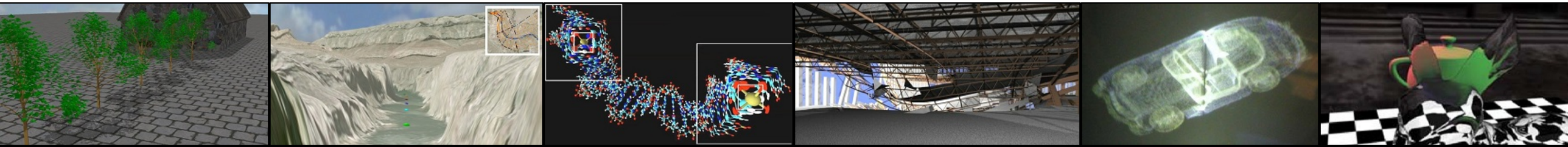
# CIS 4930/6930-002
# DATA VISUALIZATION

## Machine Learning & Data Mining

Paul Rosen

Assistant Professor

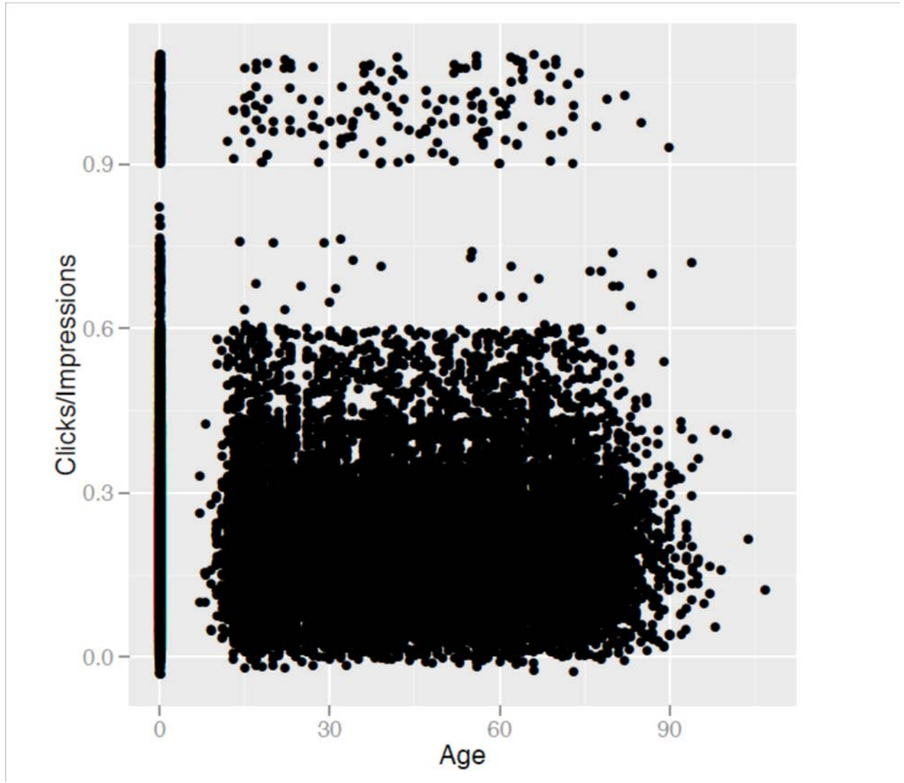
University of South Florida

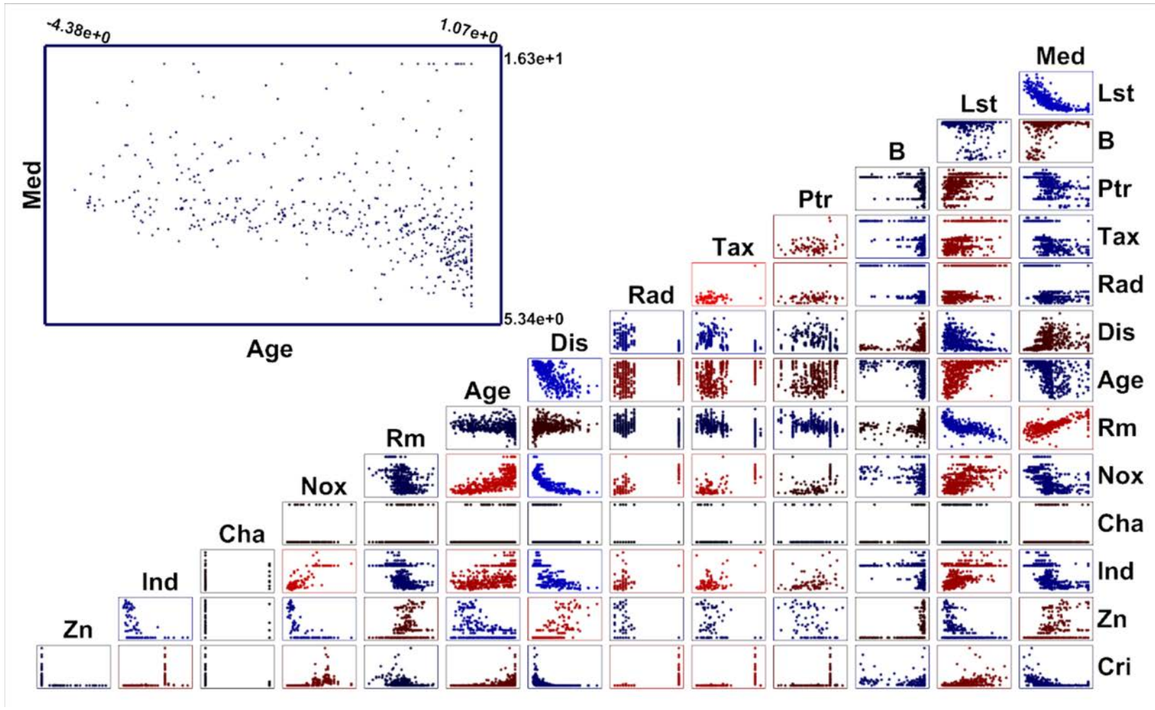


# THE PROBLEM

Number of data points and attributes are large

Limited human visual capacity



# THE SOLUTION

use the machine to do most of the work

# BUT HOW?

# STATISTICS

As we will see, statistics can help in a number of ways, but they don't give you everything you need/want

# DATA MINING TOOLS

Help to interpret data by (hopefully) reducing them to their core components/features

# WE'LL (BRIEFLY) TALK ABOUT

Dimensionality Reduction

Classifiers

Regression

Clustering

# Dimensionality Reduction

Comes in 2 basic flavors, linear and nonlinear

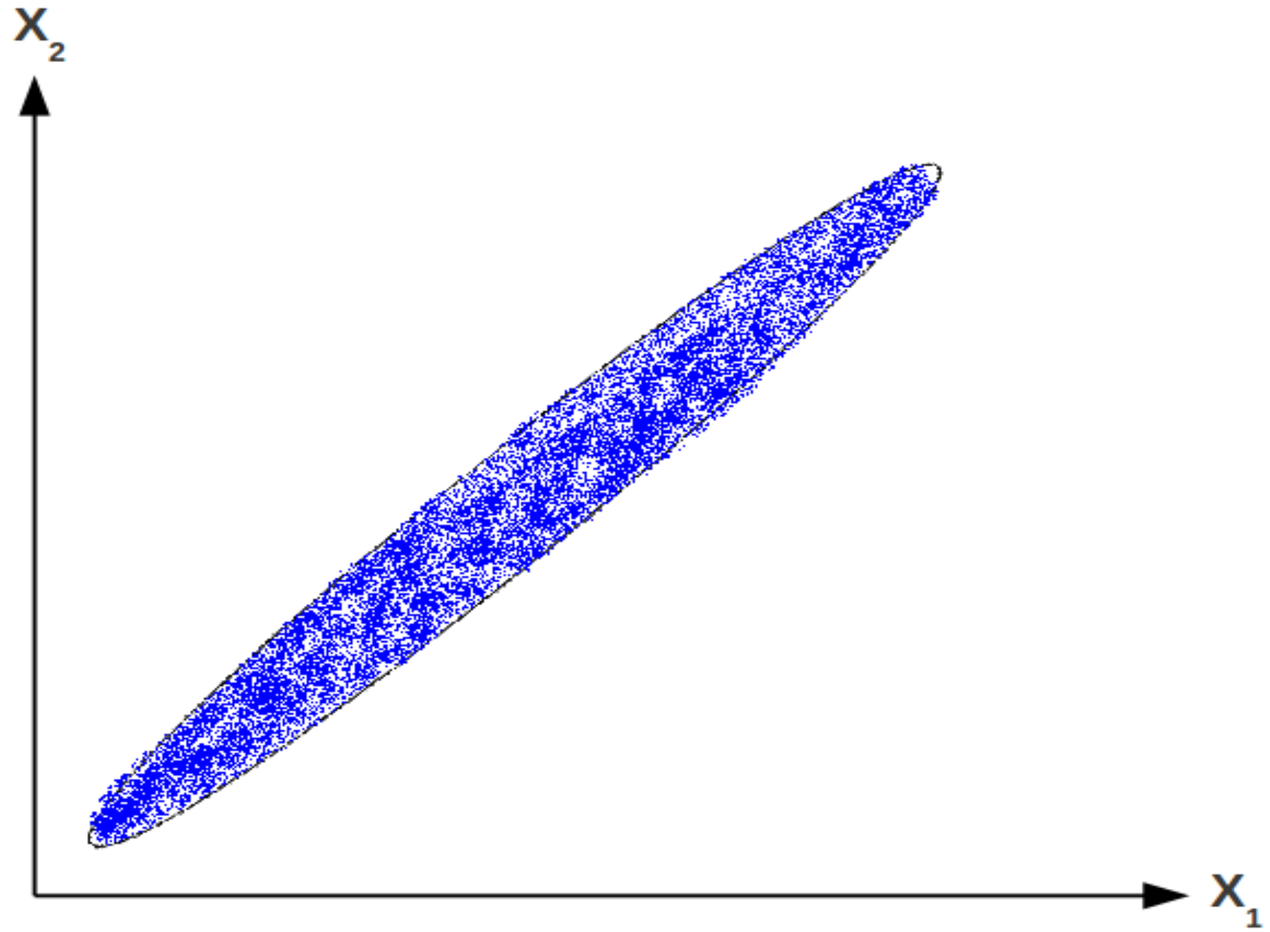
# Linear Dimensionality reduction: Principal Component Analysis (PCA)

