ASEARC Course on

Categorical Data Analysis

Professor Eric Beh

School of Mathematical & Physical Sciences
University of Newcastle

Semester 2, 2017

The weekly sequence of topics we shall (hopefully) stick to includes

Part One –Measures of Association for Categorical Data Analysis

Week 1 – A Quick Overview of the Visualisation of (Categorical) Data
Week 2 - The Contingency Table and the Chi-Squared Statistic
Week 3 - Some Common Measures of Association for 2x2 Tables
Week 4 – Some Common Measures of Association for IxJ Tables

Part Two – Graphically Summarising the Association between Categorical Data

Week 5 - Reciprocal Averaging and Methods of Decomposition
Week 6 - Simple Correspondence Analysis
Week 7 - Multiple Correspondence Analysis
Week 8 - Non-Symmetrical Correspondence Analysis

Part Three – Modelling Association between Categorical Data

Week 9 - Models of Association
Week 10 - Log-Linear Models
Week 11 - Ordinal Log-Linear Models
Week 12 - Aggregate Data Analysis

Week 13 - Revision and Questions

The assessment structure will be as follows

Assignment 1:

Measures of Association for Categorical Data Analysis (Weeks 1 – 4)

Due: Week 5 and worth 15% of final marks

Assignment 2:

Graphically Summarising the Association between Categorical Data (Weeks 5 – 8)

Due: Week 9 and worth 15% of final marks

Assignment 3:

Modelling Association between Categorical Data (Weeks 9 – 12)

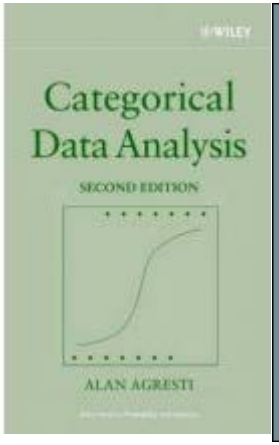
Due: Week 13 and worth 15% of final marks

Final Exam:

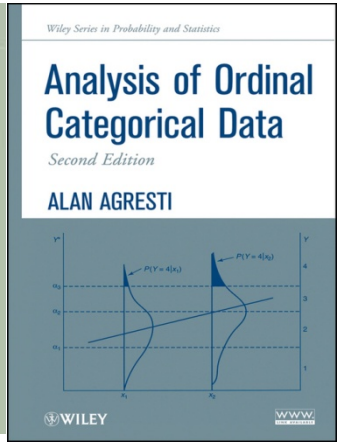
Held late during the UoN exam period and will be worth 55% of the final mark

Key Texts on Various CDA Issues

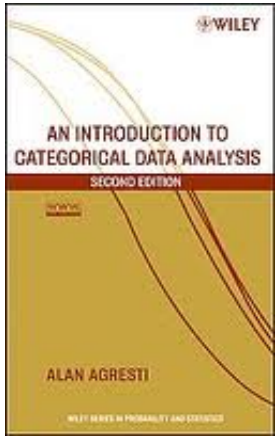
There are many great books out that describe a variety of issues concerned with categorical data analysis



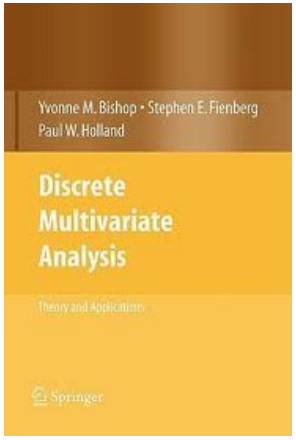
(2002)



(2010)

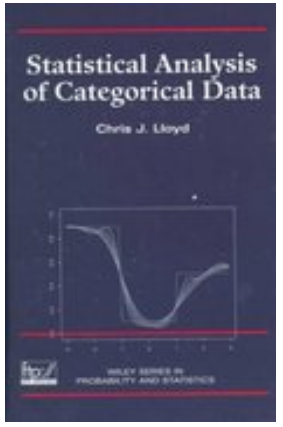


(2007)

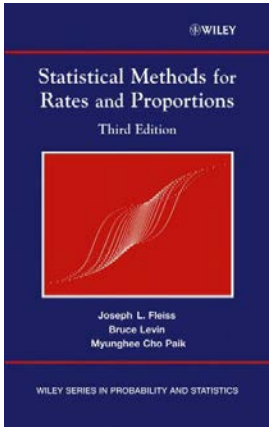


(1975)

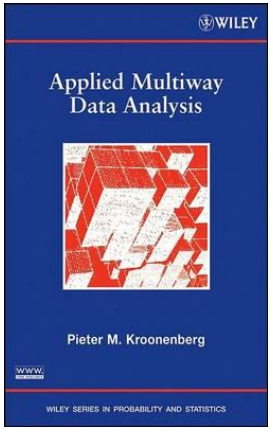
(1992)



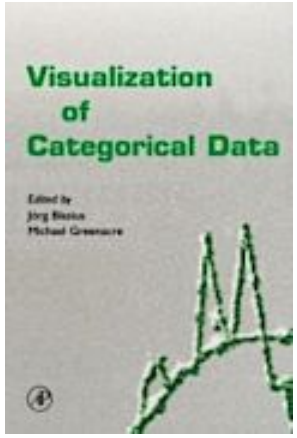
(1999)



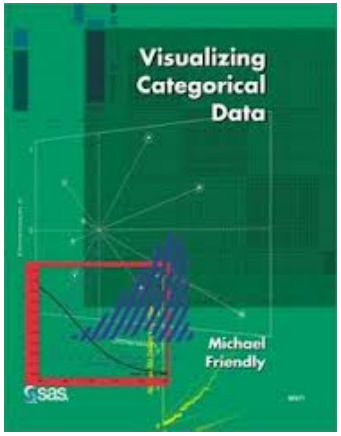
(2003)



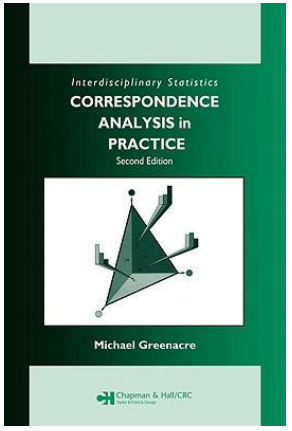
(2008)



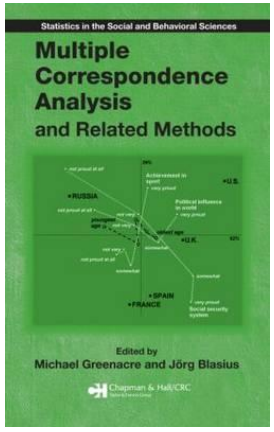
(1998)



(2000)



(2007)



(2006)

There are so many more contributors to the topic that only considered journal papers

For example, one may consider papers by (but not feel confined to)

- Leo Goodman
- Alan Agresti
- Michael Greenacre

In this course we will consider a broad spectrum of popular, and not-so-popular, issues raised in categorical data analysis. The tools by themselves are not enough to give a flavour of their development, so we will also be considering the description of techniques with a historical twist.

We will look first at the graphical display of categorical data but focus much of our attention this semester on the contingency table (two-way and multi-way)

- Week 1 –

The Visualisation of Categorical Data

Professor Eric Beh

School of Mathematical & Physical Sciences
University of Newcastle

Semester 1, 2016

What is Categorical Data?

- *Categorical data arises whenever counts are made instead of measurements*

  Placket (1974, pg vii)

- *A categorical variable has a measurement scale consisting of a set of categories*

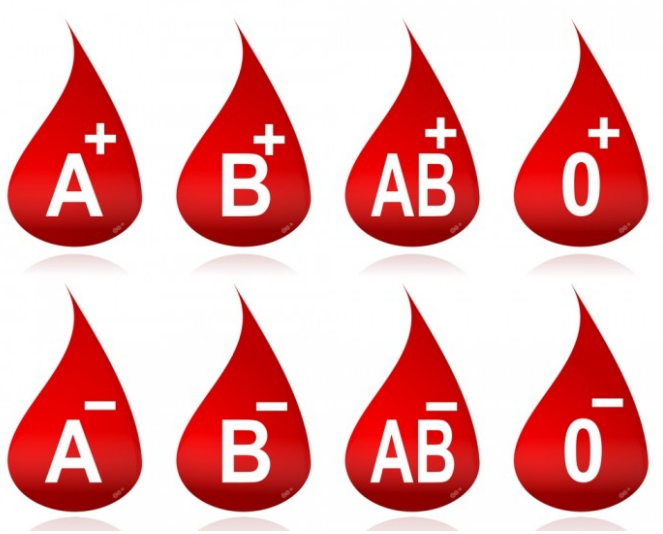
  (Agresti, 3rd ed, pg 1)

What is Categorical Data?

- For much the data you have seen before it has largely consisted of numerical data where distribution theory plays a central role.
- *Numerical data* can be
  - Summarised (eg mean, median, mode)
  - Modelled (Regression, Time series, GLM's)
- While we can graphically and numerically summarise, and model, categorical data we do so in a very different way to how we analyse numerical data
- Before we look at how to analyse categorical data, we shall consider the different forms they can take.

Different Types of Categorical Data

- *Nominal* Categorical Data
  - When the categories have NO natural order to them
  - For example, *Mode of transport to get to uni* has the following unordered categories: car, bus, bicycle, walk, train
  - Blood Type: O, A, B, AB



- *Ordinal Categorical Data*

  - When the categories DO have a natural order
  - There exists a natural "difference" between each successive category (difference not necessarily equal)
  - Patient condition: Good, fair, serious, critical
  - Level of pain: 0 – 10 on a 10 point scale (*Likert scale*)
  - Rating on a movie (on a scale from 1 to 5)

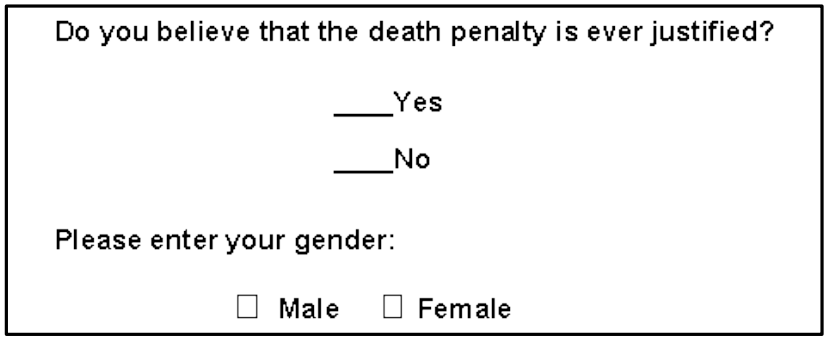
Customer Satisfaction Survey

1. Please rate how strongly you agree or disagree with each of these statements.

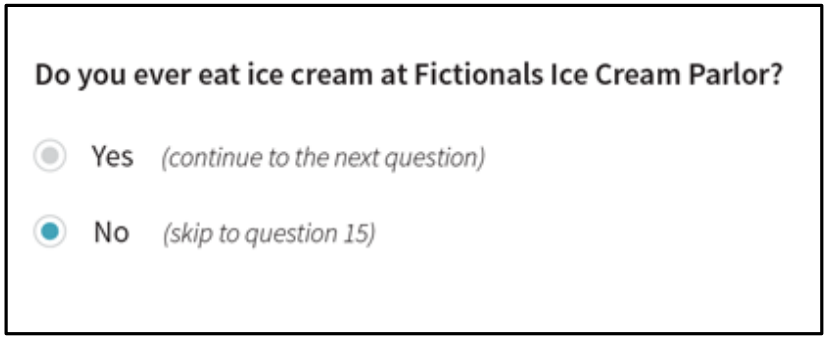
	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	N/A - I haven't used the product.
I enjoy using this product.	○	○	○	○	○	○
This product meets my needs.	○	○	○	○	○	○
I would purchase from this company again.	○	○	○	○	○	○

https://zapier.com/learn/ultimate-guide-to-forms-and-surveys/writing-effective-survey/

- *Dichotomous*, or *Binary*, Variable
  - Where there only two possible categorical responses
  - For example: yes/no, male/female, agree/disagree, smoker/non-smoker, pass/fail, undergrad/postgrad



http://www.socialresearchmethods.net/kb/questype.php



https://infoactive.co/data-design/ch04.html

- *Interval* Data (special type of *Ordinal*).

  – A continuous variable that has been "intervalised" so that we are looking at the counts in each interval category not the raw observation itself

For example . . .

– *Tax brackets*:

Residents

These rates apply to individuals who are Australian residents for tax purposes.

Tax rates 2016–17

The following rates for 2016–17 apply from 1 July 2016.

Taxable income	Tax on this income
0 – $18,200	Nil
$18,201 – $37,000	19c for each $1 over $18,200
$37,001 – $87,000	$3,572 plus 32.5c for each $1 over $37,000
$87,001 – $180,000	$19,822 plus 37c for each $1 over $87,000
$180,001 and over	$54,232 plus 45c for each $1 over $180,000

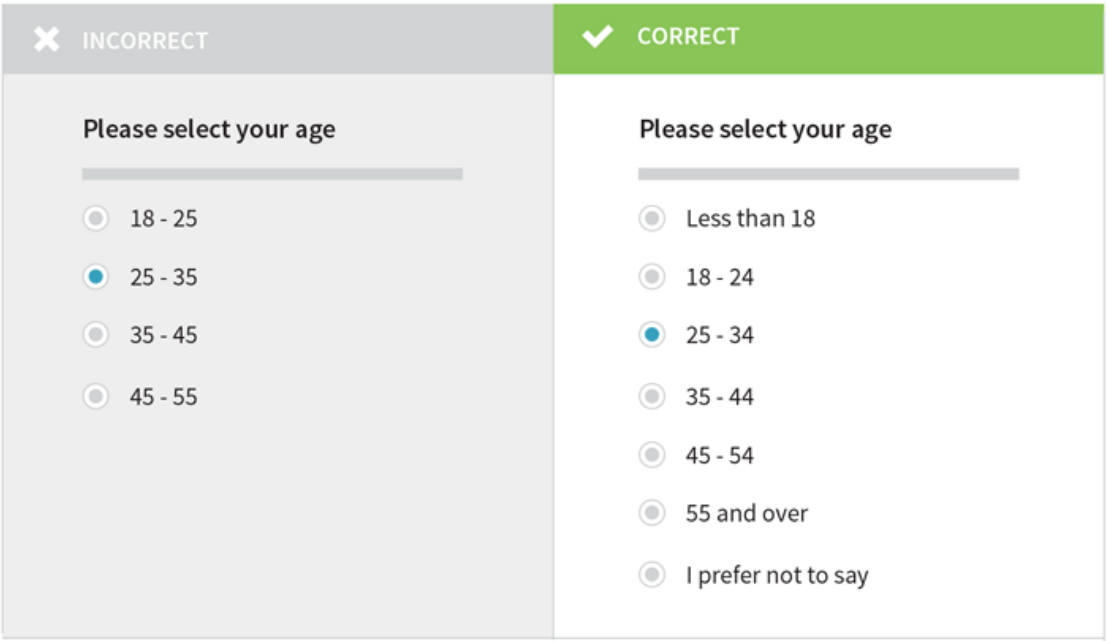
https://www.ato.gov.au/rates/individual-income-tax-rates/

- *Interval* Data (special type of *Ordinal*).

  – A continuous variable that has been "intervalised" so that we are looking at the counts in each interval category not the raw observation itself

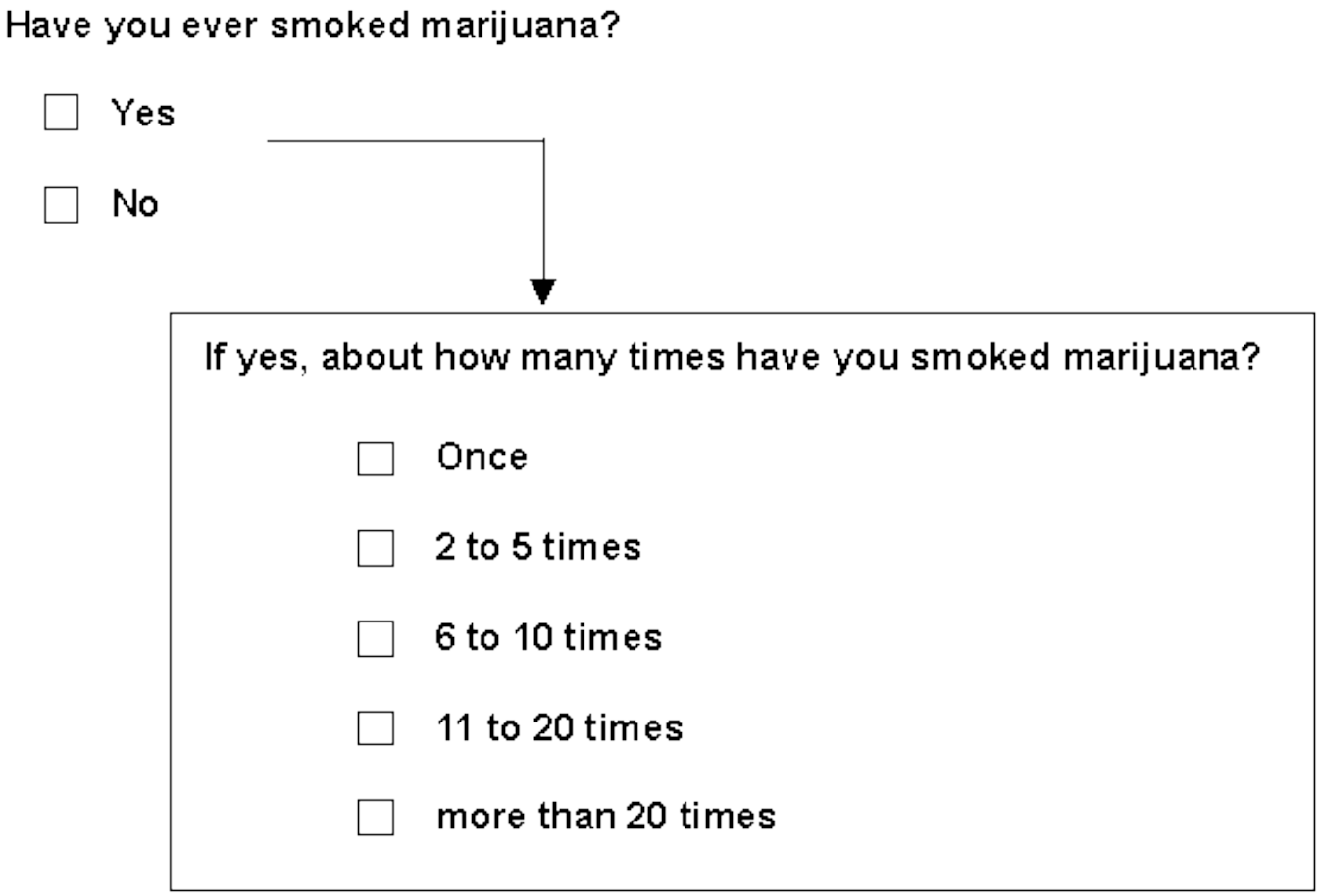
For example . . .

