t-SNE & UMAP

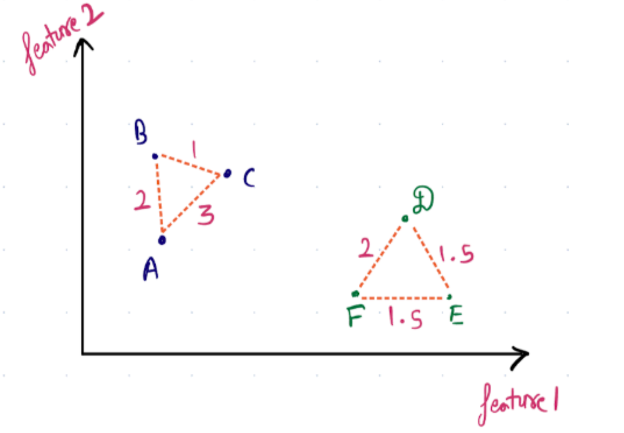
Agenda :

→ t-SNE

→ maths

→ UMAP intro.

Recap

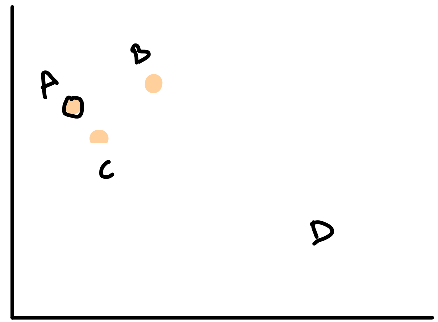


$\xrightarrow{1D}$

→ local neighbourhood

→ relative distances

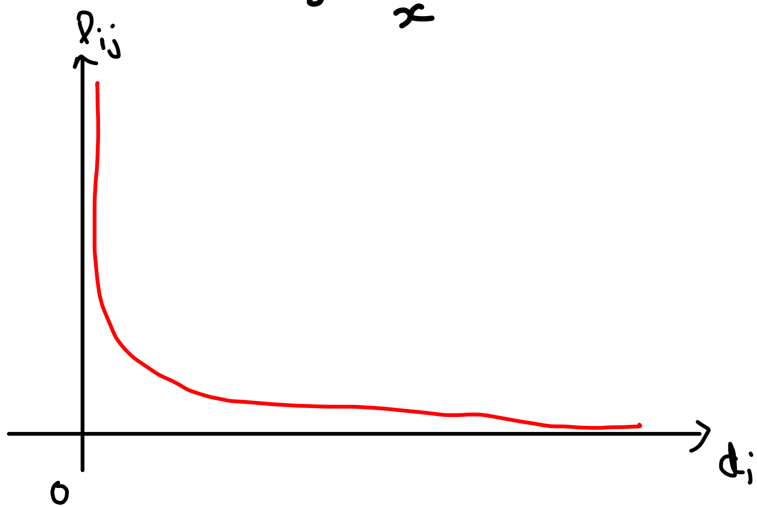
$$\left[\ \underset{\text{High Dim}^n}{\text{Probability}_N(i,j)} \approx \underset{\text{Low Dim}^n}{\text{Probability}_N(i,j)}\ \right]$$



