Bab 5: Workload Characterization Techniques

Dr. Ir. Yeffry Handoko Putra, M.T

Workload Characterisation

- Modelling process of workload because it can be repeatable in real-user environment
- the term **workload component** or **workload unit** is used instead of the user
- Example workload component:
  - *Applications*:
    If one wants to characterize the behavior of various applications, such as mail, text editing, or program development, then each application may be considered a workload component and the average behavior of each application may be characterized.
  - *Sites*:
    If one desires to characterize the workload at each of several locations of an organization, the sites may be used as workload components.
  - *User Sessions*:
    Complete user sessions from login to logout may be monitored, and applications run during the session may be combined
- The measured quantities, service requests, or resource demands, which are used to model or characterize the workload, are called **workload parameters** or **workload features**

Techniques for Workload Characterization

1. Averaging
2. Specifying dispersion
3. Single-parameter histograms
4. Multiparameter histograms
5. Principal-component analysis
6. Markov models
7. Clustering

AVERAGING

- Arithmetic mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- There are cases, however, when arithmetic mean is inappropriate and the median, mode, geometric mean, or harmonic mean should be used instead

SPECIFYING DISPERSION

- Variability is commonly specified by the variance. It is denoted by s^2 and is computed as follows:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- The ratio of the standard deviation to the mean is called the **Coefficient Of Variation (C.O.V.).**
- A zero C.O.V. implies zero variance and indicates that the measured parameter is a constant

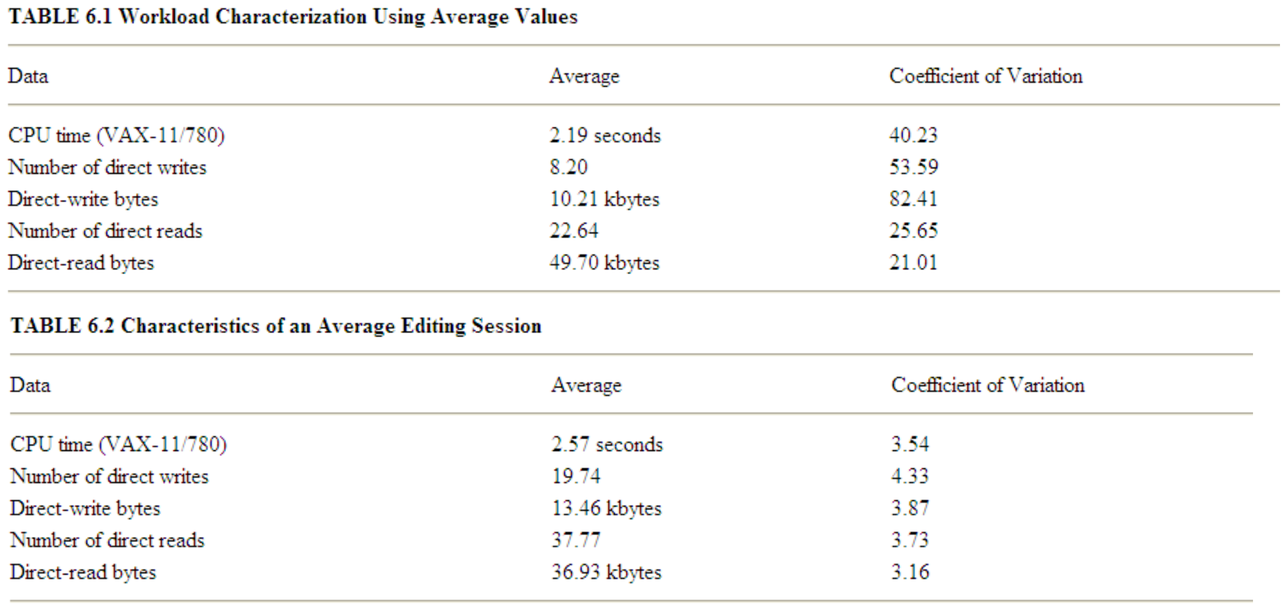


TABLE 6.1 Workload Characterization Using Average Values

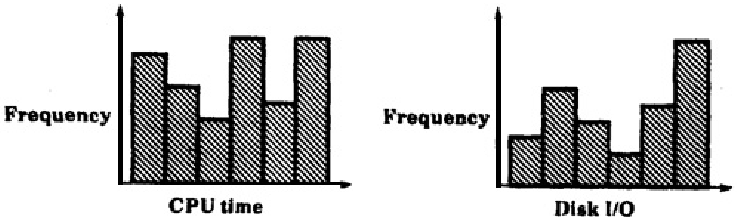
Data	Average	Coefficient of Variation
CPU time (VAX-11/780)	2.19 seconds	40.23
Number of direct writes	8.20	53.59
Direct-write bytes	10.21 kbytes	82.41
Number of direct reads	22.64	25.65
Direct-read bytes	49.70 kbytes	21.01

