

# The Influence of Data Science and AI on Visualisation Perception

Daniel Archambault<sup>1</sup>

<sup>1</sup>Swansea University

# The Role of Visualisation in Data Science

*Why is a windscreen important for a car?  
Visualisations are the windscreens for data science.*

- When humans need to understand the AI system
  - ▶ Visualisation facilitates discovery in data science
  - ▶ If something goes wrong, how can we fix it?
  - ▶ Explainable AI requires a presentation of the data to the human
- Well, perhaps the right visualisation is a table of statistics? Right?

# Have you met the Datasaurus...?

(It's kinda like Anscombe's Quartet...)

*Justin Matejka and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (ACM CHI '17). 1290-1294.*

# Same Stats, Different Graphs

- All of these graphs have the exact same statistics:
  - ▶ mean in x: 54.26, y: 47.83
  - ▶ standard deviation in x: 16.76, in y: 26.93
  - ▶ correlation: -0.06

# Same Stats, Different Graphs: Why Visualisation is Needed

- These same statistics can produce many different scatterplots (only 2D)

# Same Stats, Different Graphs: Why Visualisation is Needed

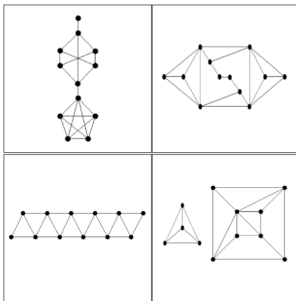
- These same statistics can produce many different scatterplots (only 2D)
  - ▶ You can get no trend
  - ▶ You can get clusters
  - ▶ You can get linear patterns
  - ▶ You can get star shapes

# Same Stats, Different Graphs: Why Visualisation is Needed

- These same statistics can produce many different scatterplots (only 2D)
  - ▶ You can get no trend
  - ▶ You can get clusters
  - ▶ You can get linear patterns
  - ▶ You can get star shapes
- You can also get a dinosaur...

Please see: <https://www.autodeskresearch.com/publications/samestats>

# Same Result in Network Analysis



*H. Chen, U. Soni, Y. Lu, R. Maciejewski and S. Kobourov, "Same Stats, Different Graphs (Graph Statistics and Why We Need Graph Drawings)," 26th Symposium on Graph Drawing (GD), 2018.*

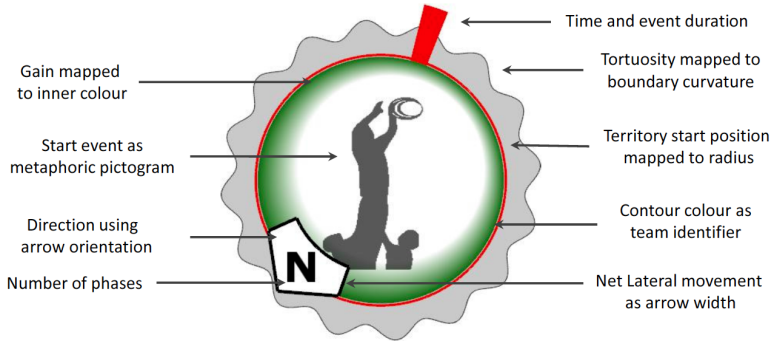
- All of these networks have the same statistics
  - ▶ vertices, edges, triangles, girth, clustering coefficient



# Moral of the Story...

- ... *make both calculations and graphs. Both sorts of output should be studied; each will contribute to understanding.*
  - ▶ F. J. Anscombe, 1973
- *Never trust summary statistics alone; always visualize your data*
  - ▶ Alberto Cairo
  - ▶ <http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html>
- Okay, Dan. Visualisation is important. But, what about encoding?