Given a set of data points embedded in any number of dimensions

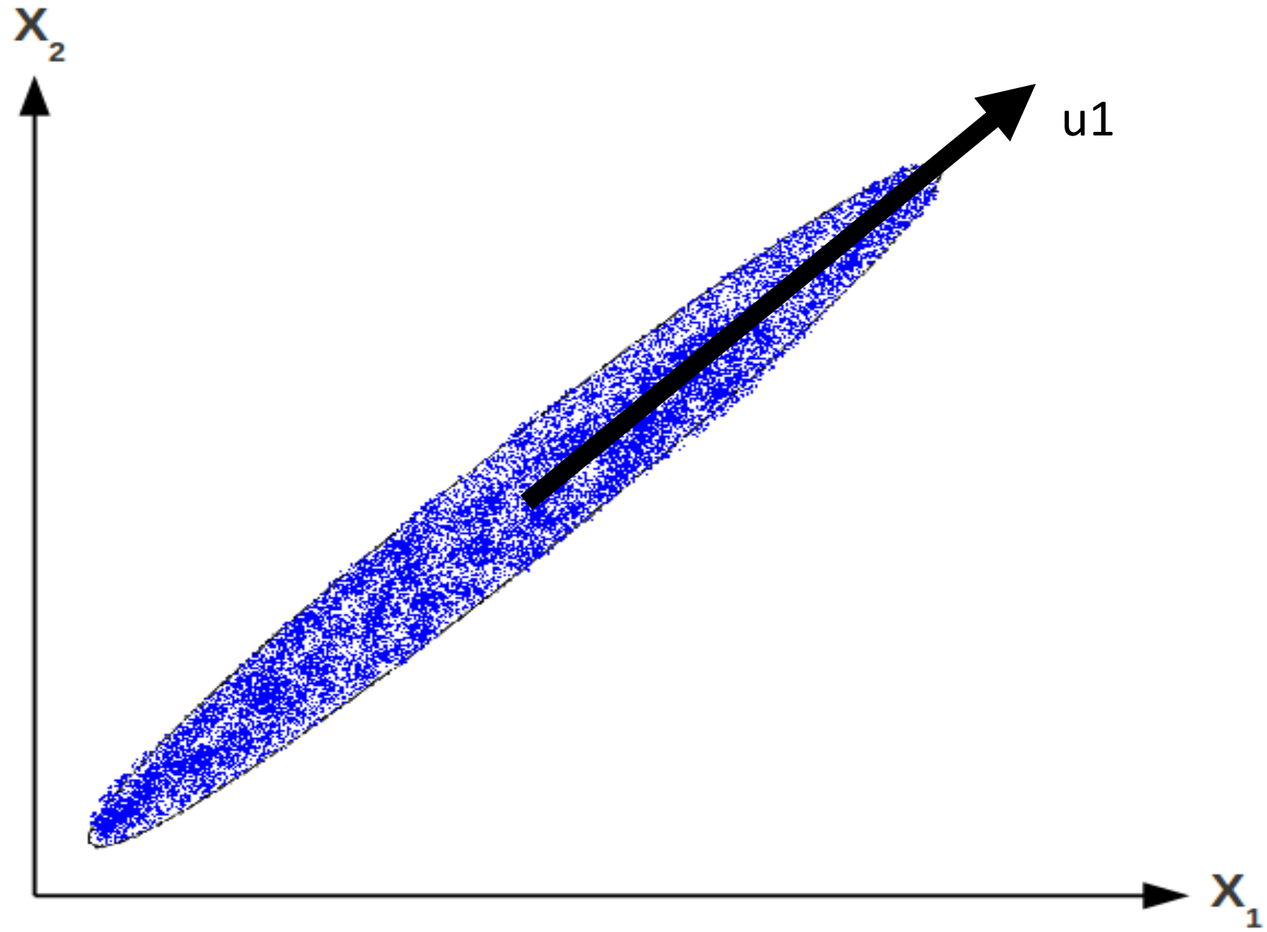
Find the vector (direction) with the largest variance

Repeat for additional vectors, finding the next largest variance orthogonal to all previous principal components

