IBM Research Data Science Course

{name:srihari sridharan}
{areas of interest:multitask learning,
neural networks,
parameter space geometry}

Overview

Learning

Unsupervised Learning

Clustering

- K-Means Clustering
- Hierarchical Clustering

Dimensionality Reduction

- Principle Component Analysis

Overview

High School Mathematics

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

Self Aware Artificial Intelligence to destroy Humanity

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

$$A = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

Machine Learning

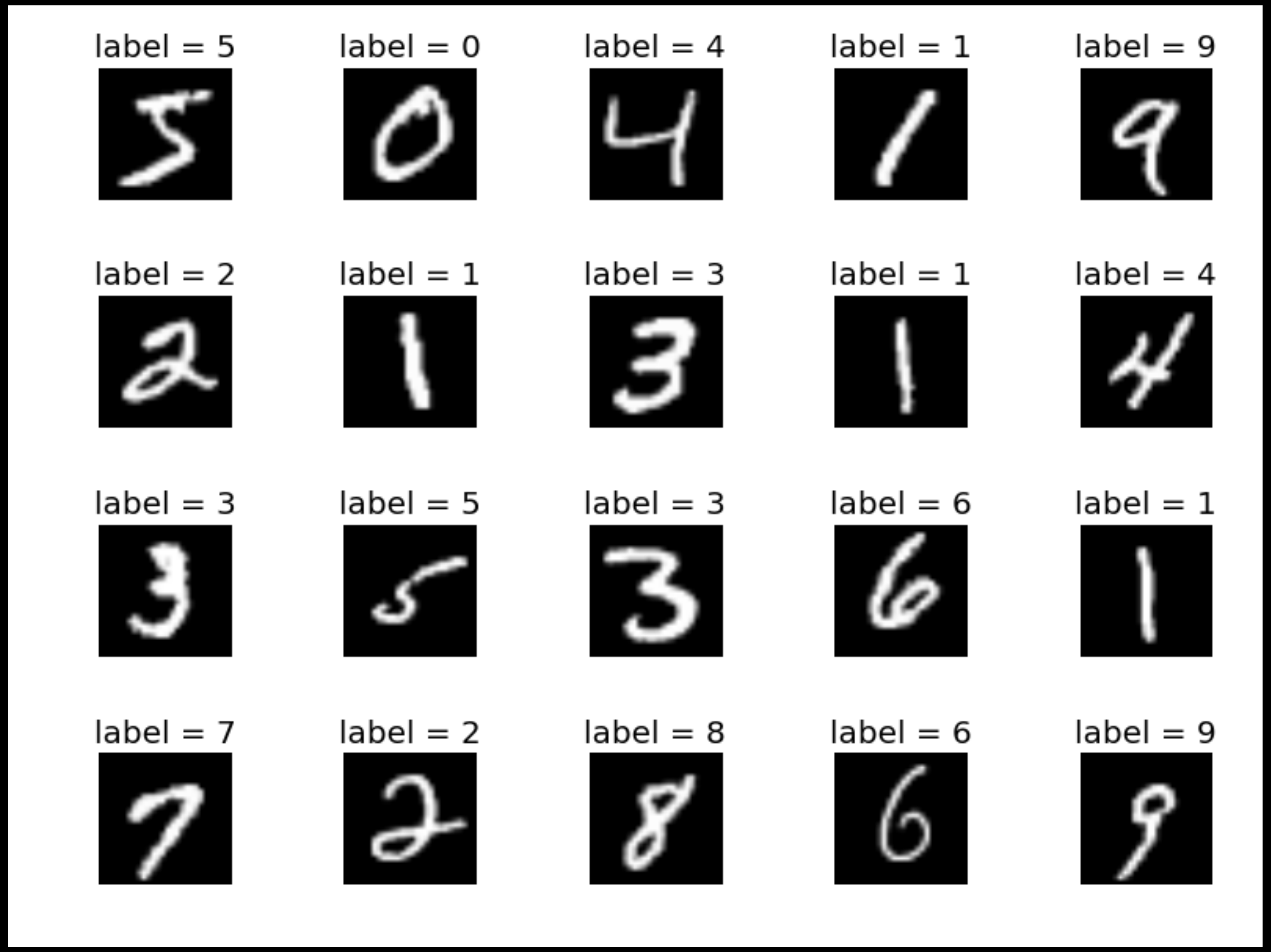
A computer program is said to **learn from experience E** with respect to some **class of tasks T** and performance measure P, if its performance at tasks in T, as measured by P, **improves with experience E**.

