

# Unsupervised Learning

Chapter 14: The Elements of Statistical Learning

Presented for 540  
by Len Tanaka

# Objectives

- Introduction
- Techniques:
  - Association Rules
  - Cluster Analysis
  - Self-Organizing Maps
  - Projective Methods
  - Multidimensional Scaling

# New Setup

- Supervised:
  - $D = \{ (x^{(i)}, y^{(i)}) \mid 1 \leq i \leq N, x \in \mathbb{R}^p, y \in \mathbb{R} \text{ or } \mathbb{D} \}$
  - $\Pr(X, Y) = \Pr(Y|X) \cdot \Pr(X)$
- Unsupervised:
  - $D = \{ (x^{(i)}) \mid 1 \leq i \leq N, x \in \mathbb{R}^p \}$
  - $Y$  is from  $X$

# Methods

- Find simple descriptions
  - Association rules
- Find distinct classes or types
  - Cluster analysis
- Find associations among  $p$  variables
  - Principal components, multidimensional scaling, self-organizing maps, principal curves

# Association Rules

- Find joint values of  $X = \{X_1, X_2, \dots, X_p\}$
- Example: “Market basket” analysis
  - $X_{ij} \in \{0, 1\}$  if product  $i$  is purchased with  $j$
- Rather than finding bumps...find regions

# Association Rules

- Let  $S_j$  be set of all values for  $j$ th variable
- $s_j \subseteq S_j$
- $\Pr[\bigcap_{j=1\dots p}(X_j \in s_j)]$  (14.2: conjunctive rule)
- $K = \sum_{j=1\dots p}|S_j|$  ( $K$  dummy variables:  $Z_1\dots Z_k$ )

$$\Pr \left[ \bigcap_{k \in \mathcal{K}} (Z_k = 1) \right] = \Pr \left[ \prod_{k \in \mathcal{K}} Z_k = 1 \right]$$

# Associative Rules

- $T(K) = \widehat{\text{Pr}} \left[ \prod_{k \in K} (Z_k = 1) \right] = \frac{1}{N} \sum_{i=1}^N \prod_{k \in K} z_{ik}$
- $T(K)$  is the prevalence of  $K$  in the data
- Set some bound  $t$  where  $\{K_i | T(K) > t\}$

# Example

$X$	Age	Sex	Employed
$i$	31	M	yes

$K$	{<30, 30+}	{M, F}	{yes, no}
-----	------------	--------	-----------

$Z$	<30	0	M	1	yes	1
	30+	1	F	0	no	0

# Apriori Algorithm

- Agrawal et al. 1995
  - $|\{K_i | T(K) > t\}|$  is small
  - Any item set of  $L$  subset of  $K$ ,  $T(L) \geq T(K)$
- Calculate  $|K| = m$ , consider  $m-1$  items
- Throw away sets  $< t$
- Each high support analyzed

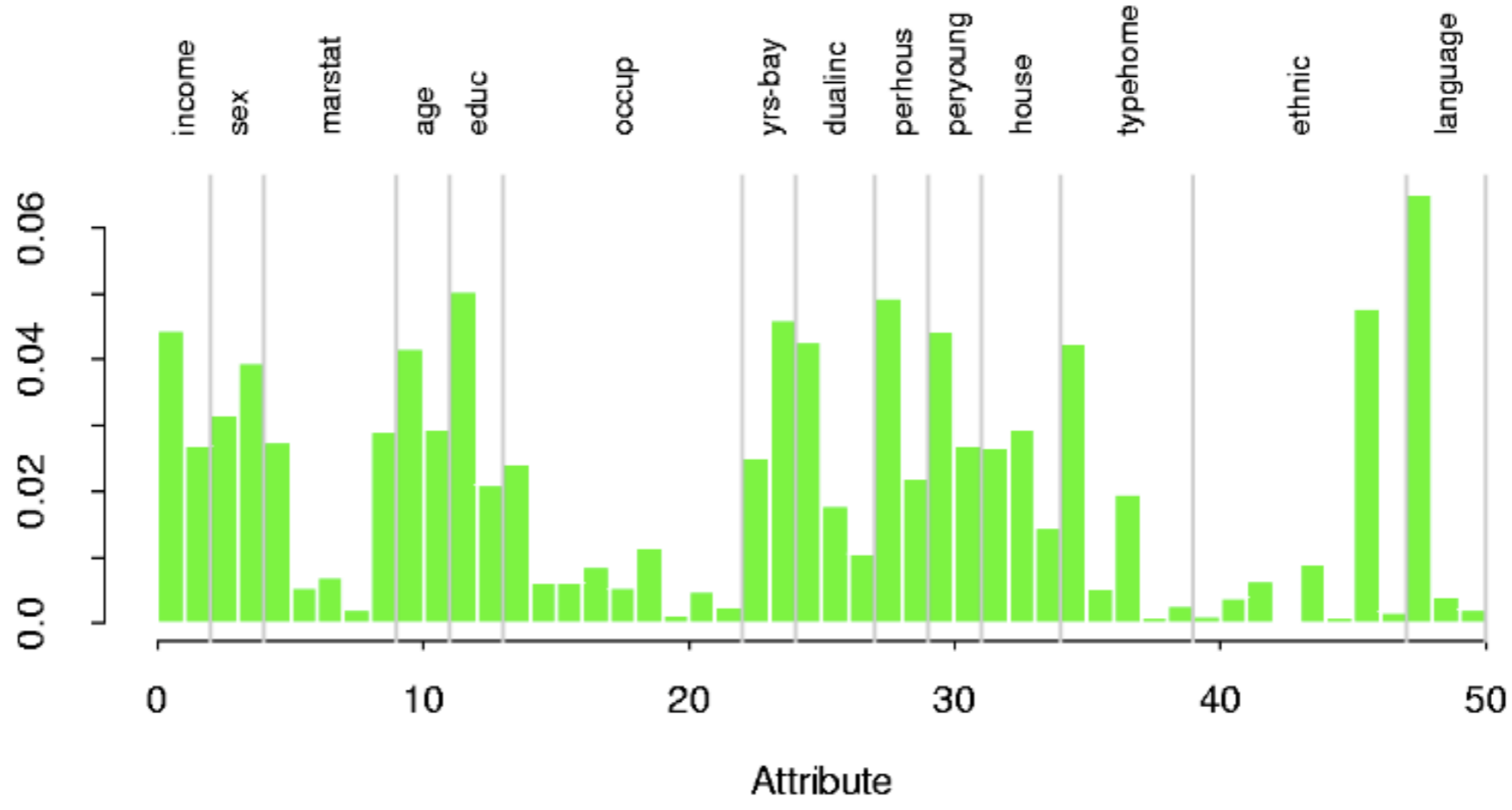
# Apriori Algorithm

- $A \Rightarrow B$
- Confidence:
  - $C(A \Rightarrow B) = T(A \Rightarrow B) / T(A)$
- Lift:
  - $L(A \Rightarrow B) = C(A \Rightarrow B) / T(B)$

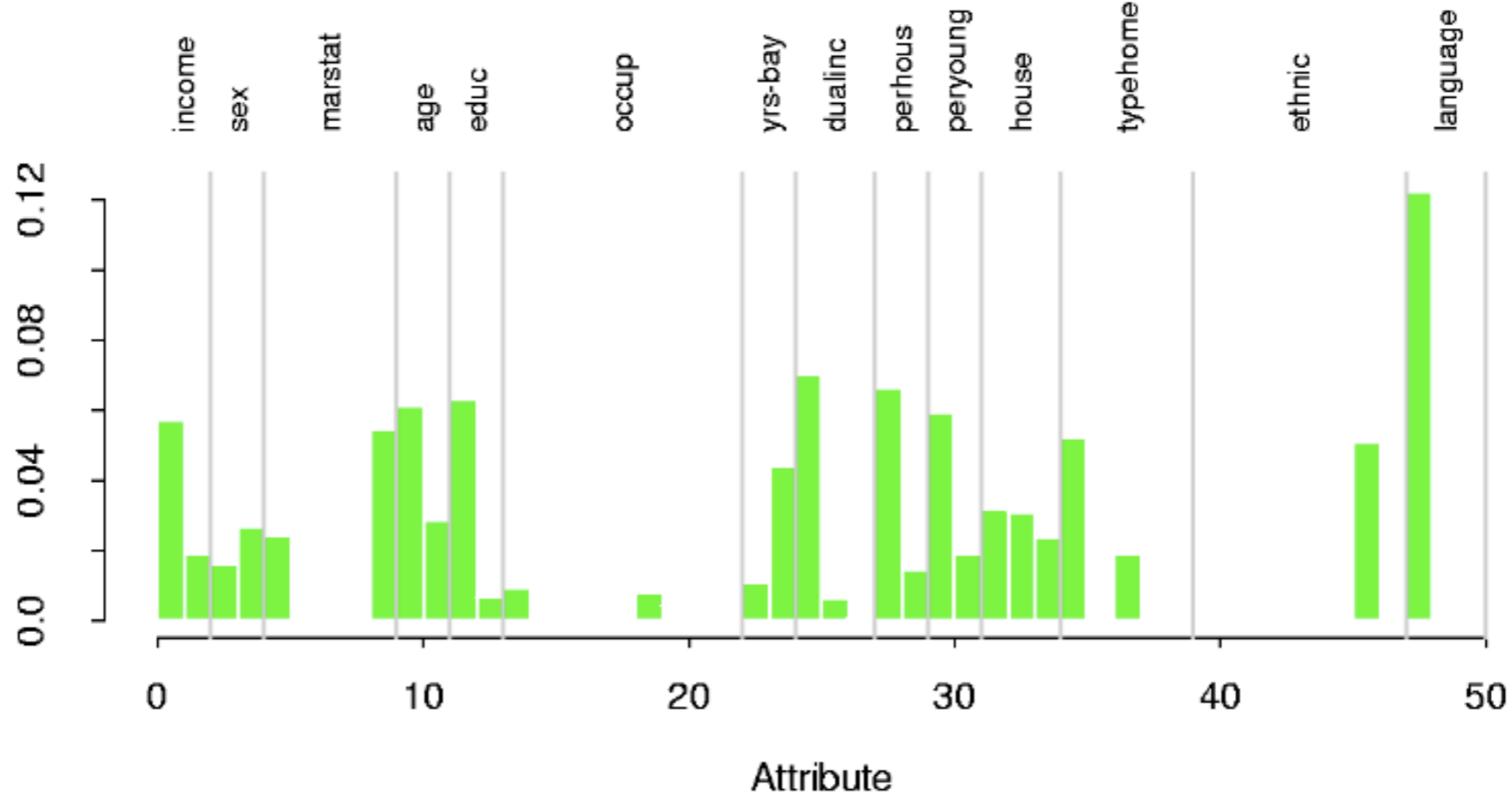
# Example:

