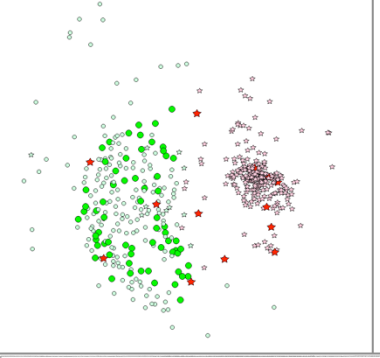
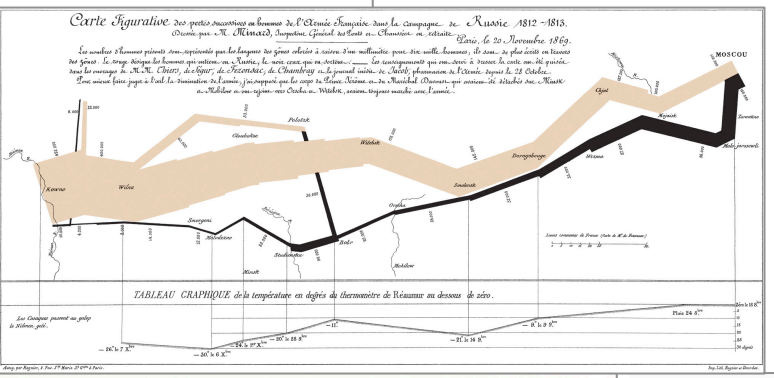
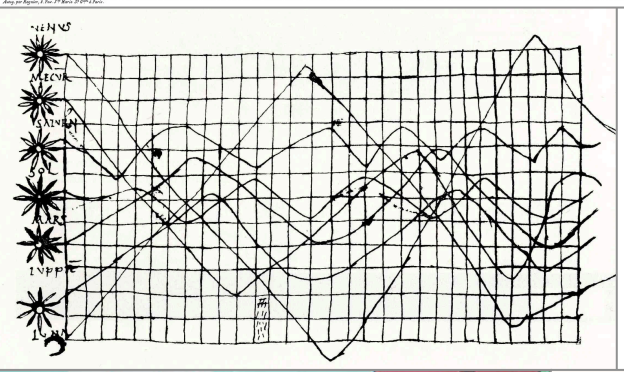
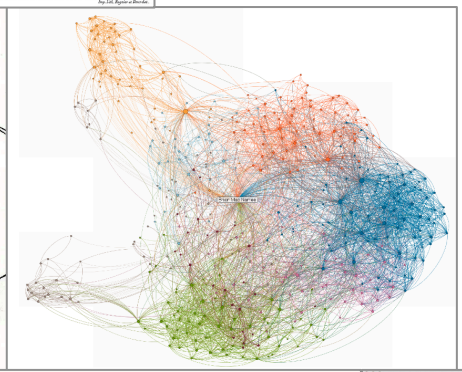


Simple Information Visualisation Summary

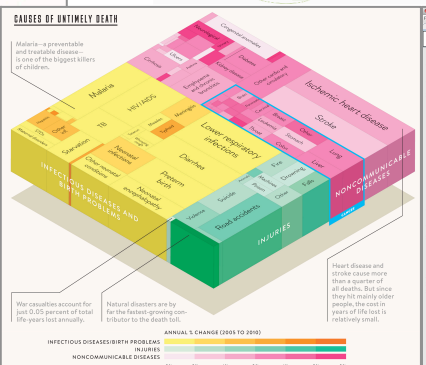






Information Visualisation COMP40610

Dr. Brian Mac Namee



Origins

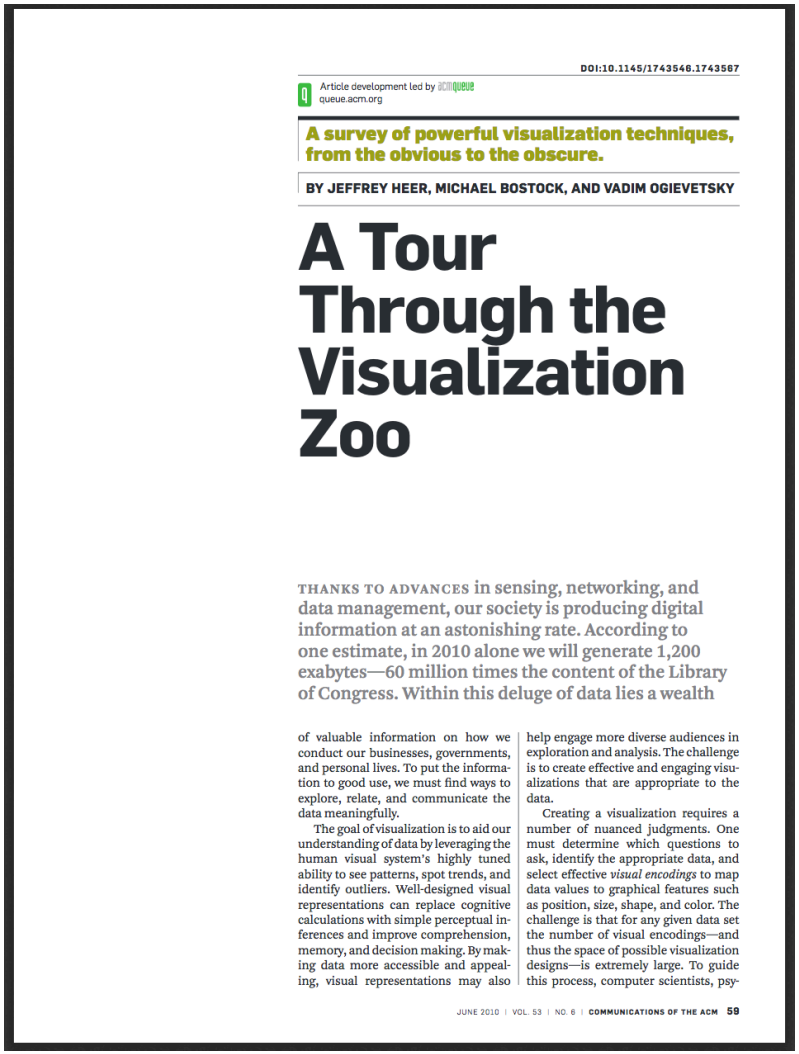
This course curates material from multiple online and published sources

When this is the case full citations will be given

The Visualistion Zoo

We can group different visualisations into the following general types:

- Nominal comparison
- Distributions
- Trends over time
- Relationships
- Data tables



DOI:10.1145/1743546.1743567

Article development led by acmqueue
queue.acm.org

A survey of powerful visualization techniques, from the obvious to the obscure.

BY JEFFREY HEER, MICHAEL BOSTOCK, AND VADIM OGIEVETSKY

A Tour Through the Visualization Zoo

THANKS TO ADVANCES in sensing, networking, and data management, our society is producing digital information at an astonishing rate. According to one estimate, in 2010 alone we will generate 1,200 exabytes—60 million times the content of the Library of Congress. Within this deluge of data lies a wealth of valuable information on how we conduct our businesses, governments, and personal lives. To put the information to good use, we must find ways to explore, relate, and communicate the data meaningfully.

The goal of visualization is to aid our understanding of data by leveraging the human visual system's highly tuned ability to see patterns, spot trends, and identify outliers. Well-designed visual representations can replace cognitive calculations with simple perceptual inferences and improve comprehension, memory, and decision making. By making data more accessible and appealing, visual representations may also help engage more diverse audiences in exploration and analysis. The challenge is to create effective and engaging visualizations that are appropriate to the data.

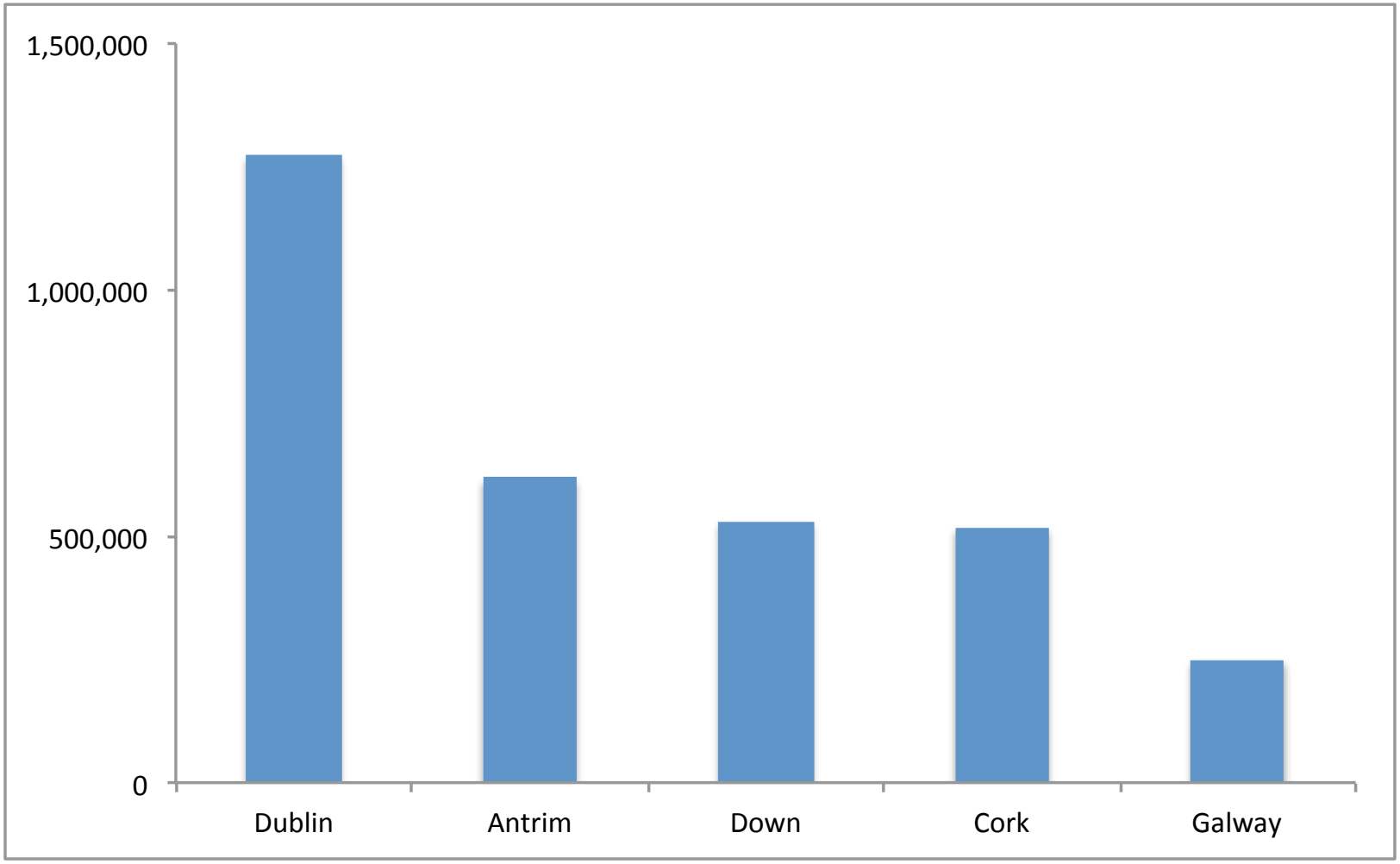
Creating a visualization requires a number of nuanced judgments. One must determine which questions to ask, identify the appropriate data, and select effective *visual encodings* to map data values to graphical features such as position, size, shape, and color. The challenge is that for any given data set the number of visual encodings—and thus the space of possible visualization designs—is extremely large. To guide this process, computer scientists, psy-

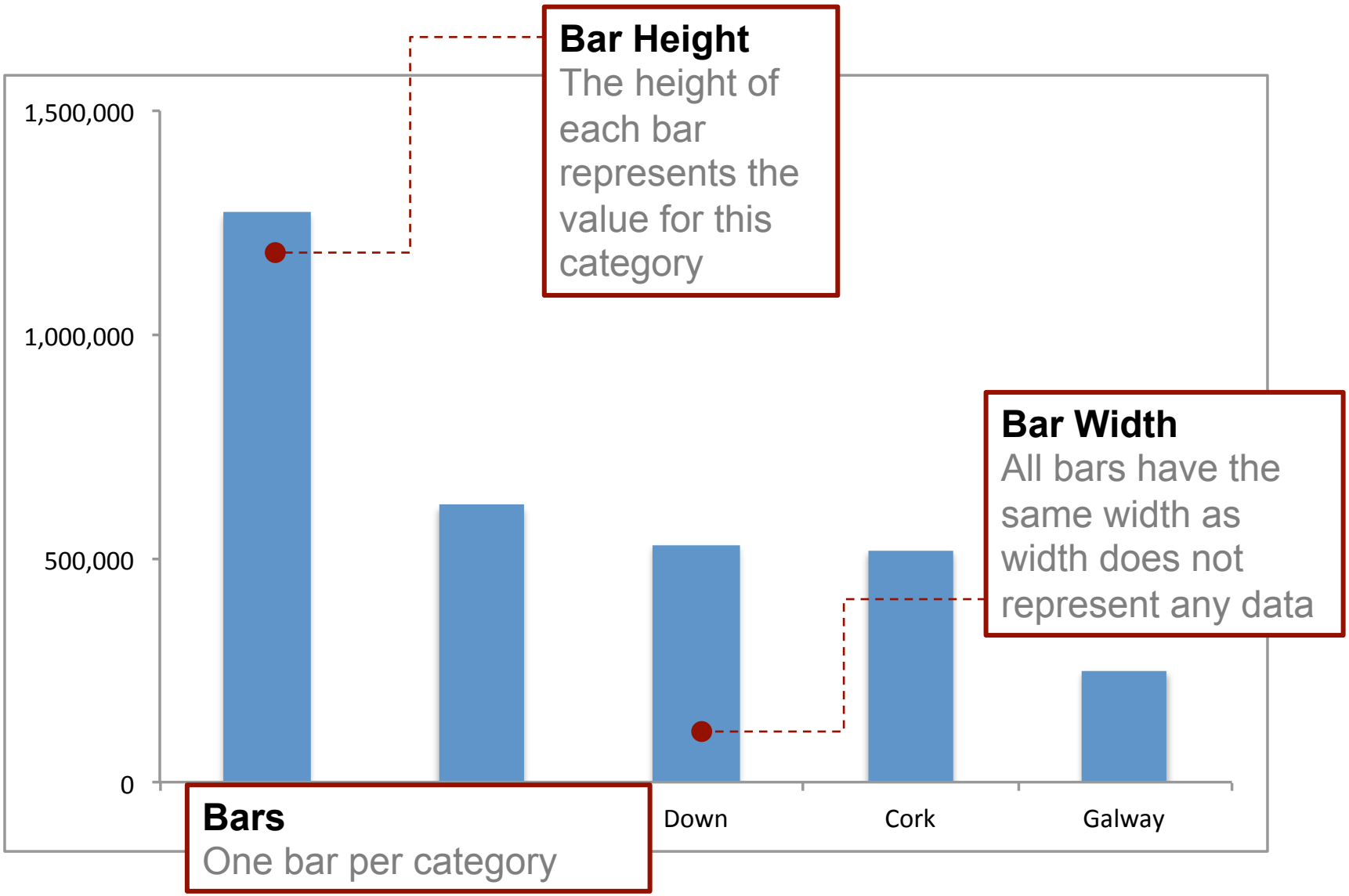
"A Tour through the Visualization Zoo", Jeffrey Heer, Michael Bostock & Vadim Ogievetsky, Queue, Volume 8 Issue 5, 2010

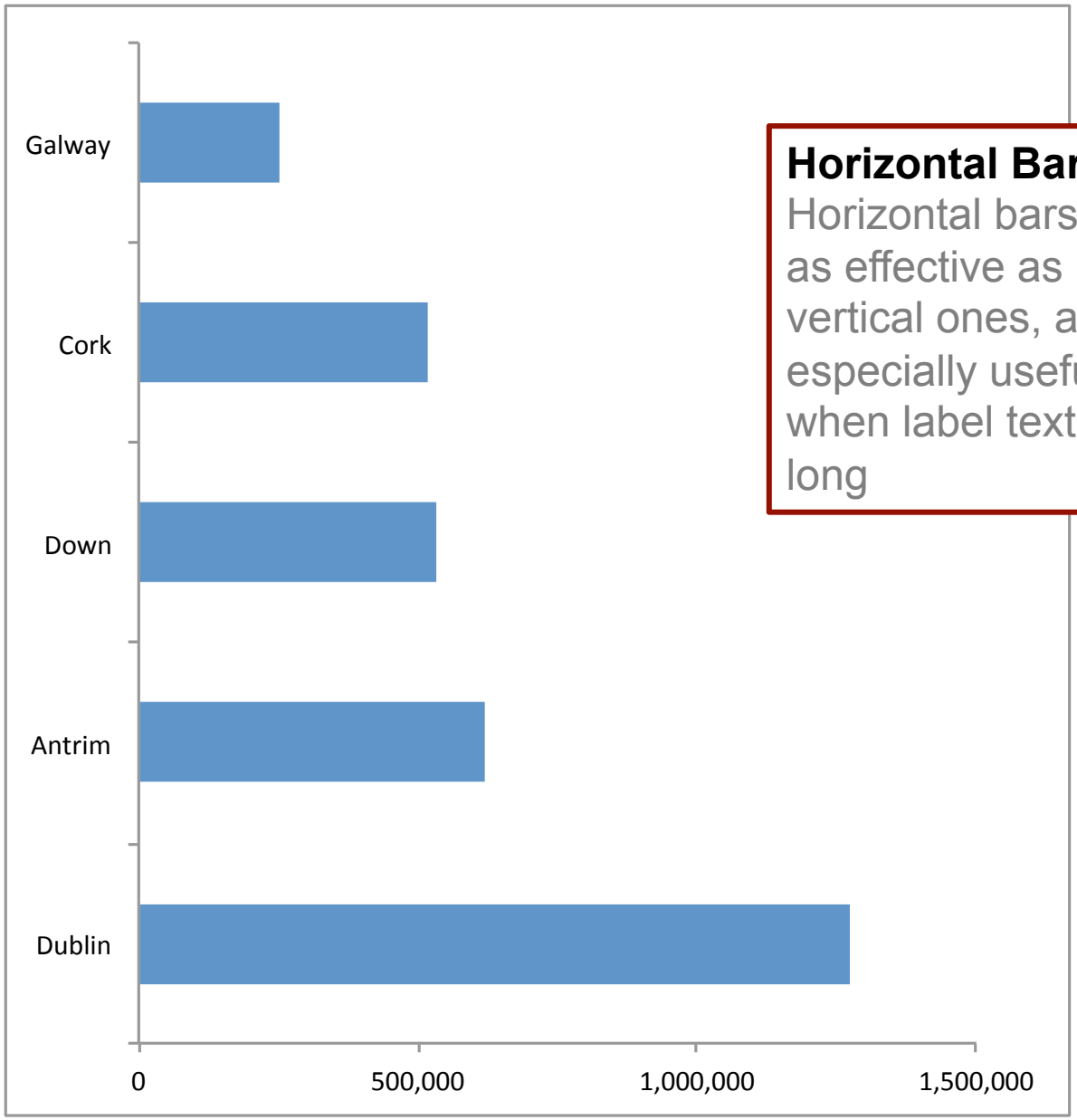
http://queue.acm.org/detail.cfm?id=1805128

NOMINAL COMPARISONS

Simple Bar Graph



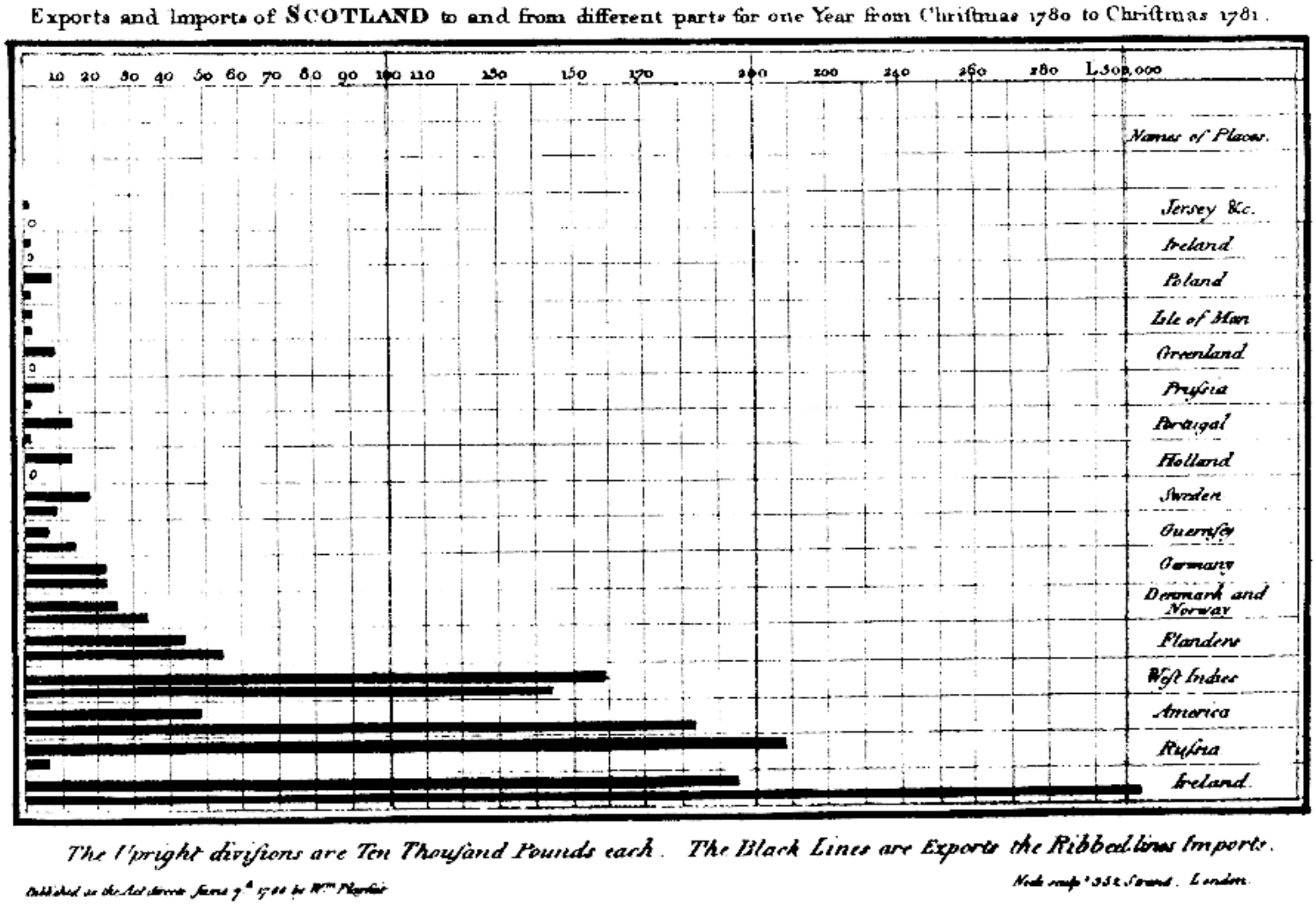

Bar Height
The height of each bar represents the value for this category
1,500,000
1,000,000
500,000
0
Bar Width
All bars have the same width as width does not represent any data
Down
Cork
Galway
Bars
One bar per category



Horizontal Bars

Horizontal bars are as effective as vertical ones, and especially useful when label text is long

The First Bar Graph!



William Playfair's "*Commercial and Political Atlas*", 1786 via Wikipedia

http://en.wikipedia.org/wiki/William_Playfair

VISUALISING DISTRIBUTIONS

Visualising Distributions

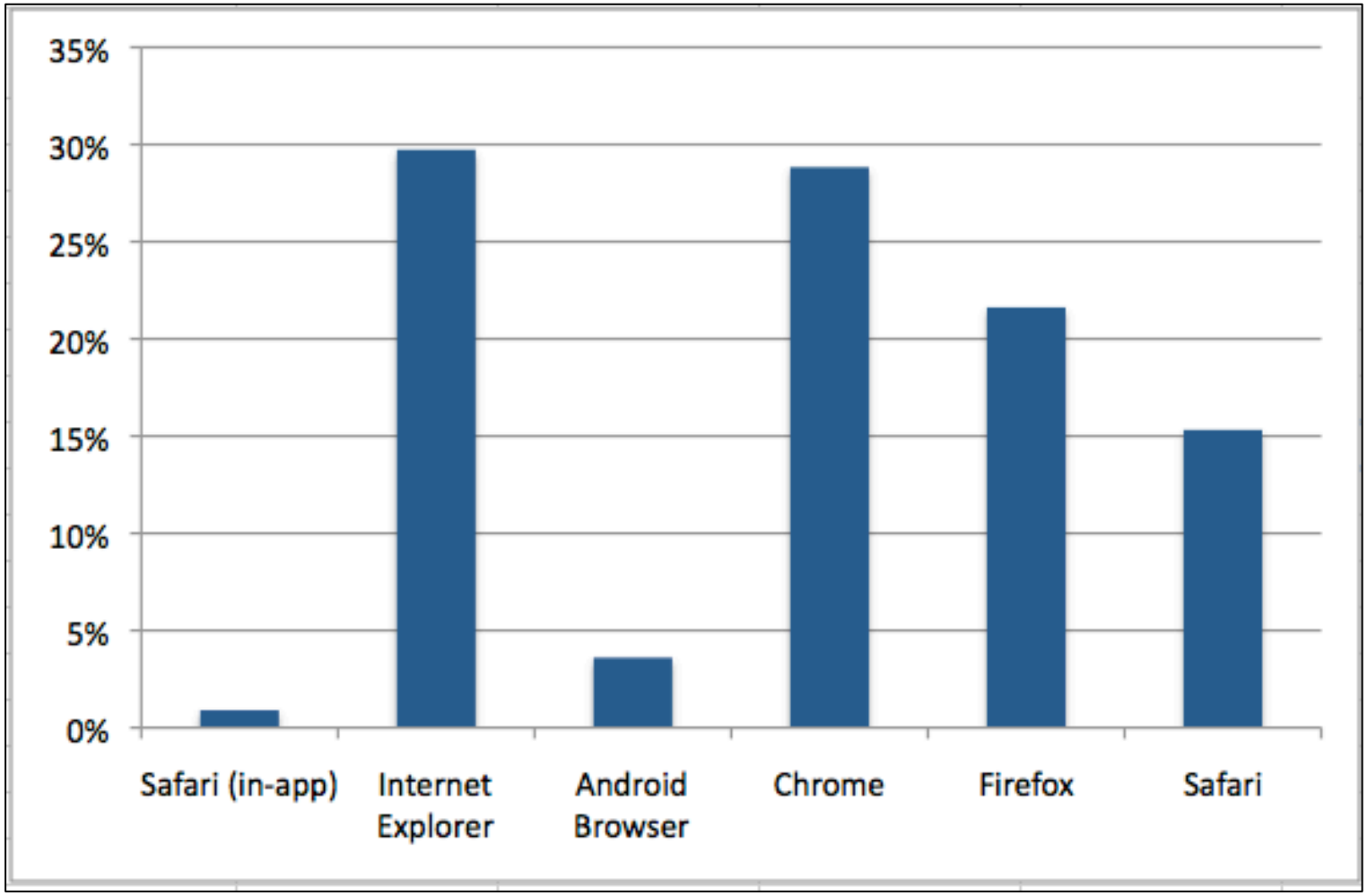
One of the simplest things we can visualise is how the values of a variable are distributed

