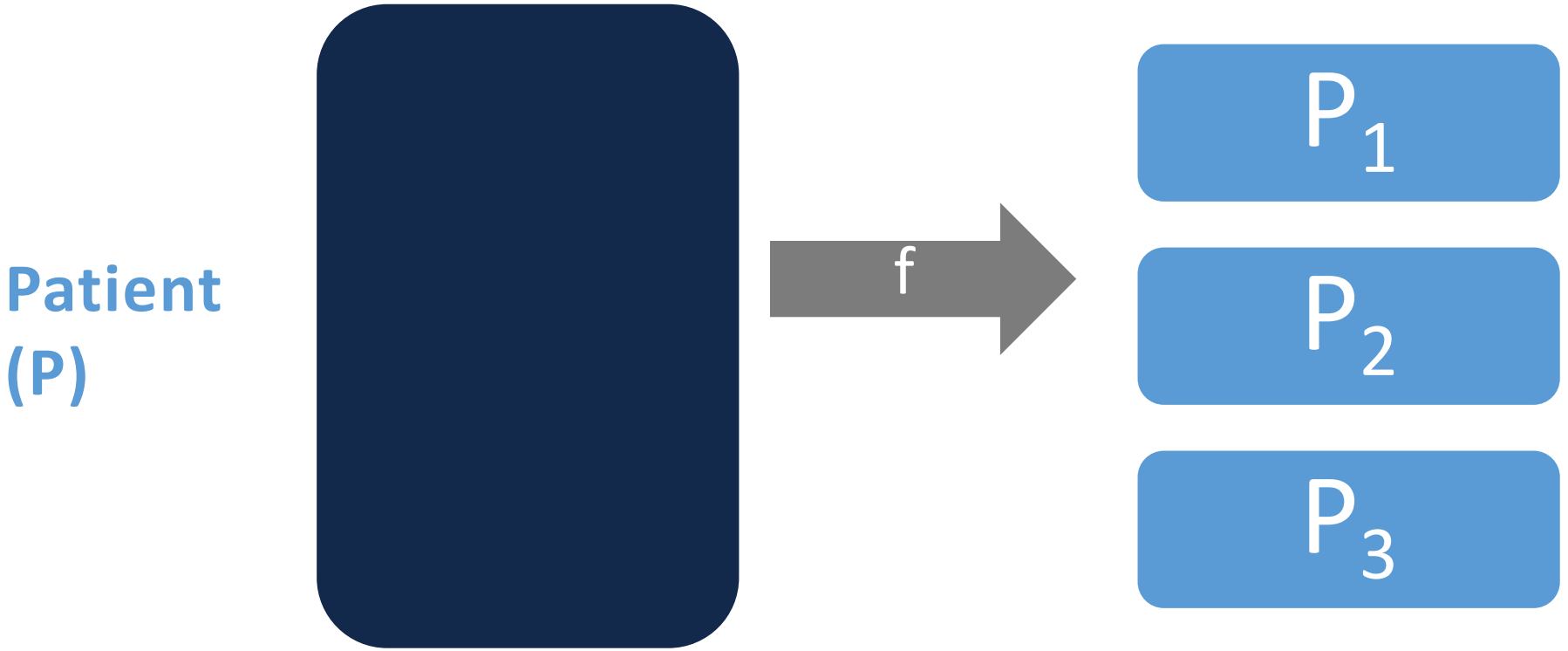
PCA INTERPRETATION



ILLINOIS

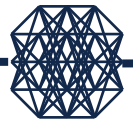
What is Clustering?

