

Medical Information Management & Mining

You Chen

Jan,15, 2013

You.chen@vanderbilt.edu

Trees

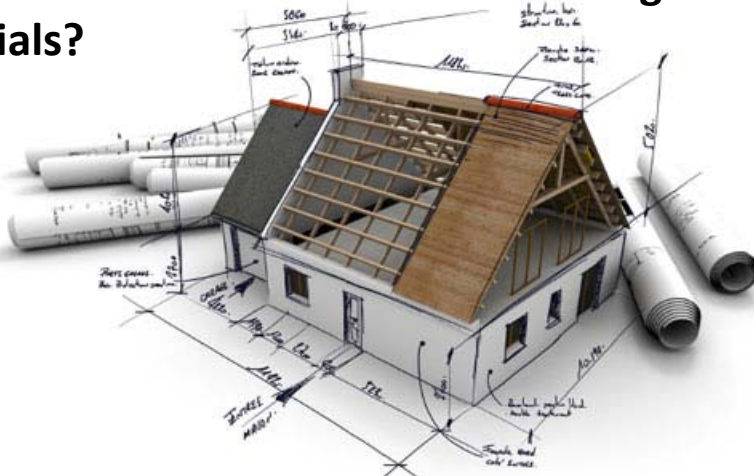


Trees cannot be used to build a house directly.
How can we transform trees to building materials?

Building Materials



Designing Houses



What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization
- Similarity Measurements
- Dimensionality Reduction

Data Representation

The real world



Information in the real world
are abstract and cannot be
directly modeled

Representation



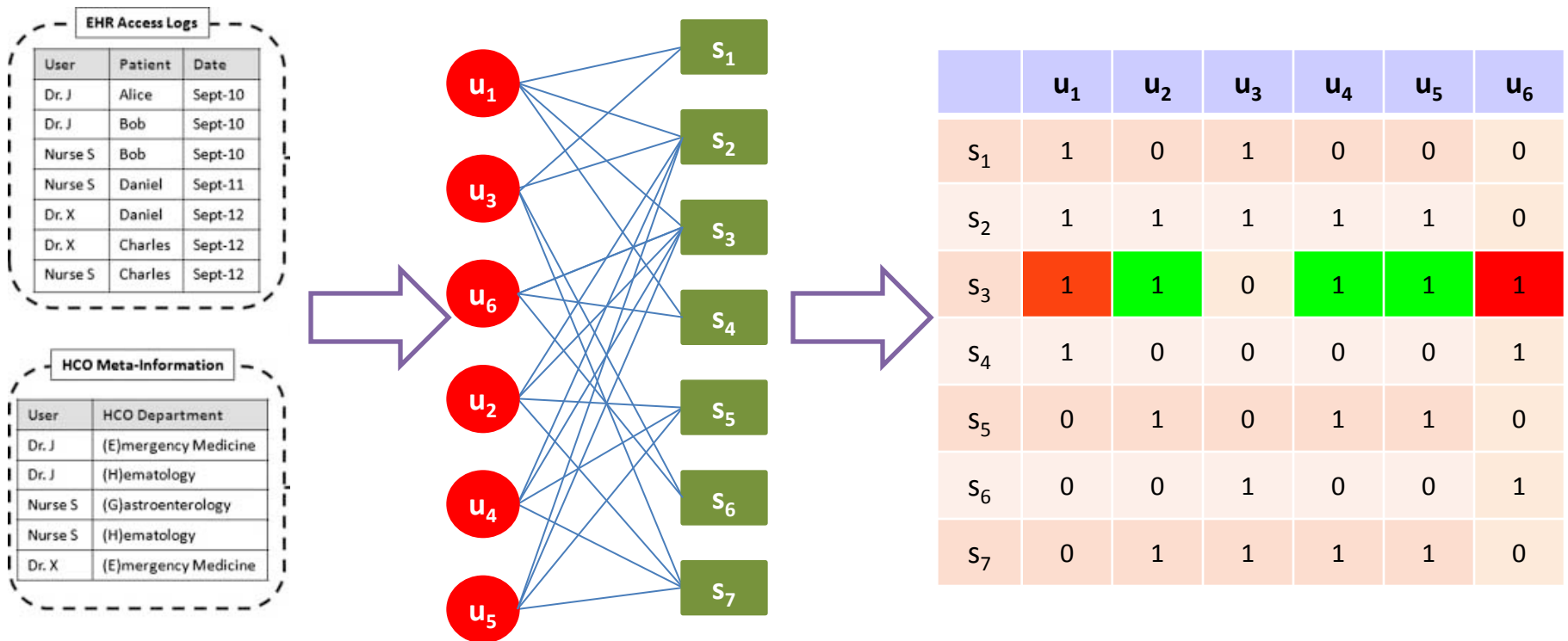
Data



Data which can be interpreted
or used by designed models

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity.

An example of data representation in access logs of EHR



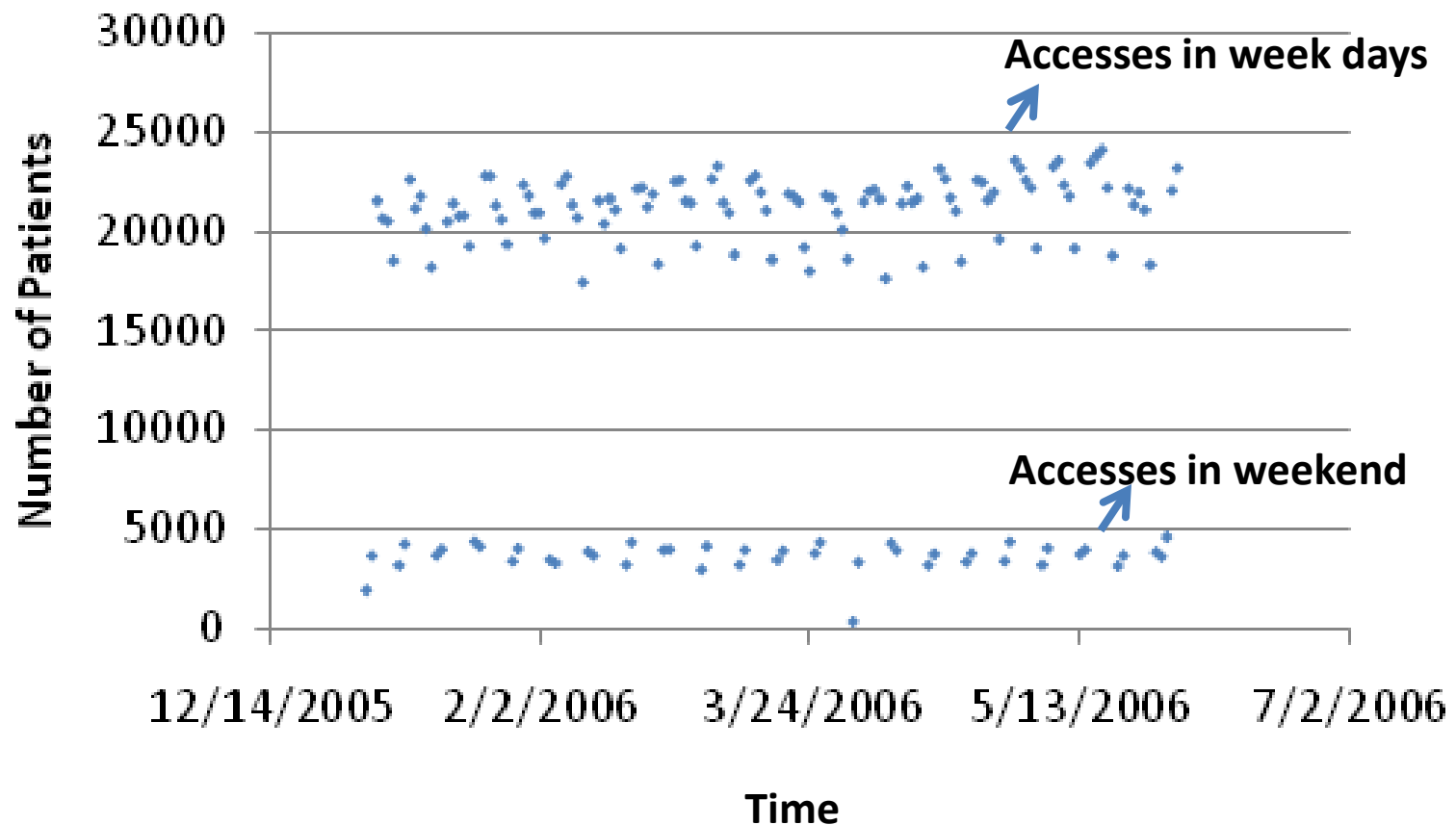
Bipartite graph of users and subjects

Binary matrix of subjects and users

Various data representations

- Flatted data
 - Data matrix or table
- Time series
- Text
- Image and video
- Workflow data

Distinct Patient Accesses across Time



A Printed Encounter Note

Subjective

Chief Complaint
Headache

Patient History

- Illnesses - asdsdfsdafsdafasd; Measles asdf
- Operations - Appendectomy - 1987
- Social History - Married; Employed
- Family History - Denies All

Allergies

- Uncategorized - Peanut-Allergen-Ingredient
- Uncategorized - Advil Liqui-Gel-Allergen-Medication
- Uncategorized - Augmentin-Allergen-Medication
- Uncategorized - Amoxicillin-Allergen-Ingredient
- Uncategorized - Tylenol-Allergen-Medication
- Uncategorized - Vytorin 10-10-Allergen-Medication - Difficulty breathing/constricting of throat, Dizziness, Congestion, Anxiety, Confusion, Drowsiness, Intolerance; comments
- Uncategorized - Dust allergy (disorder)
- Uncategorized - Allergy to animal hair (disorder)

Current Medications

- Vicodin; Date: 01/01/2008; Sig:

Image result stored in EHR

Physician Hide VTB Tools Print Help Lock X

Chart Clinical Desktop MAR (V11) Worklist

TEST, UPGRADE2 MRN: 1336731 Sex: M Pri Ins: PCP: Directives: Research
DOB: 02/02/1979 H Phone: (727)123-4567 Security: No Restricted Data
Age: 30 Years W Phone: Note: [Select] FYI: [FYI]

Select Patient i

Order Viewer Transcription Encounter 01/28/2009

[Results](#) [Results History](#) [Order Details](#) [Order Annotations](#) [ImageLink](#) [Previous Results](#) [Previous](#) [Next](#)

CHEST XRAY 2 VIEWS Resulted: Preliminary

*** CHEST XRAY 2 VIEWS Preliminary**

Images available for viewing. Please click on the Link above.

