Unsupervised and Unstructured Machine Learning

BA820 – Mohannad Elhamod

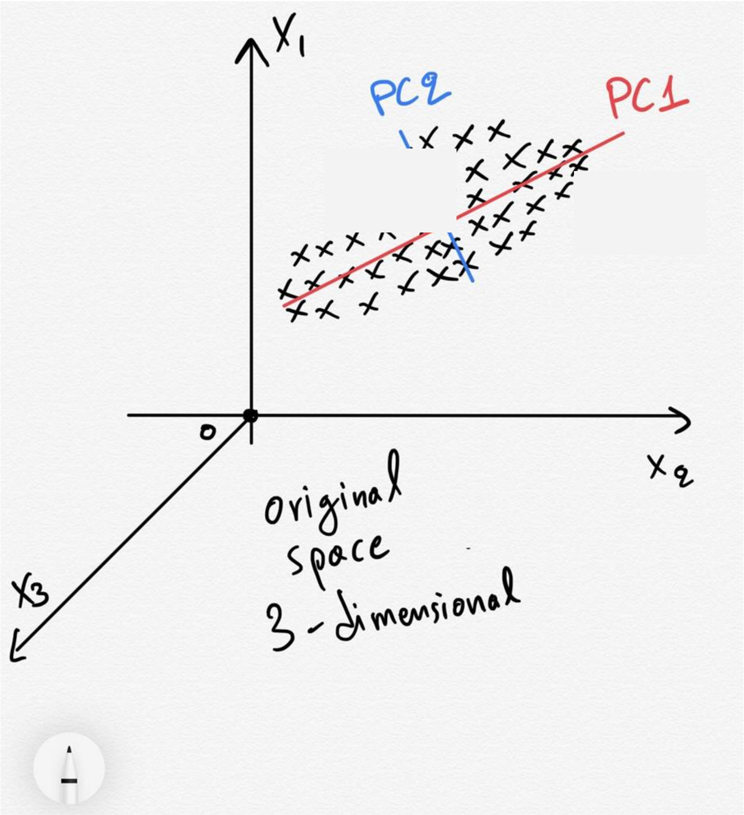
Dimensionality Reduction

Feature Reduction

Problems with having too many attributes:

- Analyses/Modeling can take a very long time
- Data can take too much space.
- Risk of correlation/redundancy amongst the variables.
  - Difficulty in interpreting the fit of our models.
  - Tend to overemphasize the underlying variable's contribution.
- Helps remove noise.
- Not easy to visualize/interpret.
- Curse of Dimensionality!

Dimensionality Reduction



towardsdatascience.com

Dimensionality Reduction Techniques

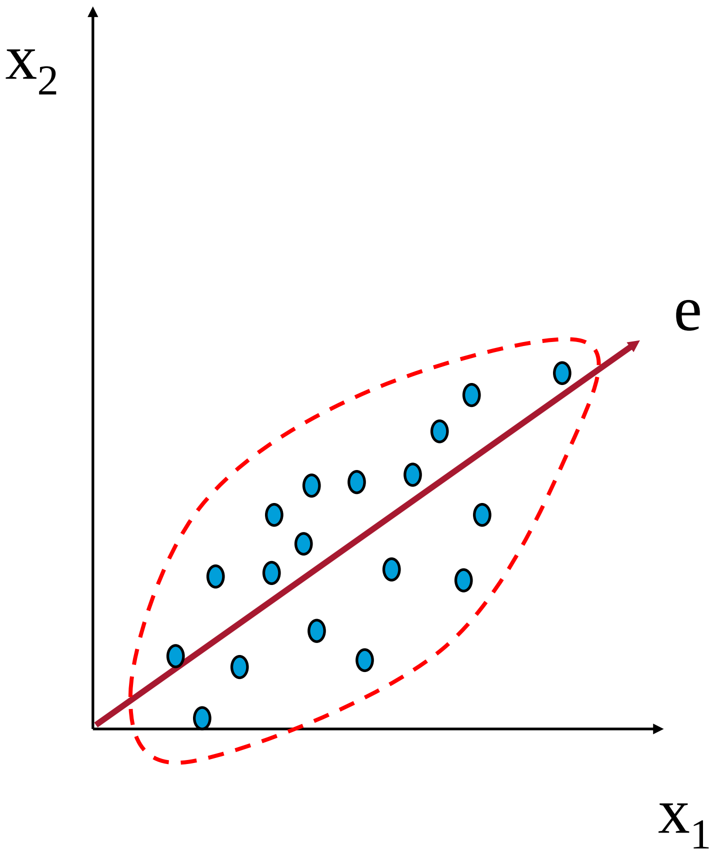
- **Linear:**
  - The new dimensions are *a linear* combination of the originals.
  - PCA is the prime example of this category.

