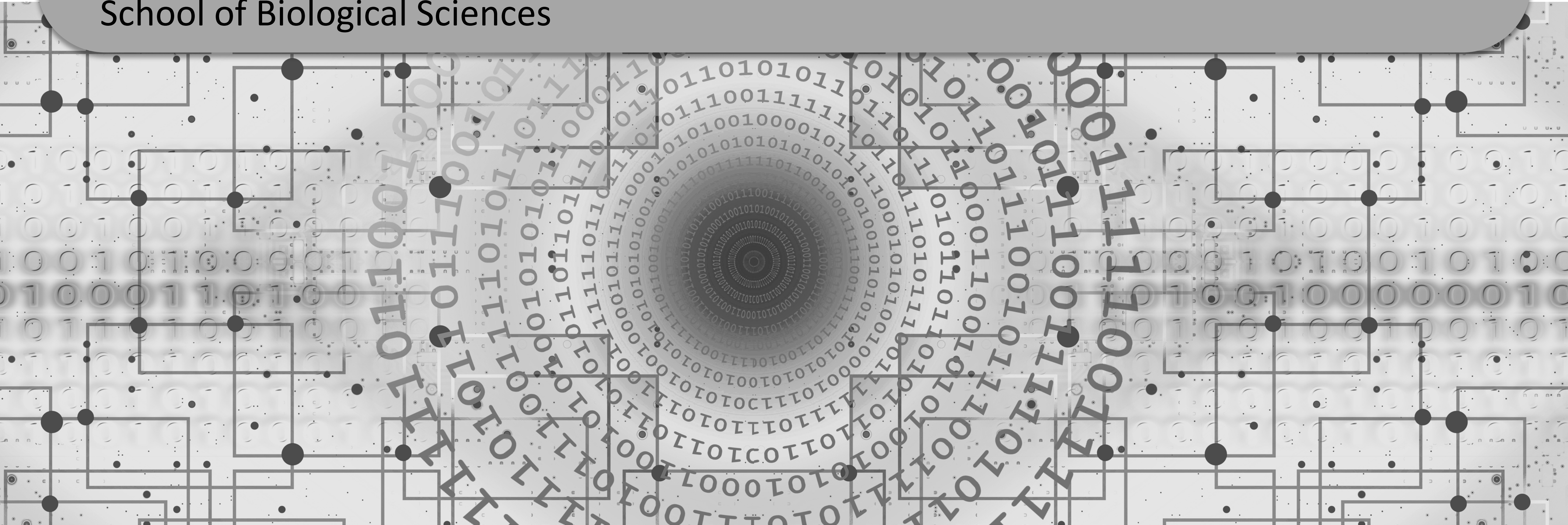
NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Graphics

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences



Learning Objectives

By the end of this topic, you should be able to:

- Explain the importance of data visualisation.
- Identify bad graphs.
- Explain how bar charts mislead.
- Explain the limitations of summary statistics using Anscombe’s quartet.

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

The Importance of Visualisation

BS0004 Introduction to Data Science

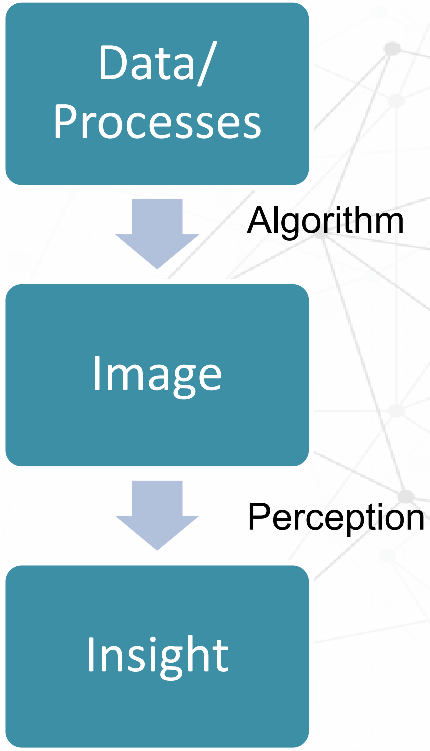
Dr Wilson Goh
School of Biological Sciences

What is data visualisation?

Data visualisation is the process of converting raw data into easily understood pictures of information that enable faster and effective exploration, discovery, insight, and decision-making.

Data → Visuals

> The “transformation from numbers to insight requires two stages.”
>
> - Jacques Bertin



Data/ Processes

Algorithm

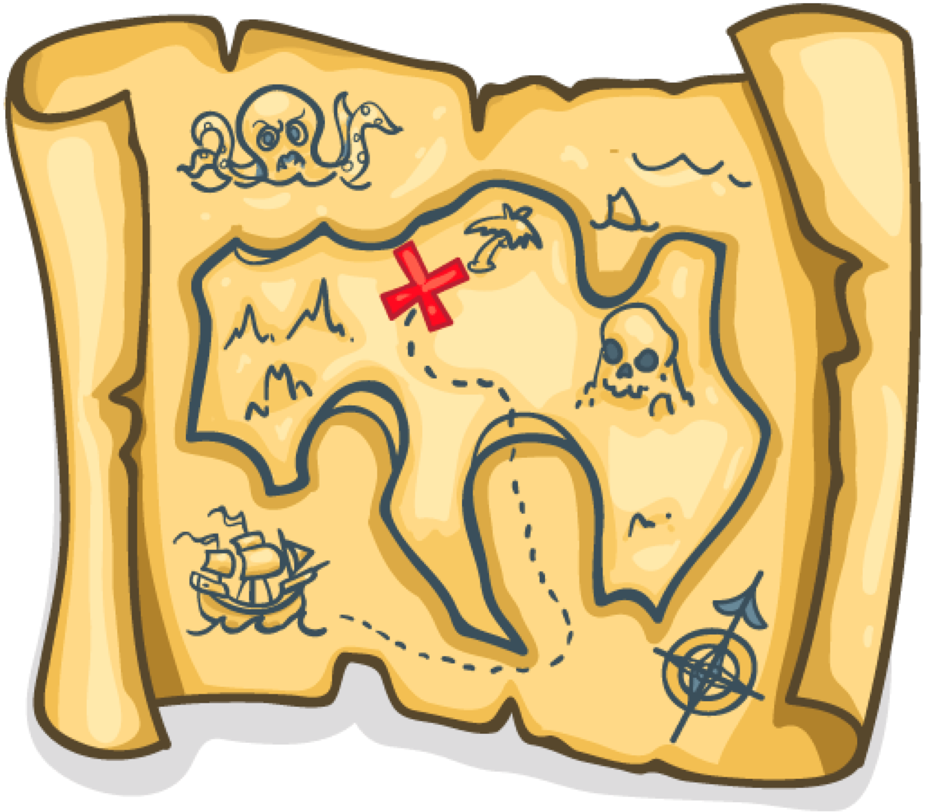
Image

Perception

Insight

X Marks the Spot

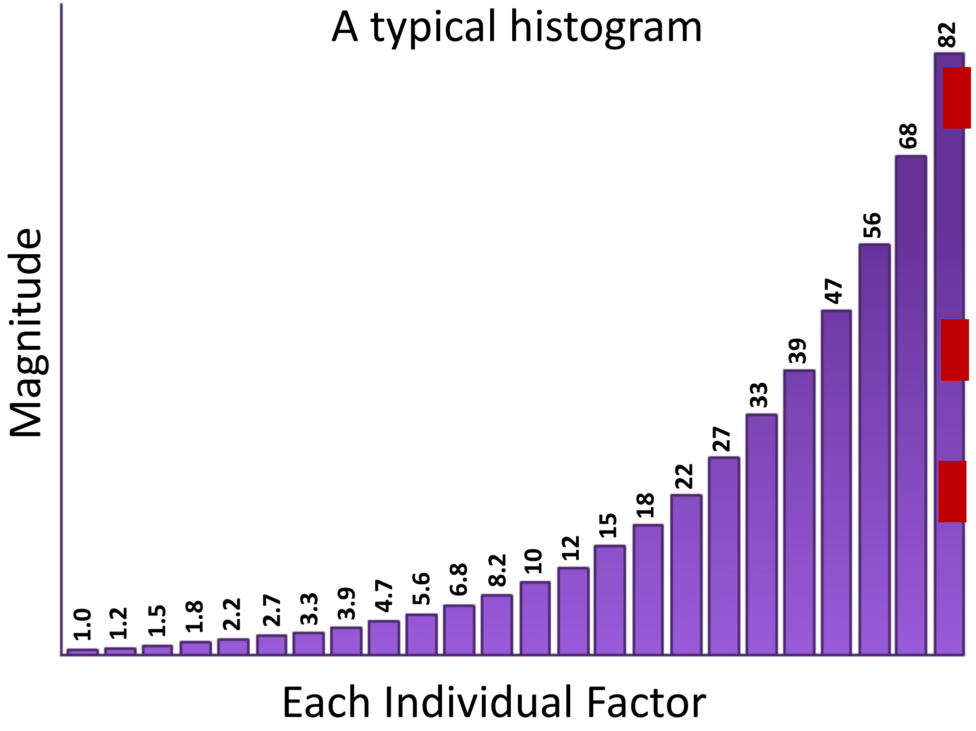
- Much of our communication is done via words.
- The specific arrangement of words conveys meaning.
- Can meaning also be conveyed via pictorial means?



Abstraction Without Words (Marks)

- A mark is made to represent some information other than itself. It is also referred to as a sign.
- Marks can be:
  - **Points** are dimensionless locations on the plane, represented by signs that obviously need to have some size, shape or colour for visualisation.
  - **Lines** represent information with a certain length, but no area and therefore no width. Again lines are visualised by signs of some thickness.
  - **Areas** have a length and a width and therefore a two-dimensional size.
  - **Surfaces** are areas in a three-dimensional space, but with no thickness.
  - **Volumes** have a length, a width and a depth. They are thus truly three-dimensional.

Adding Value to Visual Representation of Data using Perception

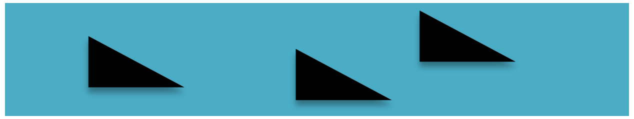
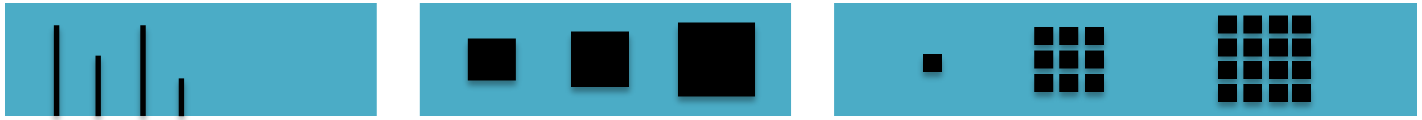
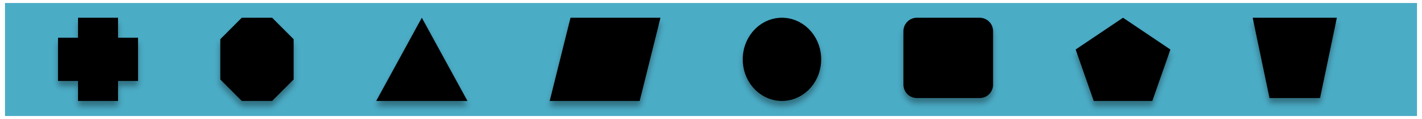
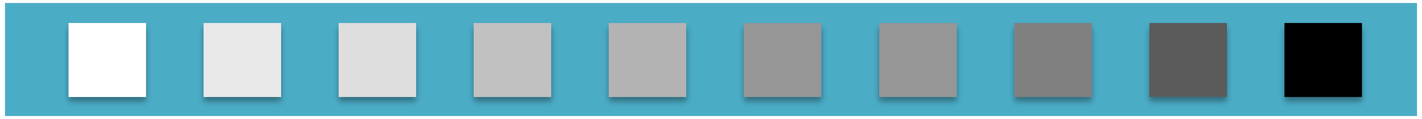
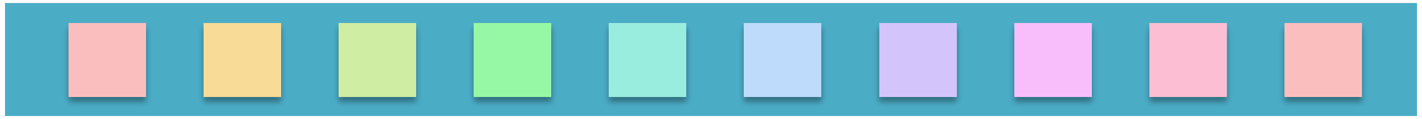
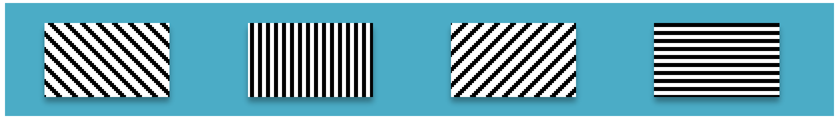
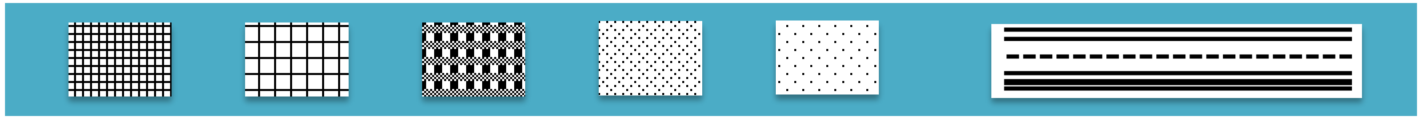


- But after using up those two dimensions, what other attributes can you use?

