Artificial Intelligence in Medicine

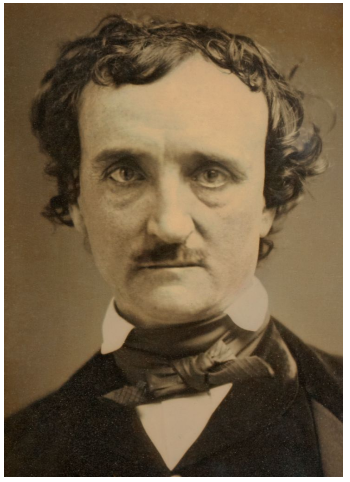
Visualizations and Dimensionality Reduction

Nir Friedman and Tommy Kaplan 12/12/22

האוניברסיטה
העברית
בירושלים
THE HEBREW
UNIVERSITY
OF JERUSALEM

"Believe half of what you see…"

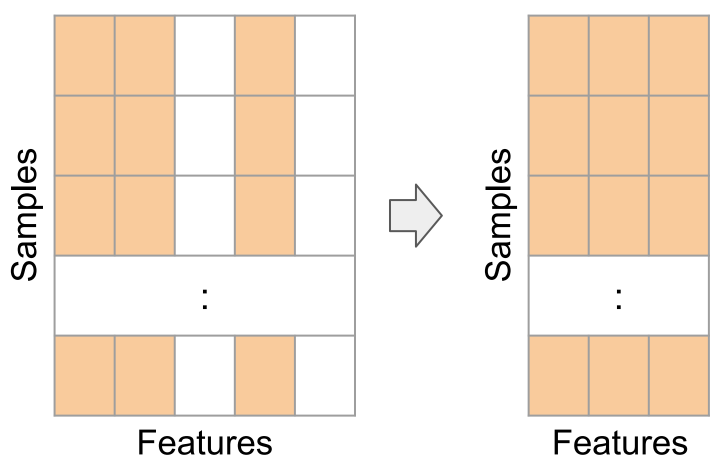
Edgar Allan Poe



Why dimensionality reduction?

- Represent data with fewer dimensions
- Remove irrelevant data features

$x \in \mathcal{R}^n$



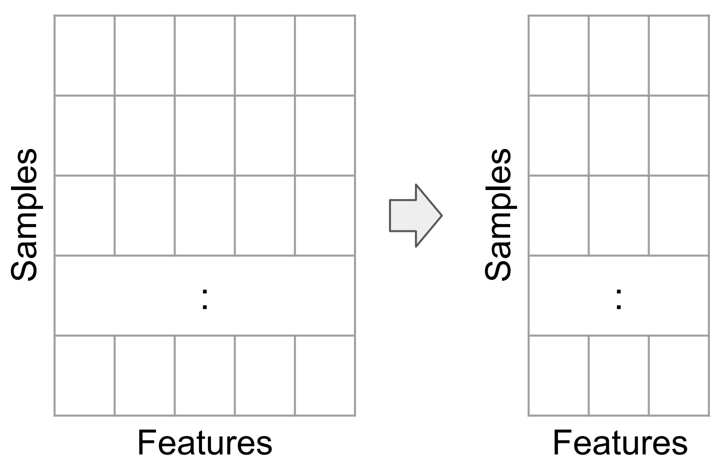
$x \in \mathcal{R}^d$ $\quad d \ll n$

Why dimensionality reduction?

- Represent data with fewer dimensions

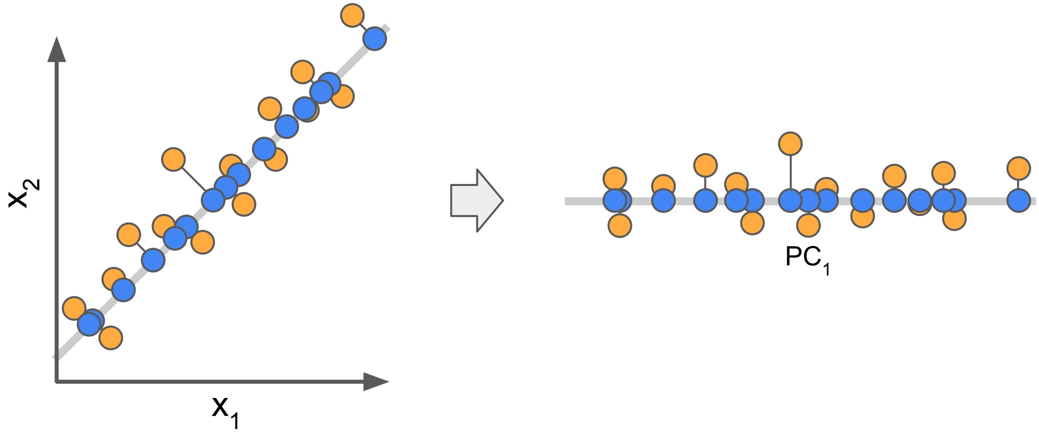
- Remove irrelevant data features
- Merge correlated/coordinated measures
- Reduce noise
- Avoid over-fitting (the Curse of Dimensionality)
- Cluster
- Find trajectory of disease, continuous state

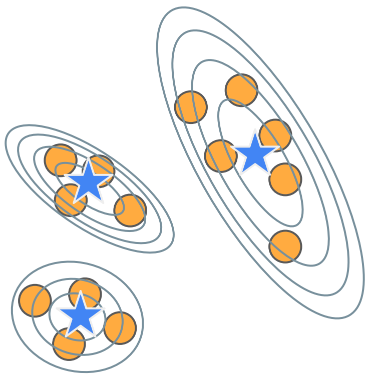
$$x \in \mathcal{R}^n$$



$$x \in \mathcal{R}^d \qquad d \ll n$$

Why dimensionality reduction?

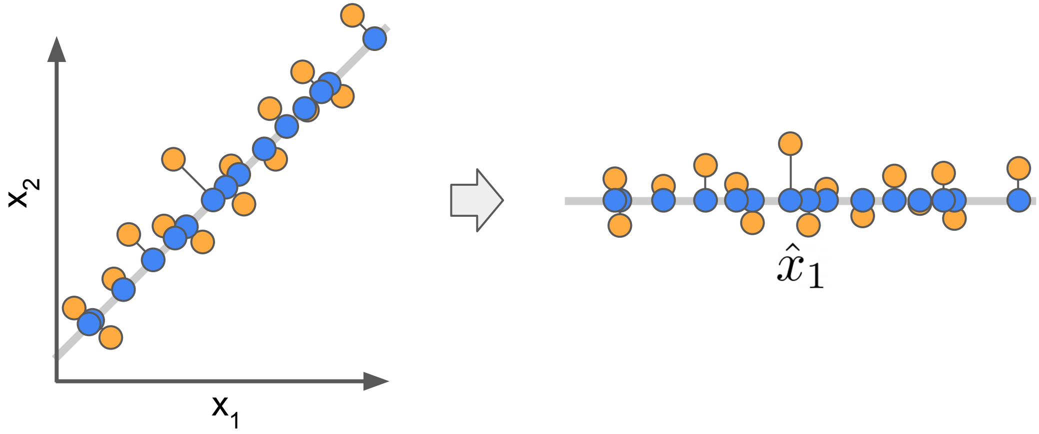




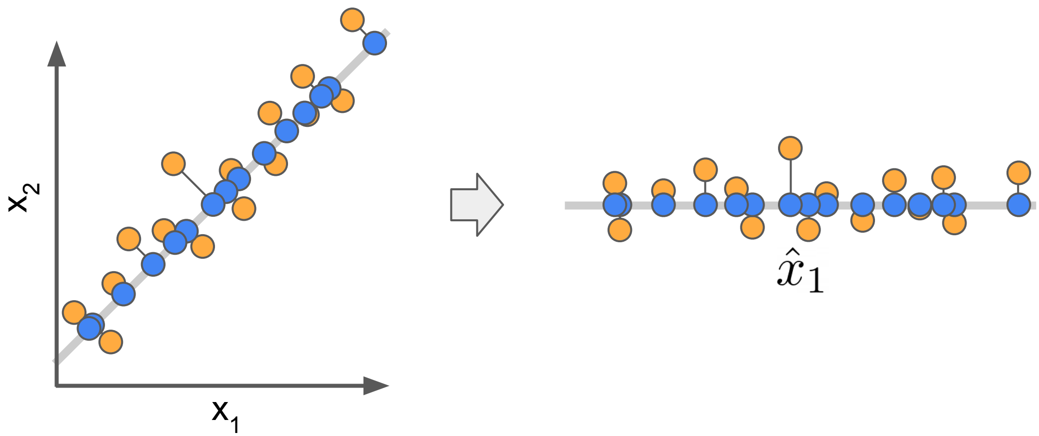
Linear projections

- Use a linear function to transform data to new space

$$\hat{x}_1 = w_1 x_1 + w_2 x_2$$



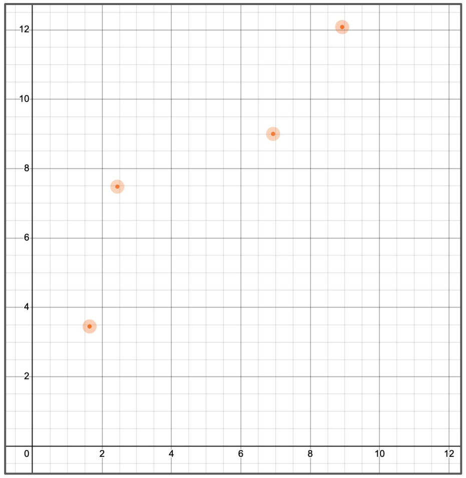
Which projection is "optimal"?

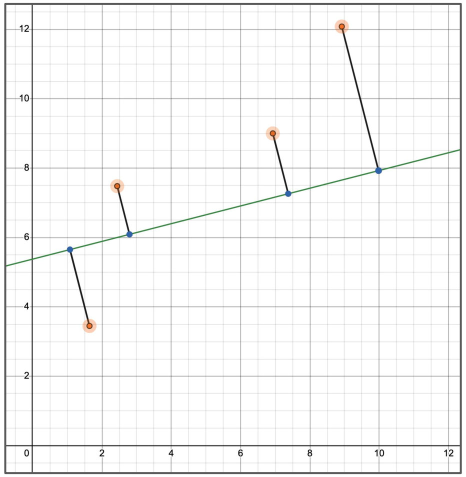


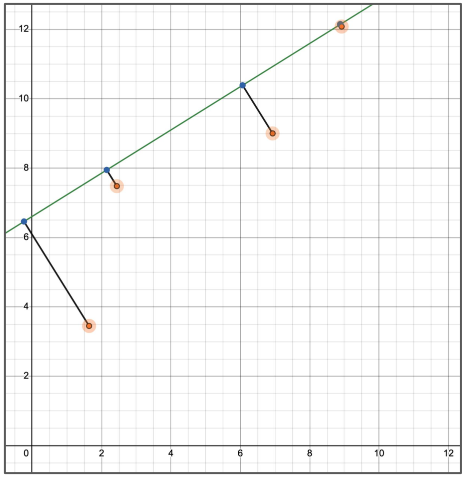
- Low-dimensional representation often introduces errors
- Aim for the most "accurate" new representation

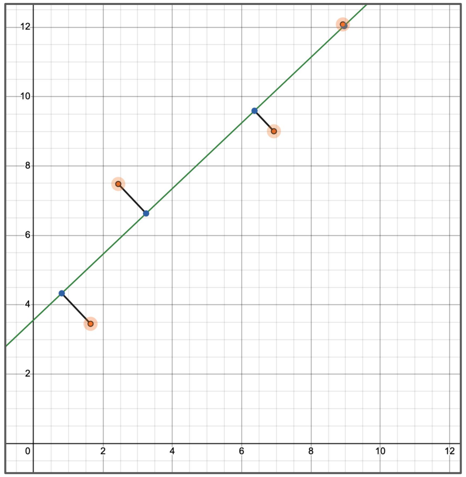
$$\arg\min \sum_i \|x_i - \hat{x}_i\|_2^2$$