– *Age brackets*:



https://infoactive.co/data-design/ch04.html

- Various combinations of these can arise.



- Various combinations of these can arise.

Did you buy anything in the hotel shop?

☐ Yes ➪ Go to Q19

☐ No ➪ Go to Q21

What did you buy in the shop (Please tick as many boxes as apply)?

☐ Toiletries

☐ Souvenirs

☐ Clothing

☐ Stationery

☐ Other

Please specify ☐

Appendix

Questionnaire on Job Satisfaction

I hereby undertake that the information obtained through this questionnaire will exclusively be used for research purpose and in no case will be disclosed to anybody.

Mohammed Masum Iqbal Researcher

Please Mention the Name of your Bank: _____

Profile of Respondent (Please tick (✓) the right answers of the following)

i. Gender: a. Male b. Female

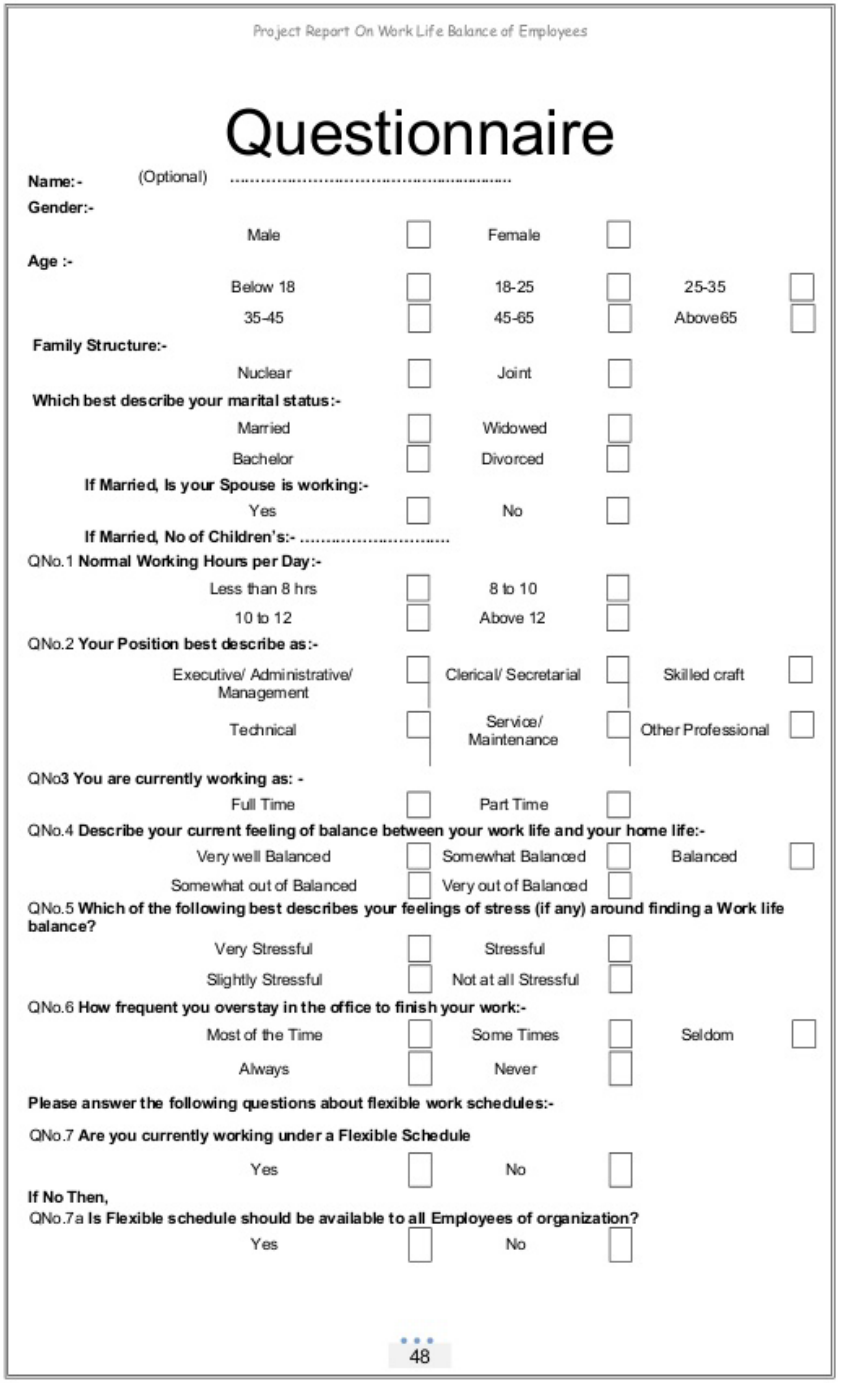
ii. Age: a. Below 30 b. 30 - 40 c. 41 - 50 d. 51 and above

iii. Tenure: a. Less than 5 years b. 5-10 years c. 11-15 years d. 15years and above

Please Tick (✓) the Response that Indicates Your Level of Satisfaction.	Highly Satisfied	Satisfied	Neither Satisfied nor Dissatisfied	Dissatisfied	Highly Dissatisfied
All things considered, how satisfied are you with your job?					

Please Tick (✓) the Responses that Indicate Your Level of Satisfaction.		Highly Satisfied	Satisfied	Neither Satisfied nor Dissatisfied	Dissatisfied	Highly Dissatisfied
	JOB RELATING FACTORS					
1.	Job Security					
2.	Fair Job Design					
3.	Challenging Work					
4.	Status					
5.	Authority and Responsibility					
6.	Autonomy					
7.	Personal Growth					
8.	Balance between Life and Work					
9.	Opportunities to Use Skills					

- Various combinations of these can arise in surveys, questionnaires.
- So how do we go about analysing information that has been collected in this manner?



Project Report On Work Life Balance of Employees

Questionnaire

Name:- (Optional) .....

Gender:-

Male ☐ Female ☐

Age :-

Below 18 ☐ 18-25 ☐ 25-35 ☐

35-45 ☐ 45-65 ☐ Above65 ☐

Family Structure:-

Nuclear ☐ Joint ☐

Which best describe your marital status:-

Married ☐ Widowed ☐

Bachelor ☐ Divorced ☐

If Married, Is your Spouse is working:-

Yes ☐ No ☐

If Married, No of Children's:- .....

QNo.1 **Normal Working Hours per Day:-**

Less than 8 hrs ☐ 8 to 10 ☐

10 to 12 ☐ Above 12 ☐

QNo.2 **Your Position best describe as:-**

Executive/ Administrative/ Management ☐ Clerical/ Secretarial ☐ Skilled craft ☐

Technical ☐ Service/ Maintenance ☐ Other Professional ☐

QNo3 **You are currently working as: -**

Full Time ☐ Part Time ☐

QNo.4 **Describe your current feeling of balance between your work life and your home life:-**

Very well Balanced ☐ Somewhat Balanced ☐ Balanced ☐

Somewhat out of Balanced ☐ Very out of Balanced ☐

QNo.5 **Which of the following best describes your feelings of stress (if any) around finding a Work life balance?**

Very Stressful ☐ Stressful ☐

Slightly Stressful ☐ Not at all Stressful ☐

QNo.6 **How frequent you overstay in the office to finish your work:-**

Most of the Time ☐ Some Times ☐ Seldom ☐

Always ☐ Never ☐

Please answer the following questions about flexible work schedules:-

QNo.7 **Are you currently working under a Flexible Schedule**

Yes ☐ No ☐

If No Then,

QNo.7a **Is Flexible schedule should be available to all Employees of organization?**

Yes ☐ No ☐

48

Some Simple Visualisations of Categorical Data

One of the first things we like to do in order to get a picture of what any form of data looks like is to visualise it. From such visual summaries we can start to form hypotheses, and generally get a "feel" for what the data is trying to tell us.

For categorical data there are numerous visualisation tools we can use

Single Categorical Variable

- Pie Chart
- Pareto Chart
- Bar Chart

Two variables – one continuous variable, one categorical variable

- Boxplot

Two variables – two categorical variables

- Mosaic plot
- 3FCD
- Correspondence plot

So, in no particular order, lets look at the development of some of the key data visualisation tools

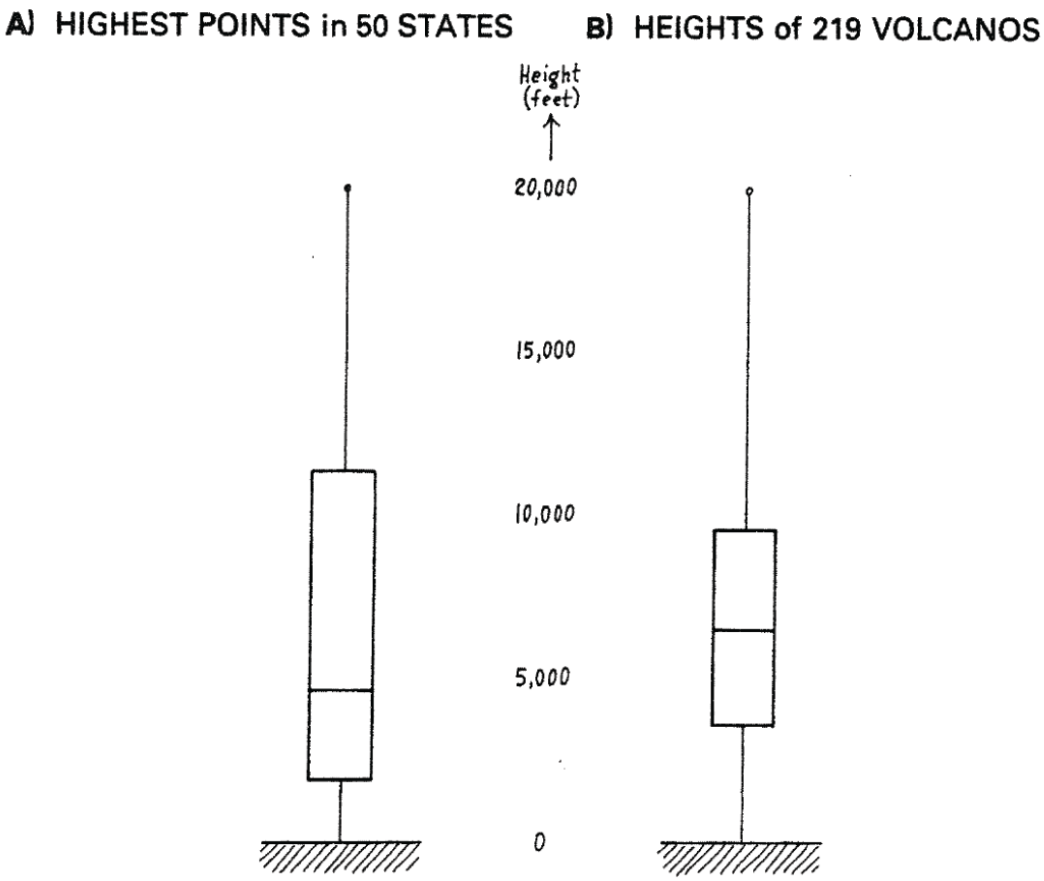
The Boxplot

The boxplot can be used to visually summarise numerical data. When such data is collected across different regions, areas, species or categories we can construct a side-by-side boxplot.

The boxplot on the left is the highest point (in feet) of each of the 50 states of the USA. The boxplot on the right looks at the height of 219 volcanos around the world.

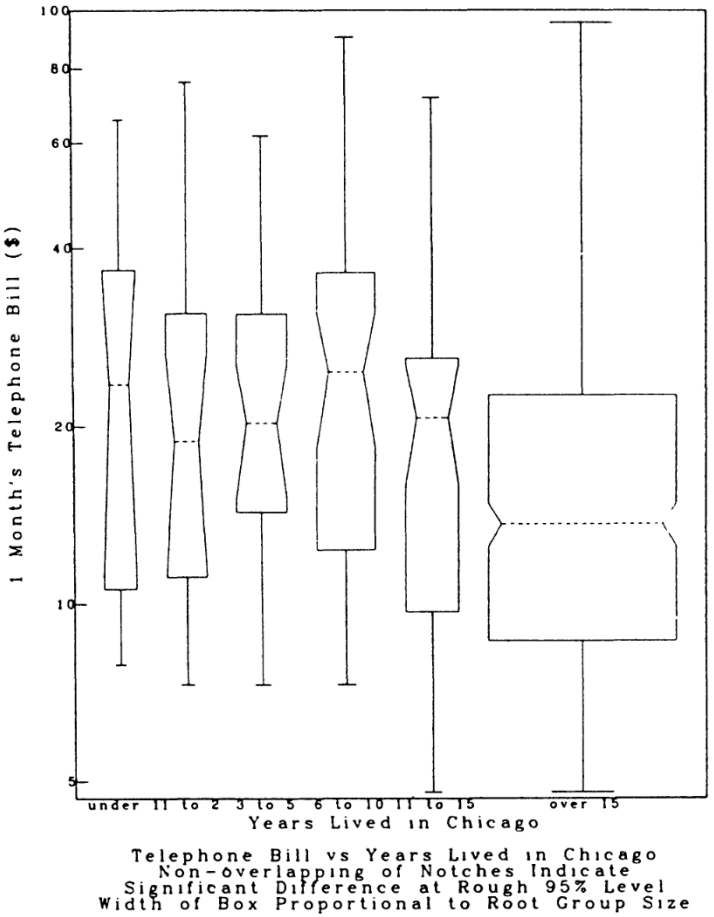
```
> boxplot(x)
```

John W. Tukey
(1915 – 2000)



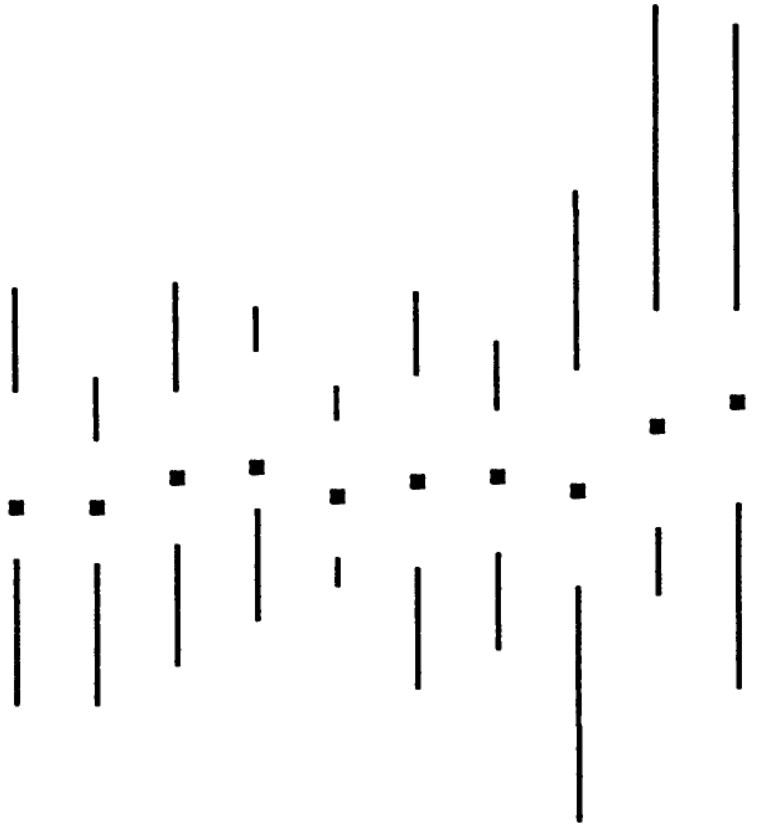
This is the first boxplot published and was by John Tukey (1977, pg 40).

