Machine Learning: Advanced

Dimensionality reduction, visualization, clustering

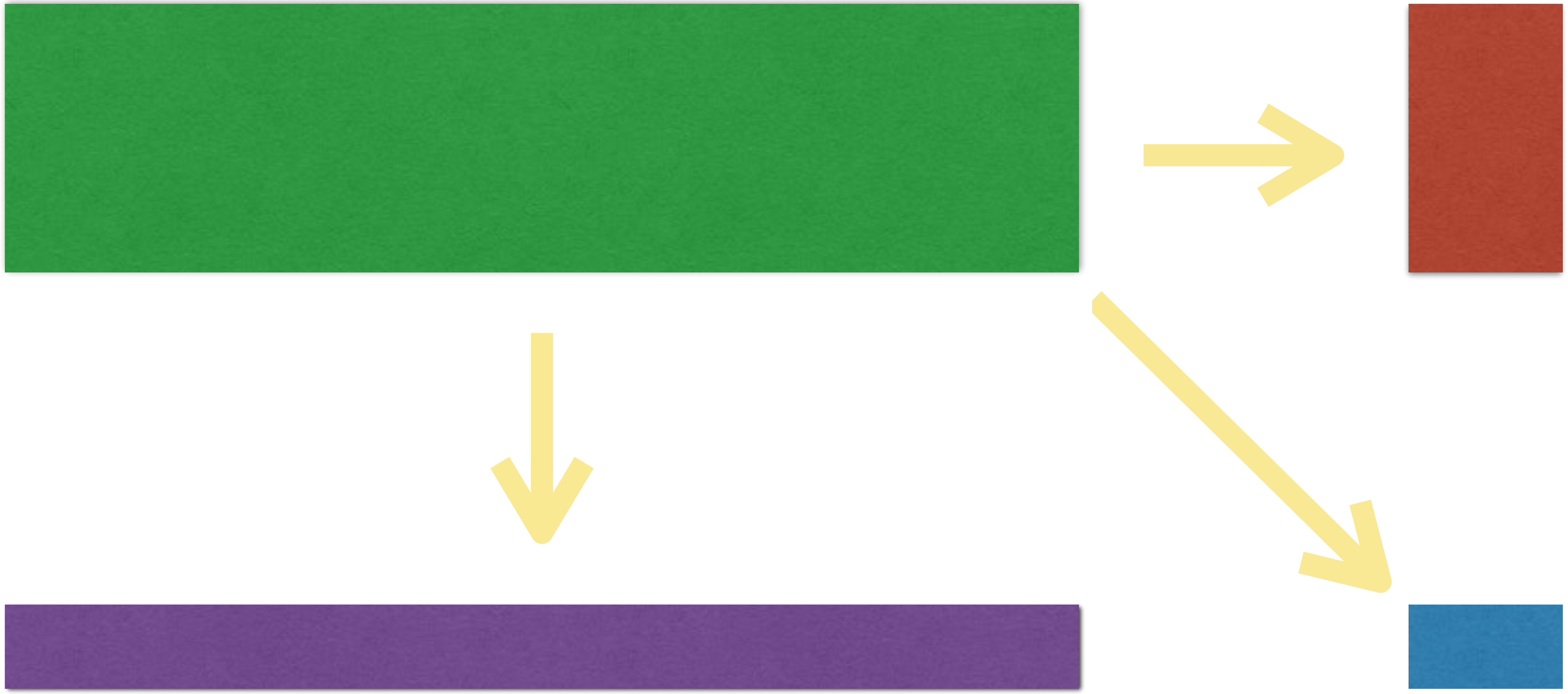
SUVRIT SRA

Massachusetts Institute of Technology

ml.mit.edu

Dimensionality reduction

(we'll use term ***broadly****: both to reduce 'd/p' and N)*



Foundational tool: PCA

(Working with data where we only have $(x_1, \ldots, x_n)$ *instead of* (x_i, y_i) *pairs!)*

What are “principal” components?

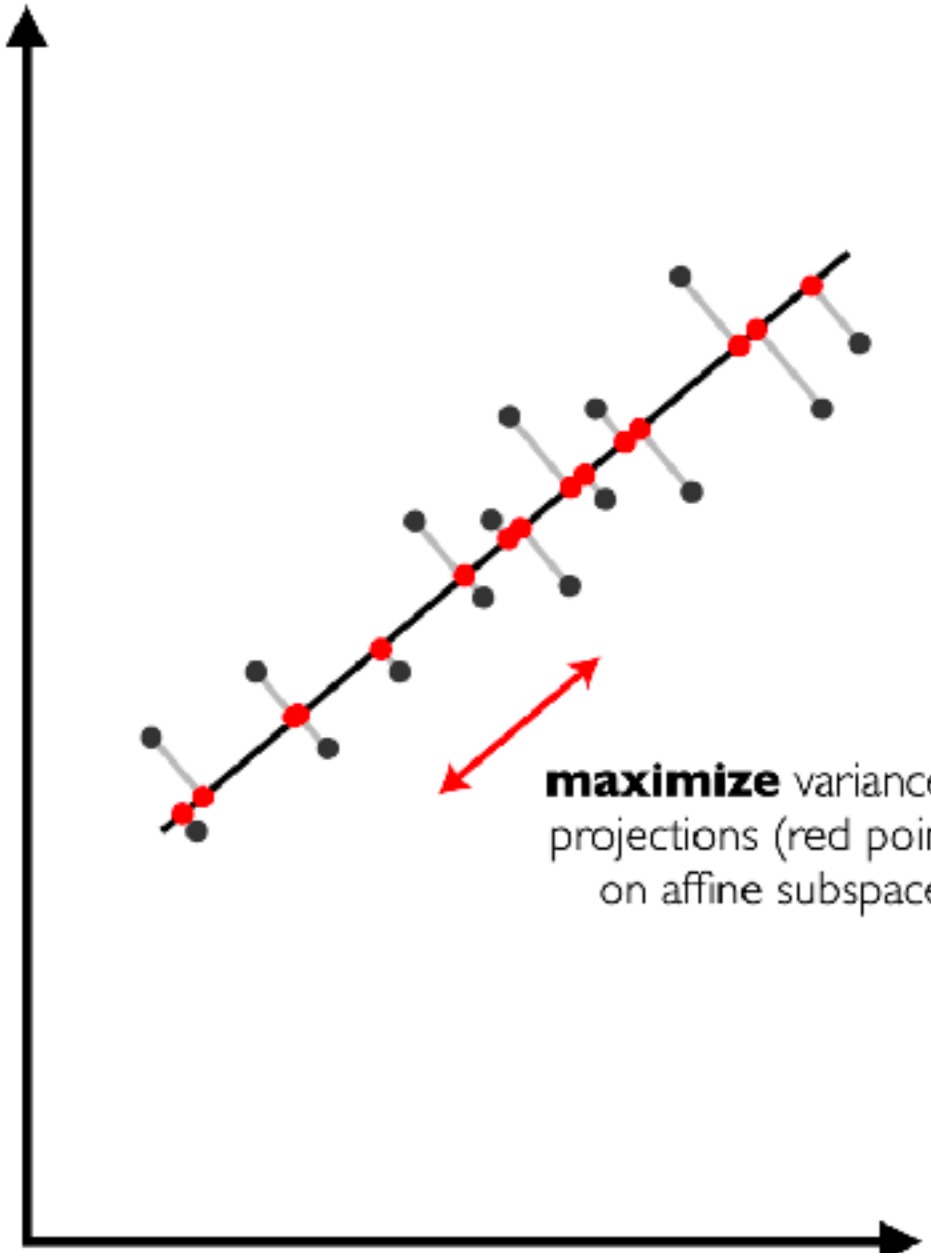
Goal of PCA Identify “principal” directions in data

Question: Why might we want such directions?

[Hotelling, 1933] Project data onto lower-dim affine subspace
Seek to maximize variance of projected data
Thereby capture directions of max spread in data

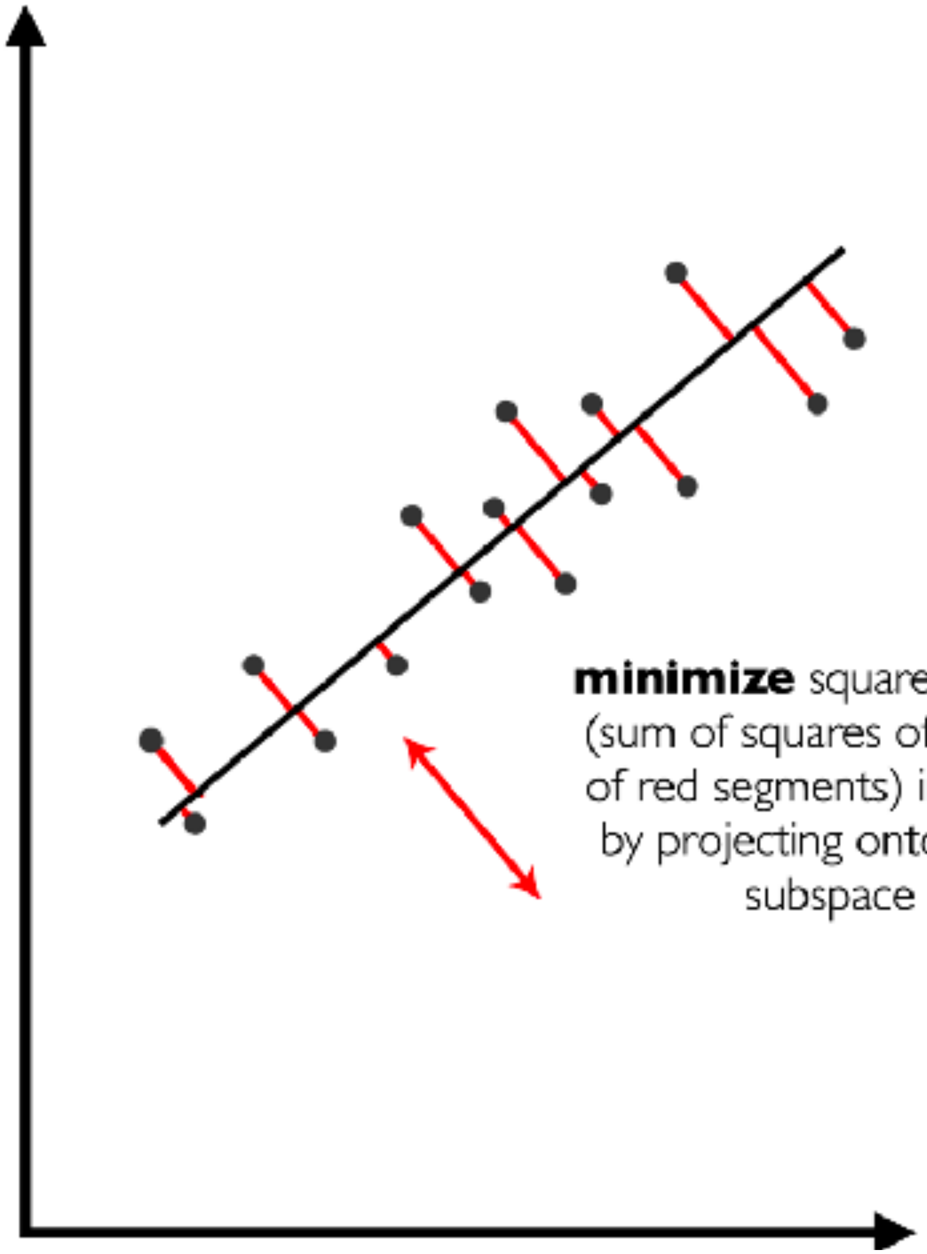
[Pearson, 1901] Project data onto lower-dim affine subspace
Seek to minimize projection error (“movement” error)
Thereby captures “most informative” directions

