Presentation of Scientific Results

Arnaud Legrand and Jean-Marc Vincent

Scientific Methodology and Performance Evaluation
ENS Lyon, November 2016

Outline

1. Data Visualization
   - Motivation
   - Jain, Chapter 10

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$

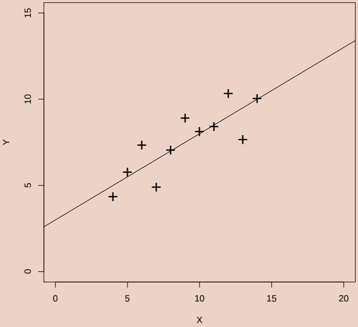
Correlation $= 0.816$

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.$
Mean of $Y = 7$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Scatter plot

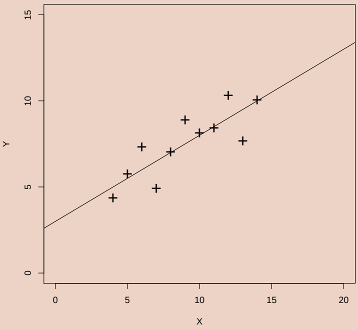


Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9$
Mean of $Y = 7$
Intercept = 3
Slope = 0.5
Res. stdev = 1.237
Correlation = 0.816

Scatter plot



1. The data set "behaves like" a linear curve with some scatter;
2. There is no justification for a more complicated model (e.g., quadratic);
3. There are no outliers;
4. The vertical spread of the data appears to be of equal height irrespective of the X-value; this indicates that the data are equally-precise throughout and so a "regular" (that is, equi-weighted) fit is appropriate.

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(2)}$	$Y^{(2)}$
10.00	9.14
8.00	8.14
13.00	8.74
9.00	8.77
11.00	9.26
14.00	8.10
6.00	6.13
4.00	3.10
12.00	9.13
7.00	7.26
5.00	4.74

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(3)}$	$Y^{(3)}$
10.00	7.46
8.00	6.77
13.00	12.74
9.00	7.11
11.00	7.81
14.00	8.84
6.00	6.08
4.00	5.39
12.00	8.15
7.00	6.42
5.00	5.73

$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(4)}$	$Y^{(4)}$
8.00	6.58
8.00	5.76
8.00	7.71
8.00	8.84
8.00	8.47
8.00	7.04
8.00	5.25
19.00	12.50
8.00	5.56
8.00	7.91
8.00	6.89

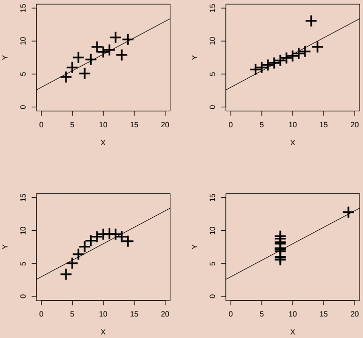
$N = 11$ samples
Mean of $X = 9.0$
Mean of $Y = 7.5$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.$
Mean of $Y = 7$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Scatter plot



$X^{(2)}$	$Y^{(2)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(3)}$	$Y^{(3)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(4)}$	$Y^{(4)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Why do we need to visualize ? The Anscombe's Quartet

$X^{(1)}$	$Y^{(1)}$
10.00	8.04
8.00	6.95
13.00	7.58
9.00	8.81
11.00	8.33
14.00	9.96
6.00	7.24
4.00	4.26
12.00	10.24
7.00	4.82
5.00	5.68

$N = 11$ samples
Mean of $X = 9.$
Mean of $Y = 7$
Intercept $= 3$
Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(2)}$	$Y^{(2)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

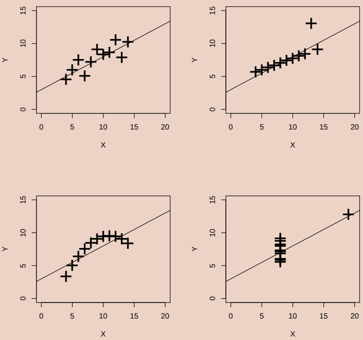
$X^{(3)}$	$Y^{(3)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

$X^{(4)}$	$Y^{(4)}$

Slope $= 0.5$
Res. stdev $= 1.237$
Correlation $= 0.816$

Scatter plot



1. data set 1 is clearly linear with some scatter.
2. data set 2 is clearly quadratic.
3. data set 3 clearly has an outlier.
4. data set 4 is obviously the victim of a poor experimental design with a single point far removed from the bulk of the data "wagging the dog".

