

Visualizing Data

Manoel Horta Ribeiro
manoel@cs.princeton.edu

COS 424 / Fall 2026



Quizz Time!

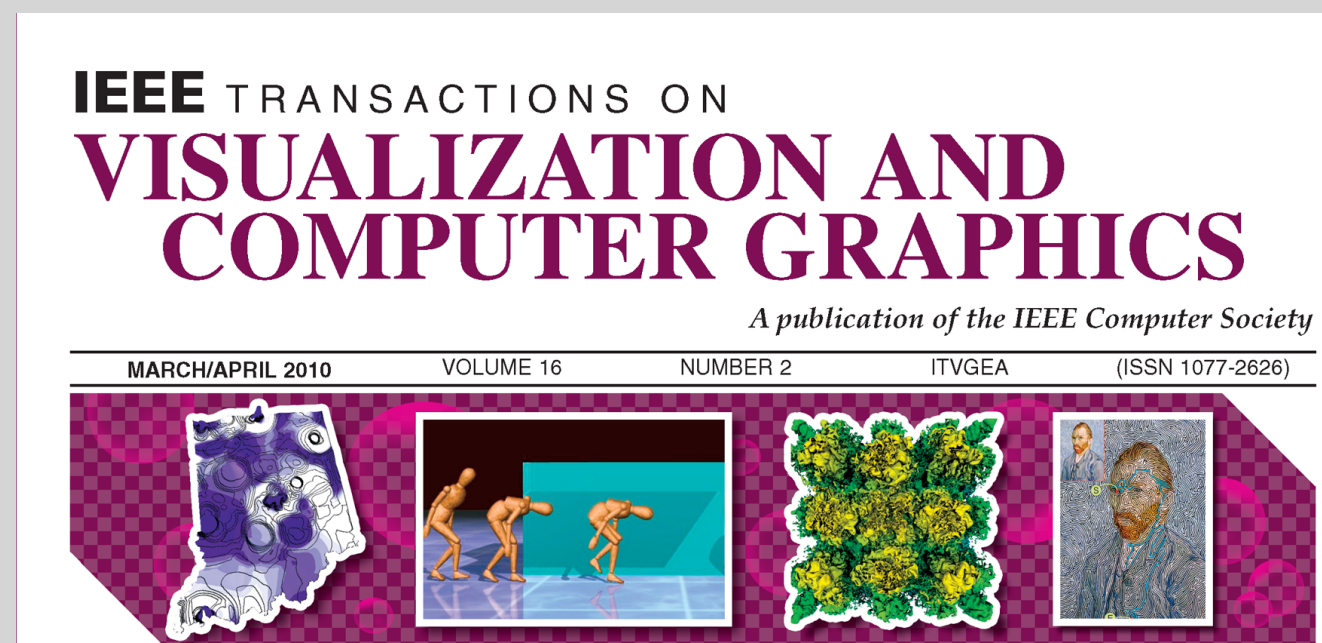


Why visualize data?