- $K = \{\text{peanut butter, jelly, bread}\}$
- $T(\text{peanut butter, jelly} \Rightarrow \text{bread}) = 0.03$
- $C(\text{peanut butter, jelly} \Rightarrow \text{bread}) =$   
 $T(\text{pb, jelly, bread}) / T(\text{pb, jelly}) = 0.82$   
 $L(\text{pb, jelly} \Rightarrow \text{bread}) = 0.82 / T(\text{bread}) = 1.95$

Relative Frequency in Data



Relative Frequency in Association Rules



# Problems

- As threshold  $t$  decreases, solution grows exponentially
- Restrictive form of data
- Rules with high confidence or lift but low support will be lost

# Unsupervised as Supervised

- Find  $g(x)$  in terms of  $g_0(x)$ 
  - Uniform density over  $x$
  - Gaussian with same mean and covariance
- Assign  $Y = 1$  for training sample
- Randomly generate  $g_0(x)$  assign  $Y = 0$

# Convert to Supervised

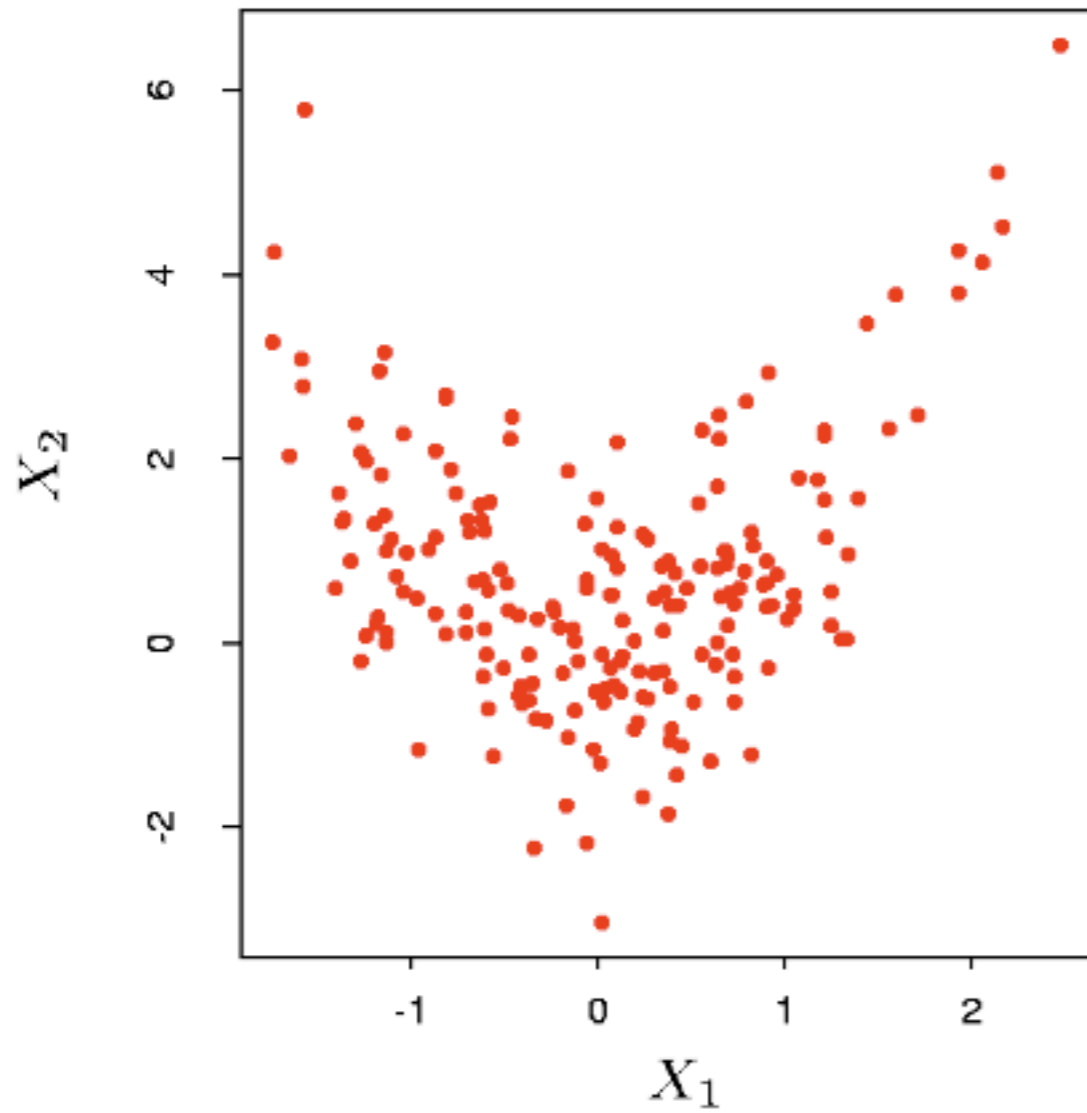
$$\begin{aligned}\mu(x) = E(Y | x) &= \frac{g(x)}{g(x) + g_0(x)} \\ &= \frac{g(x)/g_0(x)}{1 + g(x)/g_0(x)}\end{aligned}$$

$$\hat{g}(x) = g_0(x) \frac{\hat{\mu}(x)}{1 - \hat{\mu}(x)}$$

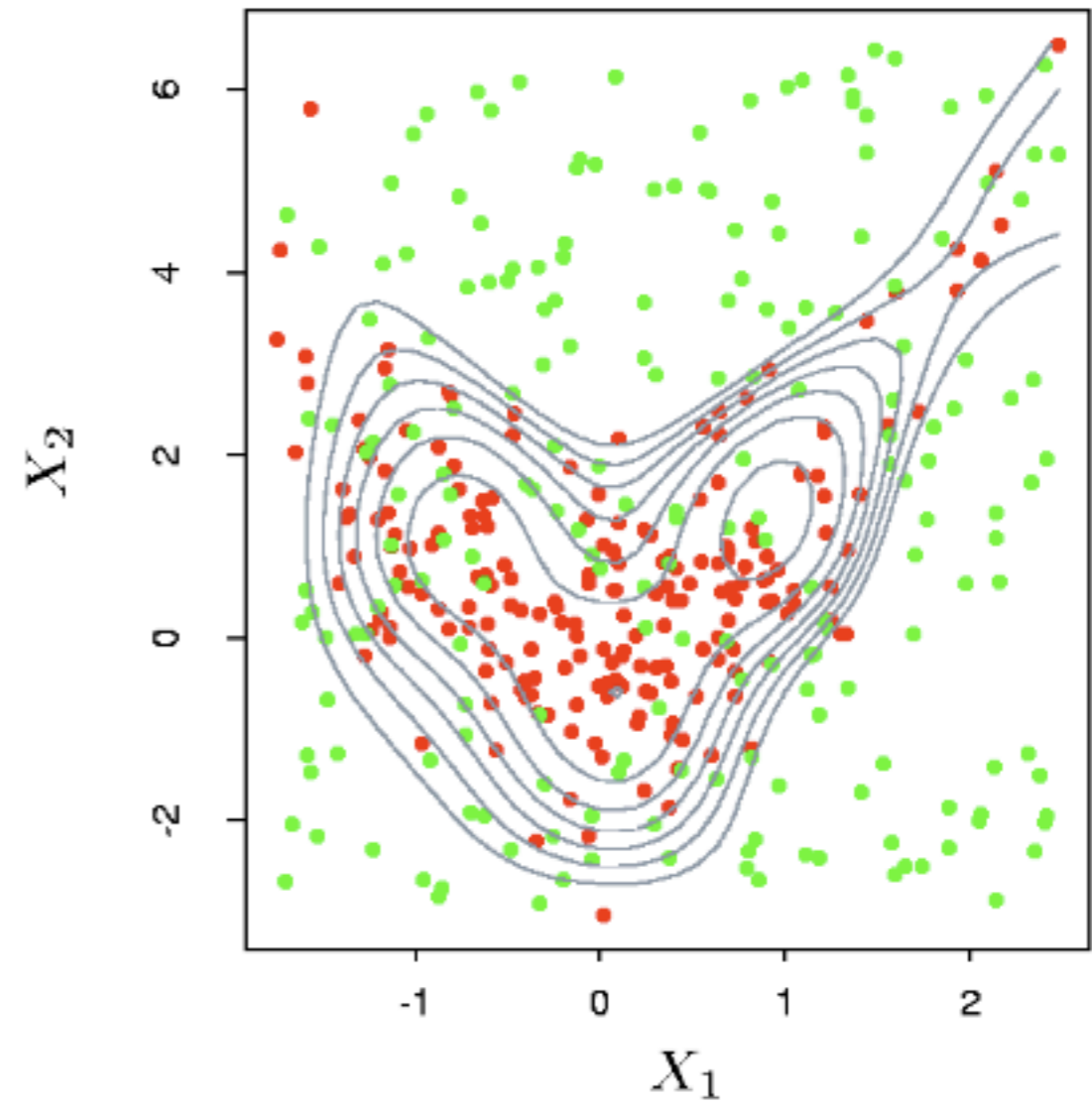
$$f(x) = \log \frac{g(x)}{g_0(x)}$$

$$\hat{g}(x) = g_0(x) e^{\hat{f}(x)}.$$

# Figure 14.3



Training classified red



Reference uniform green

# Generalized Association Rules

- $g(x)$  can be used to find data density regions
- Eliminate Apriori problem of locating low support but highly associated items

$$R = \bigcap_{j \in \mathcal{J}} (X_j \in s_j)$$

$$T(R) = \int_{x \in R} g(x) dx$$

# We have methods

- Convert unsupervised space to regions of high density
- CART
  - Decision tree terminal nodes are regions
- PRIM
  - Find the bump maximizing average value

