Visualizing quantitative data with R and RStudio

ME 447/547 Visualizing Data

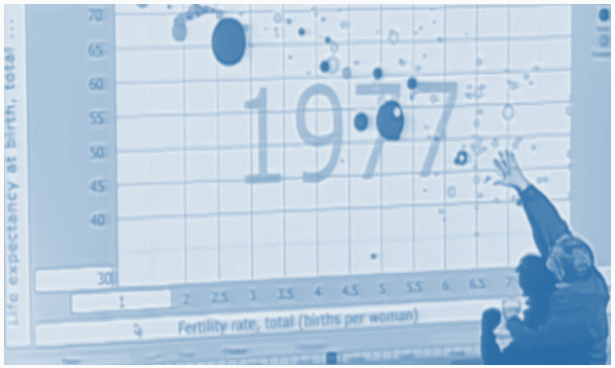
Richard Layton

December 2018

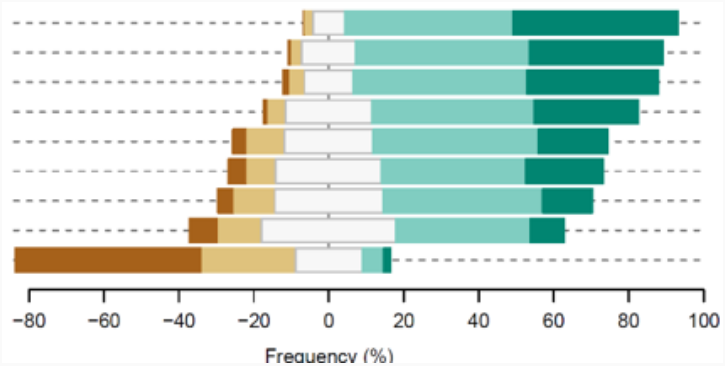
Rose-Hulman Institute of Technology

The course is designed to develop your skills in three areas

Rhetoric

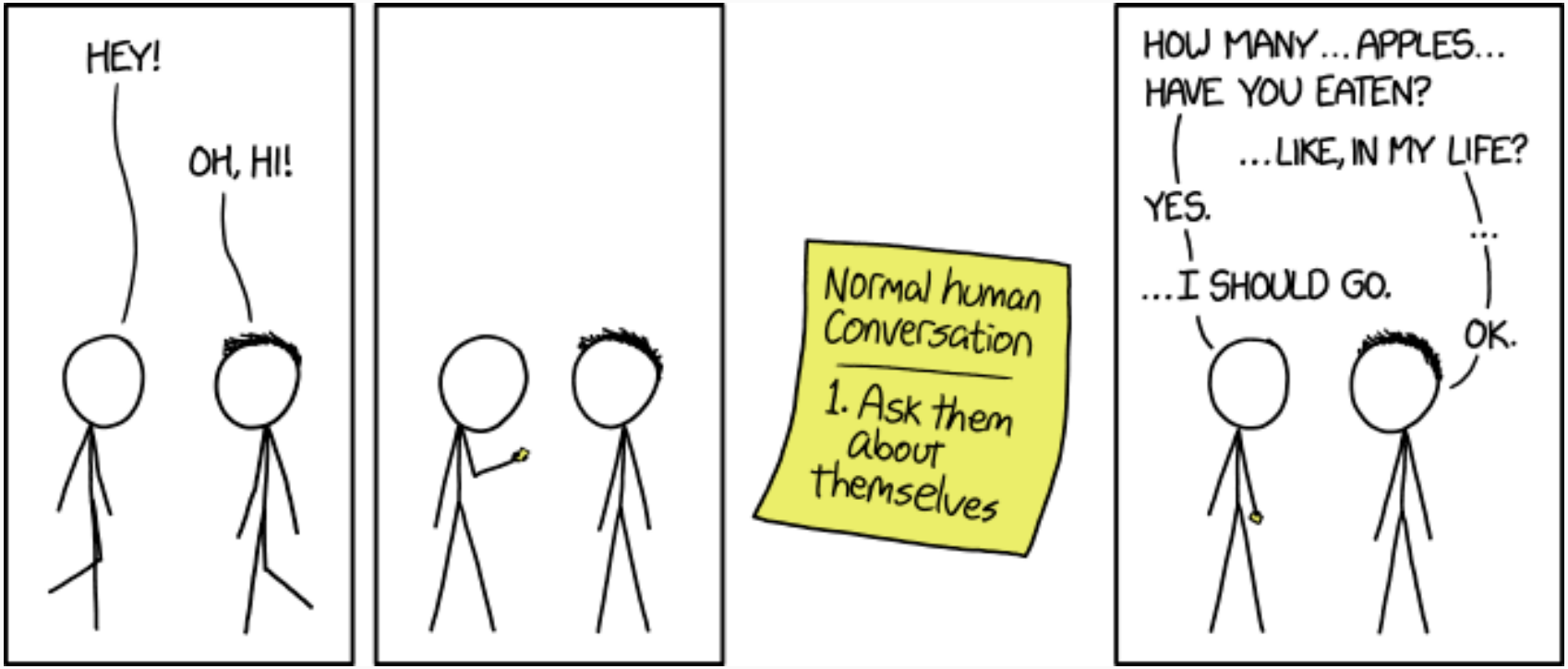


Repertoire



Means

Please find someone you don't know and introduce yourself