Problem statement

- All analysis we perform rely on (sometimes implicit) assumptions. If these assumptions do not hold, the analysis will be a complete nonsense.
- Checking these assumptions is not always easy and sometimes, it may even be difficult to list all these assumptions and formally state them.

A visualization can help to check these assumptions.

- Visual representation resort to our cognitive faculties to check properties.

  The visualization is meant to let us detect expected and unexpected behavior with respect to a given model.

Using the “right” representations

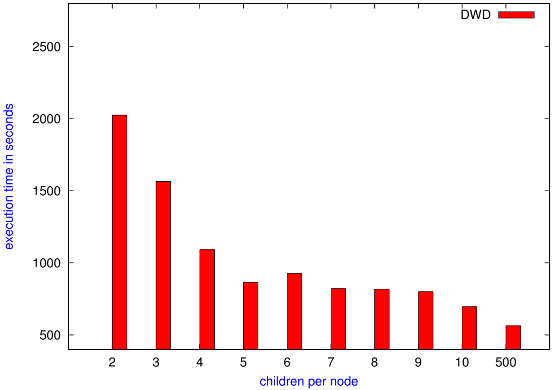
- The problem is to represent on a limited space, typically a screen with a fixed resolution, a meaningful information about the behavior of an application or system.
- $\rightsquigarrow$ need to aggregate data and be aware of what information loss this incurs.
- Every visualization emphasizes some characteristics and hides others. Being aware of the underlying models helps choosing the right representation.

Visualization and intuition

- Visualization can also be used to guide your intuition.
  Sometimes, you do not know exactly what you are looking for and looking at the data just helps.
- Some techniques (Exploratory Data Analysis) even build on this and propose to summarize main characteristics in easy-to-understand form, often with visual graphs, without using a statistical model or having formulated a hypothesis.
- **Use with care**, visualizations always have underlying models: when visualization is not adapted, what you may observe may be meaningless. Such approaches may help formulating hypothesis but these hypothesis have then to be tested upon new data-sets.

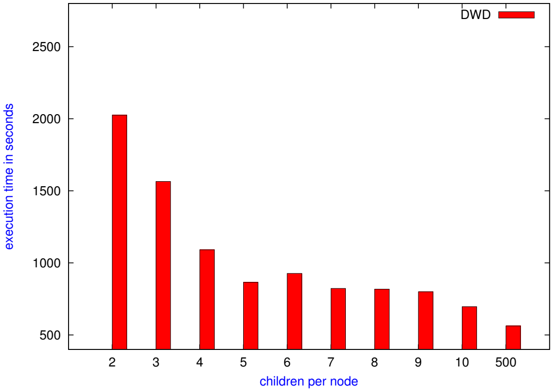
A “simple” graphical check for investigating scalability

Plotting T_p versus p.



A “simple” graphical check for investigating scalability

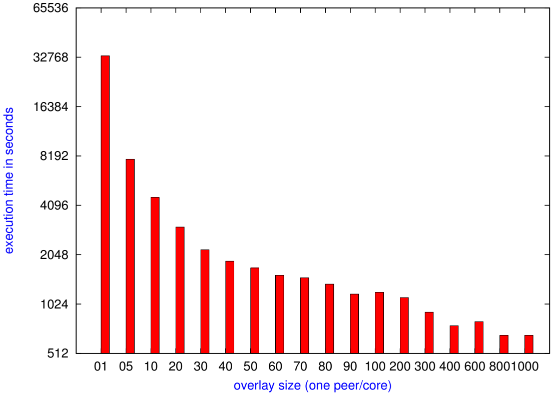
Plotting T_p versus p.