$$P_{ij} \propto \frac{1}{dist(i,j)}$$

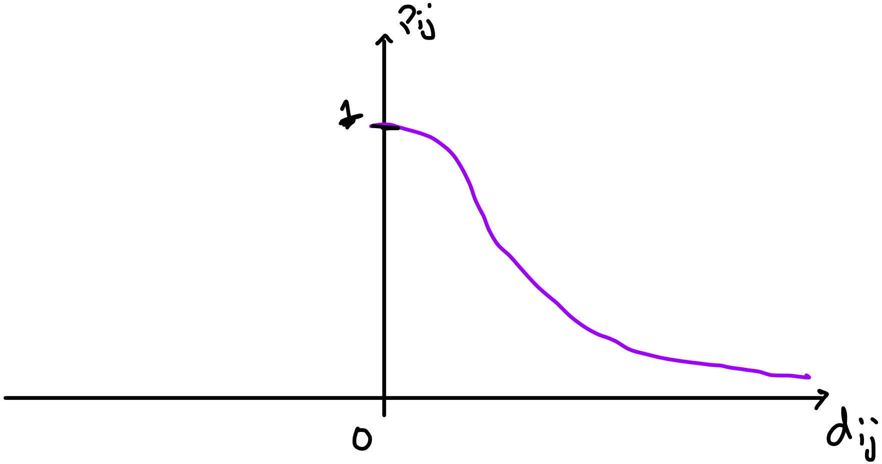
$$P_{ij} = \frac{1}{||x_i - x_j||} = \frac{1}{d_{ij}}$$

$$y = \frac{1}{x}$$

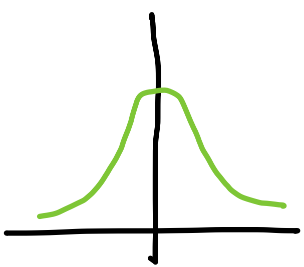


Disadvantage

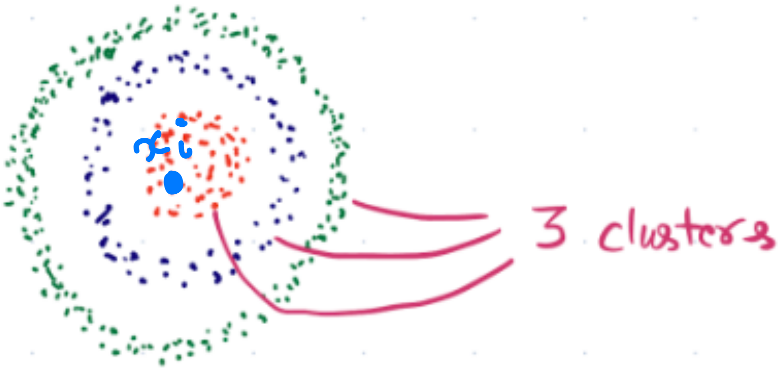
1. dis small $P \to \infty$
2. Prob falling sharply
3. later Doesn't consider relative Probabilities

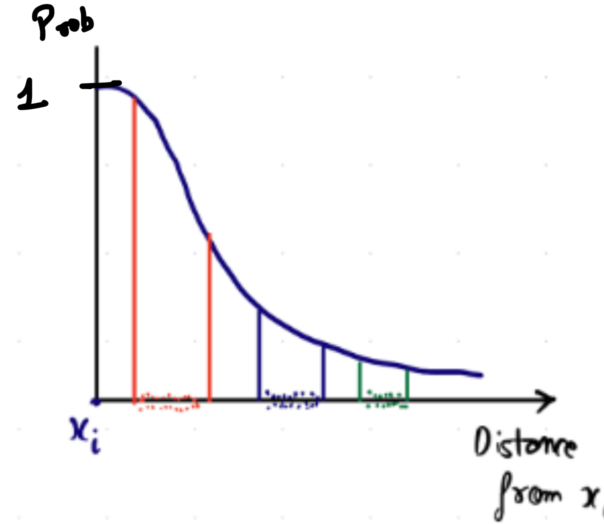


$$p_{ij} = e^{-\frac{\overbrace{\|x_i - x_j\|}^{d_{ij}}{}^2}{2\sigma_i^2}}$$



$$P(x) = \cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} \; e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$