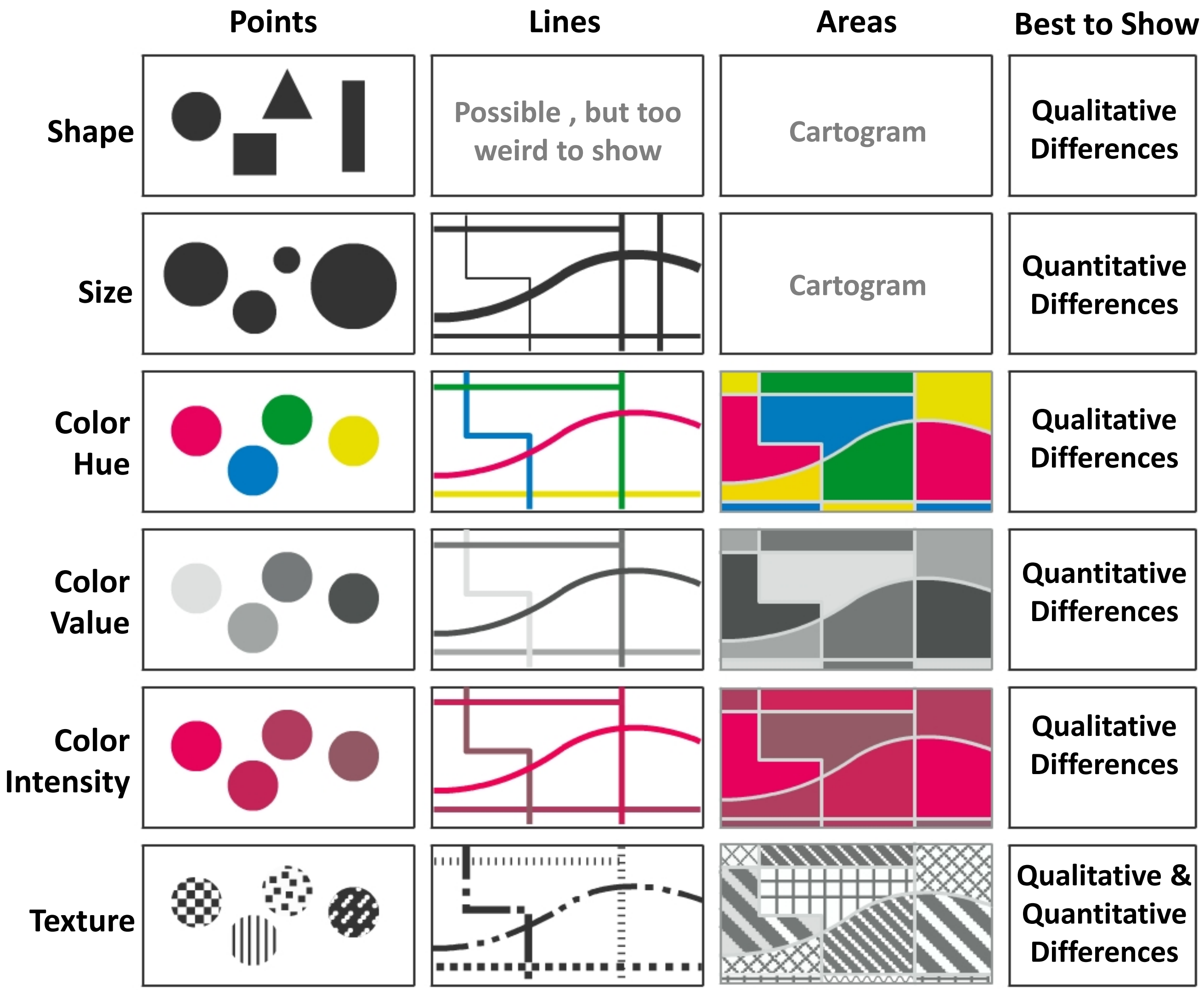
- For depicting additional factors, you then have to choose between size, color, value, texture, line orientation or shape (Bertin’s 7 retinal variables).

- Not all retinal variables are equally effective in their ability to represent information.

The 7 Retinal Variables

Bertin’s Original Visual Variables	
Position Changes in the x, y location	
Size Change in length, area or repetition	
Shape Infinite number of shapes	
Value Changes from light to dark	
Colour Changes in hue at a given value	
Orientation Changes in alignment	
Texture Variation in ‘grain’	

Idealising Bertin’s Visual (Retinal) Variables



	Points	Lines	Areas	Best to Show
Shape		Possible , but too weird to show	Cartogram	Qualitative Differences
Size			Cartogram	Quantitative Differences
Color Hue				Qualitative Differences
Color Value				Quantitative Differences
Color Intensity				Qualitative Differences
Texture				Qualitative & Quantitative Differences

Making Maps: A Visual Guide to Map Design for GIS by John Krygier and Denis Wood

Benefits of Data Visualisation

Data visualisation allows users see several different **perspectives** of the data.

Data visualisation makes it possible to **interpret vast amounts** of data.

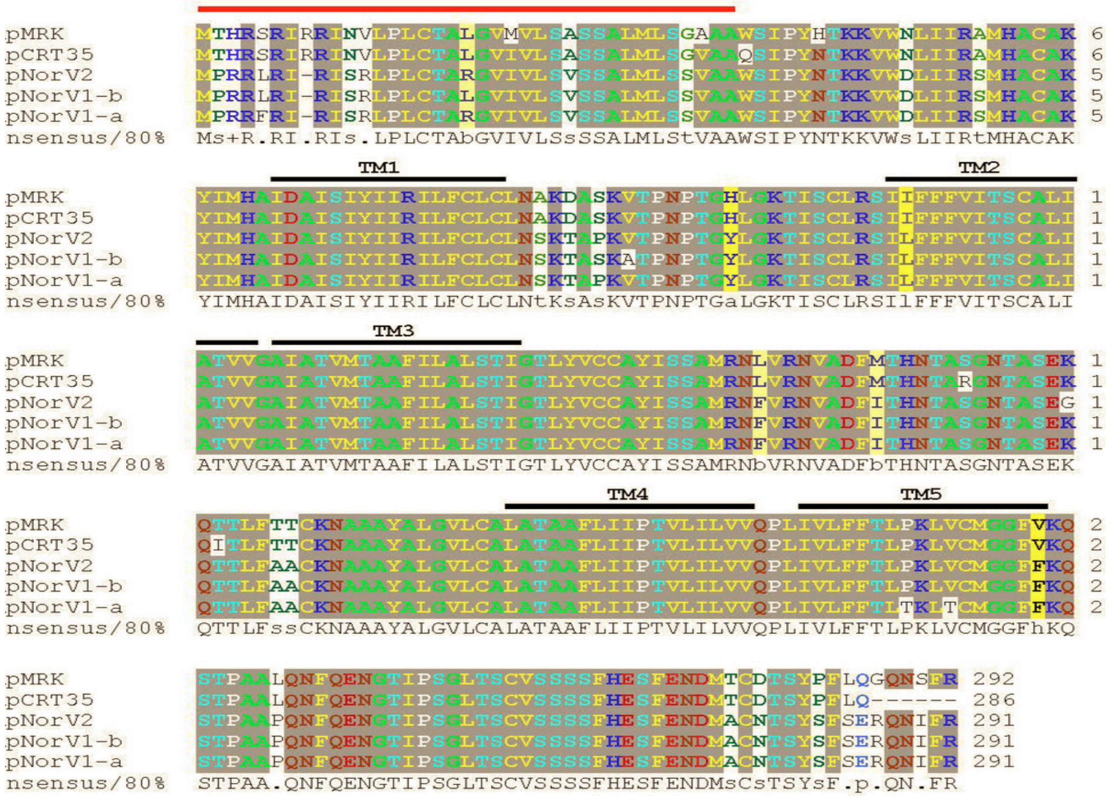
Data visualisation offers the ability to note **exceptions** in the data.

Data visualisation allows the user to **analyse visual patterns** in the data.

Exploring trends within a database through visualisation by letting analysts **navigate through data** and visually orient themselves to the patterns in the data.

A Picture Paints a Thousand Words

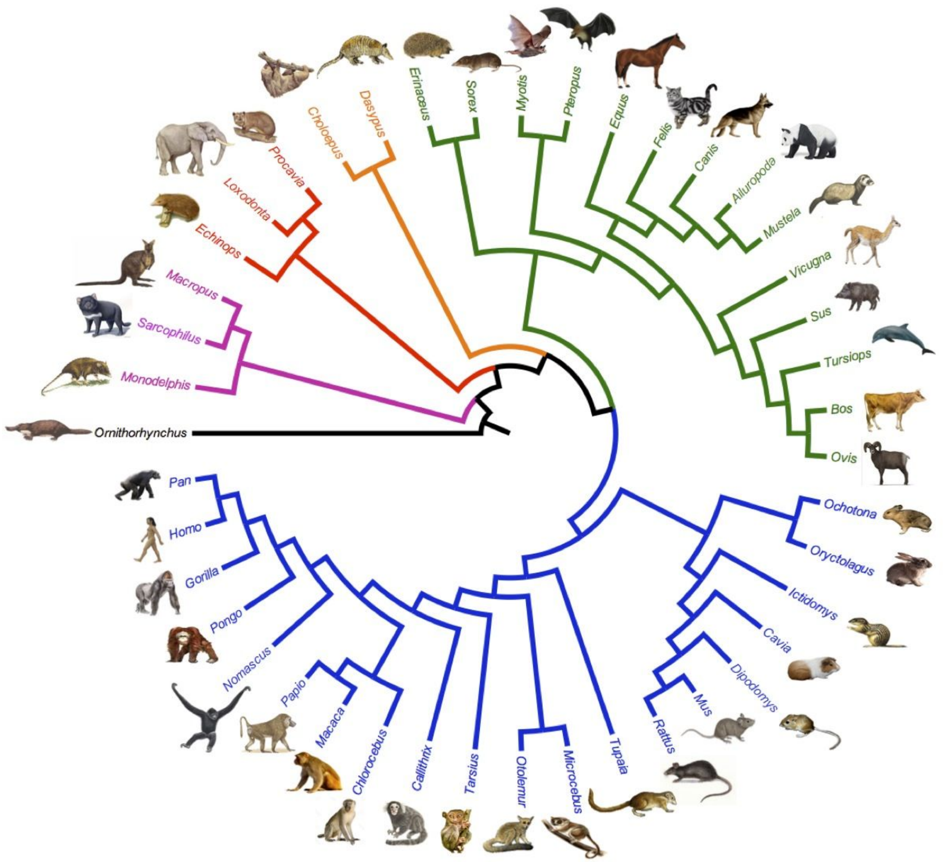
From this…



```
                                TM1                                          TM2
pMRK          MTHRSRIRRINVLPLCTALGVMVLSASSALMLSGAAAWSIPYHTKKVWNLIIRAMHACAK 6
pCRT35        MTHRSRIRRINVLPLCTALGVIVLSASSALMLSGVAAQSIPYNTKKVWNLIIRAMHACAK 6
pNorV2        MPRRLRI-RISRLPLCTARGVIVLSVSSALMLSSVAAWSIPYNTKKVWDLIIRSMHACAK 5
pNorV1-b      MPRRLRI-RISRLPLCTALGVIVLSVSSALMLSSVAAWSIPYNTKKVWDLIIRSMHACAK 5
pNorV1-a      MPRRFRI-RISRLPLCTARGVIVLSVSSALMLSSVAAWSIPYNTKKVWDLIIRSMHACAK 5
nsensus/80%   Ms+R.RI.RIs.LPLCTAbGVIVLSsSSALMLStVAAWSIPYNTKKVWsLIIRtMHACAK

                    TM1                                          TM2
pMRK          YIMHAIDAISIYIIRILFCLCLNAKDASKVTPNPTGHLGKTISCLRSIIFFFVITSCALI 1
pCRT35        YIMHAIDAISIYIIRILFCLCLNAKDASKVTPNPTGHLGKTISCLRSIIFFFVITSCALI 1
pNorV2        YIMHAIDAISIYIIRILFCLCLNSKTAPKVTPNPTGYLGKTISCLRSILFFFVITSCALI 1
pNorV1-b      YIMHAIDAISIYIIRILFCLCLNSKTASKATPNPTGYLGKTISCLRSILFFFVITSCALI 1
pNorV1-a      YIMHAIDAISIYIIRILFCLCLNSKTAPKVTPNPTGYLGKTISCLRSILFFFVITSCALI 1
nsensus/80%   YIMHAIDAISIYIIRILFCLCLNtKsAsKVTPNPTGaLGKTISCLRSIlFFFVITSCALI

                     TM3
pMRK          ATVVGAIATVMTAAFILALSTIGTLYVCCAYISSAMRNLVRNVADFMTHNTASGNTASEK 1
pCRT35        ATVVGAIATVMTAAFILALSTIGTLYVCCAYISSAMRNLVRNVADFMTHNTARGNTASEK 1
pNorV2        ATVVGAIATVMTAAFILALSTIGTLYVCCAYISSAMRNFVRNVADFITHNTASGNTASEG 1
pNorV1-b      ATVVGAIATVMTAAFILALSTIGTLYVCCAYISSAMRNFVRNVADFITHNTASGNTASEK 1
pNorV1-a      ATVVGAIATVMTAAFILALSTIGTLYVCCAYISSAMRNFVRNVADFITHNTASGNTASEK 1
nsensus/80%   ATVVGAIATVMTAAFILALSTIGTLYVCCAYISSAMRNbVRNVADFbTHNTASGNTASEK

                                       TM4                  TM5
pMRK          QTTLFTTCKNAAAYALGVLCALATAAFLIIPTVLILVVQPLIVLFFTLPKLVCMGGFVKQ 2
pCRT35        QITLFTTCKNAAAYALGVLCALATAAFLIIPTVLILVVQPLIVLFFTLPKLVCMGGFVKQ 2
pNorV2        QTTLFAACKNAAAYALGVLCALATAAFLIIPTVLILVVQPLIVLFFTLPKLVCMGGFFKQ 2
pNorV1-b      QTTLFAACKNAAAYALGVLCALATAAFLIIPTVLILVVQPLIVLFFTLPKLVCMGGFFKQ 2
pNorV1-a      QTTLFAACKNAAAYALGVLCALATAAFLIIPTVLILVVQPLIVLFFTLTKLTCMGGFFKQ 2
nsensus/80%   QTTLFssCKNAAAYALGVLCALATAAFLIIPTVLILVVQPLIVLFFTLPKLVCMGGFhKQ

pMRK          STPAALQNFQENGTIPSGLTSCVSSSSFHESFENDMTCDTSYPFLQGQNSFR 292
pCRT35        STPAALQNFQENGTIPSGLTSCVSSSSFHESFENDMTCDTSYPFLQ------ 286
pNorV2        STPAAPQNFQENGTIPSGLTSCVSSSSFHESFENDMACNTSYSFSERQNIFR 291
pNorV1-b      STPAAPQNFQENGTIPSGLTSCVSSSSFHESFENDMACNTSYSFSERQNIFR 291
pNorV1-a      STPAAPQNFQENGTIPSGLTSCVSSSSFHESFENDMACNTSYSFSERQNIFR 291
nsensus/80%   STPAA.QNFQENGTIPSGLTSCVSSSSFHESFENDMsCsTSYsF.p.QN.FR
```

Raw sequences and alignments of individual genes.

