

Effective Data Visualization

CMU Data Science Club

Jerzy Wiecezorek

4/11/17

Follow along

<http://www.stat.cmu.edu/~jwieczor/>

has these slides, R code examples, and a summary checklist

Overview

There is a spectrum between **exploratory** graphics
(for your own understanding as you analyze the data)
and **explanatory** graphics
(for communicating results to an audience).
Today's advice applies to both, but we'll focus on
explanatory graphs.

Overview

- ▶ What should I graph?
(let function constrain form, show both raw data and summaries, show precision/uncertainty)
- ▶ Can people read my graphs?
(guides and captions, image format and quality, text and color)
- ▶ Can people understand my graphs?
(comparisons, grouping and search, cognition, consistency)
- ▶ Where can I learn more about dataviz?
(books, blogs to follow)
- ▶ R examples

Principles

In general:

- ▶ Design each graph with a **clear task** in mind
- ▶ Use **common sense** to help audience read the graph
- ▶ Use knowledge of **visual perception** to help audience perform the task efficiently

For scientific communication:

- ▶ Show the data foremost, and overlay summaries as appropriate
- ▶ Show statistical precision of any summaries
- ▶ Use tables to show exact values; use graphs to show patterns

Design for a clear task

General principles:

- ▶ **Design each graph with a clear task in mind**
- ▶ Use common sense to help audience read the graph
- ▶ Use knowledge of visual perception to help audience perform the task efficiently

Design for a clear task

A visualization is a tool for showing order and patterns in data. As the creator, you can “generate order before people’s brains try to do it on their own.”

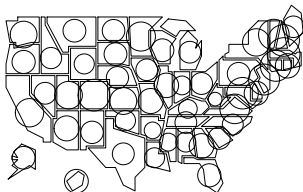
Cairo, *The Functional Art*: Cairo read a claim that *higher education is related to lower obesity*, but no evidence was shown. He found a US state-level dataset for rates of obesity and of college education. What graphical form can help show evidence for or against this claim?

Choice of graphical form

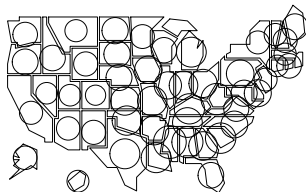
Bubbles on a map?

(larger = higher percentage)

**Percentage of people
with a BA degree or higher**



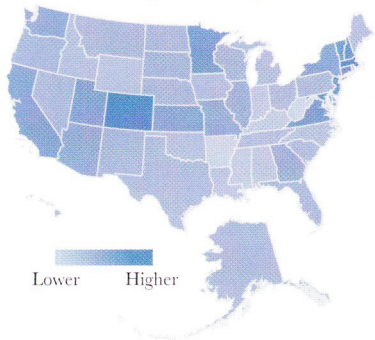
Percentage of obese people



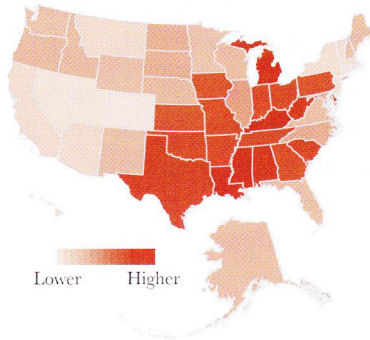
Choice of graphical form

Choropleth (colored thematic map)?
(darker = higher percentage)

Percentage of people with a
BA degree or higher



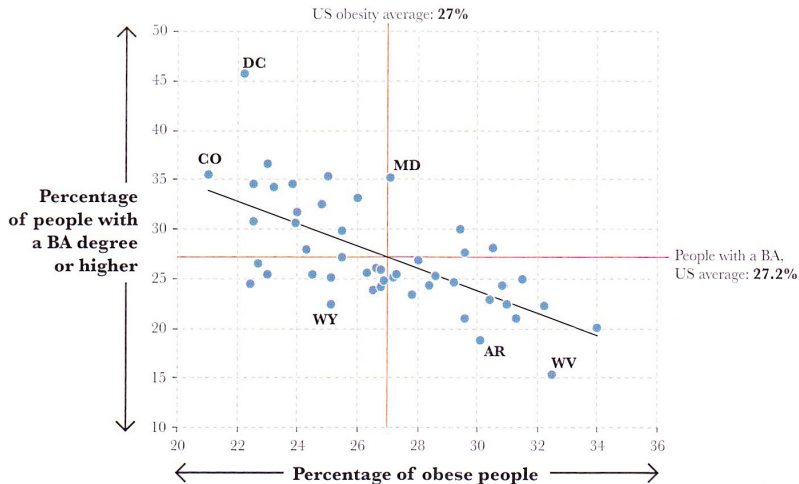
Percentage of
obese people



Choice of graphical form

Scatterplot shows the relationship directly.

Doesn't show spatial patterns, but that wasn't our goal, so OK.



Choice of graphical form

Lessons

- ▶ Decide on visual task, and choose a form that supports it.
- ▶ Just because you have certain variables (geography, time, social ties) doesn't mean it's helpful to show them (map, time series, network diagram).

