Data Pre-Processing

Cleaning, integration, exploration, reduction/transformation, visualization….

- **Readings:**
  - **KJ Ch 3, 19**

Also see:

- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- Garcia, Luengo, Herrera, "Data Preprocessing in Data Mining", Springer 2015.

- **Explore** segmentationOriginal (KJ, 3.1) and German Credit Card datasets

Types of Data

- A data set is a collection of data objects/records
  - Each object is described by several features/attributes
- Data Types
  - Nominal
    - eye color, hobby
    - Binary: special case
  - Ordinal
    - rankings (e.g., taste of potato chips on a scale from 1-10), grades
  - Interval
    - speed, temperature in Celsius

- Categorical: Nominal or Ordinal
- Others: ID, DATE, text/strings, graphs,…

Different data types often need different ways of handling/modeling.

e.g. Proportional odds model for ordinal regression.

Why Preprocess Data

- **GIGO!**
  - data may be incomplete, inconsistent, noisy; have outliers, or simply too large
- **Why is data dirty?**
  - Incomplete data may come from
    - Not available or "Not applicable" data value when collected
    - Thoughtless entry (e.g. 0 vs. missing)
  - Noisy data (incorrect values) may come from
    - Faulty data collection instruments
    - Human or computer error at data entry
      - HEB, shoulder surgery, ..
    - Out-of-date
  - Inconsistent data may come from
    - Different data sources; formats
    - Inconsistent rules e.g. hotel price on phone vs. internet
  - Duplicate records need to be eliminated

Major Preprocessing Steps

1) Data cleaning; sanity checks, consistency (already done by ETL tools if data is from warehouse)
2) Exploratory Data Analysis (Often based on a sample)
    1) Fill missing values, remove noise and outliers
    2) transformation/scaling
3) Data reduction
    1) Of records (sampling)
    2) Of attributes (feature selection/extraction)
4) Visualization

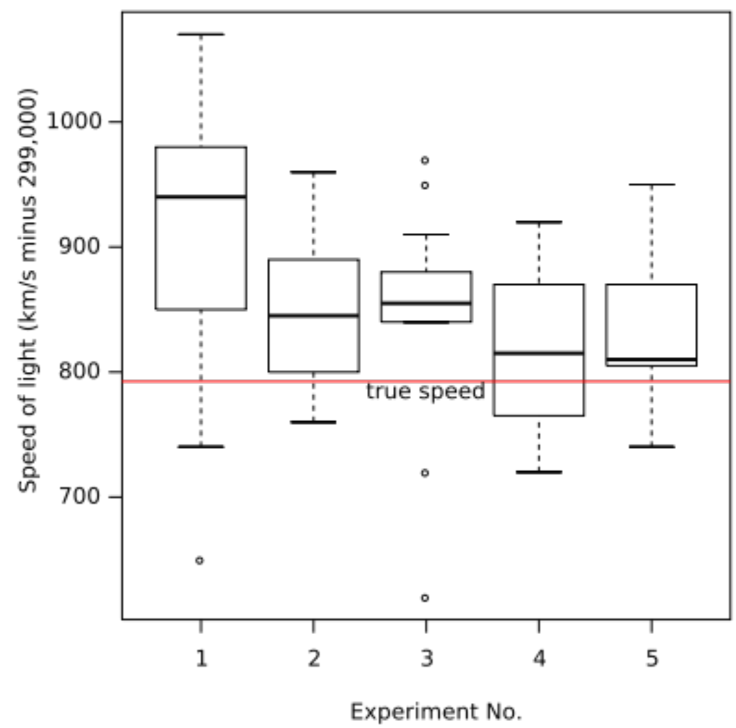
- Often takes over 90% of a project's time!
- steps 2-4 often revisited after modeling.

Before You Clean the Data..

- .. Do a quick summarization/visualization
  - Single "input" variable summaries
    - Variable type, mean, range, %missing, skewness, histograms, boxplots, ….
  - Bivariate (X_i vs. Y or X_i vs. X_k) visuals
    - (scatter plots, correlation,..)

Boxplot

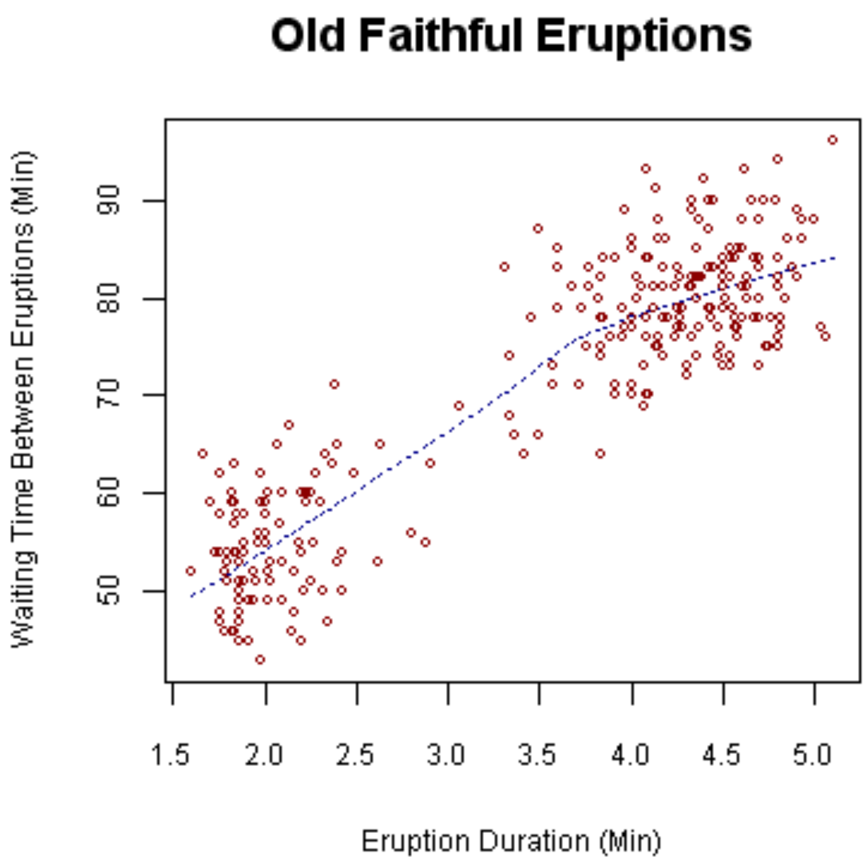
- Shows median, 1st and 3rd quartile (Q), non-outlier extreme points; outliers
- Outliers, <1.5 IQR below 1st Q or >1.5IQR above 3rd Q.



See Wikipedia

Scatterplot

- Old Faithful Example from Wikipedia

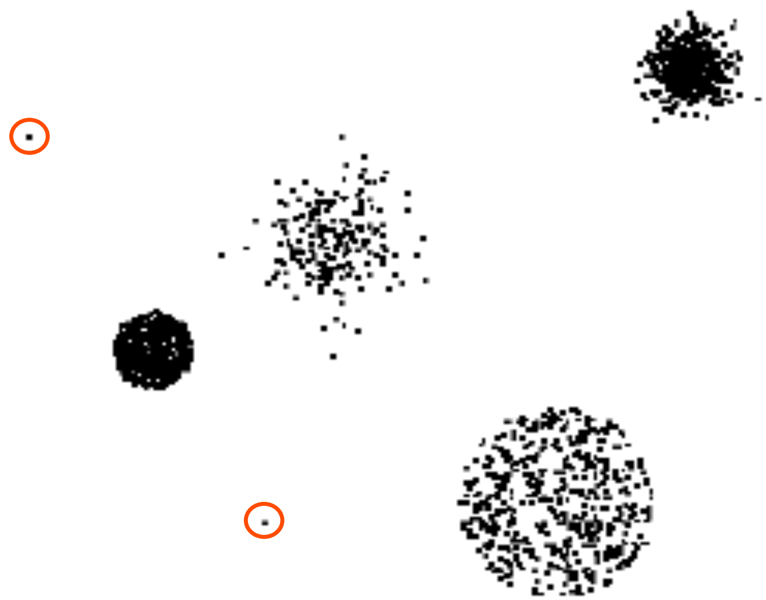


Data Cleaning I

- Dealing with Missing Values (Imputation)
  - Missing Completely at Random (MCAR)?
    - Vs. "informative missingness" (e.g doctor's choices)
  - ignore record or attribute (often missing values are concentrated in a few instances or attributes)
  - Fill in missing values
    - fill with constant, mean or mode
    - conditional mean/ mode
      - Condition on values of a set of related variable
    - Use K-NN
  - "mathematically optimal" way?

Cleaning II: Handling Outliers

- Outliers are data objects with characteristics that are considerably different than the vast majority of the other data objects in the data set



Dealing with Outliers in “X”

- Probability based (old):
  - Estimate pdf of X, using e.g. Parzen windows or mixture of Gaussians
  - Identify low p(x) points

- Discrimination based
  - Rule based, e.g.
    - » less than 1% for categorical variables
    - » Outside 3 sigma for gaussian looking numeric variables
  - Distance based: see if outlier score is > threshold or not
    - » Score could be av. Distance of k-nearest neighbors; distance to the kth neighbor, etc.