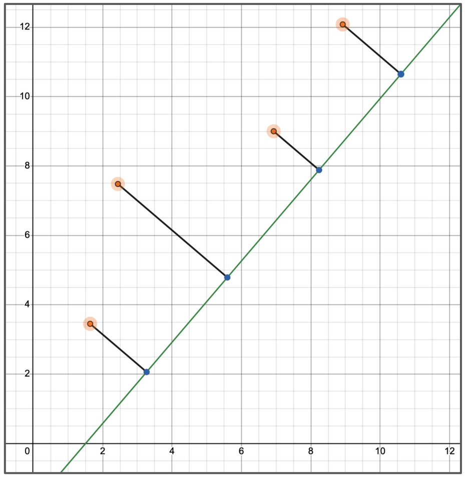
Find the axis (PC) that minimizes the error





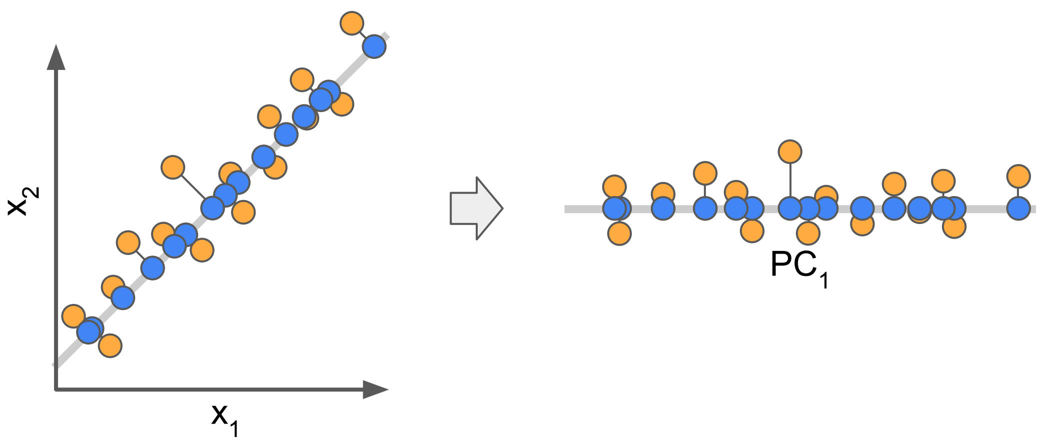






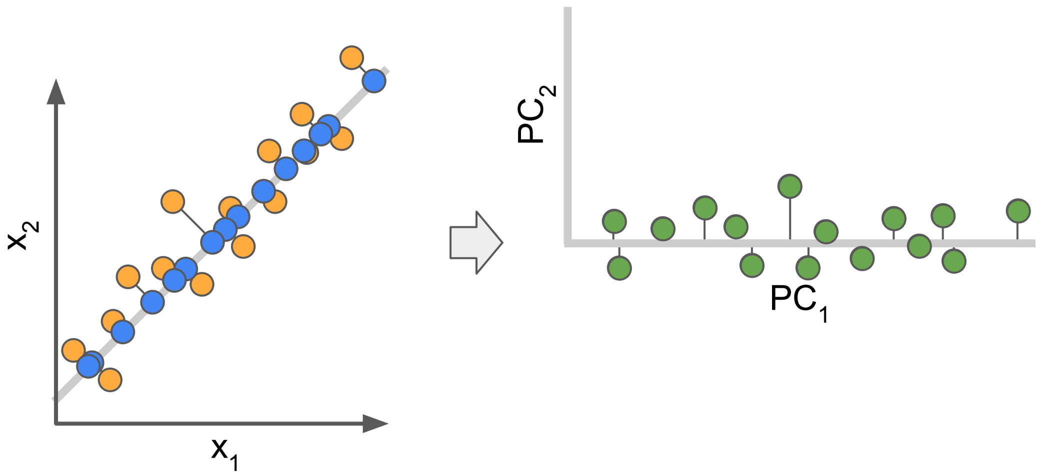
PCA - principal component analysis

- PC_1 is the direction that minimizes the errors on a 1-dimensional projection
- ⇒ PC_1 maximizes the variance
- ⇒ More information is conserved
- ⇒ **The “optimal” 1D projection**



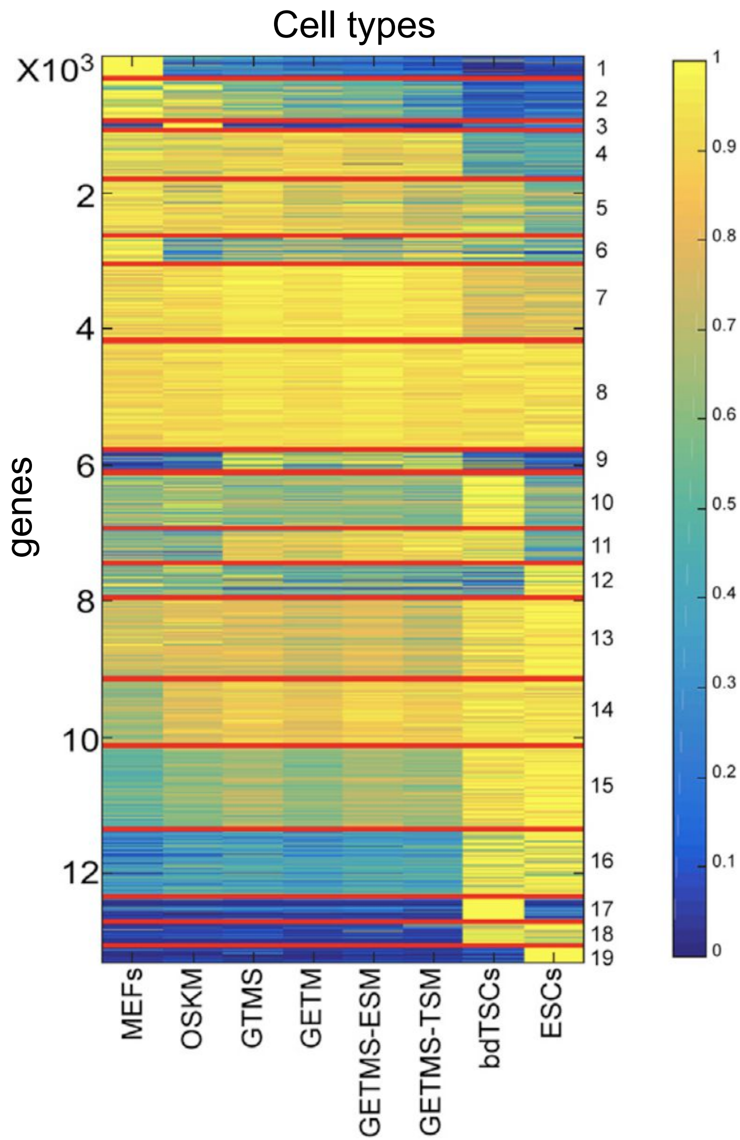
PCA - principal component analysis

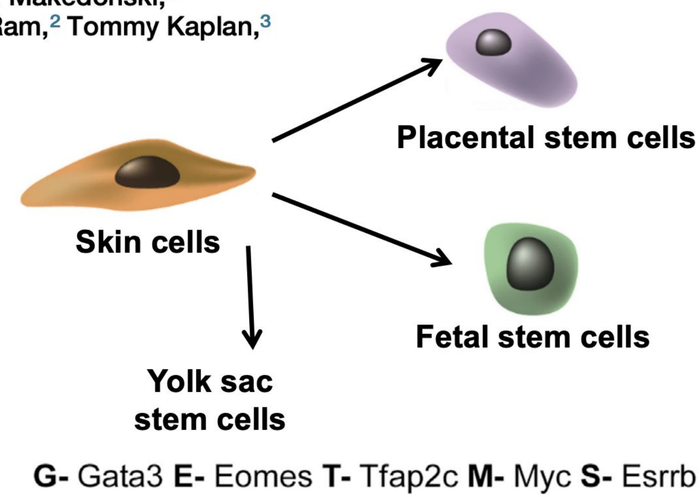
- PC_1 is the direction that minimizes the errors on a 1-dimensional projection
- PC_2 minimizes the remaining errors



Direct Induction of the Three Pre-implantation Blastocyst Cell Types from Fibroblasts

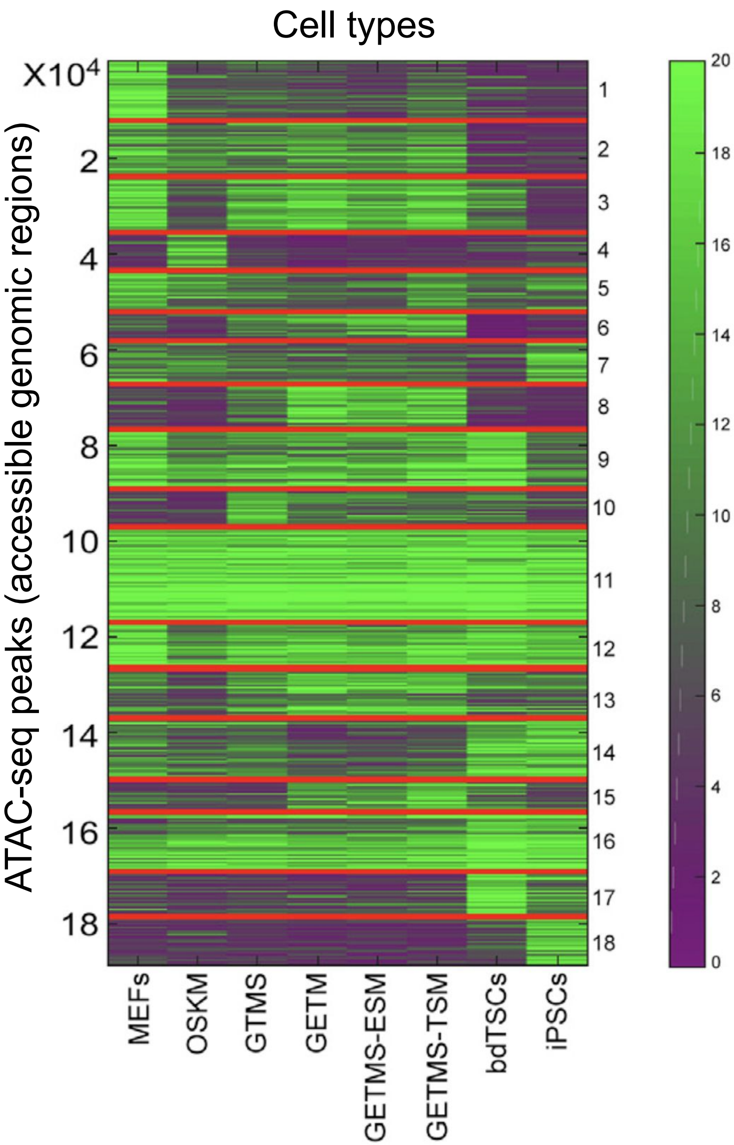
Hana Benchetrit,[1,4] Mohammad Jaber,[1,4] Valery Zayat,[1,4] Shulamit Sebban,[1] Avital Pushett,[1] Kirill Makedonski,[1] Zvi Zakheim,[1] Ahmed Radwan,[1] Noam Maoz,[1] Rachel Lasry,[1] Noa Renous,[1] Michal Inbar,[1] Oren Ram,[2] Tommy Kaplan,[3] and Yosef Buganim[1,5,*]