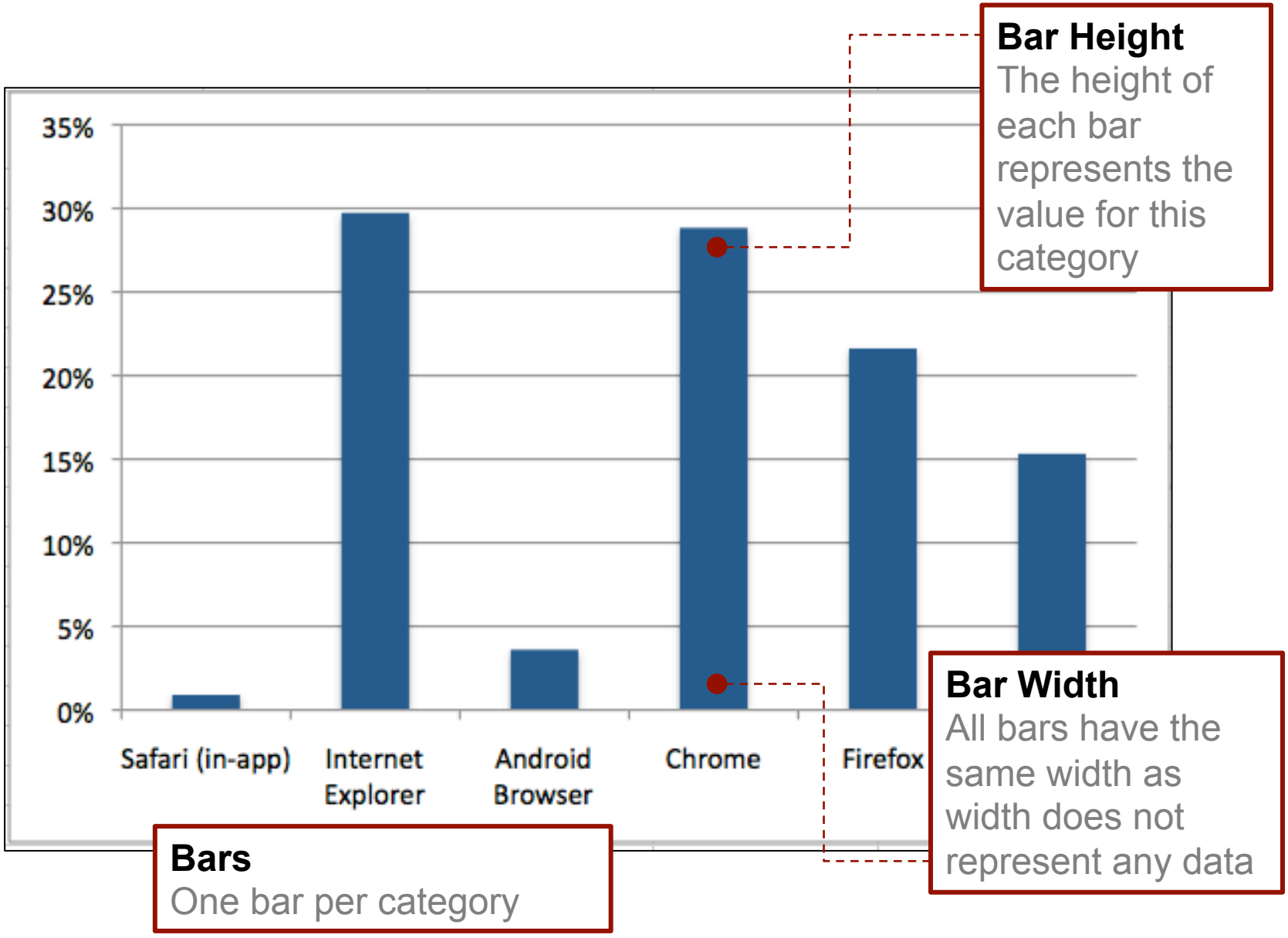
There are a number of useful ways in which we can do this

discrete 离散的

- Discrete variables (parts of a whole)
  - Simple bar graphs
  - Pie/donut charts
- Continuous variables
  - Histograms
  - Density plots
  - Box plots

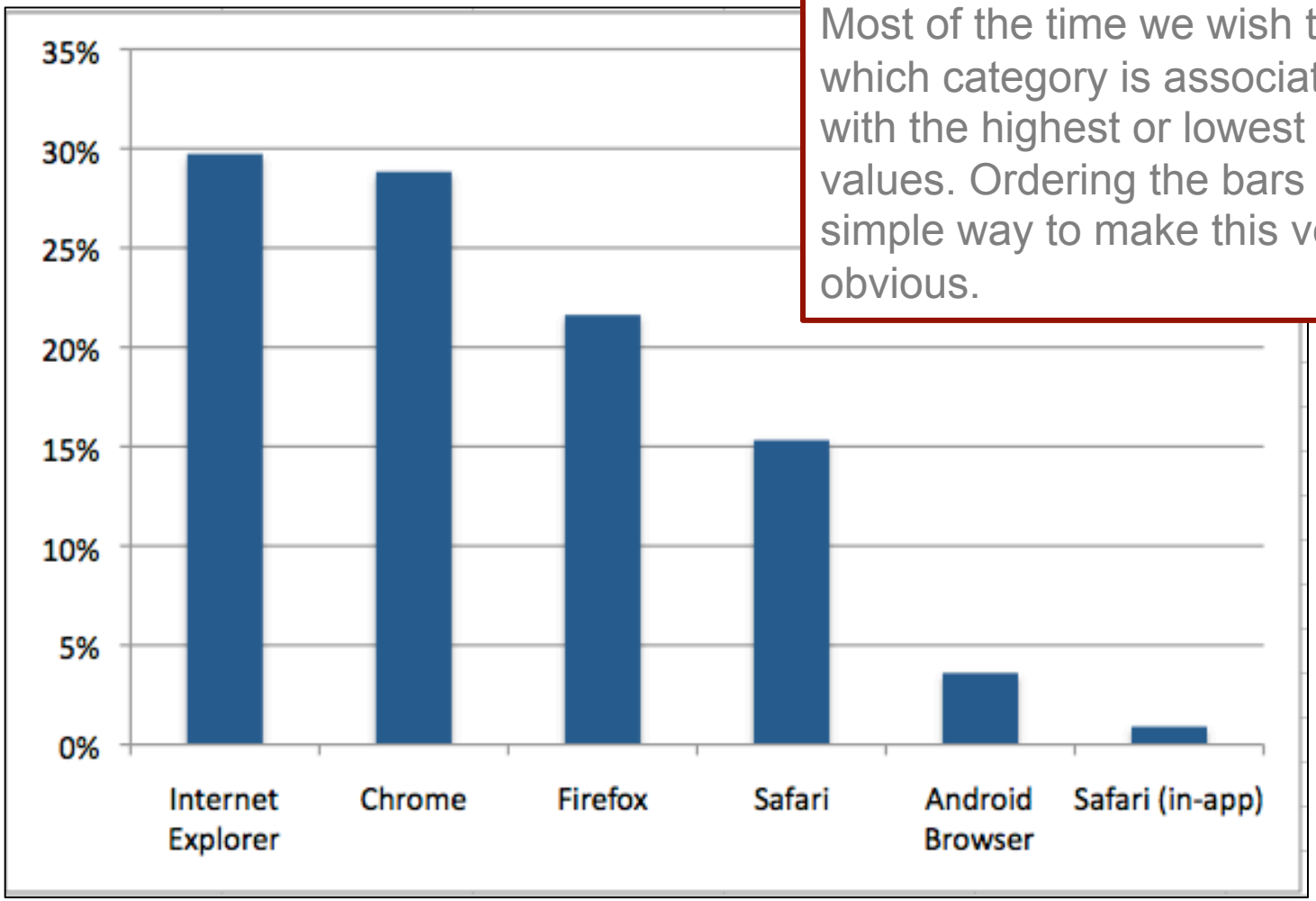
Simple Bar Graph



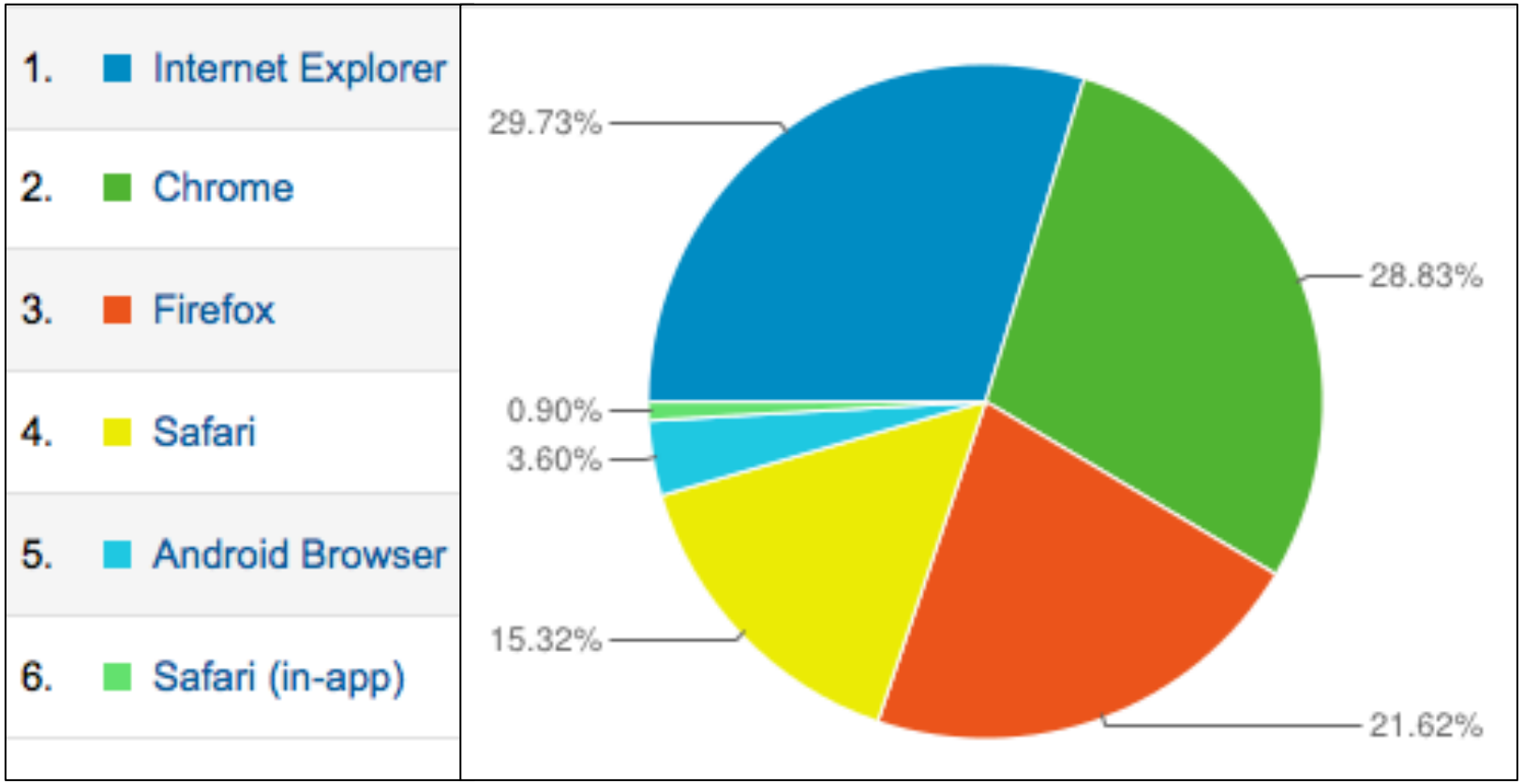

35%
30%
25%
20%
15%
10%
5%
0%
Safari (in-app)
Internet Explorer
Android Browser
Chrome
Firefox
Bar Height
The height of each bar represents the value for this category
Bar Width
All bars have the same width as width does not represent any data
Bars
One bar per category

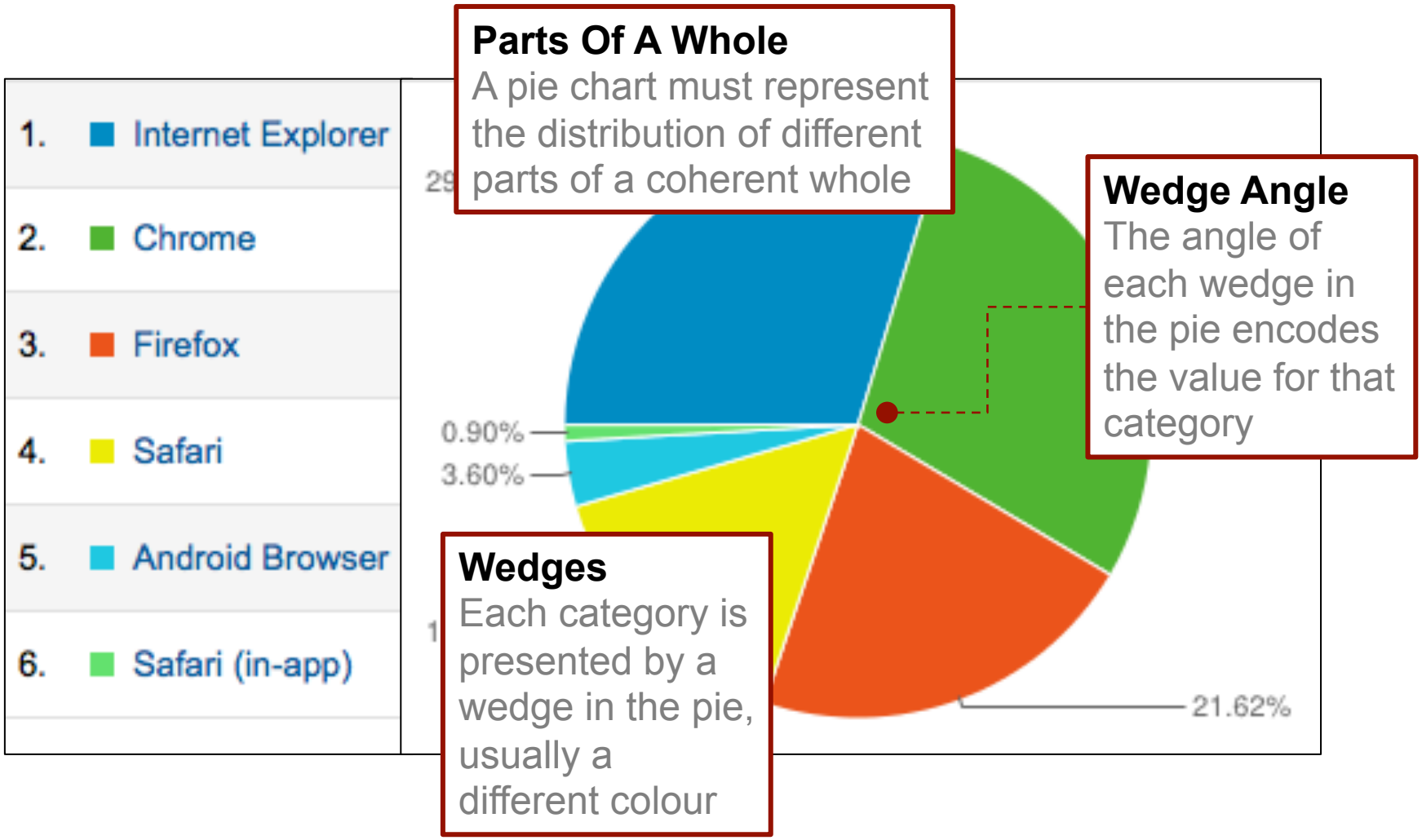
Bar Ordering

Most of the time we wish to see which category is associated with the highest or lowest values. Ordering the bars is a simple way to make this very obvious.

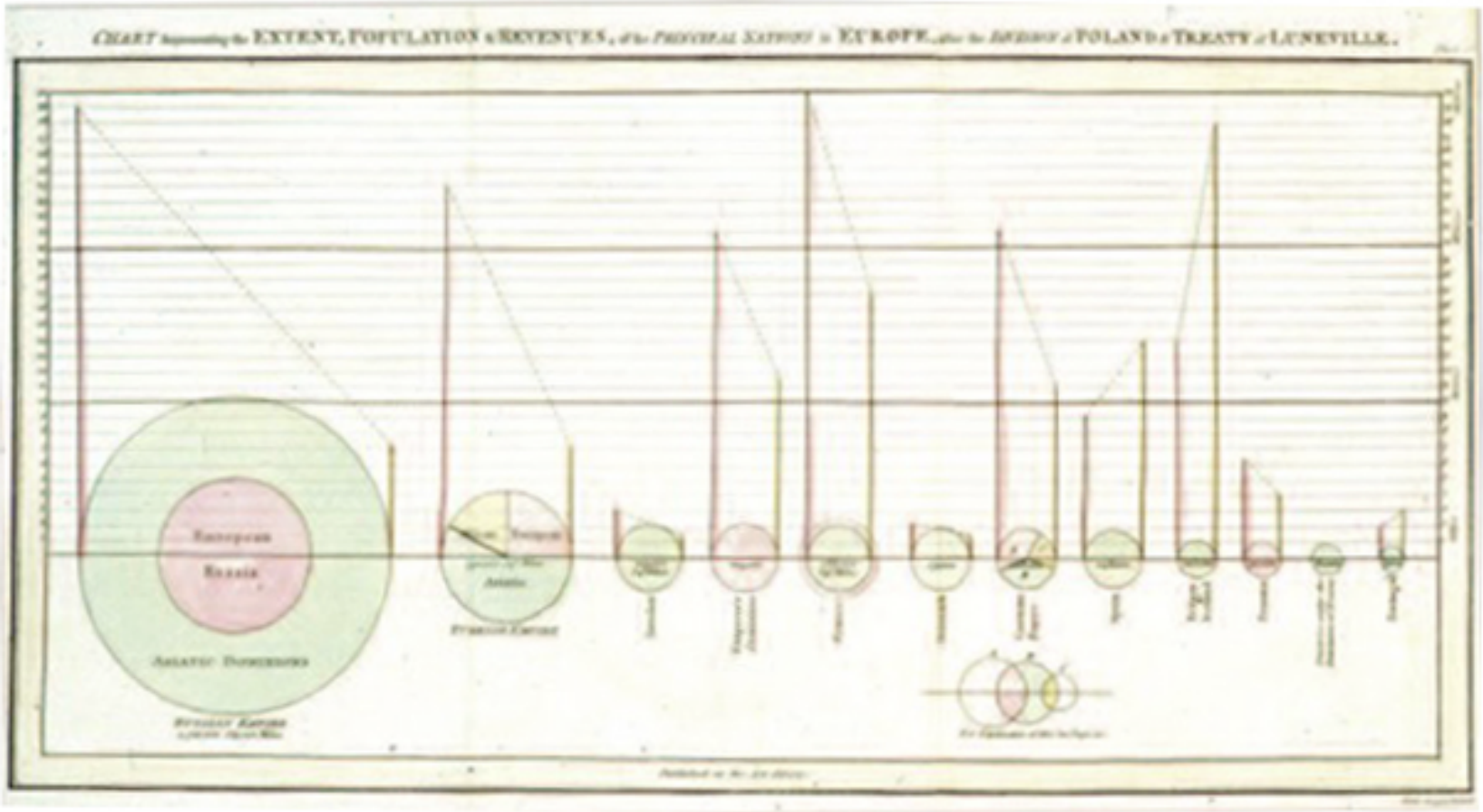


Pie Charts


1. Internet Explorer
2. Chrome
3. Firefox
4. Safari
5. Android Browser
6. Safari (in-app)
29.73%
28.83%
0.90%
3.60%
15.32%
21.62%


1. Internet Explorer
2. Chrome
3. Firefox
4. Safari
5. Android Browser
6. Safari (in-app)
0.90%
3.60%
21.62%
Parts Of A Whole
A pie chart must represent the distribution of different parts of a coherent whole
Wedge Angle
The angle of each wedge in the pie encodes the value for that category
Wedges
Each category is presented by a wedge in the pie, usually a different colour

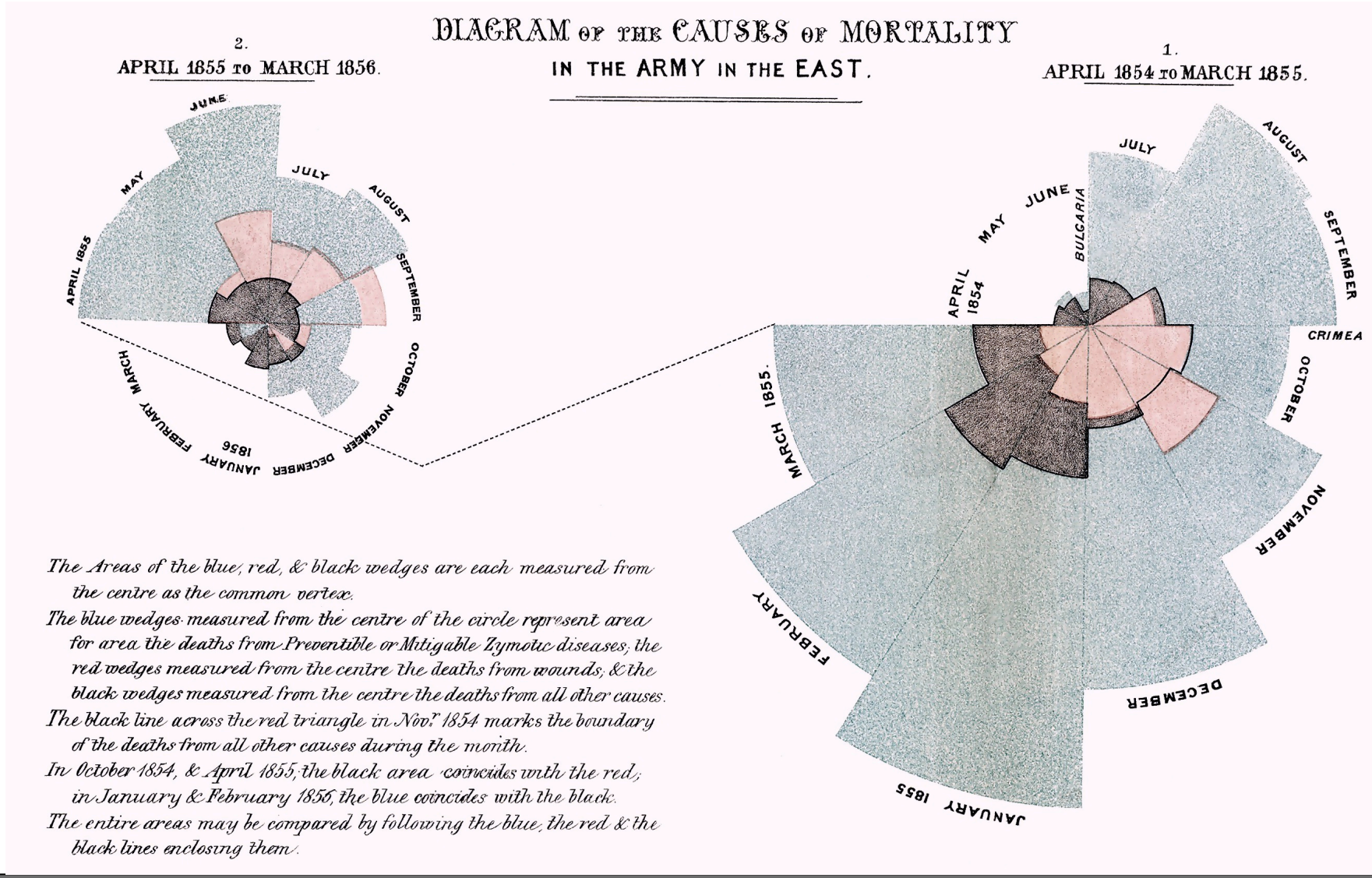
The First Pie Chart!



William Playfair's "Statistical Breviary", 1801 via The New York Times

http://www.nytimes.com/2012/04/22/magazine/who-made-that-pie-chart.html?_r=0

Florence Nightingale's Pie Charts



Florence Nightingales' Crimean War Death Charts via:
http://www.uh.edu/engines/epi1712.htm

Pie Charts

Pie charts are the subject of a lot of negative comment

- The main reason is that their descriptive power is based on our ability to interpret differences in angle - which we are not very good at!

Pie charts might be useful when:

- We have a small number of categories (< 6)
- The values sum to a meaningful whole
- The differences are coarse

A bar plot is almost always a better solution

Histograms

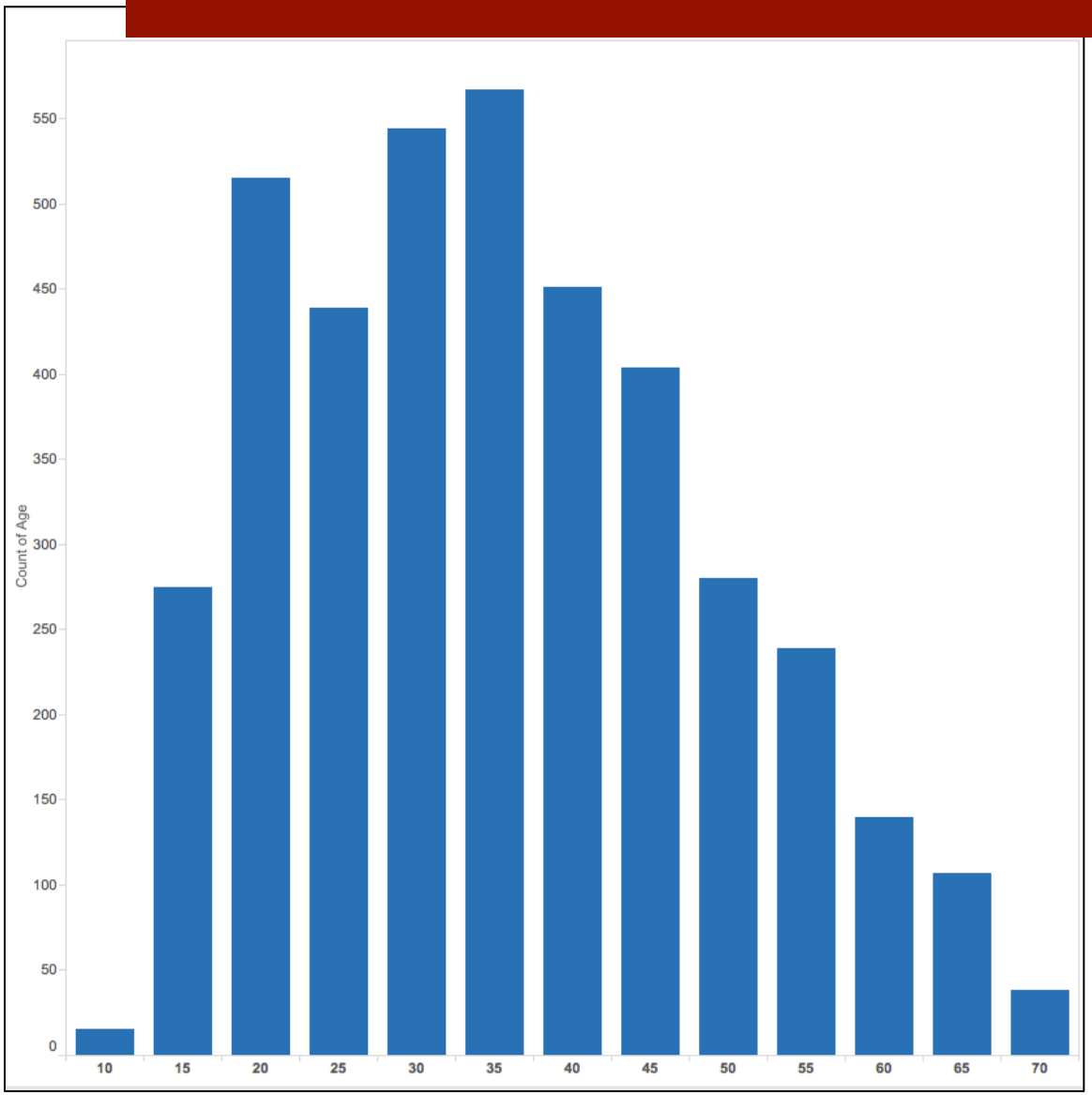
A histogram gives us an in-depth view of a single numeric variable