Most useful graphic forms and visual variables

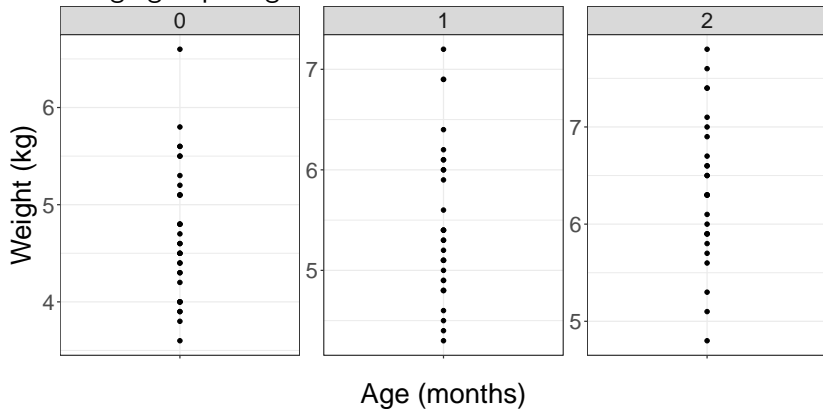
Data forms: point, line, bar
(and small multiples)

Statistical summaries: histogram, boxplot, smoothing line, error bar

Visual variables: color, point shape, point size, line width, line type

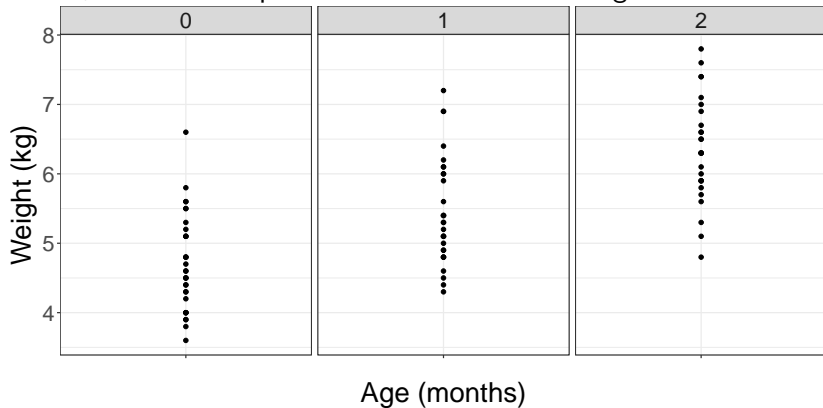
Consistency

Which age group weighs the least?



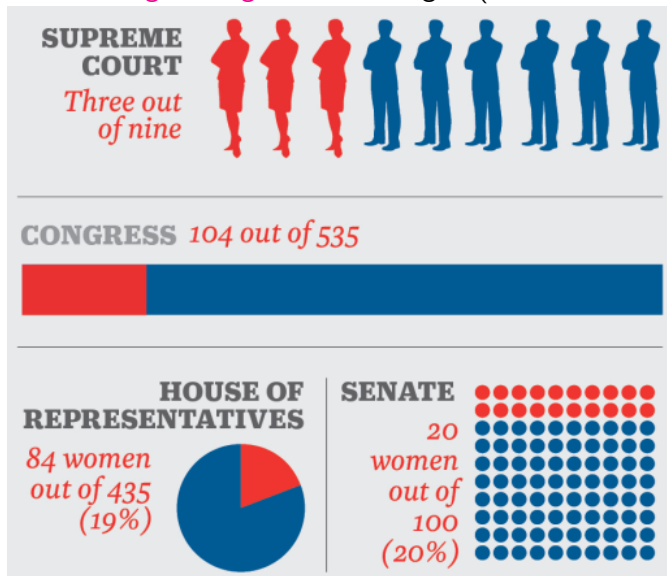
Consistency

Give all small multiples the same structure, usually **including axis limits**, to make comparisons easier and reduce cognitive load



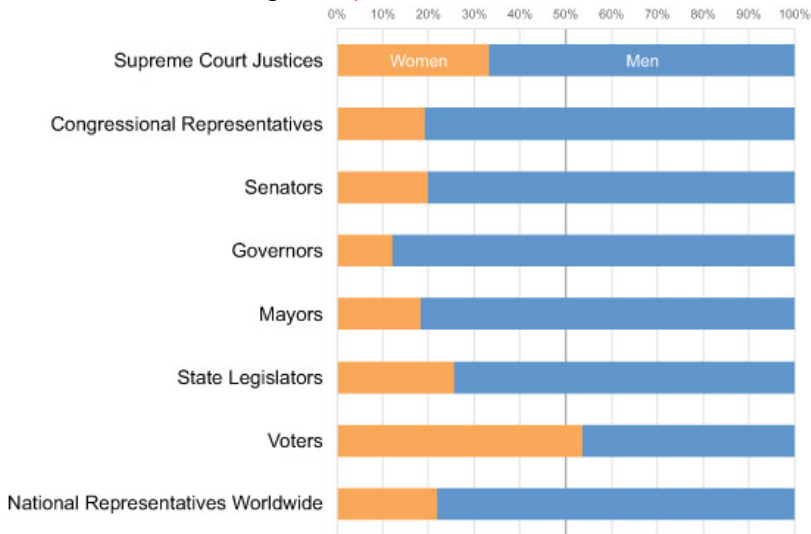
Consistency

Ensure **design changes** are meaningful (tied to data changes)



Consistency

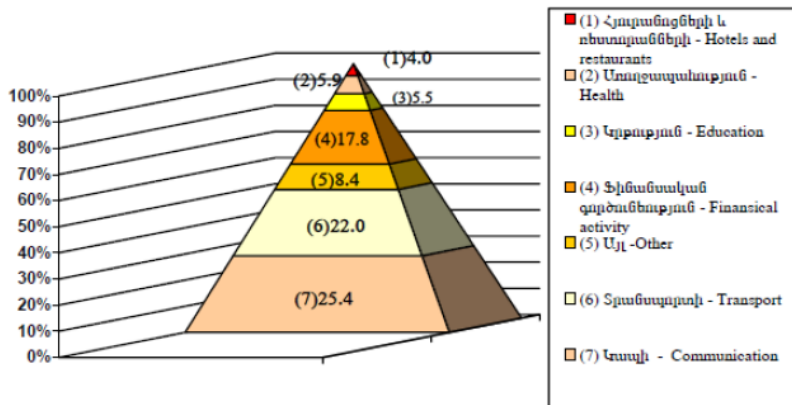
More consistent redesign, **Stephen Few**



