

Hands-on session: Python Research Data Visualisation Workshop



**The James
Hutton
Institute**

Leighton Pritchard^{1,2,3}

¹Information and Computational Sciences,

²Centre for Human and Animal Pathogens in the Environment,

³Dundee Effector Consortium,

The James Hutton Institute, Invergowrie, Dundee, Scotland, DD2 5DA



Acceptable Use Policy

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

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



What I do

- Computational biologist (1996-present)
 - Protein sequence-structure-function (1996-1999)
 - Yeast metabolism (1999-2003)
 - Plant pathology (2003-present)
- Large datasets
 - metabolomics
 - genomics
 - geographical
- Visualisation as communication to (wet) biologists
 - protein structures
 - metabolic flux
 - comparative genomics/evolution



GenomeDiagram

a b

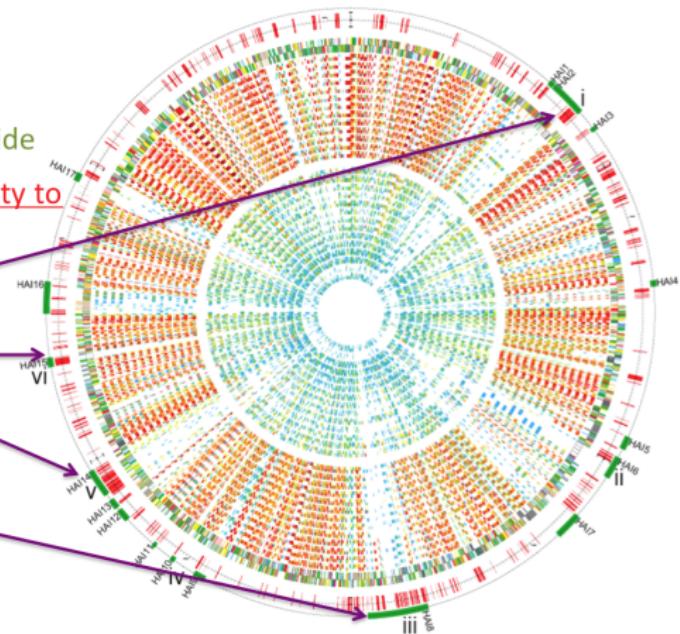
^aPritchard et al. (2006) *Bioinformatics* doi:10.1093/bioinformatics/btk021

^bToth et al. (2006) *Ann. Rev. Phytopath.* doi:10.1146/annurev.phyto.44.070505.143444



The James
Hutton
Institute

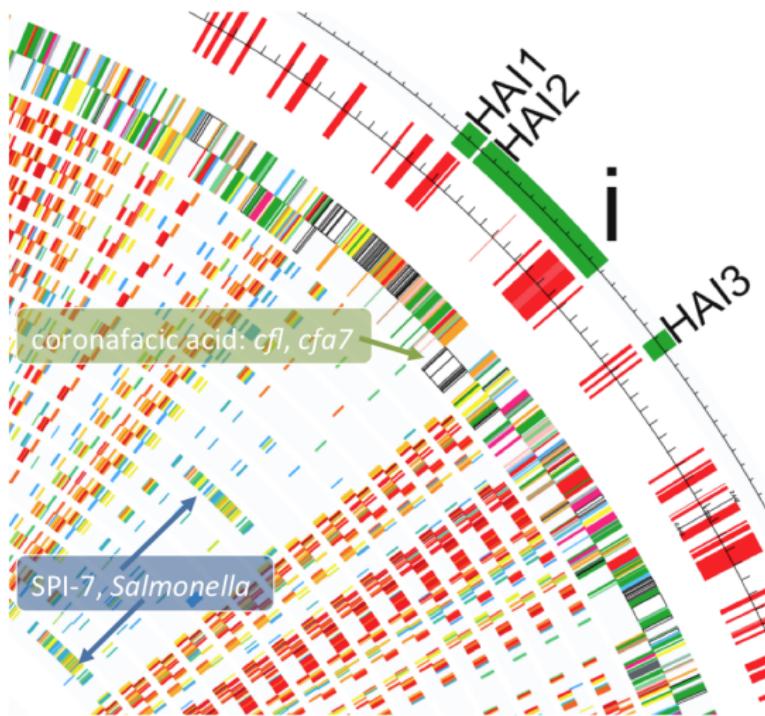
- Comparison against plant- (13) and animal-associated (14) bacteria
- Plant-associated in centre
- Animal-associated on outside
- Red marks: greater similarity to plant-associated bacteria
- **HAI2:** Phytotoxin
- **HAI15:** Adherence
- **HAI14:** Nitrogen fixation
- **HAI8:** T3SS





Functional adaptation in Pba^a

^aToth et al. (2006) Ann. Rev. Phytopath. doi:10.1146/annurev.phyto.44.070505.143444



Coronatine (*P. syringae*) interferes with jasmonate responses in host, as a jasmonate mimic

Coronafacic acid –
Pseudomonas syringae phytotoxin precursor
(coronatine)
- payload

SPI-7 -
Salmonella Typhi
Pathogenicity island
- delivery system



GenomeDiagram/SciArt ^{a b c}

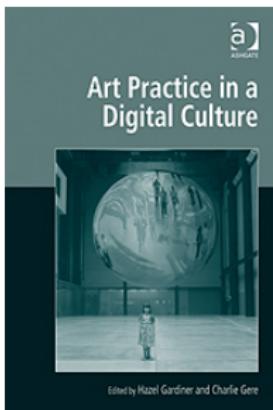


The James
Hutton
Institute

^aPritchard *et al.* (2006) *Bioinformatics* doi:10.1093/bioinformatics/btk021

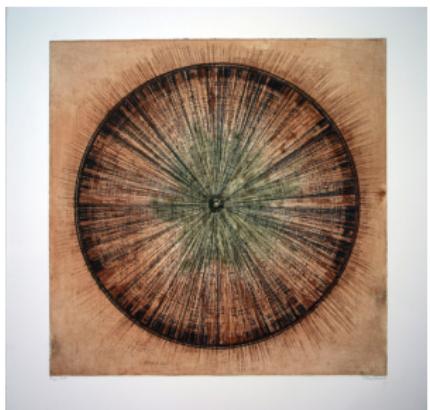
^bShemilt (2009) in "Digital Visual Culture: Theory and Practice" ISBN 978-1-84150-248-9