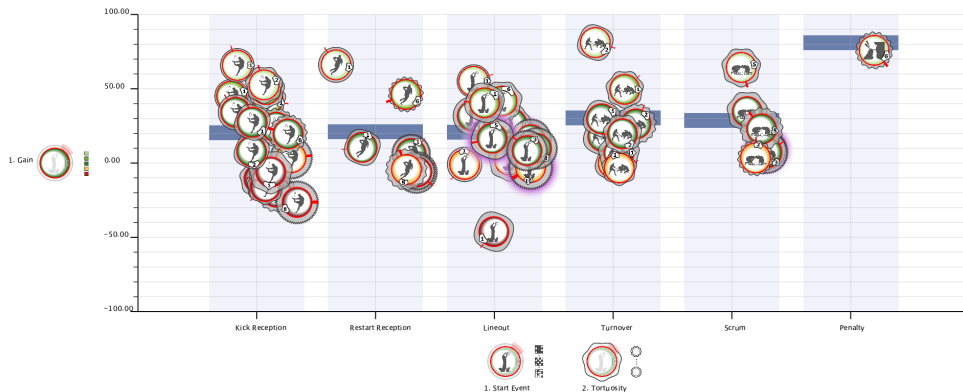
# Is the perception of order affected by encoding?



- Motivation comes from sports visualisation with Welsh Rugby
- Glyphs encode ten variables in different ways

Chung, D.H.S., Archambault, D., Borgo, R., Edwards, D.J., Laramée, R.S. and Chen, M. (2016), How Ordered Is It? On the Perceptual Orderability of Visual Channels. *Computer Graphics Forum*, 35: 131-140.

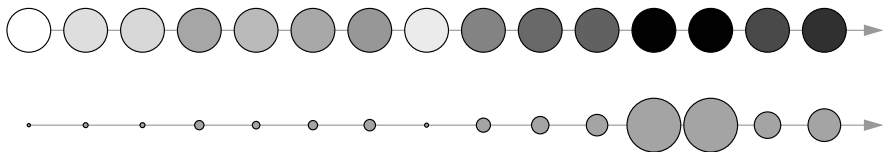
# Sorting Glyphs for Patterns



- Analyst sorts glyphs, observing patterns in other variables
- Often the task is to discover sorted subsequences along x or y
  - ▶ suggest correlation with these directions
- Does the chosen encoding matter for detecting these patterns?

# Chosen Encoding Matters

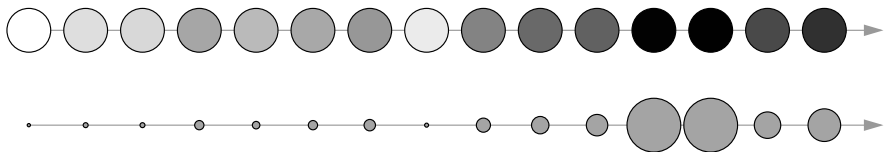
- Hypothesis: visual encoding effects perception of order



- Which of these sequences is more ordered?

# Chosen Encoding Matters

- Hypothesis: visual encoding effects perception of order



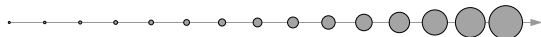
- Which of these sequences is more ordered?
  - ▶ sequences encode the exact same sequence of numbers
  - ▶ encoding influences the perception of order

# Bertin Retinal Variables

- Bertin retinal variables considered in our experiment along with numbers



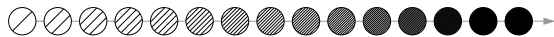
(a) Value



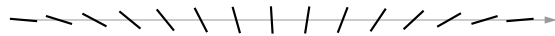
(b) Size



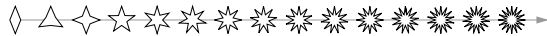
(c) Hue



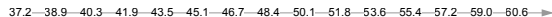
(d) Texture



(e) Orientation



(f) Shape



(g) Numeric

# Data Generation

- Start with an ordered sequence
  - ▶ Body Mass Index Used
- Inject *disorder* into sequence by swapping elements

$$f(x_i, x_j, d) = \begin{cases} \text{swap}(x_i, x_j) & \text{if } ||i - j|| \leq d \\ \text{null} & \text{else} \end{cases} \quad (1)$$

