Manifold learning: achievements and challenges

Balázs Kégl

University of Montreal

Workshop on
Principal manifolds for data cartography and dimension reduction

August 24, 2006

Outline

- Definition
- The classics (k-means, PCA, MDS)
- The first nonlinear methods (SOM, ANN, HS principal curves)
- Solving some of the problems (polygonal principal curves, LLE, ISOMAP)
- The remaining challenges

The definition

- Learn a compact representation $\mathbf{y}$ of data $\mathbf{x}$ that preserves important information
  - compactness usually means dimensionality reduction: $|\mathbf{y}| \ll |\mathbf{x}|$
  - important information: whatever is needed for performing a given task
  - usually a trade-off

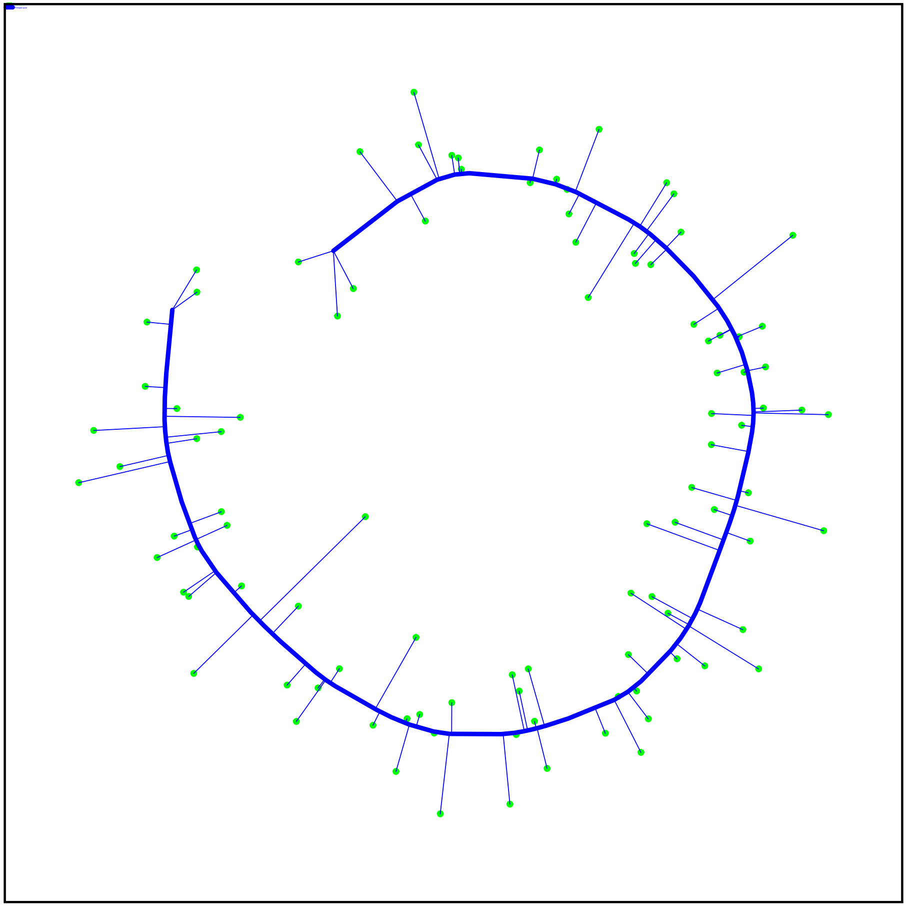
The definition

- Positive side effects
  - filtering noise
  - find the underlying hidden causes that “explain” the data
  - visualization
  - sometimes the goal, sometimes “just happens”, but not necessarily the same problems

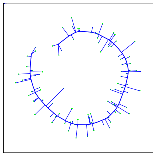
The coding view

- $\mathbf{x} \underset{encoder}{\longrightarrow} \mathbf{y} \underset{decoder}{\longrightarrow} \widehat{\mathbf{x}}$
- more general than manifold learning
- manifold is not explicit but can be traced or interpolated
- we often want “good” representation, not only efficient coding

The geometric view



The geometric view



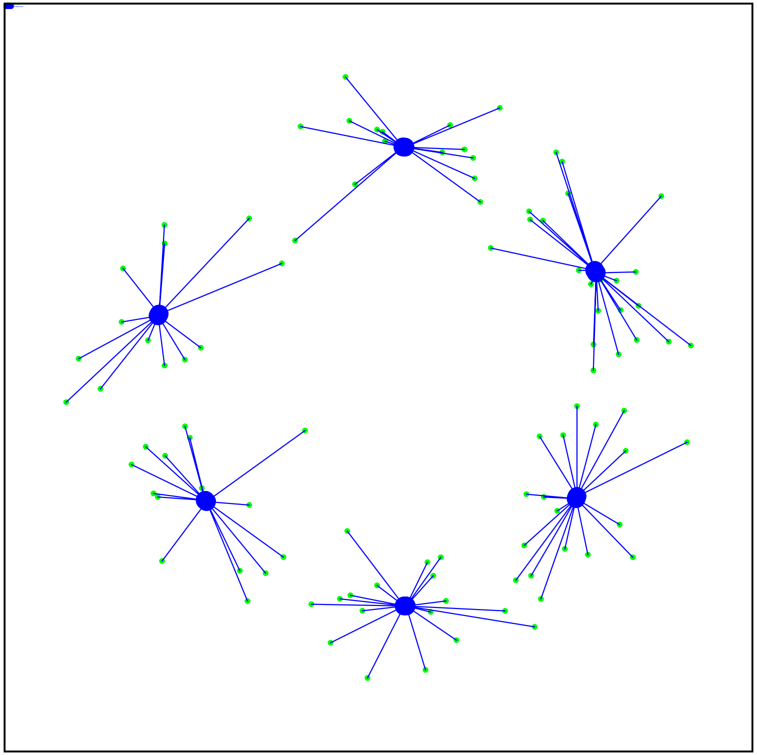
- encoding is based on a projection to a subspace
- decoding is a simple “reading out” of the coordinates
- the goal is to find or form the subspace
- information preservation can be measured by
  - the expected distance of a point and its projection to the subspace
  - topology preservation

The classics

- K-means
- Principal Component Analysis
- Multidimensional Scaling
- a lot of direct applications, but also sources of inspiration

