CS 147: Computer Systems Performance Analysis

Mistakes in Graphical Presentation

2015-06-15

CS147

CS 147:
Computer Systems Performance Analysis
Mistakes in Graphical Presentation

Overview

- Common Mistakes in Graphics
  - Excess Information
  - Multiple Scales
  - Symbols for Text
  - Poor Scales
  - Bad Line Usage
- Pictorial Games
  - Non-Zero Origins
  - Double Whammy
  - No Confidence Intervals
  - Height Scaling
  - Histogram Problems
- Graphical Integrity
- Special-Purpose Charts
- A Few Examples

Overview
Common Mistakes in Graphics
Excess Information
Multiple Scales
Symbols for Text
Poor Scales
Bad Line Usage
Pictorial Games
Non-Zero Origins
Double Whammy
No Confidence Intervals
Height Scaling
Histogram Problems
Graphical Integrity
Special-Purpose Charts
A Few Examples

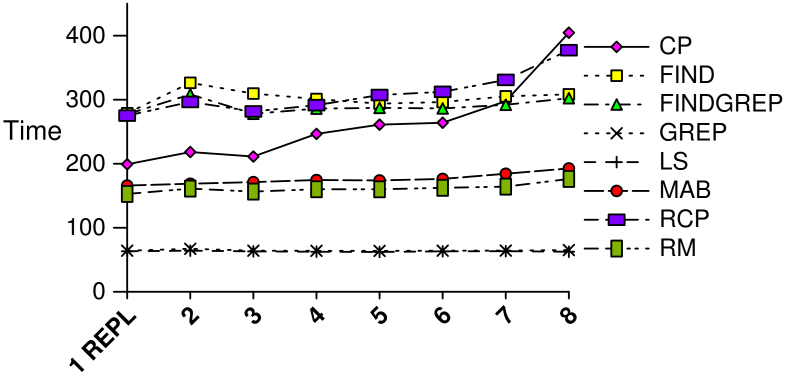
Excess Information

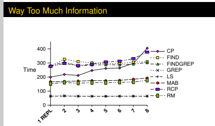
- Sneaky trick to meet length limits
- Rules of thumb:
  - 6 curves on line chart
  - 10 bars on bar chart
  - 8 slices on pie chart
    - (But note that Tufte *hates* pie charts)
- Extract essence; don't cram things in

Excess Information

- Sneaky trick to meet length limits
- Rules of thumb:
  - 6 curves on line chart
  - 10 bars on bar chart
  - 8 slices on pie chart
    - (But note that Tufte *hates* pie charts)
- Extract essence; don't cram things in

Way Too Much Information

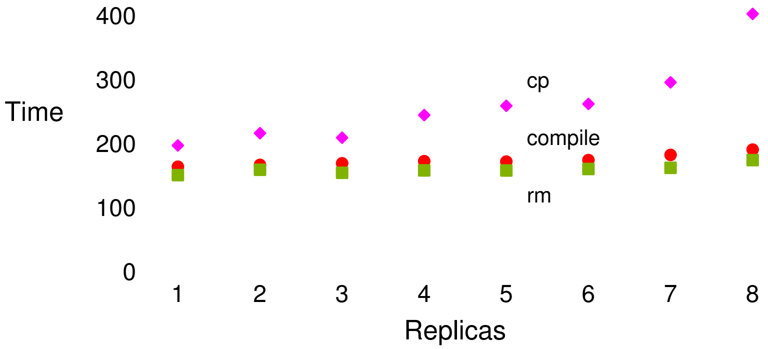




What's important on that chart?

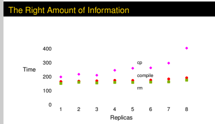
- Times for cp and rcp rise with number of replicas
- Most other benchmarks are near constant
- Exactly constant for rm

The Right Amount of Information



2015-06-15

CS147
└Common Mistakes in Graphics
　└Excess Information
　　└The Right Amount of Information



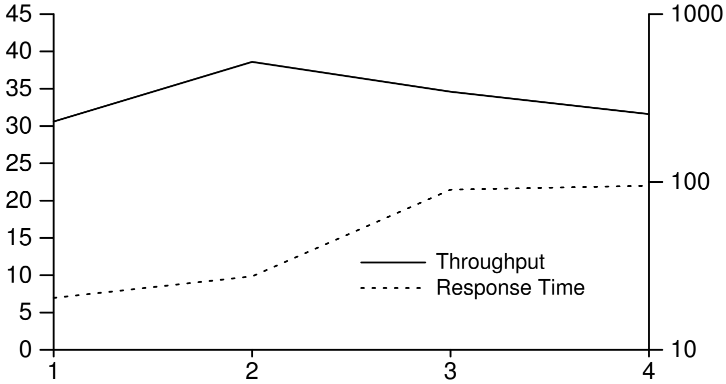
Multiple Scales

- Another way to meet length limits
- Basically, two graphs overlaid on each other
- Confuses reader (which line goes with which scale?)
- Misstates relationships
  - Implies equality of magnitude that doesn't exist