PCA: maximizing variance view

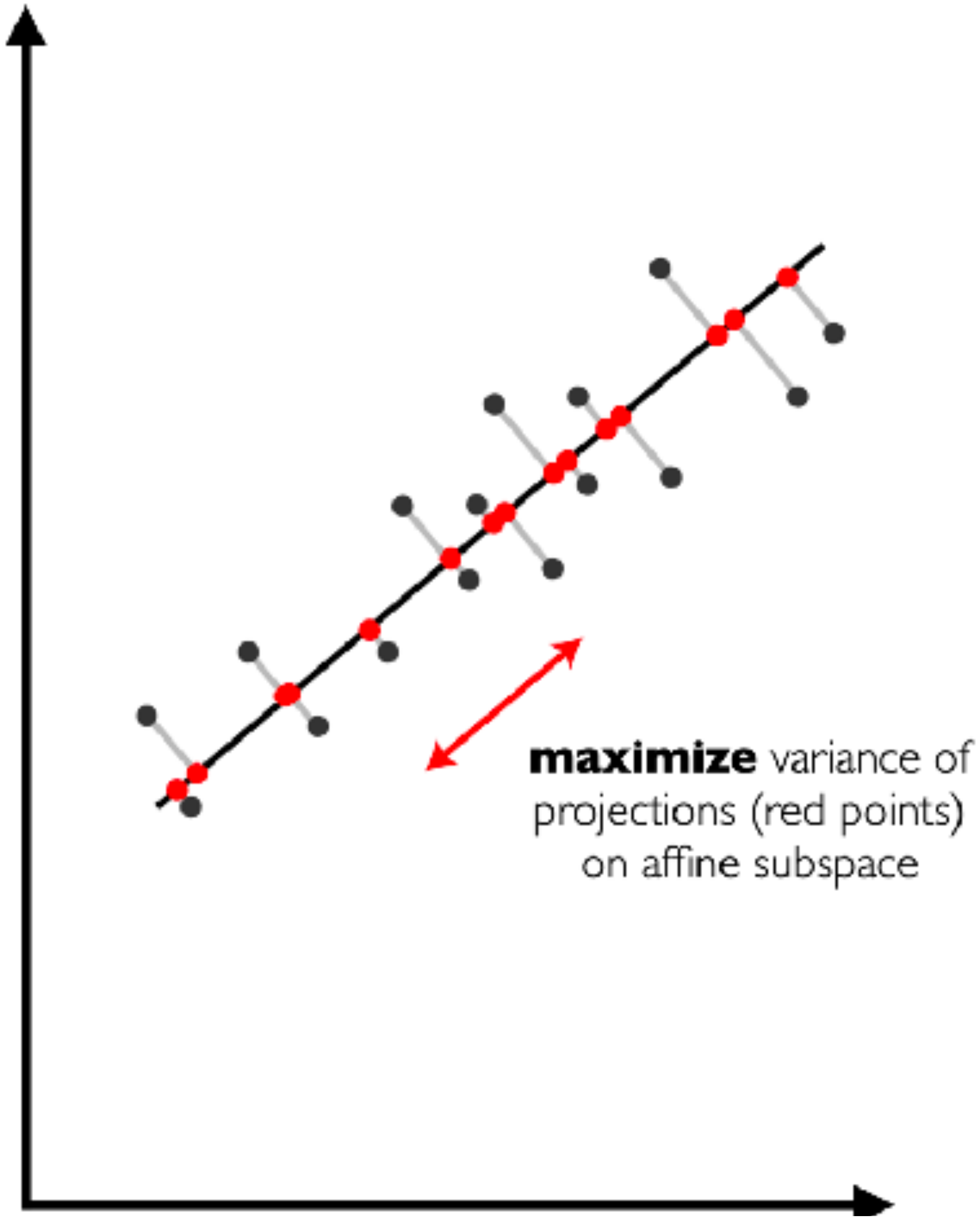


PC1: direction of largest variance in data

PCA: minimizing projection error

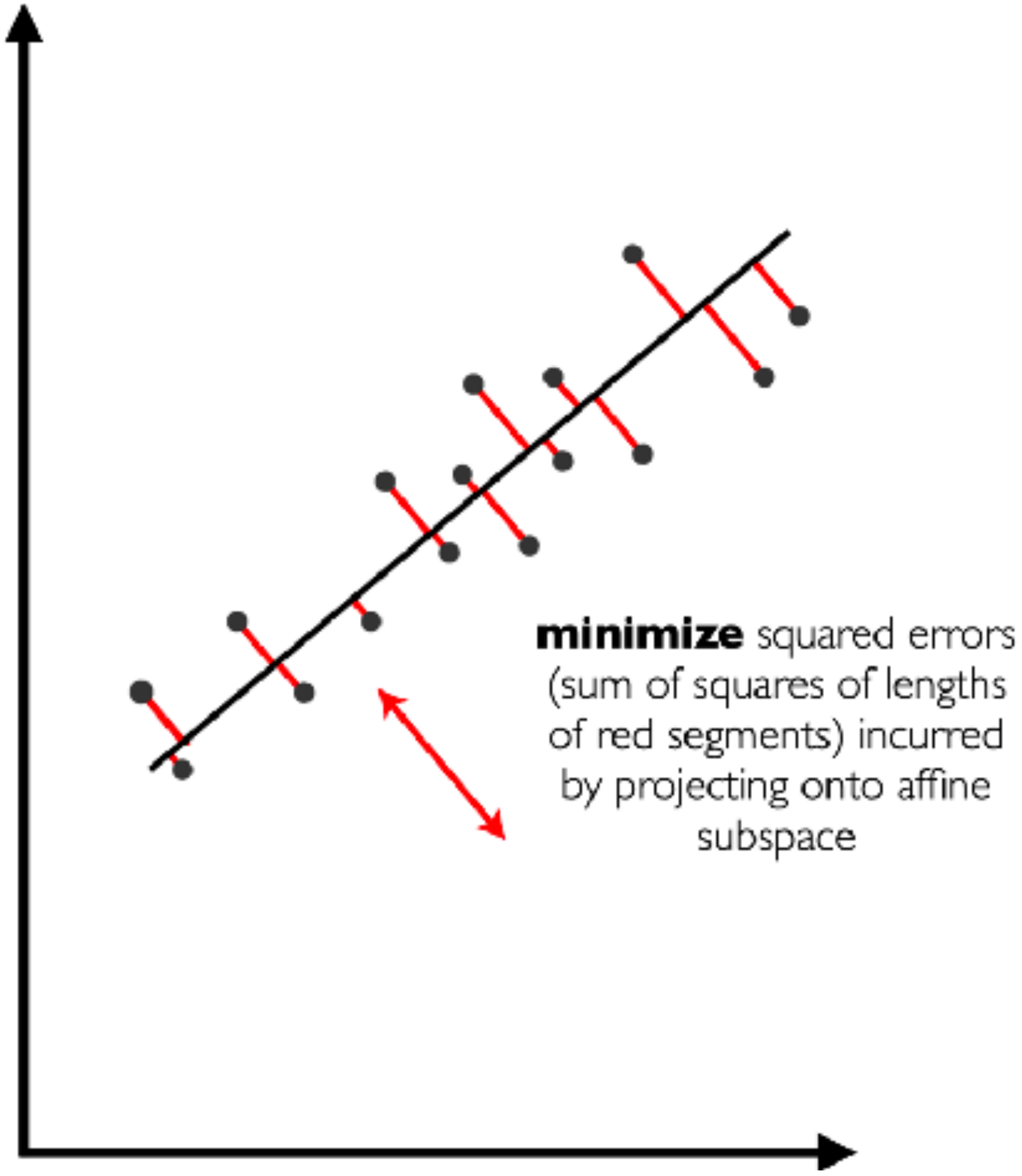


minimize squared errors (sum of squares of lengths of red segments) incurred by projecting onto affine subspace



maximize variance of projections (red points) on affine subspace

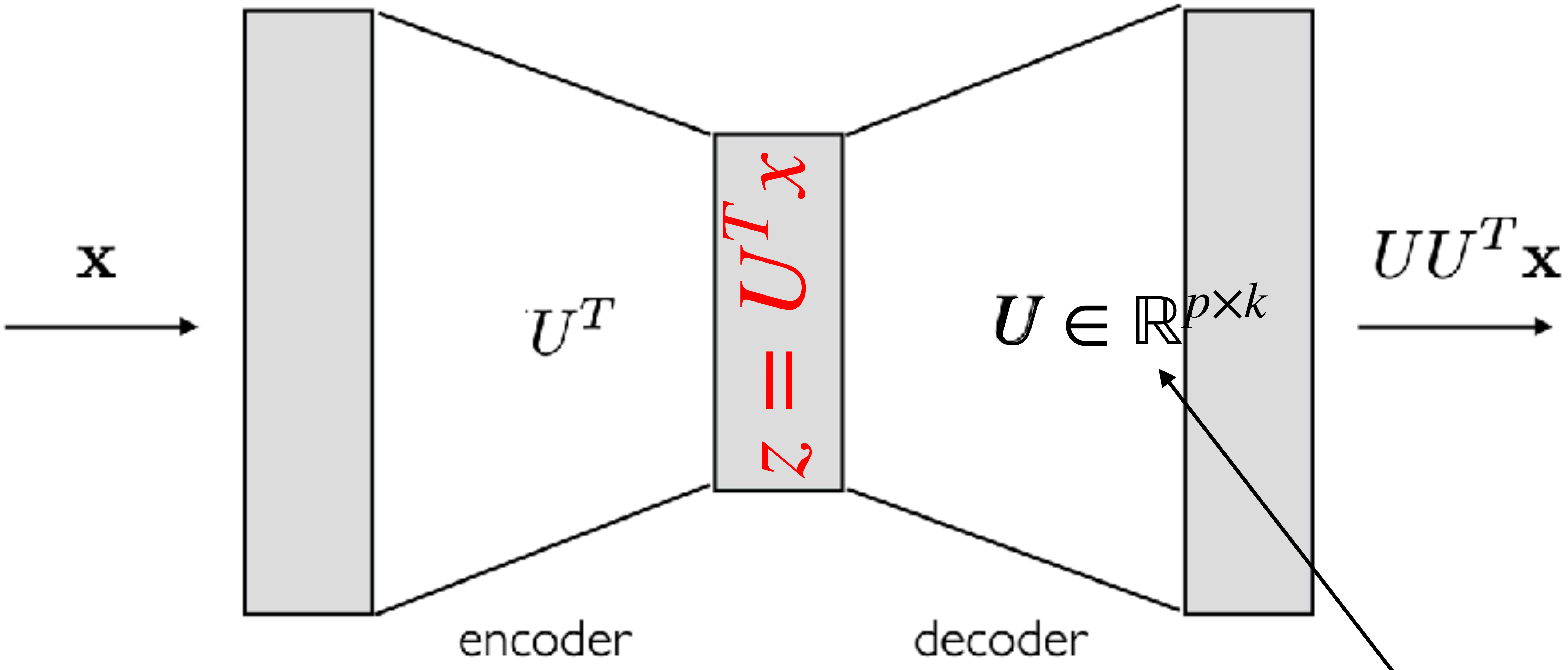
$$\approxeq$$



minimize squared errors (sum of squares of lengths of red segments) incurred by projecting onto affine subspace