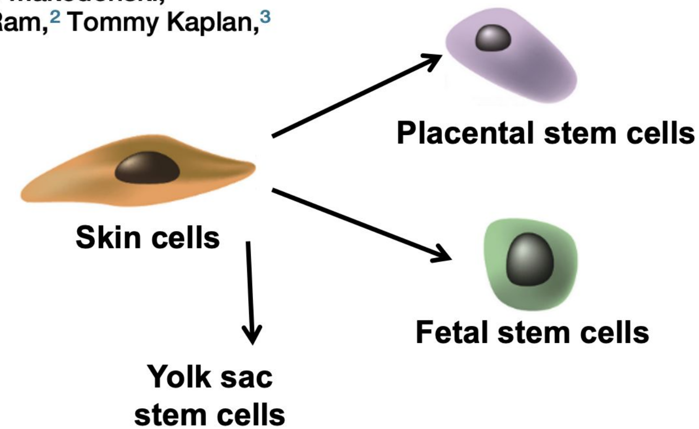




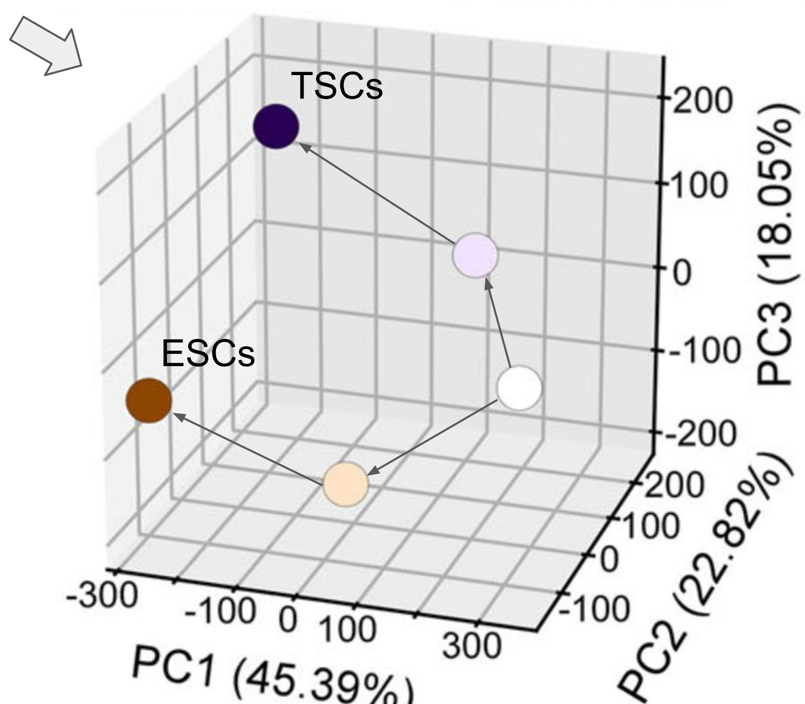
Direct Induction of the Three Pre-implantation Blastocyst Cell Types from Fibroblasts

Hana Benchetrit,[1,4] Mohammad Jaber,[1,4] Valery Zayat,[1,4] Shulamit Sebban,[1] Avital Pushett,[1] Kirill Makedonski,[1] Zvi Zakheim,[1] Ahmed Radwan,[1] Noam Maoz,[1] Rachel Lasry,[1] Noa Renous,[1] Michal Inbar,[1] Oren Ram,[2] Tommy Kaplan,[3] and Yosef Buganim[1,5,*]



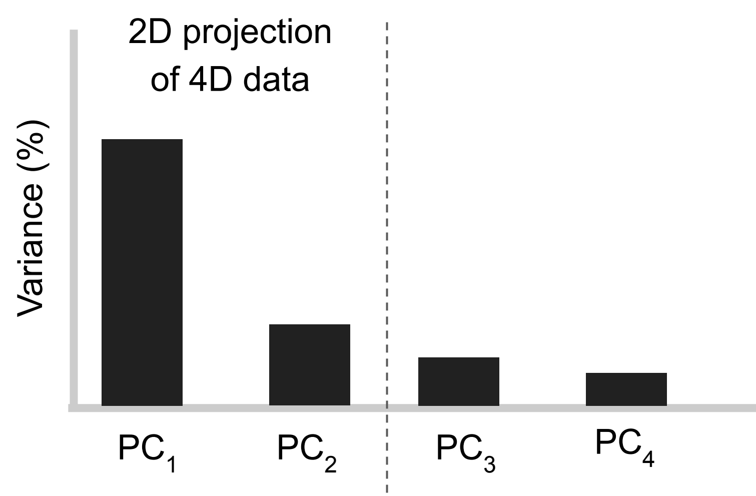


G- Gata3 **E-** Eomes **T-** Tfap2c **M-** Myc **S-** Esrrb



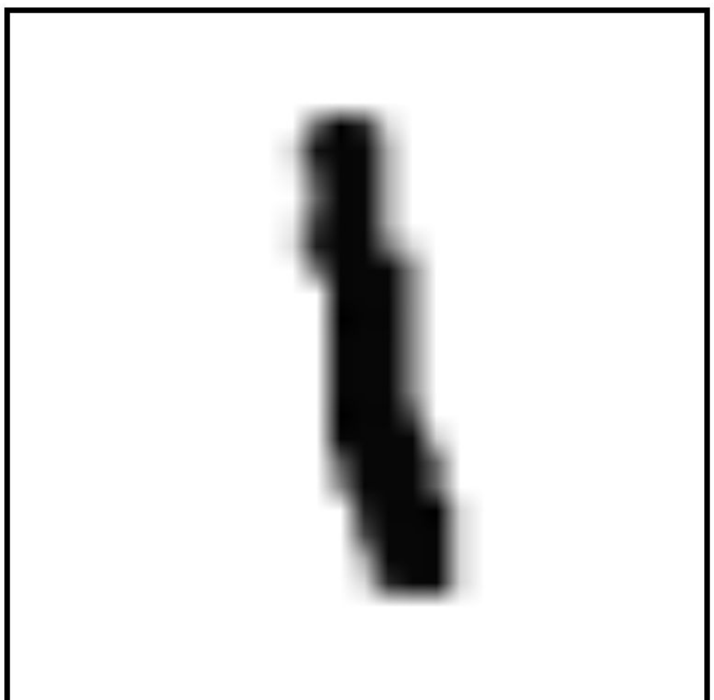
PCA - principal component analysis

- PC_1 is the direction that minimizes the errors on a 1-dimensional projection
- PC_2 minimizes the remaining errors
- PC_3 minimizes the remaining errors
- and so on…
- We then ignore components that are less important



PCA can be applied to different types of data

- Images are points in high-dimensional space
- A hand-drawn digit “1”, shown as a 14x14 matrix
- A point in a 196-dimensional space



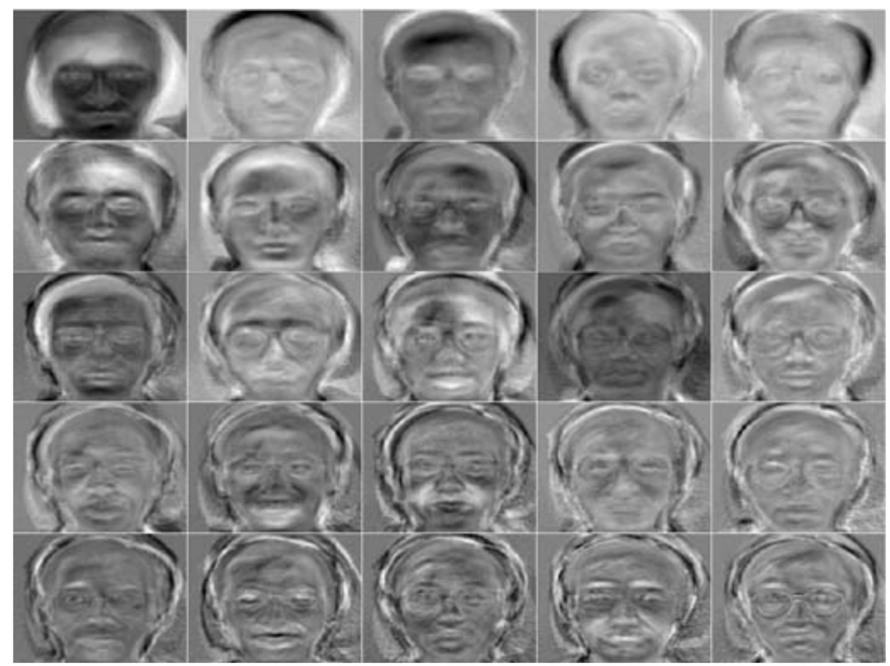
$$\simeq \begin{bmatrix}
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\
0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\
0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\
0&0&0&0&0&0&0&1&1&0&0&0&0&0\\
0&0&0&0&0&0&0&.9&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&.3&1&.1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0
\end{bmatrix}$$

Every value actually limited to one of 256 binary values encoded using 8 bits

PCA can be applied to different types of data

- Faces, for images
- Each picture is a 256x256 pixel image

Input faces



Principal components

Face Recognition Using Eigenfaces

Matthew A. Turk and Alex P. Pentland

Slide from David Sontag, NYU

PCA can be applied to different types of data

PC 1 | PC1 + PC2 | PCs 1-3 | PCs 1-4 | PCs 1-5

PCs 1 through 13

(13D projection of 256x256 data)

Face Recognition Using Eigenfaces

Matthew A. Turk and Alex P. Pentland

Slide from David Sontag, NYU

Similar methods applied to text analysis

- Each document viewed as a weighted list of words

Gene expression profiling predicts clinical outcome of breast cancer

Laura J. van 't Veer[*†], Hongyue Dai[†‡], Marc J. van de Vijver[*†], Yudong D. He[‡], Augustinus A. M. Hart[*], Mao Mao[‡], Hans L. Peterse[*], Karin van der Kooy[*], Matthew J. Marton[‡], Anke T. Witteveen[*], George J. Schreiber[‡], Ron M. Kerkhoven[*], Chris Roberts[‡], Peter S. Linsley[‡], René Bernards[*] & Stephen H. Friend[‡]

* *Divisions of Diagnostic Oncology, Radiotherapy and Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, 121 Plesmanlaan, 1066 CX Amsterdam, The Netherlands*
‡ *Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034,*

https://www.nature.com/articles/415530a

Word List

964 words in your word list

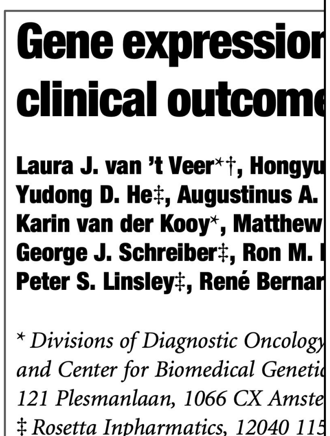
Frequency	Word
56	genes
54	tumours
46	patients
42	prognosis
32	expression
30	breast
30	er

☑ Remove Stop Words ☐ Remove Numbers ☑ Remove Special Characters

Latent Dirichlet Allocation, Blei, Ng, and Jordan (2003)

Similar methods applied to text analysis

- Each document viewed as a weighted list of words



Gene expressior
clinical outcome

Laura J. van 't Veer*†, Hongyu
Yudong D. He‡, Augustinus A.
Karin van der Kooy*, Matthew
George J. Schreiber‡, Ron M. I
Peter S. Linsley‡, René Bernar

** Divisions of Diagnostic Oncology*
and Center for Biomedical Geneti
121 Plesmanlaan, 1066 CX Amste
‡ Rosetta Inpharmatics, 12040 115

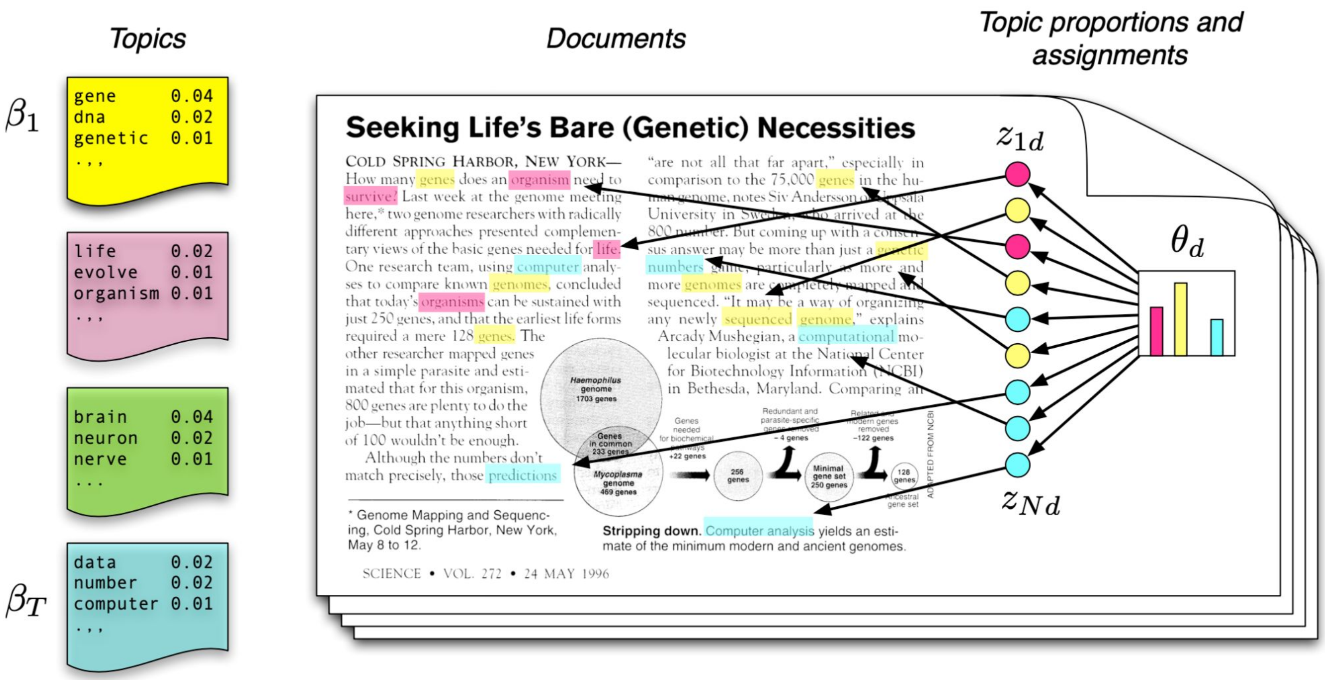
https://www.nature.con

Latent Dirichlet Allocation, Blei, Ng, and Jordan (2003)