To this…

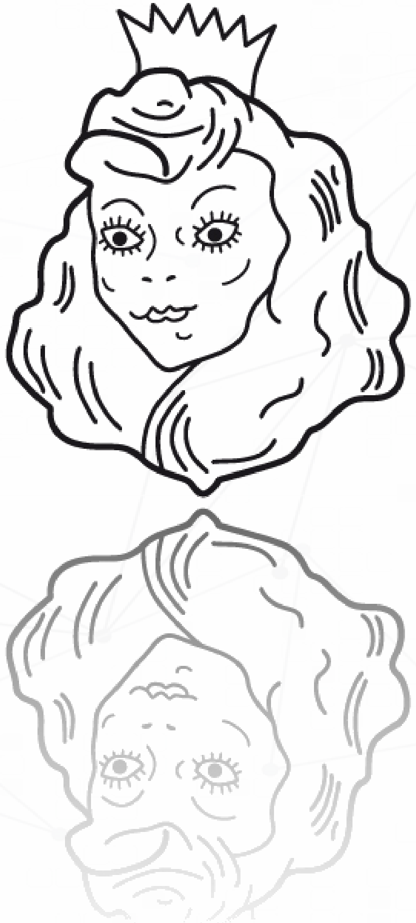


Overall evolutionary relationships.

A Picture Paints a Thousand Words

What you infer: A young beautiful princess.

- Data presentation forms the foundation of our collective scientific knowledge.
- A picture may paint a thousand words, BUT a picture can also mislead.



Reality: An old wrinkled woman.

Readings

How NOT to Lie with Visualisation (https://pdfs.semanticscholar.org/058e/2e38420b61d8d870590d971d4e7d1cd078c2.pdf)

14 Ways to Say Nothing with Scientific Visualisation (http://crack.seismo.unr.edu/ftp/vis/14ways.pdf)

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

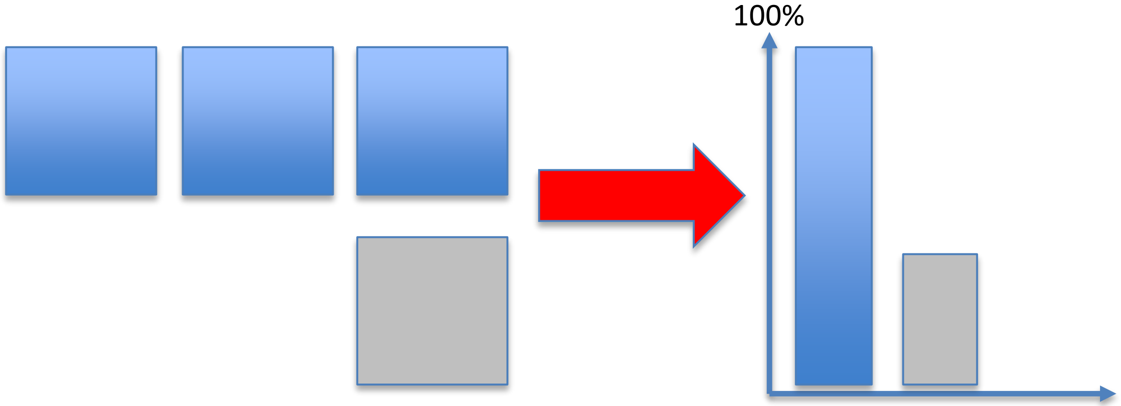
Good and Bad Graphs

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Small Data Size

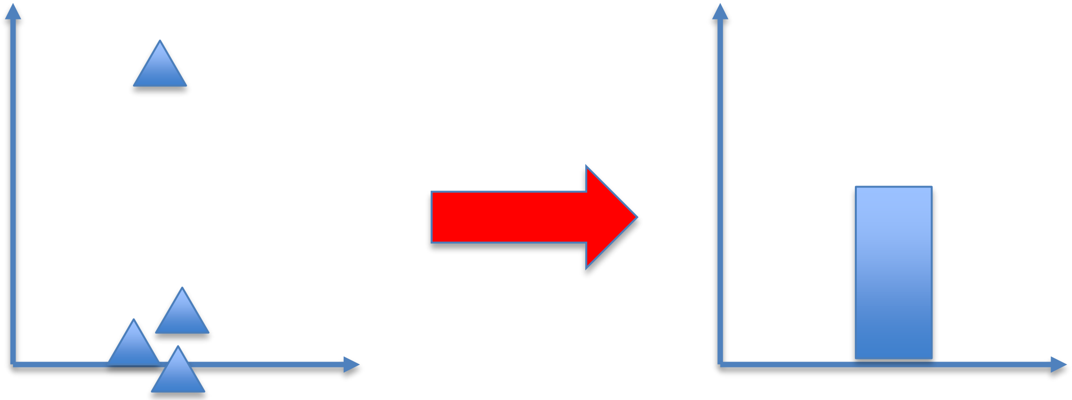
- It does not make sense to use graphs to display very small amounts of data.
- The human brain is quite capable of grasping one two, or even three values.



Data Quality

- Graphs are only as good as the data they display.
- No amount of creativity can produce a good graph from dubious data.
- Anything less would be trying to lie via misleading representation.

Data Quality



Data with no obvious signal
and 1 outlier

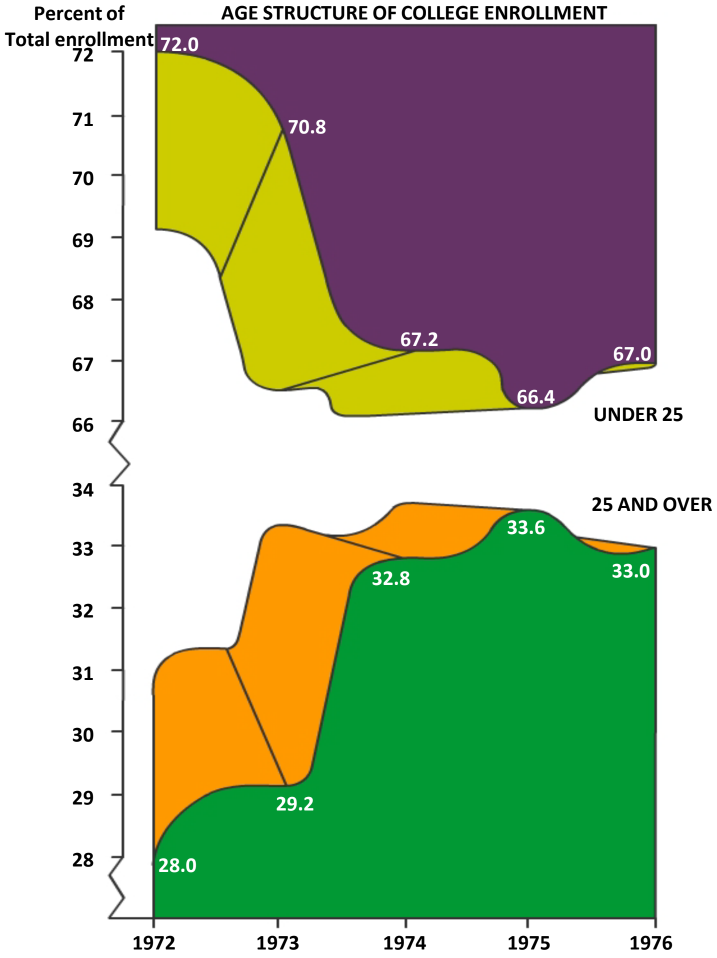
Summarisation as a barchart
to hide that obvious fact

Complexity

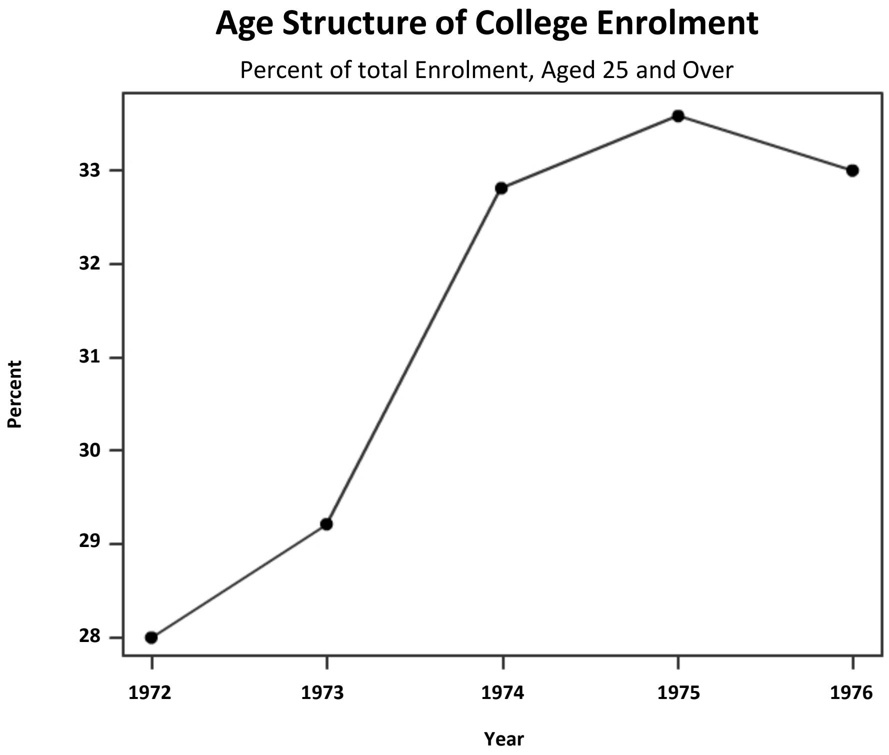
- Graphs should be no more complex than the data which they portray.

- Unnecessary complexity can be introduced by:
    - Irrelevant Decoration
    - Colour
    - 3D Effects

Complexity



Age Structure of College Enrolment (1972-1976)



Both graphs present the same data

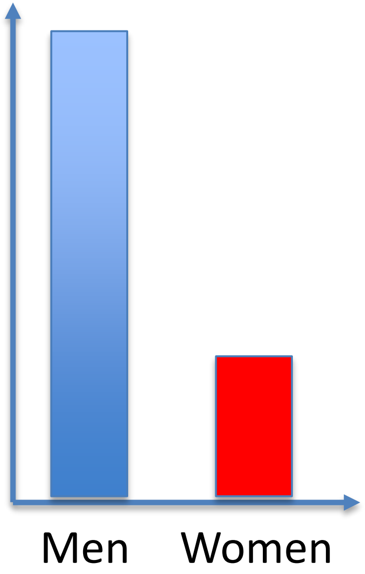
Distortions

- Graphs should not provide a distorted picture of the values they portray.

- Distortion can be either deliberate or accidental (especially if one do not really understand the data).

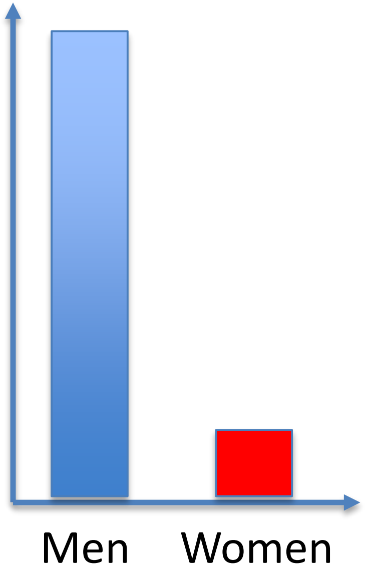
Distortions (Uncovering Hidden Context)

Unequal gender representation in an office implies gender discrimination?



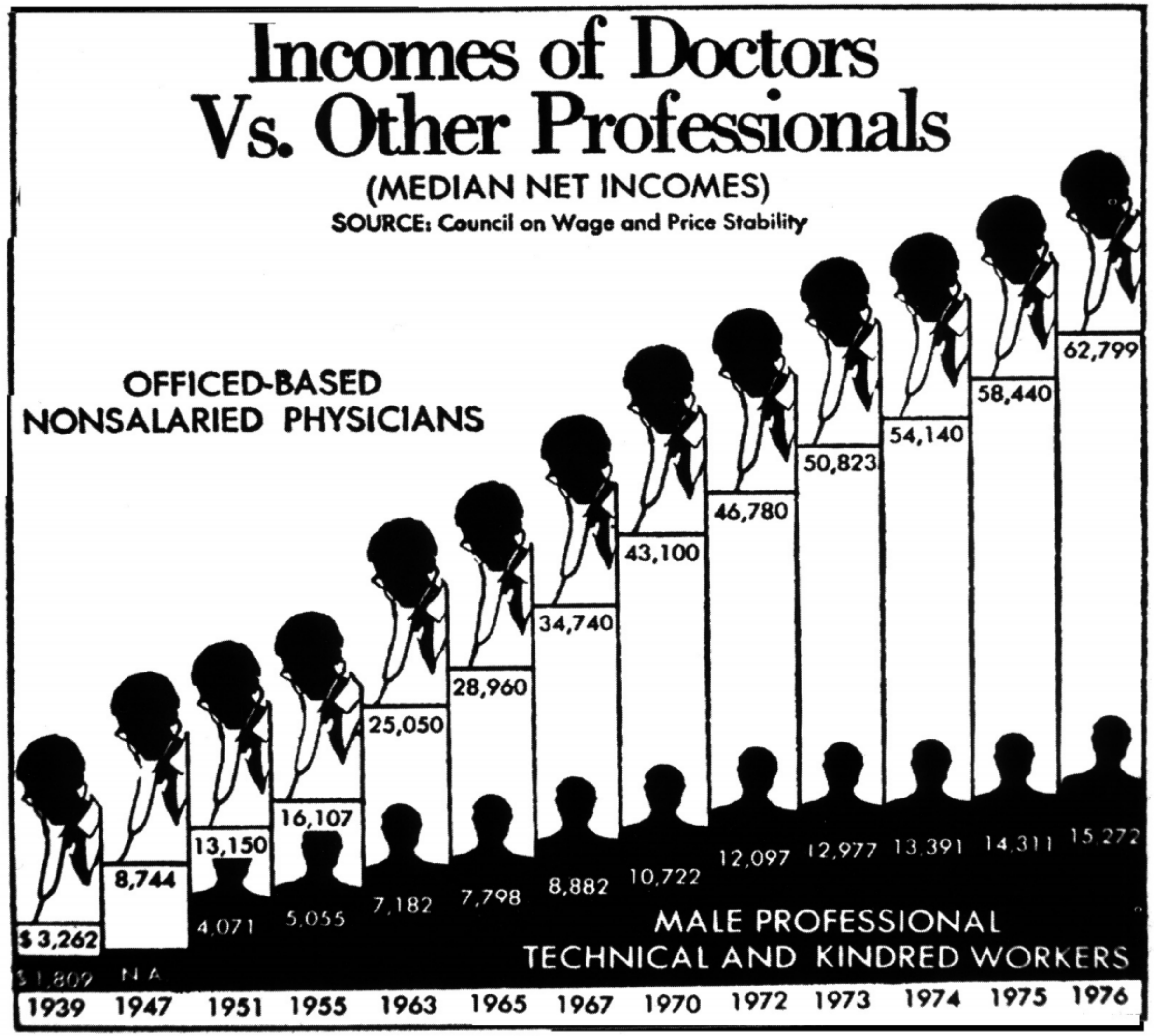
Seems like we need to do something!