- y-axis does not start at 0, which makes speedup look more impressive
- x-axis is linear with an outlier.

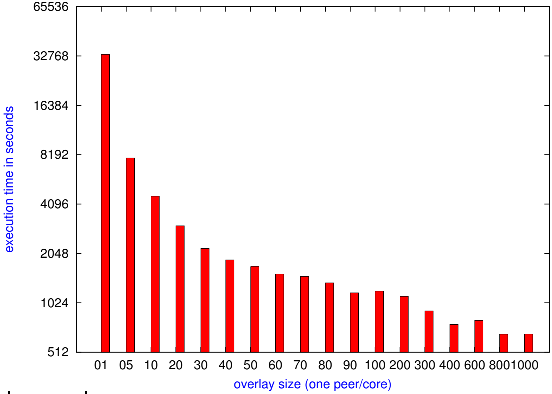
A "simple" graphical check for investigating scalability

Plotting T_p versus p.



A “simple” graphical check for investigating scalability

Plotting T_p versus p.

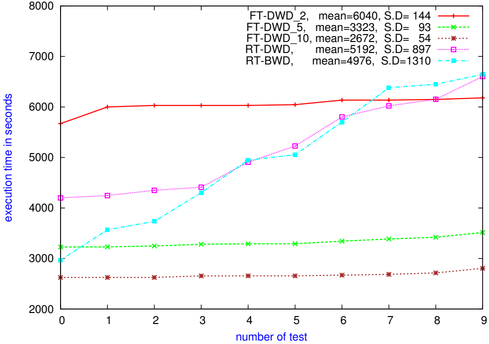


- y-axis uses log-scale
- x-axis is neither linear nor logarithmic so we cannot reason about the shape of the curve

Say, we want to test for Amhdal’s law. Propose a better representation.

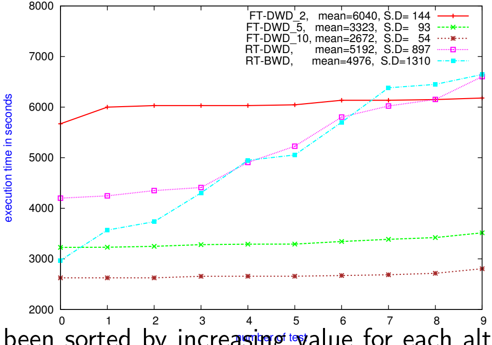
Graphically checking which alternative is better ?

5 different alternatives (FT-DWD_2, FT-DWD_5, FT-DWD_10, RT-DWD, RT-BWD), each tested 10 times.



Graphically checking which alternative is better ?

5 different alternatives (`FT-DWD_2`, `FT-DWD_5`, `FT-DWD_10`, `RT-DWD`, `RT-BWD`), each tested 10 times.



Outcomes have been sorted by increasing value for each alternative and are then linked together

- The shape of the lines do not make any sense. The lines group related values
- Experiment order does not make any sense and makes it look like alternatives have been evaluated in 10 different settings (, which suggests the values can be compared with each others for each setting)

Propose a better representation

Outline

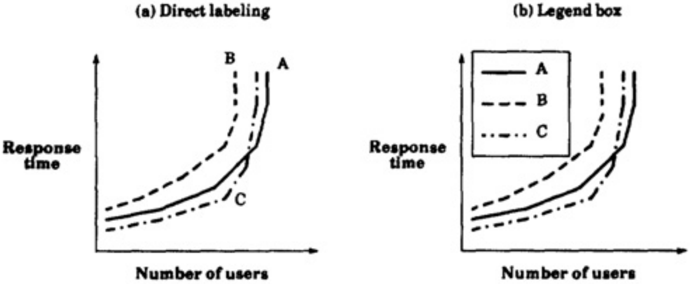
1. Data Visualization
   - Motivation
   - Jain, Chapter 10