- **Non-linear:** such, as t-SNE and UMAP.

Principal Component Analysis

Intuition

- Goal is to find direction(s) that captures the largest amount of variation in data.
- We call these direction(s) *principal component(s).*



Properties of Principal Components

- First component lies along the direction of the data’s largest variance/spread.
- Each component is perpendicular to all other components (independence).
- The components are ordered in terms of their ability of explaining the data (i.e., in order of how much variance in the data they capture).

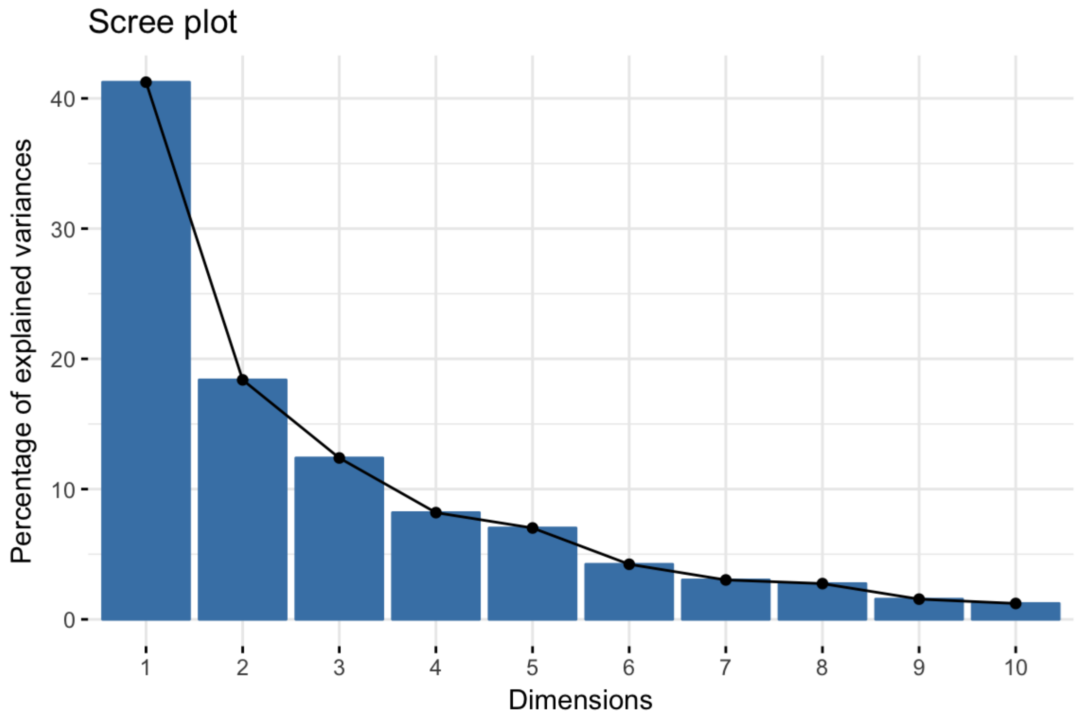
- Let’s play with this

Dimensionality Reduction with PCA

- The number of components available is equal to the number of attributes being analyzed.
- However, in *most* analyses, only the first few components account for meaningful amounts of variance (>90%), so only these first few components are retained, interpreted, and used in subsequent analyses.
- When you remove dimensions. You lose some information!

How Many Components do we select?

- Method1: Remove dimensions when reaching a sufficient cumulative explained variance.
- Method2: Elbow method could be used.



Considerations

- PCA assumes the data has linear patterns in terms of the original attribute.
- PCA can be used for Feature Engineering (the new features can be used for down-stream tasks).