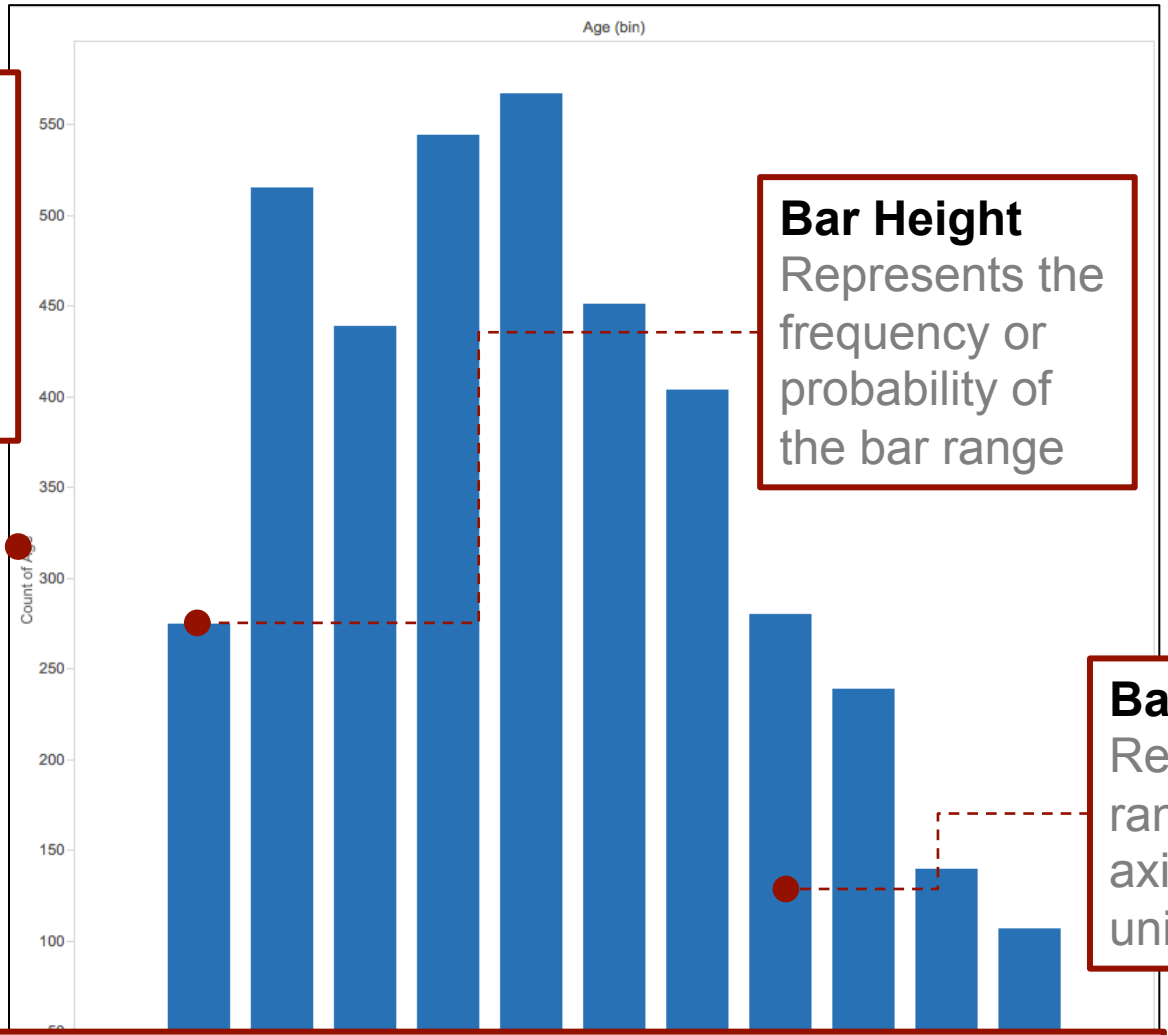
The histogram is one of your most important visual data exploration tools

To construct a histogram:

- Divide the data range into bins
- Count the occurrence frequency of each bin within the data
- Normalize the frequency counts
- Plot a bar graph to show the normalised count for each bin

Histogram


Count of Age
550
500
450
400
350
300
250
200
150
100
50
0
10
15
20
25
30
35
40
45
50
55
60
65
70



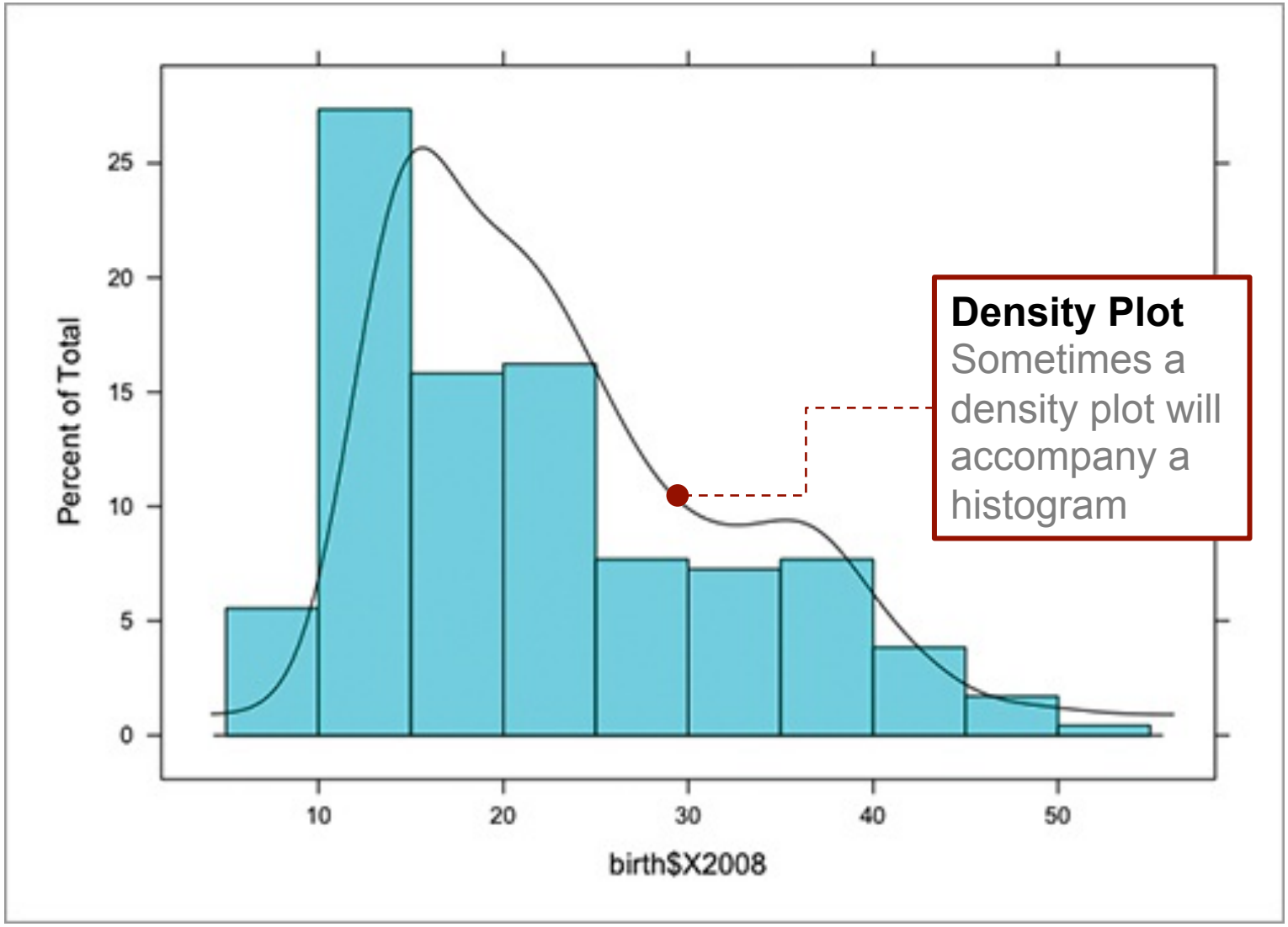
Y-Axis
Contains either probability or frequency for proportion variable

Bar Height
Represents the frequency or probability of the bar range

Bar Width
Represents an range on the x-axis, the width is uniform

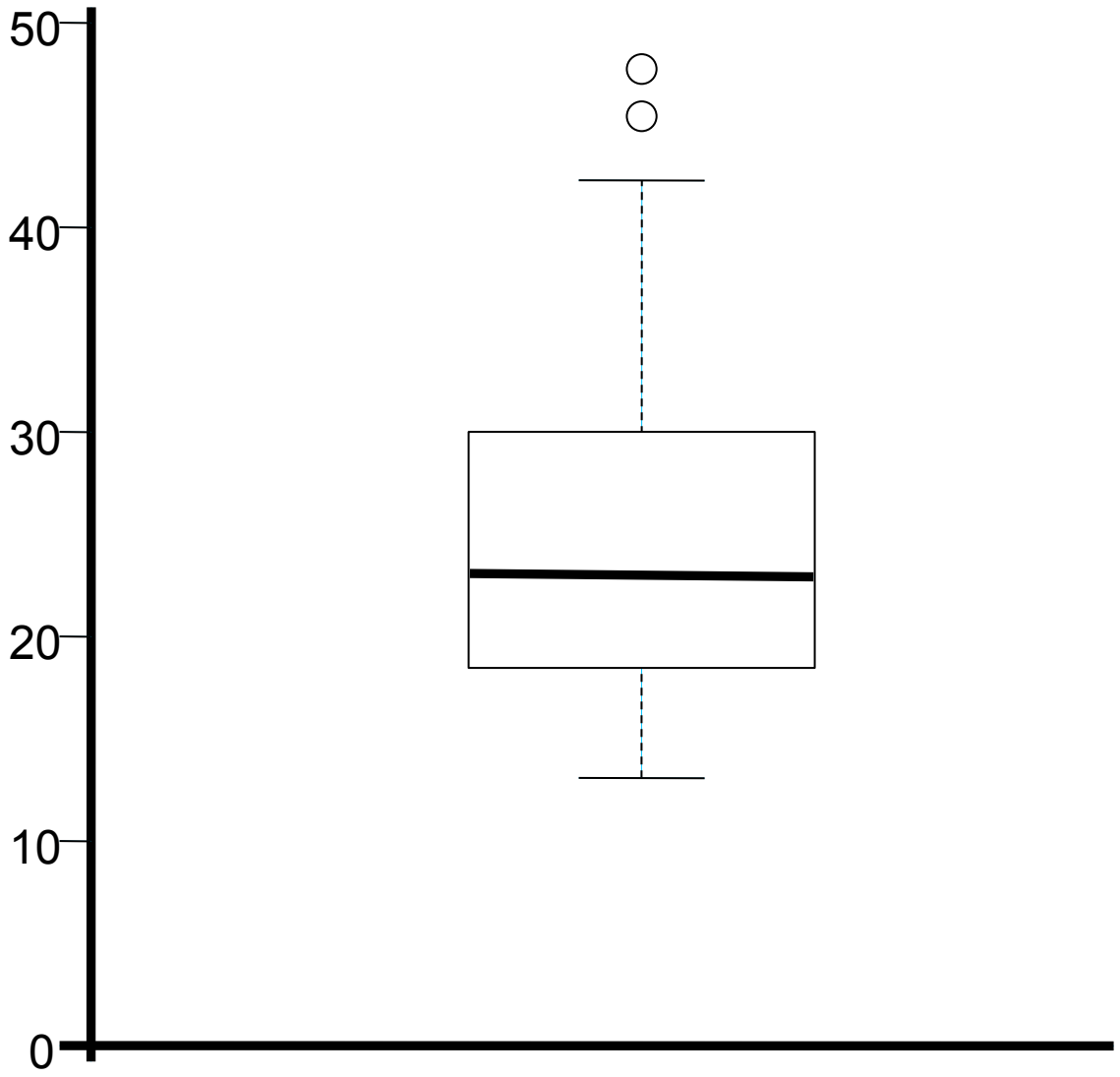
X-Axis
Contains the values for a single contiguous variable

Histogram & Density Plot Combined

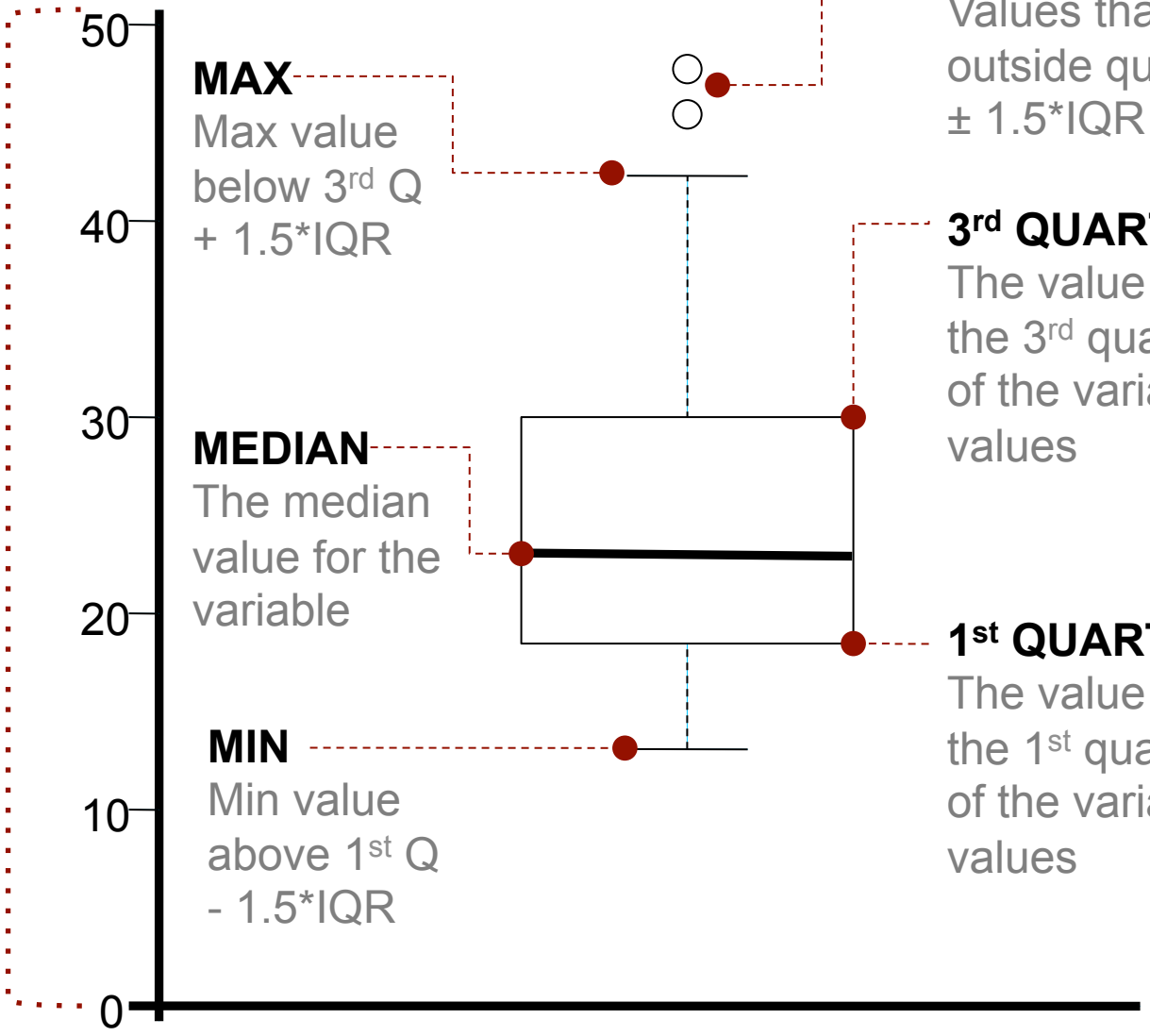


"Visualize This", N. Yau, Wiley, 2011
http://shop.oreilly.com/product/0636920022060.do

Box Plot



VARIABLE VALUES
Values displayed for a single variable



MAX
Max value below 3rd Q + 1.5*IQR

OUTLIERS
Values that fall outside quartile ± 1.5*IQR

3rd QUARTILE
The value for the 3rd quartile of the variable values

MEDIAN
The median value for the variable

1st QUARTILE
The value for the 1st quartile of the variable values

MIN
Min value above 1st Q - 1.5*IQR

Box Plot

The components of a box plot are:

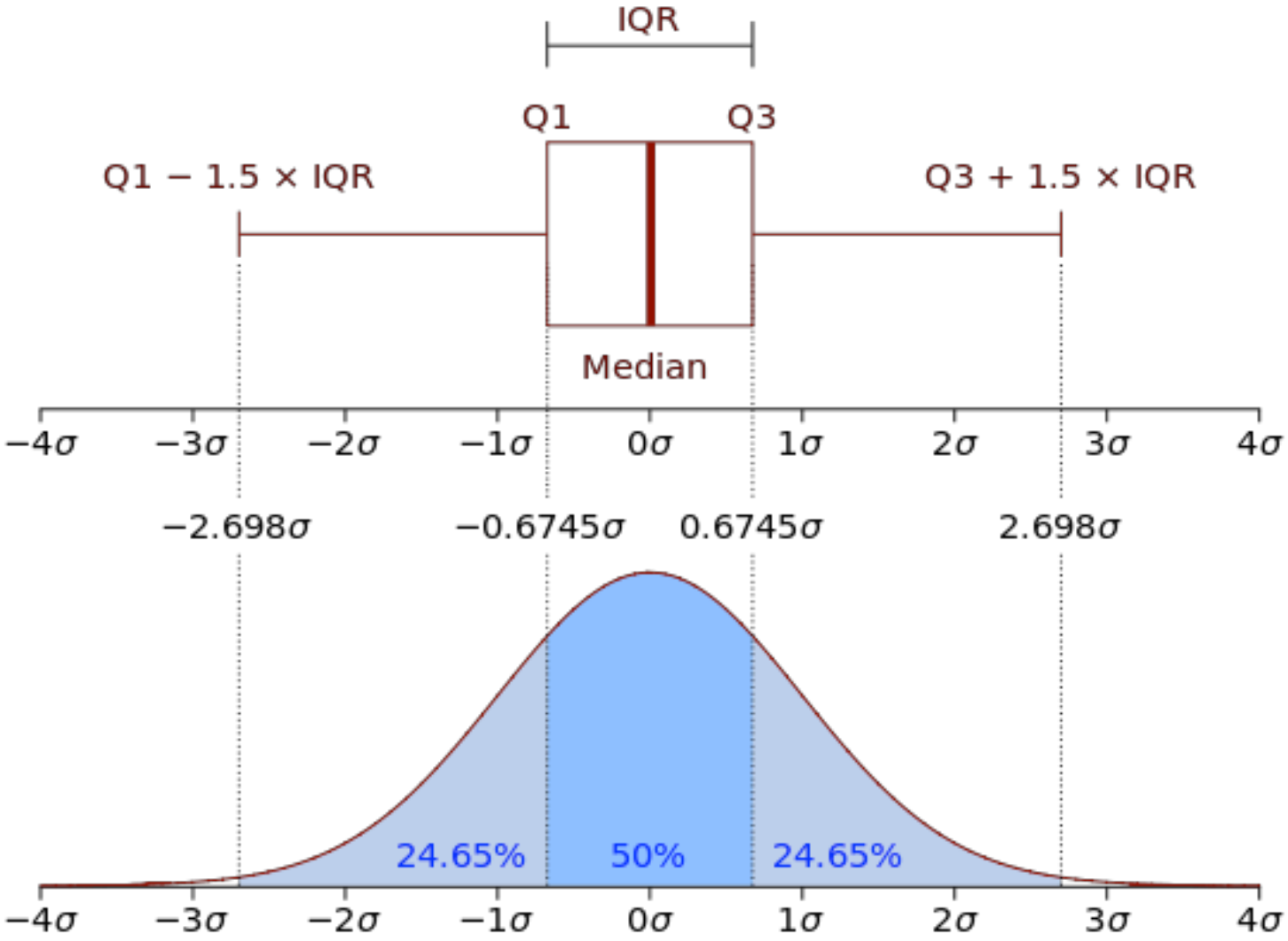
- A thick dark line at the minimum
- A horizontal lines at the 1st quartiles
- A horizontal lines at the 3rd quartiles
- A whisker down to the low value
    - Multiply the IQR by 1.5 to calculate the step
    - The low value is the lowest value above the 1st quartile minus the step
- A whisker up to the high value
    - The high value is the highest value above the 3rd quartile plus the step
- Any values outside low and high are marked as outliers

Box Plot

Some important points about a box plot:

- 50% of the data occurs between the lower and upper edges of the box
- The lower 50% of the data occurs below the median
- The upper 50% of the data occurs above the median line in the box.
- The lower 25% of the data occurs between the bottom edge of the box and the bottom edge of the lower whisker
- The upper 25% of the data occurs above the top edge of the box and the top edge of the upper whisker

Box Plots & Density Functions



From Wikipedia! http://en.wikipedia.org/wiki/Probability_density_function

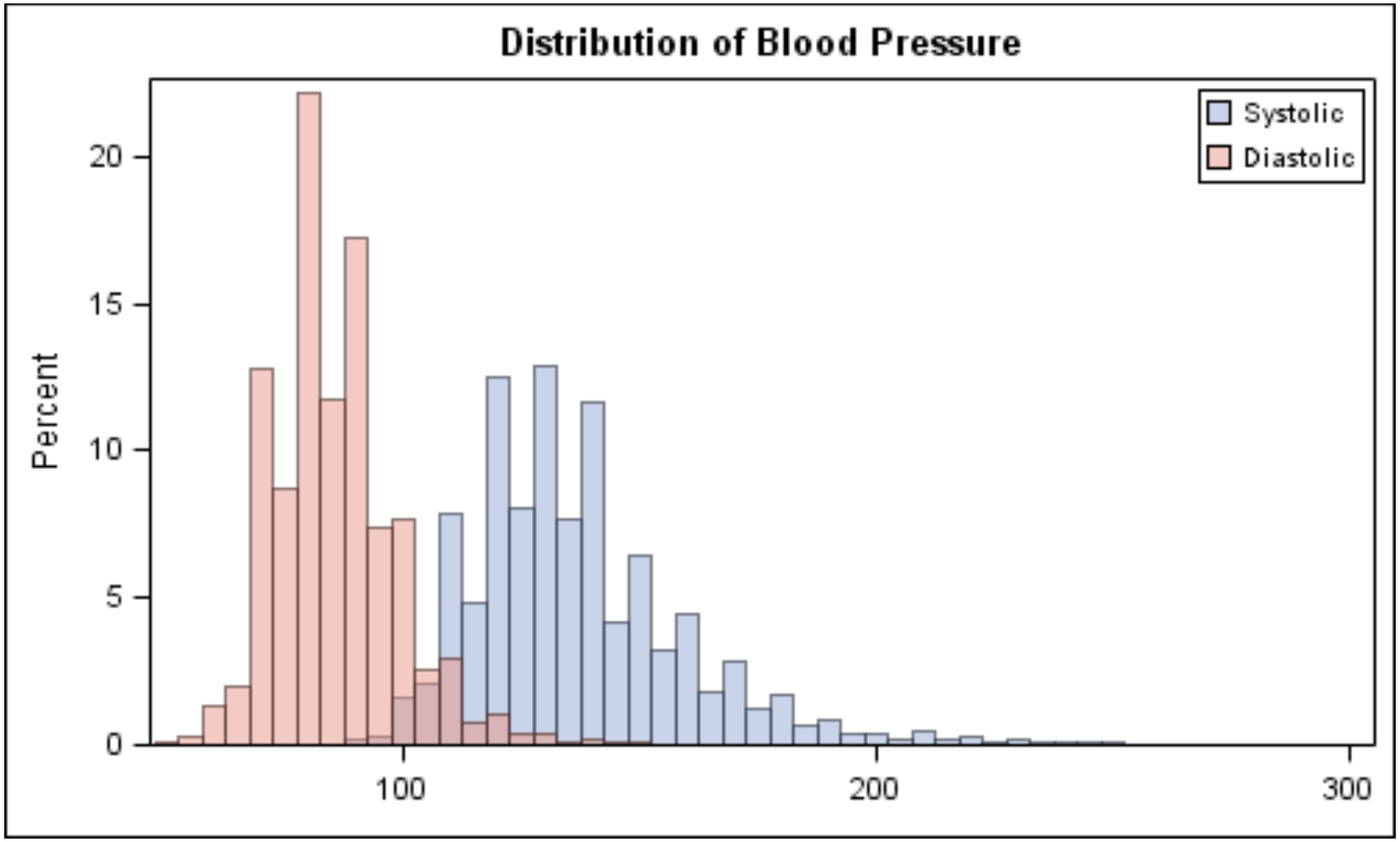
Comparing Multiple Distributions

Comparing multiple distributions is often important, but can be a little tricky – getting so much information into a chart can be difficult