Outliers in Y (robust statistics)

Identify outliers and eliminate before applying model

OR

Use models that are little affected by presence of a few outliers

- trimmed means instead of means

- alternatives to "squared error" loss functions
  - e.g. Huber's loss (quad → linear)

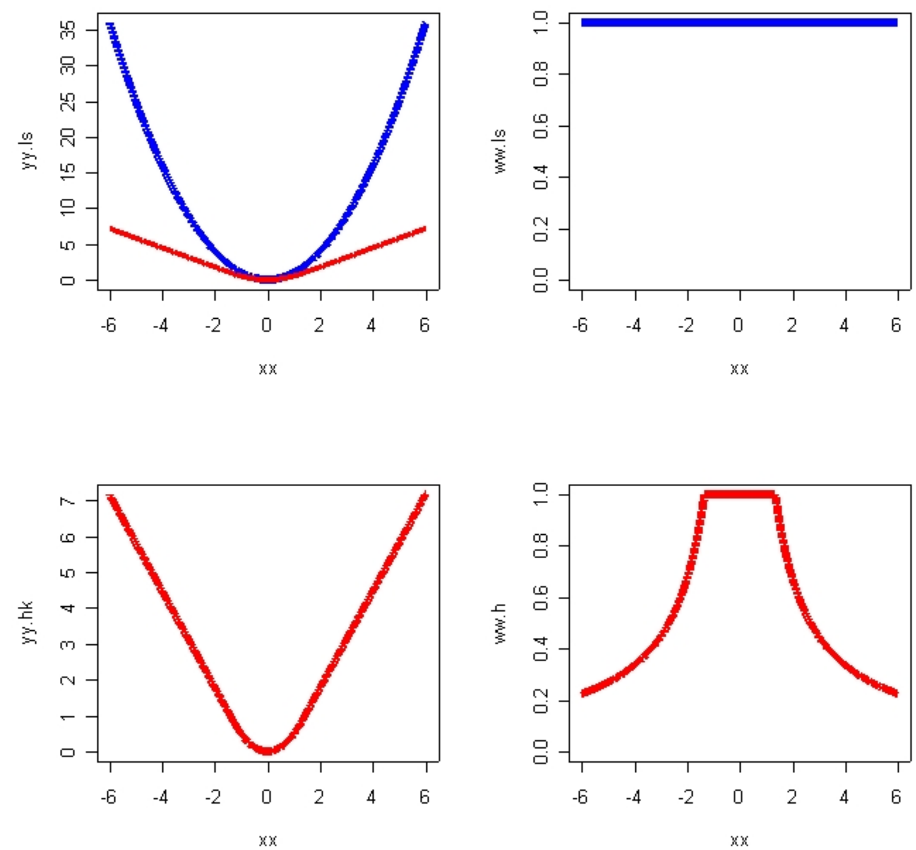


Fig: Plotted as a function of residual ($r = y - \hat{y}$):

Blue: Sq. error loss (left) and "equivalent" weights (right)

Red: Huber loss and equivalent weights if Sq. loss was used (right)

Data Transformation

- Scaling

  Normalization by Linear scaling

  - Linear [min, max] → [0,1]
  - Centering (e.g. Z-scoring: Normal/Gaussian → N (0, 1))

- (non-linear) transformation, e.g. to reduce skew or to show a simpler relationship between x and y (for example a power law shows up as a linear relationship in the log space).
  - Log;
  - square;
  - exponential

Data Reduction Methods

- **Why?**
  - get quicker answers
  - Reducing number of features may (substantially) improve results !!
    - Reduces **"curse-of-dimensionality"**
      - When dimensionality increases, (randomly distributed) data becomes increasingly sparse in the space that it occupies
        - » Problematic for many types of analysis.
    - Collinearity a problem with MLR
      - Tools, e.g. compute all pairwise correlations ("pairs" in R)
      - Heuristics, e.g. eliminate variables till max pairwise correlation < threshold

How to Reduce Data

- Reduce # of records or instances
- Reduce # of attributes or features
- Aggregate (in data cube)
- Reduce resolution of an attribute e.g. discretization of interval variable.

- Note: Data reduction technique will affect quality as well as speed.

Sampling

A recent Texas Public Employees Association (TPEA) survey found that 11.7 percent of state employee households received public assistance in the past year. More than 16,000 state employees responded to our survey, and because our sample size was so large, our results can be considered representative of all general state government — approximately 149,000 employees — with a 99 percent confidence level and a 1 percent margin of error.

From AAS, April 24, 2015

Sampling

- Methods :
  - random with/without replacement
  - Stratified
    - Keep proportions, or
    - biased - change priors

- Nature of results
  - Uncertainty $\rightarrow$ confidence levels
    - Probably Approximately Correct
      - E.g. National surveys
  - (un) biased estimate?

Some Theory (Uniform Sampling)

- Binary outcomes, p = (true but unknown) probability of success/trial
  - Expected # successes in n trials?
    - Binomial distribution of empirically observed "p^" (estimate of p)
      - Normal for n large enough (n> 30 and np, n(1-p) > 5)
    - Then p^ is unbiased and with $\sigma^2 = p(1-p)/n$

Want: **within ε of mean** with high **probability (1- α)**

- Normal: 90% of probability within +/- 1.65 σ of mean
  - 95% of probability within +/- 1.96 σ of mean
  - 99% of probability within +/- 2.58 σ of mean
  - Margin of error is ε; critical value (for standardized curve) is denoted by $z_{\alpha/2}$
    - » If α =0.05, then $z_{\alpha/2}$ is 1.96

# of samples required depends on "epsilon" and "alpha"

- $n >= p(1-p) (z_{\alpha/2} / \varepsilon)^2$
  - independent of N!!
  - Use p^ for p in above Eqn; if p^ is unknown, use 0.5 for safe answer.

Web Resources

- Many good web resources to understanding sampling, confidence intervals, etc.

Understanding confidence intervals:
http://www.lordsutch.com/pol251/schacht-08-web.pdf

Introduction to Probability (Undergrad course-notes from MIT).
http://ocw.mit.edu/OcwWeb/Mathematics/18-05Spring-2005/LectureNotes/index.htm

Other Sampling Issues

- Very effective when applicable
- good for estimate answer to aggregate query; but not for "needle in haystack" problems
- expensive! Not well supported in databases
- natural choice for progressive refinement; hypothesis testing

Reducing # of (Derived) Attributes/Features

- Feature selection

vs

Feature extraction

- parametric e.g. Principal Component Analysis (PCA), linear regression
  - assume a model, then estimate its parameters
- Vs. Non-parametric
  - histograms (aggregate info; hence classically popular)

Why is feature selection often preferred to feature extraction?

Attribute Subset Selection

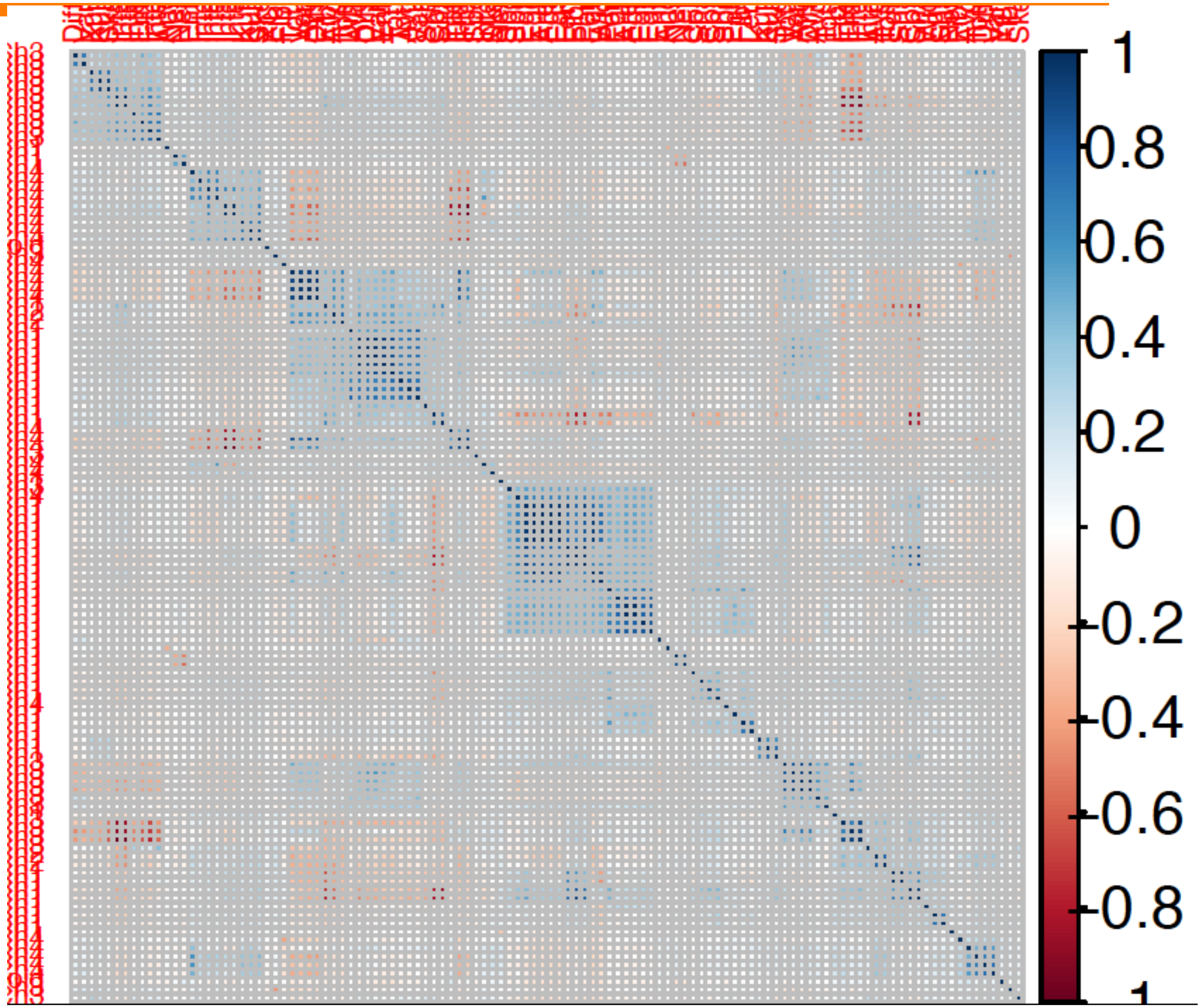
- NP-complete, so use heuristics (evaluation + search strategy)
  - Evaluation:
    - filters : use intrinsic quality measure
      e.g. correlation with other predictors (cor(data) in R); correlation/Chi-sq with target; mutual info with target…