Interviewee proportion (split by gender)



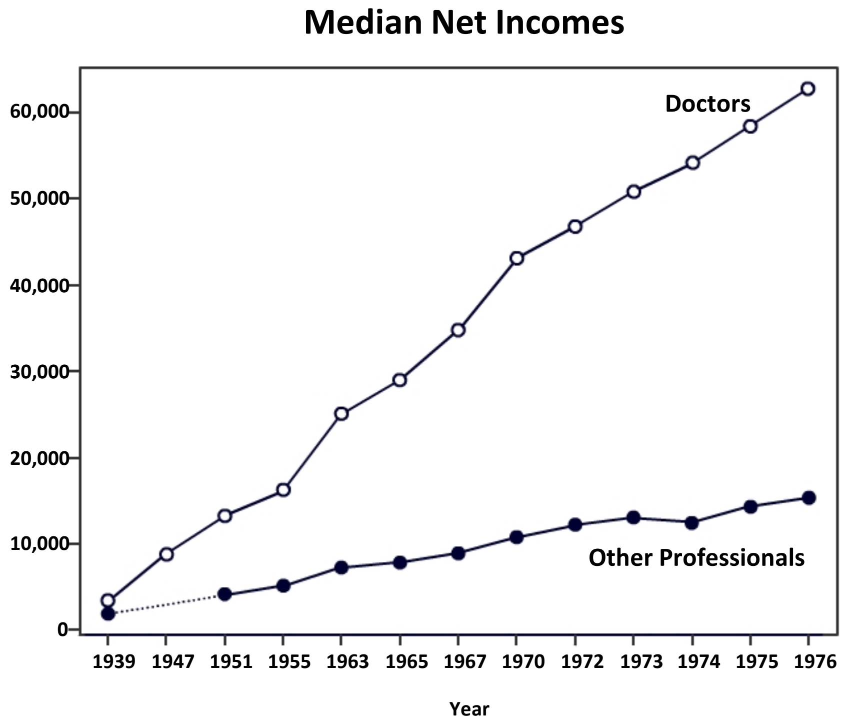
It seems that there is deliberate attempts to increase female representation despite the very low gender representation amongst interviewees

Distortions (Inappropriate Use of Linear Scaling)

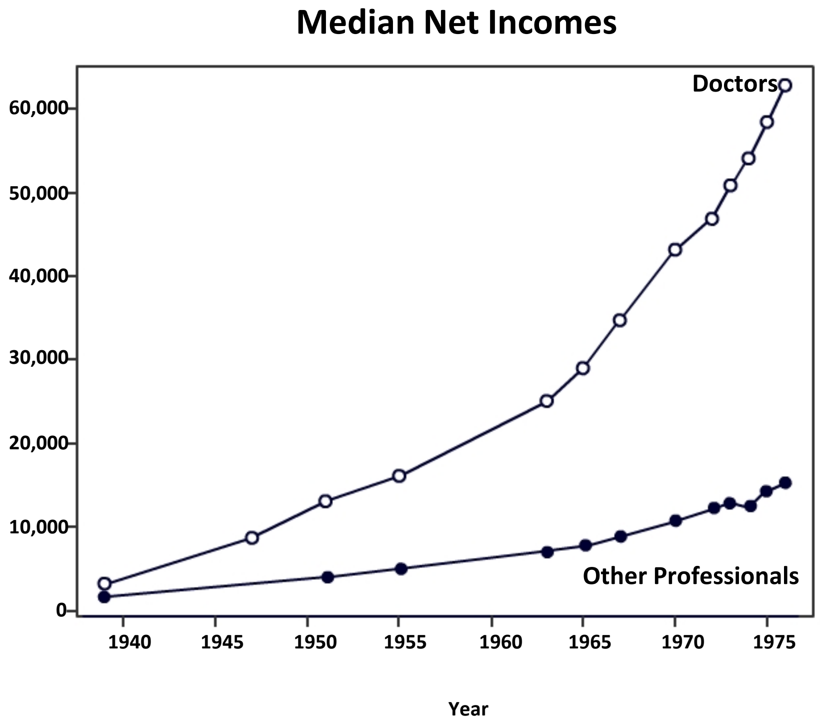


Source (under creative commons): http://maths.nayland.school.nz/Year_11/biased_graphs.htm

Distortions (Inappropriate Use of Linear Scaling)



Seems pretty innocuous. But look carefully again at the x-axis intervals.



Once the intervals are now pretty aligned. You can see a disturbing trend.

Generic Guides for Good Graphing

Draw the graph with an aim to communicate.

- If the “story” is simple, keep it simple.
- Ensures that axes, legends, annotations are fully visible.

If the “story” is complex, make it look simple.

- The aim is to draw insight quickly and accurately. If the graph is as complex as the data, it is of limited use.

Avoid distorting the data.

- Don’t use aesthetic features unless it serves useful purpose.
- Understand the context of the date you are representing.
- Don’t “hide” or “lie” by using inconsistent intervals or other visual tricks.

How Bar Charts Mislead

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Simple Tools for Representation of Data

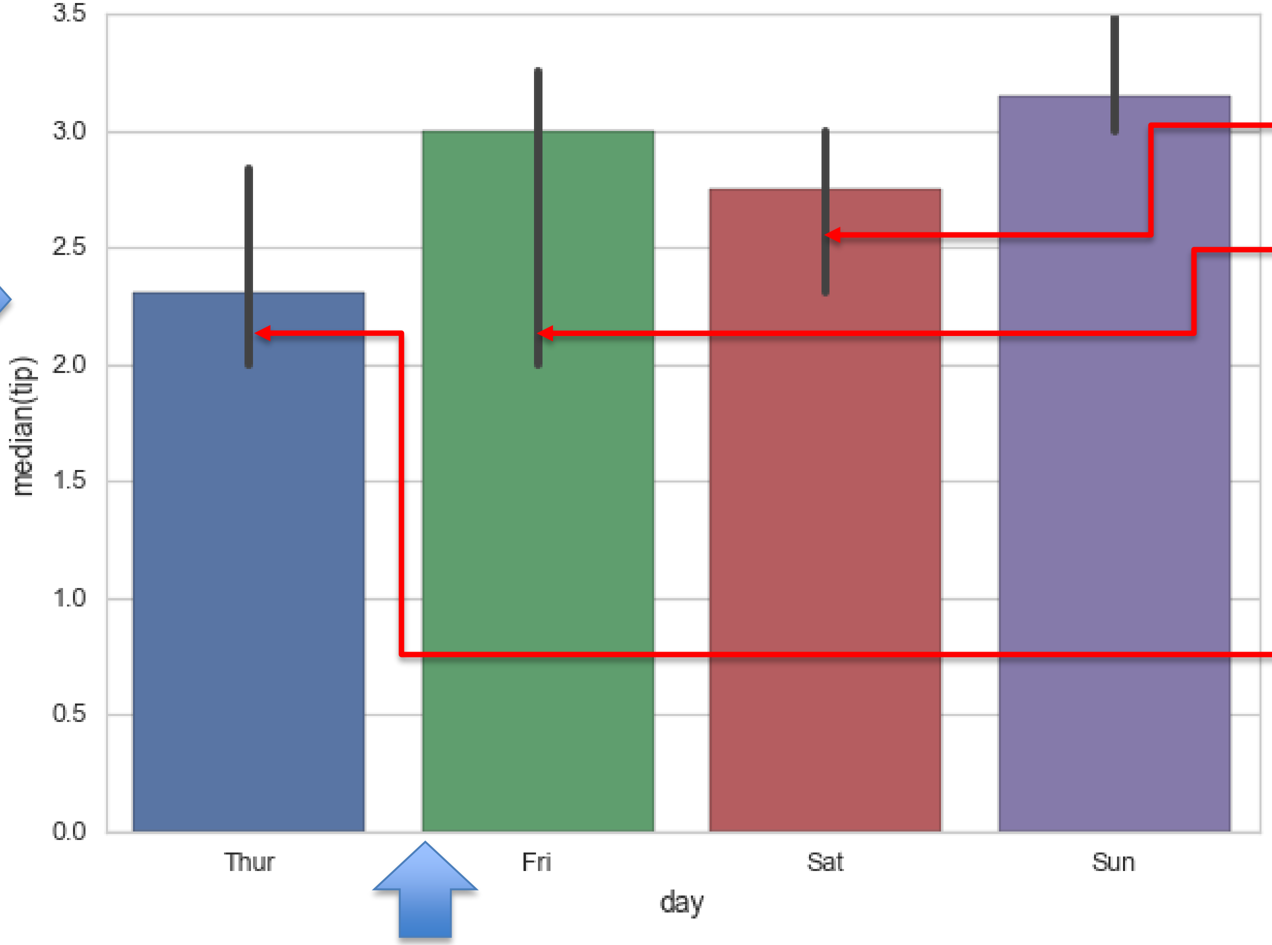
- Although the focus is on the most common way of representing data, the objective is to generalise beyond, and think carefully about how insufficiently rigorous ways of summarising data can lead to misinterpretation.

- This is as true for barcharts, as it is for other data representation tools (including pie-charts, line graphs, and so on).

Bar Charts --- 3 Issues

Elements of a Bar Chart

The y-axis holds numerical data.



Standard Error/ Deviation (optional)

Usually Mean (Sometimes Median)

The x-axis holds categorical variables e.g. control and test, days of a week.

Bar Charts --- 3 Issues

Quick refresher on mean/ median and standard error/ standard deviation

The Arithmetic Mean

$$\bar{X} = \frac{\sum X}{N}$$

The Standard Deviation

$$SD = \sqrt{\frac{\sum(x - \bar{x})^2}{N}}$$

The Median

$$md = x_{\frac{(N+1)}{2}}$$

If N is odd

$$md = \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\frac{(N)}{2}+1}\right)$$

If N is even

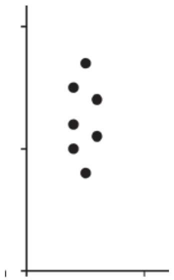
The Standard Error

$$SE = \frac{SD}{\sqrt{N}}$$

Bar Charts --- 3 Issues

The bar chart is guilty of over-simplification (but ~90% publications use it). The bar chart is really just showing summary statistics (the mean/median) and/or s.d./s.e. inferred from the entire data!

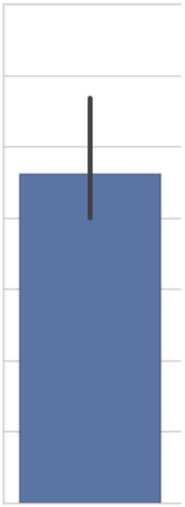
Original



Summary Statistics

$$\overline{X} = \frac{\sum X}{N}$$

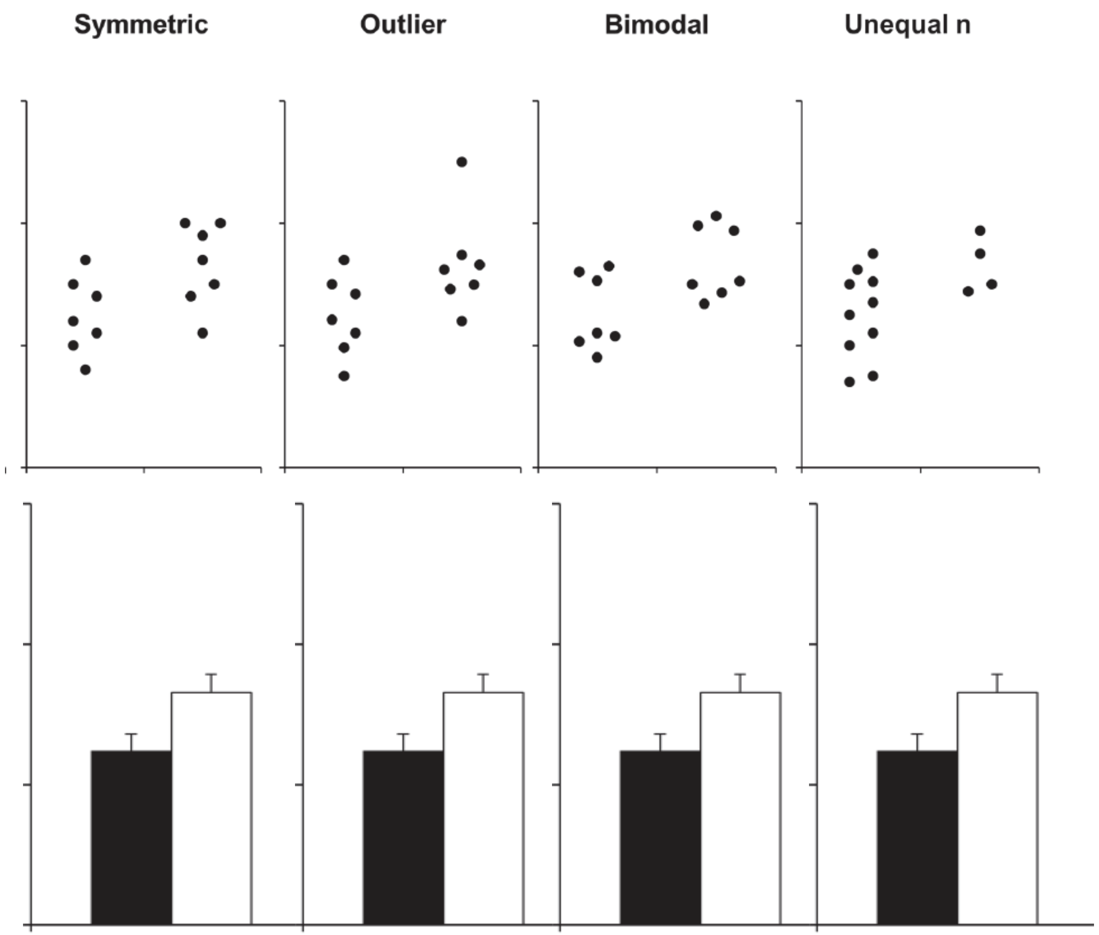
$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$$



Do summary statistics really need to be shown as graphs? Are we losing critical information? Are we creating false information? Let’s see.