Notched Boxplot

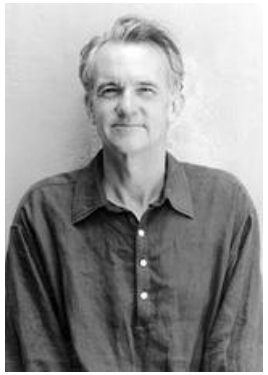


McGill, R., Tukey, J.W. and Larsen, W.A. (1978), Variations of Boxplots, *The American Statistician*, 32, 12-16

Quartile Plot



Tufte, E. R.(2001), *The Visual Display of Quantitative Information*, Graphics Press.



```
> boxplot(x, notched = TRUE)
```

Vaseplot

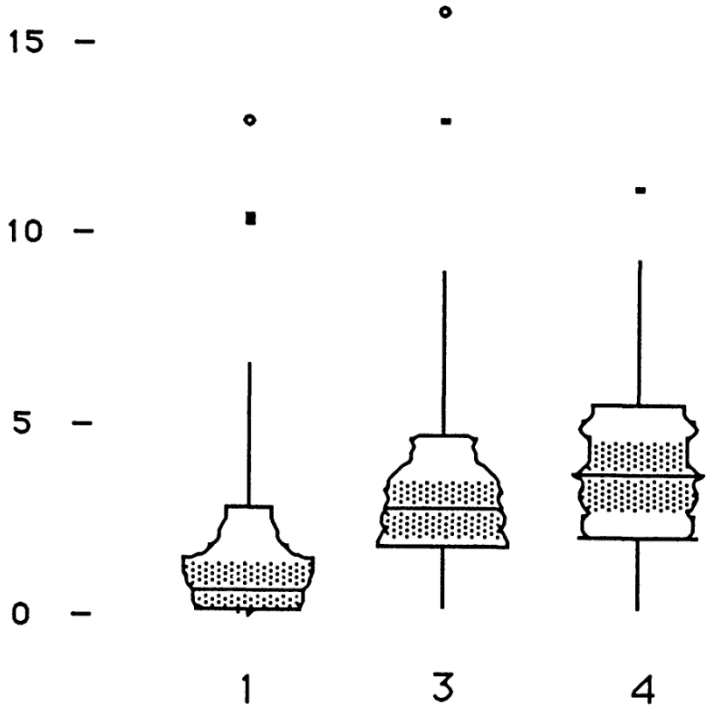


Figure 5. Vaseplots of Three Samples of Size 40 From Linearly Transformed Chi-Squared Distributions With 1, 3, and 4 Degrees of Freedom. The gray bars are confidence intervals for the medians.

Benjamini, Y. (1988), Opening of the box of a boxplot, *The American Statistician*, 42, 257-262

Violine plot

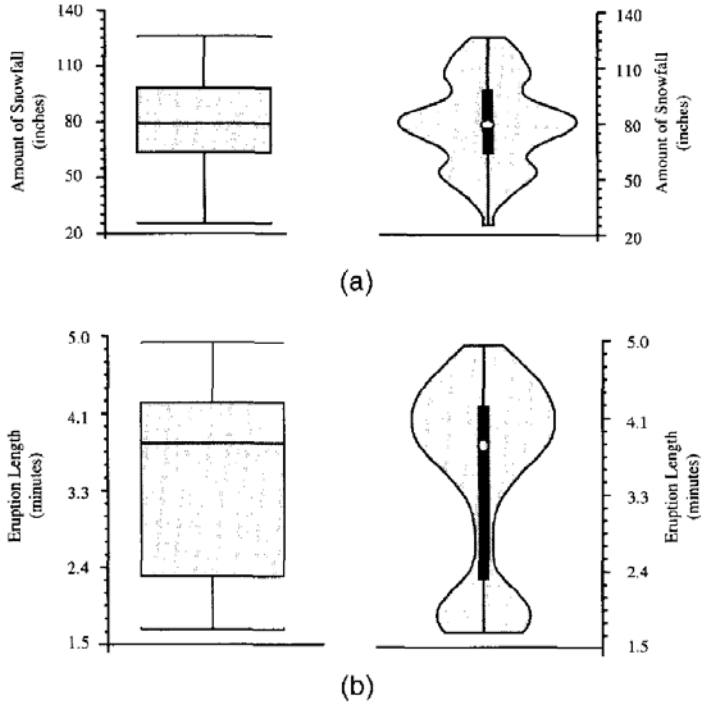


Figure 3. Additional Information in Violin Plots. Two examples from the density estimation literature: (a) annual snowfall for Buffalo, NY, 1910—1972; (b) Old Faithful eruption length.

Hintze, J.L. and Nelson, R. D. (1998), Violin plots: A box plot-density trace synergism, *The American Statistician*, 52, 181-184

However . . . .

The Range Bar Chart

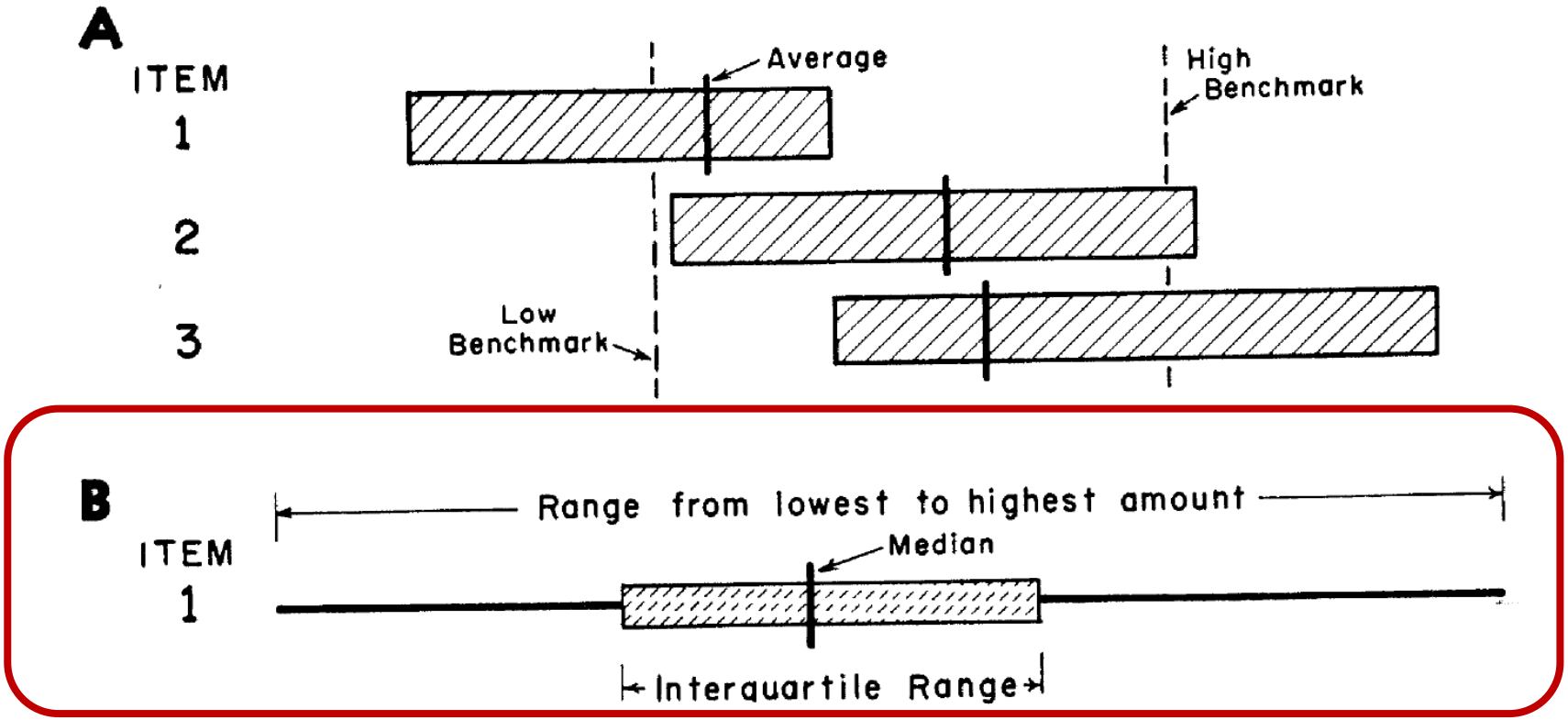
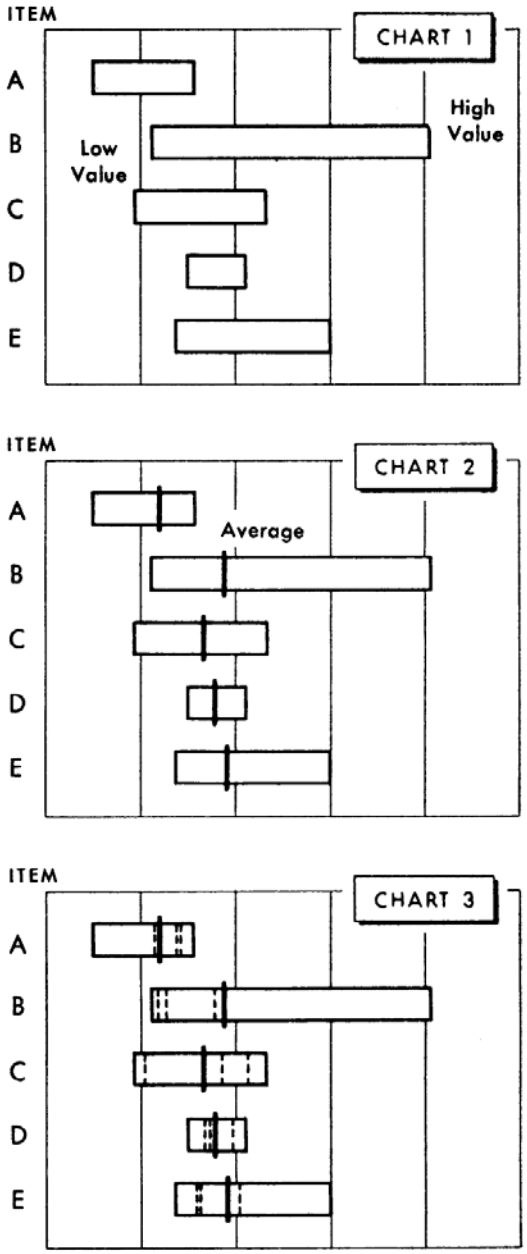
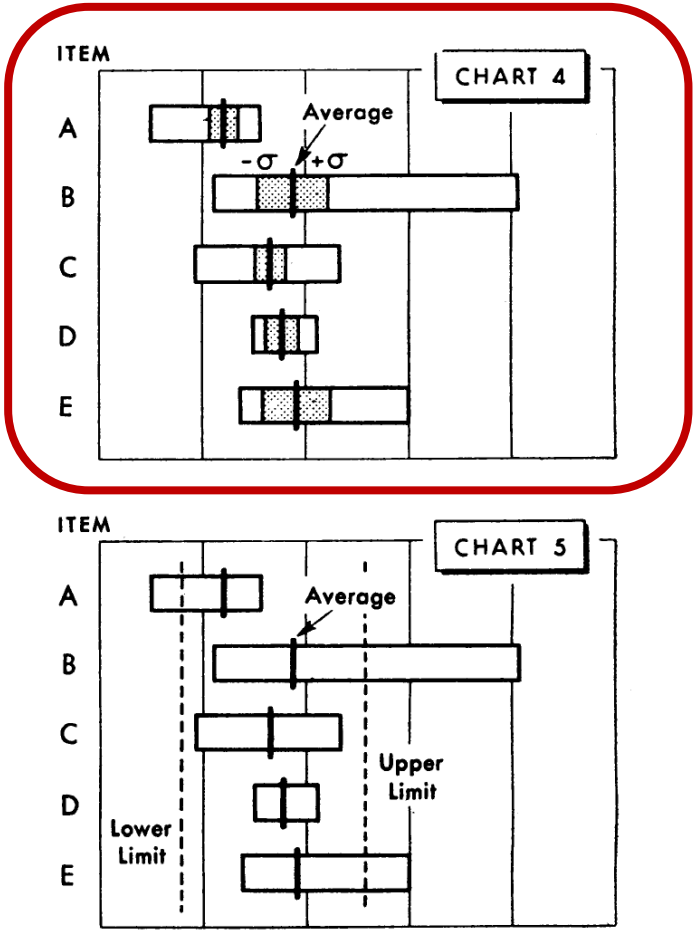


Fig. 6-24. Various uses of the range bar.

The range-bar chart can also be used to determine a bench mark for judging performance (Fig. 6-24, item *A*). Interquartile range can be indicated clearly on this type of chart (item *B*).

Spear, Mary Eleanor (1952), *Charting Statistics*, McGraw-Hill. pp. 164-166





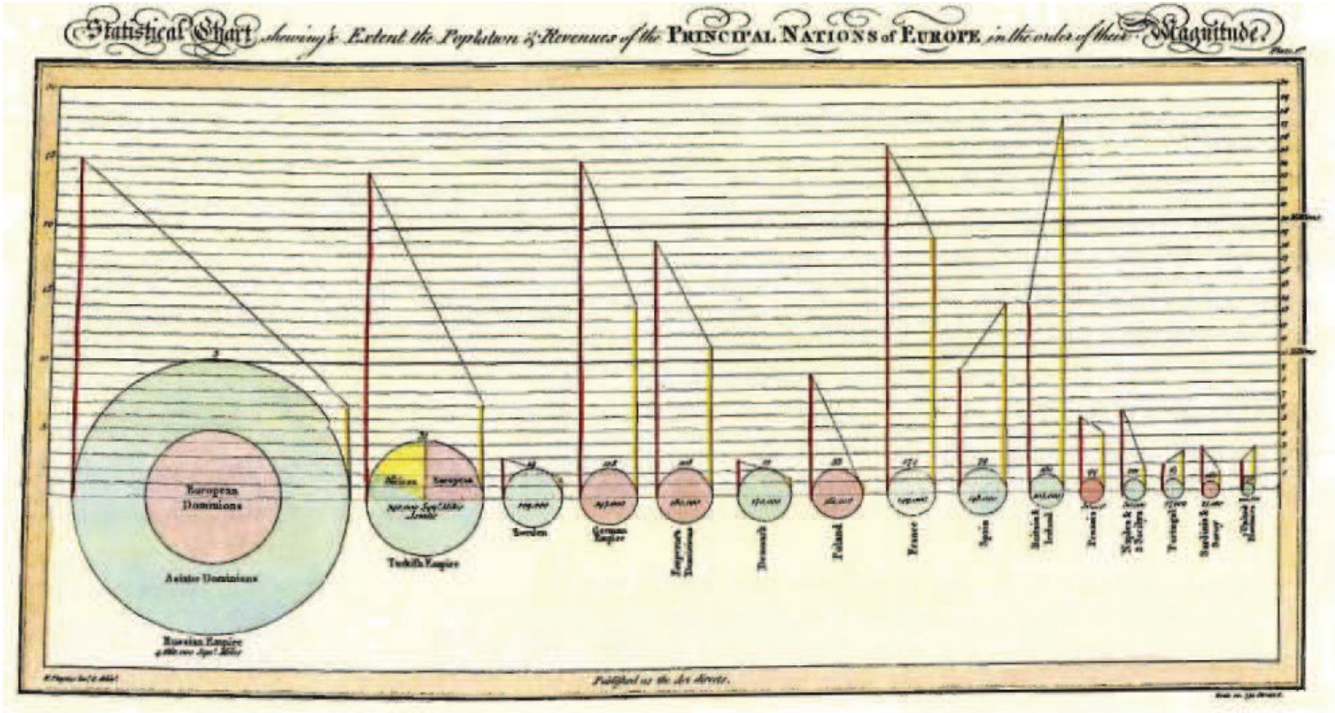
Here the "box" consists of the
- mean (not the median) and
- 1 SD either side of the mean, not the IQR

The smallest and largest value still depicted.

Haemer, K.W. (1948), Range-bar charts, *The American Statistician*, 2, 23.

The Pie Chart

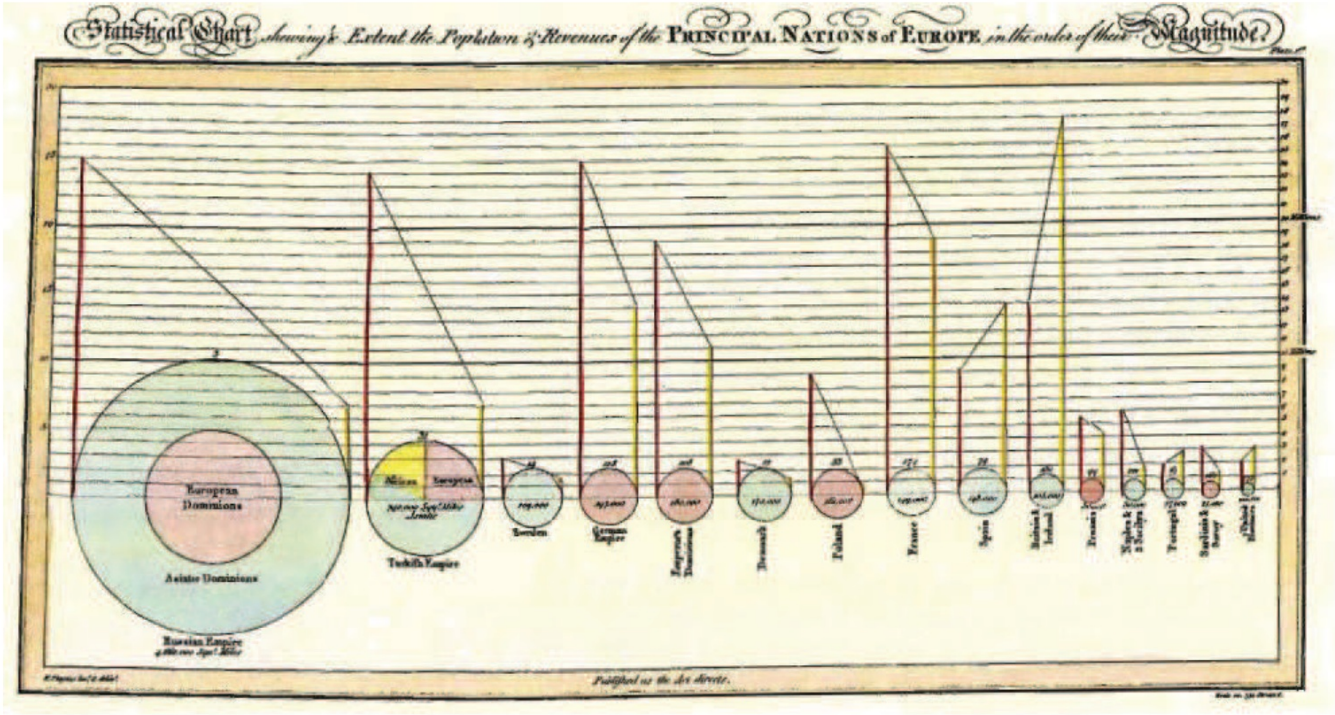
The story of the *pie chart begins with William Playfair (1759 1823).* After proposing the *line graph and barchart in 1786, Playfair (in 1801) constructed the pie chart* as a visual aid to compare the geographical size of each of European regions, and the areas around the world they occupied.