x_i

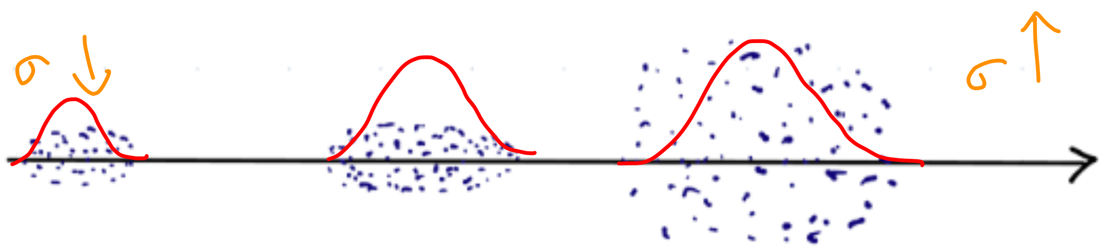
$$P_{i1} + P_{i2} + P_{i3} + \dots + P_{in} = 1 \quad \cancel{?} \quad \text{No.}$$

2, 5, 9, 1, 3

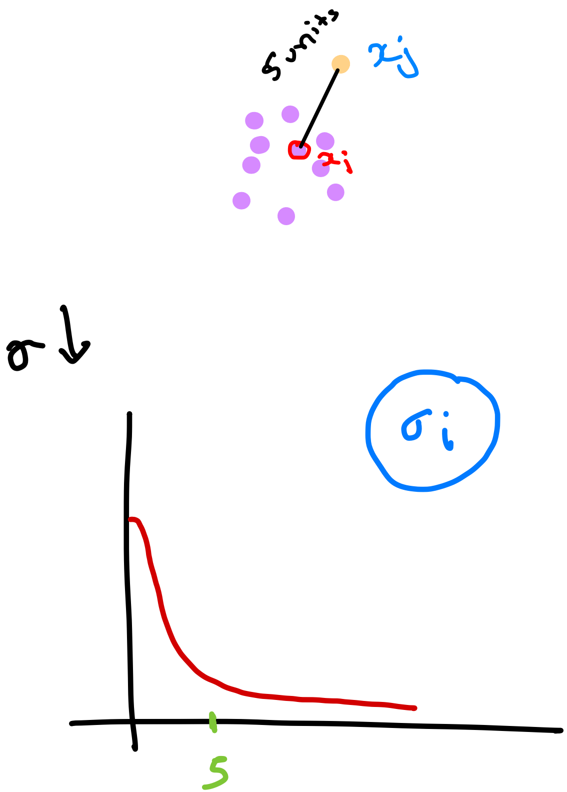
$$\frac{2}{20} = 0.1, \quad \frac{5}{20} = 0.25, \quad \frac{9}{20} = 0.48, \quad \frac{1}{20} = 0.05, \quad \frac{3}{25} =$$

$$p_{i,j} = \frac{e^{-\frac{\|x_i - x_j\|^2}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{\|x_i - x_k\|^2}{2\sigma_i^2}}}$$

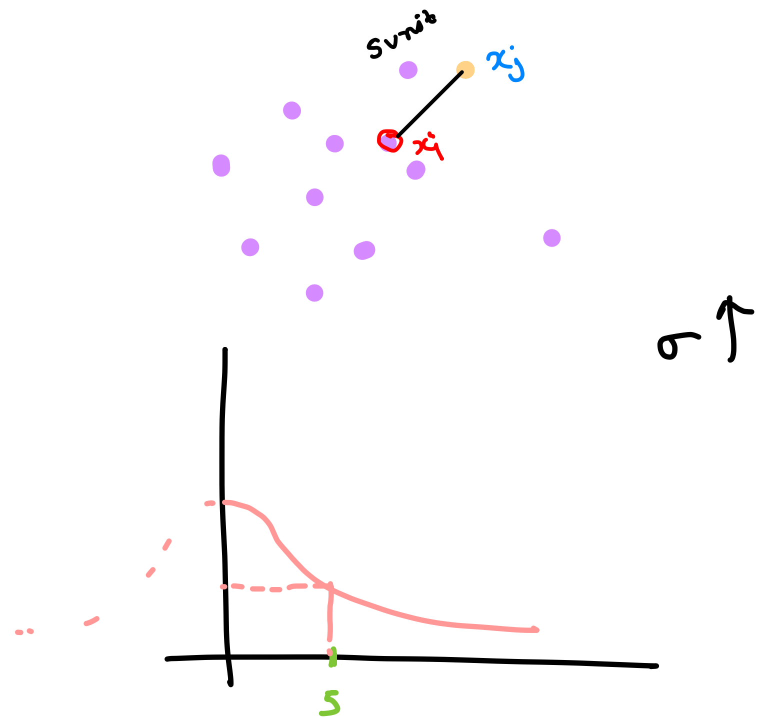
$\Rightarrow$ Calculate Prob. x_i & x_j being Neighs. in High Dimⁿ