Similar methods applied to text analysis

- Each document viewed as a weighted list of words
- Train on huge corpus, define typical “topics”

  *a bit different than PCA, as topics are not orthogonal, but idea is similar



Document seen as an N-dimensional datapoint

Axes = topics (types)

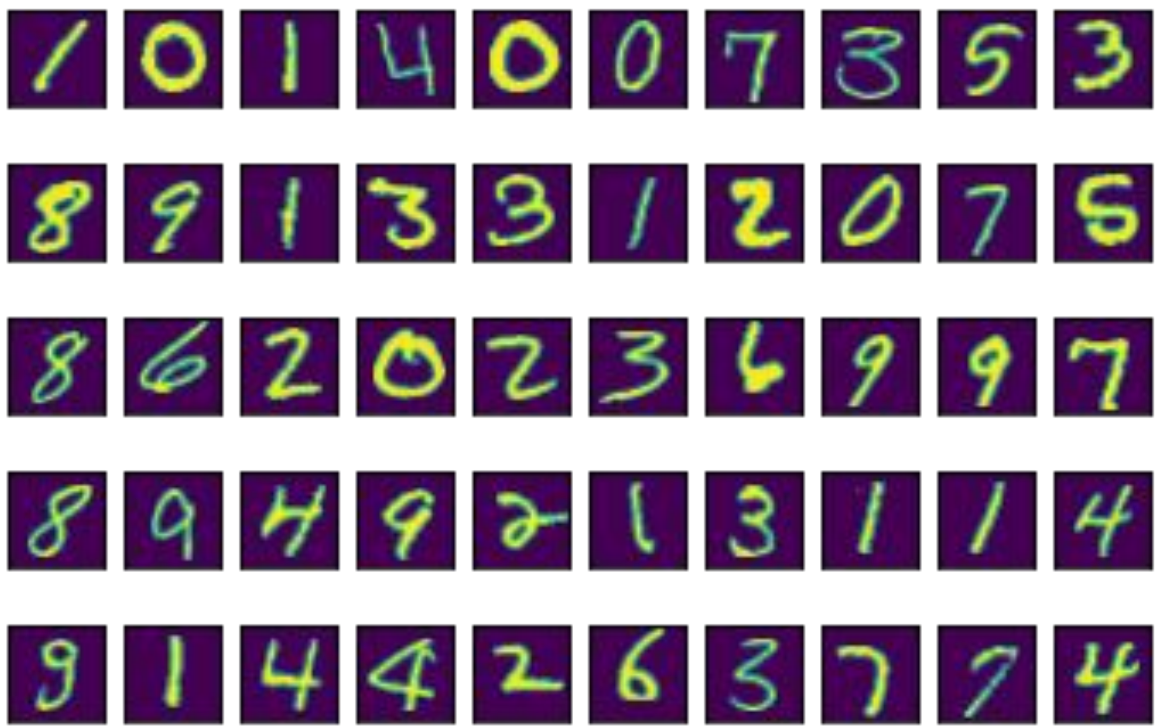
Coordinate = relevance

Latent Dirichlet Allocation, Blei, Ng, and Jordan (2003)

Slide from David Sontag, NYU

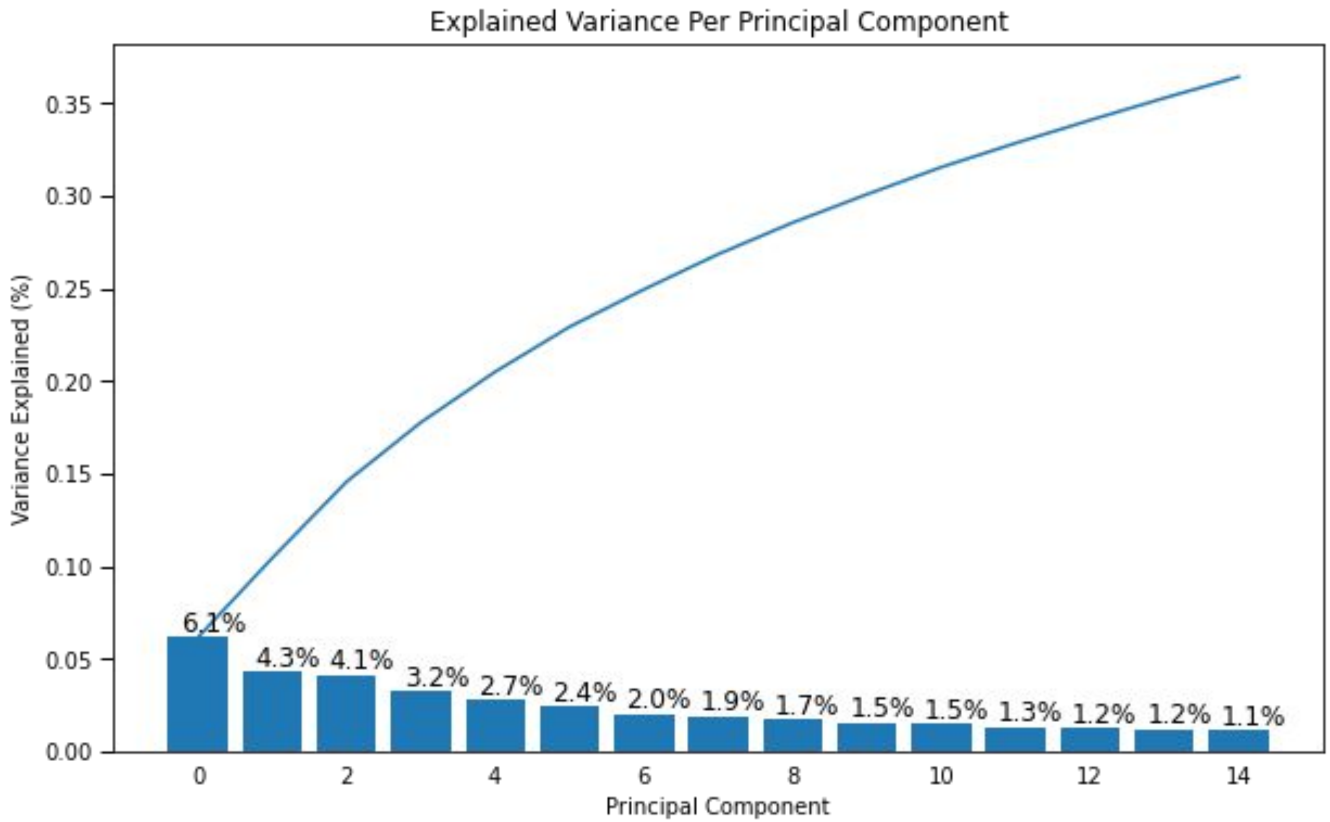
PCA can be applied to different types of data

- MNIST dataset of handwritten digits
- Each sample is a 28x28 pixel image



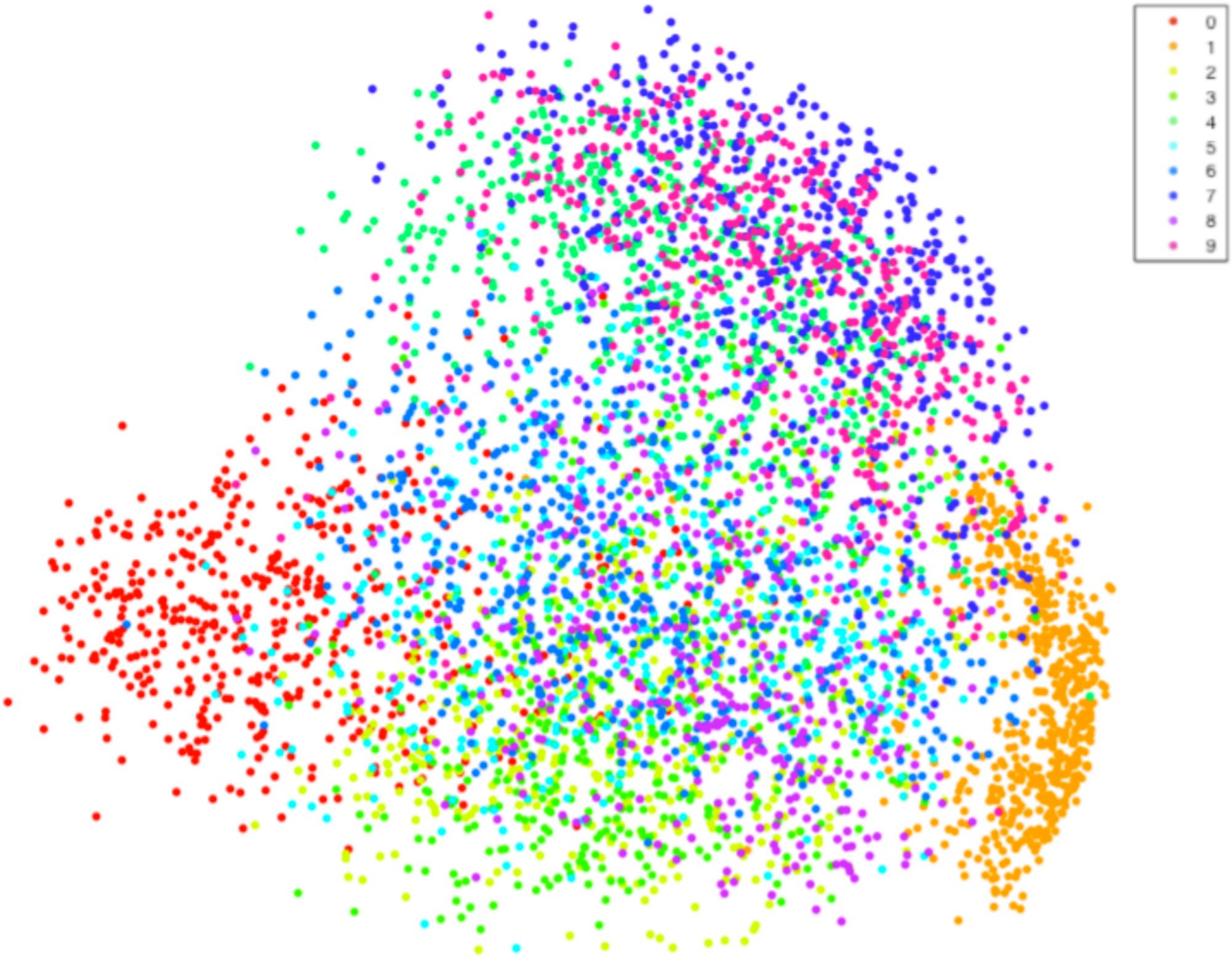
PCA applied to MNIST

- First projections do not capture much of variance



https://ryanwingate.com/intro-to-machine-learning/unsupervised/pca-on-mnist/

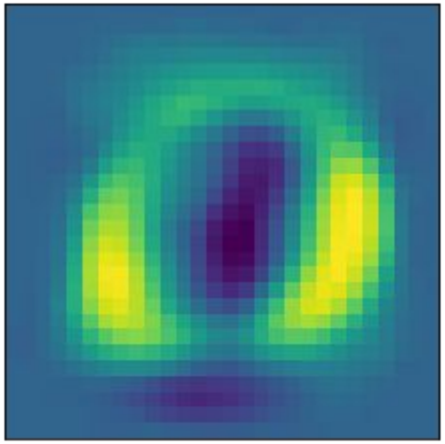
PCA applied to MNIST



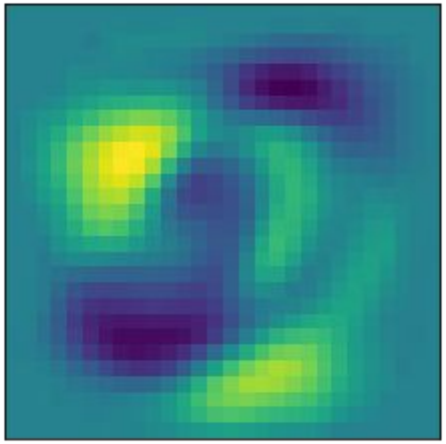
From tSNE lecture notes by Ethan Fetaya, James Lucas and Emad Andrews