    OR

    - Wrappers (extrinsic evaluation)
      - use model (decision tree, neural net etc)
  - Search:
    - Forward inclusion
    - Backward elimination
    - Stepwise (forward, but may remove predictors that no longer meet criterion)
    - Branch and bound
- Advanced Methods : http://featureselection.asu.edu/index.php

Feature Selection Using Corrplot package

- See KJ pg 55-56



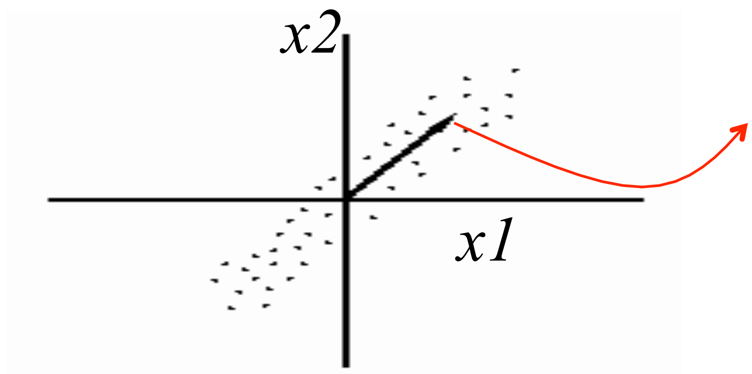
Feature Extraction Choices

- Linear
  - Unsupervised : PCA
    - 5 functions to do Principal Components Analysis in R

    http://gastonsanchez.com/blog/how-to/2012/06/17/PCA-in-R.html
  - Supervised:
    - Fisher's Linear Discriminant (classification)
    - Canonical Correlation (regression)
- Non-Linear
  - Unsupervised : Principal Curves, Sammon's Map, Kohonen's SOM
  - Supervised: Nonlinear discriminant analysis, e.g. using a multi-layered perceptron.

PCA

- Principal Components:
  - (see demos at: http://www.cs.mcgill.ca/~sqrt/dimr/dimreduction.html)
  - Reduce dimensions while retaining info about original data
    - PCA finds the best "subspace" that captures as much data variance as possible
  - optimal linear projection/reconstruction in MSE sense
  - Based on eigen-decomposition of data covariance matrix
  - Called "Latent Semantic Indexing" in Info retrieval community; KL-Transform in signal processing



1st Eigen-Vector

Example scree plot

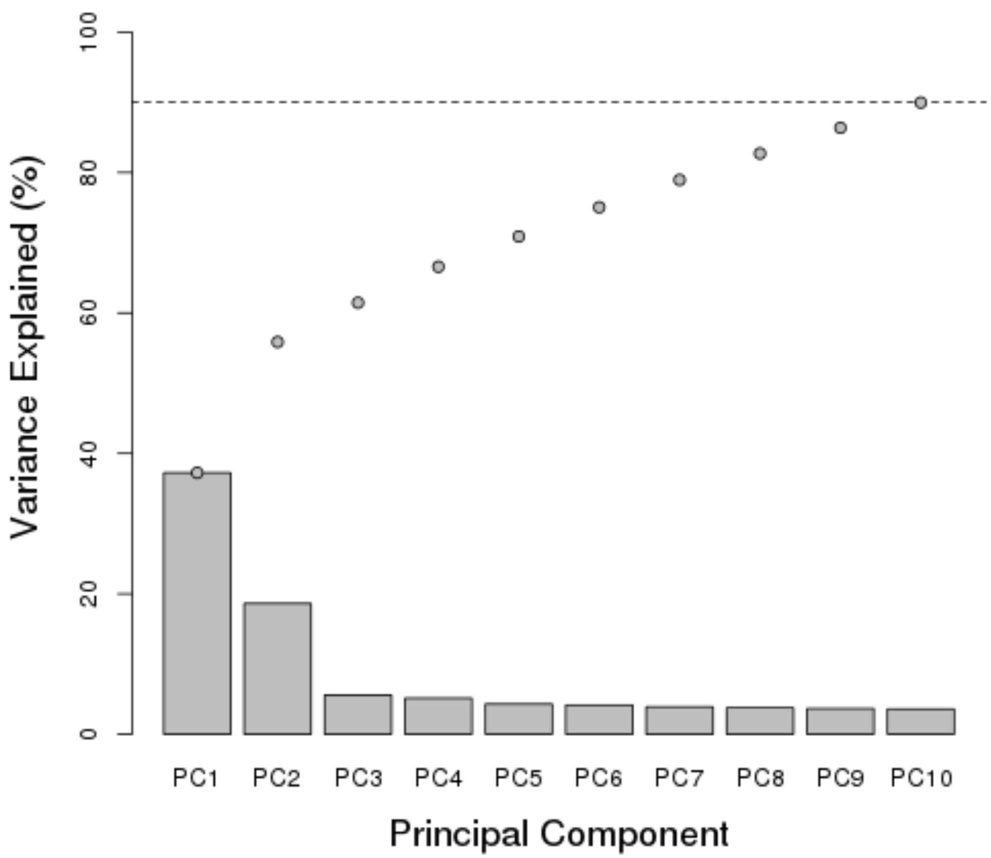
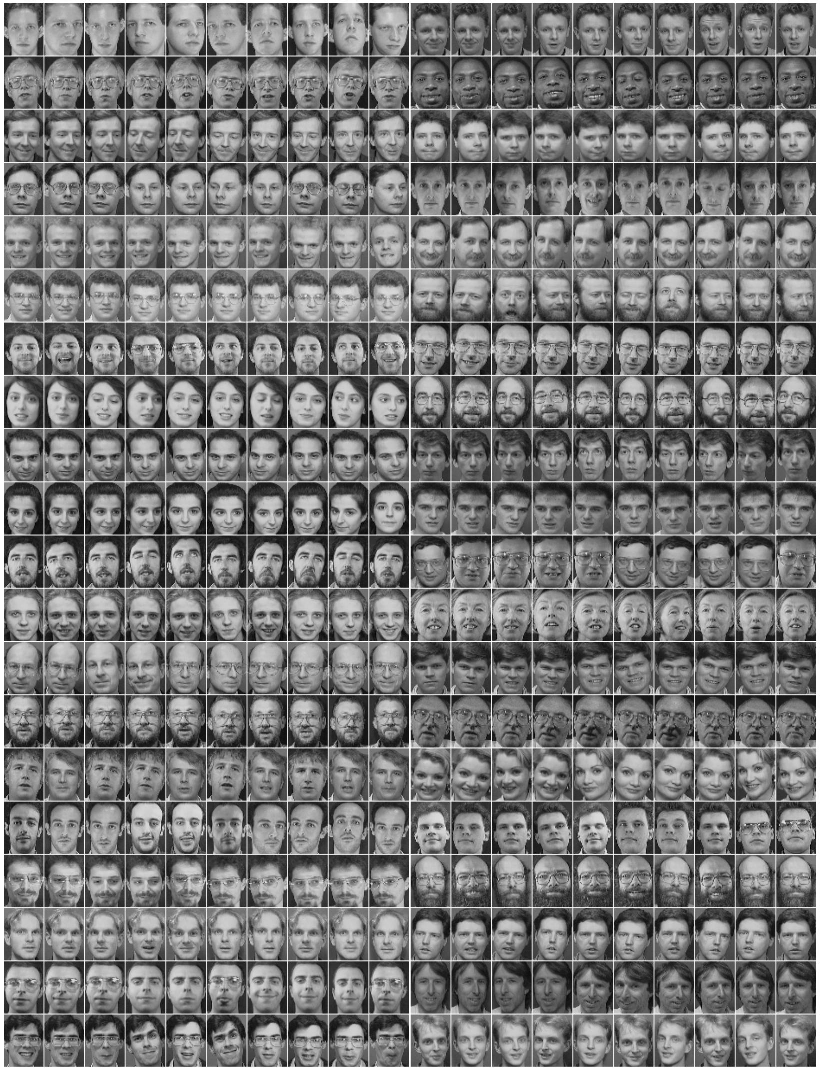
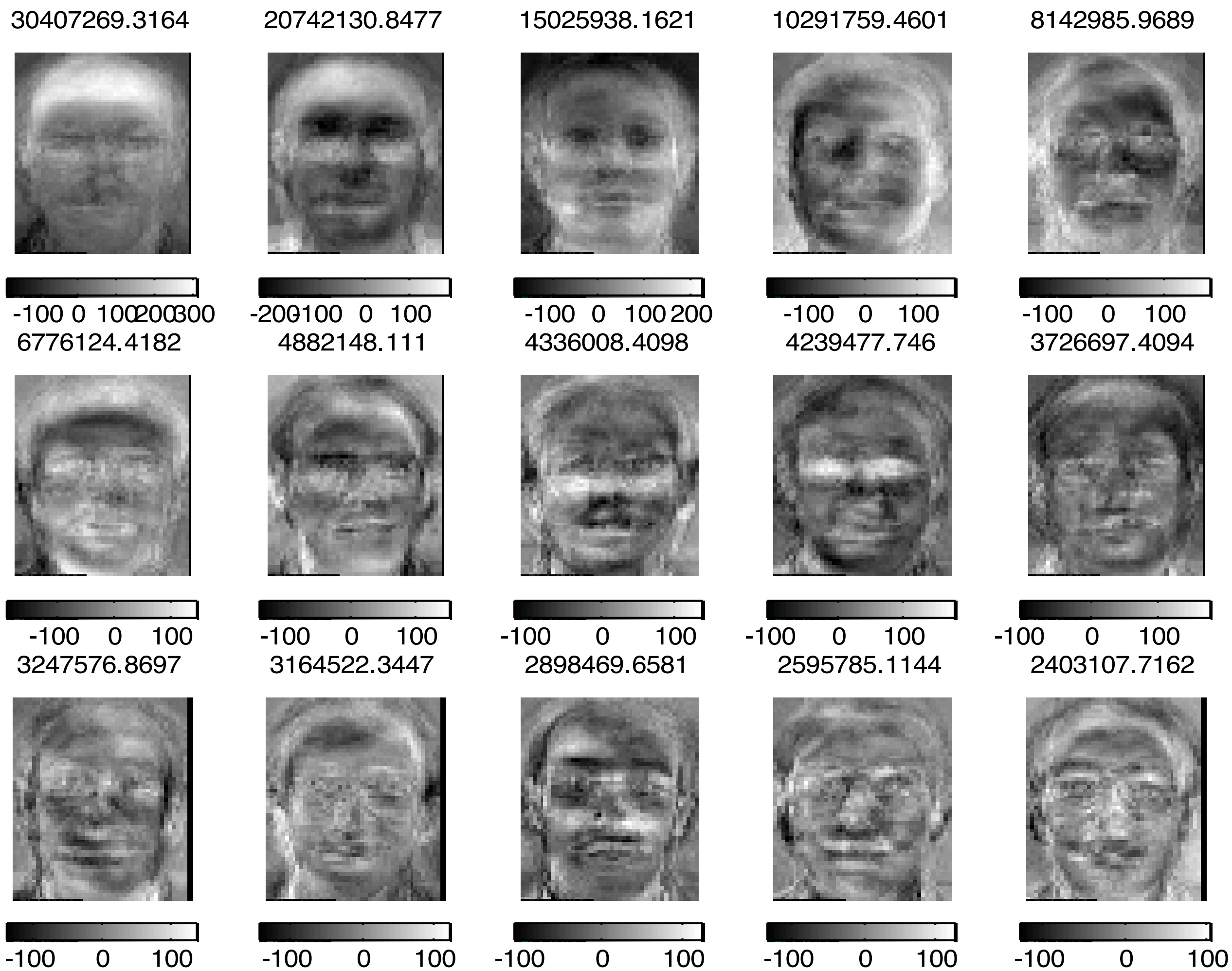