Bar Charts --- 3 Issues

Issue 1



Consider the following:
What do you think the
corresponding bar charts look like?

**Different distributions
give rise to the same bar charts.
Distribution information is lost
because the bar chart only shows
the summary statistics.**

Beyond bar and line graphs: time for a new data presentation paradigm. Weissgerber TL, Milic NM, Winham SJ, Garovic VD. PLoS Biol. 2015 Apr 22;13(4):e1002128. doi: 10.1371/journal.pbio.1002128. eCollection 2015 Apr.

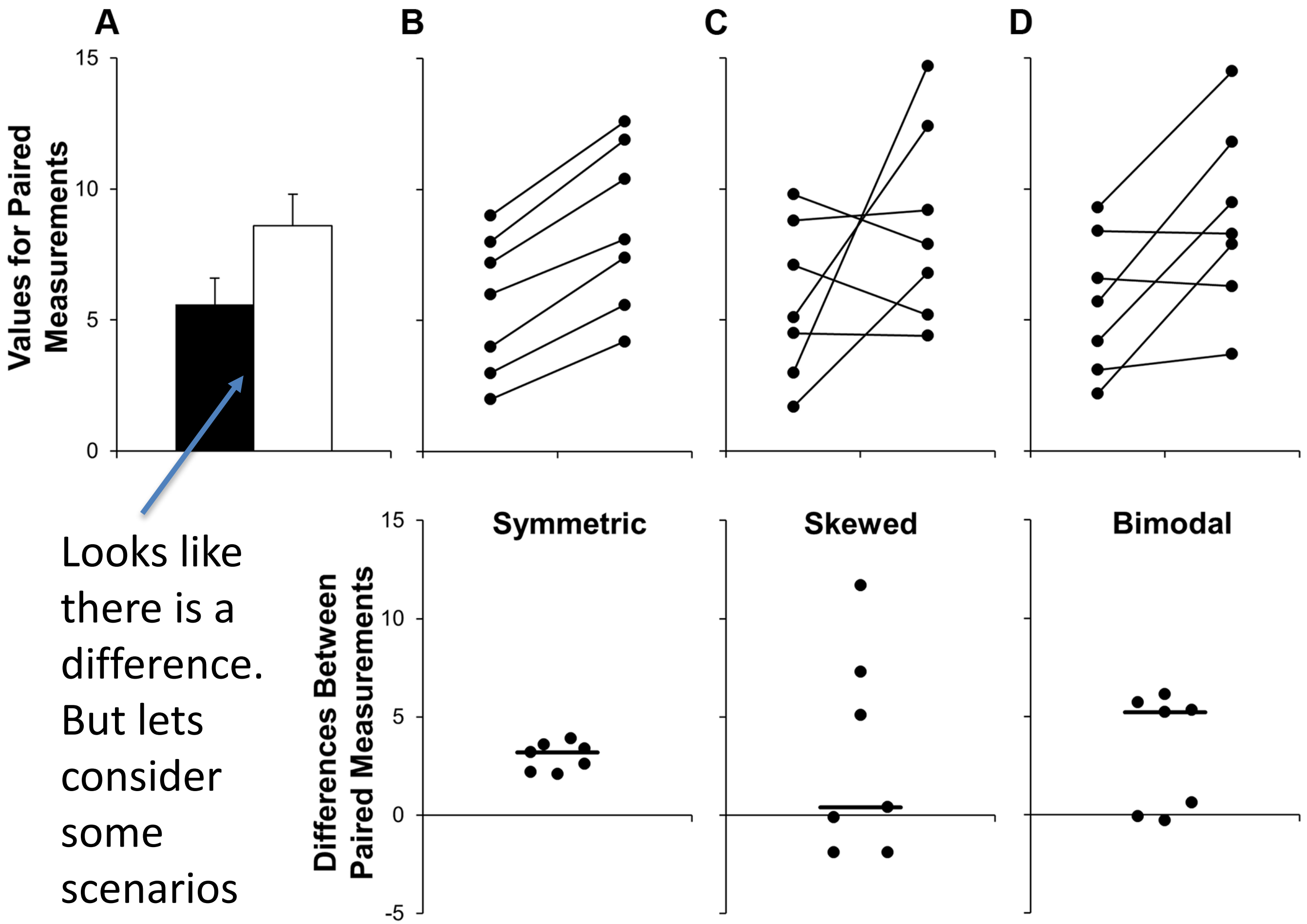
Bar Charts --- 3 Issues

Issue 1

- Many different data distributions leads to the same bar chart.
- The full data suggests different conclusions from the summary statistics.
- If you relied solely on the bar chart and never checked the full data distribution prior, you may be in for a bad surprise later.
- Always check your univariate data distribution first before relying on summary statistics.

Bar Charts --- 3 Issues

Issue 2



Looks like there is a difference. But lets consider some scenarios

Pairable means there is a one-to-one correspondence between two sets of measurements.

Beyond bar and line graphs: time for a new data presentation paradigm. Weissgerber TL, Milic NM, Winham SJ, Garovic VD. PLoS Biol. 2015 Apr 22;13(4):e1002128. doi: 10.1371/journal.pbio.1002128. eCollection 2015 Apr.

Bar Charts --- 3 Issues

Issue 2

- Separate bar charts should not be used on pairable data.

- Bar charts of paired data falsely suggest that the groups being compared are independent and provide no information about whether changes are consistent across individuals.

- Instead, you should plot the distribution of the deltas (paired differences) for individuals.

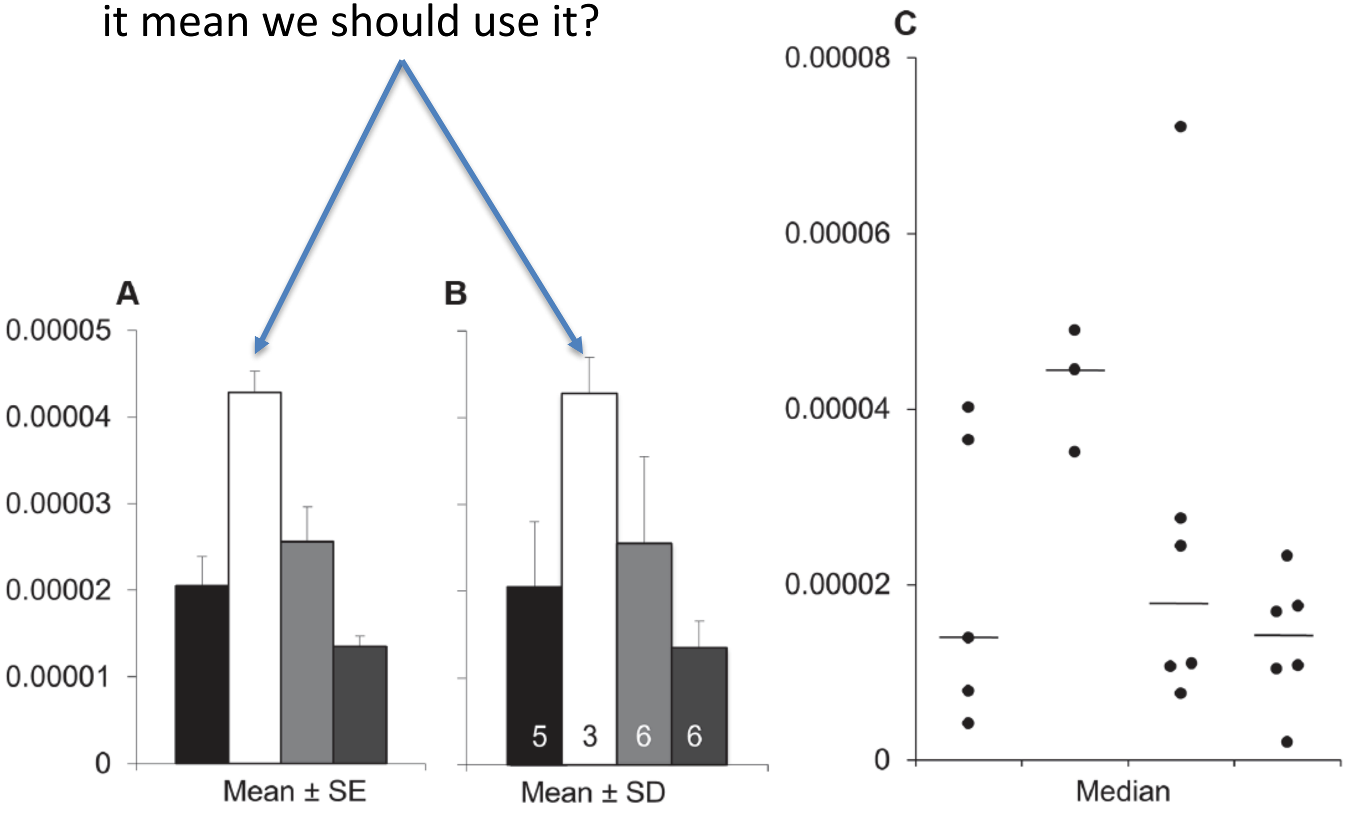
Bar Charts --- 3 Issues

Issue 3

SE will always be smaller than SD (and looks better) but does it mean we should use it?

The bar charts also look symmetrical but…

Your bar charts can be used to lie about your data centers and distributions. It can also be used to hide small sample sizes.



Beyond bar and line graphs: time for a new data presentation paradigm. Weissgerber TL, Milic NM, Winham SJ, Garovic VD. PLoS Biol. 2015 Apr 22;13(4):e1002128. doi: 10.1371/journal.pbio.1002128. eCollection 2015 Apr.

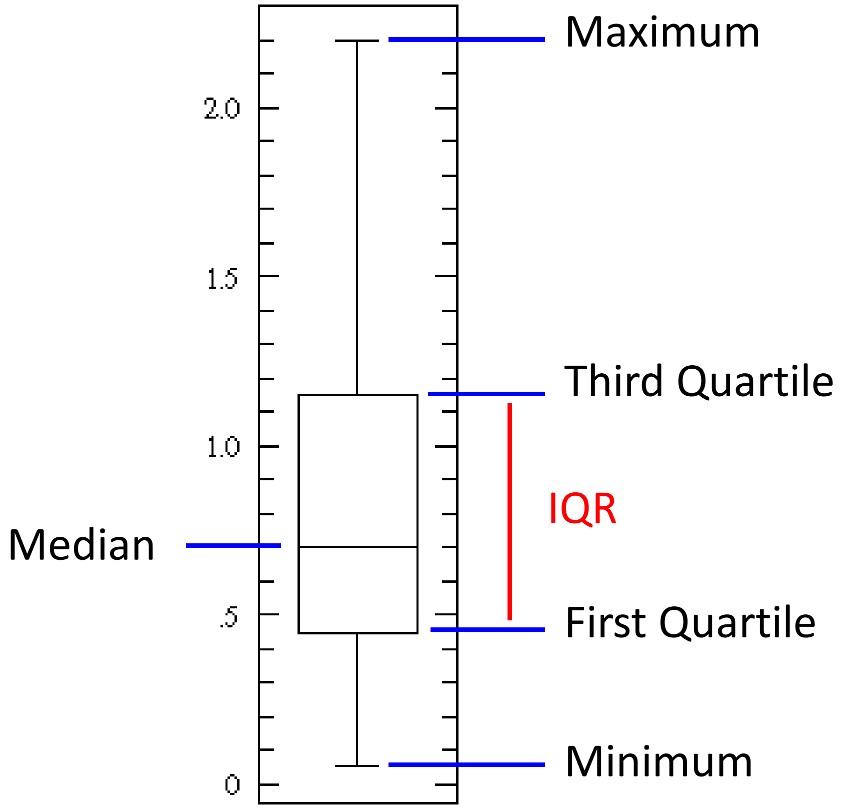
Bar Charts --- 3 Issues

Issue 3

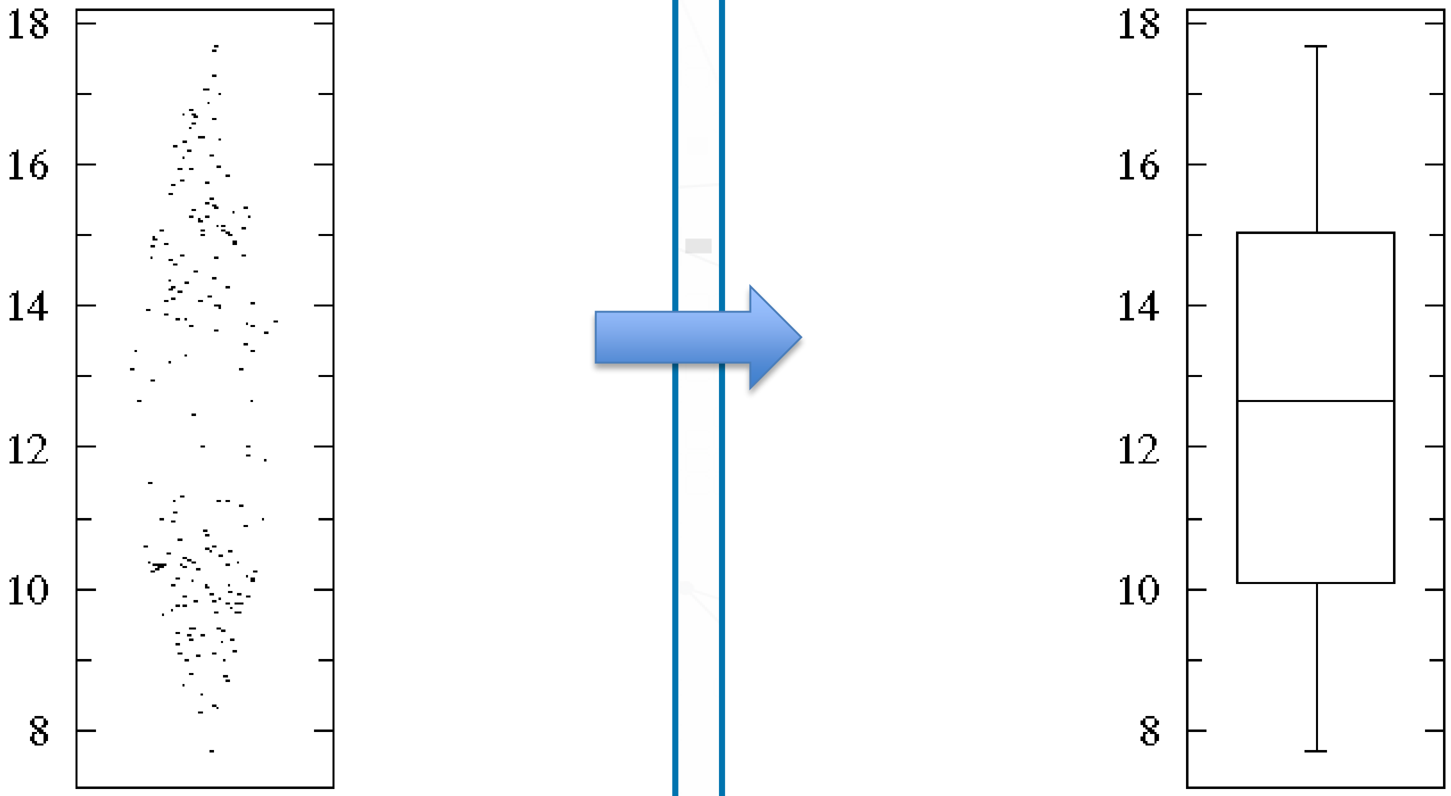
- False impressions: Showing the SE (rather than the SD) magnifies the apparent visual differences between groups.
- This effect is exacerbated when the groups being compared have different sample sizes.
- The bar chart also makes the data appear symmetrical, when in fact it is not.