Both views essentially the same

PCA as encoder-decoder*

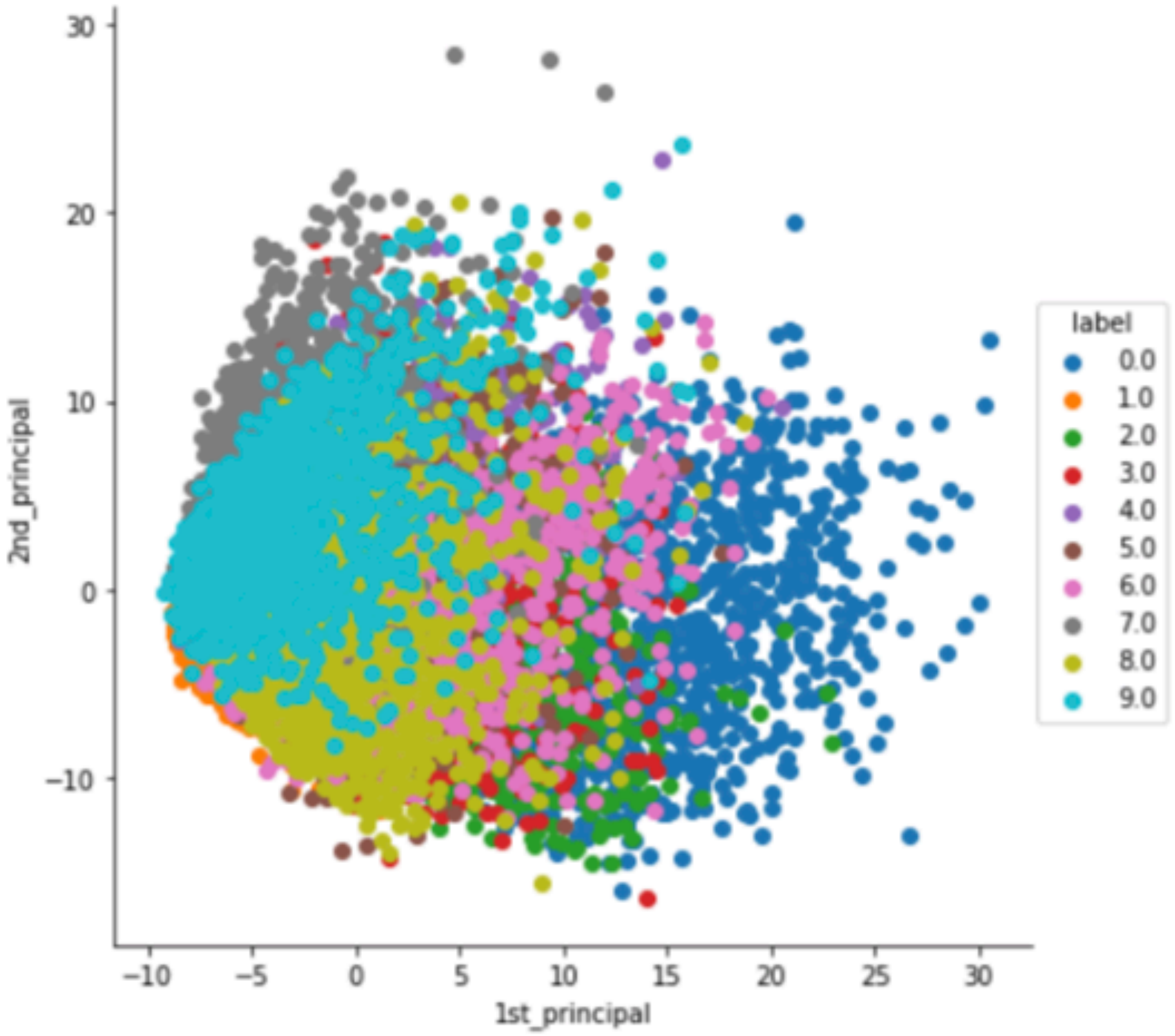


top-k eigenvectors of *S*

Ignoring the mean-shift to zero, roughly view PCA as

- Here $x \in \mathbb{R}^p$ is the original data
- $z \in \mathbb{R}^k$ is the compressed data (latent code)
- $y = UU^T x$ "decodes" *z* back into $\mathbb{R}^p$

PCA based visualization of MNIST

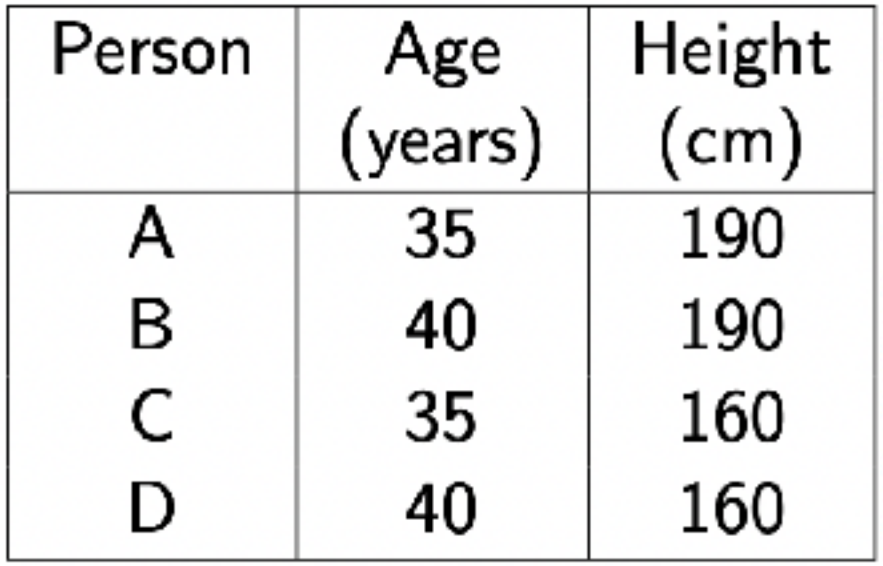


Question: What have we actually plotted here? (Which "space"?)

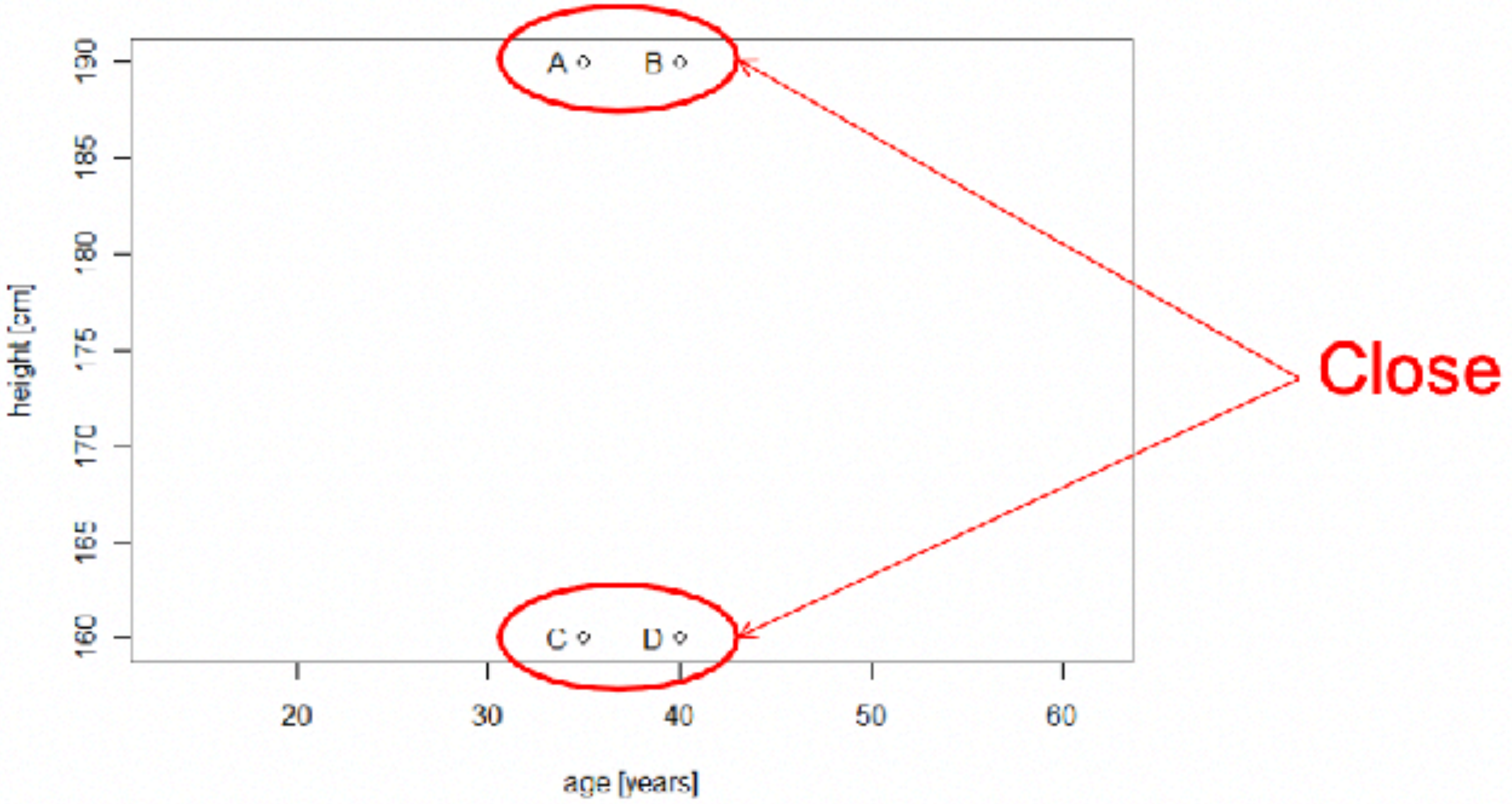
https://medium.com/analytics-vidhya/pca-vs-t-sne-17bcd882bf3d

PCA: Covariance vs Correlation?

- Using covariance PCA finds variable with largest spread as 1st PC
- Use correlation if different units are being compared



Person	Age (years)	Height (cm)
A	35	190
B	40	190
C	35	160
D	40	160

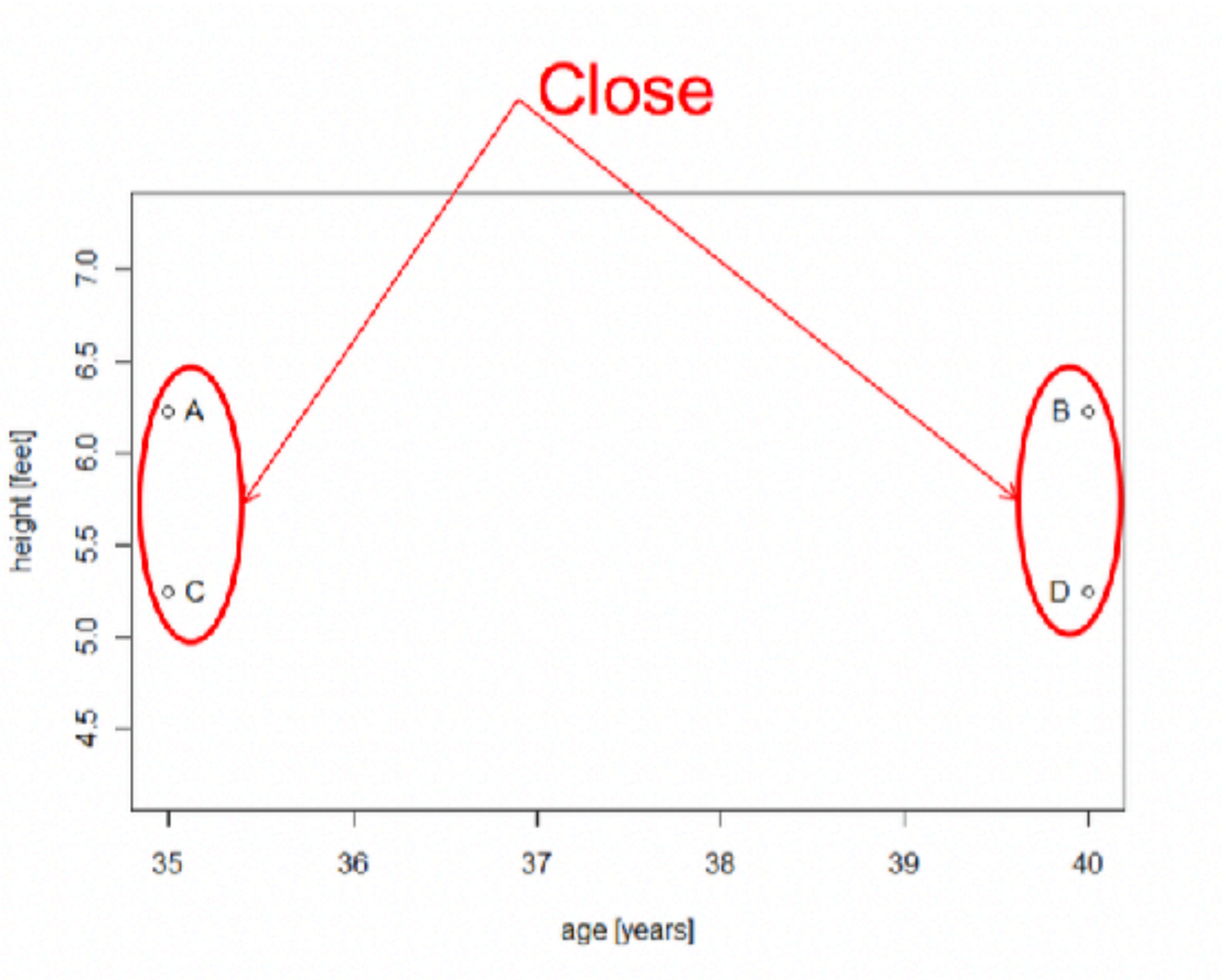


Close

PCA: Covariance vs Correlation?