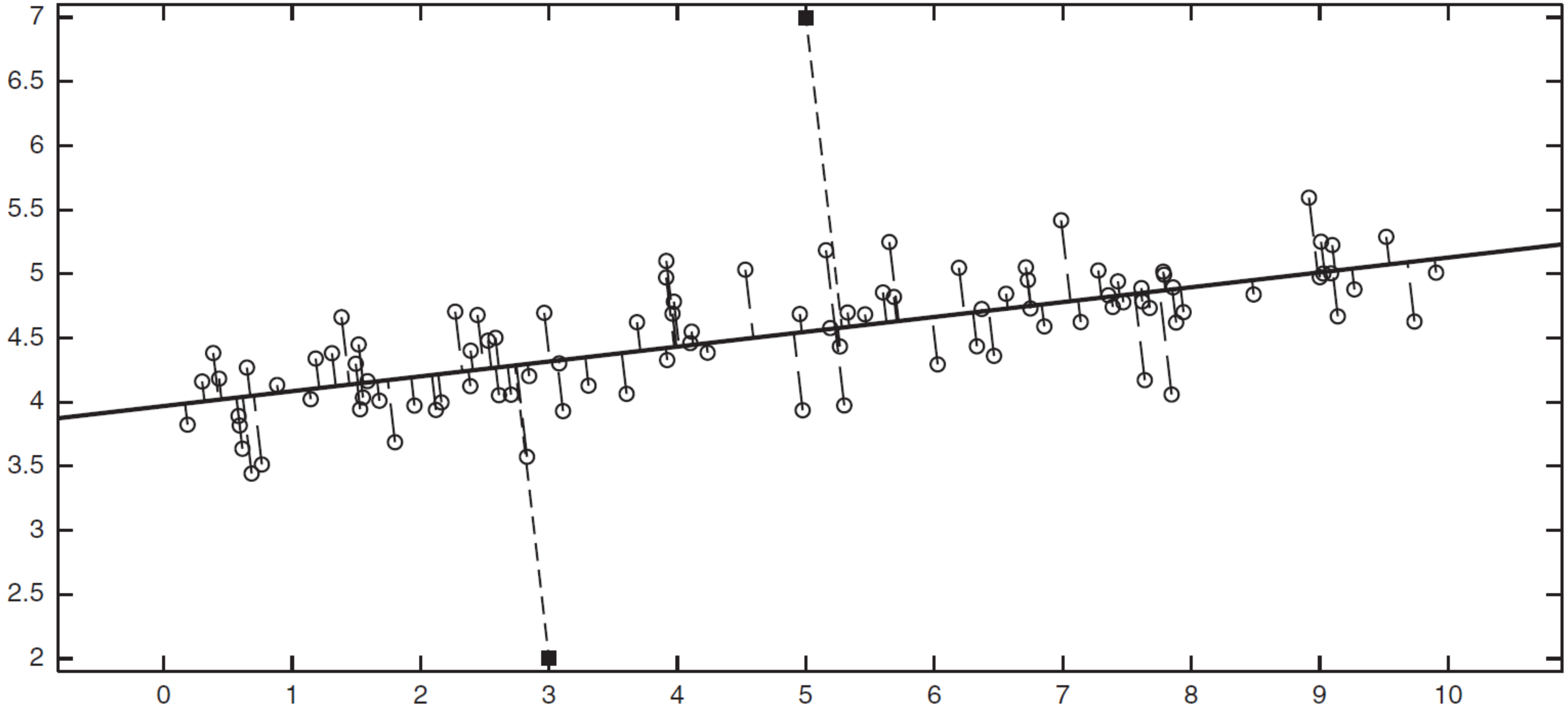
Reconstruction Error

- Let $\mathbf{x}$ be the original data point.
- Using PCA, project the point to a lower dimensionality.
- Project the object back to the original space. Call this object $\hat{\mathbf{x}}$

$$\textbf{Reconstruction Error}(\mathbf{x}) = ||\mathbf{x} - \hat{\mathbf{x}}||$$