Image Database



EigenFaces

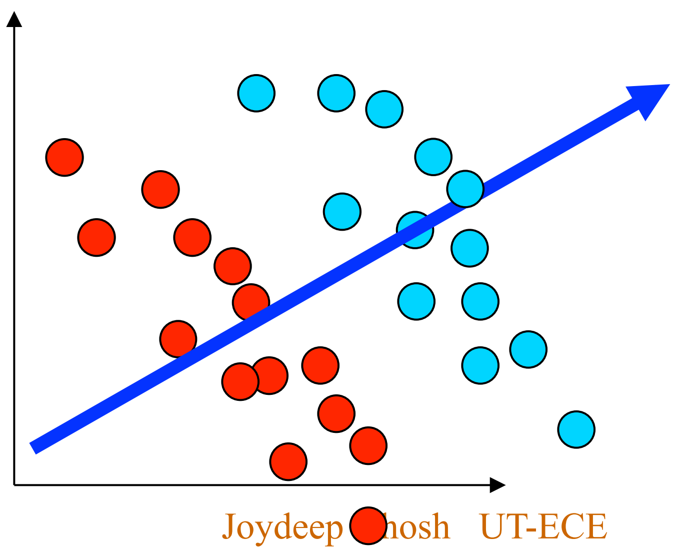
http://www.cs.princeton.edu/~cdecoro/eigenfaces/



Linear Supervised Method: Fisher's Linear Discriminant (FLD)

- FLD finds the projection direction that best separates the two classes
- Multiple discriminant analysis (MDA) extends LDA to multiple classes
- For fun: Fisherfaces vs. Eigenfaces https://www.youtube.com/watch?v=x8W_htbct3U

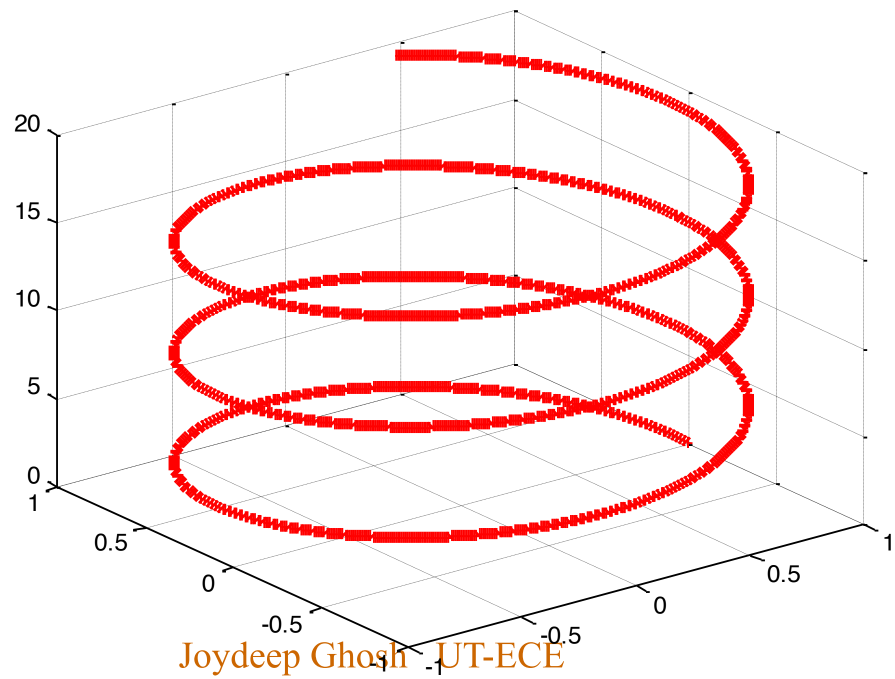
(David Mumford at 6:30)



Best projection direction for classification

Deficiencies of Linear Methods

- Data may not be best summarized by linear combination of features
  - Example: PCA cannot discover 1D structure of a helix



Multi-dimensional Scaling (MDS); Also see Perceptual Mapping

– When only pairwise distances (or similarities are known)
  - Objects may not be in Euclidean space
– Minimize a distortion measure ("stress")

Table 1 Flying Mileages Between 10 American Cities

Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington, DC	
0	587	1212	701	1936	604	748	2139	2182	543	Atlanta
587	0	920	940	1745	1188	713	1858	1737	597	Chicago
1212	920	0	879	831	1726	1631	949	1021	1494	Denver
701	940	879	0	1374	968	1420	1645	1891	1220	Houston
1936	1745	831	1374	0	2339	2451	347	959	2300	Los Angeles
604	1188	1726	968	2339	0	1092	2594	2734	923	Miami
748	713	1631	1420	2451	1092	0	2571	2408	205	New York
2139	1858	949	1645	347	2594	2571	0	678	2442	San Francisco
2182	1737	1021	1891	959	2734	2408	678	0	2329	Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington, DC

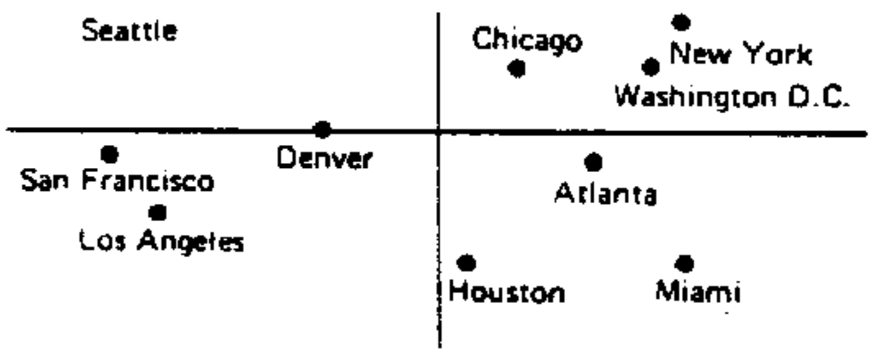
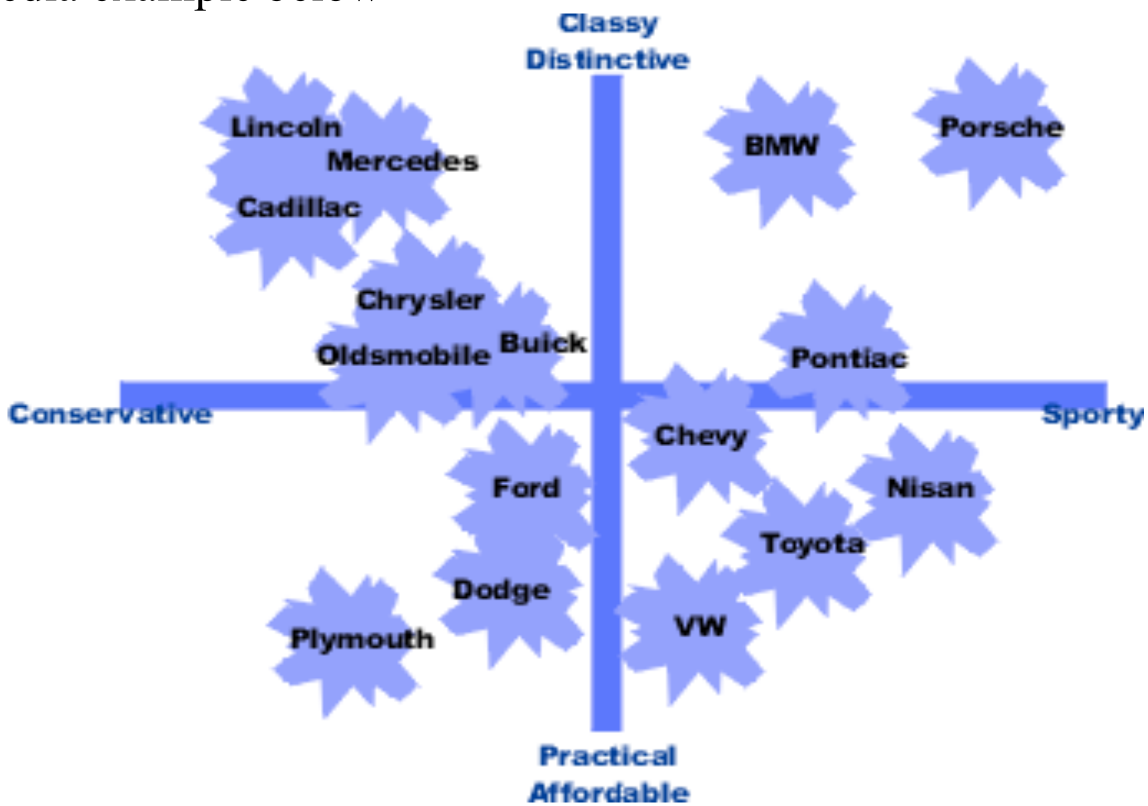


