Graphical Data Analysis: foolish displays and fully informative displays, how can you tell the difference?

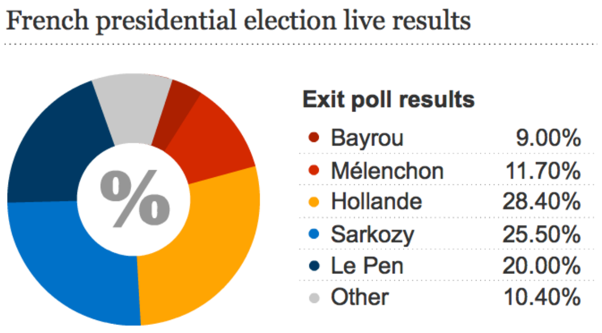
Antony Unwin
University of Augsburg
unwin@math.uni-augsburg.de

PolBeRG/ELECDEM Workshop Budapest 27th April, 2012

Warning: Thinking required

Some graphics examples

- Do you like the graphic?
- What can you see?
- What is the message?
- What other information might there be in the data?
- Is the graphic ‘foolish’ or ‘fully informative’?

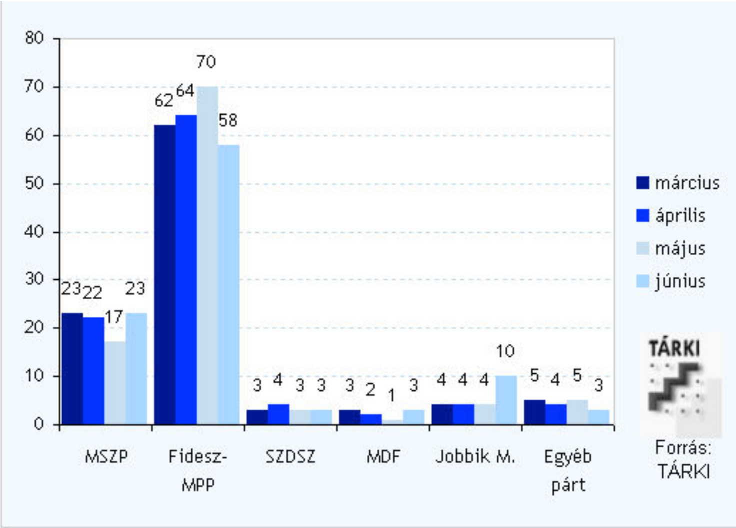


French presidential election live results

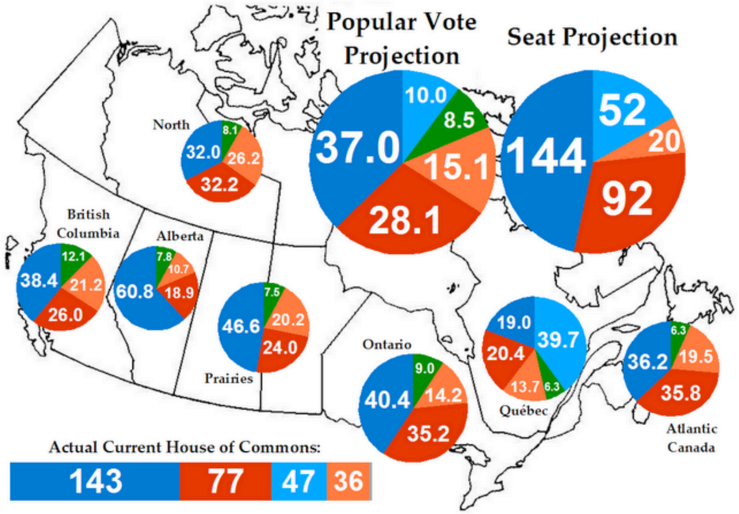
Exit poll results

● Bayrou	9.00%
● Mélenchon	11.70%
● Hollande	28.40%
● Sarkozy	25.50%
● Le Pen	20.00%
● Other	10.40%