- Points with large reconstruction errors are anomalies

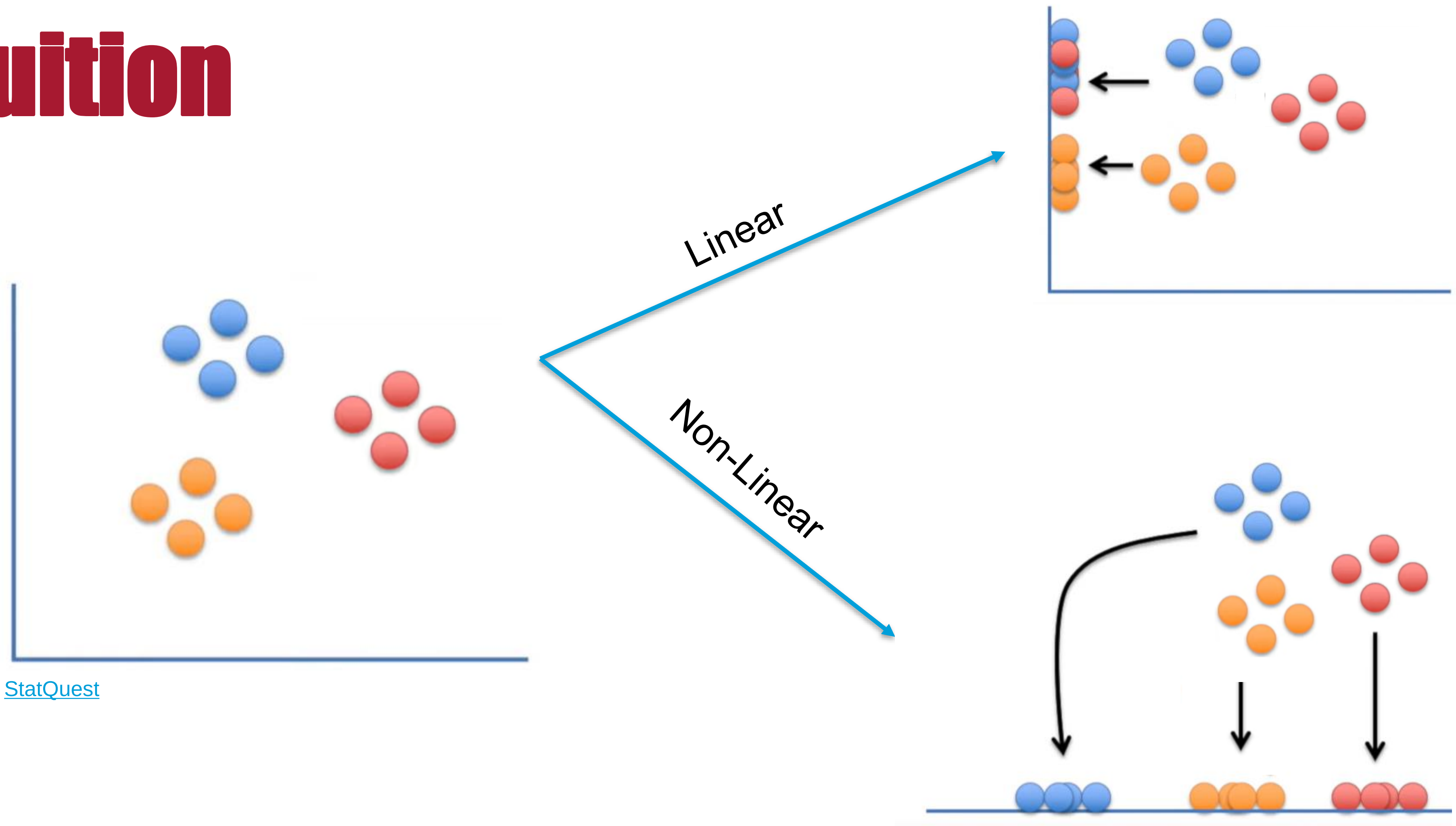
Reconstruction of two-dimensional data



Non-linear Dimensionality Reduction (t-SNE)