Figure 1 CMDS of flying mileages between 10 American cities.

MDS also called Perceptual Mapping

- In the marketing literature
  - Used for brand positioning etc.
  - Wikipedia example below



http://faculty.msb.edu/homak/homahelpsite/WebHelp%20061813/Multidimensional_Scaling_Illustration.htm

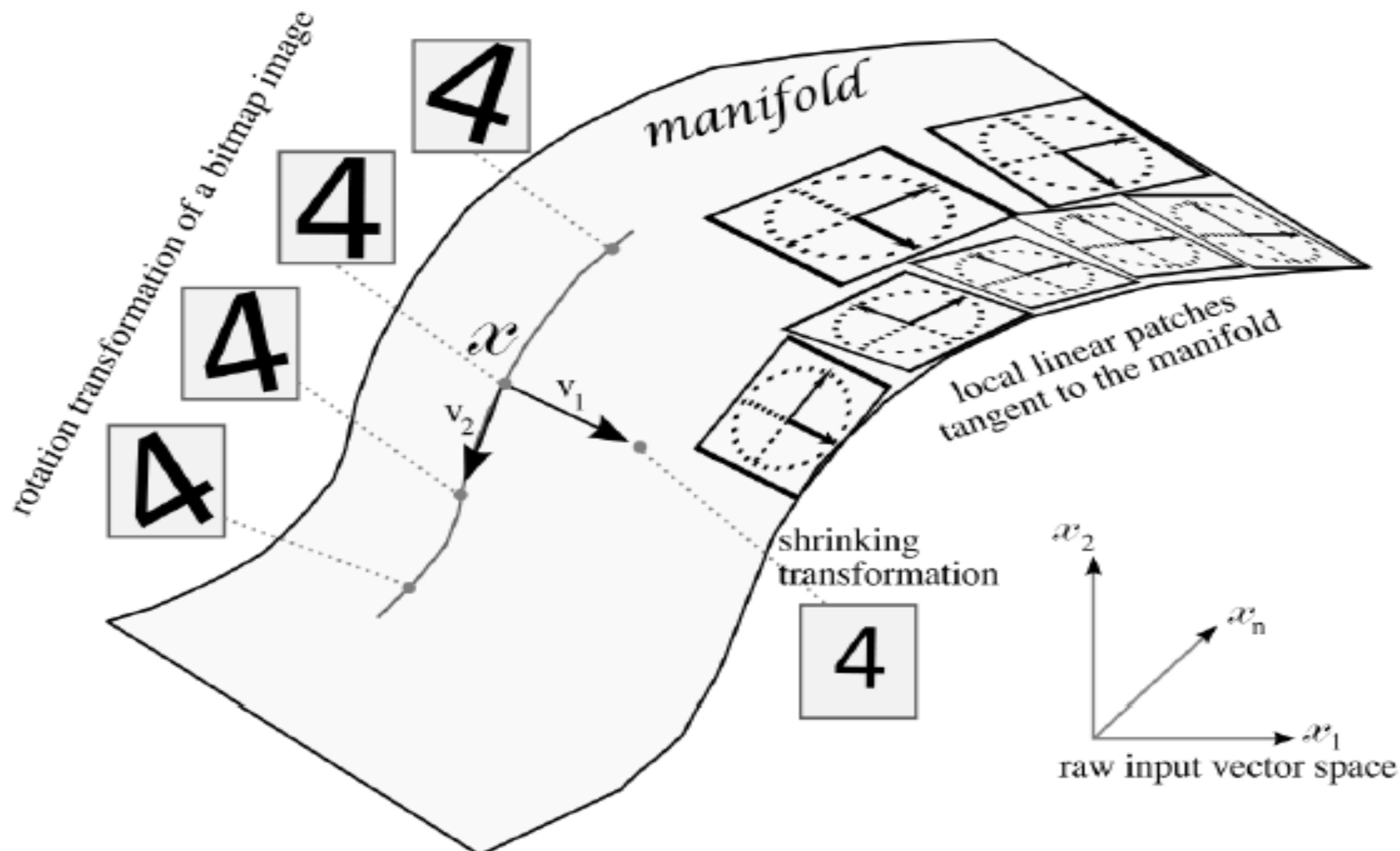
Deficiencies of Linear Methods

- Useful characteristics for real world data are often not linear combination of features
  - Example: poses of faces

  - Example: when shall I get hit (from motion data)

Figures from ISOMAP paper and Jerkins *et. al*

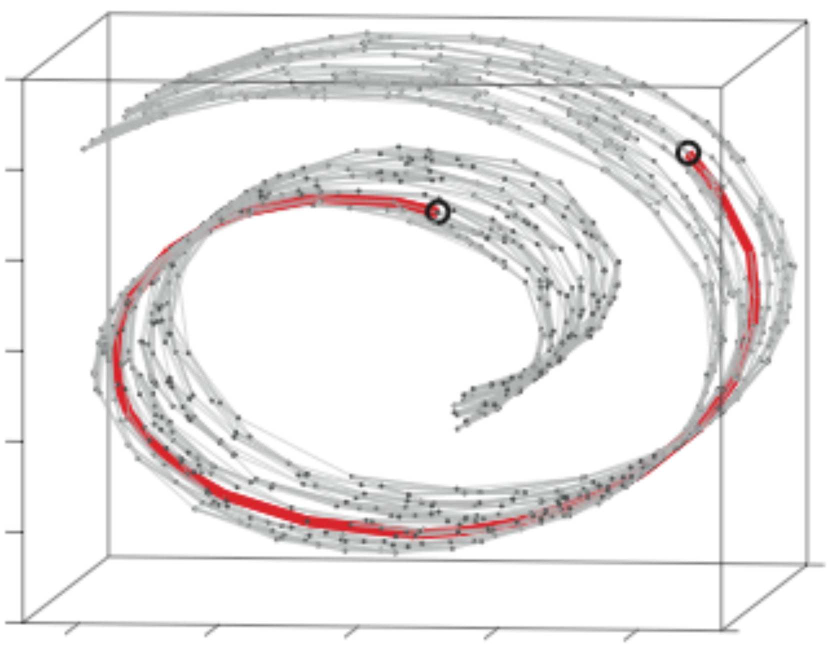
Handwritten Digits lie near a low-D manifold



Non-Linear Dimensionality Reduction

- Manifold based (ISOMAP, SOM,…)
  - The “swiss roll” below is an example of a manifold
  - Distance should be Measured on the Manifold and not in original space
- Multi-dimensional Scaling (in general)

B



Which Technique is best?

- Data Set characteristics
  – Pairwise distance (or similarity) only? (multi-dimensional scaling)
  – attributes ordered? Hierarchical?…
  – sparse data? Skewed data?
  – Dimensionality
- Metrics:
  – accuracy vs. reduction
  – progressive resolution refinement
  – incremental maintenance