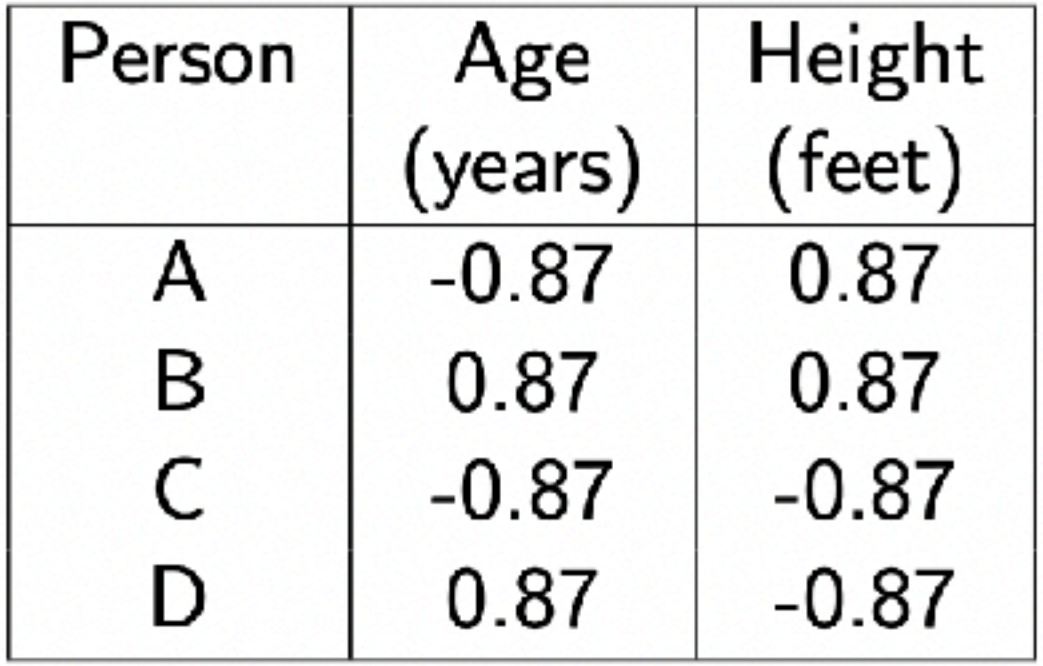
- Using covariance PCA finds variable with largest spread as 1st PC
- Use correlation if different units are being compared

Person	Age (years)	Height (feet)
A	35	6.232
B	40	6.232
C	35	5.248
D	40	5.248

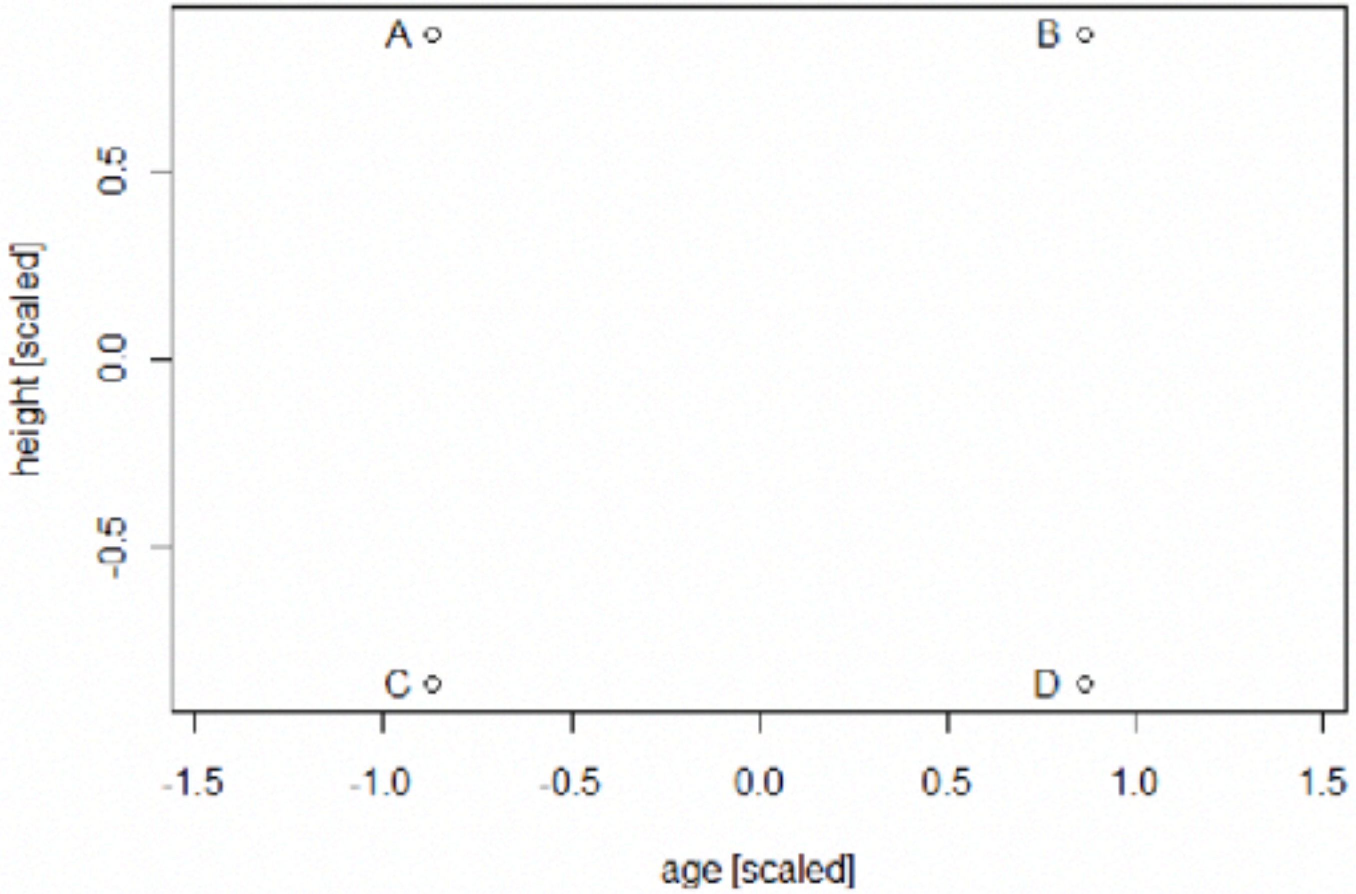


PCA: Covariance vs Correlation?

- Using covariance PCA finds variable with largest spread as 1st PC
- Use correlation if different units are being compared



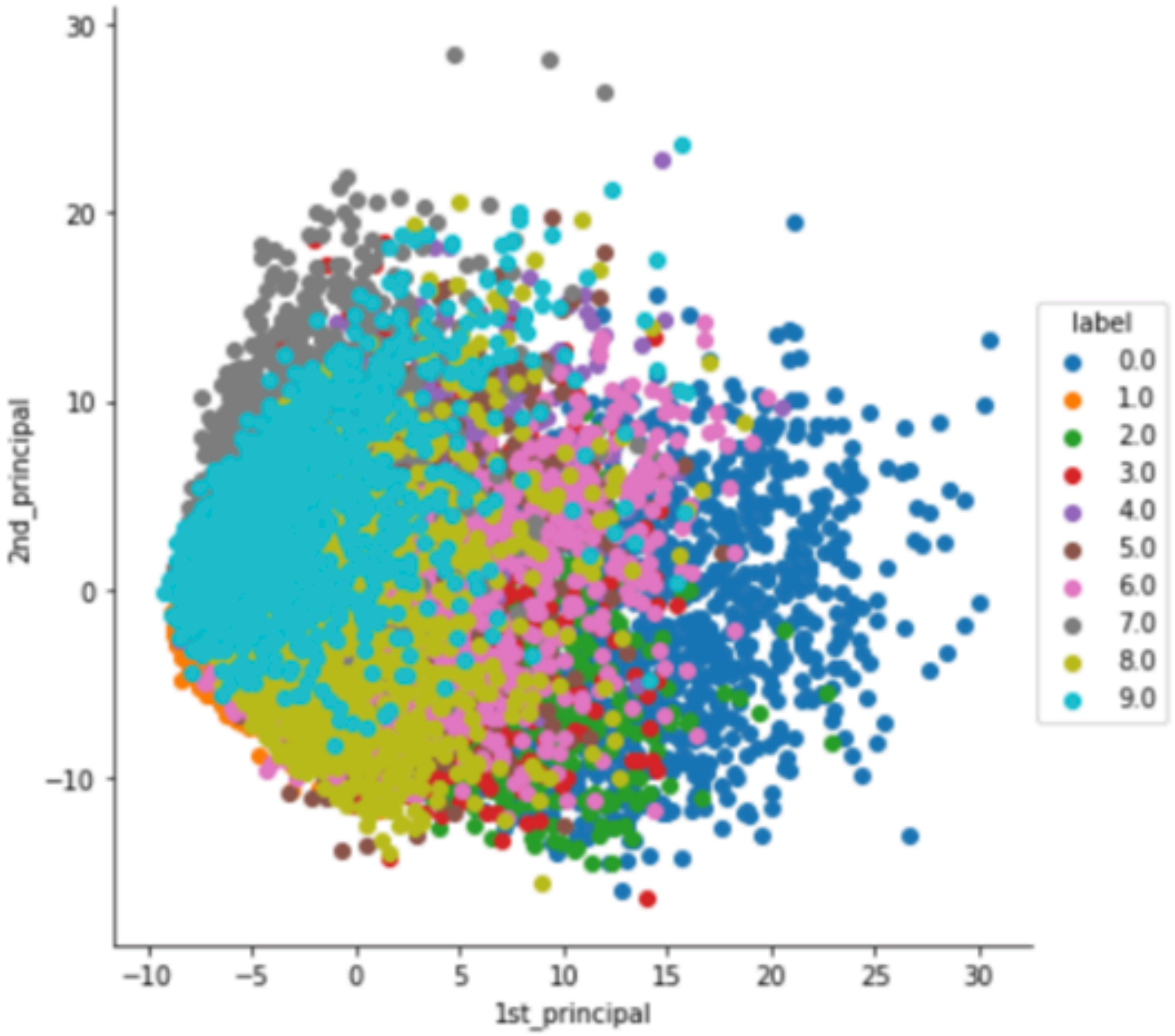
Person	Age (years)	Height (feet)
A	-0.87	0.87
B	0.87	0.87
C	-0.87	-0.87
D	0.87	-0.87



Stochastic Neighbor Embedding (SNE)

(Working with data where we only have $(x_1, \ldots, x_n)$ *instead of* (x_i, y_i) *pairs!)*

Recall PCA visualization of MNIST



https://medium.com/analytics-vidhya/pca-vs-t-sne-17bcd882bf3d

Basics of SNE

PCA does global similarity, potential for suffering from outliers, missing out local structure, however other than ‘k’, it is parameter free and easy to use on “new” data

Want a method sensitive to local structure, possibly by doing nonlinear dim-redux
(structure: local neighbors in high-d space should remain neighbors in low-d)

Key ideas

1. Convert Euclidean distance into conditional probabilities that encode “similarity”
2. For each point, pretend there’s a Gaussian centered at it, and probability of picking a neighbor scales according to euclidean distance

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)};$$

i.e., the prob that point x_i would pick x_j as its neighbor

Where does 't'-SNE come in?