The area of the circles allows for a comparison of the land area (in square miles) each region occupied, given the geopolitical situation in Europe in 1801.

- The green coloured regions indicate those countries that were adjudged a maritime power,
- The red colour regions reflected those countries with no maritime power.

Playfair also incorporated **vertical lines** to compare the population of each region (by the red line) while the green line reflects the **tax revenues**;



Friendly (2008, pg 24) said that such a joint comparison is

"today considered a sin in statistical graphics (you can easily jiggle either scale to show different things)".

Interestingly, Playfair's contribution to data visualisation in Great Britain remained relatively unknown. Spence (2005) discusses that this may be because Playfair was involved in

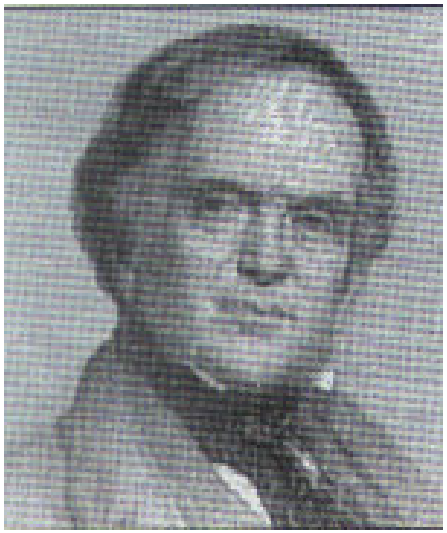
"failed and sometimes fraudulent business ventures in London and Paris since the early 1780's".

Friendly and Denis (2005, pg 106) add that

"Playfair was indeed a sinner, graphic and otherwise".

Tufte (2001, pg 178) said of the pie chart

"A table is nearly always better than a dumb pie chart; the only worse design that a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies. . ."



William Playfair
(1759 - 1823)

Debates on the usefulness of a pie chart began early. See, for example, Eells (1926).

Despite the reputation of his pie charts, Playfair is honoured in various ways in his hometown of Edinburgh, Scotland . . . .

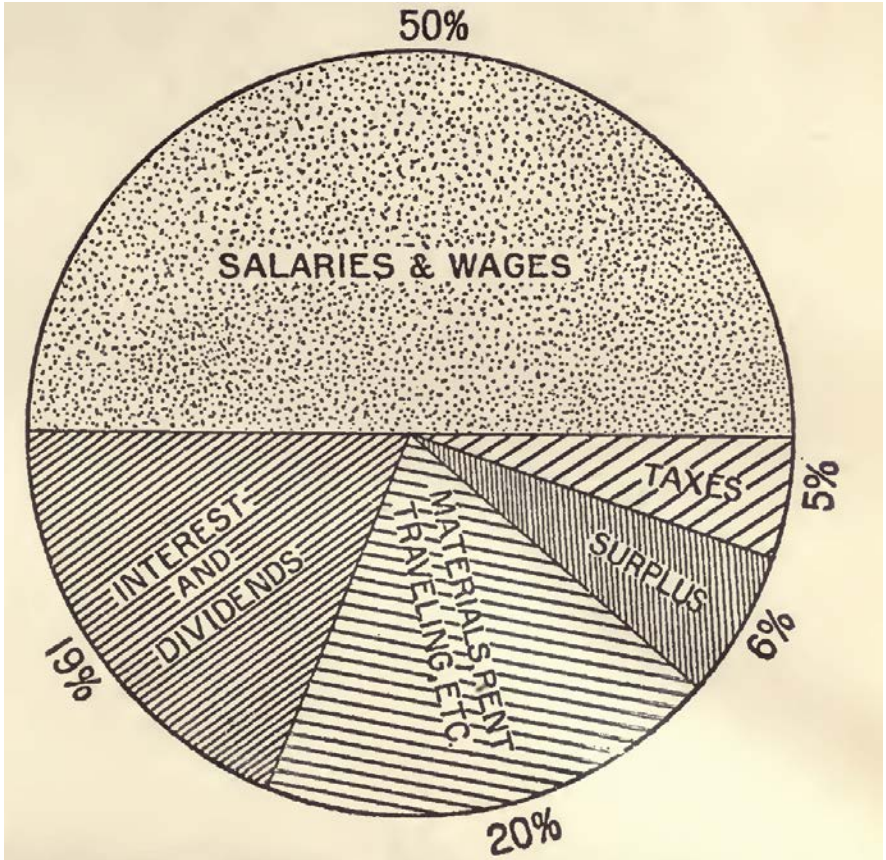


Fig. 2. Disposition of the Gross Revenue of the Bell Telephone System for the Year 1911

This chart was taken from the annual report to the stockholders of the American Telephone and Telegraph Company for the year ending December 31, 1911

Brinton (1914, pg 5)

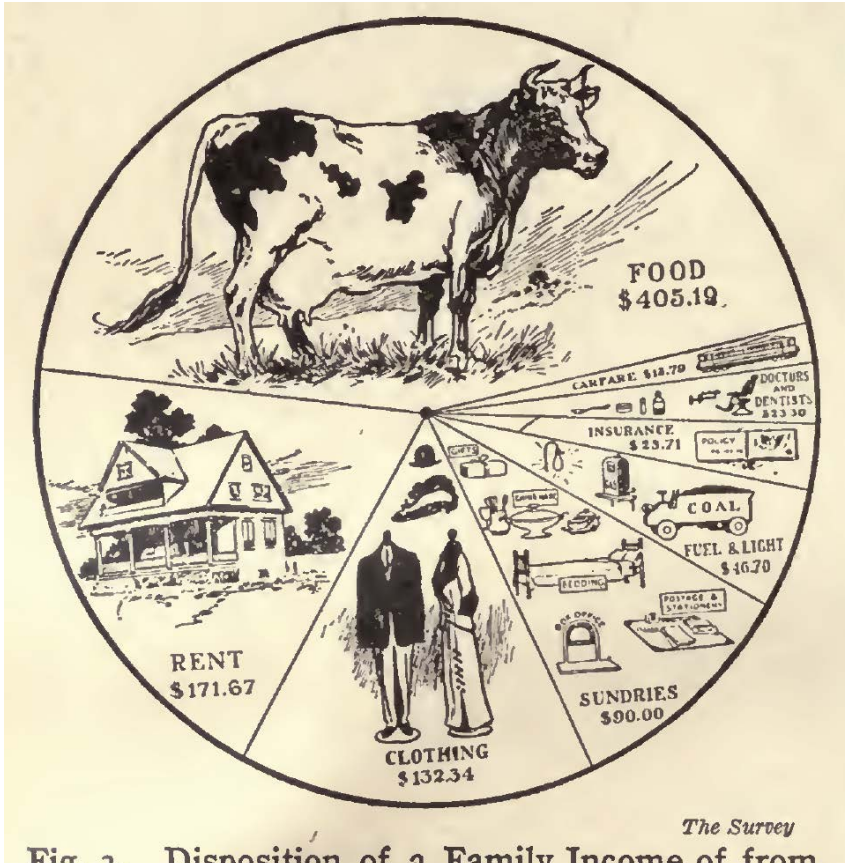


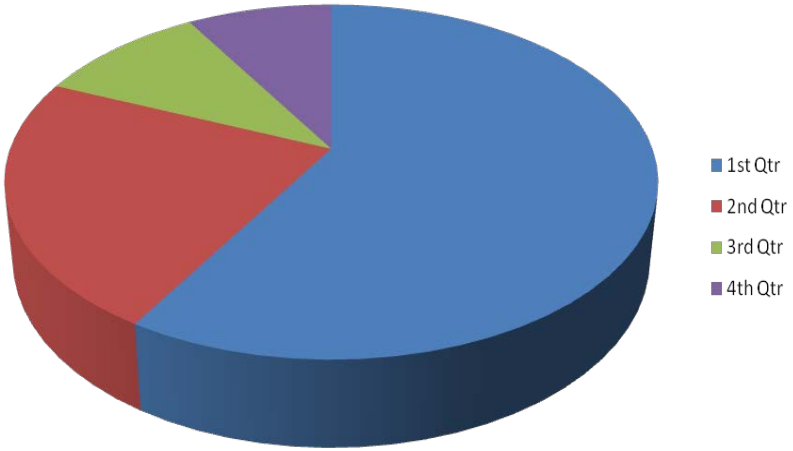
Fig. 3. Disposition of a Family Income of from $900 to $1000

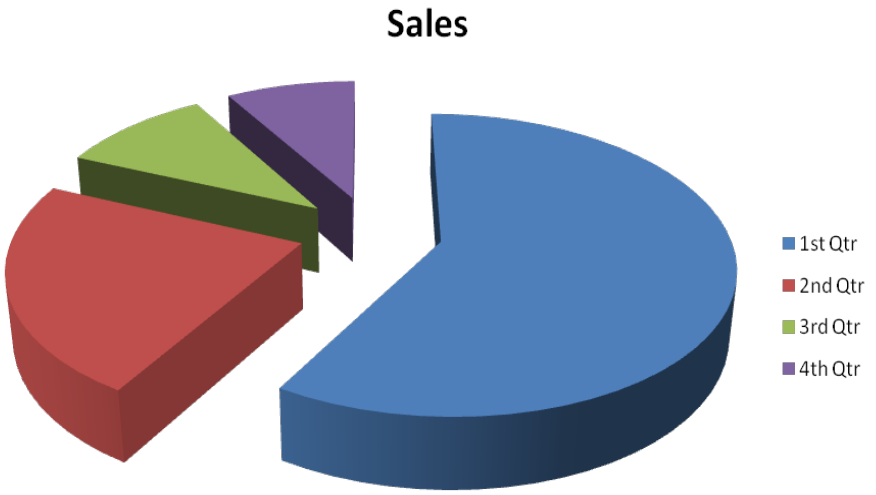
This cut shows an attempt to put figures in popular form. The eye is likely to judge by the size of the pictures rather than by the angles of the sectors

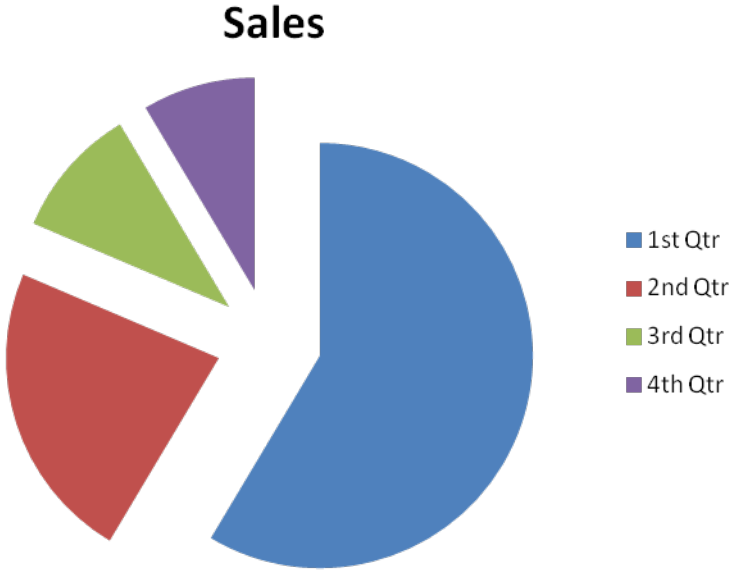
Brinton (1914, pg 6)

Haskell (1922, pg 9) calls it a version of the pie chart a *circular percentage chart*. Brinton (1939, pg 81) notes that alternative names of the pie chart include the *sector chart* and *divided circle*. He dedicates a whole chapter (Chp 9) to their study.

"A table is nearly always better than a dumb pie chart"

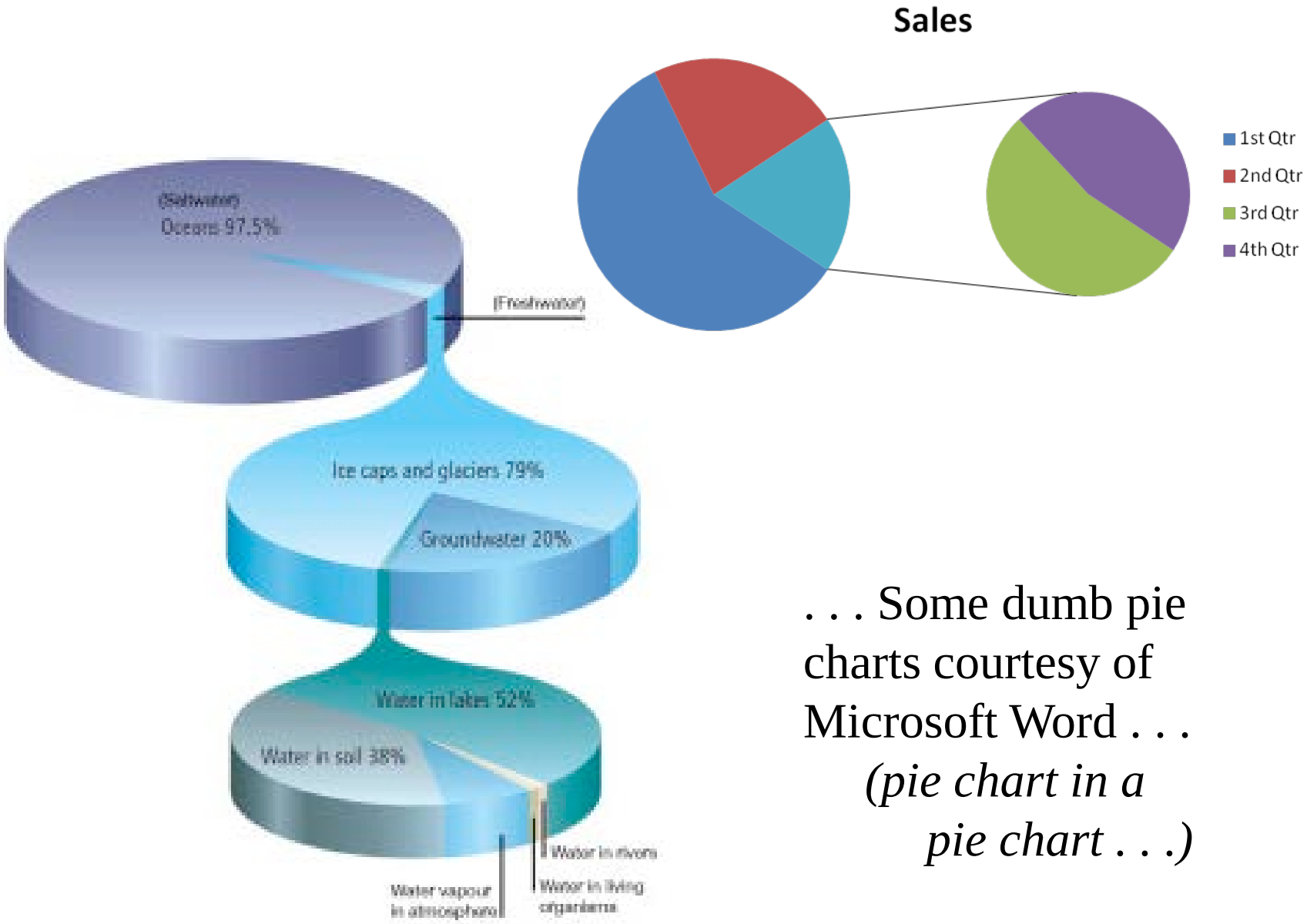






. . . Some dumb pie charts courtesy of Microsoft Word . . .