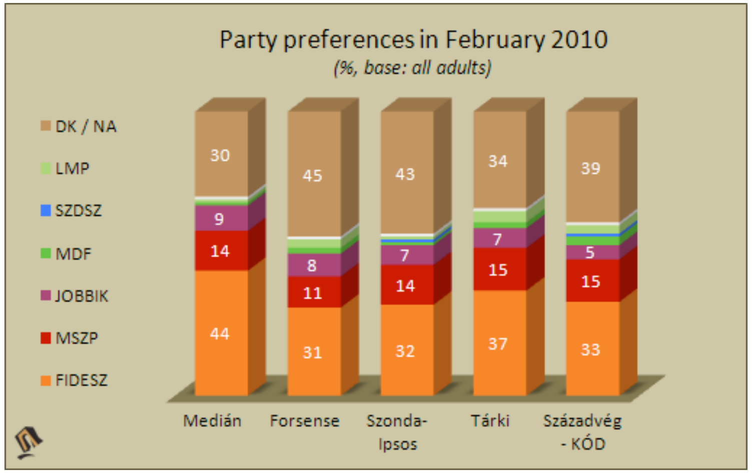
BBC website c. 21.25 on 23. April 2012



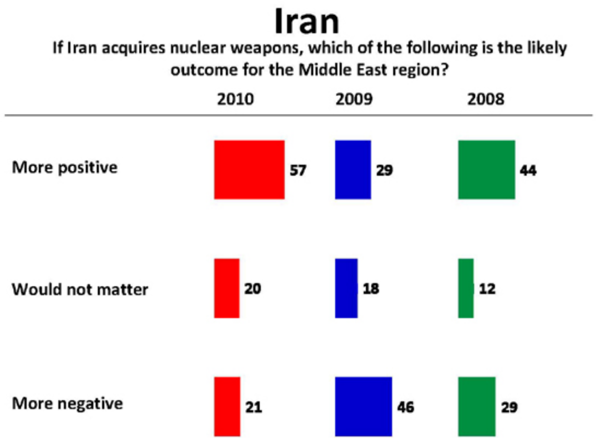
Hungarian Spectrum June 2009



Canadian Projection 18th February 2011



RiskandForecast.com 9th March 2010



Arab Opinion Survey June 29-July 20, 2010

US Census 1890

63 million Hollerith cards

Ordering does not have to be alphabetic

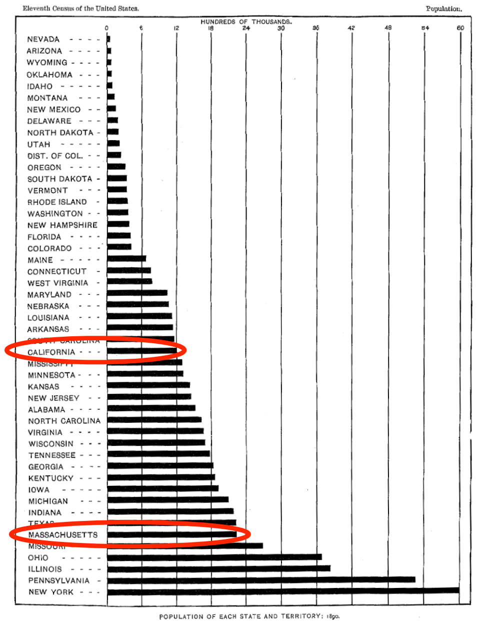
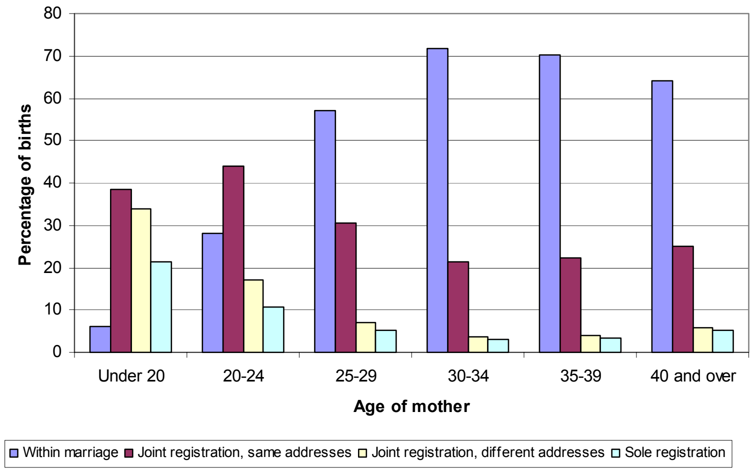
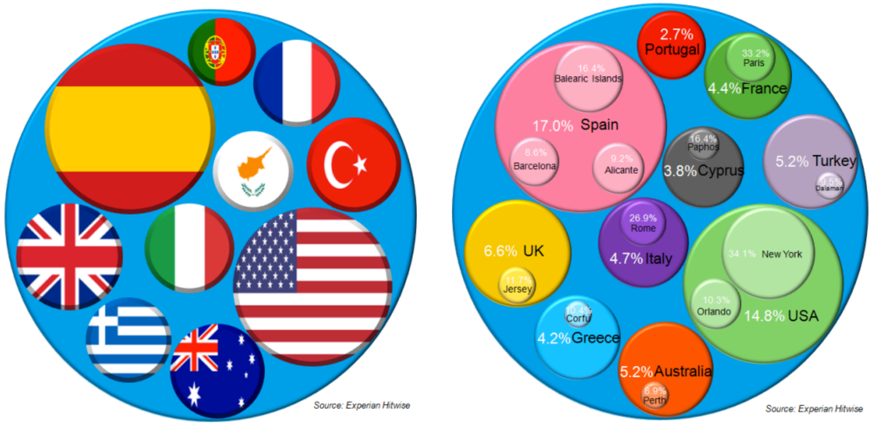


Figure 2 – Percentage of live births by mother's age and type of registration, England and Wales, 2008

