Data Preprocessing

Lecturer: Dr. Bo Yuan

E-mail: yuanb@sz.tsinghua.edu.cn

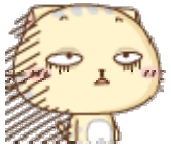
Outline

- Data Cleansing
- Data Transformation
- Data Description
- Feature Selection
- Feature Extraction

Where are data from?

Why Data Preprocessing?

- Real data are notoriously dirty!
  - The biggest challenge in many data mining projects
- Incomplete
  - Occupancy = " "
- Noisy
  - Salary = "-100"
- Inconsistent
  - Age = "42" vs. Birthday = "01/09/1985"
- Redundant
  - Too much data or too many features for analytical analysis
- Others
  - Data types
  - Imbalanced datasets



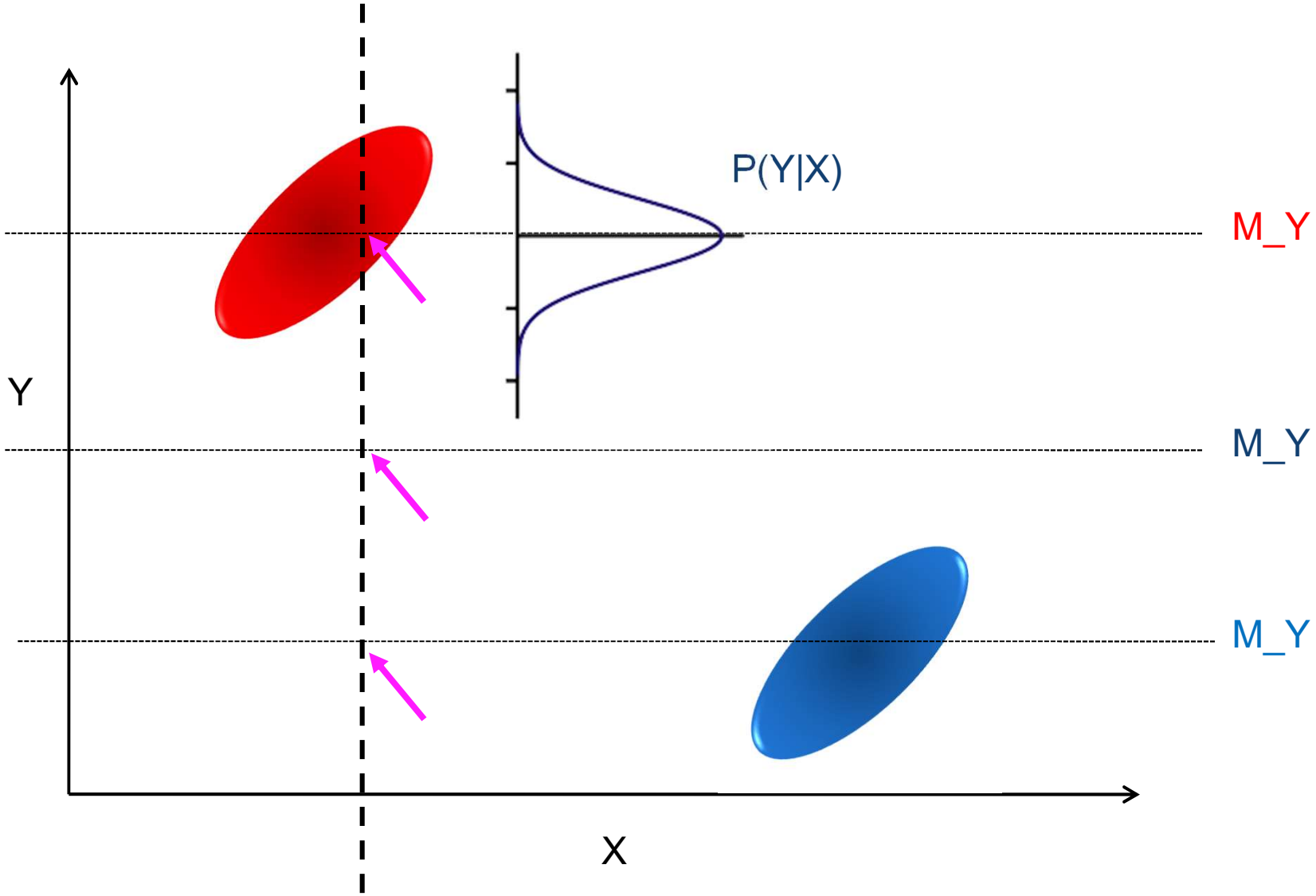
Missing Data

- Data are not always available.
  - One or more attributes of a sample may have empty values.
  - Many data mining algorithms cannot handle missing values directly.
  - May cause significant troubles.
- Possible Reasons
  - Equipment malfunction
  - Data not provided
  - Not Applicable (N/A)
- Different Types
  - Missing completely at random
  - Missing conditionally at random
  - Not missing at random

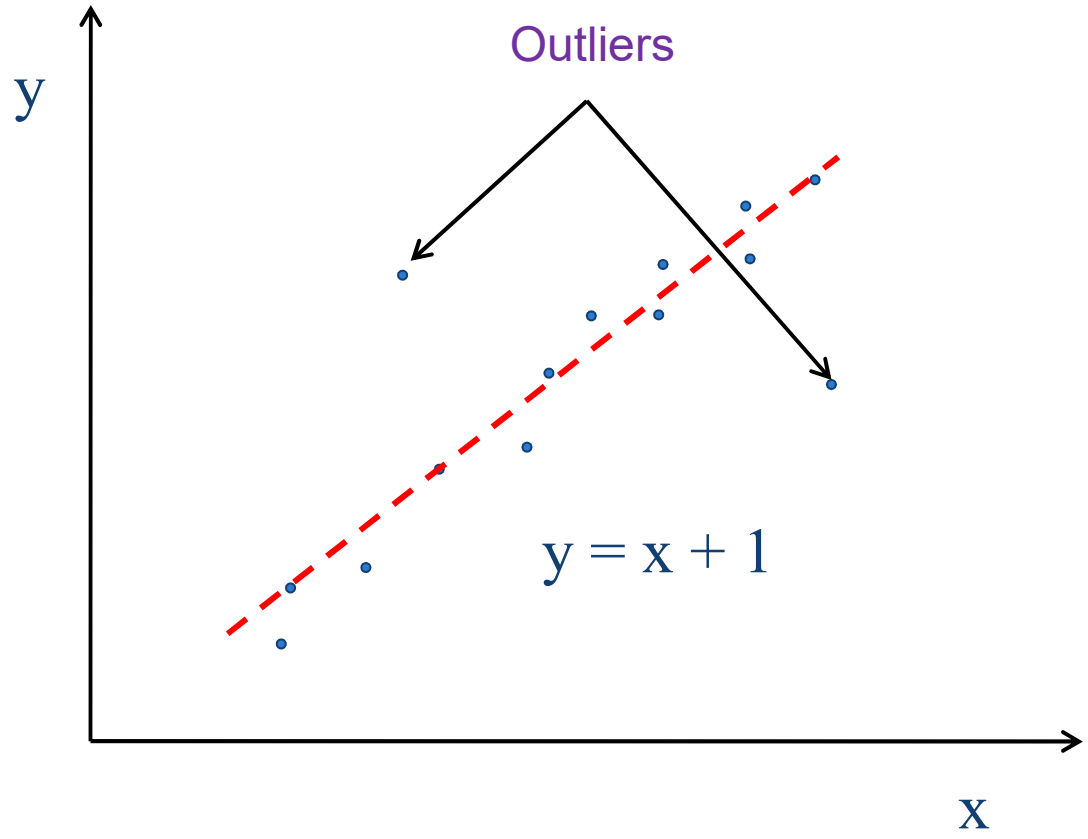
How to handle missing data?

- Ignore
  - Remove samples/attributes with missing values
  - The easiest and most straightforward way
  - Work well with low missing rates
- Fill in the missing values manually
  - Recollect the data
  - Domain knowledge
  - Tedious/Infeasible
- Fill in the missing values automatically
  - A global constant
  - The mean or median
  - Most probable values
- More art than science

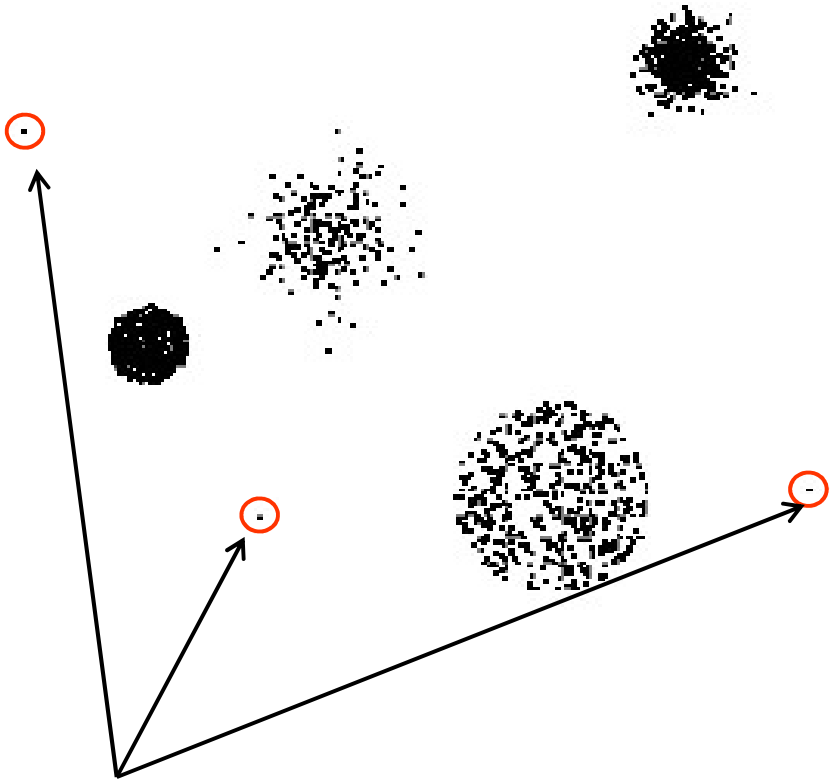
Missing Data: An Example



Outliers



Outliers

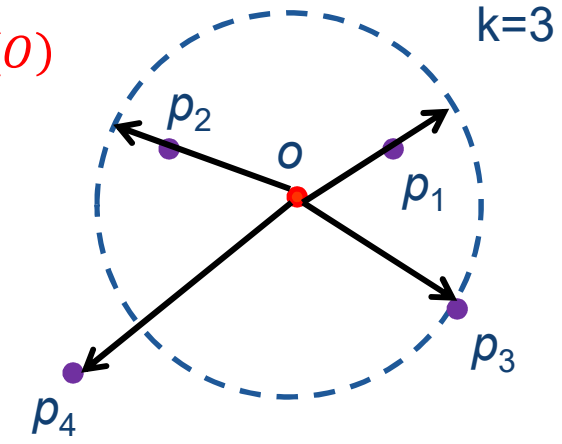


Outliers

Anomaly vs. Outlier

Local Outlier Factor

$distance_k(O)$

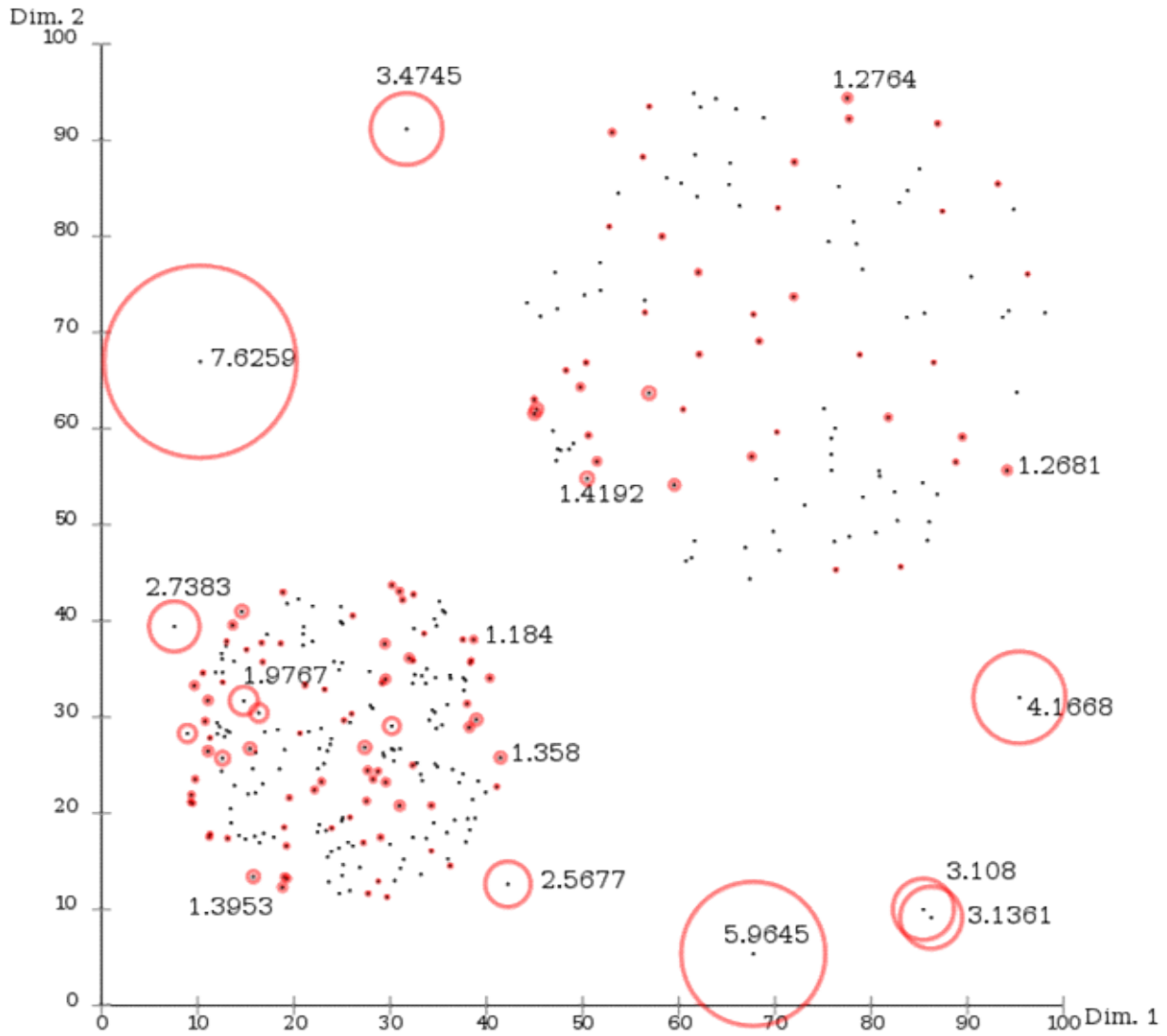


$$distance_k(A,B) = max\{distance_k(B), d(A,B)\}$$

$$lrd(A) = 1/\left(\frac{\sum_{B\in N_k(A)} distance_k(A,B)}{|N_k(A)|}\right)$$

$$LOF_k(A) = \frac{\sum_{B\in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} = \frac{\sum_{B\in N_k(A)} lrd(B)}{|N_k(A)|} / lrd(A)$$

Local Outlier Factor



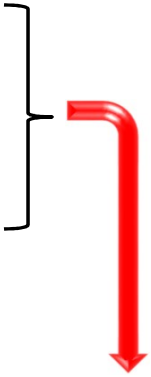
Duplicate Data

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

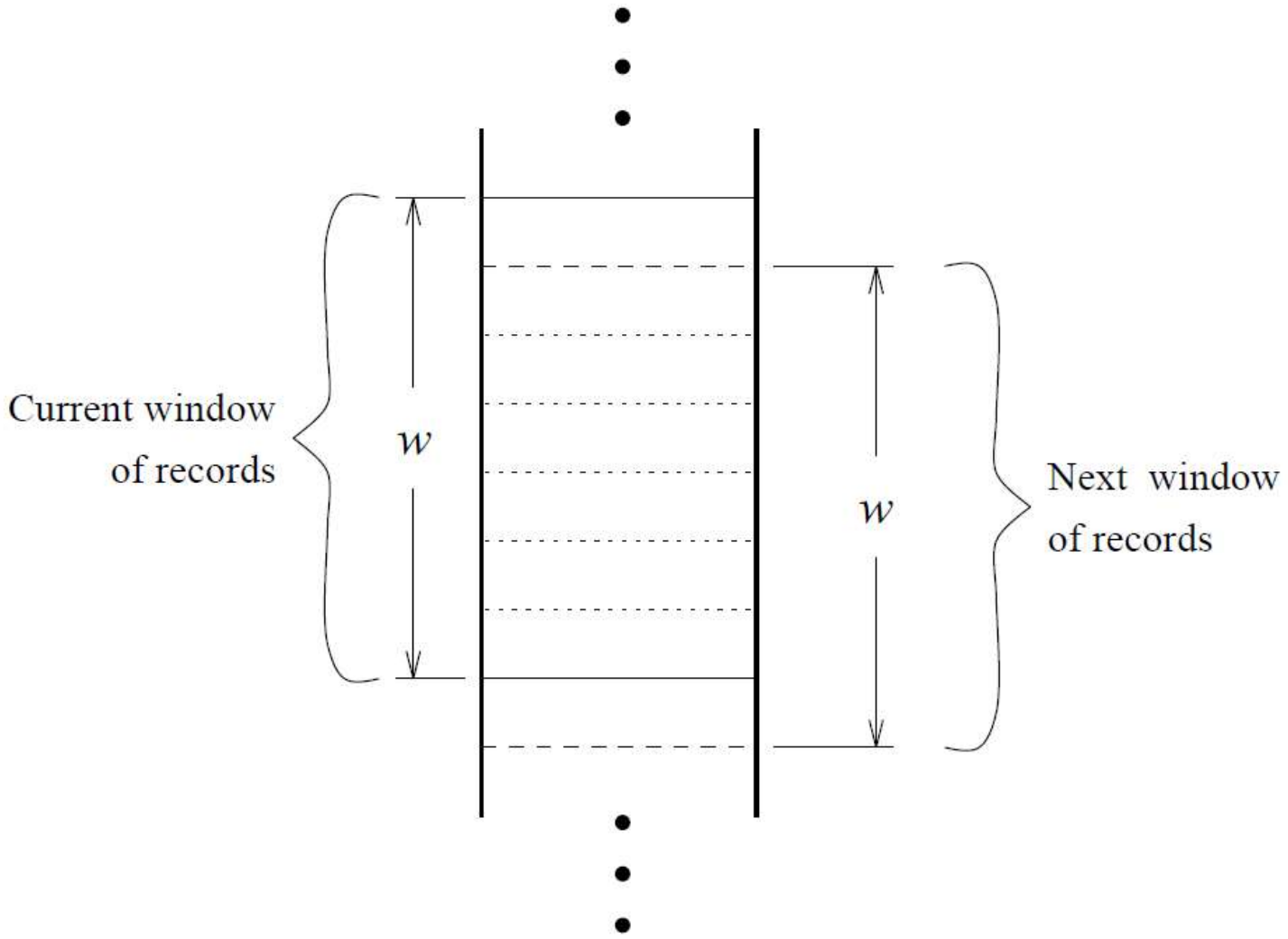
Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666