The conditional prob $p_{j|i}$ also very sensitive to outliers (**Why?**)

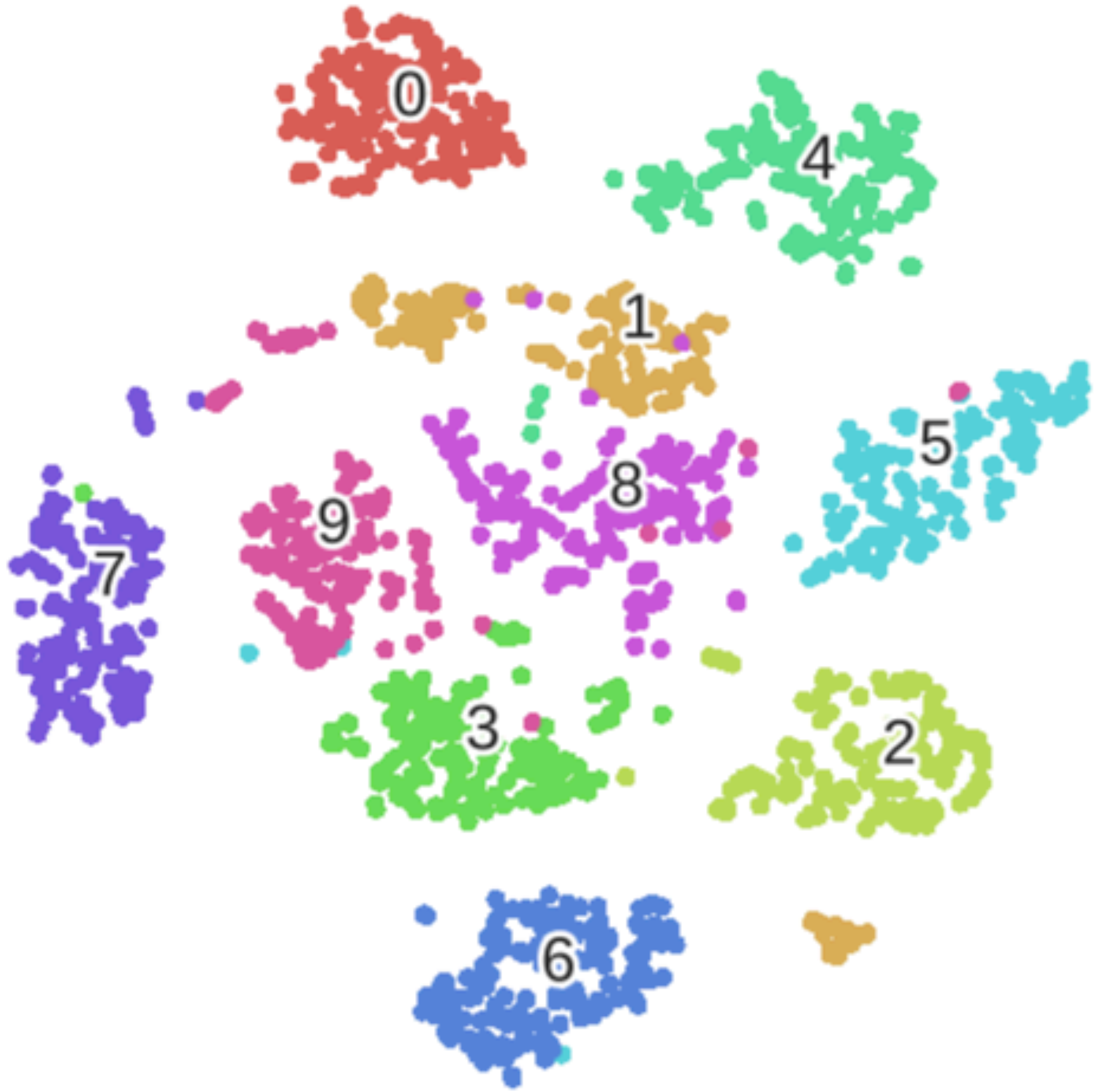
For x_i an outlier, all pairwise distances $\|x_i - x_j\|^2$ large, and $p_{j|i}$ values extremely small, so location of low-dim y_i has little effect on cost function. So location of y_i not well-determined by other points.

Use Student-t distribution instead of Gaussian in mapped (low-d) space

$$q_{j|i} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq i}(1 + \|y_i - y_k\|^2)^{-1}}$$

Key reason: Allows moderate distance in high-d space to be faithfully modeled by a much larger distance in the mapped space, and thereby, eliminates unwanted attraction of points in mapped space that are moderately dissimilar

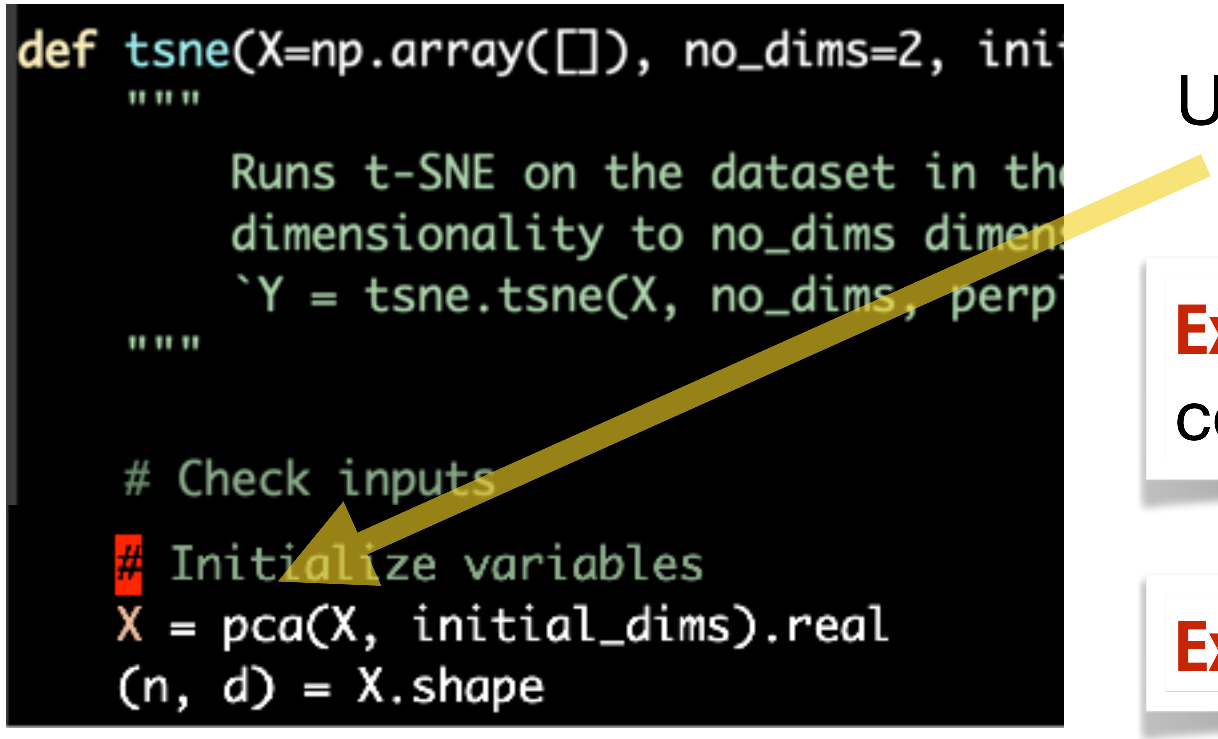
Nonlinear dimension reduction: t-SNE



https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

t-SNE: Some remarks

Less known fact about t-SNE



```
def tsne(X=np.array([]), no_dims=2, ini
    """
        Runs t-SNE on the dataset in th
        dimensionality to no_dims dimen
        `Y = tsne.tsne(X, no_dims, perp
    """

    # Check inputs

    # Initialize variables
    X = pca(X, initial_dims).real
    (n, d) = X.shape
```

Uses PCA to initialize

from original tsne.py implementation

Exercise: Init. t-SNE using k-means and compare visualizations with PCA choice

Exercise: Discuss Pros and Cons of t-SNE

Exercise*: What could go wrong if input dimensionality is quite high?

Clustering

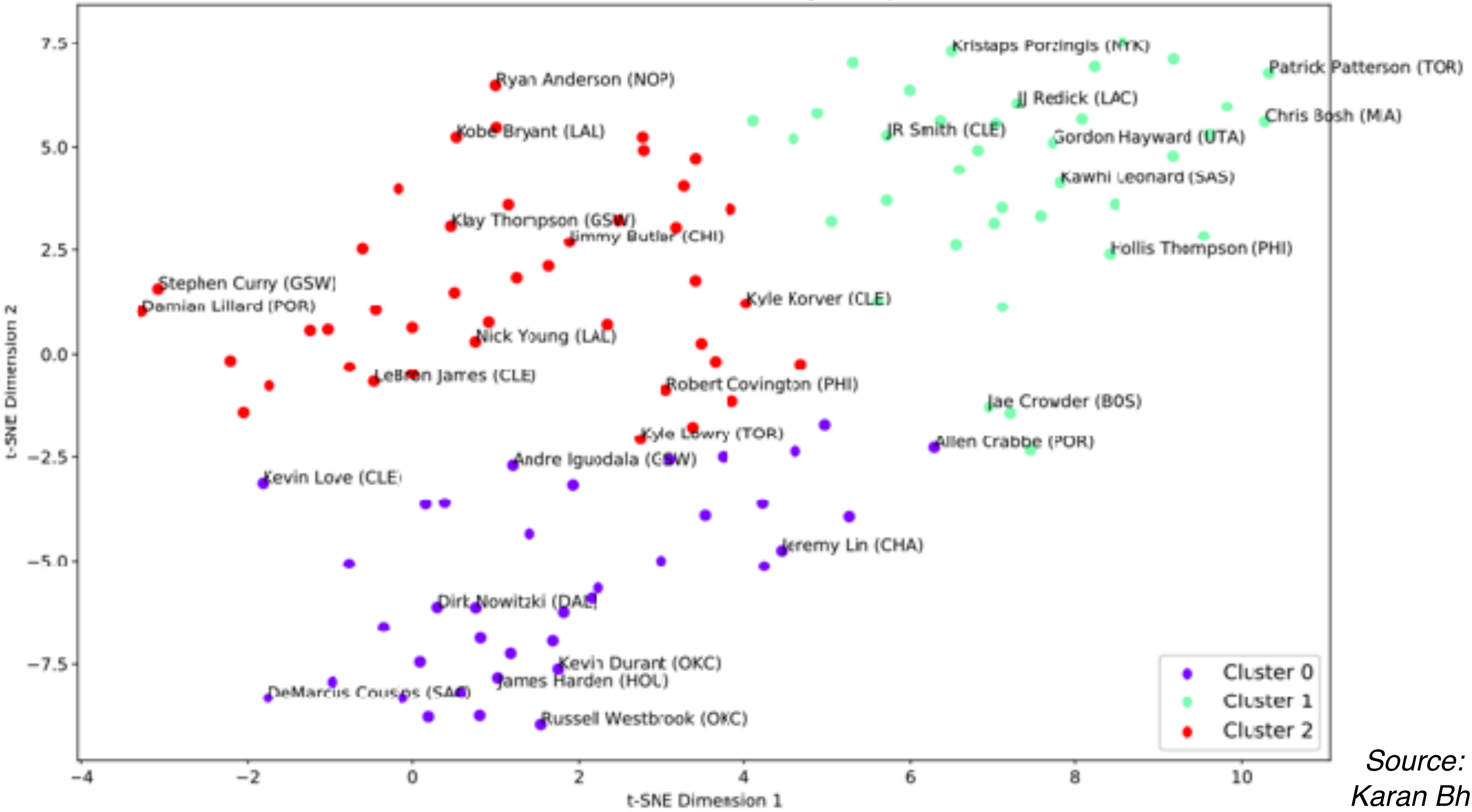
(Working with data where we only have $(x_1, \ldots, x_n)$ *instead of* (x_i, y_i) *pairs!)*

Clustering examples

- News recommendation
- Finding groups of similar customers / players
- Compression
- Creating features
- Semi-supervised learning