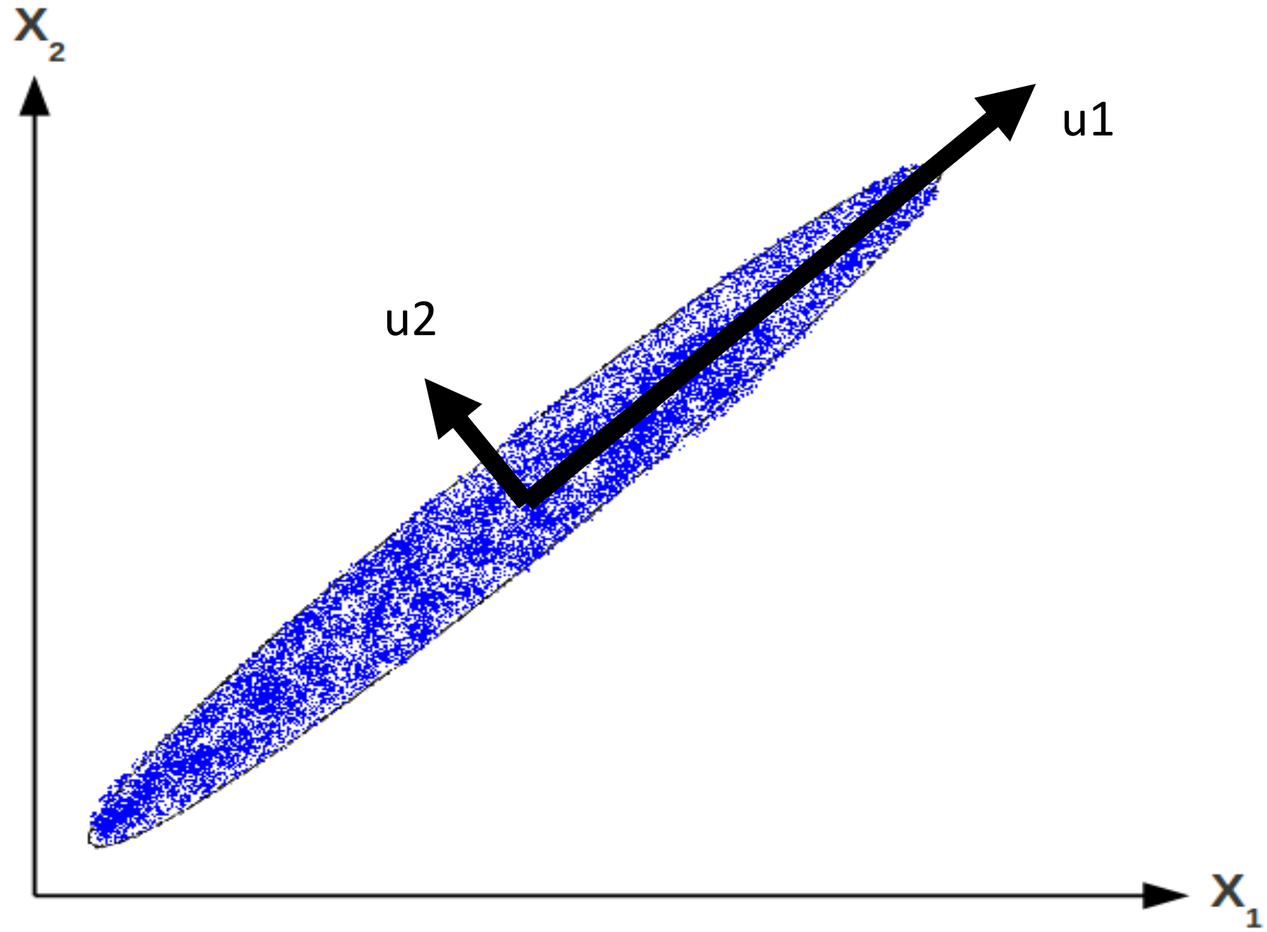
# PCA: AN EXAMPLE



# PCA: AN EXAMPLE



# PCA: AN EXAMPLE

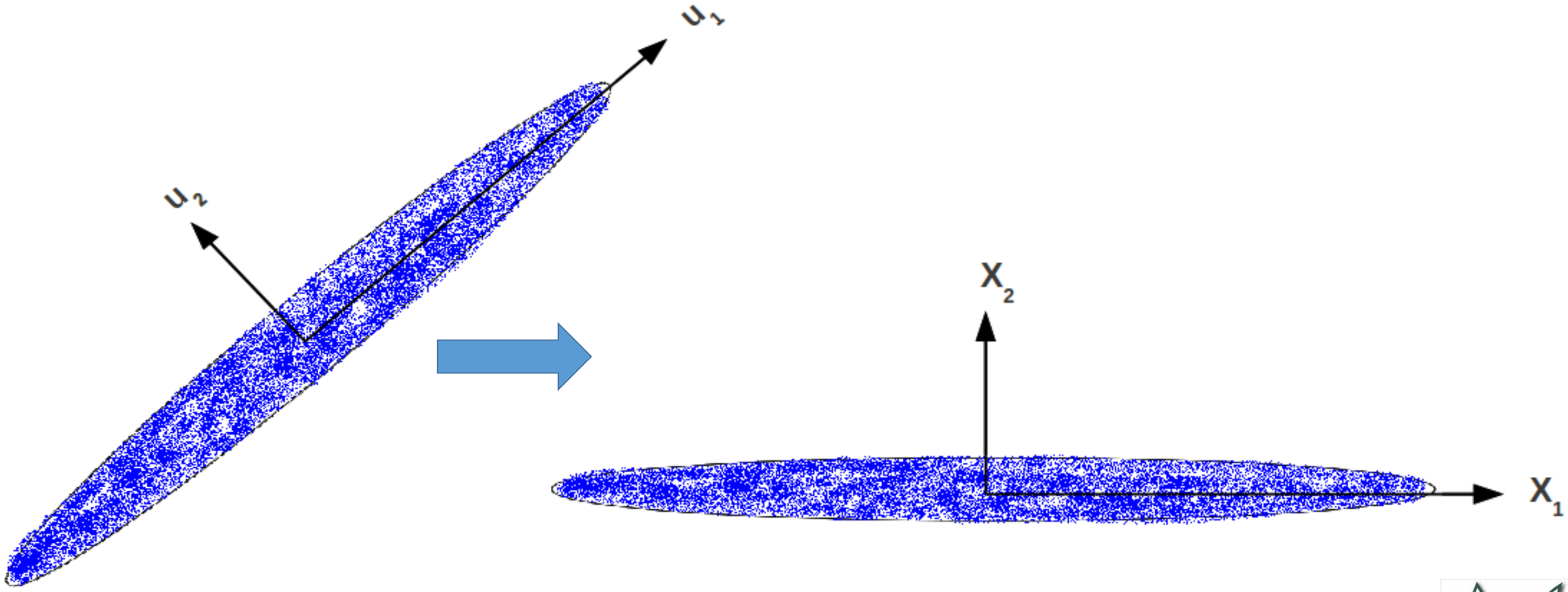


# PCA Provides

A new basis that maximizes variance, where the length of the vectors indicates variance in that direction

Common practice is to reproject points by subtracting the mean and adjusting basis

# PCA: AN EXAMPLE



# PCA Provides

If data is higher than 2D, the first 2 principal components can be selected (or 3 for 3D)

# WHAT DO WE GAIN/LOSE…

# WHEN IS THIS A GOOD/BAD IDEA…

In the 2D case?

In an nD case?

# Principal Component Analysis (PCA)

Can be calculated by finding the eigenvectors $O(d^3)$ of the covariance matrix $O(nd^2)$

Practically speaking, the covariance matrix finding dominates computation, since $n \gg d$

Subsampling points can improve performance, as long as the sampling is *good*

# Nonlinear Dimensionality Reduction: Multidimensional Scaling (MDS)

Given pairwise distances/dissimilarities

Find a map that preserves distances

# MDS

Many “flavors”

We’ll look at classical MDS

# Basic MDS Process

Given a dissimilarity (distance) matrix $D = (d_{ij})$

MDS seeks to find an embedding in $P$ dimensions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in R^P$, so that $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$

For certain types of dissimilarity matrices, this exists for some large $P$

More generally, we want $d_{ij} \approx \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$, as close as possible

# Basic MDS Process

The mapping from matrix $D = (d_{ij})$ to $d_{ij} \approx \|\boldsymbol{x_i} - \boldsymbol{x_j}\|$ is found using eigenvector/eigenvalue decomposition of the dissimilarity matrix

Although this is a linear operation, the effect to data is nonlinear

# METRIC VS. NON-METRIC SPACE

Distance and dissimilarity are defined for any pair of objects in any space. A distance function is also called metric if it satisfies…

| | |
|---|---|
| $d(x, y) \geq 0,$ | non-negativity |
| $d(x, y) = 0 \; iff \; x = y$ | identity |
| $d(x, y) = d(y, x)$ | symmetry |
| $d(x, z) \leq d(x, y) + d(y, z)$ | triangle inequality |

# Metric vs. Non-metric space

Given a set of dissimilarities, one can ask whether these values are distances, and whether the space they define is metric or not

# Why does this matter?

Non-metric data requires special treatment

# CLASSICAL MDS EXAMPLE: AIRLINE DISTANCES

**TABLE 13.2.** *Airline distances (km) between 18 cities. Source: Atlas of the World, Revised 6th Edition, National Geographic Society, 1995, p. 131.*

| | Beijing | Cape Town | Hong Kong | Honolulu | London | Melbourne |
|---|---|---|---|---|---|---|
| Cape Town | 12947 | | | | | |
| Hong Kong | 1972 | 11867 | | | | |
| Honolulu | 8171 | 18562 | 8945 | | | |
| London | 8160 | 9635 | 9646 | 11653 | | |
| Melbourne | 9093 | 10338 | 7392 | 8862 | 16902 | |
| Mexico | 12478 | 13703 | 14155 | 6098 | 8947 | 13557 |
| Montreal | 10490 | 12744 | 12462 | 7915 | 5240 | 16730 |
| Moscow | 5809 | 10101 | 7158 | 11342 | 2506 | 14418 |
| New Delhi | 3788 | 9284 | 3770 | 11930 | 6724 | 10192 |
| New York | 11012 | 12551 | 12984 | 7996 | 5586 | 16671 |
| Paris | 8236 | 9307 | 9650 | 11988 | 341 | 16793 |
| Rio de Janeiro | 17325 | 6075 | 17710 | 13343 | 9254 | 13227 |
| Rome | 8144 | 8417 | 9300 | 12936 | 1434 | 15987 |
| San Francisco | 9524 | 16487 | 11121 | 3857 | 8640 | 12644 |
| Singapore | 4465 | 9671 | 2575 | 10824 | 10860 | 6050 |
| Stockholm | 6725 | 10334 | 8243 | 11059 | 1436 | 15593 |
| Tokyo | 2104 | 14737 | 2893 | 6208 | 9585 | 8159 |

| | Mexico | Montreal | Moscow | New Delhi | New York | Paris |
|---|---|---|---|---|---|---|
| Montreal | 3728 | | | | | |
| Moscow | 10740 | 7077 | | | | |
| New Delhi | 14679 | 11286 | 4349 | | | |
| New York | 3362 | 533 | 7530 | 11779 | | |
| Paris | 9213 | 5522 | 2492 | 6601 | 5851 | |

# CLASSICAL MDS EXAMPLE: AIRLINE DISTANCES

Negative eigenvalues indicate Airline distance is non-metric

Should only use at most first 3 eigenvectors (negative eigenvalues produce resulting in the imaginary domain)