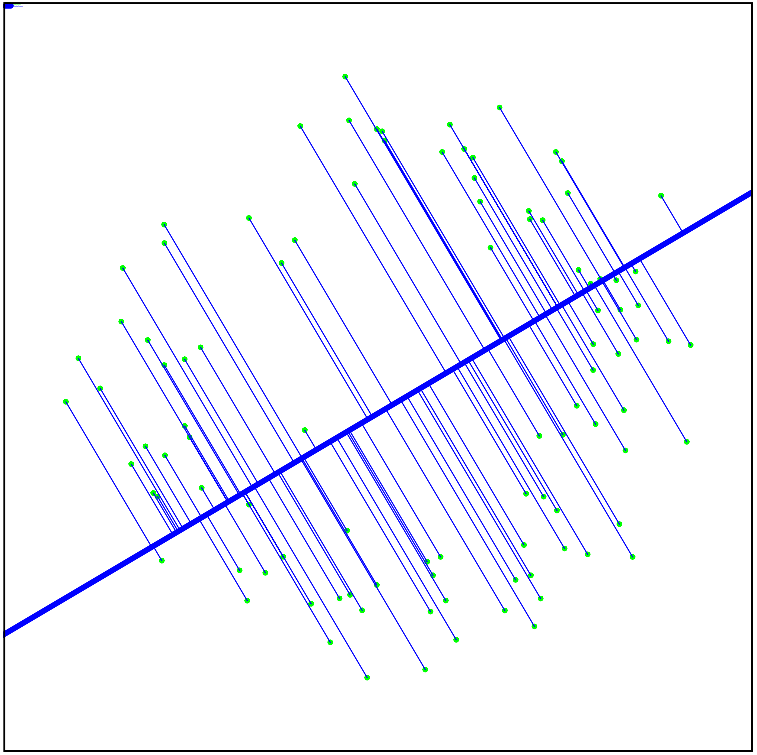
The classics

- K-means
  - singular manifold: find the nearest k points



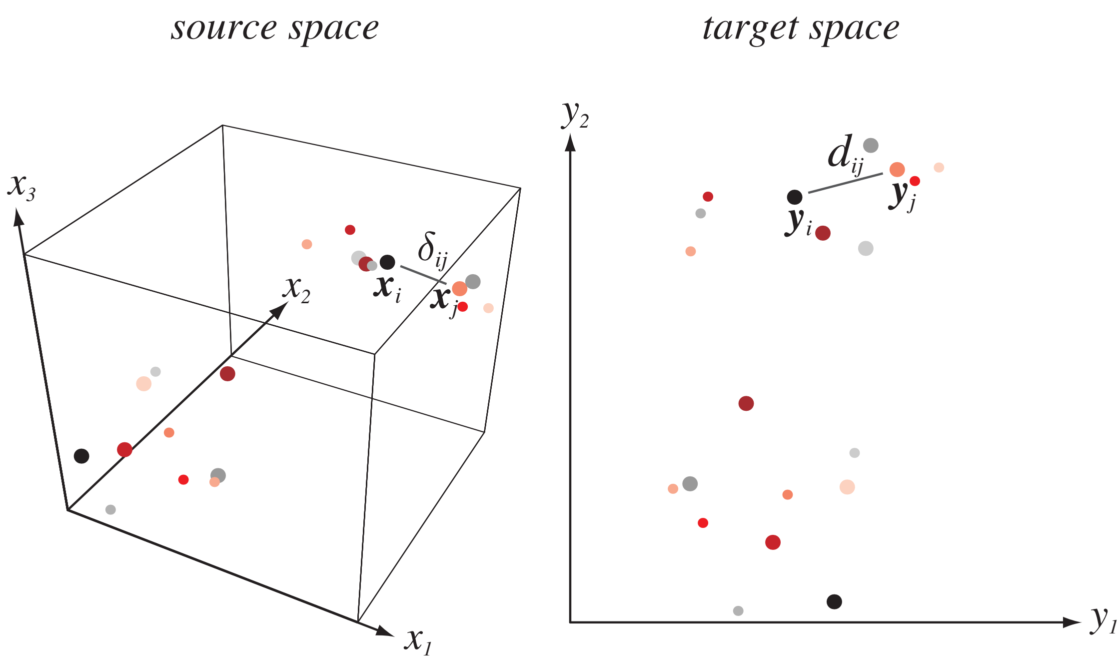
The classics

- Principal Component Analysis
  - linear manifold: find the nearest linear subspace

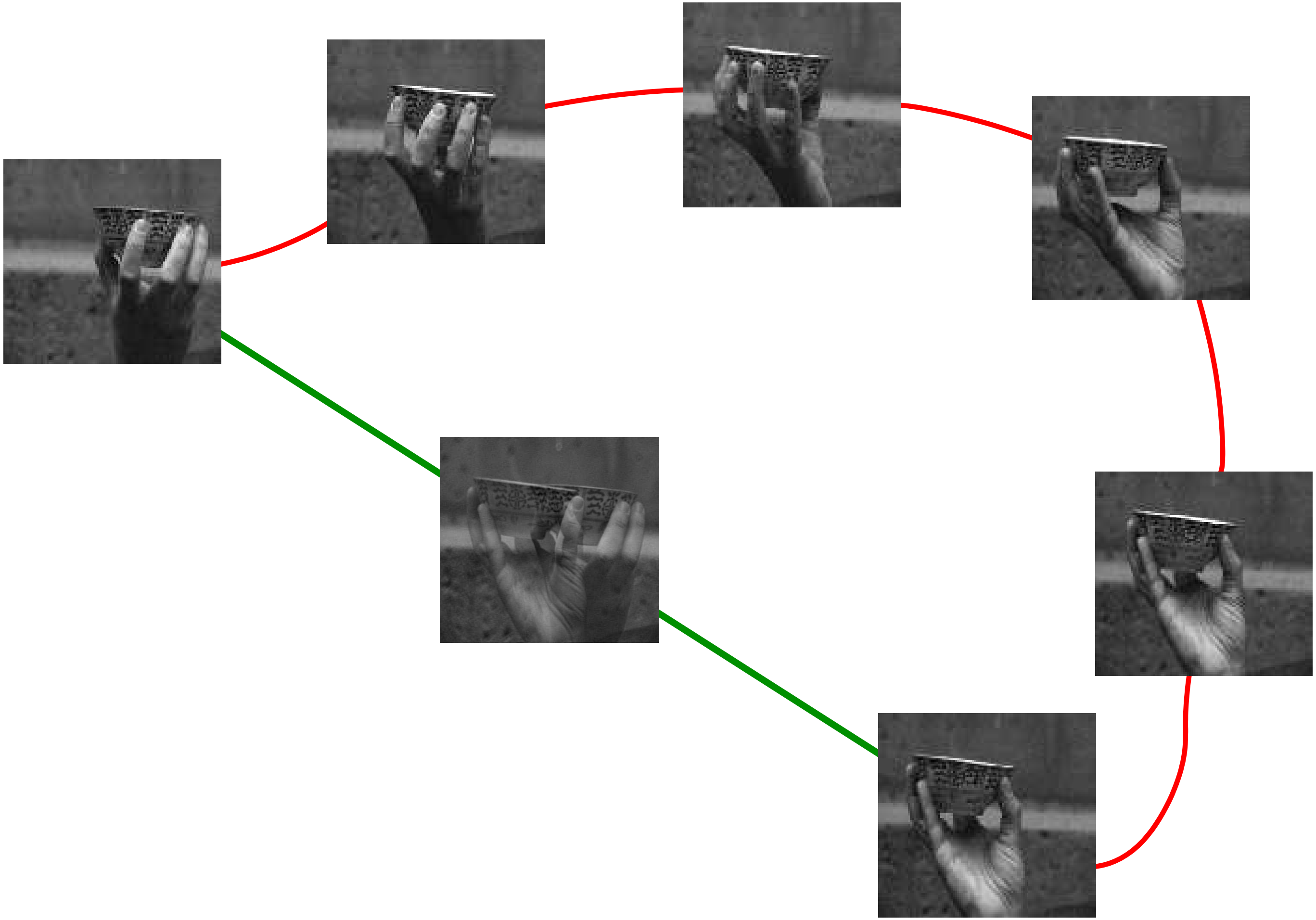


The classics

- Multidimensional Scaling
  - distance preserving manifold: find the linear subspace that preserves pairwise distances the best



Why go nonlinear?



The first nonlinear methods

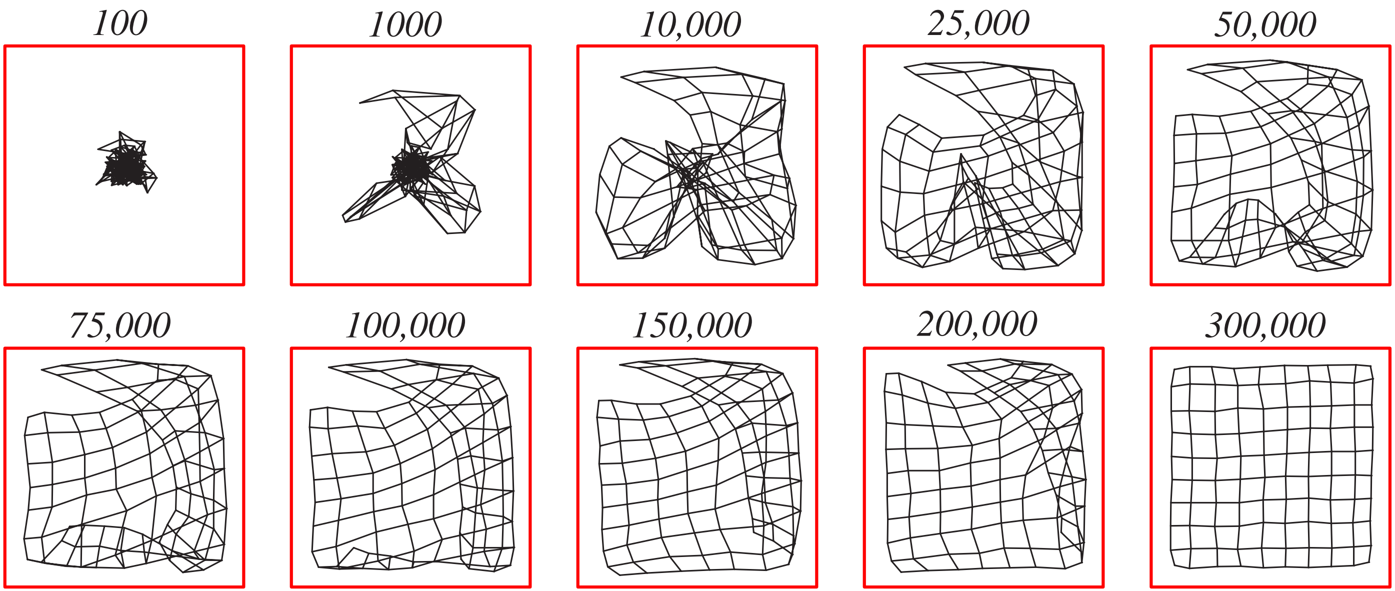
- Self-Organizing Maps and Generative Topographic Mapping