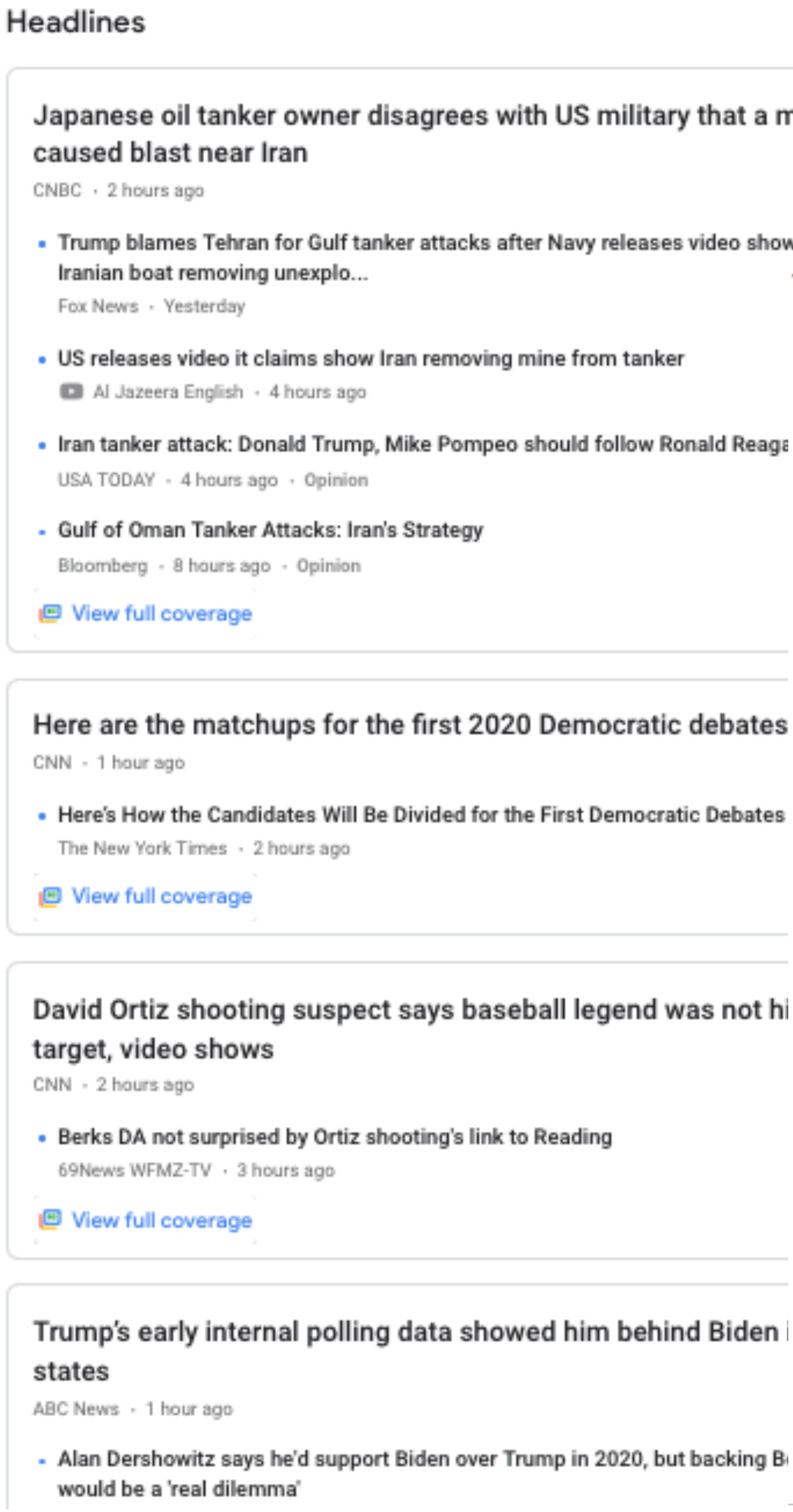
Clustering examples

2015-2016 3-Point Shooters Clustered by Attempt Features



Source: IDS.012 class project by Nate Bailey, Karan Bhuwalka, Hin Lee, Tim Zhong

Clustering examples



Headlines

More Headlines

Japanese oil tanker owner disagrees with US military that a mine caused blast near Iran

CNBC · 2 hours ago

- Trump blames Tehran for Gulf tanker attacks after Navy releases video showing Iranian boat removing unexplo...

  Fox News · Yesterday

- US releases video it claims show Iran removing mine from tanker

  Al Jazeera English · 4 hours ago

- Iran tanker attack: Donald Trump, Mike Pompeo should follow Ronald Reaga

  USA TODAY · 4 hours ago · Opinion

- Gulf of Oman Tanker Attacks: Iran's Strategy

  Bloomberg · 8 hours ago · Opinion

View full coverage

Here are the matchups for the first 2020 Democratic debates

CNN · 1 hour ago

- Here's How the Candidates Will Be Divided for the First Democratic Debates

  The New York Times · 2 hours ago

View full coverage

David Ortiz shooting suspect says baseball legend was not hi target, video shows

CNN · 2 hours ago

- Berks DA not surprised by Ortiz shooting's link to Reading

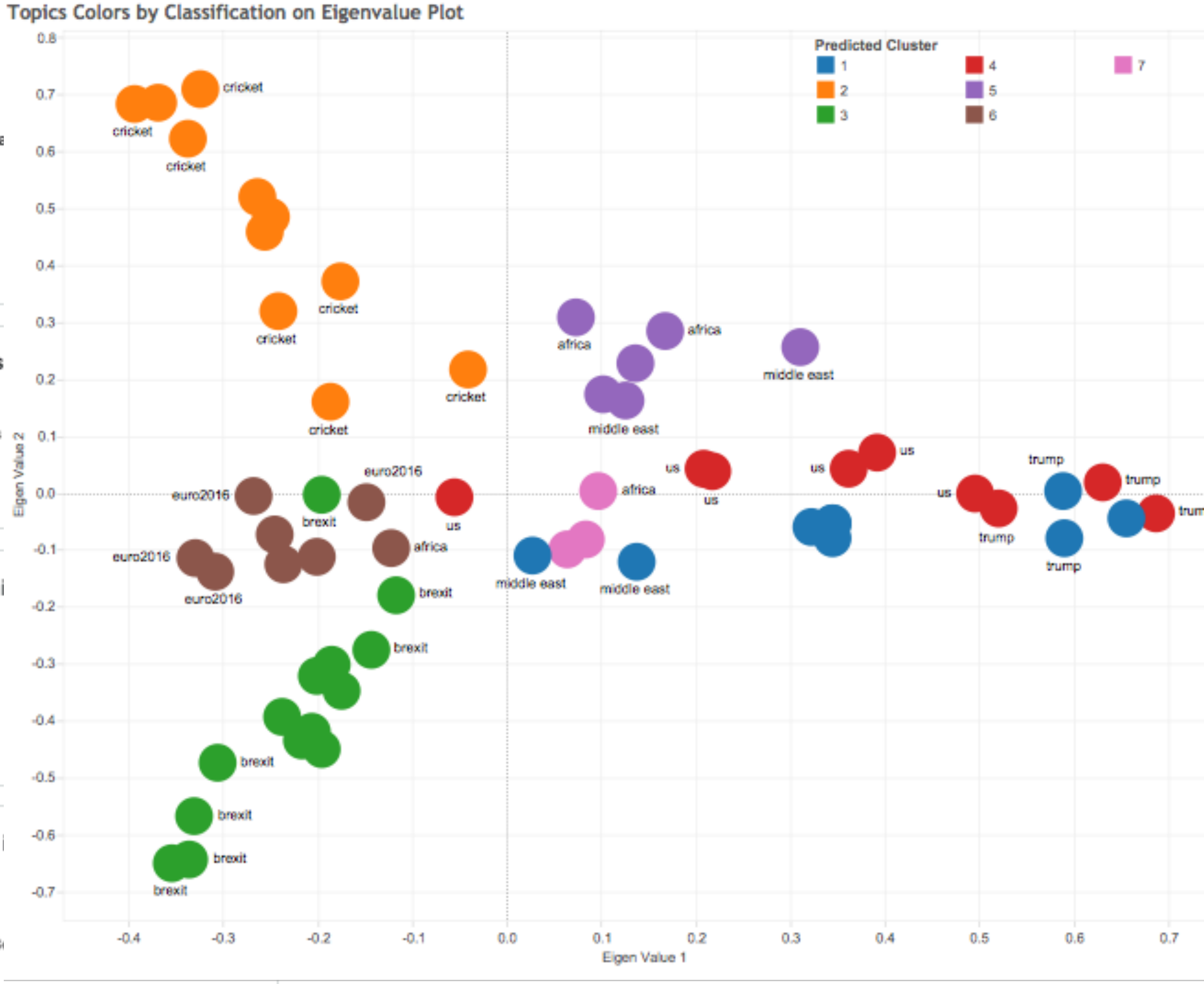
  69News WFMZ-TV · 3 hours ago

View full coverage

Trump's early internal polling data showed him behind Biden i states

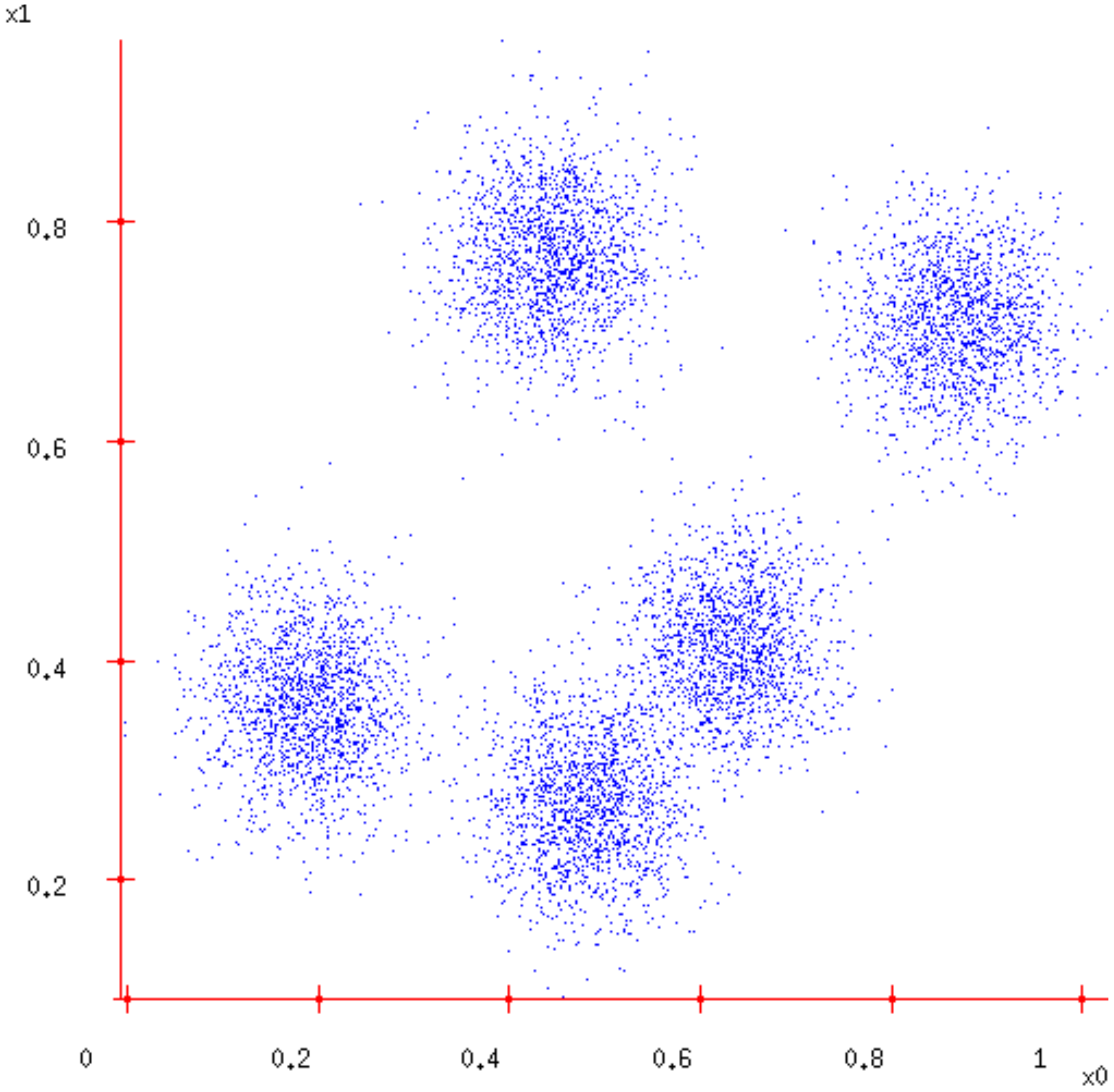
ABC News · 1 hour ago

- Alan Dershowitz says he'd support Biden over Trump in 2020, but backing B would be a 'real dilemma'