TABLE 13.6. *Eigenvalues of* B *and the eigenvectors corresponding to the first three largest eigenvalues (in red) for the airline distances example.*

| | Eigenvalues | Eigenvectors | | |
|---|---|---|---|---|
| 1 | 471582511 | 0.245 | –0.072 | 0.183 |
| 2 | 316824787 | 0.003 | 0.502 | -0.347 |
| 3 | 253943687 | 0.323 | –0.017 | 0.103 |
| 4 | –98466163 | 0.044 | –0.487 | -0.080 |
| 5 | –74912121 | –0.145 | 0.144 | 0.205 |
| 6 | –47505097 | 0.366 | –0.128 | –0.569 |
| 7 | 31736348 | –0.281 | –0.275 | –0.174 |
| 8 | –7508328 | –0.272 | –0.115 | 0.094 |
| 9 | 4338497 | –0.010 | 0.134 | 0.202 |
| 10 | 1747583 | 0.209 | 0.195 | 0.110 |
| 11 | –1498641 | –0.292 | –0.117 | 0.061 |
| 12 | 145113 | –0.141 | 0.163 | 0.196 |
| 13 | –102966 | –0.364 | 0.172 | –0.473 |
| 14 | 60477 | –0.104 | 0.220 | 0.163 |
| 15 | –6334 | –0.140 | –0.356 | –0.009 |
| 16 | –1362 | 0.375 | 0.139 | –0.054 |
| 17 | 100 | –0.074 | 0.112 | 0.215 |
| 18 | 0 | 0.260 | –0.214 | 0.173 |

# CLASSICAL MDS EXAMPLE: AIRLINE DISTANCES

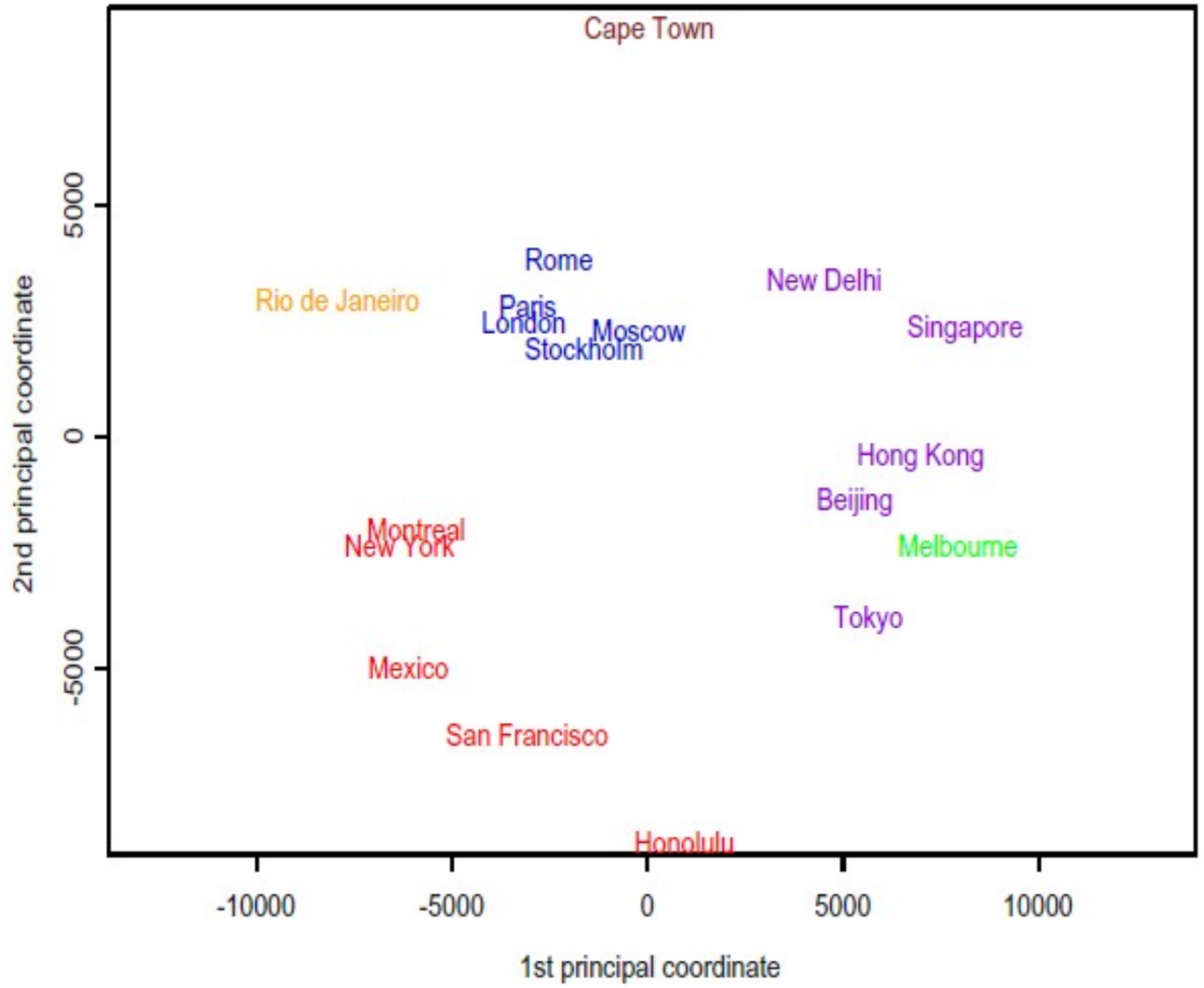



**FIGURE 13.1.** *Two-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors*

# CLASSICAL MDS EXAMPLE: AIRLINE DISTANCES

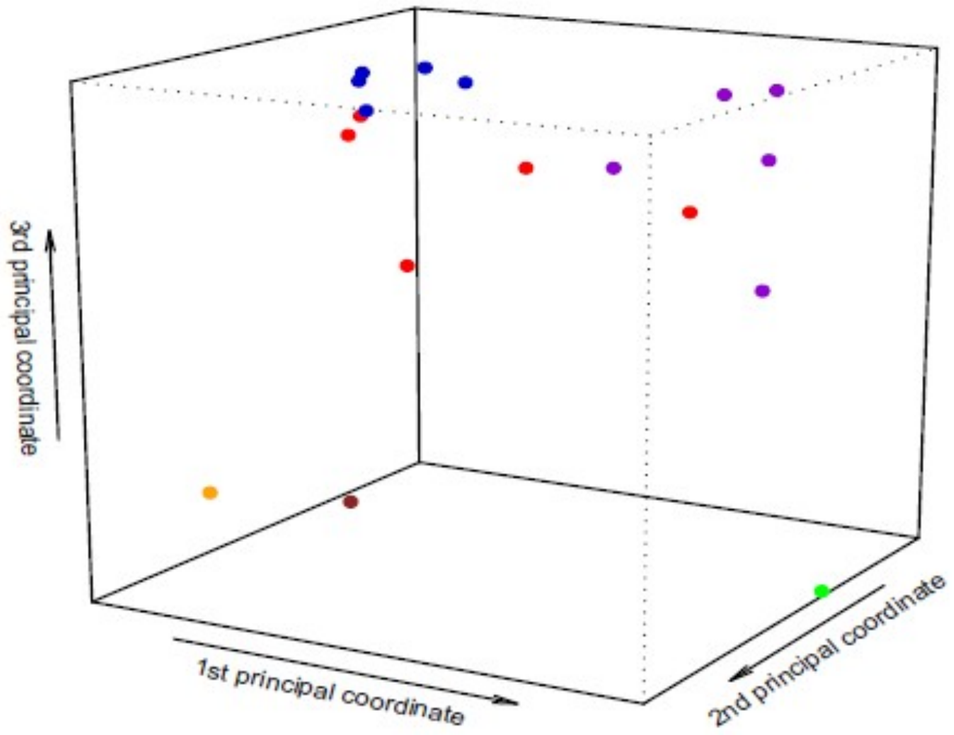



**FIGURE 13.2.** *Three-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors reflect the different continents: Asia (purple), North America (red), South America (yellow), Europe (blue), Africa (brown), and Australasia (green).*