PCA applied to MNIST

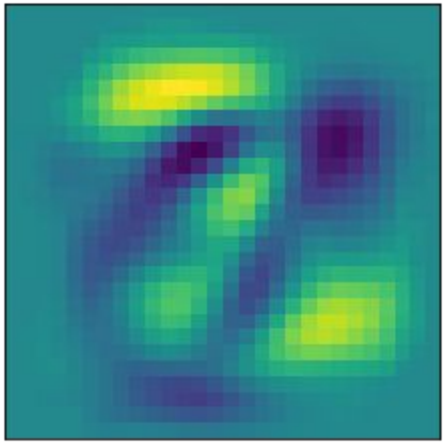
- Simple projections do not capture much of variance
- Not do they make much sense



PC_1 (PVE=6.1%)



PC_2 (PVE=4.3%)



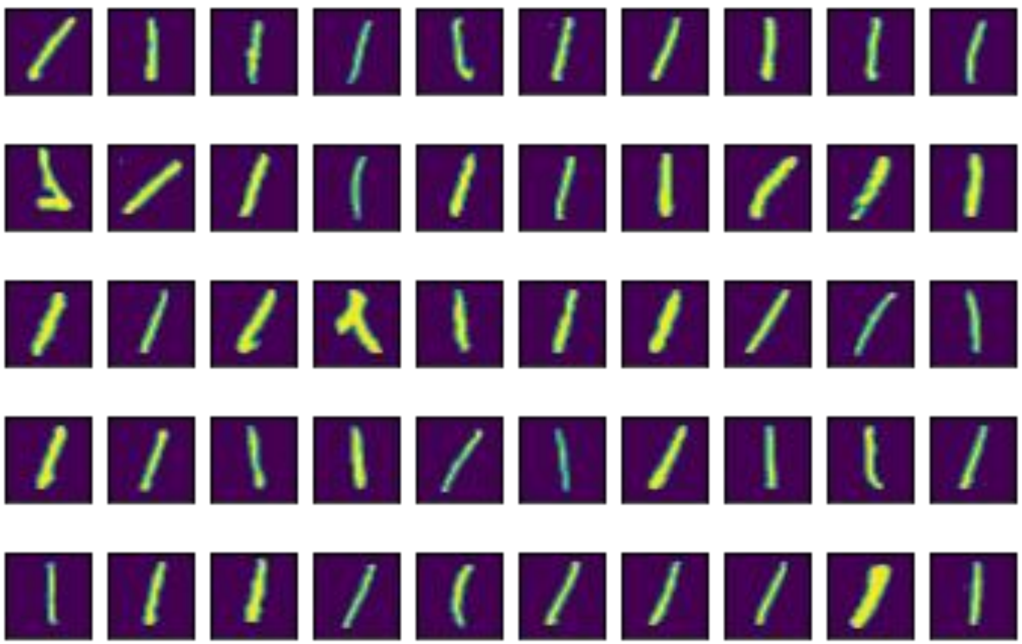
PC_3 (PVE=4.1%)

Always check percent of variance explained (PVE) by PCs

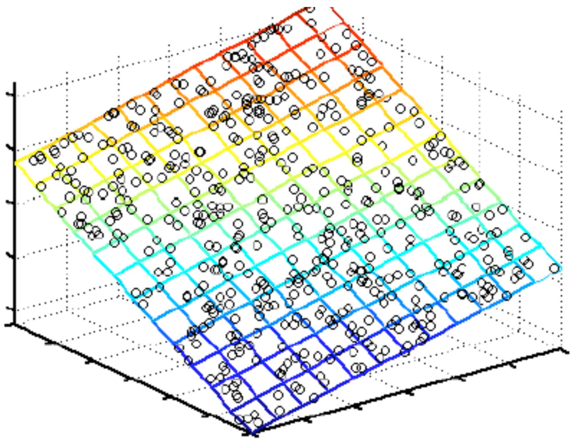
https://ryanwingate.com/intro-to-machine-learning/unsupervised/pca-on-mnist/

PCA can be applied to different types of data

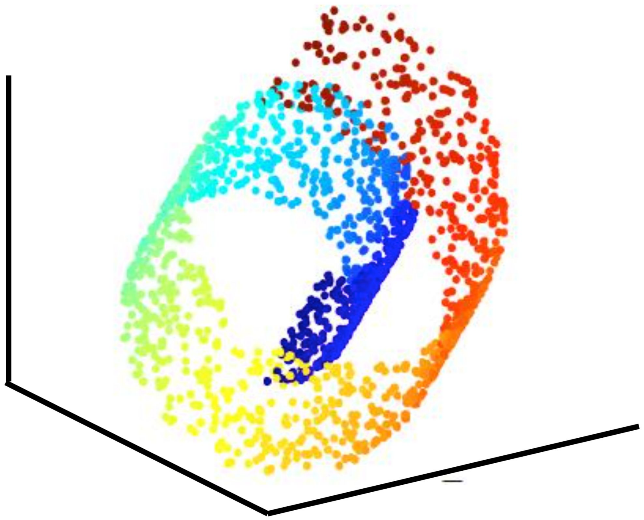
- Why has it failed?
- In this representation, digits are similar enough (different regions/pixels are “on” - no archetypes)



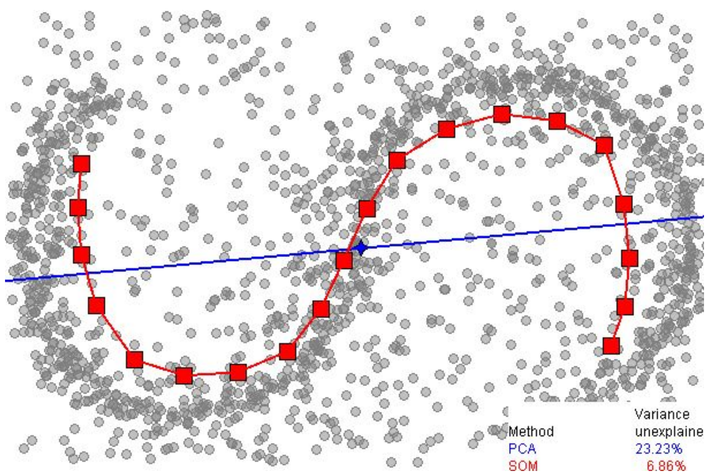
Beyond linear dimensionality reduction



Linear models are limited to hyperplanes
(or lower dimensional subspaces)



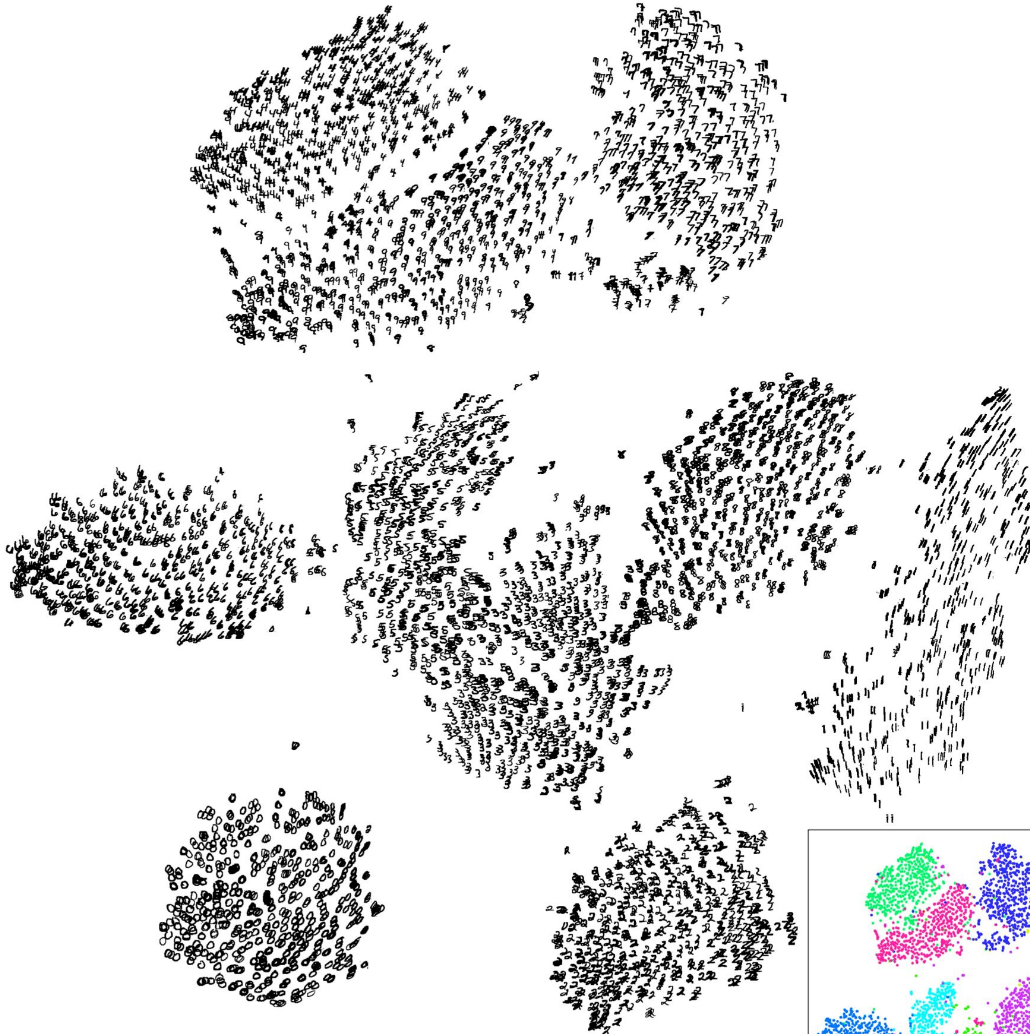
The swiss roll is a (locally linear) manifold

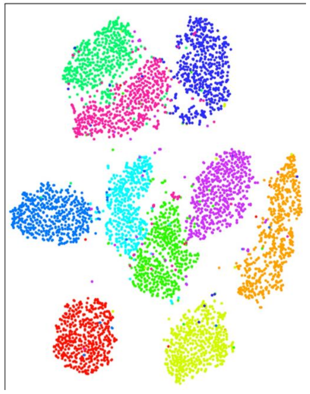


They will perform poorly on other structures

Let’s use neighbors’ adjacency to unfold nonlinear structures

2D embedding of MNIST using tSNE

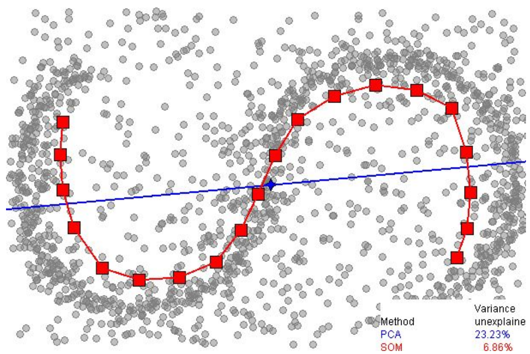


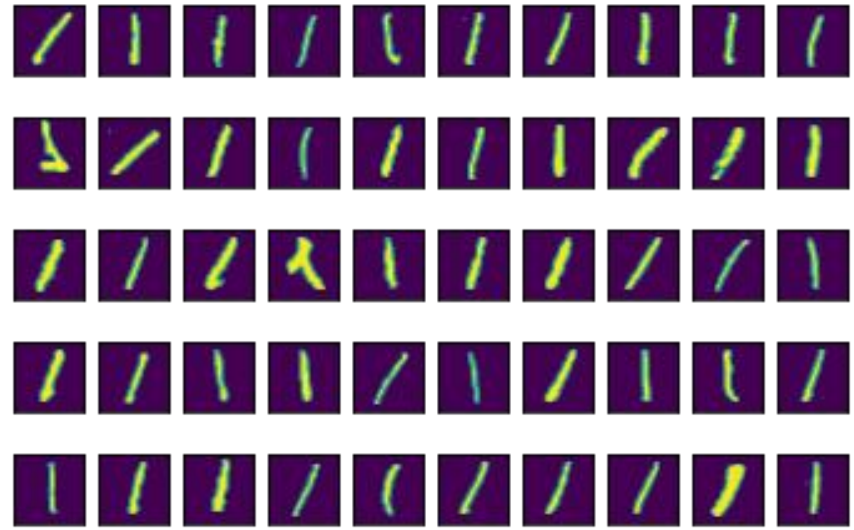


From tSNE lecture notes by Ethan Fetaya, James Lucas and Emad Andrews

How does tSNE work?