// Mitchell, 1997

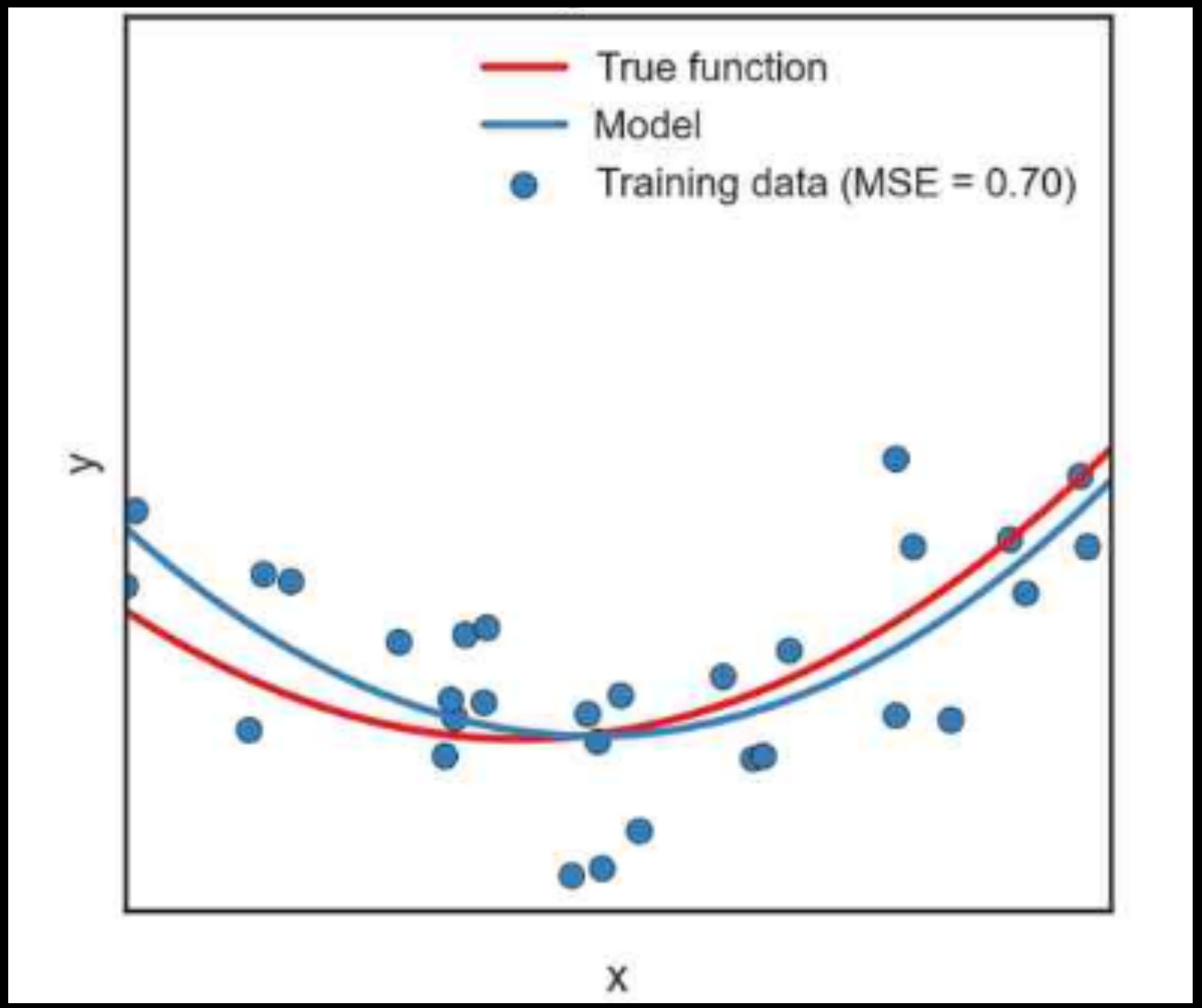
Supervised Learning

Given (x^1, y^1), (x^2, y^2), … , (x^n, y^n), choose a function $f(x) = y$

- Classification : $y_n \in$ {finite set}



- Regression : $y_n \in \mathbb{R}$ or $\mathbb{R}^d$



Classification

Features	Labels
	bird
	cat
	dog
	sportscar
	toilet-tissue

// YOLO : Real Time Object Detection, Joseph Redmon

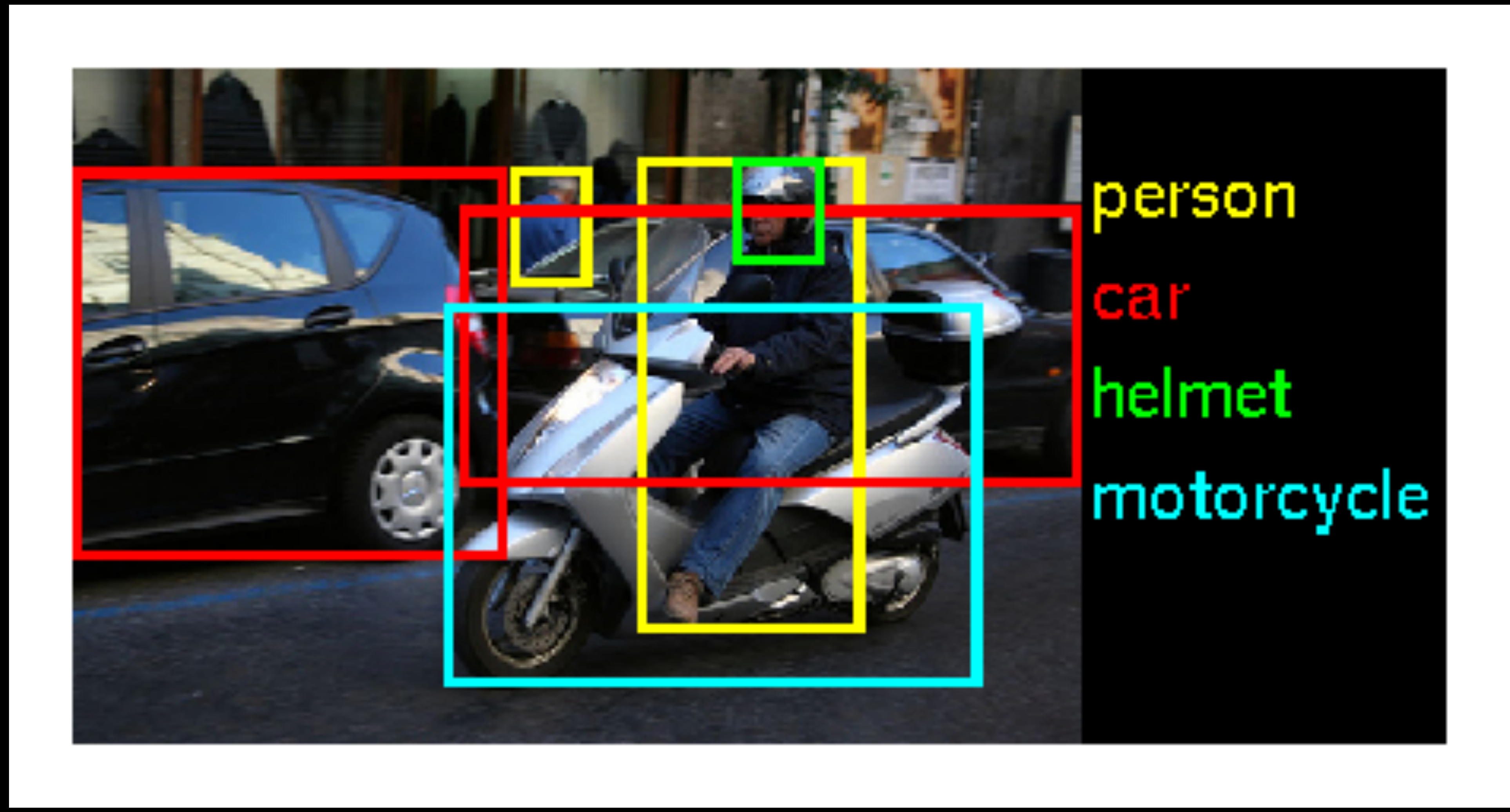
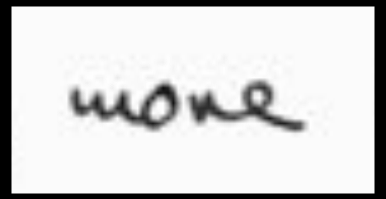
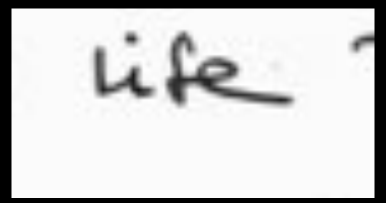
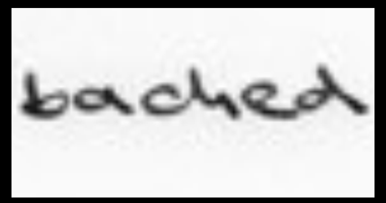
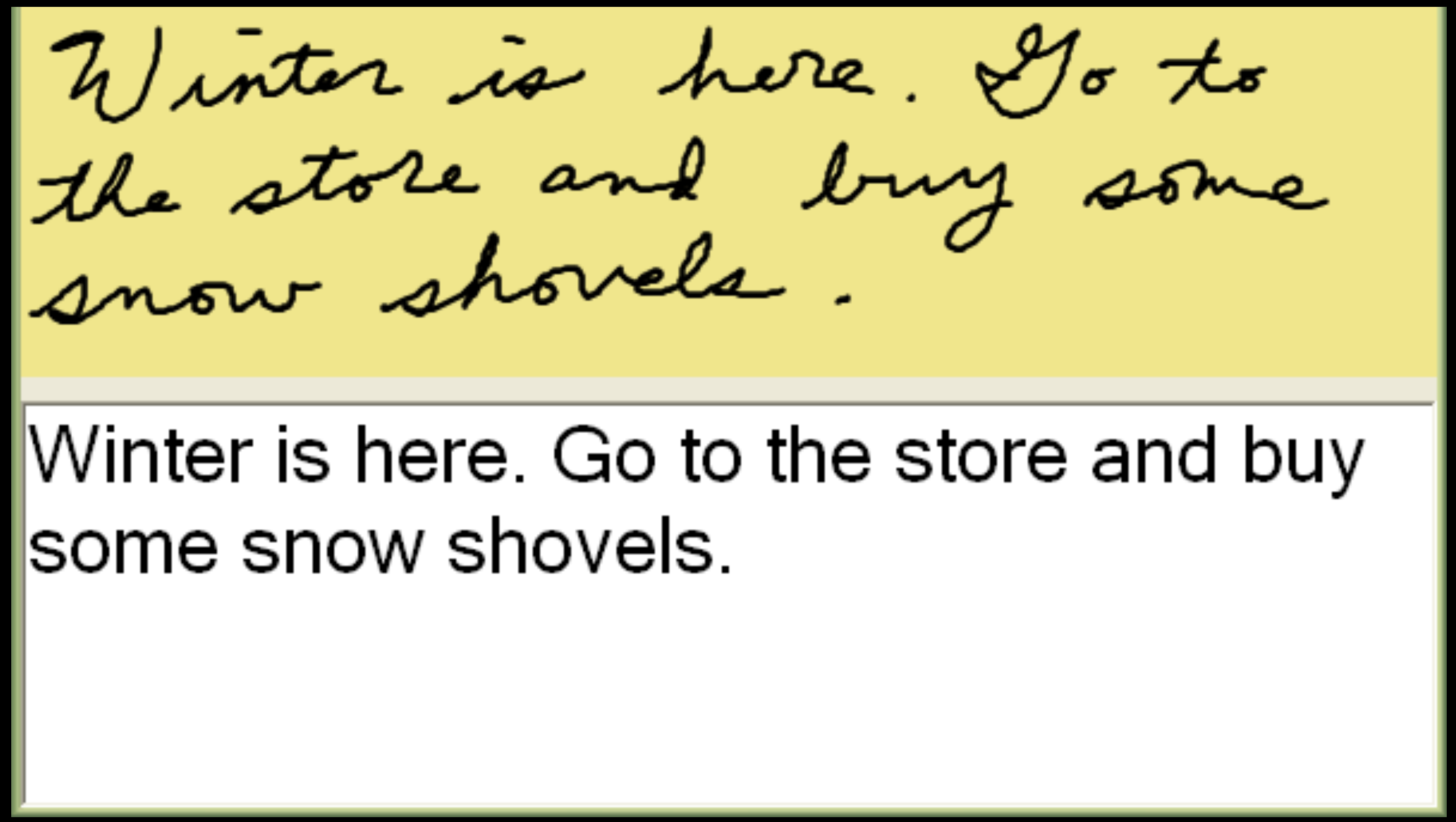


Image Recognition,
Object Detection

Classification

Features	Labels
	MOVE
	stop
	more
	life
	backed

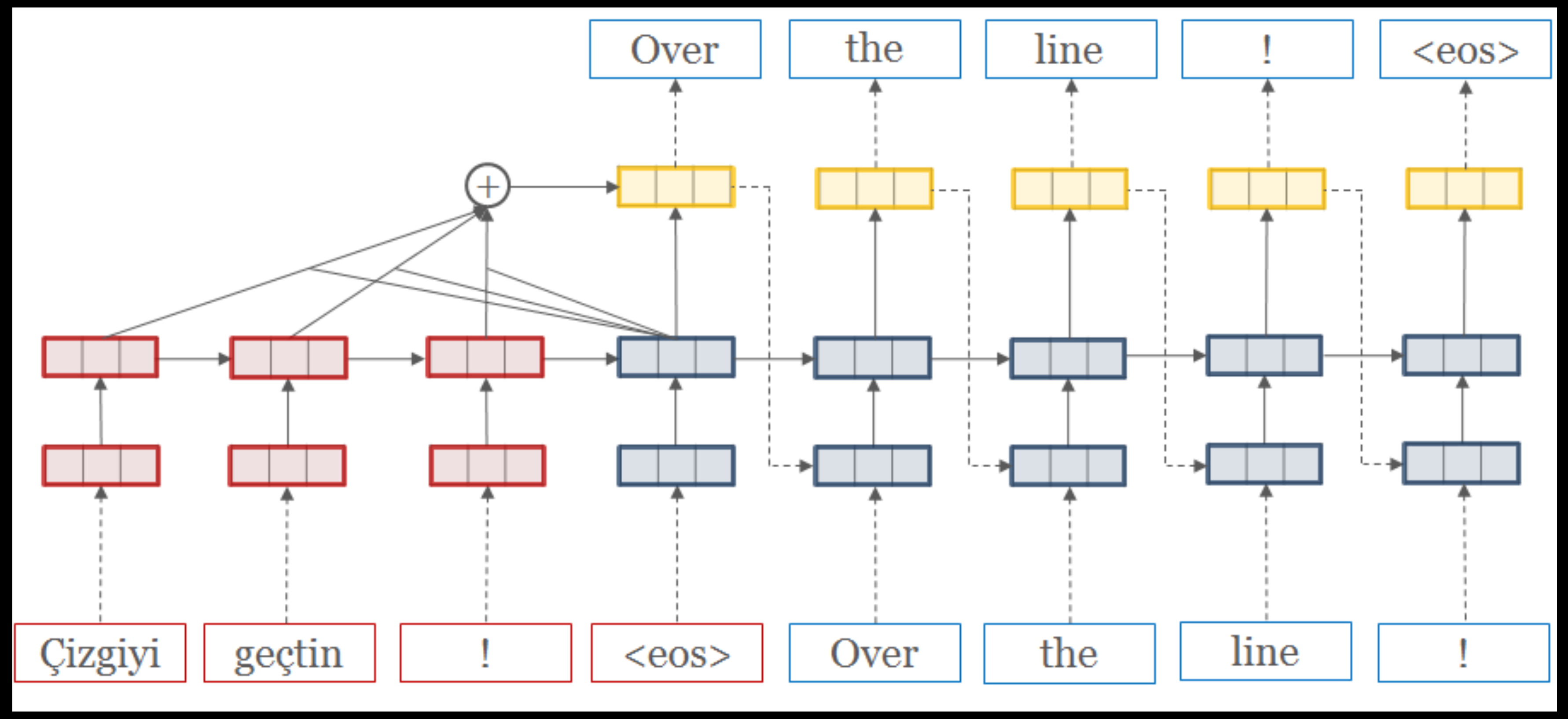


Handwriting recognition

Machine Translation

Features	Labels
Knowledge	知
	识
is	就
	是
Power	力
	量
<end>	。

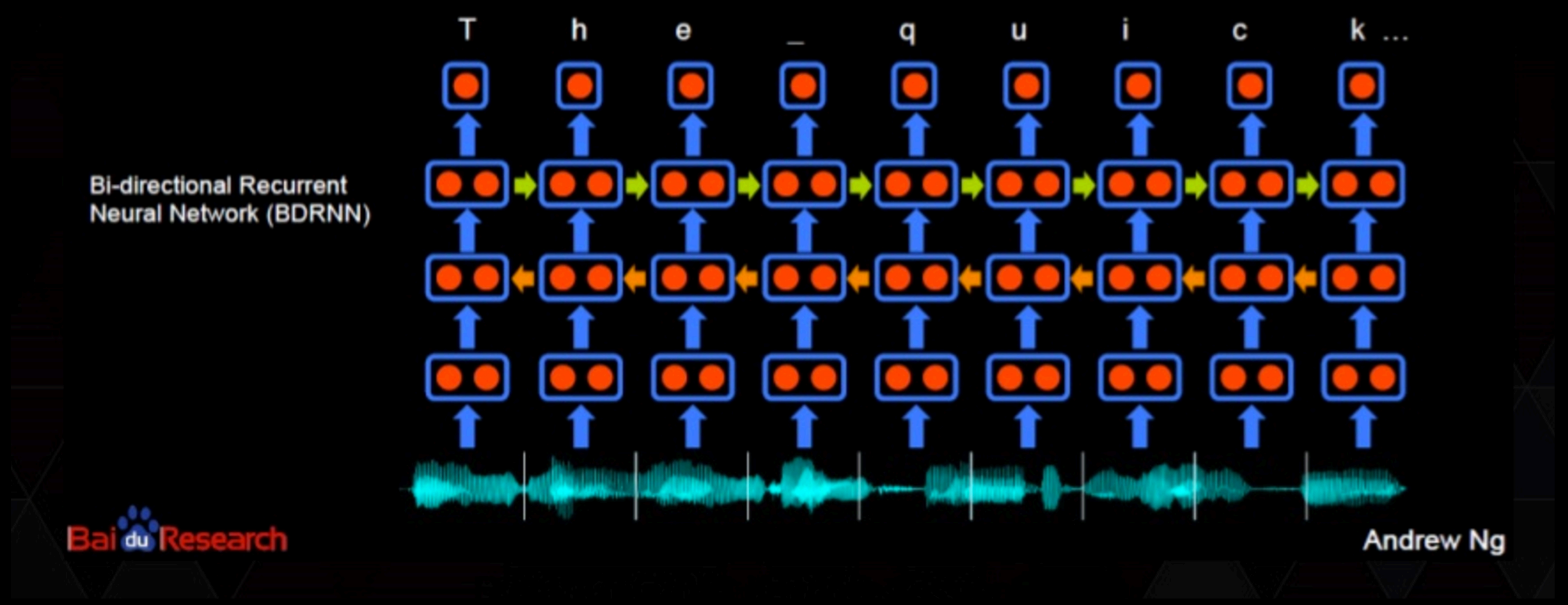
// OpenNMT (Neural Machine Translation)