Customers (integrated target with cleaned data)

No	*LName*	*FName*	*Gender*	*Street*	*City*	*State*	*ZIP*	*Phone*	*Fax*	*CID*	*Cno*
1	Smith	Kristen L.	F	2 Hurley Place	South Fork	MN	48503-5998	444-555-6666		11	493
2	Smith	Christian	M	2 Hurley Place	South Fork	MN	48503-5998			24	
3	Smith	Christoph	M	23 Harley Street	Chicago	IL	60633-2394	333-222-6542	333-222-6599		24

Duplicate Data



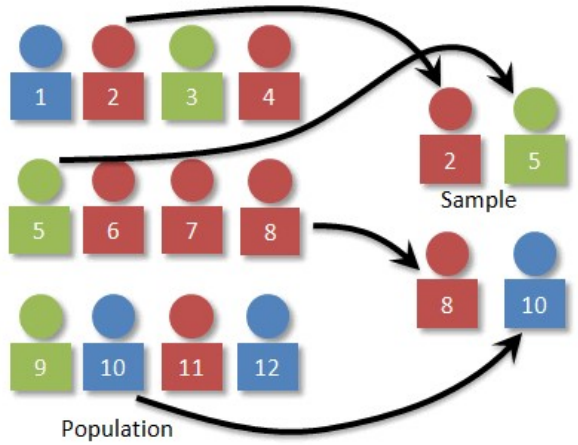
Duplicate Data

First	Last	Address	ID	Key
Sal	Stolfo	123 First Street	45678987	STLSAL123FRST456
Sal	Stolfo	123 First Street	45678986	STLSAL123FRST456
Sal	Stolpho	123 First Street	45688987	STLSAL123FRST456
Sal	Stiles	123 Forest Street	45654321	STLSAL123FRST456

```
Given two records, r1 and r2.
IF the last name of r1 equals the last name of r2,
       AND the first names differ slightly,
       AND the address of r1 equals the address of r2
THEN
       r1 is equivalent to r2.
```

Data Transformation

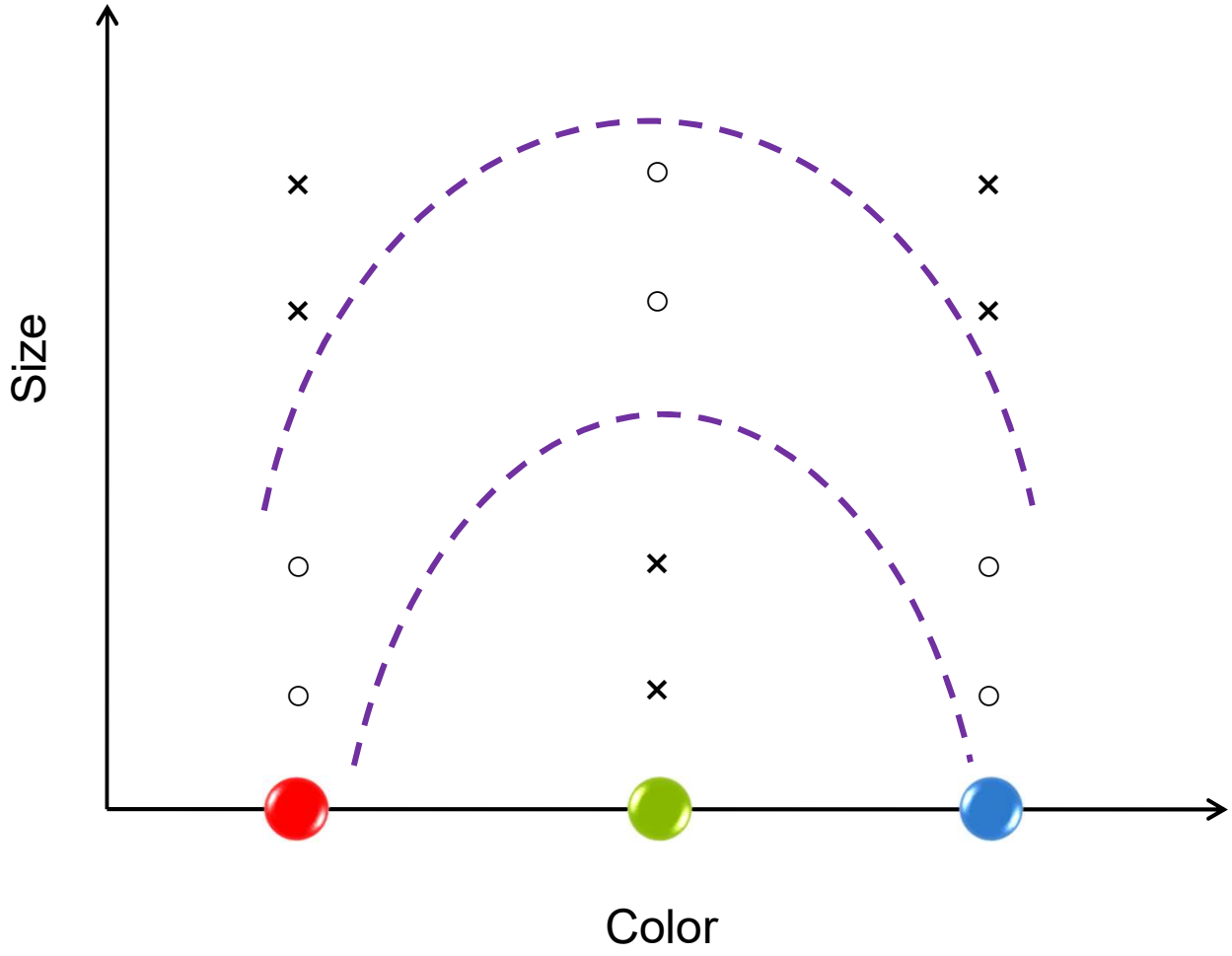
- Now we have an error free dataset.
- Still needs to be standardized.
- Type Conversion
- Normalization
- Sampling



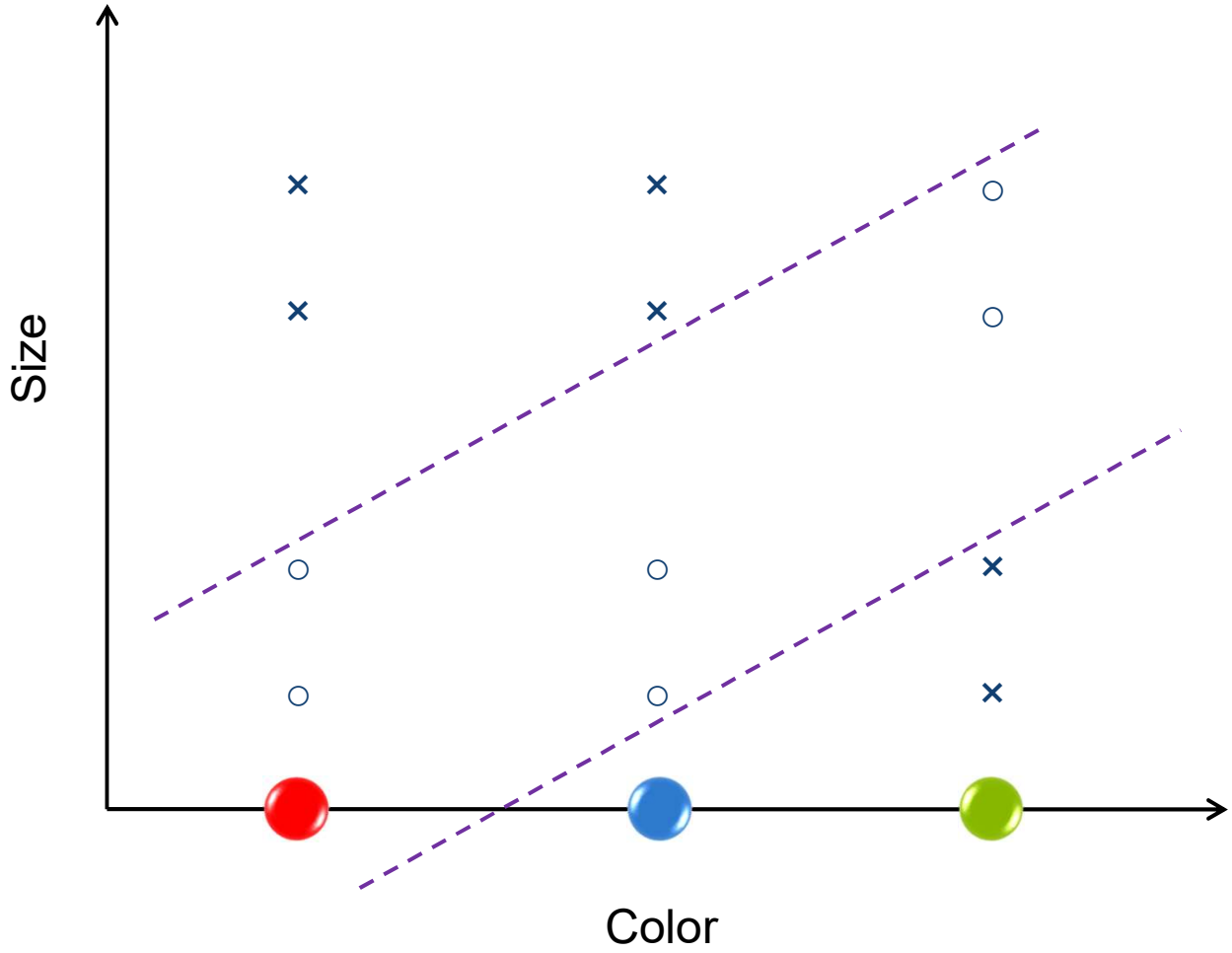
Attribute Types

- Continuous
  - Real values: Temperature, Height, Weight …
- Discrete
  - Integer values: Number of people …
- Ordinal
  - Rankings: {Average, Good, Best}, {Low, Medium, High} …
- Nominal
  - Symbols: {Teacher, Worker, Salesman}, {Red, Green, Blue} …
- String
  - Text: “Tsinghua University”, “No. 123, Pingan Avenue” …

Type Conversion



Type Conversion

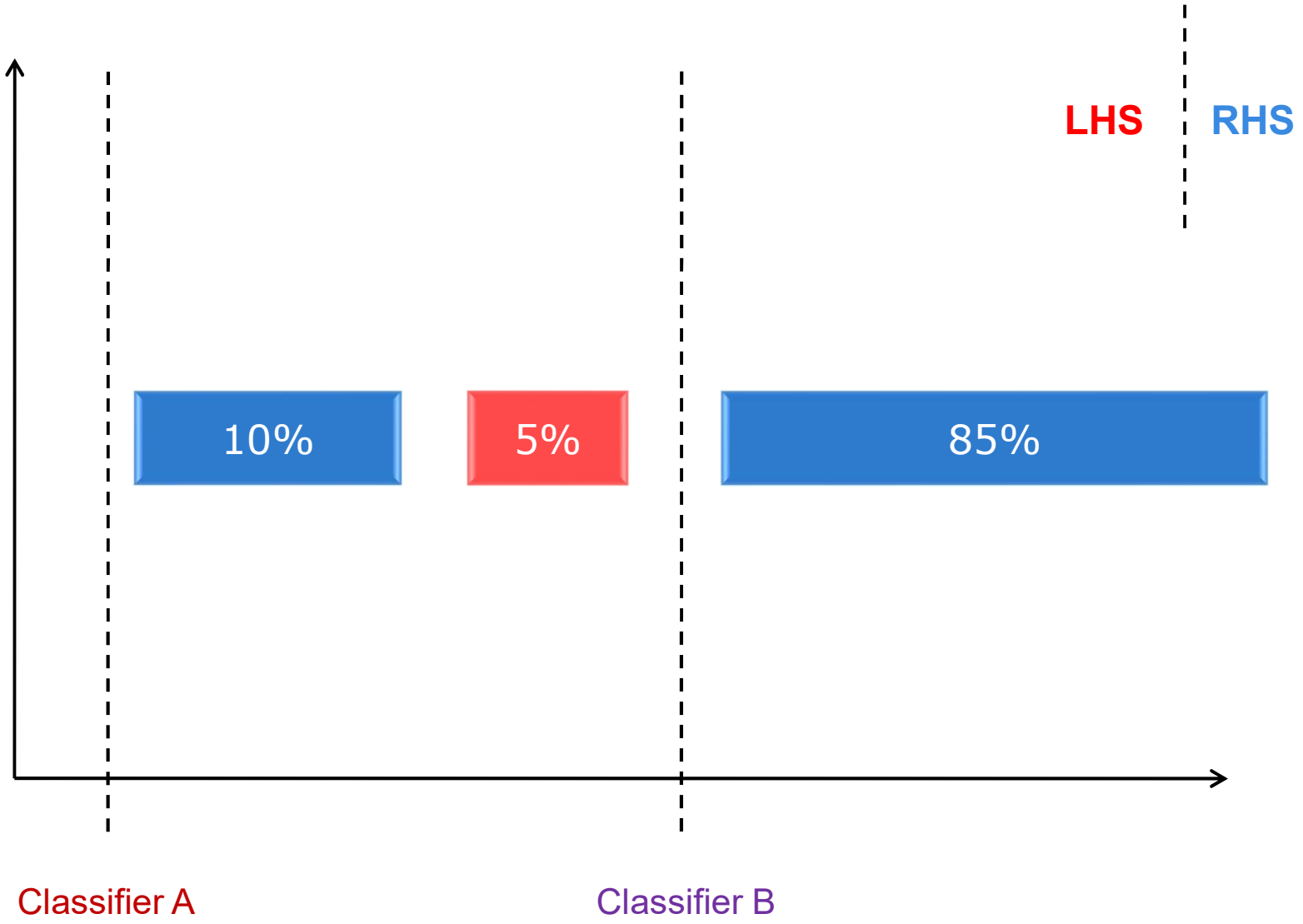


Type Conversion

Sampling

- A database/data warehouse may store terabytes of data.
- Processing limits: CPU, Memory, I/O …
- Sampling is applied to reduce the time complexity.
- In statistics, sampling is applied often because obtaining the entire set of data is expensive.
- Aggregation
  - Change of scale:
    - Cities → States; Days → Months
  - More stable and less variability
- Sampling can be also used to adjust the class distributions.
  - Imbalanced dataset

Imbalanced Datasets



Imbalanced Datasets

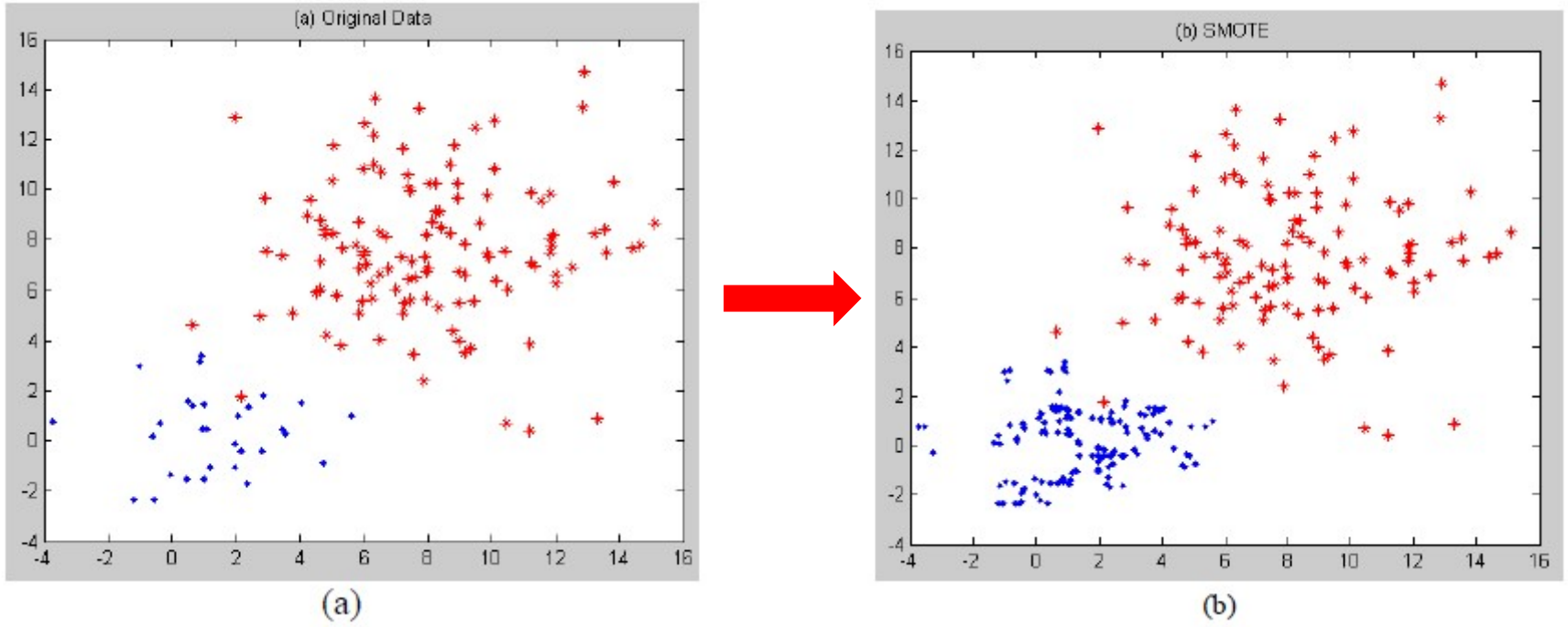
$$G - mean = (Acc^{+} \times Acc^{-})^{1/2}$$

$$where \;\; Acc^{+} = \frac{TP}{TP + FN}; \quad Acc^{-} = \frac{TN}{TN + FP}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