# CLASSICAL MDS EXAMPLE: AIRLINE DISTANCES

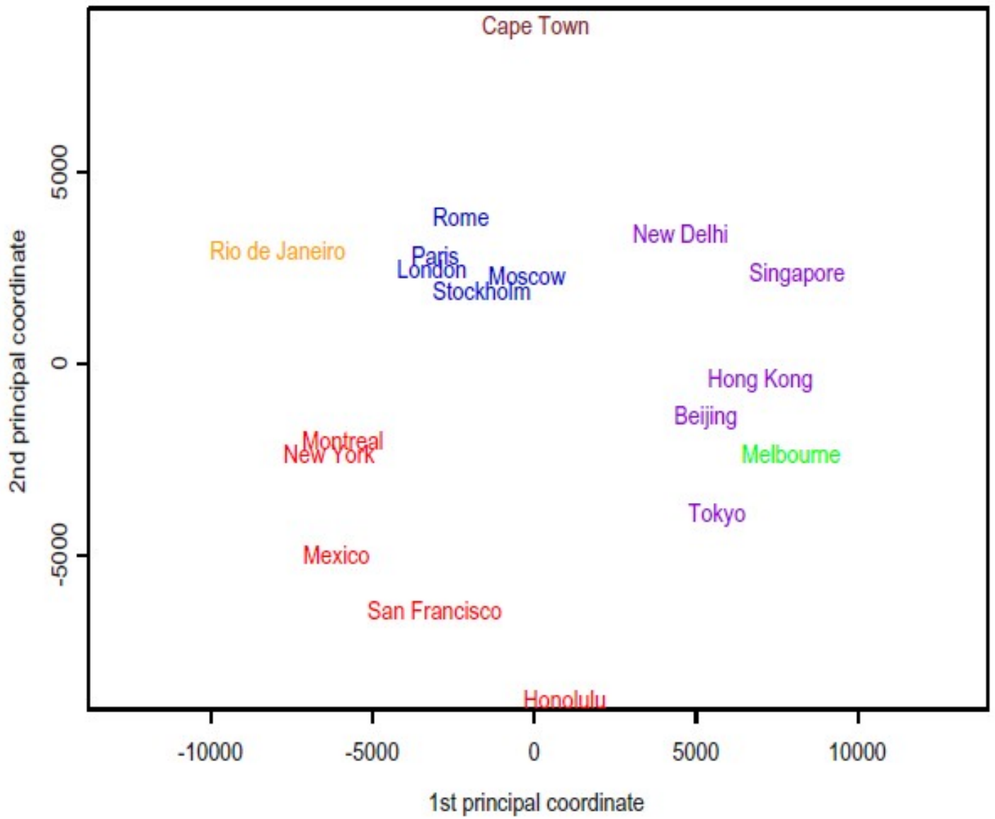



FIGURE 13.1. *Two-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors*

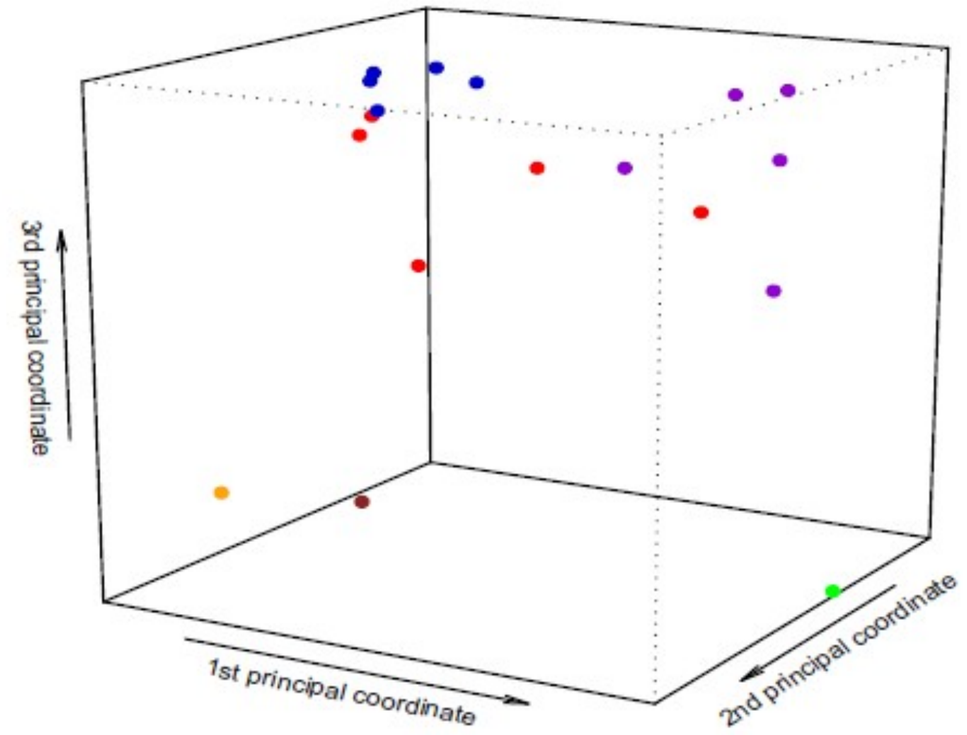
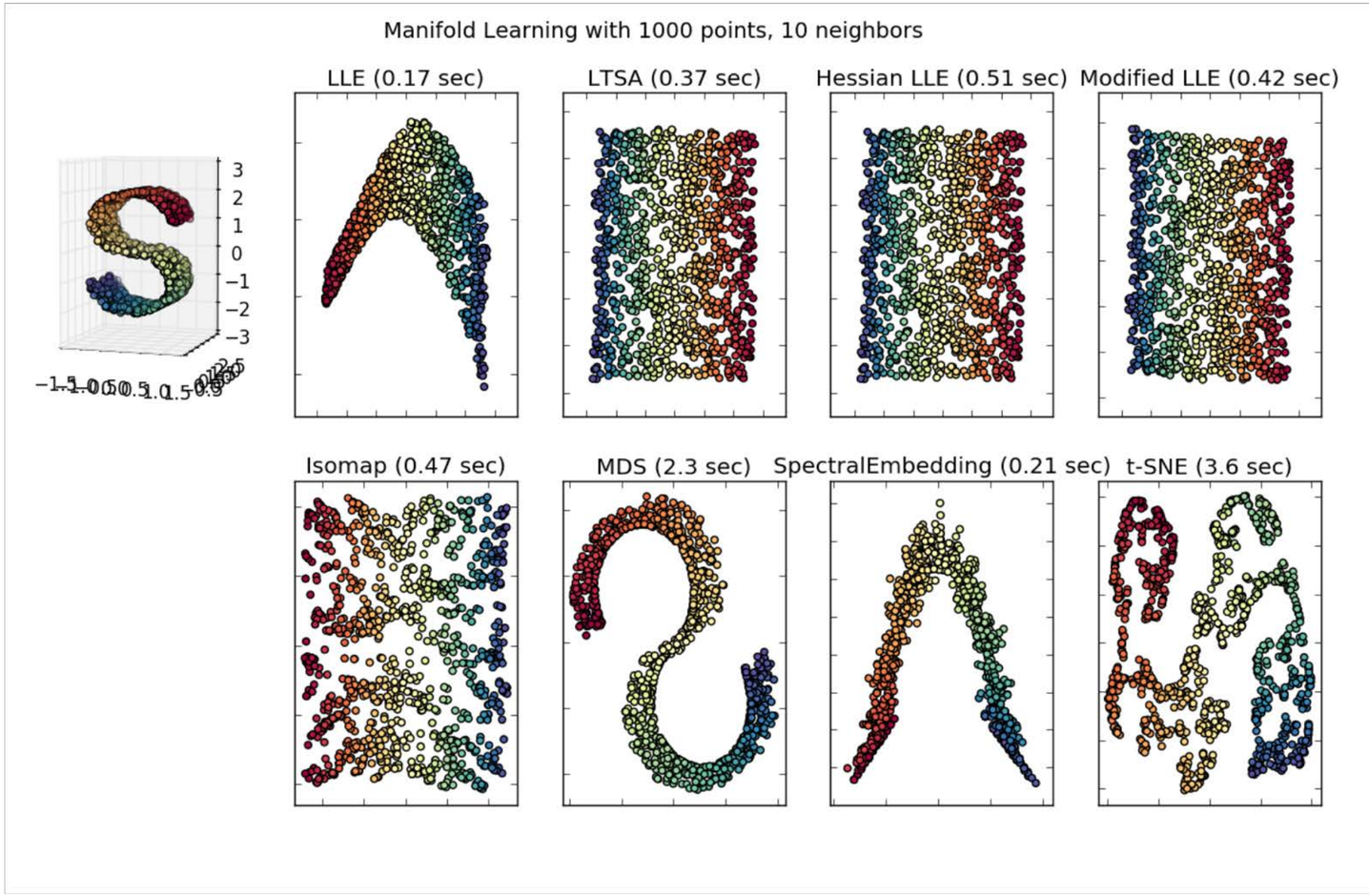



FIGURE 13.2. *Three-dimensional map of 18 world cities using the classical scaling algorithm on airline distances between those cities. The colors reflect the different continents: Asia (purple), North America (red), South America (yellow), Europe (blue), Africa (brown), and Australasia (green).*