Read the basics

- For all such kind of “general” graphs where you summarize the results of several experiments, the very least you need to read is Jain’s book: The Art of Computer Systems Performance Analysis. A new edition is expected in sept. 2015
- It has check lists for “Good graphics”, which I made more or less available on the lecture’s webpage
- It presents the most common pitfalls in data representation
- It will teach how to cheat with your figures. . .
- . . . and how to detect cheaters. ;)

Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)

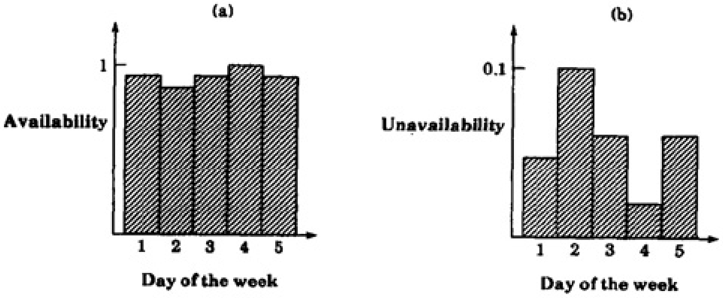


(a) Direct labeling

(b) Legend box

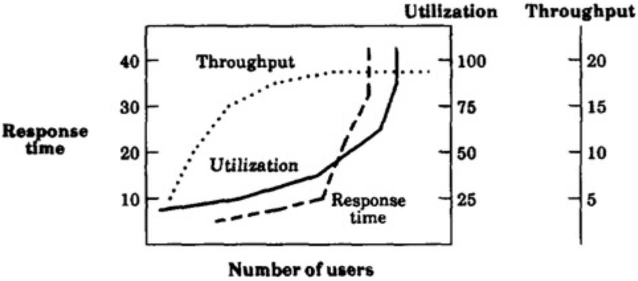
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



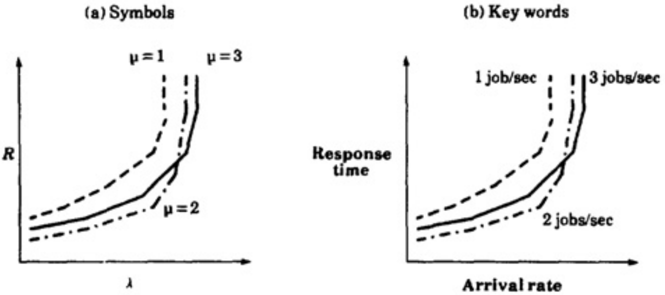
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)

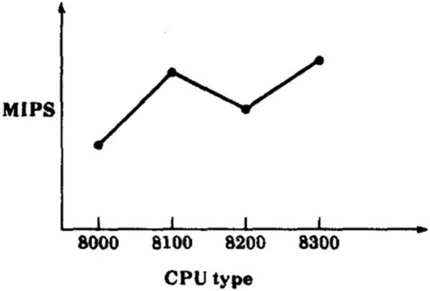


(a) Symbols

(b) Key words

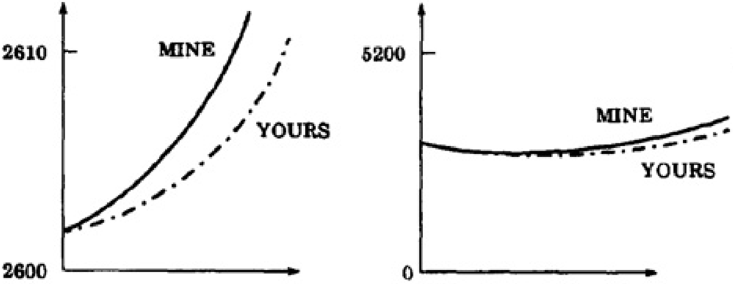
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