TABLE 6.2 Characteristics of an Average Editing Session

Data	Average	Coefficient of Variation
CPU time (VAX-11/780)	2.57 seconds	3.54
Number of direct writes	19.74	4.33
Direct-write bytes	13.46 kbytes	3.87
Number of direct reads	37.77	3.73
Direct-read bytes	36.93 kbytes	3.16

Case Study 6.1 The resource demands of various programs executed on six university sites were measured for 6 months. The average demand by each program is shown in Table 6.1. Notice that the C.O.V. of the measured values are rather high, indicating that combining all programs into one class is not a good idea. Programs should be divided into several classes. Table 6.2 shows the average demand for all editors in the same data. The C.O.V. are now much lower.

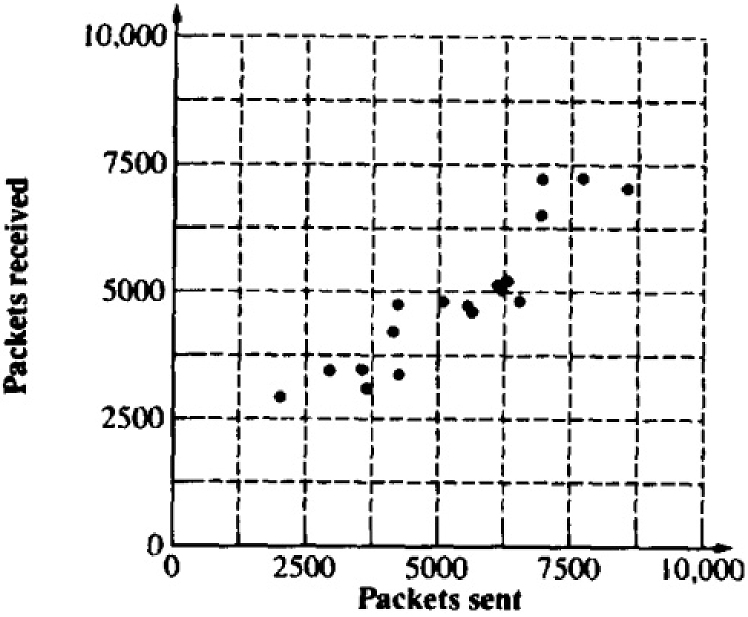
SINGLE-PARAMETER HISTOGRAMS



- A histogram shows the relative frequencies of various values of a parameter. For continuous-value parameters, this requires dividing the complete parameter range into several smaller subranges called *buckets* (or *cells*) and counting the observations that fall in each cell.
- Given *n* buckets per histogram, *m* parameters per component, and *k* components, this method requires presenting *nmk* numerical values. This may be too much detail to be useful. Thus, this should be used only if the variance is high and the averages cannot be used.
- The key problem with using individual-parameter histograms is that they ignore the correlation among different parameters. For example, short jobs (jobs with small elapsed time) may create a lower number of disk I/O and may take a smaller amount of CPU time than long jobs. A test workload based on the single-parameter histograms may generate a job with short CPU time and a large number of disk I/Os—a situation generally not possible in a real workload.

MULTIPARAMETER HISTOGRAMS

- If there is a significant correlation between different workload parameters, the workload should be characterized using multiparameter histograms
- An *n*-dimensional matrix (or histogram) is used to describe the distribution of *n* workload parameters.



PRINCIPAL-COMPONENT ANALYSIS

- One technique commonly used to classify workload components is by the weighted sum of their parameter values. Using a_j as weight for the jth parameter x_j, the weighted sum y is

$$y = \sum_{j=1}^{n} a_j x_j$$

- This sum can then be used to classify the components into a number of classes such as low demand or medium demand.
- One method of determining the weights in such situations is to use the principal-component analysis, which allows finding the weights w_i's such that y_j's provide the maximum discrimination among the components. The quantity y_j is called the **principal factor**.