- Autoassociative Neural Networks

- Hastie-Stuetzle (HS) principal curves

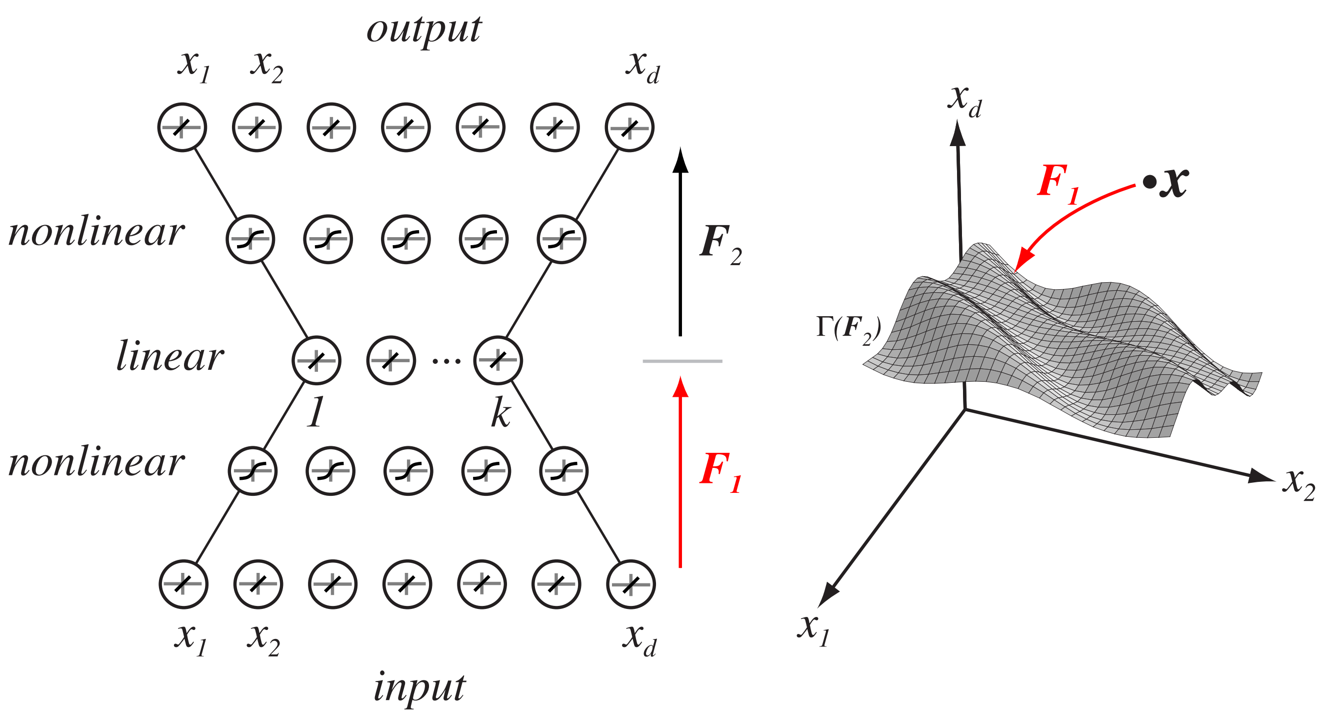
The first nonlinear methods

- Self-Organizing Maps



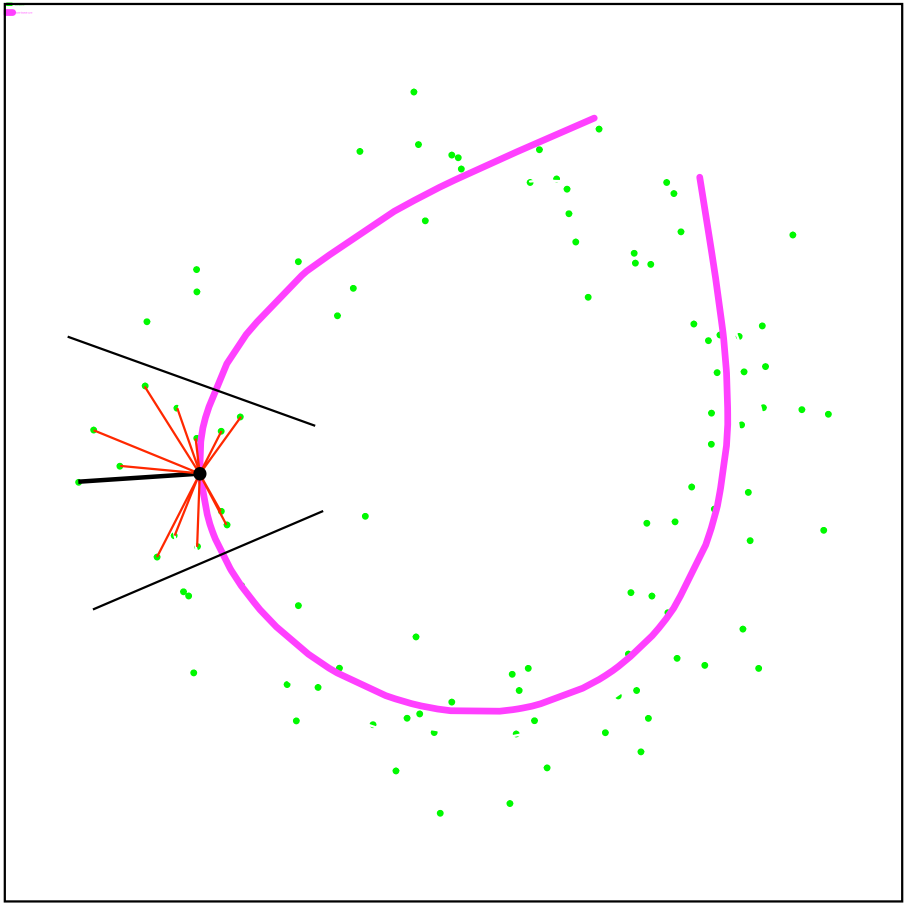
The first nonlinear methods

- Autoassociative Neural Networks