Analysis

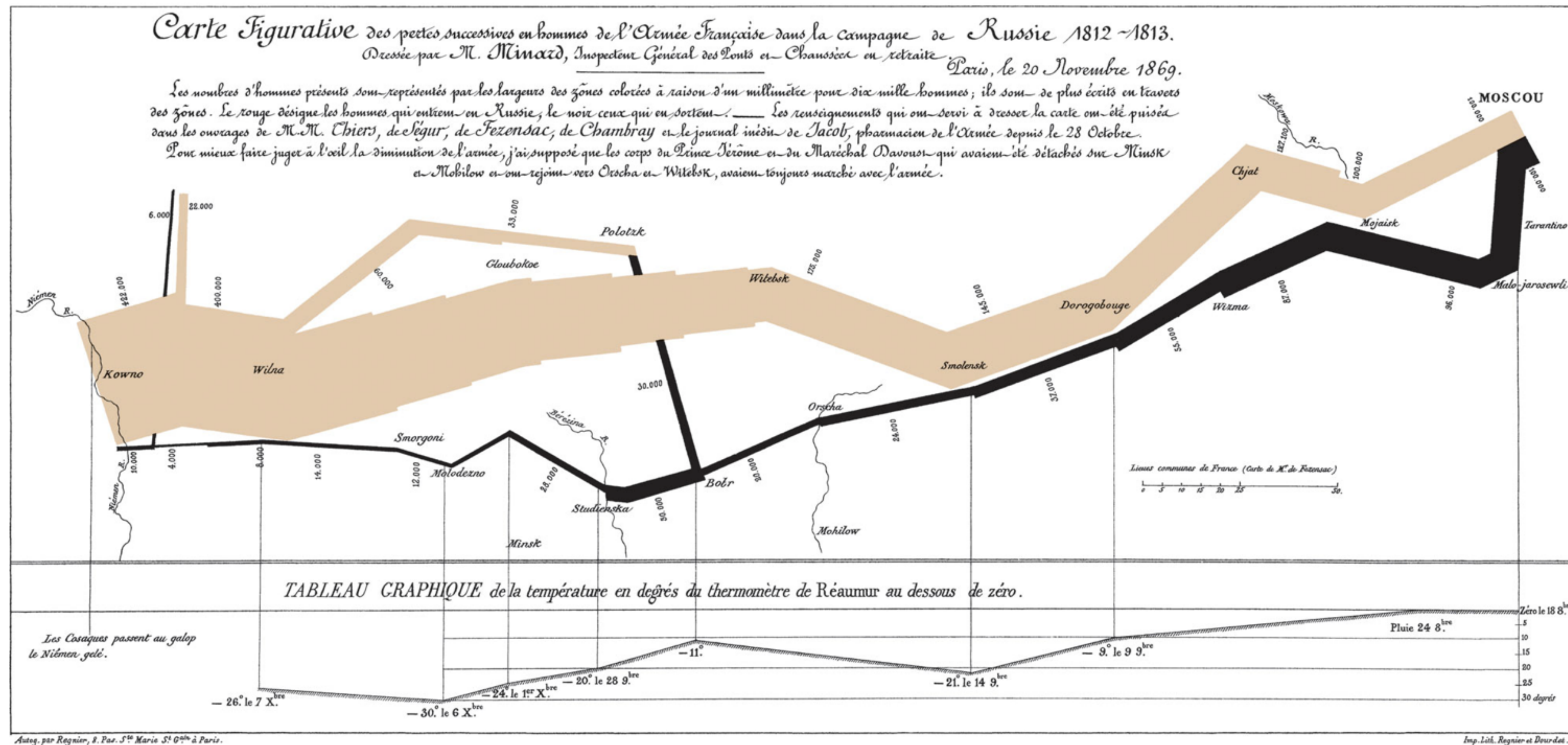
Communication

Decision making



There is a whole research community dedicated to DataVis, can be a course on its own!

“The best plot ever”



A group of people, mostly young adults, are gathered on a city street. In the foreground, a woman with blonde hair, wearing a light blue button-down shirt and a lanyard, points her right arm high into the air. She is looking upwards with an expression of surprise or excitement. Behind her, several other people are also looking up. A woman in a floral top is holding a smartphone, and another woman in a blue floral dress is looking towards the camera. The background shows a street with buildings and a red awning. A large white speech bubble is overlaid on the right side of the image, containing the text "That plot sucks because X, Y, Z".

