Hands-on session: Python

Research Data Visualisation Workshop

The James Hutton Institute

Leighton Pritchard[1,2,3]
[1]Information and Computational Sciences,
[2]Centre for Human and Animal Pathogens in the Environment,
[3]Dundee Effector Consortium,
The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA

Acceptable Use Policy

The James Hutton Institute

Recording of this talk, taking photos, discussing the content using email, Twitter, blogs, etc. is permitted (and encouraged), providing distraction to others is minimised.

These slides will be made available at http://www.slideshare.net/leightonp

Table of Contents

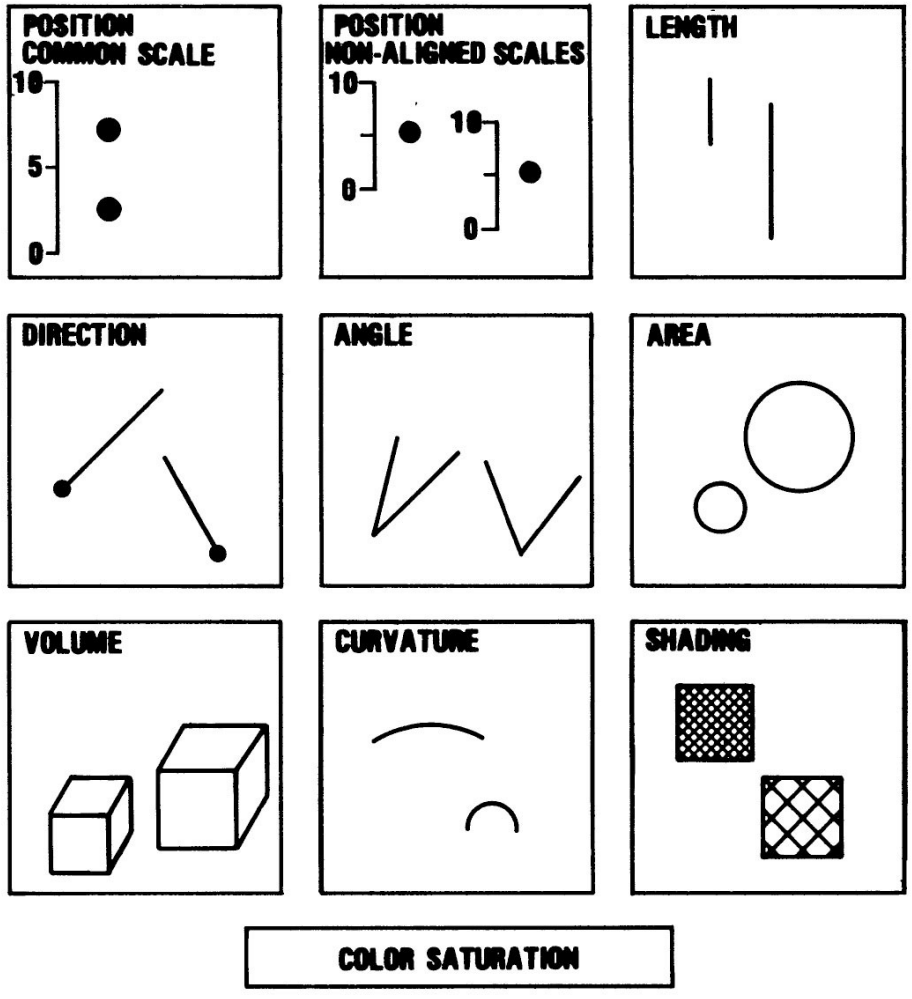
The James Hutton Institute

Elementary Perceptual Tasks [a]

[a] Cleveland & McGill (1984) *J. Am. Stat. Ass.*

The James Hutton Institute

The most basic visual tasks:

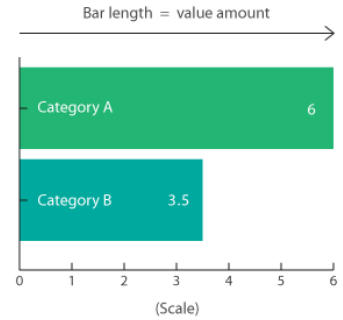


Implementations [a]

[a] http://www.datavizcatalogue.com/

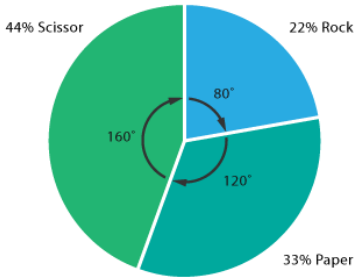
Position: common scale

- Scatterplot
- Bar Chart



Angle

- Pie Chart
- Do(ugh)nut Chart



Curvature

- Arc Diagram
- Chord Diagram

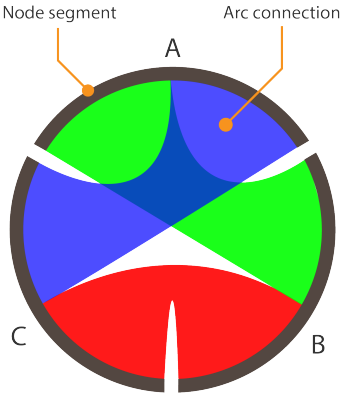


Table of Contents

The James Hutton Institute

What works best? Experiment [a] [b]

[a] Cleveland & McGill (1984) *J. Am. Stat. Ass.*
[b] Heer & Bostock (2010) *CHI 2010*

The James Hutton Institute

Empirical measurements of interpretation

- Subjects shown graphs representing same data
- ($\log_2$) Error in subjects' accuracy compared by graph type

Judgement types

- 1-3: Position on a common scale (bar chart, stacked bar chart)
- 4-5: Length encoding (stacked bar chart)
- 6: Angle (pie chart)
- 7-9: Area (bubble chart, aligned rectangles, treemap)

What works best? Result [a] [b]

[a] Cleveland & McGill (1984) *J. Am. Stat. Ass.*
[b] Heer & Bostock (2010) *CHI 2010*

The James Hutton Institute

- We have inherent biases that can distort information recovered
- Position > Angle ≈ Length > Area
- Accuracy plateaus as charts increase in size
- Gridlines improve accuracy
- Aspect ratios affect area judgements (squares worst)