BOSTON UNIVERSITY

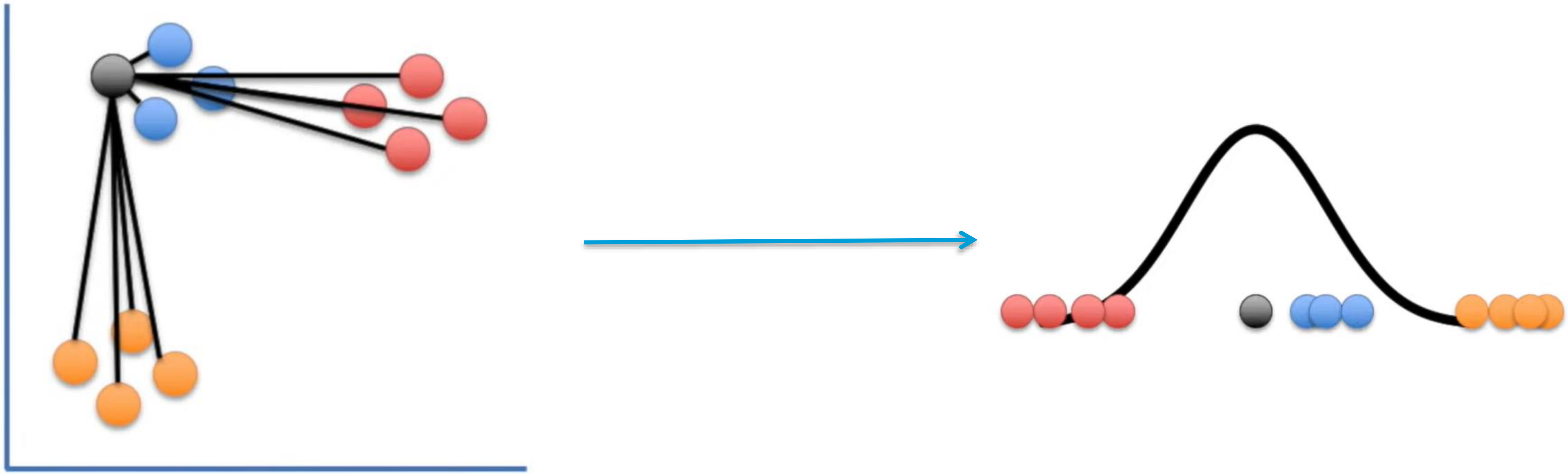
Intuition



StatQuest

Intuition

- For each datapoint, map the distances in the original space onto a t-distribution in the reduced space.