Multiple Scales

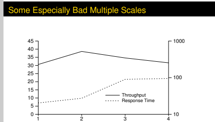
- Another way to meet length limits
- Basically, two graphs overlaid on each other
- Confuses reader (which line goes with which scale?)
- Misstates relationships
  - Implies equality of magnitude that doesn't exist

Some Especially Bad Multiple Scales



2015-06-15

CS147
└Common Mistakes in Graphics
  └Multiple Scales
    └Some Especially Bad Multiple Scales



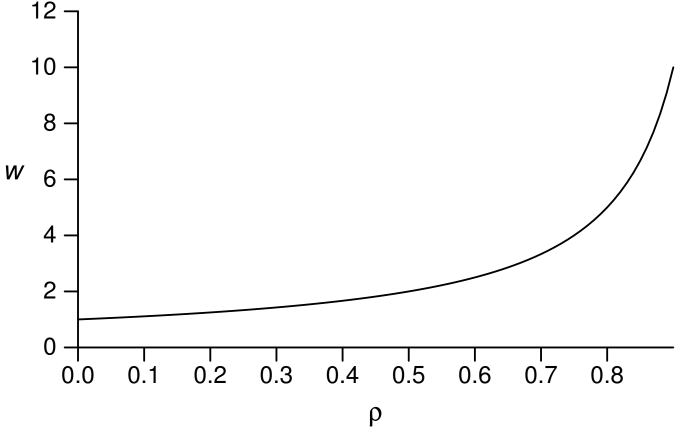
Using Symbols in Place of Text

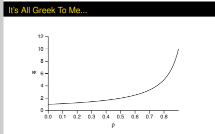
- Graphics should be self-explanatory
  - Remember that the graphs often draw the reader in
- So use explanatory text, not symbols
- This means no Greek letters!
  - Unless your conference is in Athens...

Using Symbols in Place of Text

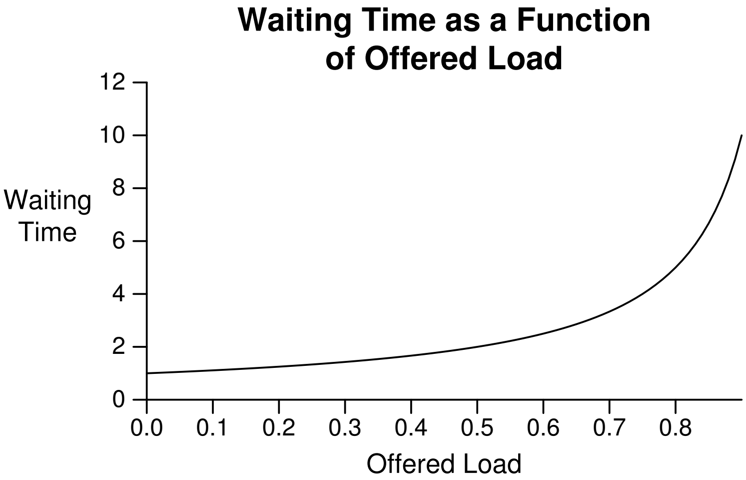
- Graphics should be self-explanatory
  - Remember that the graphs often draw the reader in
- So use explanatory text, not symbols
- This means no Greek letters!
  - Unless your conference is in Athens...

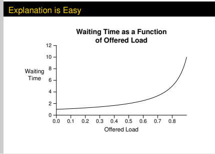
It’s All Greek To Me...





Explanation is Easy





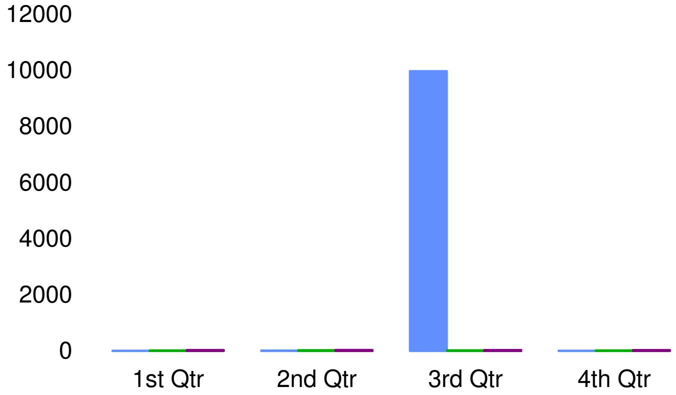
Poor Scales

- Fiddle with axis ranges (and logarithms) to get your message across
  - But don't lie or cheat
- Sometimes trimming off high ends makes things clearer
  - Brings out low-end detail

Poor Scales

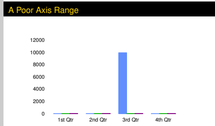
- Fiddle with axis ranges (and logarithms) to get your message across
  - But don't lie or cheat
- Sometimes trimming off high ends makes things clearer
  - Brings out low-end detail

