The Art of Data Presentation

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

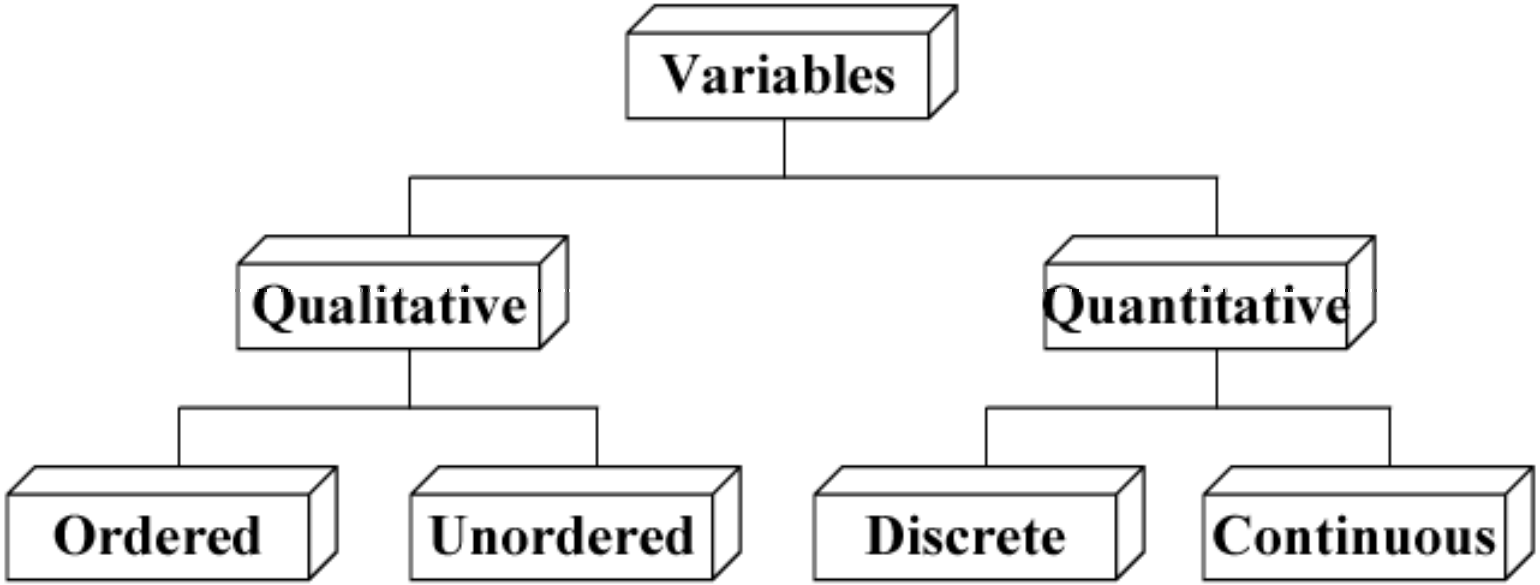
These slides are available on-line at:

http://www.cse.wustl.edu/~jain/cse567-06/

Overview

- Types of Variables
- Guidelines for Preparing Good Charts
- Common Mistakes in Preparing Charts
- Pictorial Games
- Special Charts for Computer Performance
  - Gantt Charts
  - Kiviat Graphs
  - Schumacher Charts
- Decision Maker's Games

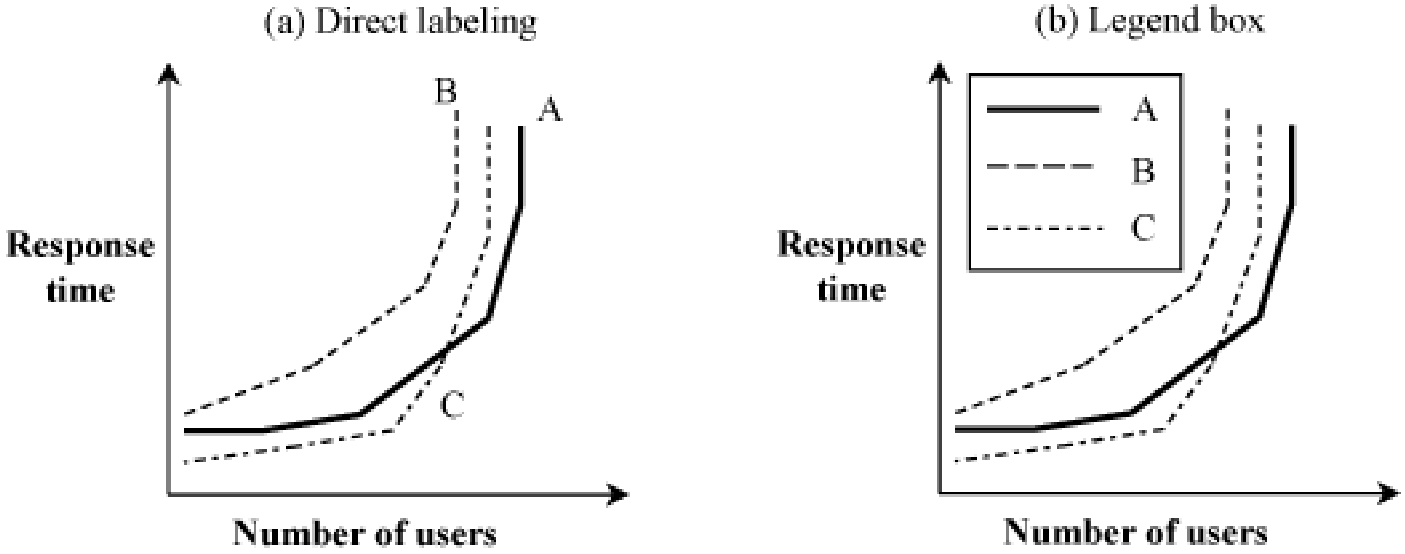
Types of Variables



- Type of computer: Super computer, minicomputer, microcomputer
- Type of Workload: Scientific, engineering, educational
- Number of processors
- Response time of system

Guidelines for Preparing Good Charts

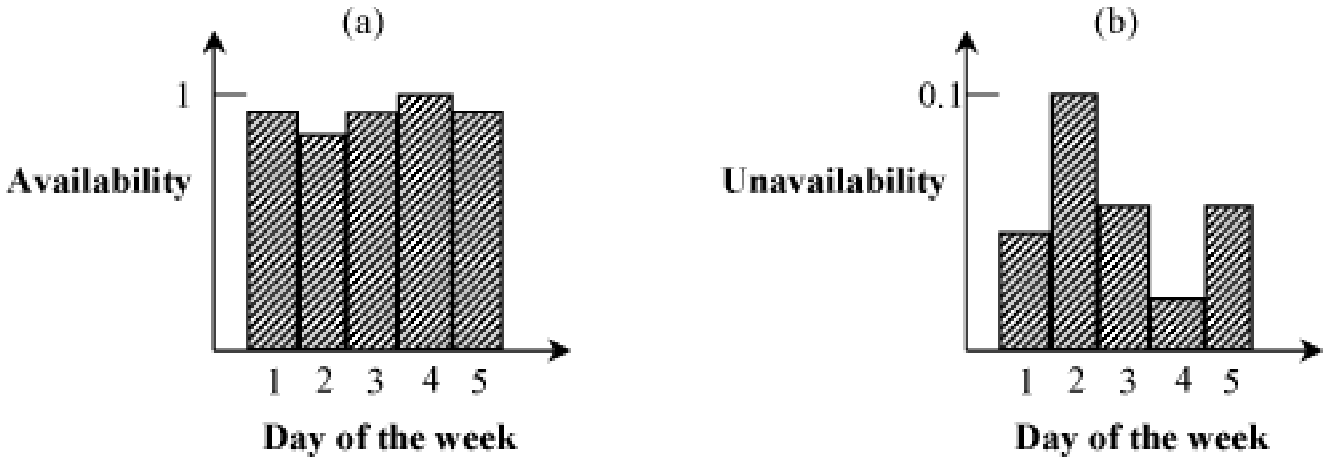
- Require minimum effort from the reader
  Direct labeling vs. legend box