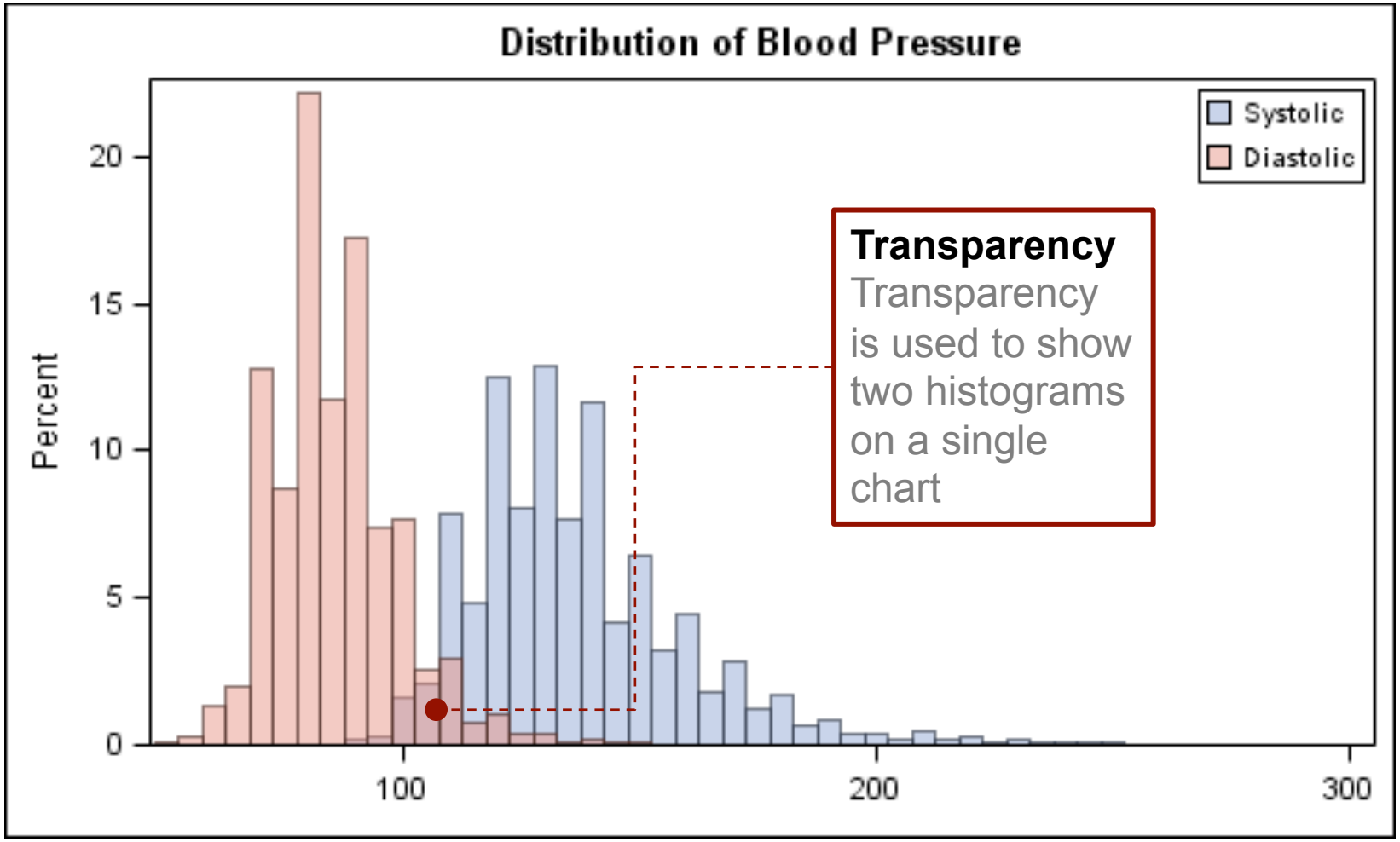
Useful techniques include

- Multiple histograms
- Multiple box plots
- Small multiples

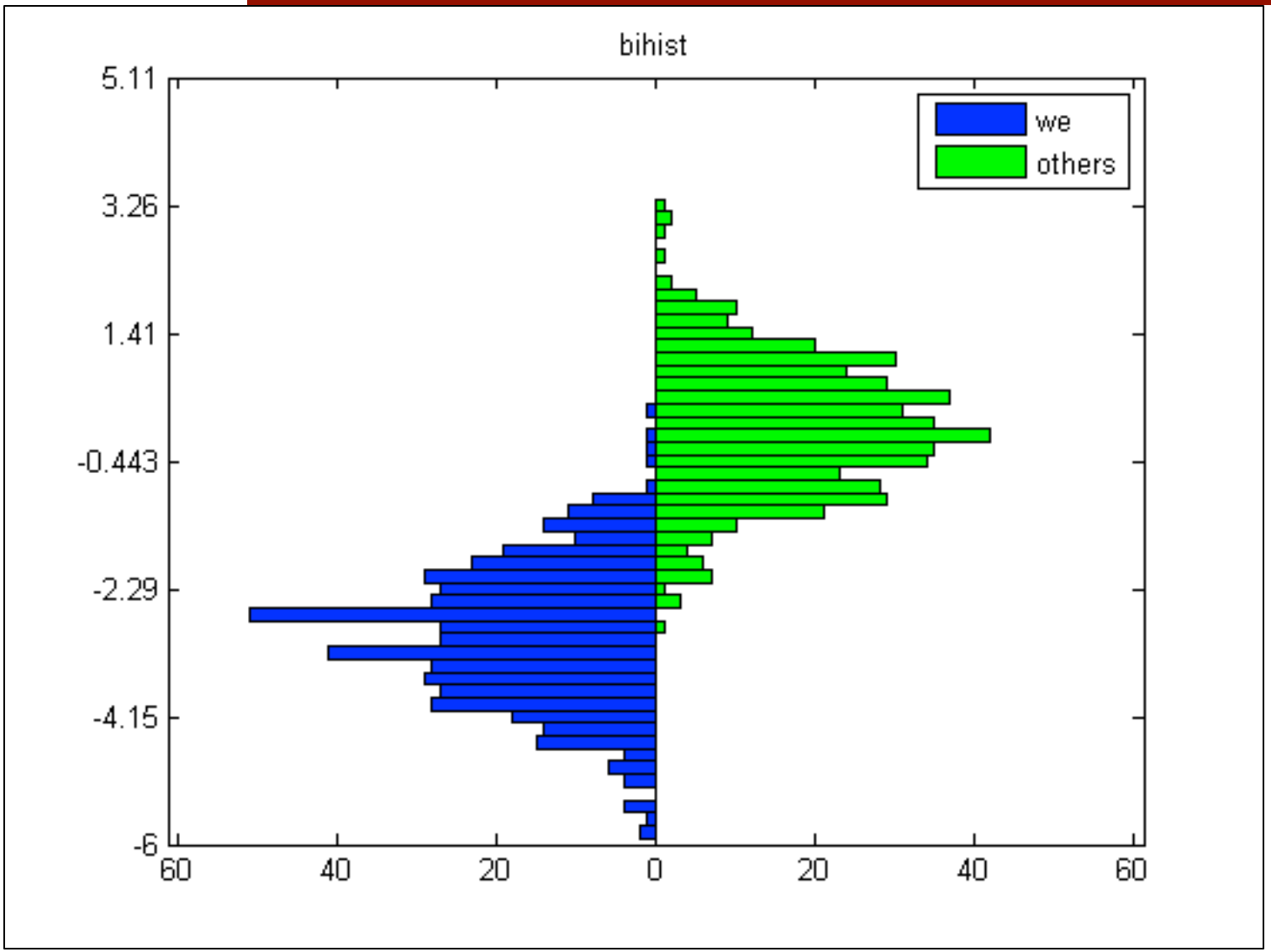
Overlaid Histograms



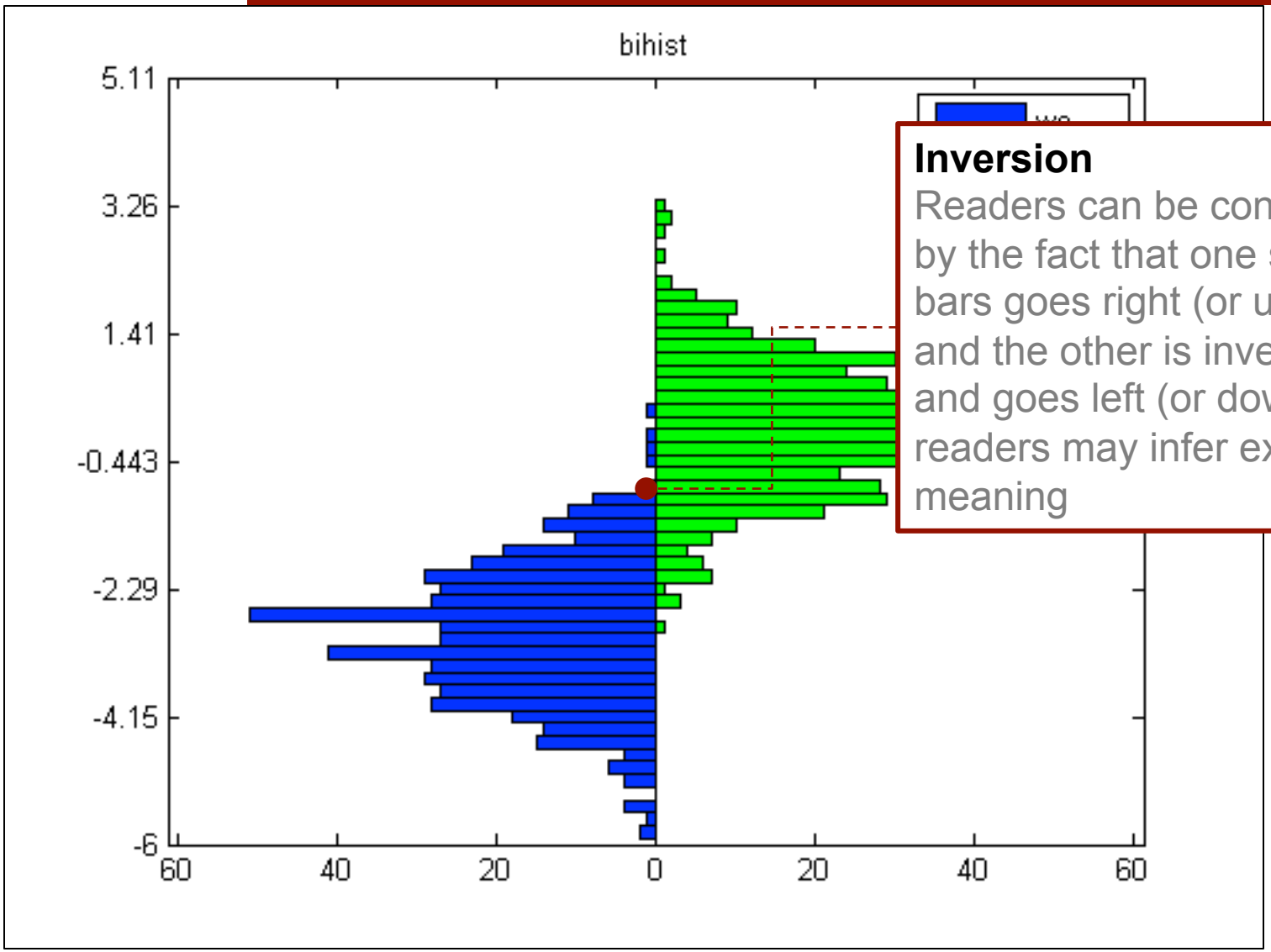
Overlaid Histograms



Back-to-Back Histograms



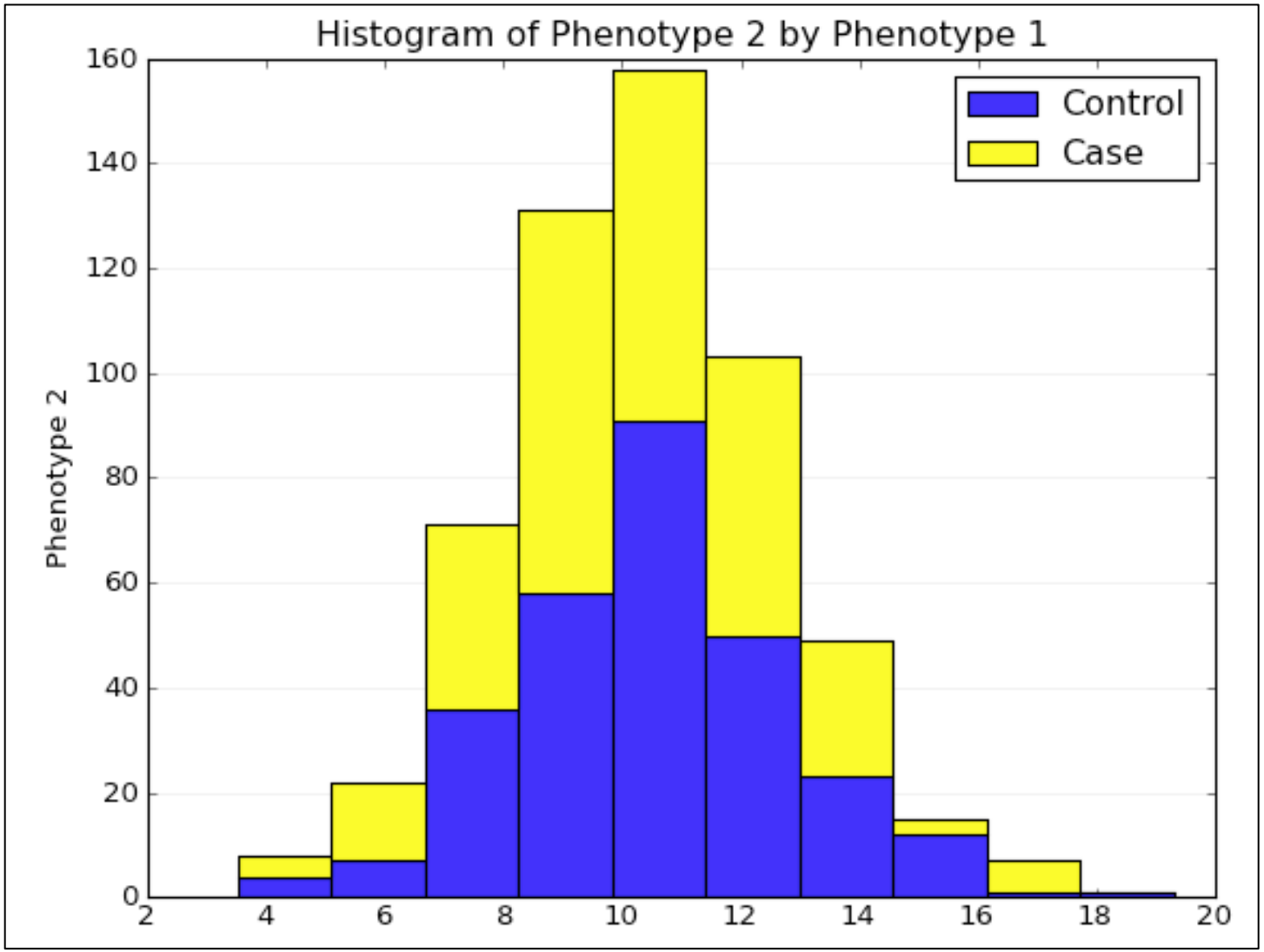
Back-to-Back Histograms

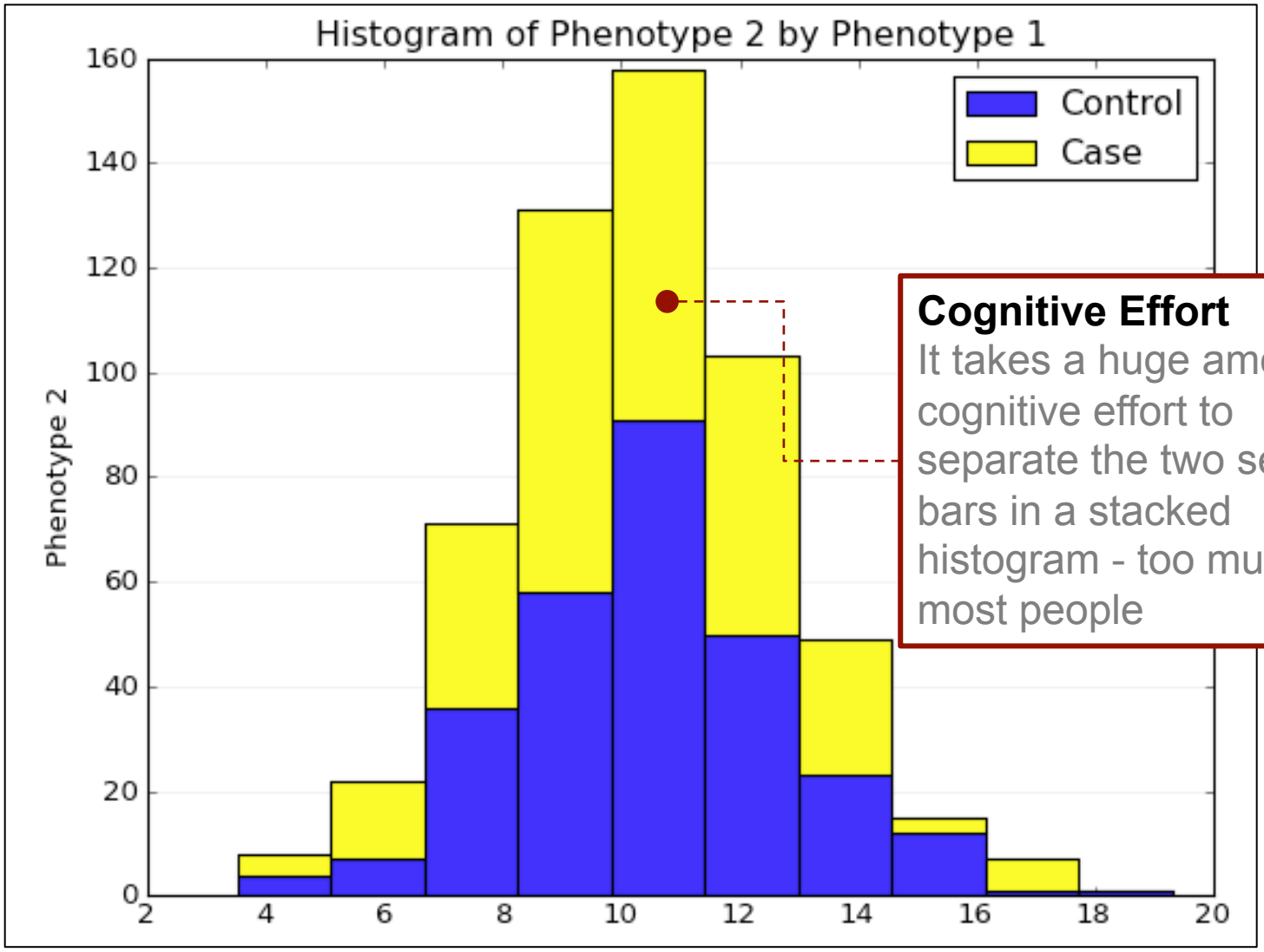


Inversion

Readers can be confused by the fact that one set of bars goes right (or up) and the other is inverted and goes left (or down) - readers may infer extra meaning

Beware Of Stacked Histograms

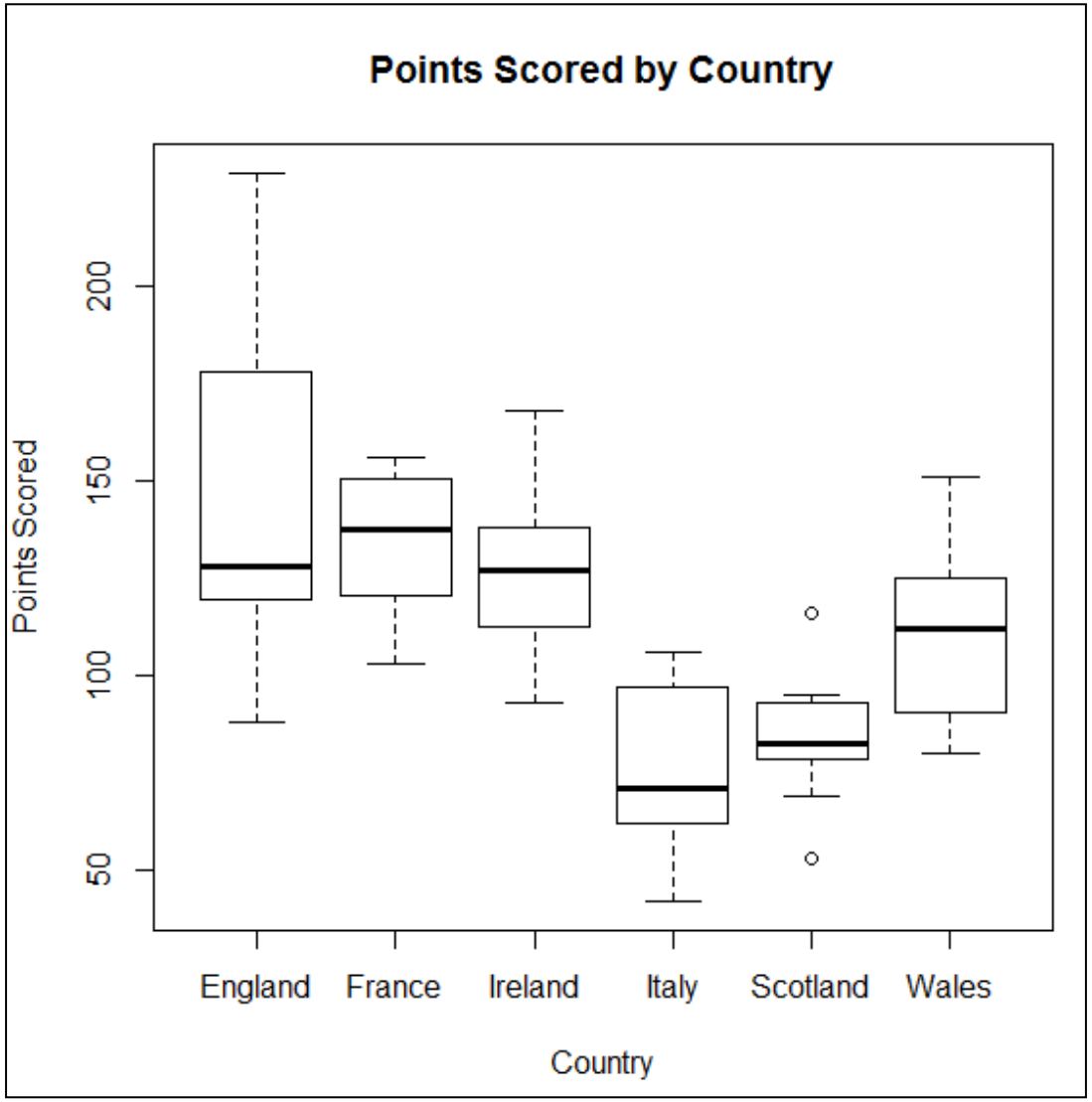




Cognitive Effort

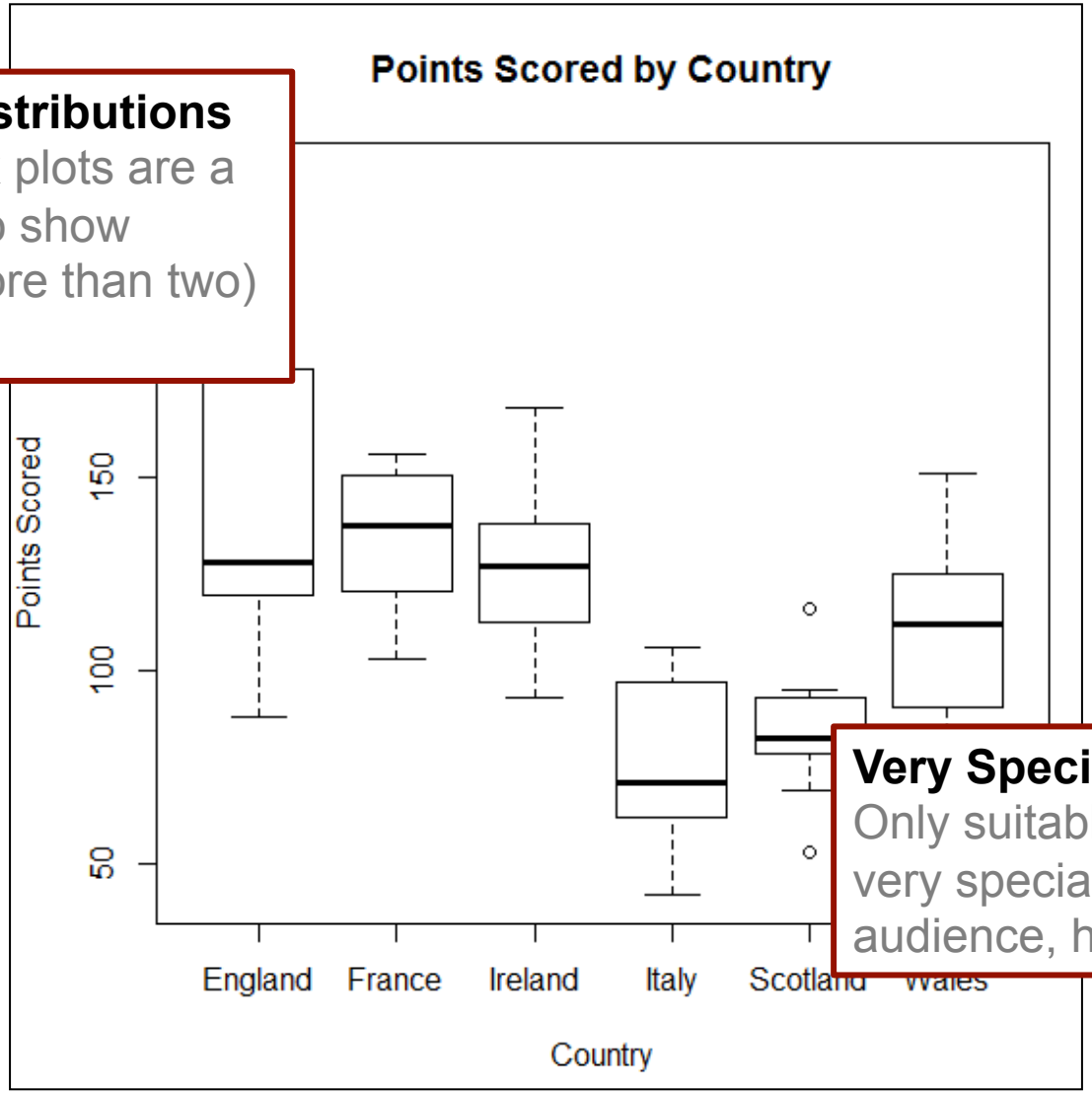
It takes a huge amount of cognitive effort to separate the two sets of bars in a stacked histogram - too much for most people

Multiple Box Plots



Multiple Box Plots

Multiple Distributions
Multiple box plots are a great way to show multiple (more than two) distributions



Very Specialised
Only suitable for a very specialised audience, however

Small Multiples

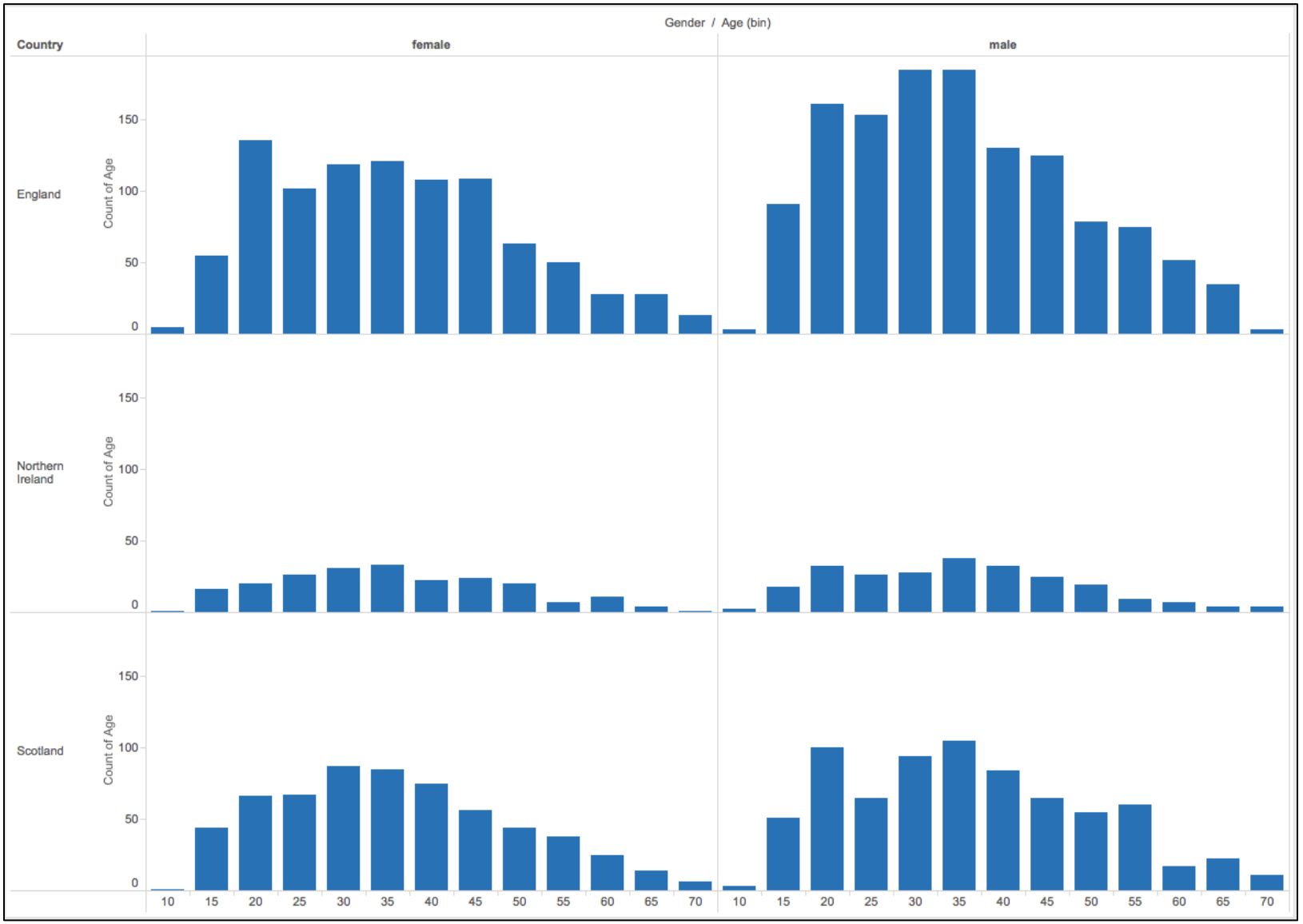
One solution to illustrating time is to use **small multiples** to show multiple snapshots of the data at different points in time

- The term was popularised by Edward Tufte

This is sometimes overlooked as people try to pack multiple information into a single chart but is a very clear way to ensure "*all of the data and just the data*" is visualised

The disadvantage of small multiples is the cognitive load placed on readers to move between multiple visualisations

Small Multiples



Consistency

Whenever small multiples are used consistency is key. Axes, colours titles etc must be kept consistent to allow readers to compare across charts

Layout

Think about how to layout the small multiple graphs. Is there a layout that will aid cognition?

Summary

We often need to create visualisations to compare values

There are a range of ways to do this

Key things to keep in mind are:

- Are you comparing values or proportions?
- Are you comparing single values or distributions?
- Are you comparing across one or many dimensions?

VISUALISING TRENDS OVER TIME

What To Look For

The most common things we look for in a time series data is trends

- Is something increasing or decreasing?
- Are there noticeable cycles?
- Are there any outliers?