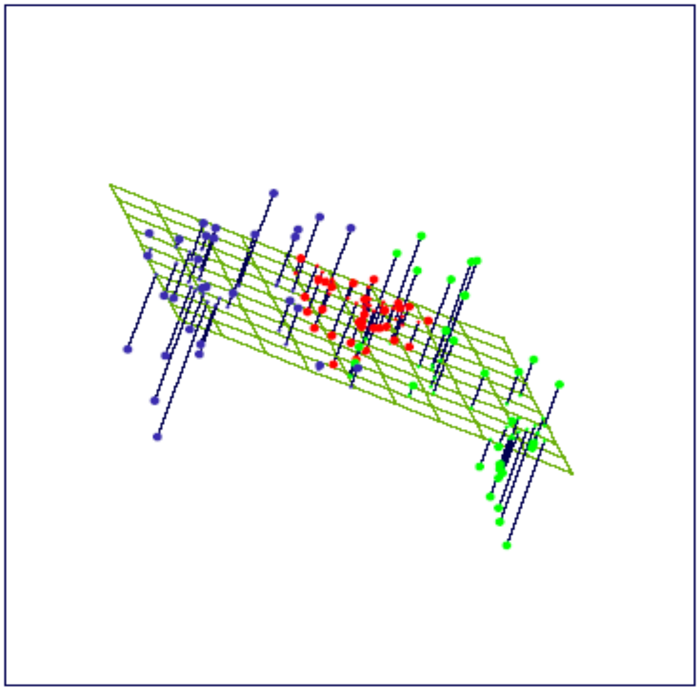
Data and Results Validation

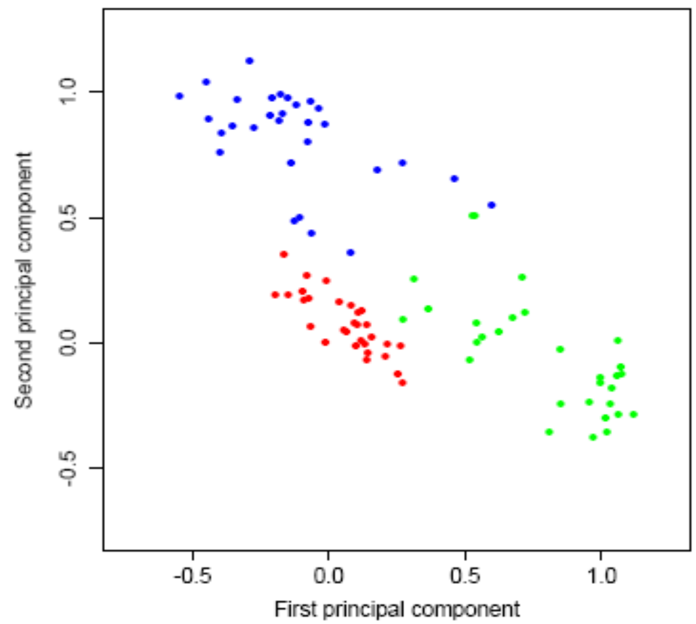
- **Bonferroni's Theorem**: if there are too many possible conclusions to draw, some will be true for purely statistical reasons, with no physical validity
- If possible, see that the entries make sense and data was collected properly
  - Ex: milk study at Lanarkshire, Scotland
- <u>Data is often observational and not experimental</u>
- Results validation vs. data dredging, snooping, fishing
  - E.g. S&P index almost perfectly predicted by butter, cheese production and sheep population in US and Bangladesh
  - "parapsychologist" David Rhine found (1950's) found about .1% guessed all 10 card colors correctly, but failed in next round.
    - Concluded that "telling people they have ESP causes them to lose it"!
    - www.tylervigen.com

Visualization

- Of data; process; results
- Motivation
  - For data-driven hypotheses human interaction is necessary
    - Humans can quickly analyze complex systems
    - Humans are good at pattern recognition
    - Humans are flexible
  - Exploratory Data Analysis
  - And communication! http://www.gapminder.org/

Geometric Projection using PCA

- 



- "Half-Sphere Example, HTF Fig 14.21
- Often one projects along multiple pairs of Principal Components.

Principal Surfaces (Non-Linear)

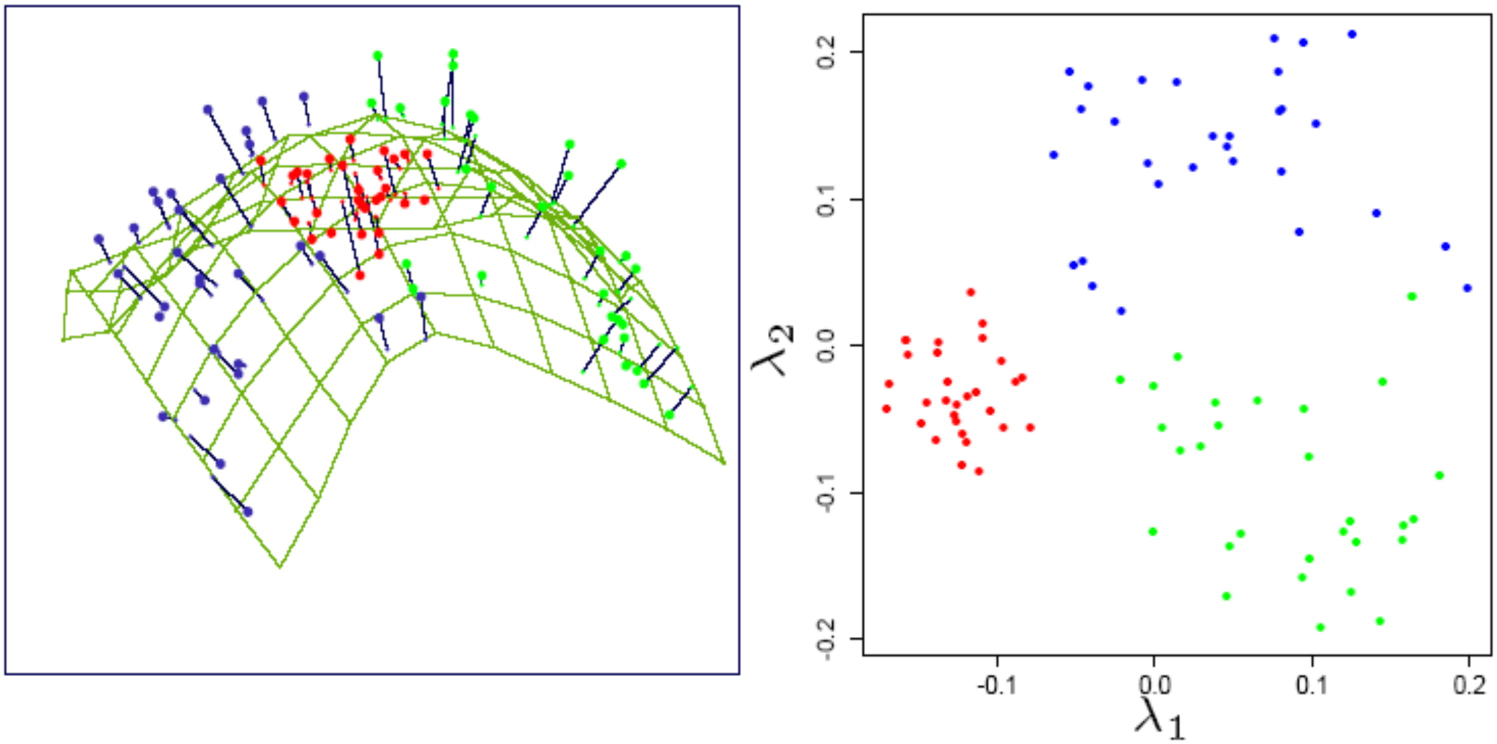


Figure 14.26: *Principal surface fit to half-sphere data.*

Actual Half-Sphere Data

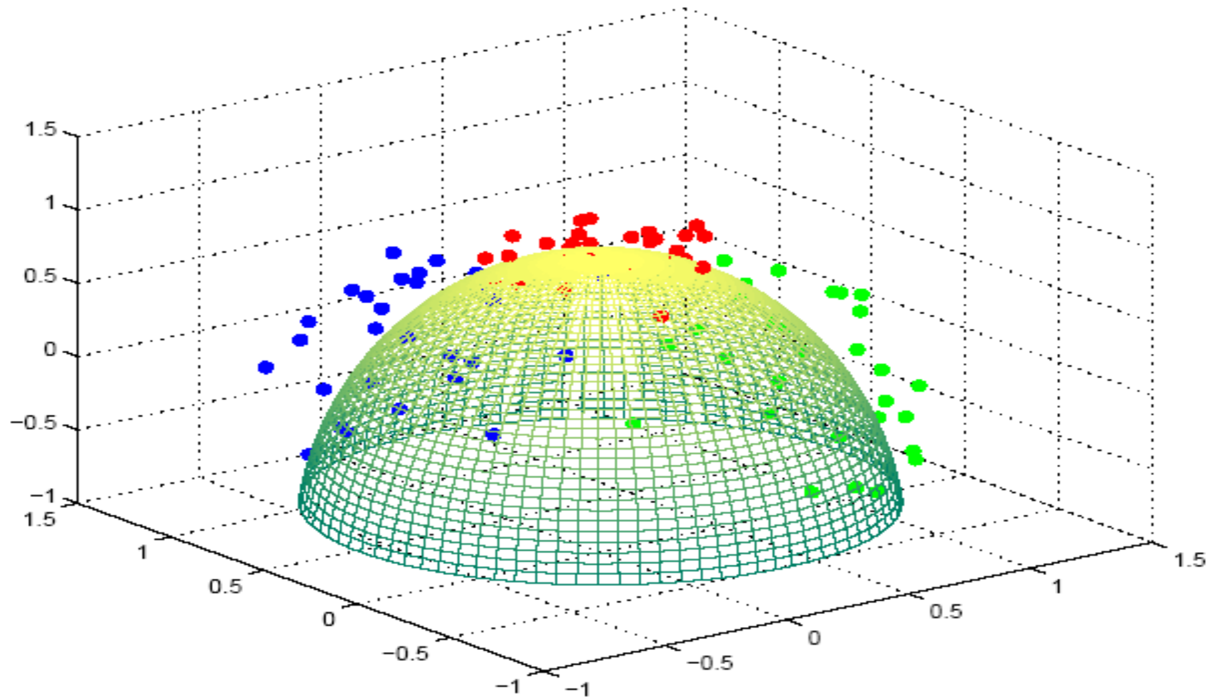


Figure 14.15: *Simulated data in three classes, near the surface of a half-sphere.*

Modern Web-Based Visualization

- Interactive, often Javascript based
- http://d3js.org/
- **D3.js** is a JavaScript library for manipulating documents based on data. **D3** helps you bring data to life using HTML, SVG and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.
- Gallery at: https://github.com/mbostock/d3/wiki/Gallery
- Webinar and ebook: http://it-ebooks.info/book/1265/
  - **D3** allows you to bind arbitrary data to a Document Object Model (DOM), and then apply data-driven transformations to the document. For example, you can use D3 to generate an HTML table from an array of numbers. Or, use the same data to create an interactive *Scalable Vector Graphics* (*SVG*) bar chart with smooth transitions and interaction.
- http://nvd3.org/
  - Simpler than D3.js