**That plot sucks
because X, Y, Z**

OBAMACARE ENROLLMENT

6,000,000

**AS OF
MARCH 27**

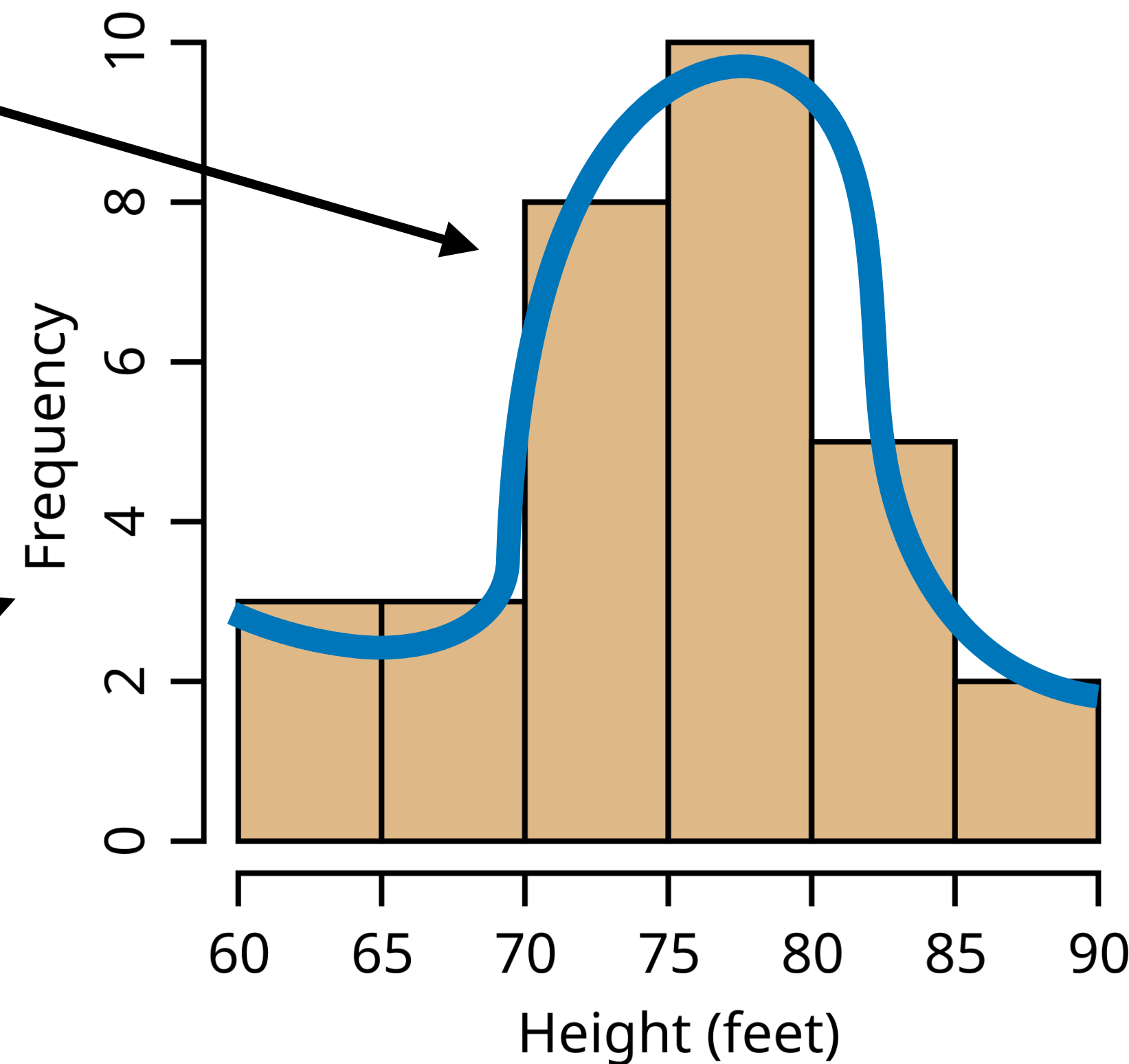
7,066,000

**MARCH 31
GOAL**

One Variable: Histograms

Kernel density estimate:
generate a smooth
approximation of the
distribution with a kernel
function (a “window”).

Heights of Black Cherry Trees



Potential options:

- Frequency (heights sum to n)
- Percentage (heights sum to 1)
- Density (area sums to 1)

- Let P_{75} be the height of the 75th largest tree;
- Let P_{25} be the height of the 25th largest tree;

Freedman–Diaconis rule:

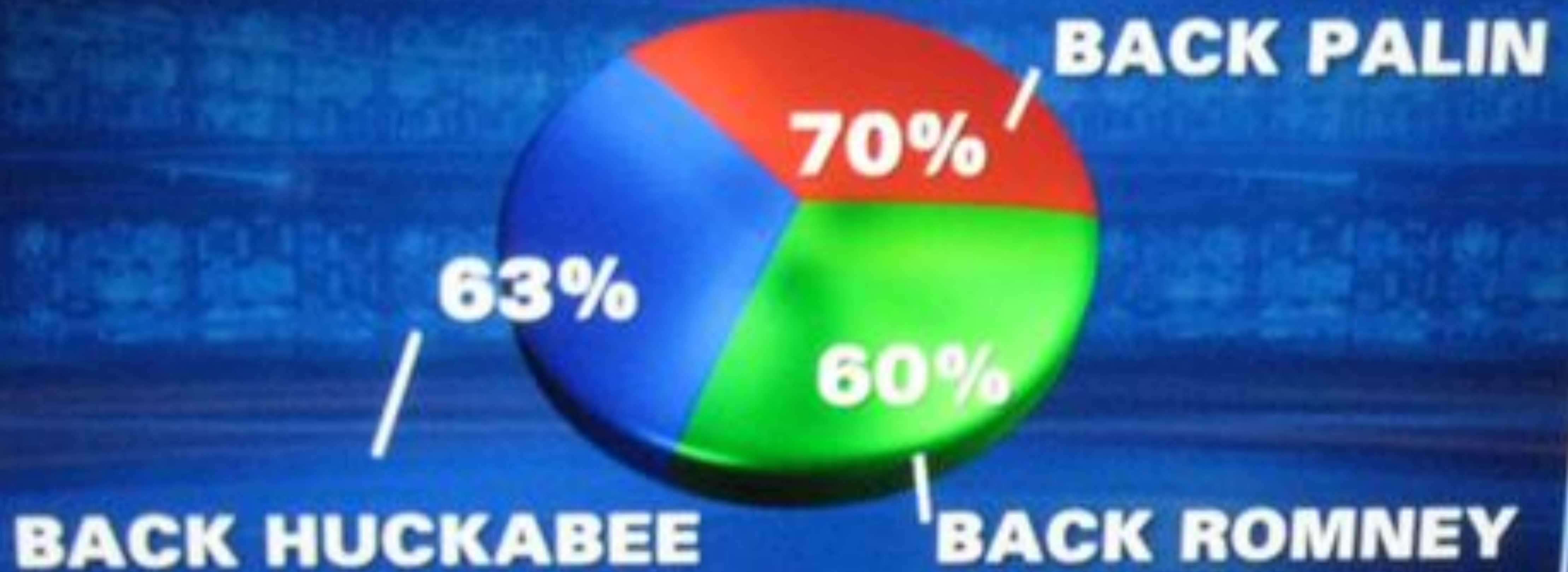
Choose the bin width equal to:

$$h = 2 \frac{P_{75} - P_{25}}{n^{1/3}}$$

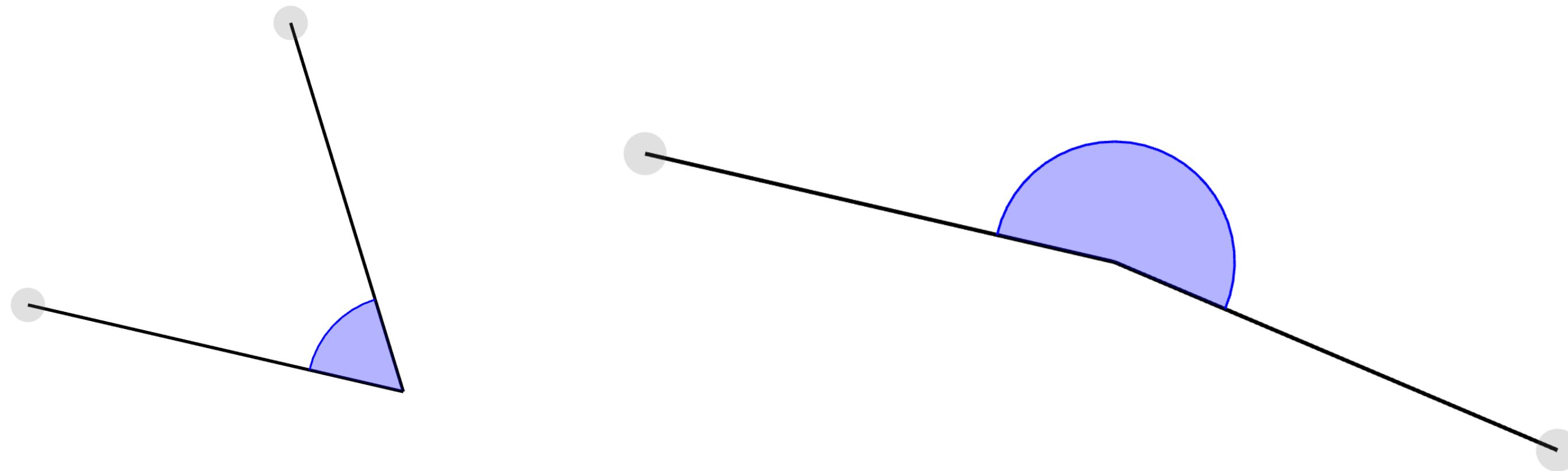
Bin sizes are consequential!
How to choose them?