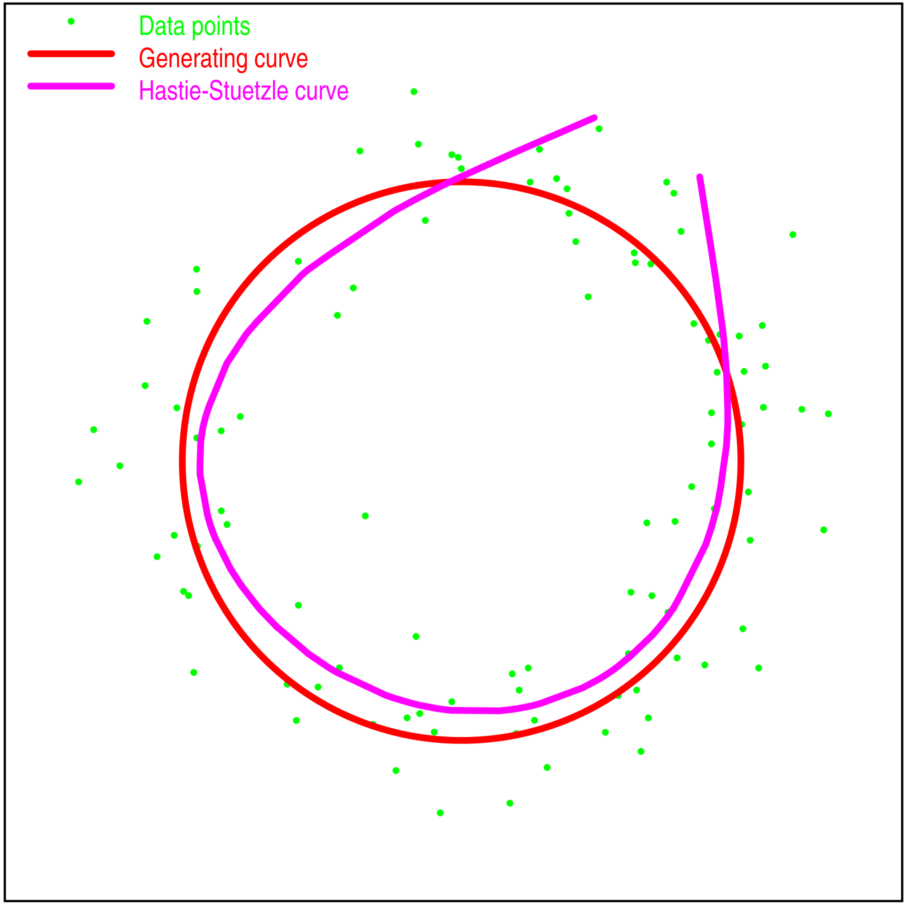
The first nonlinear methods

- HS principal curves



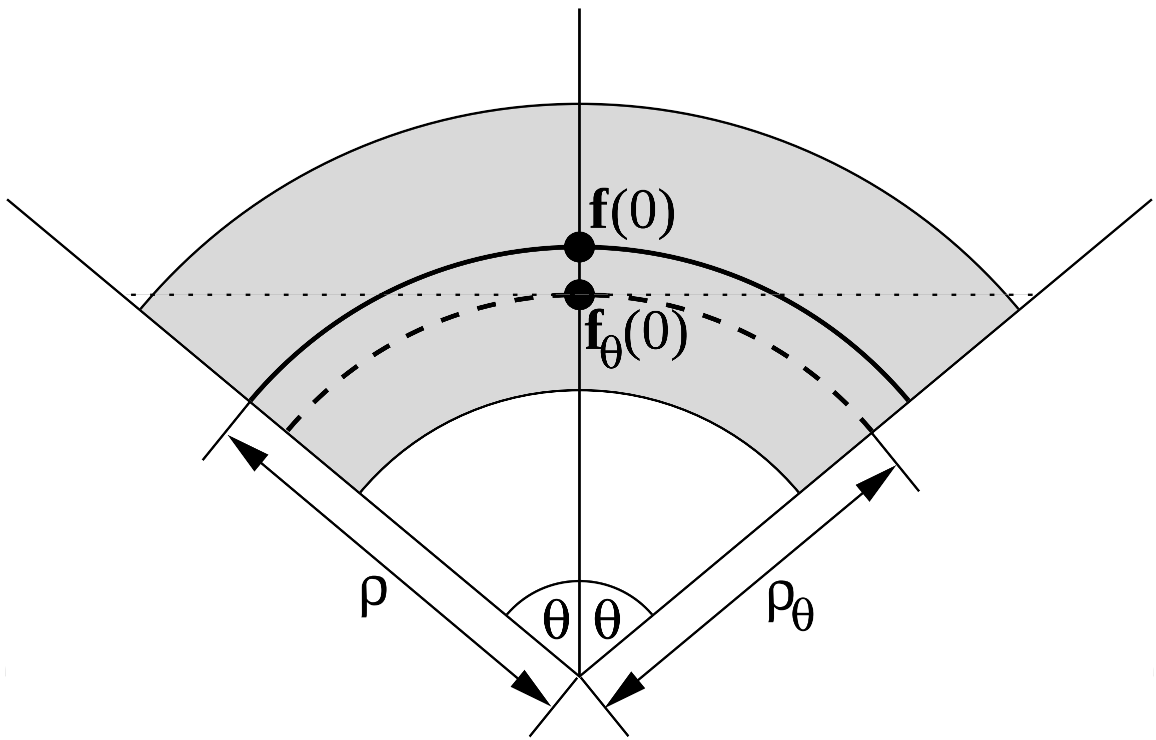
Challenge #1: The estimation bias

- "Cutting turns"



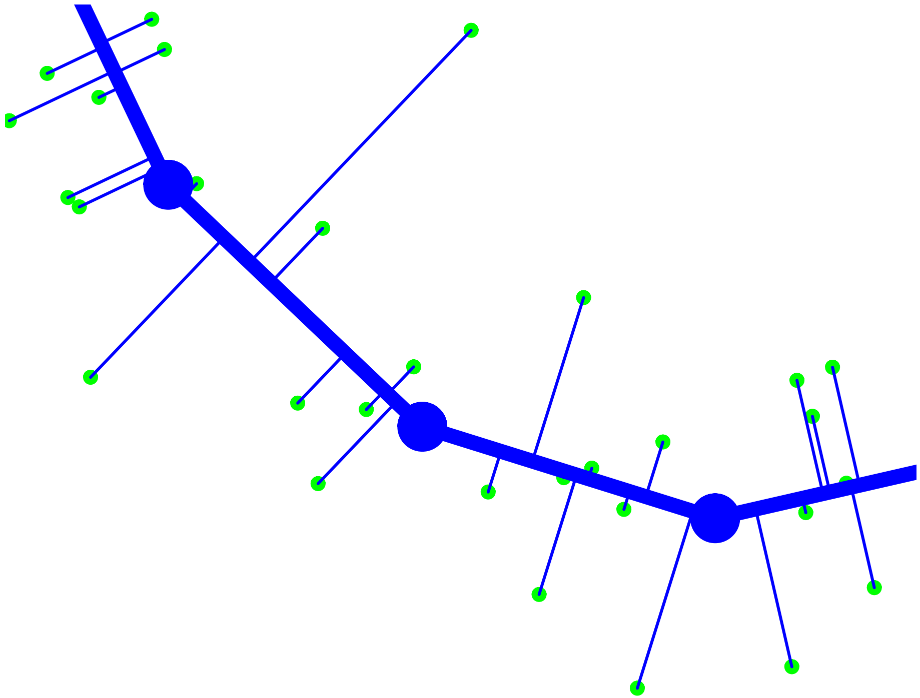
Challenge #1: The estimation bias

- “Cutting turns”