https://www.xkcd.com/1976/

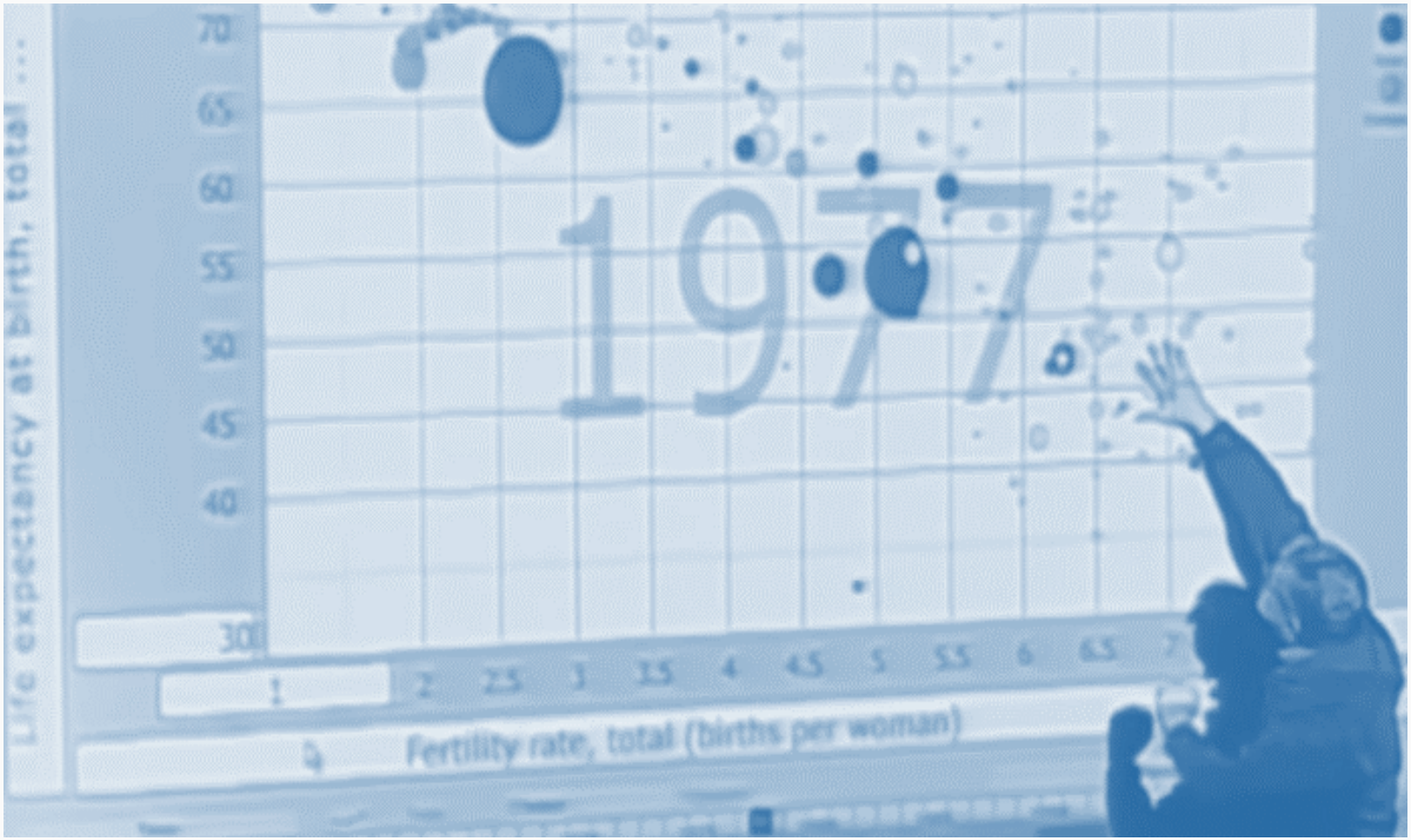
Have a seat together

Find out what non-academic interest or hobby your partner has

Prepare to introduce your partner and their interest

Visual rhetoric

Designers shape information visually for rhetorical ends



Hans Rosling 2006 TED Talk

Consider the argument

How did Hans shape the information visually?

