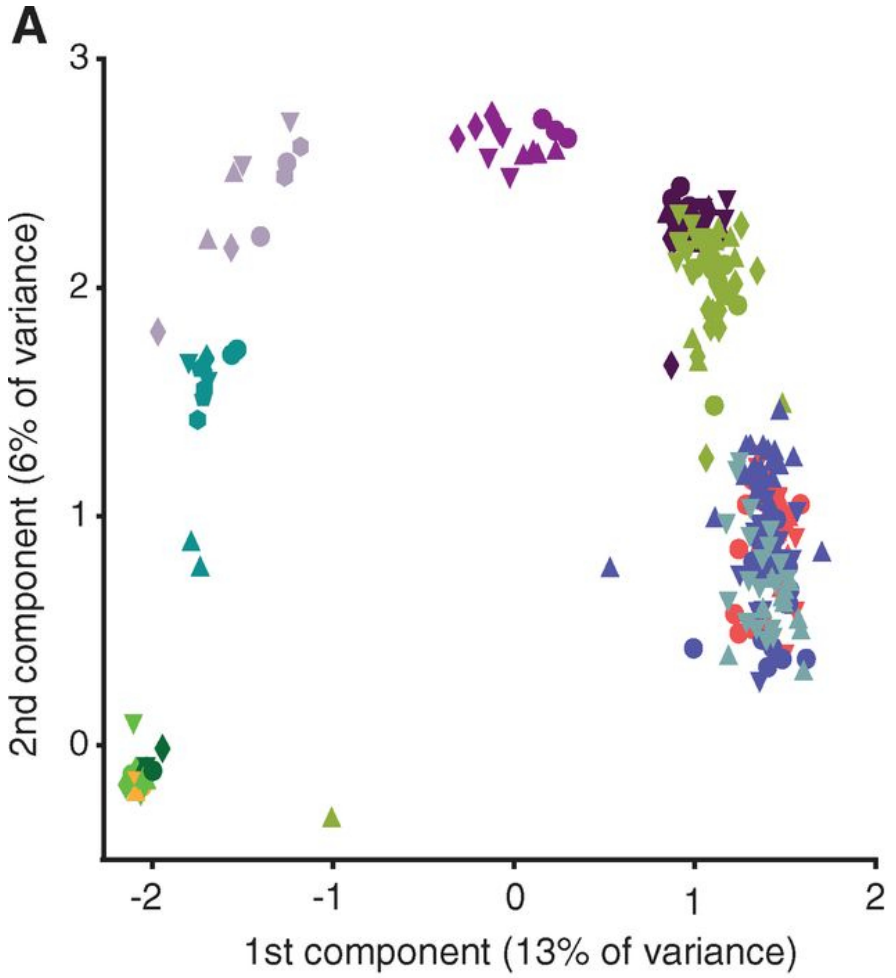
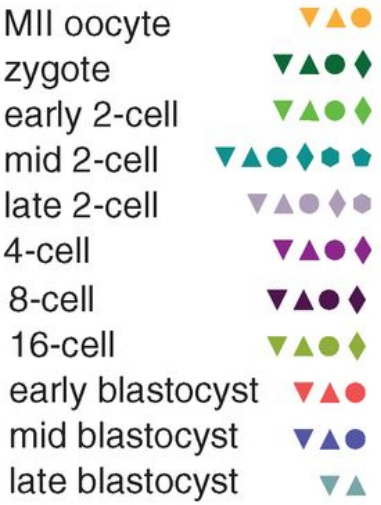
PCA and clustering

Principal Component Analysis (PCA)

- Dimension reduction technique
- Idea is to preserve most of the variation when reduced to lower dimensions.
- Does not work well with the data
  - having non linear relationship with the variables
  - having low variation

PCA (*Cont.*)





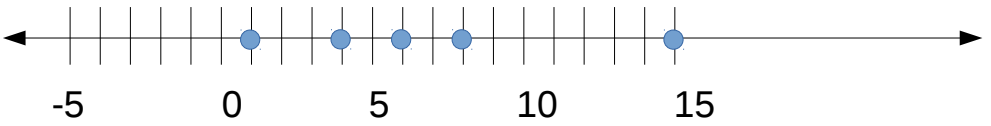
How does expression of thousand of genes reduced to a single dot (cell)?