# Example

- Married, own home, not apartment = 24%
- <24yo, single, not homemaker or retired, rent or live with family = 24%
- Own home, not apartment  $\Rightarrow$  married
  - $C = 95.9\%$ ,  $L = 2.61$
- Apriori can't do  $X \neq$  value

# Cluster Analysis

- Segment data
- Subsets are closely related
- Find natural hierarchy
- Form descriptive statistics

# Measuring Similarity

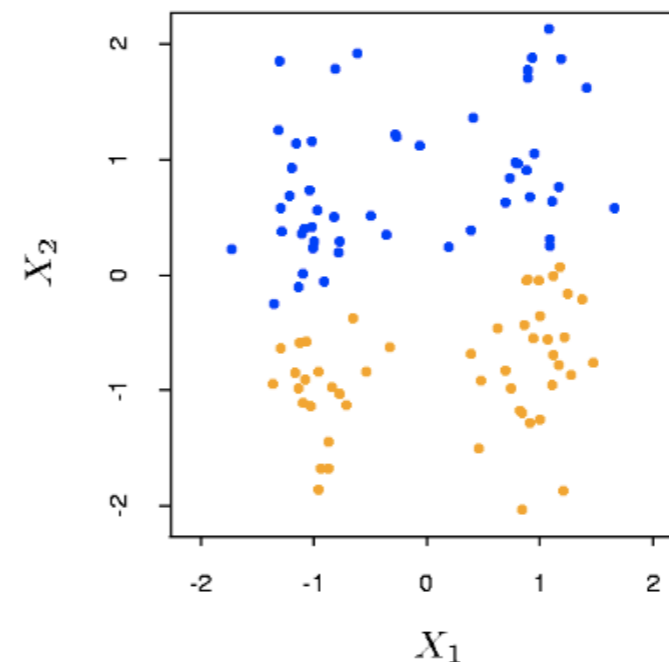
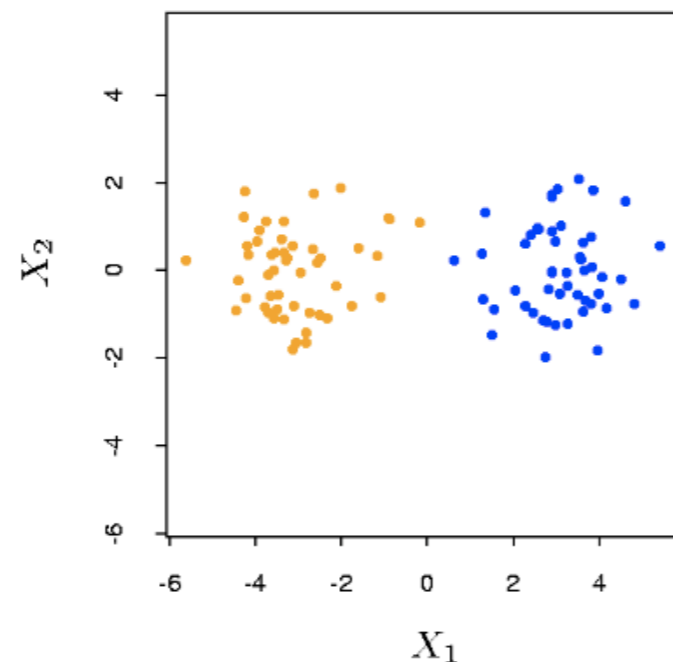
- Proximity matrices
  - $N \times N$  matrix  $\mathbf{D}$  where  $d_{ii'} = \text{proximity}$
  - Diagonal is 0, values positive, usually symmetric
- Dissimilarities based on attributes
  - $j = 1 \dots p$
  - $D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_{ij}, x_{i'j}) \quad d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2.$

# Measuring Dissimilarity

- Object dissimilarity

- $$D(x_i, x_{i'}) = \sum_{j=1}^p w_j \cdot d_j(x_{ij}, x_{i'j}); \quad \sum_{j=1}^p w_j = 1.$$

- Weights can be adjusted to highlight variables with greater dissimilarity



$$w = 1/[2(\text{var}(X_j))]$$

# Clustering Algorithms

- Combinatorial algorithms
- Mixture modeling
  - Kernel density estimation, ex: section 6.8
- Mode seekers
  - PRIM

# Combinatorial Algorithms

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})$$

$$T = \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N d_{ii'} = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \left( \sum_{C(i')=k} d_{ii'} + \sum_{C(i') \neq k} d_{ii'} \right)$$

$$B(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i') \neq k} d_{ii'}$$

$$T = W(C) + B(C)$$

Minimize

Maximize

# Clustering Algorithms

- K-means
- Vector Quantization
- K-medoids
- Hierarchical Clustering
  - Agglomerative
  - Divisive

# K-means Clustering

1. For a given cluster assignment  $C$ , the total cluster variance (14.33) is minimized with respect to  $\{m_1, \dots, m_K\}$  yielding the means of the currently assigned clusters (14.32).
2. Given a current set of means  $\{m_1, \dots, m_K\}$ , (14.33) is minimized by assigning each observation to the closest (current) cluster mean. That is,

$$C(i) = \operatorname{argmin}_{1 \leq k \leq K} \|x_i - m_k\|^2. \quad (14.34)$$

3. Steps 1 and 2 are iterated until the assignments do not change.

# Vector Quantization



**FIGURE 14.9.** *Sir Ronald A. Fisher (1890-1962) was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a  $1024 \times 1024$  grayscale image at 8 bits per pixel. The center image is the result of  $2 \times 2$  block VQ, using 200 code vectors, with a compression rate of 1.9 bits/pixel. The right image uses only four code vectors, with a compression rate of 0.50 bits/pixel*

# K-medoids Clustering

1. For a given cluster assignment  $C$  find the observation in the cluster minimizing total distance to other points in that cluster:

$$i_k^* = \operatorname{argmin}_{\{i: C(i)=k\}} \sum_{C(i')=k} D(x_i, x_{i'}). \quad (14.35)$$

Then  $m_k = x_{i_k^*}$ ,  $k = 1, 2, \dots, K$  are the current estimates of the cluster centers.

2. Given a current set of cluster centers  $\{m_1, \dots, m_K\}$ , minimize the total error by assigning each observation to the closest (current) cluster center:

$$C(i) = \operatorname{argmin}_{1 \leq k \leq K} D(x_i, m_k). \quad (14.36)$$

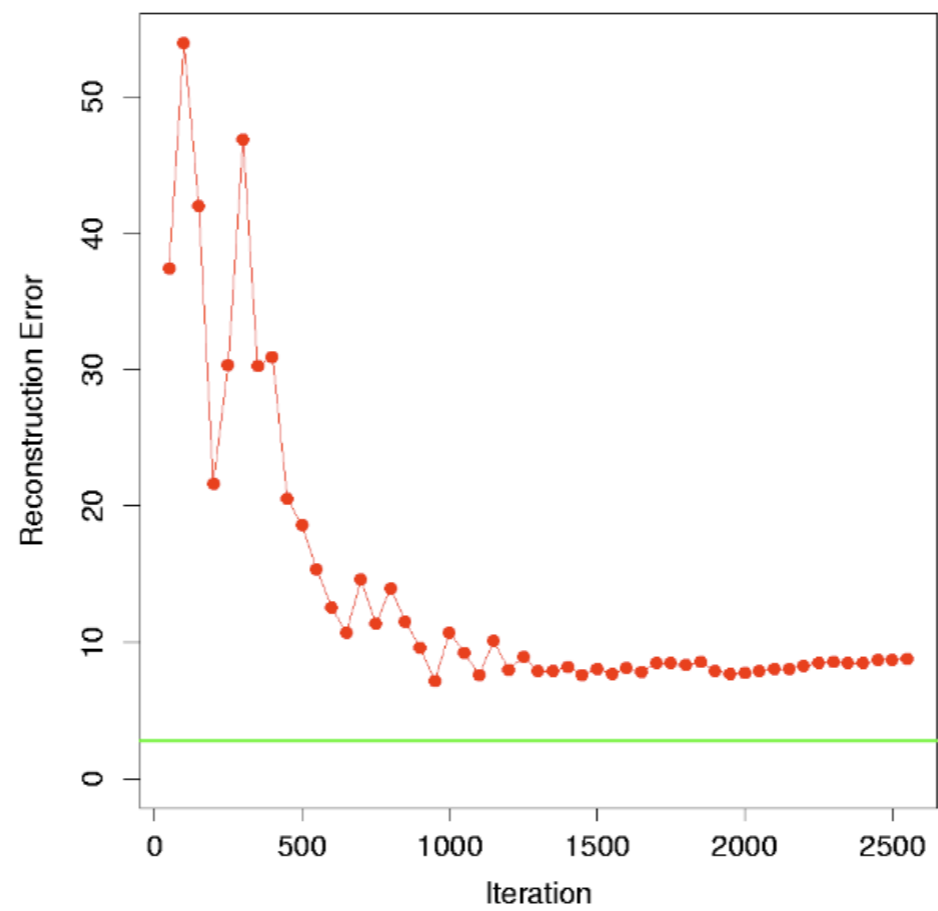
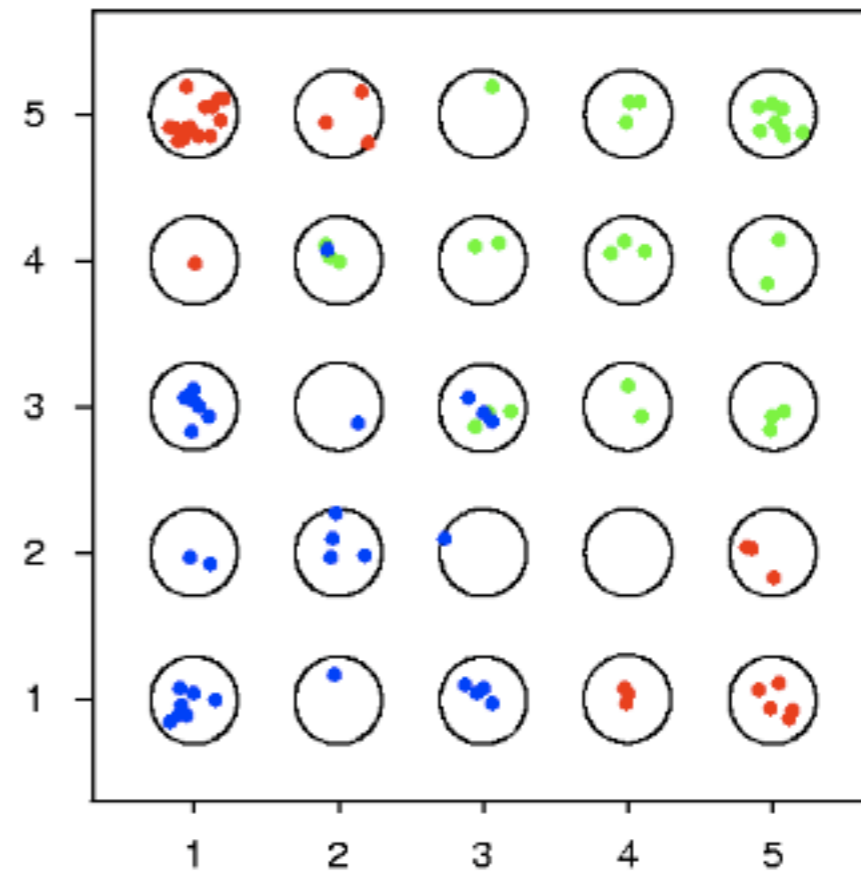
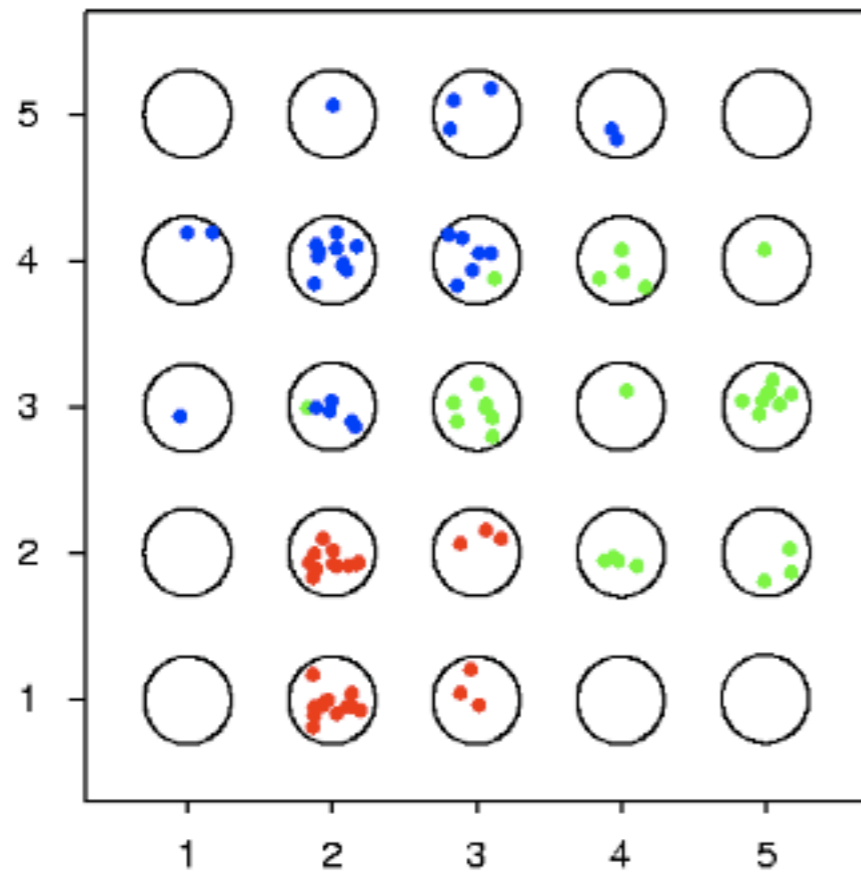
3. Iterate steps 1 and 2 until the assignments do not change.

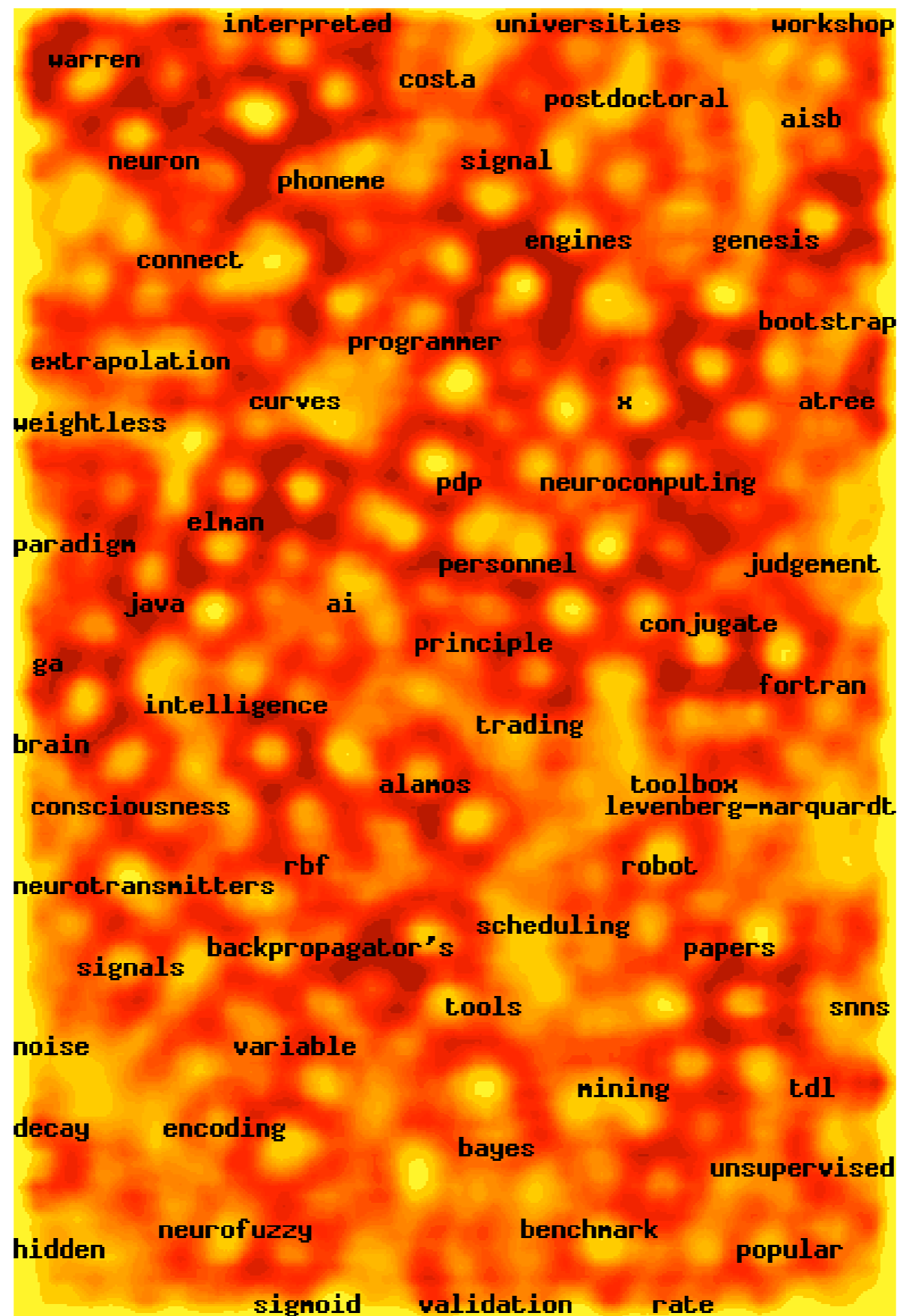


# Self-Organizing Maps

- Fit  $K$  vertices of grid to data
  - Grid: rectangular, hexagonal, ...
- Constrained K-means versus principal curves
- Updated by minimizing  $m_k$  Euclidean distance
- Parameters  $r$  and  $\alpha$ :
  - Decline from 1 to 0 over 1000 iterations

$$m_k \leftarrow m_k + \alpha(x_i - m_k)$$





<http://websom.hut.fi/websom/comp.ai.neural-nets-new/html/root.html>

# Projective Methods

- Principal Component Analysis
- Principal Curve/Surface Analysis
- Independent Component Analysis