Keep your neighbors close

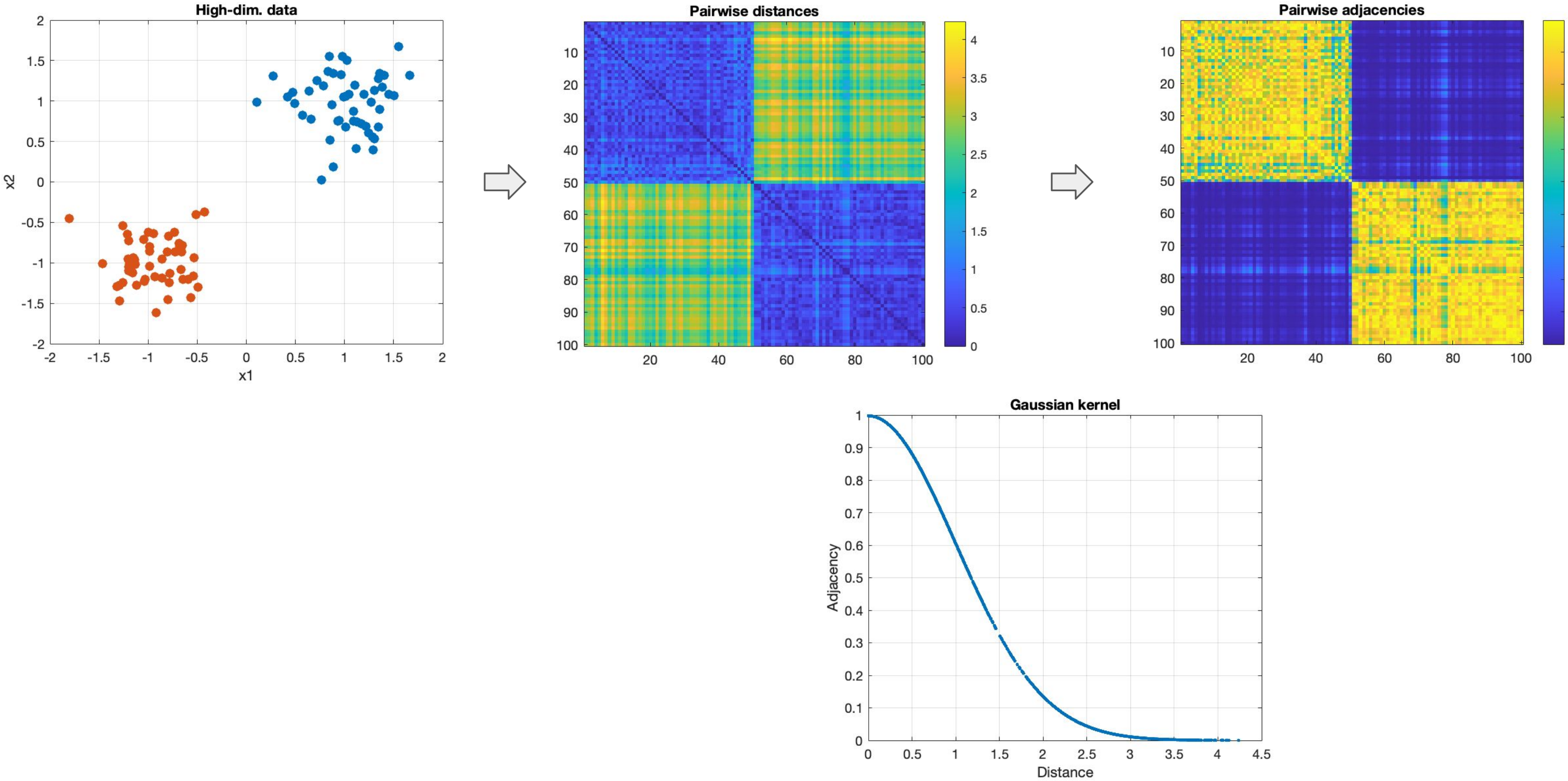




- tSNE stochastically projects data points to a lower dimension
- Iterates until converges to a local solution
- Close neighbors remain close
- Distant pairs could be distorted (less important)

Stochastic Neighbor Embedding, Hinton and Roweis (2002); Visualizing Data using t-SNE, van der Maaten and Hinton (2008)

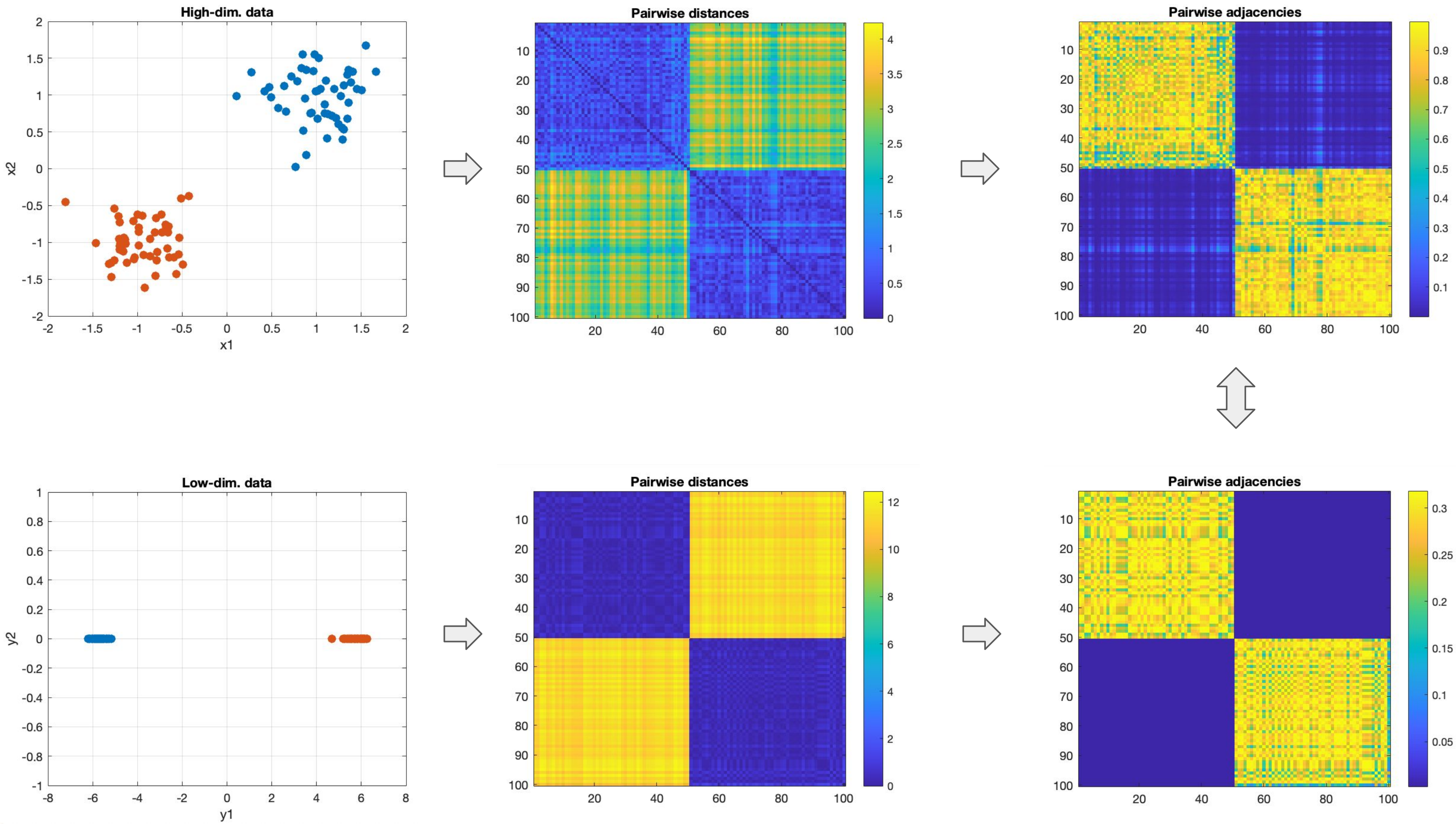
How does tSNE work?



- Look for low-dimensional embedding with similar adjacencies

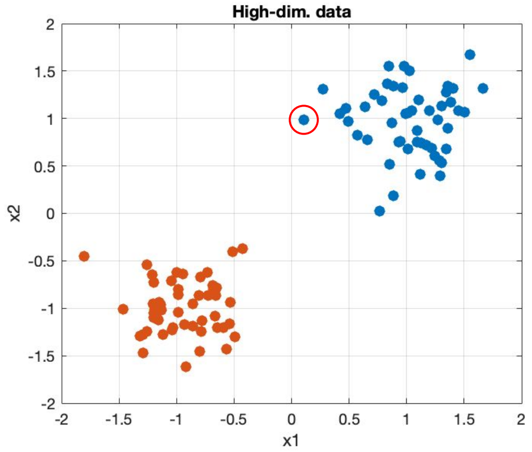
Stochastic Neighbor Embedding, Hinton and Roweis (2002); Visualizing Data using t-SNE, van der Maaten and Hinton (2008)

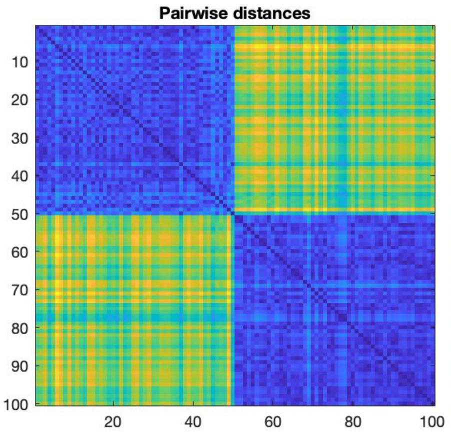
How does tSNE work?