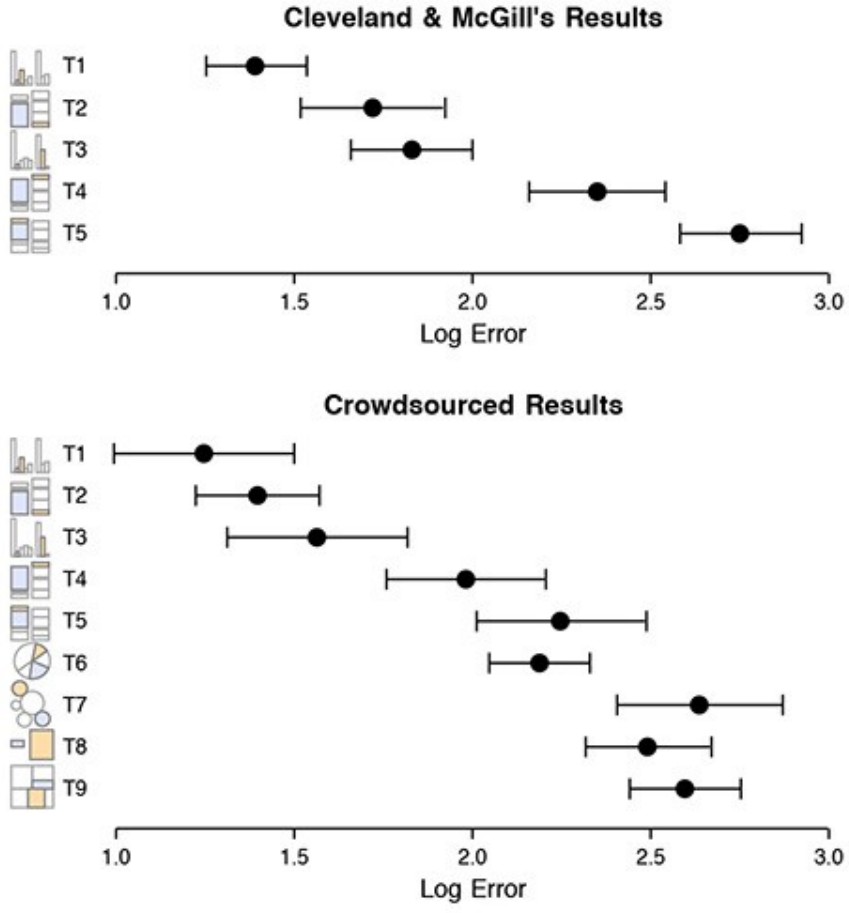


Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.

Table of Contents

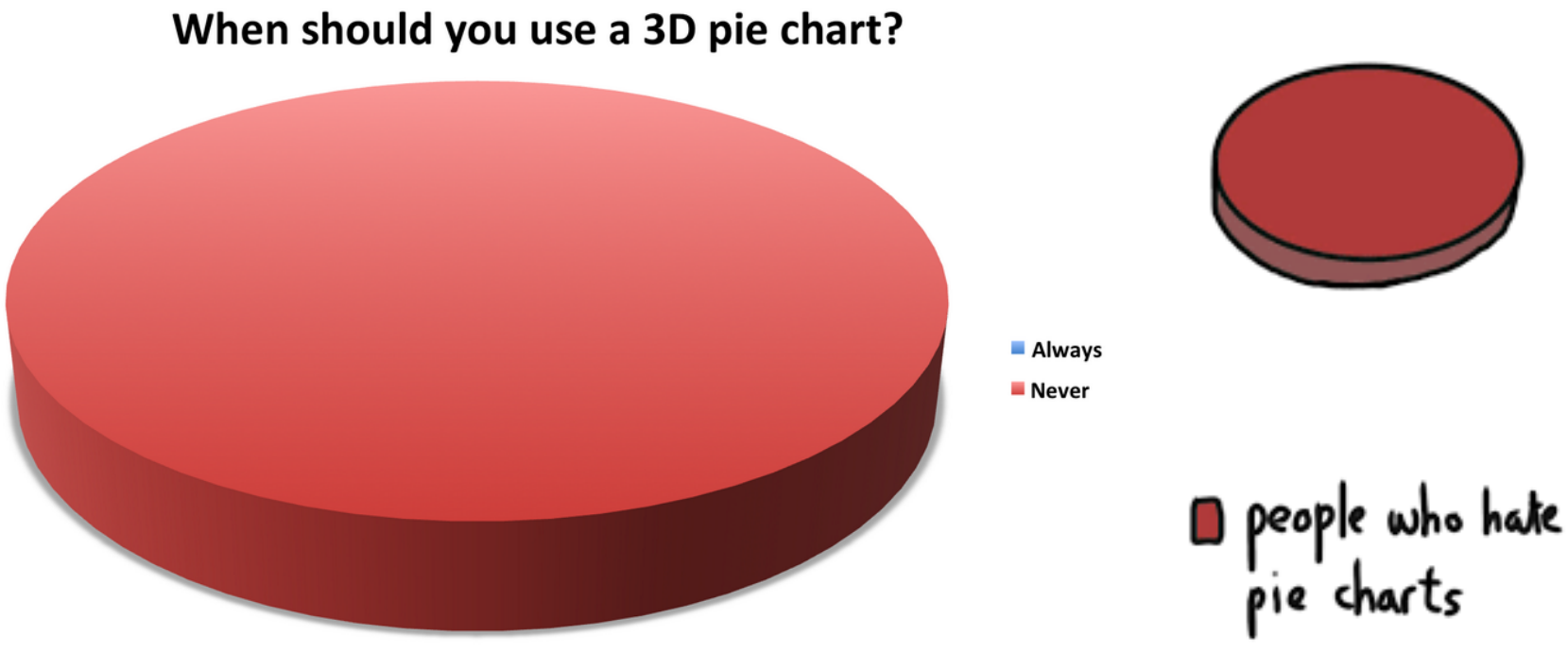
The James Hutton Institute

People hate pie charts

http://www.storytellingwithdata.com/blog/2011/07/death-to-pie-charts

especially Edward Tufte

A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them[...] pie charts should never be used. - *"The Visual Display of Quantitative Information"*



"E pur si muove..." [a] [b]

[a] Eells (1926) *J Am. Stat. Ass.*

[b] Simkin & Hastie (1987) *J Am. Stat. Ass.*

For proportions of a whole:

- Pie charts read as accurately as bar charts
- As number of components in the chart increases, bars are less efficient than pie charts