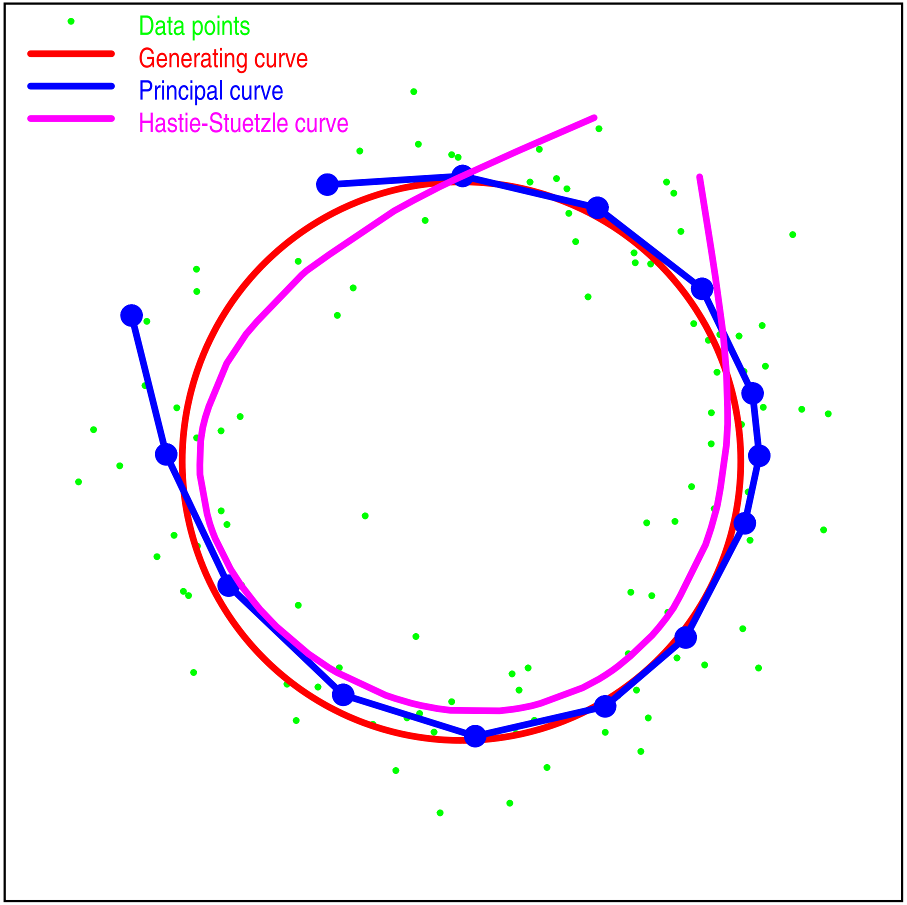
Challenge #1: The estimation bias

- Solution: project on line segments

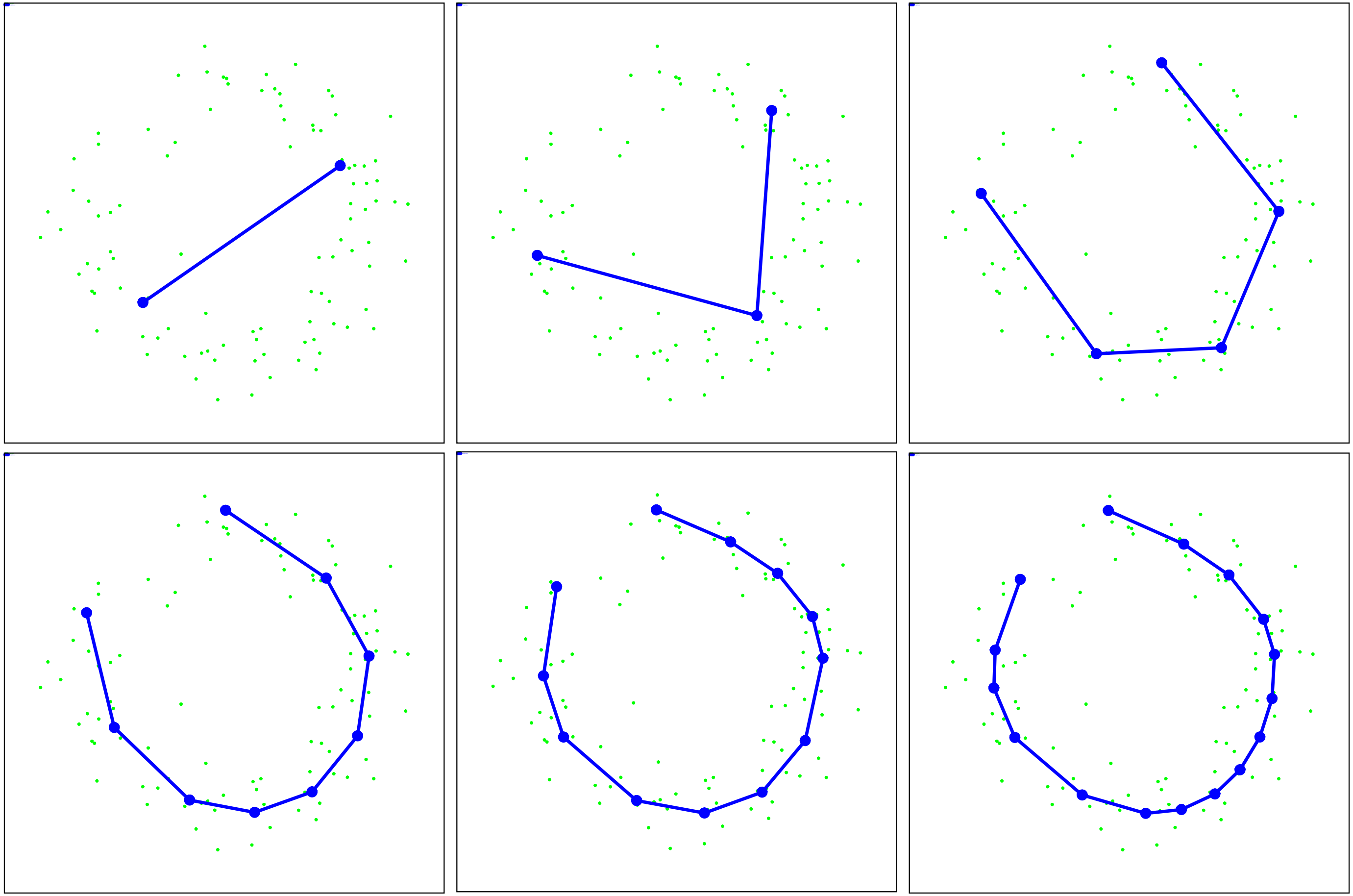


Challenge #1: The estimation bias

- Solution: project on line segments



The polygonal line algorithm



Principal Curves

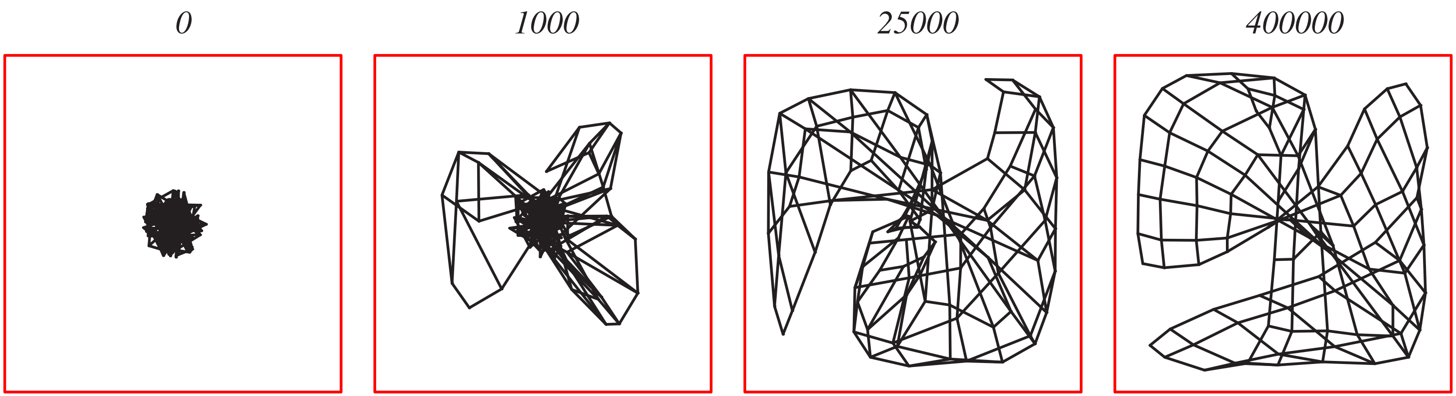
- HS (theory is flawed, algorithm is slow and non-robust, high estimation bias)

- Polygonal line algorithm and length constraint

- Regularized principal manifolds, principal curves with bounded turn, k-segments algorithm, elastic principal graphs and manifolds, local principal curves