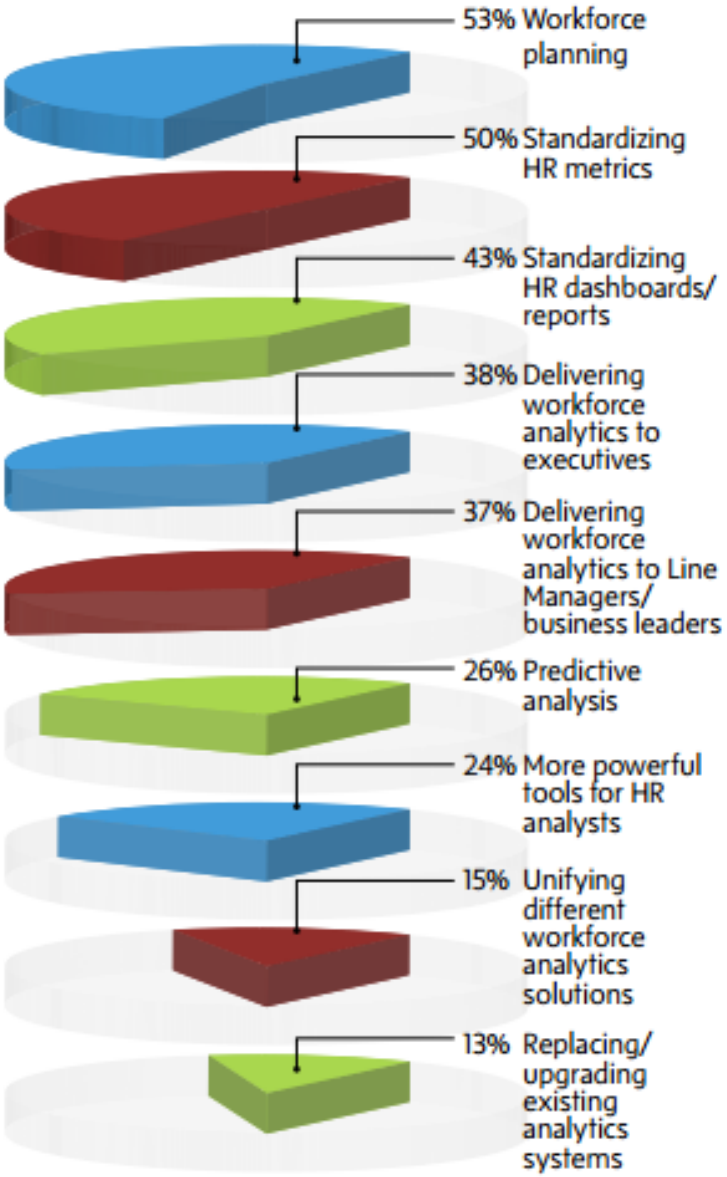
"A table is nearly always better than a dumb pie chart"



. . . Some dumb pie charts courtesy of Microsoft Word . . .

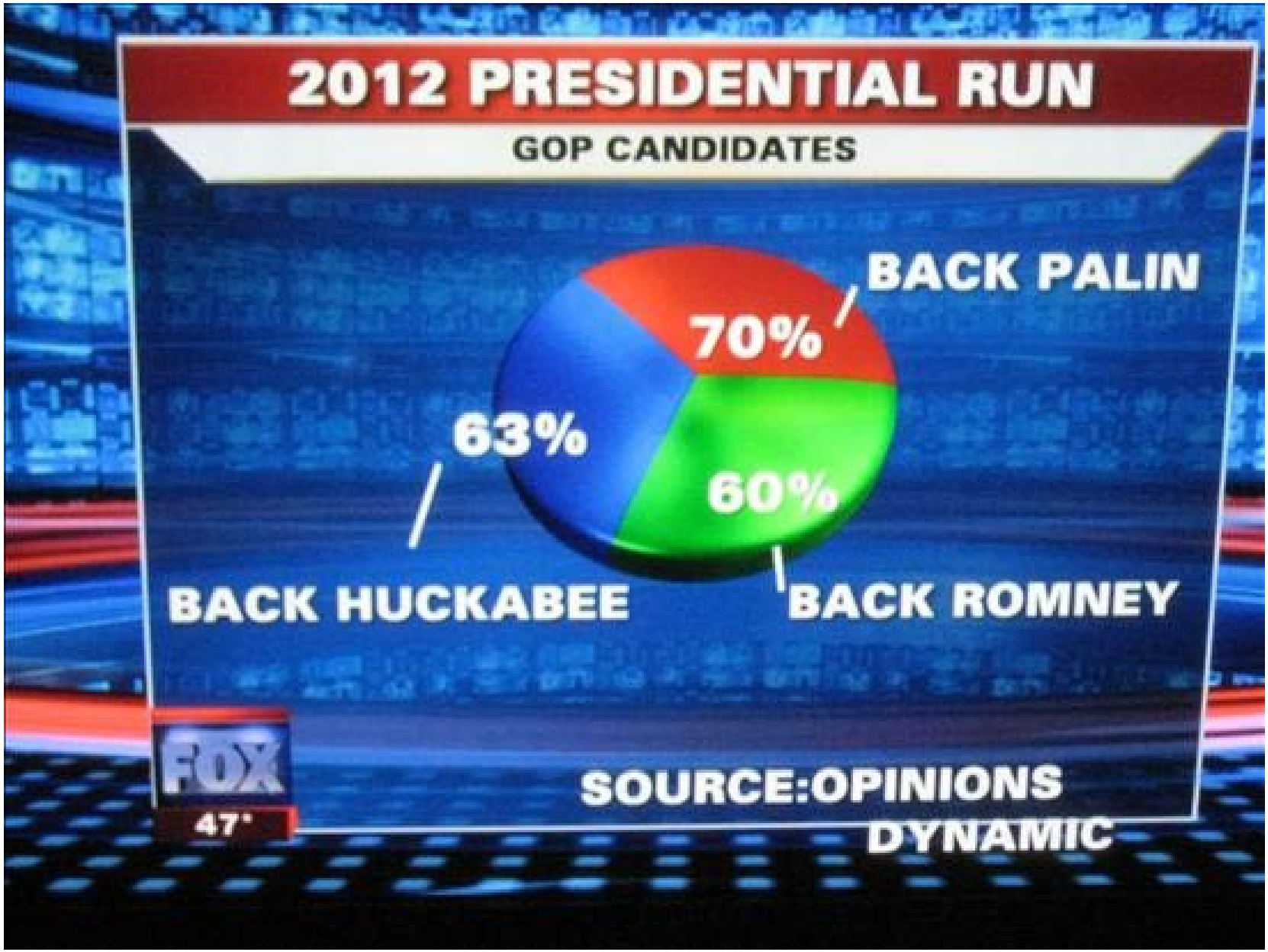
(pie chart in a pie chart . . .)

“A table is nearly always better than a dumb pie chart”

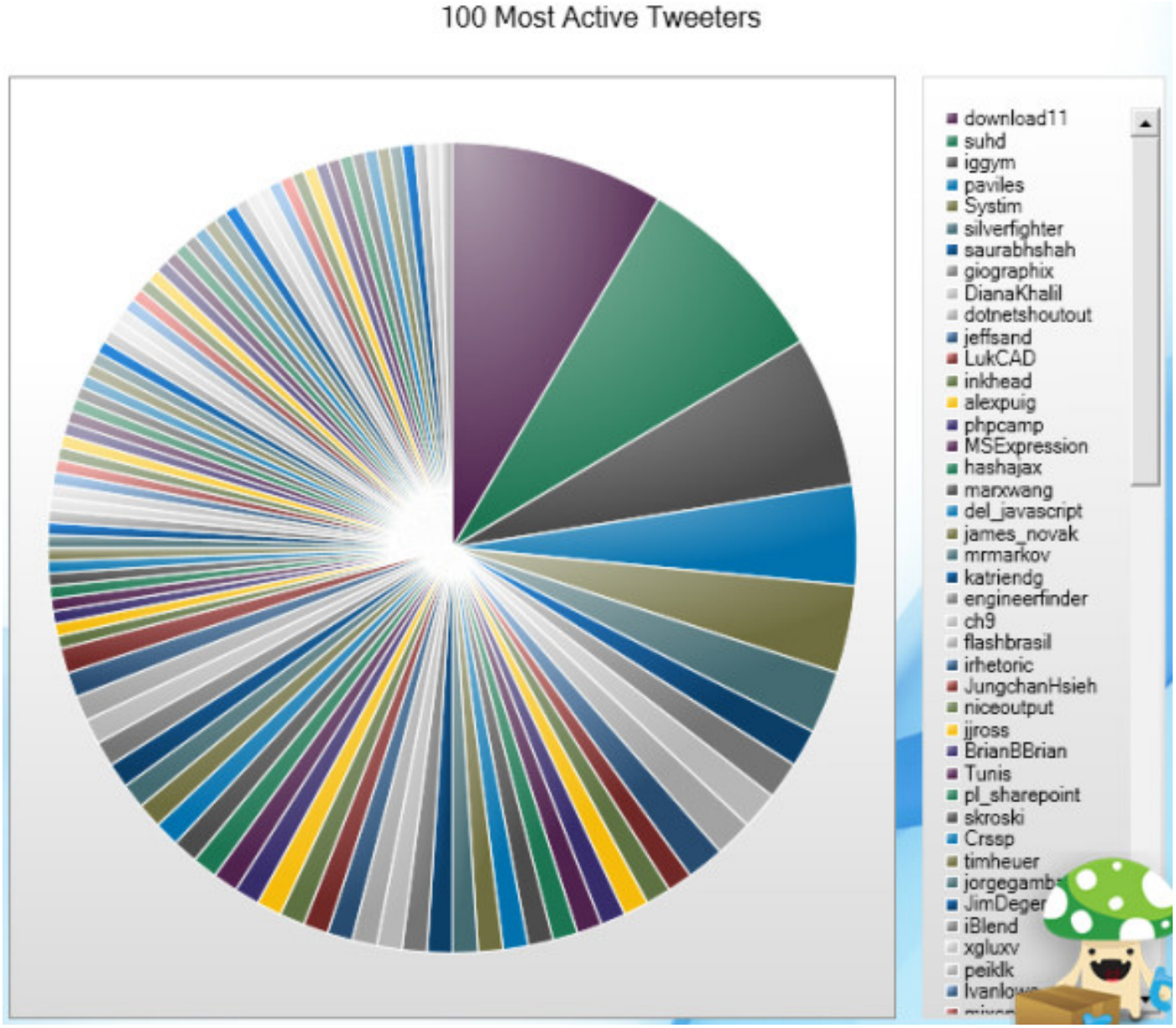


Exact source unknown . . . But from someone who describes himself as "an innovation leader in delivering analytics."

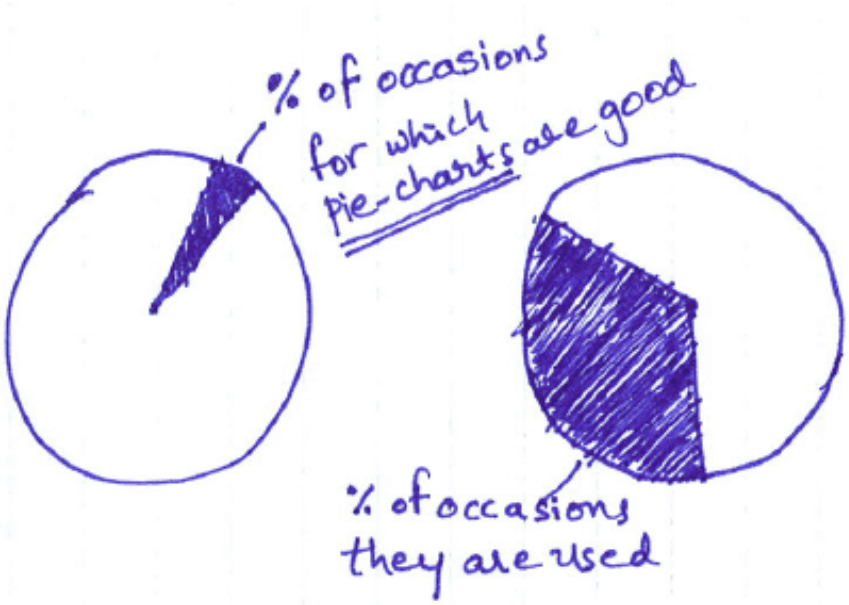
Some are just inaccurate . . .

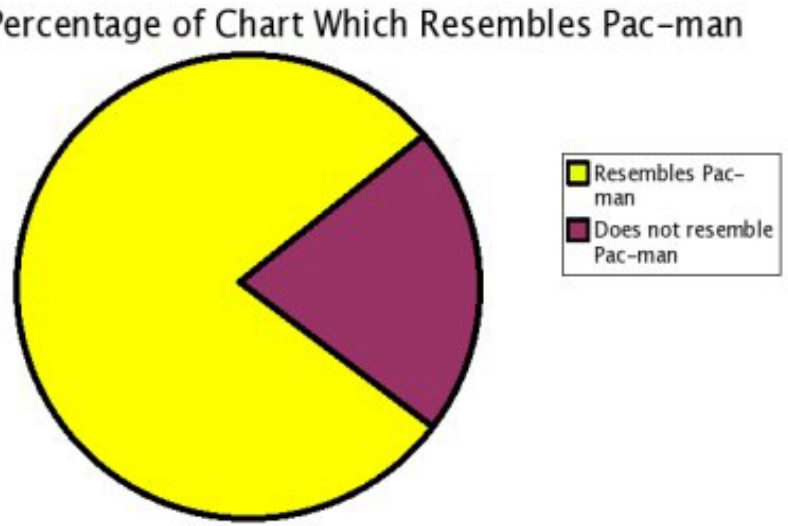


Some are just not informative . . .

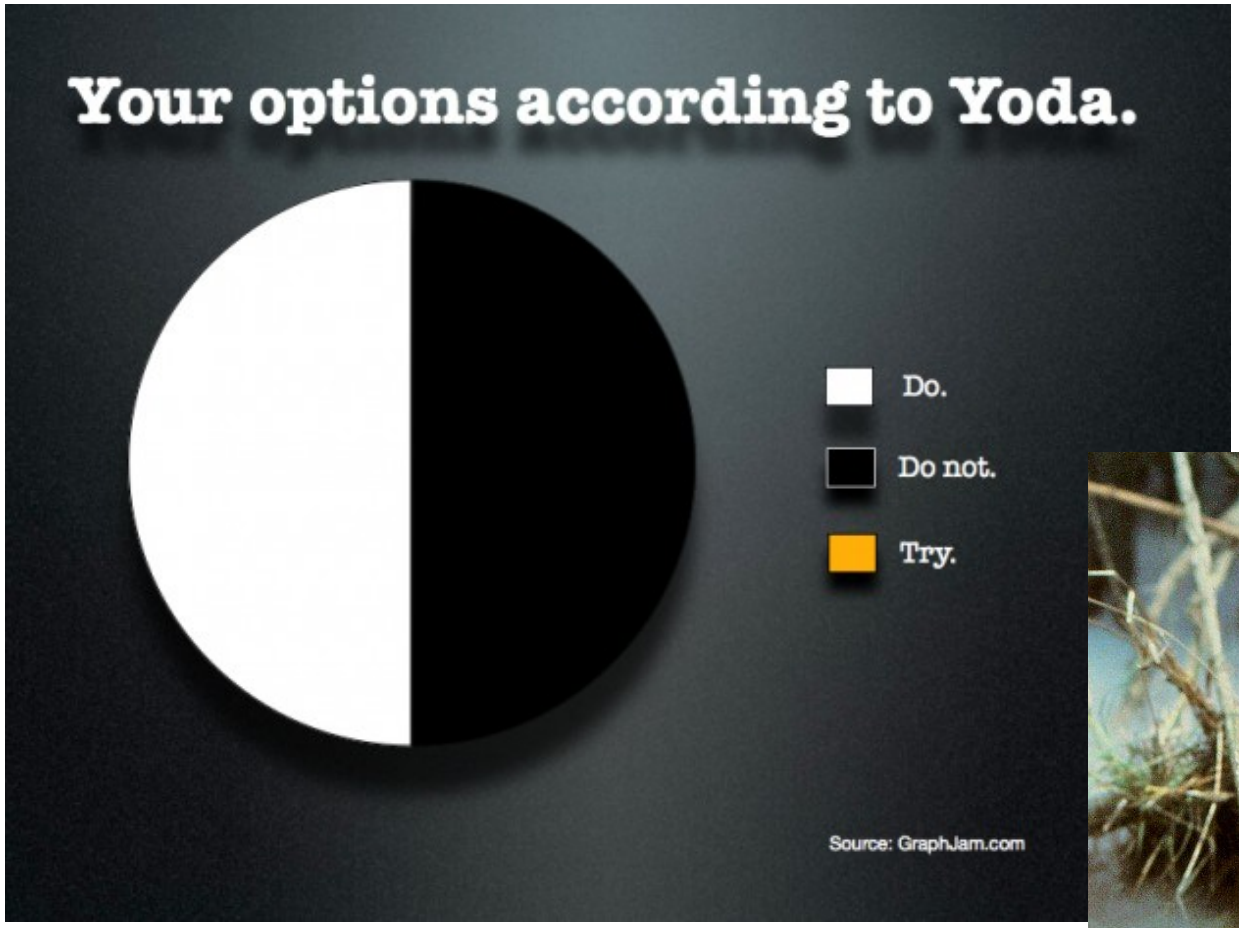


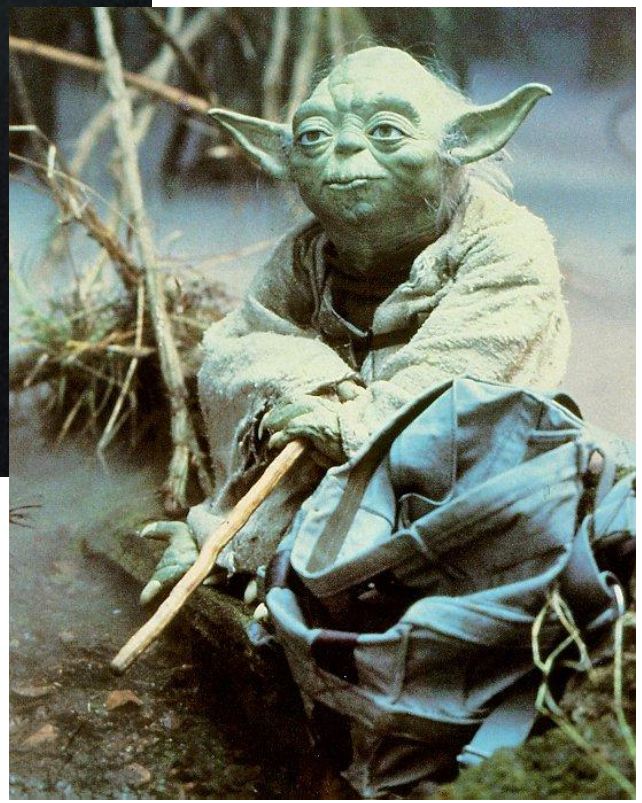
Some are just more fun than serious . . .