2012 PRESIDENTIAL RUN

GOP CANDIDATES



Why not pie charts?



How much bigger is the angle on the right?

A) $\times 3.16$

B) $\times 3.74$

C) $\times 4.12$

Why not pie charts?



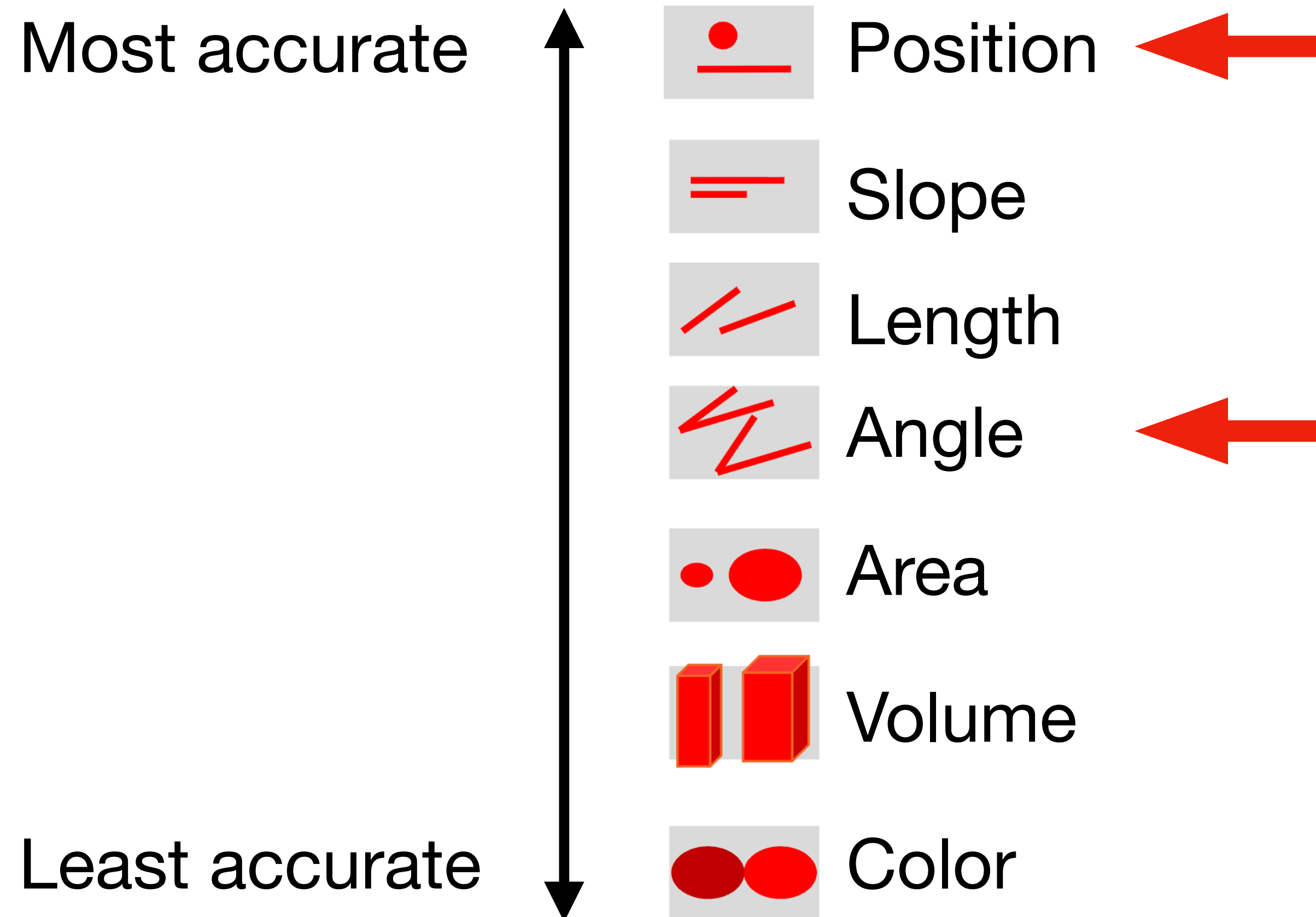
How much bigger is the top bar?