- Maximize Information: Words in place of symbols
  Cleary label the axes

Guidelines (cont)

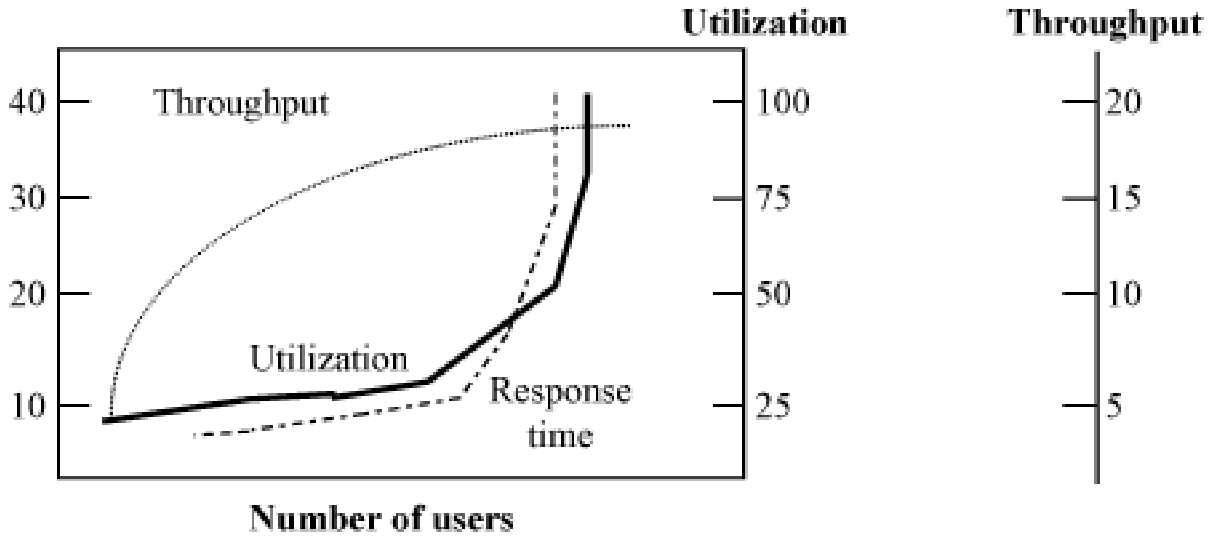
- Minimize Ink: No grid lines, more details



- Use Commonly accepted practices: origin at (0,0) Independent variable (cause) along x axis, linear scales, increasing scales, equal divisions
- Avoid ambiguity: Show coordinate axes, scale divisions, origin. Identify individual curves and bars.
- See checklist in Box 10.1

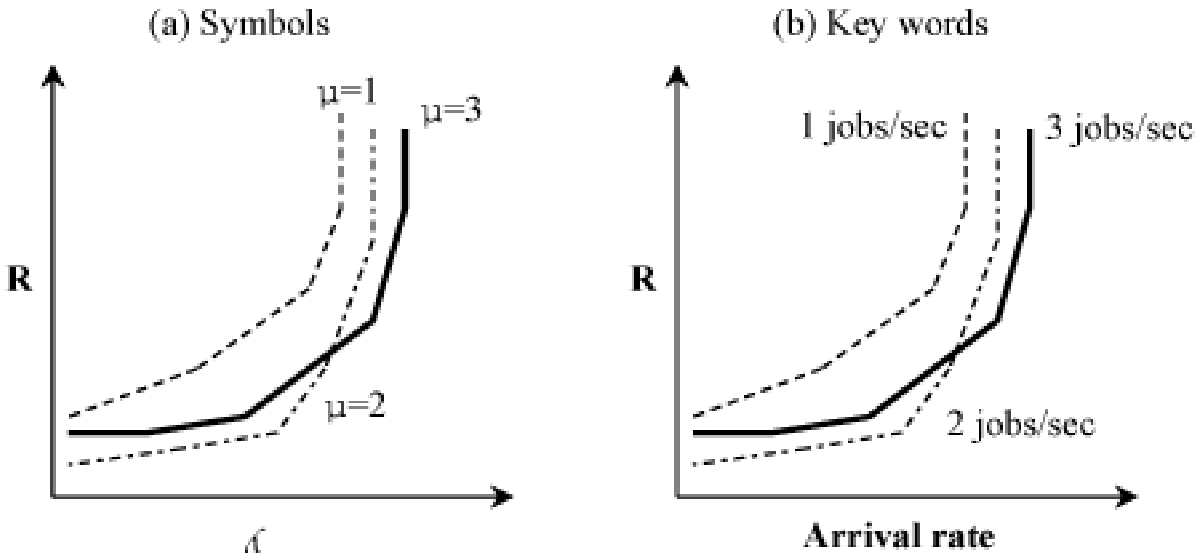
Common Mistakes in Preparing Charts

- Presenting too many alternatives on a single chart
  Max 5 to 7 messages => Max 6 curves in a line charts, no more than 10 bars in a bar chart, max 8 components in a pie chart
- Presenting many y variables on a single chart



Common Mistakes in Charts (Cont)

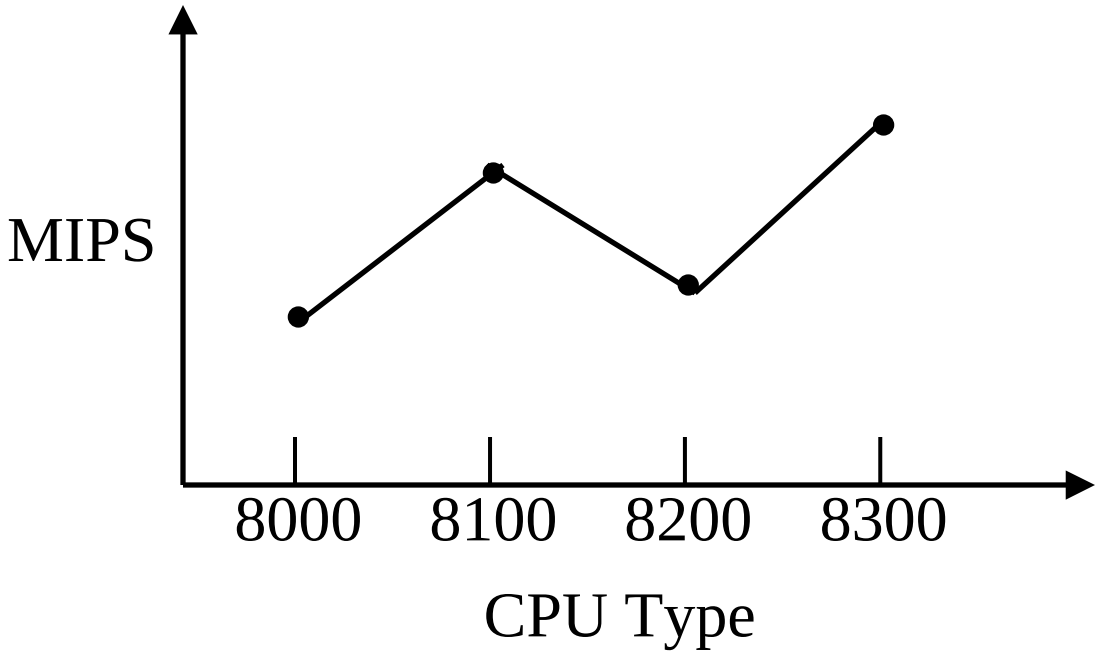
- Using symbols in place of text



- Placing extraneous information on the chart: grid lines, granularity of the grid lines
- Selecting scale ranges improperly: automatic selection by programs may not be appropriate