A Poor Axis Range

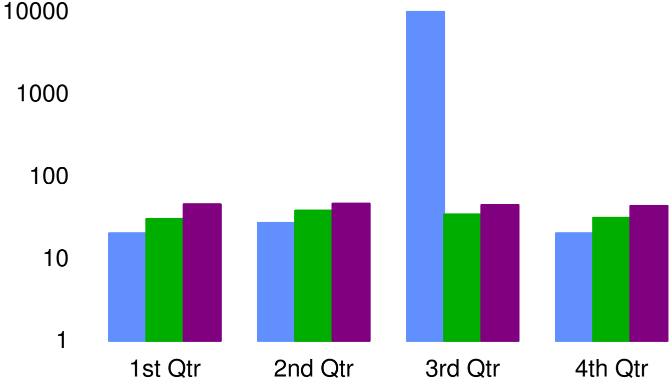


2015-06-15

CS147
└Common Mistakes in Graphics
└Poor Scales
└A Poor Axis Range

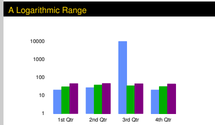


A Logarithmic Range

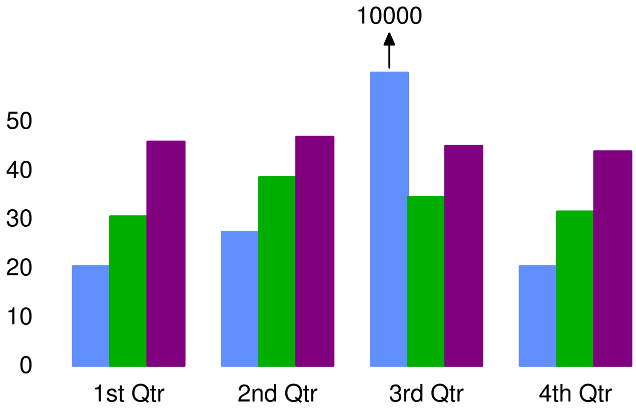


2015-06-15

CS147
└Common Mistakes in Graphics
 └Poor Scales
  └A Logarithmic Range

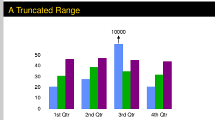


A Truncated Range



2015-06-15

CS147
└Common Mistakes in Graphics
  └Poor Scales
    └A Truncated Range



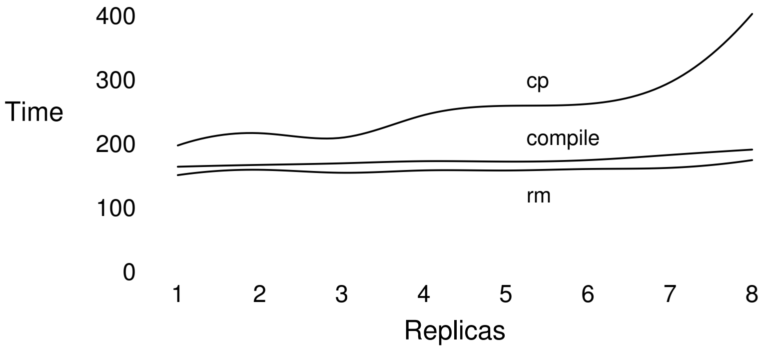
Using Lines Incorrectly

- Don’t connect points unless interpolation is meaningful
- Don’t smooth lines that are based on samples
  - Exception: fitted non-linear curves

Using Lines Incorrectly

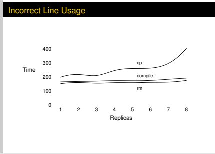
- Don’t connect points unless interpolation is meaningful
- Don’t smooth lines that are based on samples
  - Exception: fitted non-linear curves

Incorrect Line Usage



2015-06-15

CS147
└Common Mistakes in Graphics
 └Bad Line Usage
  └Incorrect Line Usage

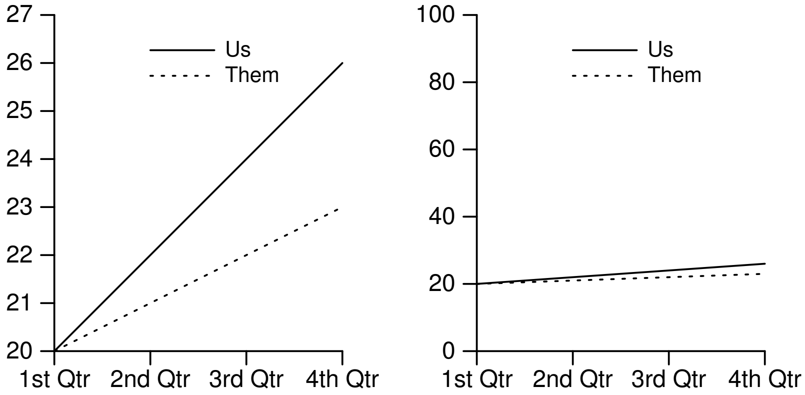


Non-Zero Origins and Broken Scales

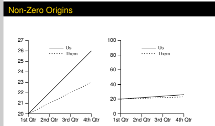
- People expect (0,0) origins
  - Subconsciously
- So non-zero origins are great way to lie
- More common than not in popular press
- Also very common to cheat by omitting part of scale
  - “Really, Your Honor, I included (0,0)”