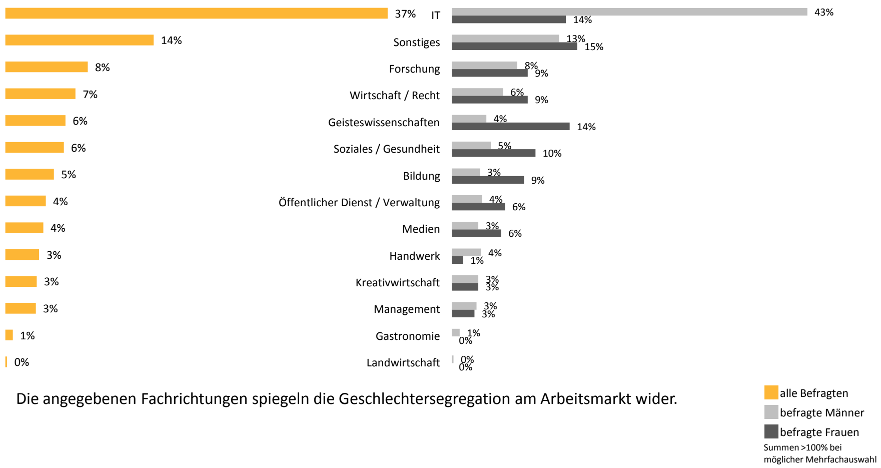




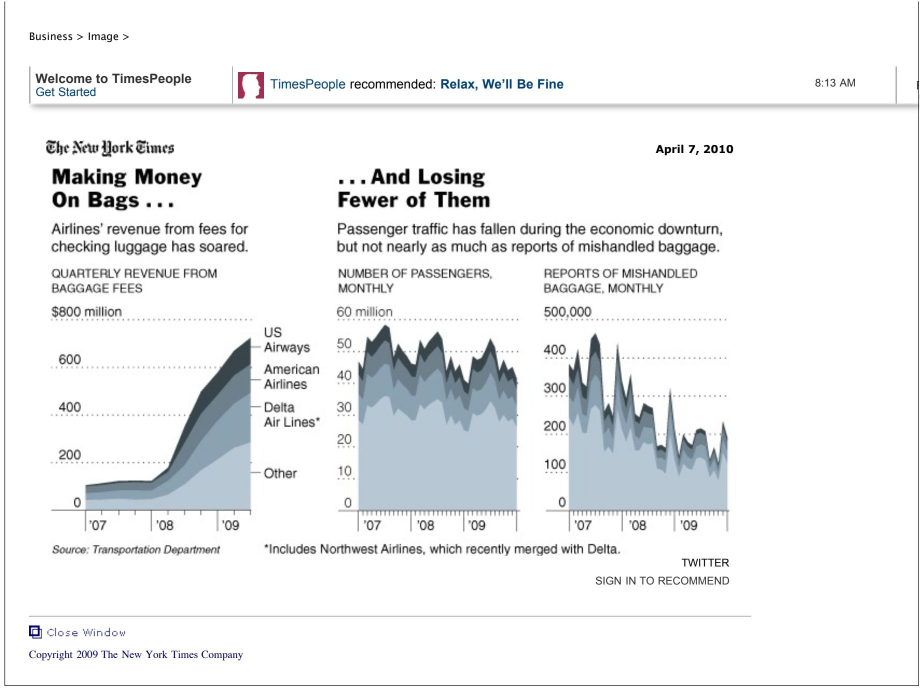
Flight searches by the UK Internet population

weblogs.hitwise.com/james-murray/2011/09/flight_search_infographic_new.html

German Pirates' Survey



Die angegebenen Fachrichtungen spiegeln die Geschlechtersegregation am Arbeitsmarkt wider.



Business > Image >

Welcome to TimesPeople
Get Started

TimesPeople recommended: **Relax, We'll Be Fine**

8:13 AM

The New York Times

April 7, 2010

Making Money On Bags . . .

Airlines' revenue from fees for checking luggage has soared.

QUARTERLY REVENUE FROM BAGGAGE FEES

. . . And Losing Fewer of Them

Passenger traffic has fallen during the economic downturn, but not nearly as much as reports of mishandled baggage.

NUMBER OF PASSENGERS, MONTHLY

REPORTS OF MISHANDLED BAGGAGE, MONTHLY

Source: Transportation Department

*Includes Northwest Airlines, which recently merged with Delta.