Fig: Single-cell gene expression profiles projected onto the first two principal components. Cells from different stages and embryos are designated by colors and symbols (Deng et. al 2014).

PCA (*Cont.*)

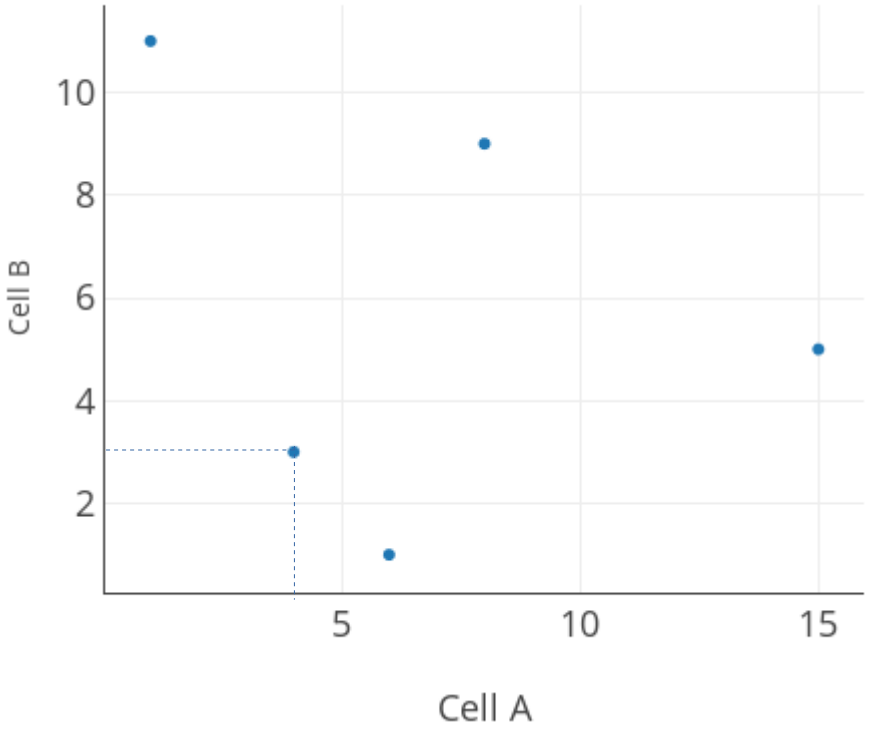
- Little bit about dimension.

Gene	Cell A
g1	4
g2	6
g3	8
g4	1
g5	15



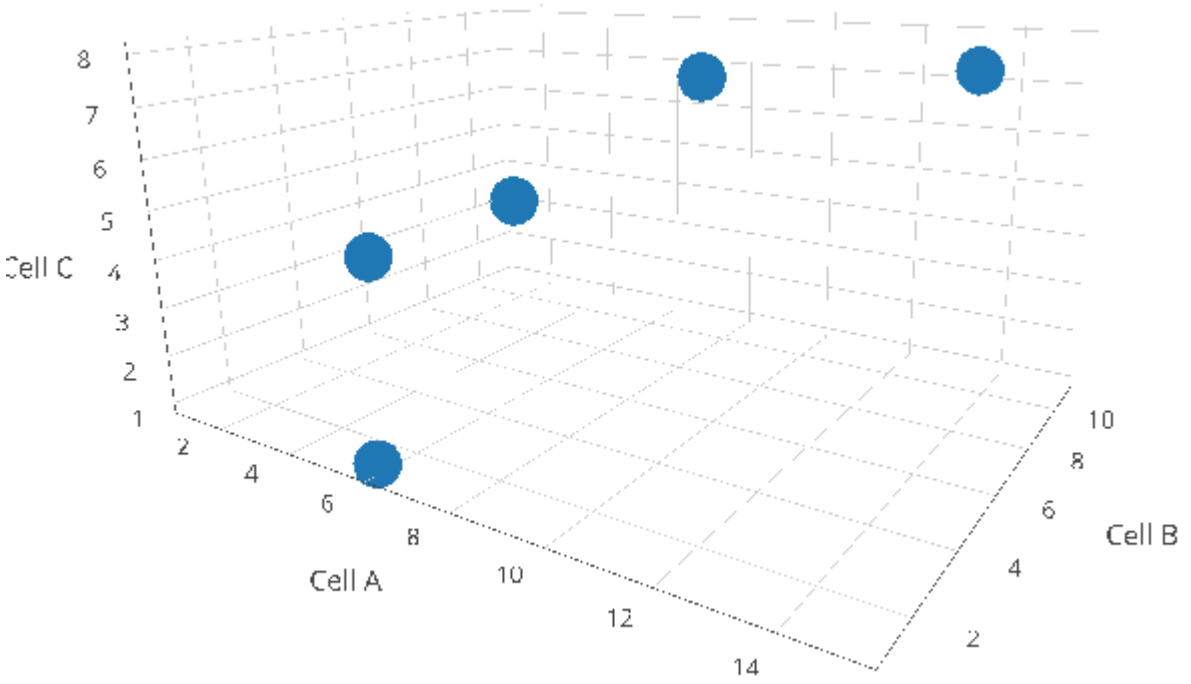
PCA (*Cont.*)