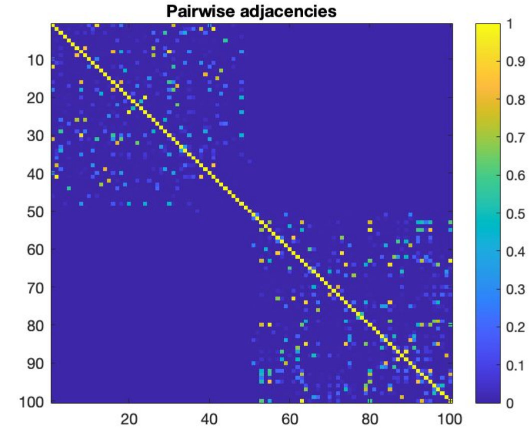


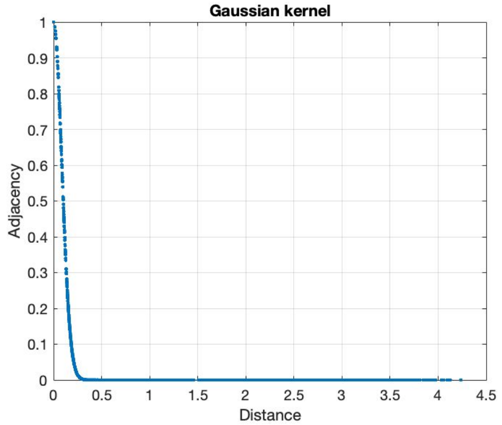
Stochastic Neighbor Embedding, Hinton and Roweis (2002); Visualizing Data using t-SNE, van der Maaten and Hinton (2008)

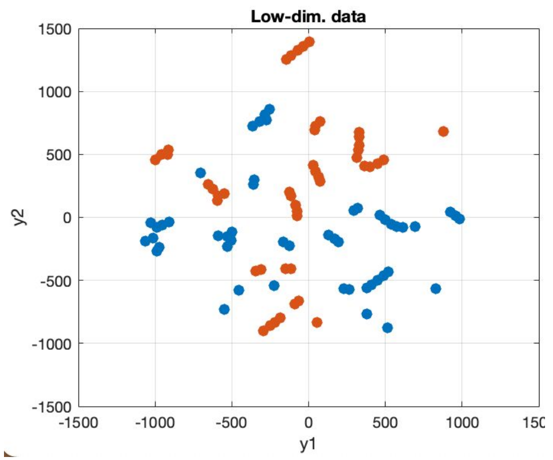
How does tSNE work?







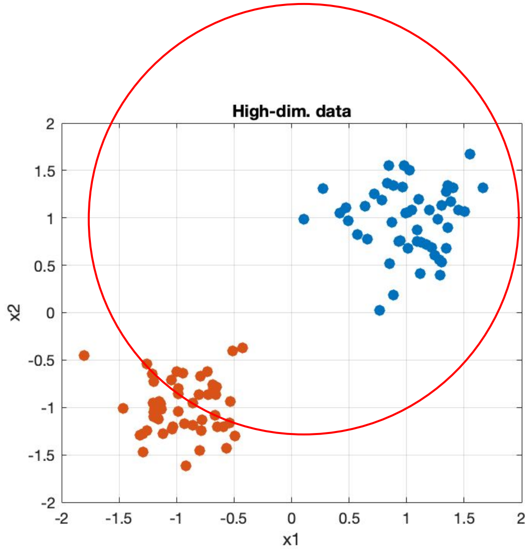


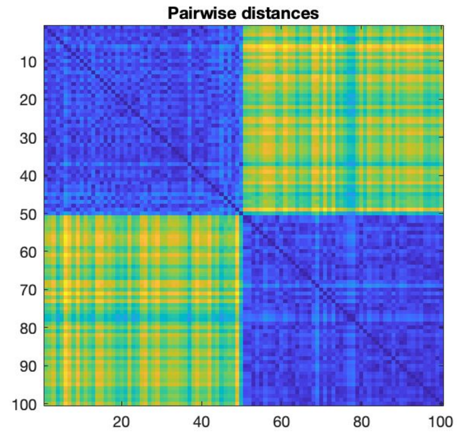


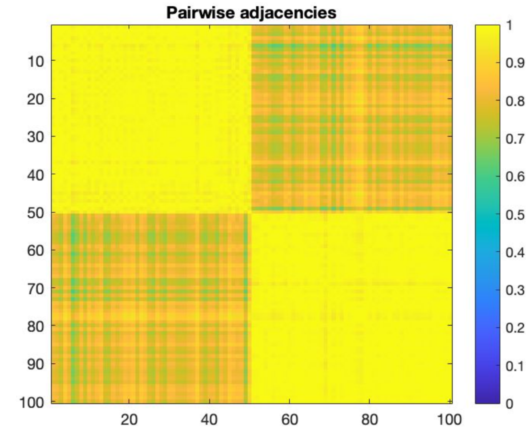
Too few neighbors (tight neighborhood) and it’ll fail

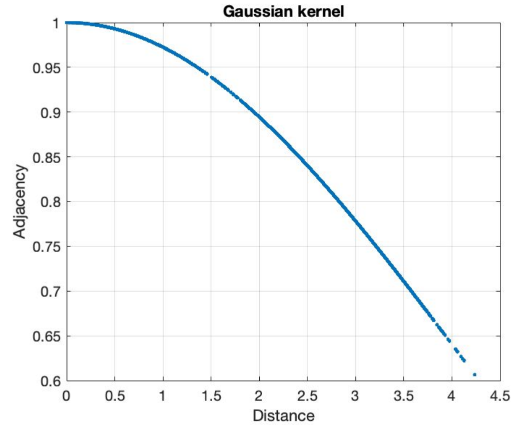
Stochastic Neighbor Embedding, Hinton and Roweis (2002); Visualizing Data using t-SNE, van der Maaten and Hinton (2008)

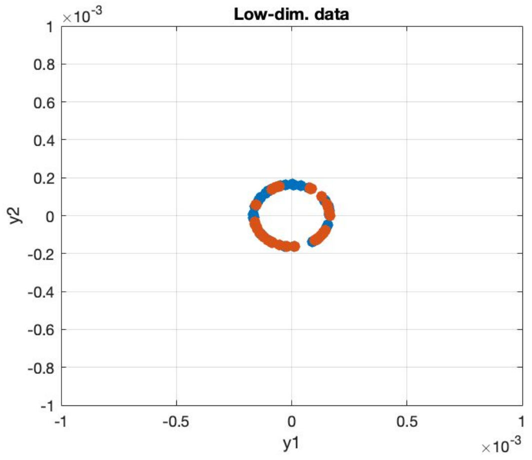
How does tSNE work?









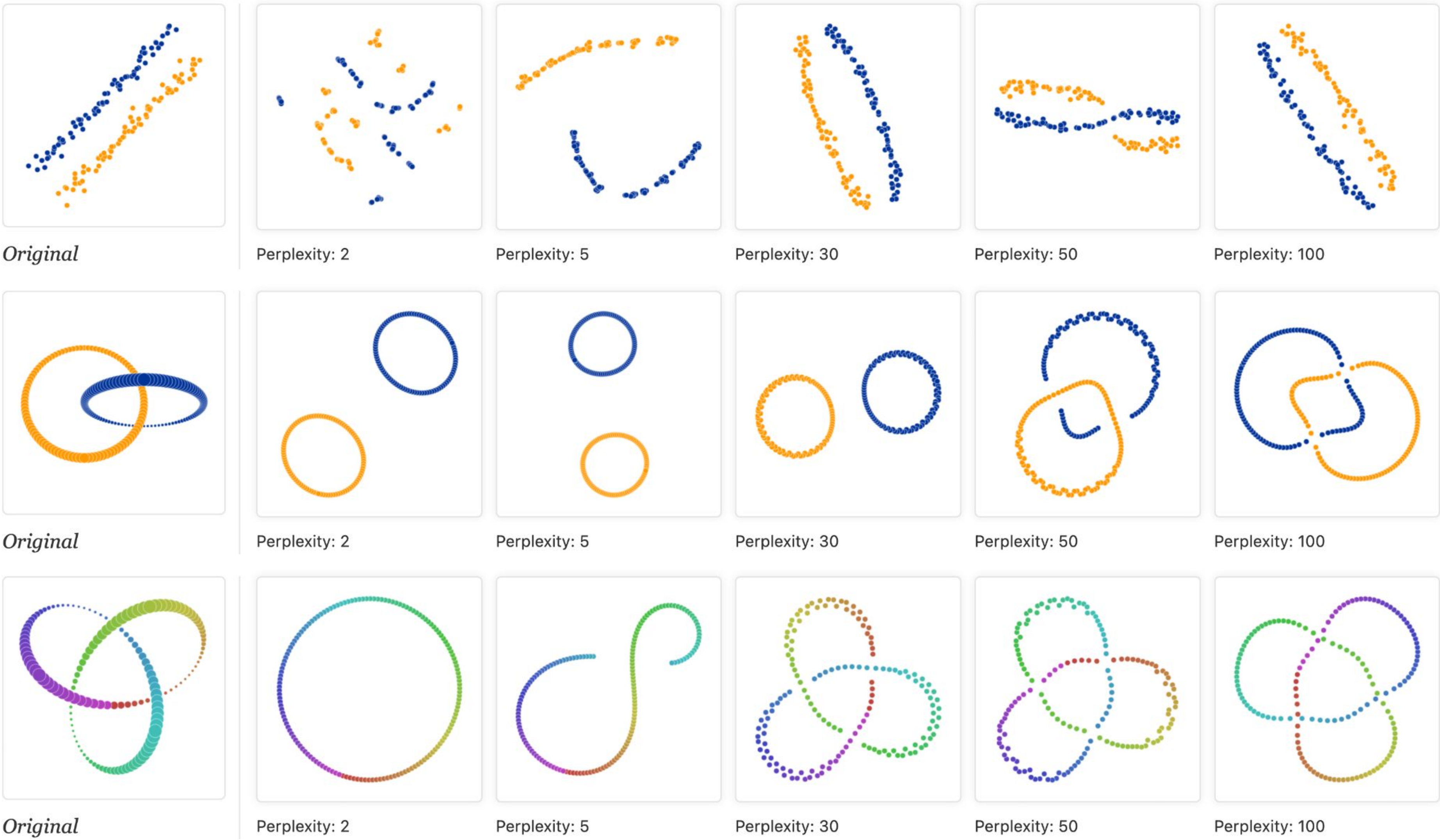


Too many neighbors (wide neighborhood) and it'll fail

Stochastic Neighbor Embedding, Hinton and Roweis (2002); Visualizing Data using t-SNE, van der Maaten and Hinton (2008)

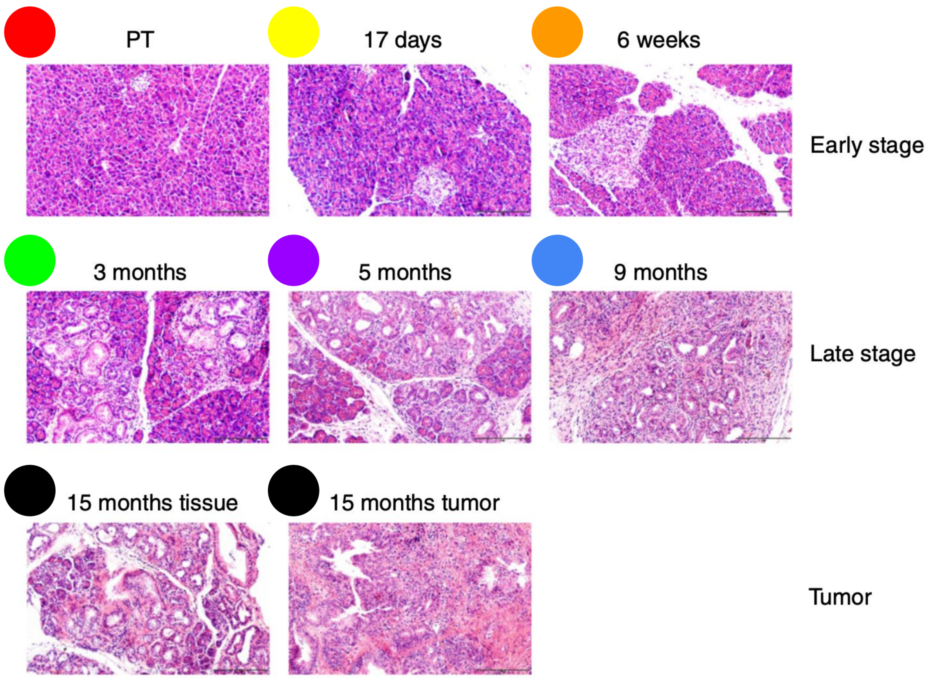
How does tSNE work?