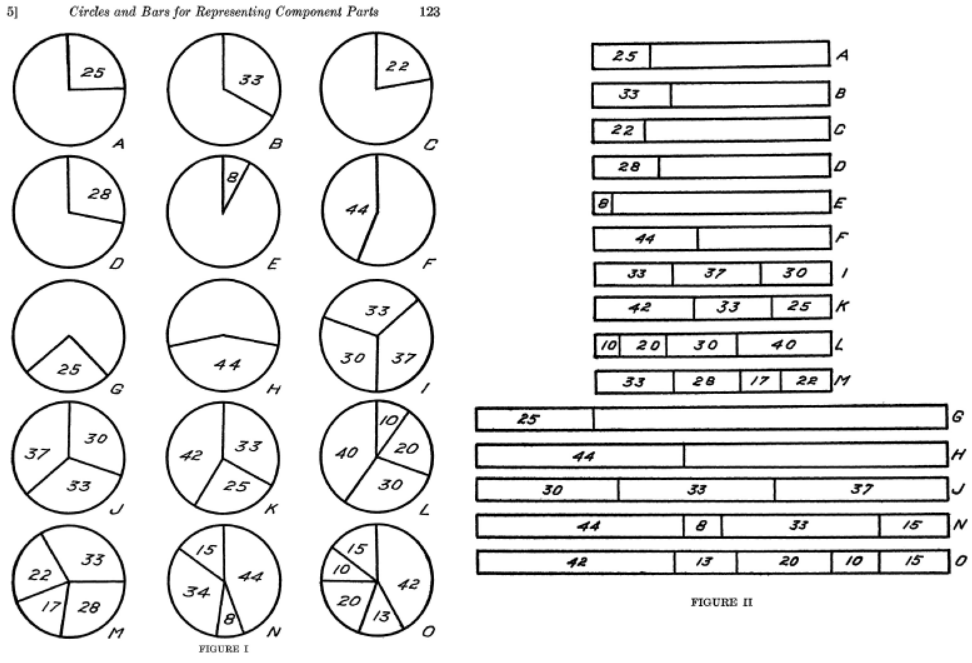


Table of Contents

The James Hutton Institute

Bar charts are bad...mmmkay?

The James Hutton Institute

There is an ongoing backlash against bar charts

(and I'm not picking on Nick, he just tweets a lot...)

But are they really that bad?

Interpretation of bars and lines [a]

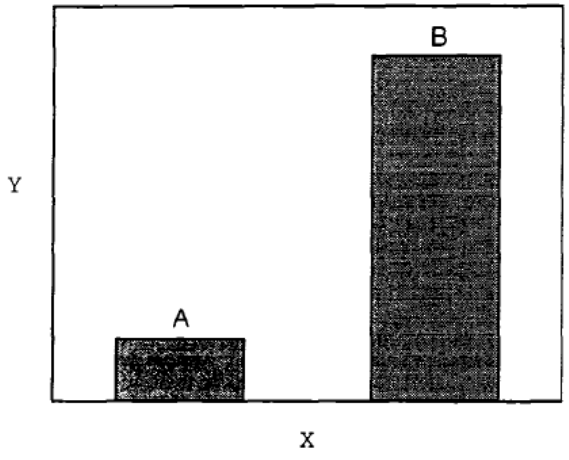
[a] Zacks & Tversky (1999) *Mem. Cognit.*

The James Hutton Institute

People interpret bars and lines differently

Experiment 1: In absence of context (arbitrary X, Y)

- **bars**: discrete comparison (24:0)
- **lines**: trend assessment (0:35)



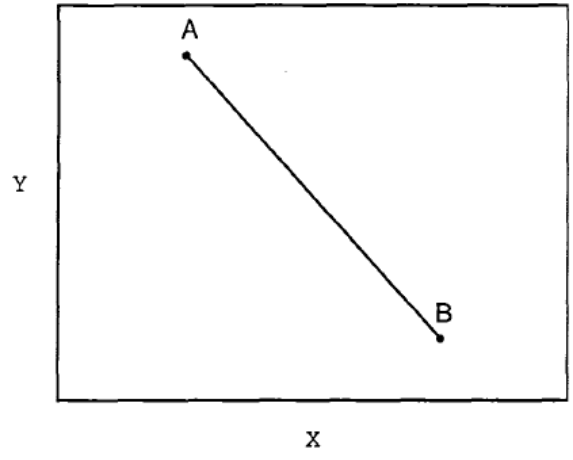


Figure 1: Examples of the bar and line graph stimuli used in Experiment 1.

Interpretation of bars and lines [a]

[a] Zacks & Tversky (1999) *Mem. Cognit.*

People interpret bars and lines differently

Experiment 2: With context (discrete or continuous data)

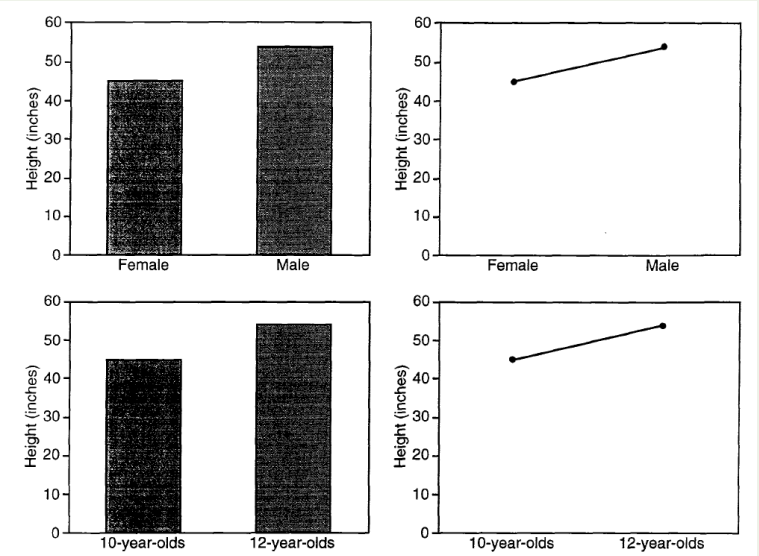


Figure 2: Examples of the bar and line graph stimuli, and the continuous and discrete conceptual domains used in Experiment 2.

	Gender (discrete domain)		Age (continuous domain)	
	Bar graph	Line graph	Bar graph	Line graph
Discrete comparison	28	22	28	9
Trend assessment	0	3	2	14

Table 2: Frequency of data characterization responses as a function of graph type (bar graph or line graph) and conceptual domain (gender or age).