Boxplot

- The boxplot is another common way of summarising data.
- It shows more information than a bar chart.
- Is it effectively better than the bar chart?



The boxplot is also susceptible to misleading representation



Although it shows more, the boxplot is no substitute for the univariate scatterplot.

Boxplot

Advantages

- Handles large data easily.
- Shows more summary statistics than a bar chart (median, non-symmetry, IQR, and potential outliers).

Disadvantages

- May hide true data distribution (e.g. bimodal data).
- Does not work well with small sample size.

Some Generic Good Analytical Practices (GAPs)

- For **small sample size** (< 5), summary statistics are not meaningful -> Use scatterplots.
- Check the **actual distribution** of individual data points (do not skip right to summary statistics).
- Use the **median** rather than the mean to identify the center of your data.
- Always check for **outliers, non-symmetry, hidden subpopulations**, and handle them accordingly.
- Never apply **statistical tests** before checking the data distribution.

To learn how to draw univariate scatterplots in Excel go to:
https://www.ctspedia.org/do/view/CTSpedia/TemplateTesting

Readings/ References

Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLoS Biol 13(4): e1002128. doi:10.1371/journal.pbio.1002128

Cooper RJ, Schriger DL, Close RJ (2002) Graphical literacy: the quality of graphs in a large-circulation journal. Annals of emergency medicine 40: 317–322

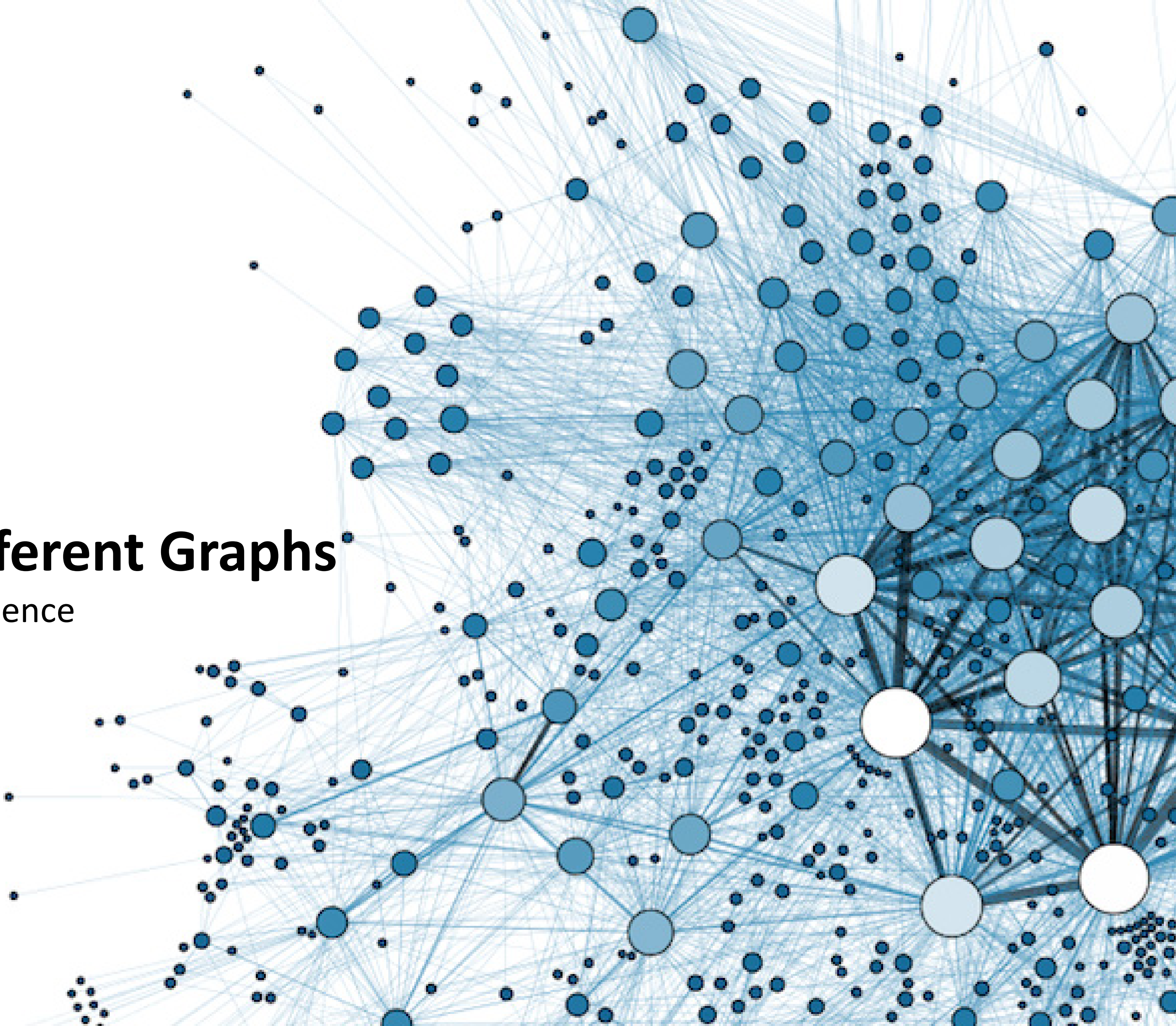
Schriger DL, Sinha R, Schroter S, Liu PY, Altman DG (2006) From submission to publication: a retrospective review of the tables and figures in a cohort of randomised controlled trials submitted to the British Medical Journal. Annals of emergency medicine 48: 750–756

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Same Statistics, Different Graphs

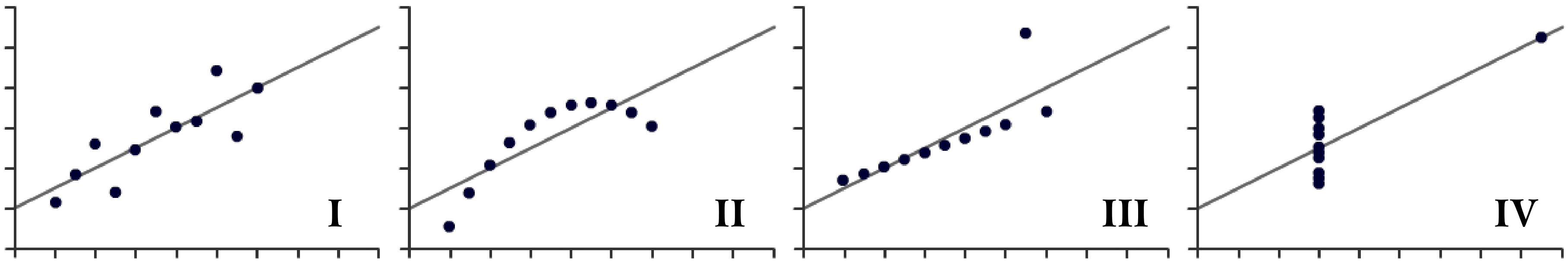
BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences



Anscombe’s Quartet

- Anscombe’s Quartet is a set of four distinct datasets each consisting of 11 (*x*,*y*) pairs.
- Each dataset produces the same summary statistics (mean, standard deviation, and correlation) while producing vastly different plots.



Anscombe, F.J. (1973). Graphs in Statistical Analysis. *The American Statistician 27*, 1, 17–21

Anscombe’s Quartet

- This dataset is frequently used to illustrate the importance of graphical representations when exploring data.
- Four *clearly different* and *identifiably distinct* datasets are producing the same statistical properties.

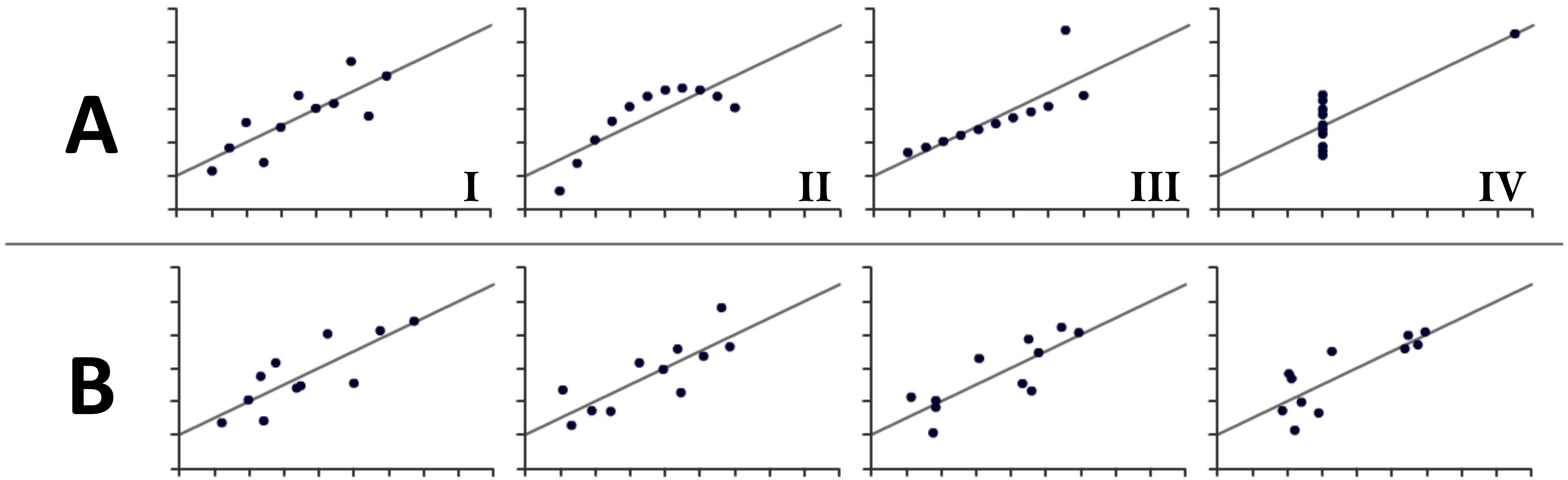
https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet

- The implication is that you cannot rely only on numerical summaries to interpret your data.
- Plotting and checking your data distributions visually is important and should constitute a important part of good analytical practice.

Source: https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet



Generating datasets with varied appearance and identical statistics through simulated annealing.

Series A are the Anscombe’s quartet. Series B are randomly generated data points taken from the summary statistics.

$$(\bar{x} = 54.02, \bar{y} = 48.09, sdx = 14.52, sdy = 24.79, Pearson's\ r = \ +0.32)$$

But series B does not exhibit any clear substructure while series A distributions are quite limited in variety. Can this point be made clearer?

Anscombe’s Quartet

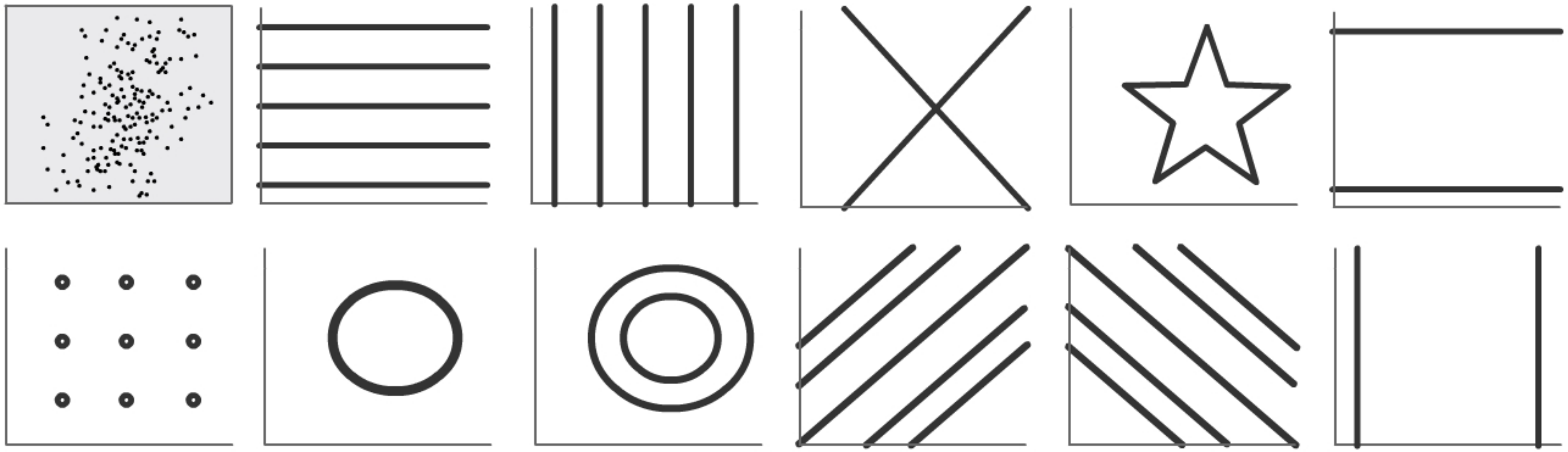
Generating datasets with varied appearance and identical statistics through simulated annealing. It is relatively *easy* to take an existing dataset, modify it slightly, and maintain (nearly) the same statistical properties.

```
current_ds ← initial_ds
for x iterations, do:
    test_ds ← PERTURB(current_ds, temp)
    if ISERROROK(test_ds, initial_ds):
        current_ds ← test_ds

function PERTURB(ds, temp):
    loop:
        test ← MOVERANDOMPOINTS(ds)
        if FIT(test) > FIT(ds) or temp > RANDOM():
            return test
```

https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet

You can re-fit the dataset to almost any “template” while largely preserving the overall summary statistics.