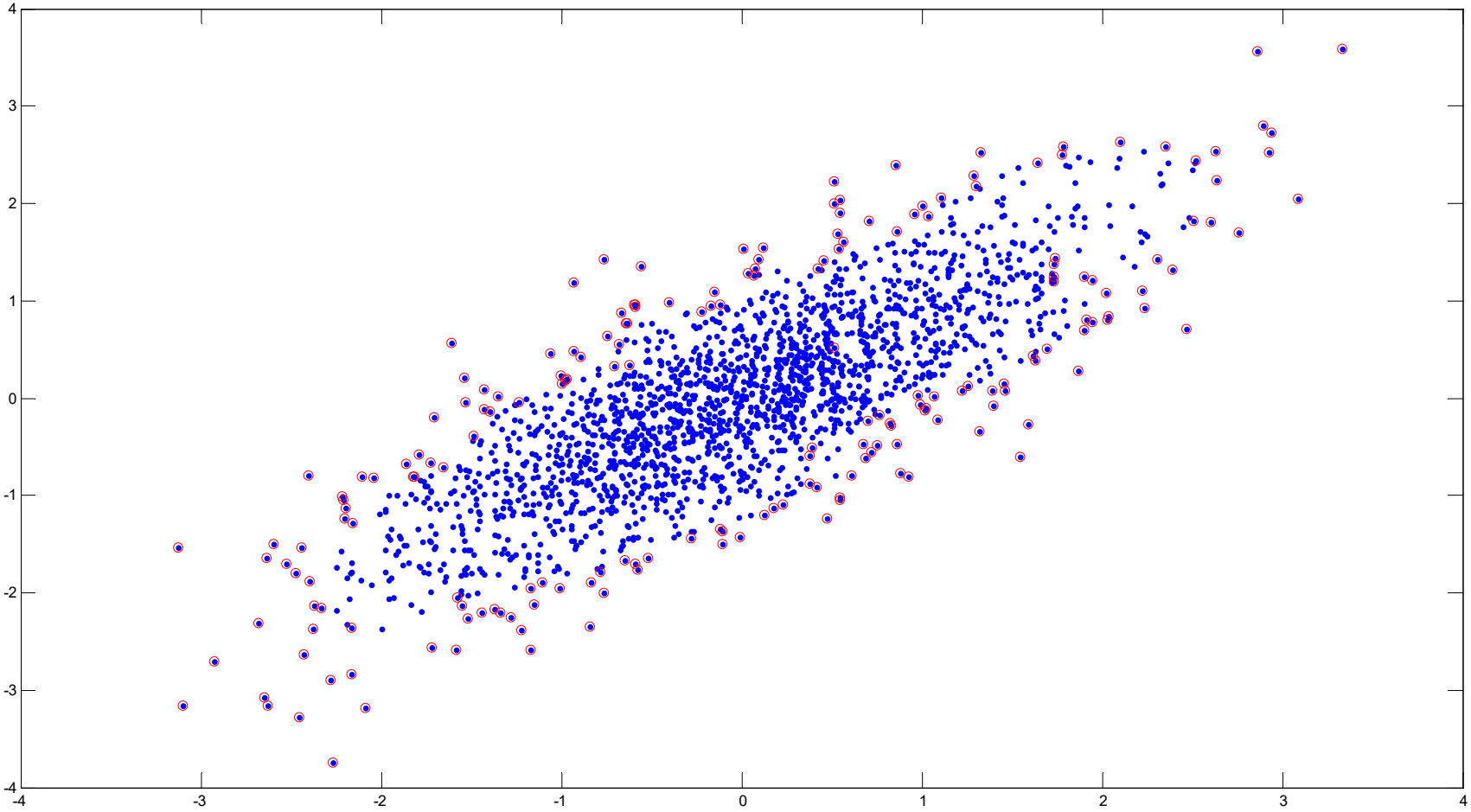
$$where \quad Precision = \frac{TP}{TP + FP}; \quad Recall = \frac{TP}{TP + FN} = Acc^{+}$$

Over-Sampling



SMOTE

Boundary Sampling



Normalization

- The height of someone can be 1.7 or 170 or 1700 …
- Min-max normalization:

$$v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$$

  - Let income range \$12,000 to \$98,000 be normalized to [0.0, 1.0]. Then \$73,600 is mapped to

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$$

- Z-score normalization (μ: mean, σ: standard deviation):

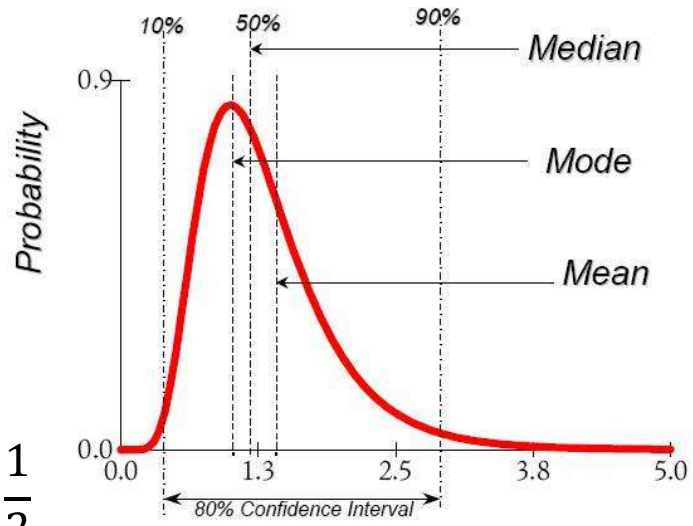
$$v' = \frac{v - \mu}{\sigma}$$

  - Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

Data Description

- Mean
  - Arithmetic mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n)$$



- Median

$$P(X \le m) = P(X \ge m) = \int_{-\infty}^{m} f(x)dx = \frac{1}{2}$$

- Mode
  - The most frequently occurring value in a list
  - Can be used with non-numerical data.

- Variance
  - Degree of diversity

$$Var(X) = E[(X - \mu)^2] \qquad Var(X) = \int (x - \mu)^2 f(x)dx$$

Data Description

- Pearson's product moment correlation coefficient

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\overline{A}\,\overline{B}}{(n-1)\sigma_A \sigma_B}$$

- If $r_{A, B} > 0$, A and B are positively correlated.
- If $r_{A, B} = 0$, no linear correlation between A and B.
- If $r_{A, B} < 0$, A and B are negatively correlated.

Data Description

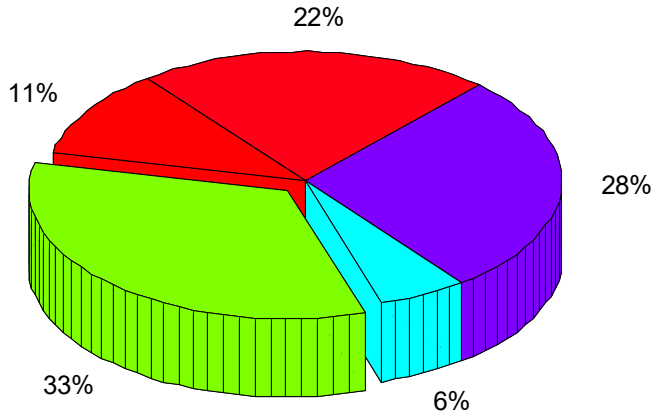
- Pearson's chi-square (χ^2) test

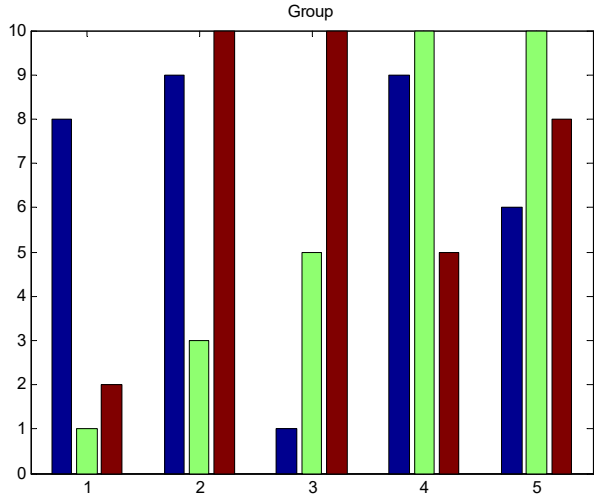
$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

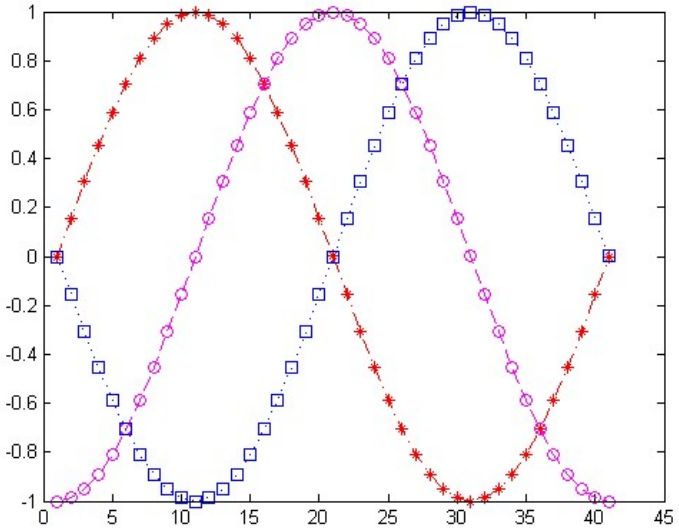
	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum (col.)	300	1200	1500

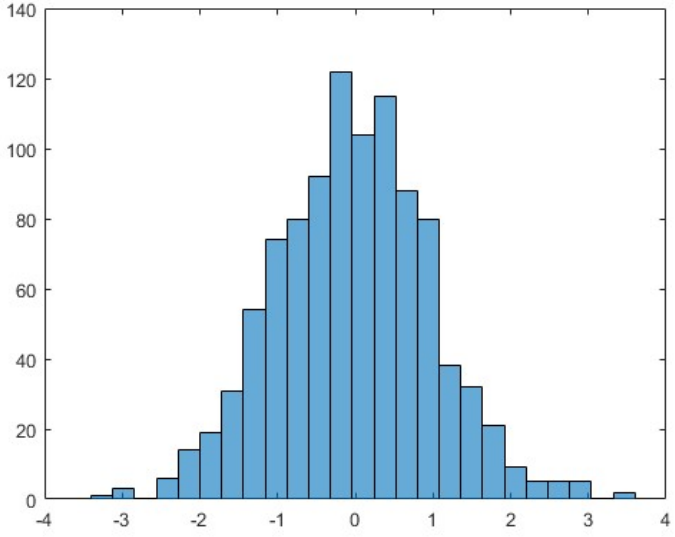
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

Data Visualization (1D)

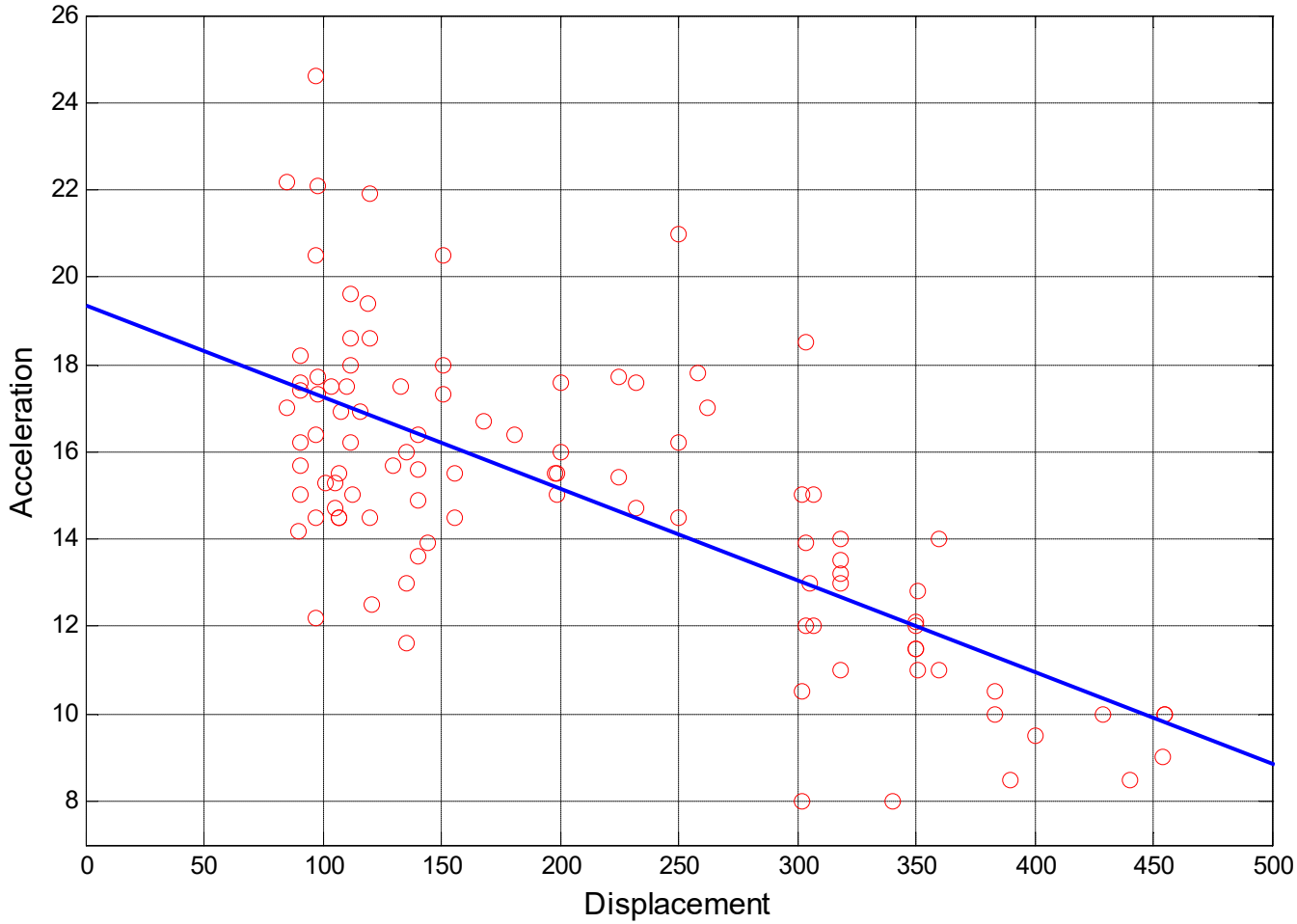




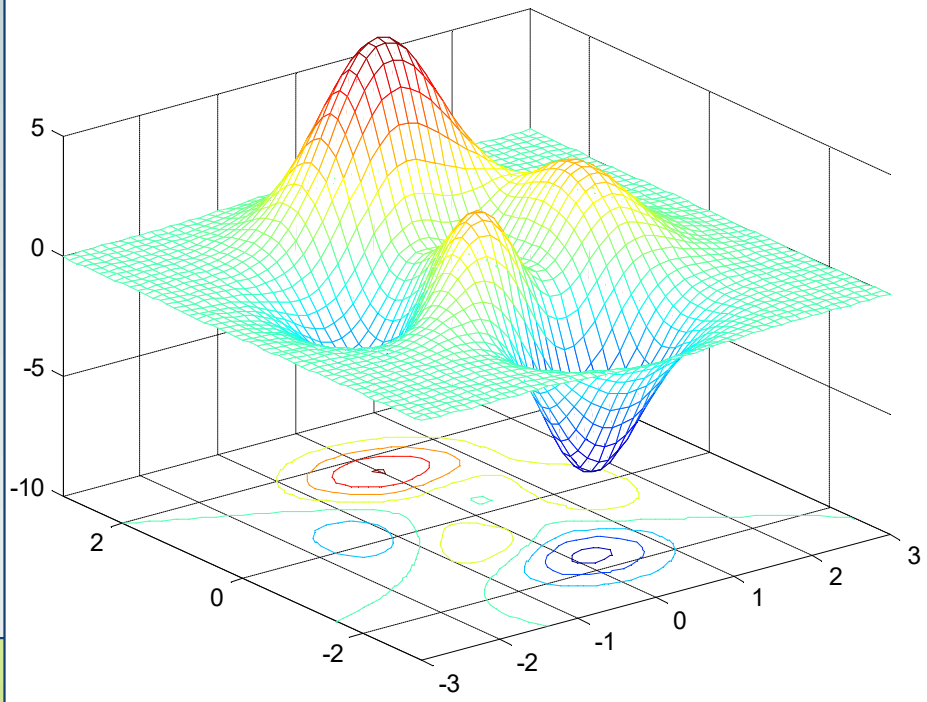


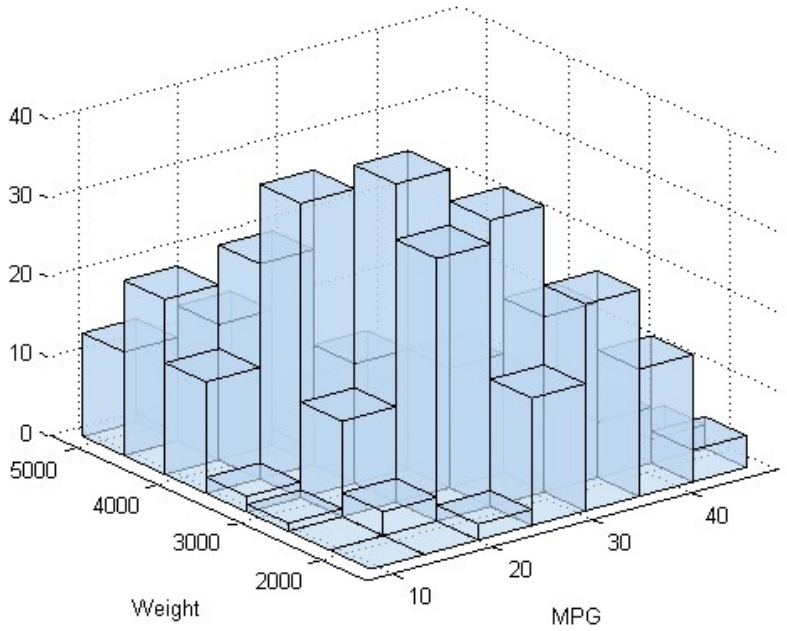


Data Visualization (2D)