Machine Translation

Classification

Features	Labels
	T h e _ q



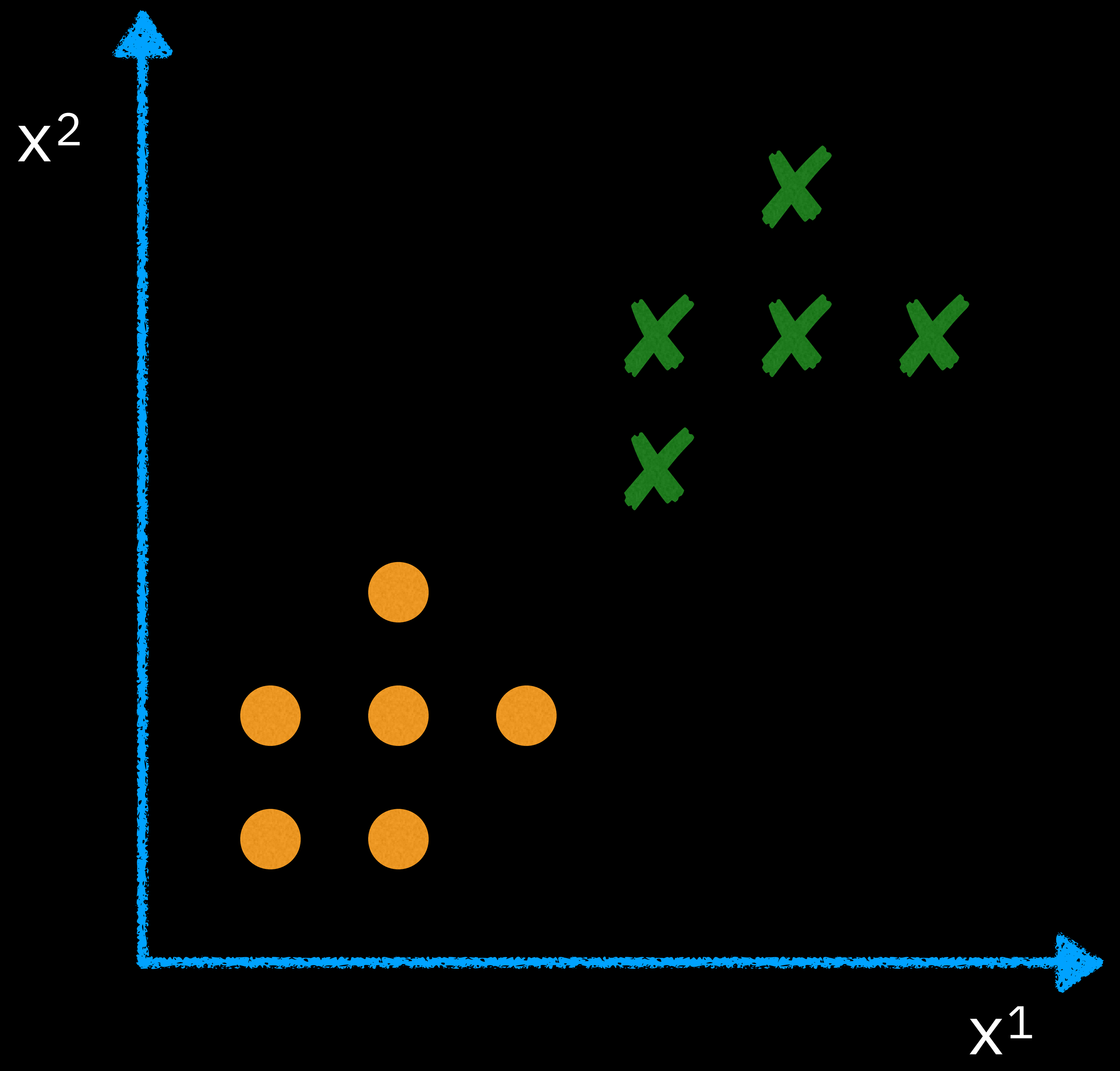
Speech Recognition

// Baidu Research

Supervised Learning

Data :

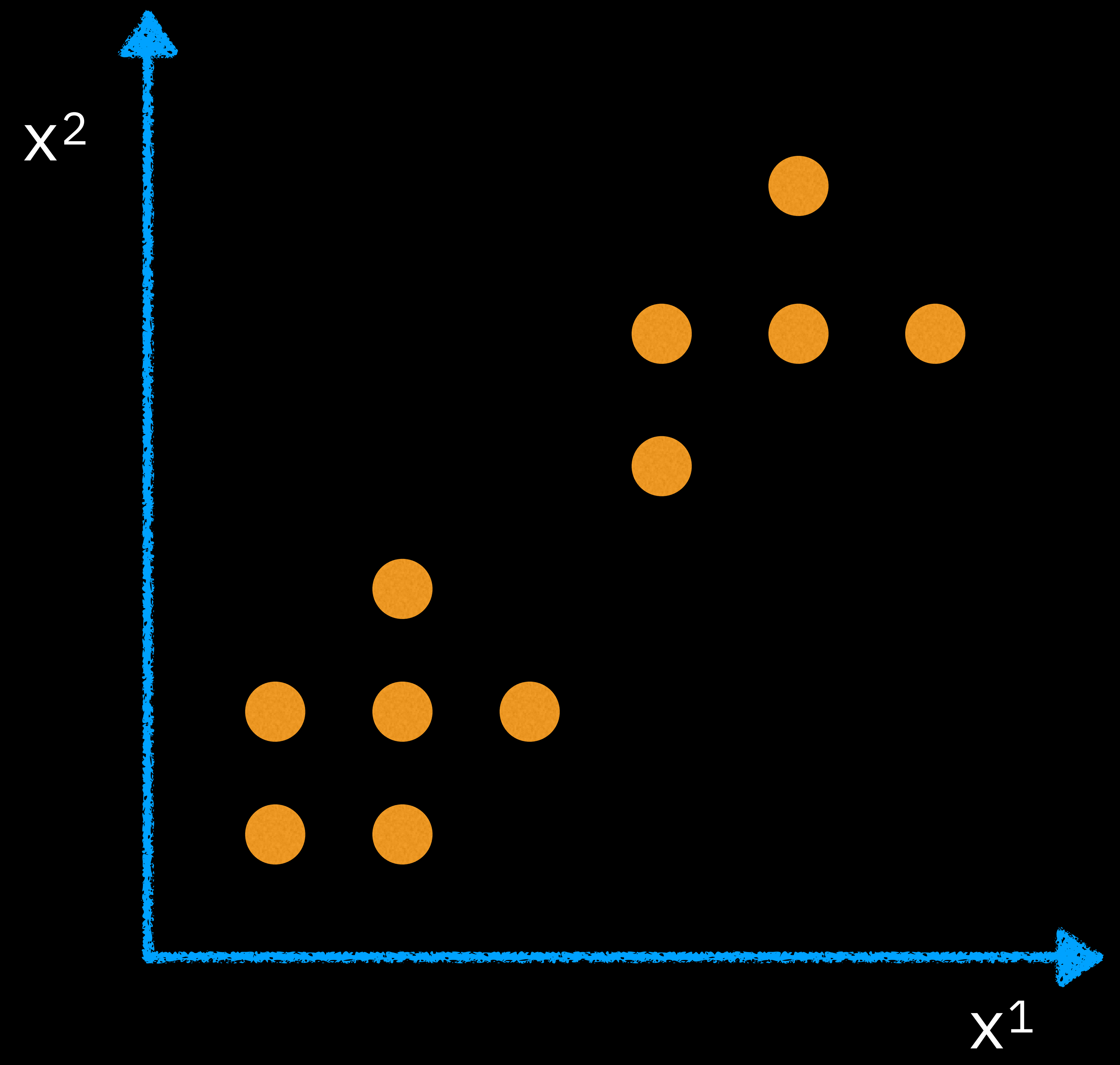
{1,1} : No
{2,1} : No
{1,2} : No
{2,2} : No
{3,2} : No
{2,3} : No
{4,4} : Yes
{4,5} : Yes
{5,5} : Yes
{5,6} : Yes
{6,5} : Yes



Unsupervised Learning

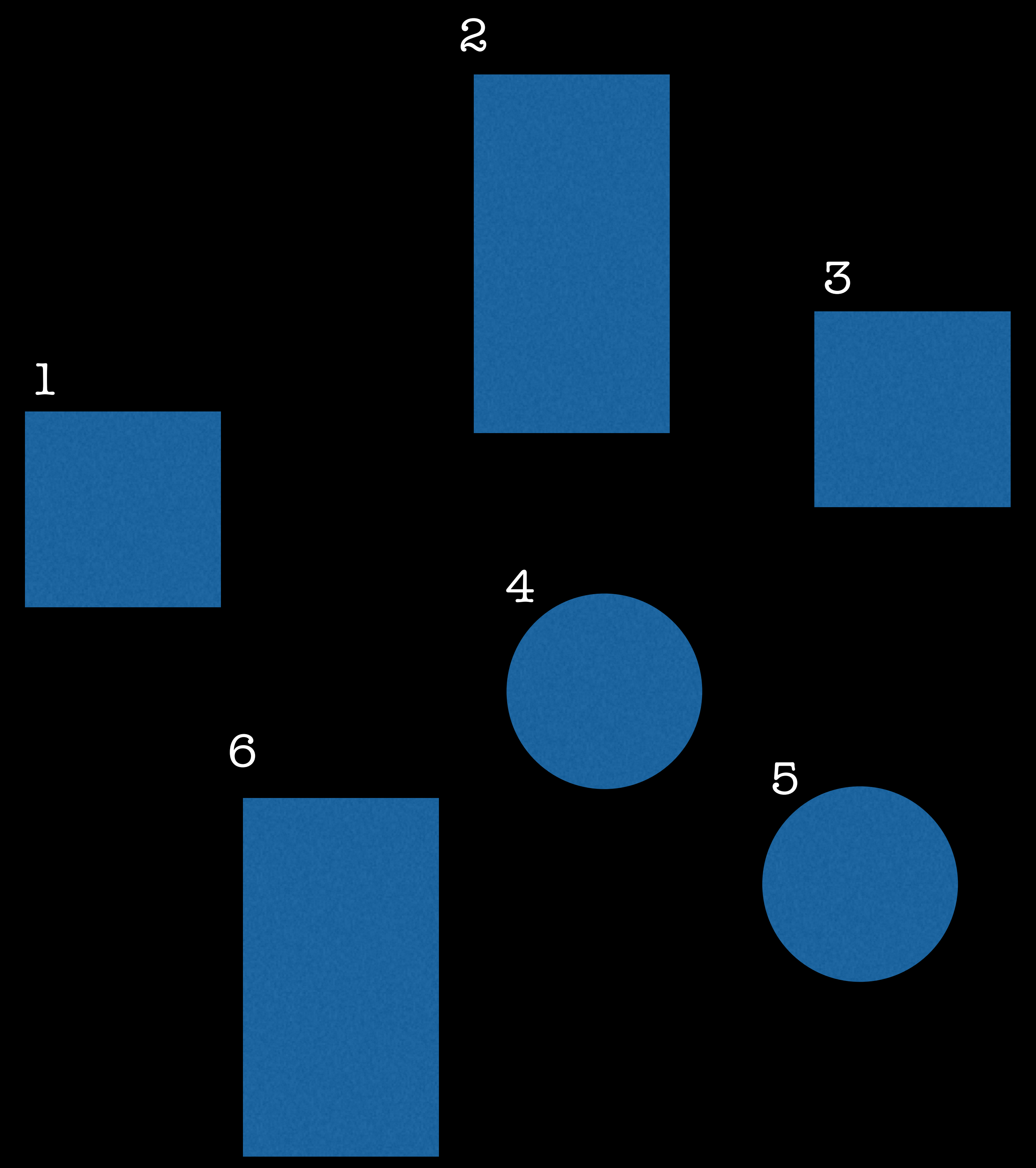
Data :