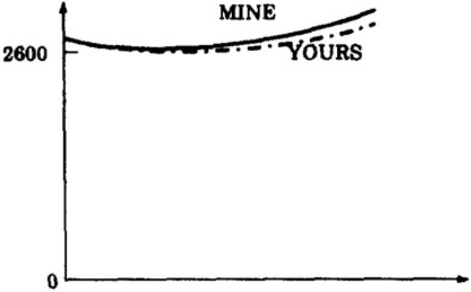
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



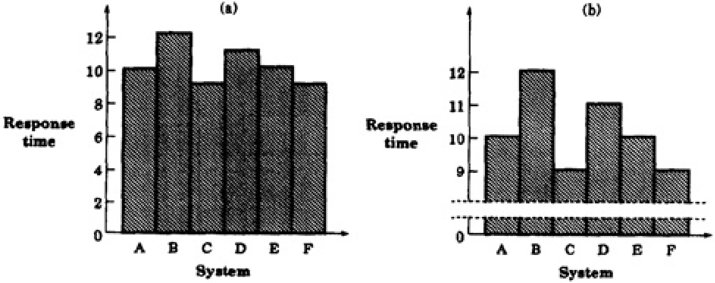
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



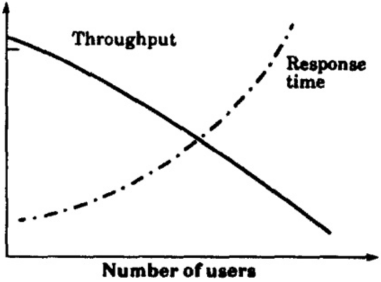
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



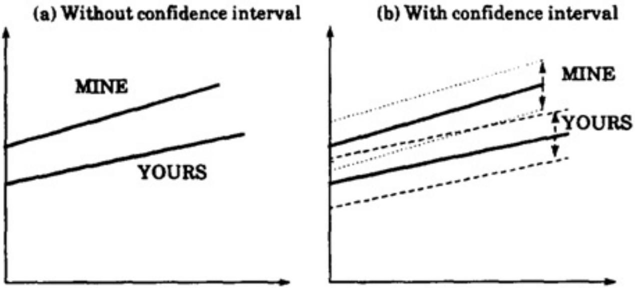
Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)

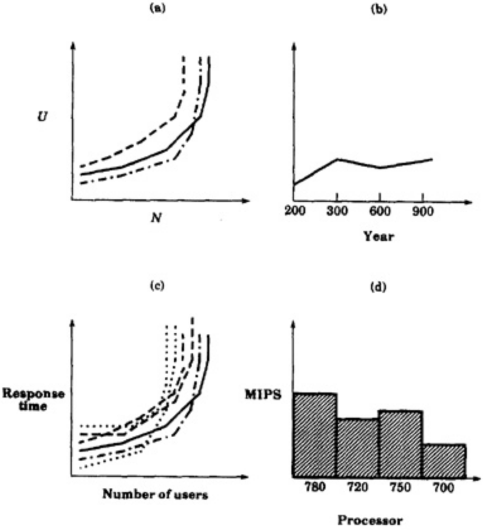


Guidelines

1. Require minimum effort to the reader: get the message (legends, labels, trends, annotations, . . .)
2. Maximize information (self-sufficient, clear labels, units, . . .)
3. Minimize Ink (avoid cluttered information. . .)
4. Use commonly accepted practices (effect along the y-axis, scales)
5. Avoid Ambiguity (coordinates, scales, colors, only one variable, . . .)



What about these ones ?



Use the right tools

R is a system for statistical computation and graphics.

- Avoid programming with R. Most things can be done with one liners.
- Excellent graphic support with **ggplot2**.
- `knitr` allows to mix R with LaTeX or Markdown. Literate programming to ease reproducible research.

Rstudio is an IDE a system for statistical computation and graphics. It is easy to use and allows publishing on **rpubs**.

Org-mode Allows to mix sh, perl, R, ... within plain text documents and export to LaTeX, HTML, ...