Gene	Cell A	Cell B
g1	4	3
g2	6	1
g3	8	9
g4	1	11
g5	15	5



PCA (*Cont.*)

Gene	Cell A	Cell B	Cell C
g1	4	3	4
g2	6	1	1
g3	8	9	7
g4	1	11	3
g5	15	5	8



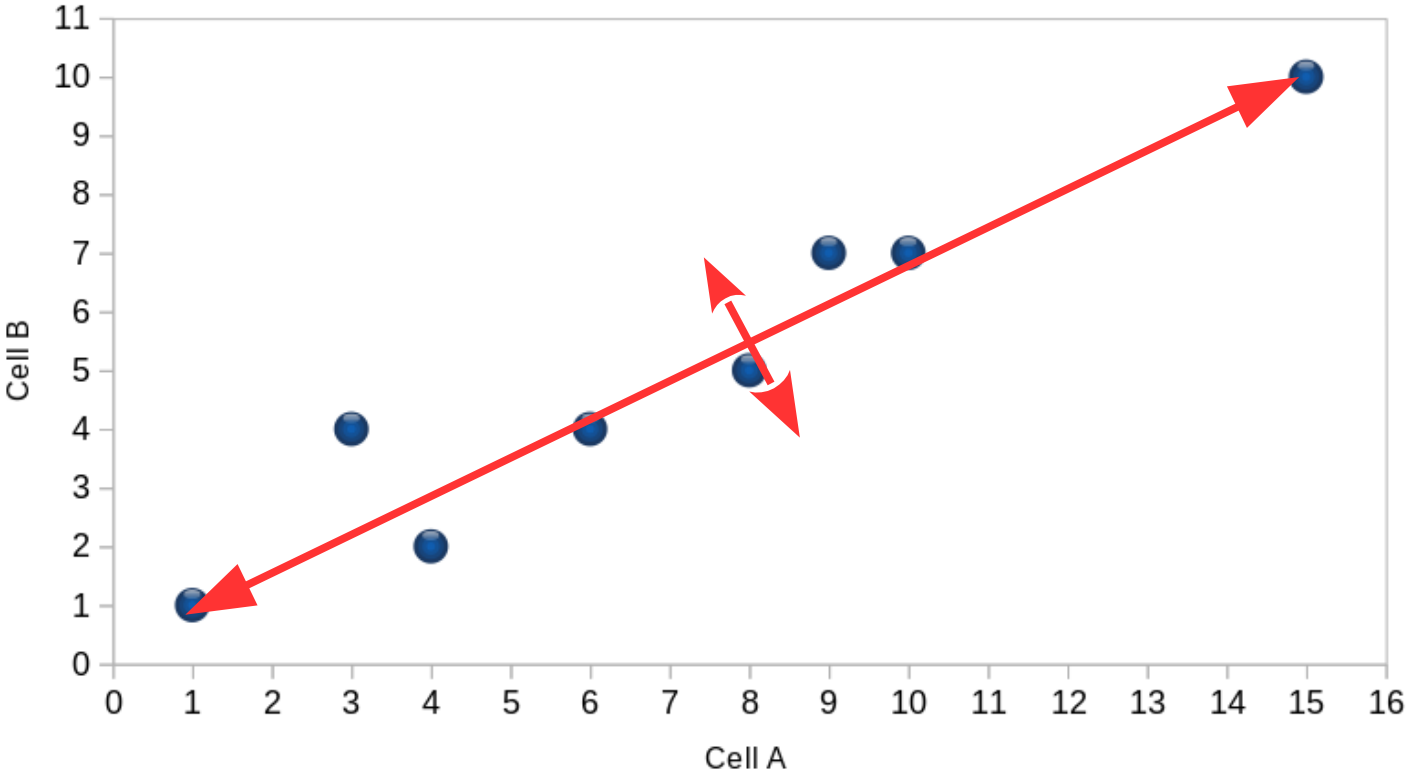
PCA (*Cont.*)