{1,1}
{2,1}
{1,2}
{2,2}
{3,2}
{2,3}
{4,4}
{4,5}
{5,5}
{5,6}
{6,5}



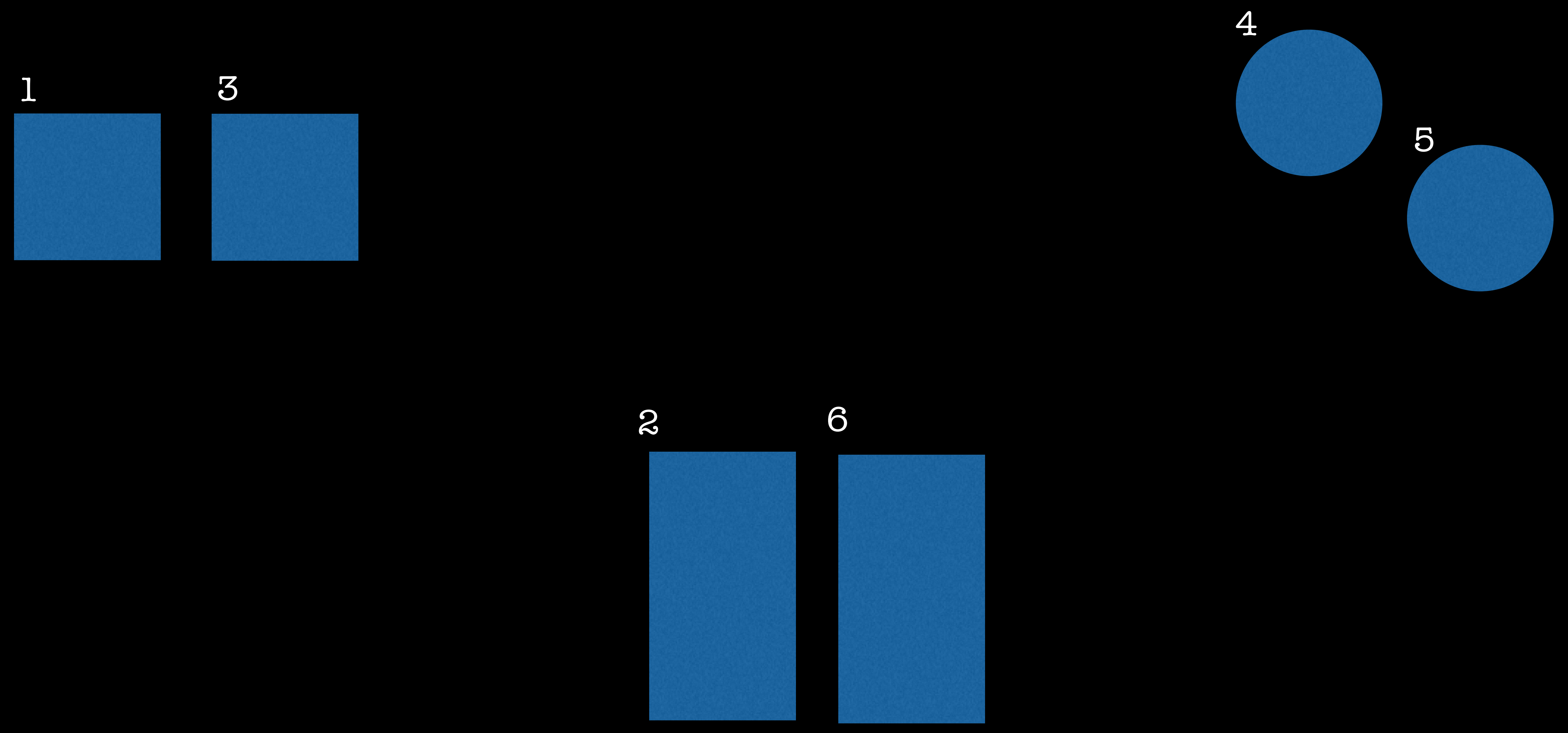
Intuition

Split into 3 categories



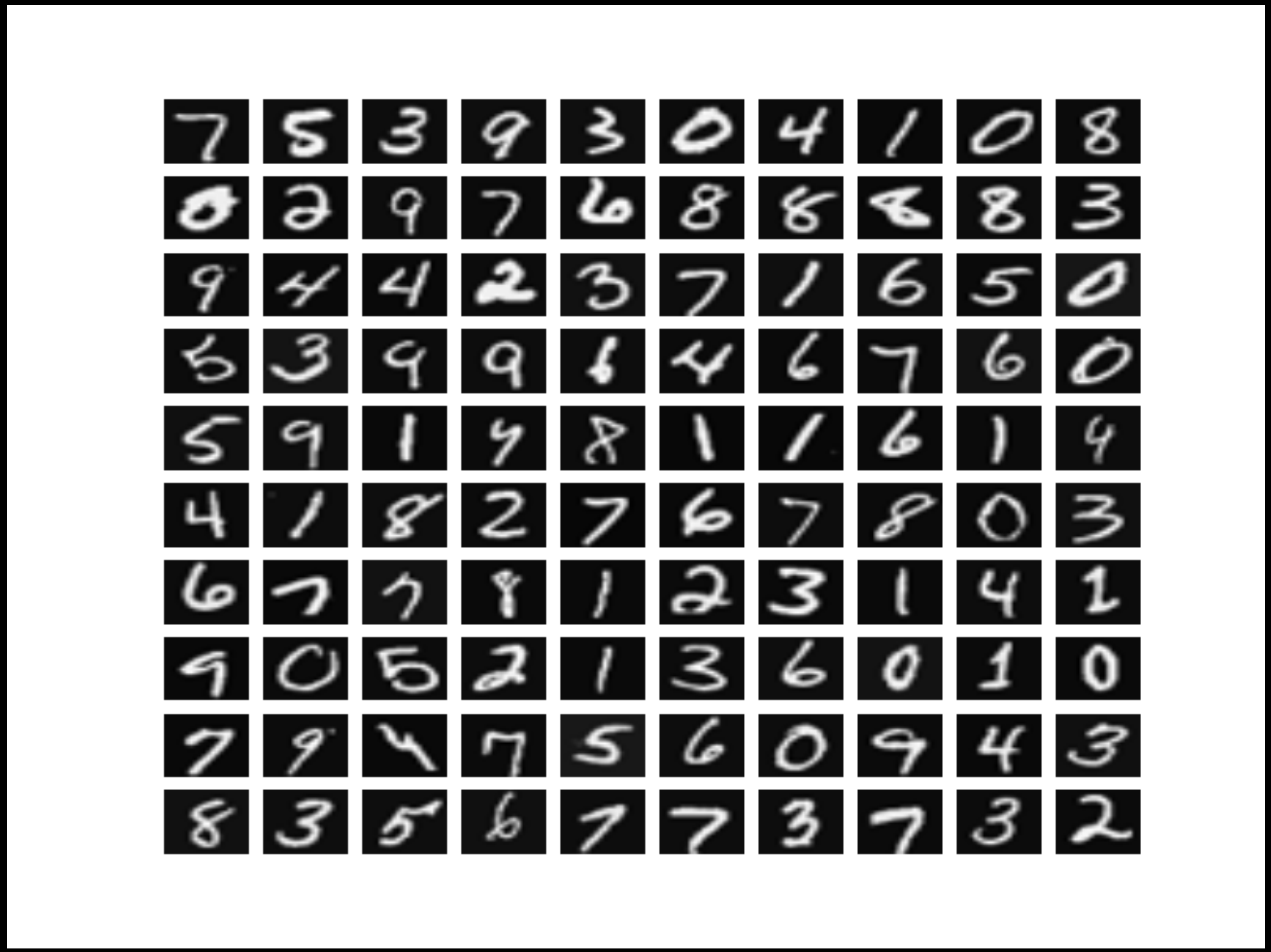
Intuition

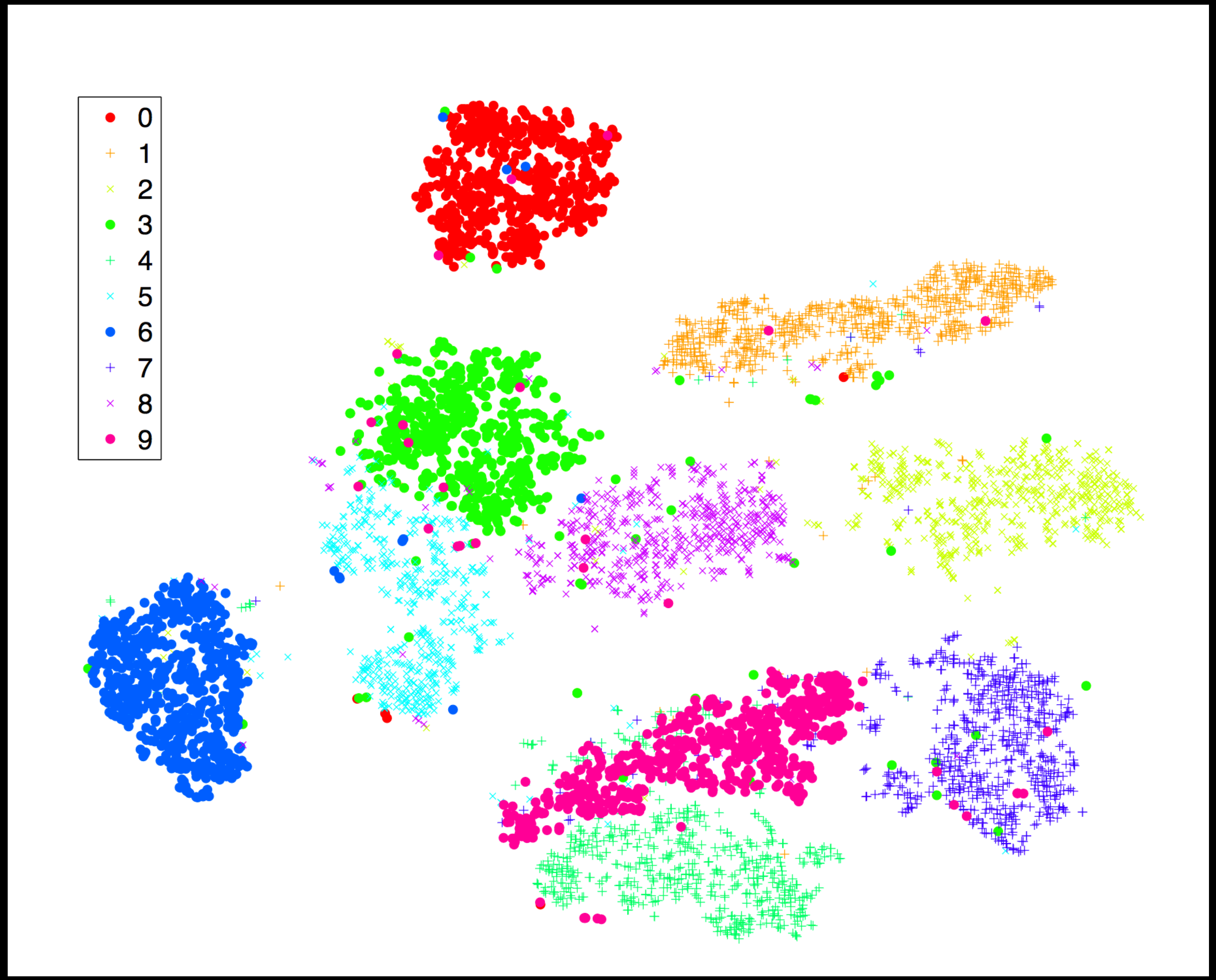
Split into 3 categories



Intuition

Split based on their shapes (Pixels)





// t-SNE, Laurens van der Maaten, 2008

Unsupervised Learning

Given $(x^1, \dots, x^n)$, find some patterns in the data

How do you find the underlying structure of a dataset? How do you summarise it and group it most usefully? How do you effectively represent data in a compressed format?