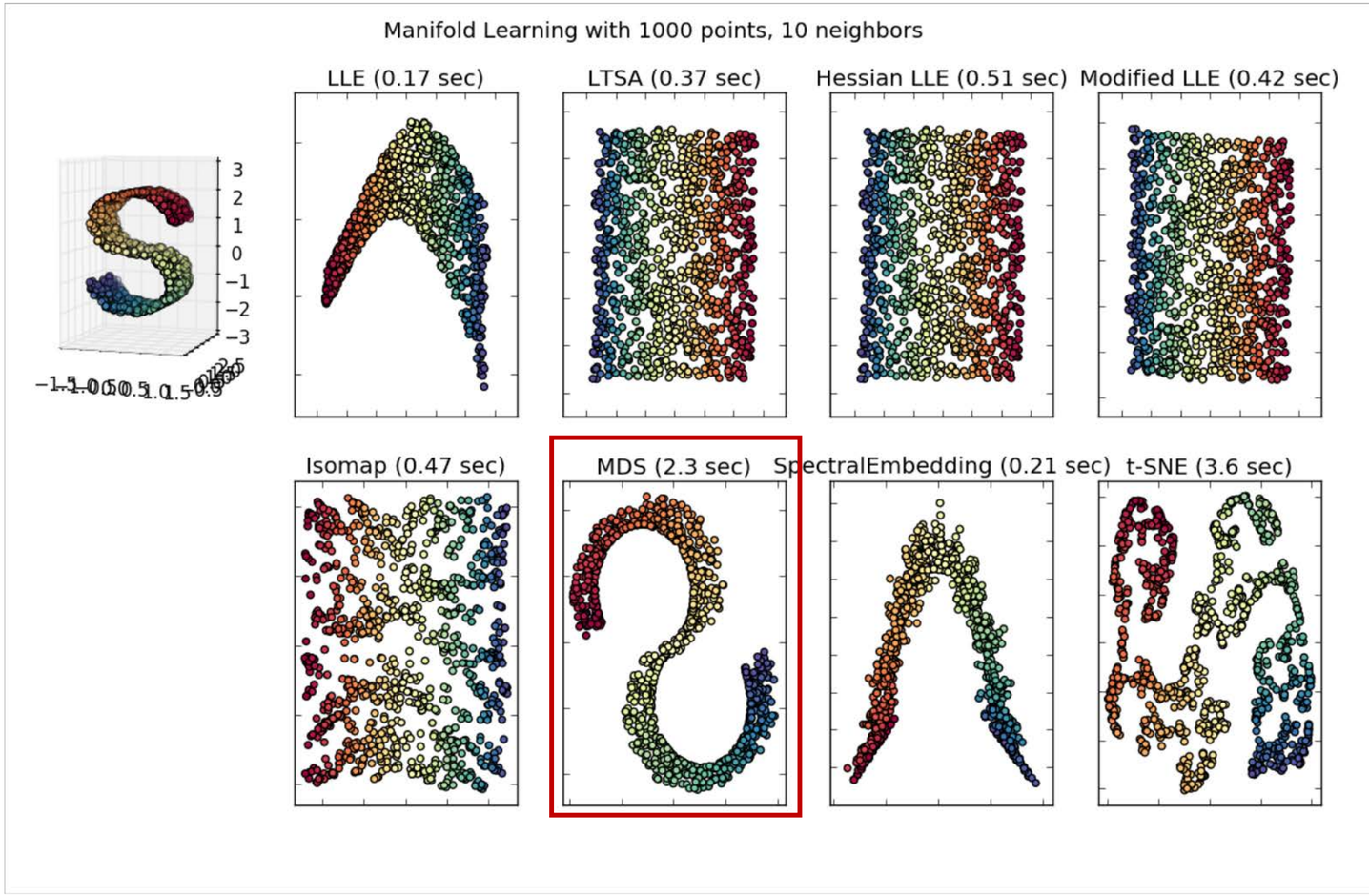
# Nonlinear Dimension Reduction

A lot of options, all preserve different features of the high dimensional space



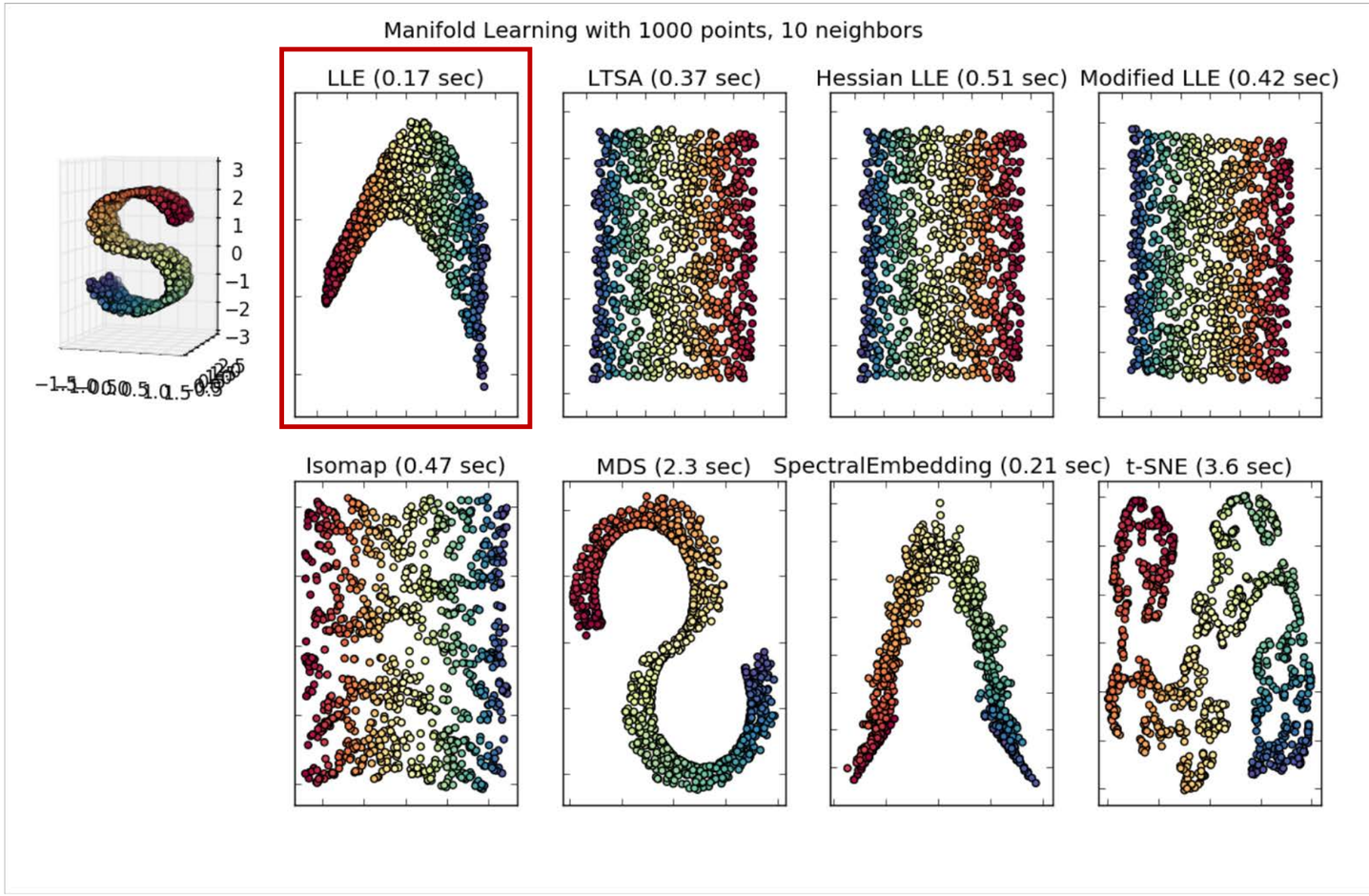
# MULTIDIMENSIONAL SCALING (MDS)

We've already seen



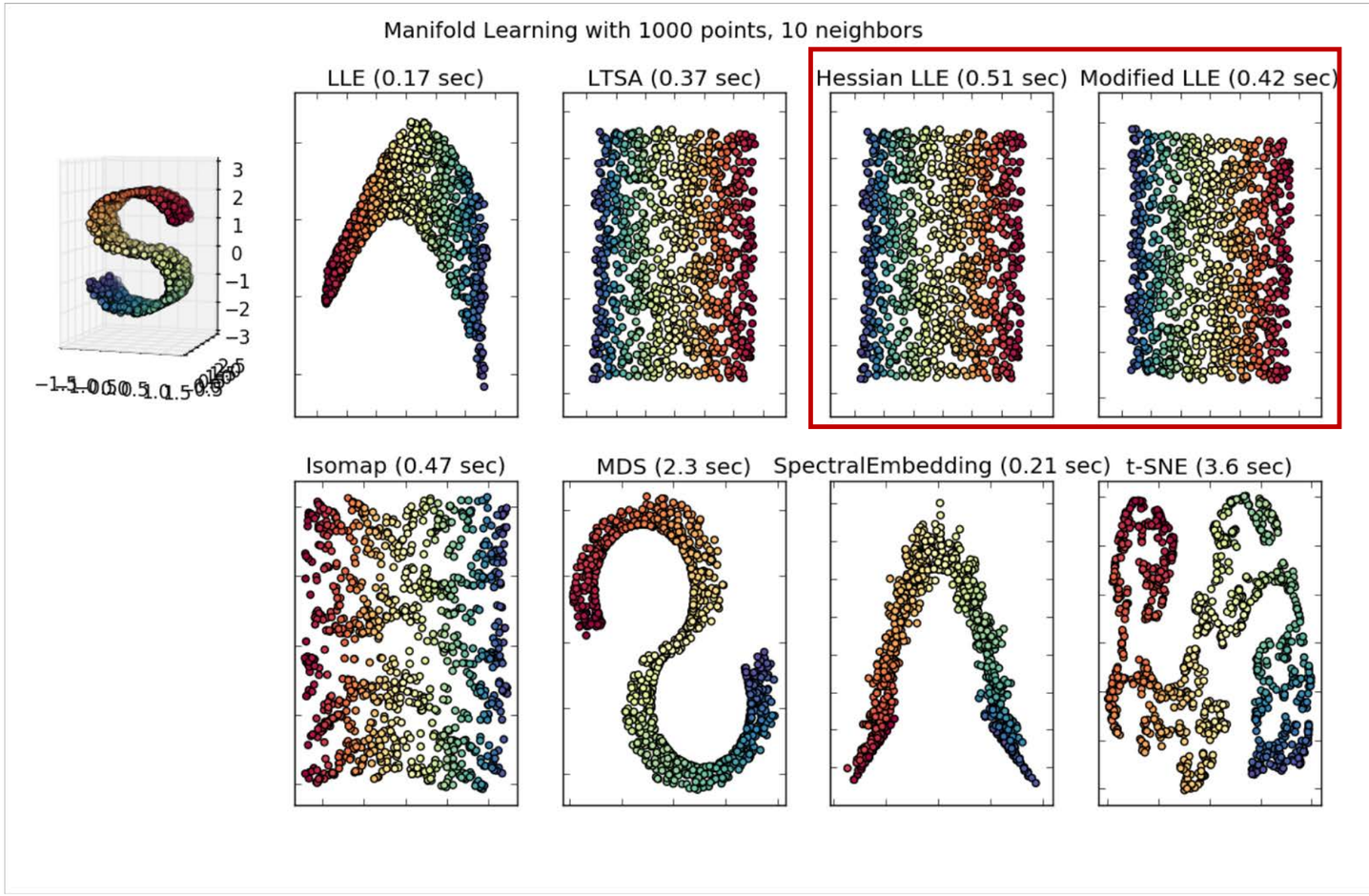
# Local Linear Embedding (LLE)