Ordered by: **Provider,Test** Collected/Examined: **03Nov2008 02:12PM**
Verified by: **Provider,Test**
Result Communication: No patient communication needed at this time;
Stage: **Preliminary**
Resulted: **03Nov2008 02:12PM** Last Updated: **28Jan2009 07:30AM** Accession: **175_RISIC**

Results History

12Feb2009 03Nov2008

CHEST XRAY 2 VIEWS [Report 1](#) [Report 2](#)

Report #1 - 12Feb2009 02:10PM - CHEST XRAY 2 VIEWS

Addendum Begins
This is an addendum to test tasking
Addendum Ends
Result to check on tasking

Report #2 - 03Nov2008 02:12PM - CHEST XRAY 2 VIEWS

Images available for viewing. Please click on the Link above.

Order Details

Status: **Resulted: Preliminary** Requested to be done: **03Nov2008**
Requested Performing Location: **Radiology North** Priority: **Routine** Order #: **TW18803598** Requisition #: **122548**
Overdue after: **13Nov2008**
Ordered by: **Provider,Test** Supervised by: **Provider,Test** Authorization: **Not Required**
Financial Authorization: **Not Needed**

Edit Print Print Result Fax Task Audit Copy Copy Selected Annotate Mark As Reviewed

GE Centricity 3.0 -- Web Page Dialog

TEST, UPGRADE2
ID: 1336731

CHEST XRAY 2 VIEWS
04/07/2008 2:57:03 PM 1 series
Chest 2 Images

TEST, UPGRADE2 (1336731) 04/07/2008 2:57:03 PM (2 of 2 images)

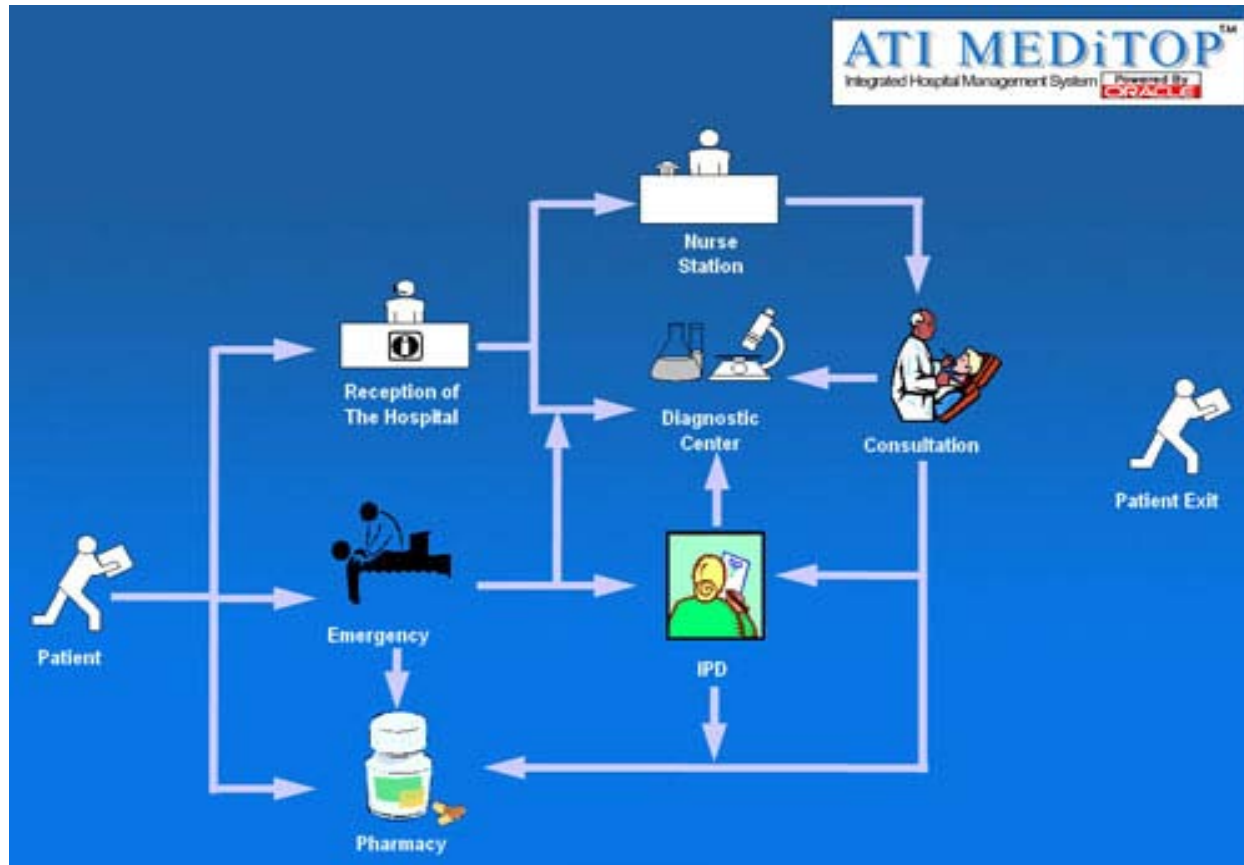
Se:3618 [H] TEST, UPGRADE2 Se:3618 [H] TEST, UPGRADE2
Im:2 Study Date:04/07/... Im:3 Study Date:04/07/...
Study Time:2:57:0... MRN:1336731 Study Time:2:57:0... MRN:1336731

Warning: not enough gray scales (32) to display the image properly C8192 W16384

Patient info **Exam info** **Key Image notes** **Exam notes** **Report**

Respiration	<input type="checkbox"/>	16			12
Respiration Quality	<input type="checkbox"/>	Norm			
Height	<input type="checkbox"/>	65 in		65.5 in	72 in
Weight	<input type="checkbox"/>	125.125 lb		141 lbs2	171 lbs2
Pain Scale	<input type="checkbox"/>	8			

Clinical Workflow



Types of attribute scales

- Nominal Scale
- Ordinal Scale
- Numerical Scale
 - Ratio Scale
 - Interval Scale

Nominal Scale

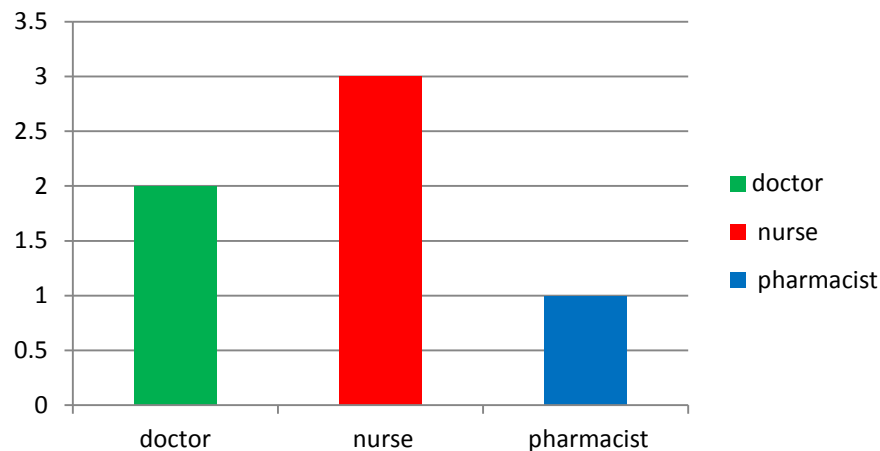
- The values of the attribute are only “labels”, which is used to distinguish each other
 - Finite number of values
 - No order information
 - No algebraic operation could be conducted, except those related to frequency
- An example
 - {1,2,3} -> {doctor, nurse, pharmacist}
 - > {Medical Information Service, Clinical Trials Center , Breast Center }

A table describing accesses of roles and users on a patient across time

Nominal
Attribute →

Accessed Role	Number of Accesses	Number of Users	Time Span
docotr	10	3	t1-t2
nurse	20	5	t2-t3
pharmacist	15	6	t3-t4
doctor	16	4	t4-t5
nurse	22	5	t5-t6
nurse	6	2	t6-t7

Frequency of
different values
in nominal
attribute



Ordinal Scale

- The values of the attribute is to indicate certain ordering relationship resided in the attribute
 - Order is more important than value
 - No algebraic operation could be conducted, except those related to sorting
- An example
 - Richter's scale on earthquake

A heartquake of 5.5 magnitudes is more important than one of 3 but less than one 9. Which indicates that there is an order between the data

Richter magnitudes	Description	Earthquake effects
Less than 2.0	Micro	Micro earthquakes, not felt.
2.0-2.9	Minor	Generally not felt, but recorded.
3.0-3.9		Often felt, but rarely causes damage.
4.0-4.9	Light	Noticeable shaking of indoor items, rattling noises. Significant damage unlikely.
5.0-5.9	Moderate	Can cause major damage to poorly constructed buildings over small regions. At most slight damage to well-designed buildings.
6.0-6.9	Strong	Can be destructive in areas up to about 160 kilometers (100 mi) across in populated areas.
7.0-7.9	Major	Can cause serious damage over larger areas.
8.0-8.9	Great	Can cause serious damage in areas several hundred miles across.
9.0-9.9		Devastating in areas several thousand miles across.
10.0+	Epic	Never recorded; see below for equivalent seismic energy yield.