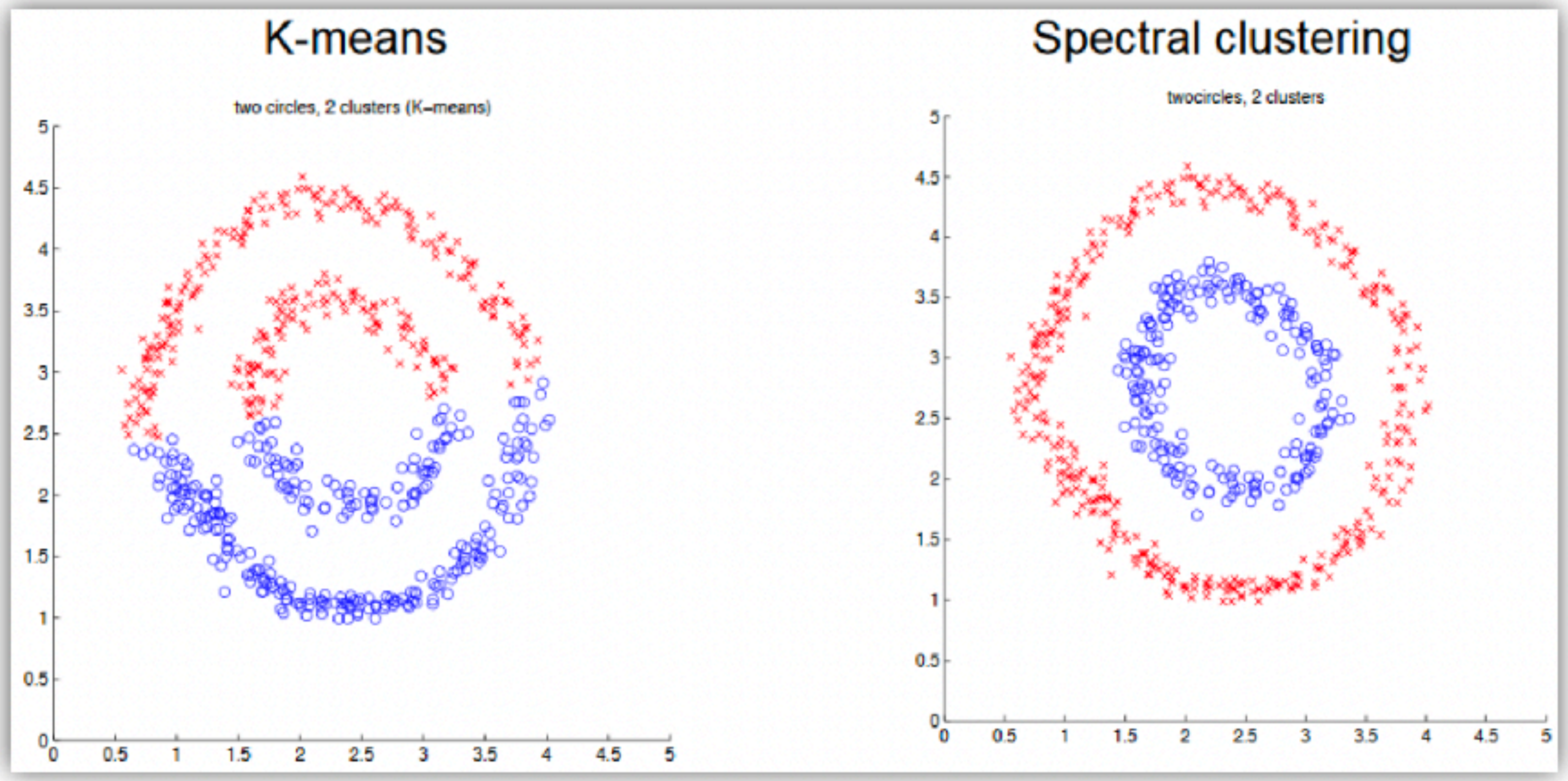
Topics Colors by Classification on Eigenvalue Plot

What is a good clustering?

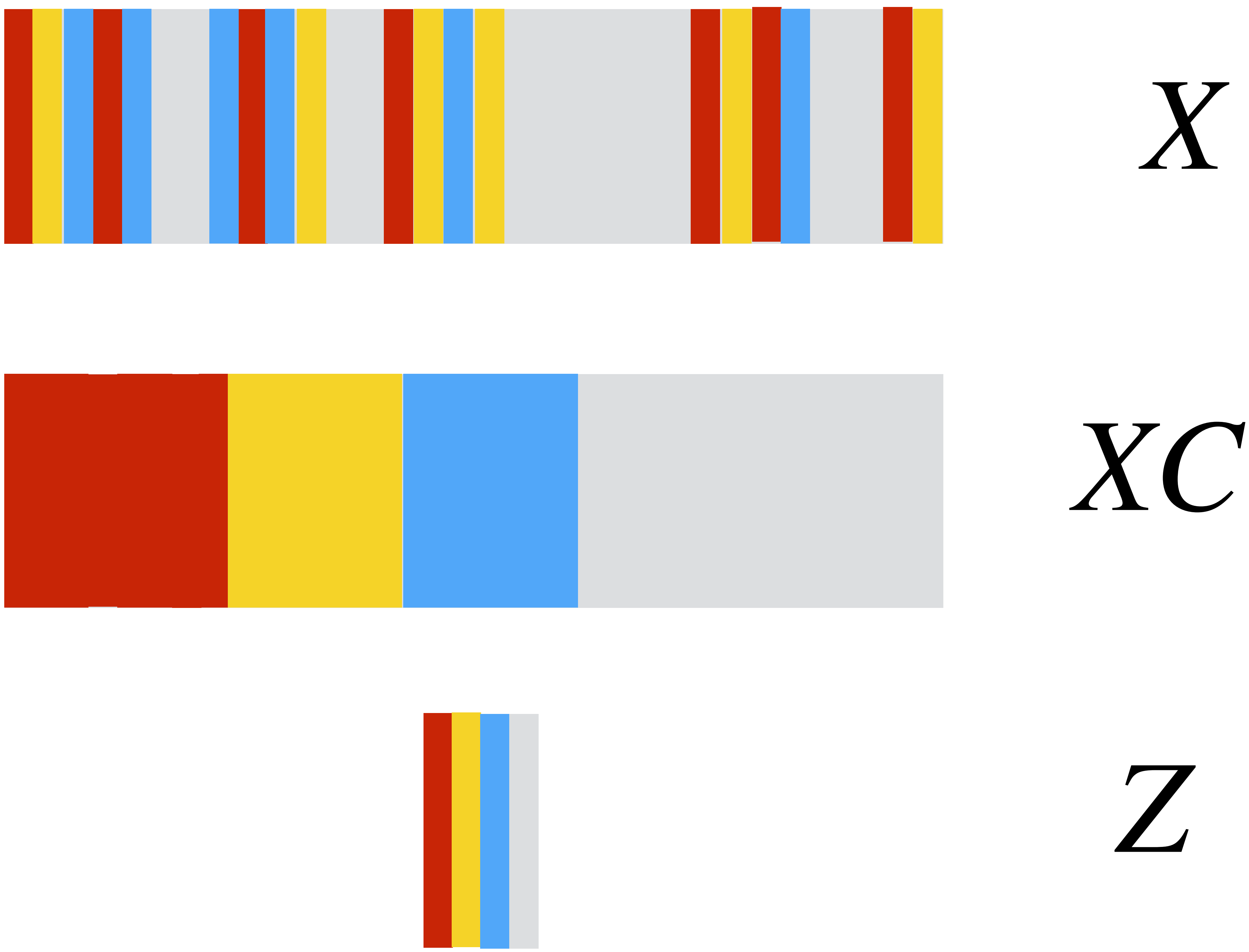


What is a good clustering?

http://cs.nyu.edu/~dsontag/courses/ml13/slides/lecture16.pdf



Clustering columns of a matrix



Clustering: Formal Setup

- Unlabeled data points $x_1, \ldots, x_n$
- Find: clusters $C_1, \ldots, C_k$
  and one representative for each cluster: $z_1, \ldots, z_k$
- Optimization formulation for clustering:

$$\sum_{j=1}^{k} \sum_{x \in C_j} \operatorname{dist}(x, z_j)$$

- **Important:** how do we measure distance?

Common distance measures

- squared ℓ_2-distance

$$\mathrm{dist}(x_i, x_j) = \|x_i - x_j\|^2$$

- ℓ_1-distance (more robust)

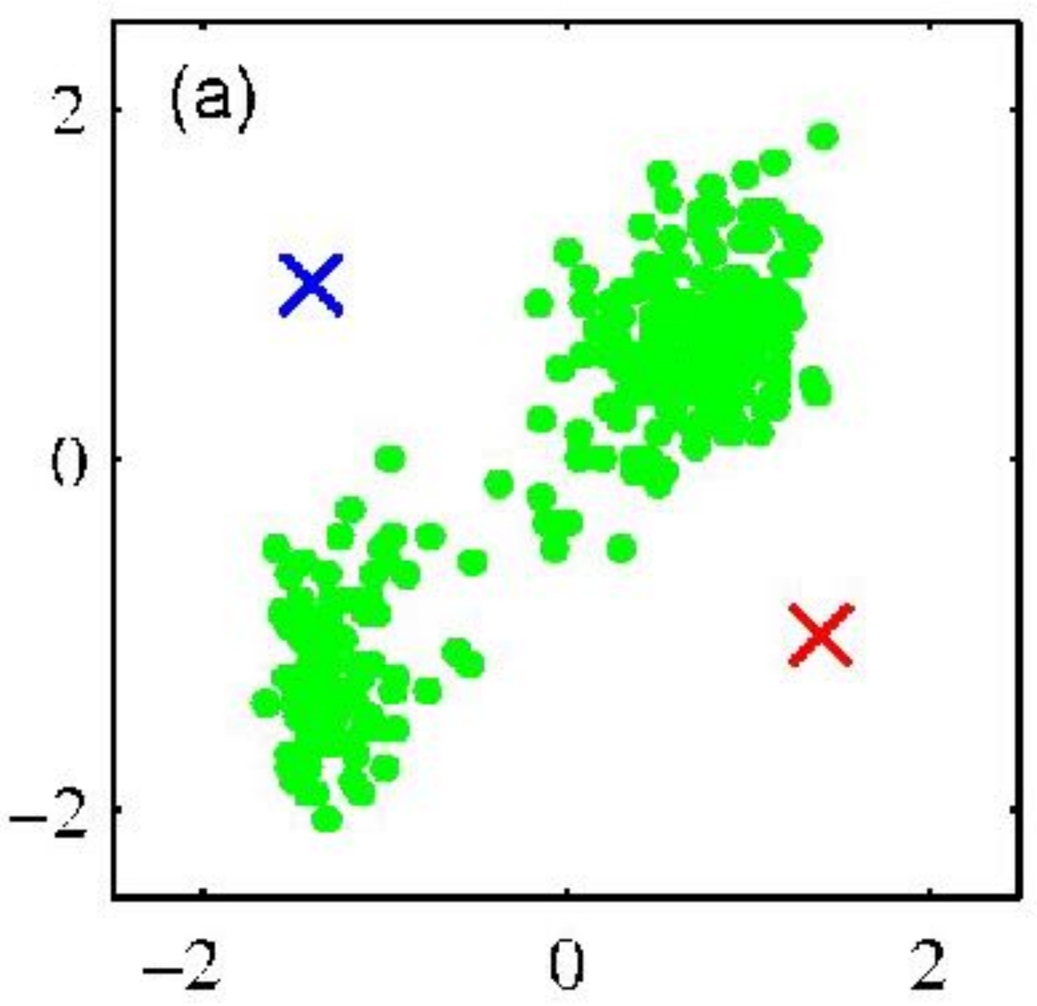
$$\mathrm{dist}(x_i, x_j) = \|x_i - x_j\|_1$$