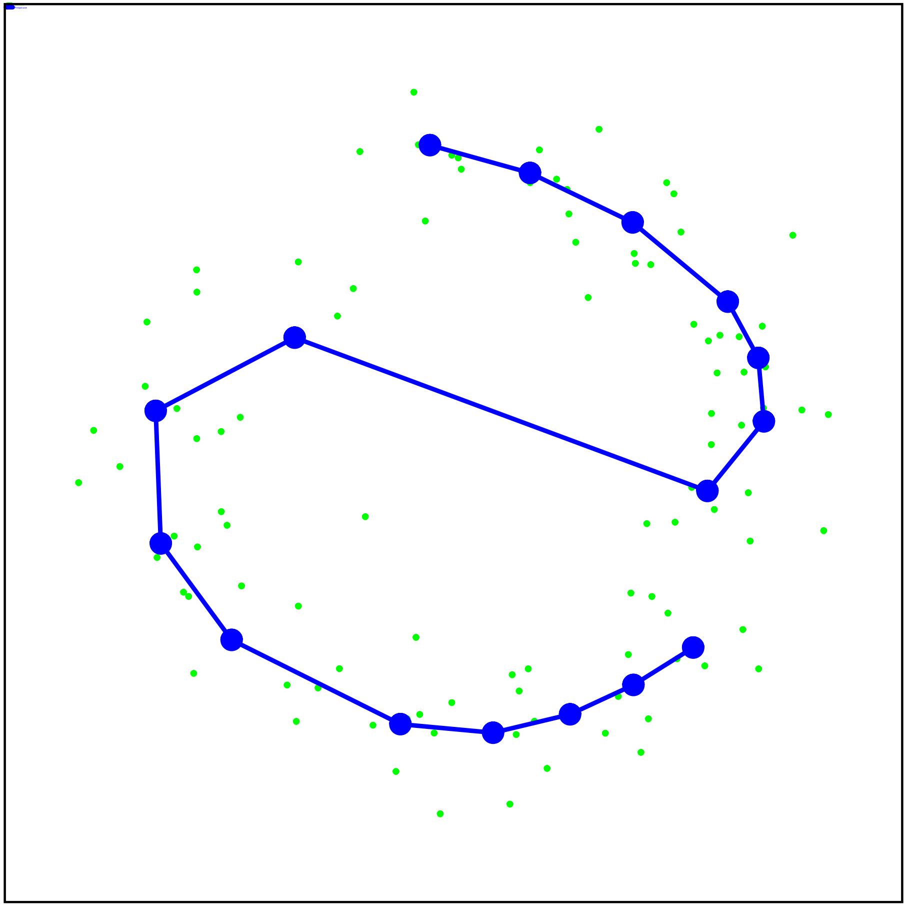
Challenge #2: The warping

- “Bad” initialization, local minima



Challenge #2: The warping

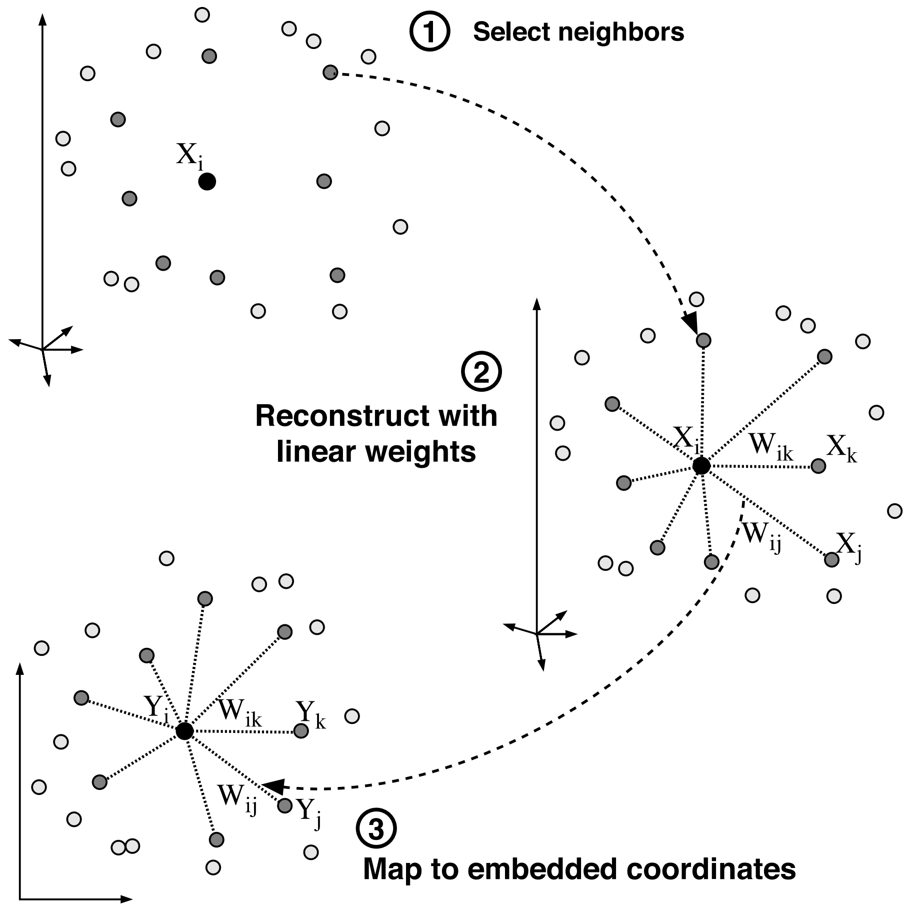
- “Bad” initialization, local minima



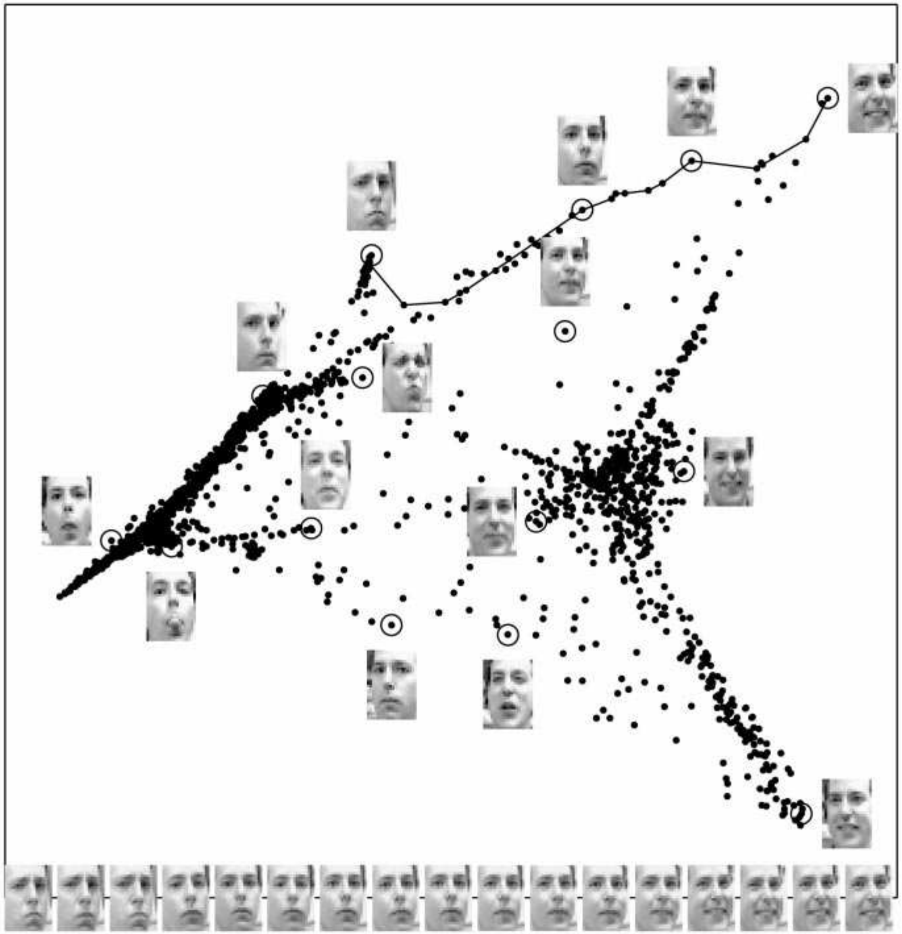
Challenge #2: The warping

- Solution: “one shot” methods
  - Local Linear Embedding, ISOMAP, Kernel PCA, and other spectral methods
  - non-geometric (implicit manifolds)
  - handle complex structures but break down with noise
  - relatively slow