- Clustering
- Density Estimation
- Dimensionality Reduction

Unsupervised Learning

Given $(x^1, \dots, x^n)$, find some patterns in the data

How do you find the underlying structure of a dataset? How do you summarise it and group it most usefully? How do you effectively represent data in a compressed format?

- **Clustering**
- Density Estimation
- **Dimensionality Reduction**

K-Means Clustering

- Create groups of data points such that **points in different clusters are dissimilar** while **points within a cluster** are similar.

- Here are the steps to k-means clustering:

  1. **Define the k centroids.** Initialise these at random (there are also fancier algorithms for initialising the centroids that end up converging more effectively).
  2. **Find the closest centroid & update cluster assignments.** Assign each data point to one of the k clusters. Each data point is assigned to the nearest centroid's cluster. Here, the measure of "nearness" is a hyperparameter — often Euclidean distance.
  3. **Move the centroids to the centre of their clusters.** The new position of each centroid is calculated as the average position of all the points in its cluster.
  4. Keep repeating steps 2 and 3 until the centroid stop moving a lot at each iteration (i.e., until the algorithm converges).

// Machine Learning for Humans, Vishal Maini, 2017
// Demo Time : https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-Means Clustering

Pros :

- **Simple** : Easy to implement
- **Efficient** : Time Complexity : O(tkn) {n : no. of data points, k : no. of clusters, t : no. of iterations}
- Most popular clustering algorithm

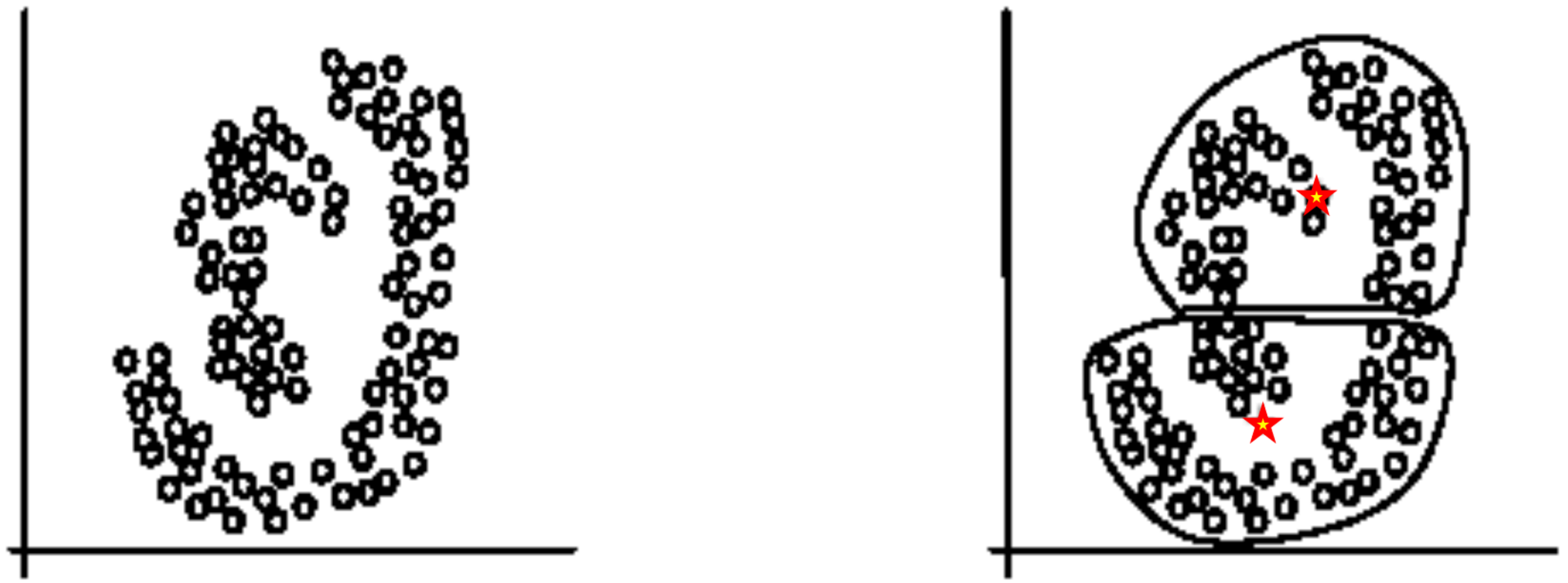
Cons :

- User needs to specify **k**
- Sensitive to outliers (far away points)
- Sensitive to initial seeds
- Only applicable if a mean is defined

// Machine Learning for Humans, Vishal Maini, 2017
// https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

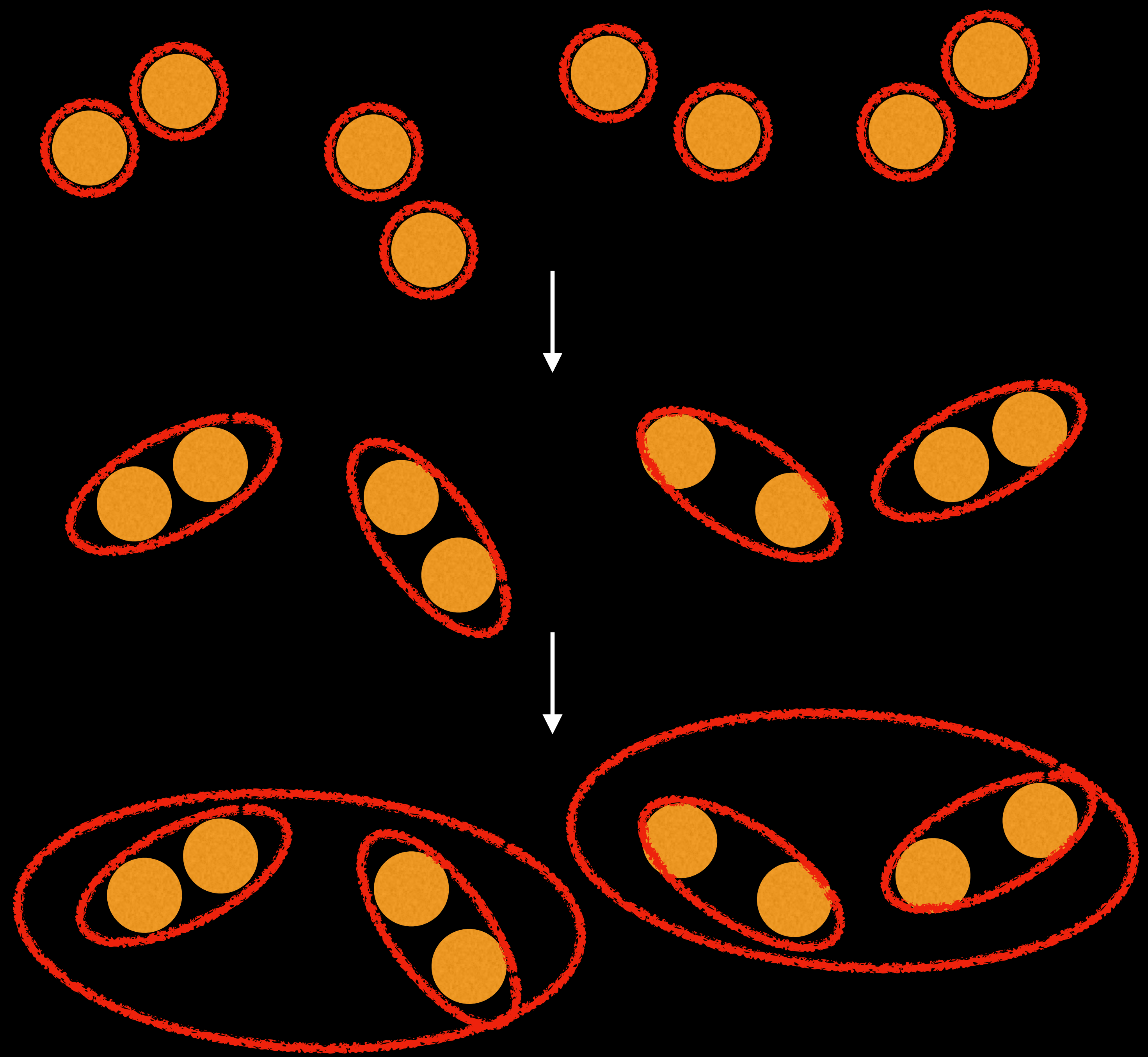
K-Means Clustering

- Doesn't work for clusters that are not hyper-ellipsoids (or hyper-spheres)



(A): Two natural clusters (B): *k*-means clusters

Hierarchical Clustering (Agglomerative)

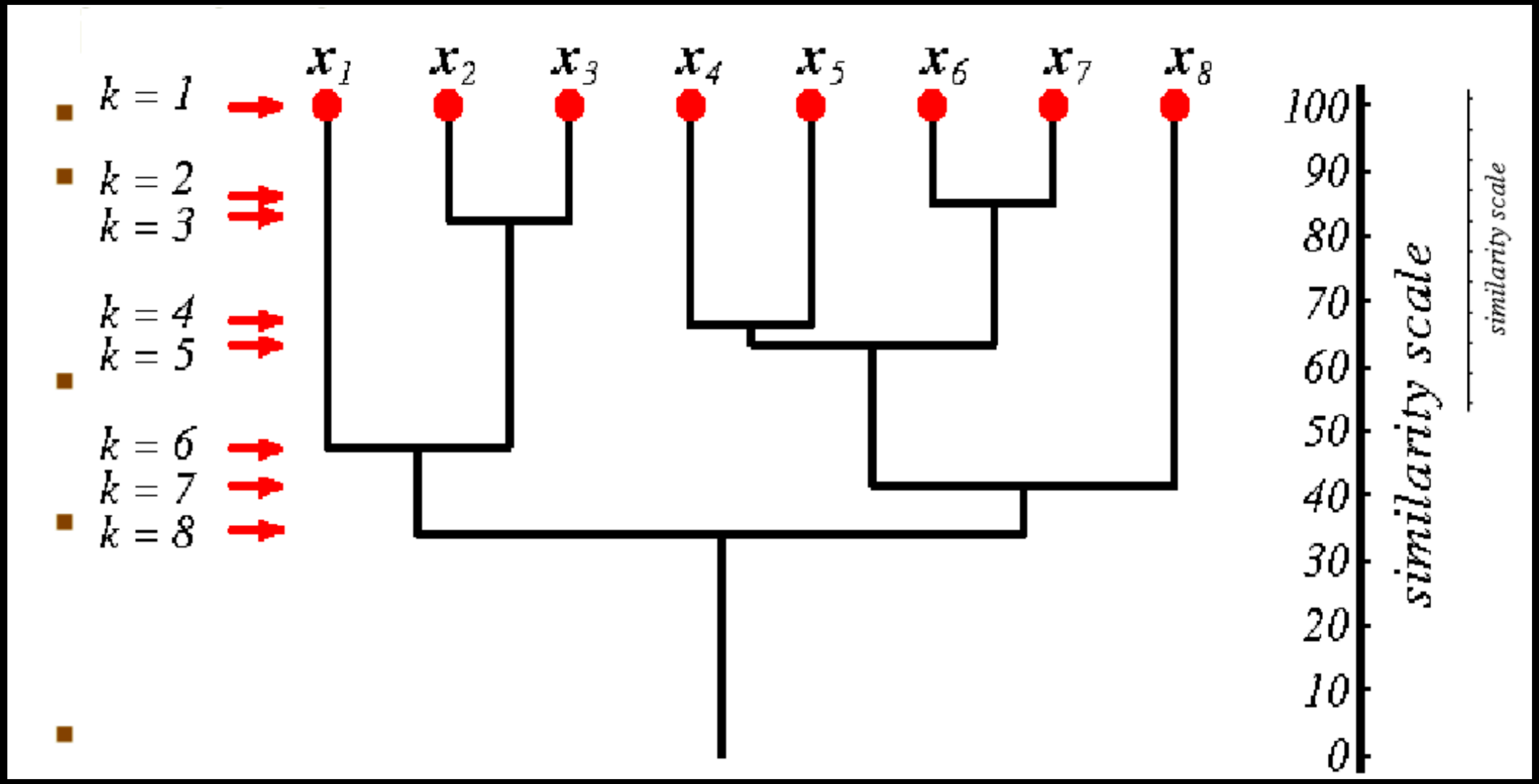


- **Start with N clusters,** one for each data point.
- **Merge the two clusters that are closest to each other**. Now you have N-1 clusters.
- **Recompute the distances between the clusters.** There are several ways to do this. One of them is to consider the distance between two clusters to be the average distance between all their respective members.
- **Repeat steps 2 and 3 until you get one cluster of N data points.** You get a dendrogram.
- **Pick a number of clusters and draw a horizontal line in the dendrogram.** In general, the number of clusters you get is the number of intersection points of your horizontal line with the vertical lines in the dendrogram.