What were his rhetorical goals?

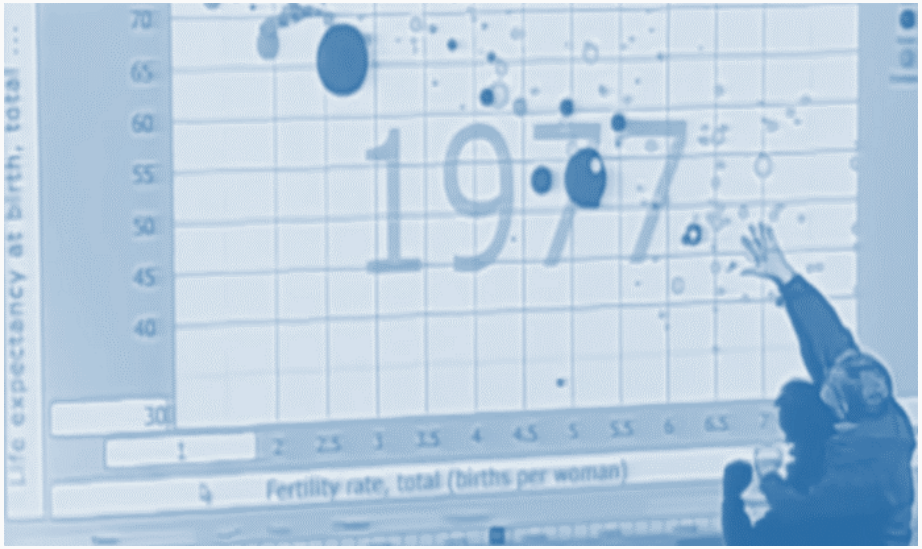


Image: TED2006

Consider this visual argument. True or false?