Despite having the same interval of 0.5, the difference from one point to another in the scale in Joule is not uniform

Richter Approximate Magnitude	Joule equivalent
0.0	63.1 kJ
0.5	355 kJ
1.0	2.00 MJ
1.5	11.2 MJ
2.0	63.1 MJ
2.5	355 MJ
3.0	2.00 GJ
3.5	11.2 GJ

Difference of 0.5

Difference of 9.2 MJ

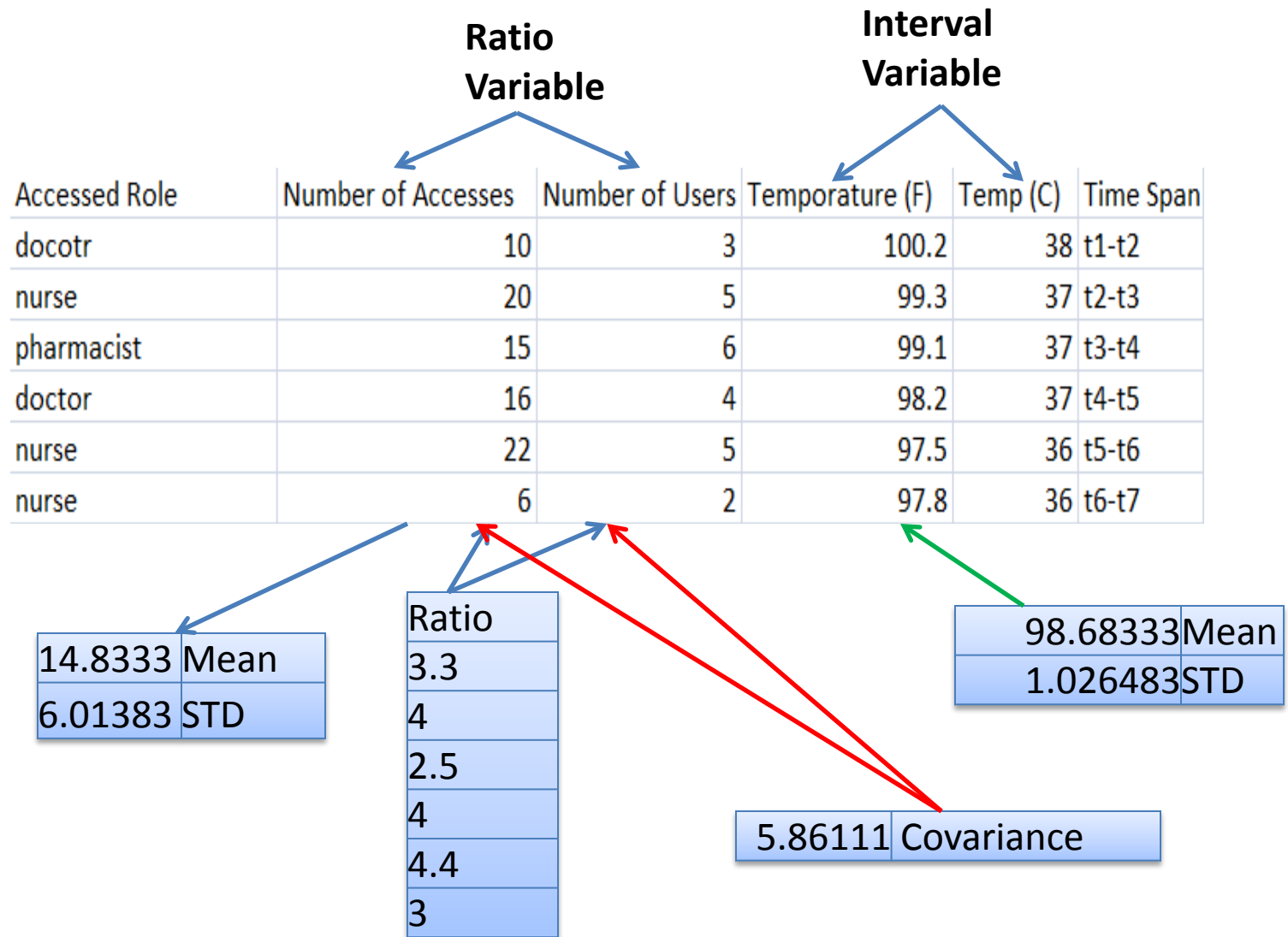
Difference of 51.9 MJ

Numerical Scale

- The values of the attribute is to indicate quantity of some predefined unit
 - There should be a basic unit, which can be transferred to another one
 - The value is how many copies of the basic unit
 - Some algebraic operations could be conducted
- Two types of numerical scale
 - Interval scale
 - Ratio scale

Differences of Ratio Scale and Interval Scale

- An interval variable is a measurement **where the difference between two values is meaningful**.
 - The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.
- A ratio variable, **has all the properties of an interval variable, and also has a clear definition of 0.0**. When the variable equals 0.0, there is none of that variable.
 - Variables like number of accesses on a patient, number of accesses of a user are ratio variables.
 - Temperature, expressed in F or C, is not a ratio variable. A temperature of 0.0 on either of those scales does not mean 'no temperature'.
 - However, temperature in Kelvin is a ratio variable, as 0.0 Kelvin really does mean 'no temperature'.



Which Computing Operations Could be Done?

Algebraic Operation	Nominal	Ordinal	Interval	Ratio
Frequency distribution	Yes	Yes	Yes	Yes
Median and Percentiles	No	Yes	Yes	Yes
Add or Subtract	No	No	Yes	Yes
Mean, Standard Deviation	No	No	Yes	Yes
Ratio, or Coefficient of Variation	No	No	No	Yes

What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization and Discretization
- Similarity Measurements
- Dimensionality Reduction

Why we need data normalization?

	15,000	3.5	1,2
User	Access	Age	Experience
1	25,000	24	4
2	40,000	27	5
3	55,000	32	7
4	27,000	25	5
5	53,000	30	5

Euclidean distance



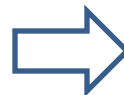
	1	2	3	4	5
1	0	15000	30000	2000	28000
2	15000	0	15000	13000	13000
3	30000	15000	0	28000	2000
4	2000	13000	28000	0	26000
5	28000	13000	2000	26000	0

(a) Euclidean distance before normalization

Min-Max Normalization



User	Access	Age	Experience
1	0.000033	0.100	0.200
2	0.500000	0.400	0.400
3	1.000000	0.900	0.800
4	0.070000	0.200	0.400
5	0.930000	0.700	0.400



	1	2	3	4	5
1	0.000000	0.616414	1.414166	0.233332	1.1273
2	0.616414	0.000000	0.812383	0.477234	0.5270
3	1.414166	0.812383	0.000000	1.233286	0.4521
4	0.233332	0.477234	1.233286	0.000000	1.0005
5	1.127384	0.527022	0.452154	1.000505	0.0000

(b) Euclidean distance after normalization

Three types of normalization

- The attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, 0.0 to 1.0
 - Min-max normalization
 - Z-score normalization
 - Decimal scaling normalization

Min-Max Normalization

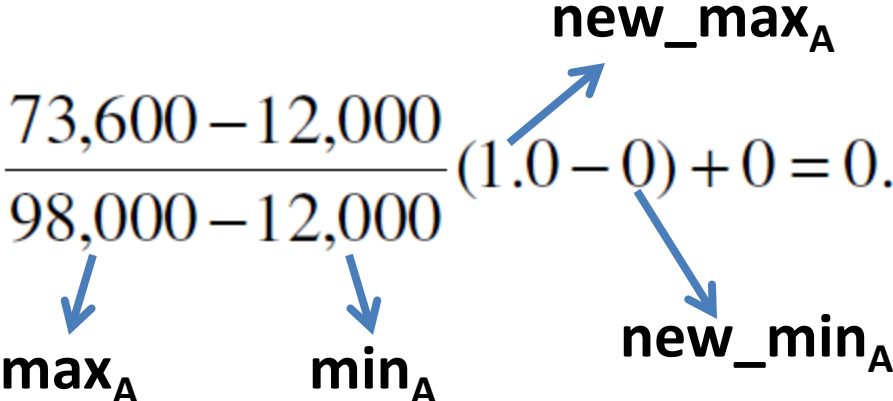
- Performs a linear transformation on the original data
- Support: \min_A and \max_A are the minimum and maximum values of an attribute, A.
- Min-max normalization maps a value, v , of A to v' in the range $[\text{new_min}_A, \text{new_max}_A]$ by computing:

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

An Example of Min-Max Normalization

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

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



max_A min_A new_max_A new_min_A

Z-Score Normalization

- Change the original data quite a bit
- The values for an attribute, A, are normalized based on the mean (\bar{A}) and standard deviation (σ_A) of A.
- A value, v , of A is normalized to v' by computing:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

distance is represented by units of standard deviations from the mean

Accessed Role	Number of Accesses	Number of Users	Temperature (F)	Temp (C)	Time Span
docotr	10	3	100.2	38	t1-t2
nurse	20	5	99.3	37	t2-t3
pharmacist	15	6	99.1	37	t3-t4
doctor	16	4	98.2	37	t4-t5
nurse	22	5	97.5	36	t5-t6
nurse	6	2	97.8	36	t6-t7

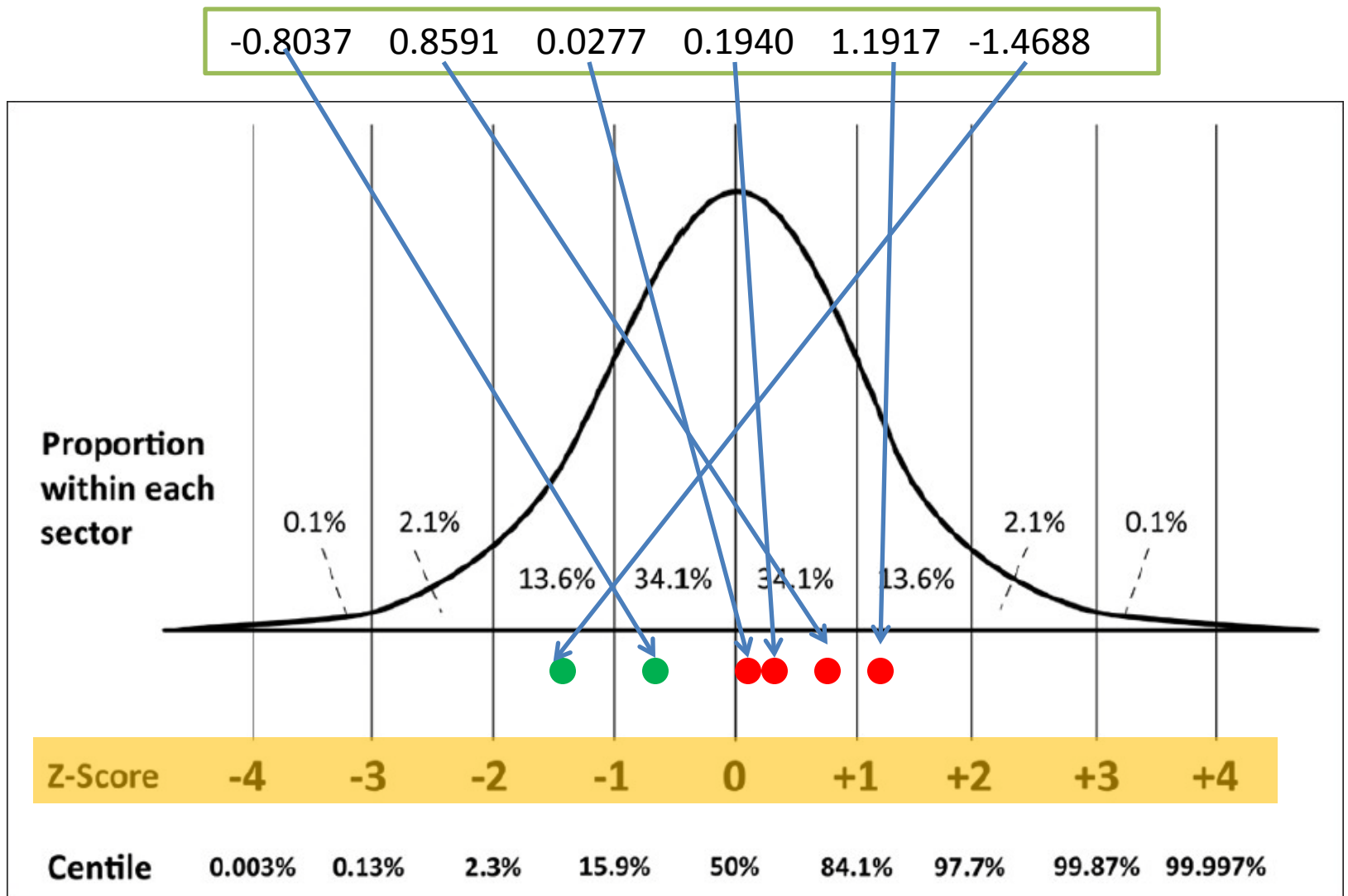
14.8333	Mean
6.01383	STD

Z-score

-0.8037 0.8591 0.0277 0.1940 1.1917 -1.4688

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

The role nurse during time t2-t3 has accesses above average— a distance of 0.8591 above the average accesses