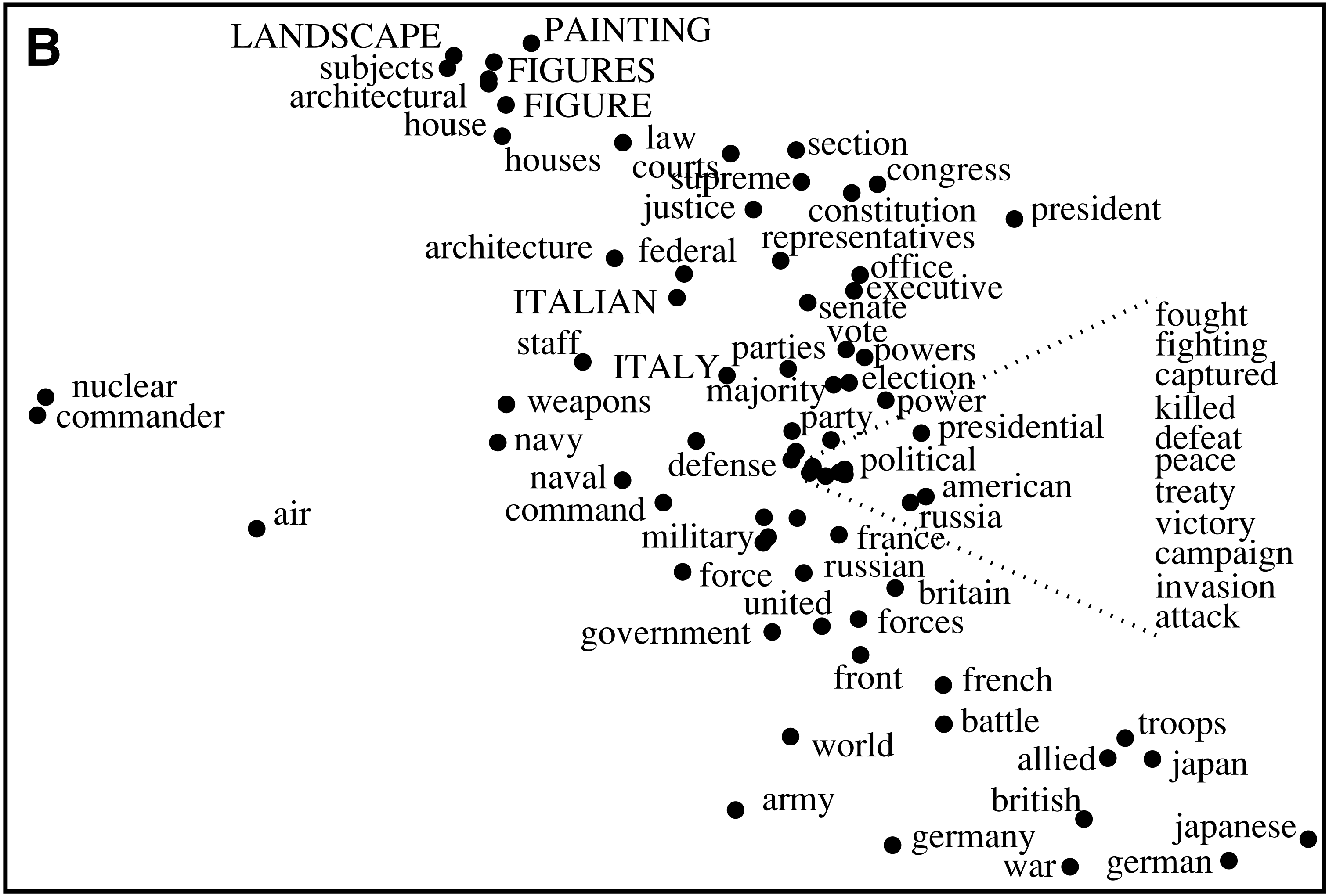
Local Linear Embedding



Local Linear Embedding

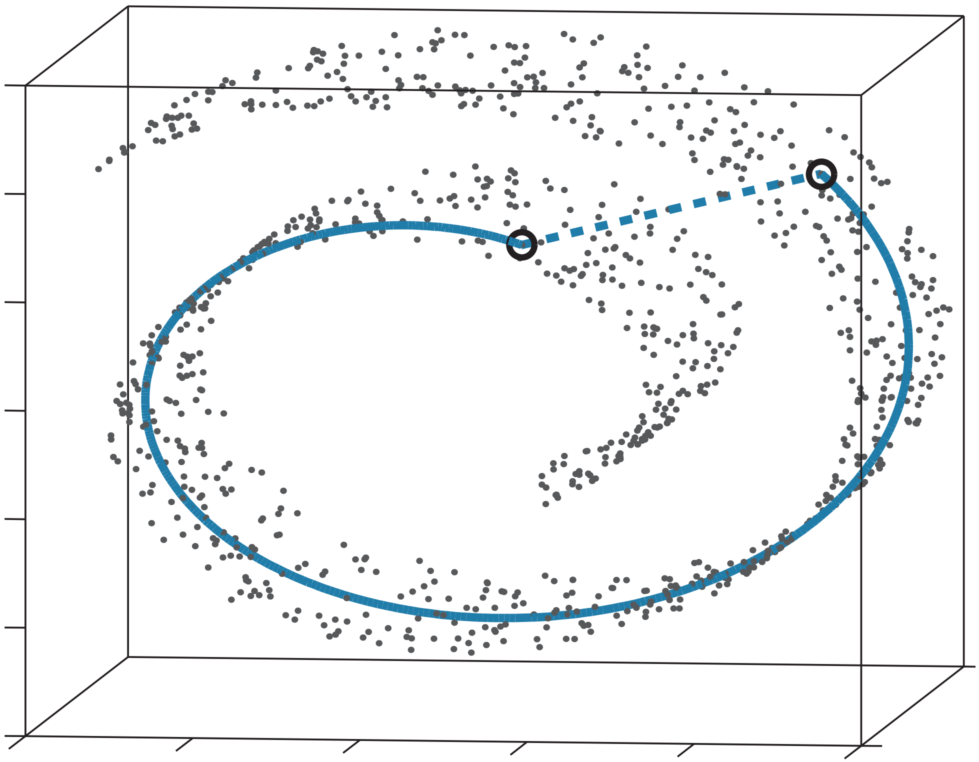


Local Linear Embedding



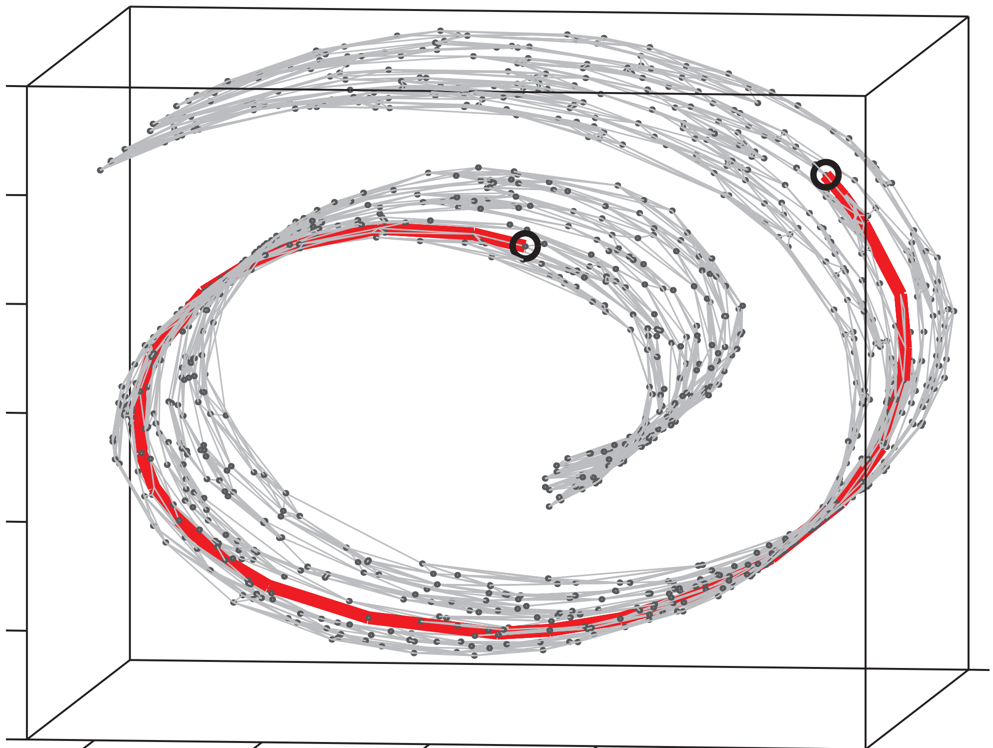
ISOMAP

- Problem: the geodesic distance (distance on the manifold) is much longer than the euclidien distance



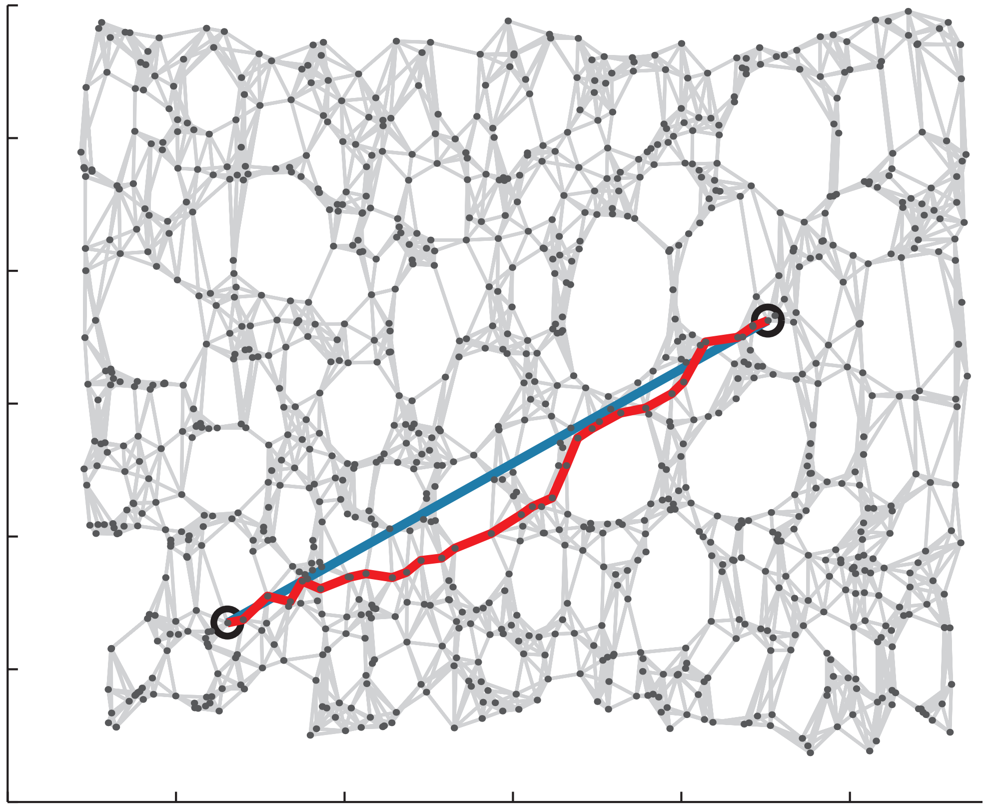
ISOMAP

- Solution: construct the neighborhood graph, and find the shortest path between each pair of points.