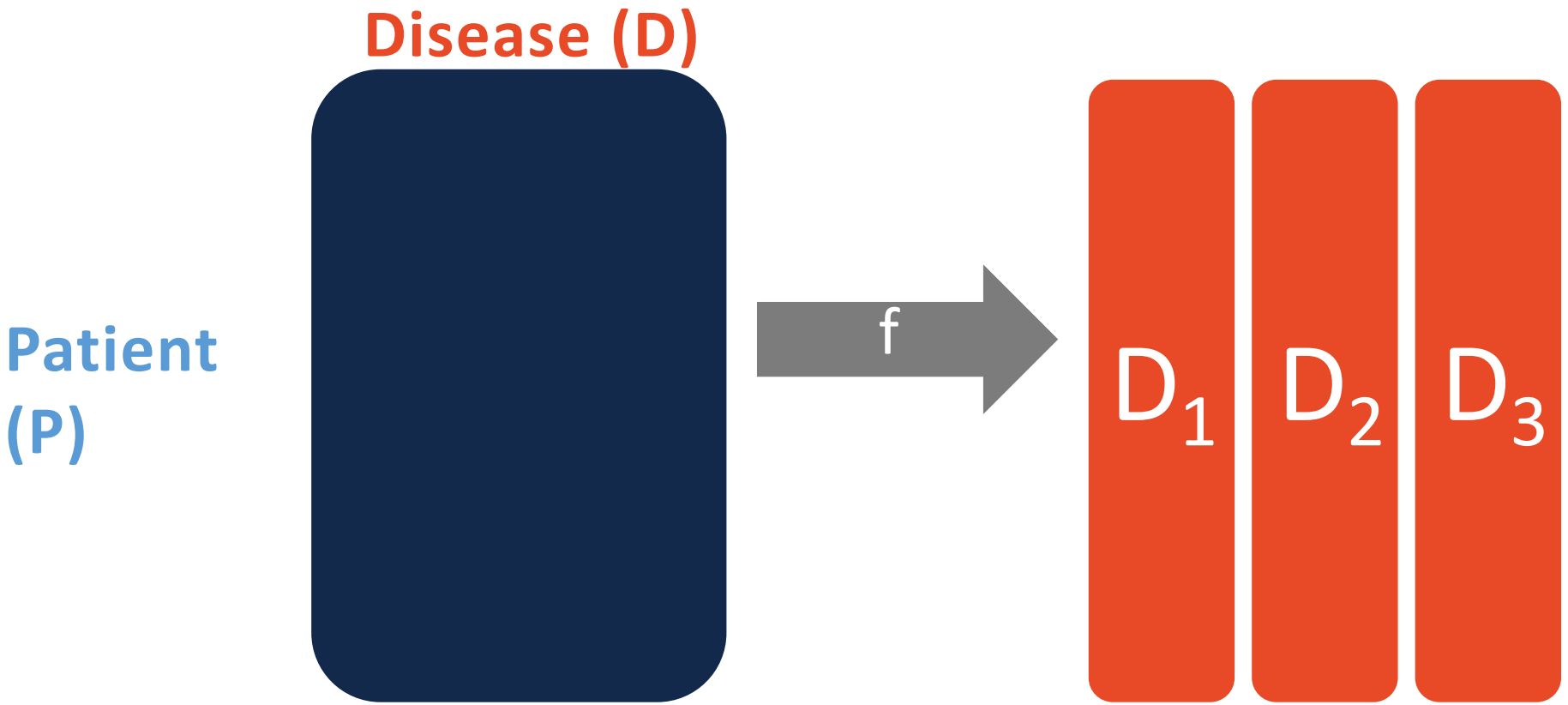




ILLINOIS

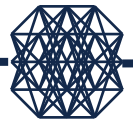
What is Clustering?





ILLINOIS

K-means



INPUT

Data points $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$,
# of clusters **k**

OUTPUT

k clusters $S_1, S_2, \ldots, S_k$

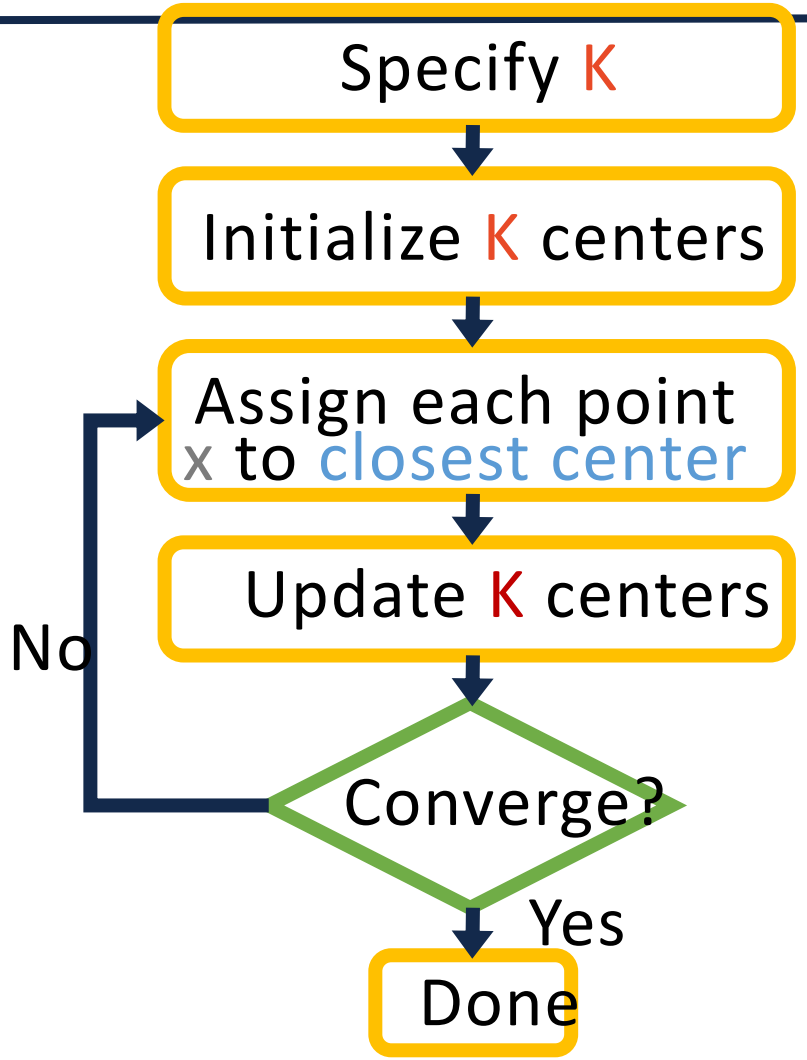
Objective:

$$min \sum_{i=1}^{k} \sum_{x \in Si} \|x - \mu_i\|^2$$

μ_i is the center in S_i

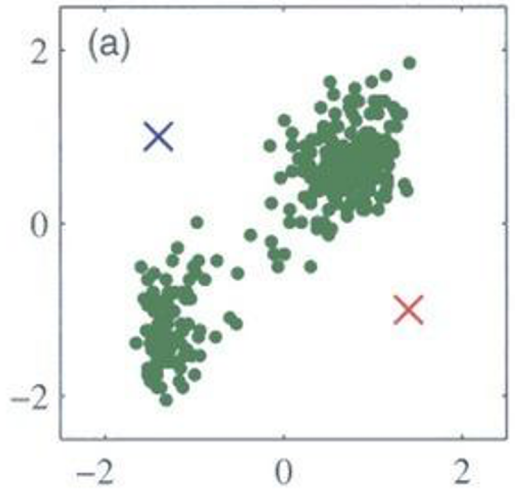
ILLINOIS

K-means Algorithm



$$\mu_i = \frac{1}{|Si|} \sum_{x \in Si} x$$

K-means Example



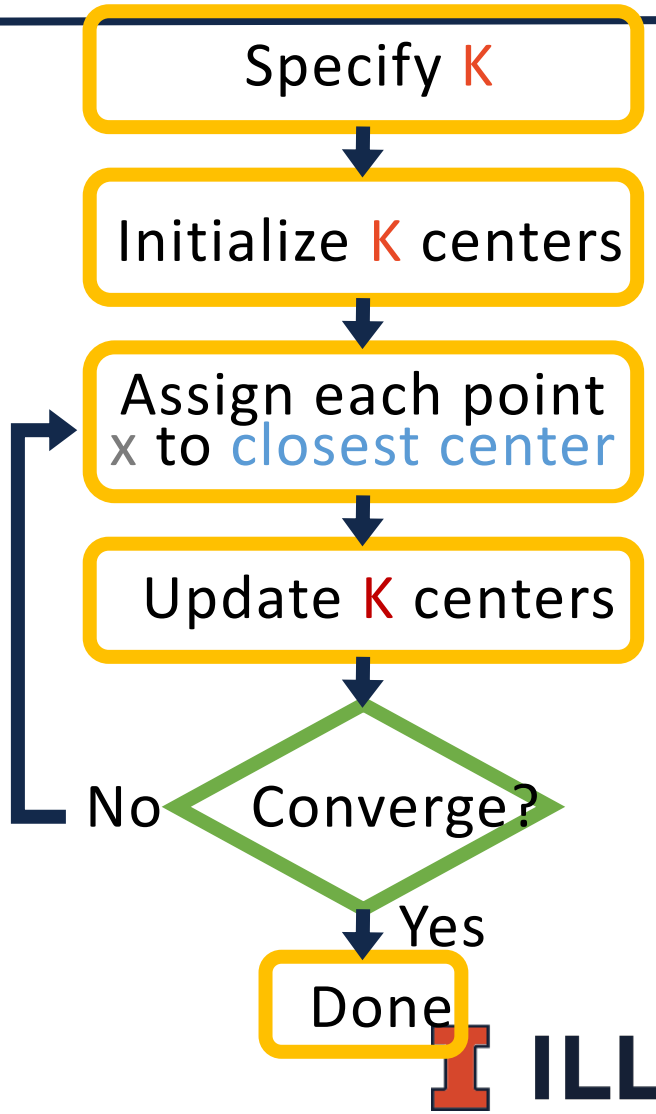
K-means Quiz

Given

- n: # of points
- k: clusters
- d: dimensionality of each point
- i: number of iterations

What is the computational complexity?

O(n•k•d•i)



ILLINOIS