^cShemilt (2010) in "Art Practice in a Digital Culture", ISBN 978-0-7546-7623-2



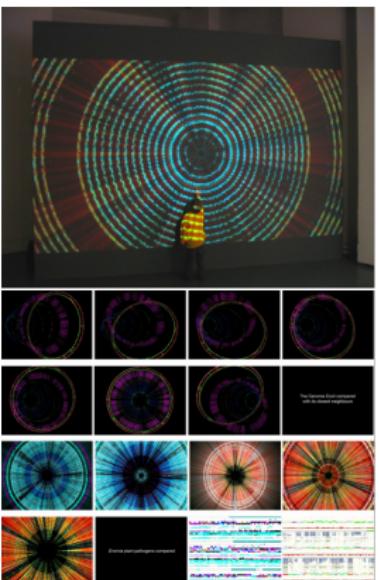
Influence

Free open-source comparative genomics visualisation library



Impact

Artwork (prints, audio-visual installation) exhibited in UK and internationally





Comparative metabolism

a *b* BIOPYTHON

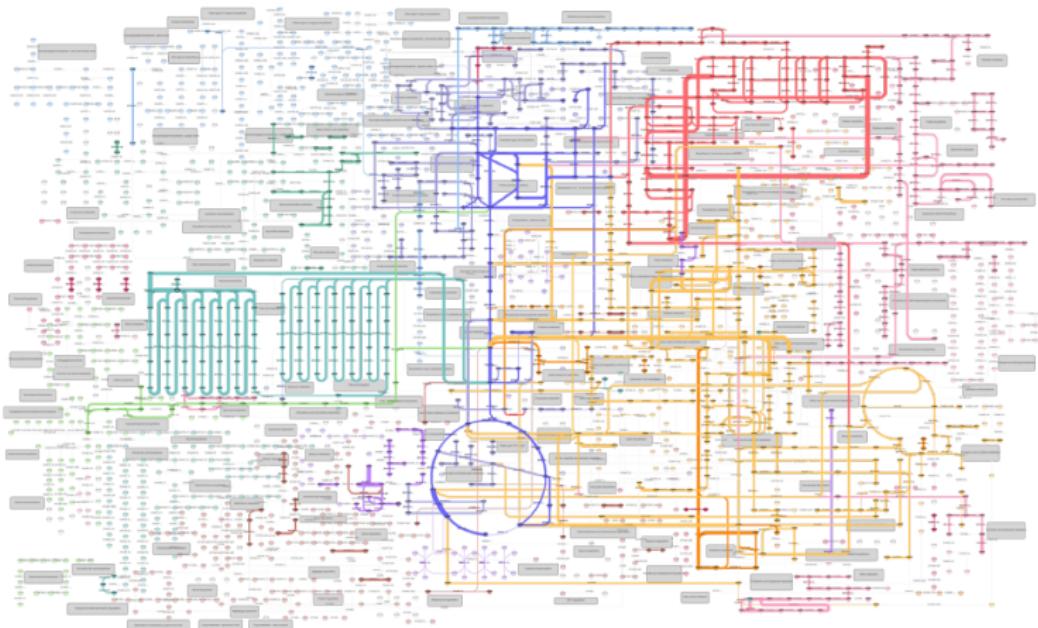
a Biopython KGML/KEGG visualisation module

b <https://github.com/widdowquinn/Notebooks-Bioinformatics>



The James
Hutton
Institute

Metabolic potential is a clue to community function & host range



PyANI: prokaryote classification ^{a b}

^a<http://widdowquinn.github.io/pyani/>

^bPritchard *et al.* (2016) *Anal. Methods* doi:10.1039/C5AY02550H

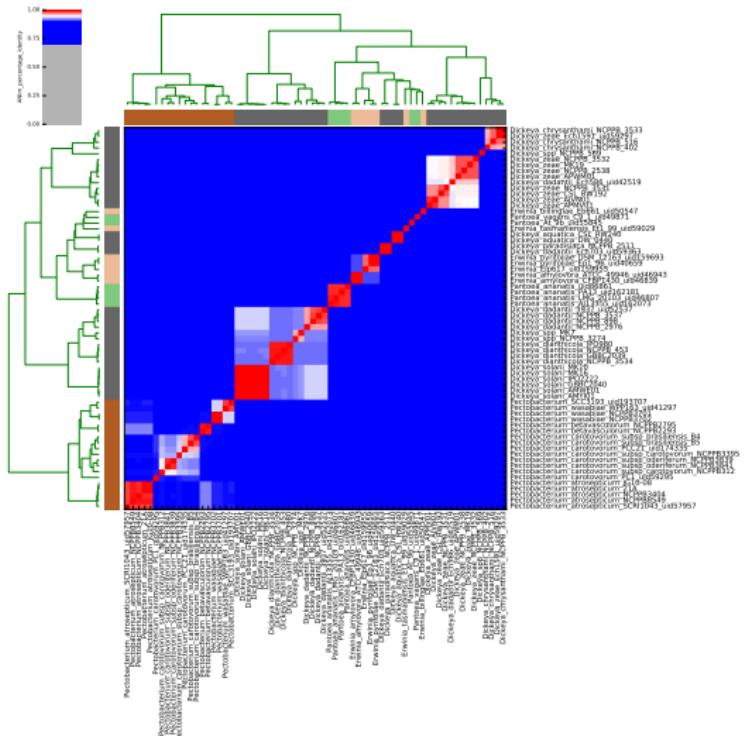




Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



Datavis is art and science

. . . storytelling in pictorial or graphical format

- Stories to yourself

- Stories to others

- Cautionary tales



Datavis is art and science

. . . storytelling in pictorial or graphical format

- Stories to yourself

- summarise big stories quickly
- data exploration and mining
- identify areas/items of importance
- find relationships and patterns