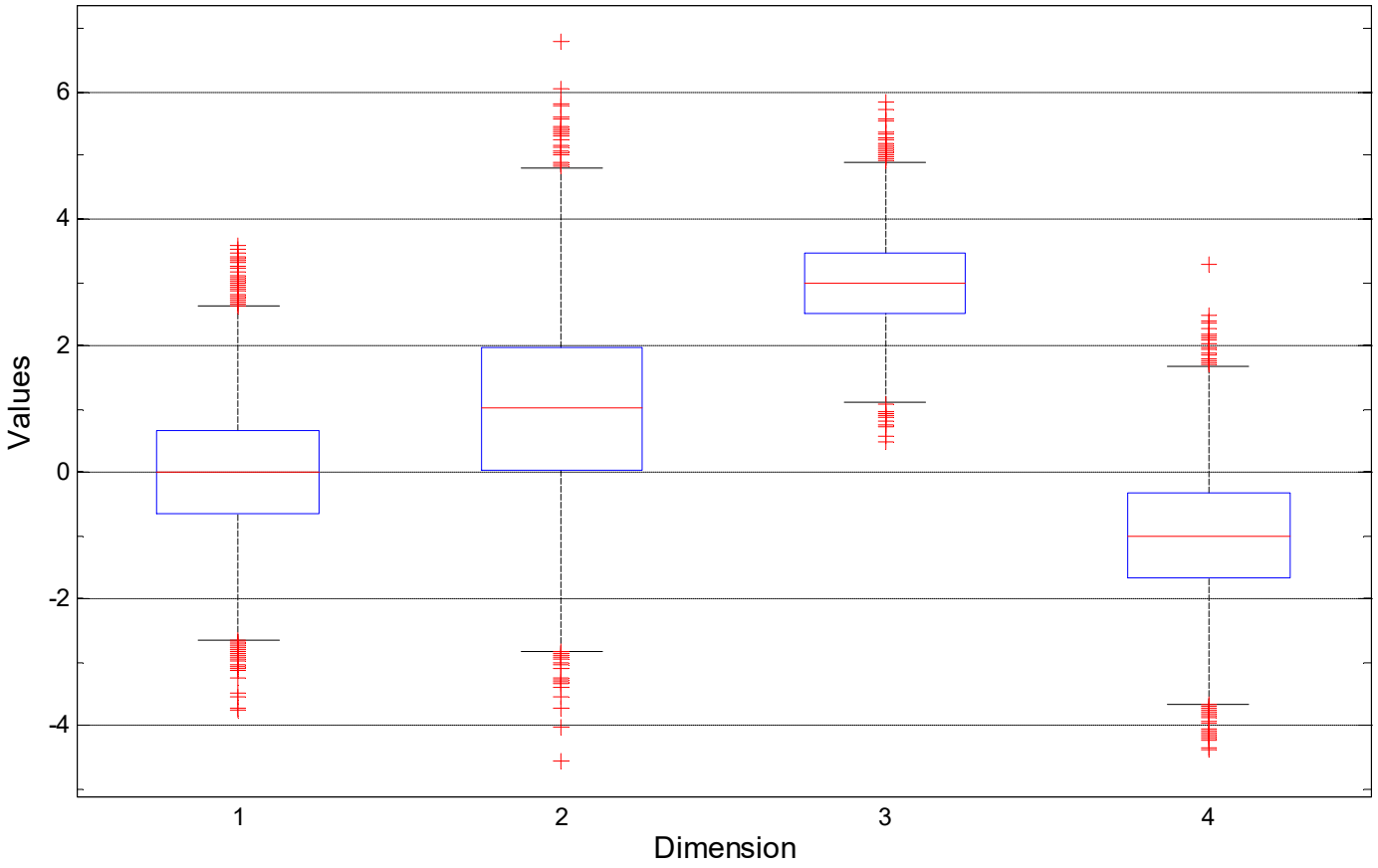


Surf Plots (3D)

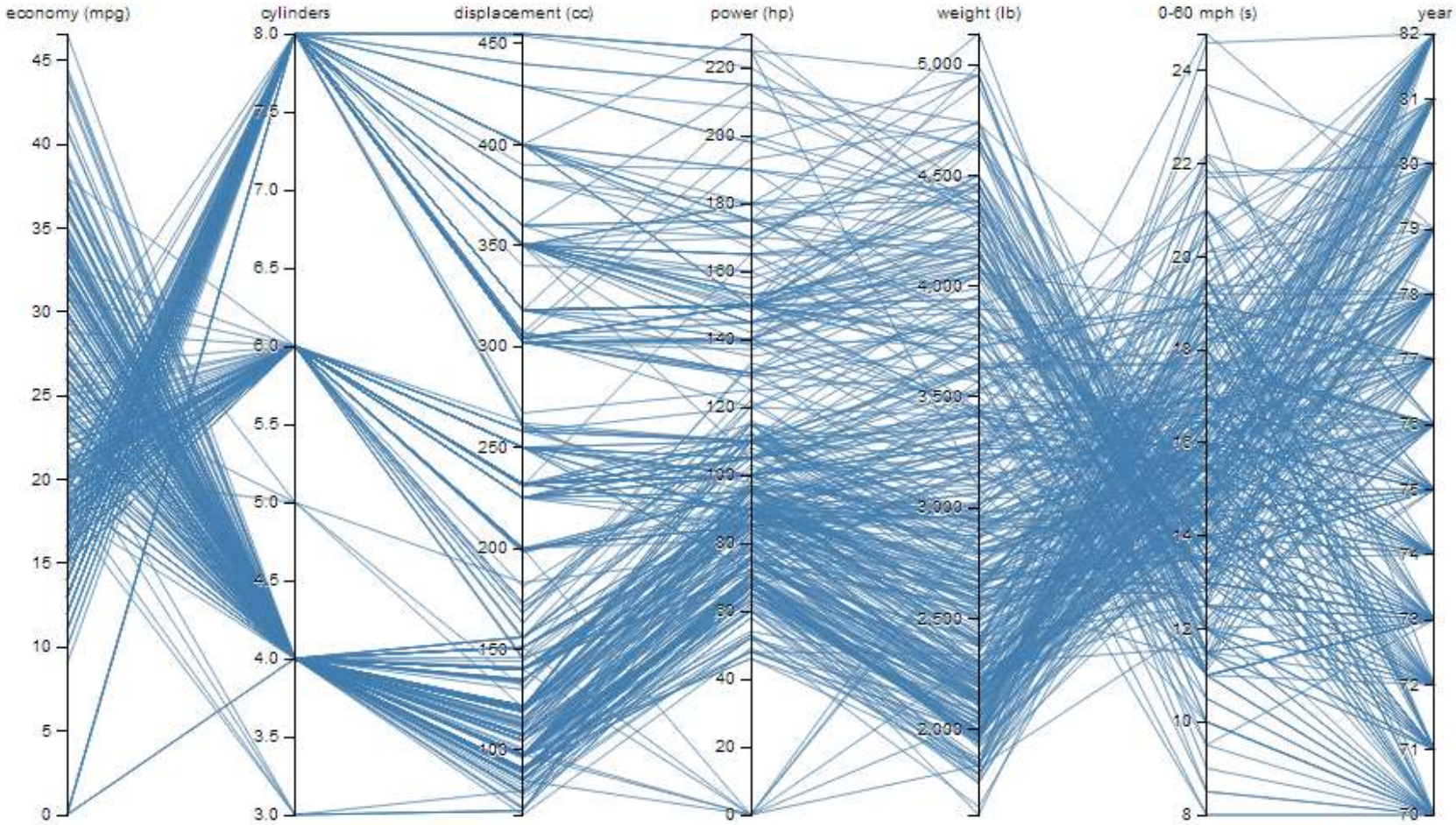




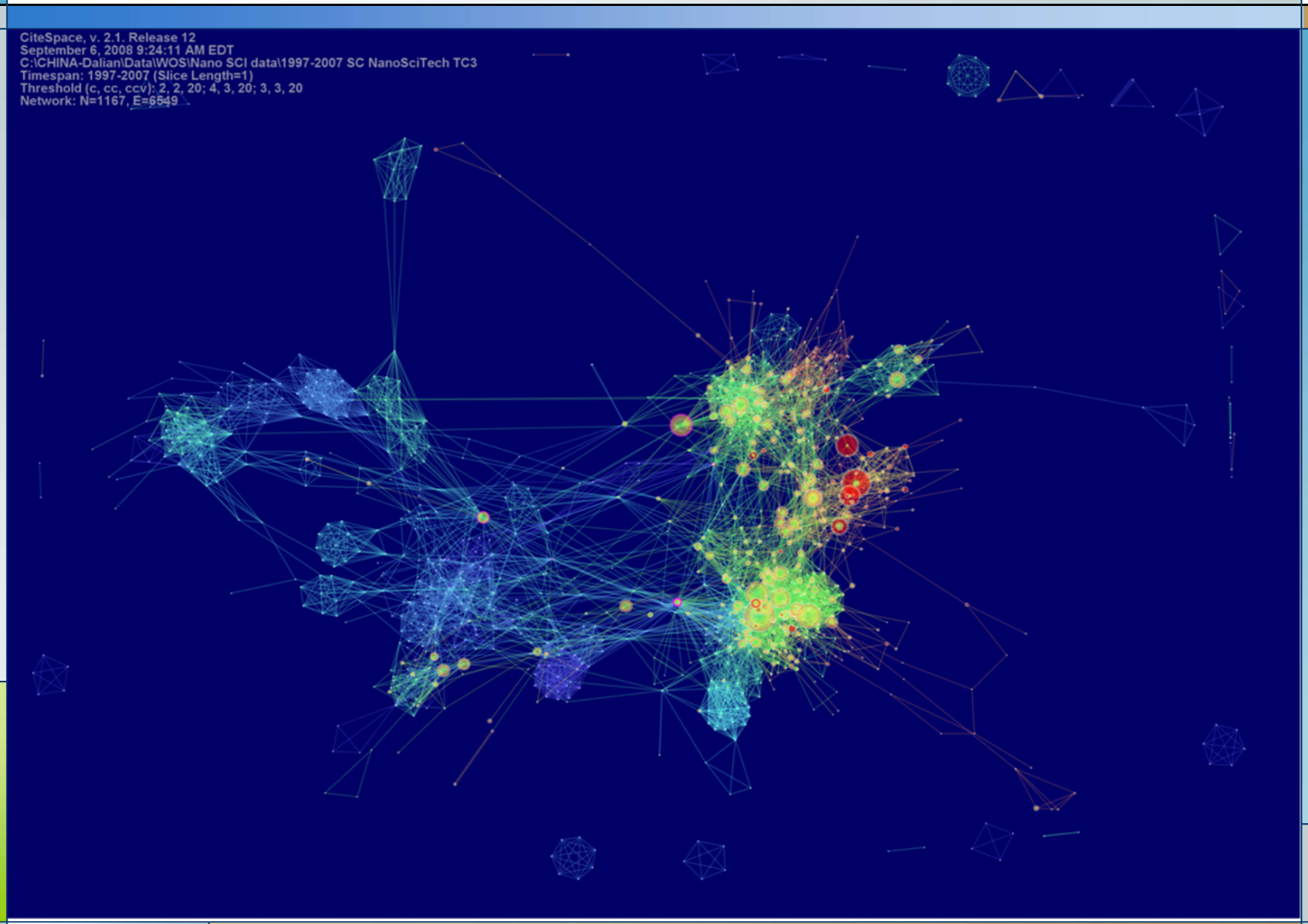
Box Plots (High Dimensional)



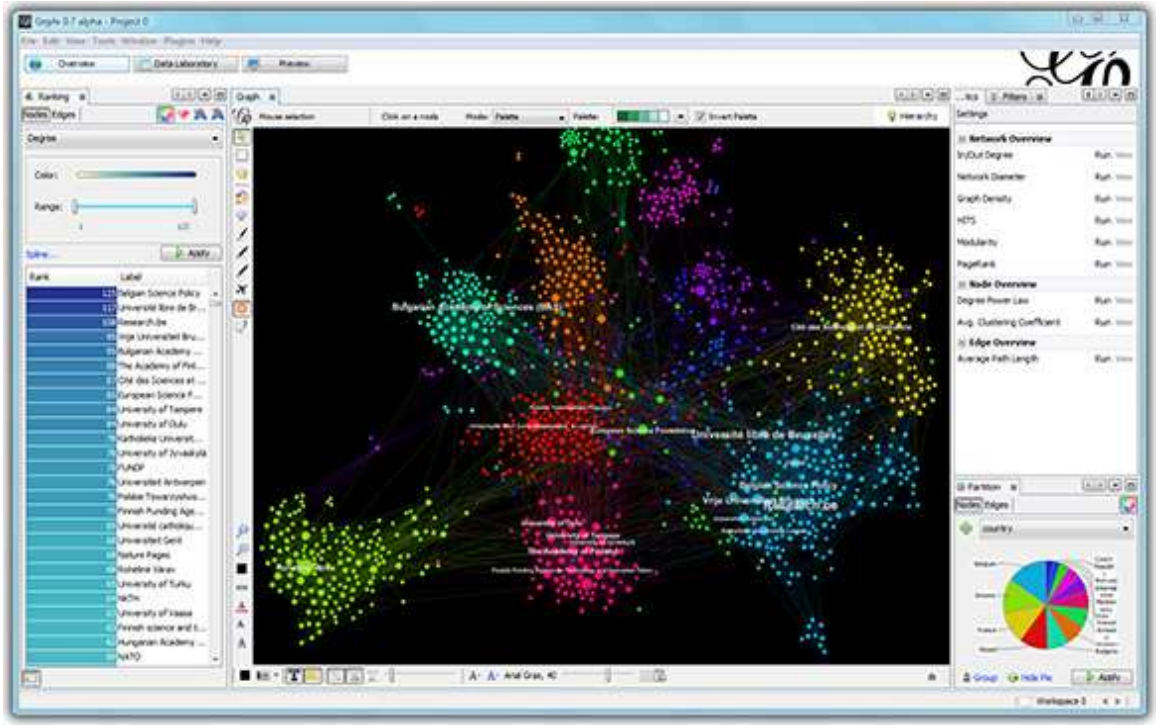
Parallel Coordinates (High Dimensional)

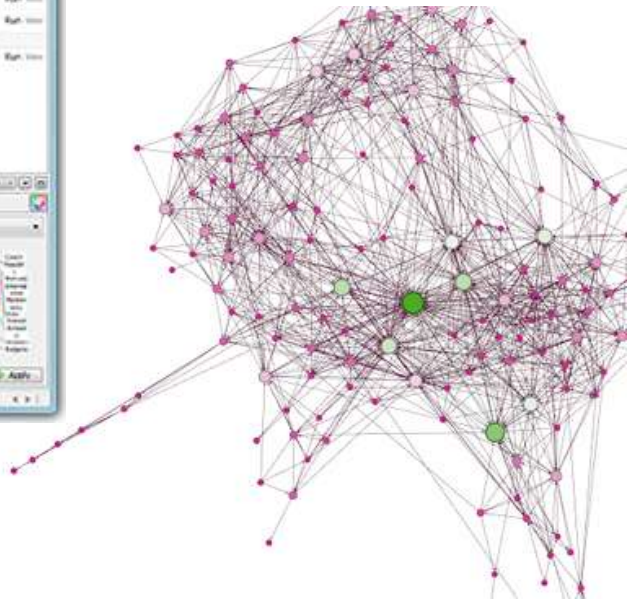


CiteSpace


CiteSpace, v. 2.1. Release 12
September 6, 2008 9:24:11 AM EDT
C:\CHINA-Dalian\Data\WOS\Nano SCI data\1997-2007 SC NanoSciTech TC3
Timespan: 1997-2007 (Slice Length=1)
Threshold (c, cc, ccv): 2, 2, 20; 4, 3, 20; 3, 3, 20
Network: N=1167, E=6549