Consistency

Avoid meaningless visual variables like shadow or 3D

STRUCTURE OF SERVICES 2007



Consistency

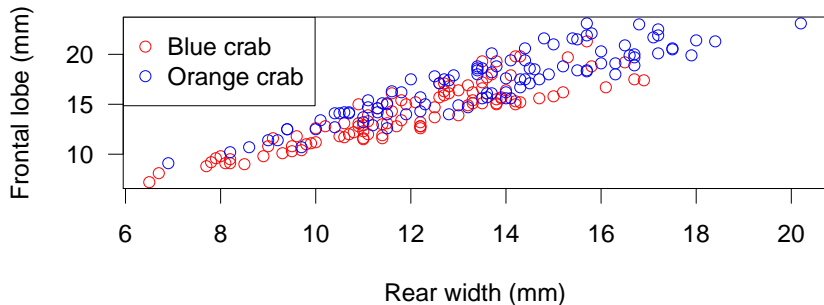
Lessons

- ▶ Use consistent mappings (colors and shapes, axis limits) across graphs.
- ▶ Don't reuse the same mappings across different data variables.
- ▶ Avoid meaningless variety in design.
- ▶ Avoid shadow, 3D, and other variables not mapped to data.

Semantic associations

Orange vs blue crab species: I've actually seen this in a talk
(**crabs dataset**)

Crab dimensions by species



Semantic associations

Lessons

- ▶ Use meaningful mappings:
orange vs blue crab species = orange and blue symbols.
- ▶ Use conventional mappings: blue = cold, red = hot.
- ▶ “More = more”:
deeper saturation or larger size = higher value of variable.

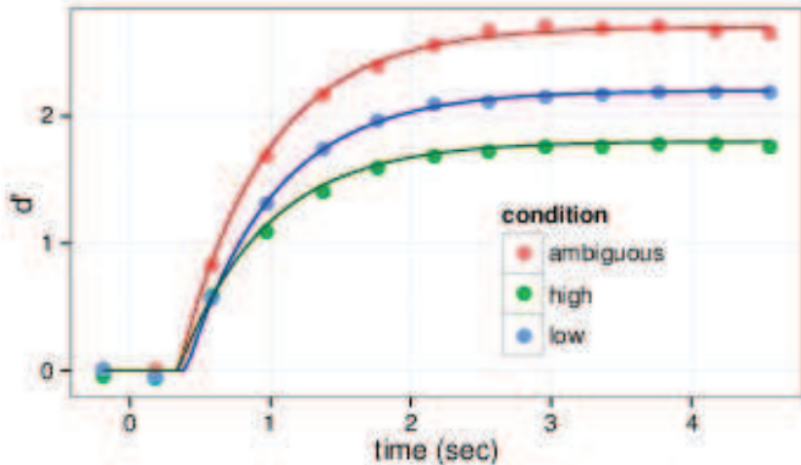
Legibility

General principles:

- ▶ Design each graph with a clear task in mind
- ▶ **Use common sense to help audience read the graph**
- ▶ Use knowledge of visual perception to help audience perform the task efficiently

Legibility

What is done well here? What could be improved?



Guides

Title or caption: explain anything not already on axes or legend.

What are the individual plotted points (raw data by subject, vs. averages by group)? Is this the full dataset or a subset?

Axis labels: give units (“number of events”, “\$”, “%”, “m/s”). Give readable labels, not default variable names (Pretest Score, not PRE_TEST_SCORE).

Tick marks: use round, readable numbers and avoid scientific notation. If “Dollars” axis has ticks like “1e6”, plot “Dollars in millions” instead.

Grid lines: omit or make very light. Do not overwhelm the data.

Legend: use same ordering as on plot. Or, if possible, omit and use direct labels instead.

Color

Ensure your colors can be distinguished from each other:

- ▶ Around 10% of men and 1% of women are colorblind.
- ▶ Some color palettes do not photocopy well.

Use **Color Brewer** to choose a well-tested palette.

Image quality

Understand when to use bitmap vs. vector formats.

Ensure graph's text size is similar to surrounding body text.

Don't stretch graphs! Create & save them at the right size in the first place.

Vector vs bitmap

Vector vs bitmap explained

Bitmap: common formats

- ▶ **jpg/jpeg** is lossy, designed for photos but not text/charts
- ▶ **png** is lossless, good for text/charts, common on web

Vector: common formats

- ▶ **svg** can display in browser, common on web
- ▶ **pdf** is for standalone doc (or to put inside another pdf)

Recommended formats and resolutions

Software	Recommended graphics device
Illustrator	svg
pdflatex	pdf, png (600 ppi)
Office	png (600 ppi)
web	png (72 ppi)

(ppi = Pixels Per Inch)

From Wickham, [ggplot2](#), Table 8.3

Save images at intended final size

Decide on target size and create the graph that way from the start. Gives better quality than changing size after saving, and avoids stretching/distorting the graph.

For example:

- ▶ Default textwidth in a LaTeX article is around 5.4 inches
- ▶ Blog templates may have a default column width of 500 pixels

Visual Perception

In general:

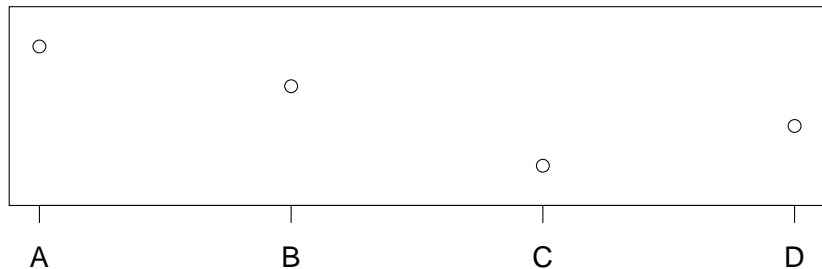
- ▶ Design each graph with a clear task in mind
- ▶ Use common sense to help audience read the graph
- ▶ **Use knowledge of visual perception to help audience perform the task efficiently**