- Stories to others

- Cautionary tales



Datavis is art and science

. . . storytelling in pictorial or graphical format

■ Stories to yourself

- summarise big stories quickly
- data exploration and mining
- identify areas/items of importance
- find relationships and patterns

■ Stories to others

- present your interpretation of data
- make a specific point
- assert a relationship or pattern
- demonstrate significance

■ Cautionary tales



Datavis is art and science

. . . storytelling in pictorial or graphical format

■ Stories to yourself

- summarise big stories quickly
- data exploration and mining
- identify areas/items of importance
- find relationships and patterns

■ Stories to others

- present your interpretation of data
- make a specific point
- assert a relationship or pattern
- demonstrate significance

■ Cautionary tales

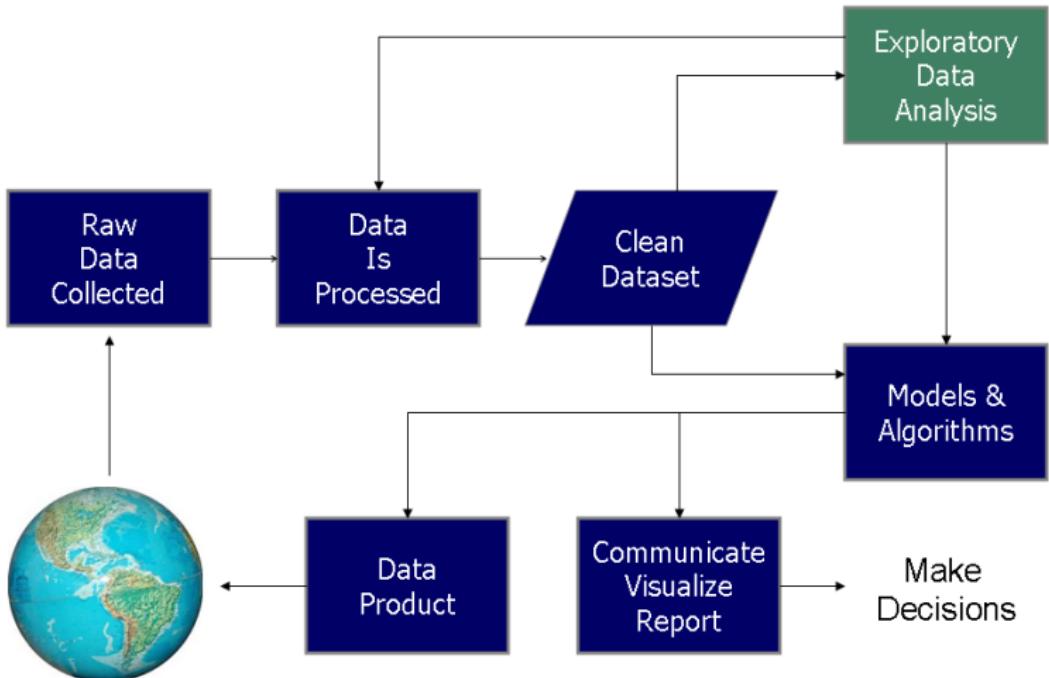
- avoid distortion
- make the reader think about the data, not the presentation
- avoid *chartjunk* (excessive decoration)
- aim for high data:ink ratio



The point of data visualisation ^a

^ahttps://en.wikipedia.org/wiki/Data_visualization

Where does visualisation belong?





Communicating effectively

- Understand the data

- Know (or be receptive to) the message

- Know your audience



Communicating effectively

- Understand the data
 - size
 - cardinality
 - meaning
 - relationships
- Know (or be receptive to) the message
- Know your audience



Communicating effectively

- Understand the data
 - size
 - cardinality
 - meaning
 - relationships
- Know (or be receptive to) the message
 - what does pictorial representation mean?
 - match graphical relationships to data relationships
- Know your audience



Communicating effectively

- Understand the data
 - size
 - cardinality
 - meaning
 - relationships
- Know (or be receptive to) the message
 - what does pictorial representation mean?
 - match graphical relationships to data relationships
- Know your audience
 - how do people process pictorial information
 - how does your audience process information
 - domain-specific representations



Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

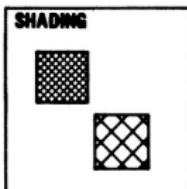
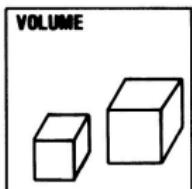
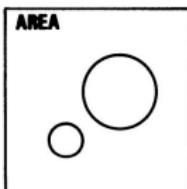
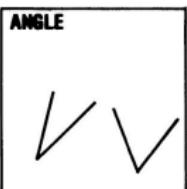
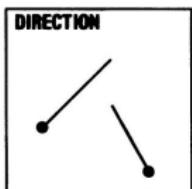
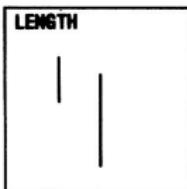
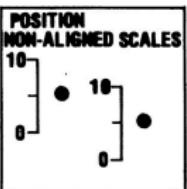
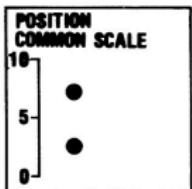
4 Acknowledgements