Gene	Cell A	Cell B	Cell C	Cell D
g1	4	3	4	2
g2	6	1	1	1
g3	8	9	7	0
g4	1	11	3	4
g5	15	5	8	9

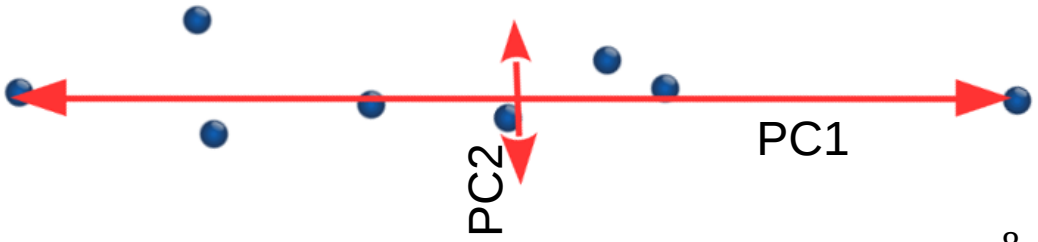
How does it look?
I have even 2000 cells.

PCA (*Cont.*)

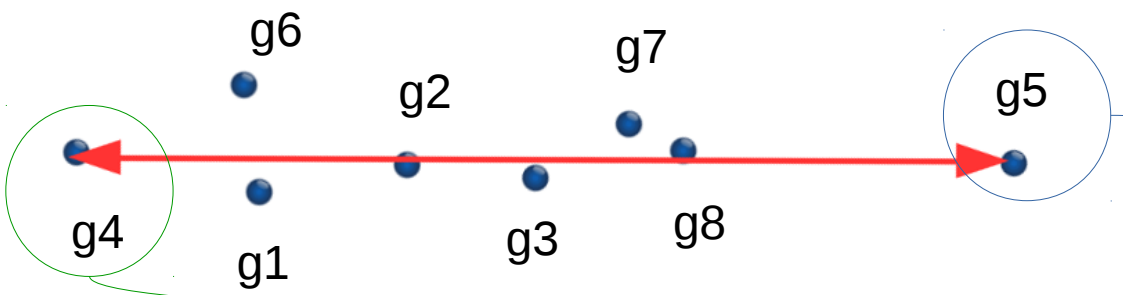
Gene	Cell A	Cell B
g1	4	2
g2	6	4
g3	8	5
g4	1	1
g5	15	10
g6	3	4
g7	9	7
g8	10	7



- There is principal component for each dimension.
- **BUT** here we are plotting genes not cells.



PCA (*Cont.*)