Quantitative comparisons

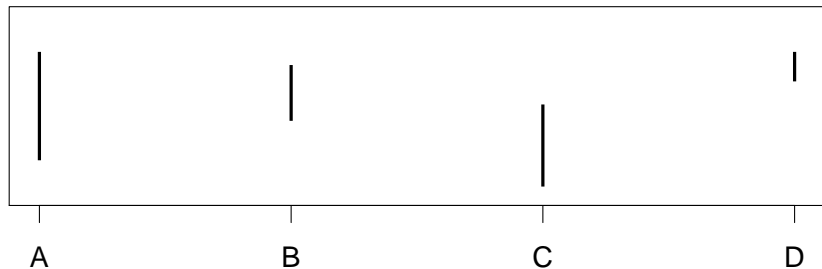
We'll try an experiment on next few slides:

	A	B	C	D
Positions	1	?	?	?
Lengths	1	?	?	?
Angles	1	?	?	?
Areas	1	?	?	?

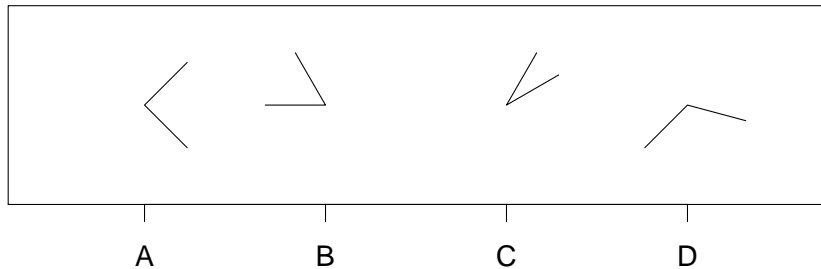
Quantitative perceptual tasks: position, aligned



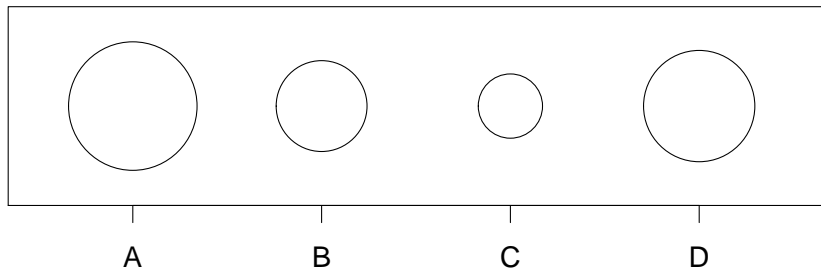
Quantitative perceptual tasks: length



Quantitative perceptual tasks: angle



Quantitative perceptual tasks: area



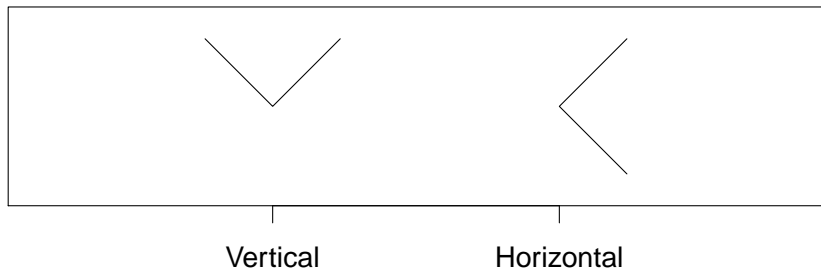
Quantitative perceptual tasks: answers

	A	B	C	D
Positions	1	3/4	1/4	2/4
Lengths	1	2/4	3/4	1/4
Angles	1	2/3	1/3	4/3
Areas	1	2/4	1/4	3/4

Cleveland and McGill (1984)

Cleveland, *The Elements of Graphing Data*

Quantitative perceptual tasks: effect of angle orientation



Same angle looks wider when bisector is horizontal.

Ordering of perceptual tasks

Cleveland and McGill's ordering (split over 2 slides)

Allows more
accurate comparisons



2D position along common, aligned scale



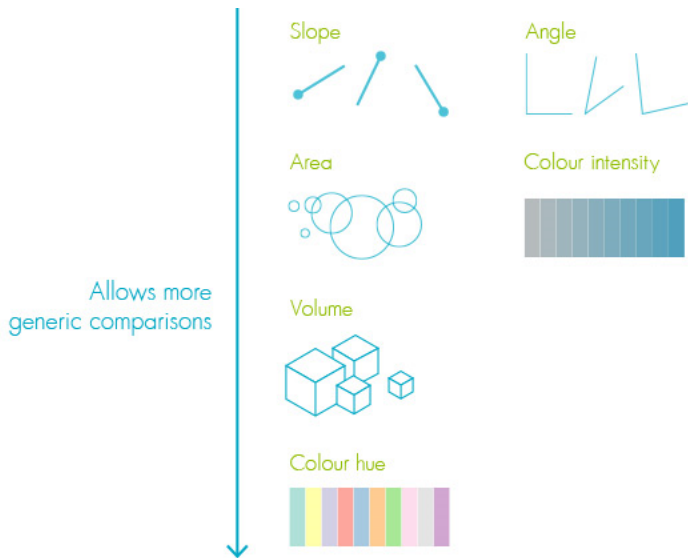
2D position along common, but unaligned scales



Length



Ordering of perceptual tasks



Distance

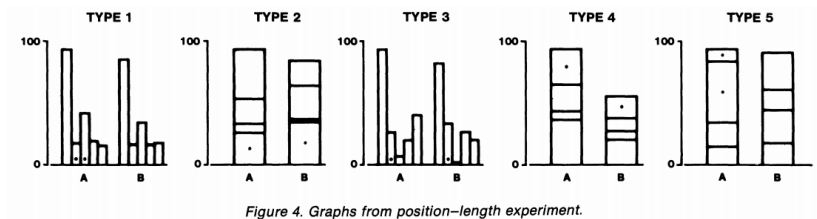


Figure 4. Graphs from position-length experiment.