# Principal Components

$$f(\lambda) = \mu + \mathbf{V}_q \lambda,$$

$$\min_{\mu, \{\lambda_i\}, \mathbf{V}_q} \sum_{i=1}^N \|x_i - \mu - \mathbf{V}_q \lambda_i\|^2.$$

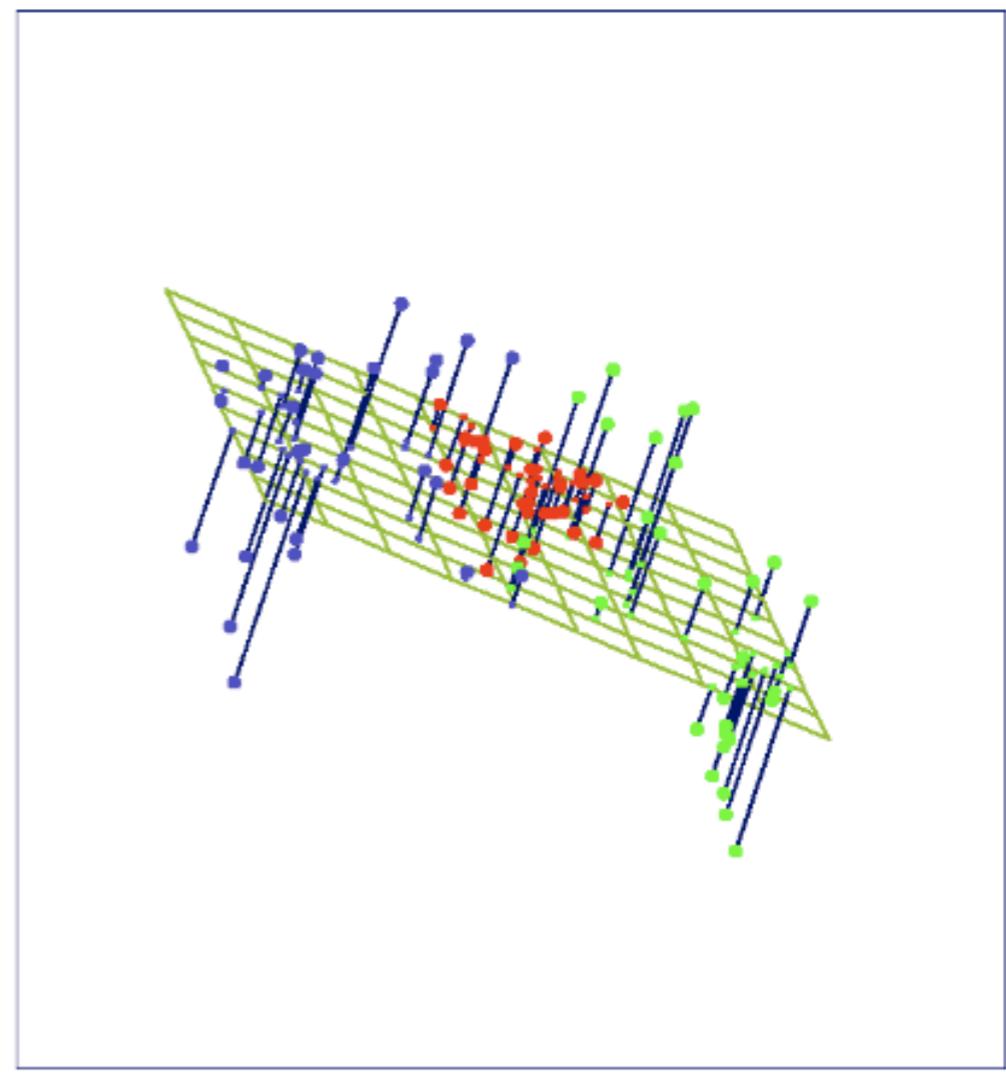
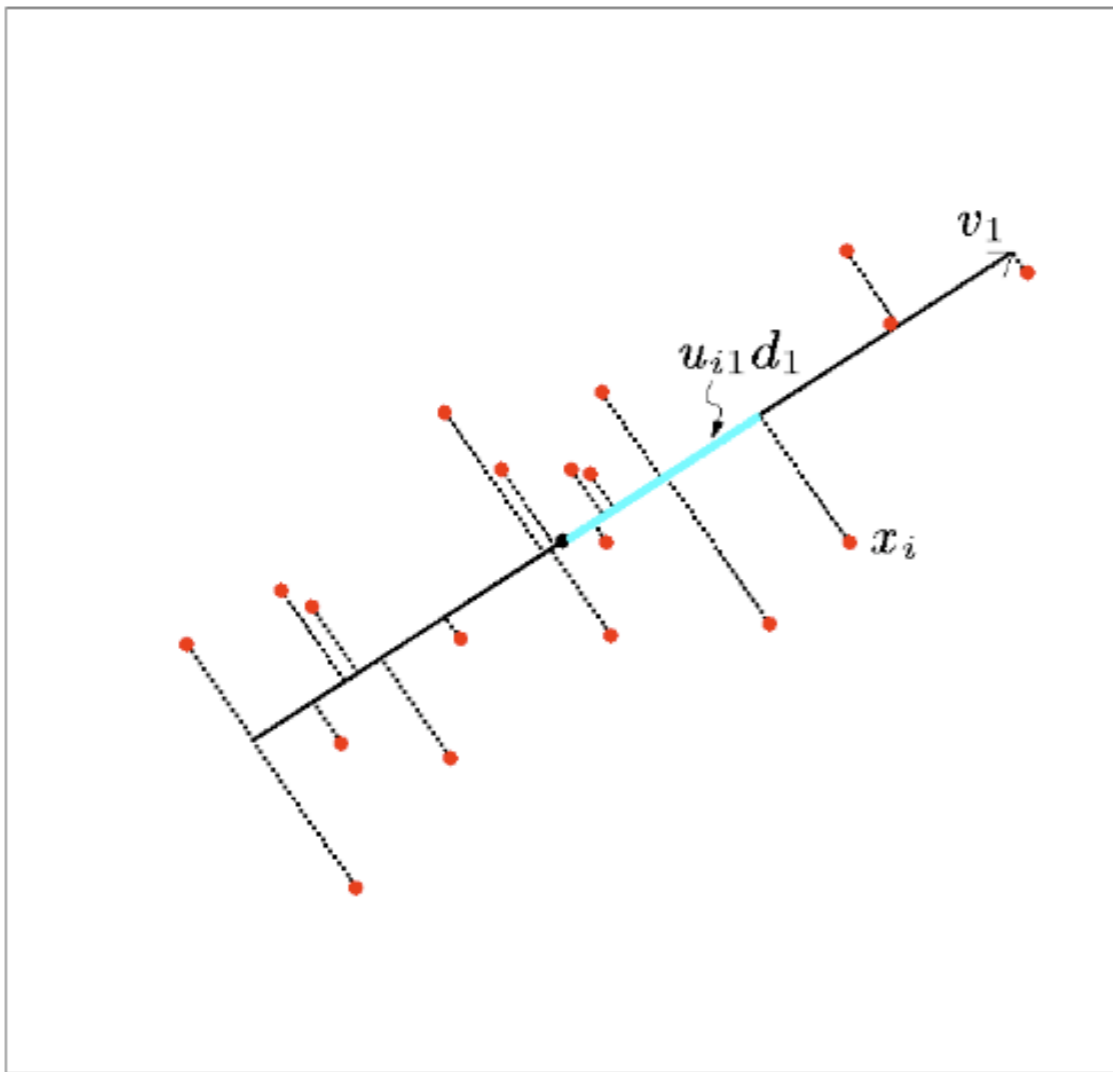
$$\begin{aligned}\hat{\mu} &= \bar{x}, \\ \hat{\lambda}_i &= \mathbf{V}_q^T (x_i - \bar{x}).\end{aligned}$$

$$\min_{\mathbf{V}_q} \sum_{i=1}^N \|(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})\|^2.$$

# Principal Components

- Singular value decomposition:
  - $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T$
  - $\mathbf{U}$ : left singular vectors,  $N \times p$  orthogonal
  - $\mathbf{V}$ : right singular vectors,  $p \times p$  orthogonal
  - $\mathbf{D}$ : singular values,  $p \times p$  diagonal