Non-Zero Origins and Broken Scales
- People expect (0,0) origins
  - Subconsciously
- So non-zero origins are great way to lie
- More common than not in popular press
- Also very common to cheat by omitting part of scale
  - “Really, Your Honor, I included (0,0)”

Non-Zero Origins

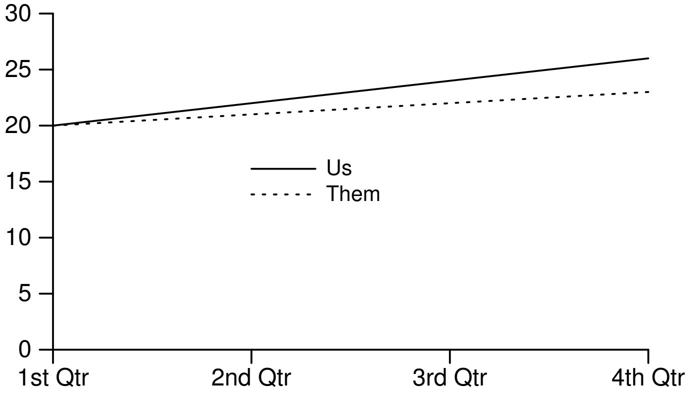


2015-06-15



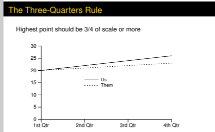
The Three-Quarters Rule

Highest point should be 3/4 of scale or more



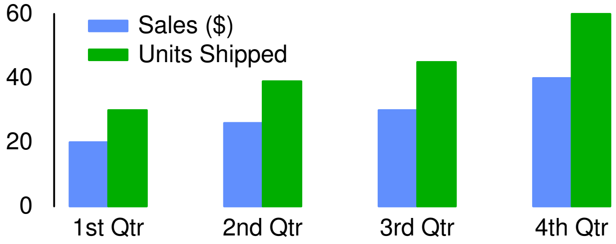
2015-06-15

CS147
└Pictorial Games
└Non-Zero Origins
└The Three-Quarters Rule



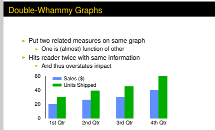
Double-Whammy Graphs

- Put two related measures on same graph
  - One is (almost) function of other
- Hits reader twice with same information
  - And thus overstates impact



2015-06-15

CS147
└Pictorial Games
  └Double Whammy
    └Double-Whammy Graphs

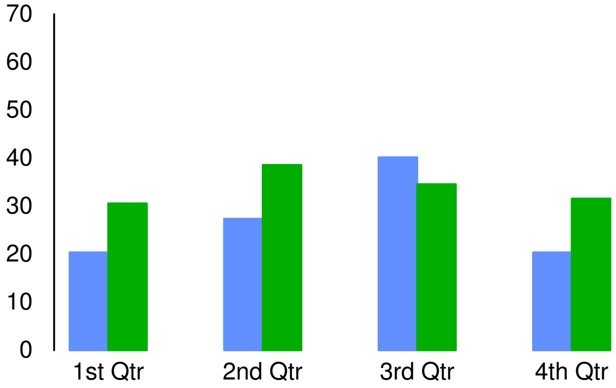


Omitting Confidence Intervals

- Statistical data is inherently fuzzy
- But means appear precise
- Giving confidence intervals can make it clear there's no real difference
  - So liars and fools leave them out

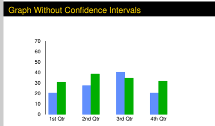
Omitting Confidence Intervals
- Statistical data is inherently fuzzy
- But means appear precise
- Giving confidence intervals can make it clear there's no real difference
  - So liars and fools leave them out

Graph Without Confidence Intervals

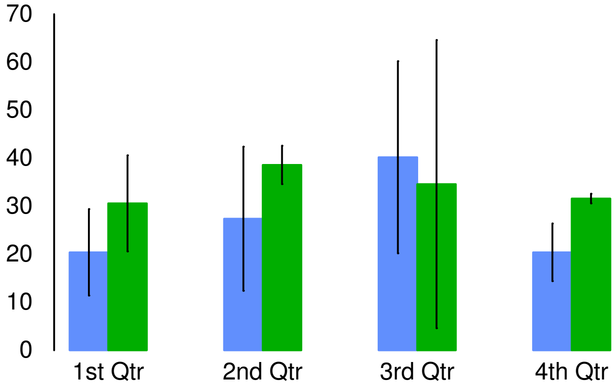


2015-06-15

CS147
└Pictorial Games
 └No Confidence Intervals
  └Graph Without Confidence Intervals