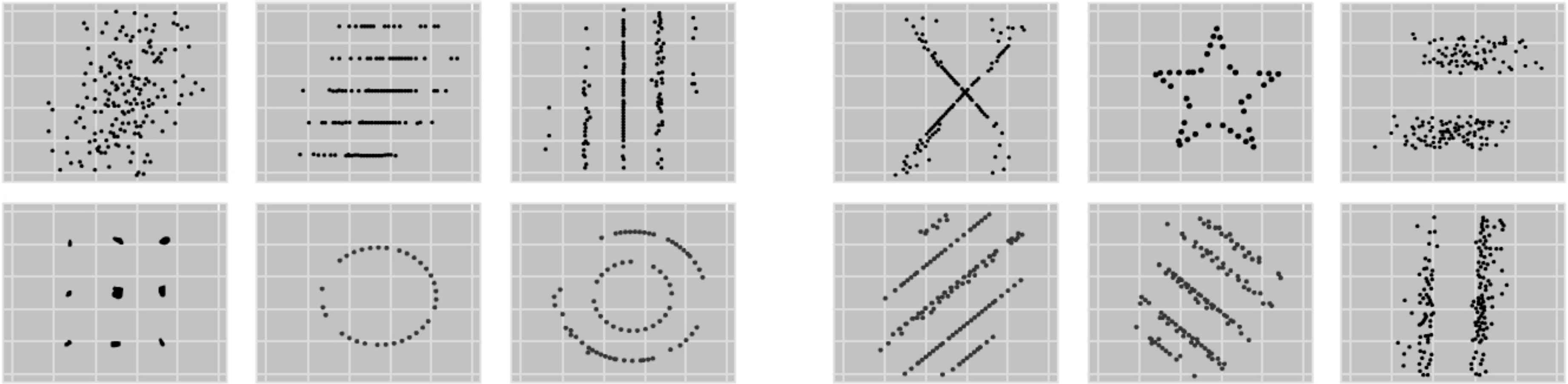


The initial data set (top-left), and line segment collections used for directing the output towards specific shapes.

https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet

Data is surprisingly malleable while preserving the global summary statistics. Algorithm ran for 200,000 iterations to achieve the final results.

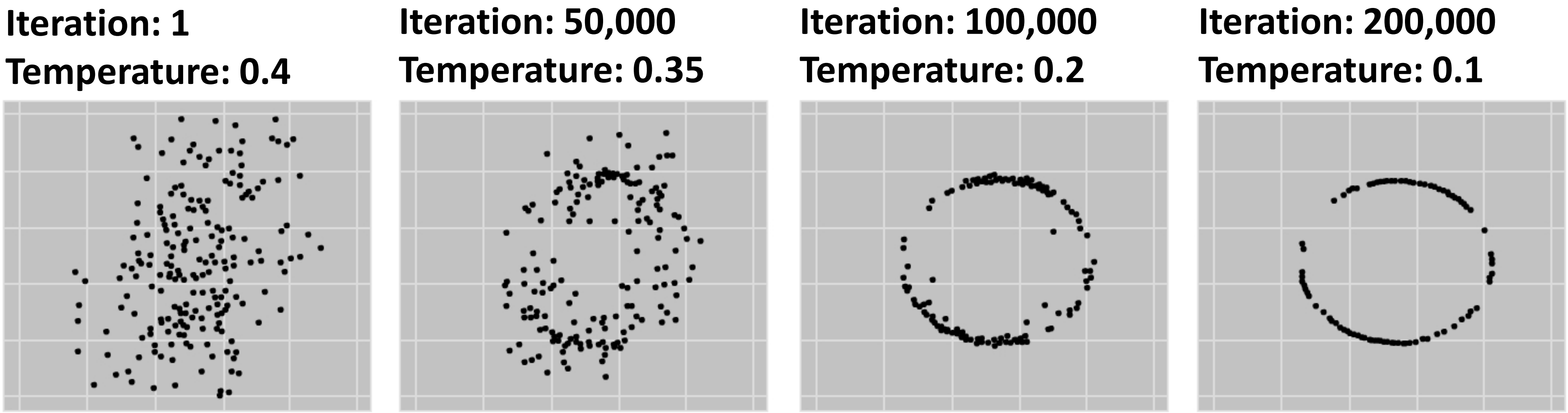


While different in appearance, each has the same summary statistics (mean, std. deviation, and Pearson’s corr.) to 2 decimal places.

https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet

Repeated iterations are needed to improve the fit of the data to the template.



Iteration: 1
Temperature: 0.4

Iteration: 50,000
Temperature: 0.35

Iteration: 100,000
Temperature: 0.2

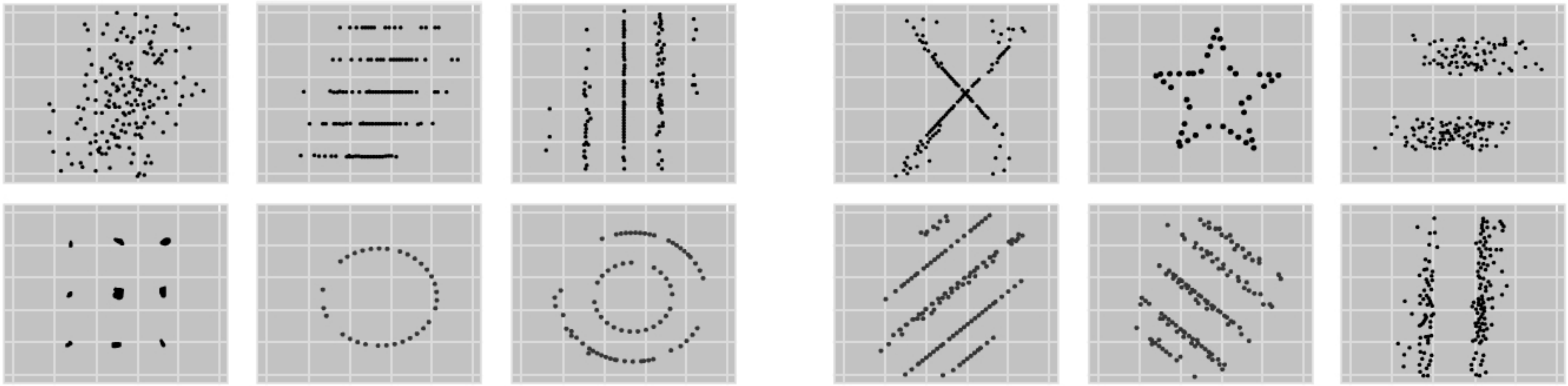
Iteration: 200,000
Temperature: 0.1

You can basically make the data look like anything you want while retaining the same overall statistical measures not just the typical parametric measure such as mean and sd., but also including non-parametric measures of *x/y median*, *x/y interquartile range (IQR)*, and *Spearman’s rank correlation coefficient*.

https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet

https://youtu.be/It4UA75z_KQ



https://dl.acm.org/citation.cfm?id=3025912

Anscombe’s Quartet

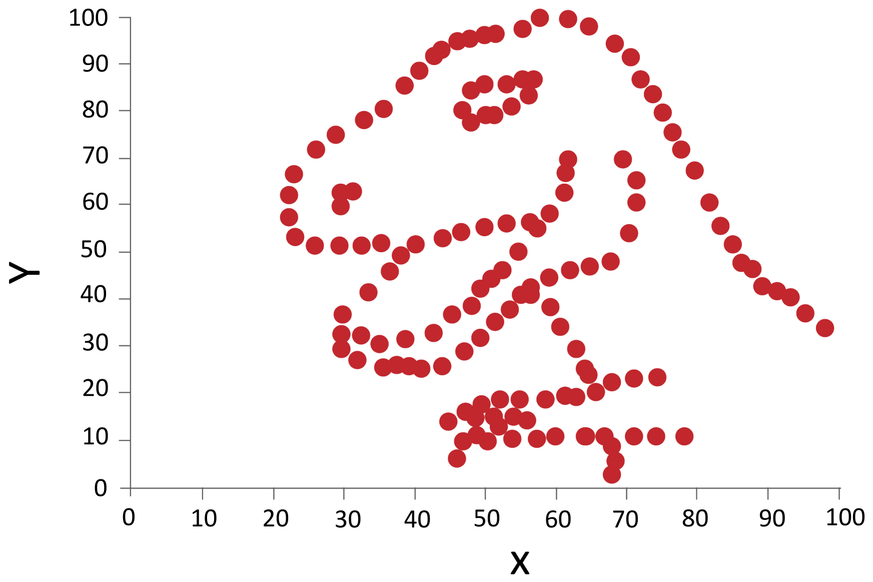
It is not just “generic” looking data that is amendable to this approach. This approach of completely changing the visuals of a data works not just with generic looking data. But can also be applied on data with a very specific “look”.

Produced by Alberto Cairo. The Anscombosaurus Rex generates “normal” summary statistics, but is actually a “dinosaur”.

Plot it for yourself [here](#).

Introducing the Anscombosaurus Rex

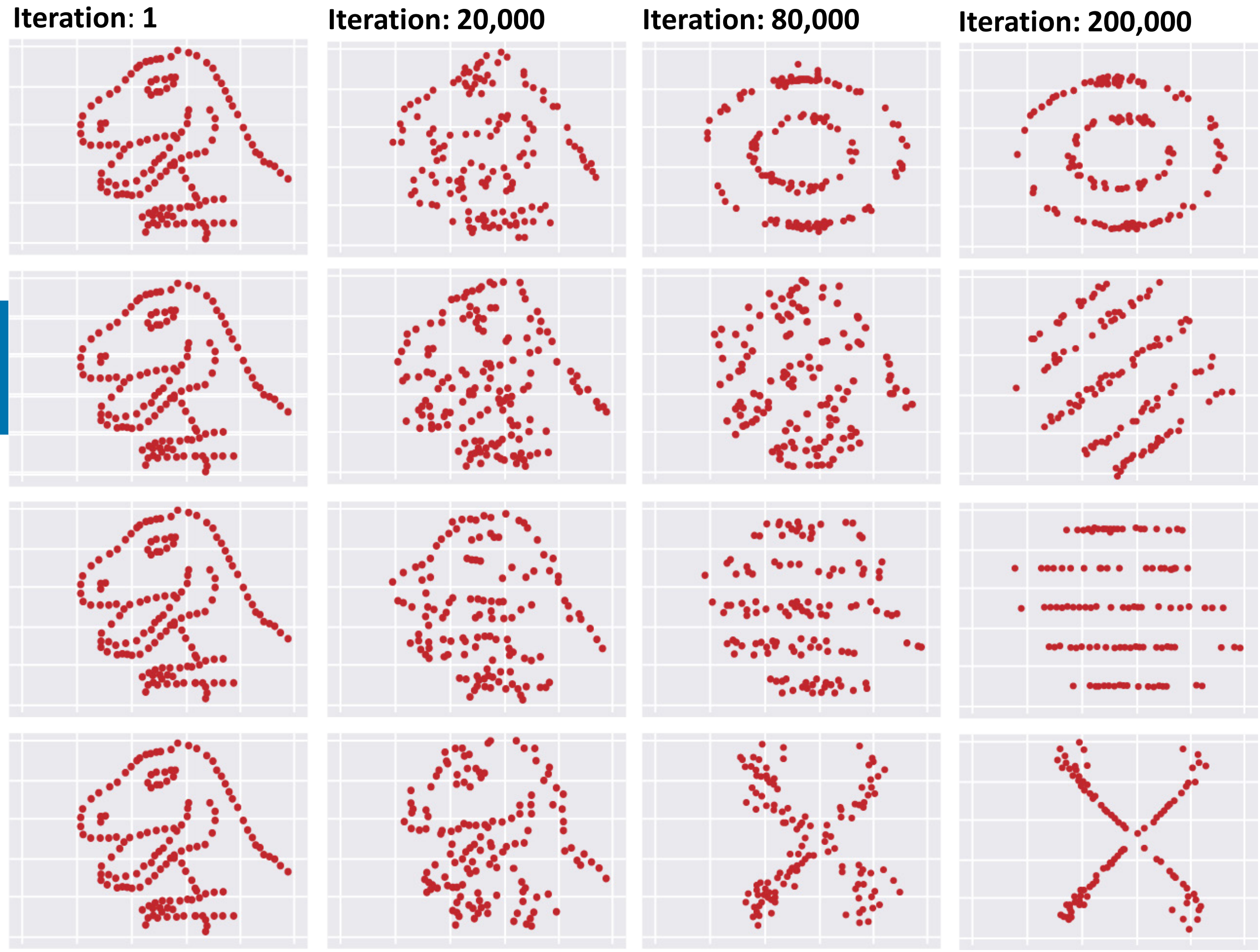
N = 142 ; X mean = 54.2633 ; X SD = 16.7651 ; Y mean = 47.8323 ; Y SD = 26.9354 ; Pearson correlation = -0.0645