Tries to embed such that local neighborhood distances are preserved as well as possible



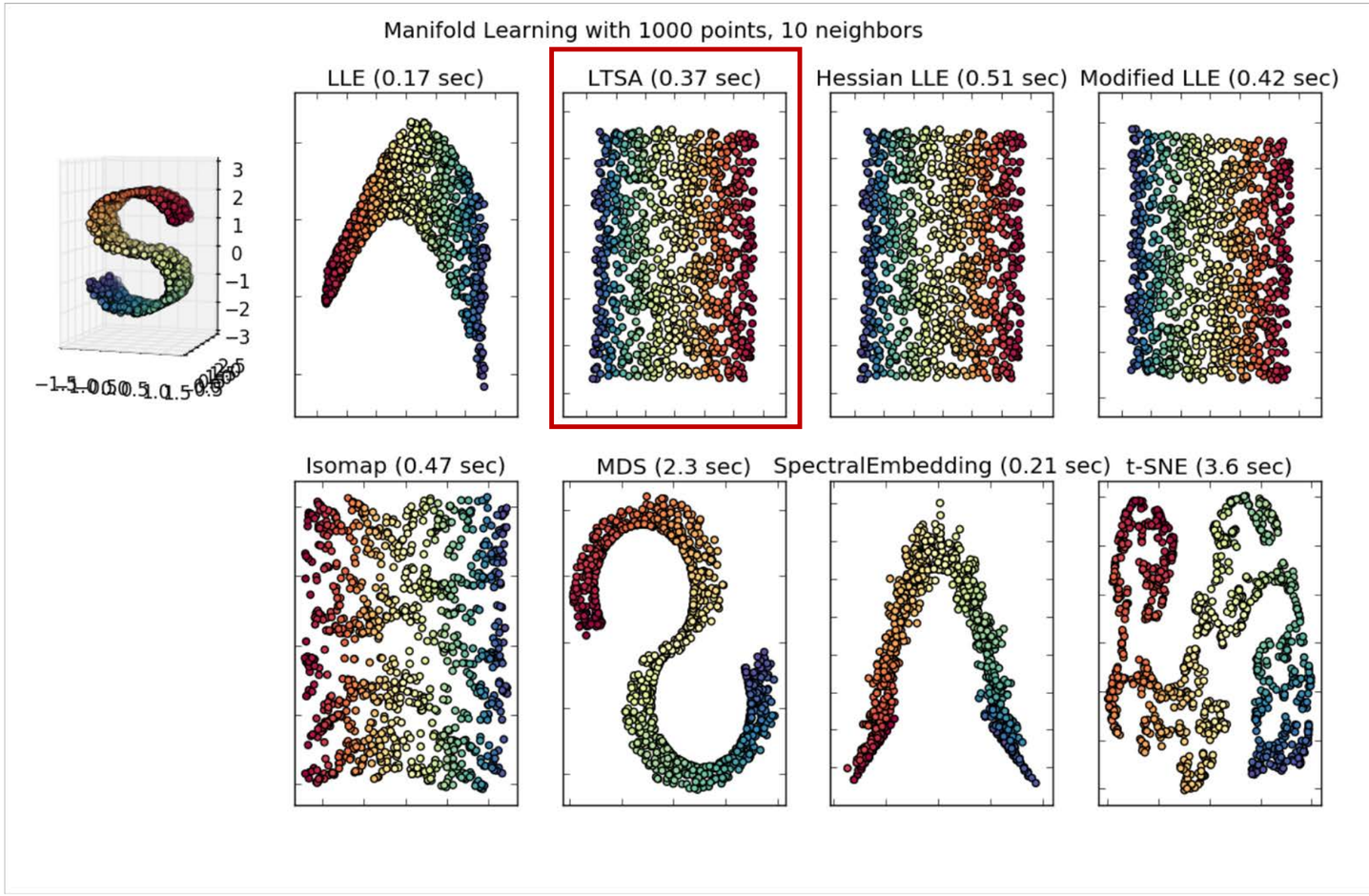
# LLE Extensions

Fix various issues associated with regularization of the domain in LLE



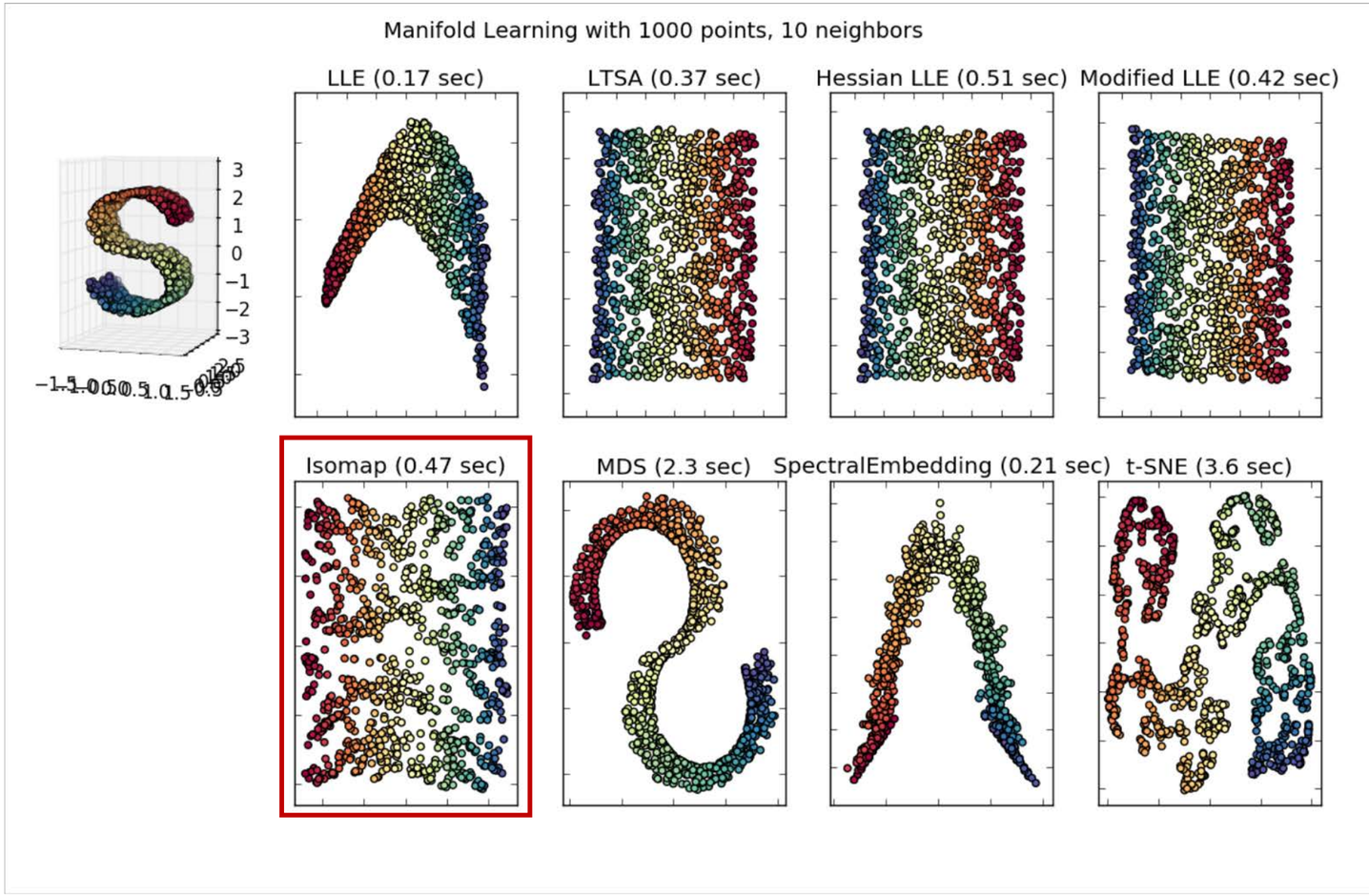
# Local Tangent Space Alignment (LTSA)