StatQuest

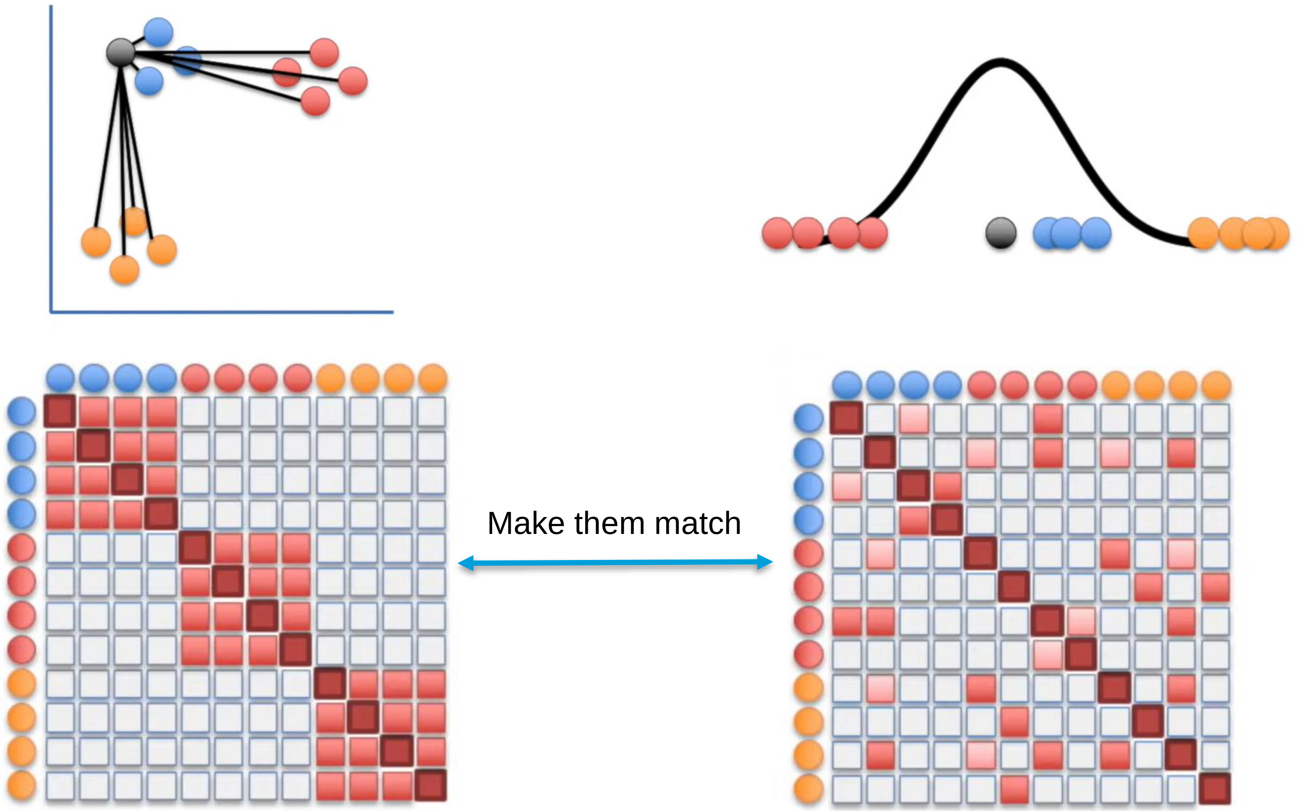
Original dimensions

Fewer dimensions

Intuition

- Objective: Make the similarity matrices correlate.
- It's iterative
- Has a cost/objective function.

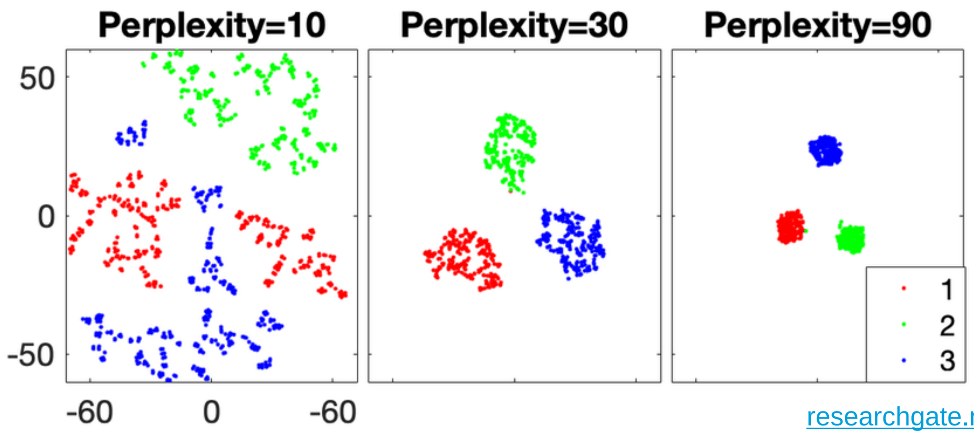
StatQuest



T-SNE (t-distributed Stochastic Neighbor Embedding)

- Most suitable for visual exploration.
- Preserves ‘nearness’ between samples:
    - Maximizes the distance in the new space when points are relatively apart in the high dimensional space, and vice versa.
    - Thus, it will naturally *cluster* the points that are close to each other in the high dimensional space.