Decimal Scaling Normalization

- normalizes by moving the decimal point of values of attribute A.
- The number of decimal points moved depends on the maximum absolute value of A.
- A value, v , of A is normalized to v' by computing :

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\text{Max}(|v'|) < 1$

An Example of Decimal Scaling

- Suppose that the recorded values of A range from -986 to 917.
- The maximum absolute value of A is 986.
- To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., $j = 3$) so that

Data Discretization

- Dividing the range of a continuous attribute into intervals
- Interval labels can then be used to replace actual data values
- Reduce the number of values for a given continuous attribute

Why Data Discretization?

- Some classification algorithms only accept categorical attributes
 - many learning methods –like association rules, Bayesian networks can handle only discrete attributes
- This leads to a concise, easy-to-use, knowledge-level representation of mining results

The goal of discretization is to reduce the number of values for a continuous attribute

grouping them into a number, n , of intervals (bins).

Typical Methods of Discretization

- Binning
 - Top-down split, unsupervised
- Clustering analysis
 - Either top-down split or bottom-up merge, Supervised
- Interval merging by χ^2 Analysis
 - Supervised, bottom-up merge

Discretization

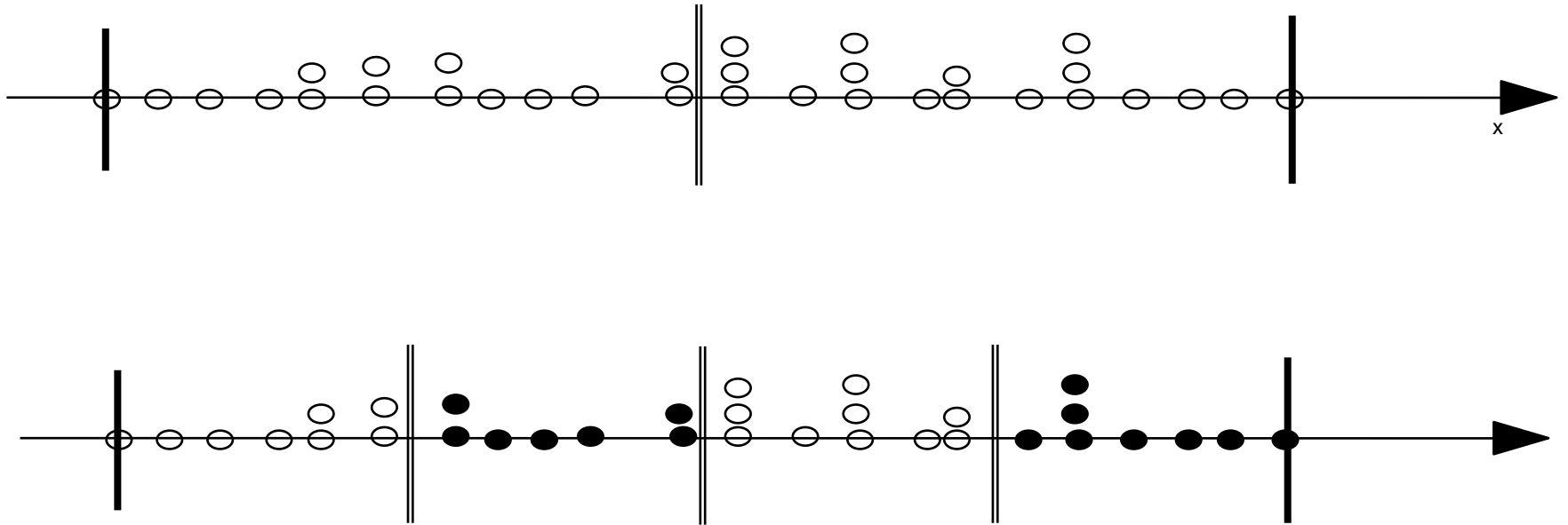


Illustration of the supervised vs. unsupervised discretization

Binning

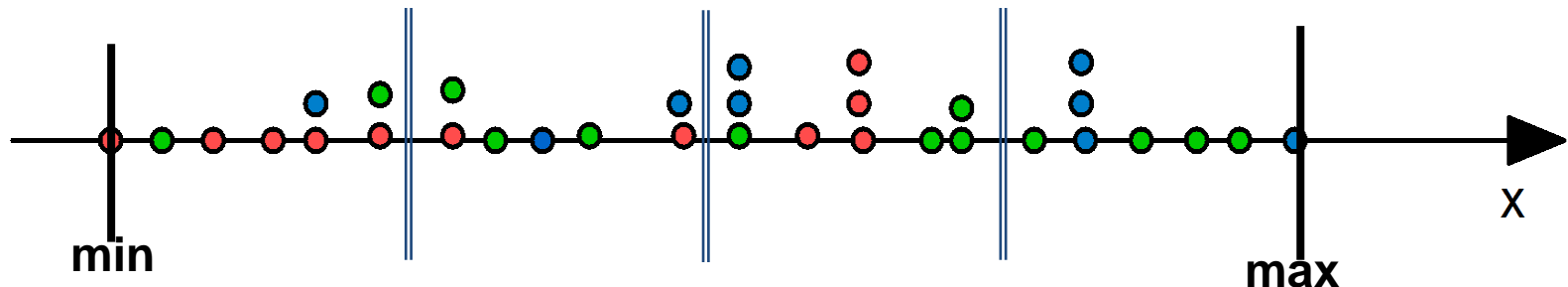
- The sorted values are distributed into a number of buckets, or bins, and then replacing each bin value by the bin mean or median
- Binning is a top-down splitting technique based on a specified number of bins
- Binning is an unsupervised discretization technique, because it does not use class information

Two Methods of Binning

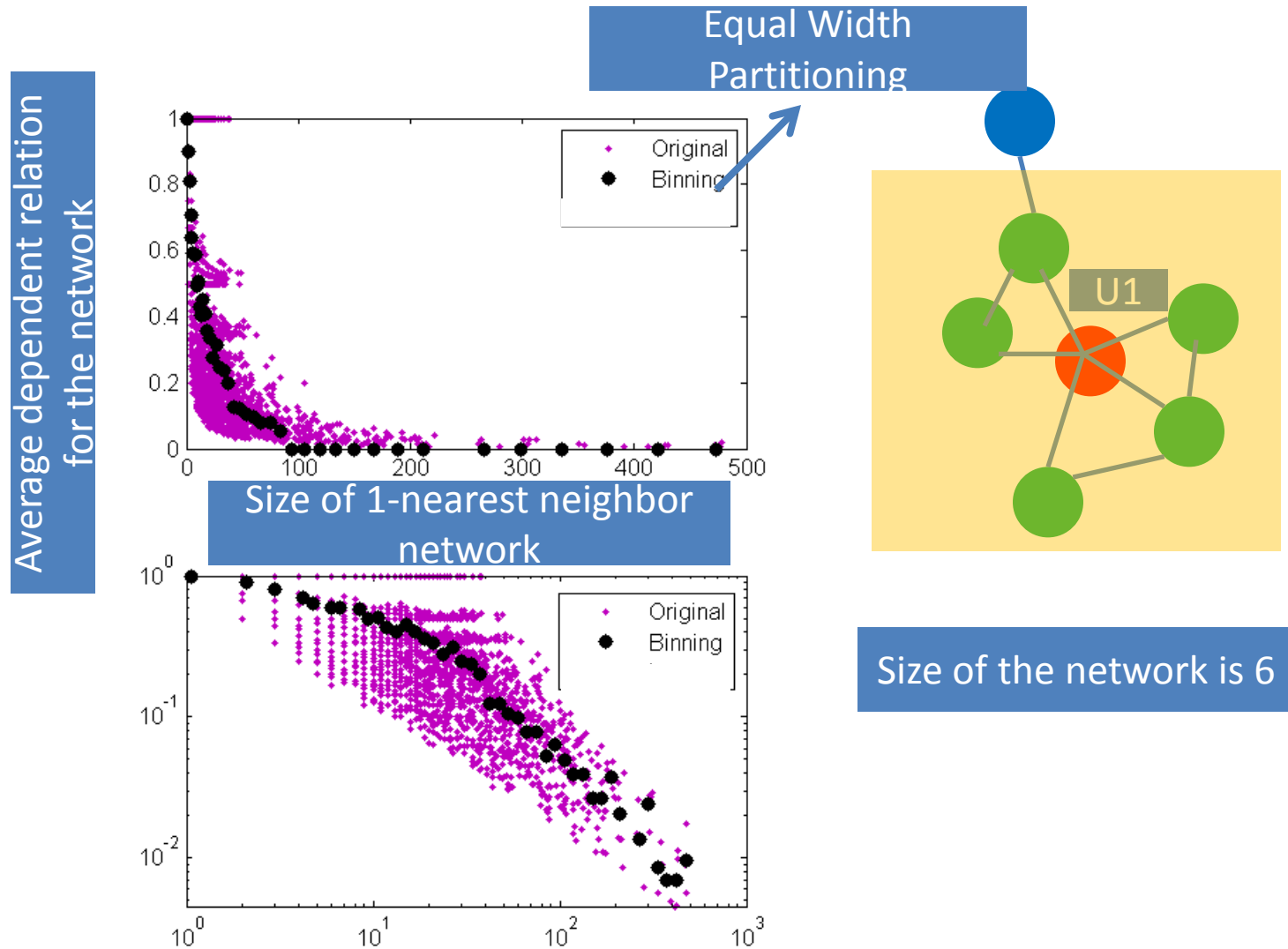
- Equal-width (distance) partitioning
 - Divides the range into N intervals of equal size
 - if A and B are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$
 - The most straightforward, but outliers may dominate presentation
- Equal-depth (frequency) partitioning