Elementary Perceptual Tasks ^a

^aCleveland & McGill (1984) *J. Am. Stat. Ass.*

The most basic visual tasks:



COLOR SATURATION

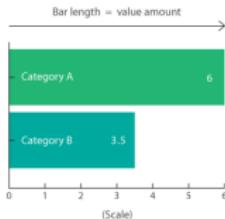


Implementations ^a

^a<http://www.datavizcatalogue.com/>

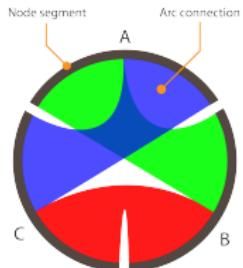
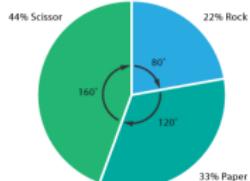
Position: common scale

- Scatterplot
- Bar Chart



Angle

- Pie Chart
- Do(ugh)nut Chart



Curvature

- Arc Diagram
- Chord Diagram



Table of Contents



The James
Hutton
Institute

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



What works best? Experiment ^{a b}

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

^b Heer & Bostock (2010) *CHI 2010*

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}

^aCleveland & McGill (1984) *J. Am. Stat. Ass.*

^bHeer & Bostock (2010) *CHI 2010*

- 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)

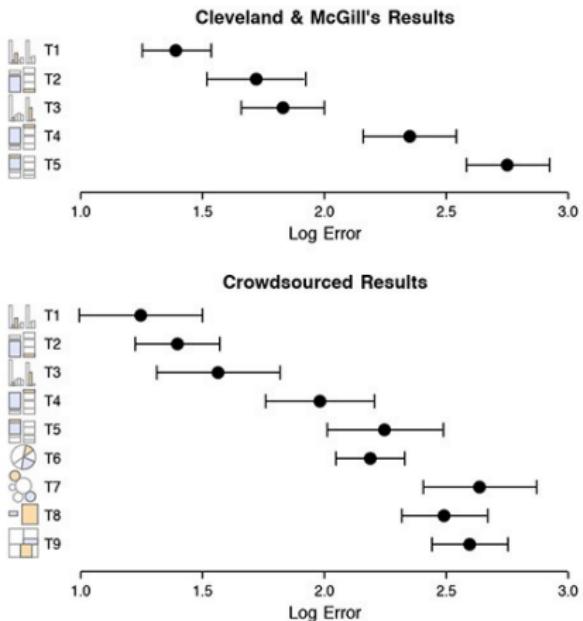


Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.



Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



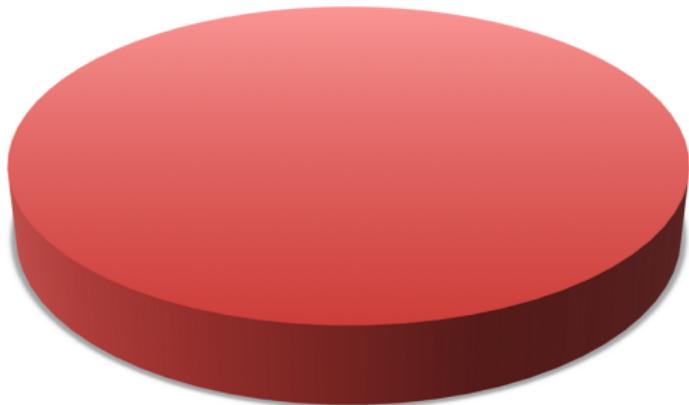
People hate pie charts

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

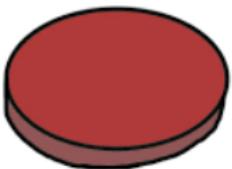
especially Edward Tufte

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"

When should you use a 3D pie chart?



- Always
- Never



■ people who hate
pie charts



"E pur si muove..." a b

^aEells (1926) *J Am. Stat. Ass.*

^bSimkin & Hastie (1987) *J Am. Stat. Ass.*



The James
Hutton
Institute

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

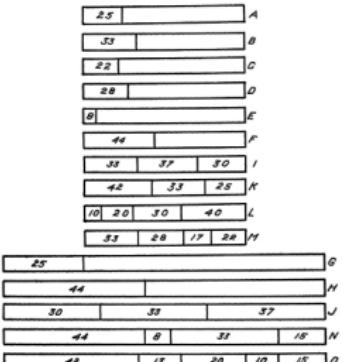
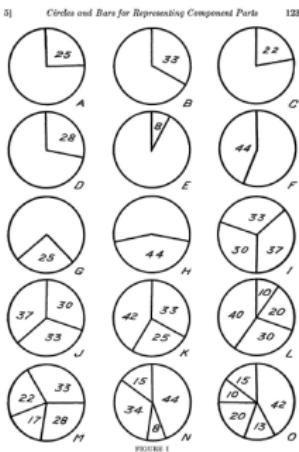


FIGURE II



Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



Bar charts are bad...mmmkay?



There is an ongoing backlash against bar charts
(and I'm not picking on Nick, he just tweets a lot...)



Nick Loman
@pathogenomenick



Following

Call out barplots when you see them in talks!!

Nikolai Slavov @slavovLab

The preference for bar plots over boxplot is as misguided as it is widespread.
[twitter.com/PracheeAC/stat...](https://twitter.com/PracheeAC/stat/)

RETWEETS 8
LIKES 8



9:56 AM - 7 Jul 2016

···



Nick Loman
@pathogenomenick



Following

The stacked bar chart -- a visualisation crime
that won't die.