Bars vs. lines

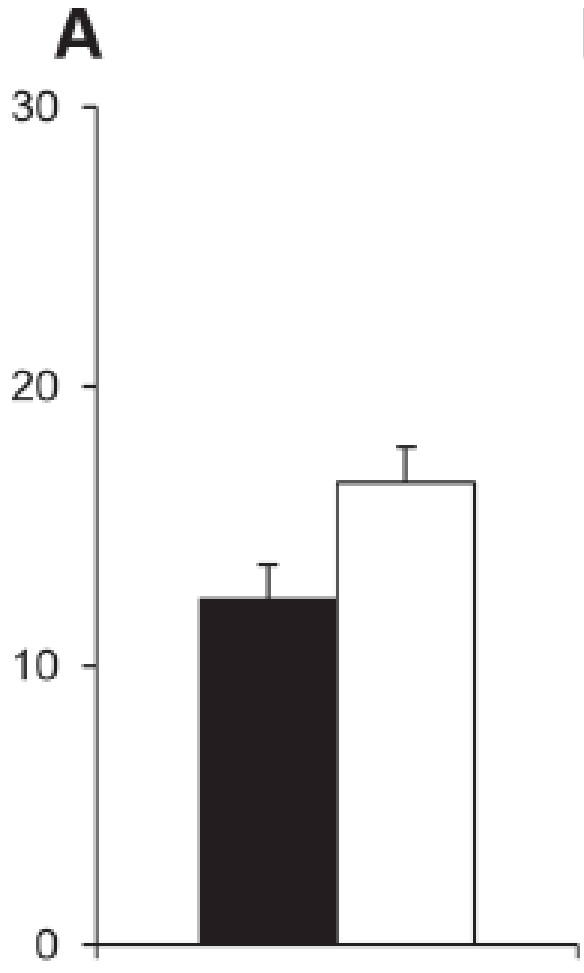
- People naturally interpret bar charts as categorical data
- People naturally interpret line graphs as trends
- Using bars for trend data or lines for categorical data can mislead the reader

Bar charts can mislead [a]

[a] Weissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

The James Hutton Institute

- Do these bars differ in value?

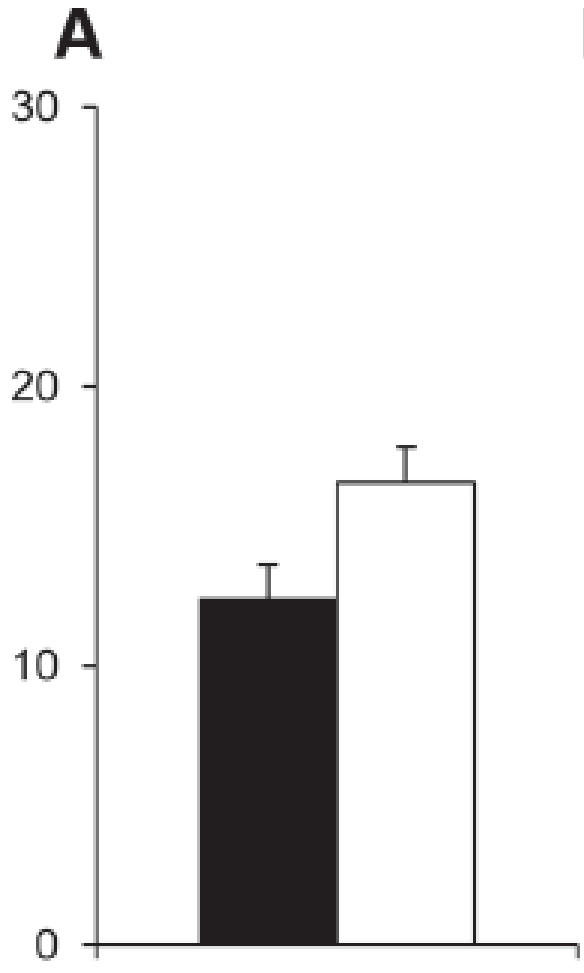


Bar charts can mislead [a]

[a] Weissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

The James Hutton Institute

- Do these bars differ in value?
- Bar charts represent data as a single point: lossy compression.
- Could different datasets give the same bar chart?



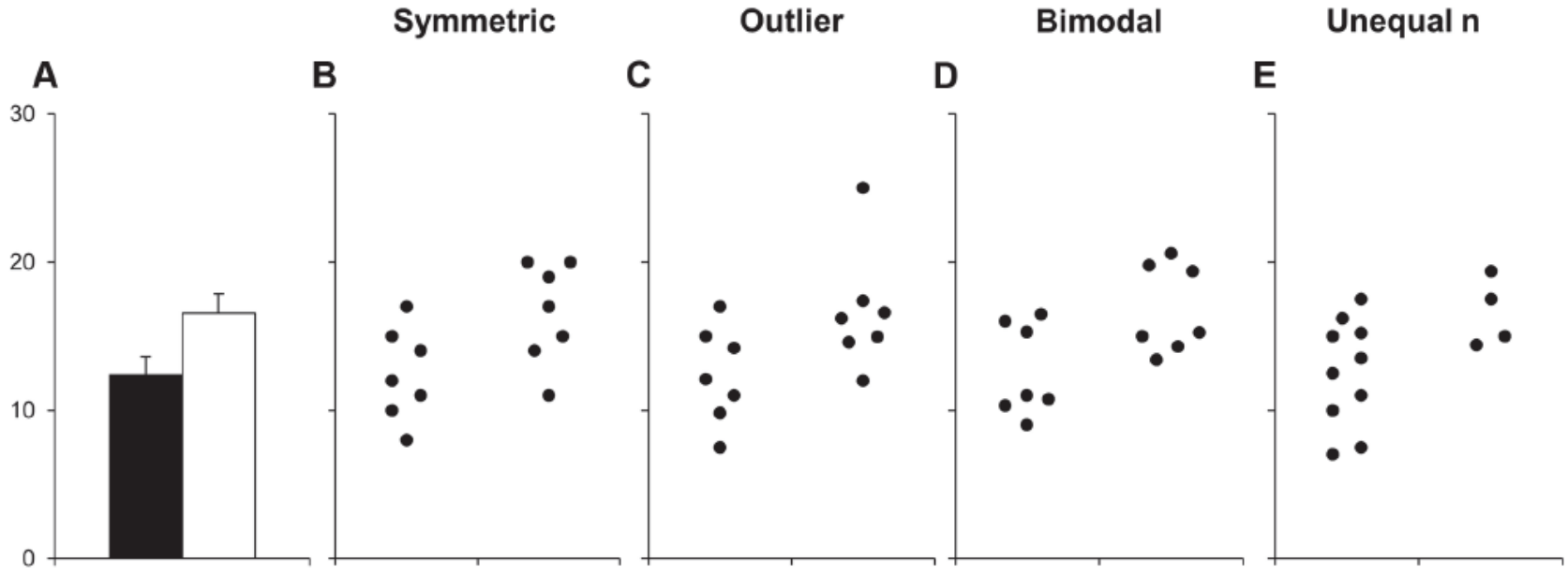
Bars are lossy compression [a]

[a] Weissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

The James Hutton Institute

Bars hide detail:

- Number of data points
- Variance of data points
- Distribution of data points (outliers, etc.)