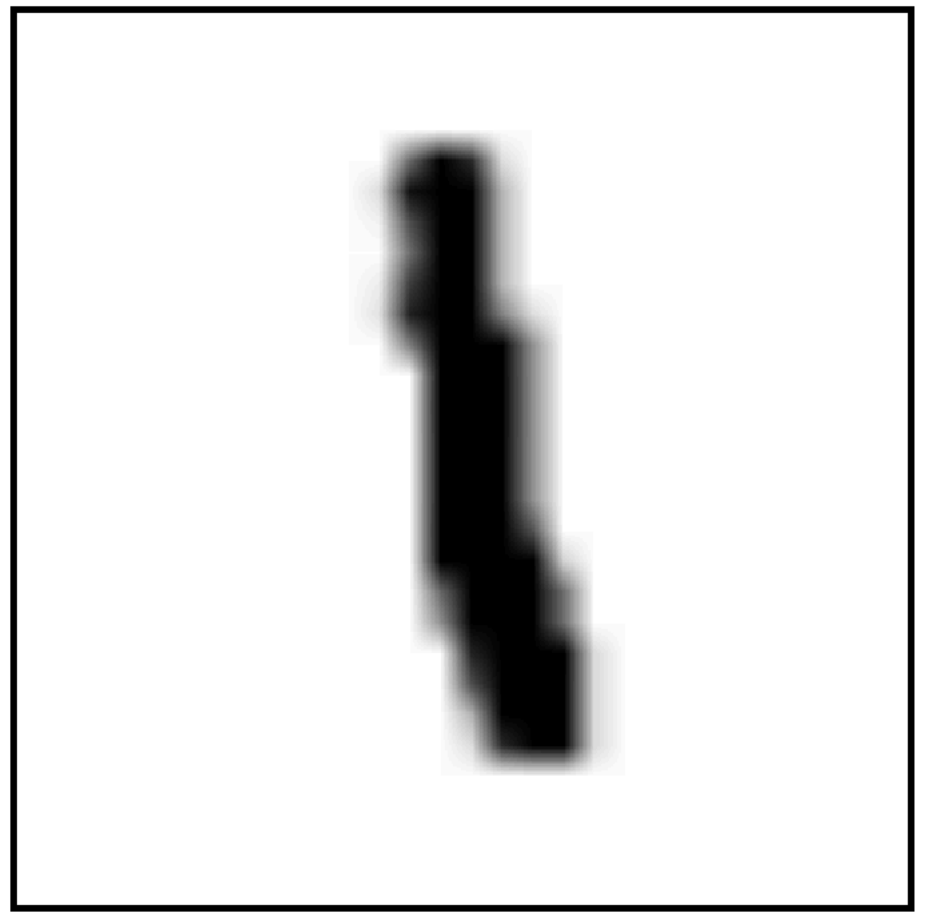
// Machine Learning for Humans, Vishal Maini, 2017
// MIT – Unsupervised Learning

Hierarchical Clustering (Agglomerative)

- A dendrogram structure built from bottom up.
- Merging the most similar pair of clusters (Distance)
- Stopping when all the clusters are merged to 1 cluster.



Dimensionality Reduction



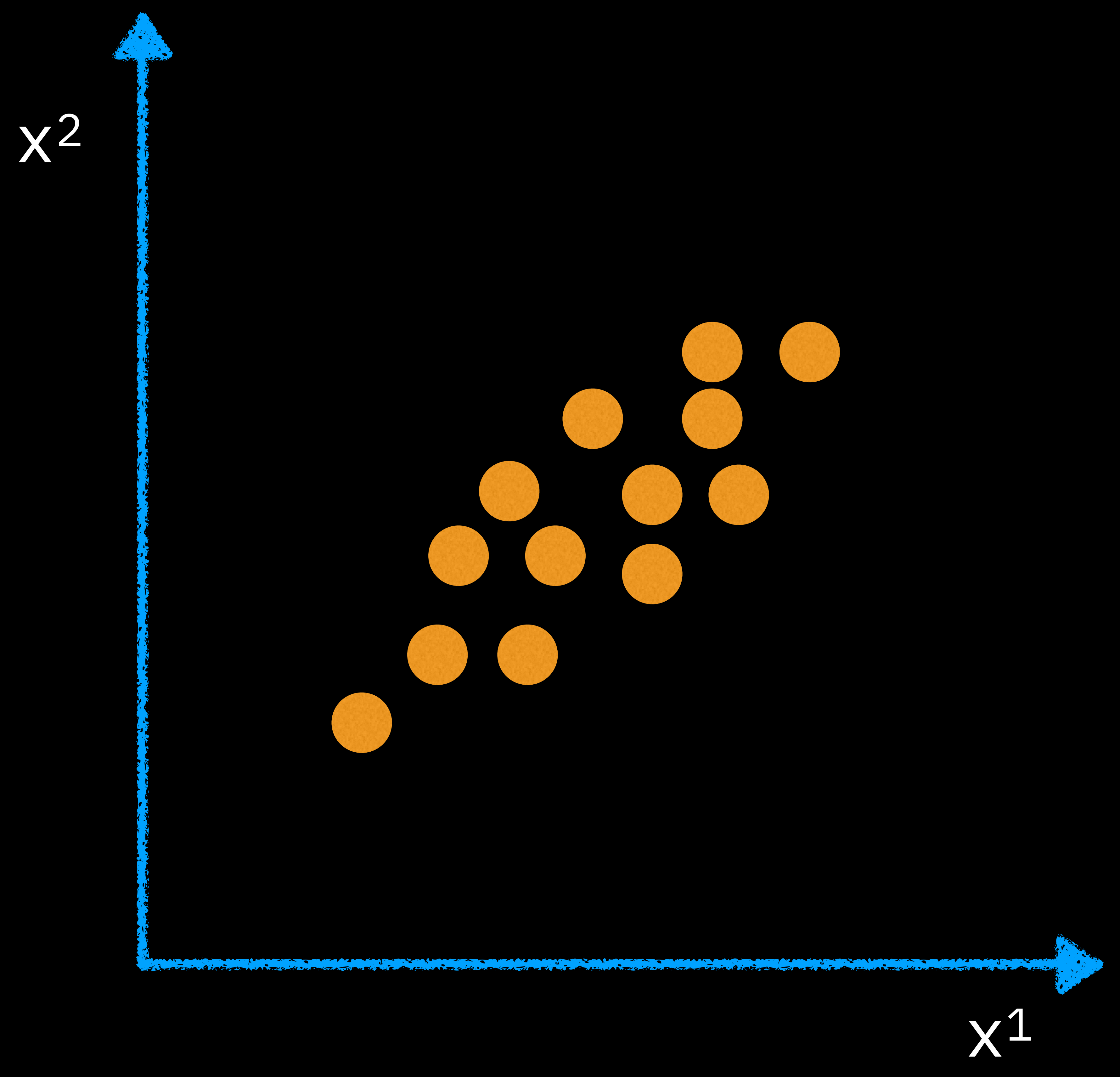
$$\cong \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\ 0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\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

{ 0,0,0, …. , 0.6, 0.8, 0, … 0 } //28*28 = 784 : **{1}**

- Time and Space Complexity
- Not all features are relevant to problem. (Leads to overfitting)

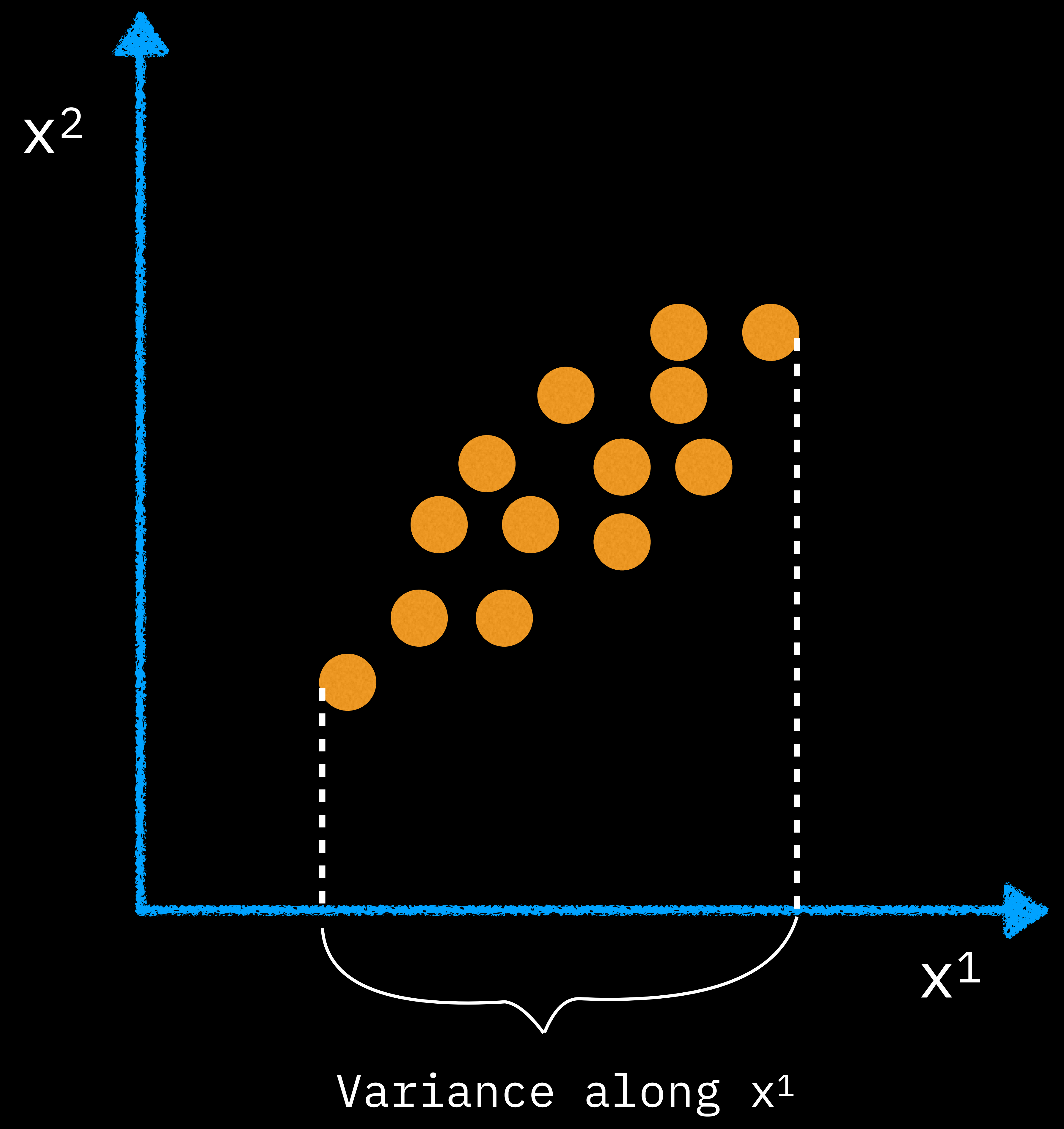
// Chris Olah, Introduction to Dimensionality Reduction