$N_{\text{people on welfare}} > N_{\text{people with a full time job}}$

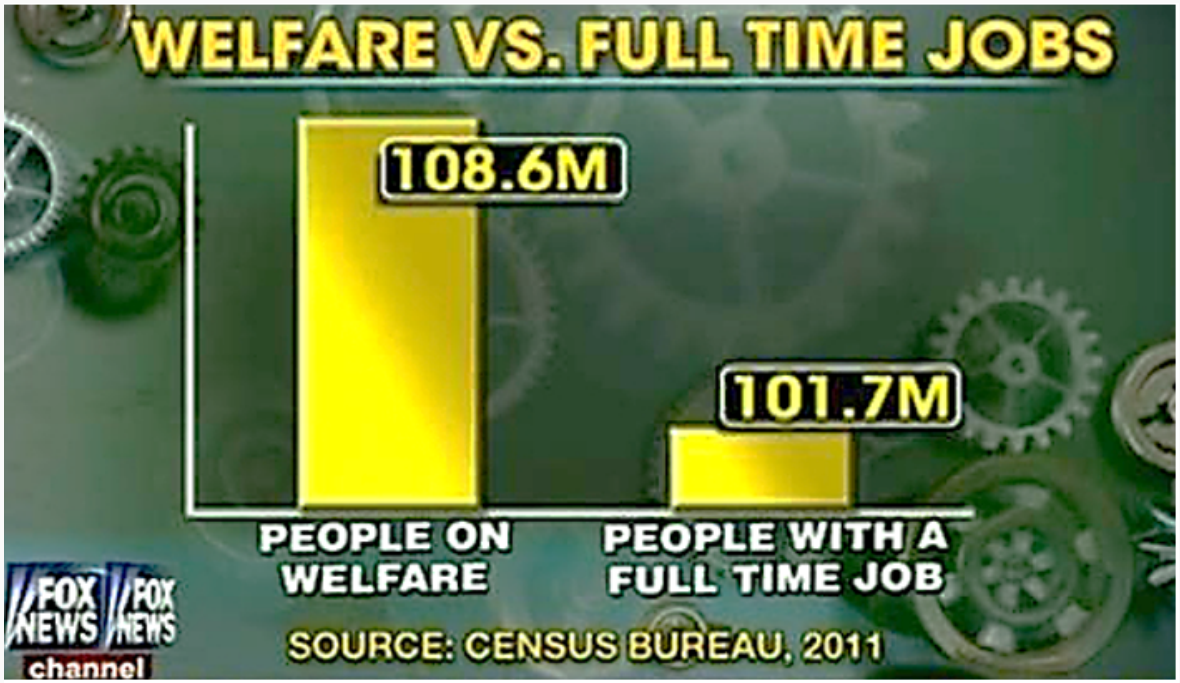


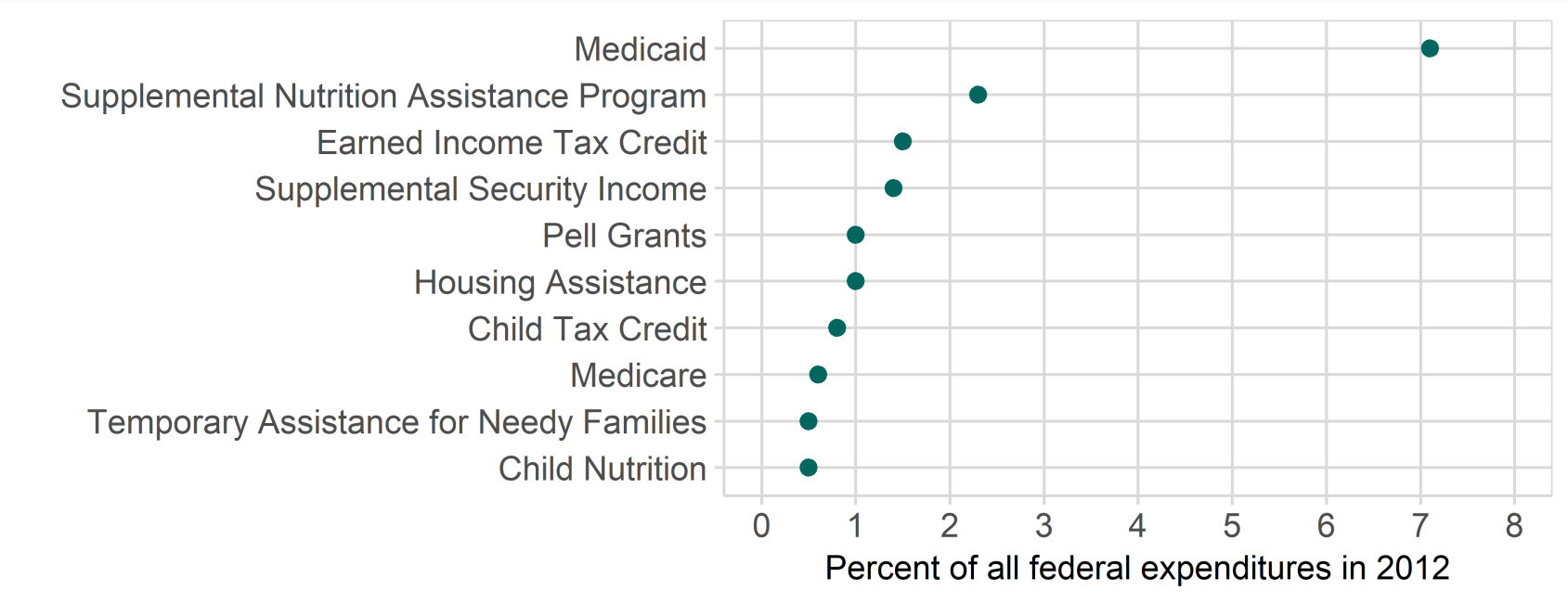
Image: Media Matters

False. One count is artificially high; the other is artificially low.

The counts use different definitions of “people”.

To avoid ambiguity, let's define "welfare"