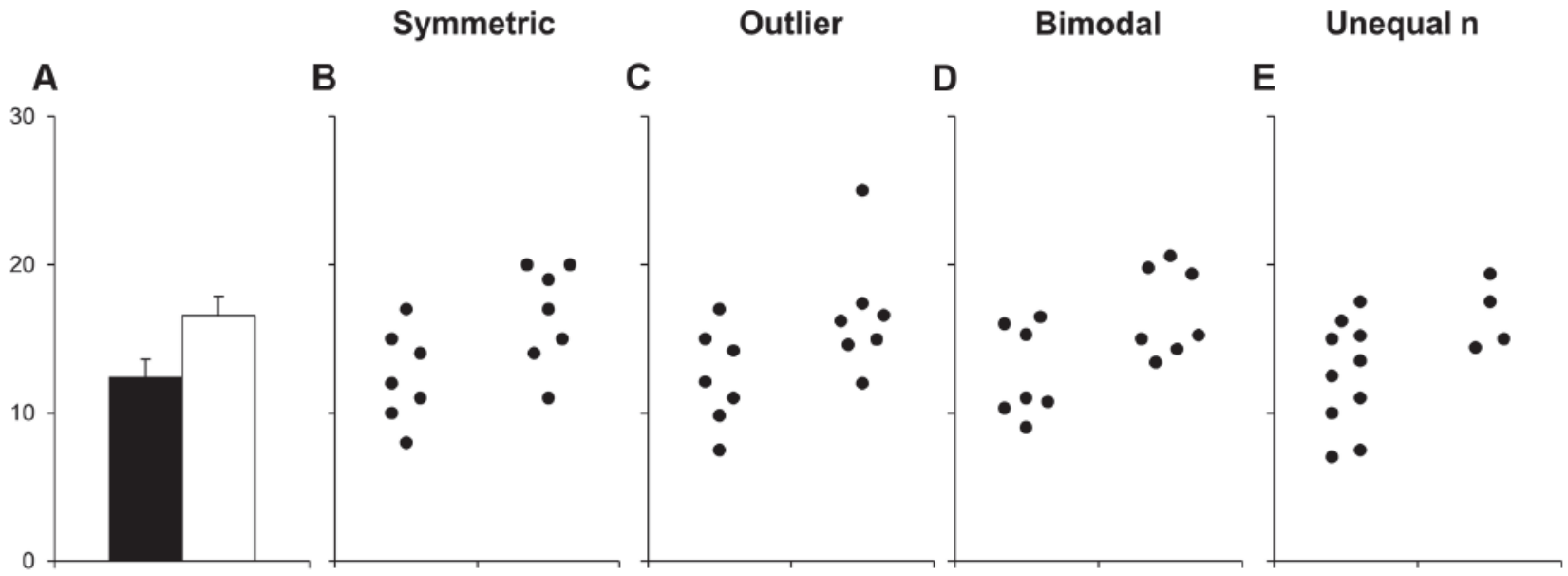
Bars may mislead on statistics [a]

[a] Weissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

The James Hutton Institute

Bars may imply incorrect test statistics:

Overlaps, outliers, covariates, sample sizes masked



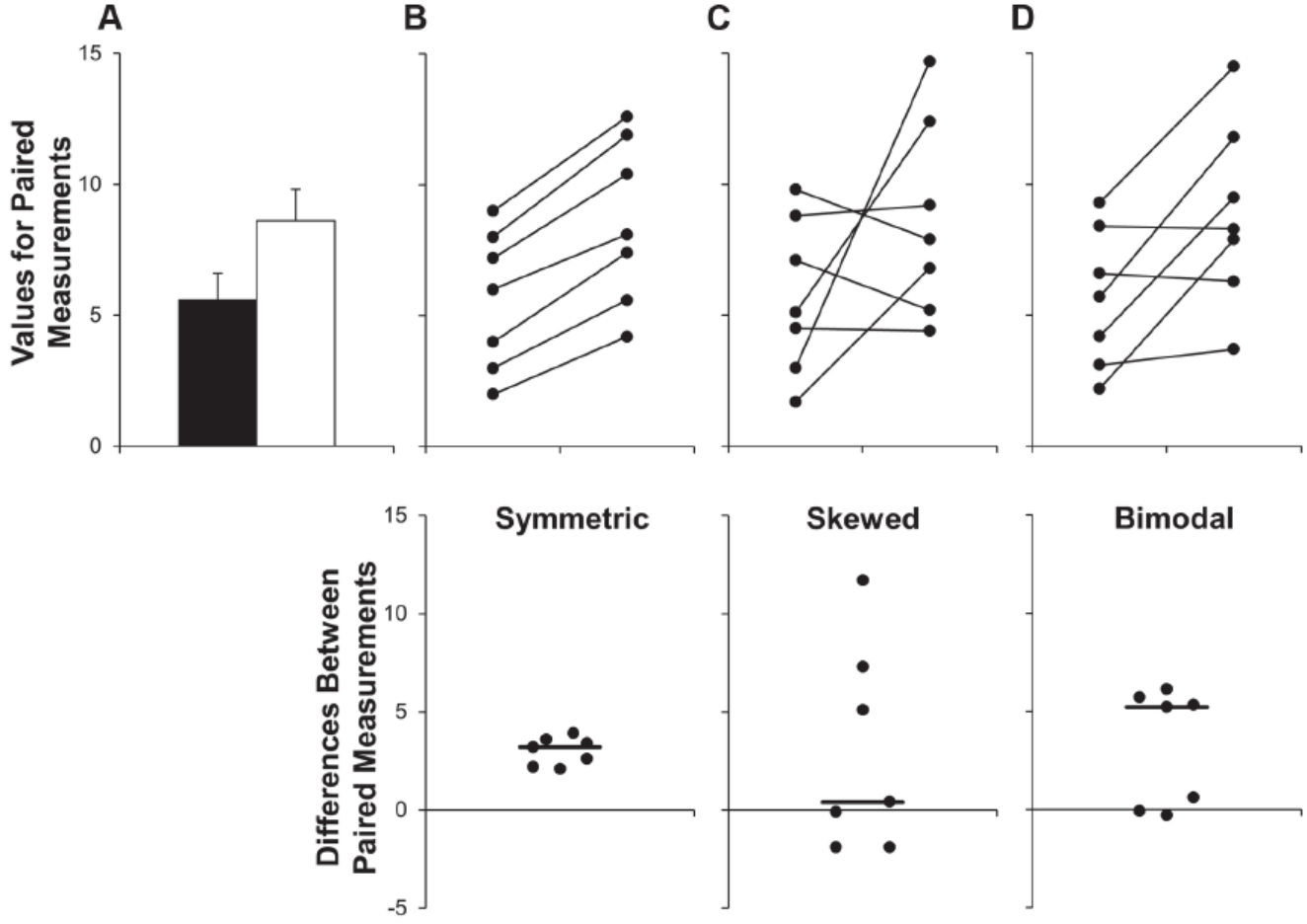
Test				
	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