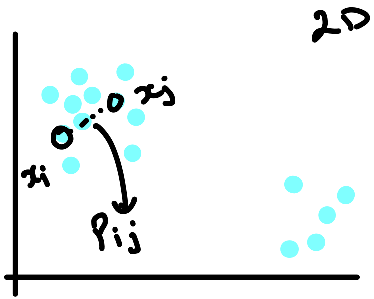


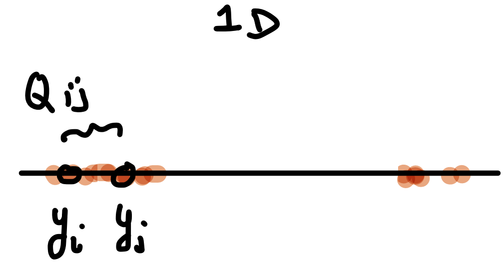
Case I



Case II





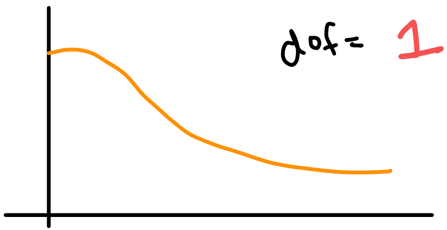


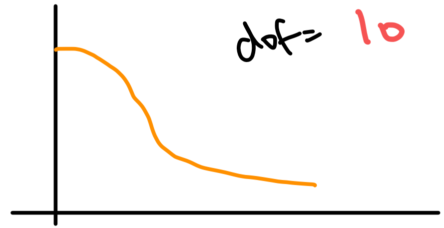
$Q_{ij} \rightarrow$ Prob. of 2 datapoints (y_i, y_j) $\rightarrow$ embedding of x_i & x_j being Neigh in Low Dim^n

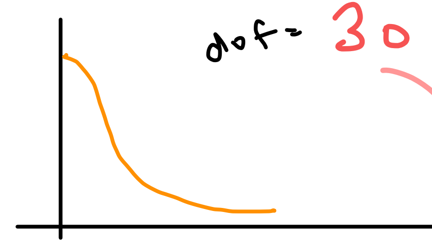
$$Q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \|y_i - y_k\|^2\right)^{-1}}$$

$$\frac{\frac{1}{1+d_{ij}^2}}{\sum_{k \neq i}\left(\frac{1}{1+d_{ij}^2}\right)} = t\text{-dist}_{\,dof=1}$$

Student t-dist → fatter tail distribution

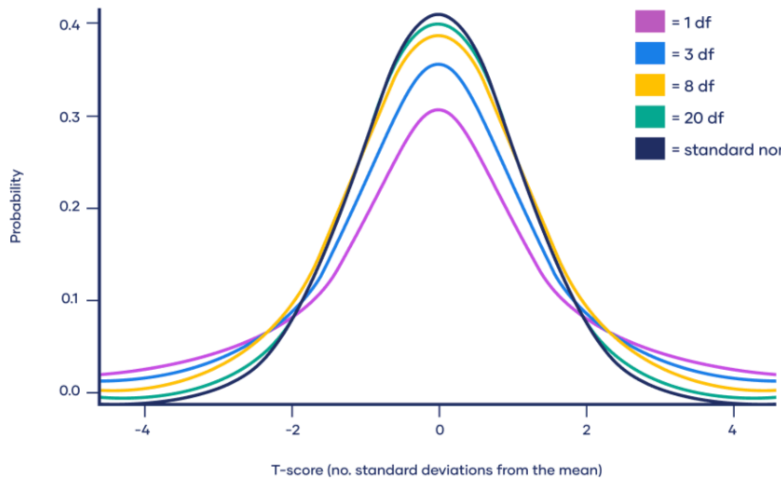






Normal dist.