Tries to characterize geometry via local tangent directions



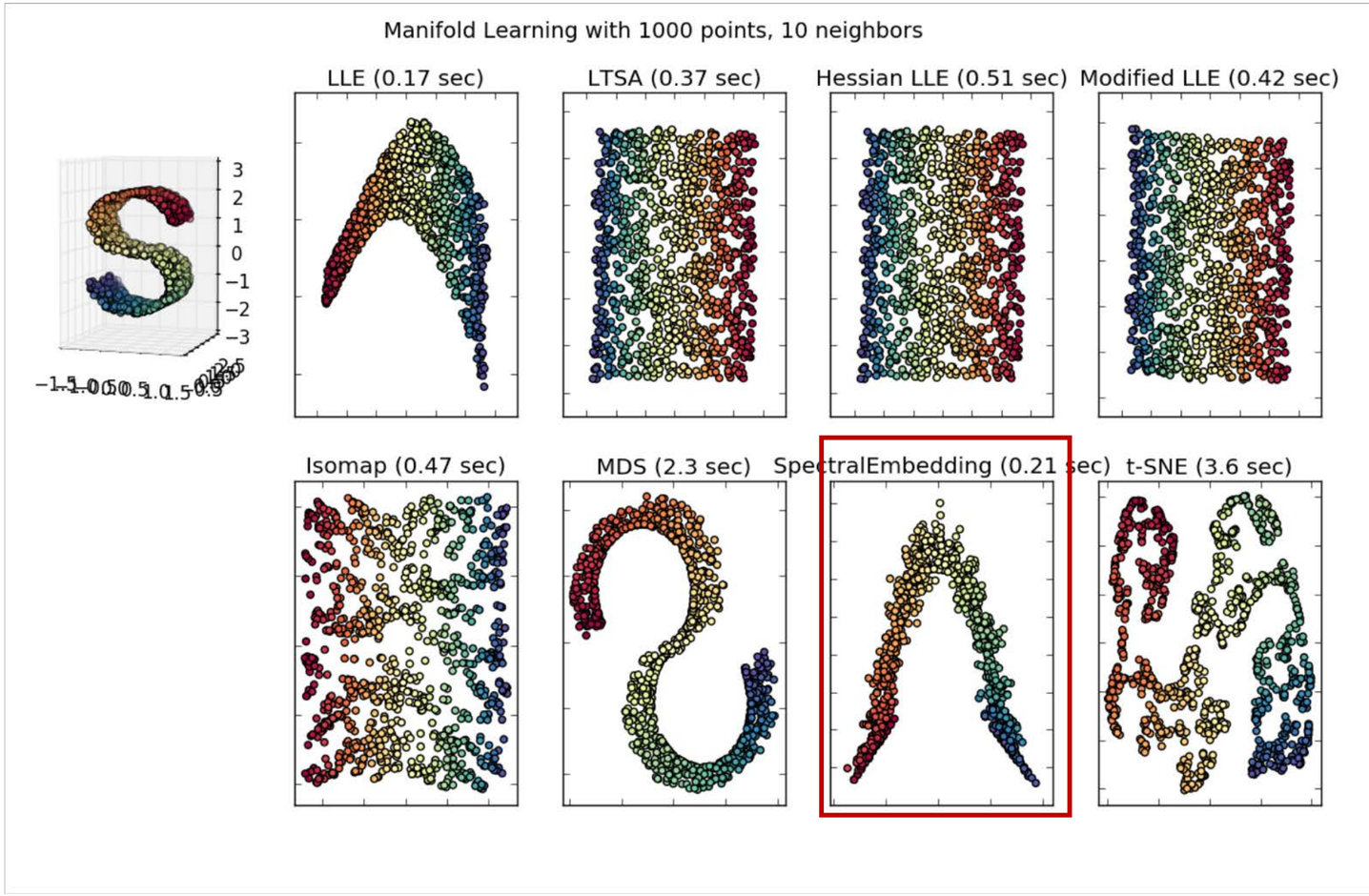
# ISOMAP

Extension of MDS. Tries to preserve geodesic distance as well as possible



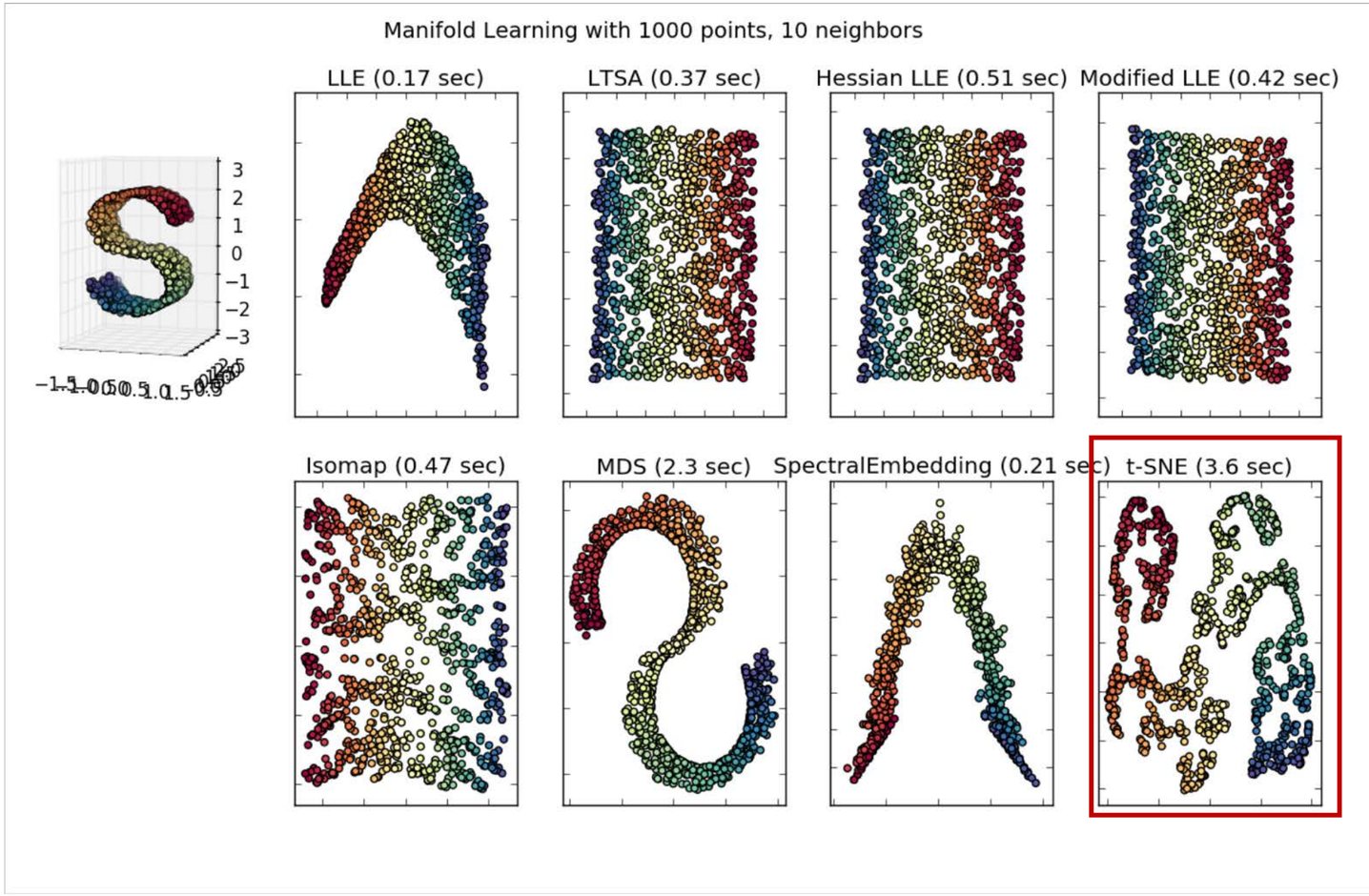
# SPECTRAL EMBEDDING

Uses a spectral decomposition of the graph Laplacian



# T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE)

Converts the points probability distribution and samples that distribution



# SO…VISUALIZING DATA, WHAT DO YOU USE AND WHEN?