Bars for paired data [a]

[a] Weissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

The James Hutton Institute

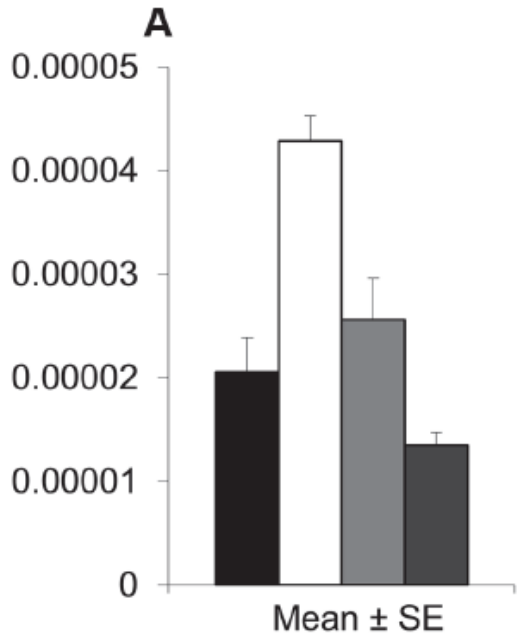
Bars imply independence of data:



Better than bar charts?

The James Hutton Institute

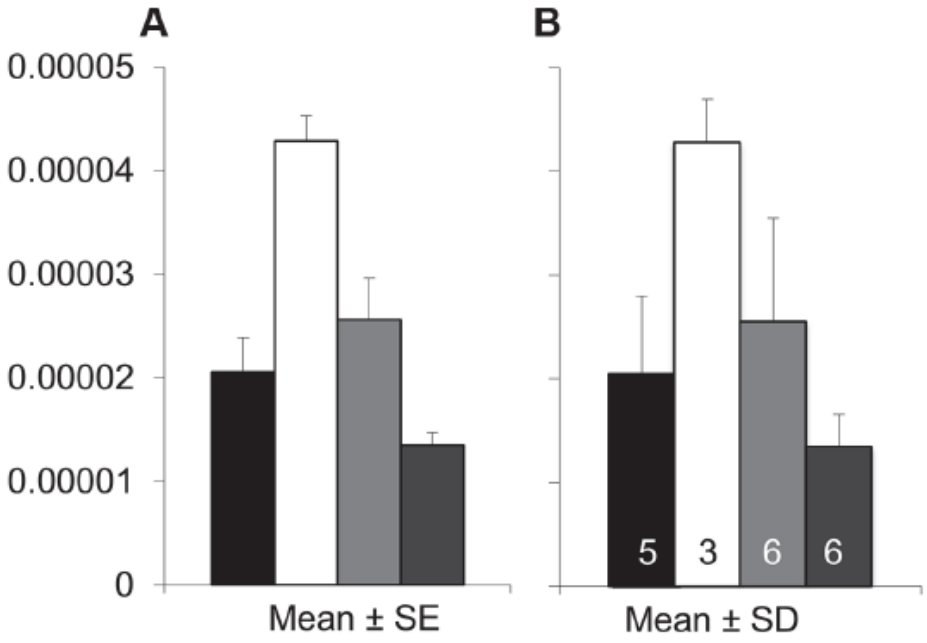
Bar chart with SE bars suggests group 2 is highest



Better than bar charts?

The James Hutton Institute

Bar chart with SD bars suggests there is overlap



Better than bar charts?