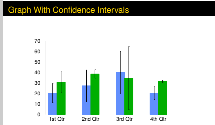


Graph With Confidence Intervals



2015-06-15

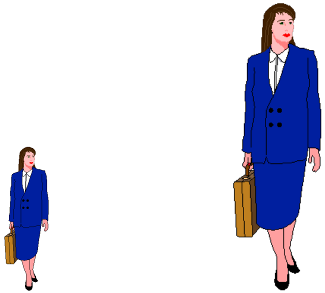
CS147
└Pictorial Games
 └No Confidence Intervals
  └Graph With Confidence Intervals


Graph With Confidence Intervals

Scaling by Height Instead of Area

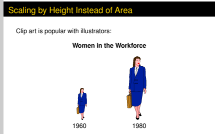
Clip art is popular with illustrators:

Women in the Workforce



1960 1980

2015-06-15



The Trouble with Height Scaling

- Previous graph had heights of 2:1
- But people perceive areas, not heights
  - So areas should be what's proportional to data
- Tufte defines *lie factor:* size of effect in graphic divided by size of effect in data
  - Not limited to area scaling
  - But especially insidious there (quadratic effect)

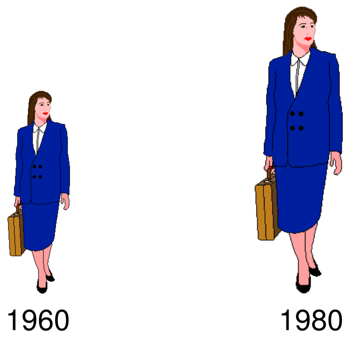
The Trouble with Height Scaling

- Previous graph had heights of 2:1
- But people perceive areas, not heights
  - So areas should be what's proportional to data
- Tufte defines *lie factor:* size of effect in graphic divided by size of effect in data
  - Not limited to area scaling
  - But especially insidious there (quadratic effect)

Scaling by Area

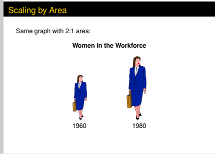
Same graph with 2:1 area:

Women in the Workforce



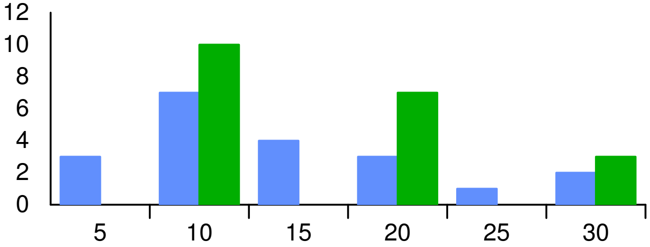
2015-06-15

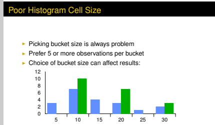
CS147
└Pictorial Games
  └Height Scaling
    └Scaling by Area



Poor Histogram Cell Size

- Picking bucket size is always problem
- Prefer 5 or more observations per bucket
- Choice of bucket size can affect results:





Note that green bars are steadily decreasing, but blue bars rise, fall, and rise again.

It’s not clear which is correct (given small counts in the smaller buckets).

Principles of Graphics Integrity (Tufte)

- Proportional representation of numbers
- Clear, detailed, thorough labeling
- Show data variation, not design variation
- Use deflated money units
- Don't have more dimensions than data has
- Don't quote data out of context

Principles of Graphics Integrity (Tufte)

- Proportional representation of numbers
- Clear, detailed, thorough labeling
- Show data variation, not design variation
- Use deflated money units
- Don't have more dimensions than data has
- Don't quote data out of context

Proportional Representation of Numbers

- Maintain lie factor of 1.0
- Use areas, not heights, with clip art
- Avoiding “decorative” graphs will do wonders
  - Not too hard for most engineers!

Proportional Representation of Numbers
- Maintain lie factor of 1.0
- Use areas, not heights, with clip art
- Avoiding “decorative” graphs will do wonders
  - Not too hard for most engineers!

Clear, Detailed, Thorough Labeling

- Goal is to defeat distortion and ambiguity
- Write explanations on graphic itself
- Label important events in the data

Clear, Detailed, Thorough Labeling
- Goal is to defeat distortion and ambiguity
- Write explanations on graphic itself
- Label important events in the data

Show Data Variation, Not Design Variation

- Use one design for entire graphic
- In papers, try to use one design for *all* graphs
- Again, artistic license is big culprit

Show Data Variation, Not Design Variation

- Use one design for entire graphic
- In papers, try to use one design for *all* graphs
- Again, artistic license is big culprit

Use Deflated Money Units

- Often necessary to show money over time
  - Even in computer science
  - E.g., price/performance over time
  - Or expected future cost of a disk
- Nominal dollars are meaningless
- Derate by some standard inflation measure
  - That's what the WWW is for!