- Cosine similarity (relates to Pearson correlation coefficient)

$$cos(x^{(i)}, x^{(j)}) = \frac{x^{(i)} \cdot x^{(j)}}{\|x^{(i)}\| \|x^{(j)}\|} = \frac{\sum_{l=1}^{d} x_l^{(i)} x_l^{(j)}}{\sqrt{\sum_{l=1}^{d} (x_l^{(i)})^2} \sqrt{\sum_{l=1}^{d} (x_l^{(j)})^2}}$$

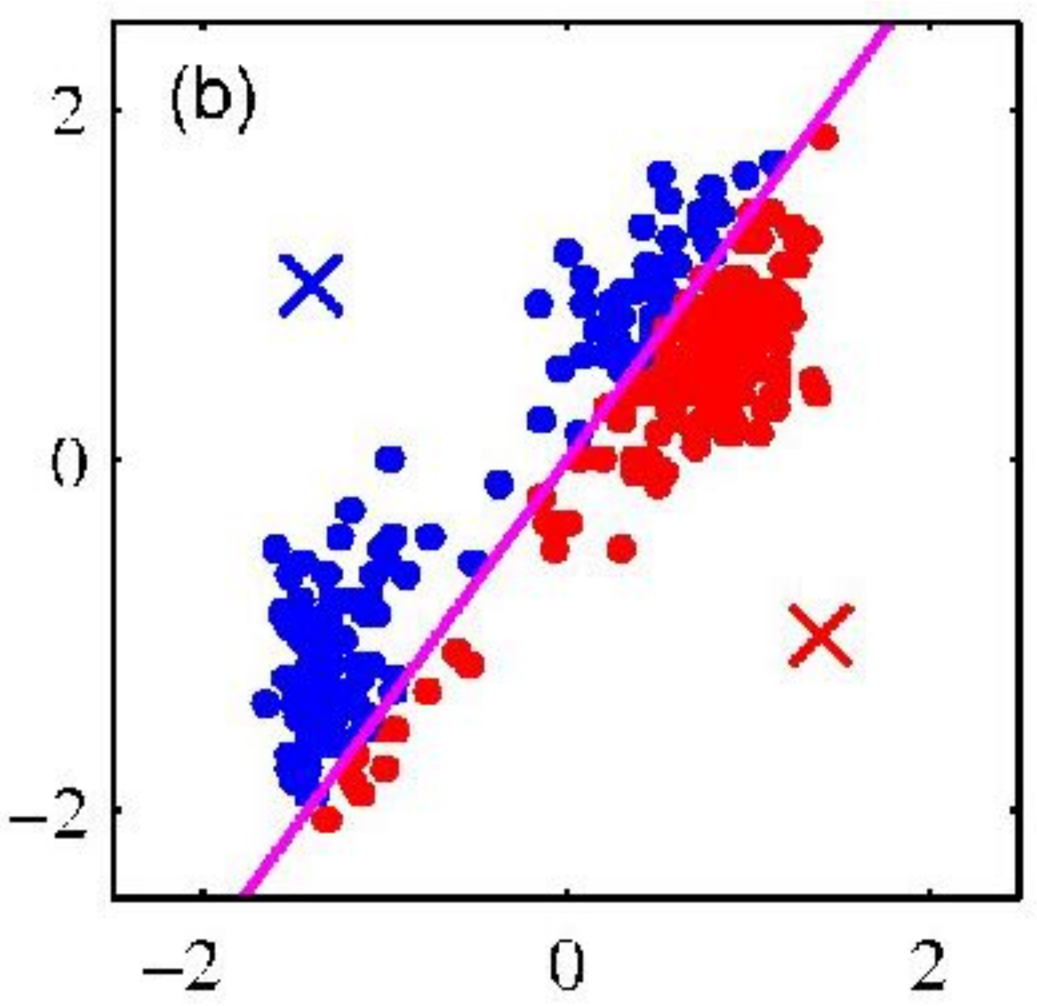
Find best cluster assignment

$$\min \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2$$



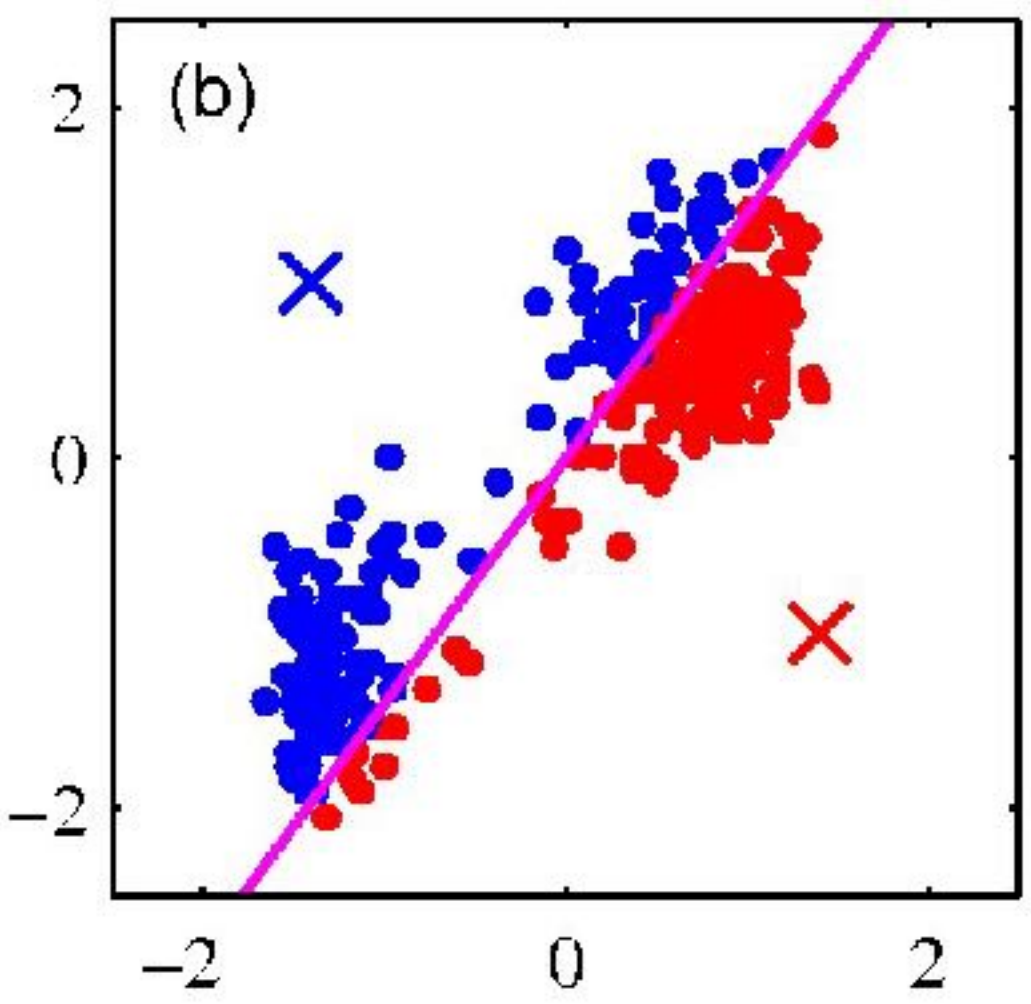
Find best cluster assignment

$$\min \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2$$



Find the best centroid for each cluster

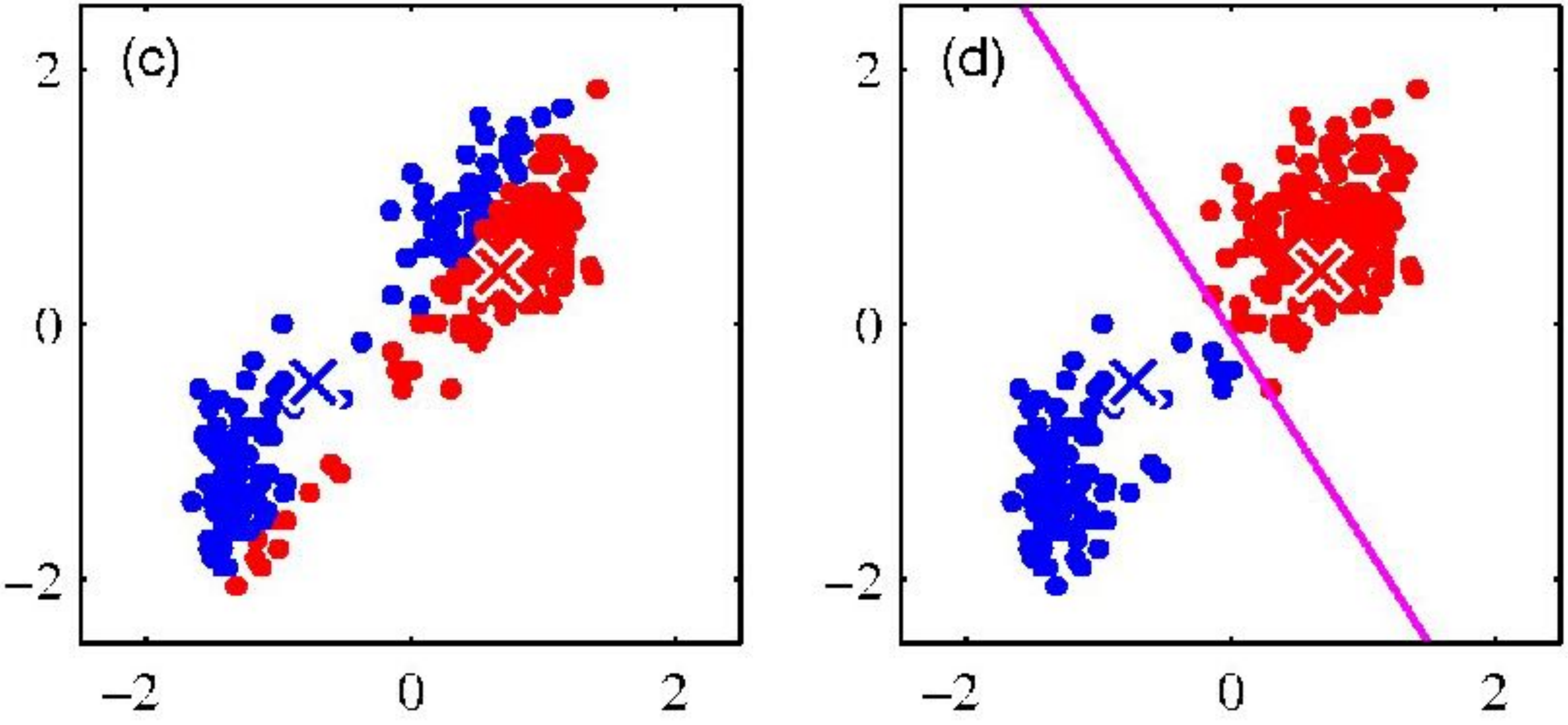
$$\min \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2$$



Find the best centroid for each cluster

$$\min \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2$$

After reassigning points



K-means algorithm

1. Initialize centroids $z_1, \ldots, z_k$
2. Repeat until there is no more change in cost:
   - Given $z_1, \ldots, z_k$, find best cluster assignments of points
   - Given cluster assignments, find the best centroids

Finding the best centroid for a cluster

$$\min \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2$$

- *With fixed cluster assignments, best centroid is the cluster mean / average of the points in the cluster*

$$z_j = \frac{1}{|C_j|} \sum_{i \in C_j} x_i$$

K-means Algorithm

1. Initialize centroids $z_1, \ldots, z_k$

2. Repeat until there is no more change in cost:

   - Given $z_1, \ldots, z_k$, find best cluster assignments of points: *assign each point to its closest centroid*

   - Given cluster assignments, find the best centroids: *cluster means*

Question: *What happens to the cost function during the algorithm?*