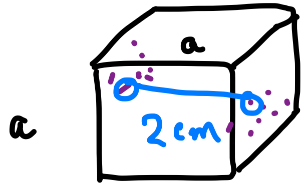
Cauchy Distribution



= 30

Why cauchy ?

(3D)



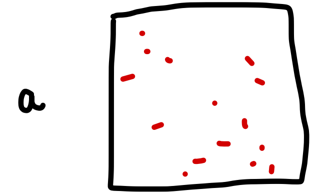
$a = 5$

$a^3 = 125$

(1m dataPoints)

transform $\longrightarrow$

$a = 5$



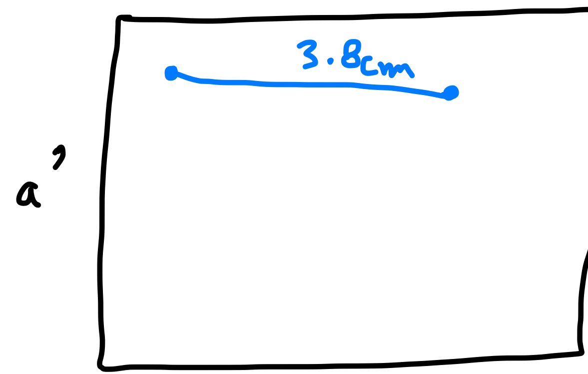
$a^2 = 25$

→ less space available

"Crowding Problem"

↓ strech

a'

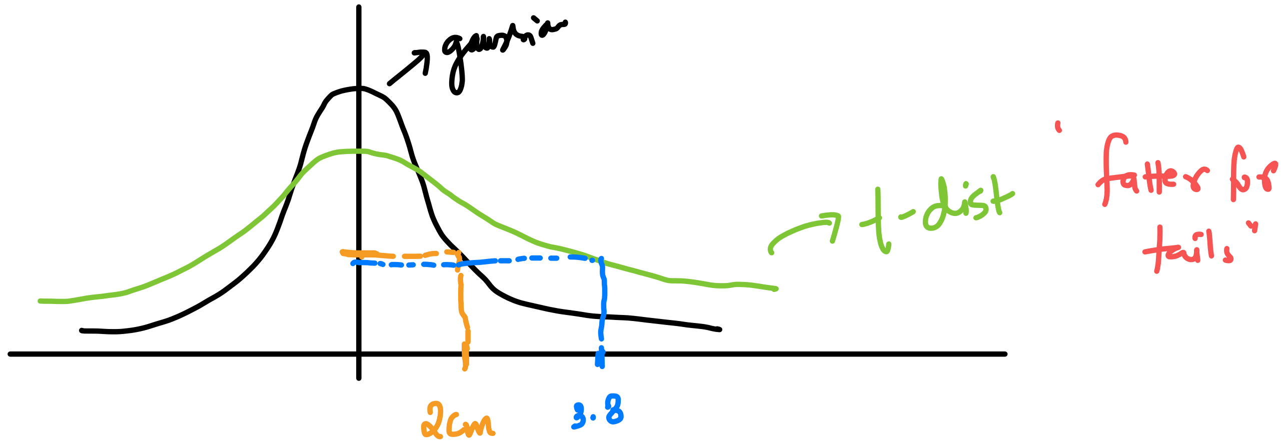


More space available

$$dis(x_i, x_j) < dis(y_i, y_j)$$

High Dim^n low Dim^n

$$P_{ij} \sim Q_{ij}$$



In t-SNE, why is the t-distribution preferred over the normal distribution for lower-dimensional data?