- Statistically, given a set of n parameters $\{x_1,x_2,\ldots,x_n\}$, the principal-component analysis produces a set of **factors** $\{y_1,y_2,\ldots,y_n\}$ such that the following holds:

1. The y's are linear combinations of x's:

$$y_i = \sum_{j=1}^{n} a_{ij} x_j$$

   Here a_{ij} is called the **loading** of variable x_j on factor y_i.

2. The y's form an orthogonal set, that is, their inner product is zero:

$$\langle y_i, y_j \rangle = \sum_{k} a_{ik} a_{kj} = 0$$

   This is equivalent to stating that the y_i's are uncorrelated to each other.

3. The y's form an ordered set such that y_1 explains the highest percentage of the variance in resource demands, y_2 explains a lower percentage, y_3 explains a still lower percentage, and so forth. Thus, depending upon the level of detail required, only the first few factors can be used to classify the workload components

Example 6.1 The number of packets sent and received, denoted by x_s and x_r, respectively, by various stations on a local-area network were measured. The observed numbers are shown in the second and third columns of Table 6.4. A scatter plot of the data is shown in Figure 6.2. As seen from this figure, there is considerable correlation between the two variables. The steps in determining the principal factors are as follows:

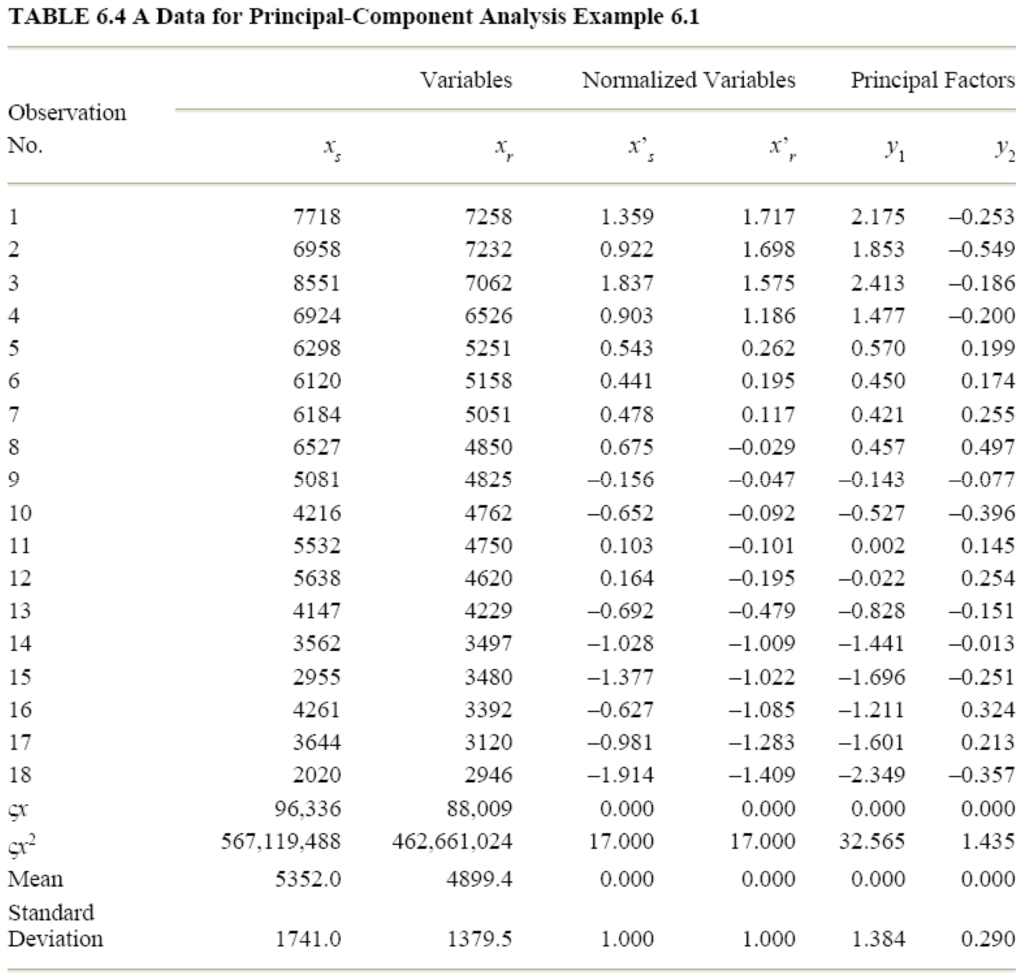


TABLE 6.4 A Data for Principal-Component Analysis Example 6.1

Observation No.	Variables		Normalized Variables		Principal Factors	
	x_s	x_r	x'_s	x'_r	y_1	y_2
1	7718	7258	1.359	1.717	2.175	–0.253
2	6958	7232	0.922	1.698	1.853	–0.549
3	8551	7062	1.837	1.575	2.413	–0.186
4	6924	6526	0.903	1.186	1.477	–0.200
5	6298	5251	0.543	0.262	0.570	0.199
6	6120	5158	0.441	0.195	0.450	0.174
7	6184	5051	0.478	0.117	0.421	0.255
8	6527	4850	0.675	–0.029	0.457	0.497
9	5081	4825	–0.156	–0.047	–0.143	–0.077
10	4216	4762	–0.652	–0.092	–0.527	–0.396
11	5532	4750	0.103	–0.101	0.002	0.145
12	5638	4620	0.164	–0.195	–0.022	0.254
13	4147	4229	–0.692	–0.479	–0.828	–0.151
14	3562	3497	–1.028	–1.009	–1.441	–0.013
15	2955	3480	–1.377	–1.022	–1.696	–0.251
16	4261	3392	–0.627	–1.085	–1.211	0.324
17	3644	3120	–0.981	–1.283	–1.601	0.213