Cleveland and McGill (1984)

Quantitative perceptual tasks

Lessons:

- ▶ Best to show quantitative variables with position or length. Use points, lines, or bars.
- ▶ Bars encode length, so start bars at 0. If you must zoom in, use dotplots (encoding position) instead.
- ▶ Avoid stacked bars (not aligned).
Use dots or lines (aligned baselines) instead.
- ▶ Avoid pies, area, and volume entirely.
- ▶ Choose and order hues sensibly. Use [Color Brewer](#).
- ▶ Place things-to-be-compared near each other.

Preattentive processing example: find and count the 6s

9	5	9	2	9	9	0	7	9	6	2	2	6	0	6	9	6	7	2	0
9	7	3	0	1	8	0	5	5	6	6	0	2	2	7	2	6	2	3	9
6	5	4	2	5	9	4	8	5	4	7	4	3	2	2	6	7	7	5	5
3	8	7	1	9	9	8	5	9	2	3	4	0	5	9	3	8	8	0	7
0	1	8	4	1	6	3	7	4	0	8	4	0	2	6	5	5	6	8	2

Preattentive processing example: find and count the 6s

9	5	9	2	9	9	0	7	9	6	2	2	6	0	6	9	6	7	2	0
9	7	3	0	1	8	0	5	5	6	6	0	2	2	7	2	6	2	3	9
6	5	4	2	5	9	4	8	5	4	7	4	3	2	2	6	7	7	5	5
3	8	7	1	9	9	8	5	9	2	3	4	0	5	9	3	8	8	0	7
0	1	8	4	1	6	3	7	4	0	8	4	0	2	6	5	5	6	8	2

Preattentive processing

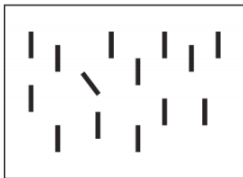
We automatically process and notice certain features, while others require conscious thought to find

We process faster when there are few categories to distinguish

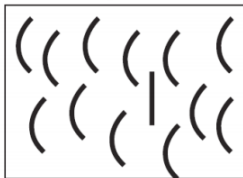
Preattentive processing: features

Colin Ware, *Information Visualization*

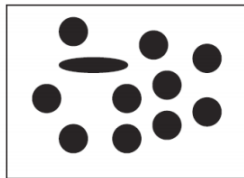
Orientation



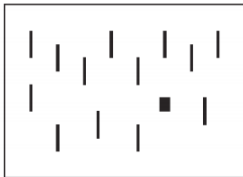
Curved/straight



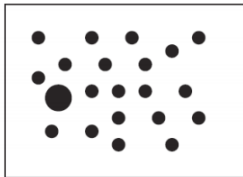
Shape



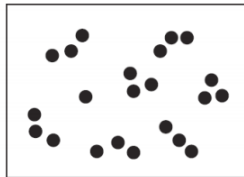
Shape



Size

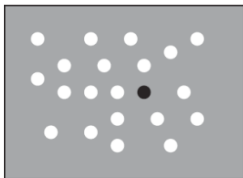


Number

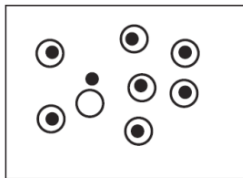


Preattentive processing: features

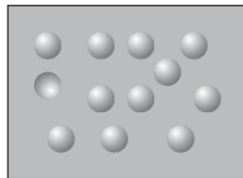
Gray/value



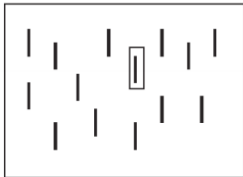
Enclosure



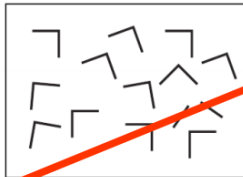
Convexity/concavity



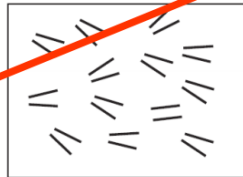
Addition



Juncture



Parallelism



Preattentive processing

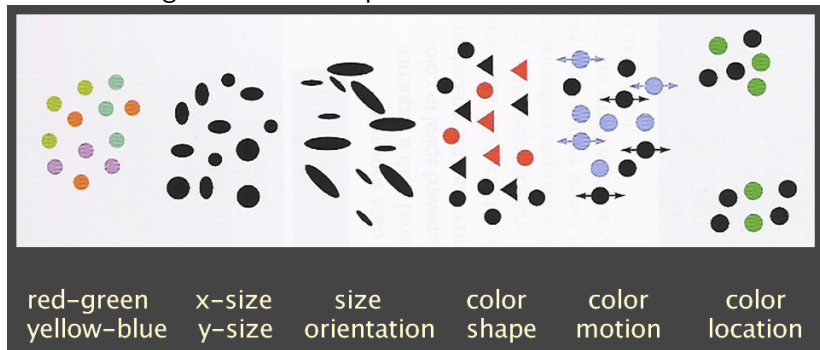
Lessons

- ▶ Distinguish categorical groups by features like hue & shape.
- ▶ Hue also lets you use direct labels instead of a legend.
- ▶ Don't try to show too many groups on one plot. Use small multiples to show more sub-groups.
- ▶ If highlighting one group, use a preattentive attribute.