Gephi





Feature Selection

ID

Weight

Age

Gender

Height

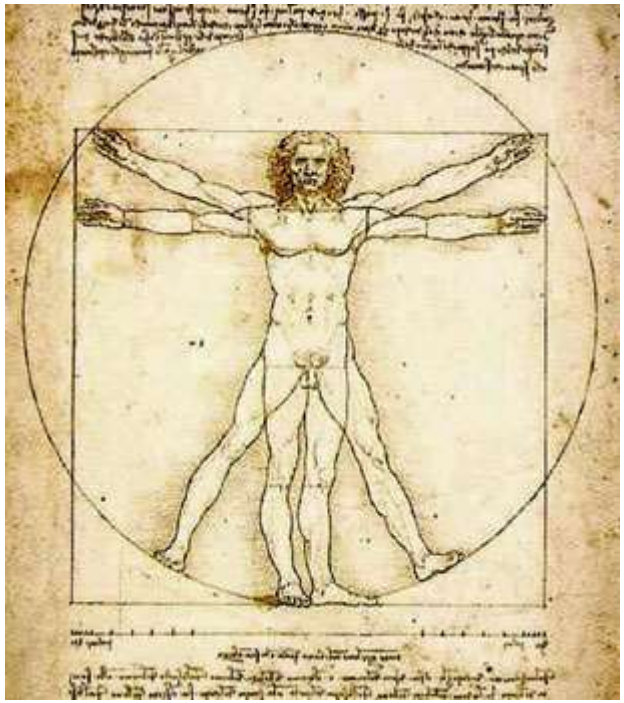
Education

Income

Address

Occupation

Location



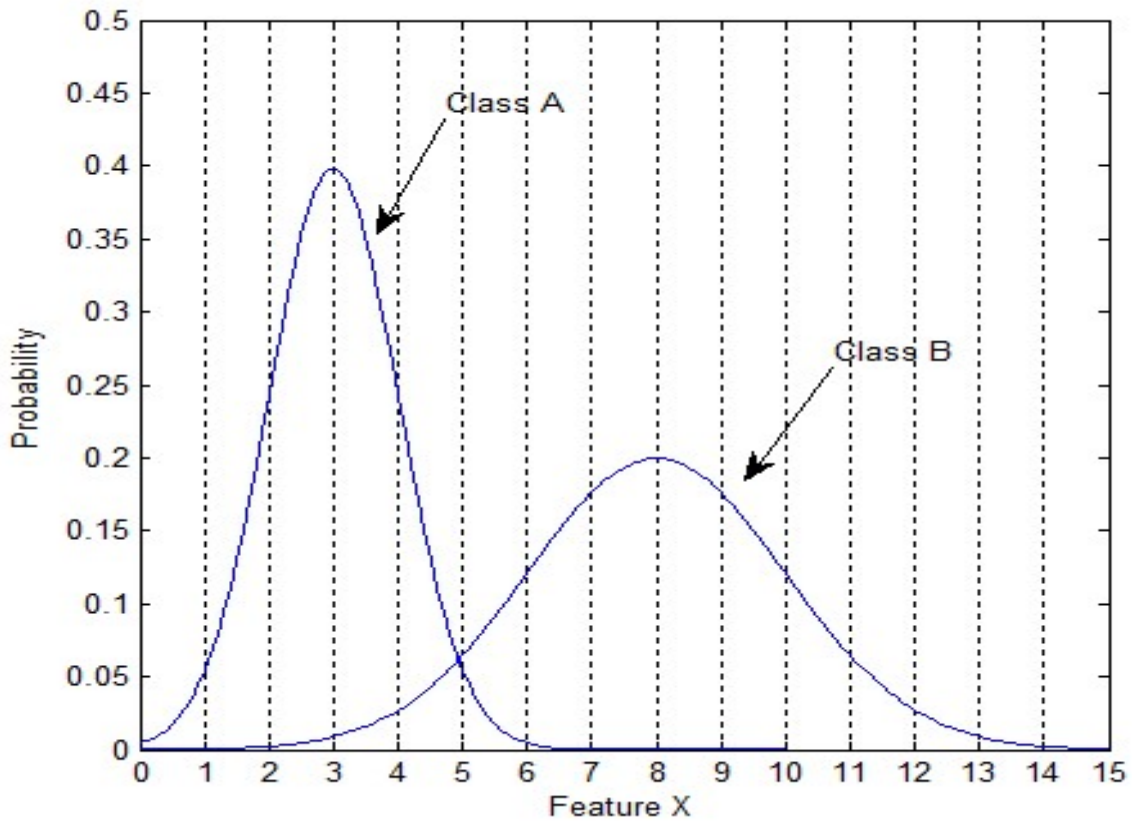
Noisy?

Irrelevant?

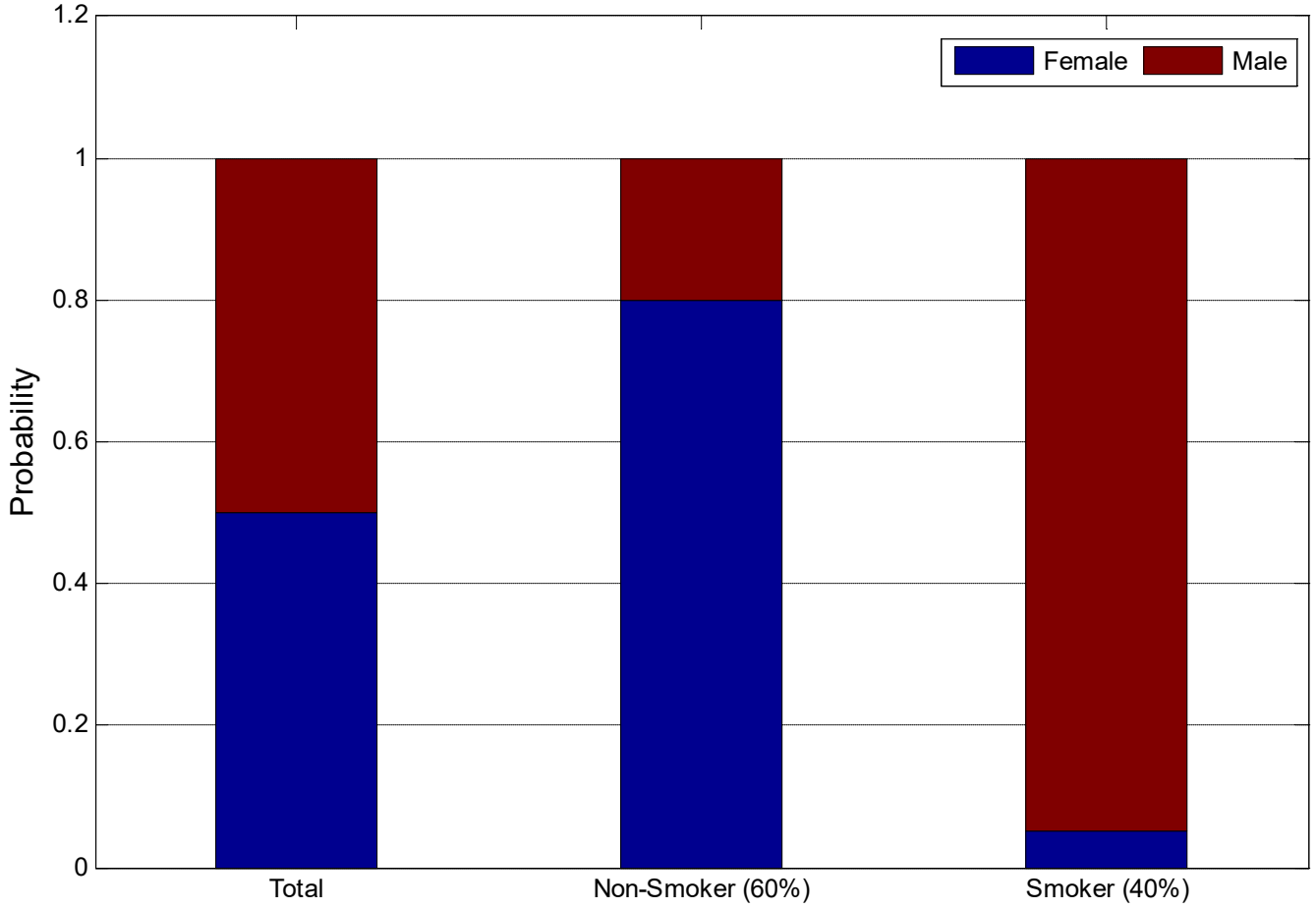
Duplicate?

Complexity?

Class Distributions

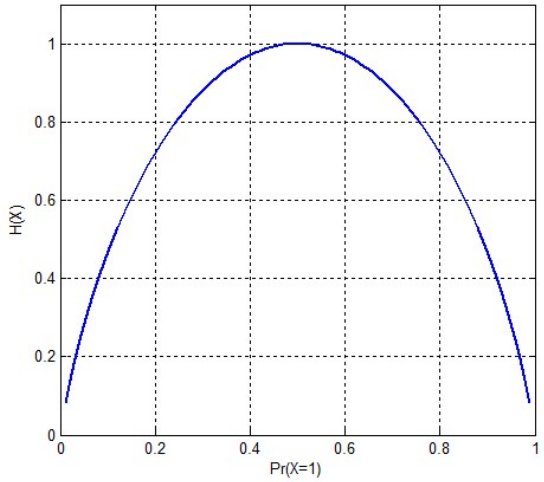


Class Distributions



Entropy

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$



$$H(S) = -0.5 \cdot \log_2 0.5 - 0.5 \cdot \log_2 0.5 = 1.0$$

X: {a = "Non−Smoker"; b = "Smoker" }

$$H(S \mid X = a) = -0.8 \cdot \log_2 0.8 - 0.2 \cdot \log_2 0.2 = 0.7219$$

$$H(S \mid X = b) = -0.05 \cdot \log_2 0.05 - 0.95 \cdot \log_2 0.95 = 0.2864$$

$$H(S \mid X) = 0.6 \cdot H(S \mid X = a) + 0.4 \cdot H(S \mid X = b) = 0.5477$$

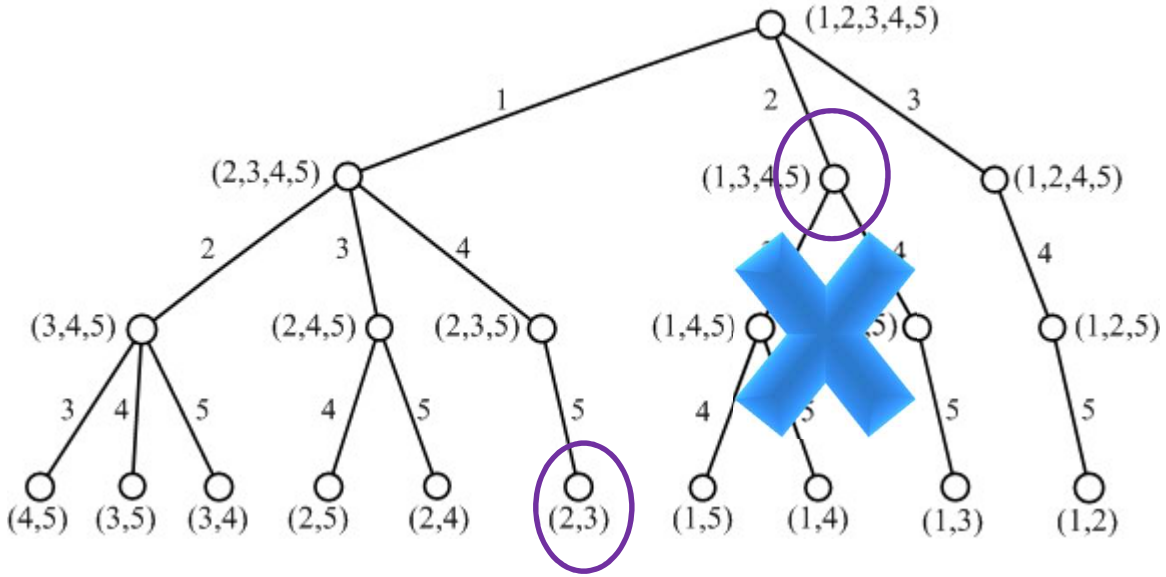
$$Gain(S, X) = H(S) - H(S \mid X) = 0.4523$$ Information Gain

Feature Subset Search

- Exhaustive
  - All possible combinations

$$C_{10}^{3} = \frac{10!}{(10-3)!3!} = 120 \qquad C_{20}^{5} = \frac{20!}{(20-5)!5!} = 15504$$

- Branch and Bound



$$S_1 \supset S_2 \supset S_3 \quad \Rightarrow \quad J(S_1) > J(S_2) > J(S_3)$$

Feature Subset Search

- Top K Individual Features

$$J(X_k) = \{J(x_1), \cdots, J(x_k)\},\ J(x_1) > J(x_2) > \ldots > J(x_k)$$

- Sequential Forward Selection

$$J(X_k + x_1) > J(X_k + x_2) > \cdots > J(X_k + x_{D-k}),\ x_i \notin X_k$$

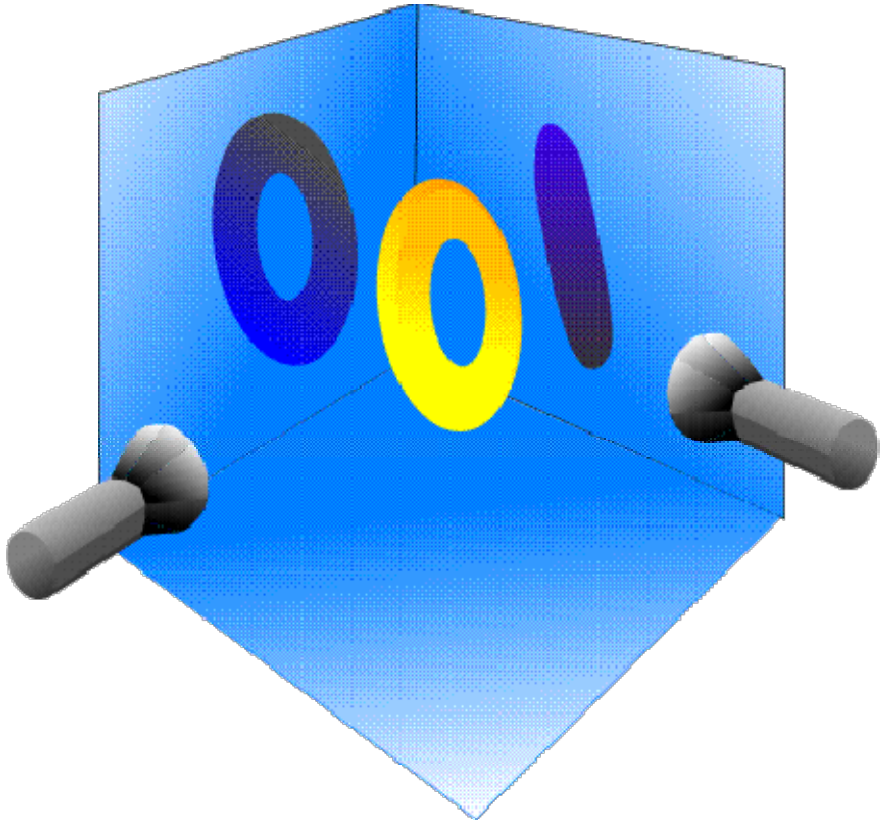
- Sequential Backward Selection

$$J(X_k - x_1) > J(X_k - x_2) > \cdots > J(X_k - x_k),\ x_i \in X_k$$

- Optimization Algorithms
  - Simulated Annealing
  - Tabu Search
  - Genetic Algorithms