Challenges

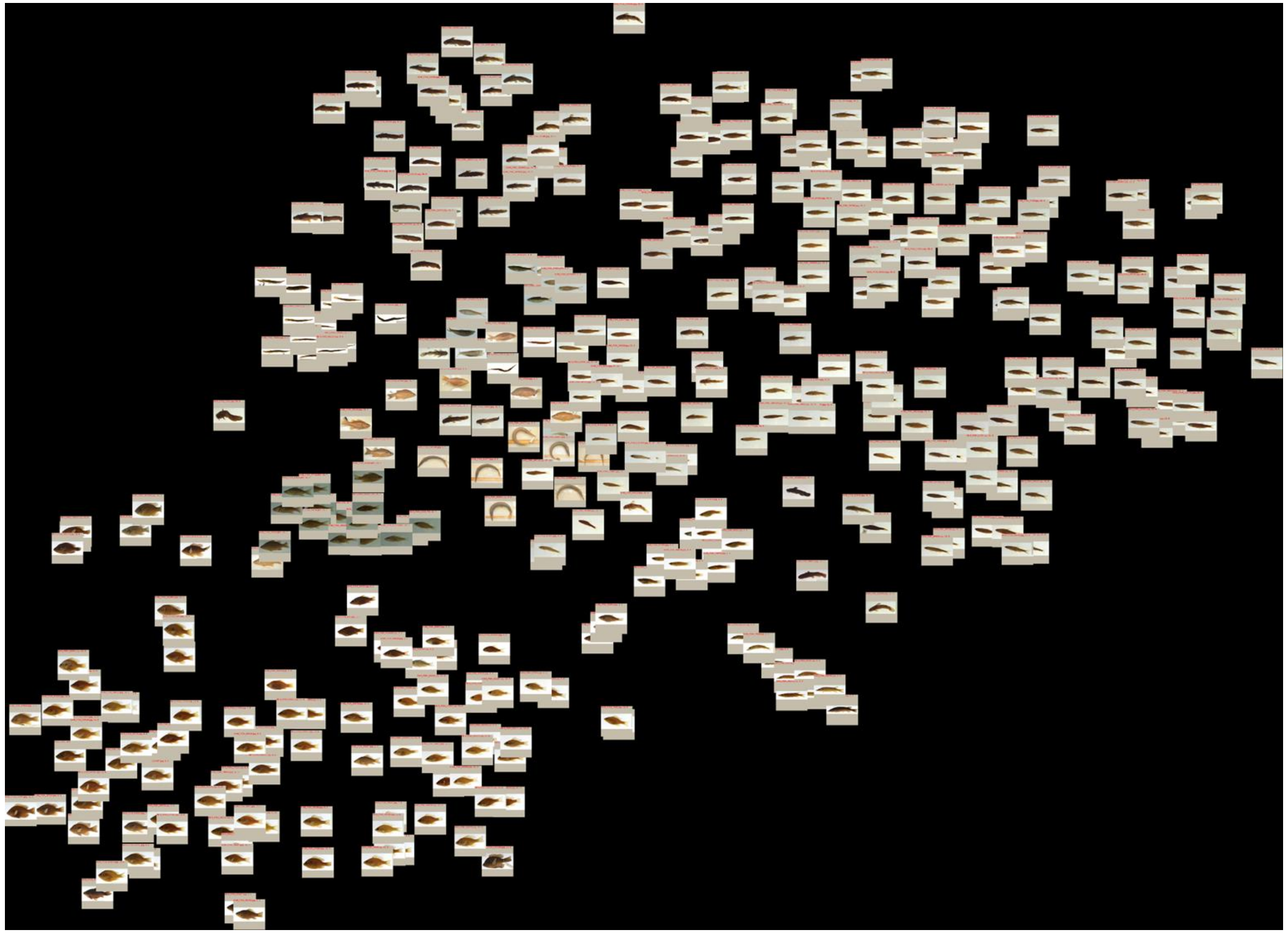


researchgate.net

- Stochastic.
- Relatively slow.
- Cannot be used on new data.
- “Has some hyper-parameters (usually safe to set to `auto`):
  - *Learning rate:* experiment with values at logarithmic scale.
  - *Perplexity:* related to the variance of the data in the new space. A higher number leads to more distinct clustering.
  - *Number of iteration.*

Honorable Mention

- Preliminary results from my research during PhD.



Metrics of "Goodness"

BOSTON UNIVERSITY

PCA vs t-SNE

- PCA:
  - Reconstruction error.
  - Explained variance.
  - Statistical metric (pca.score in sklearn. *Higher is better*).
- t-SNE
  - Statistical metric (tsne.kl_divergence_ in sklearn. *Lower is better*).