The James Hutton Institute

Univariate scatterplots show sample sizes, outliers, variance

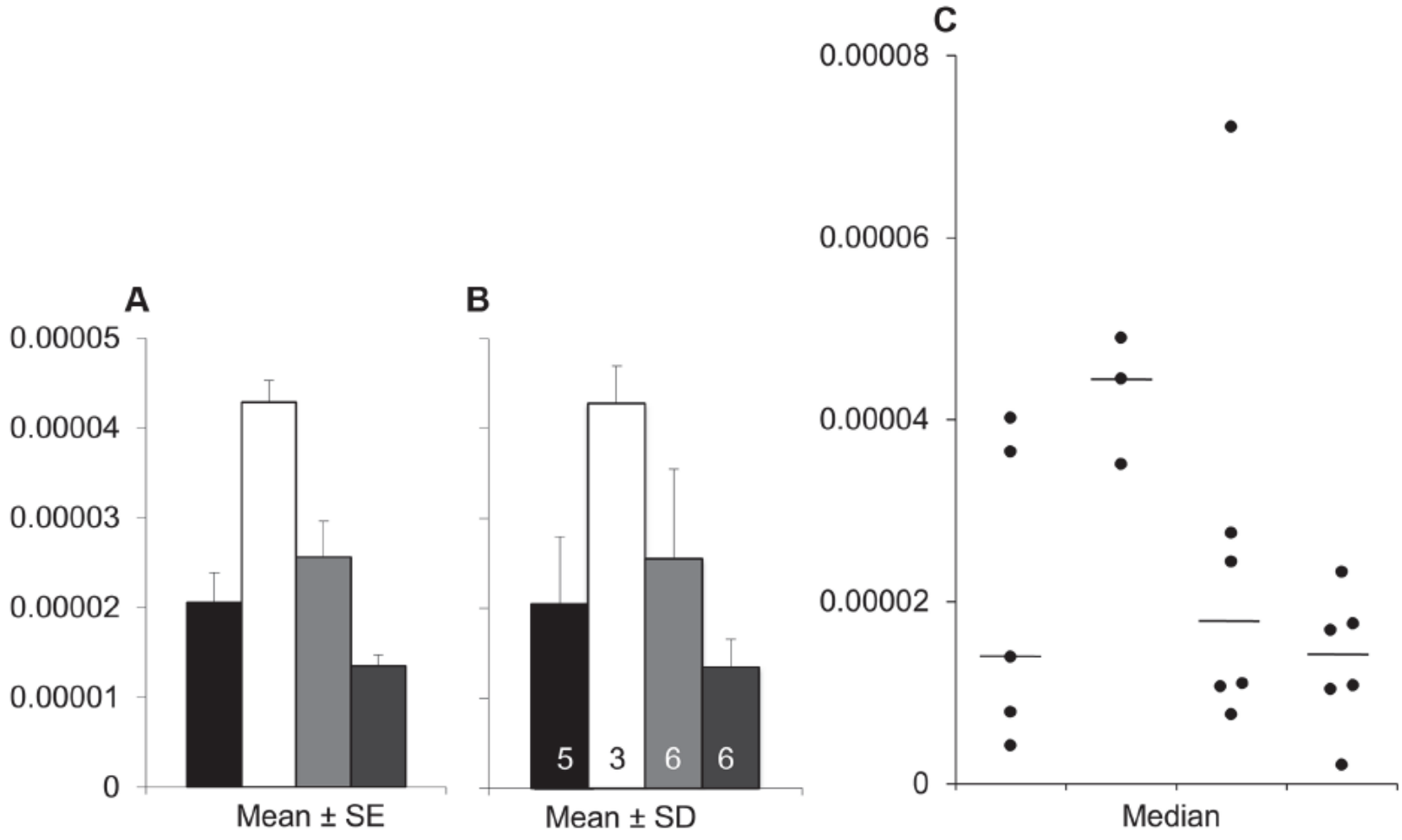


Table of Contents

The James Hutton Institute

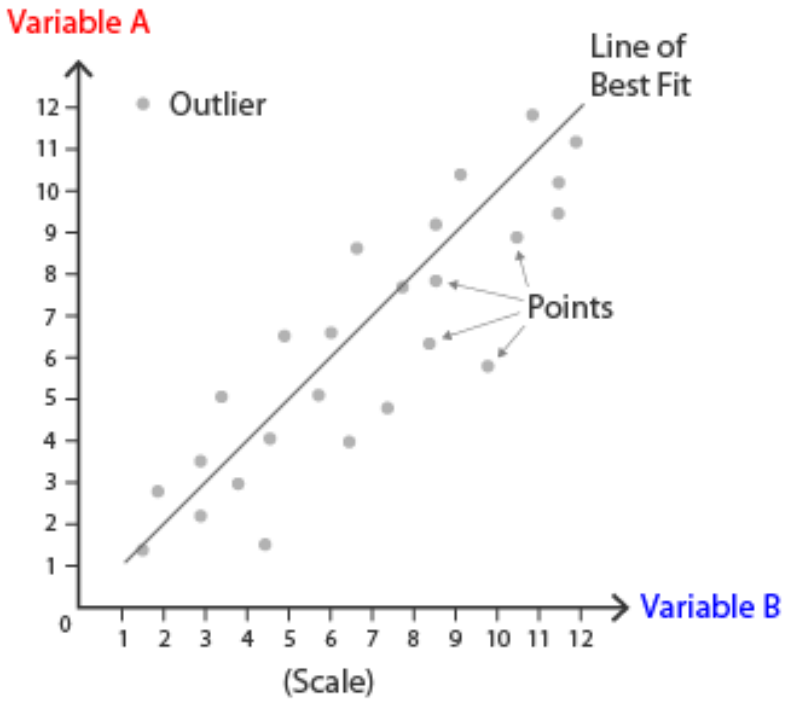
Scatterplots [a]

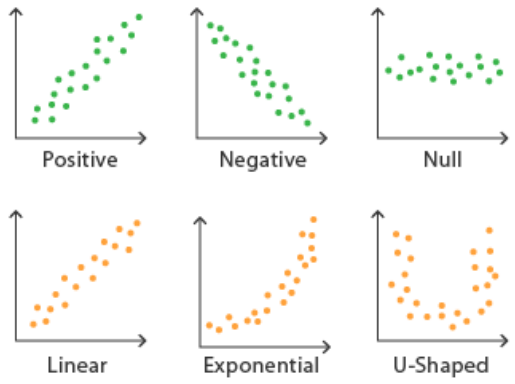
[a] http://www.datavizcatalogue.com/

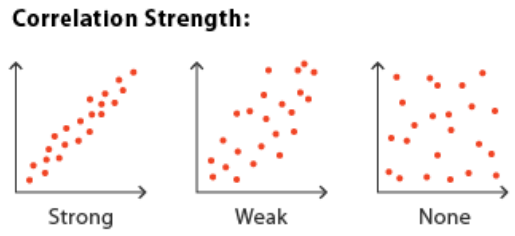
The James Hutton Institute

Scatterplots should be awesome:

- Positions on common scale (lowest error representation)
- Show all data: outliers, sample sizes, trends, etc.







Framing affects interpretation [a]

[a] Cleveland *et al.* (1982) *Science* doi:10.1126/science.216.4550.1138

The James Hutton Institute

Point cloud size affects interpretation of correlation (more diffuse interpreted as lower correlation coefficient)