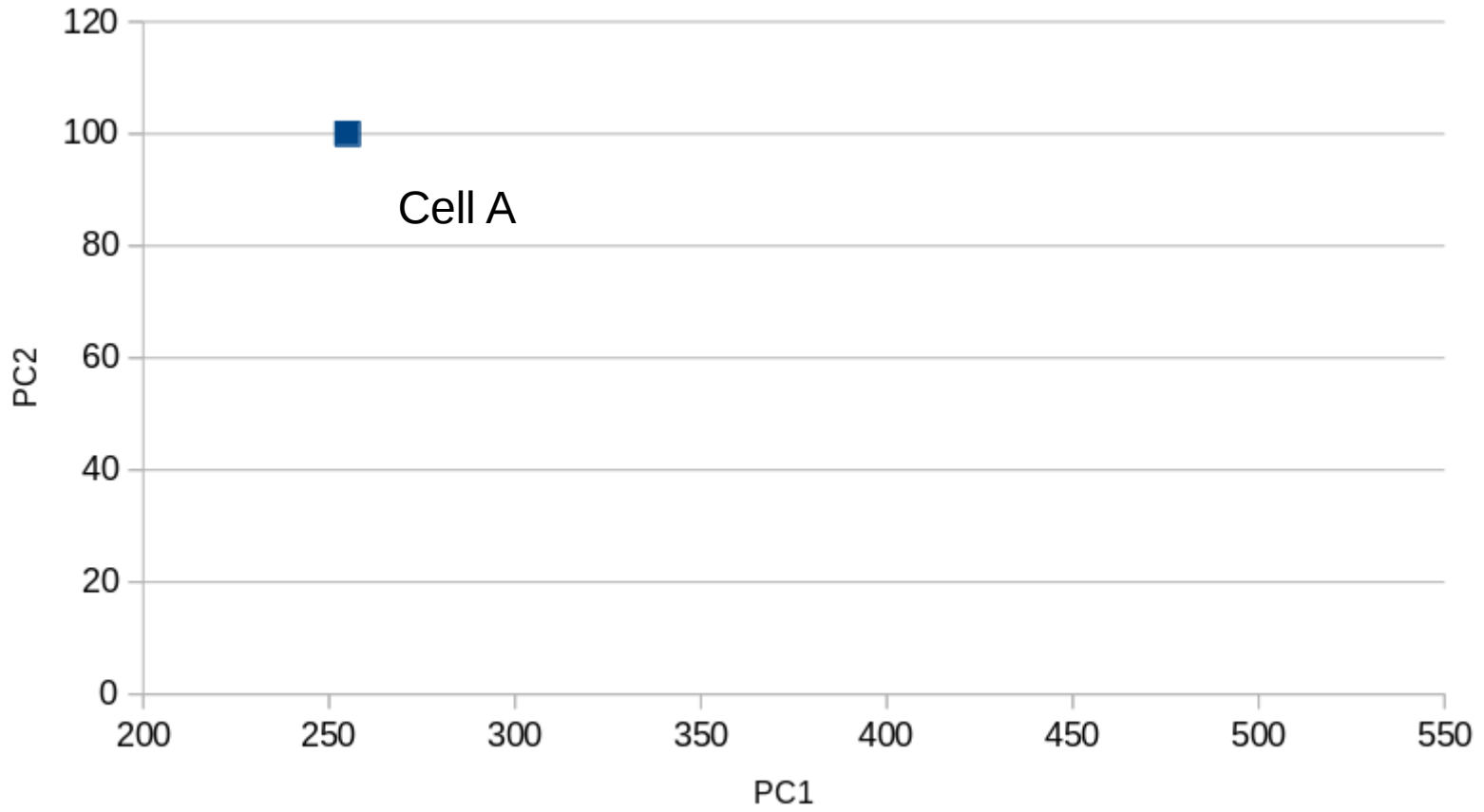
Gene	Influence in PC1	In numbers
g1	high	-9
g2	low	4
g3	low	0
g4	high	-14
g5	high	15
g6	high	-10
g7	low	4
g8	low	5

Original data

Gene	Cell A	Cell B
g1	4	2
g2	6	4
g3	8	5
g4	1	1
g5	15	10
g6	3	4
g7	9	7
g8	10	7

PC1 score = (4 * -9) + (6 * 4) …… = 255
PC2 score = (4 * influence of g1 in PC2)
+ (6 * influence of g2 in PC2) …
= 50 (lets say)

PCA (*Cont.*)


120
100
80
60
40
20
0
PC2
Cell A
200
250
300
350
400
450
500
550
PC1

PCA (*Cont.*)

Raw data

PCA reduced data

Fig: PCA analogy

Graph based clustering

Table: Expression matrix toy data.