To find these patterns, we have to look beyond individual data points to view the overall trend

Visualising Patterns Over Time

Visualising patterns over time

- Discrete points in time
  - Bar graph
- Continuous points in time
  - Line chart
  - Step chart

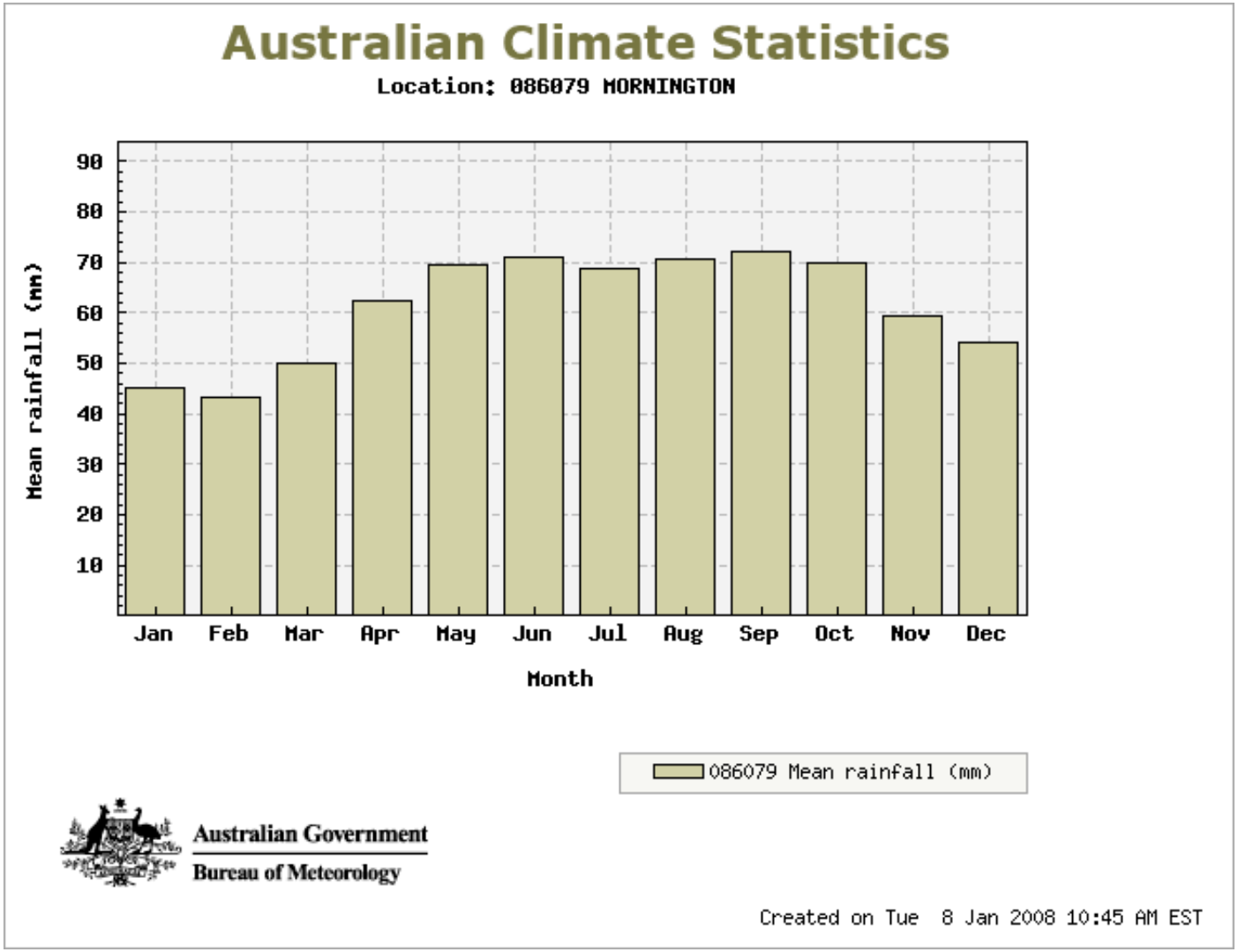
Discrete Points In Time

Often we have datasets that contain single measurements for a reasonably small number of discrete points in time

- Profit per year
- Rainfall per month

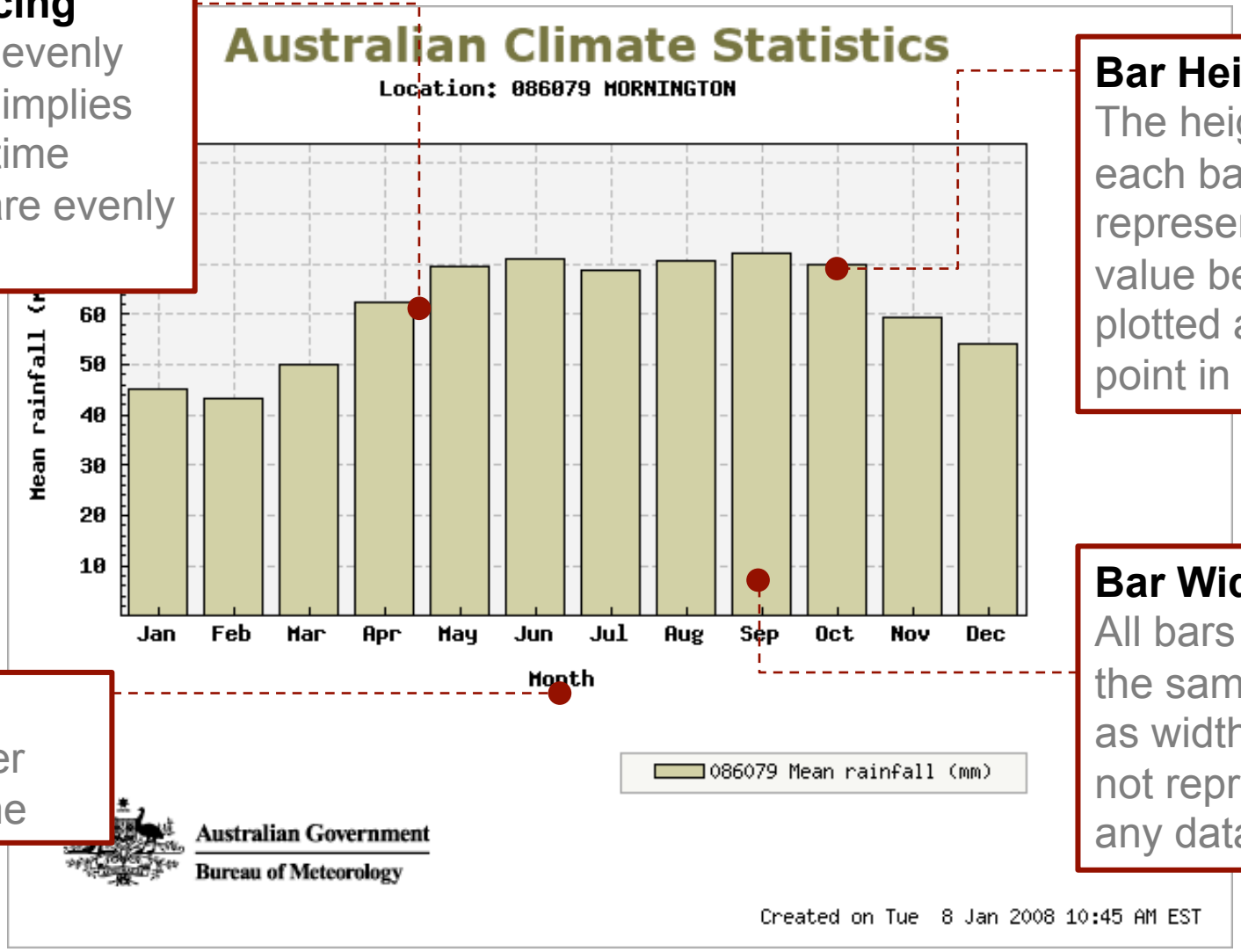
In these cases a simple bar graph is often appropriate

Simple Bar Graph



Australian Bureau of Meteorology
http://www.mornpen.vic.gov.au/page/imageThumbnail.asp?C_Id=820

Bar Spacing
Bars are evenly spaces - implies discrete time periods are evenly spaced

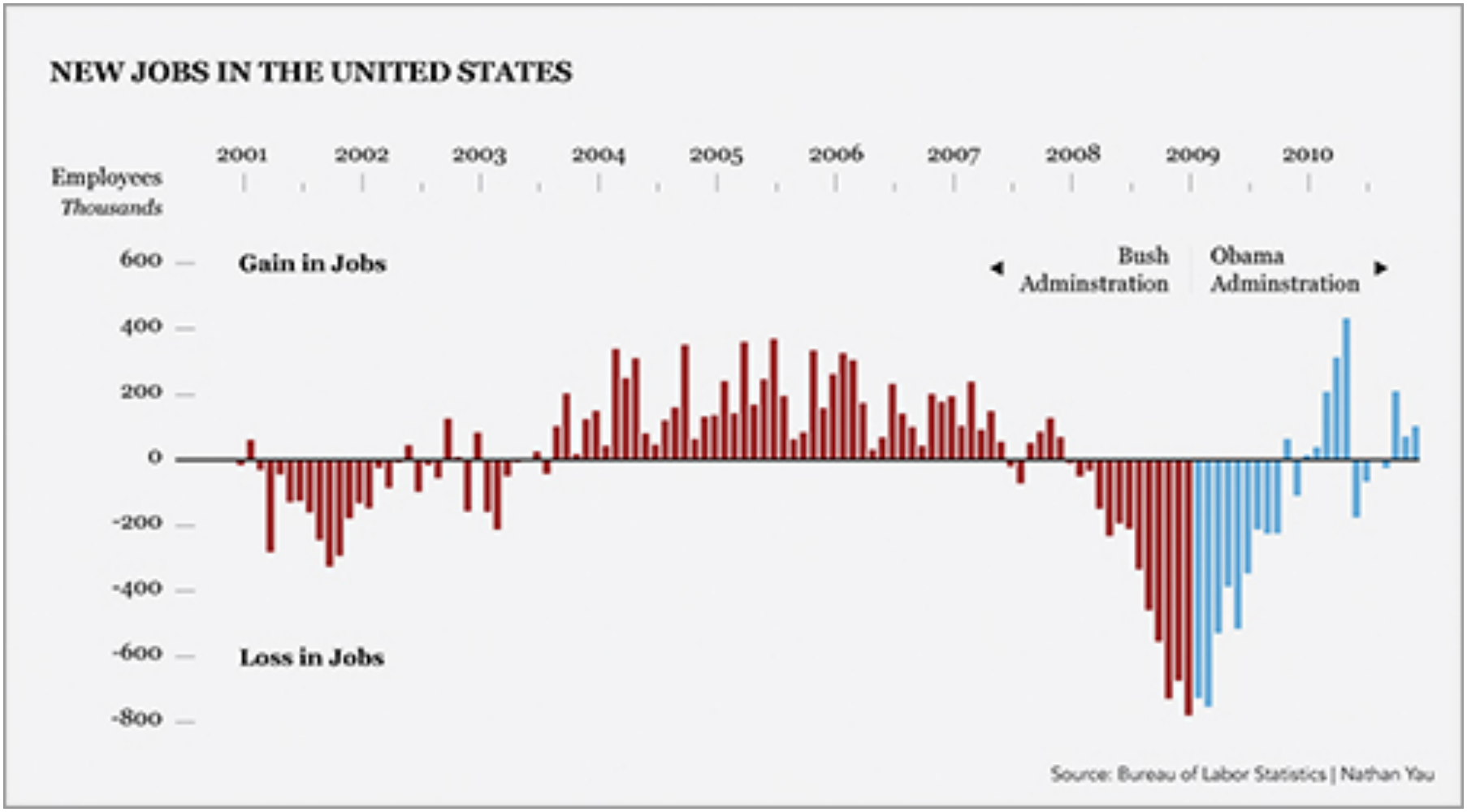


Bar Height
The height of each bar represents the value being plotted at a point in time

Bars
One bar per point in time

Bar Width
All bars have the same width as width does not represent any data

http://www.mornpen.vic.gov.au/page/imageThumbnail.asp?C_Id=820



"Visualize This", N. Yau, Wiley, 2011

http://shop.oreilly.com/product/0636920022060.do

Continuous Points In Time

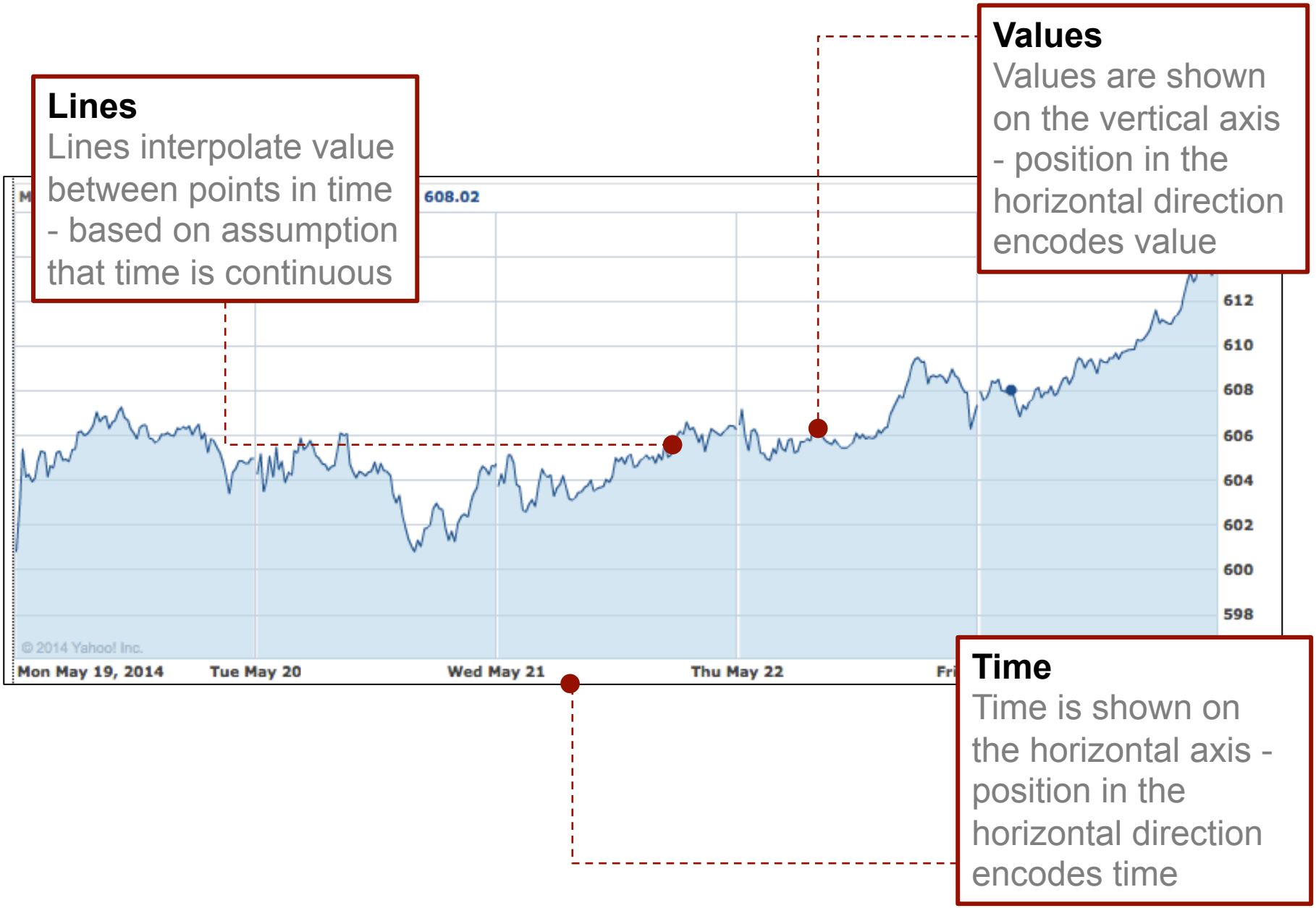
Often we have datasets that contain single measurements for a number of continuous points of time

- Stock prices over time
- Internet traffic over time

In these cases a line charts are most appropriate

Line Charts

Apple (AAPL) stock price chart from Yahoo! Finance finance.yahoo.com



Apple (AAPL) stock price chart from Yahoo! Finance finance.yahoo.com

Mixing Discrete And Continuous Time

Apple (AAPL) stock price chart from Yahoo! Finance finance.yahoo.com

Continuous
Stock prices are measured in continuous time and so prices are interpolated using a line chart

Discrete
Volume is measured in discrete intervals - per day and so a bar chart is most appropriate

Apple (AAPL) stock price chart from Yahoo! Finance finance.yahoo.com

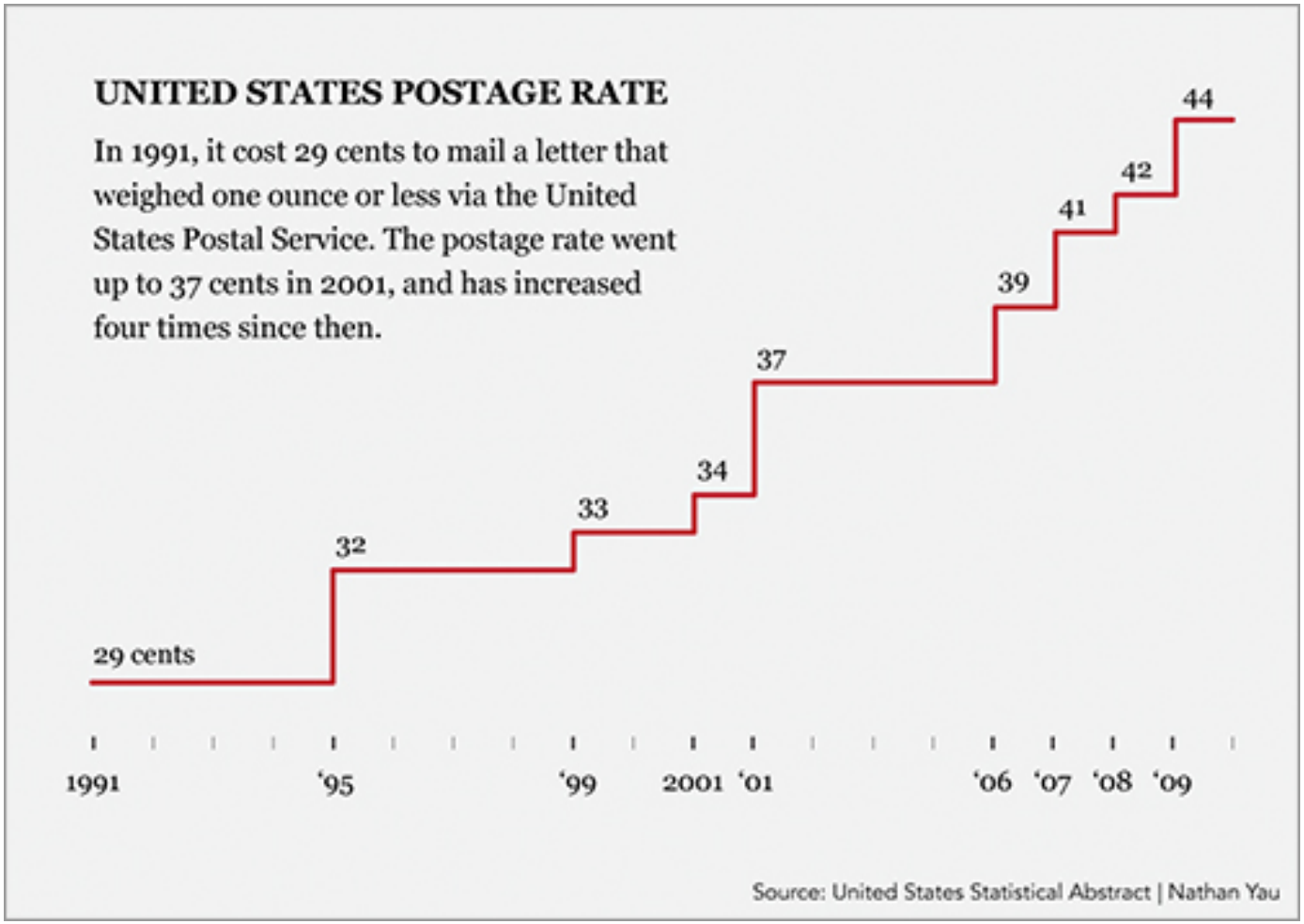
Step Chart

Basic line charts interpolate the change in measurement from one point in time to another

Sometimes this is not appropriate as values remain constant in between measurements

In these cases step charts are more appropriate

Step Chart



"Visualize This", N. Yau, Wiley, 2011

http://shop.oreilly.com/product/0636920022060.do

Visualising Multiple Trends Over Time

Can be a little tricky

There are some fairly standard ways to visualising more than one trend over time

- Multiple lines
- Small multiples
- Stacked bar/area charts
- Stream graph
- Animation

Multiple Lines

Adding multiple lines to a line chart is by far the easiest way to show a number of different trends over time

The lines should be distinguished by some other encoding - typically colour or pattern

Multiple Lines

Apple (AAPL), Google (GOOG) and Hewlett Packard (HPQ) stock price chart from Yahoo! Finance
finance.yahoo.com

Small Multiples

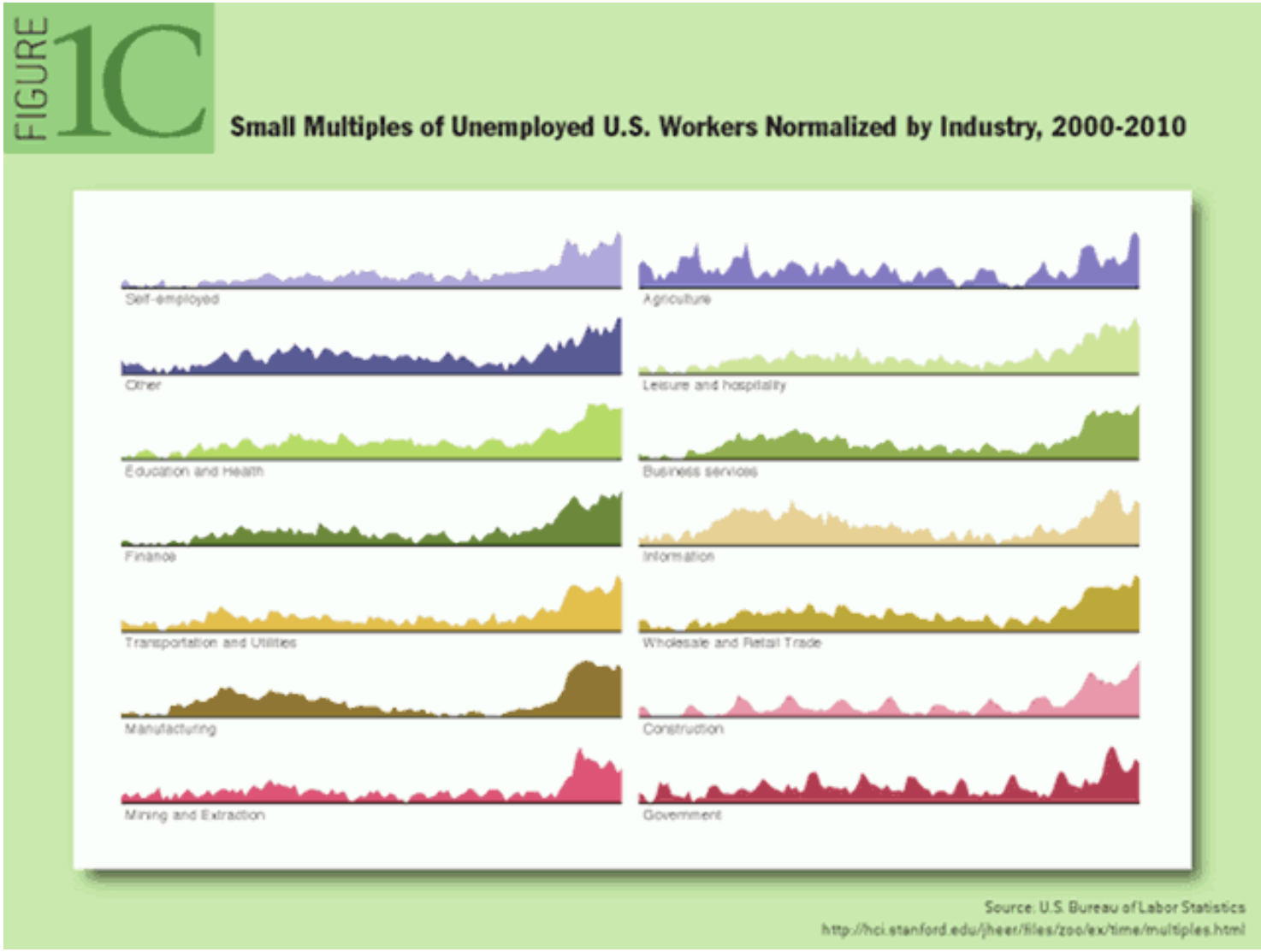
One solution to illustrating time is to use small multiples to show multiple snapshots of the data at different points in time

- The term was popularised by Edward Tufte

This is sometimes overlooked as people try to pack multiple information into a single chart but is a very clear way to ensure “all of the data and just the data” is visualised

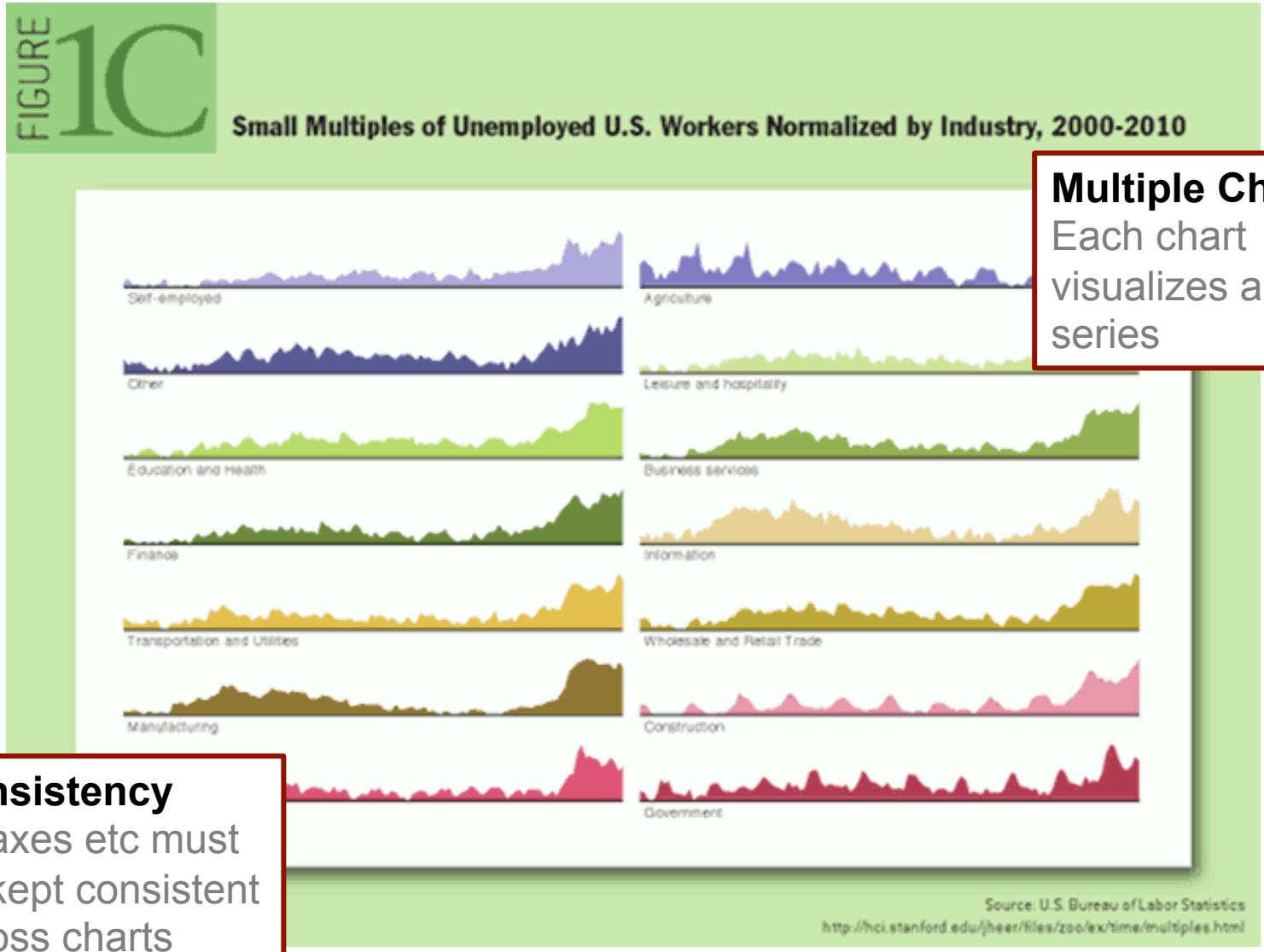
The disadvantage of small multiples is the cognitive load placed on readers to move between multiple visualisations

Small Multiples



"A Tour Through the Visualization Zoo", Jeffrey Heer, Michael Bostock & Vadim Ogievetsky, Queue, Volume 8 Issue 5, 2010