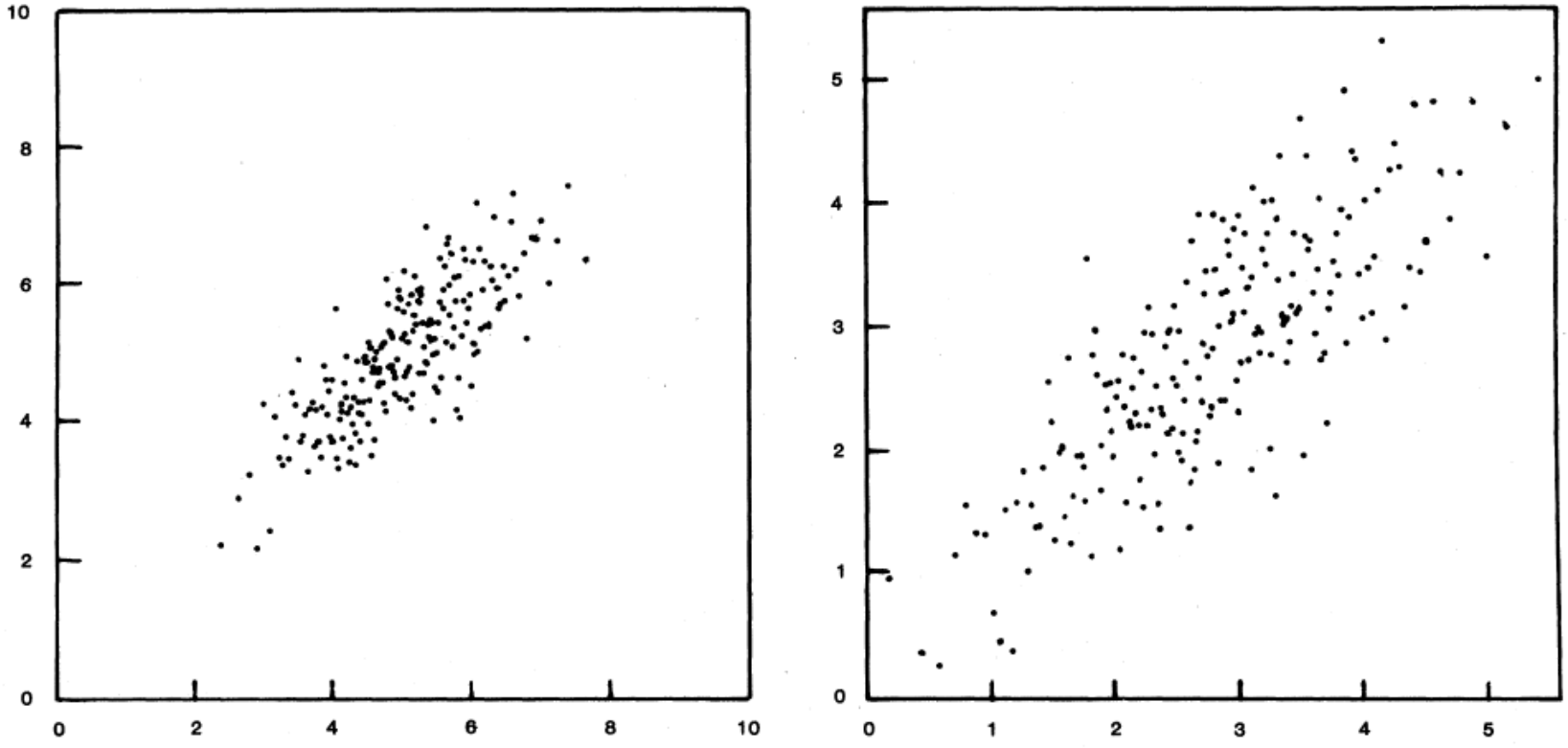


Fig. 1. Reductions of two scatterplots used in the three types of experiments. The left panel is point-cloud size 2 and the right panel is point-cloud size 4. In both panels $w(r) = .4$ and $r = .8$.

Interpreting correlation is difficult [a]

[a] Fisher *et al.* (2014) *PeerJ* doi:10.7717/peerj.589

The James Hutton Institute

People don't judge significance well

- 47.4% of significant relationships correctly classified
- 74.6% of non-significant relationships correctly classified

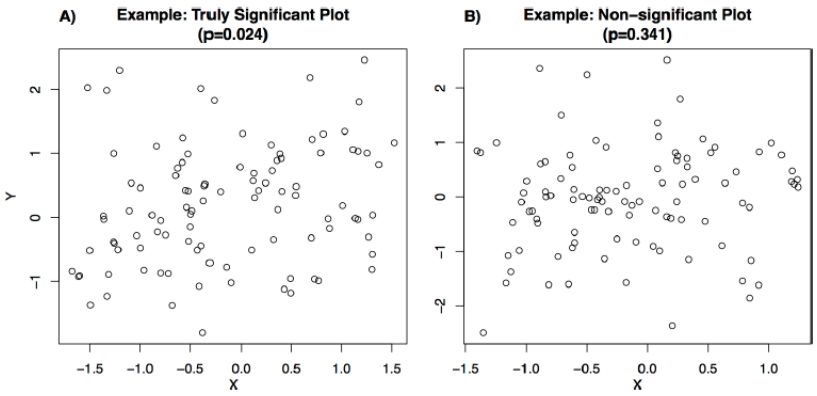


Figure 1 Examples of plots shown to users.

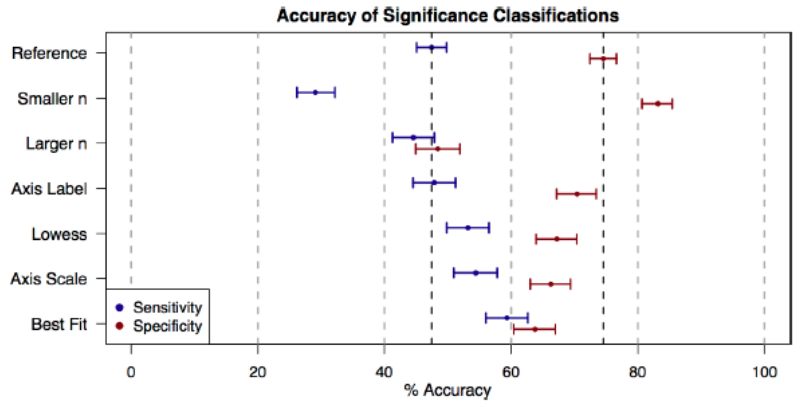


Table of Contents

The James Hutton Institute

Latency affects usage

The James Hutton Institute

Increasing latency to 0.5s:

- decreases user activity
- decreases dataset coverage
- reduces rate of hypothesis generation
- changes data exploration strategy
- reduces future interaction with other graphics

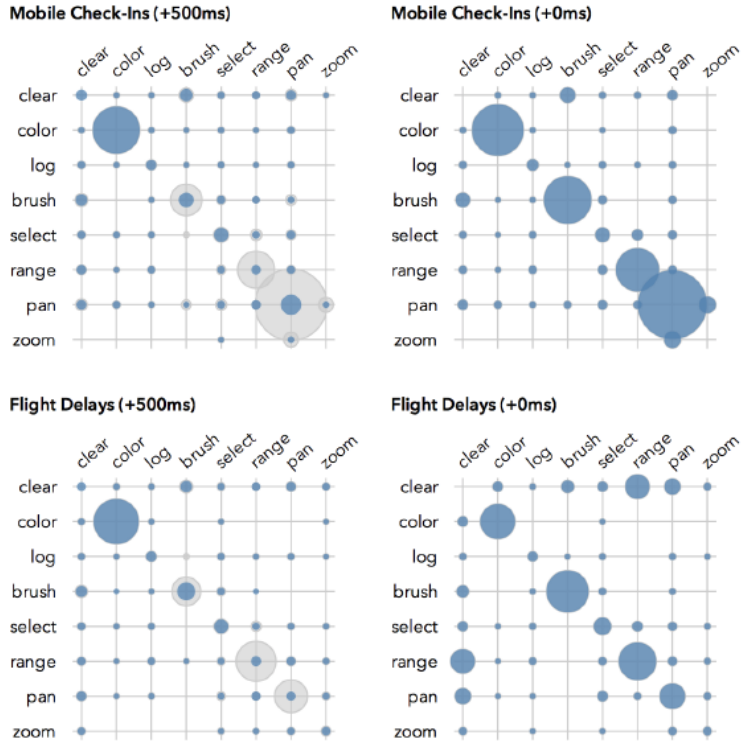


Fig. 4. Transitions between application events by analysis scenario and latency condition. Circular area represents the number of transitions between pairs of event types. Gray circles represent transitions between triggered events; blue circles between processed events. Rows represent source nodes and columns represent target nodes.

Licence: CC-BY-SA

The James Hutton Institute

By: Leighton Pritchard