http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

Anscombe’s Quartet

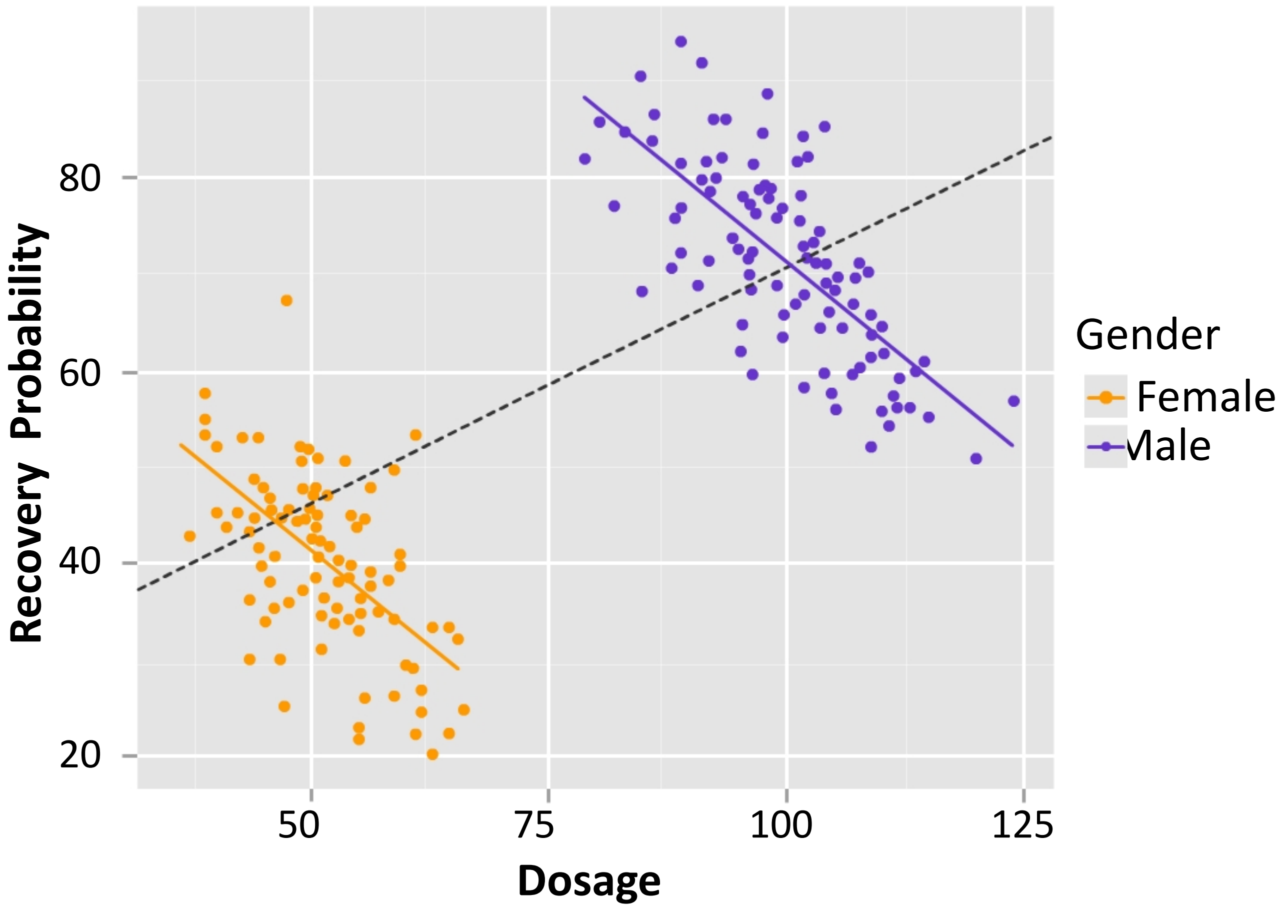
Refitting the Anscombosaurus to any of the templates also works.



https://dl.acm.org/citation.cfm?id=3025912

Simpson’s Paradox

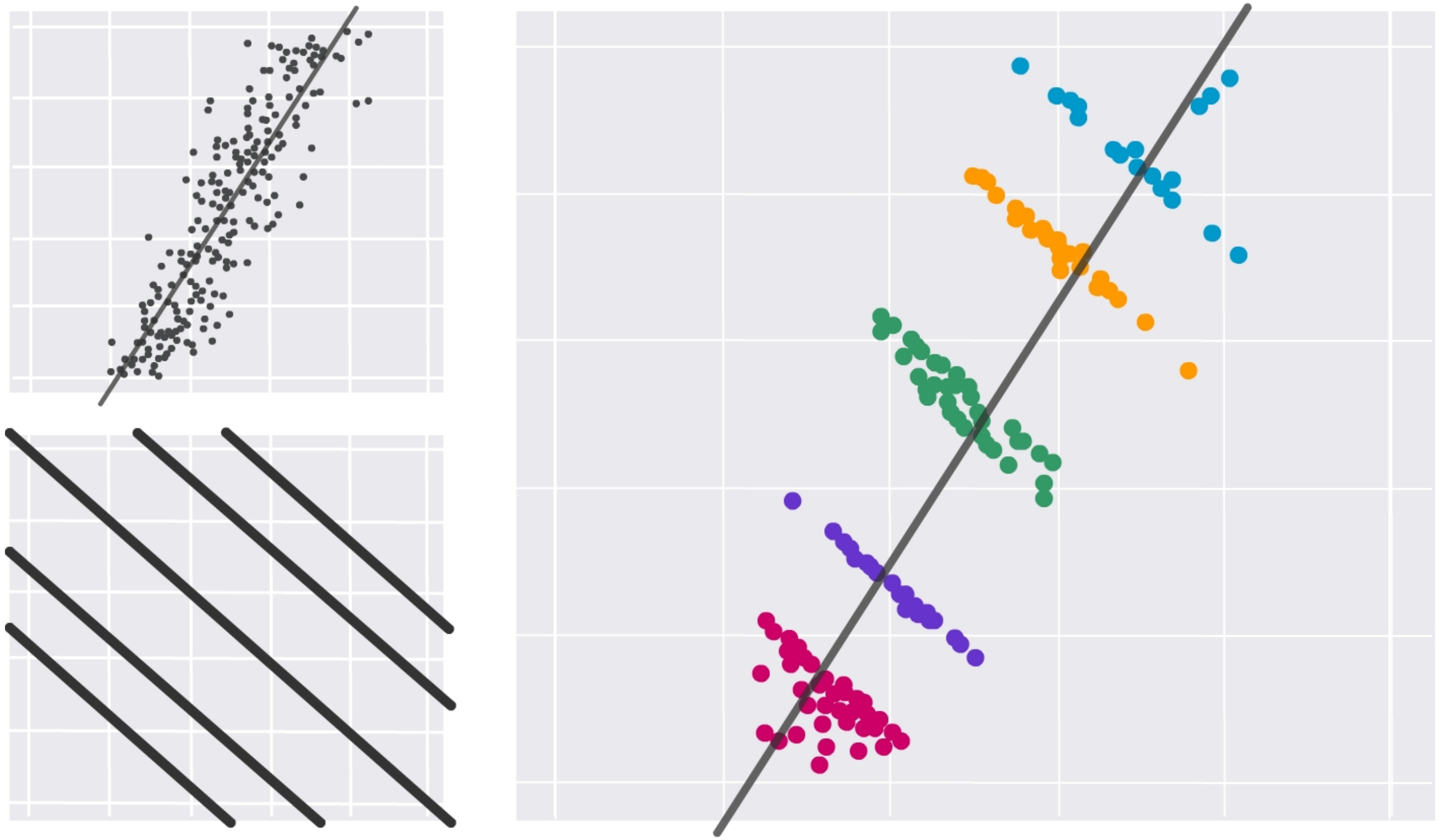
Simpson’s paradox occurs with data sets where a trend appears when looking at individual groups in the data, but disappears or reverses when the groups are combined.



https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00513/full

Simpson’s Paradox

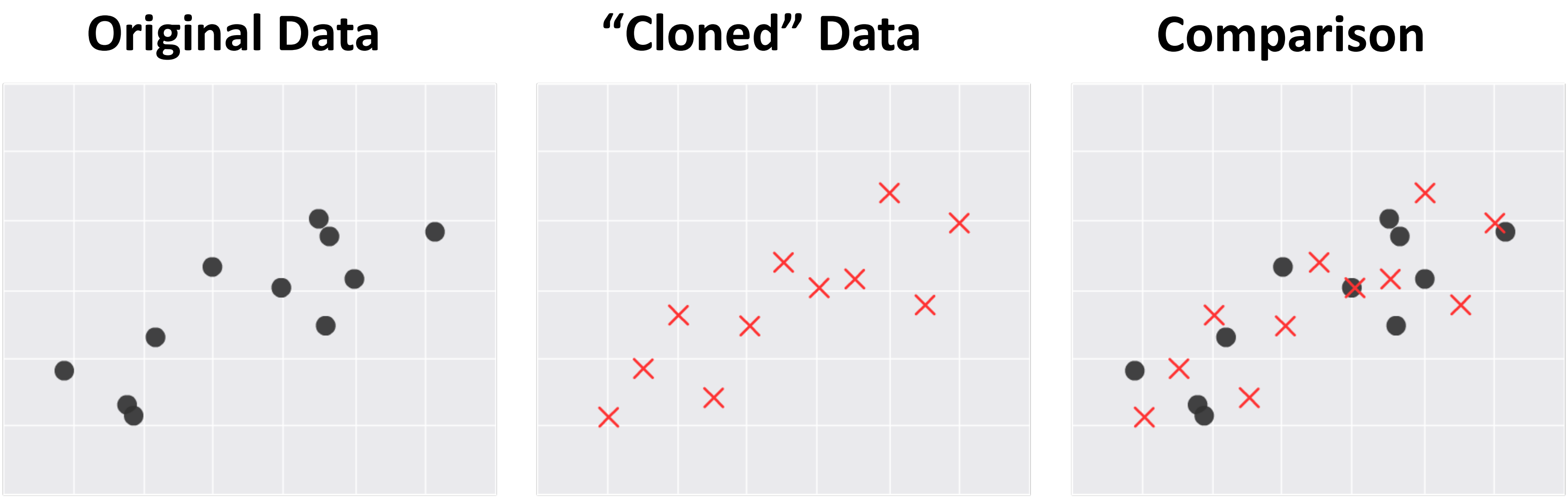
Both datasets (A and C) have the same overall Pearson's correlation of +0.81.



A: Original B: Template C: Simulated Outcome

https://dl.acm.org/citation.cfm?id=3025912

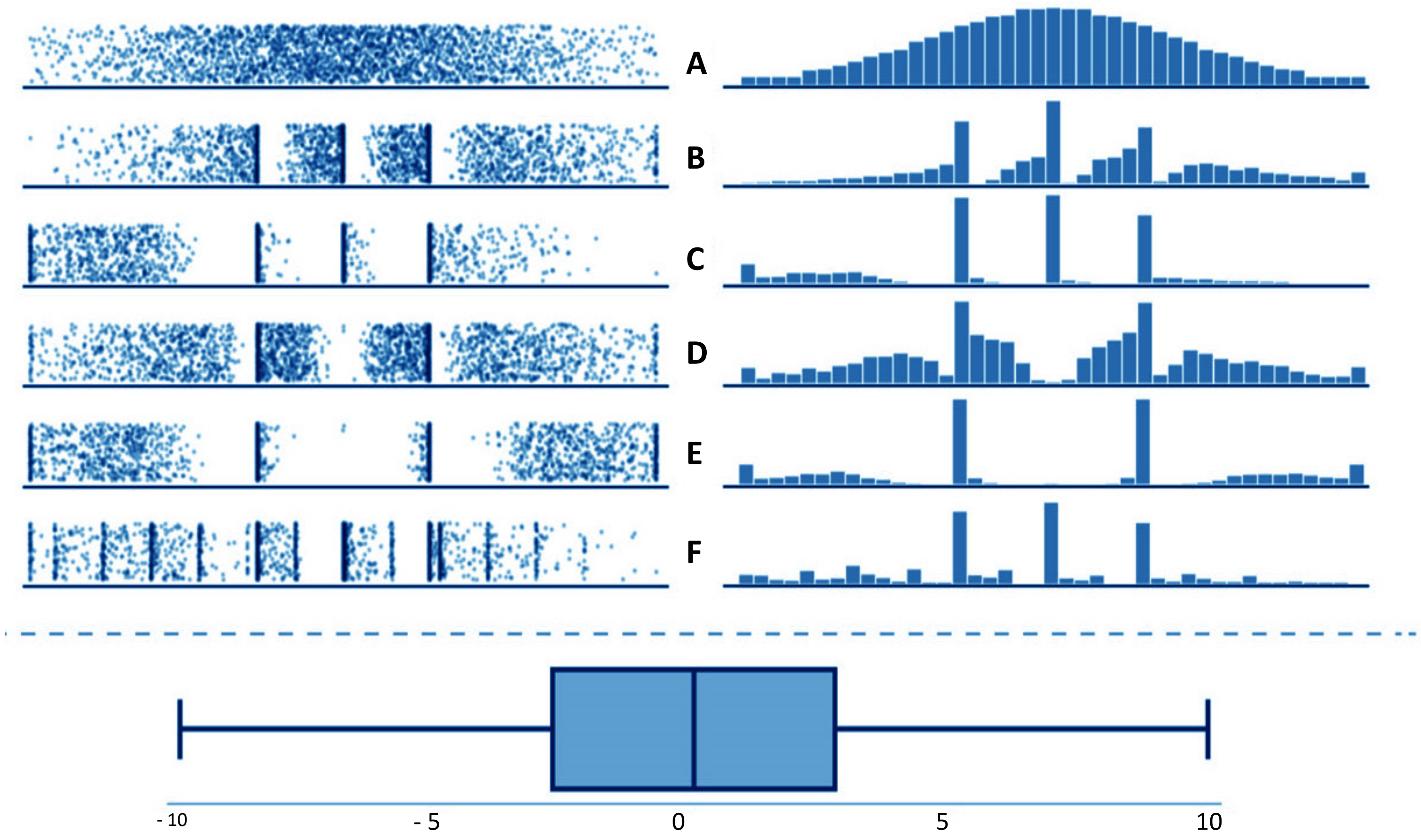
Generating Null or Cloning Datasets



“Cloning” datasets to anonymise sensitive data.

Using this approach to generate more data points.

1-Dimensional Simulations



Six data distributions, each with the same 1st quartile, median, and 3rd quartile values, as well as equal locations for points 1.5 IQR from the 1st and 3rd quartiles. Each dataset produces an identical boxplot.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Summary

BS0004 Introduction to Data Science

Dr Wilson Goh
School of Biological Sciences

Summary

1. It is not difficult for data with very different distributions to generate very similar global statistical measures.

2. We should always inspect the data visually before deciding what to do with it analytically.