# Principal Components

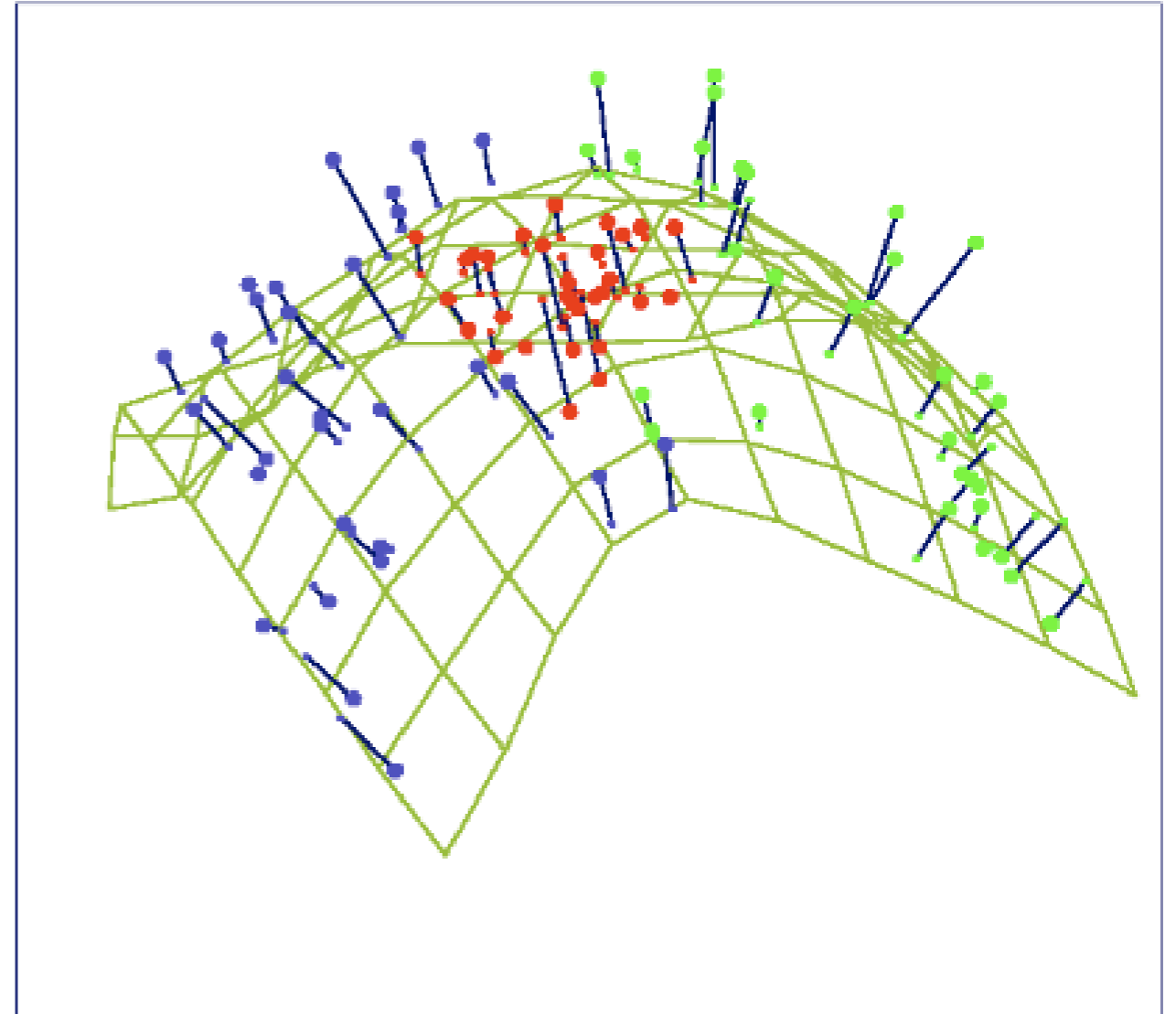
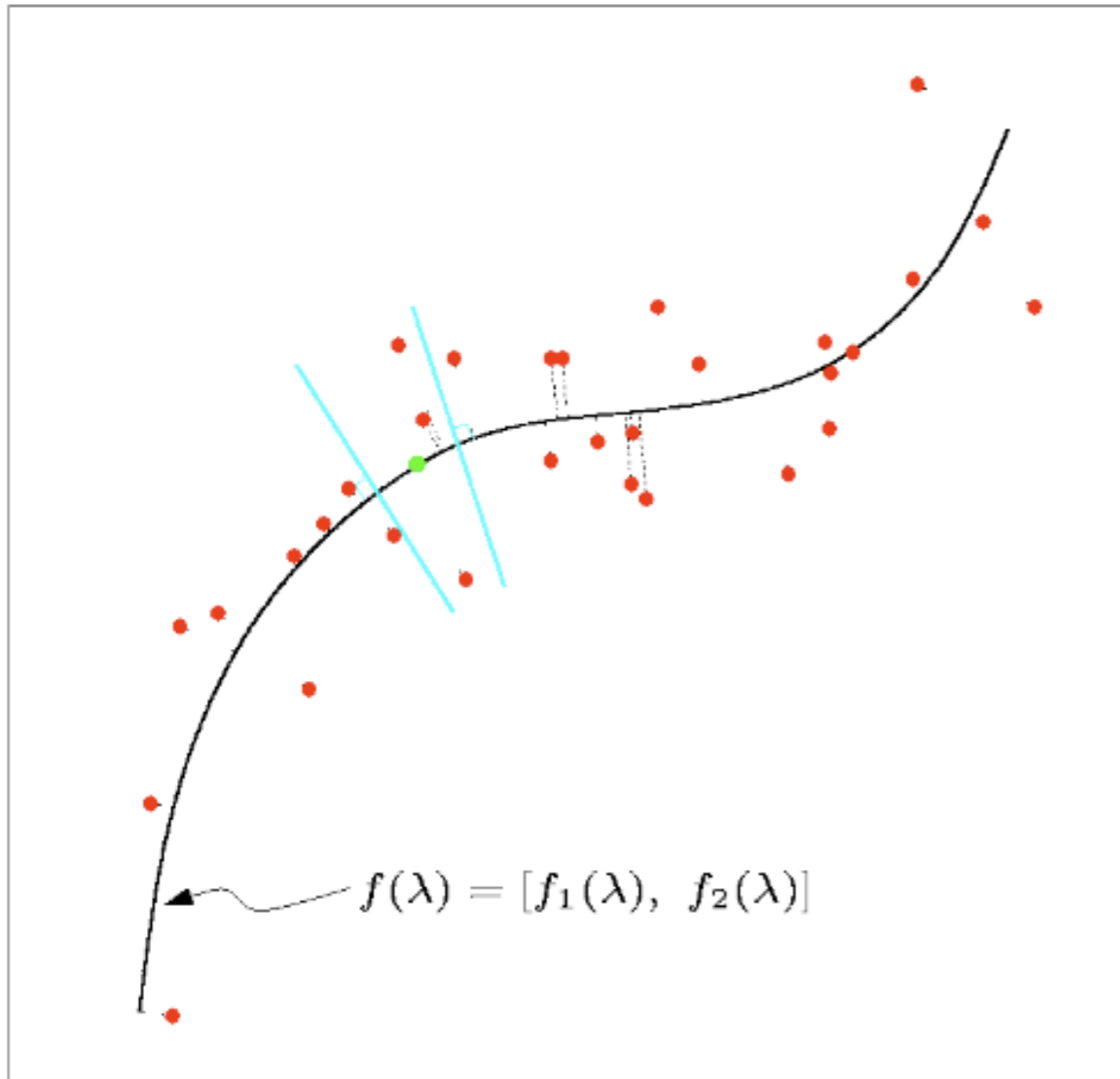


# Principal Curve

To find a principal curve  $f(\lambda)$  of a distribution, we consider its coordinate functions  $f(\lambda) = [f_1(\lambda), f_2(\lambda), \dots, f_p(\lambda)]$  and let  $X = (X_1, X_2, \dots, X_p)$ . Consider the following alternating steps:

$$\begin{aligned} \text{(a)} \quad \hat{f}_j(\lambda) &\leftarrow \text{E}(X_j | \lambda(X) = \lambda); \quad j = 1, 2, \dots, p, \\ \text{(b)} \quad \hat{\lambda}_f(x) &\leftarrow \text{argmin}_{\lambda'} \|x - \hat{f}(\lambda')\|^2. \end{aligned} \tag{14.56}$$

# Principal Curves



# Versus SOM

- Principal curves and surfaces share similarities to self-organizing maps
- As SOM prototypes increase, closer match to principal curves
- Principal curves provide smooth parameterization versus discrete

# Independent Components

- Goal is source separation
- Example in audio removing noise
- Find statistically independent signals where distribution not normal with constant variance