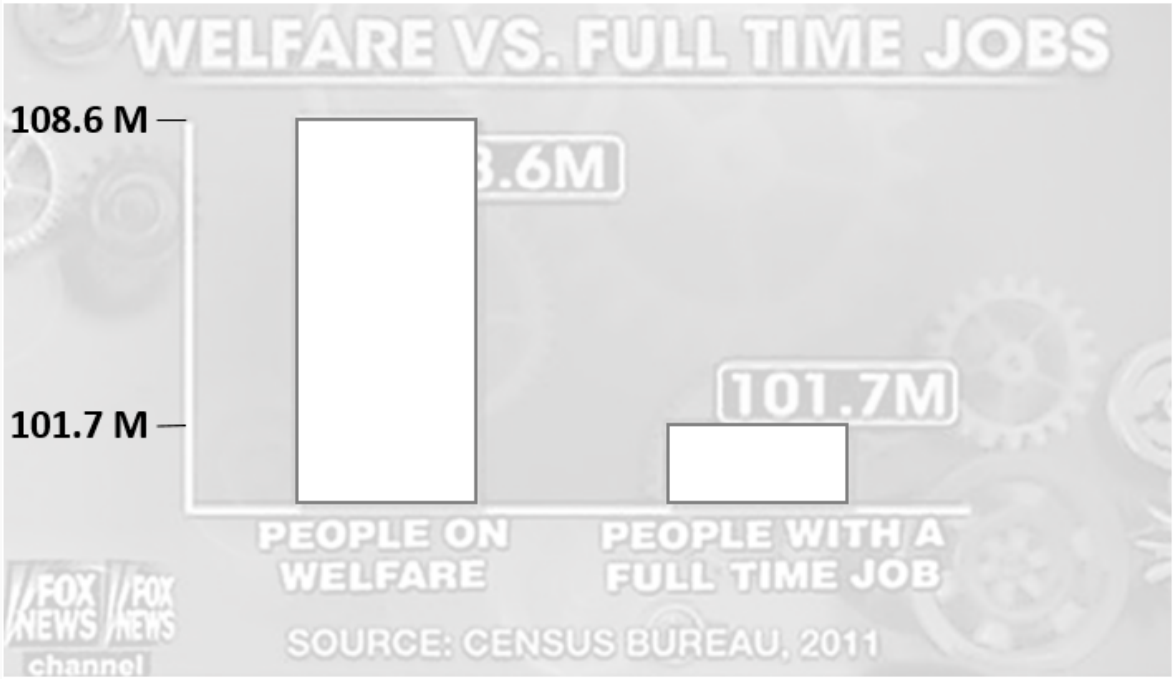
Federal means-tested programs and tax credits



In total, 17% of the 2012 US federal budget ($590 B / $3540 B).

In addition, the visual and textual arguments are in conflict

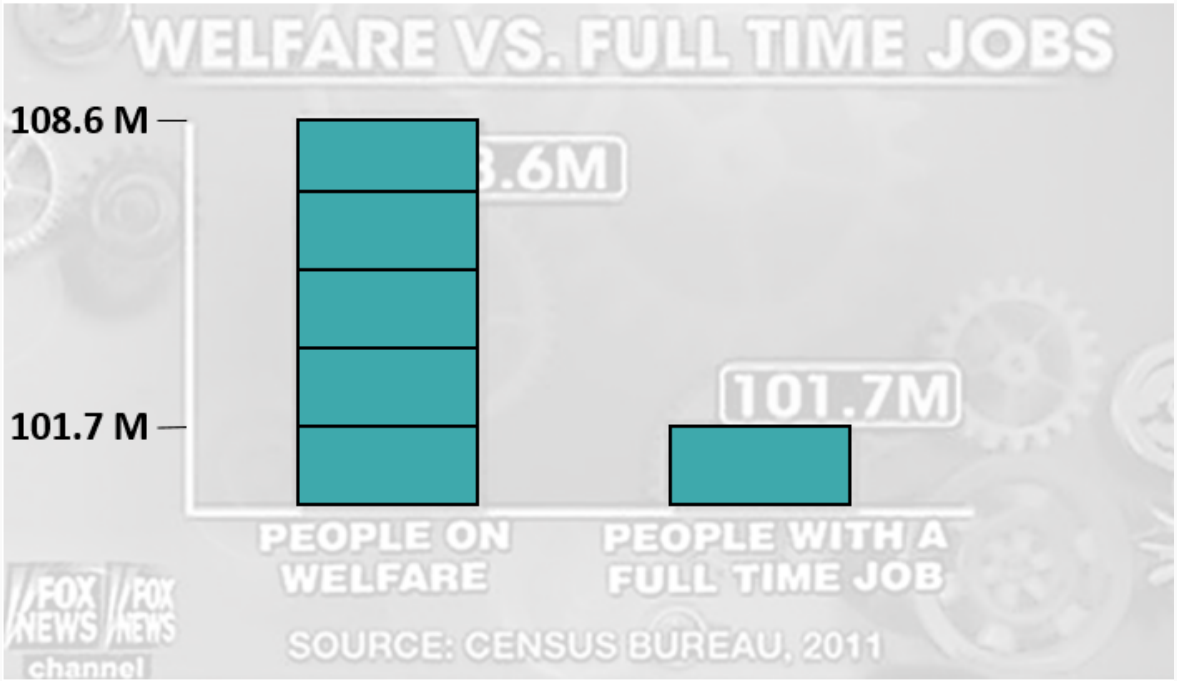
What is the visual lie?