Use Deflated Money Units

- Often necessary to show money over time
  - Even in computer science
  - E.g., price/performance over time
  - Or expected future cost of a disk
- Nominal dollars are meaningless
- Derate by some standard inflation measure
  - That's what the WWW is for!

Don’t Have More Dimensions Than Data Has

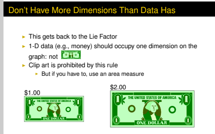
- This gets back to the Lie Factor
- 1-D data (e.g., money) should occupy one dimension on the graph: not 
- Clip art is prohibited by this rule
  - But if you have to, use an area measure

$1.00

$2.00

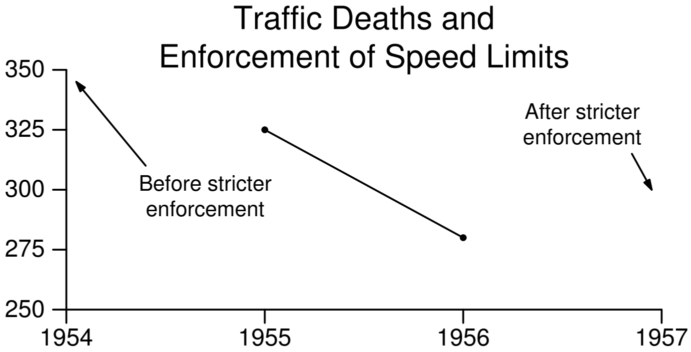
2015-06-15

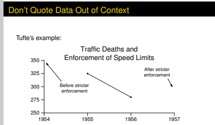
CS147
└Graphical Integrity
└Don’t Have More Dimensions Than Data Has



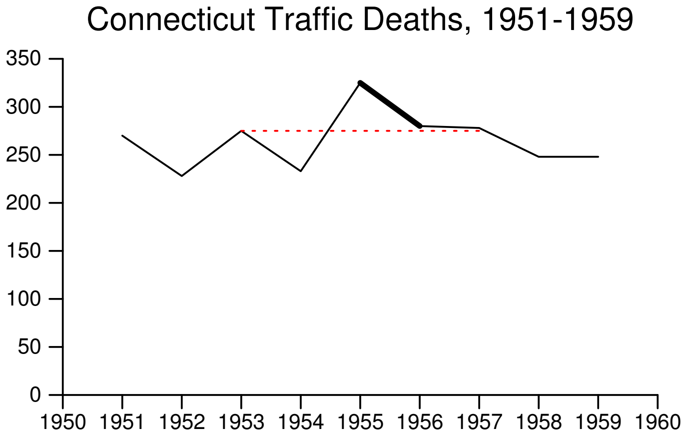
Don’t Quote Data Out of Context

Tufte’s example:

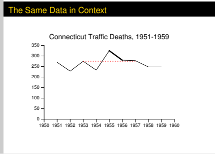




The Same Data in Context



2015-06-15



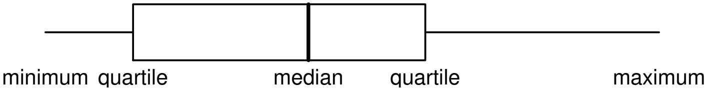
Special-Purpose Charts

- Tukey's box plot
- Histograms
- Scatter plots
- Gantt charts
- Kiviat graphs

Special-Purpose Charts
- Tukey's box plot
- Histograms
- Scatter plots
- Gantt charts
- Kiviat graphs

Tukey’s Box Plot

- Shows range, median, quartiles all in one:



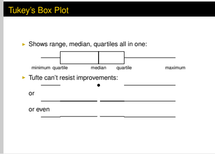
- Tufte can’t resist improvements:

or

or even

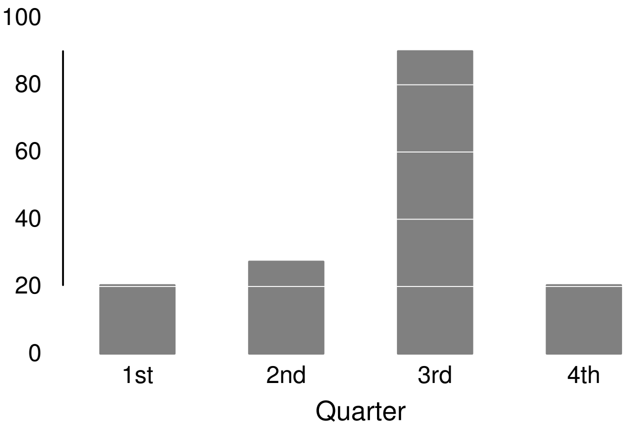
2015-06-15

CS147
└Special-Purpose Charts
└Tukey’s Box Plot



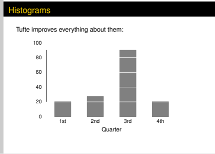
Histograms

Tufte improves everything about them:



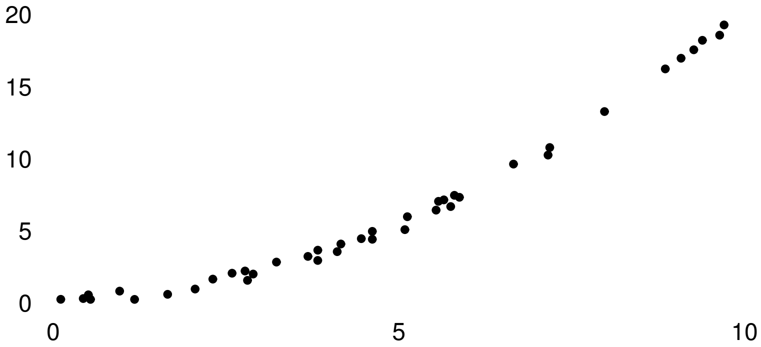
2015-06-15

CS147
└Special-Purpose Charts
└Histograms



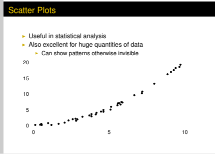
Scatter Plots

- Useful in statistical analysis
- Also excellent for huge quantities of data
  - Can show patterns otherwise invisible



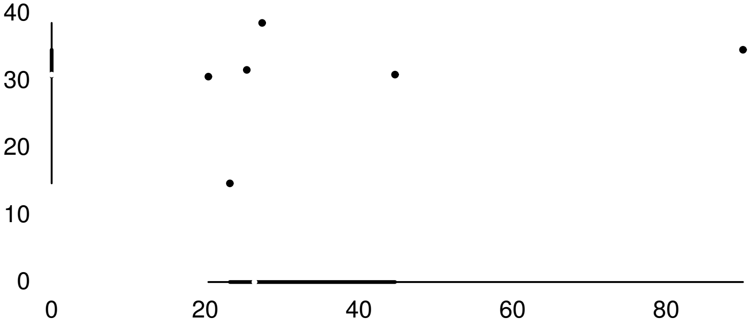
2015-06-15

CS147
└Special-Purpose Charts
└Scatter Plots



Better Scatter Plots

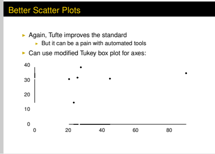
- Again, Tufte improves the standard
  - But it can be a pain with automated tools
- Can use modified Tukey box plot for axes:



2015-06-15

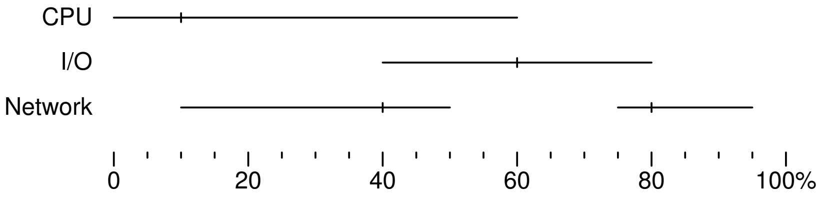
CS147
└Special-Purpose Charts

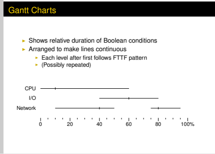
└Better Scatter Plots



Gantt Charts

- Shows relative duration of Boolean conditions
- Arranged to make lines continuous
  - Each level after first follows FTTF pattern
  - (Possibly repeated)

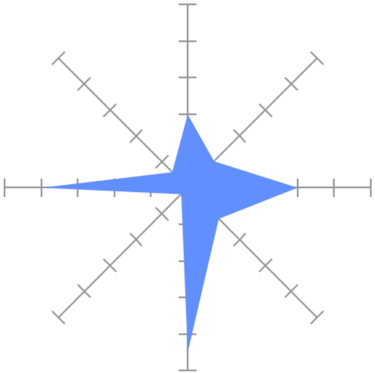




Gantt charts are any chart with horizontal lines showing spans on the X axis.
Also useful for scheduling; shows simultaneous tasks.
Lines are divided in mid-true; any vertical line shows one unique combo of conditions.
Length of line with particular condition shows percentage of time system spends in that state.