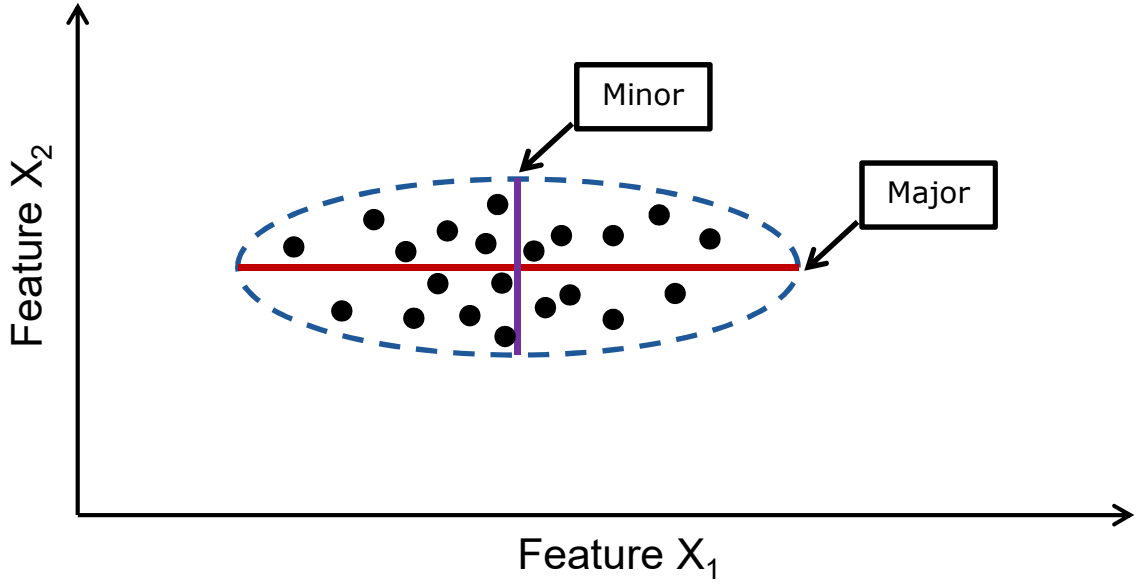
Feature Extraction

Principal Component Analysis

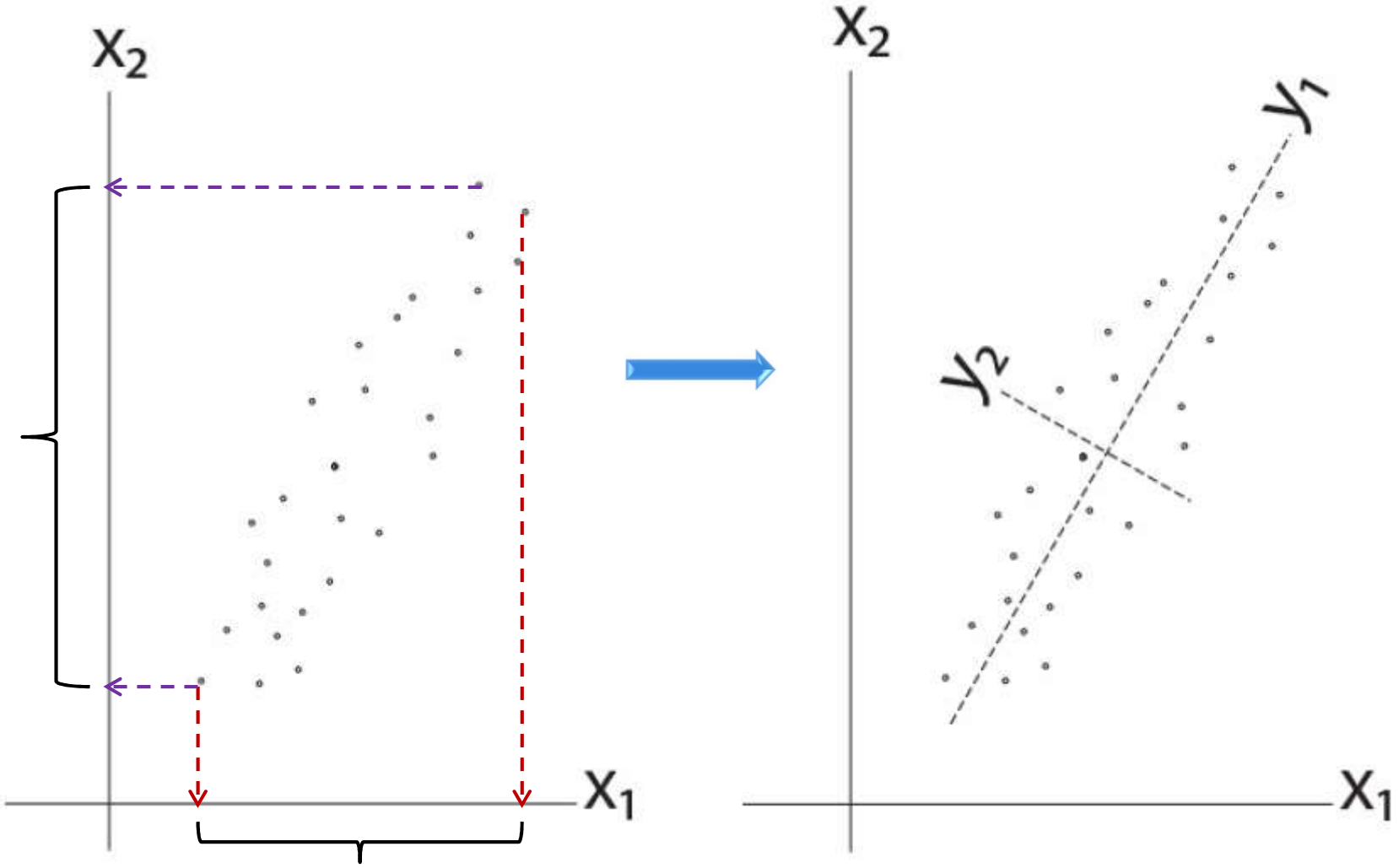


2D Example

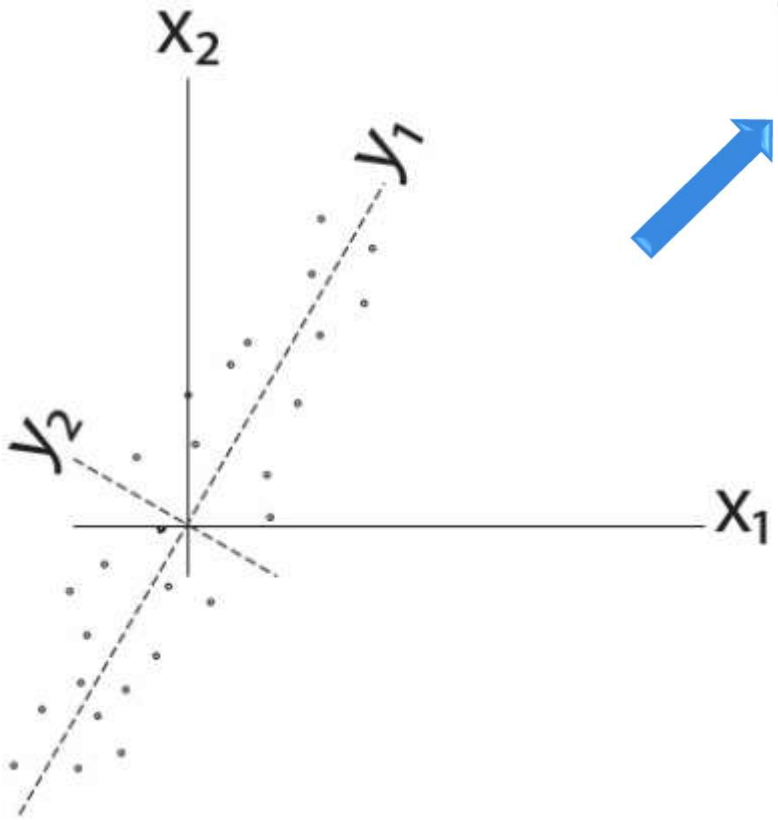
- Data: Gaussian Distribution
- Variance: Information
- Ellipse: Major Axis vs. Minor Axis
- Select the attribute corresponding to the Major Axis.

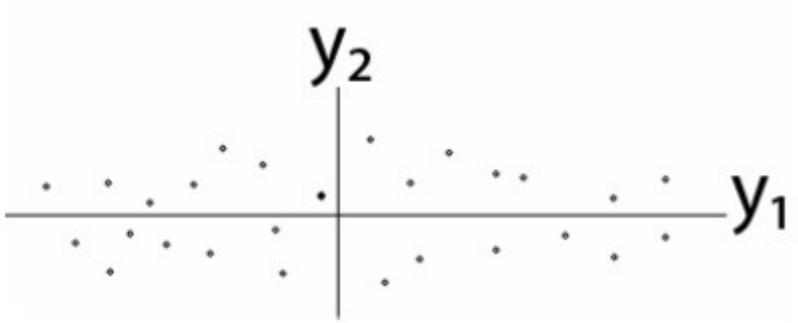


2D Example



2D Example





$$S(X) = \frac{1}{n-1} XX^T$$

Remove correlation

$$S(Y) = \frac{1}{n-1} YY^T$$

Some Math

Goal: S(Y) has nonzero diagonal entries and all off-diagonal elements are zero.

$$Y = PX \rightarrow S(Y) = \frac{1}{n-1} YY^T \rightarrow YY^T = (PX)(PX)^T = PXX^TP^T.$$

$$XX^T = QDQ^T \leftarrow (n-1)S(Y) = PXX^TP^T = PQDQ^TP^T = (PQ)D(PQ)^T \leftarrow P = Q^T$$

Eigendecomposition

A Different View

$$J(e) = \sum_{i=1}^{n} \left\| x_k{}' - x_k \right\|^2$$

$$= \sum_{i=1}^{n} \|\alpha_k e - x_k\|^2$$

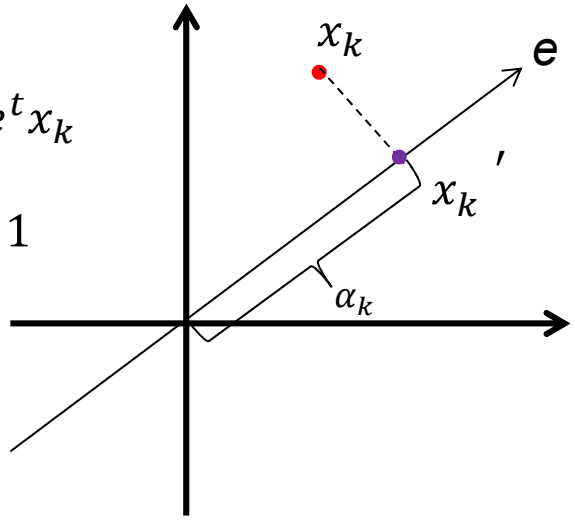
$$= \sum_{i=1}^{n} \alpha_k{}^2 \|e\|^2 - 2\sum_{i=1}^{n} \alpha_k e^t x_k + \sum_{i=1}^{n} \|x_k\|^2$$

$$= -\sum_{i=1}^{n} \alpha_k{}^2 + \sum_{i=1}^{n} \|x_k\|^2$$

$$= -\sum_{i=1}^{n} e^t x_k x_k{}^t e + \sum_{i=1}^{n} \|x_k\|^2$$

$$\alpha_k = e^t x_k$$

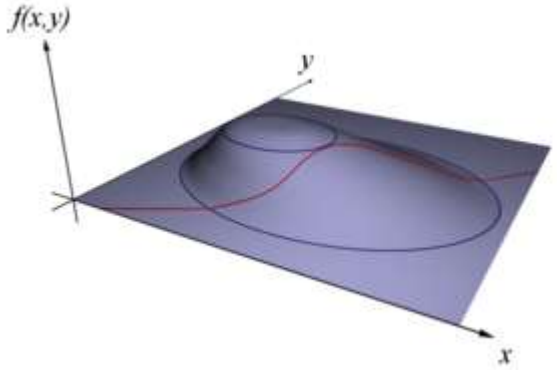
$$\|e\| = 1$$

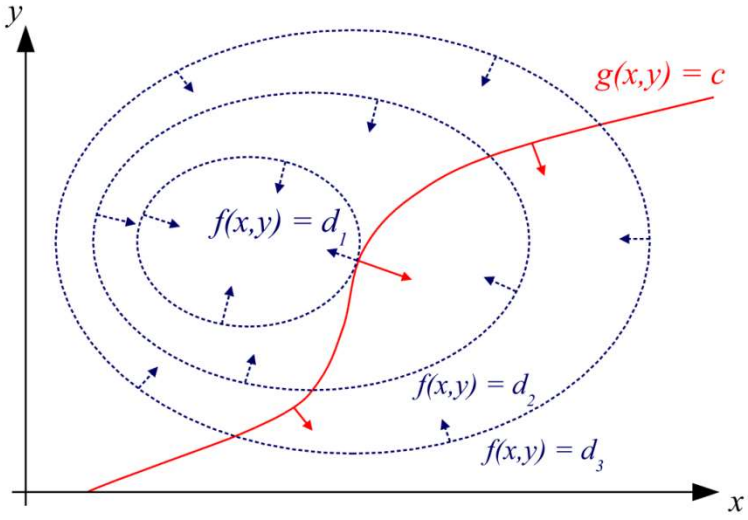


$$\max_e e^t S e \quad s.t. \ \|e\|=1$$

$$S = \sum_{i=1}^{n} x_k x_k{}^t$$

Lagrange Multipliers





$$\max_{x,y} f(x,y) = 3xy \quad s.t.\ 2x + y = 8$$

$$F(x,y,\lambda) = 3xy - \lambda(2x + y - 8)$$

$$F_x = 3y - 2\lambda$$
$$F_y = 3x - \lambda$$
$$F_\lambda = -(2x + y - 8)$$

$$\lambda = 6$$
$$x = \frac{\lambda}{3} = 2$$
$$y = \frac{2}{3}\lambda = 4$$
$$f(2,4) = 3 \cdot 2 \cdot 4 = 24$$

More Math …

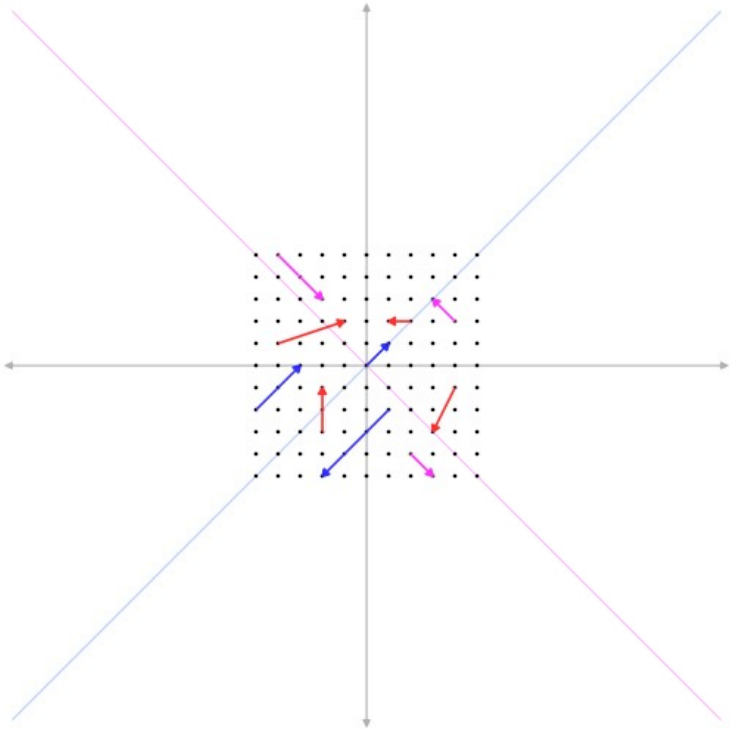
$$u = e^t Se - \lambda\left(e^t e - 1\right)$$

$$\frac{\partial u}{\partial e} = 2Se - 2\lambda e$$

$Se = \lambda e$ ← eigenvector

↑ eigenvalue

$$e^t Se = \lambda e^t e = \lambda$$



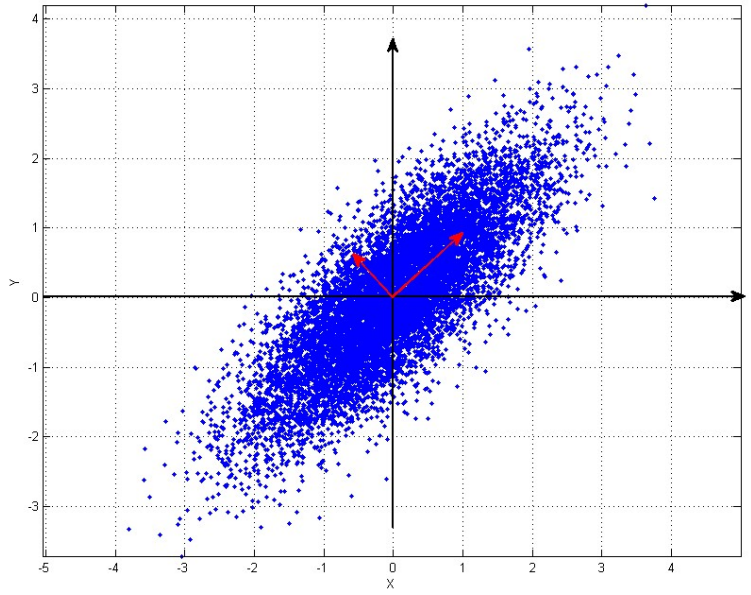
PCA

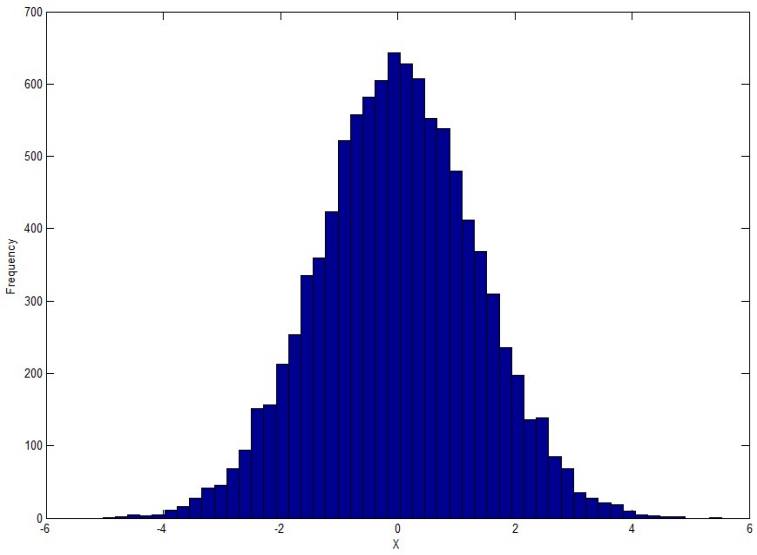
To project the original data to the eigenvectors of S with the largest eigenvalues.