Separable dimensions

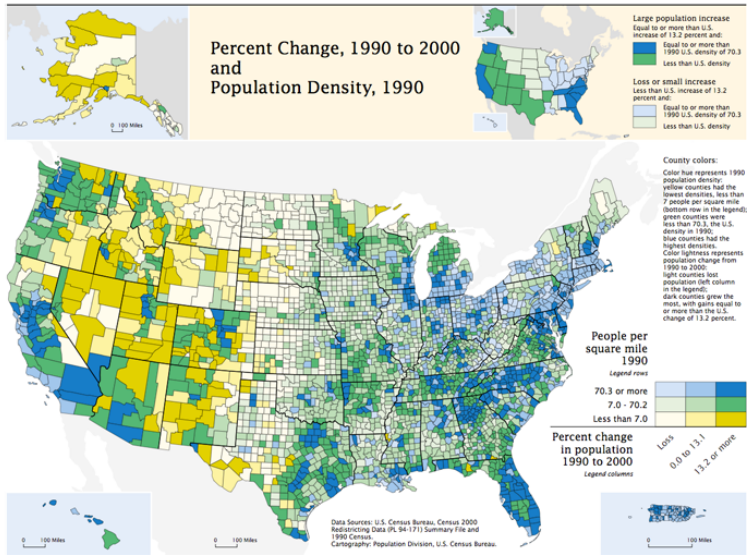
Some examples from Colin Ware, *Information Visualization*

<- More integral ... More separable ->



Integral dimensions example

US Census Bureau map using hue and saturation



Separable dimensions

Lessons

- ▶ Use color and another variable (shape, size, orientation, motion).
- ▶ Use small multiples rather than different plotting symbols.
- ▶ Avoid mixing 2 aspects of color, or 2 aspects of size.
- ▶ Don't show too many grouping variables at once.

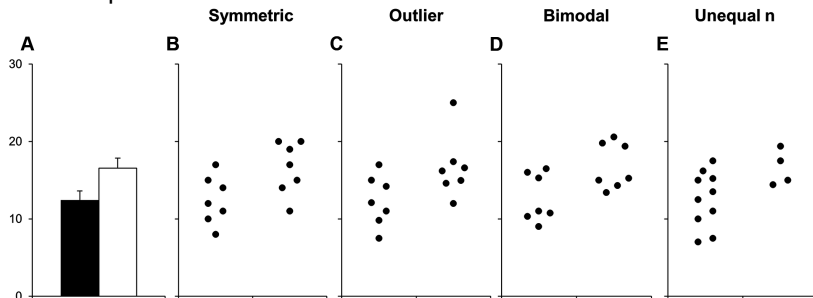
Show data, then summaries

Scientific communication principles:

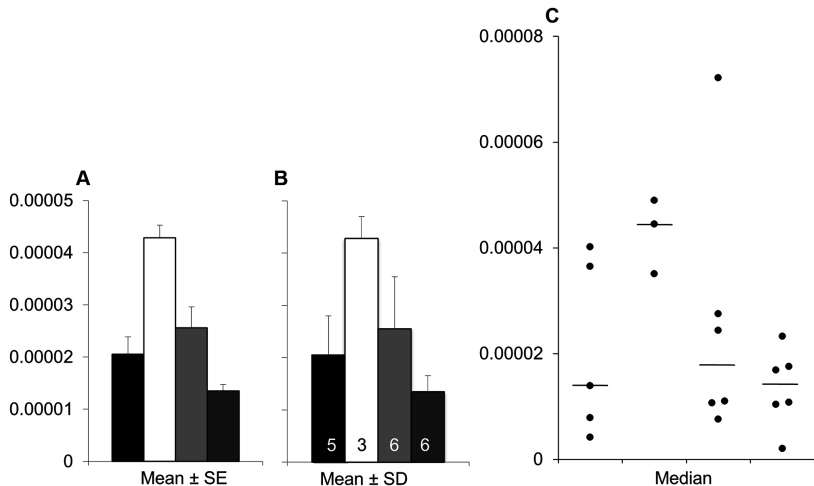
- ▶ **Show the data foremost, and overlay summaries as appropriate**
- ▶ Show statistical precision of any summaries
- ▶ Use tables to show exact values; use graphs to show patterns

Show data foremost

Weissgerber et al. (2015): summaries hide important data structure and sample size



Show data first, then add summaries



Show statistical precision

Scientific communication principles:

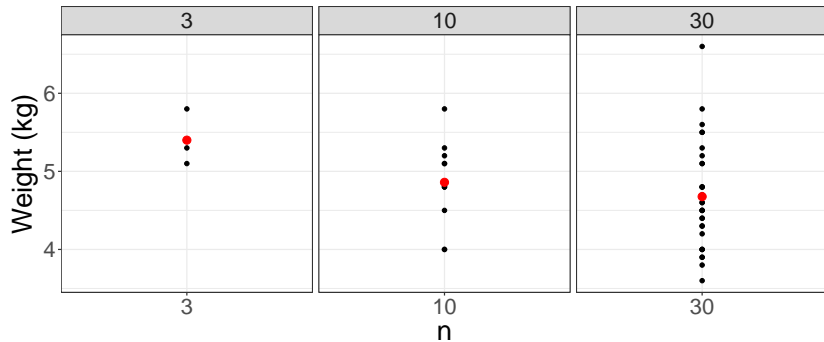
- ▶ Show the data foremost, and overlay summaries as appropriate
- ▶ **Show statistical precision of any summaries**
- ▶ Use tables to show exact values; use graphs to show patterns

Show statistical precision

How well do $n=3$ observations estimate the population mean?