18	2020	2946	–1.914	–1.409	–2.349	–0.357
çx	96,336	88,009	0.000	0.000	0.000	0.000
çx^2	567,119,488	462,661,024	17.000	17.000	32.565	1.435
Mean	5352.0	4899.4	0.000	0.000	0.000	0.000
Standard Deviation	1741.0	1379.5	1.000	1.000	1.384	0.290

1. *Compute the mean and standard deviations of the variables:*

$$\overline{x}_s = \frac{1}{n}\sum_{i=1}^{n} x_{si} = \frac{96,336}{18} = 5352.0$$

$$\overline{x}_r = \frac{1}{n}\sum_{i=1}^{n} x_{ri} = \frac{88,009}{18} = 4889.4$$

$$s^2_{x_s} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{si} - \overline{x}_s)^2$$

$$= \frac{1}{n-1}\left[\left(\sum_{i=1}^{n} x^2_{si}\right) - n\overline{x}^2_s\right]$$

$$= \frac{567,119,488 - 18 \times 5352^2}{17} = 1741.0$$

Similarly

$$s^2_{xr} = \frac{462,661,024 - 18 \times 4889.4^2}{17} = 1379.5$$

2. *Normalize the variables to zero mean and unit standard deviation*. The normalized values x'_s and x'_r, are given by

$$x'_s = \frac{x_s - \overline{x_s}}{s_{Xs}} = \frac{x_s - 5352}{\sqrt{1741}}$$

$$x'_r = \frac{x_r - \overline{x_r}}{s_{xr}} = \frac{x_r - 4889}{\sqrt{1380}}$$

The normalized values are shown in the fourth and fifth columns of Table 6.4.

3. *Compute the correlation among the variables:*

$$R_{x_s,x_r} = \frac{(1/n)\sum_{i=1}^{n}(x_{si} - \overline{x}_s)(x_{ri} - \overline{x}_r)}{s_{x_s}s_{x_r}} = 0.916$$

4. *Prepare the correlation matrix:*

$$\mathbf{C} = \begin{bmatrix} 1.000 & 0.916 \\ 0.916 & 1.000 \end{bmatrix}$$

5. *Compute the eigenvalues of the correlation matrix*. This is done by solving the characteristic equation. Using *I* to denote an identity matrix,

$$|\lambda I - C| = \begin{vmatrix} \lambda - 1 & -0.916 \\ -0.916 & \lambda - 1 \end{vmatrix} = 0$$

or $(\lambda - 1)^2 - 0.916^2 = 0$. The eigenvalues are 1.916 and 0.084.

6. *Compute the eigenvectors of the correlation matrix*. The eigenvector $\mathbf{q_1}$ corresponding to $\lambda_1 = 1.916$ is defined by the following relationship:

$\mathbf{Cq_1 = \lambda_1 q_1}$ or $\begin{bmatrix} 1.000 & 0.916 \\ 0.916 & 1.000 \end{bmatrix} \times \begin{bmatrix} q_{11} \\ q_{21} \end{bmatrix} = 1.916 \begin{bmatrix} q_{11} \\ q_{21} \end{bmatrix}$

or $q_{11} = q_{21}$

7.