TWITTER
SIGN IN TO RECOMMEND

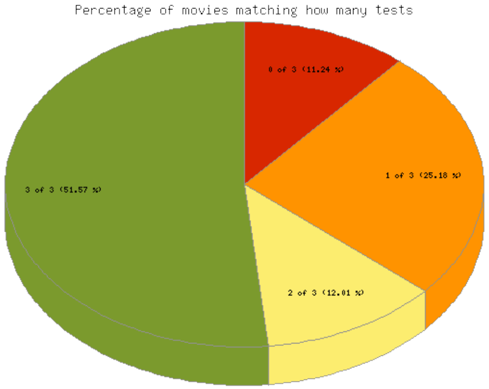
Close Window

Bechdel Test for Films

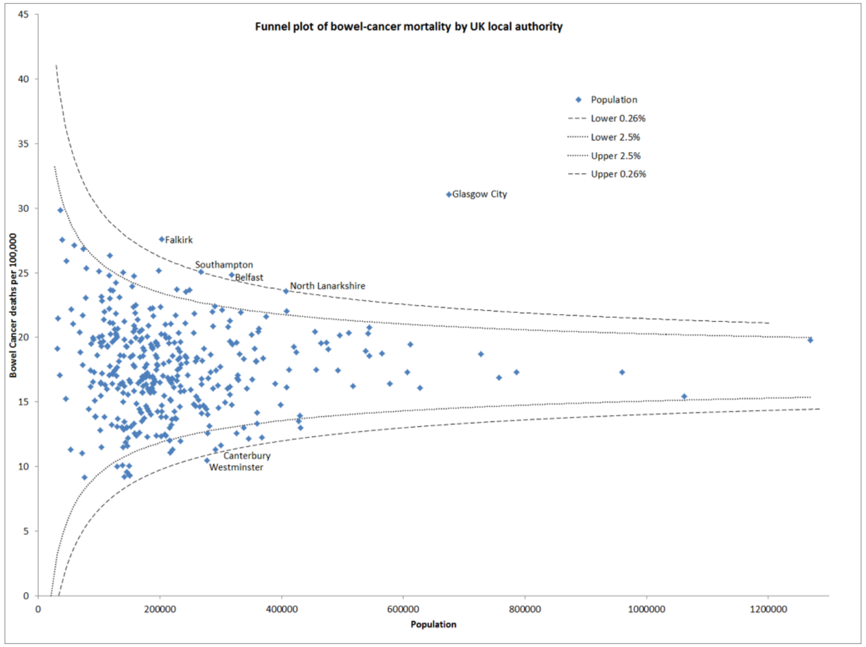
1. It has to have at least two women in it

2. Who talk to each other

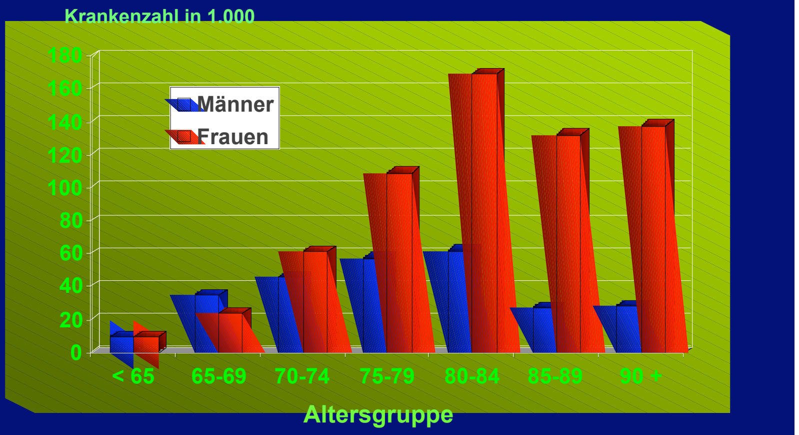
3. About something besides a man



bechdeltest.com/statistics/

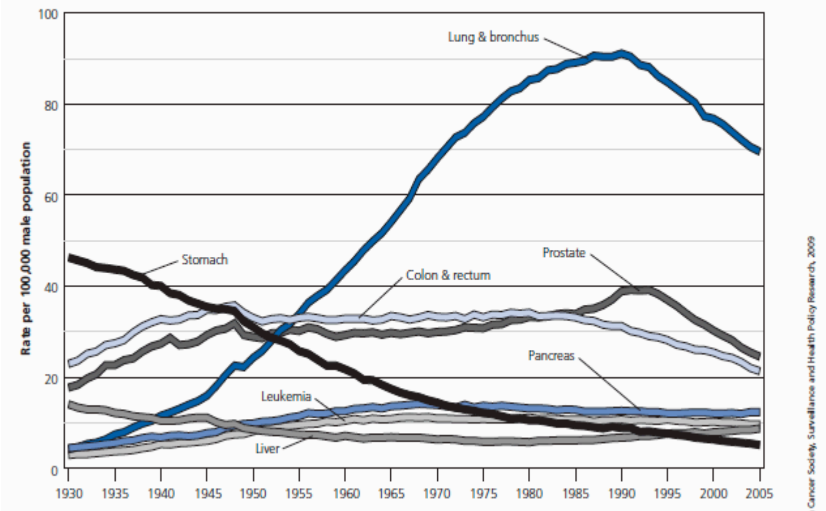


Geschätzte Verteilung der Demenzkranken in Deutschland zum Ende des Jahres 2002 nach Geschlecht und Alter *



* EURODEM-Daten; Lobo et al. (2000) Neurology 54, Suppl. 5: 4-9

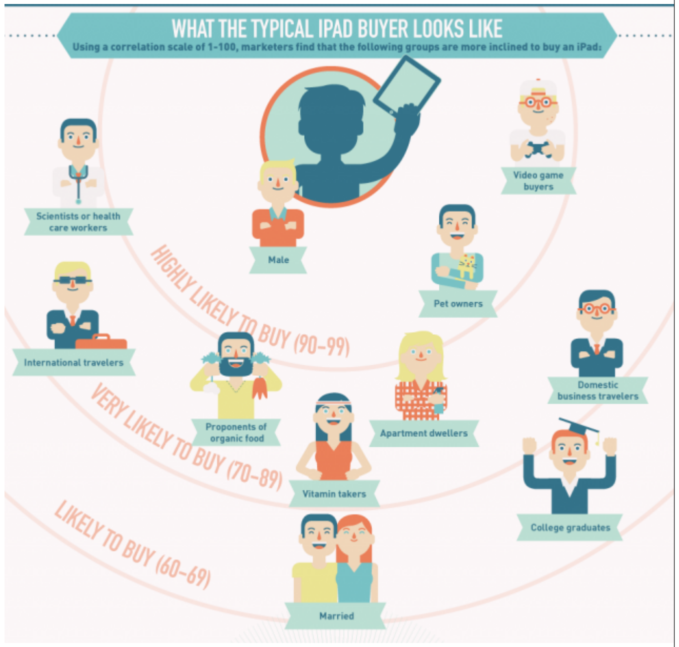
Age-adjusted Cancer Death Rates,* Males by Site, US, 1930-2005