Principle Component Analysis



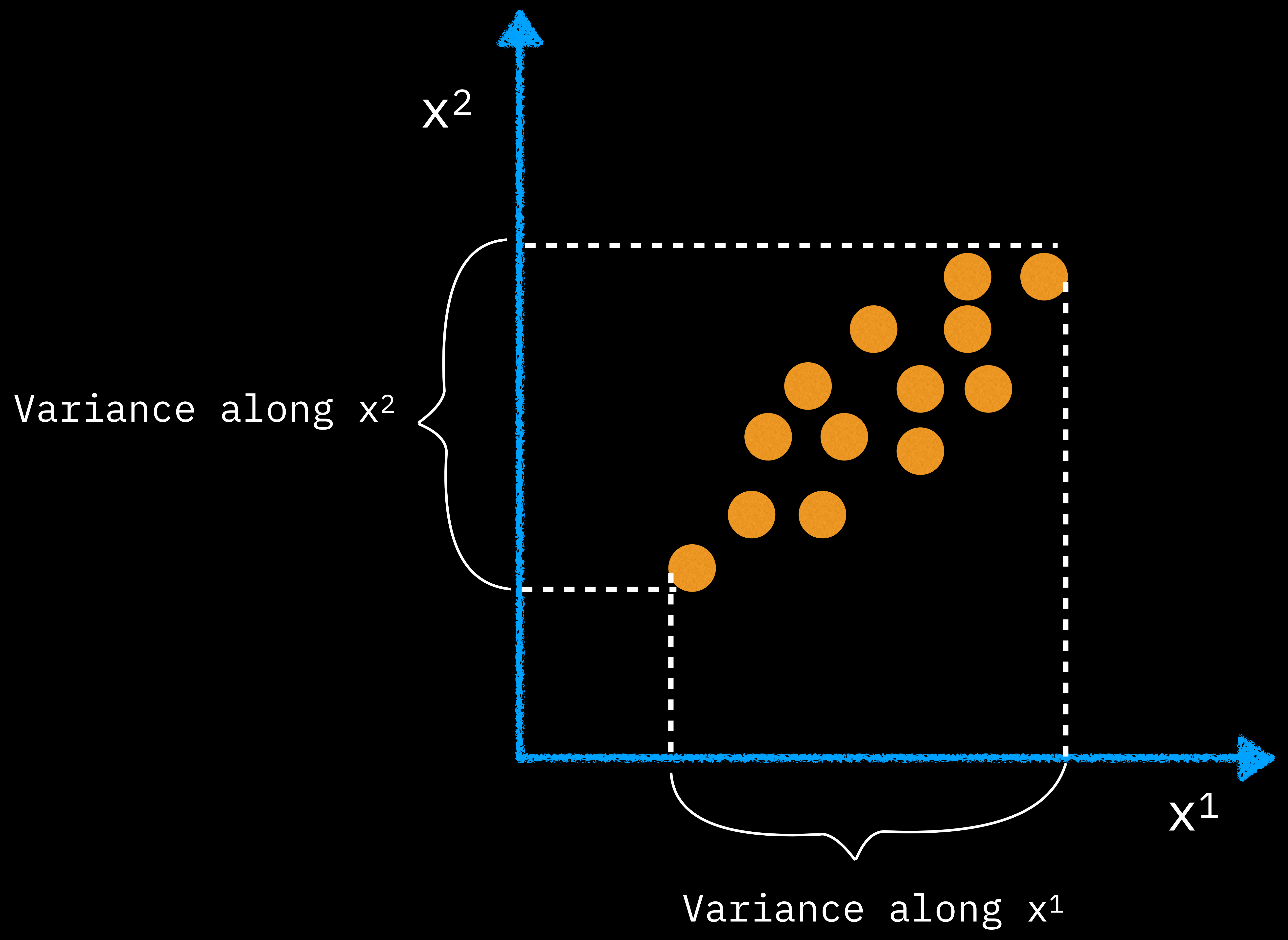
- Components (Features) : {x1, x2)

Principle Component Analysis



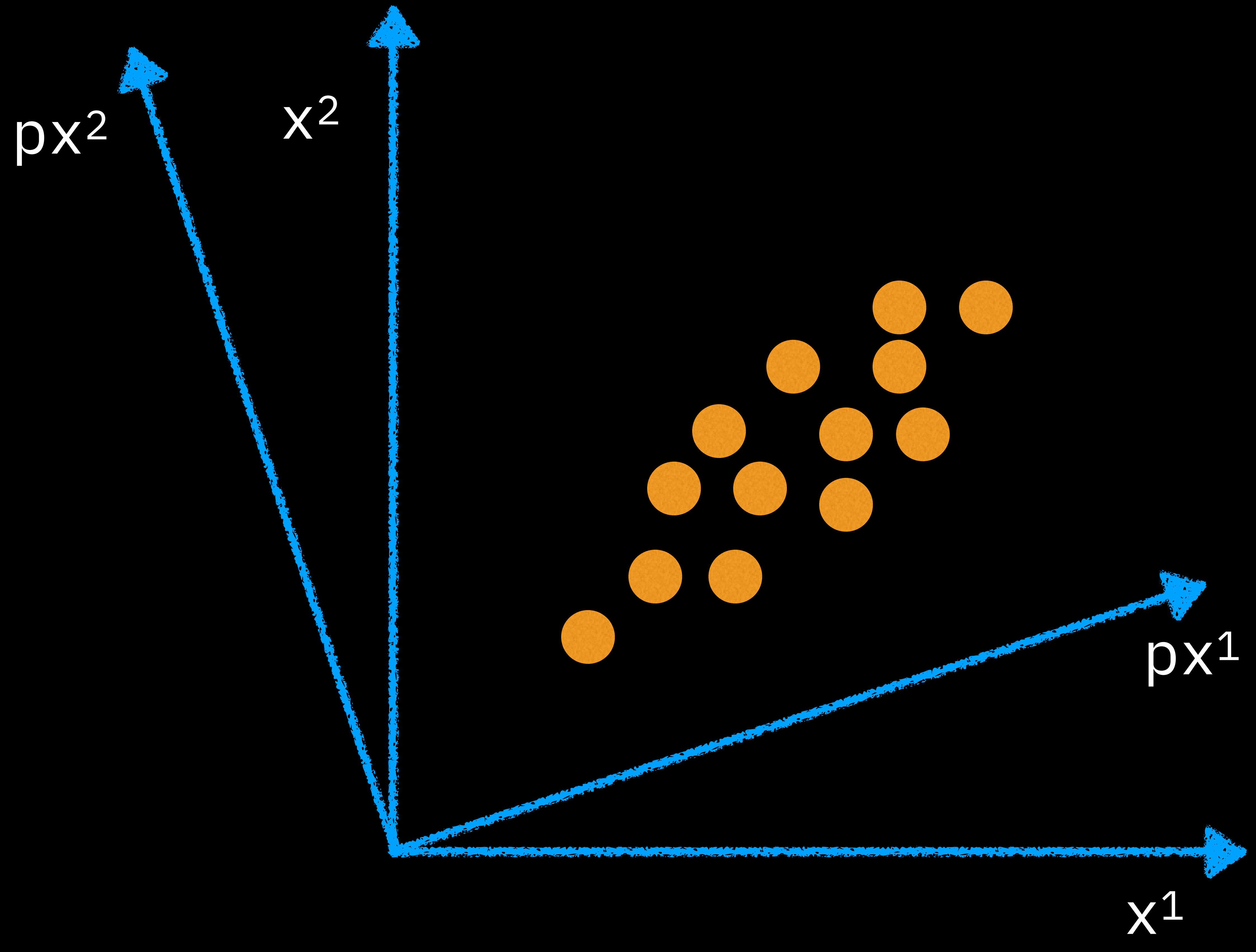
- Components (Features) : {x1, x2)

Principle Component Analysis



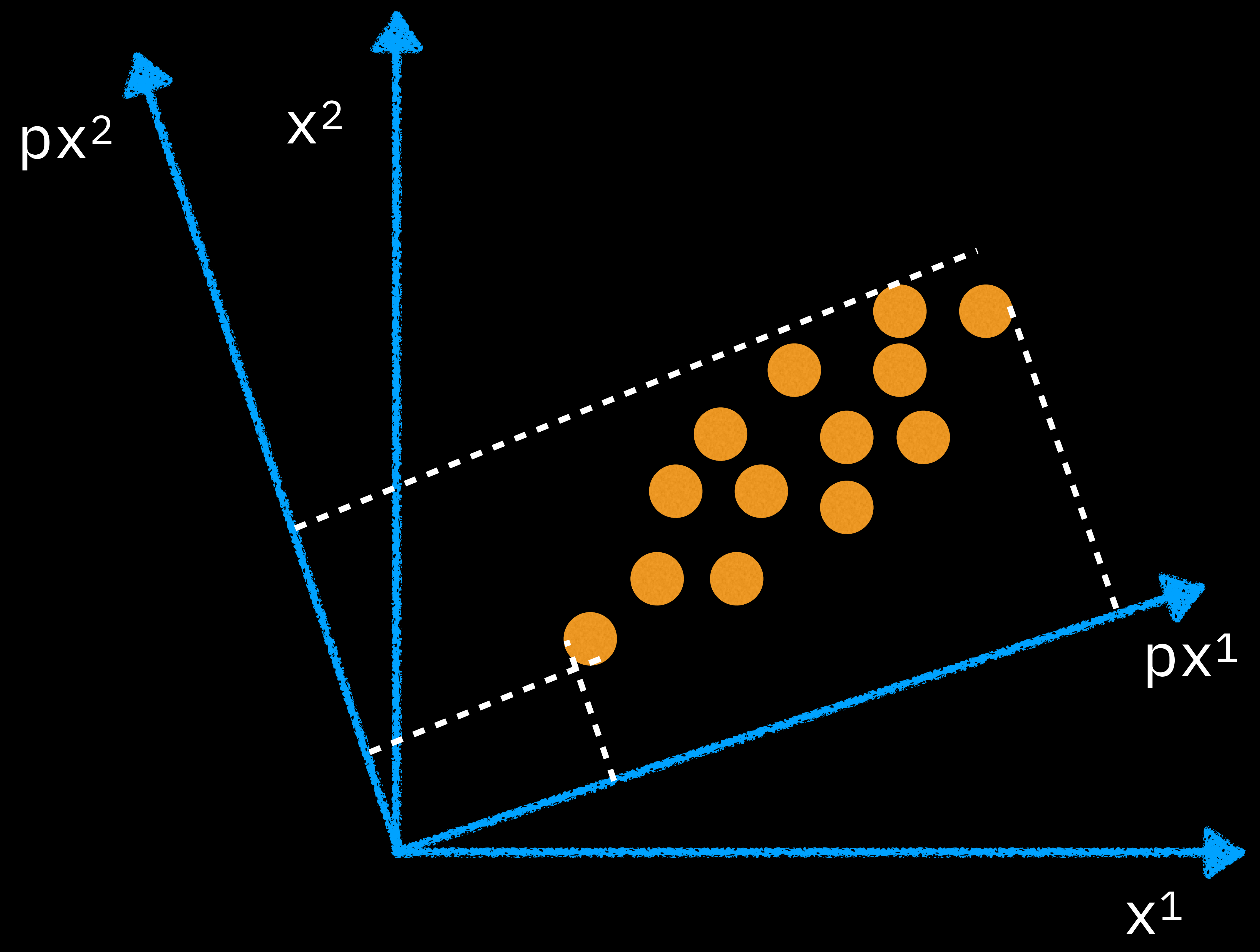
- Components (Features) : $\{x^1, x^2\}$