- Restricting the length of the eigenvector to 1, the following vector is the first eigenvector:

$$q_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Similarly, the second eigenvector is

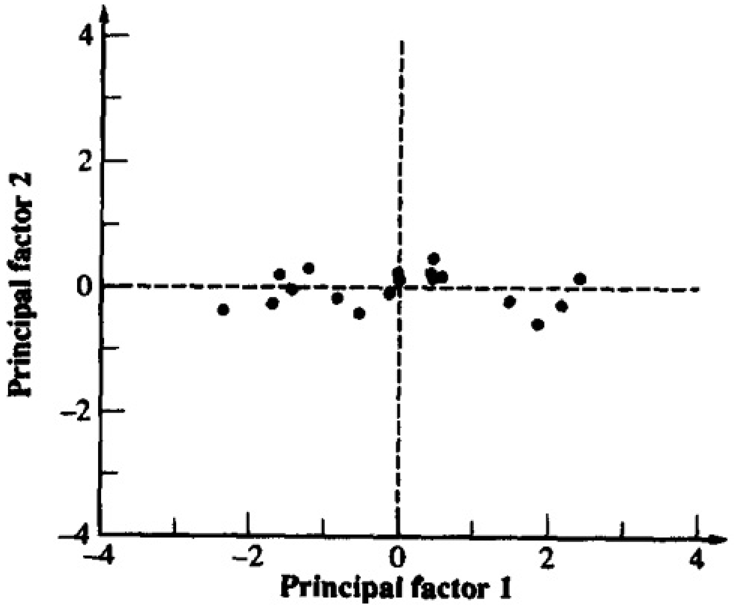
$$q_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}$$

7. *Obtain principal factors by multiplying the eigenvectors by the normalized vectors*

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} q_1^T q_2^T \end{bmatrix} \begin{bmatrix} x_s' \\ x_r' \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{x_s - 5352}{\sqrt{1741}} \\ \frac{x_r - 4889}{\sqrt{1380}} \end{bmatrix}$$

8. *Compute the values of the principal factors*. These are shown in the last two columns of Table 6.4.

9. *Compute the sum and sum of squares of the principal factors*. The sum must be zero. The sum of squares give the percentage of variation explained. In this case, the sums of squares are 32.565 and 1.435. Thus, the first factor explains 32.565/(32.565 + 1.435), or 95.7%, of the variation. The second factor explains only 4.3% of the variation and can thus be ignored.
10. *Plot the values of principal factors*.



MARKOV MODELS

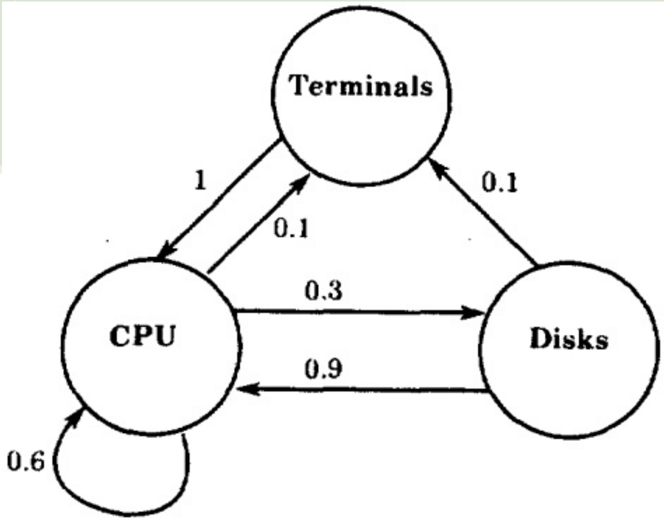
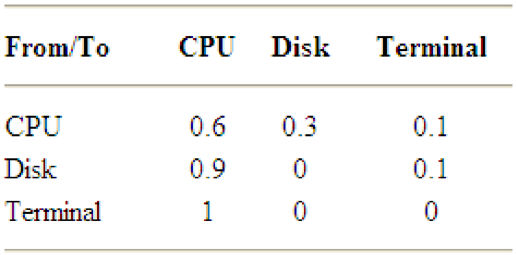