	Cell A	Cell B	Cell C
Gene A	4	1	5
Gene B	3	4	1
Gene C	10	0	2
Gene D	6	1	7

Euclidean distance $d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$

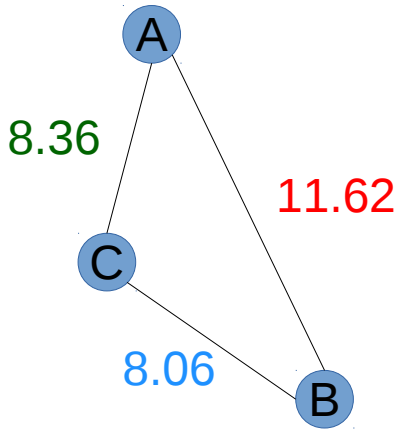


Fig: Graph

AB = $\sqrt{(4-1)^2 + (3-4)^2 + (10-0)^2 + (6-1)^2}$ = 11.62
AC = 8.36
BC = 8.06

Graph based clustering (*Cont.*)

- Shared Nearest Neighbor Graph (SNN)
  - Threshold parameter (τ) = minimum shared neighbors

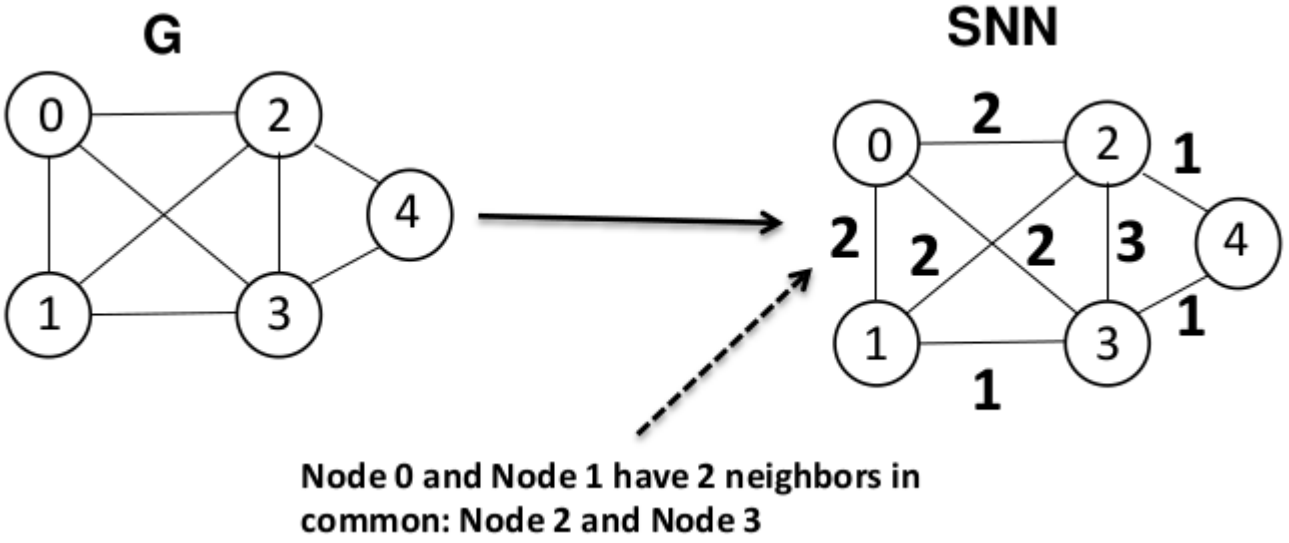


Fig: SNN (from Mahmud S.)

Graph based clustering (*Cont.*)

- We remove edges having weight less than τ (tau).
- Nodes connected by edges are in same cluster.