As the designer well knows, a visual argument prevails

Textual argument: one number is 7% larger than the other

Visual argument: one group is 5 times larger than the other



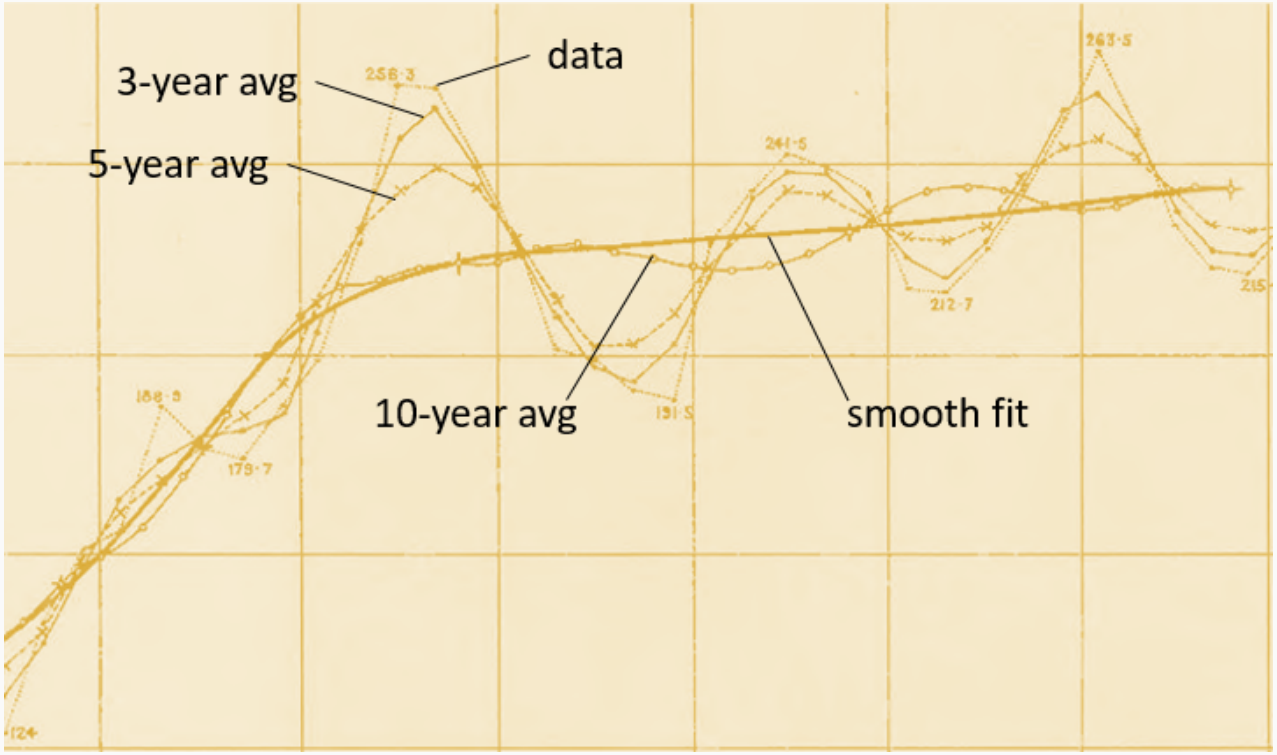
What were the designer's rhetorical goals?

Conclusion: Design entails an ethical obligation

Design is choice

Choice is power

Power confers responsibility



At issue was whether exports had become stationary. The three- and five-year moving averages show strong evidence of an approximately 10 year cycle. From Bowley (1901) *Elements of Statistics*.

Image: (Friendly, 2008)

Repertoire

Effective graph design begins by knowing the prior art

62

strip plot

box and whisker plot

multiway

scatterplot

dot plot

line graph

conditioning plot

63

scatterplot matrix

parallel coordinate plot

cycle plot

mosaic plot

financial (OHLC) plot

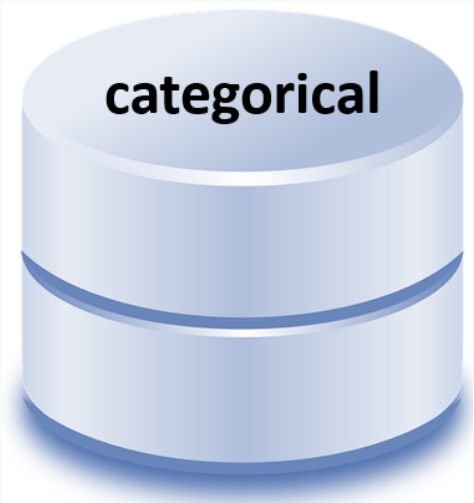
linked micromaps

diverging stacked bar

The data structure determines which designs are suitable

Number of variables?