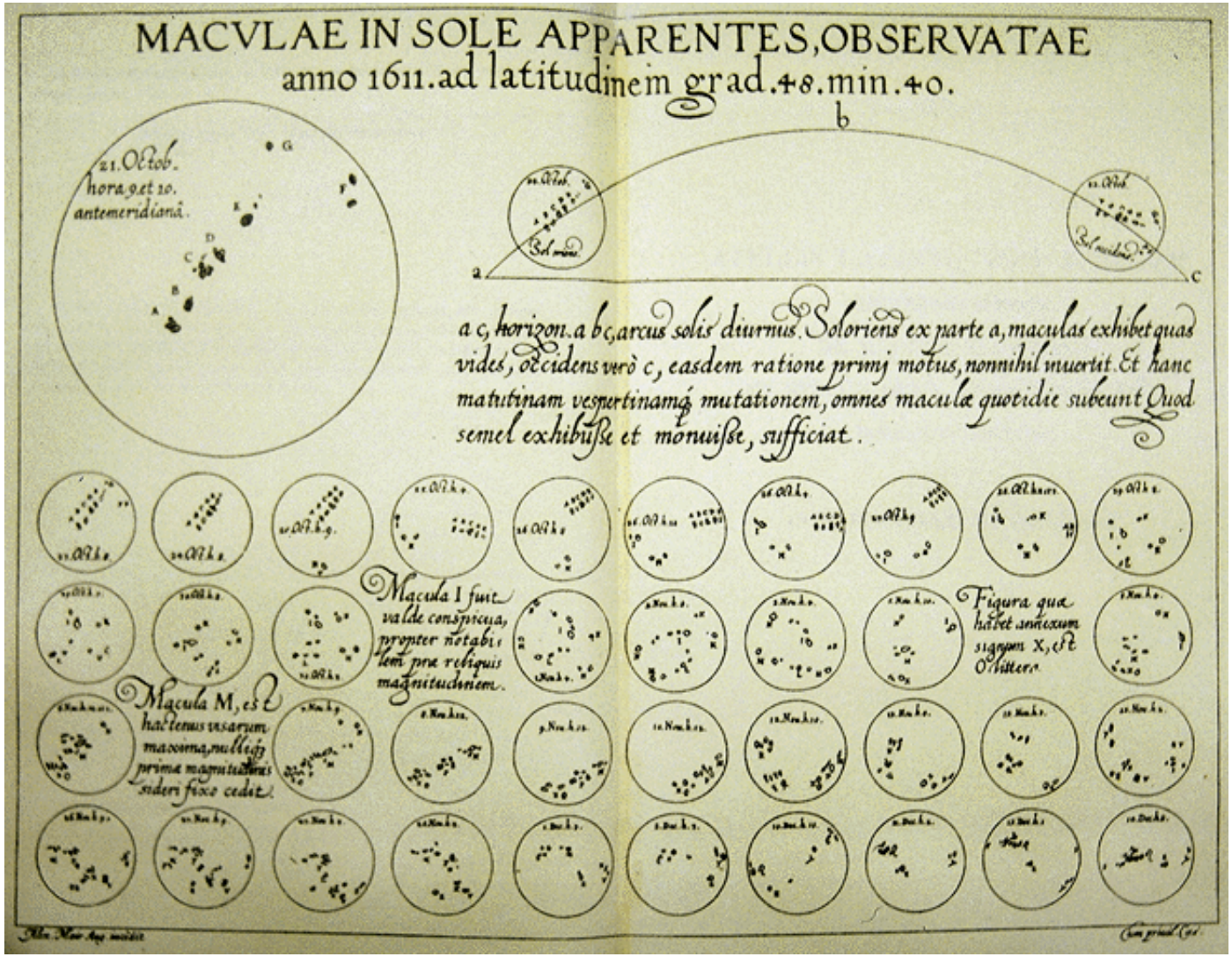
http://queue.acm.org/detail.cfm?id=1805128



Multiple Charts
Each chart visualizes a single series

Consistency
All axes etc must be kept consistent across charts

Small Multiples



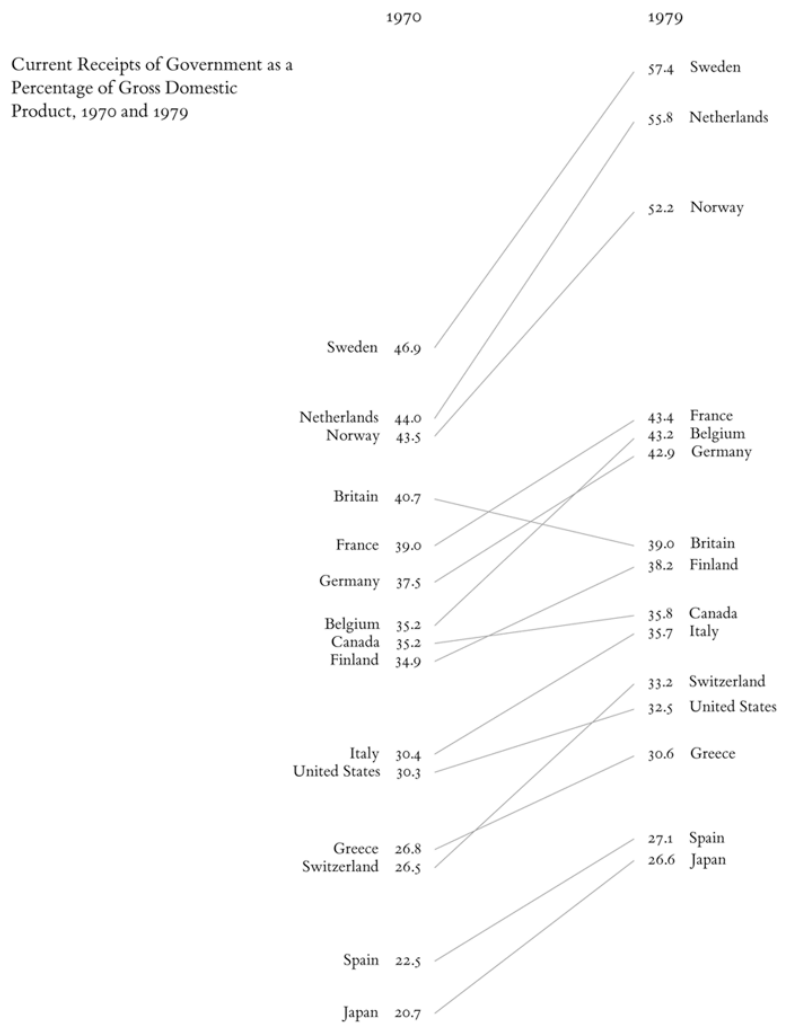
Tres Epistolae de Maculis Solaribus Scriptae ad Marcum Welserum, Christoph Scheiner, 1612

http://galileo.rice.edu/sci/observations/sunspots.html

Slope Graphs

Useful for making comparison across few (usually 2) points in time

Esepcially useful when a ranking is important



Charlie Park on SlopeGraphs http://charliepark.org/slopegraphs/ http://charliepark.org/a-slopegraph-update/

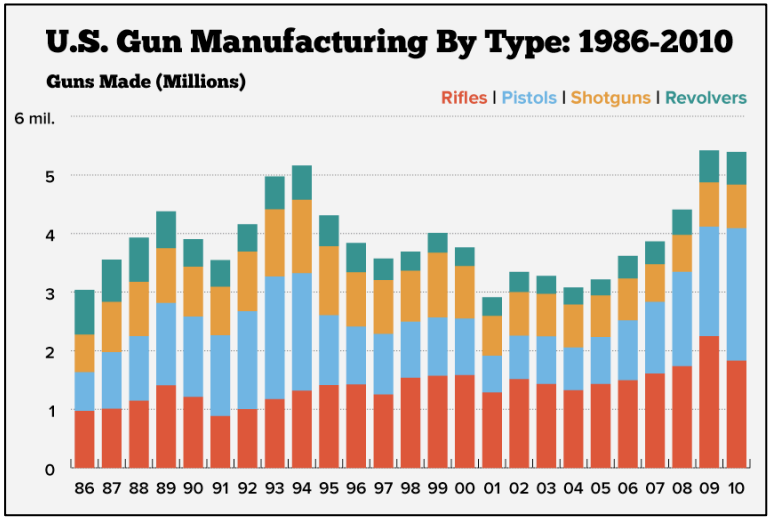
The Visual Display of Quantitative Information, Edward Tufte, 1983 http://www.amazon.com/gp/product/0961392142

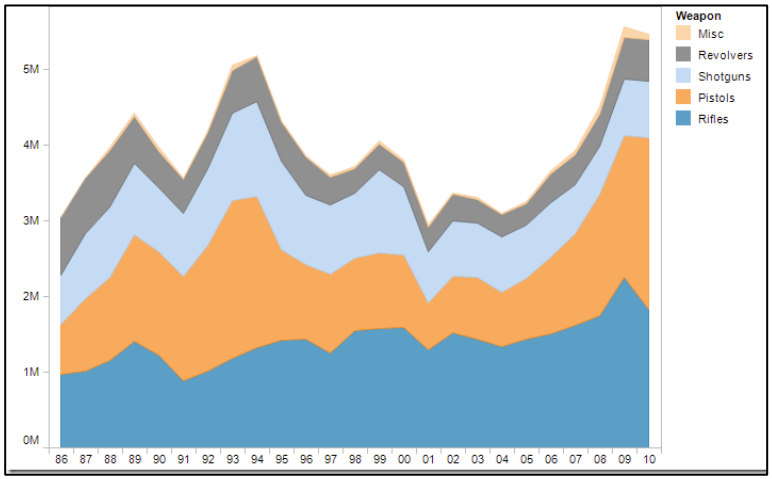
Current Receipts of Government as a Percentage of Gross Domestic Product, 1970 and 1979

Stacked Bar/Area Charts

People often depict multiple time series in stacked area or bar charts

Interpreting multiple stacked time series can be really difficult - the upper series are totally distorted by the lower series

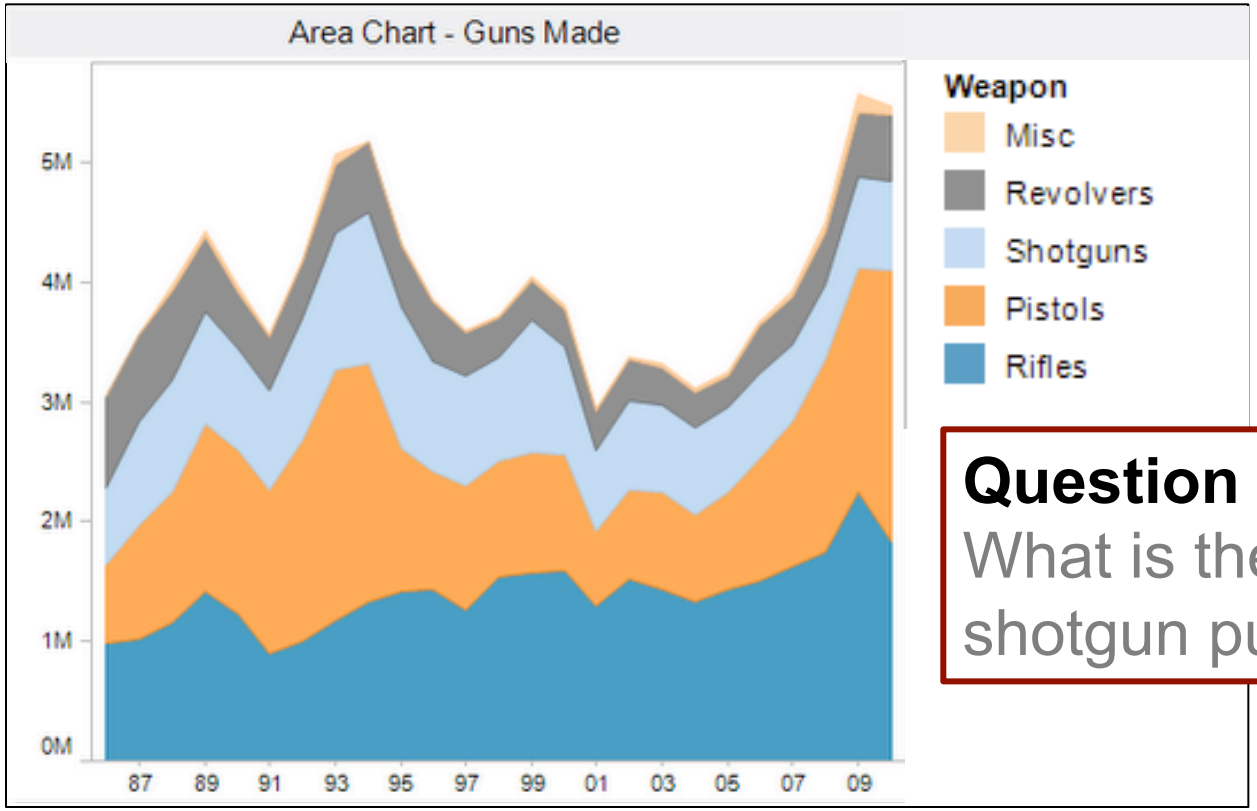




Displaying time-series data: Stacked bars, area charts or lines…you decide!, VizWiz, 2012
http://vizwiz.blogspot.ie/2012/08/displaying-time-series-data-stacked.html

Charting U.S. Gun Manufacturing, Matt Stiles, The Daily Viz, 2012
http://thedailyviz.com/2012/08/07/charting-u-s-gun-manufacturing/

Stacked Area Chart

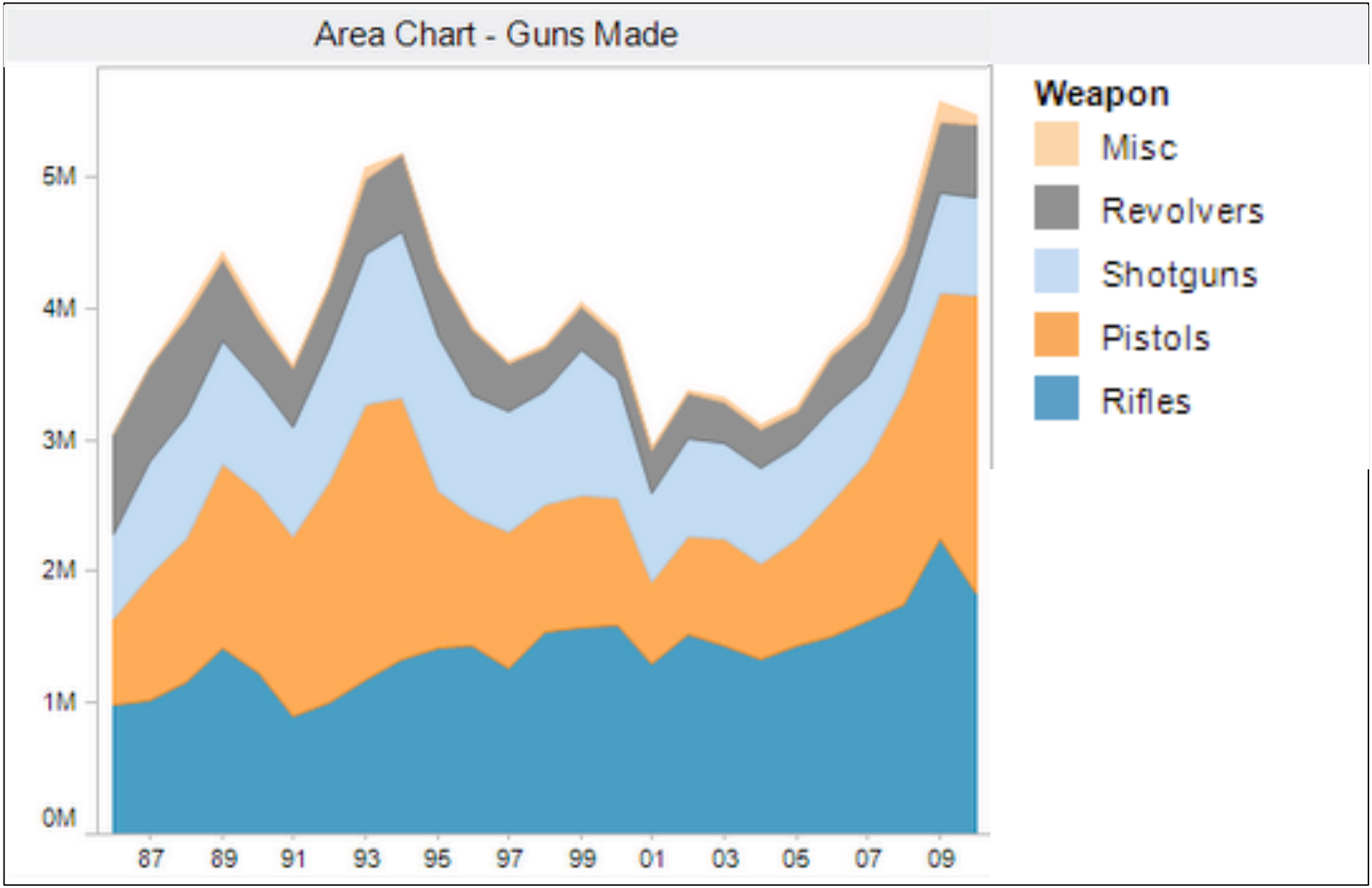


Question

What is the trend in shotgun purchases?

Displaying time-series data: Stacked bars, area charts or lines…you decide!, VizWiz, 2012

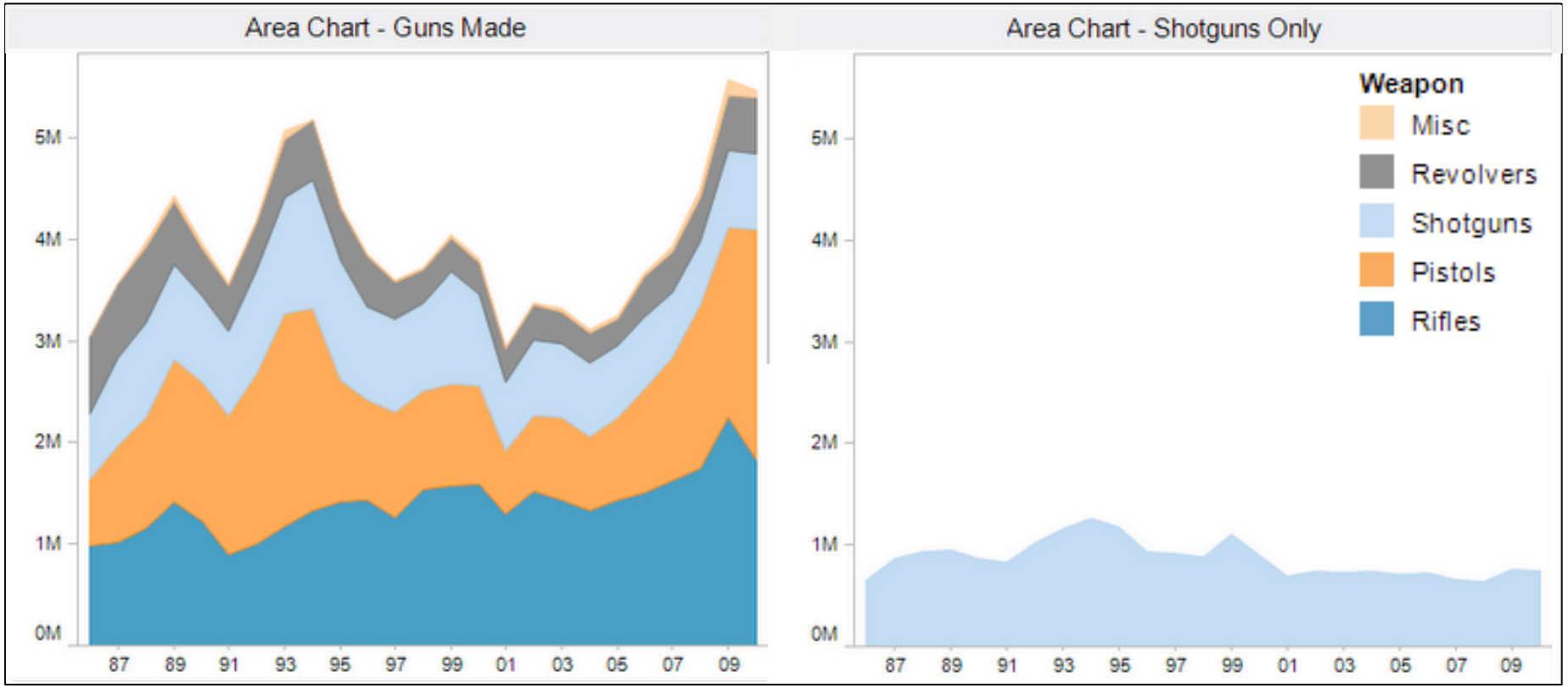
http://vizwiz.blogspot.ie/2012/08/displaying-time-series-data-stacked.html



Displaying time-series data: Stacked bars, area charts or lines…you decide!, VizWiz, 2012

http://vizwiz.blogspot.ie/2012/08/displaying-time-series-data-stacked.html

Stacked Area Chart

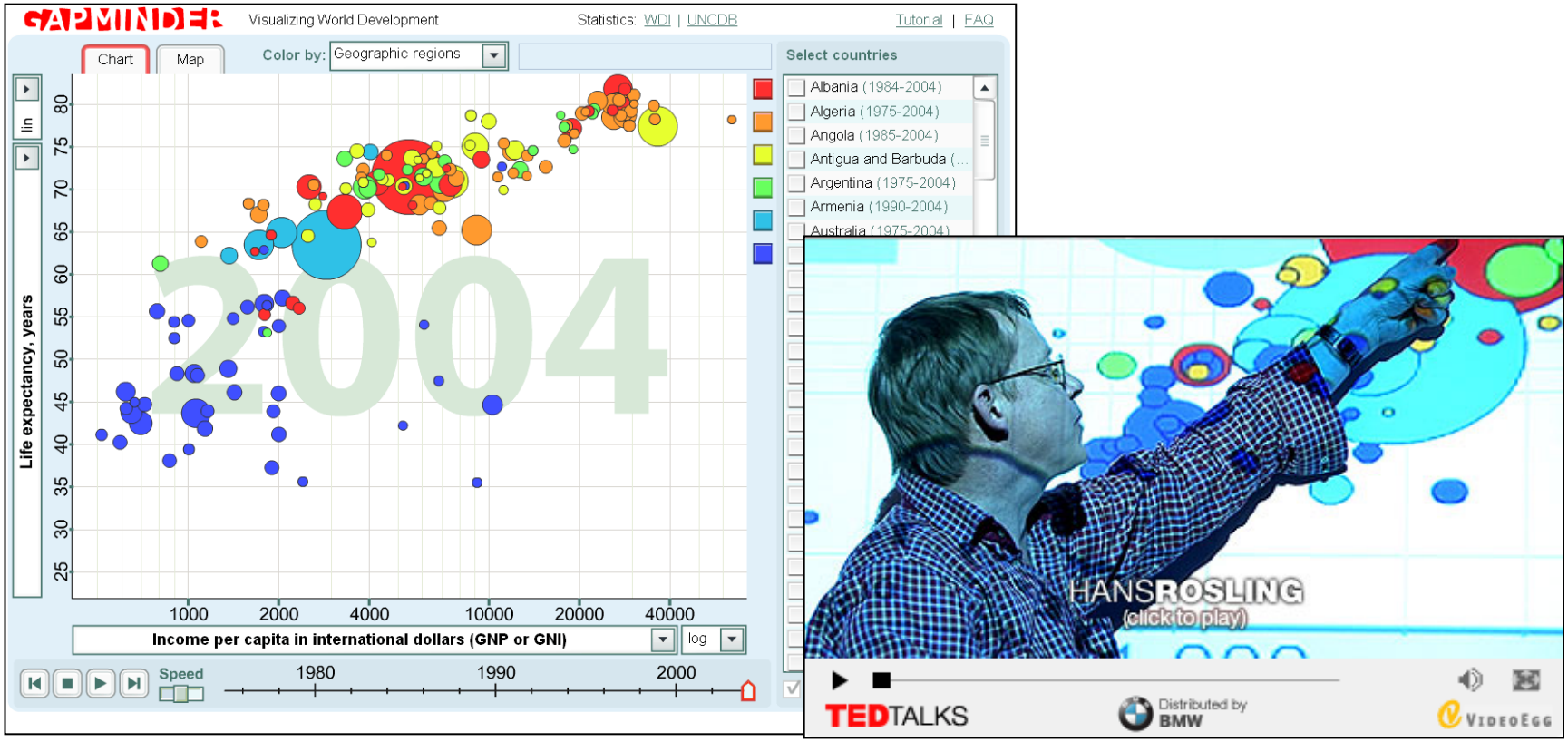


Displaying time-series data: Stacked bars, area charts or lines…you decide!, VizWiz, 2012

http://vizwiz.blogspot.ie/2012/08/displaying-time-series-data-stacked.html

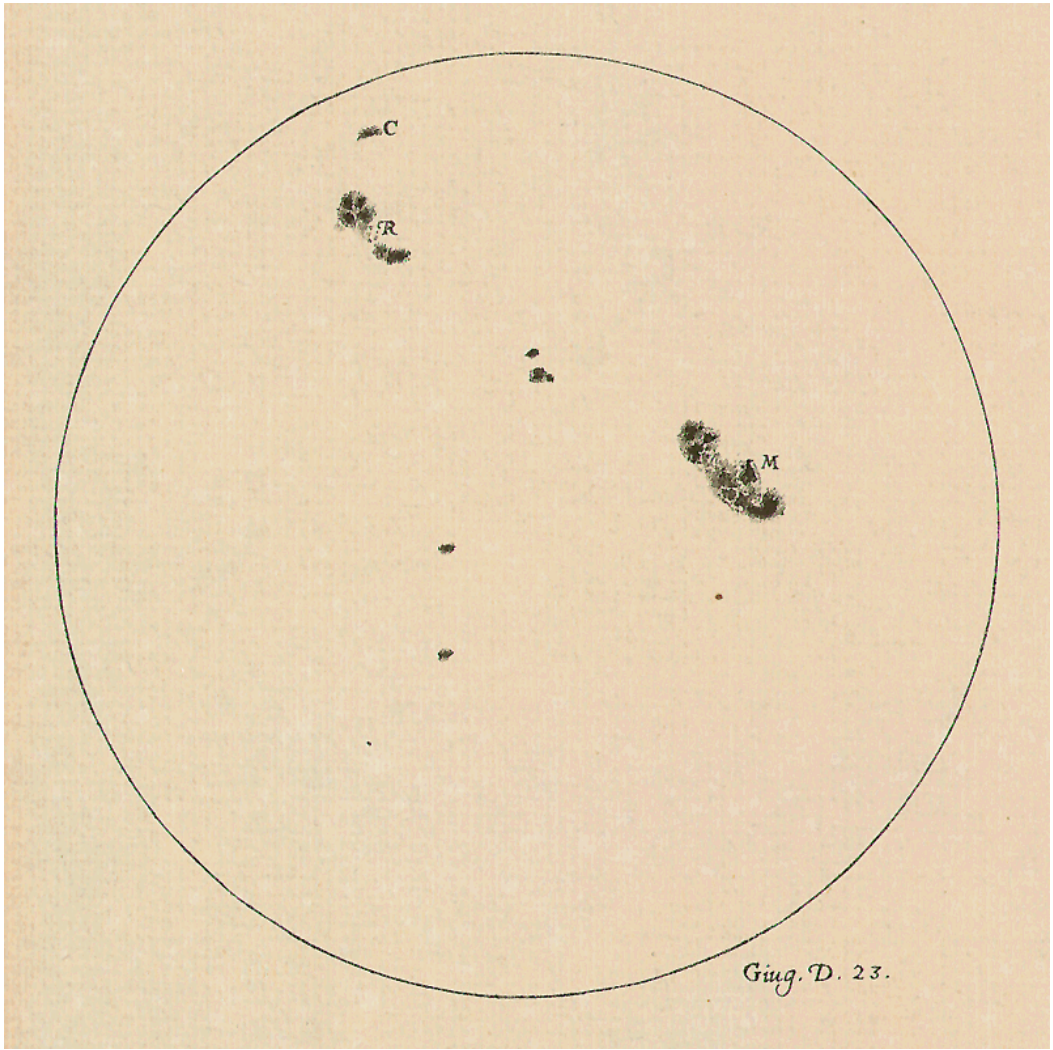
Animation

Animation is obviously a great way to illustrate changes over time - for on-screen delivery



www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

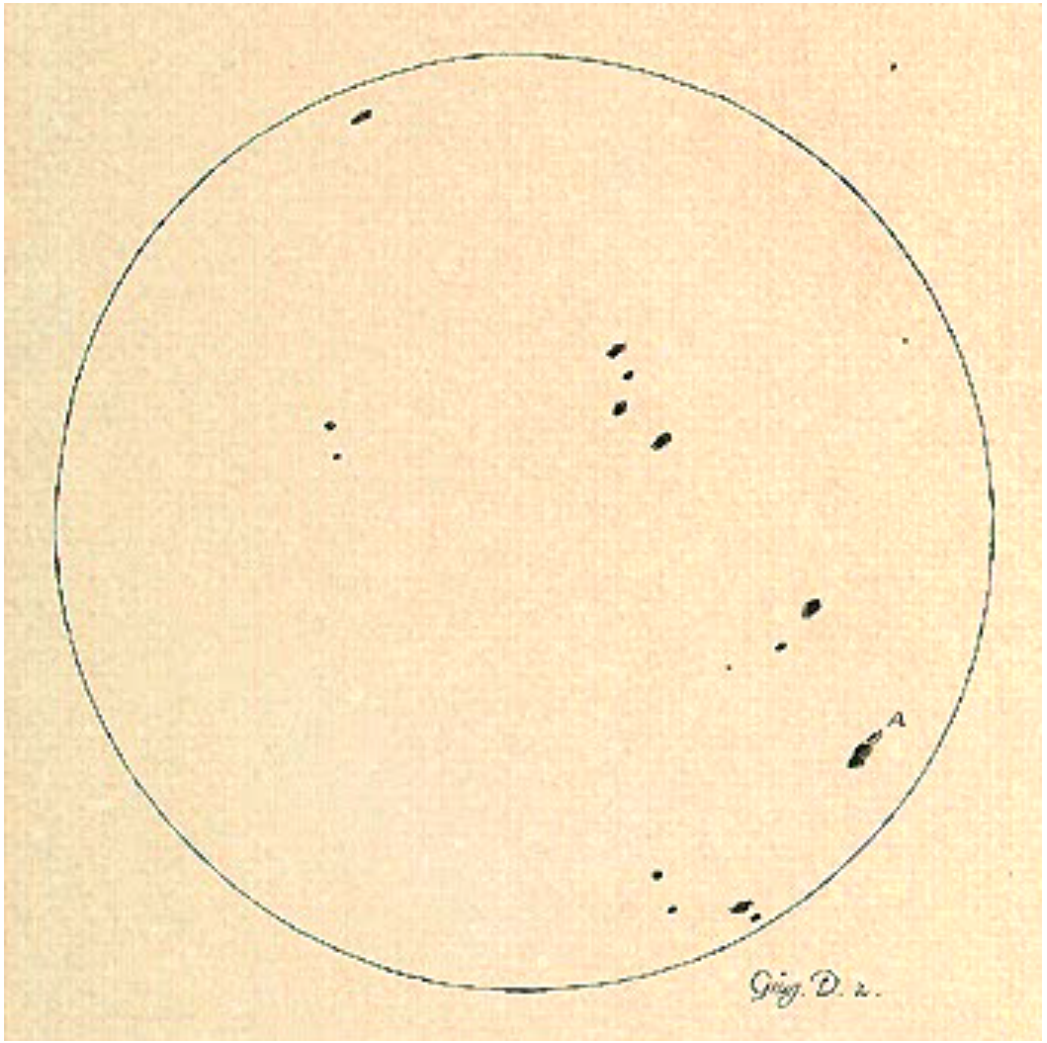
Animation



Galileo made *flip books* of sunspot images to show how they moved over time

Galileo's Sunspot Drawings
http://galileo.rice.edu/sci/observations/ssm_slow.mpg

Animation



Galileo made *flip books* of sunspot images to show how they moved over time

Galileo's Sunspot Drawings
http://galileo.rice.edu/sci/observations/ssm_slow.mpg

Summary

Visualising across time is a key methodology

There are a number of approaches that are common and useful

- Discrete points in time
- Continuous points in time
- Multiple times series

VISUALISING RELATIONSHIPS

Visualising Relationships

It is common, important and useful to visualise relationships between different variables

- As something goes up, does another thing go down?