# ICA

$$H(Y) = - \int g(y) \log g(y) dy.$$

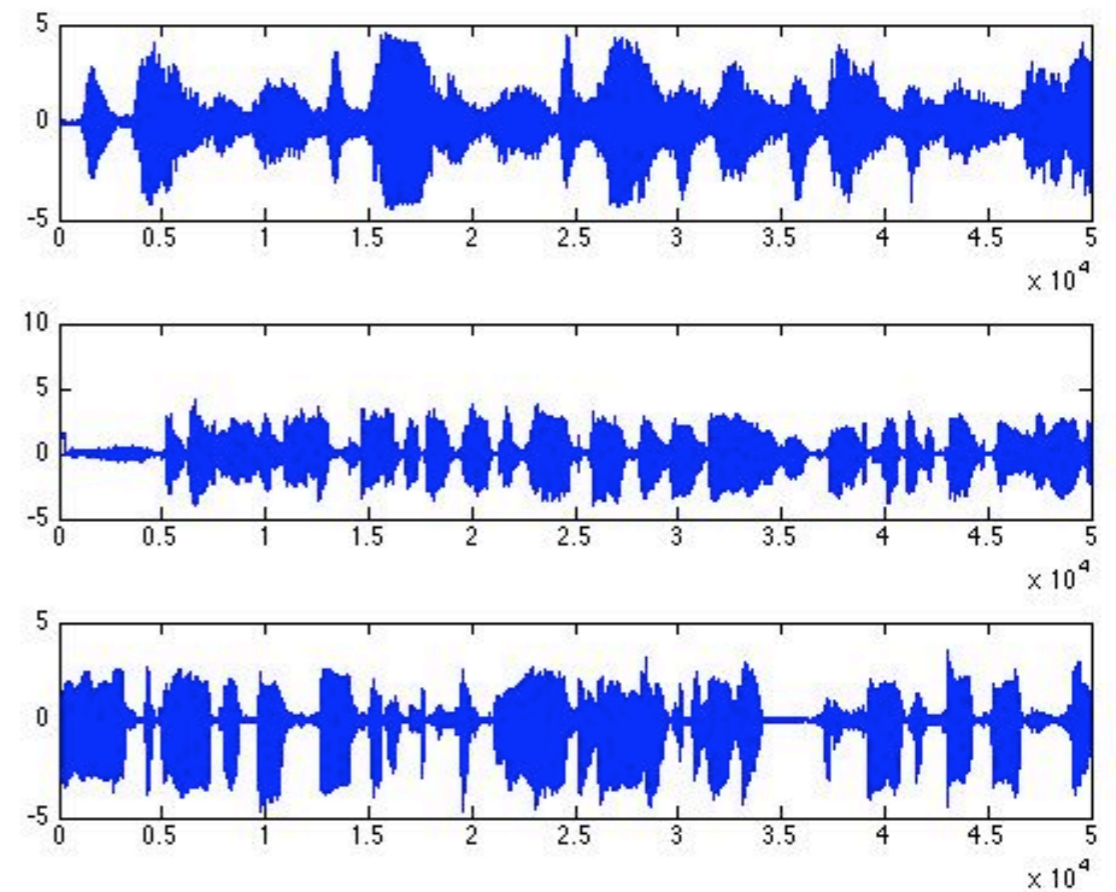
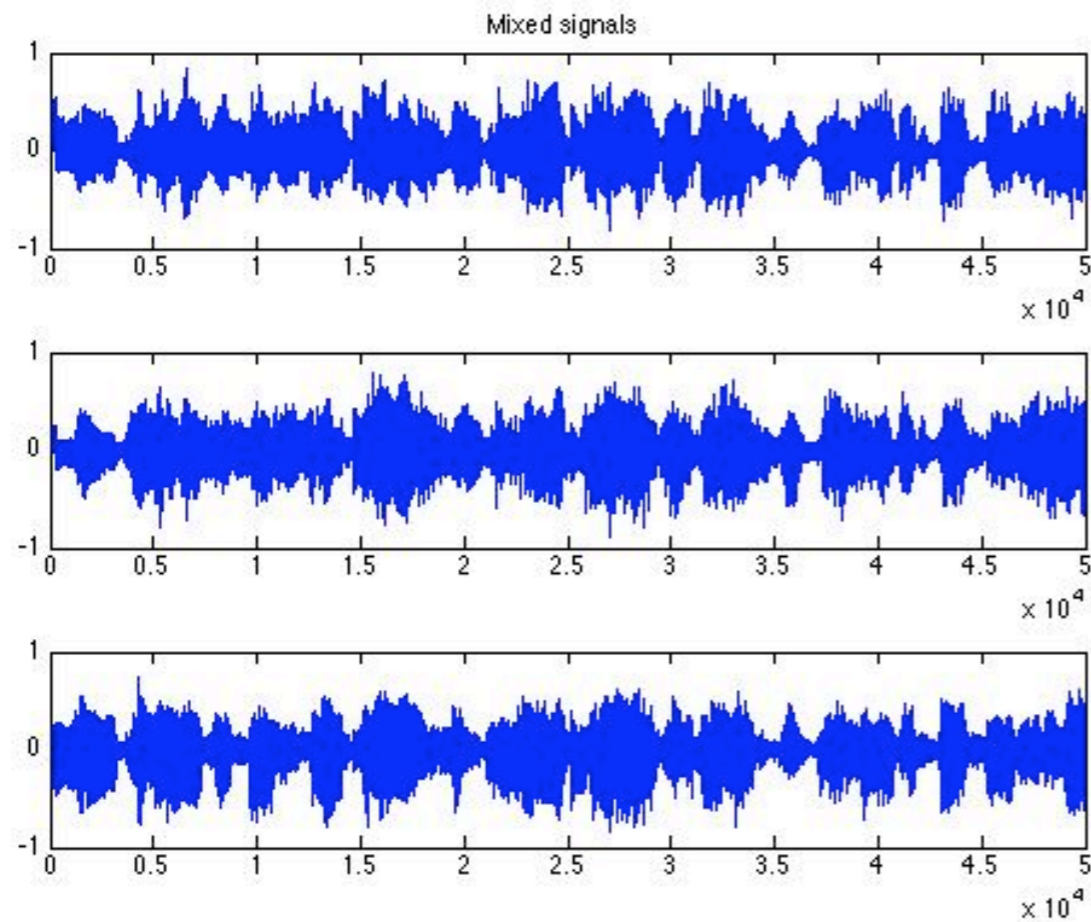
$$Y = \mathbf{A}^T X$$

$$I(Y) = \sum_{j=1}^p H(Y_j) - H(Y)$$

$$I(Y) = \sum_{j=1}^p H(Y_j) - H(X) - \log |\det \mathbf{A}|$$

$$= \sum_{j=1}^p H(Y_j) - H(X).$$

# ICA Example



# Multidimensional Scaling

- Given  $d$  as distance or dissimilarity measure
- Minimize stress function:

- Least squares:  $S_D(z_1, z_2, \dots, z_N) = \left[ \sum_{i \neq i'} (d_{ii'} - \|z_i - z_i'\|)^2 \right]^{1/2}$

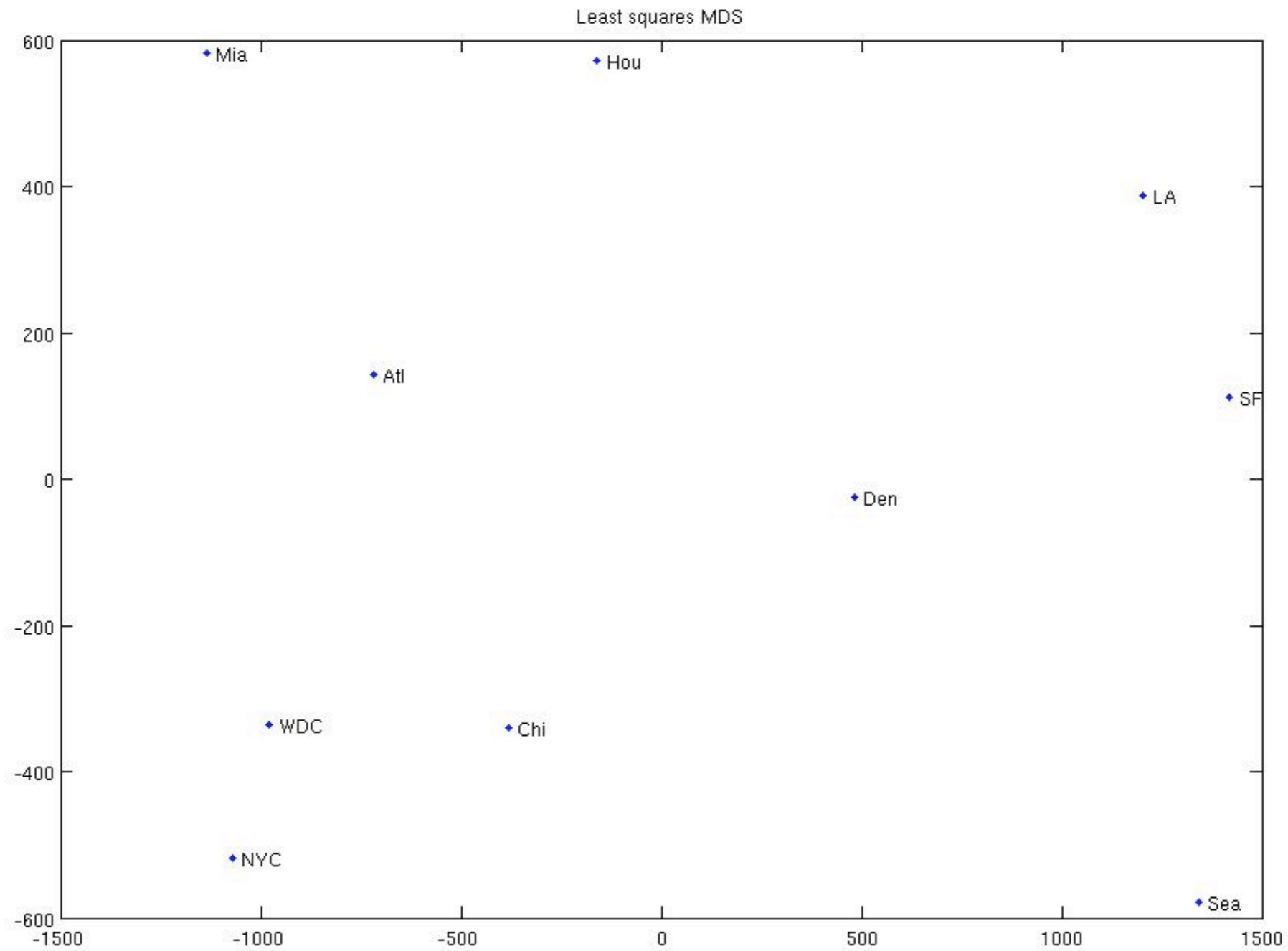
- Sammon mapping:  $\sum_{i \neq i'} \frac{(d_{ii'} - \|z_i - z_i'\|)^2}{d_{ii'}}$