A) $\times 1.56$

B) $\times 2.24$

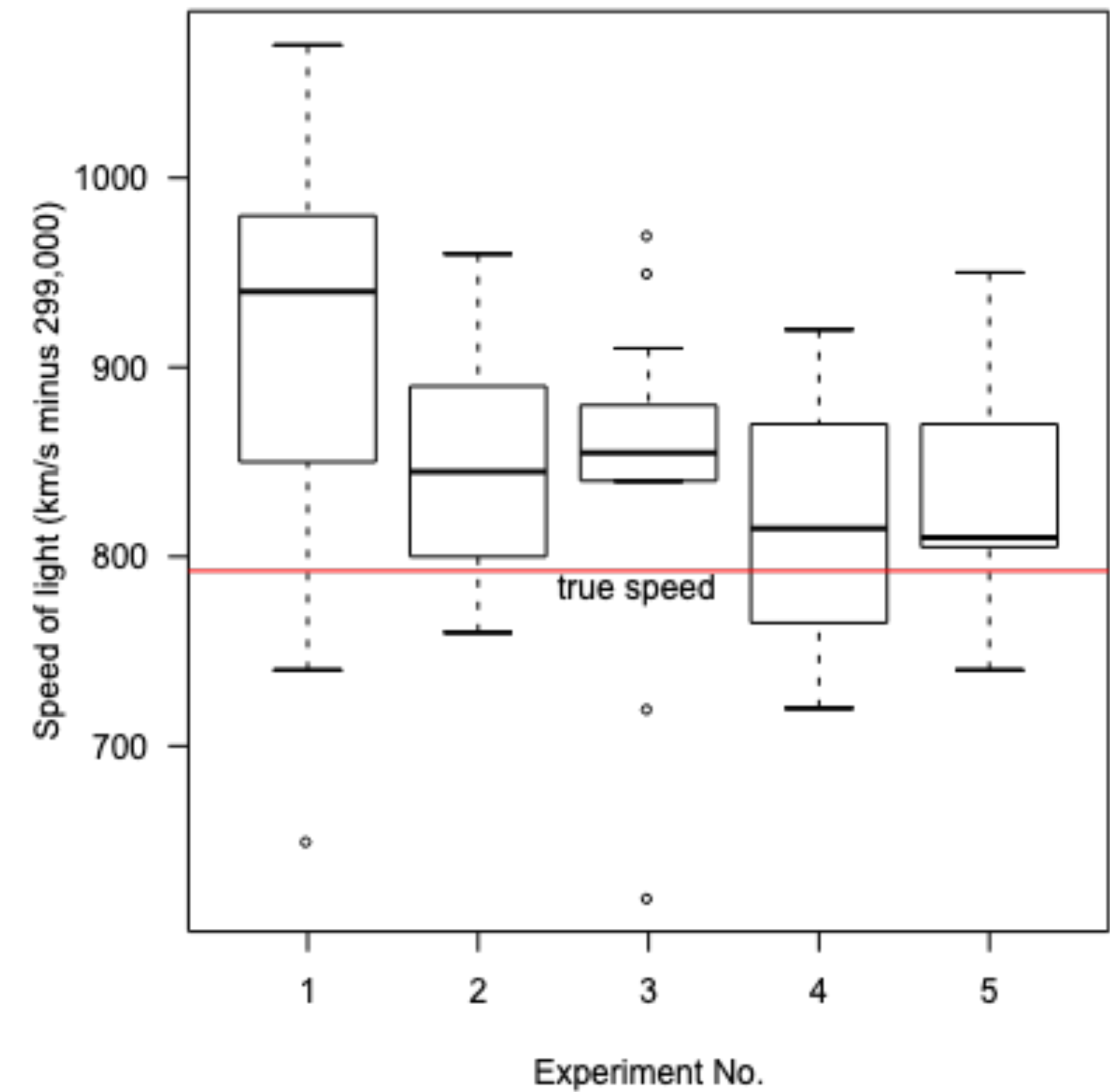