An Example of Equal-width Partitioning

- Sorted data for price (in dollars):
 - 4, 8, 15, 21, 21, 24, 25, 28, 34
- $W = (B - A) / N = (34 - 4) / 3 = 10$
 - Bin 1: 4-14, Bin2: 15-24, Bin 3: 25-34
- Equal-width (distance) partitioning:
 - Bin 1: 4, 8
 - Bin 2: 15, 21, 21, 24
 - Bin 3: 25, 28, 34



Distribution of users in hospital on two measurements



- Equal-depth (frequency) partitioning
 - Divides the range into N intervals, each containing approximately same number of samples
 - Good data scaling
- Example
 - Sorted data for price (in dollars):
 - 4, 8, 15, 21, 21, 24, 25, 28, 34
 - Equal-depth (frequency) partitioning:
 - Bin 1: 4, 8, 15
 - Bin 2: 21, 21, 24
 - Bin 3: 25, 28, 34

Typical Methods of Discretization

- Binning
 - Top-down split, unsupervised
- Clustering analysis
 - Either top-down split or bottom-up merge, Supervised
- Interval merging by χ^2 Analysis
 - Supervised, bottom-up merge

Clustering Analysis

- A clustering algorithm can be applied to discretize a numerical attribute, A , by partitioning the values of A into clusters or groups.
- Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results.

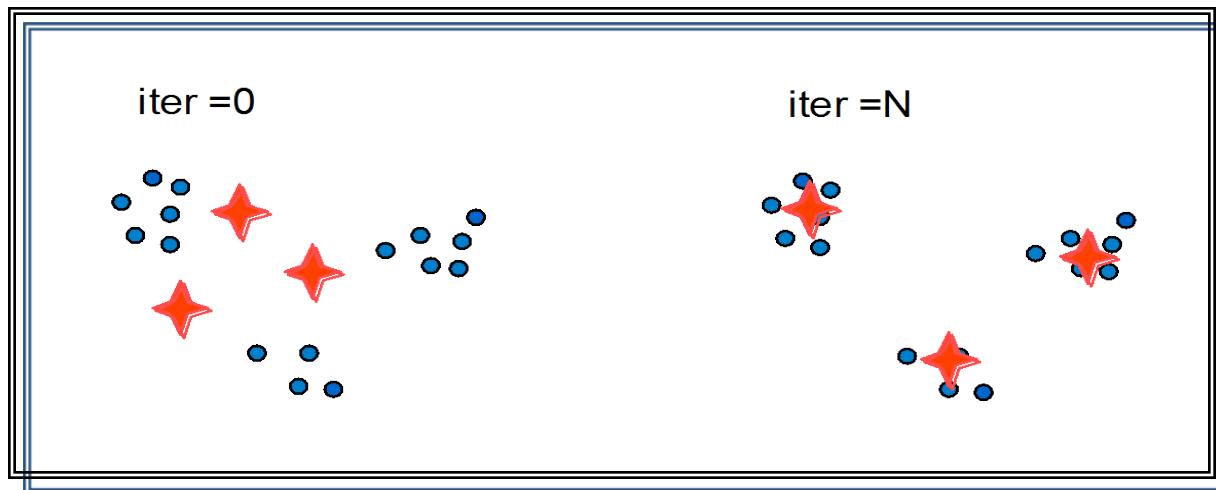
Generate a Concept Hierarchy for Attribute A

- By following either a top-down splitting strategy or a bottom-up merging strategy, where each cluster forms a node of the concept hierarchy
 - In the former, each initial cluster or partition may be further decomposed into several sub-clusters, forming a lower level of the hierarchy
 - In the latter, clusters are formed by repeatedly grouping neighboring clusters in order to form higher level concepts.

Example: K-means Clustering Discretization

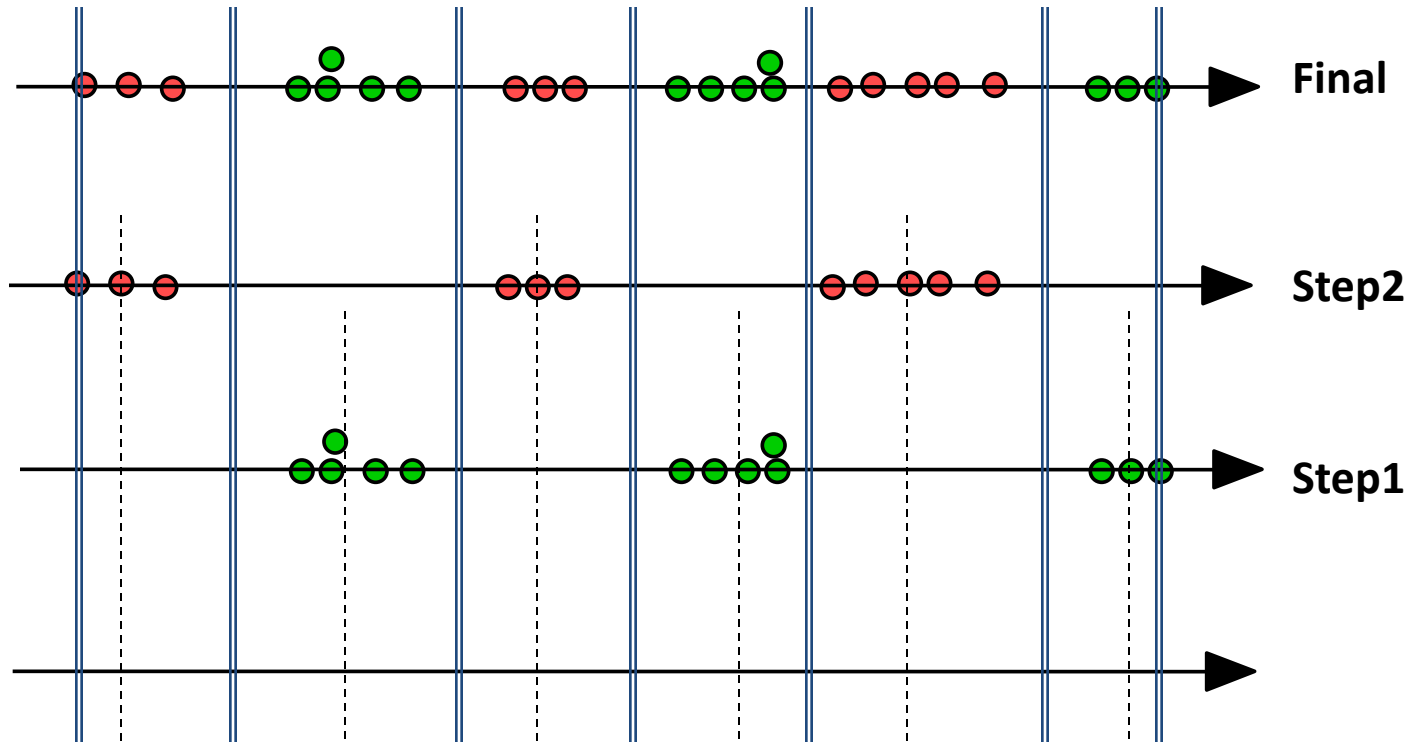
K-means clustering is an iterative method of finding clusters in multidimensional data; the user must define:

- number of clusters for **each** feature
- similarity function
- performance index and termination criterion



K-means Clustering Discretization

Example:



cluster centers

=====
interval's boundaries/midpoints (min value, midpoints, max value)

K-means Clustering Discretization

- **The clustering must be done for all attribute values for each class separately.**

The final boundaries for this attribute will be all of the boundaries for all the classes.

- **Specifying the number of clusters is the most significant factor influencing the result of discretization:**

to select the proper number of clusters, we cluster the attribute into several intervals (clusters), and then calculate some measure of goodness of clustering to choose the most “correct” number of clusters

Typical Methods of Discretization

- Binning
 - Top-down split, unsupervised
- Clustering analysis
 - Either top-down split or bottom-up merge, Supervised
- Interval merging by χ^2 Analysis
 - Supervised, bottom-up merge

Chi Merge

- It is a bottom-up method
- Find the best neighboring intervals and merge them to form larger intervals recursively
- The method is supervised in that it uses class information
- The basic notion is that for accurate discretization, the relative class frequencies should be fairly consistent within an interval
- Therefore, if two adjacent intervals have a very similar distribution of classes, then the intervals can be merged. Otherwise, they should remain separate




ChiMerge Technique Example

Sample	Feature	Class
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

	K=1	K=2	Σ
Interval [7.5, 8.5]	$A_{11}=1$	$A_{12}=0$	$R_1=1$
Interval [8.5, 9.5]	$A_{21}=1$	$A_{22}=0$	$R_2=1$
Σ	$C_1=2$	$C_2=0$	$N=2$

- Interval points for feature F are: 0, 2, 5, 7.5, 8.5, 9.5, etc.

ChiMergeTechnique (Example)

	K=1	K=2	Σ
Interval [7.5, 8.5]	$A_{11}=1$	$A_{12}=0$	$R_1=1$  Row Sum
Interval [8.5, 9.5]	$A_{21}=1$	$A_{22}=0$	$R_2=1$
Σ	$C_1=2$  Column Sum	$C_2=0$	$N=2$  Total Sum

- Based on the table's values, we can calculate expected values:

$$E_{11} = 2/2 = 1, E_{12} = 0/2 \approx 0.1,$$

$$E_{21} = 2/2 = 1, \text{ and } E_{22} = 0/2 \approx 0.1$$

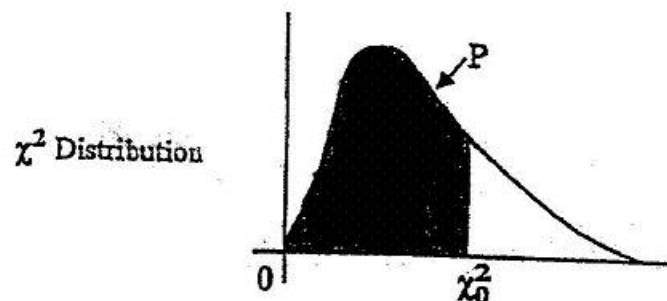
$$\text{Expected} = (\text{rowSum} * \text{colSum}) / \text{TotalSum}$$

- And corresponding χ^2 score: $\chi^2 = \text{Sum}((\text{Original} - \text{Expected})^2 / \text{Expected})$

$$\chi^2 = (1 - 1)^2 / 1 + (0 - 0.1)^2 / 0.1 + (1 - 1)^2 / 1 + (0 - 0.1)^2 / 0.1 = 0.2$$

For the degree of freedom $d=1$, and $\chi^2 = 0.2 < 2.706$ (P value 0.90) (MERGE !)

$$DF = (\text{rowNum} - 1) * (\text{colNum} - 1) = (2 - 1) * (2 - 1)$$



The table below gives the value χ_0^2 for which $P[\chi^2 < \chi_0^2] = P$ for a given number of degrees of freedom and a given value of P .