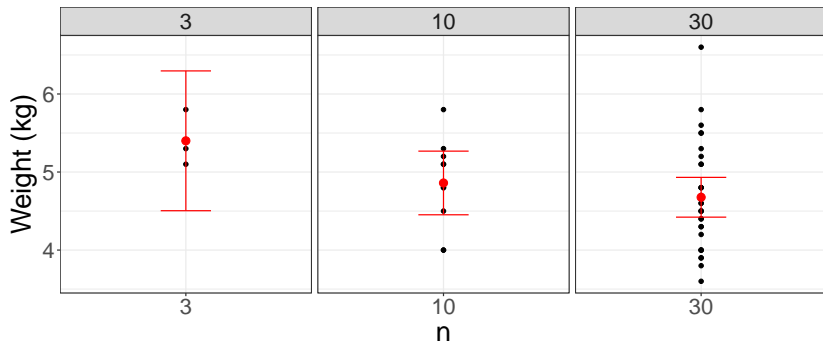
$n=10$? $n=30$?

Sampled newborn weights,
with estimated means in red



Show statistical precision

Sampled newborn weights,
with 95% CIs for the means in red



Show statistical precision

Be clear about whether error bars show SD, SE, or CI.

- ▶ **SD** (standard deviation) summarizes data spread. No need to show it if you already show the data.
- ▶ **SE** (standard error) summarizes precision of the sample mean estimate. Useful for calculations, but not directly interpretable.
- ▶ **CI** (confidence interval) summarizes plausible values of the unknown population mean, based on our observed sample.

CIs are most interpretable & best to plot.

Specify confidence level, e.g. “Error bars show 95% confidence intervals.”

Use tables for lookup

Scientific communication principles:

- ▶ Show the data foremost, and overlay summaries as appropriate
- ▶ Show statistical precision of any summaries
- ▶ **Use tables to show exact values;
use graphs to show patterns**

Use tables for lookup

Graphs are better for making comparisons and finding patterns.
Tables are better for showing precise values, if audience needs them:

- ▶ Experimental design: sample size, number of repeated measurements, or variable settings in each group
- ▶ Summary statistics for further calculation: group means and SDs allow reader to test differences between specific groups of interest
- ▶ Reference tables: Celsius-to-Fahrenheit conversions, normal distribution critical values, etc.

Principles

In general:

- ▶ Design each graph with a **clear task** in mind
- ▶ Use **common sense** to help audience read the graph
- ▶ Use knowledge of **visual perception** to help audience perform the task efficiently

For scientific communication:

- ▶ Show the data foremost, and overlay summaries as appropriate
- ▶ Show statistical precision of any summaries
- ▶ Use tables to show exact values; use graphs to show patterns

Further resources

- ▶ *My checklist* for the material covered today
- ▶ Cairo, *The Functional Art*: great overview from a data journalism perspective
- ▶ Robbins, *Creating More Effective Graphs*: short, accessible summary of classic advice by Tufte and Cleveland; includes a graph design checklist
- ▶ Few, *Show Me The Numbers*: examples in a business context; excellent advice on table design
- ▶ Ware, *Information Visualization*: thorough textbook on insights from perception research
- ▶ Weissgerber et al. (2015): article on showing all the data, not just statistical summaries
- ▶ Donahue, *Fundamental Statistical Concepts in Presenting Data*: real case studies from a statistical consultant

R code examples

Follow along using code and data at

<http://www.stat.cmu.edu/~jwieczor/>

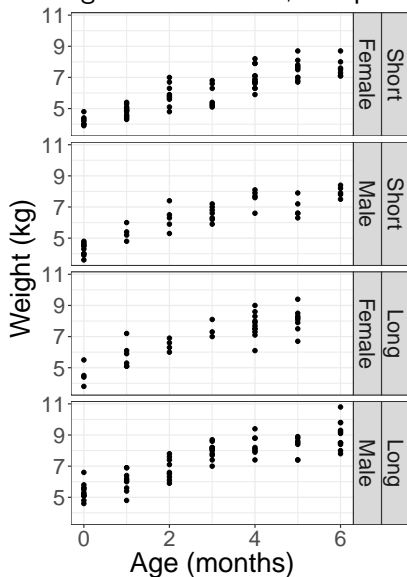
Thanks!

A few more principles, if there's time

- ▶ Align small multiples for the task
- ▶ Rank/order informatively
- ▶ Show derived variables directly
- ▶ Use Gestalt principles