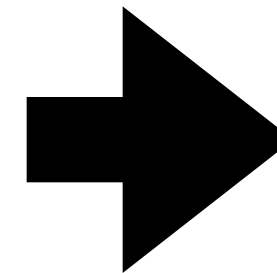
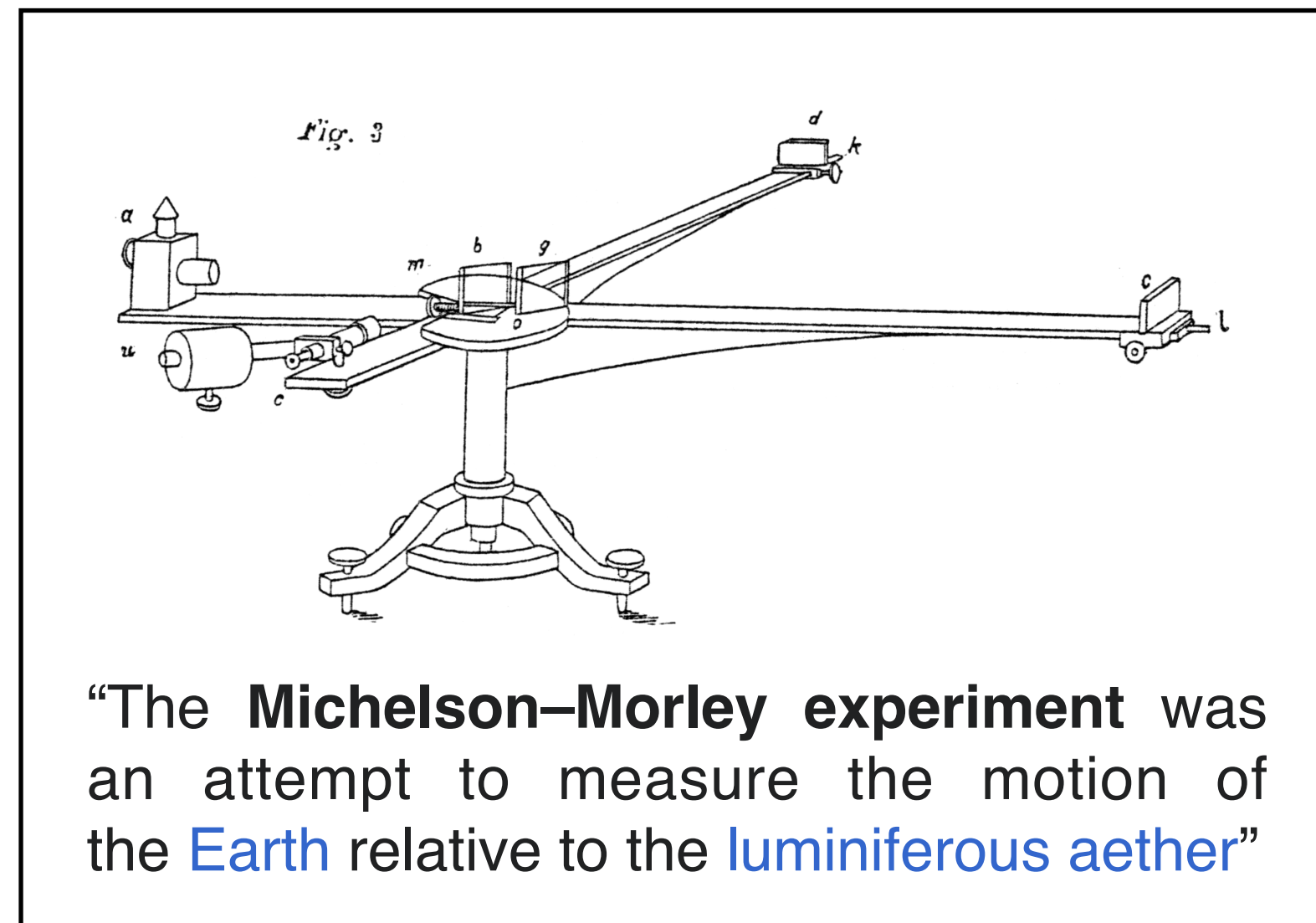
C) $\times 2.99$