Degrees of Freedom	Values of P									
	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.01	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597

ChiMergeTechnique (Example)

	K=1	K=2	Σ
Interval [0, 10.0]	$A_{11}=4$	$A_{12}=1$	$R_1=5$
Interval [10.0, 42.0]	$A_{21}=1$	$A_{22}=3$	$R_2=4$
Σ	$C_1=5$	$C_2=4$	$N=9$

- $E_{11}= 2.78$, $E_{12}=2.22$, $E_{21}= 2.22$, $E_{22}= 1.78$, and $\chi^2= 2.72 > 2.706$
(NO MERGE !)
- Final discretization: [0, 10], [10, 42], and [42, 60]

The ChiMerge Method

- Initially, each distinct value of a numerical attribute A is considered to be one interval
- Chi-Square tests are performed for every pair of adjacent intervals
- Adjacent intervals with the least Chi-Square values are merged together, since **low Chi-Square** values for a pair indicate similar class distributions
- This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max interval, max inconsistency, etc.)

What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization and Discretization
- Similarity Measurements
- Dimensionality Reduction

Why care about similarity?

- Represent the internal relationship between data objects
- It is essential to many data mining algorithms

Distance Measurements

- Distance measure can be used to characterize the concept of “similarity”
- Distance or Metric should satisfy
 - Non-negativity: $d(i, j) \geq 0$ and $d(i, j) = 0$ iff $i = j$
 - Symmetry $d(i, j) = d(j, i)$ for all i, j
 - Triangle inequality

$$d(i, j) \leq d(i, k) + d(k, j) \text{ for all } i, j \text{ and } k$$

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^d |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, d is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) of data objects p and q .

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. “supremum” (L_{\max} norm, L_{∞} norm) distance
 - This is the maximum difference between any component of the vectors

Euclidean Distance

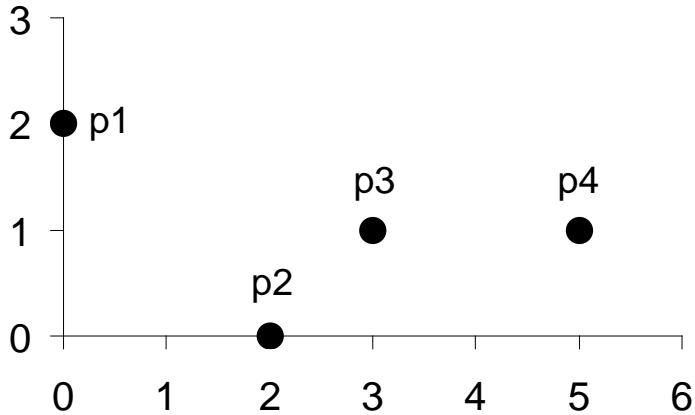
- Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^d (p_k - q_k)^2}$$

Normalization is necessary, if scales differ.

Euclidean Distance-Examples

$$\text{Dis}(p1,p2)=\text{Sqrt}((0-2)^2+(2-0)^2)=\text{sqrt}(8)=2.828$$



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Weighted Minkowski distance

$$d(\mathbf{x}_1, \mathbf{x}_2) = \left(\sum_{k=1}^d w_k (\mathbf{x}_1(k) - \mathbf{x}_2(k))^p \right)^{\frac{1}{p}}$$

Reflects the importance of each attribute

In both weighted and unweighted Minkowski distance, each attribute contribute independently to the measure of distance

Mahalanobis Distance

- Mahalanobis distance standardizes data not only in the direction of each attributes but also based on the covariance between attributes

$$mahalanobis(p, q) = \sqrt{(p - q) \Sigma^{-1} (p - q)^T}$$

Where p and q are two data points in d dimensions

Σ is the covariance matrix of the input data X , *the size of it is d by d. “d” is the number of attributes or variables*

$$\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

An Example of Mahalanobis Distance for Two Data Points

Step1: There are three data points in two attributes

A: (0.5, 0.5);

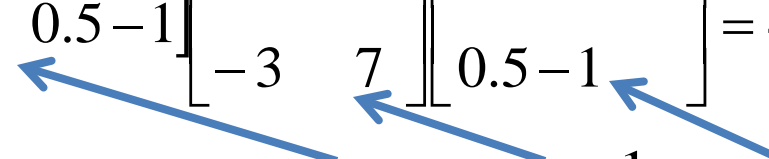
B: (0, 1);

C: (1.5, 1.5)

Step2: Covariance Matrix Calculation for two variables

$$\Sigma = \begin{bmatrix} 0.58 & 0.25 \\ 0.25 & 0.25 \end{bmatrix} \quad \Sigma^{-1} = \begin{bmatrix} 3 & -3 \\ -3 & 7 \end{bmatrix}$$

Step3: Distance calculation

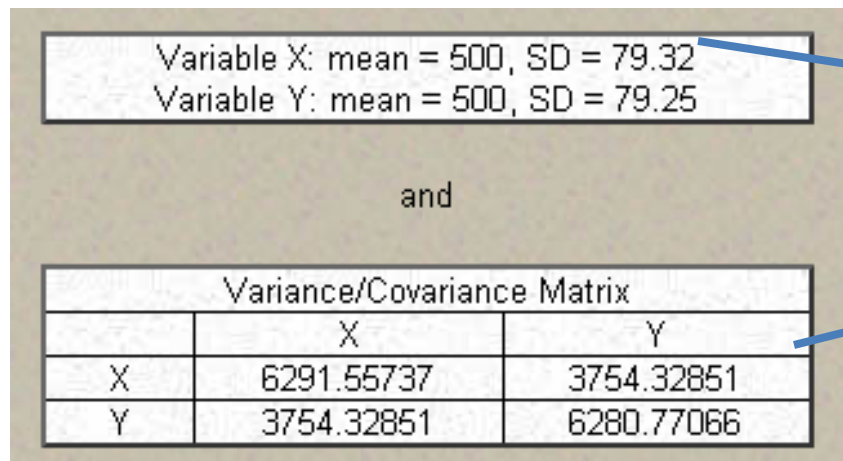
$$Mahal(A, B) = [0.5 - 0 \quad 0.5 - 1] \begin{bmatrix} 3 & -3 \\ -3 & 7 \end{bmatrix} \begin{bmatrix} 0.5 - 0 \\ 0.5 - 1 \end{bmatrix} = 4$$


$$mahalanobis(p, q) = (p - q) \Sigma^{-1} (p - q)^T$$

An Example of Mahalanobis Distance for One Data Point

Step1: A single observation $x(410, 400)$ in two dimensions: d_1 (named X) and d_2 (named Y)

Step2: Covariance Matrix Calculation of dataset



The screenshot displays statistical results for two variables, X and Y. At the top, a box contains the mean and standard deviation for each variable. Below this, the word 'and' is centered. At the bottom, a table titled 'Variance/Covariance Matrix' shows the relationships between X and Y. Blue arrows point from the text annotations to the mean/SD box and the covariance matrix table.

Variable X: mean = 500, SD = 79.32 Variable Y: mean = 500, SD = 79.25		
and		
Variance/Covariance Matrix		
	X	Y
X	6291.55737	3754.32851
Y	3754.32851	6280.77066

For d_1 , the mean value is 500, and standard deviation is 79.32

Covariance Matrix of variables X, and Y on all dataset

Step3: Distance calculation

$$\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)$$

Single point x in two dimensions X and Y

**x=(410
400)**

Mean vector of all attributes (X and Y)

**m=(500
500)**

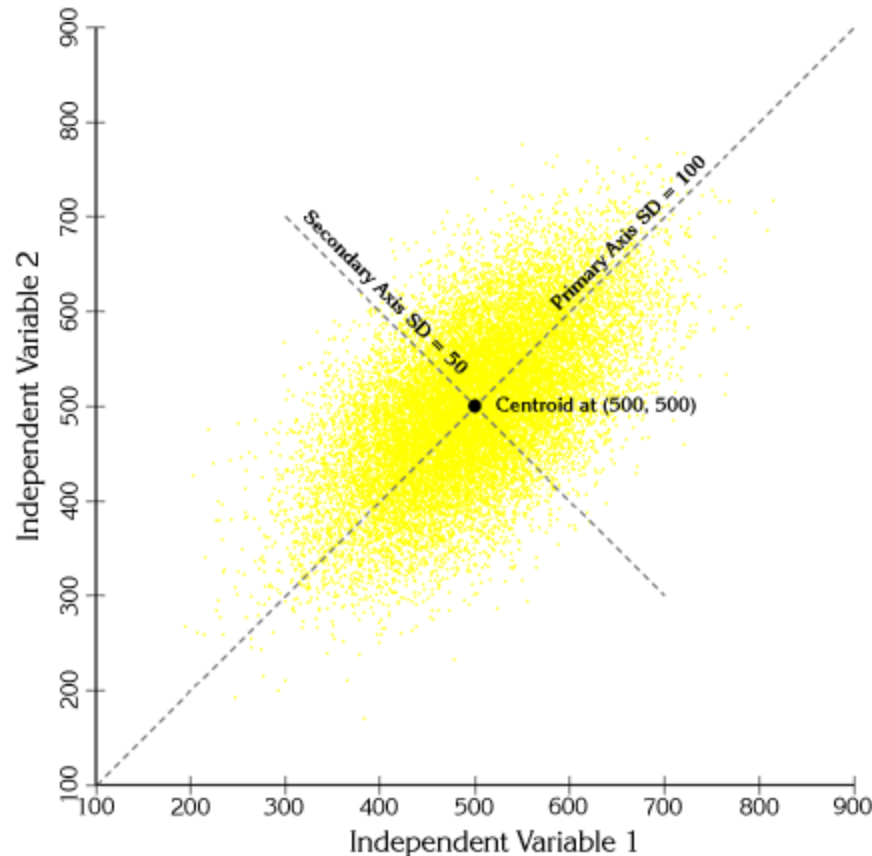
Given that Mahalanobis Distance $D^2 = (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$

$$(\mathbf{x} - \mathbf{m}) = \begin{pmatrix} 410 - 500 \\ 400 - 500 \end{pmatrix} = \begin{pmatrix} -90 \\ -100 \end{pmatrix}$$

$$\mathbf{C}^{-1} = \begin{pmatrix} 6291.55737 & 3754.32851 \\ 3754.32851 & 6280.77066 \end{pmatrix}^{-1} = \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix}$$

$$\begin{aligned} \text{Therefore } D^2 &= (-90 \quad -100) \times \begin{pmatrix} 0.00025 & -0.00015 \\ -0.00015 & 0.00025 \end{pmatrix} \times \begin{pmatrix} -90 \\ -100 \end{pmatrix} \\ &= 1.825 \end{aligned}$$