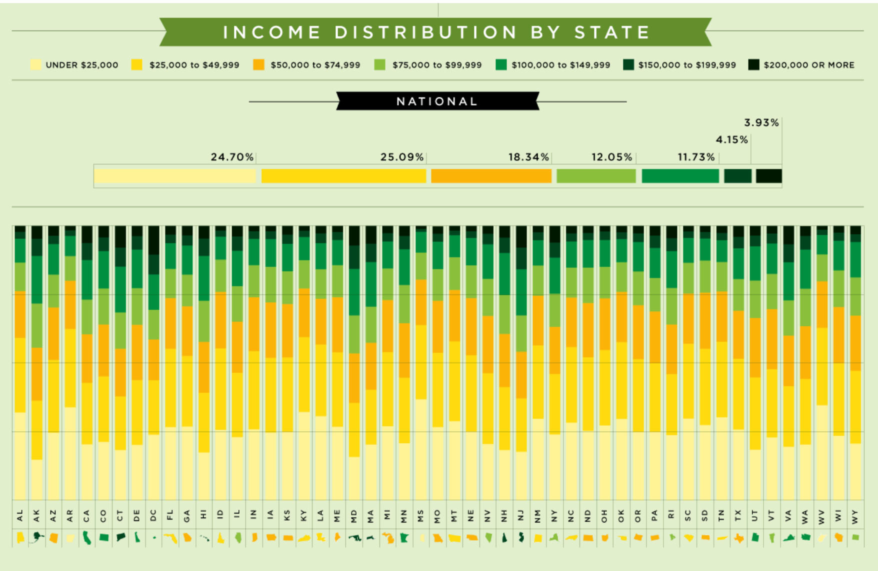
*Per 100,000, age adjusted to the 2000 US standard population.

Note: Due to changes in ICD coding, numerator information has changed over time. Rates for cancer of the liver, lung and bronchus, and colon and rectum are affected by these coding changes.

Source: US Mortality Data, 1960 to 2005, US Mortality Volumes, 1930 to 1959, National Center for Health Statistics, Centers for Disease Control and Prevention, 2008.



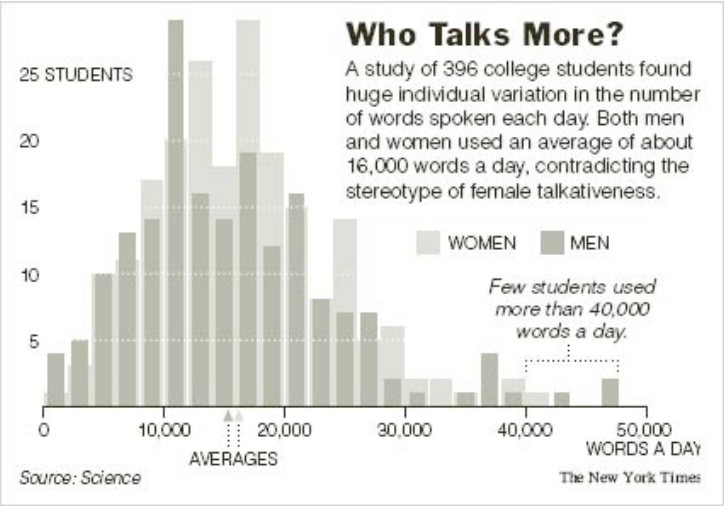
WHAT THE TYPICAL IPAD BUYER LOOKS LIKE
Using a correlation scale of 1-100, marketers find that the following groups are more inclined to buy an iPad:
Scientists or health care workers
Male
Video game buyers
Pet owners
International travelers
HIGHLY LIKELY TO BUY (90–99)
Domestic business travelers
Proponents of organic food
Apartment dwellers
VERY LIKELY TO BUY (70–89)
Vitamin takers
College graduates
LIKELY TO BUY (60–69)
Married