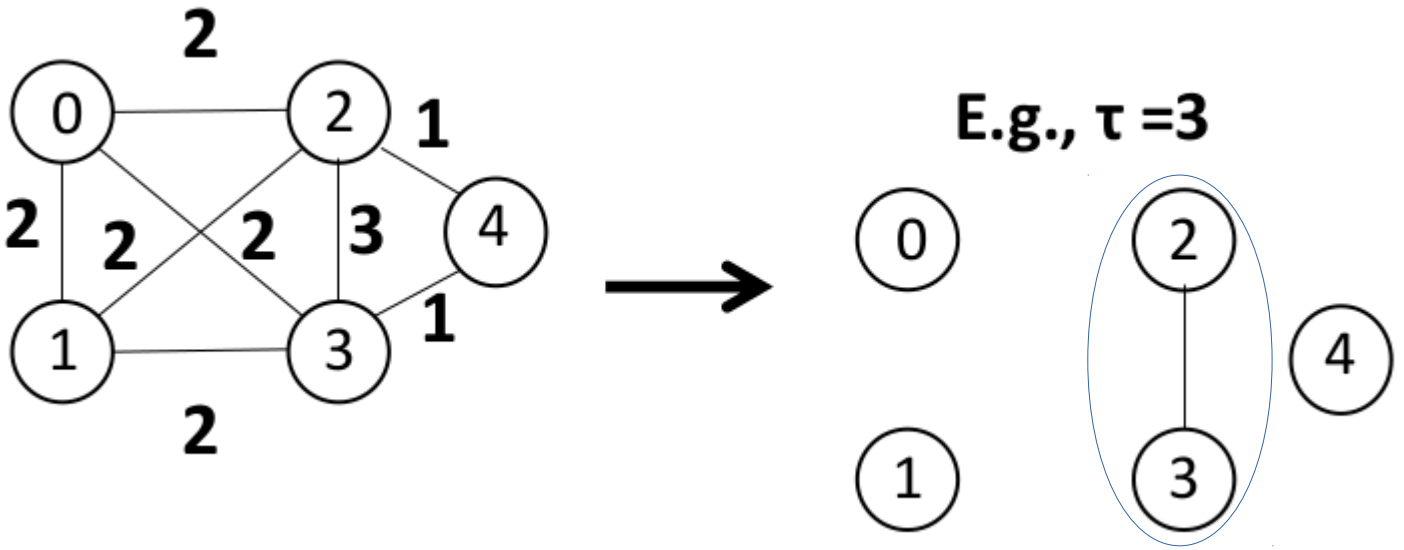


Fig: SNN (from Mahmud S.)

Visualization

- tSNE
  - Dimensional reduction technique
  - Suits for high dimensional data
  - Takes into account non-linear relationship

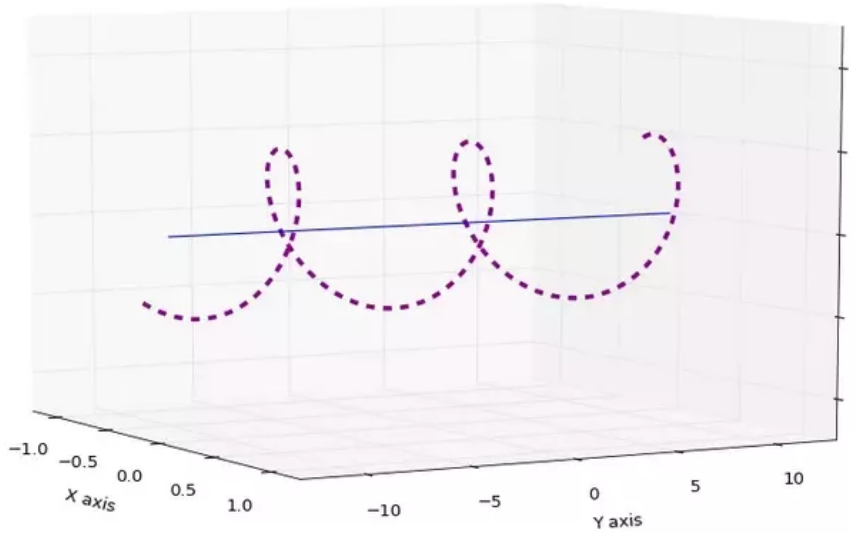


Fig: PCA vs. tSNE (from *quora.com*)