RETWEETS 3
LIKES 3



6:30 PM - 2 Sep 2014

···

But are they really that bad?



Interpretation of bars and lines

^aZacks & Tversky (1999) *Mem. Cognit.*

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)

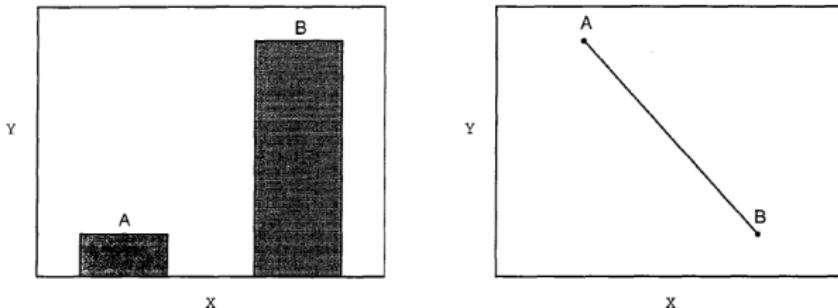


Figure 1: Examples of the bar and line graph stimuli used in Experiment 1.



Interpretation of bars and lines ^a

^aZacks & Tversky (1999) *Mem. Cognit.*

People interpret bars and lines differently

Experiment 2: With context (discrete or continuous data)

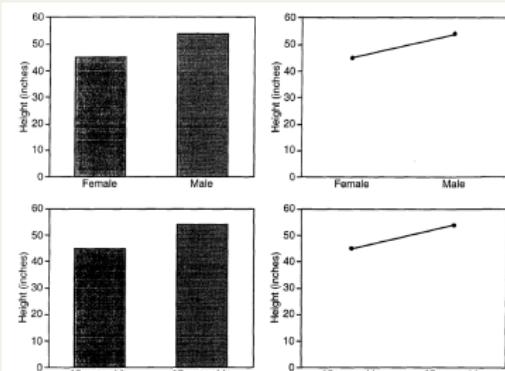


Figure 2: Examples of the bar and line graph stimuli, and the continuous and discrete conceptual domains used in Experiment 2.

	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

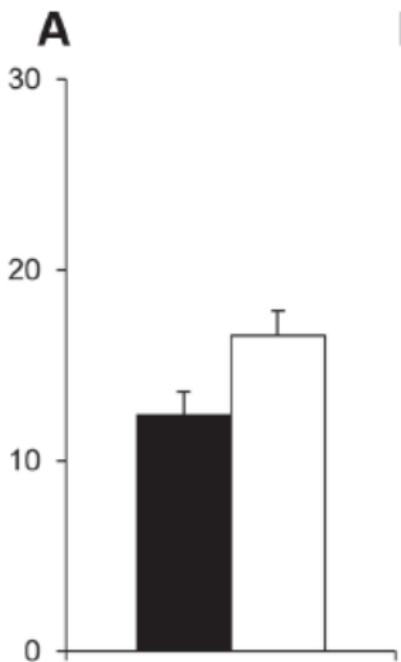
- People naturally interpret bar charts as categorical data
- People naturally interpret line graphs as trends
- Using bars for trend data or lines for categorical data can mislead the reader



Bar charts can mislead ^a

^aWeissgerber et al. (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

- Do these bars differ in value?

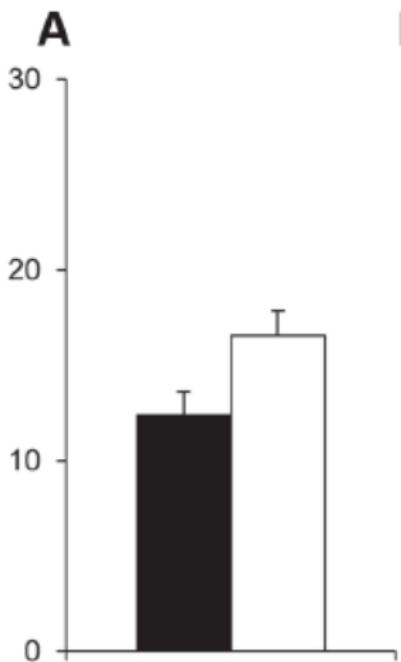




Bar charts can mislead ^a

^aWeissgerber et al. (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

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



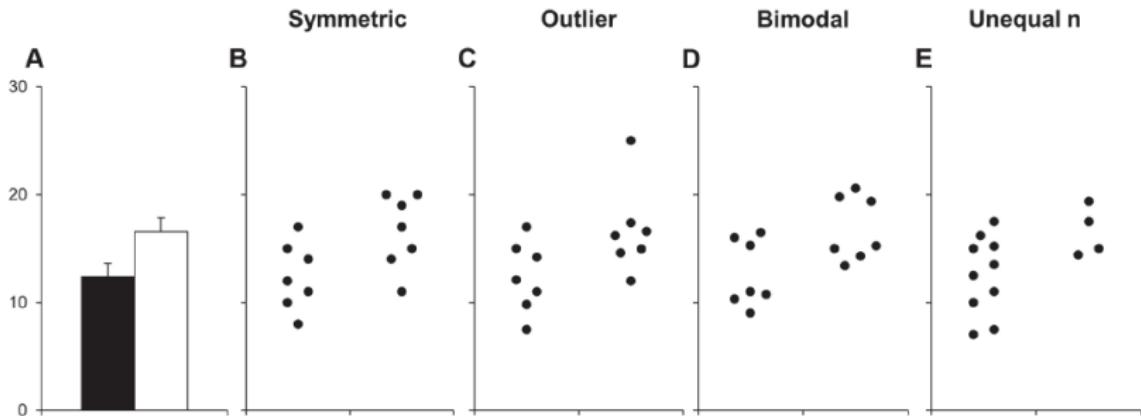


Bars are lossy compression ^a

^aWeissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

Bars hide detail:

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



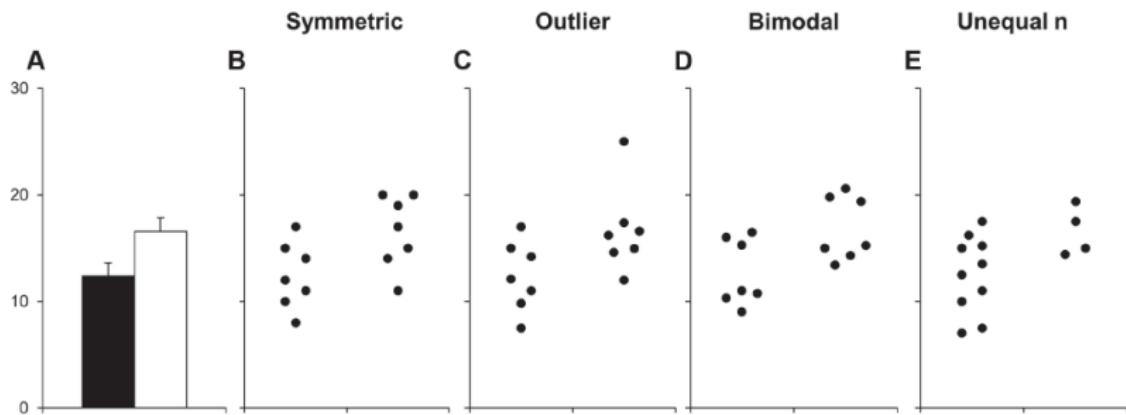


Bars may mislead on statistics ^a

^aWeissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

Bars may imply incorrect test statistics:

Overlaps, outliers, covariates, sample sizes masked



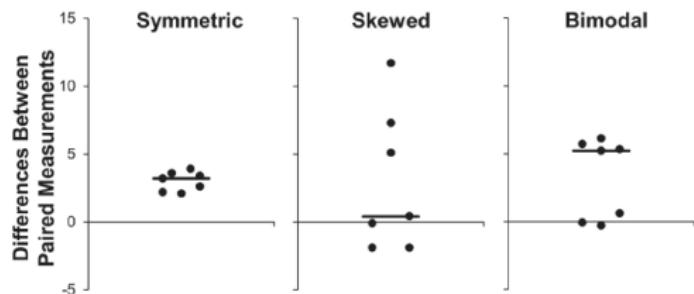
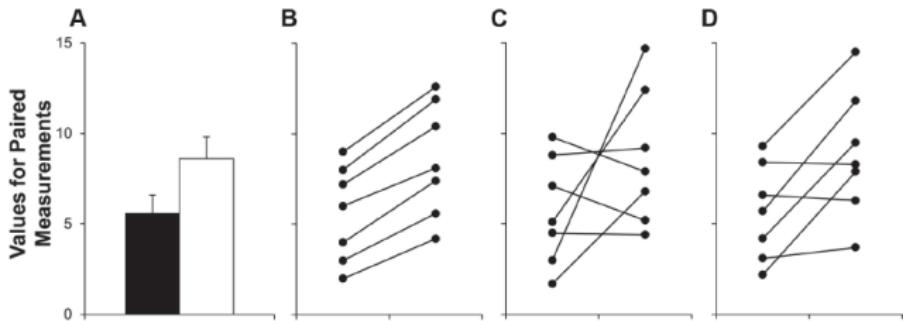
Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103



Bars for paired data ^a

^aWeissgerber *et al.* (2015) *PLoS Biol.* doi:10.1371/journal.pbio.1002128

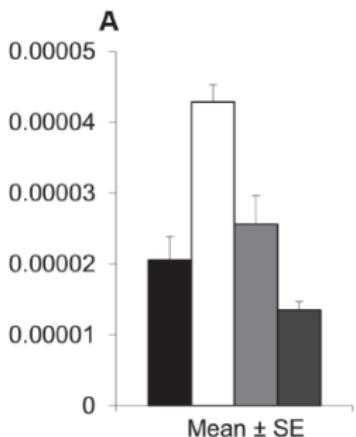
Bars imply independence of data:





Better than bar charts?

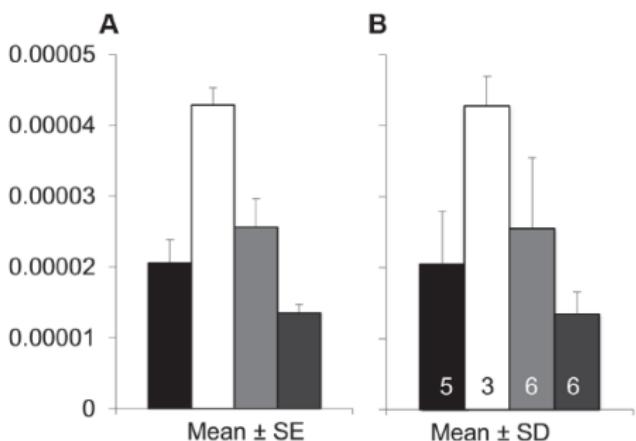
Bar chart with SE bars suggests group 2 is highest





Better than bar charts?

Bar chart with SD bars suggests there is overlap





Better than bar charts?