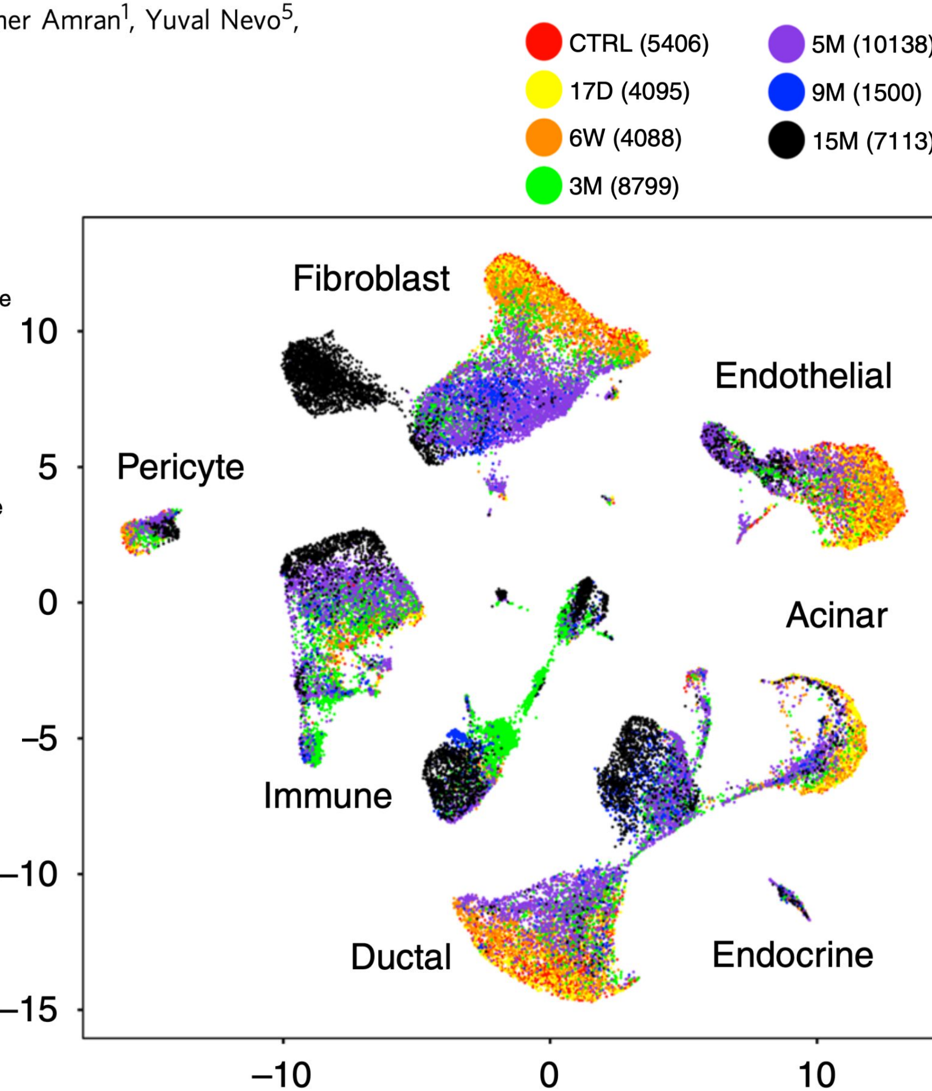
https://distill.pub/2016/misread-tsne



Single-cell transcriptomes of pancreatic preinvasive lesions and cancer reveal acinar metaplastic cells' heterogeneity

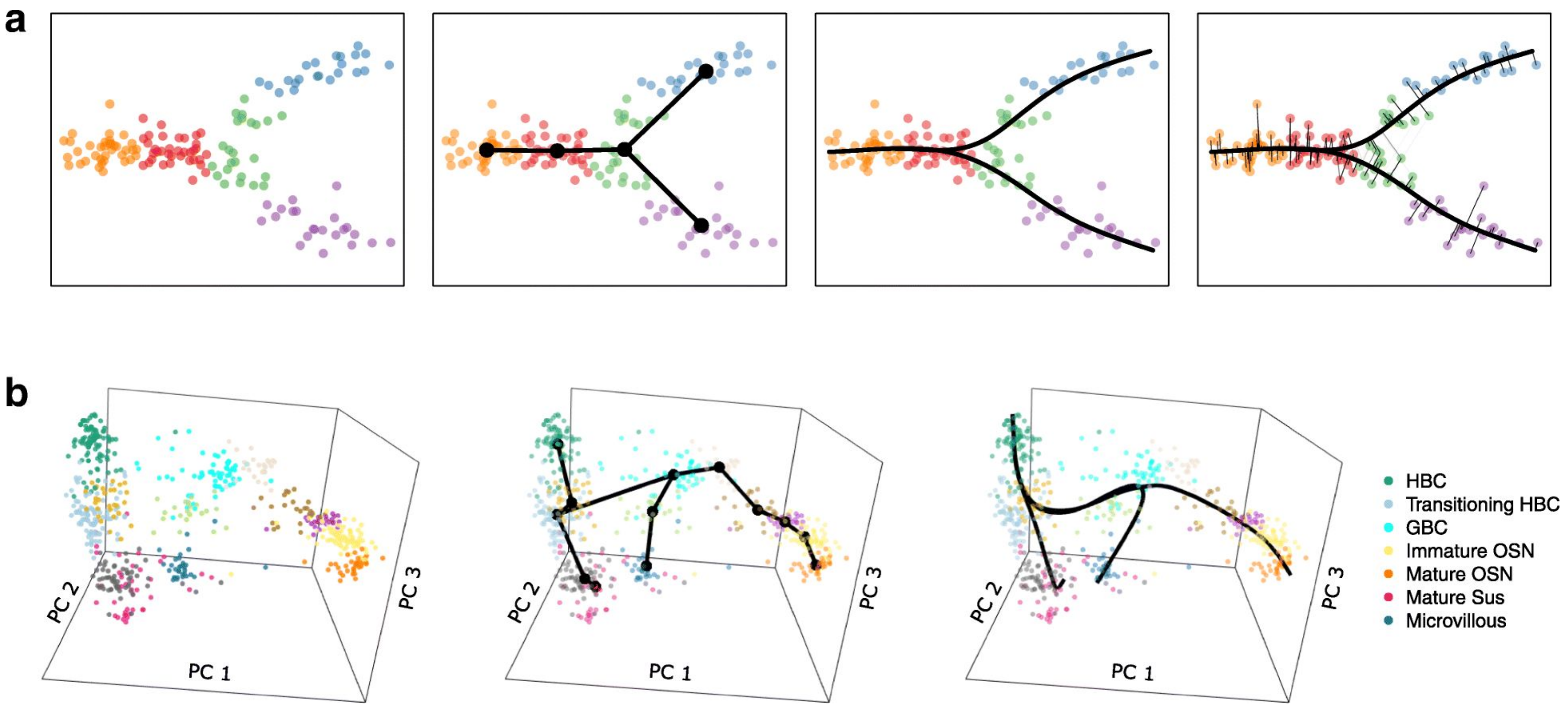
Yehuda Schlesinger[1,7], Oshri Yosefov-Levi[1,7], Dror Kolodkin-Gal[2,3,7], Roy Zvi Granit[1], Luriano Peters[1], Rachel Kalifa[2,3], Lei Xia[1], Abdelmajeed Nasereddin[4], Idit Shiff[4], Osher Amran[1], Yuval Nevo[5], Sharona Elgavish[5], Karine Atlan[6], Gideon Zamir[2]✉ & Oren Parnas[1]✉





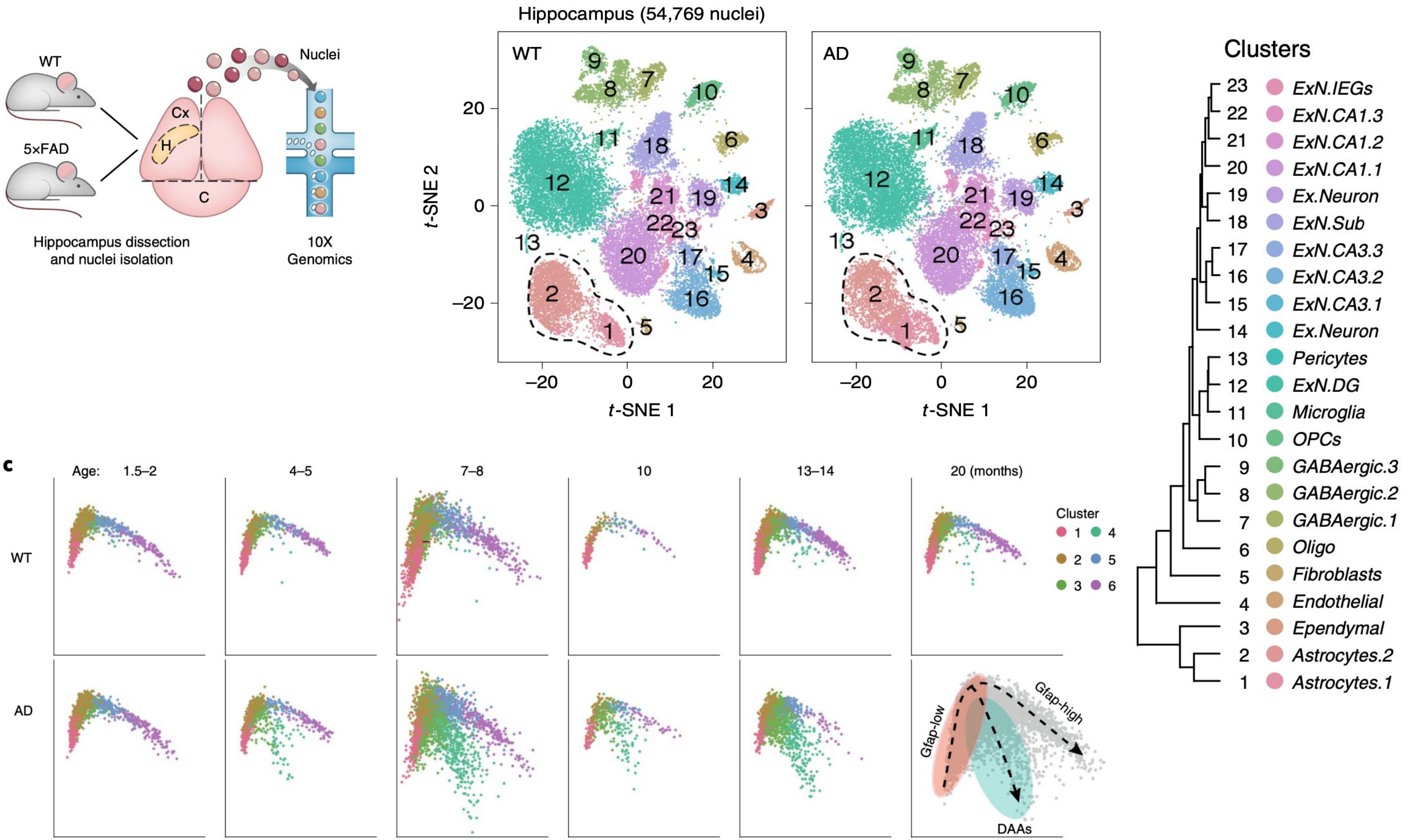
Trajectories and Pseudotime

- Many other alternatives were developed for analysis of single-cell RNA-seq data.
- Slingshot, Monocle and others track gradual changes in cells
- Most embed in a low-dim, cluster, and connect clusters



Disease-associated astrocytes in Alzheimer's disease and aging

Naomi Habib[1,2]✉, Cristin McCabe[2,8], Sedi Medina[3,8], Miriam Varshavsky[1,8], Daniel Kitsberg[1], Raz Dvir-Szternfeld[3], Gilad Green[1], Danielle Dionne[2], Lan Nguyen[2], Jamie L. Marshall[2], Fei Chen[2], Feng Zhang[2,4,5], Tommy Kaplan[6], Aviv Regev[2,7]✉ and Michal Schwartz[2,3]✉



What have we learned?

- Dimensionality reduction increases interpretability
- Projection onto important features
- PCA minimizes distortion, maximizes variance (information).
  - Each new coordinate is a linear function of old coords.
  - Each new axis is orthogonal to previous ones
  - Spans the maximal variance
- Non-linear embedding
  - tSNE converts pairwise distances to adjacencies
  - Stochastically seeks low-dim embedding that mimics original adjacencies. Converges to local optimum.
  - Other similar methods also used, allow inference of trajectories or pseudotime.

Syllabus

1. Introduction
2. Classification
3. Learning 1
4. AI in ophthalmology (Prof. Itay Chowers)
5. Learning 2
6. Regression
7. Clustering
8. Visualization (and dimensionality reduction)
9. Deep learning in image analysis (Prof. Leo Joskowicz)
10. Missing data, statistical dependencies
11. Natural language in medicine (Dr. Gabi Stanovsky)
12. Decisions (utility)
13. Longitudinal Data / Project