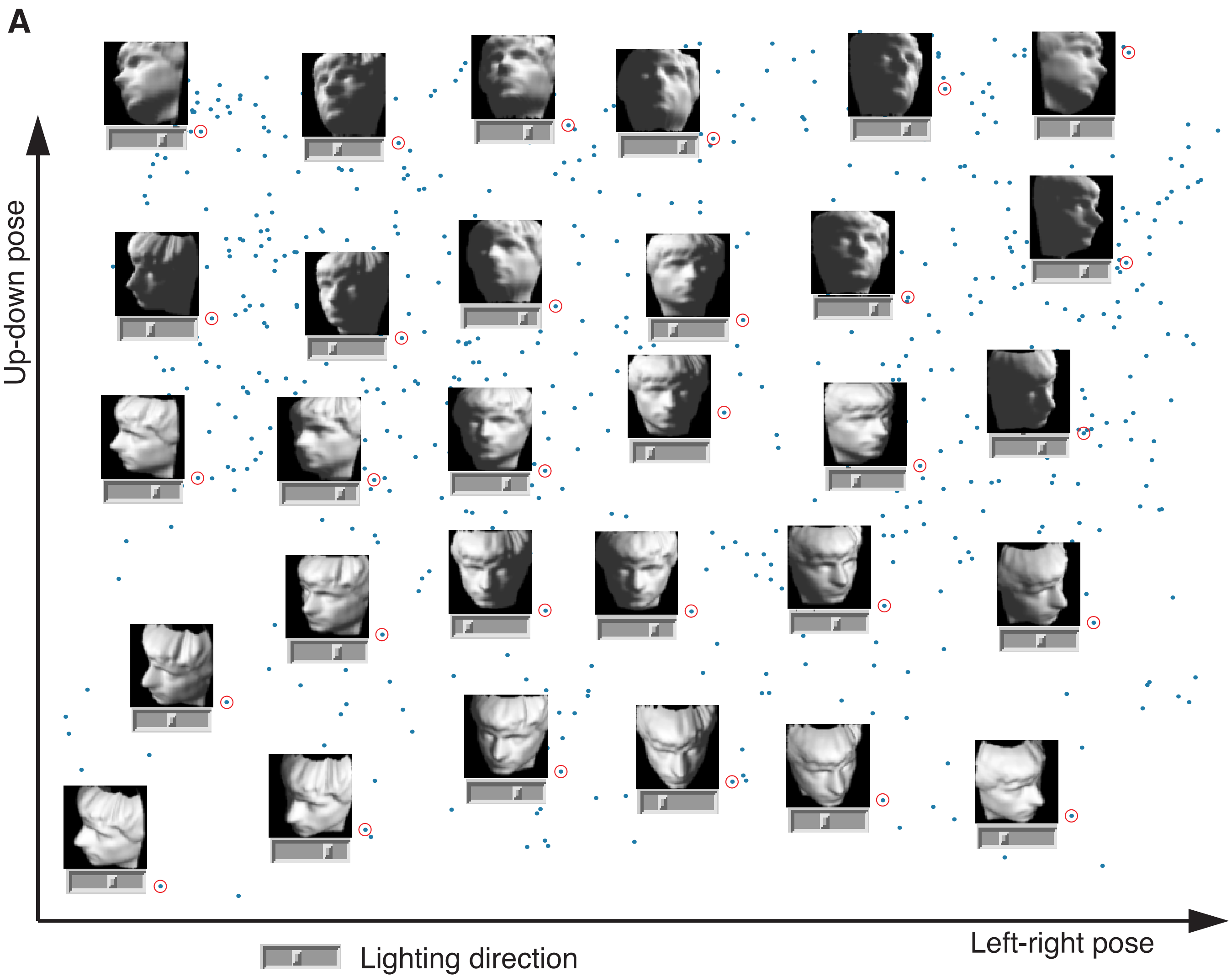


ISOMAP

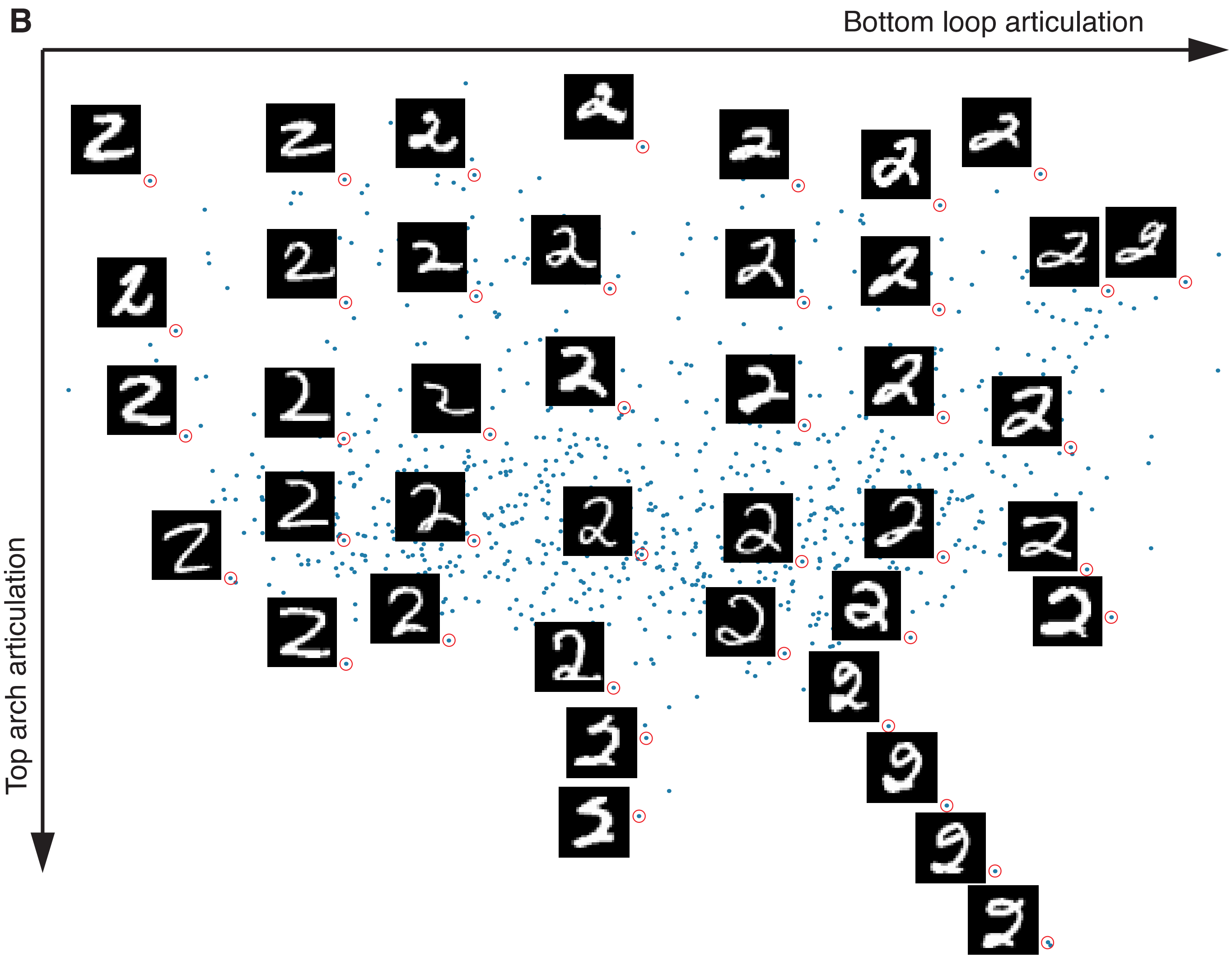
- Solution: use multidimensional scaling to map the data



ISOMAP



ISOMAP



Challenges

- More or less solved
  - nonlinearity
  - complex structures (on which “global” methods fail)
  - noise

Challenges

- Unsolved
  - noise combined with high curvature or complex structures
  - noise combined with relatively high intrinsic dimensionality – data sparseness – curse of dimensionality
  - non-smooth manifolds
  - proposed solution: non-local manifold learning, hierarchical models