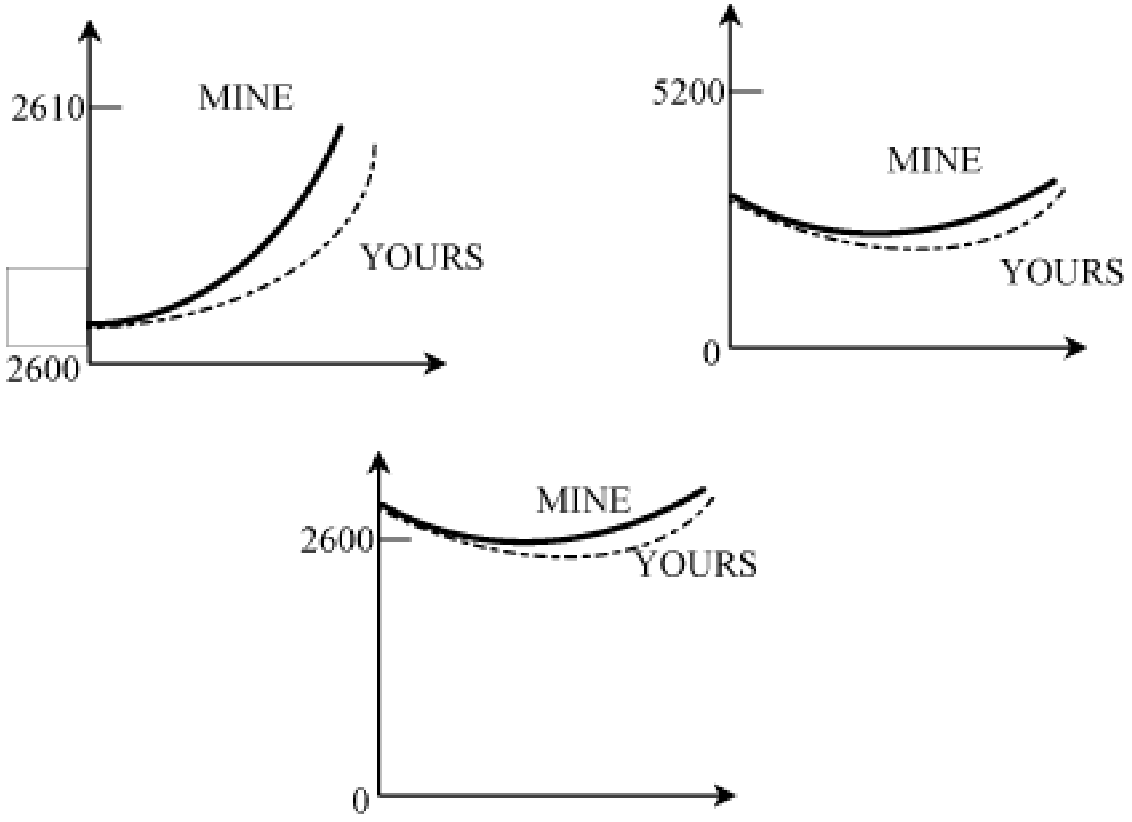
Common Mistakes in Charts (Cont)

- Using a line chart in place of column chart: line => Continuity



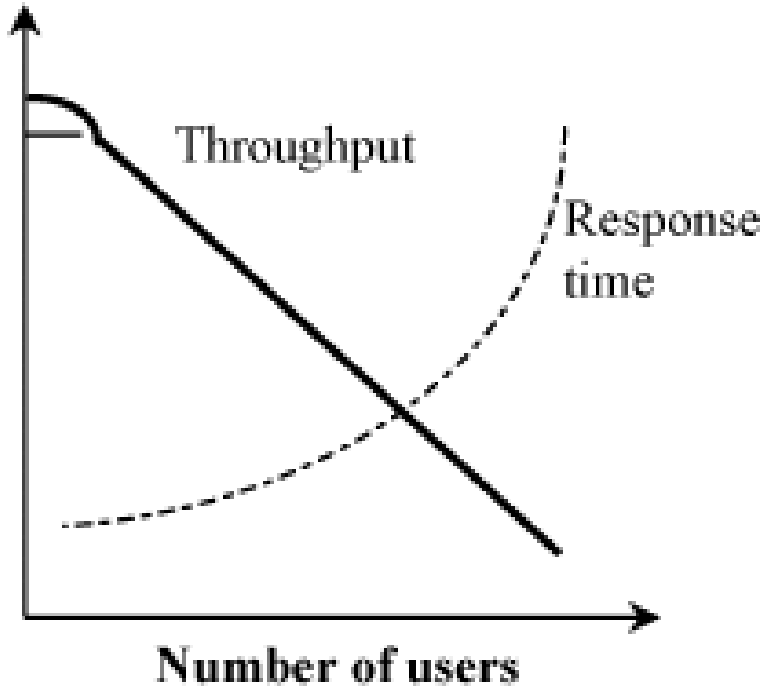
Pictorial Games

- Using non-zero origins to emphasize the difference
  Three quarter high-rule => height/width > 3/4



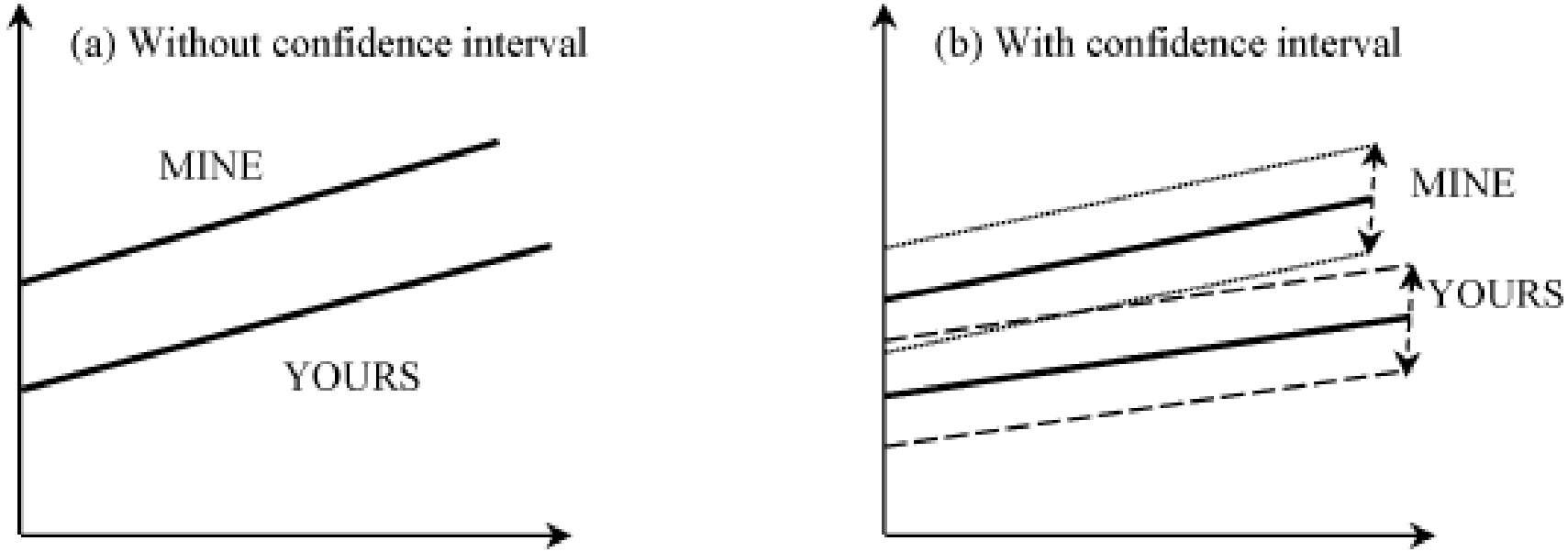
Pictorial Games (Cont)