Some are just more fun than serious . . .





The Impact of Florence Nightingale

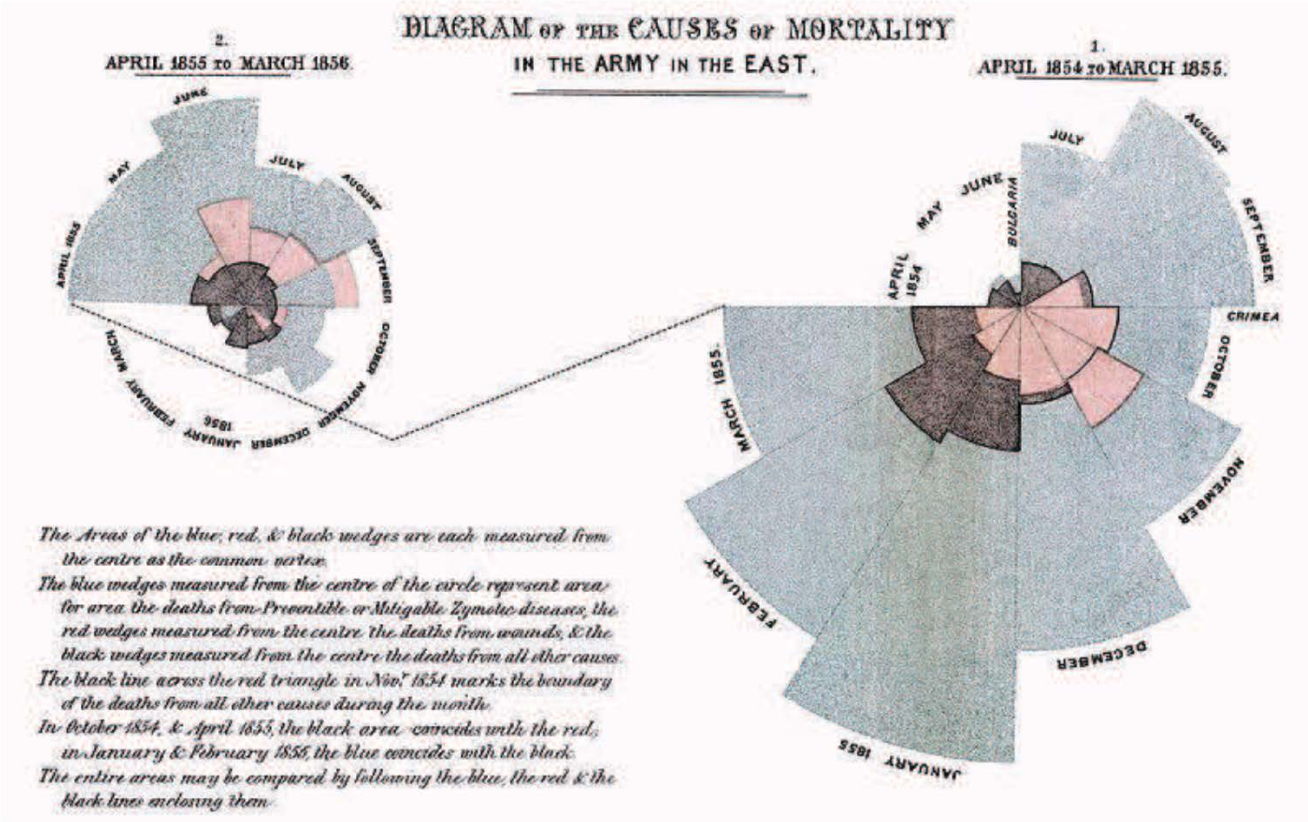
An early variation of the pie chart was called the *coxplot* – see the diagram on the next page. It was proposed by Famous English nurse Florence Nightingale who was also a highly accomplished statistician

Nightingale became enraged at the deplorable sanitary conditions of army hospitals during the Crimean War* (1853 – 1856) and fought for reforms to improve their conditions. Her literary and mathematical skills were instrumental in these reforms, and many of her arguments were highly statistical.

* The Crimean War was fought along the Crimean peninsula that is located on the northern coast of the Black Sea (now part of modern day Ukraine). It was between the Russian Empire and the alliance formed between the French, British, Ottoman Empires and Sardinia.

Florence Nightingale
(1820 – 1910)

In 1859 Nightingale was elected the first female member of the Royal Statistical Society and she later became an honorary member of the American Statistical Association



Florence Nightingale
(1820 – 1910)

This coxplot appears in Nightingales 1858

Notes on matters affecting the Health, Efficiency and Hospital Administration of the British Army

It highlights from April 1855 to March 1856 those **death rates** as a result of battle wounds, preventable diseases and deaths due to other causes. The death rates are reflected by the area of each bar of the *coxplot* not its length.

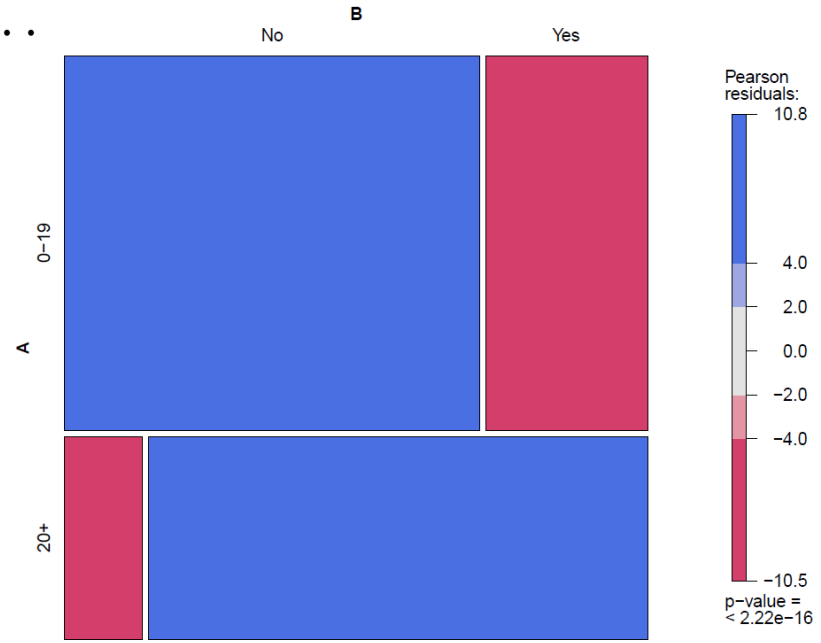
The Mosaic Display

Michael Friendly
(1945 -)

. . . is a fairly popular way of visualising the association between two categorical variables. It was extensively developed (although not originated) by Michael Friendly.

The mosaic display is similar to a grouped bar chart, where the widths of the bars show the relative frequencies of one variable, and the heights of the sections in each bar show the relative frequencies of the second variable. It is applicable for any two-way contingency table. For our 2x2 contingency table of asbestos data, the mosaic display is given by . . .

Occupational Exposure (yrs)	Asbestosis No	Yes	Total
0-19	522	203	725
20+	53	339	392
Total	575	542	1117



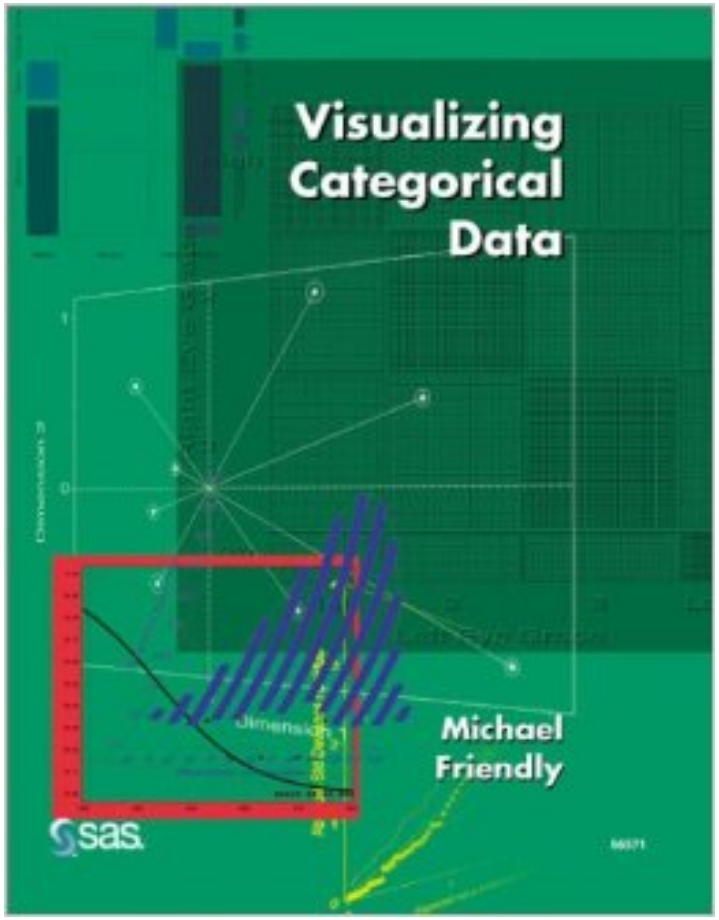
Conditional on *Occupational Exposure*

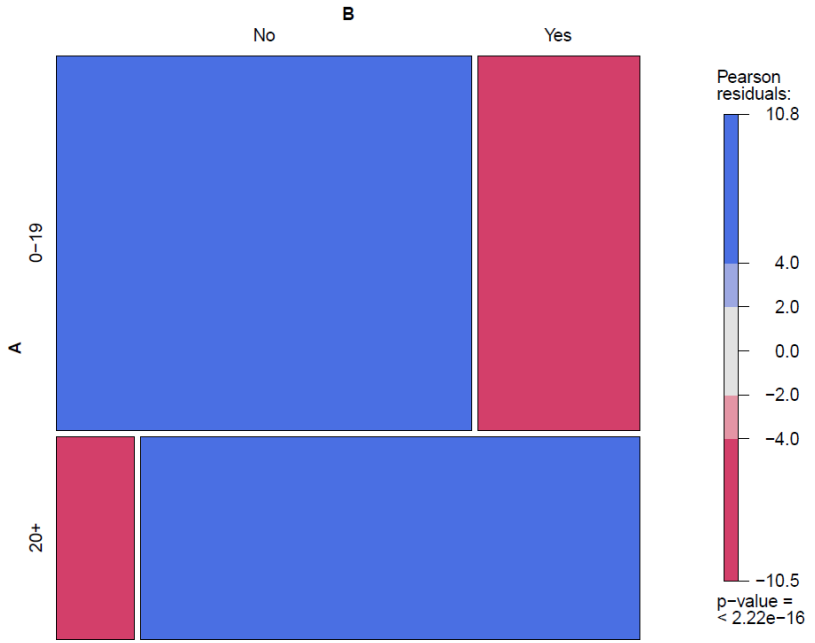
In R, mosaic displays may be constructed using the `mosaic` function in the `vcd` library.

The Mosaic Display

. . . is a fairly popular way of visualising the association between two categorical variables. It was extensively developed (although not originated) by Michael Friendly.

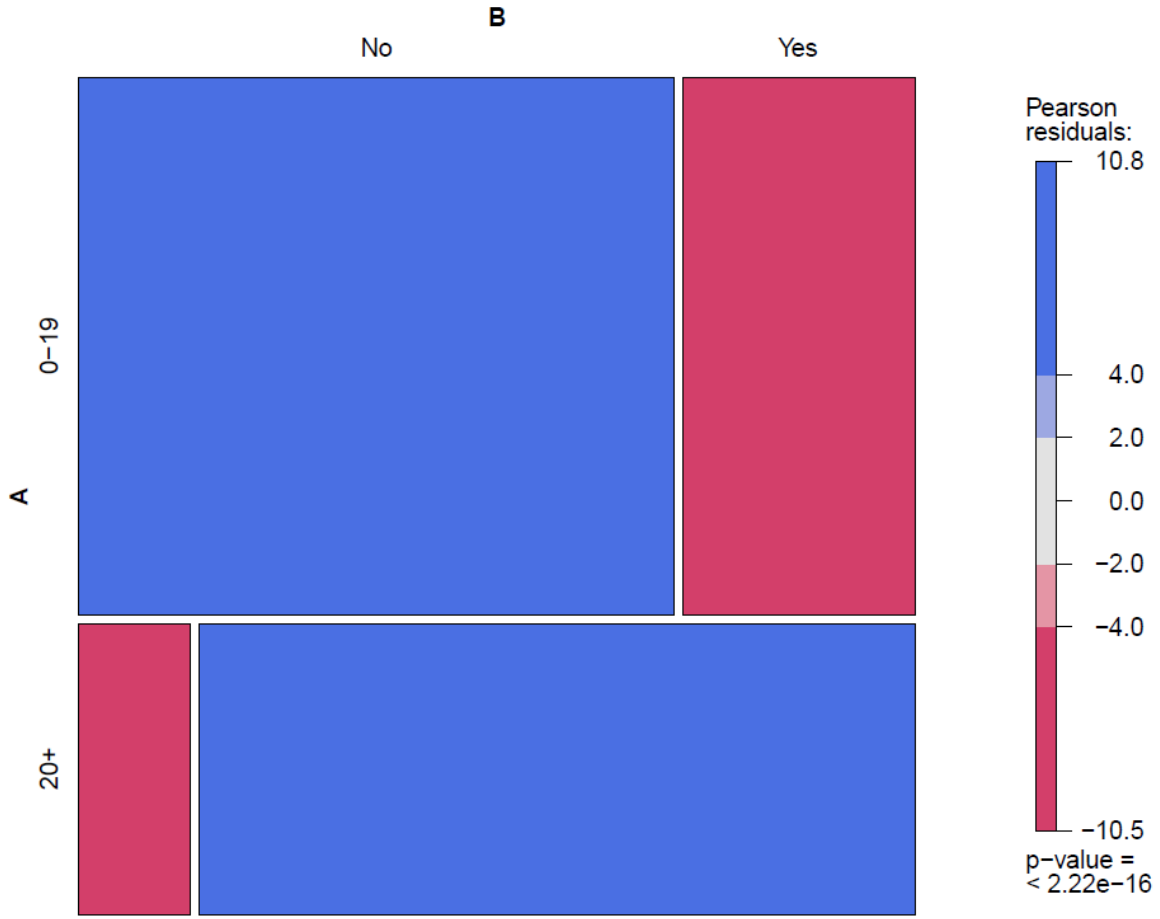
Michael Friendly
(1945 -)



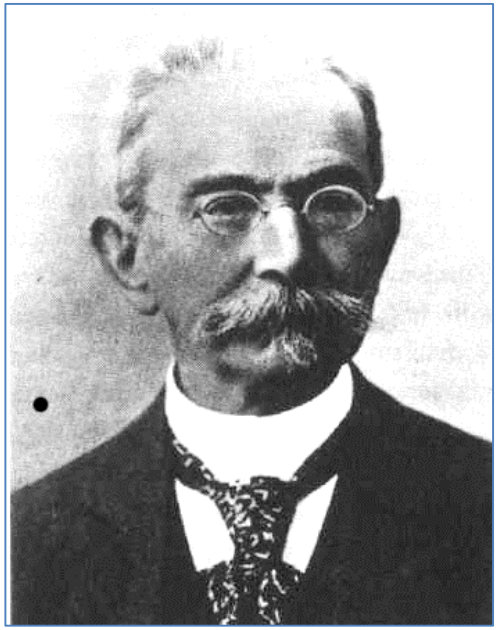


Conditional on *Occupational Exposure*

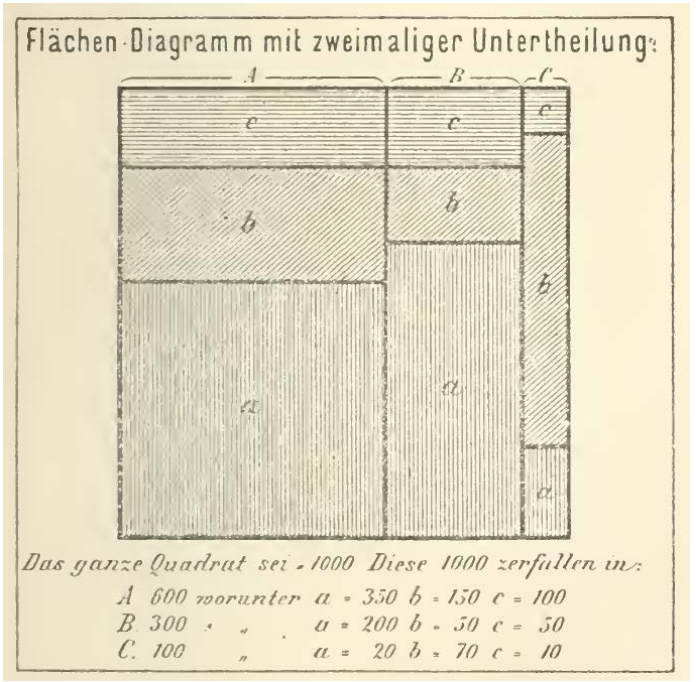
Who Developed the Mosaic Display?



Who Developed the Mosaic Display?