Continuous or discrete?



Number of variables?

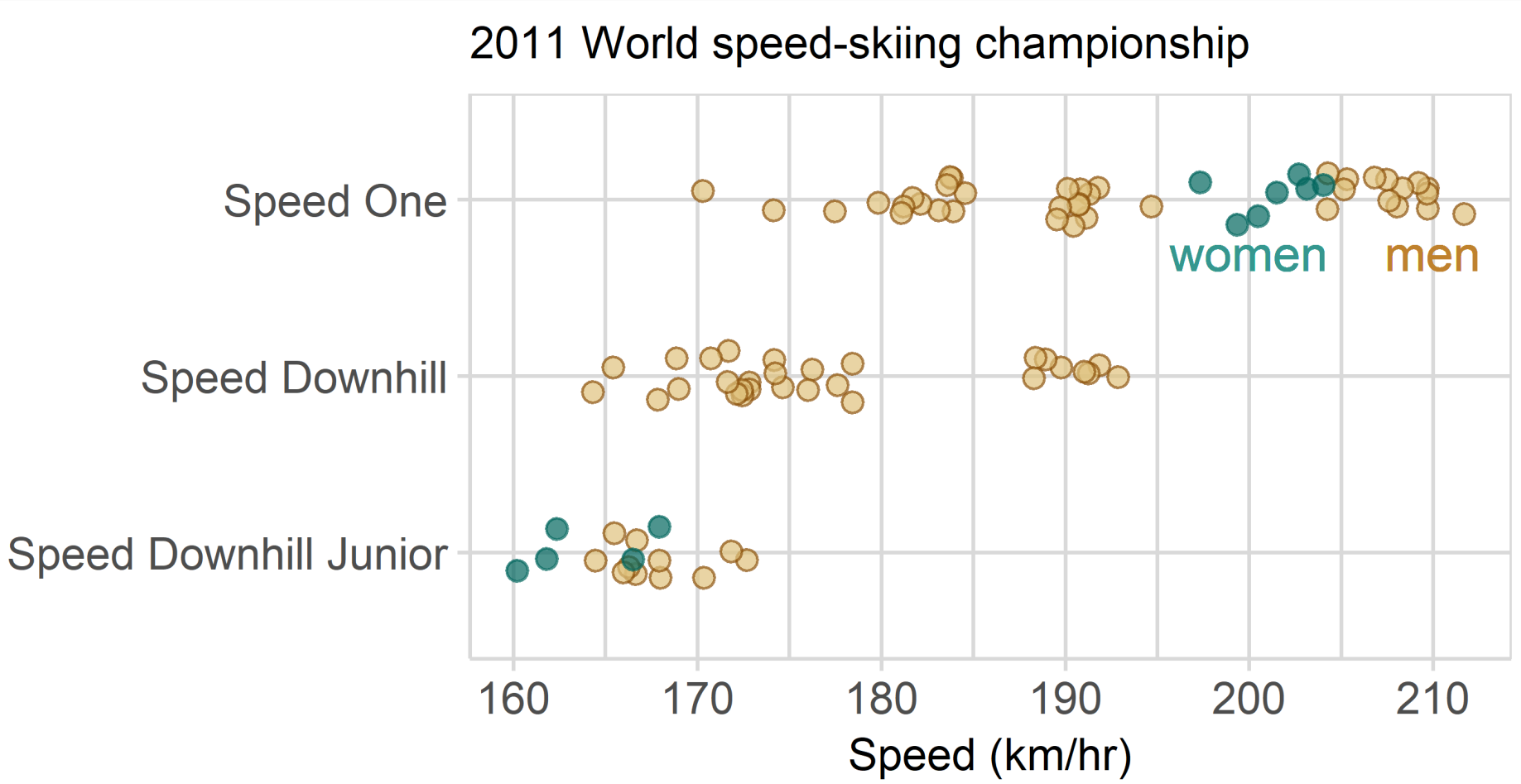
Nominal or ordinal?

Number of levels each?