8 users have participated

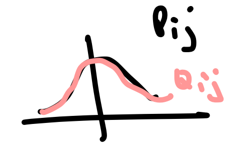
A	The t-distribution has a simpler mathematical form.	0%
B	The t-distribution helps reduce crowding by assigning heavier tails to the data.	75%
C	The t-distribution guarantees a uniform distribution of data points in lower-dimensional space.	25%
D	The normal distribution is not suitable for modeling lower-dimensional data.	0%

why?

End Quiz Now

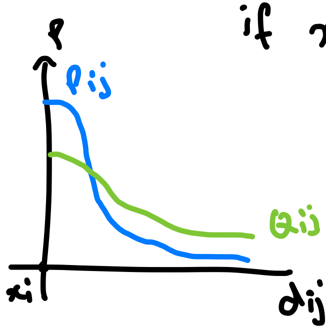
☐ $P_{ij} \sim Q_{ij}$. How to Measure similarity of 2 dist.?

New Loss $\Rightarrow$ KL Divergence



$$KL_{Dir}(P, Q) = \sum_{i}^{n} \sum_{j}^{n} P_{ij} \cdot \log\left(\frac{P_{ij}}{Q_{ij}}\right) \geq 0$$

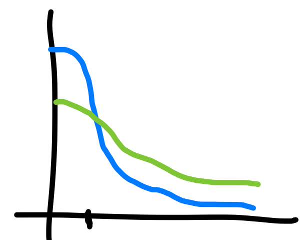
if x_j is far



$P_{ij} \sim 0$

$KL = 0$

x_j is close



$P_{ij} > Q_{ij}$

$$\frac{P_{ij}}{Q_{ij}} > 1$$

$\log\left(\frac{P}{Q}\right) > \log(1)$

~ 0

Limitations?

look → 1 hour

1. Super slow

learnable params = n

'n-embeddings'

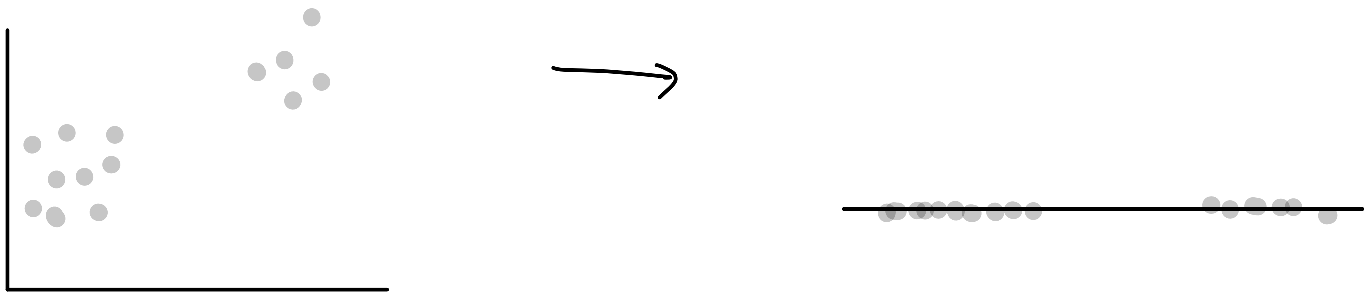
2. Hyper-param tuning

UMAP [2020]

→ Uniform manifold Approximation & Projection

tSNE → UMAP

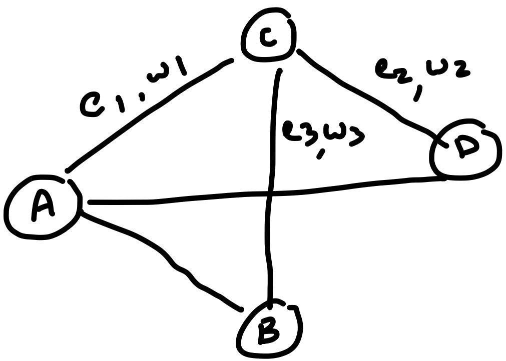
Algo:



1. n-neighbours (x_i) → Hyper param
2. Build a graph (x_i)

   → Nodes & edges

   e.g - Google maps

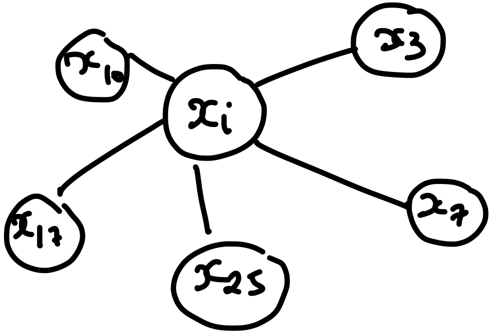


Nodes : landmarks, places

edges : road

weights : traffic

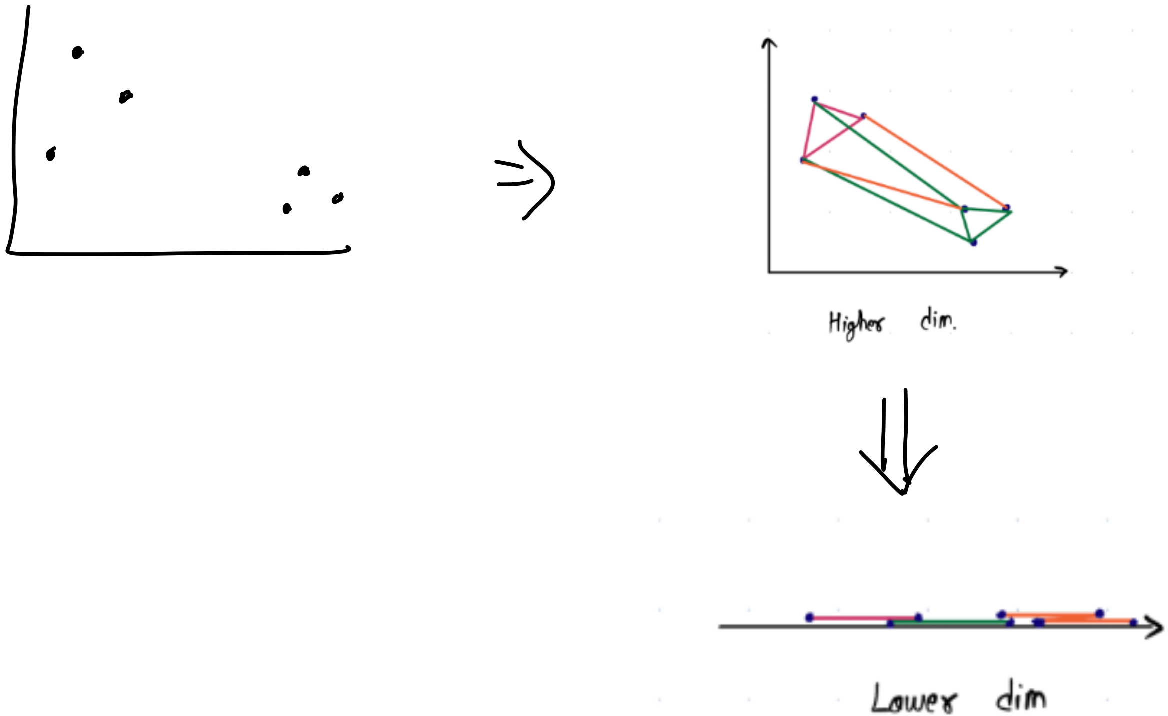
(n=5 neighbours)



③ We want $\boxed{\text{Graph}(x_i) \approx \text{Graph}(y_i)}$ $\forall_{i:1\to N}$

Graph(x_i) → High Dim^n

Graph(y_i) → Low Dim^n


Higher dim.
Lower dim

PCA → global variance / structure

t-SNE → local structure

UMAP → local + global