junkcharts.typepad.com/junk_charts/infographics/

The New York Times

July 6, 2007

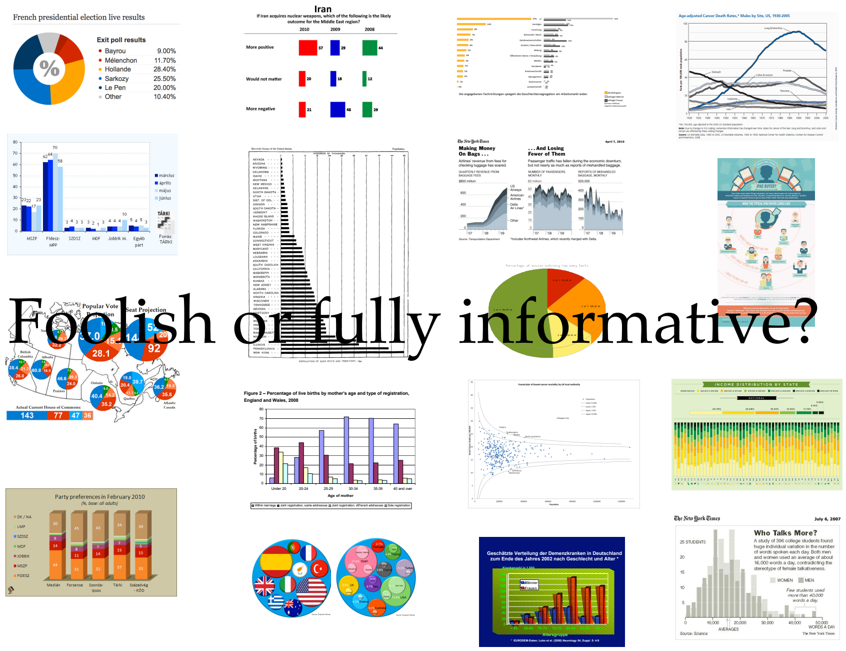


Who Talks More?

A study of 396 college students found huge individual variation in the number of words spoken each day. Both men and women used an average of about 16,000 words a day, contradicting the stereotype of female talkativeness.

Source: Science

The New York Times



Foolish or fully informative?

Interpreting graphics

- Does everyone see the same things?
- How can the information be described verbally?
  - "a picture is worth a 1000 words"
  - twitter is limited to 140 characters (how many for a graphic?)
- How important are
  - background knowledge
  - scales and labelling
  - title, caption, legend, guides, annotations, accompanying text?
- How can the information be assessed statistically?

Presentation Graphics: Questions and Principles

Questions

- What variables and data are shown?
- What is the source of the data? Is it reliable?
- How much data? Could more data be obtained?
- Data quality? Likely accuracy, reliability
- Graphic quality? Appropriate form, distortion, …
- Coherency: Do the title, caption, labels, scales, legend, annotations, accompanying text all tell the same story?
- Does the story match the data?

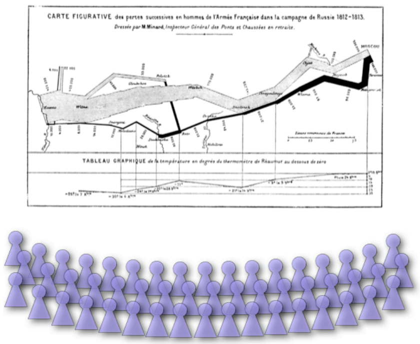
Principles

- Graphics are for displaying or uncovering (qualitative) information not for presenting exact (quantitative) data. Present data in tables.
- Several simple graphics may be better than one complex one.
- Colour should be used with care and good taste.
- Scales are important (min, max, zero, units, orders)
- Size, aspect ratio, frames, grids make a difference.
- Consider: Content, Context, Construction

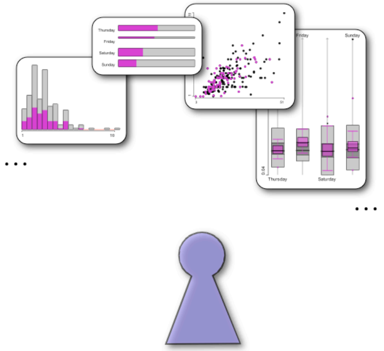
Software

- (Whatever you can work well with…)
- R and its packages
  - *ggplot2*
  - *lattice*
  - *vcd*
  - …
  - and then get a designer to help

Presentation Graphics / Exploratory Graphics



Presentation



Exploration

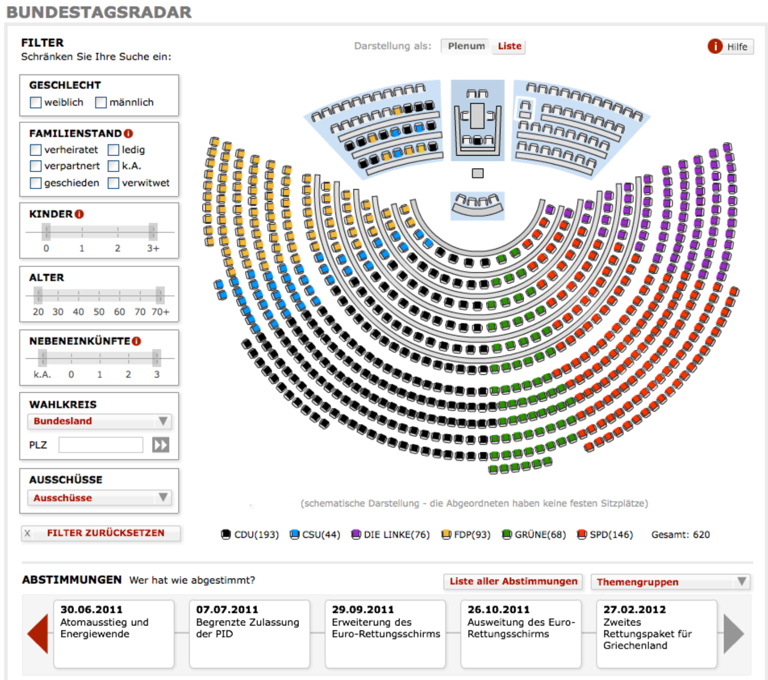
Presentation v. Exploration

- Presentation graphics usually involve only one graphic for viewing by a huge number of people
- Exploratory graphics usually involve a huge number of graphics for viewing by only one person
- Presentation graphics convey known information
- Exploratory graphics are used to find information
- Presentation graphics should attract attention
- Exploratory graphics should direct attention

Why visualize to explore?