- Using double-whammy graph for dramatization
  Using related metrics



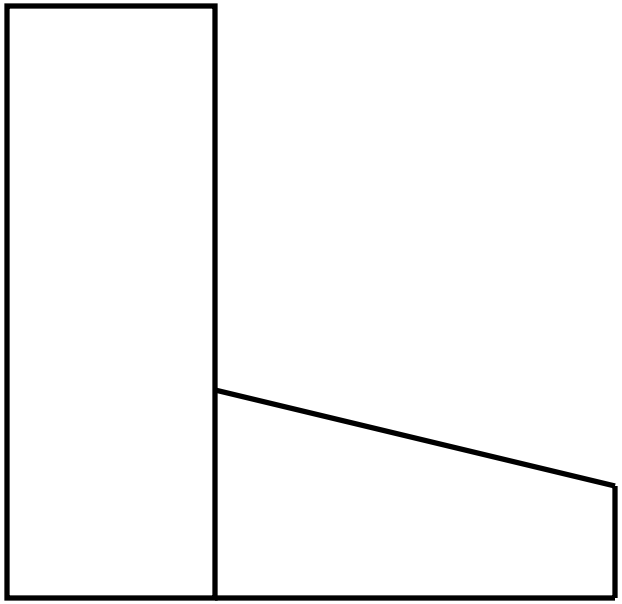
Pictorial Games (Cont)

- Plotting random quantities without showing confidence intervals

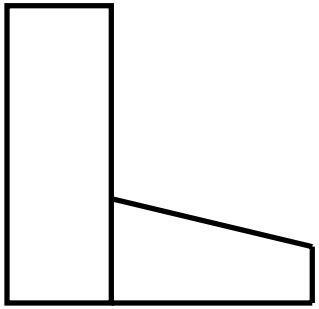


Pictorial Games (Cont)

- Pictograms scaled by height



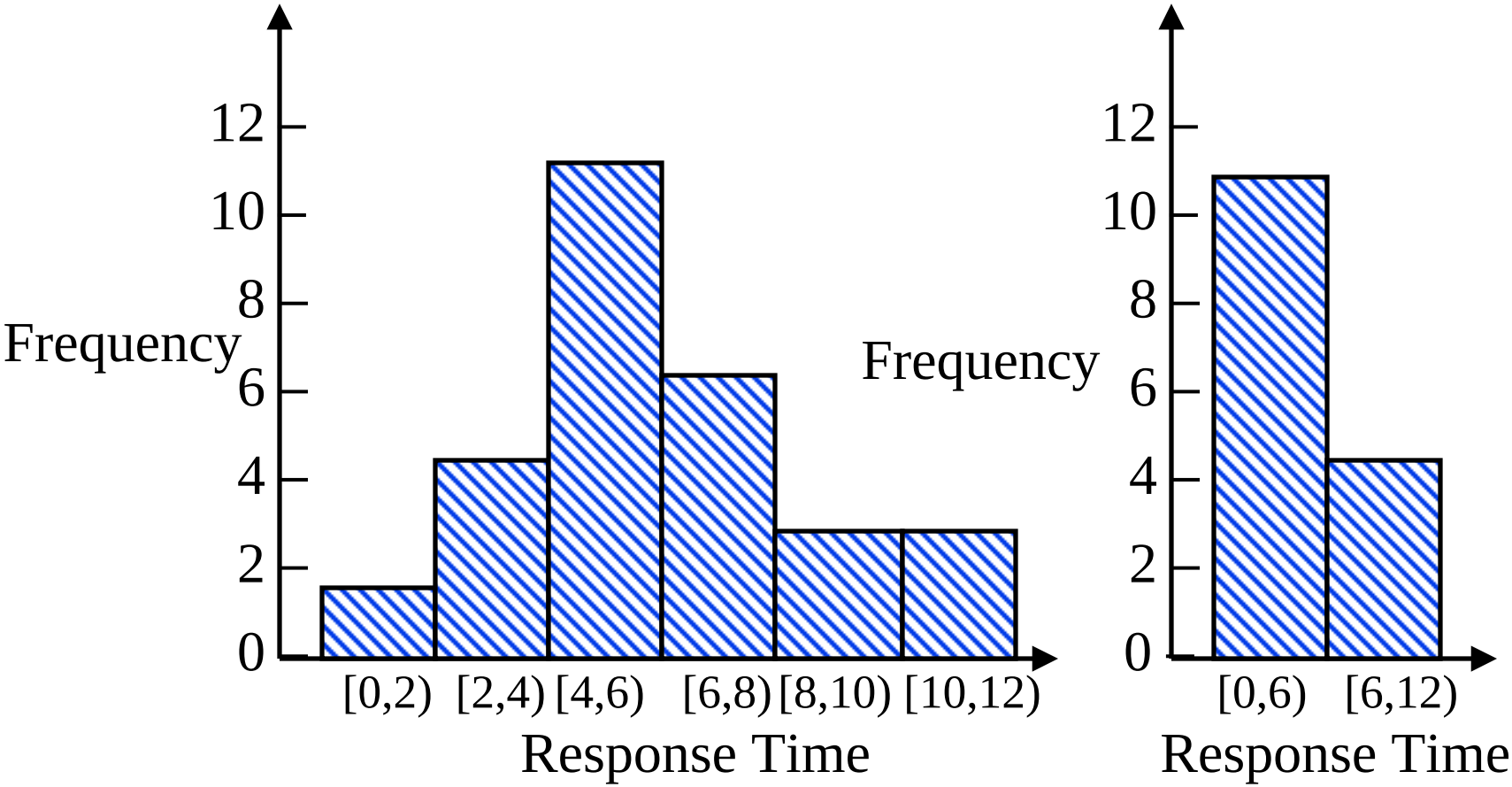
Mine
Performance = 2



Yours
Performance = 1

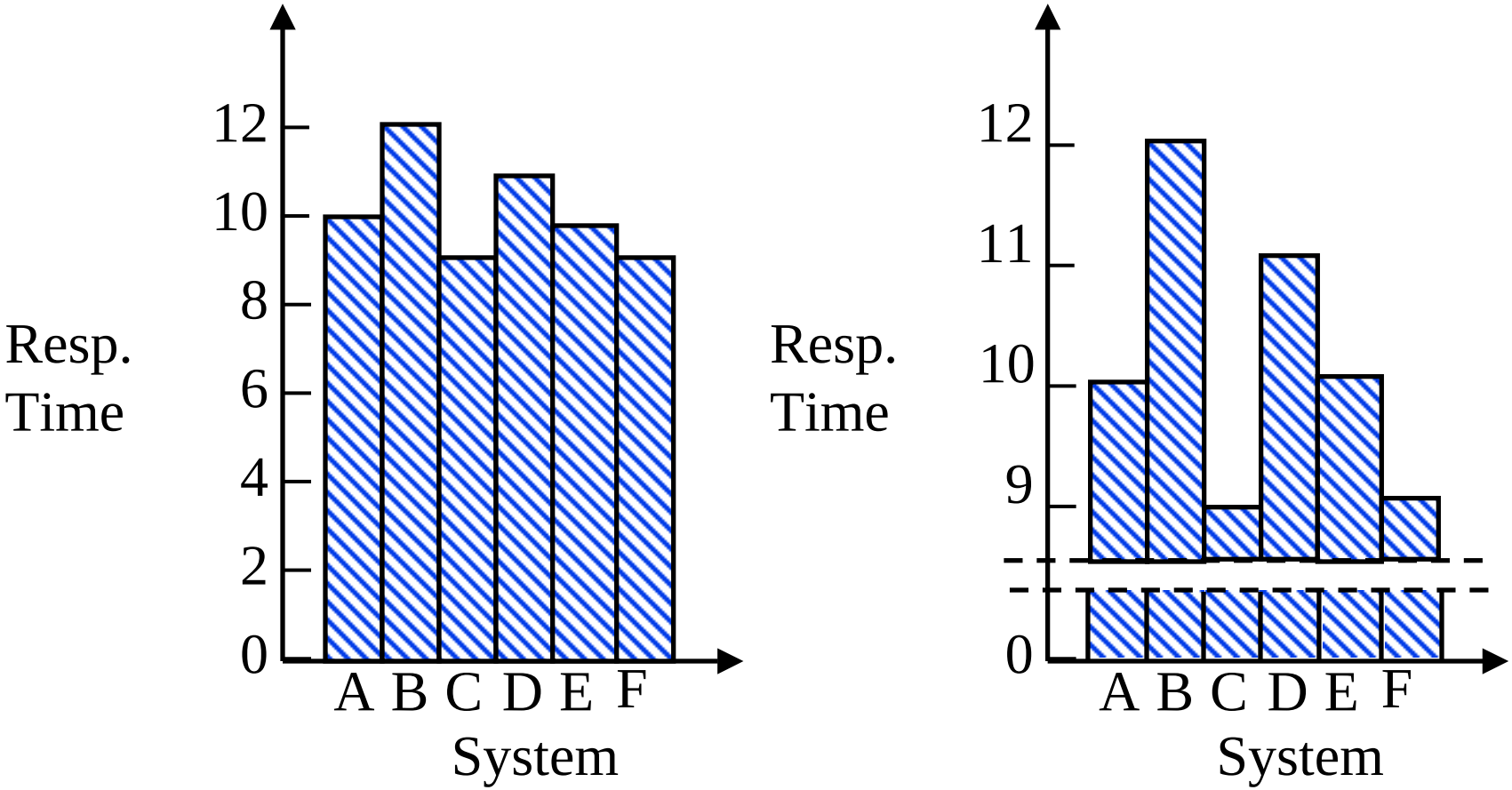
Pictorial Games (Cont)