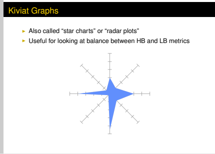
Kiviat Graphs

- Also called “star charts” or “radar plots”
- Useful for looking at balance between HB and LB metrics

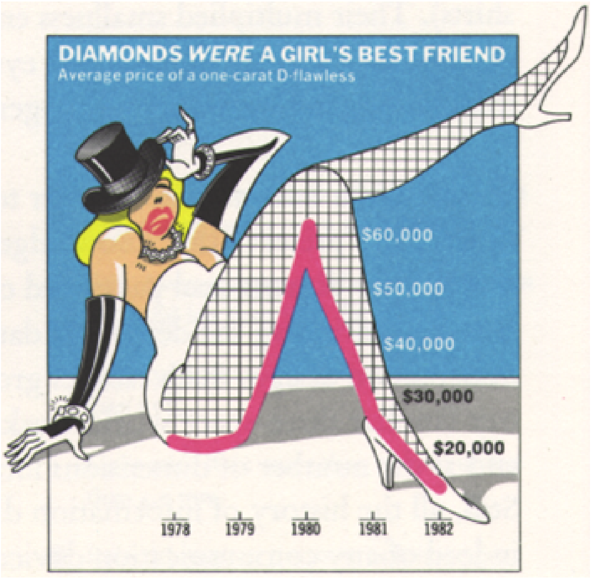


2015-06-15

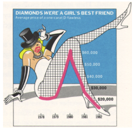
CS147
└Special-Purpose Charts
└Kiviat Graphs



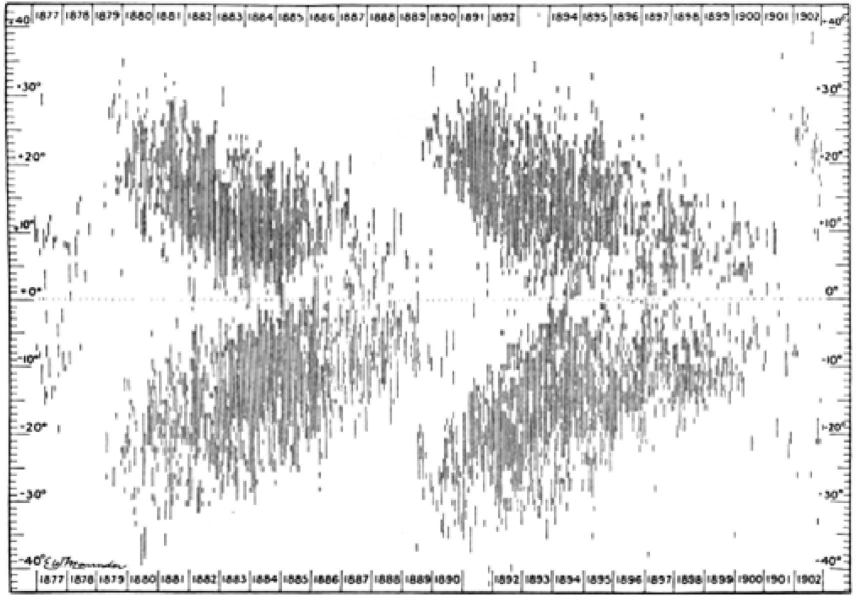
A Very Bad Graph

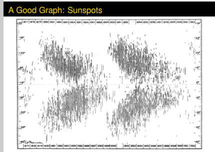


A Very Bad Graph



A Good Graph: Sunspots



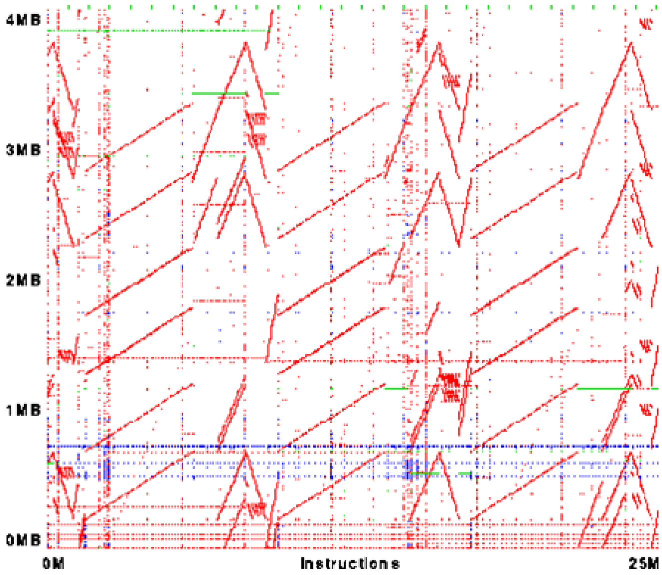


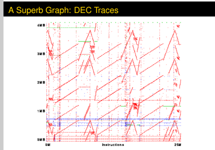
Vertical scale is latitude of sunspot; length of bar is extent of latitude width of sunspot (longitude width is not in the graph). The 11-year cycle is easily visible.

The horizontal scale is empty in a few places where sunspot data extends into it.

This graph was drawn in 1904 by Edward Walter Maunder (1851-1928). It is commonly called a “butterfly diagram” for obvious reasons.

A Superb Graph: DEC Traces





X axis is time (instructions executed). Y axis is memory address referenced, modulo 4 MB. Red lines are data accesses, blue instructions. Green is perhaps stack? Note how parallel access to arrays is easy to see, as well as occasional faster access and reverse-order access.