

# Outline

## Administrivia and Introduction

Course Structure

Syllabus

Introduction to Data Mining

## Dimensionality Reduction

Introduction

Principal Components Analysis

Singular Value Decomposition

Multidimensional Scaling

**Isomap**

## Clustering

Introduction

Hierarchical Clustering

K-means

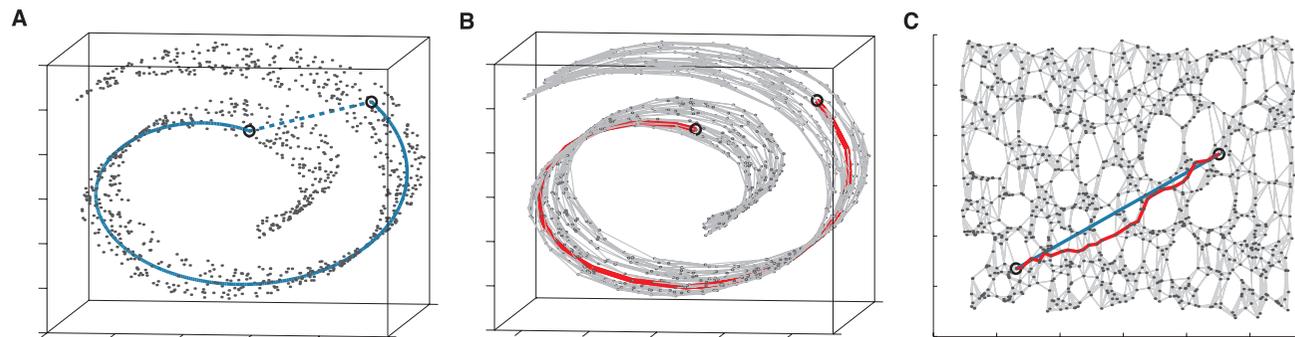
Vector Quantisation

Probabilistic Methods

# Isomap

Isomap is useful for non-linear dimension reduction

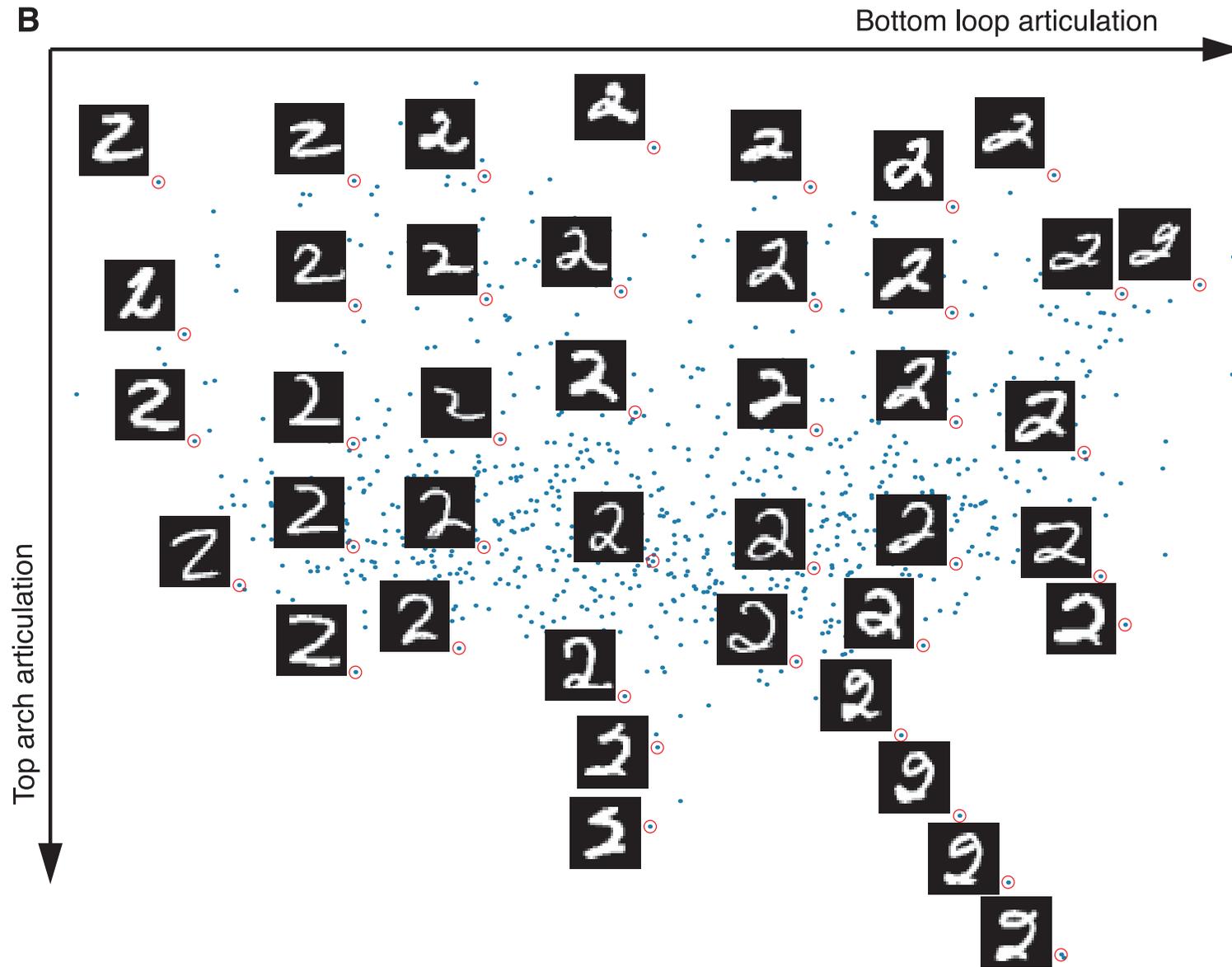
1. Calculate distances  $d_{ij}$  for  $i, j = 1, \dots, n$  between all data points, using the Euclidean distance.
2. Form a graph  $G$  with the  $n$  samples as nodes, and edges between the respective  $K$  nearest neighbors (in Euclidean metric).
3. Replace distances  $d_{ij}$  by 'shortest-path' distance  $d_{ij}^G$ <sup>2</sup> and perform classical MDS, using these distances.



Examples from Tenenbaum et al. (2000)

<sup>2</sup>The path-distance in the graph is, for a given path  $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i_m$  between two nodes  $i_1$  and  $i_m$  that follows the edges of the graph, the sum of the original distances  $\sum_{k=1}^{m-1} d_{i_k i_{k+1}}$ . The shortest path distance between two points  $i$  and  $j$  is the minimal path distance along all paths starting in  $i$  and ending in  $j$ .

# Embedding Handwritten Characters



# Embedding Faces