- Using inappropriate cell size in histograms



Pictorial Games (Cont)

- Using broken scales in column charts

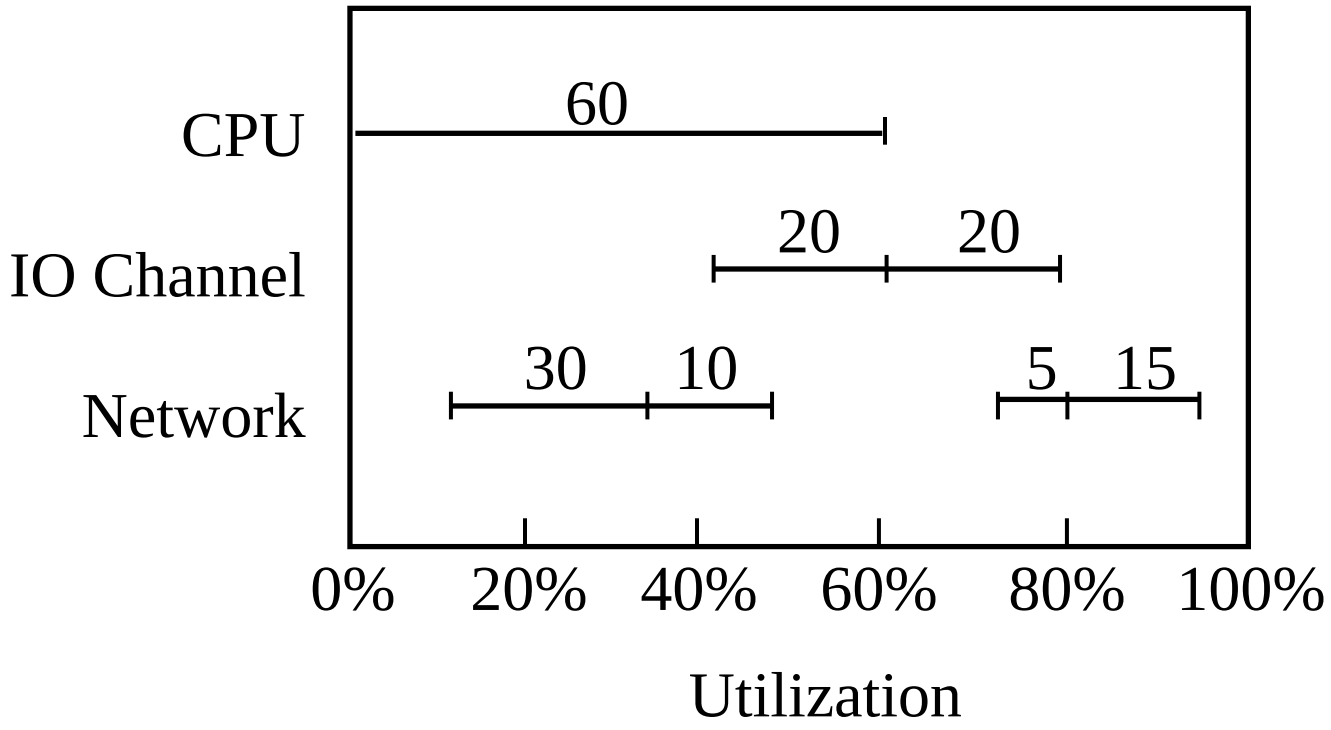


Special Charts for Computer Performance

- Gantt charts
- Kiviat Graphs
- Schumacher's charts

Gantt Charts

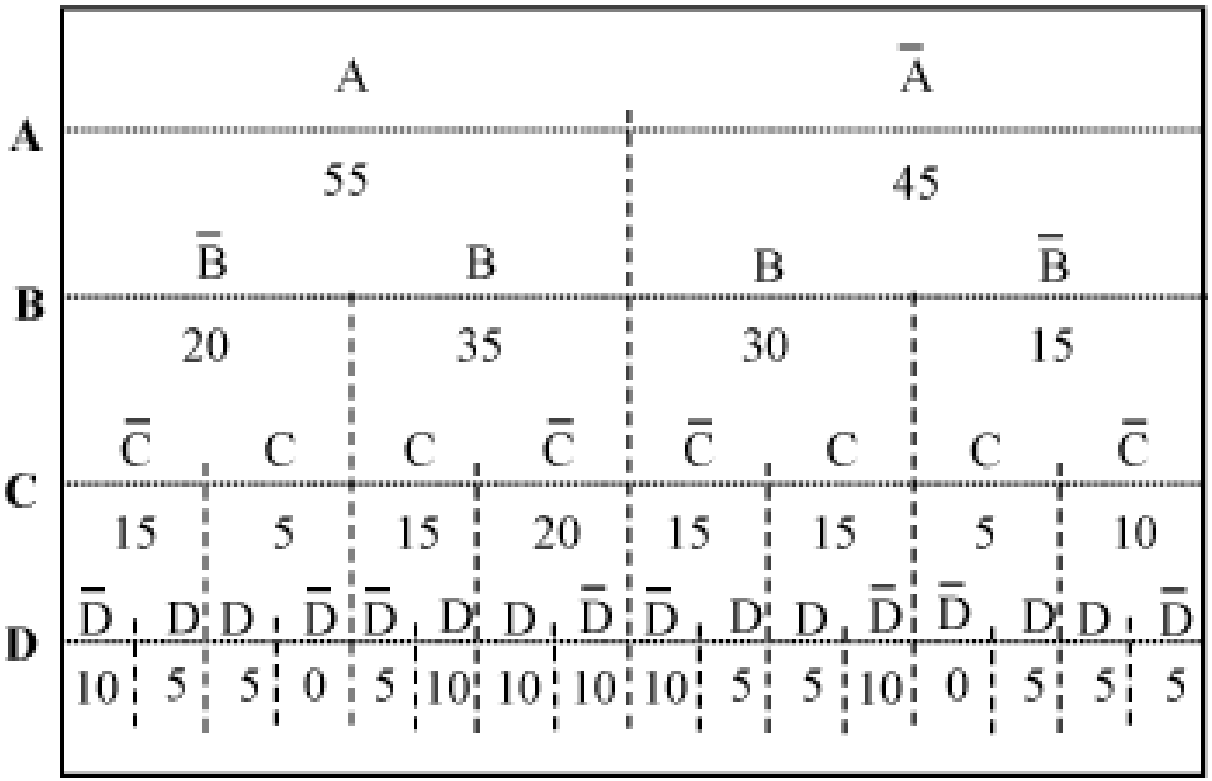
- Shows relative duration of a number of conditions