Design gallery — strip plot (or jitter plot, 1D scatterplot)

Quantitative variable: Speed

Categorical variables: Event (3 levels) and sex (2 levels)



Design gallery — box and whisker

Movie franchises (Kirk Figure 6.19)

Minutes at 12 exposition (Robbins Figure 4.11)

Skiing (Unwin Figure 1.2)

Olive oil (Unwin Figure 10.4)

Design gallery — multiway

Population data by county (Robbins Figure 8.15)

Livestock (Cleveland Figure 6.1 and 6.2)

Any of the midfieldr data

Design gallery — scatterplot

Life expectancy by country (Kirk Figure 6.30)

Crime rates by state (Kirk Figure 6.31) bubble plot

Weight and height by sport (Unwin Figure 5.11)

Olive oil (Unwin Figure 10.12 and10.14)

regression, linear and loess

Design gallery — Cleveland dot plot

State areas (Robbins Figure 4.3) with log base2 scale

midfieldr graphs, e.g., grad rate, starters, etc.

Brain and body mass by species (Cleveland/Elements Fig. 1.7)

Language speakers (Cleveland/Elements Fig. 1.9)

Fraction of journal space in graphs (Cleveland/Elements Figure 3.22)

Design gallery — line graph

Energy data (Robbins Figure 5.3)

Stock market (Robbins Figure 6.10)

Blood level data (Robbins Figure 7.21)

Car production (Robbins Figure 8.13)

Design gallery — conditioning plot

Rubber properties (Cleveland Figure 4.4)

NOx (Cleveland Figure 4.6)

Solar radiation (Cleveland Figure 5.5)

Design gallery — scatterplot matrix

Energy data (Robbins Figure 5.6)

Blood level data (Robbins Figure 7.22)

Rubber properties (Cleveland Figure 4.1)

Solar radiation (Cleveland Figure 5.1)

Pima Indians diabetes (Unwin Figure 1.9)

Crime rates (Unwin Figure 5.12)

Design gallery — parallel coordinate

Nutrient contents (Kirk Figure 6.32)

Food data set (Unwin Figure 6.2)

Design gallery — cycle plot

St Louis Science Center attendance (Robbins Figure 4.18)

CO2 (Robbins Figure 4.20) or (Cleveland Figure 3.75))

Arctic ice (?)

Design gallery — mosaic plot

Monterrey Bay Aquarium (Robbins Figure 5.10)

Titanic (Unwin Figure 7.2)

Treatment (Unwin Figure 7.9)

Design gallery — financial (OHLC) plot

Dow Jones (Robbins Figure 5.16)

Gold price (Robbins Figure 8.2)

Design gallery — linked micromaps

Soybean data (Robbins Figure 5.11)

Design gallery — proportional symbol map

Election funding raised (Kirk Figure 6.50)

Design gallery — dot map

Race and location (Kirk Figure 6.52)

Tornado data (Brunsdon Figure 5.1)

CATME world map

Broad Street cholera map (Bivand Figure 4.7)

Design gallery — diverging stacked bar

Literacy proficiency (Kirk Figure 6.24)

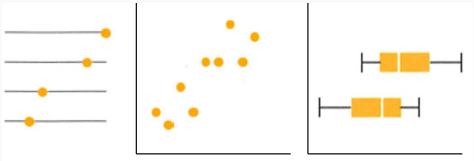
Student volunteers (Evergreen Figure 5.2)

My teaching evaluations

One from the Robbins paper

Implications for the designer

Grasp the data structure



Explore using suitable designs

Refine the logic of your argument

Consider original designs if required by the story

Meet the needs of the audience

Means

Use the right tool for the job

 RStudio	primary interface, integrates all our software
 R	tidying data and creating graphs
 R markdown	writing the portfolio, interleaving prose with code
 Git	local version control
 GitHub	collaborating and publishing the portfolio

The main topical threads weave through the calendar

data

software

visual rhetoric

repertoire of graphs

portfolio

calendar

paper reprint, with permission
e-copy on Moodle, with permission

w	d	agenda & assignments
1	M	Course goals and outcomes Sign-out two reprints
	T	Introduction to visual rhetoric Install software
	R	Relating data structure to graph design Doumont (2009) Designing the graph
	F	Software lab
2	M	Graph basics with ggplot2 Practice
	T	Tufte (1997) Decision to launch Challenger
	R	Data basics Practice
	F	Data lab with file management Return reprints

https://github.com/DSR-RHIT/me447-visualizing-data

References

Friendly M (2008) A brief history of data visualization. Chen C-h, Härdle W, and Unwin A eds. *Handbook of Data Visualization*. Springer-Verlag, Berlin, 15–56