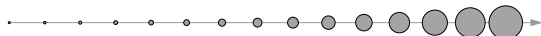
- Produce stimuli with various Pearson correlation coefficients
- Gives an objective way to measure order (correlation) and compare it to perceived order (answer entered by the user)

# Visual Mapping of Elements

- We use the visual channels described by Bertin in our experiment



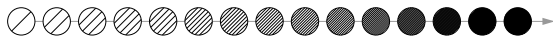
(h) Value



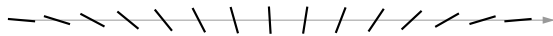
(i) Size



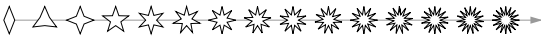
(j) Hue



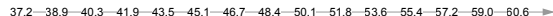
(k) Texture



(l) Orientation



(m) Shape



(n) Numeric

- For each sequence, need to ensure that difference between elements can be perceived




# Task and Experimental Procedure

- Task is to rate how ordered the sequence is from 1 to 5

**How ordered is it?** Worker ID: 485672  
Job Progress:

An image is shown containing a sequence of elements to be read from left to right.  
For each image, rate how ordered the sequence of elements is.



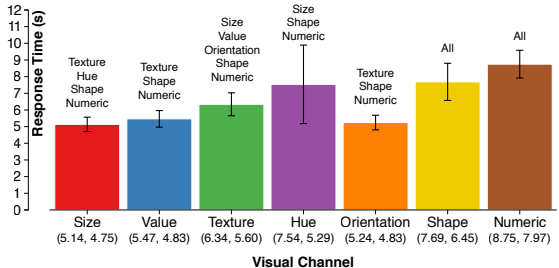
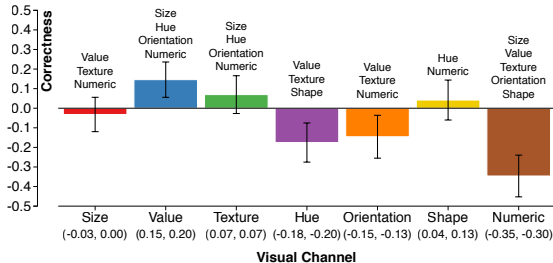
Unordered ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ Ordered

**Submit**

- This measurement is an error measure
  - Compare the distance between entered answer and level of disorder ( $N_*$ )
- Response time is also measured to determine difficulty of encoding

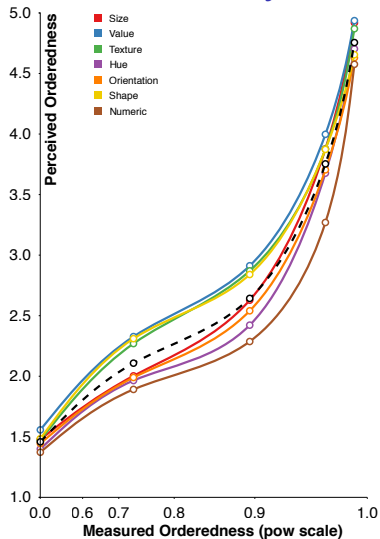
# Results

- 110 participants (62 male and 48 female) each paid \$1.00
- Error rate is difference between entered answer and actual answer
- Response time is measured in time taken to answer (difficulty of encoding)



- Order underestimated for some visual channels and overestimated for others

# Results Divided by Level of Disorder



- Order in sequence either overestimated or underestimated depending on encoding (gap is statistically significant)
- Overestimation of order
  - ▶ value, texture, and shape
- Underestimation of order
  - ▶ orientation, hue, and numeric
- Size seems to follow the order in the data closely
- Depending on encoding chosen, order is overestimated or underestimated

# Conclusion

- Visualisation is important: same stats produce different graphs
- Selection of the visualisation is important: different encodings of the same data are perceived differently
- For effective data science the visualisation and the analytics/machine learning technique cannot be considered separately