R/Python visual interfacts

- Ggplot2 / matplotlib: static
- Python: http://bokeh.pydata.org/en/latest/
  - Gallery shows source code
- R: Shiny from Rstudio
- http://shiny.rstudio.com/
  - Again gallery shows code: (server.R, ui.R)

- Also see
  - Google charts https://developers.google.com/chart/?hl=en
  - Google bubble chart is similar to Gapminder video:
  https://www.youtube.com/watch?v=jbkSRLYSojo
  - http://setosa.io/ (e.g Simpson's paradox, PCA visuals etc)
  - http://www.highcharts.com/

Extras

Singular Value Decomposition (SVD)

- Practical way of obtaining Principal components

customer \ day	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

$$\boxed{\mathbf{A}} = \begin{bmatrix} 0.18 & 0 \\ 0.36 & 0 \\ 0.18 & 0 \\ 0.90 & 0 \\ 0 & 0.53 \\ 0 & 0.80 \\ 0 & 0.27 \end{bmatrix} \times \begin{bmatrix} 9.64 & 0 \\ 0 & 5.29 \end{bmatrix} \times \begin{bmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{bmatrix}$$

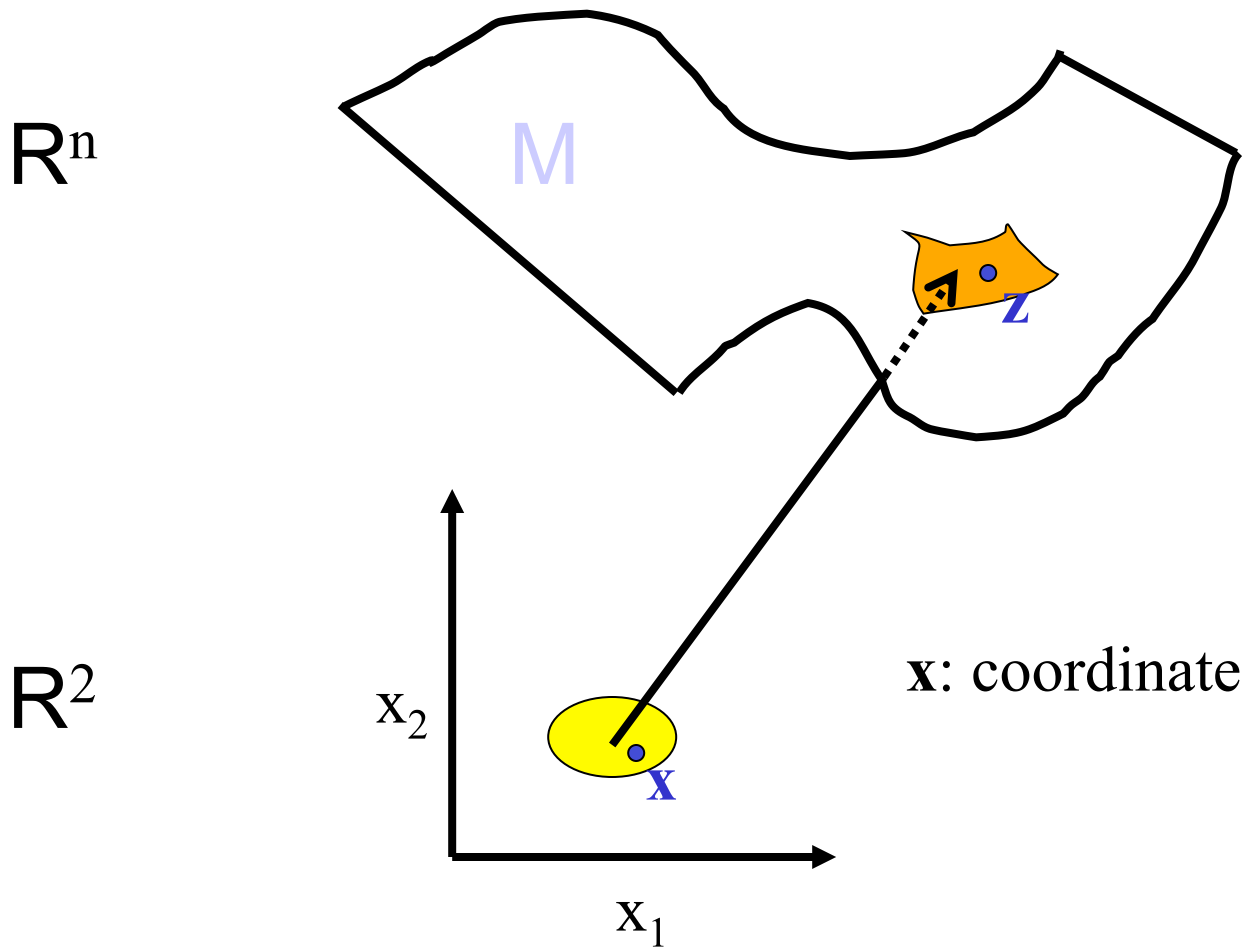
SVD

- Singular Value Decomposition (SVD)
  $A = U \times \Lambda \times V^T$
  - for A = customer -day matrix, interpret
  - U as customer-to-pattern similarity matrix
    - Columns of U are (orthonormal) eigen-"days"
      - Eigenvectors of AA^T
  - V as day-to-pattern similarity matrix
    - Rows of V are (orthonormal) eigen-"customers"
      - Eigenvectors of A^TA
  - is diagonal matrix of singular values (sorted)
    - (sq. root of eigen-values of AA^T Or A^TA)

Manifold and Dimensionality Reduction

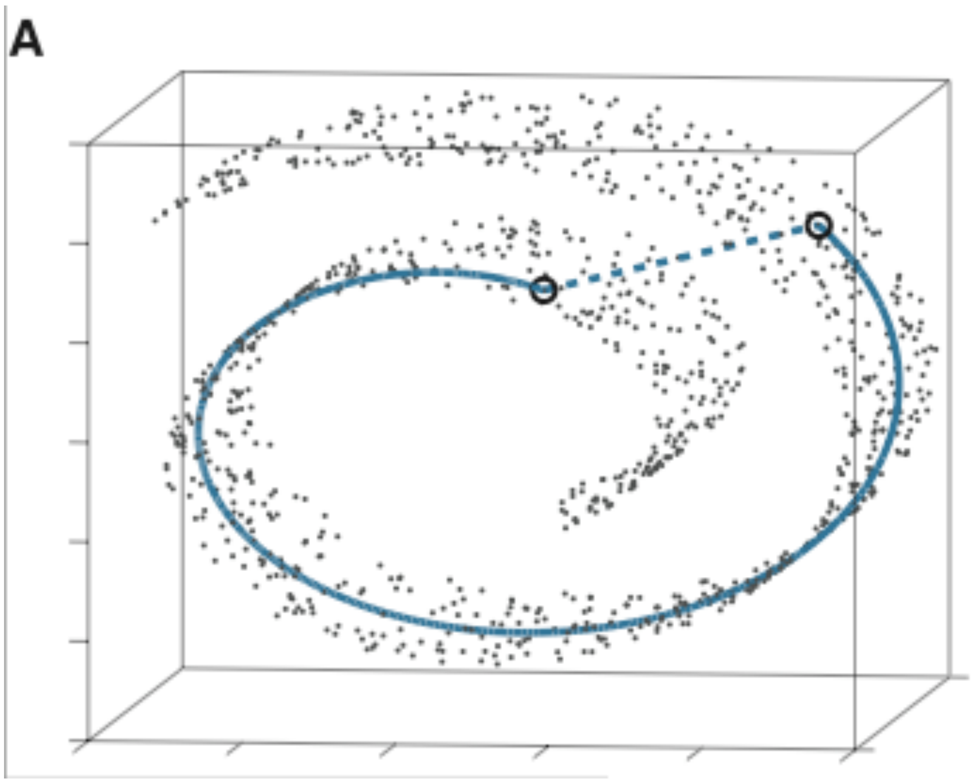
- Manifold: generalized "subspace" in R^n ($n >> 1$)
- Points in a *local* region on a manifold can be indexed by a subset of R^k
  - The value of k is usually small
  - Thus map n-dim space into local k-dim coordinates.
  - Neural approaches include SOM and GTM

Example of a Manifold



x: coordinate for **z**

Example: Manifold in Swiss Roll



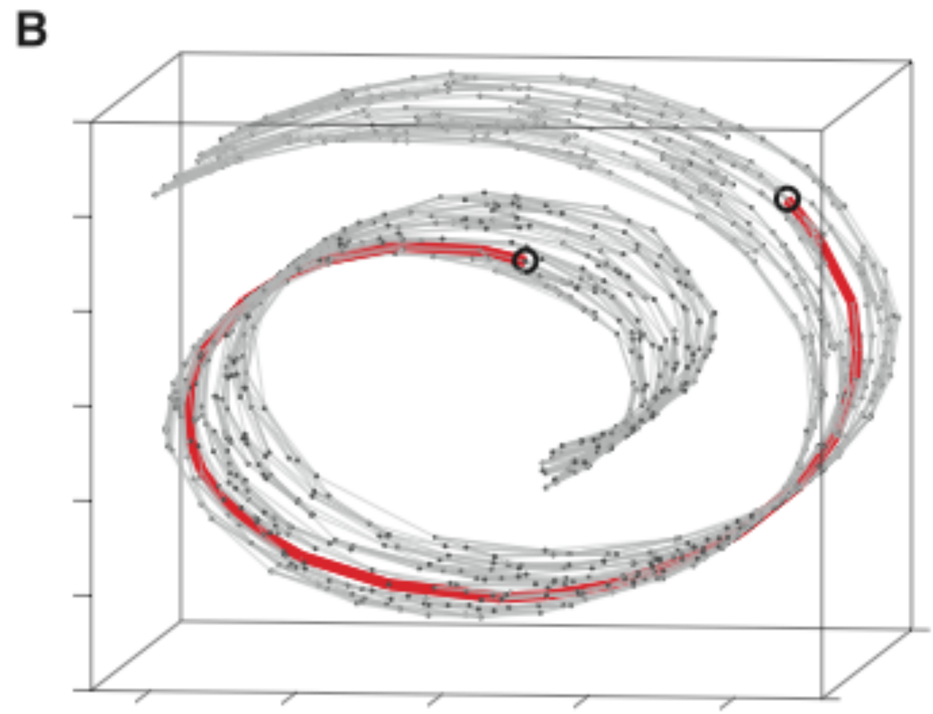
ISOMAP Algorithm

- Goal: preserve intrinsic geometry of data as captured via distances (**along the manifold**) between pairs of data points