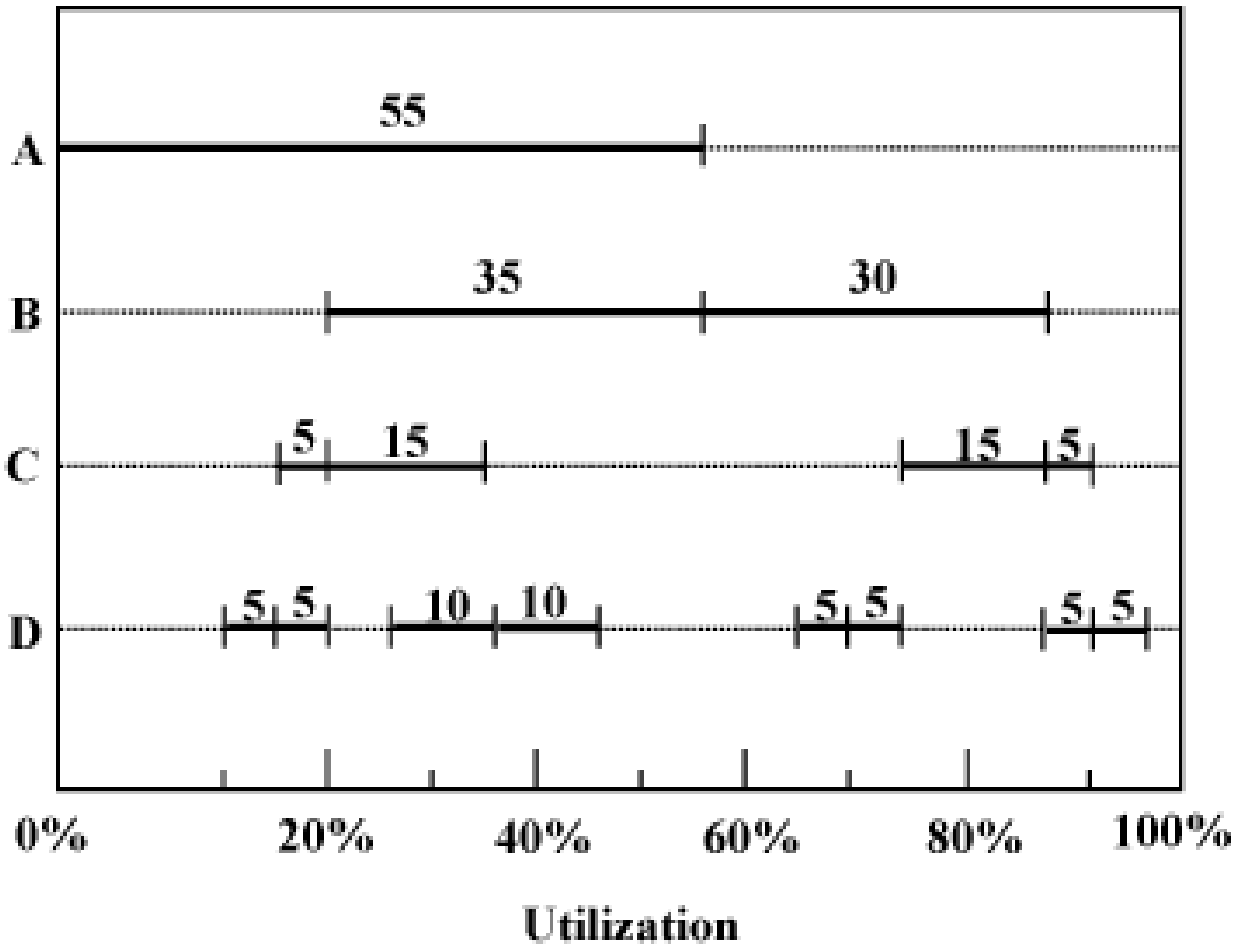
Example: Data for Gantt Chart

A	B	C	D	Time Used
0	0	0	0	5%
0	0	0	1	5%
0	0	1	0	0%
0	0	1	1	5%
0	1	0	0	10%
0	1	0	1	5%
0	1	1	0	10%
0	1	1	1	5%
1	0	0	0	10%
1	0	0	1	5%
1	0	1	0	0%
1	0	1	1	5%
1	1	0	0	10%
1	1	0	1	10%
1	1	1	0	5%
1	1	1	1	10%
			Total	100%