There are a number of simple techniques to allow us visualise relationships:

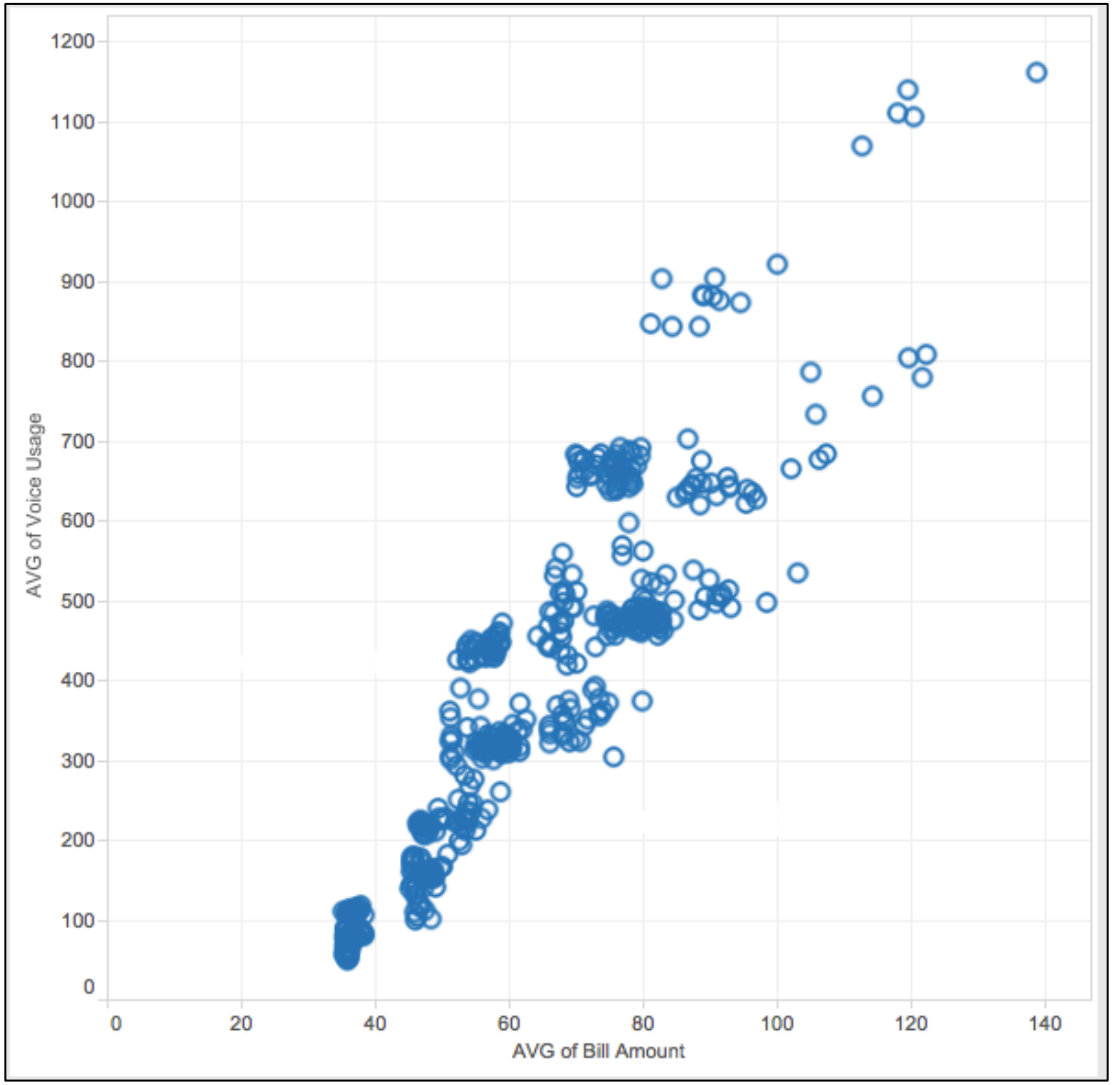
- Scatter plots
- Scatter plot matrix
- Bubble plots (channelling Hans)

Finding Correlation

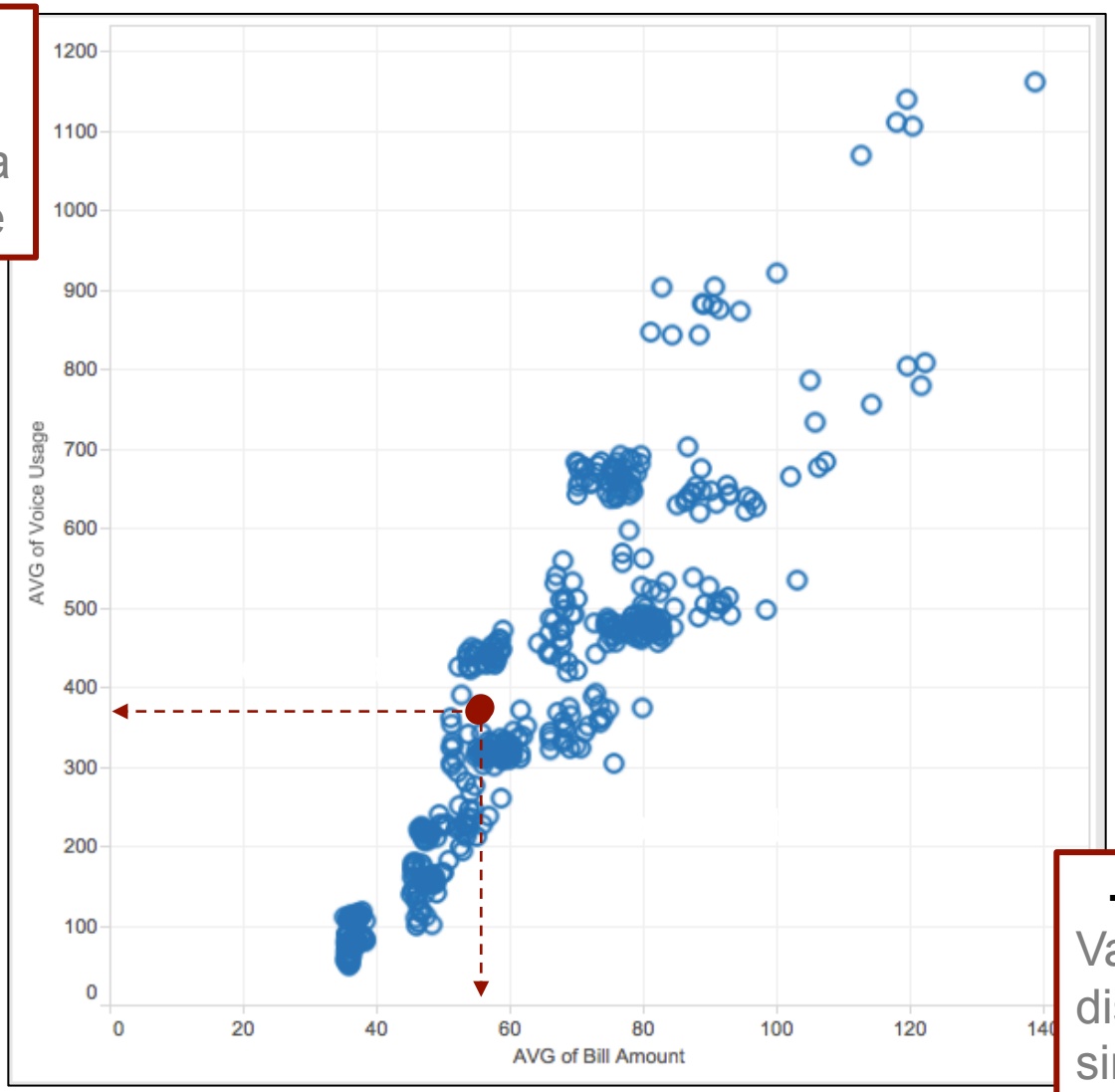
It's difficult to account for every outside, or confounding factor, which makes it difficult to prove causation

You can, however, easily find and see correlation and a **scatter plot** is our key tool for visualising it

Scatter Plot

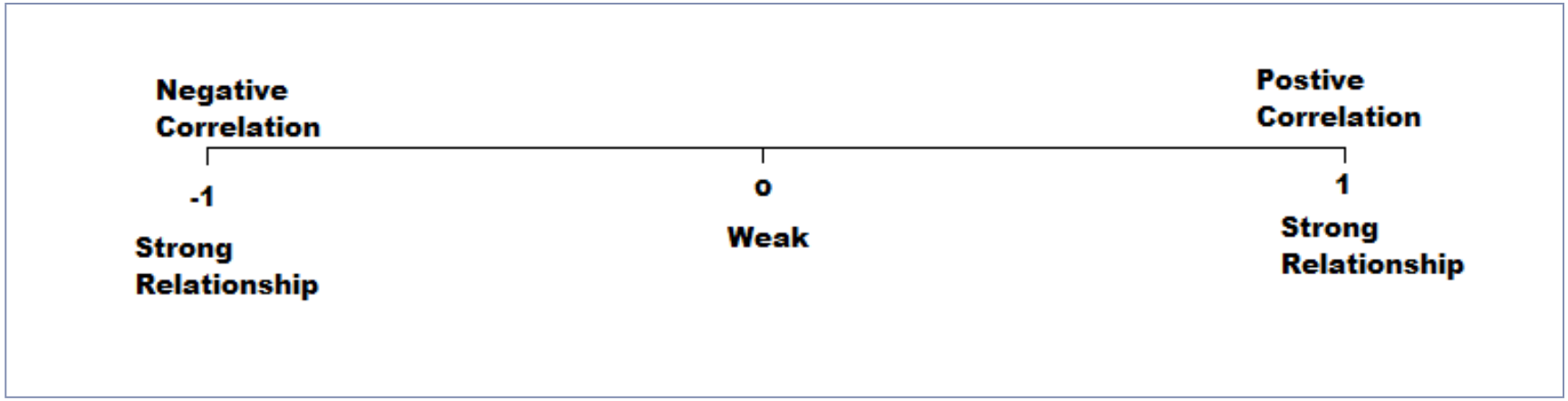


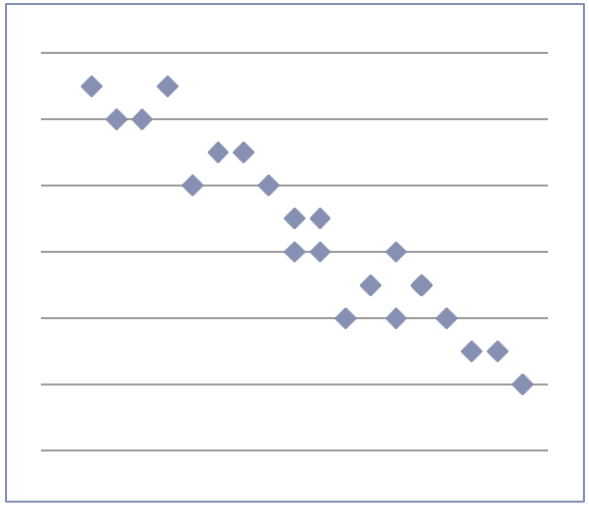
Y-Axis
Values displayed for a single variable

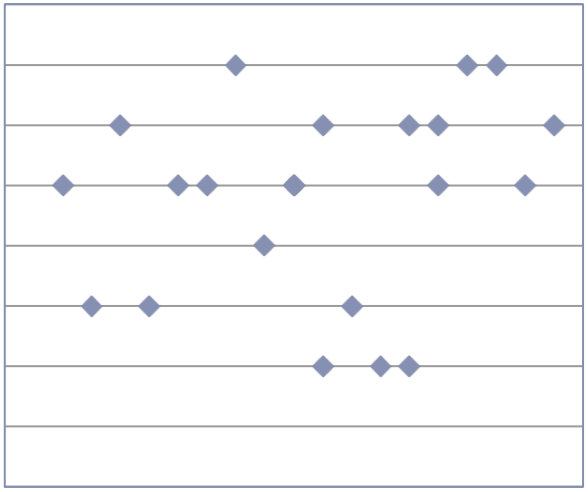


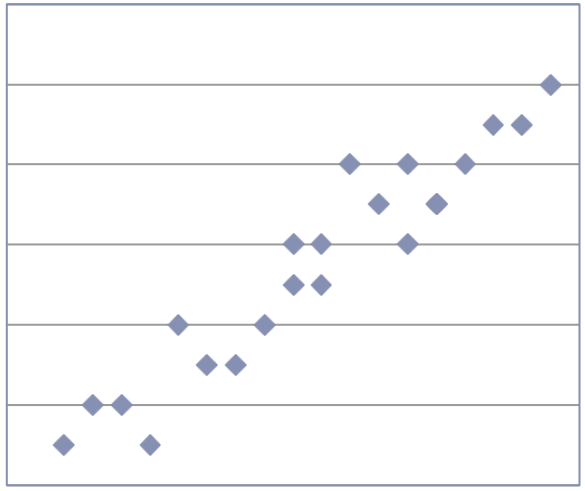
-Axis
Values displayed for a single variable

Simple Scatter Plot





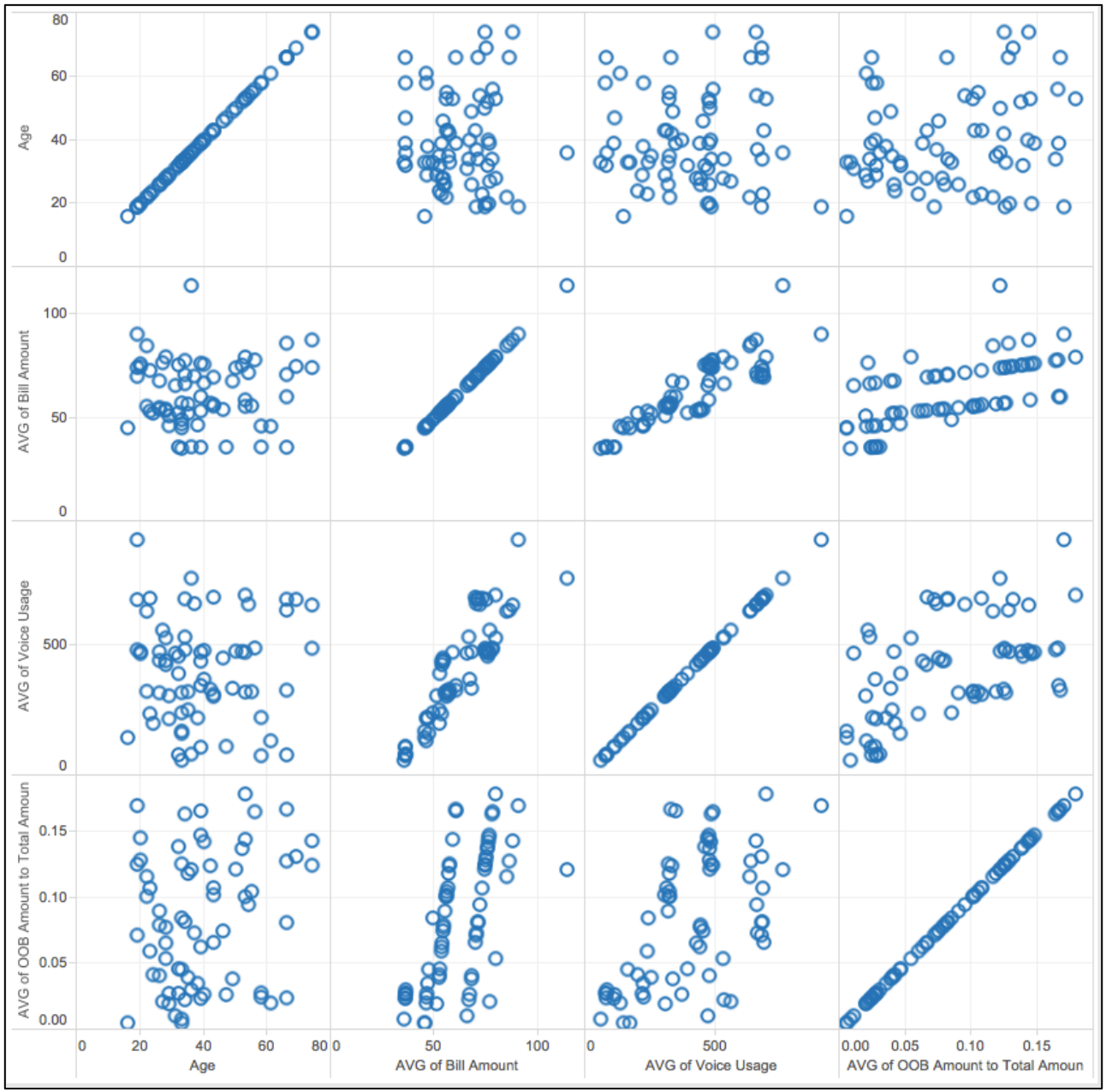


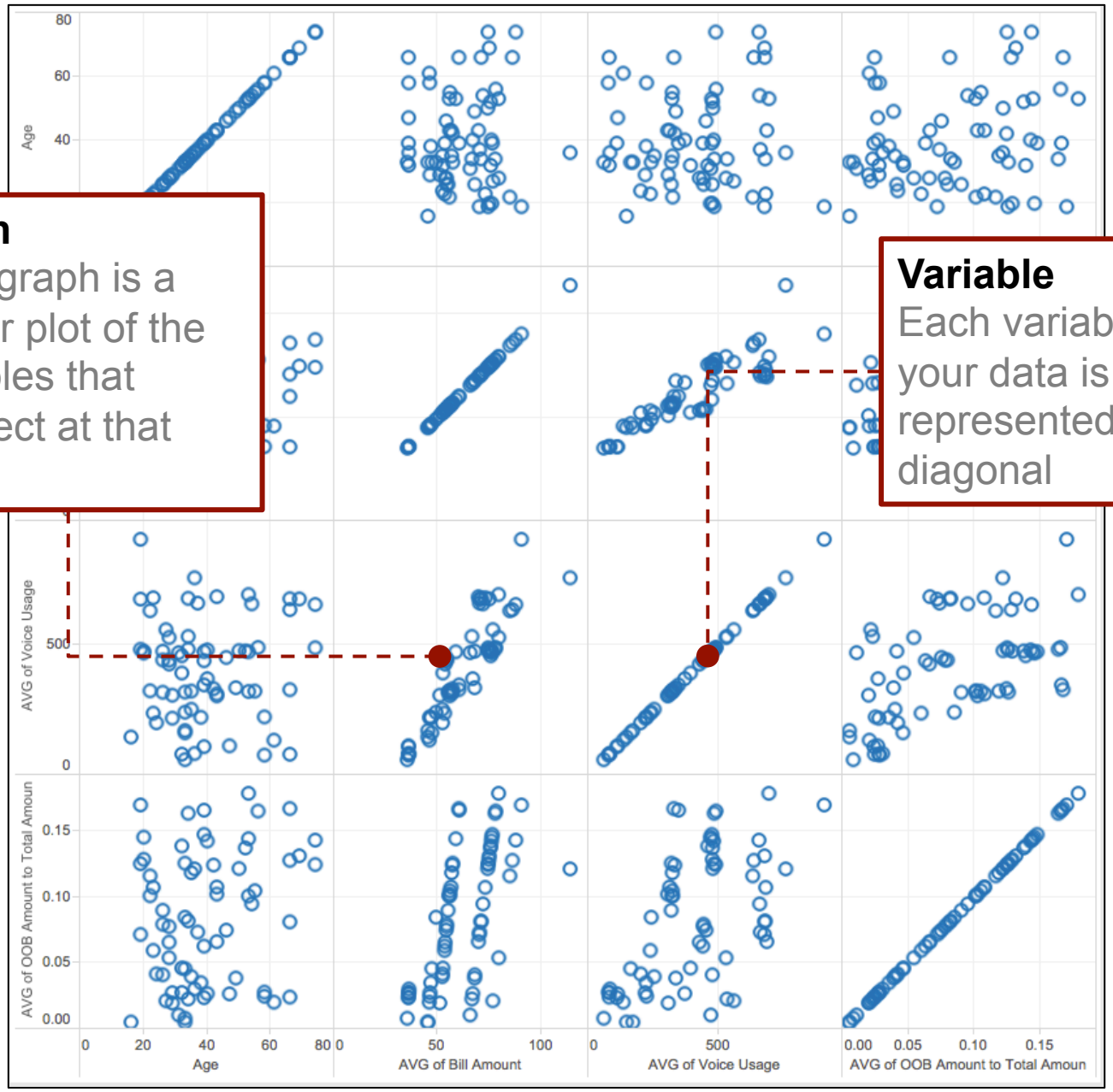


Exploring Even More Variables

You can plot every possible pair with a scatter plot matrix (**SPLOM**) to compare all variables

- A square grid with all variables on both the vertical and horizontal
- Each column represents a variable on the horizontal axis, and each row represents a variable on the vertical axis
- This provides all possible pairs


Age
AVG of Bill Amount
AVG of Voice Usage
AVG of OOB Amount to Total Amoun
0
20
40
60
80
50
100
500
0.00
0.05
0.10
0.15

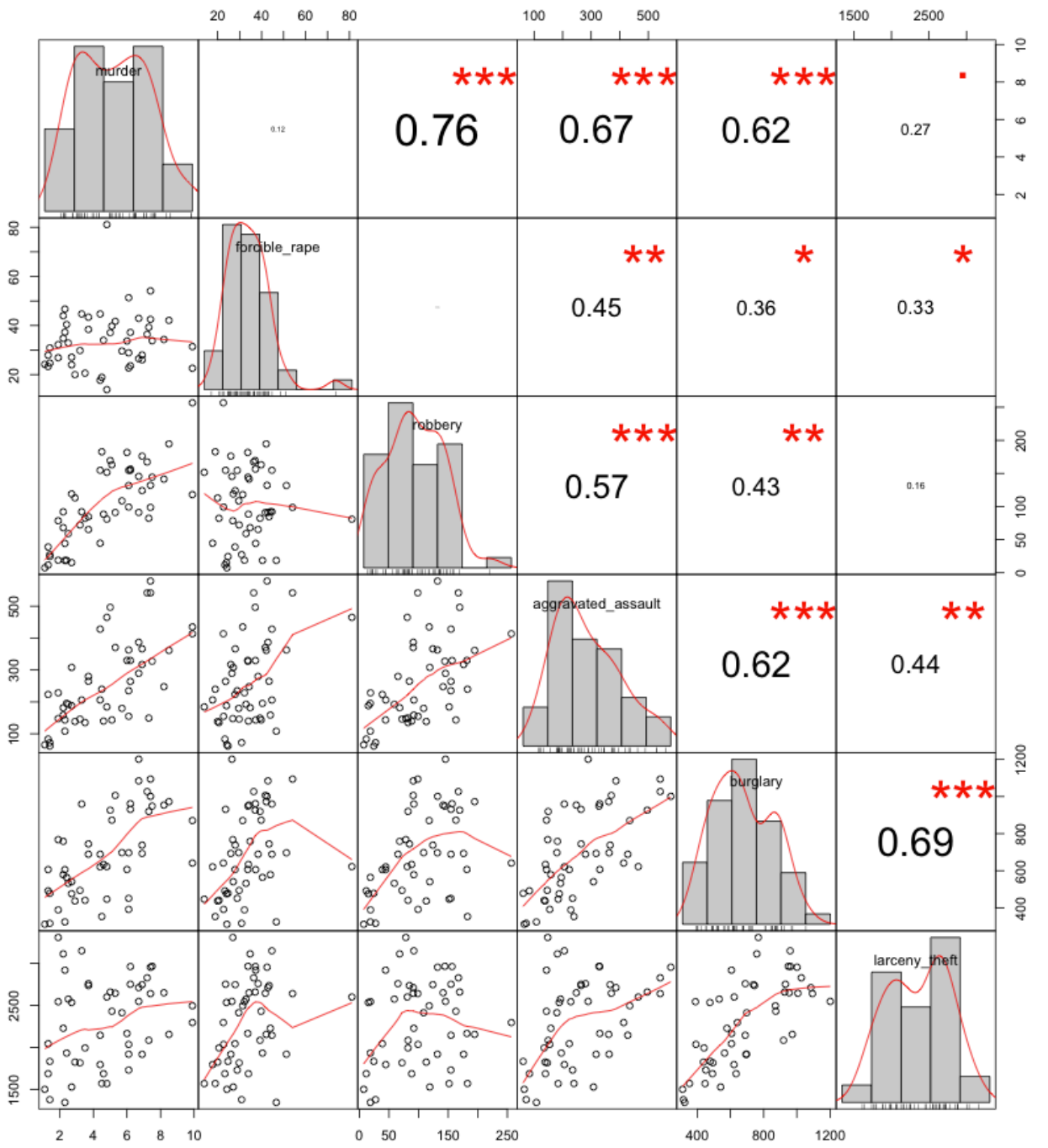


Graph

Each graph is a scatter plot of the variables that intersect at that point

Variable

Each variable in your data is represented on the diagonal


murder
forcible_rape
robbery
aggravated_assault
burglary
larceny_theft
0.12
0.76
0.67
0.62
0.27
0.45
0.36
0.33
0.57
0.43
0.16
0.62
0.44
0.69
***
**
*
.