Principle Component Analysis



- Components (Features) : $\{x^1, x^2\}$

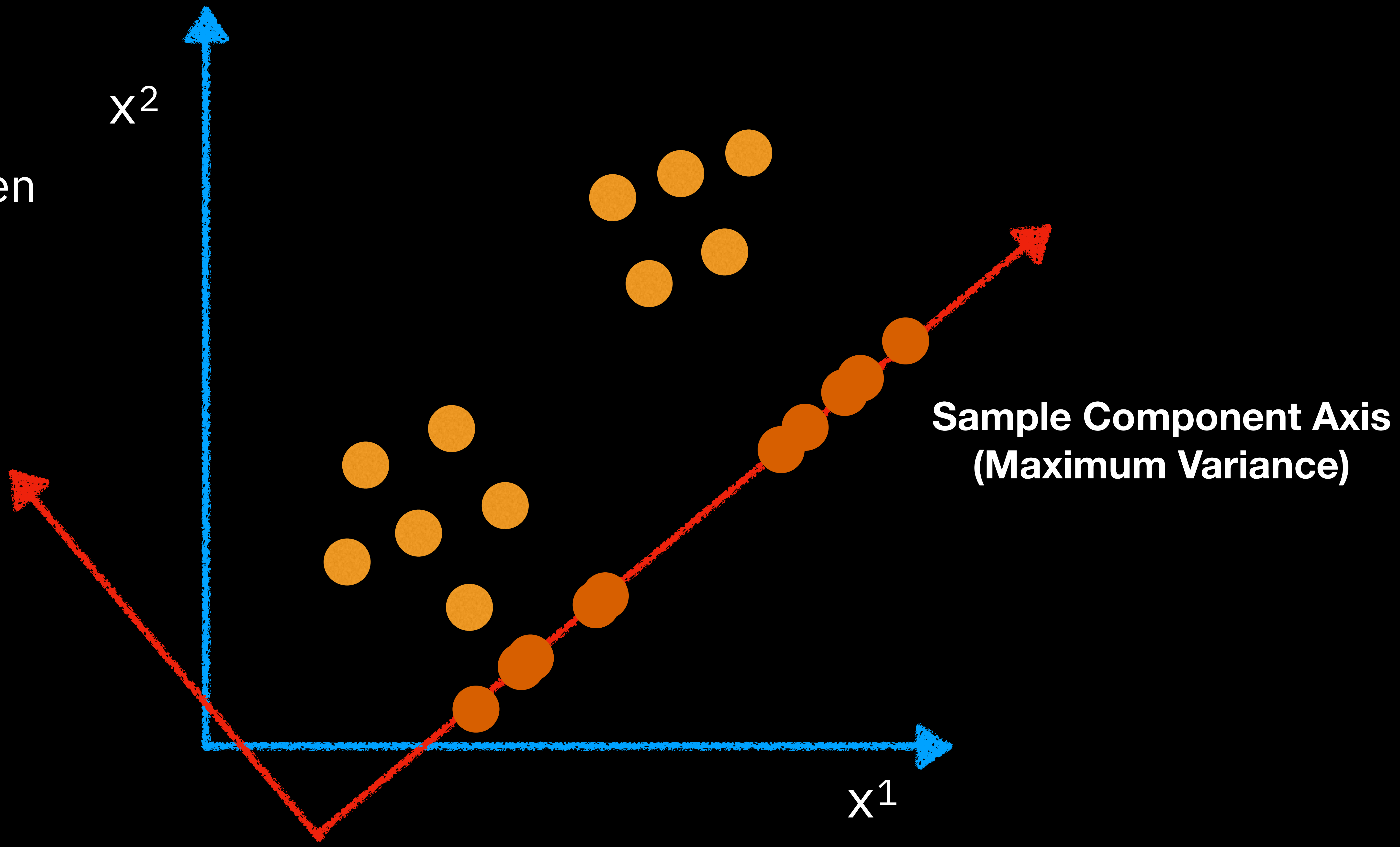
Principle Component Analysis



- Components (Features) : $\{x^1, x^2\}$
- Better component : $\{px^1, px^2\}$

Principle Component Analysis

- Maximises Variance.
- Directions which are orthogonal
- Best Reconstruction (Minimum error when moving from n to m dimensions)



Principle Component Analysis

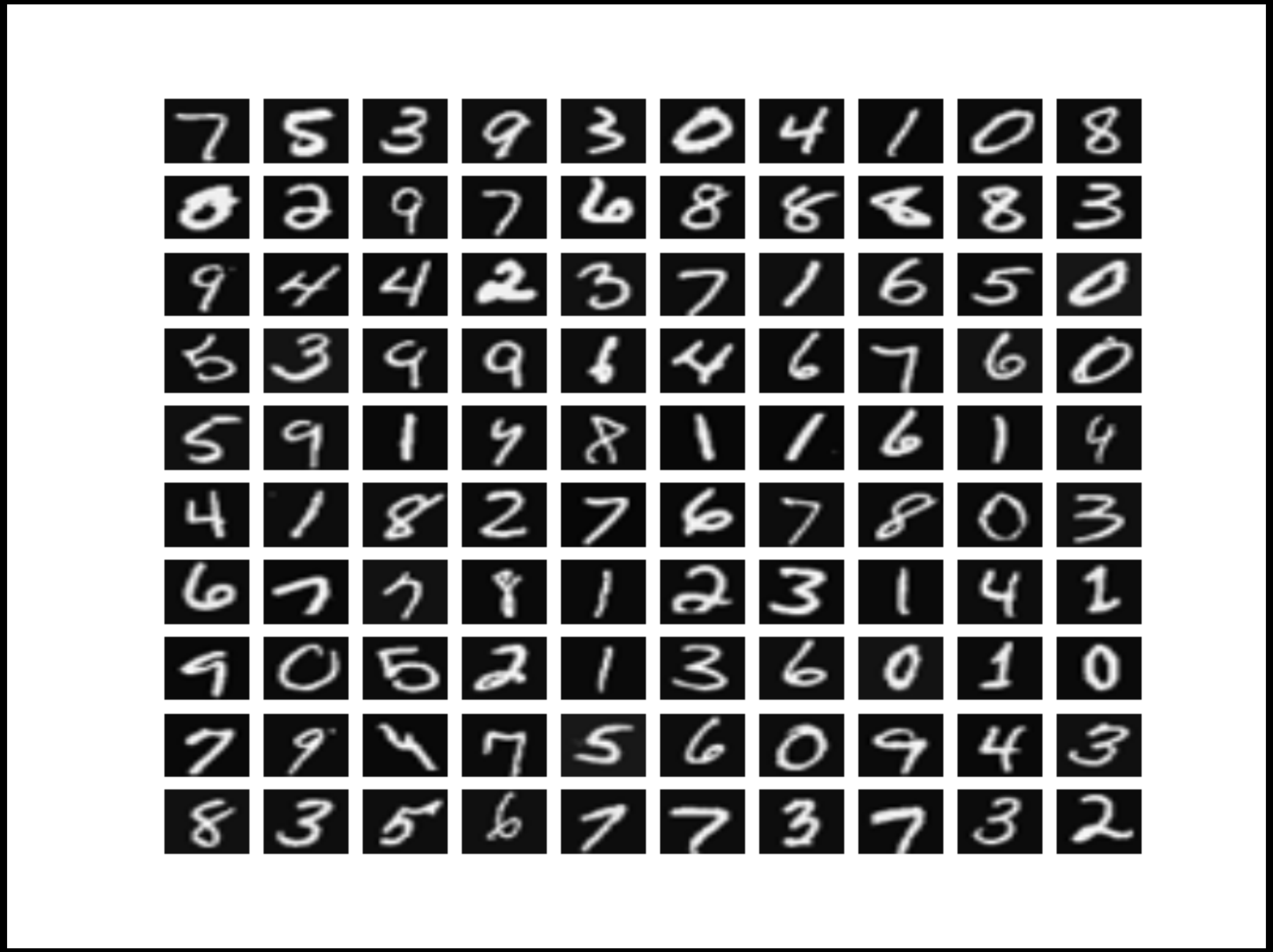
Data → Covariance Matrix → Eigenvectors, Eigenvalues → Sort and Select top k Eigenvectors → Transform to k-dimensional Eigenvector Space

(or Singular Value Decomposition)

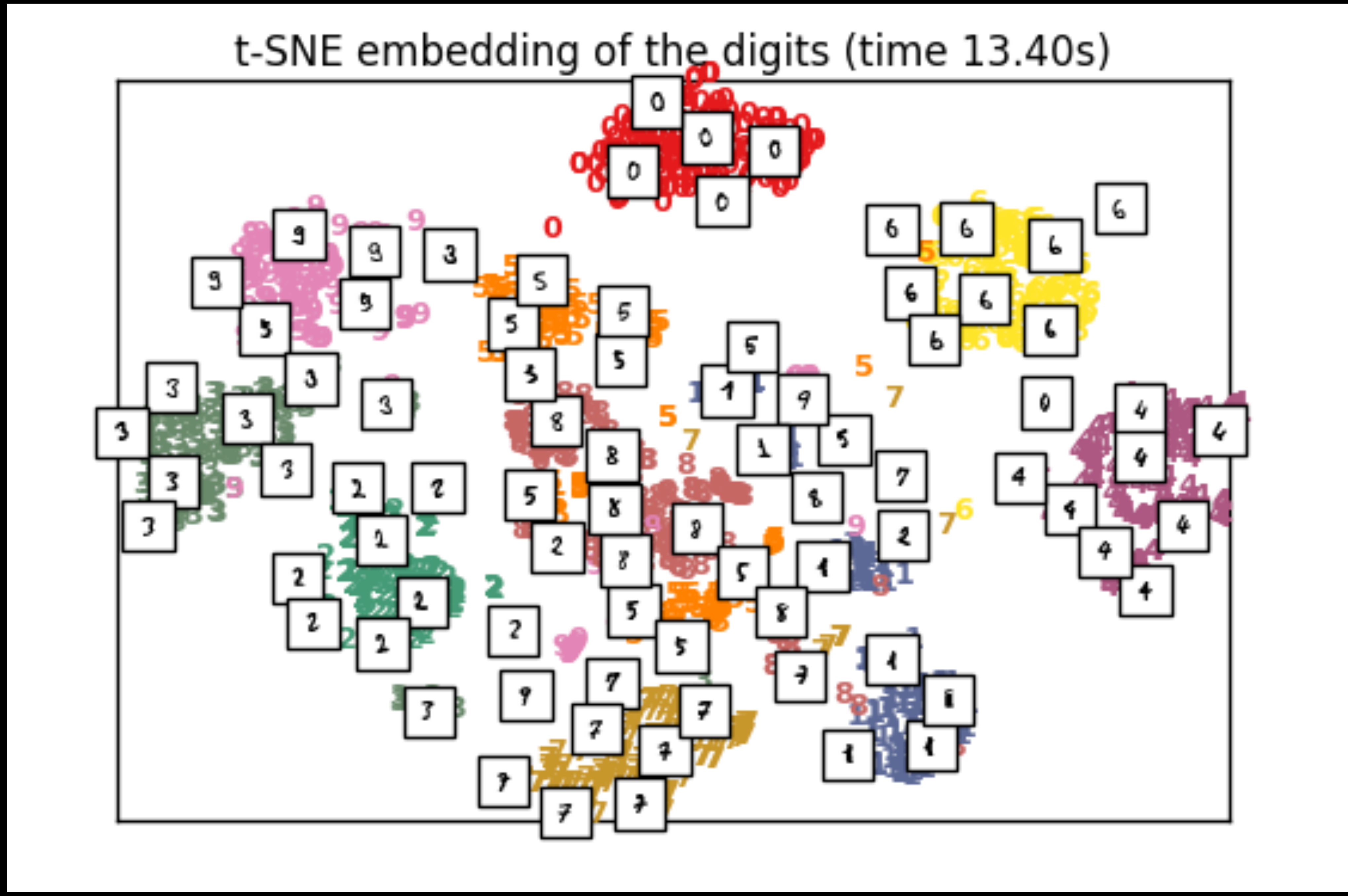
coding_time

```
// Notebook courtesy : Siraj Raval, 2017
// More on Eigenvectors, Eigenvalues checkout Youtube Channel : 3Blue1Brown
```

Structured Output



Unstructured



Structured

Any Questions?

email id : harisris@gmail.com