Draft of the Gantt Chart

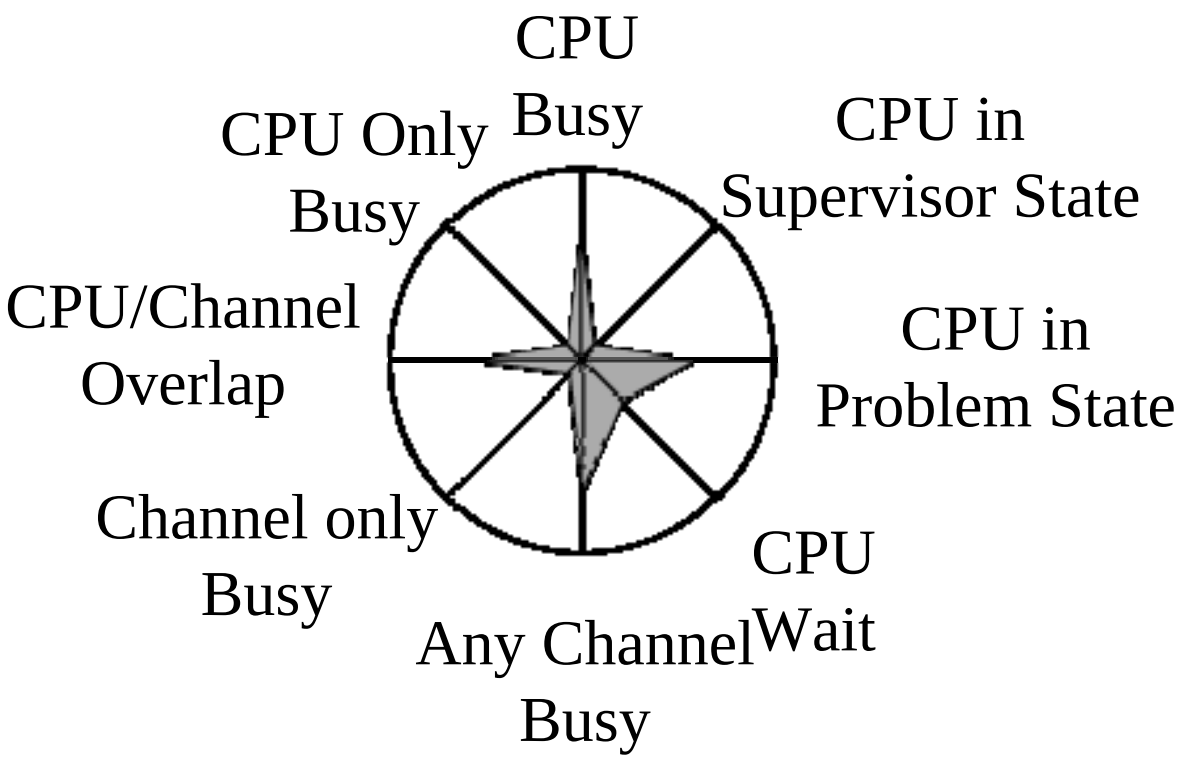


Final Gantt Chart

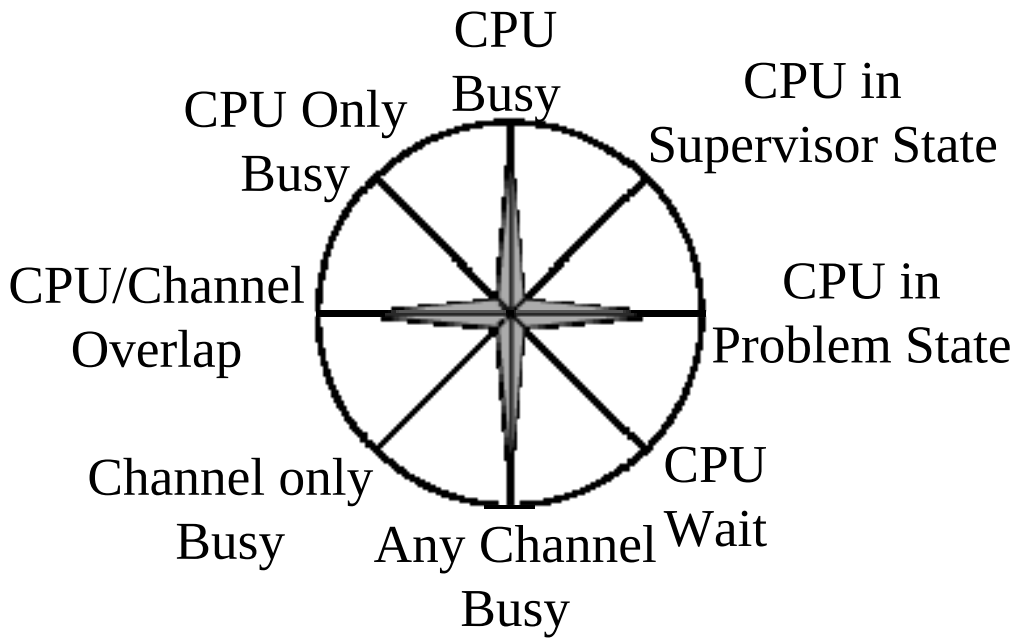


Kiviat Graphs

- Radial chart with even number of metrics
- HB and LB metrics alternate
- Ideal shape: star



Kiviat Graph for a Balanced System



- **Problem**: Inter-related metrics

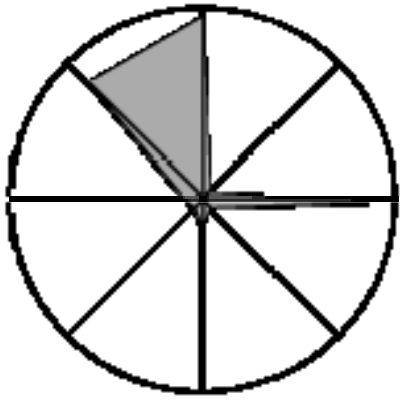
CPU busy = problem state + Supervisor state

CPU wait = 100 – CPU busy

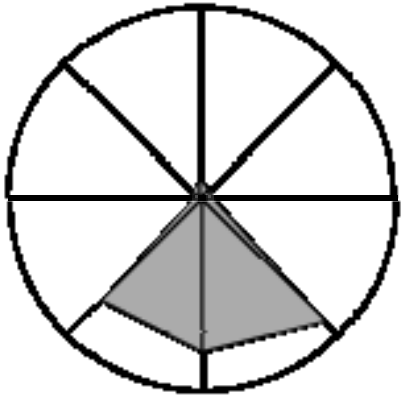
Channel only – any channel –CPU/channel overlap

CPU only = CPU busy – CPU/channel overlap

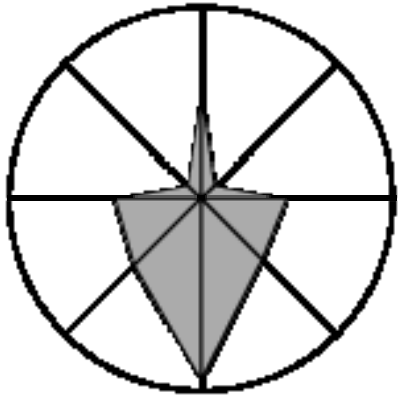
Shapes of Kiviat Graphs



CPU Keel boat



I/O Wedge



I/O Arrow

Merrill's Figure of Merit (FoM)

- Performance = $\{x_1, x_2, x_3, \ldots, x_{2n}\}$
  Odd values are HB and even values are LB

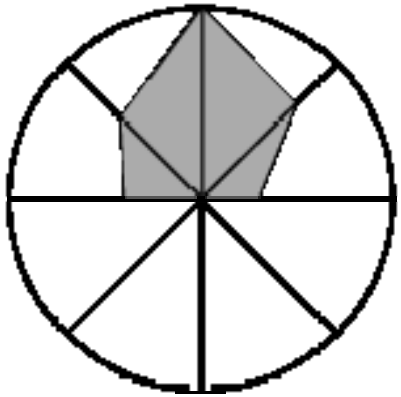
$$\text{FOM} = \left[\frac{1}{2n}\sum_{i=1}^{n}(x_{2i-1} + x_{2i+1})(100 - x_{2i})\right]^{1/2}$$

- x_{2n+1} is the same as x_1
- Average FOM = 50%

Example: FoM

- System A:

System	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A	100	60	40	0	40	0	40	60
B	70	30	40	30	70	30	40	30

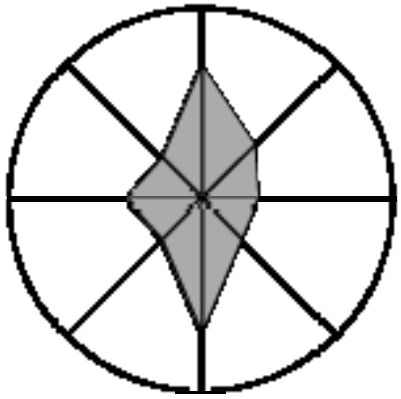


$$\begin{aligned}
\text{FOM}_A &= \Big[\frac{1}{8}\{(100+40)(100-60)+(40+40)(100-0)\\
&\quad +(40+40)(100-0)+(40+100)(100-60)\}\Big]^{1/2}\\
&= \left[\frac{5600+8000+8000+5600}{8}\right]^{1/2}\\
&= \sqrt{\frac{27,200}{8}} = 58
\end{aligned}$$

FoM Example (Cont)

- System B:

System	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
A	100	60	40	0	40	0	40	60
B	70	30	40	30	70	30	40	30



$$\begin{aligned}
\text{FOM}_B &= \left[\frac{1}{8}\{(70+40)(100-30)+(40+70)(100-30)\right. \\
&\quad \left.+(70+40)(100-30)+(40+70)(100-30)\}\right]^{1/2} \\
&= \left[\frac{7700+7700+7700+7700}{8}\right]^{1/2} \\
&= \sqrt{\frac{30,800}{8}} = 62
\end{aligned}$$

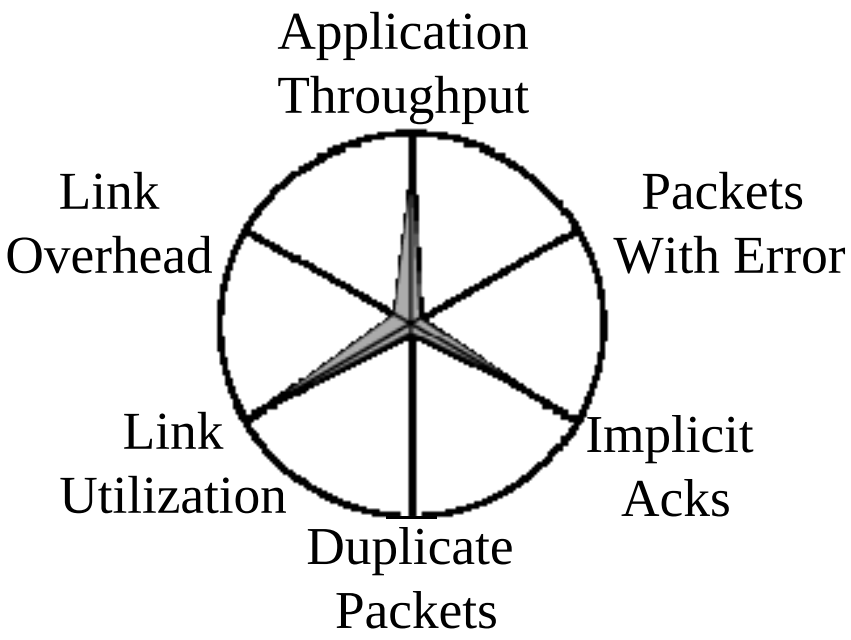
System B has a higher figure of merit and it is better.