- Look for global trends
  - overall structure
- Look for local features
  - data quality
  - groups or clusters
  - outliers, tail distributions and extremes
  - patterns of all kinds

Exploratory Examples



BUNDESTAGSRADAR

FILTER
Schränken Sie Ihre Suche ein:

GESCHLECHT
□ weiblich □ männlich

FAMILIENSTAND
□ verheiratet □ ledig
□ verpartnert □ k.A.
□ geschieden □ verwitwet

KINDER
0 1 2 3+

ALTER
20 30 40 50 60 70 70+

NEBENEINKÜNFTE
k.A. 0 1 2 3

WAHLKREIS
Bundesland
PLZ

AUSSCHÜSSE
Ausschüsse

X FILTER ZURÜCKSETZEN

Darstellung als: Plenum Liste

(schematische Darstellung - die Abgeordneten haben keine festen Sitzplätze)

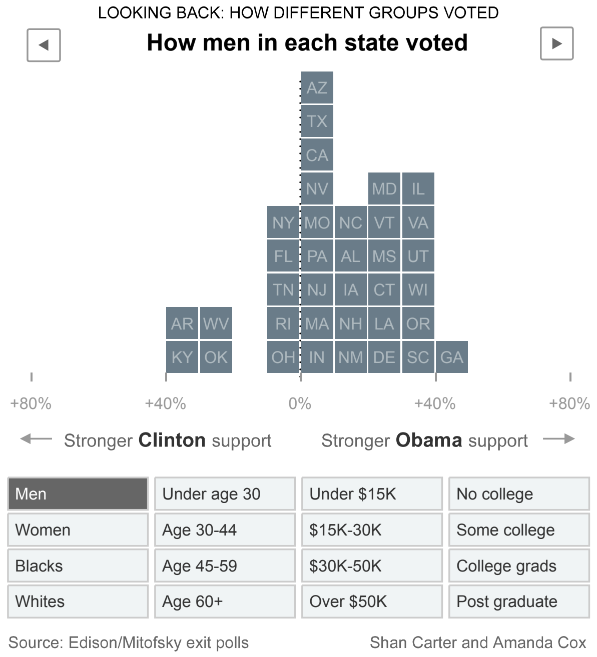
CDU(193) CSU(44) DIE LINKE(76) FDP(93) GRÜNE(68) SPD(146) Gesamt: 620

ABSTIMMUNGEN Wer hat wie abgestimmt?

30.06.2011	07.07.2011	29.09.2011	26.10.2011	27.02.2012
Atomausstieg und Energiewende	Begrenzte Zulassung der PID	Erweiterung des Euro-Rettungsschirms	Ausweitung des Euro-Rettungsschirms	Zweites Rettungspaket für Griechenland

Bundestagsradar

- www.spiegel.de/flash/flash-22868.html
- Positive
  - Colour, Querying, Filtering
- Negative
  - individual identification not comparisons
  - designed for querying not assessment
  - group sizes not shown graphically
  - dialog selection not direct selection



LOOKING BACK: HOW DIFFERENT GROUPS VOTED

How men in each state voted

Source: Edison/Mitofsky exit polls

Shan Carter and Amanda Cox

How different groups voted

- bit.ly/WrAgh
- Positive
  - animation, simple controls, querying, fixed scales
- Negative
  - no state size information (not by total or group)
  - states with insufficient information not listed
  - wide binwidths

Large survey example

Bowling Alone (DDB Lifestyle survey 1975-1998)

Cases 84989

Variables

Church
Religfun
Religion
Age
Gender
Education
Region
Income
Fistfight
… c. 400

20 CHURCH Attended church or other place of worship (freq last 12 months)

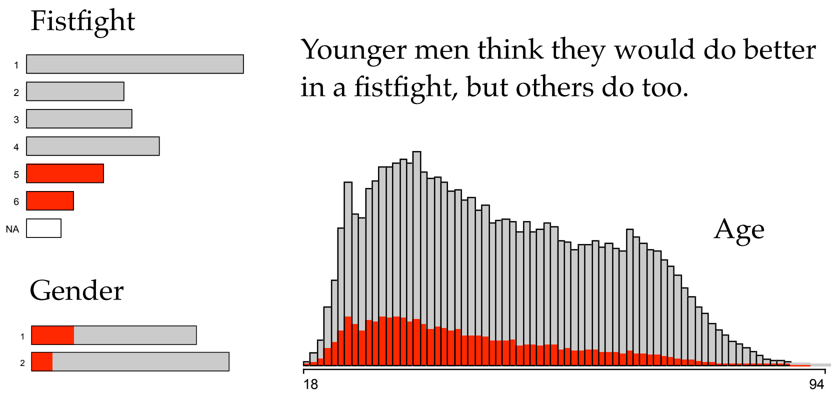
1	None
2	1-4 times
3	5-8 times
4	9-11 times
5	12-24 times
6	25-51 times
7	52+ times

208 RELIGFUN Religious fundamentalism is the greatest peril in the country today

1	Definitely Disagree
2	Generally Disagree
3	Moderately Disagree
4	Moderately Agree
5	Generally Agree
6	Definitely Agree

209 RELIGION Religion is an important part of my life

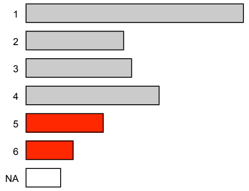
Linking of many variables



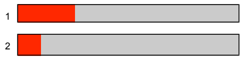
Younger men think they would do better in a fistfight, but others do too.