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

PCA Examples





$$s = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix};$$

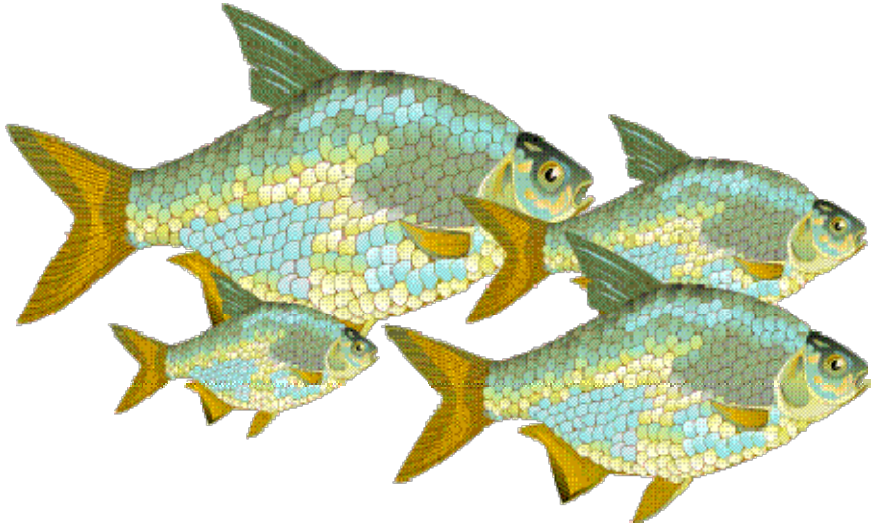
```
x = mvnrnd(zeros(10000,2), s);
plot(x(:, 1), x(:, 2), '.');
axis equal;
grid on;
[V, D] = eig(s);
```

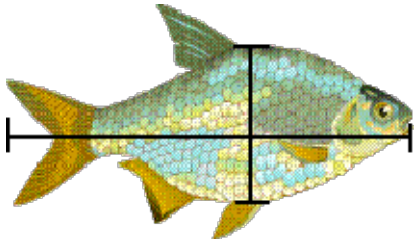
$$V = \begin{bmatrix} -0.7071 & 0.7071 \\ 0.7071 & 0.7071 \end{bmatrix};$$

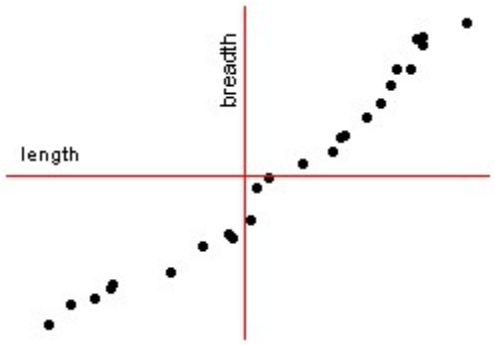
$$D = \begin{bmatrix} 0.2 & 0 \\ 0 & 1.8 \end{bmatrix};$$

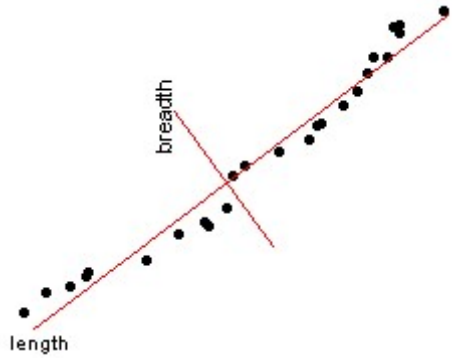
```
newx = x * V(:,2);
hist(newx, 50);
```

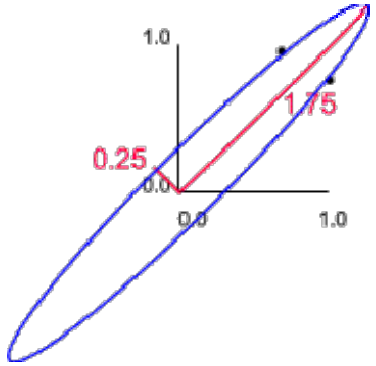
Fish



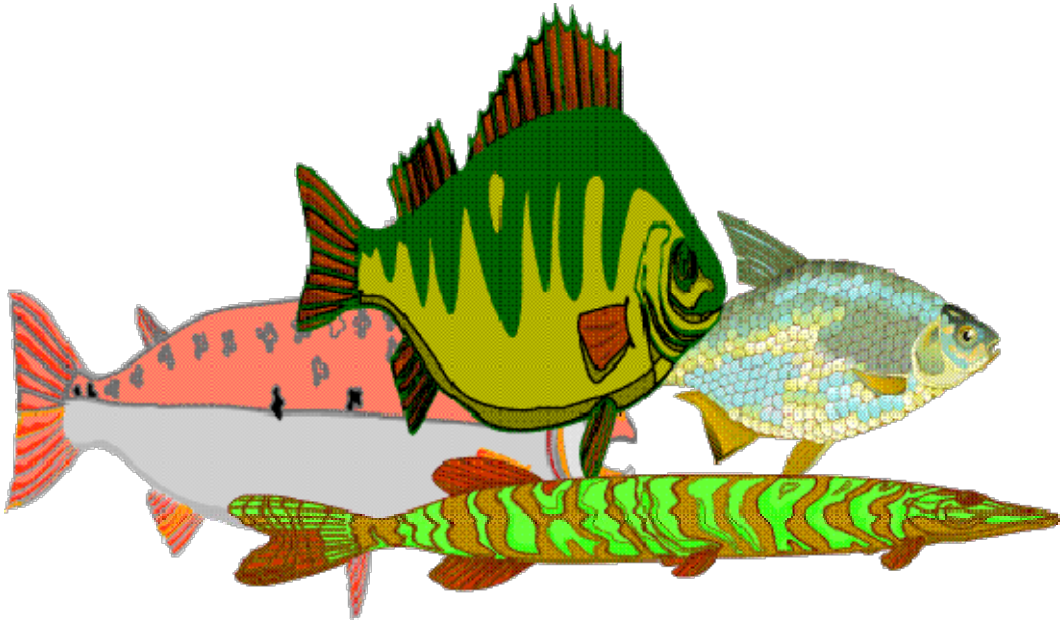


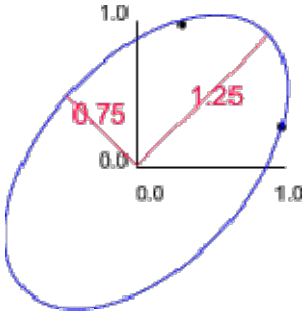




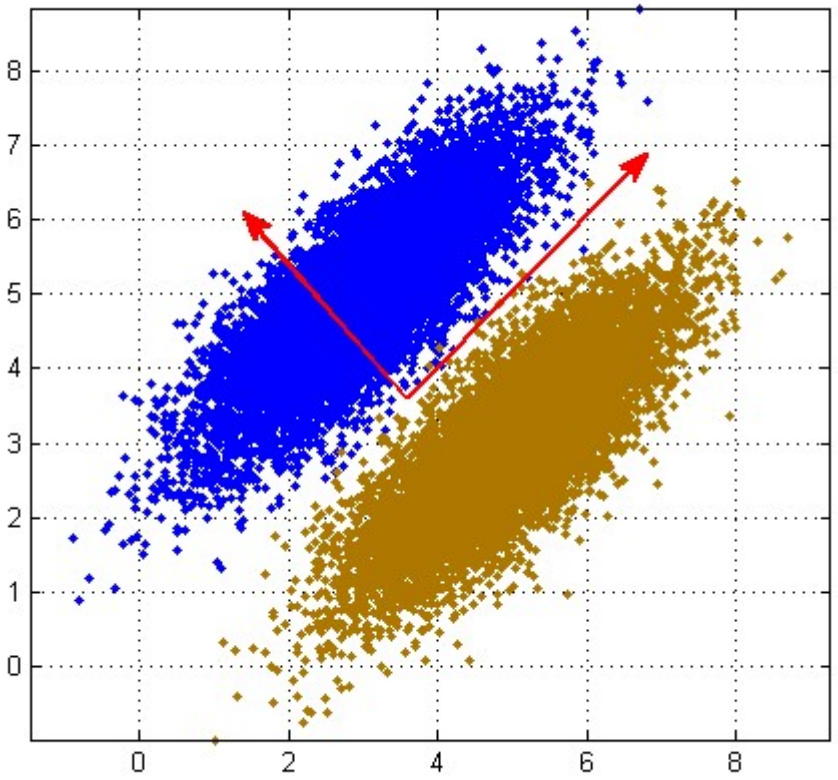


Fish





The Issue of PCA

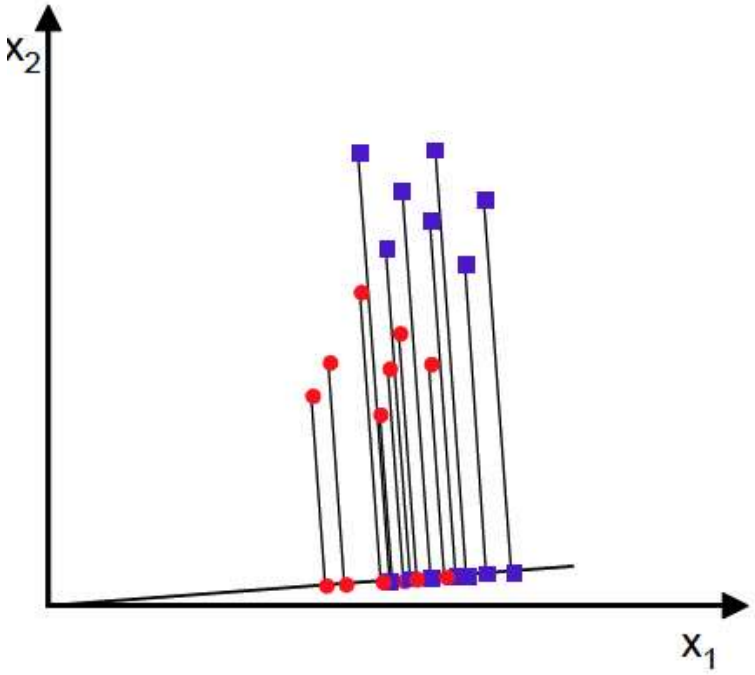


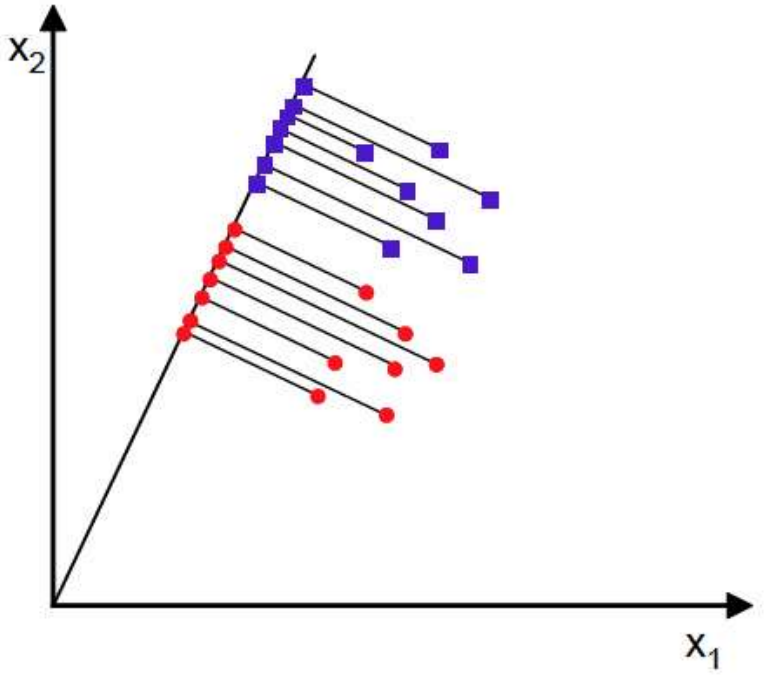
Now, let's consider class information ...

Linear Discriminant Analysis

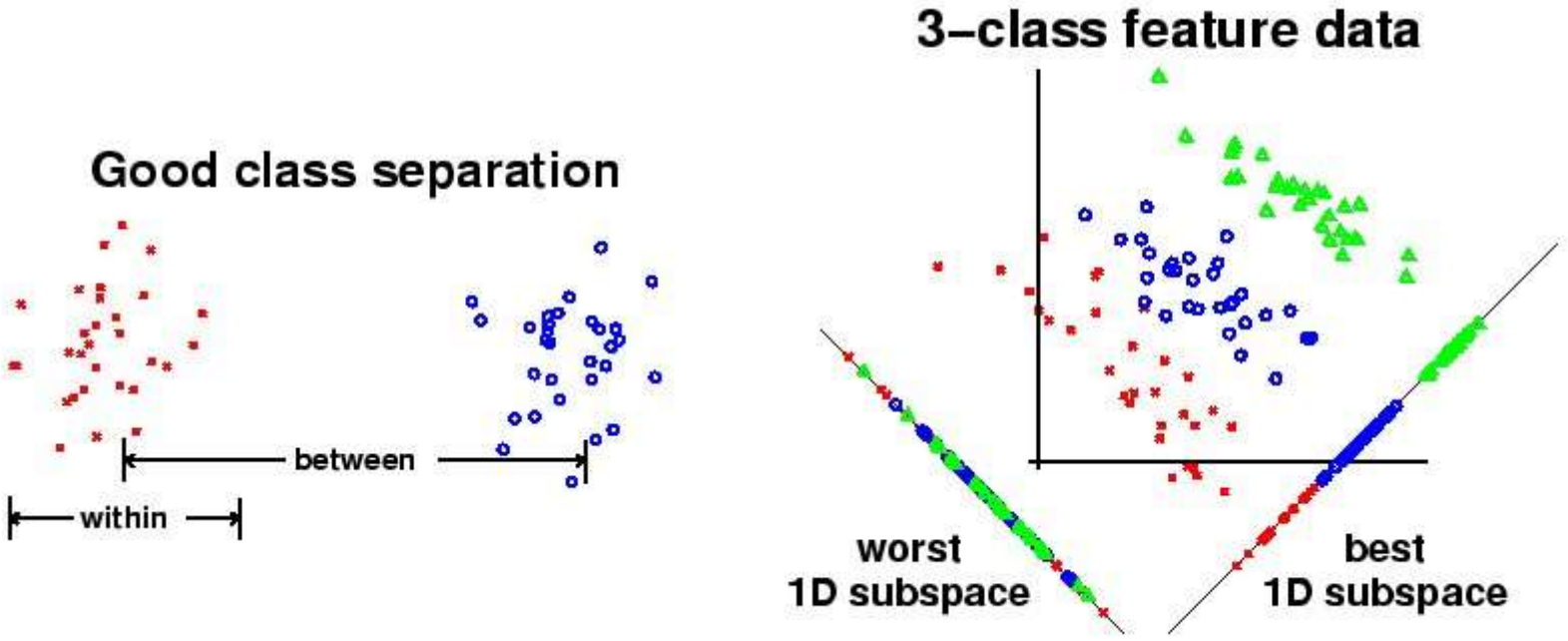
- The objective of LDA is to perform dimension reduction while preserving as much of the class discriminatory information as possible.

- Given a set of d-D vectors x_1, x_2, ..., x_n of which N_1 belong to ω_1, and N_2 to ω_2. Of all the possible projection lines $y=w^Tx$, find the one that maximizes the separability.

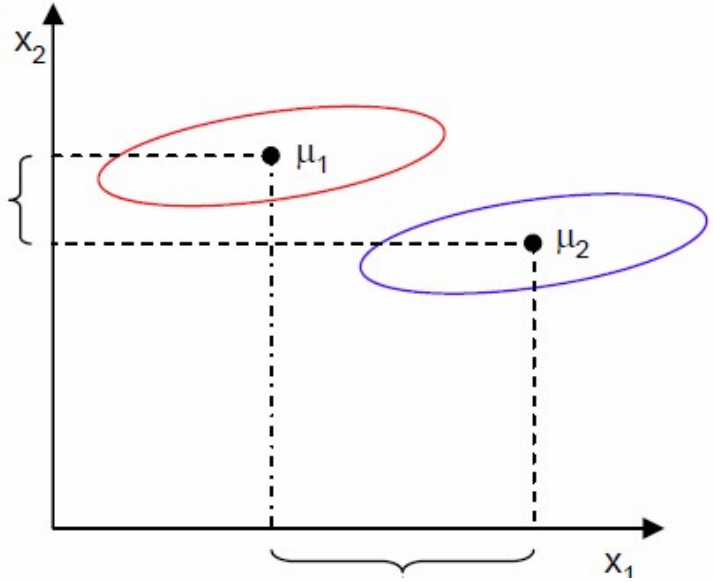


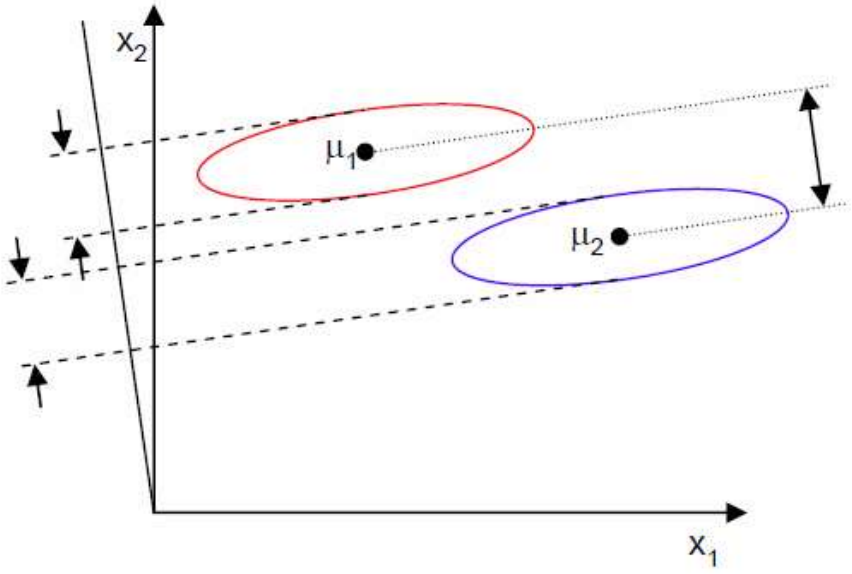


Measure of Separability



Fisher Criterion





$$J = \frac{\left|\mu_1 - \mu_2\right|^2}{S_1^2 + S_2^2}$$

⟸ Maximize the distance between classes

⟸ Minimize the scatter within each class

Some Math …

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \quad \text{and} \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x = w^T\mu_i$$

between-class scatter

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T\mu_1 - w^T\mu_2)^2 = w^T\underbrace{(\mu_1-\mu_2)(\mu_1-\mu_2)^T}_{S_B}w = w^T S_B w$$

$$S_i = \sum_{x\in\omega_i}(x-\mu_i)(x-\mu_i)^T$$

$$S_1 + S_2 = S_W$$

within-class scatter

$$J(w) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \quad \Rightarrow \quad J(w) = \frac{w^T S_B w}{w^T S_W w}$$

$$\tilde{s}_i^2 = \sum_{y\in\omega_i}(y-\tilde{\mu}_i)^2 = \sum_{x\in\omega_i}(w^T x - w^T\mu_i)^2 = \sum_{x\in\omega_i} w^T(x-\mu_i)(x-\mu_i)^T w = w^T S_i w$$

$$\tilde{s}_1^2 + \tilde{s}_2^2 = w^T S_W w$$

More Math…

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$ ← generalized Rayleigh quotient

$$\frac{d}{dw}[J(w)] = \frac{d}{dw}\left[\frac{w^T S_B w}{w^T S_W w}\right] = 0 \Rightarrow$$

$$\Rightarrow [w^T S_W w]\frac{d[w^T S_B w]}{dw} - [w^T S_B w]\frac{d[w^T S_W w]}{dw} = 0 \Rightarrow$$

$$\Rightarrow [w^T S_W w]2S_B w - [w^T S_B w]2S_W w = 0$$

$$\frac{[w^T S_W w]}{[w^T S_W w]} S_B w - \frac{[w^T S_B w]}{[w^T S_W w]} S_W w = 0 \Rightarrow$$

$$\Rightarrow S_B w - J S_W w = 0 \Rightarrow$$

$$\Rightarrow S_W^{-1} S_B w - J w = 0$$

More Math...