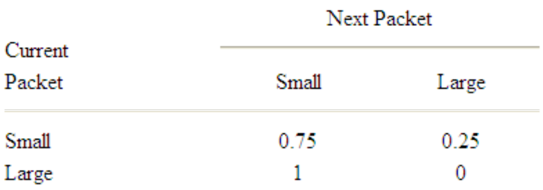
TABLE 6.5 Transition Probability Matrix



From/To	CPU	Disk	Terminal
CPU	0.6	0.3	0.1
Disk	0.9	0	0.1
Terminal	1	0	0

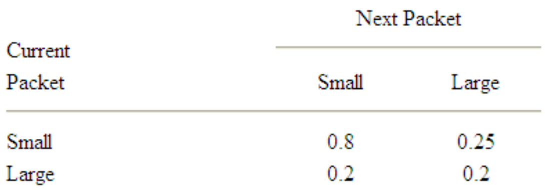
Example 6.2 Traffic monitoring on a computer network showed that most of the packets were of two sizes—small and large. The small packets constituted 80% of the traffic. A number of different transition probability matrices will result in an overall average of 80% of small packets. Two of the possibilities are as follows:

1. An average of four small packets are followed by an average of one big packet. A sample sequence, using s for small and b for big, is ssssbssssbssss. In this sequence, three of the four small packets are followed by another small packet. Also, every big packet is followed by a small packet. The corresponding transition probability matrix is x



Current Packet	Next Packet	
	Small	Large
Small	0.75	0.25
Large	1	0

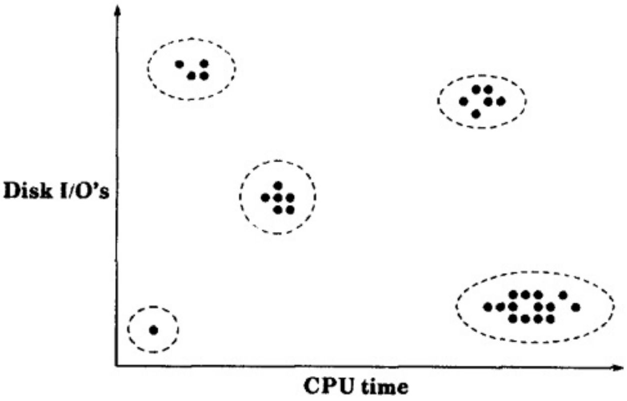
2. Another alternative is to generate a random number between 0 and 1. If the number is less than or equal to 0.8, generate a small packet; otherwise, generate a large packet. This assumes that the next packet size does not depend upon the current packet size. The transition probability matrix in this case is



Current Packet	Next Packet	
	Small	Large
Small	0.8	0.25
Large	0.2	0.2

CLUSTERING

- classify these components into a small number of classes or clusters
- To characterize measured workload data using clustering, the steps are as follows:
  1. Take a sample, that is, a subset of workload components.
  2. Select workload parameters.
  3. Transform parameters, if necessary.
  4. Remove outliers.
  5. Scale all observations.
  6. Select a distance measure.
  7. Perform clustering.
  8. Interpret results.
  9. Change parameters, or number of clusters, and repeat steps 3 to 7.
  10. Select representative components from each cluster.



1. **Sampling**
   One method of sampling is random selection.
2. **Parameter Selection**
   The two key criteria for selecting parameters are their impact on performance and their variance
3. **Transformation**
   If the distribution of a parameter is highly skewed, one should consider the possibility of replacing the parameter by a transformation or function of the parameter. e.g. logarithmic transformation
4. **Outliers**
   The data points with extreme parameter values are called outliers. nly those outlying components that do not consume a significant portion of the system resources should be excluded.
5. **Data Scaling**
   1. *Normalize to Zero Mean and Unit Variance*: $x'_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k}$
   2. *Weights*: $x'_{ik} = w_k x_{ik}$
   3. *Range Normalization* $x'_{ik} = \frac{x_{ik} - x_{\min,k}}{x_{\max,k} - x_{\min,k}}$
   4. *Percentile Normalization*

$$x'_{ik} = \frac{x_{ik} - x_{2.5,k}}{x_{97.5,k} - x_{2.5,k}}$$

Distance Metric

1. *Euclidean Distance*

$$d = \left\{ \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right\}^{0.5}$$

2. *Weighted Euclidean Distance:*

$$d = \sum_{k=1}^{n} \{ a_k (x_{ik} - x_{jk})^2 \}^{0.5}$$

3. *Chi-Square Distance*

$$d = \sum_{k=1}^{n} \left\{ \frac{(x_{ik} - x_{jk})^2}{x_{ik}} \right\}$$

Clustering Techniques

- The basic aim of clustering is to partition the components into groups so the members of a group are as similar as possible and different groups are as dissimilar as possible
- nonhierarchical approaches:
- hierarchical approaches
  - agglomerative
  - divisive