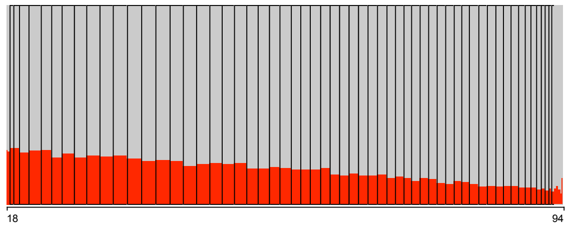
Spineplots and spinograms

Fistfight



Spineplot of gender





Spinogram of age

Example: Titanic disaster

2201 passengers and crew classified by

— gender

— age (child or adult)

— ship's class (1st, 2nd, 3rd, crew)

— survived or died

Virginian-Pilot.

THE NORFOLK LANDMARK

OVER FIFTEEN HUNDRED SANK TO DEATH WITH GIANT WHITE STAR STEAMER TITANIC

Horrible Disaster Greatest In Marine History of World

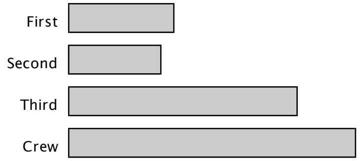
VESSEL REMAINED AFLOAT FOUR HOURS AFTER CRASH

866 PASSENGERS RESCUED

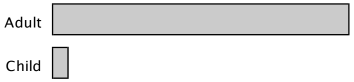
GIGANTIC TITANIC WAS VERITABLE FLOATING HOTEL

(R. J. MacG. Dawson, J. Statistics Education 3 no 3, 1995)

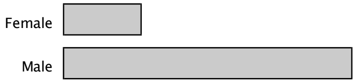
Titanic basic barcharts



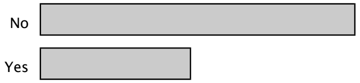
Class



Age



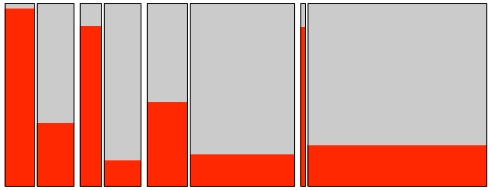
Sex



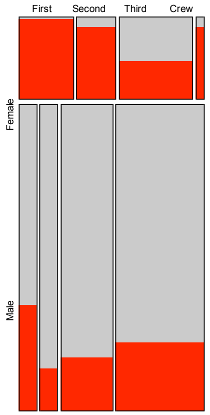
Survived

Titanic Disaster

Survival by gender and class



Gender within class



Class by gender

Titanic Comments

- It is difficult to display multivariate categorical data.
- There are several different kinds of mosaicplot and many different orderings and selections of variables.
- Which mosaicplot is best is a matter of taste.
- Choosing an effective mosaicplot requires speed and flexibility.
- Building mosaicplots up step by step helps explain them to others.

Exploratory Graphical Analysis

- Use lots of graphics
  - try different versions of the same graphic
  - use different graphics for the same data
  - use small multiples (e.g. trellis/lattice)
  - use combinations of graphics (plot ensembles)
  - use interactive graphics
- Datasets are rarely independent random samples (as is assumed in Statistics), so generalise with care

Graphics Books

- “Grammar of Graphics” L. Wilkinson
- "Interactive Graphics for Data Analysis" M. Theus, S. Urbanek
- "Graphics of Large Datasets" A. Unwin, M. Theus, H. Hofmann
- "Handbook of Data Visualization" (eds. Chen, Härdle, Unwin)
- ***Books by Edward Tufte, Bill Cleveland, Howard Wainer***

Websites (1)

- Gallery of Data Visualization
  - www.math.yorku.ca/SCS/Gallery/
- Statistical Modeling, Causal Inference, and Social Science
  - www.stat.columbia.edu/~gelman/blog/
- UK Local Government (public)
  - www.improving-visualisation.org
- Tableausoftware (commercial)
  - www.tableausoftware.com

Websites (2)

- Many Eyes
  - manyeyes.alphaworks.ibm.com/manyeyes/
- Junk Charts
  - junkcharts.typepad.com/
- Flowing Data
  - flowingdata.com
- Ask ET (Ed Tufte)
  - www.edwardtufte.com

Websites (3)

- Martin Theus Blog
  - www.theusRus.de/blog
- Guardian newspaper
  - www.guardian.co.uk/data-store
- New York Times Graphics
  - www.smallmeans.com/new-york-times-infographics/
- Name voyager and name mapper (some entertainment)
  - www.babynamewizard.com

Summary

- Presentation graphics in the media are often poor and should be interpreted with care
- Follow good graphics principles (and get design help)
- Exploratory graphics are different
  - draw many graphics
  - use multiple graphics
- Datasets contain many different kinds of information
  - graphics are good for finding and for presenting