Georg von Mayr
(1841 - 1925)



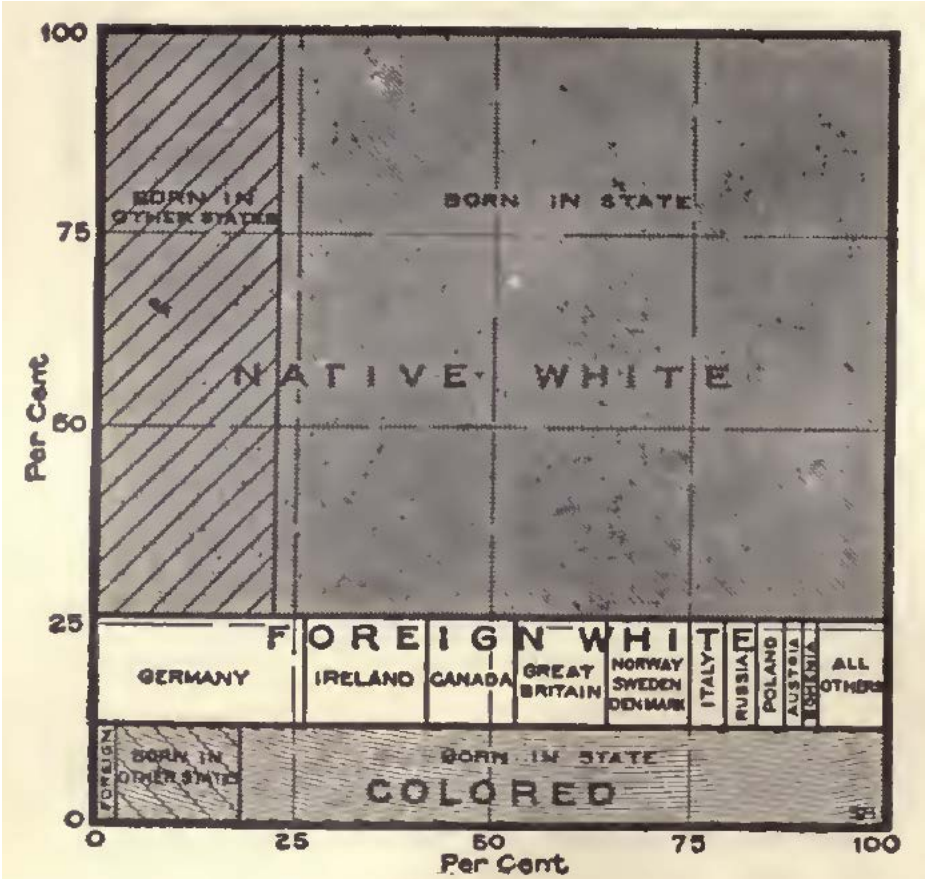
In English, he referred to them as a *"Proportional Square"*

Although often accredited to

Hartigan, J.A. and Kleiner, B. (1981), Mosaics for contingency tables, In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (ed. W.F. Eddy), Springer, pp. 268 – 273 (a follow-up paper appeared in JASA in 1984)

Chernoff Face

Who Developed the Mosaic Display?



United States Statistical Atlas for the Census of 1900

Fig. 5. Elements of the Population of the United States in 1900

Brinton (1914, pg 8)

Chernoff Face

Variations of thee Mosaic Display

Sieve Diagrams
(or Parquet diagram)

The difference between the observed and expected cell frequencies is reflected by the density of the shading.

Cells whose observed frequency is larger than its expected value appears more denser than average.

Colour reflects direction of deviation (red = negative, blue = positive)

Figure 10. Sieve diagram for hair-color, eye-color data. Observed frequencies are equal to the number squares in each cell, so departure from independence appears as variations in shading density.

Variations of thee Mosaic Display

Chernoff Face

Mosaic Matrix

- Good for multiple categorical variables
- Only represents **bivariate** associations
- Does not easily provide a clear visualisation trivariate, or "higher"-variate, association structures

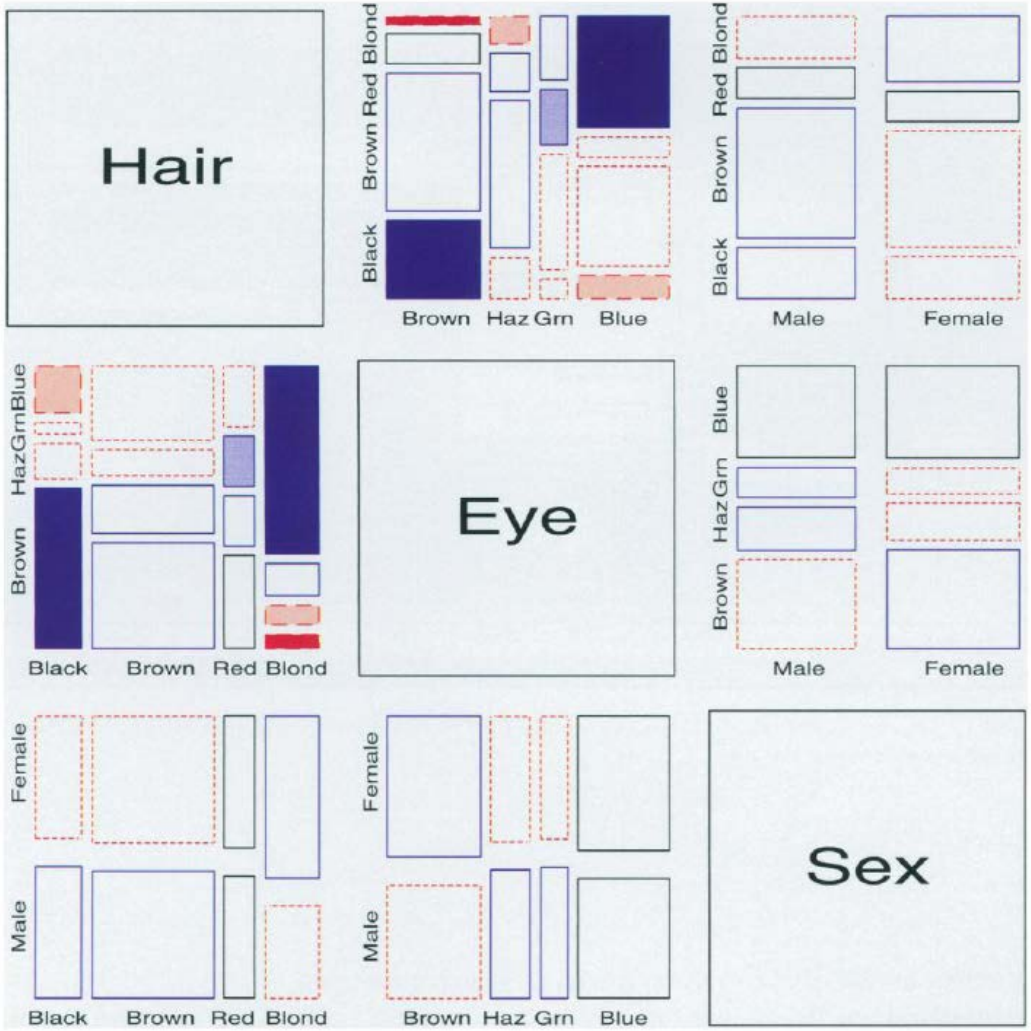


Figure 11. Mosaic matrix for hair color, eye color, and sex. Each off-diagonal panel shows the two-way mosaic for the corresponding row and column variables.

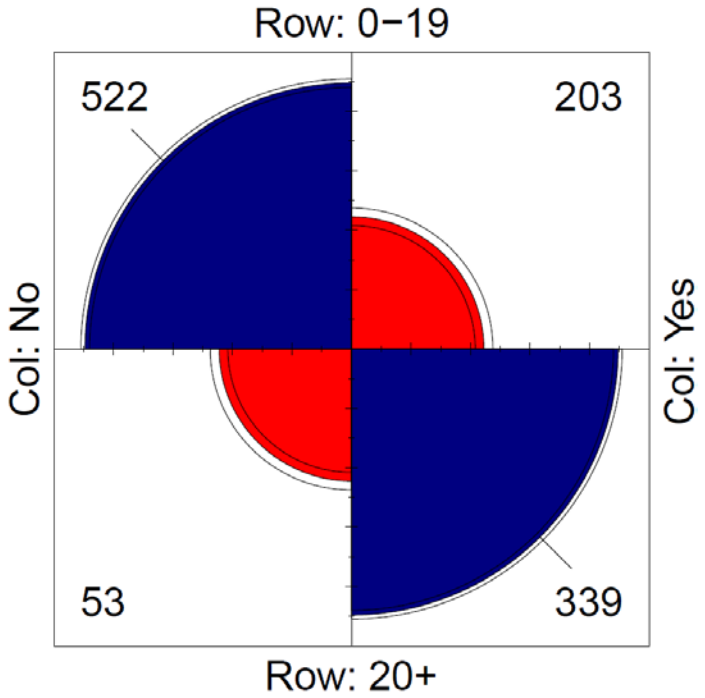
The Fourfold Display

Stephen E. Fienberg
(1942 -)

In the past, a 2x2 contingency table was also referred to as a fourfold table (since the summarised data consists of four joint frequencies). In the same vain as Florence Nightingale's coxplot, Fienberg (1975) proposed the fourfold display.

Suppose we consider the following 2x2 contingency table which is based on a sample of 1117 metropolitan New York workers during the 1960's. This data summarises the length occupational exposure to asbestos (in years) against whether they contracted asbestosis (Selikoff, 1981).

Occupational Exposure (yrs)	Asbestosis No	Yes	Total
0-19	522	203	725
20+	53	339	392
Total	575	542	1117



In R, fourfold displays may be constructed using the `fourfold` function in the `vcd` library.

The Fourfold Display

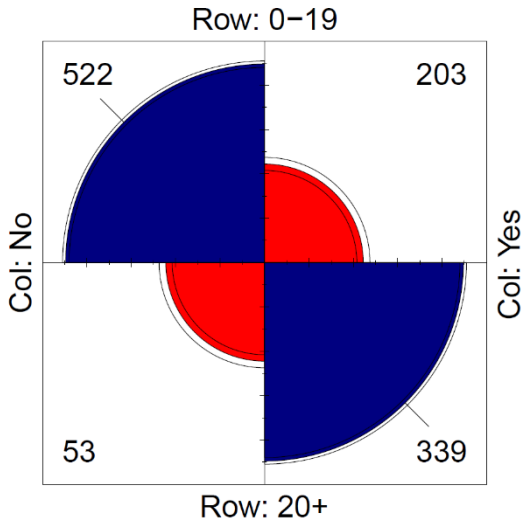
- When analysing 2x2 tables

Fienberg, S.E. (1975) 'Perspective Canada' as a social report. *Social Indicators Research*, **2**, 153 – 174.

Stephen E. Fienberg (1942 -)

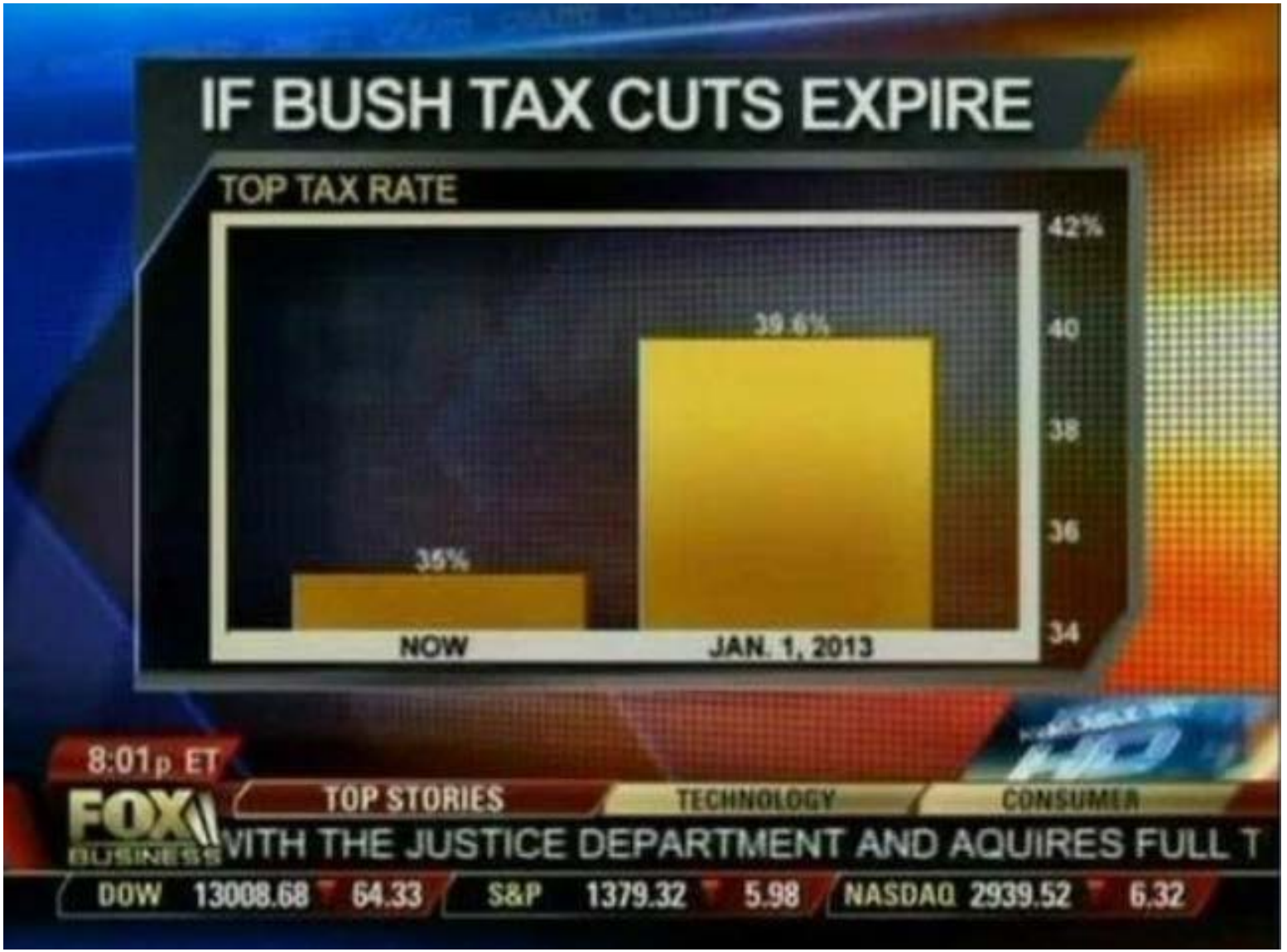
Occupational Exposure (yrs)	Asbestosis No	Yes	Total
0-19	522	203	725
20+	53	339	392
Total	575	542	1117

$$\text{OR} = \frac{522 \times 339}{53 \times 203} = 16.45$$

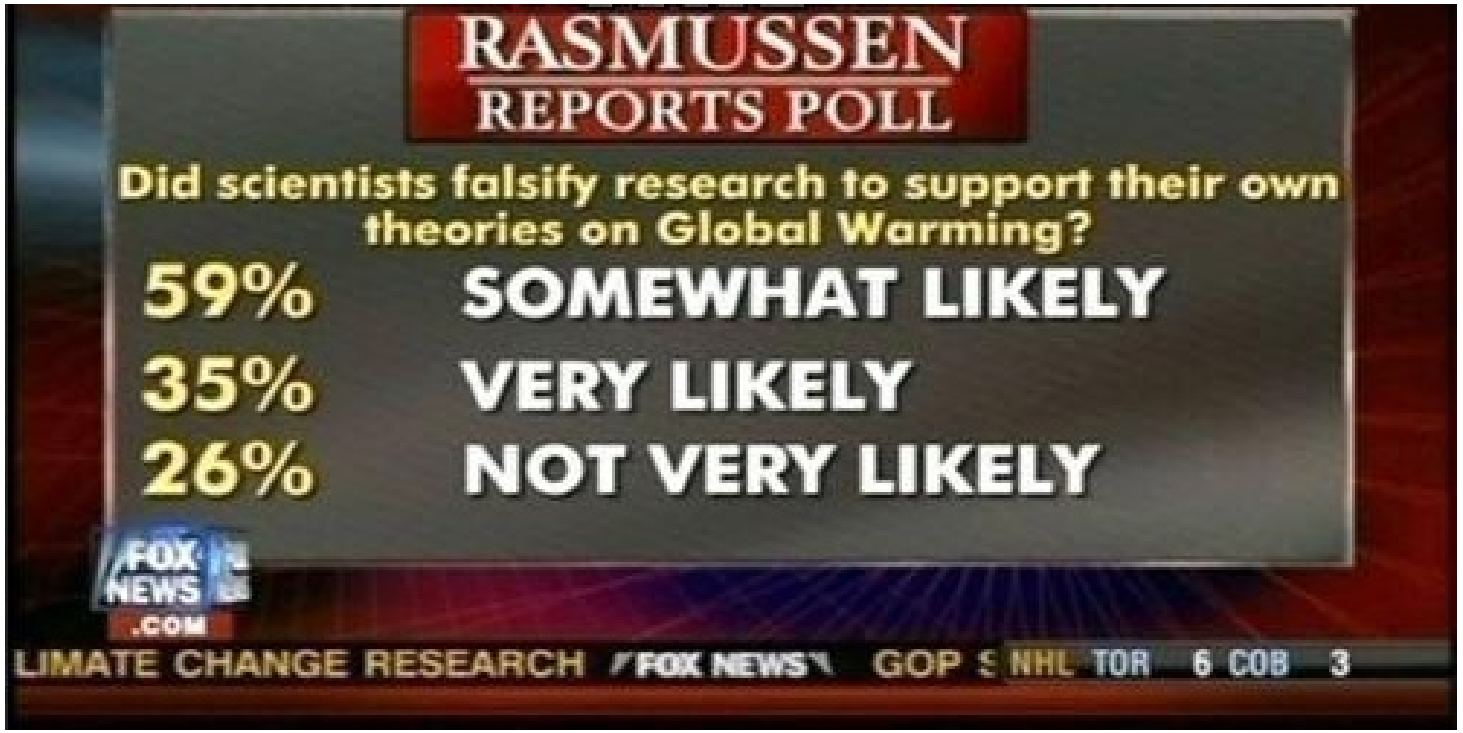


- The area of the quarter circles are proportional to the joint relative frequencies, p_{ij}
- Comparisons between adjacent quarter circles – in terms of the radii lengths – is based on a square root scale
- Shading emphasises the direction of the association
- Unlike squares, using quarter circles is meant to give a "flowing" effect