Steps:

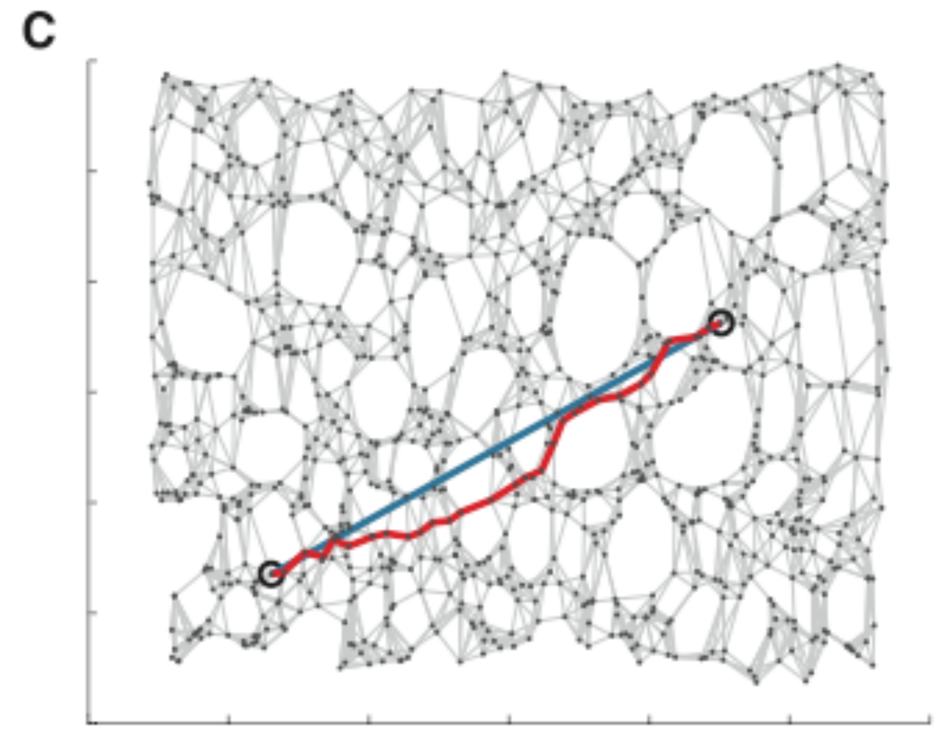
1. Determine which points are neighbors
2. Estimate geodesic distances and compute shortest path
   – For near points, measuring input space distances provides good enough approximation
   – For distant points, add up a sequence of short hops between neighboring points
3. Apply MDS to matrix of distances

- Estimating geodesic distances via shortest path



Step 3

- Apply classical MDS to matrix of distances

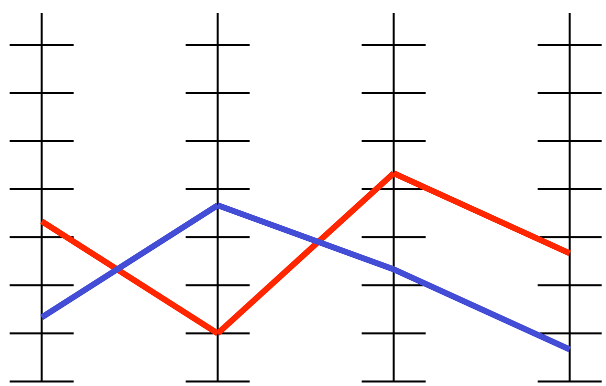


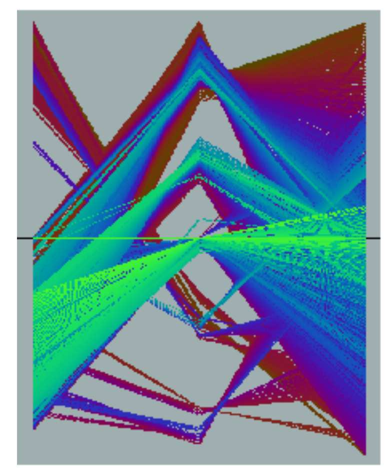
Parallel Coordinates

- Draw each point as an open polygon through *k* equidistant axes

K=4

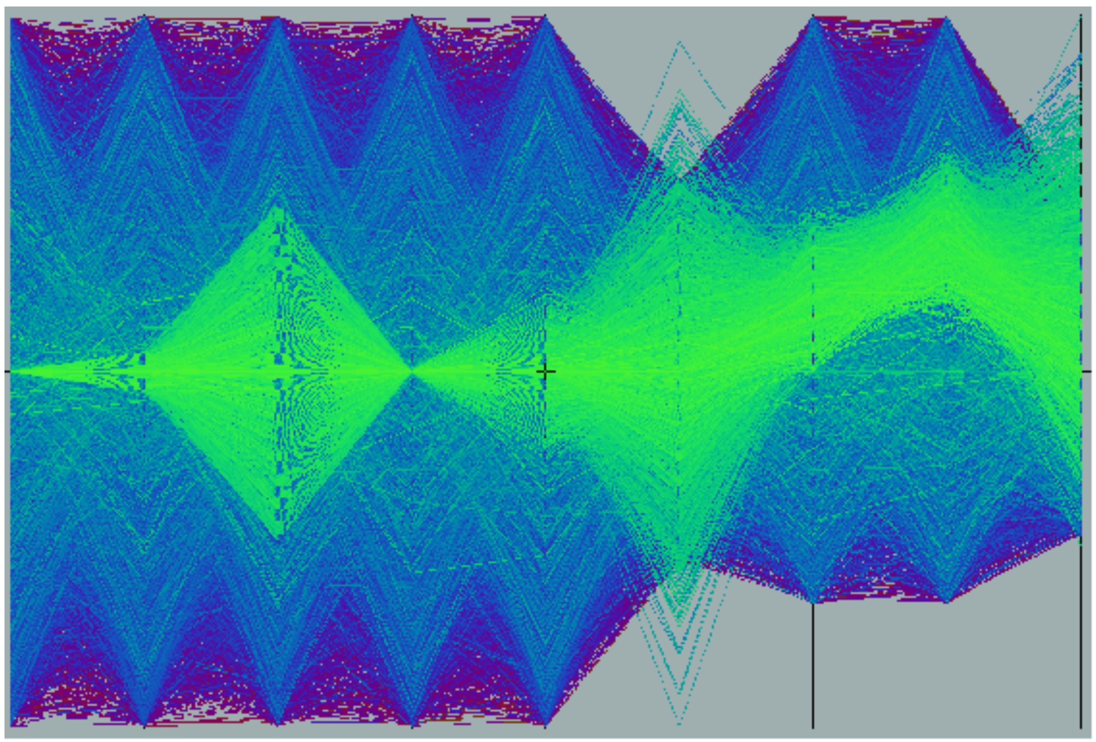
3.2	1.3
1.0	3.6
5.3	2.3
2.7	1.8





Illustration

- Parallel Coordinates good at seeing distributions in O(10) dimensions, O(1000) points



Uniform

0 mean gaussian

Non-0 mean gaussian

Histograms

- popular in commercial databases
  - often gives low error estimates, using small space
  - used mainly for selectivity estimation purposes within a query optimizer or in query execution (e.g. load balancing).
- Idea: "uniform" summary of (numeric) data within "buckets"
- choices for bucket location:
  - equi-width or equi-sum
  - V-optimal (min. variance)
  - MaxDiff
  - Compressed: big ones are singleton, others are equi-summed (e.g. DB2)

Histograms II

- Choice depends on the goal: which parameter should the histogram try to estimate quickly and accurately.

- <u>Evaluation:</u>
  - can apply to unordered attributes only if hierarchy is present
  - works well with near-uniform or highly skewed data; poorly with moderately `skewed data !!`
  - scaling to high-D??
  - progressive resolution refinement?? (only if histogram built on hierarchical data)
  - Incrementally updatable?? (depends)

Table 2 from Jersey Report

Data Type	SVD	Wavelet	Regression	Log-Linear	Histogram	Clustering	Index Tree	Sampling
Distance Only	N	N	N	N	D	Y	M	Y
Unordered Flat	Y	N	N	Y	D	N	M	Y
Unordered Hierarchical	Y	M	N	Y	M	M	M	Y
Sparse	B	F	F	F	F	B	F	D
Skewed	F	F	B	F	F	F	F	D
High Dimensional	N	F	W	W	M	D	W	W
Secondary Metrics								
Progressive Resolution	Y	Y	Y	N	M	D	Y	Y
Incremental Computation	N	Y	M	N	M	M	Y	Y

Y = Yes ; N = No ; M = Maybe;
F = Fine ; B = Better ; W = Worse ;
D = Depends (on further specification, could be better or worse).

Table 2: Applicability of data reduction techniques to different types of data