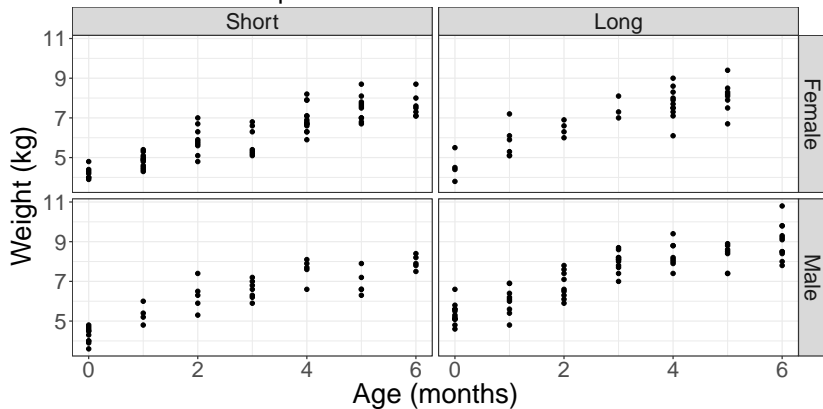
Alignment

Among male newborns, compare weights by length

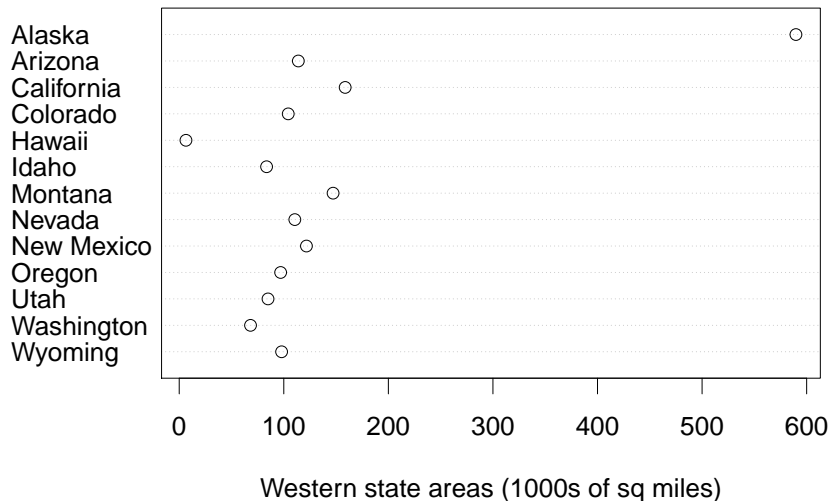


Alignment

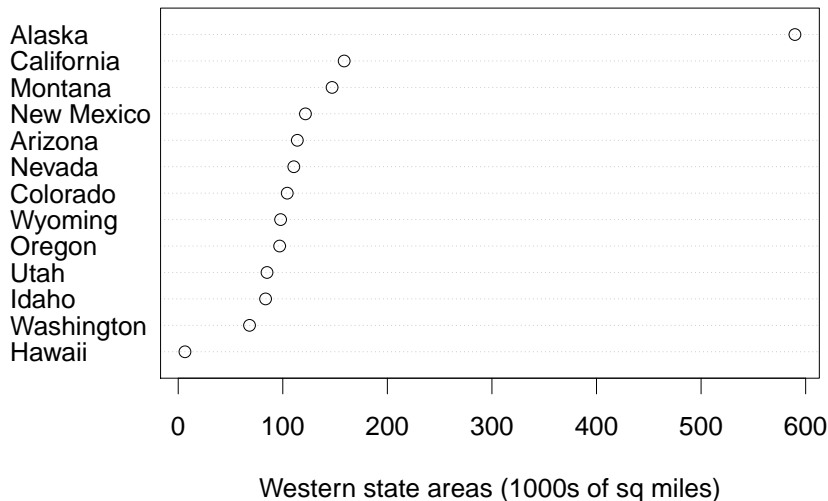
Among male newborns, compare weights by length:
easier search and comparison now



Ranking: alphabetical



Ranking: informative

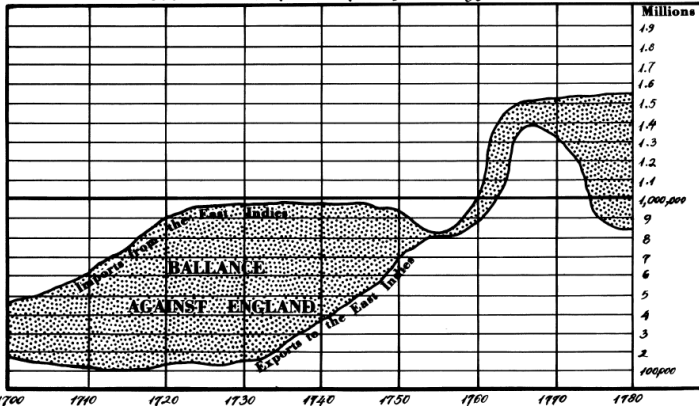


Derived variables

William Playfair, one of the earliest line charts

What does the difference look like?

CHART of EXPORTS and IMPORTS to and from the EAST INDIES
From the Year 1700 to 1780 by W. Playfair



The Bottom Line is Divided into Years the Right hand Line into HUNDRED THOUSAND POUNDS
is force Page 31st
Published in the Scot's Evening 16th Aug. 1785

Derived variables

Differences shown directly, by Cleveland and McGill

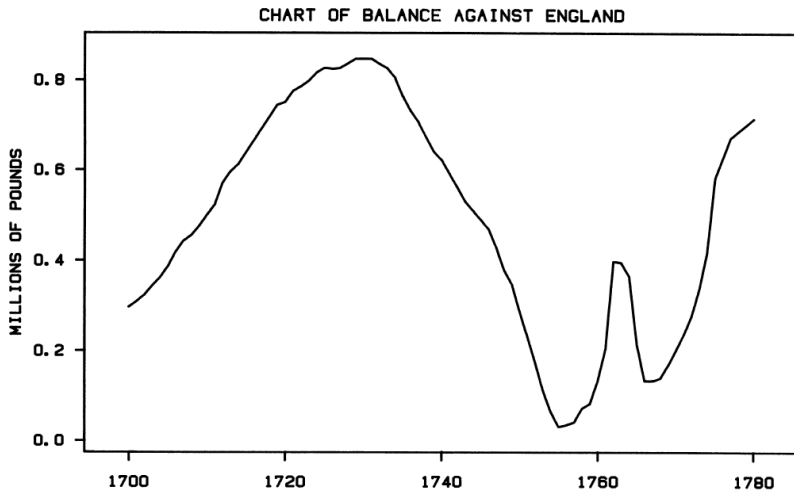


Figure 28. Playfair data.

Alignment, ranking, and derived variables

Lessons

- ▶ Decide on visual task, and helpfully align elements to be compared.
- ▶ As you explore data, try several arrangements.
- ▶ Order your dots/bars meaningfully: rank by a variable, not alphabetically.
- ▶ If differences or ratios are interesting, compute and plot them directly.

Gestalt

Gestalt = “pattern” in German

We automatically **structure data into patterns / groups** using certain features

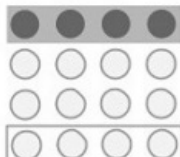
Proximity



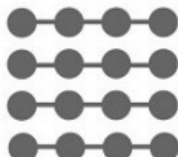
Similarity



Enclosure



Connection



Gestalt

Lessons

- ▶ Distinguish categorical groups by similarity, proximity, or enclosure.
- ▶ Use proximity to structure your layout (arrange small multiples).
- ▶ Use connection to show groups on line chart, parallel coordinates chart, or network graph.
- ▶ To highlight one group, use enclosure or similarity.