$$S_W^{-1} S_B w = Jw$$ ← eigenvector problem!

$$S_B w = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = (\mu_1 - \mu_2)R$$

$$R = (\mu_1 - \mu_2)^T w$$ ← scalar

$$Jw = S_w^{-1}(S_B w) = S_w^{-1}(\mu_1 - \mu_2)R$$

$$w = \frac{R}{J} S_w^{-1}(\mu_1 - \mu_2)$$

$$w^* = \underset{w}{\text{argmax}} \left\{ \frac{w^T S_B w}{w^T S_W w} \right\} = S_W^{-1}(\mu_1 - \mu_2)$$

LDA Example

- Dataset
  - $C1 = [(1,2); (2,3); (3,3); (4,5); (5,5)]$
  - $C2 = [(1,0); (2,1); (3,1); (3,2); (5,3); (6,5)]$
- Covariance of [c1; c2]
  - $\mathrm{Z} = \begin{bmatrix} 2.7636 & 2.2545 \\ 2.2545 & 3.0182 \end{bmatrix}$
- Eigenvectors and Eigenvalues of Z
  - $V = \begin{bmatrix} -0.7268 & 0.6869 \\ 0.6869 & 0.7268 \end{bmatrix}$
  - $D = \begin{bmatrix} 0.6328 & 0 \\ 0 & 5.1490 \end{bmatrix}$
- The direction of PCA projection: $[0.6869, 0.7268]^T$

LDA Example

- The mean of each class
  - $\mu_1 = mean(c1) = [3.0, 3.6]^T$
  - $\mu_2 = mean(c2) = [3.3, 2.0]^T$

- $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \begin{bmatrix} 0.11 & -0.53 \\ -0.53 & 2.56 \end{bmatrix}$

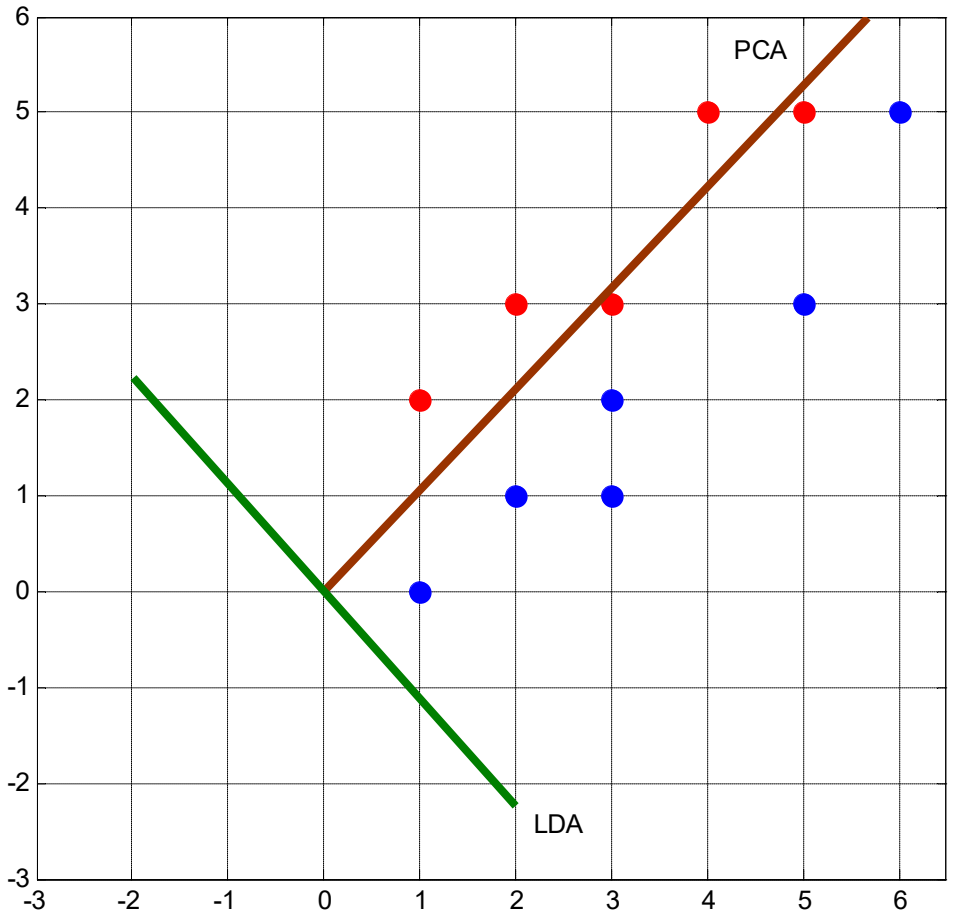
- The scatter of each class
  - $S_1 = 4 \times cov(c1) = \begin{bmatrix} 10 & 8 \\ 8 & 7.2 \end{bmatrix}$
  - $S_2 = 5 \times cov(c2) = \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix}$

- $S_w = S_1 + S_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$

LDA Example

- $S_w{}^{-1}S_B = \begin{bmatrix} 0.26 & -1.27 \\ -0.30 & 1.42 \end{bmatrix}$
- Eigenvectors and Eigenvalues of $S_w{}^{-1}S_B$
  - $V = \begin{bmatrix} -0.98 & 0.67 \\ -0.20 & -0.75 \end{bmatrix}$
  - $D = \begin{bmatrix} 0 & 0 \\ 0 & 1.69 \end{bmatrix}$
- The direction of LDA projection: $[0.6656, -0.7463]^T$
- Alternatively
  - $S_w{}^{-1}(\mu_1 - \mu_2)^T = [-0.7936, 0.8899]^T$
  - After normalization: $[-0.6656, 0.7463]^T$

LDA Example



C-Class LDA

- **Fisher's LDA generalizes very gracefully for C-class problems**
  - Instead of one projection **y**, we will now seek (C-1) projections $[y_1,y_2,\ldots,y_{C-1}]$ by means of (C-1) projection vectors w_i, which can be arranged by columns into a projection matrix $W=[w_1|w_2|\ldots|w_{C-1}]$:

$$y_i = w_i^T x \Rightarrow y = W^T x$$

- **Derivation**
  - The generalization of the within-class scatter is

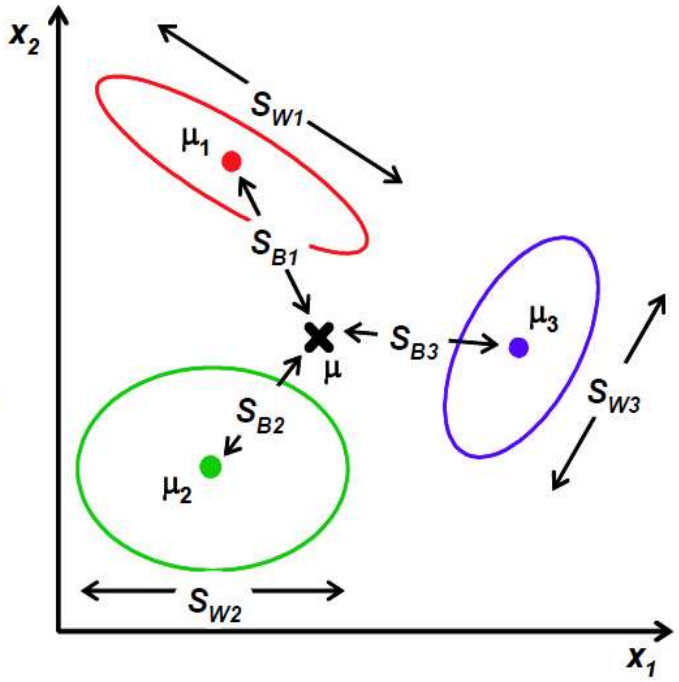
$$S_W = \sum_{i=1}^{C} S_i$$

$$\text{where } S_i = \sum_{x\in\omega_i} (x-\mu_i)(x-\mu_i)^T \text{ and } \mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x$$

  - The generalization for the between-class scatter is

$$S_B = \sum_{i=1}^{C} N_i(\mu_i-\mu)(\mu_i-\mu)^T$$

$$\text{where } \mu = \frac{1}{N}\sum_{\forall x} x = \frac{1}{N}\sum_{x\in\omega_i} N_i\mu_i$$



C-Class LDA

- For C-class LDA with C=2, S_B is defined as:

$$S_B = N_1(\mu_1 - \mu)(\mu_1 - \mu)^T + N_2(\mu_2 - \mu)(\mu_2 - \mu)^T$$

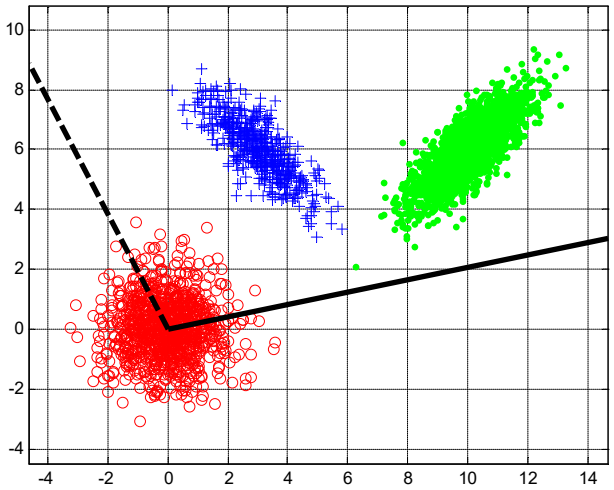
$$= N_1\left(\mu_1 - \frac{N_1\mu_1 + N_2\mu_2}{N}\right)\left(\mu_1 - \frac{N_1\mu_1 + N_2\mu_2}{N}\right)^T + N_2\left(\mu_2 - \frac{N_1\mu_1 + N_2\mu_2}{N}\right)\left(\mu_2 - \frac{N_1\mu_1 + N_2\mu_2}{N}\right)^T$$

$$= N_1\left(\frac{N_2\mu_1 - N_2\mu_2}{N}\right)\left(\frac{N_2\mu_1 - N_2\mu_2}{N}\right)^T + N_2\left(\frac{N_1\mu_2 - N_1\mu_1}{N}\right)\left(\frac{N_1\mu_2 - N_1\mu_1}{N}\right)^T$$

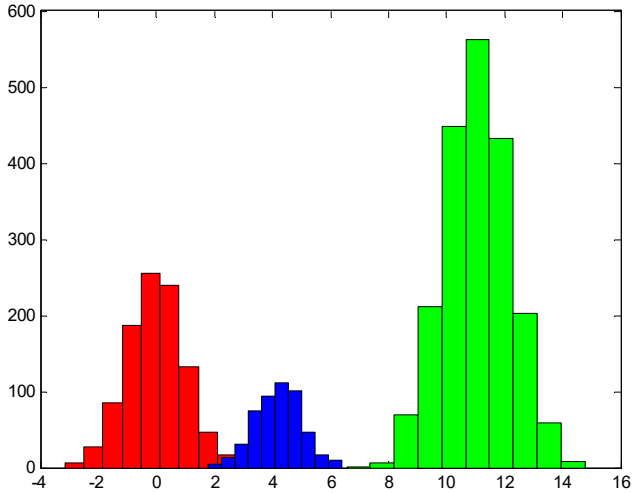
$$= \frac{N_1N_2^2}{N^2}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T + \frac{N_1^2N_2}{N^2}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$$

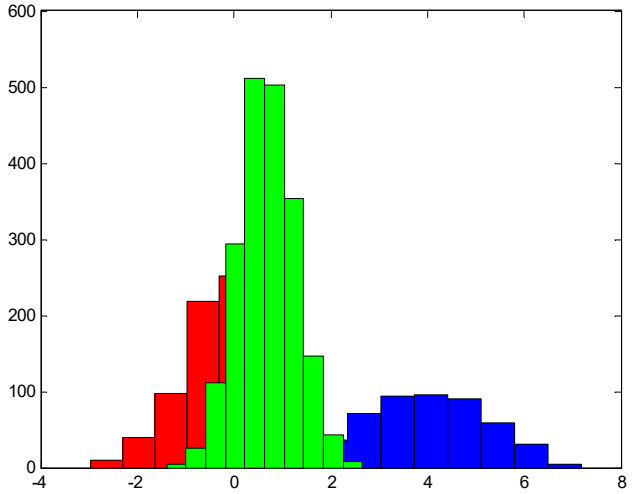
$$= \frac{N_1N_2}{N}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

C-Class LDA



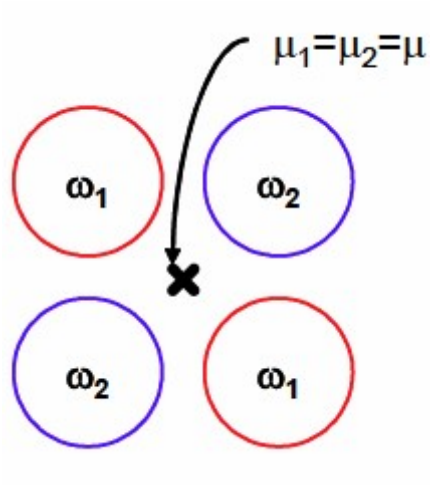

```
N1=1000; N2=500; N3=2000;
X1=mvnrnd([0,0], [1,0;0,1], N1);
X2=mvnrnd([3,6], [1,-0.8;-0.8,1], N2);
X3=mvnrnd([10,6], [1,0.8;0.8,1], N3);
S1=(N1-1)*cov(X1); S2=(N2-1)*cov(X2); S3=(N3-1)*cov(X3);
Sw=S1+S2+S3;
M1=mean(X1); M2=mean(X2); M3=mean(X3);
Mu=(N1*M1+N2*M2+N3*M3)/(N1+N2+N3);
Sb=N1*(M1-Mu)'*(M1-Mu)+N2*(M2-Mu)'*(M2-Mu) +N3*(M3-Mu)'*(M3-Mu);
J=inv(Sw)*Sb;  [V,D]=eig(J);
```

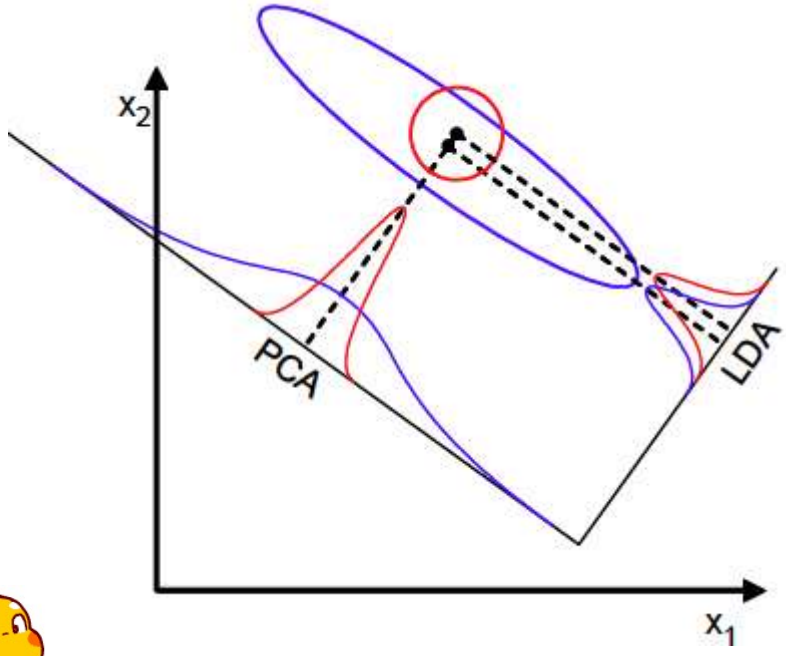


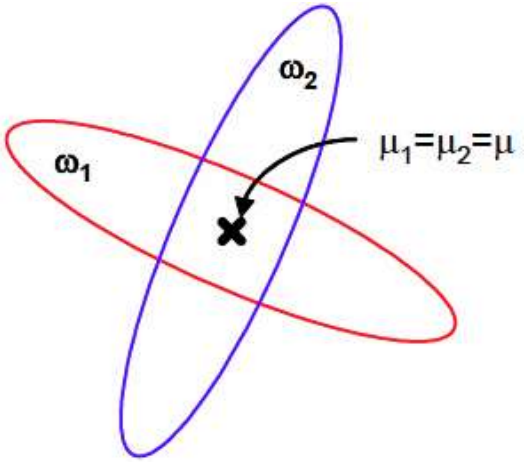


Limitations of LDA

- LDA produces at most C-1 projections
  - S_B is a matrix with rank C-1 or less.
- S_W may be singular.
- LDA does not work well when ...







Reading Materials

- M. A. Hernandez and S. J. Stolfo, "Real-World Data is Dirty: Data Cleansing and The Merge/Purge Problem," Data Mining and Knowledge Discovery, vol. 2, pp. 9–37, 1998.
- A. Donders, G. van der Heijden, T. Stijnen, and K. Moons, "Review: A Gentle Introduction to Imputation of Missing Values," Journal of Clinical Epidemiology, vol. 59, pp. 1087-1091, 2006.
- N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
- N. Japkowicz and S. Stephen, "The Class Imbalance Problem: A Systematic Study," Intelligent Data Analysis, vol. 6, pp. 429–449, 2002.
- D. Keim, "Information Visualization and Visual Data Mining," IEEE Transactions on Visualization and Computer Graphics, vol. 8, pp. 1-8, 2002.
- PCA Tutorials
  - http://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf
  - http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- Lagrange Multipliers
  - http://diglib.stanford.edu:8091/~klein/lagrange-multipliers.pdf