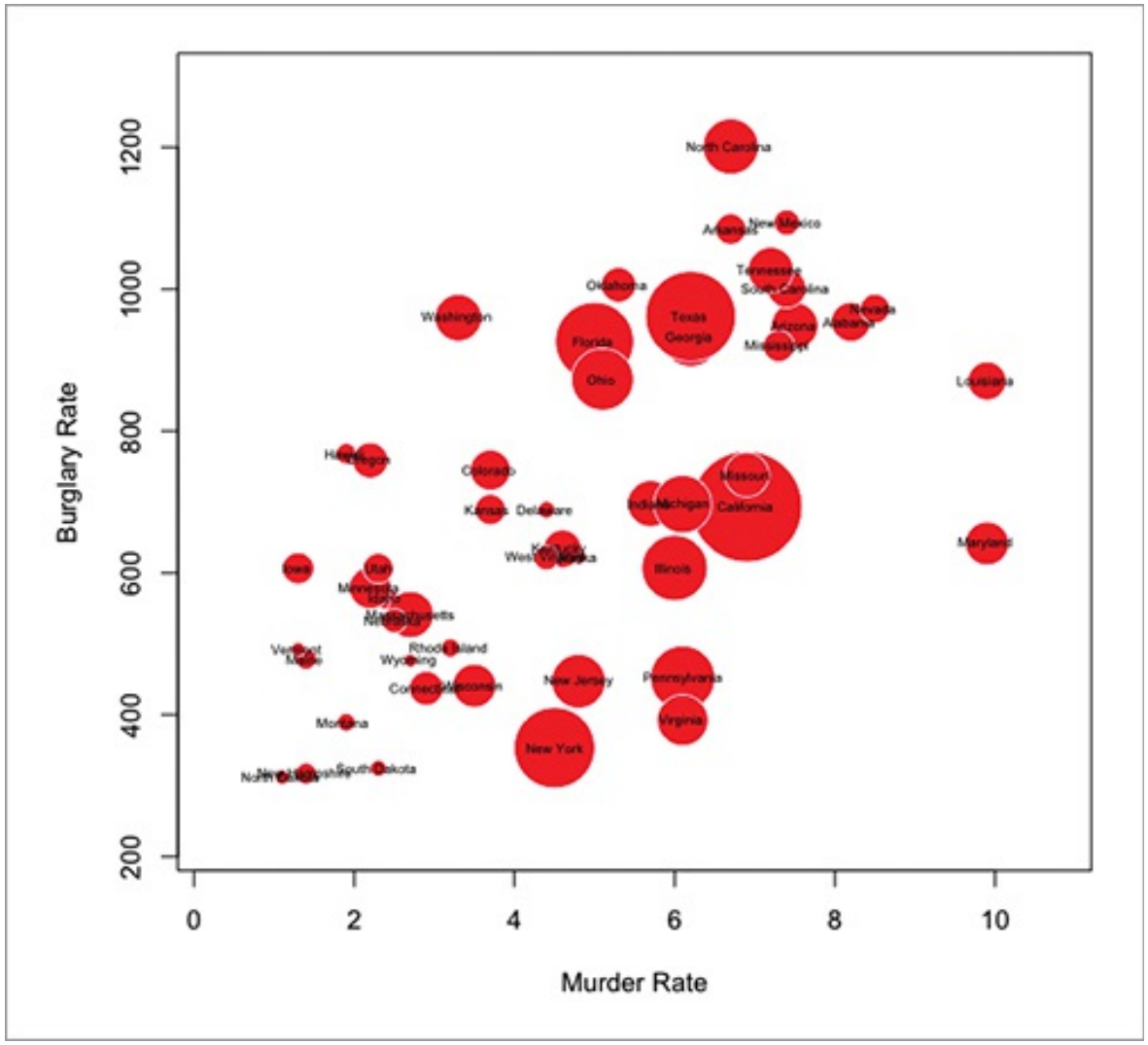
Channelling Hans

Hans Rosling, Professor of International Health at Karolinska Institute and chairman of the Gapminder Foundation popularised bubble charts

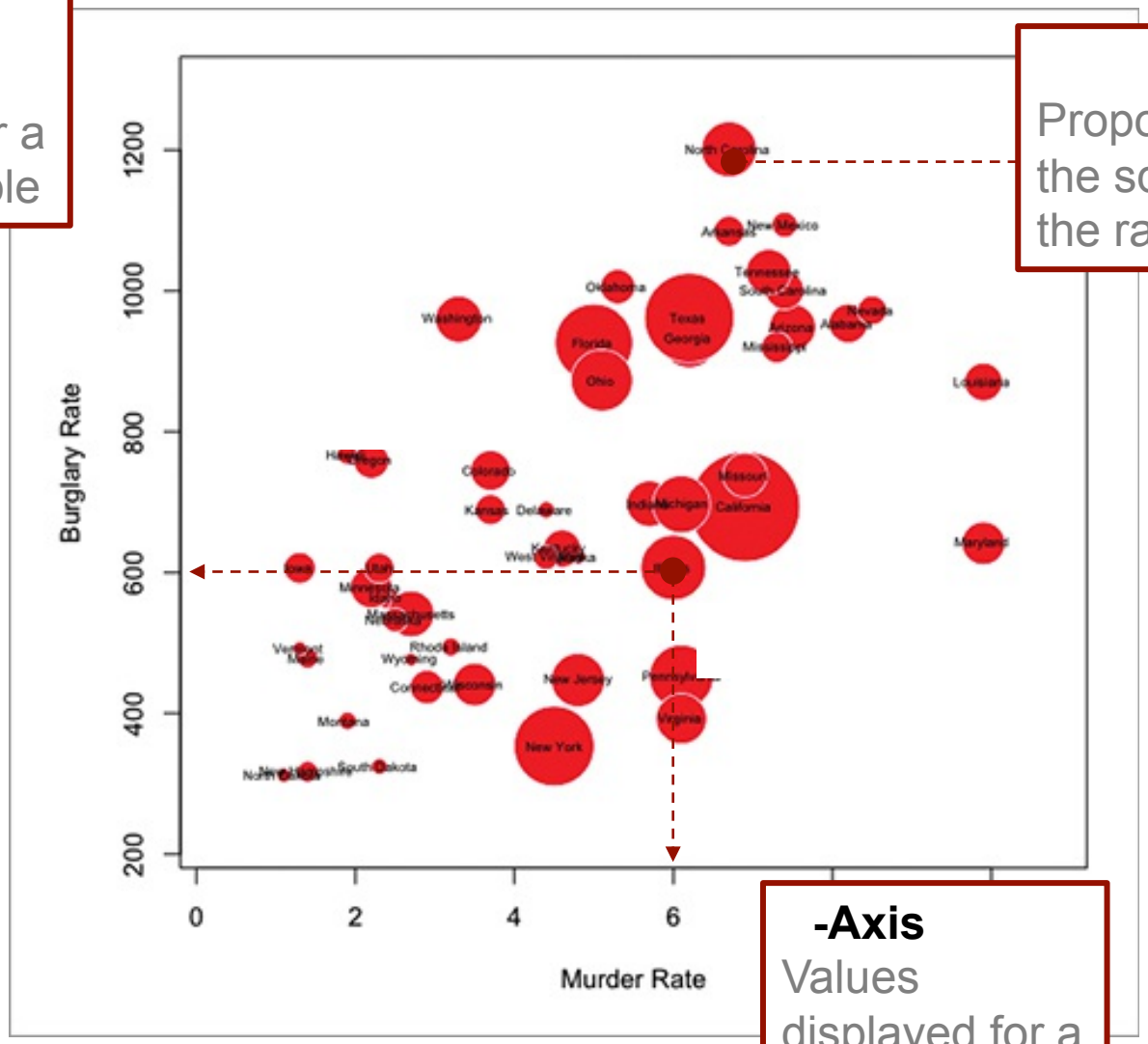
- The advantage of this chart type is that can compare three variables at one time
- One variable is on the x-axis, one is on the y-axis, and the third is represented by the **area size** of the bubbles

www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

Bubble Plot


Burglary Rate
200
400
600
800
1000
1200
0
2
4
6
8
10
Murder Rate
North Carolina
Oklahoma
Texas
Georgia
Florida
Ohio
Washington
Louisiana
Missouri
California
Michigan
Illinois
Colorado
Kansas
Maryland
New Jersey
Pennsylvania
Virginia
New York
Montana
South Dakota

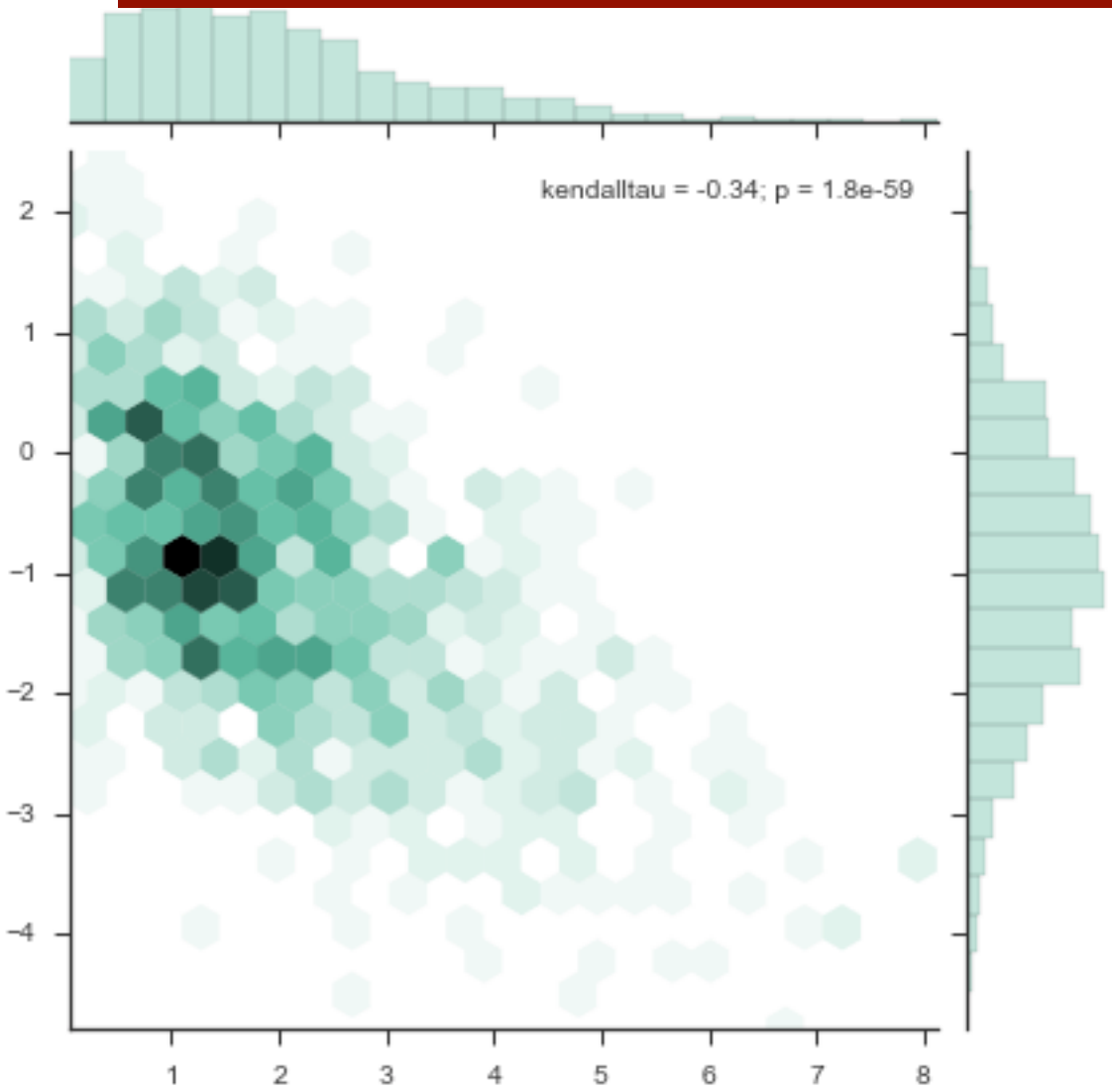
Y-Axis
Values displayed for a single variable



Proportional to the square of the radius

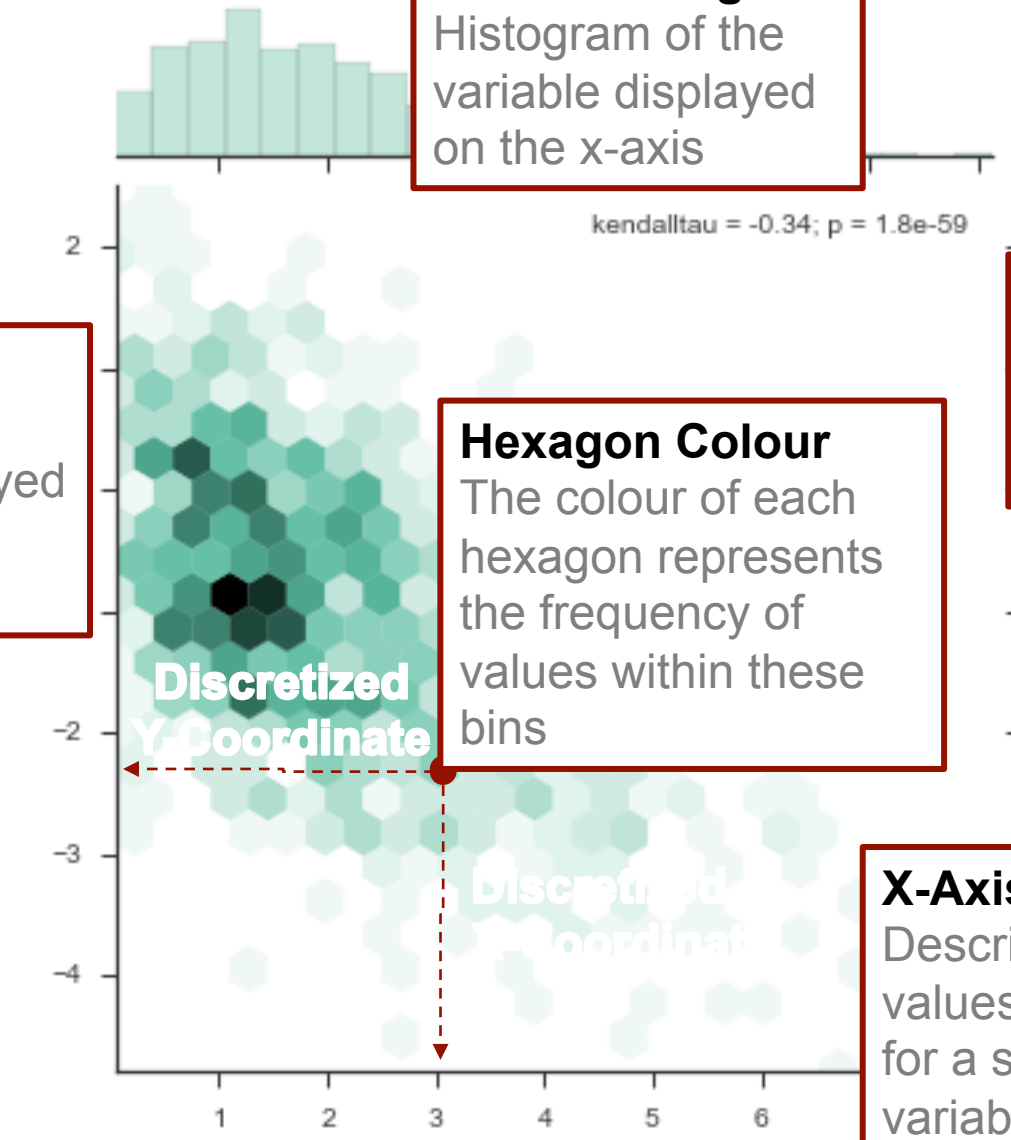
-Axis
Values displayed for a single variable

Hexbin Plot



Seaborn Python package: www.stanford.edu/~mwaskom/software/seaborn/

X-Axis Histogram
Histogram of the variable displayed on the x-axis



Y-Axis
Discretized values displayed for a single variable

Hexagon Colour
The colour of each hexagon represents the frequency of values within these bins

Y-Axis Histogram
Histogram of the variable displayed on the y-axis

X-Axis
Descritized values displayed for a single variable

Seaborn Python package: www.stanford.edu/~mwaskom/software/seaborn/

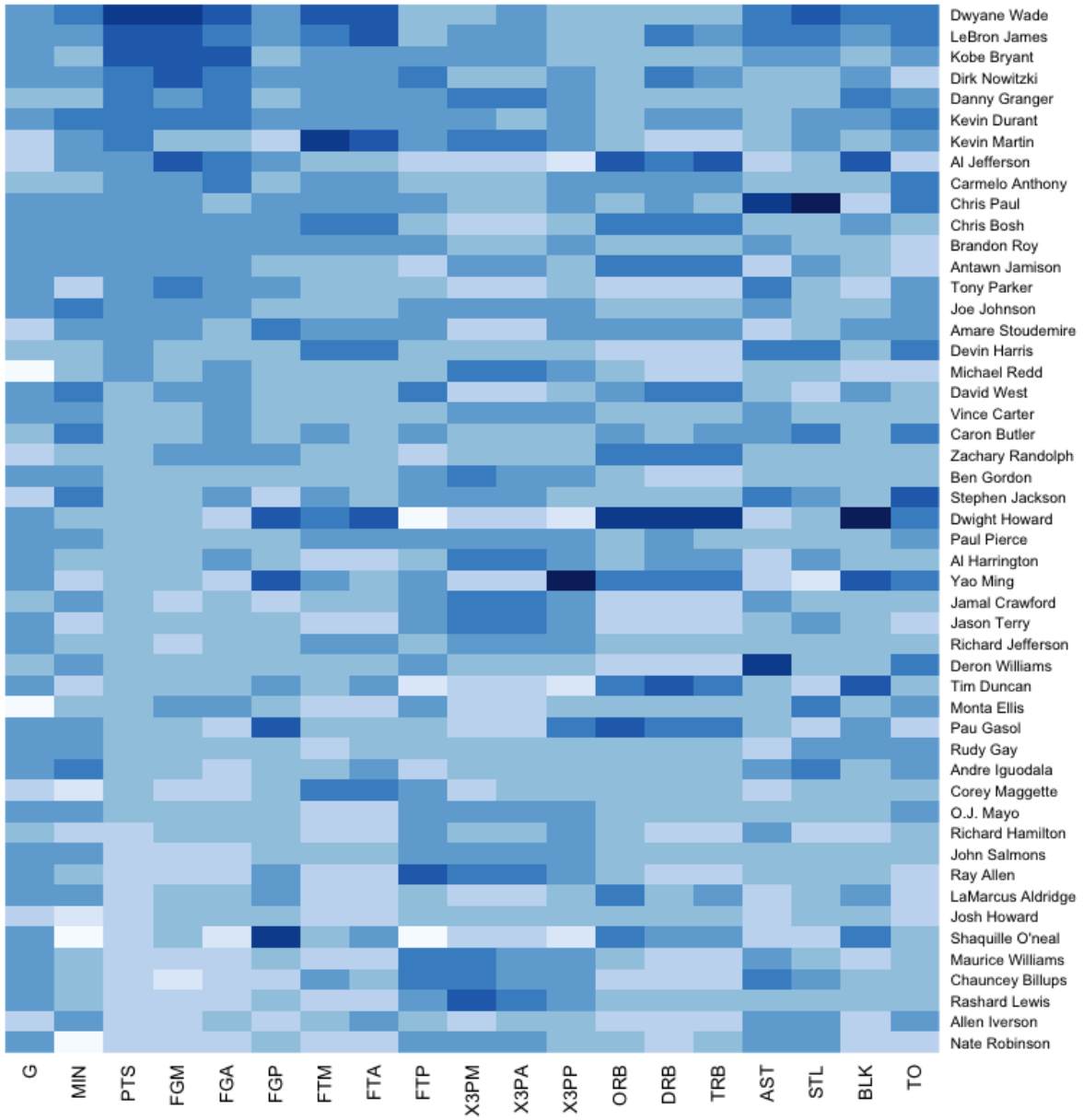
Heatmaps

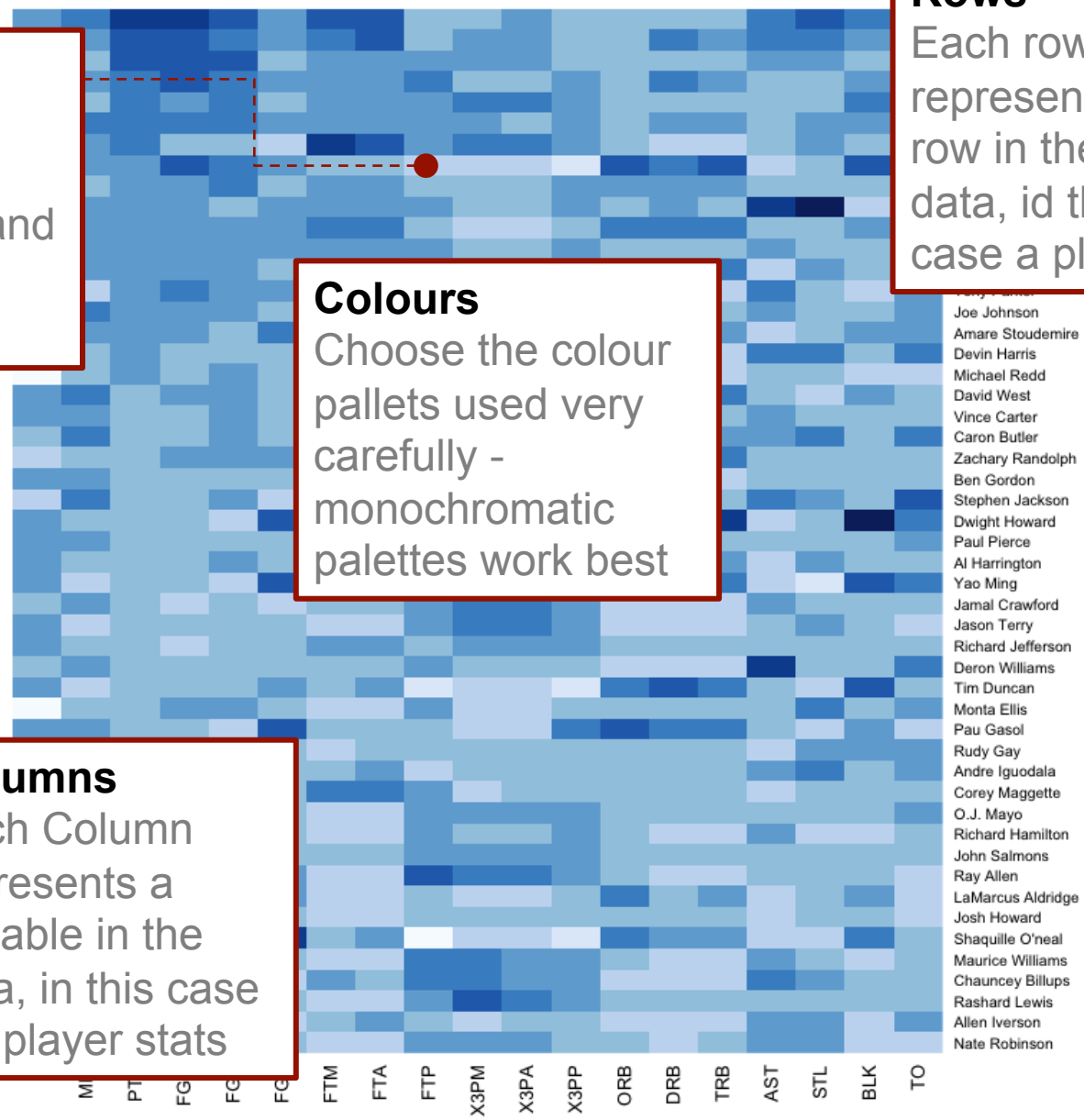
Sometimes we wish to investigate the relationships between multiple variables at once

This is a fundamentally hard problem but there are some techniques that work

Heat maps are a simple examples that can be quite effective

Heat Maps


Dwyane Wade
LeBron James
Kobe Bryant
Dirk Nowitzki
Danny Granger
Kevin Durant
Kevin Martin
Al Jefferson
Carmelo Anthony
Chris Paul
Chris Bosh
Brandon Roy
Antawn Jamison
Tony Parker
Joe Johnson
Amare Stoudemire
Devin Harris
Michael Redd
David West
Vince Carter
Caron Butler
Zachary Randolph
Ben Gordon
Stephen Jackson
Dwight Howard
Paul Pierce
Al Harrington
Yao Ming
Jamal Crawford
Jason Terry
Richard Jefferson
Deron Williams
Tim Duncan
Monta Ellis
Pau Gasol
Rudy Gay
Andre Iguodala
Corey Maggette
O.J. Mayo
Richard Hamilton
John Salmons
Ray Allen
LaMarcus Aldridge
Josh Howard
Shaquille O'neal
Maurice Williams
Chauncey Billups
Rashard Lewis
Allen Iverson
Nate Robinson
G
MIN
PTS
FGM
FGA
FGP
FTM
FTA
FTP
X3PM
X3PA
X3PP
ORB
DRB
TRB
AST
STL
BLK
TO



Each cell represents a single value and is coloured appropriately

Rows
Each row represents a row in the data, id this case a player

Colours
Choose the colour pallets used very carefully - monochromatic palettes work best

Columns
Each Column represents a variable in the data, in this case the player stats

Summary

The key relationship we can look for is correlation

- Comparing variables: Scatter plot
- Exploring more variables: Scatter plot matrix
- Channelling Hans: Bubble plot
- Comparing data points: Heat map

DON’T FORGET DATA TABLES!

The Meagre Data Table

In all the previous visualisation examples we have focused on creating charts and graphs from raw data normally stored in some sort of data table

Don’t forget that in many instances it makes sense to allow the user to actually see the data in a data table

The Meagre Data Table

With regard to data tables remember:

- In cases where precision is required showing the actual data is invaluable
- If a visualisation of data does not add anything in terms of enhancing the understanding of the data than the visualisation is redundant

A good example of this is the information contained in a balance sheet

Notes to the Financial Statements

1. Income

	2012 €000	2011 €000
Opening Balance	(3,807)	(2,061)
Received from State Authorities	89,559	106,710
Received from the Special Obstetrics Indemnity Scheme	645	1,797
Closing Balance	13,827	3,807
	100,224	**110,253**

2. Other Expenses

	2012 €000	2011 €000
State Claims Agency expenses		
- Legal fees	16,497	13,612
- Medical fees	3,317	1,386
- Engineers' fees	278	176
- Other fees (including investigation and actuary fees)	992	839
	21,084	**16,013**
Plaintiff expenses		
- Legal fees	22,271	26,474
- Other expert fees	-	2
- Travel expenses	13	8
	22,284	**26,484**
Witness expenses	9	11
	43,377	**42,508**

Effective Data Tables

When designing tables think about layout, typography, and design to make the table as effective as possible

Use the Tufte data ink ratio idea

- Remove unnessecary gridlines & borders
- Remove bolding
- Remove colour
- Left align text, right align numbers
- Ensure consitent rounding
- Use whitespace effectively

Effective Data Tables

Name	Age	Salary	Loan to Value	County
Mary	43	34876.98	0.9741	Mayo
John	56	54882	0.41772	Cork
Pauline	21	56864.1	0.0976	Kilkenny
George	34	104282.76	0.4894	Galway
Henry	54	31410	0.86	Dublin
Shane	28	84220.78	0.213	Dublin
Sarah	76	72306.97	0.5194	Leitrim
Jane	49	30675.22	0.34	Armagh
Aidan	31	103583.56	0.88088	Sligo
Simon	48	67341.39	0.1428	Cork
Aoife	32	68001.4	0.886	Waterford
Sinead	19	67502	0.89719	Kildare

Effective Data Tables

Name	Age	Salary	Loan to Value	County
Mary	43	34,877	0.974	Mayo
John	56	54,882	0.418	Cork
Pauline	21	56,864	0.098	Kilkenny
George	34	104,283	0.489	Galway
Henry	54	31,410	0.860	Dublin
Shane	28	84,221	0.213	Dublin
Sarah	76	72,307	0.519	Leitrim
Jane	49	30,675	0.340	Armagh
Aidan	31	103,584	0.881	Sligo
Simon	48	67,341	0.143	Cork
Aoife	32	68,001	0.886	Waterford
Sinead	19	67,502	0.897	Kildare

SUMMARY

Summary

The key points made in this section were:

- We can group different visualisations into the following general types:
  - Nominal comparisons
  - Distributions
  - Trends over time
  - Relationships
  - Data tables
- Think about what you want to say and choose the right visualisation type