- Classical scaling:  $\sum_{i \neq i'} (s_{ii'} - \langle z_i - \bar{z}_i, z_i' - \bar{z}_i' \rangle)^2$

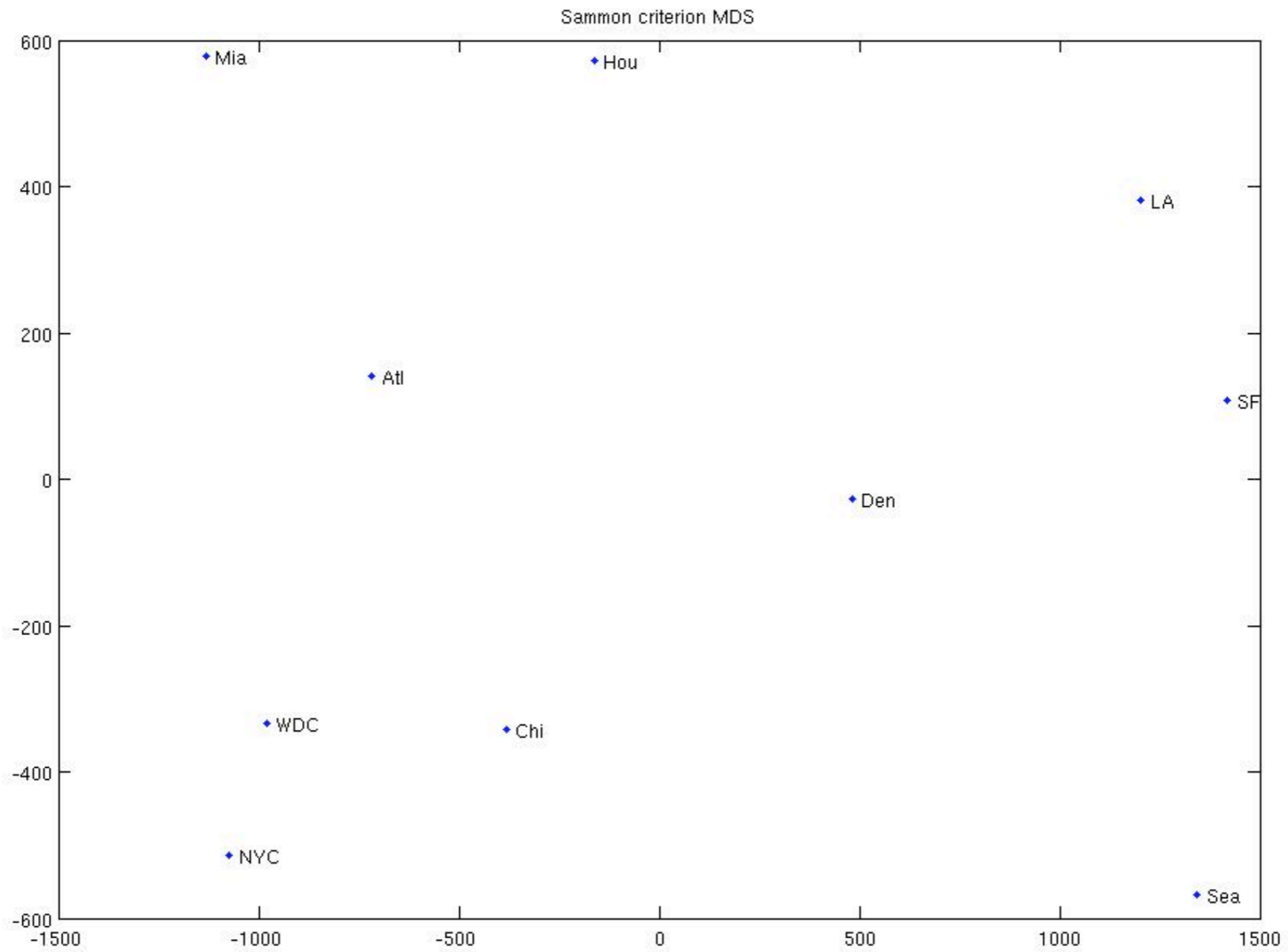
# U.S. Cities Example

	Atl	Chi	Den	Hou	LA	Mia	NYC	SF	Sea	WDC
Atl	0	587	1212	701	1936	604	748	2139	2182	543
Chi	587	0	920	940	1745	1188	713	1858	1737	597
Den	1212	920	0	879	831	1736	1631	949	1021	1494
Hou	701	940	879	0	1374	968	1420	1645	1891	1220
LA	1936	1745	831	1374	0	2339	2451	347	959	2300
Mia	604	1188	1726	968	2339	0	1092	2594	2734	923
NYC	748	713	1631	1420	2451	1092	0	2571	2408	205
SF	2139	1858	949	1645	347	2594	2571	0	678	2442
Sea	2182	1737	1021	1891	959	2734	2408	678	0	2329
WDC	543	597	1494	1220	2300	923	205	2442	2329	0

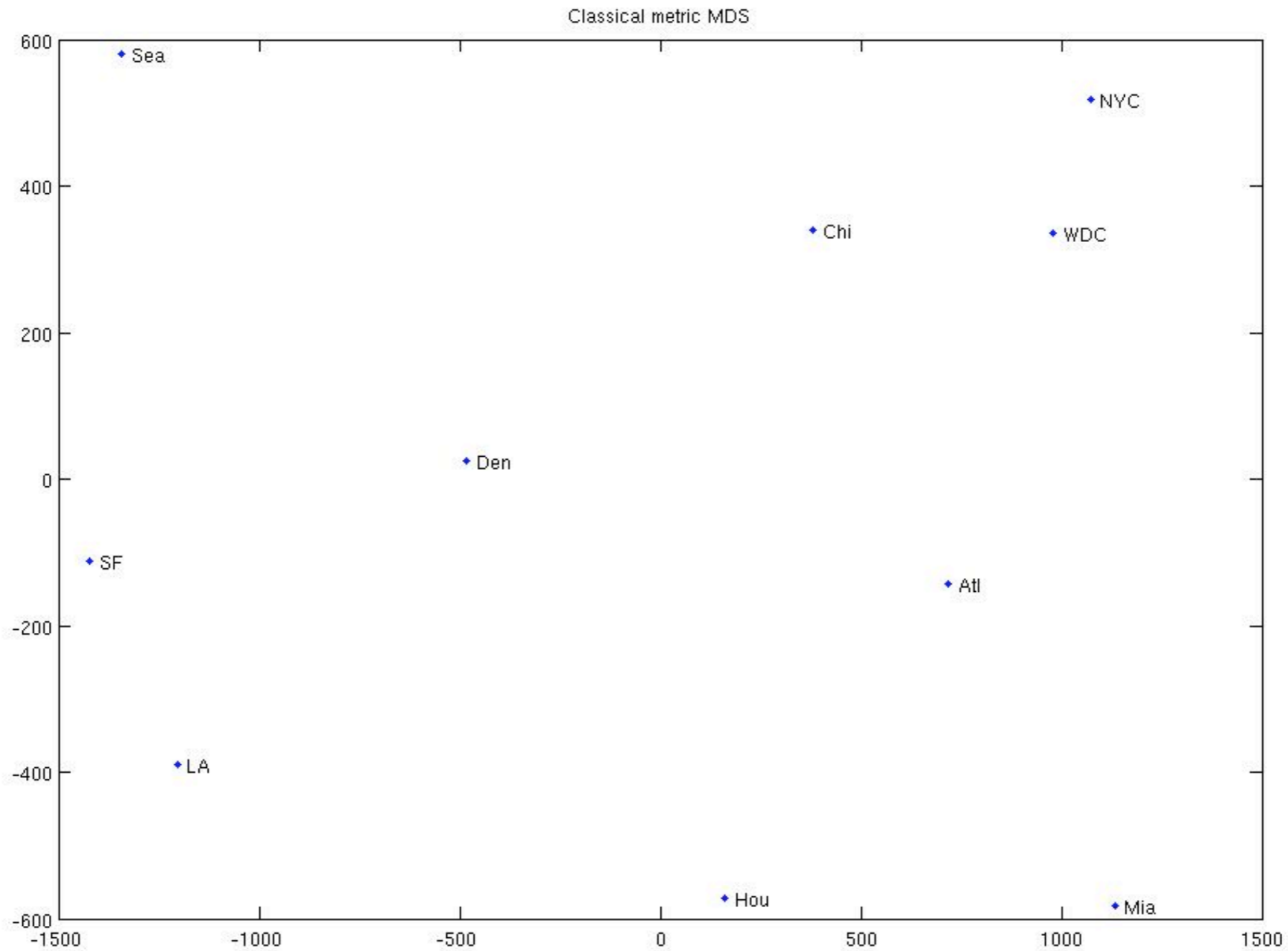
# Least Squares MDS



# Sammon MDS



# Classic MDS



# Conclusions

- Reframe our set of  $X$
- Techniques:
  - Association Rules
  - Cluster Analysis
  - Self-Organizing Maps
  - Projective Methods
  - Manifold Modeling

# References

- Burges CJC. Geometric Methods for Feature Extraction and Dimensional Reduction:A Guided Tour. Data Mining and Knowledge Discovery Handbook:A Complete Guide for Practitioners and Researchers. Eds Rokach L, Maimon O. Kluwer Academic Publishers, 2004.
- Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2001.

**Thank you**

**email:**

**Len.Tanaka@uth.tmc.edu**