We can also draw actual ellipses at regions of constant Mahalanobis values



- One interesting feature to note from this figure is that a Mahalanobis distance of 1 unit corresponds to 1 standard deviation along both primary axes of variance

Similarity Between Binary Vectors

- Common situation is that objects, p and q , have only binary attributes

- Compute similarities using the following quantities

M_{01} = the number of attributes where p was 0 and q was 1

M_{10} = the number of attributes where p was 1 and q was 0

M_{00} = the number of attributes where p was 0 and q was 0

M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Distance/Coefficients

SMC = number of matches / number of attributes

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

J = number of value-1-to-value-1 matches / number of not-both-zero attributes values

$$= (M_{11}) / (M_{01} + M_{10} + M_{11})$$

SMC versus Jaccard: Example

$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$

$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$\text{SMC} = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| ||d_2||,$$

where \bullet indicates vector dot product and $||d||$ is the length of vector d .

- Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150, \text{ distance} = 1 - \cos(d_1, d_2)$$

Distance between two values x and y of an attribute a (*Nominal*)

- Value Difference Metric (VDM)- Classes based distance measurements

The number of output classes

The number of instances in T that have value x for attribute a and output class c

constant, usually 1 or 2

$$VDM_a(x, y) = \sum_{c=1}^C \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^q$$

the number of instances in the training set T that have value x for attribute a

For example, if an attribute *color* has three values *red*, *green* and *blue*, and the application is to identify whether or not an object is an apple, *red* and *green* would be considered closer than *red* and *blue* because the former two both have similar correlations with the output class *apple*.

Complex Structure

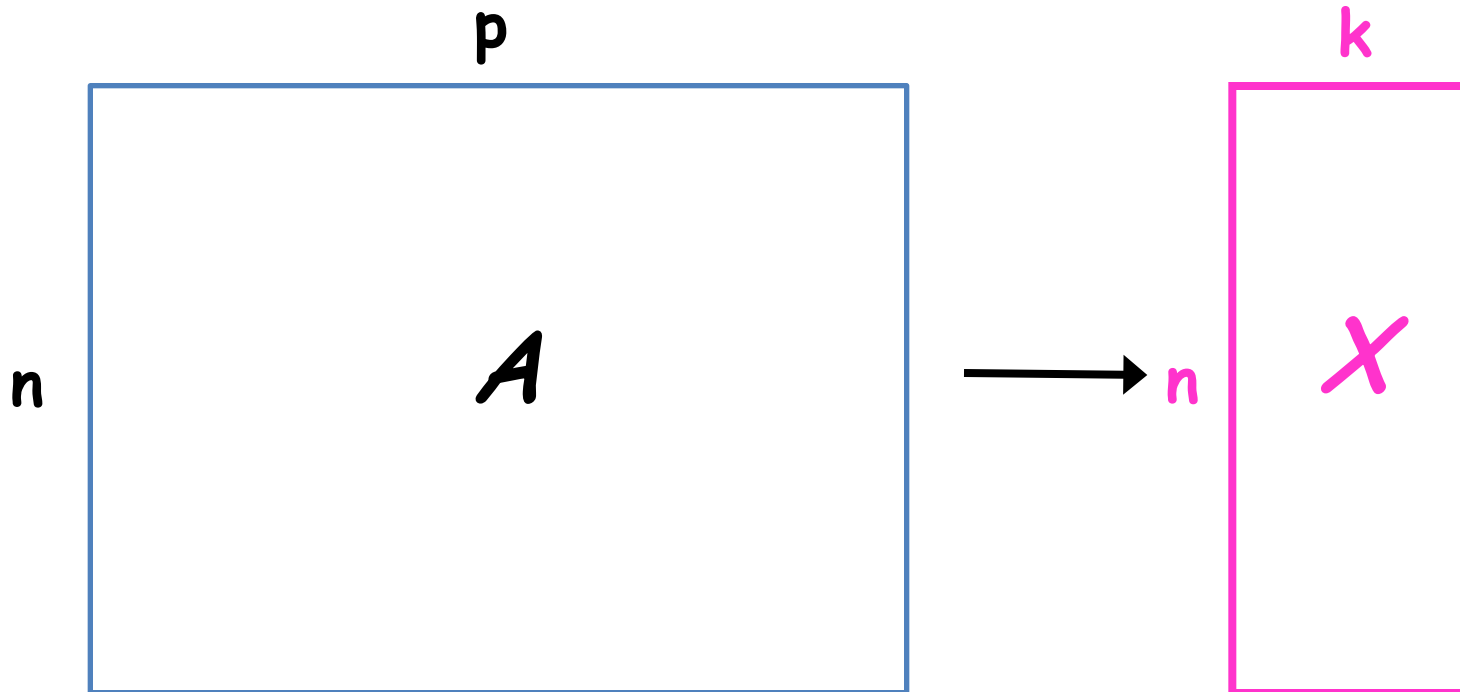
- For distribution: KL divergence, cross entropy, ...
- For trees, graphs: defining graph kernels, ...

What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization and Discretization
- Similarity Measurements
- Dimensionality Reduction

Principal Component Analysis (PCA)

- summarization of data with many (p) variables by a smaller set of (k) derived (synthetic, composite) variables.



Principal Component Analysis (PCA)

- takes a data matrix of n objects by p variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original p variables
- the first k components display as much as possible of the variation among objects.

Geometric Rationale of PCA

- objects are represented as a cloud of n points in a multidimensional space with an axis for each of the p variables
- the centroid of the points is defined by the mean of each variable
- the variance of each variable is the average squared deviation of its n values around the mean of that variable.

$$V_i = \frac{1}{n-1} \sum_{m=1}^n (X_{im} - \bar{X}_i)^2$$

Geometric Rationale of PCA

- degree to which the variables are linearly correlated is represented by their covariance

$$C_{ij} = \frac{1}{n-1} \sum_{m=1}^n (X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j)$$

Covariance of variables i and j

Sum over all n objects

Value of variable i in object m

Mean of variable i

Value of variable j in object

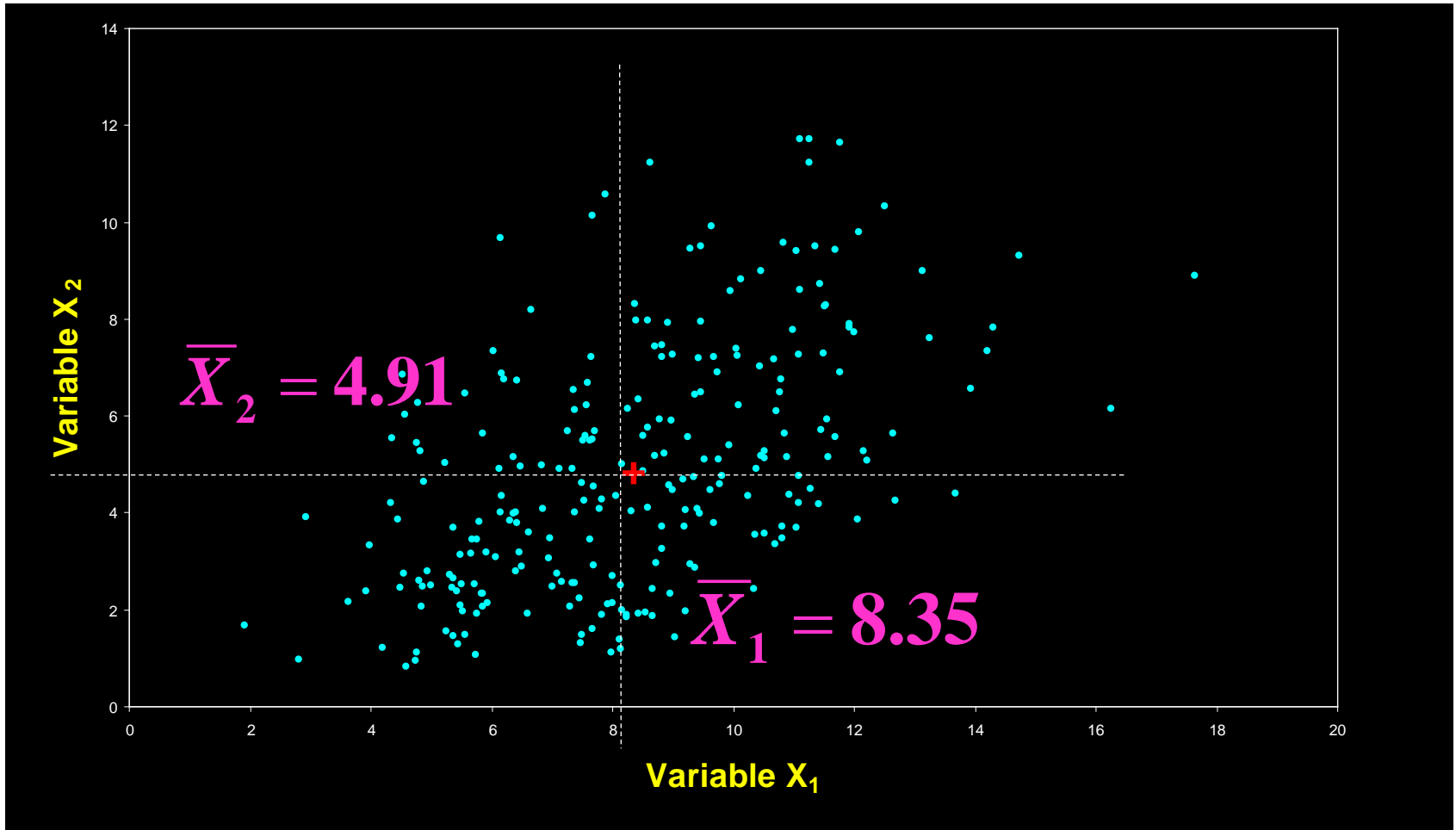
Mean of variable j

Geometric Rationale of PCA

- objective of PCA is to rigidly **rotate the axes** of this p -dimensional space to new positions (principal axes) that have the following properties:
 - ordered such that **principal axis 1 has the highest variance**, axis 2 has the next highest variance, , and axis p has the lowest variance
 - covariance among each pair of the principal axes is zero (**the principal axes are uncorrelated**).