Figure of Merit: Known Problems

- All axes are considered equal
- Extreme values are assumed to be better
- Utility is not a linear function of FoM
- Two systems with the same FoM are not equally good.
- System with slightly lower FoM may be better

Kiviat Graphs For Other Systems

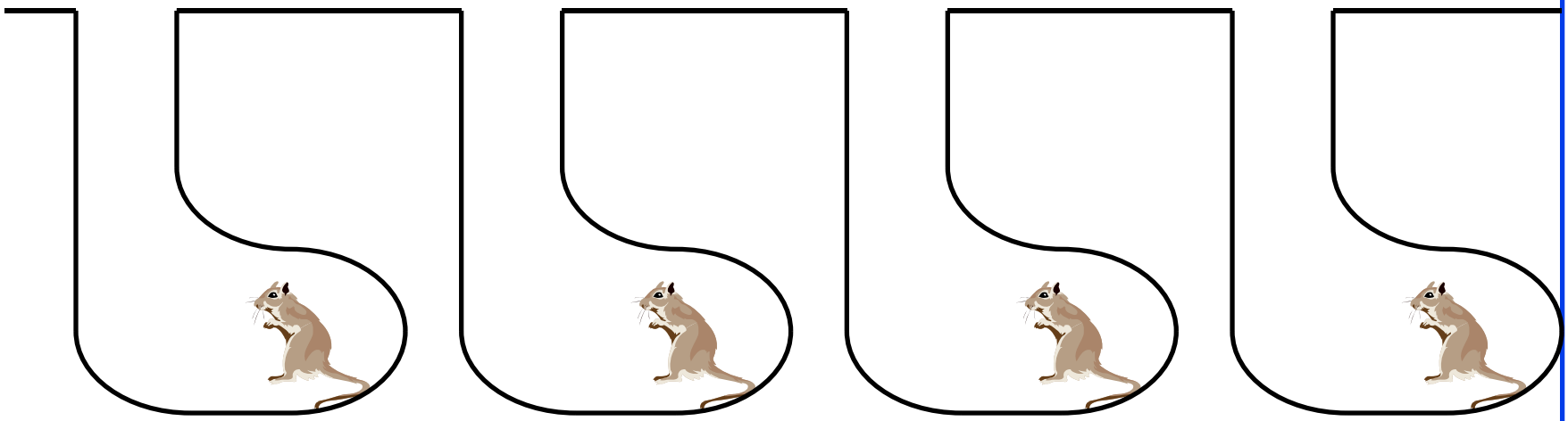
- Networks:



Schumacher Charts

- Performance matrix are plotted in a tabular manner
- Values are normalized with respect to long term means and standard deviations
- Any observations that are beyond mean ± one standard deviation need to be explained
- See Figure 10.25 in the book

Performance Analysis Rat Holes

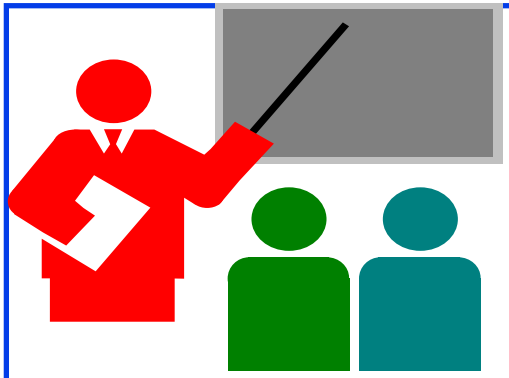


Workload Metrics Configuration Details

Reasons for not Accepting an Analysis

- This needs more analysis.
- You need a better understanding of the workload.
- It improves performance only for long IOs/packets/jobs/files, and most of the IOs/packets/jobs/files are short.
- It improves performance only for short IOs/packets/jobs/files, but who cares for the performance of short IOs/packets/jobs/files, its the long ones that impact the system.
- It needs too much memory/CPU/bandwidth and memory/CPU/bandwidth isn't free.
- It only saves us memory/CPU/bandwidth and memory/CPU/bandwidth is cheap.

See Box 10.2 on page 162 of the book for a complete list



Summary

1. Qualitative/quantitative, ordered/unordered, discrete/continuous variables
2. Good charts should require minimum effort from the reader and provide maximum information with minimum ink
3. Use no more than 5-6 curves, select ranges properly, Three-quarter high rule
4. Gantt Charts show utilizations of various components
5. Kiviat Graphs show HB and LB metrics alternatively on a circular graph
6. Schumacher Charts show mean and standard deviations
7. Workload, metrics, configuration, and details can always be challenged. Should be carefully selected.

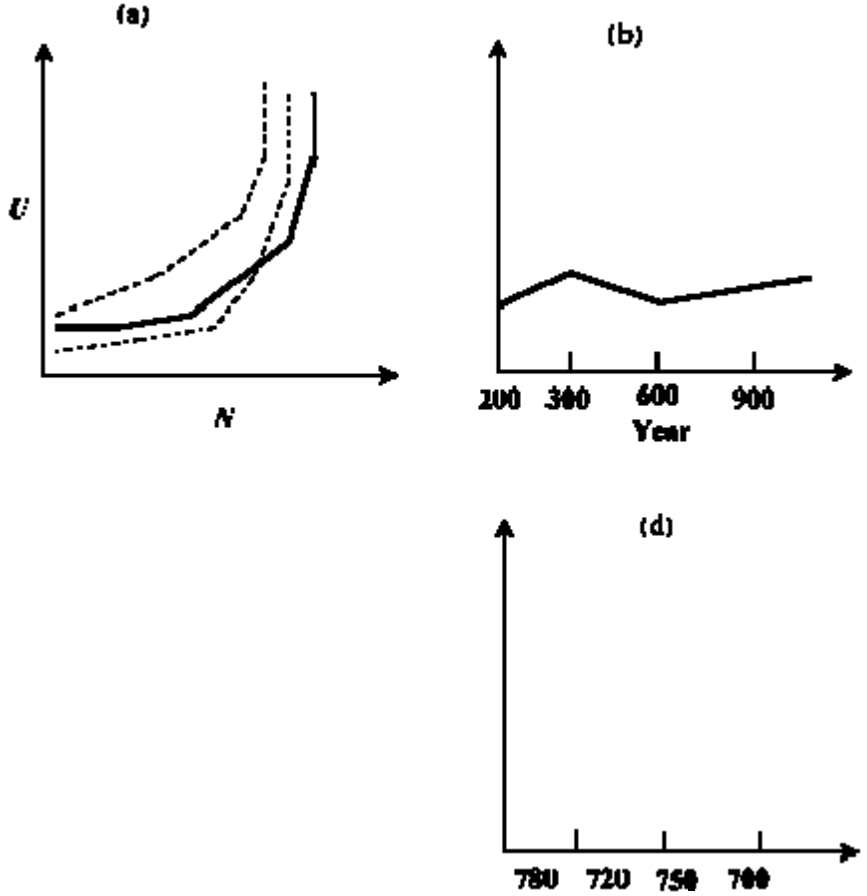
Exercise 10.1

What type of chart (line or bar) would you use to plot:

a. CPU usage for 12 months of the year
b. CPU usage as a function of time in months
c. Number of I/O's to three disk drives: A, B, and C
d. Number of I/O's as a function of number of disk drives in a system

Exercise 10.2

- List the problems with the following charts



Exercise 10.3

- On a system consisting of 3 resources, called A, B, and C. The measured utilizations are shown in the following table. A zero in a column indicates that the resource is not utilized. Draw a Gantt chart showing utilization profiles.

A	B	C	Time Used
0	0	0	25%
0	0	1	10%
0	1	0	20%
0	1	1	5%
1	0	0	5%
1	0	1	15%
1	1	0	5%
1	1	1	15%
	Total		100%

Exercise 10.4

- The measured values of the eight performance metrics listed in Example 10.2 for a system are: 70%, 10%, 60%, 20%, 80%, 30%, 50%, and 20%. Draw the Kiviat graph and compute its figure of merit.

Exercise 10.5

- For a computer system of your choice, list a number of HB and LB metrics and draw a typical Kiviat graph using data values of your choice.

Homework

- Read Chapter 10
- Submit solutions to exercises 10.3 and 10.4
  Approximate hand-drawn figures are sufficient