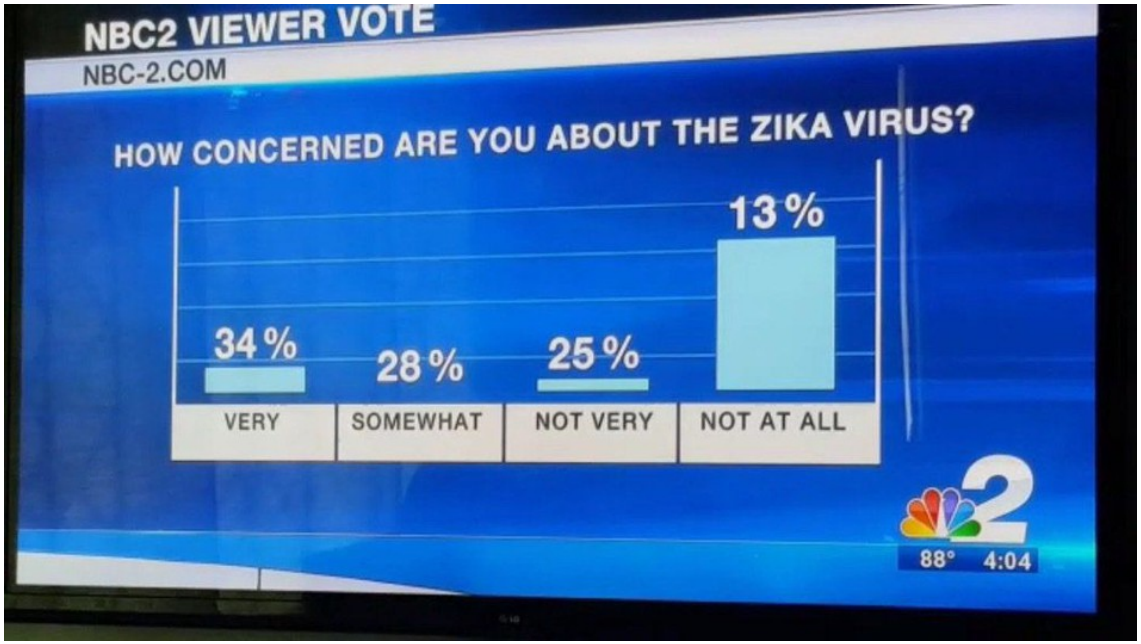
Generally . . .
Here are some bad graphics (from the US media)

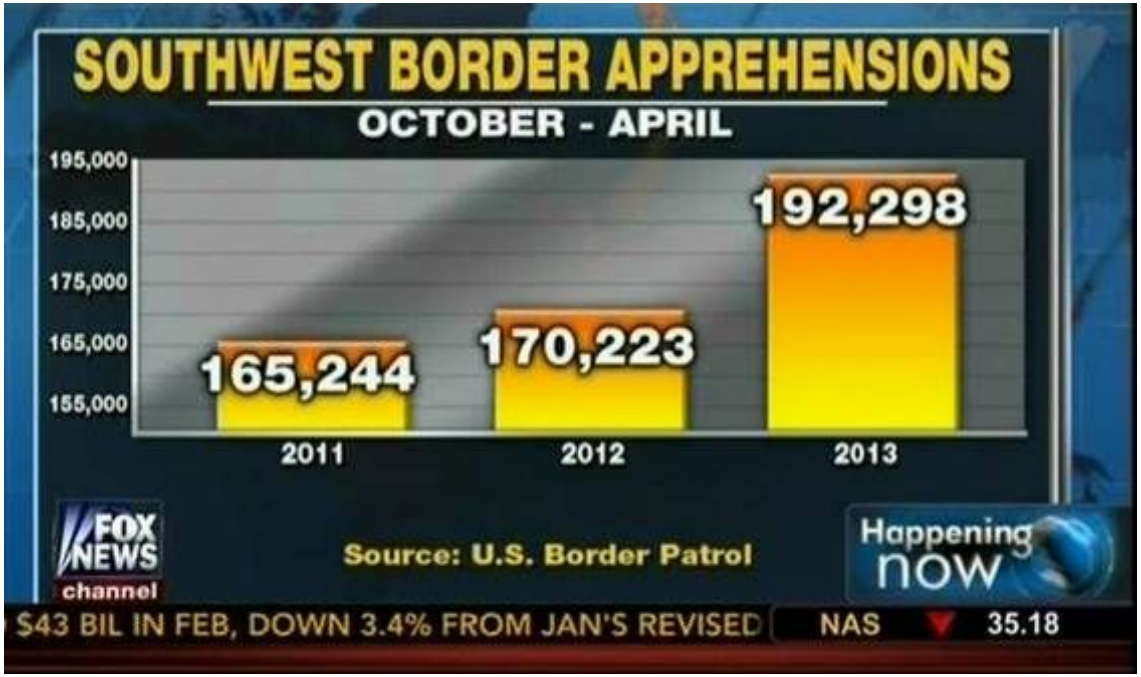


Generally . . .
Here are some bad graphics (from the US media)



- Wainer, H. (1984), How to display data badly, *The American Statistician*, 38, 137 - 147
- Carr, D. B. and Nusser, S. M. (1995), Converting tables to plots: A challenge from Iowa State, Statistical Computing & Statistical Graphics Newsletter, 6, 11 – 18.
- Gelman, A., Pasarica, C. and Dodhia, R. (2002), Let's practice what we preach: Turning tables into graphs, *The American Statistician*, 56, 121 – 130.

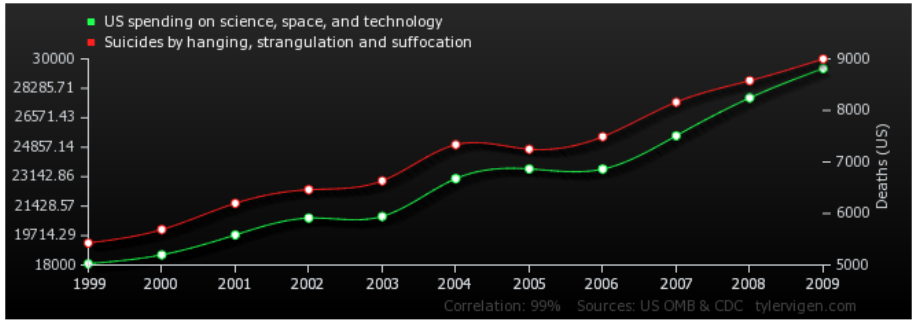




https://www.mediamatters.org/research/2012/10/01/a-history-of-dishonest-fox-charts/190225

Correlation is not Causation

US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation



Upload this chart to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
US spending on science, space, and technology *Millions of todays dollars (US OMB)*	18,079	18,594	19,753	20,734	20,831	23,029	23,597	23,584	25,525	27,731	29,449
Suicides by hanging, strangulation and suffocation *Deaths (US) (CDC)*	5,427	5,688	6,198	6,462	6,635	7,336	7,248	7,491	8,161	8,578	9,000

Correlation: 0.992082

Permalink - Mark as interesting - Not interesting

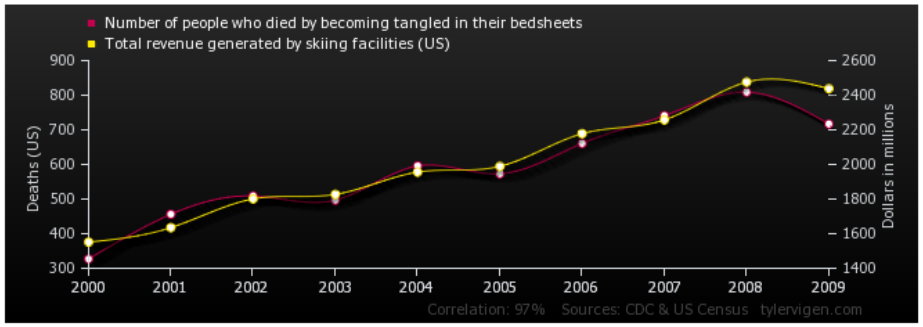
Correlation is not Causation

Correlation is NOT Causation

Number of people who died by becoming tangled in their bedsheets

correlates with

Total revenue generated by skiing facilities (US)



Upload this chart to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number of people who died by becoming tangled in their bedsheets *Deaths (US) (CDC)*	327	456	509	497	596	573	661	741	809	717
Total revenue generated by skiing facilities (US) *Dollars in millions (US Census)*	1,551	1,635	1,801	1,827	1,956	1,989	2,178	2,257	2,476	2,438

Correlation: 0.969724

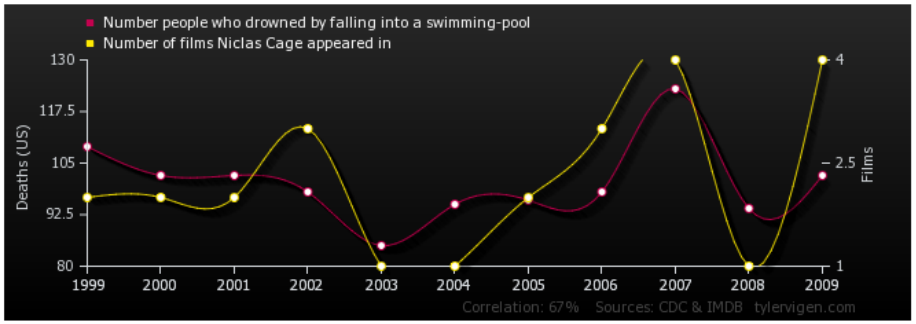
Permalink - Mark as interesting - Not interesting

Correlation is not Causation

Number people who drowned by falling into a swimming-pool

correlates with

Number of films Nicolas Cage appeared in



Upload this chart to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Number people who drowned by falling into a swimming-pool *Deaths (US) (CDC)*	109	102	102	98	85	95	96	98	123	94	102
Number of films Nicolas Cage appeared in *Films (IMDB)*	2	2	2	3	1	1	2	3	4	1	4

Correlation: 0.666004

Permalink - Mark as interesting - Not interesting

Lots more available at

http://www.tylervigen.com/

Correlation is NOT Causation

Science alert Australia & New Zealand

Home | Life | Health | Environment | Technology | Culture | Space

News

Female-named hurricanes kill more than male-named ones

GABRIELLA MUNOZ
TUESDAY, 03 JUNE 2014

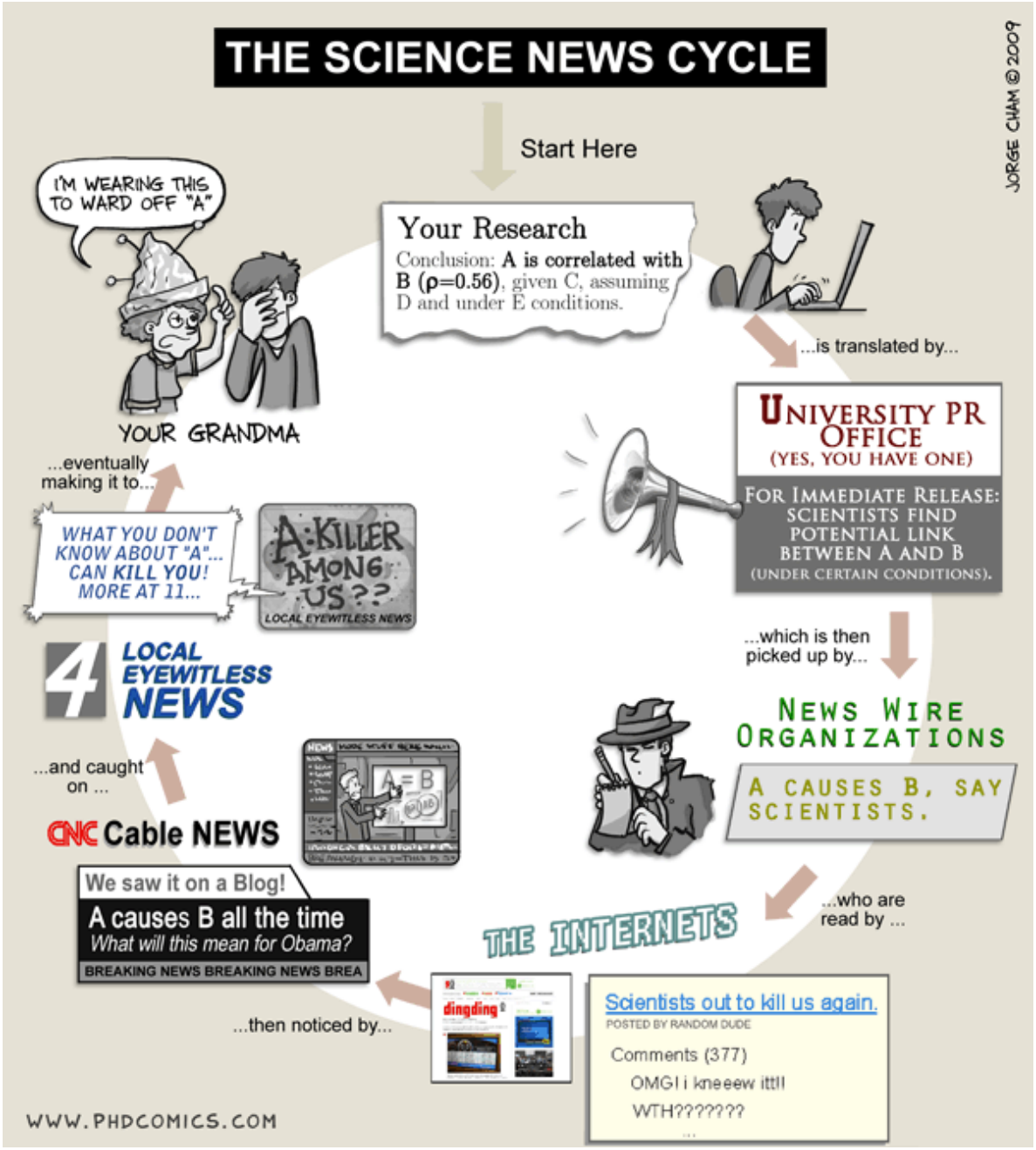
Like 1.2k Tweet 5 Send

New research suggests that lady hurricanes cause more havoc because people don't take them seriously.

Tropical storm Olivia as seen from space.

Image: NASA

http://sciencealert.com.au/news/20140306-25605.html



http://www.phdcomics.com/comics/archive.php?comicid=1174

Next Week

- Next week (Week 2) we shall look at a bit of the history and development of Pearson's famous chi-squared statistic
- In doing so we shall also explore the impact of Lambert Quetelet, Francis Galton and measures of association that involve the chi-squared statistic

- *Week 3* – measures of association for two dichotomous categorical variables in the form of a 2x2 contingency table
- *Week 4* – measures of association for IxJ tables

References

- Becketti, S. and Gould, W. (1987), Rangefinder boxplots: A note, *The American Statistician*, 41, 149.
- Benjamini, Y. (1988), Opening of the box of a boxplot, *The American Statistician*, 42, 257-262
- Chernoff H 1973 The use of faces to represent points in kdimensional space graphically. *Journal of the American Statistical Association 68, 361 – 368.*
- Eells WC 1926 The relative merits of circles and bars for representing component parts. *Journal of the American Statistical Association 21, 119 – 132.*
- Fienberg SE 1975 “Perspective Canada" as a social report. *Social Indicators Research, 2, 153 – 174.*
- Friendly M 2008 A brief history of data visualization. In *Handbook of Data Visualization (eds Chen C, HÄadle W.* and Unwin A), pp. 15 – 56. Springer
- Friendly M and Denis D 2005 The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences 41, 103 – 130.*
- Funkhouser HG 1936 A note on a tenth century graph. *Osiris, 1, 260 – 262.*
- Funkhouser HG 1937 Historical development of the graphical representation of statistical data. *Osiris, 3, 269 404.*
- Galton F 1886 Regression towards mediocrity in hereditary stature. *he Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246 263*
- Goldberg, K.M. and Iglewicz, B. (1992), Bivariate extensions of the boxplot, *Technometrics*, 34, 307-320.

References

- Haemer, K.W. (1948), Range-bar charts, *The American Statistician*, 2, 23.
- Haskell, A. C. (1922), *Graphic Charts in Business*, New York: Codex
- Hintze, J.L. and Nelson, R. D. (1998), Violin plots: A box plot-density trace synergism, *The American Statistician*, 52, 181-184
- McGill, R., Larsen, W.A. and Tukey, J.W. (1978), Variations of Boxplots, *The American Statistician*, 32, 12-16
- Pearson K 1895 Contributions to the Mathematical Theory of Evolution II: Skew variation in homogeneous material.
- *Philosophical Transactions of the Royal Society of London, Series A 186, 343 – 414.*
- Playfair W 1801 *The Statistical Breviary: Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe. T. Bensley.*
- Rousseeuw, P.J., Ruts, I. And Tukey, J.W. (1999), The bagplot: A bivariate boxplot, *The American Statistician*, 53, 382-387.
- Selikoff IJ 1981 Household risks with inorganic fibers, *Bulletin of the New York Academy of Medicine 57, 947 – 961*.
- Spear, M. E. (1952), *Charting Statistics*, McGraw-Hill. pp. 164-166
- Spence I 2005 No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics 30, 353 – 368.*
- Tufte, E. R.(2001), *The Visual Display of Quantitative Information*, Graphics Press.
- Tukey JW 1977 *Exploratory Data Analysis. AddisonWesley*