2D Example of PCA

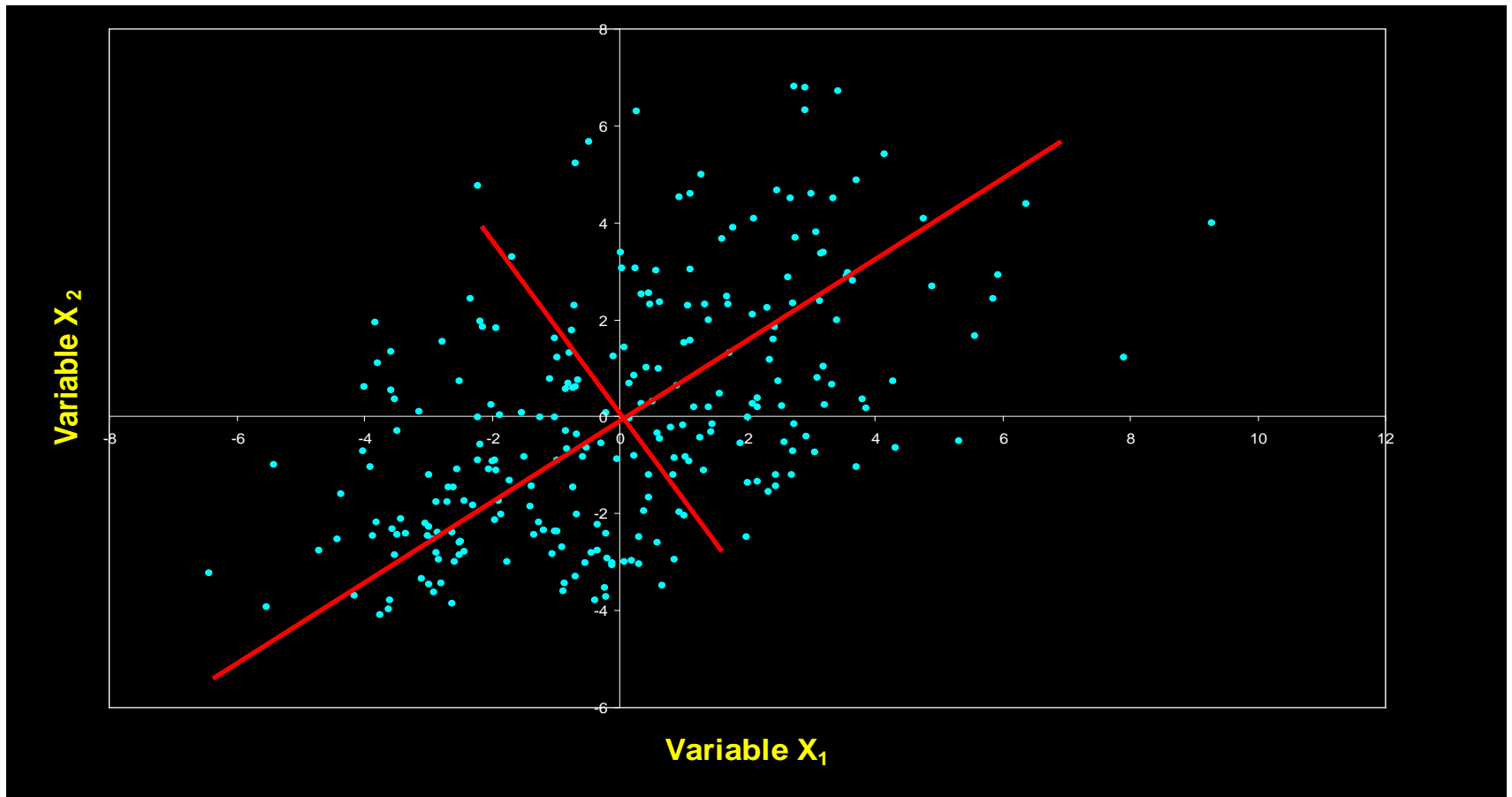
- variables X_1 and X_2 have positive covariance & each has a similar variance.



$$V_1 = 6.67 \quad V_2 = 6.24 \quad C_{1,2} = 3.42$$

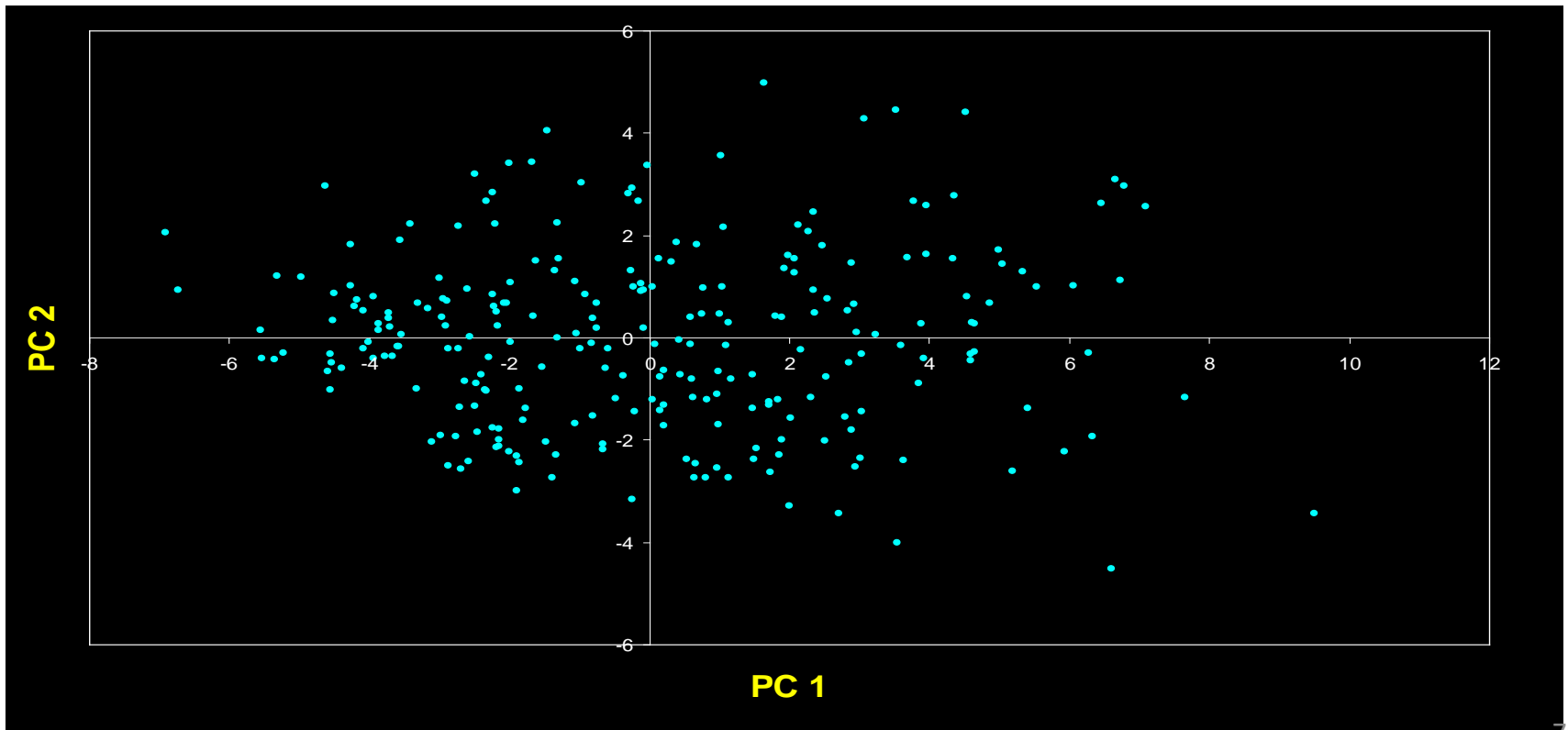
Configuration is Centered

- each variable is adjusted to a mean of zero (by subtracting the mean from each value).



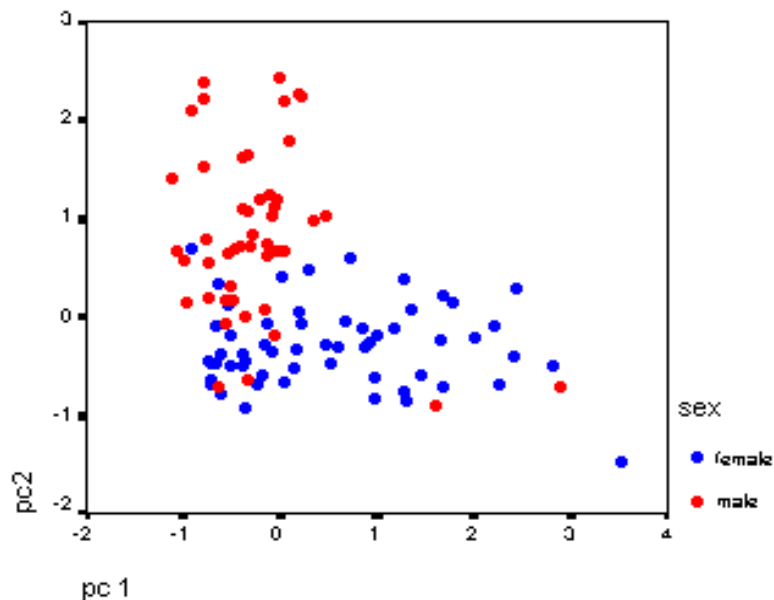
Principal Components are Computed

- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03
- PC 1 and PC 2 have zero covariance.



An Example

case	ht (x_1)	wt(x_2)	age(x_3)	sbp(x_4)	heart rate (x_5)
1	175	1225	25	117	56
2	156	1050	31	122	63
...
n	202	1350	58	154	67



Thanks!