tSNE

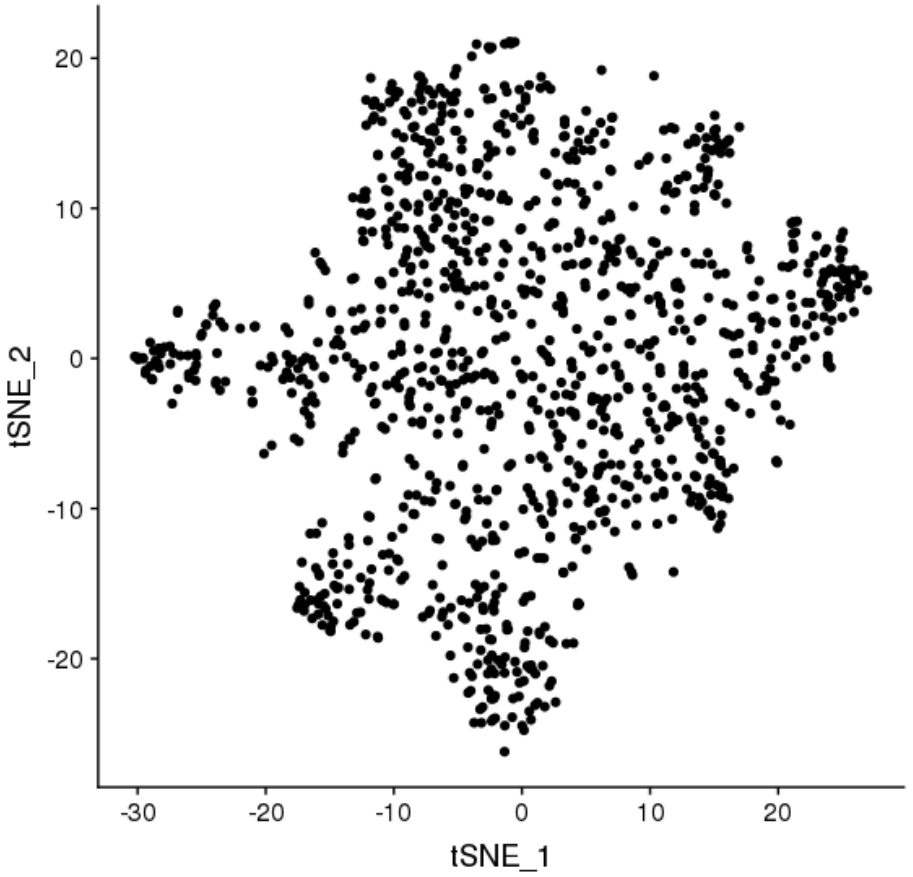


Fig: tSNE plot

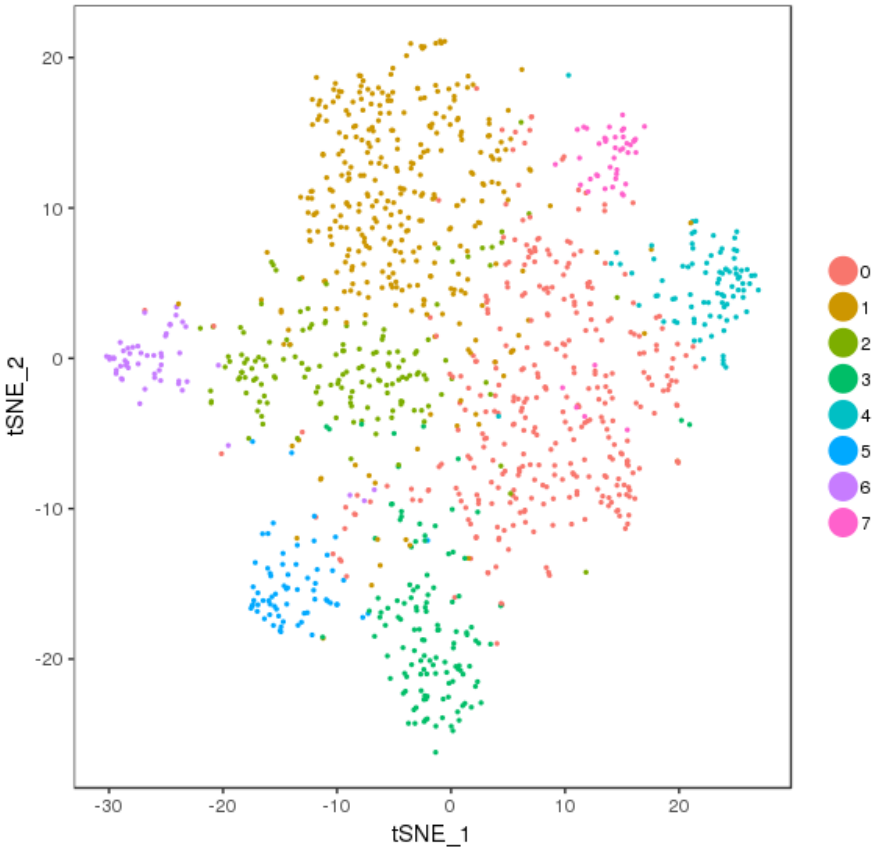


Fig: tSNE plot with cluster information