Univariate scatterplots show sample sizes, outliers, variance

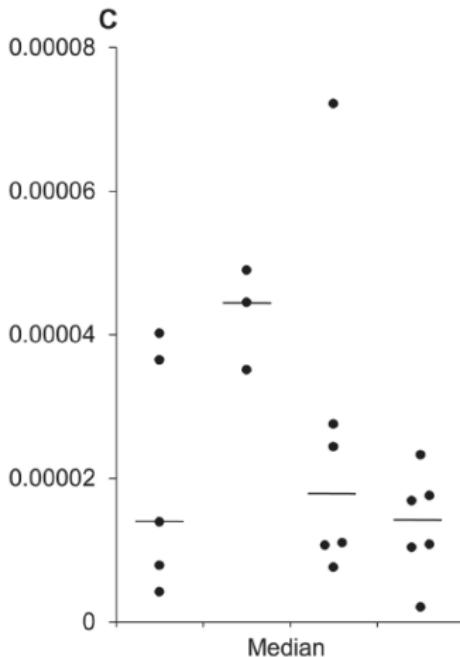
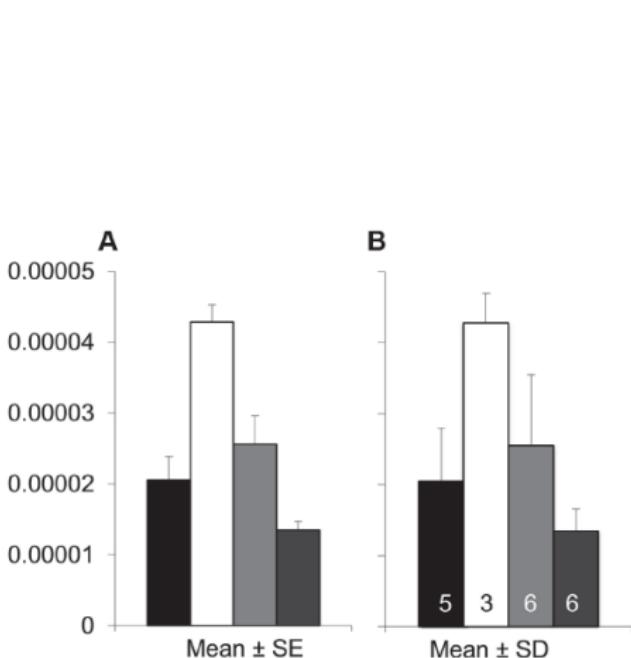




Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements

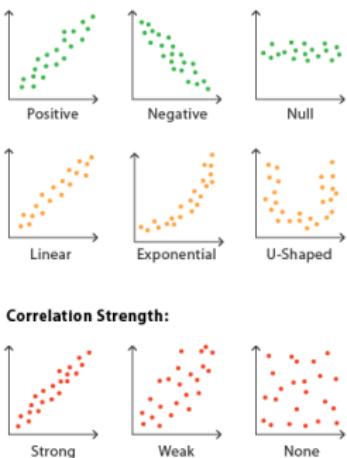
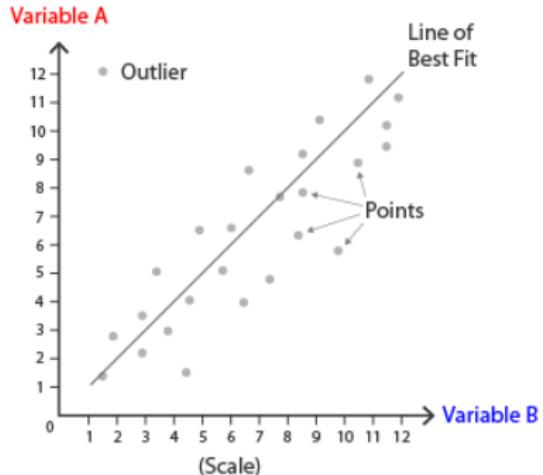


Scatterplots ^a

^a<http://www.datavizcatalogue.com/>

Scatterplots should be awesome:

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





Framing affects interpretation ^a

^aCleveland et al. (1982) *Science* doi:10.1126/science.216.4550.1138

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

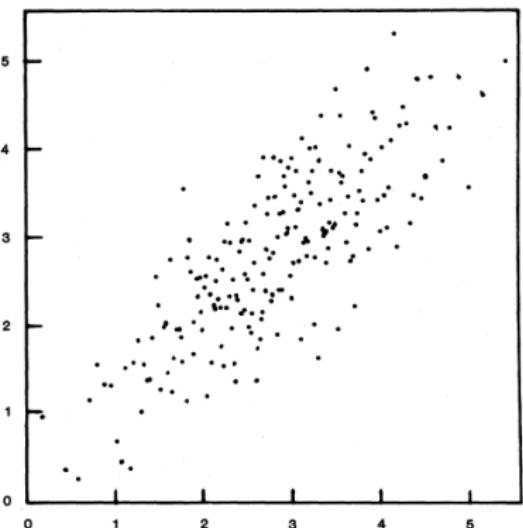
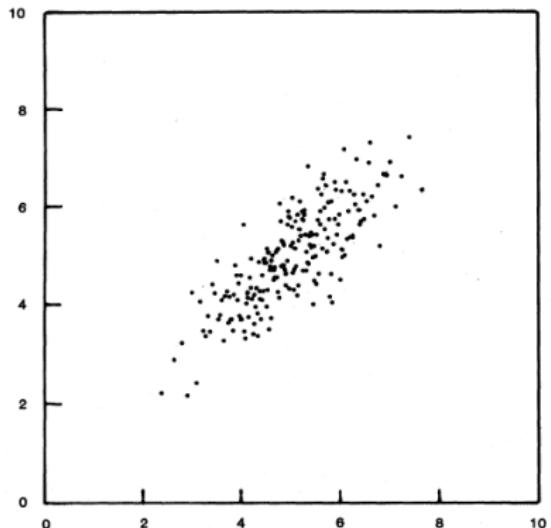


Fig. 1. Reductions of two scatterplots used in the three types of experiments. The left panel is point-cloud size 2 and the right panel is point-cloud size 4. In both panels $w(r) = .4$ and $r = .8$.



Interpreting correlation is difficult ^a

^aFisher et al. (2014) PeerJ doi:10.7717/peerj.589

People don't judge significance well

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

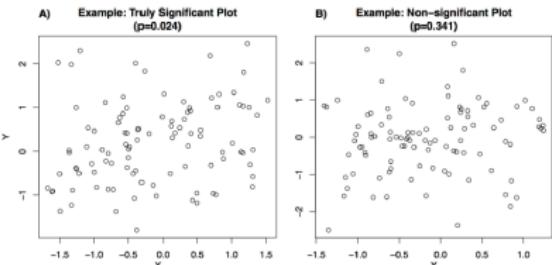


Figure 1 Examples of plots shown to users.

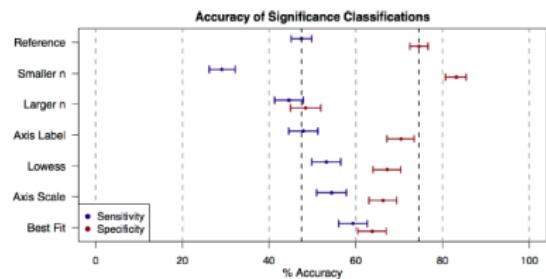




Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



Latency affects usage

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

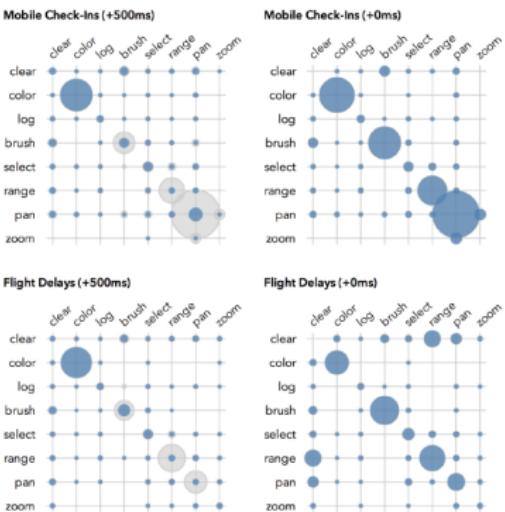


Fig. 4. Transitions between application events by analysis scenario and latency condition. Circular area represents the number of transitions between pairs of event types. Gray circles represent transitions between triggered events; blue circles between processed events. Rows represent source nodes and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



Table of Contents

1 Introduction

- Why listen to me?
- What is visualisation?
- Elementary perceptual tasks

2 Evidence-based representation

- What representations work best?
- Pie charts
- Bars and lines
- Scatterplots
- Interactive plots

3 Conclusions

- General advice

4 Acknowledgements



A word of caution

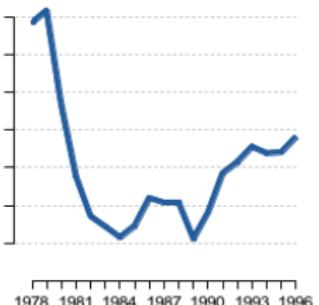
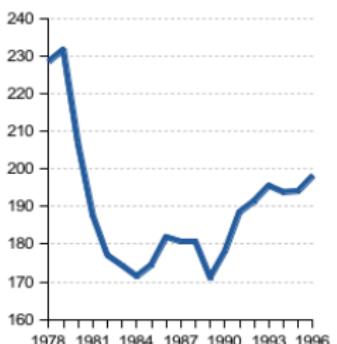
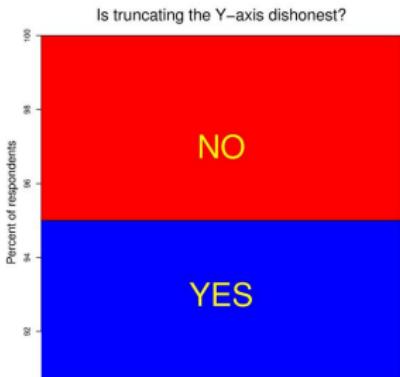
Rules are there to be broken

There are good reasons for breaking nearly every rule of presentation at one time or another.



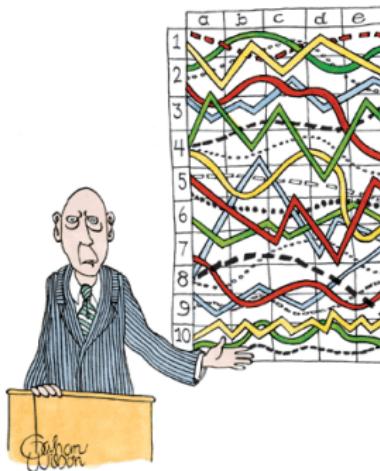
Use the full axis

but does difference from initial value matter more than absolute value?

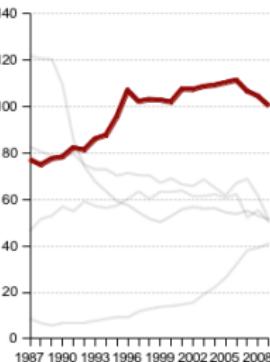
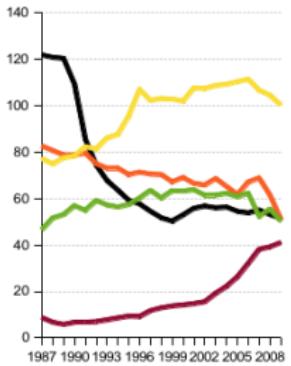




Simplify to your message



"I'll pause for a moment so you can let this information sink in."





Licence: CC-BY-SA

By: Leighton Pritchard

This presentation is licensed under the Creative Commons Attribution ShareAlike license

<https://creativecommons.org/licenses/by-sa/4.0/>