Perception of magnitudes

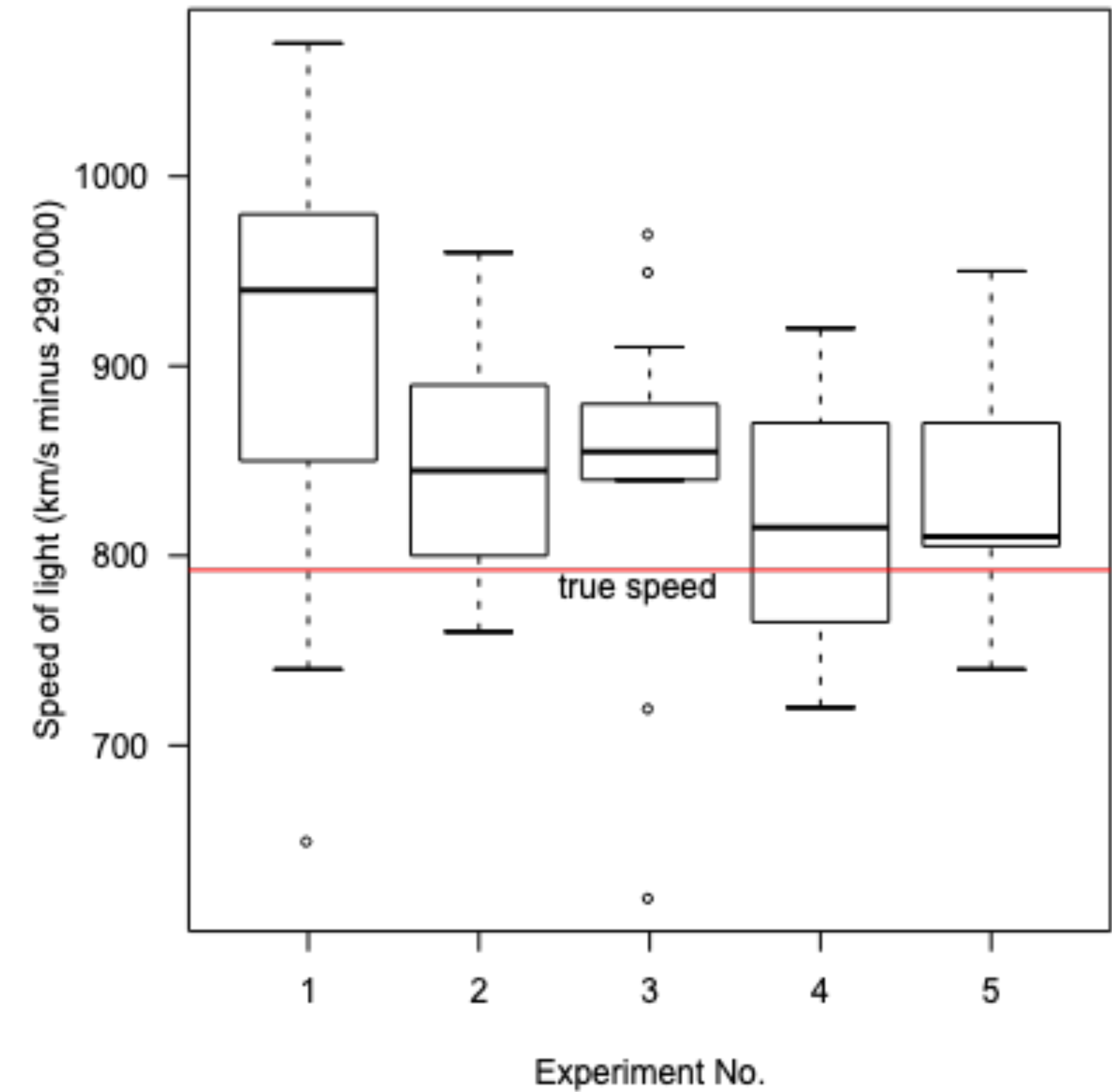
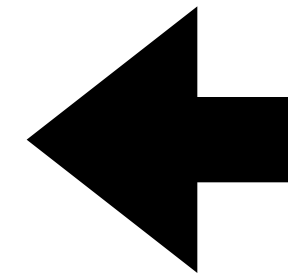
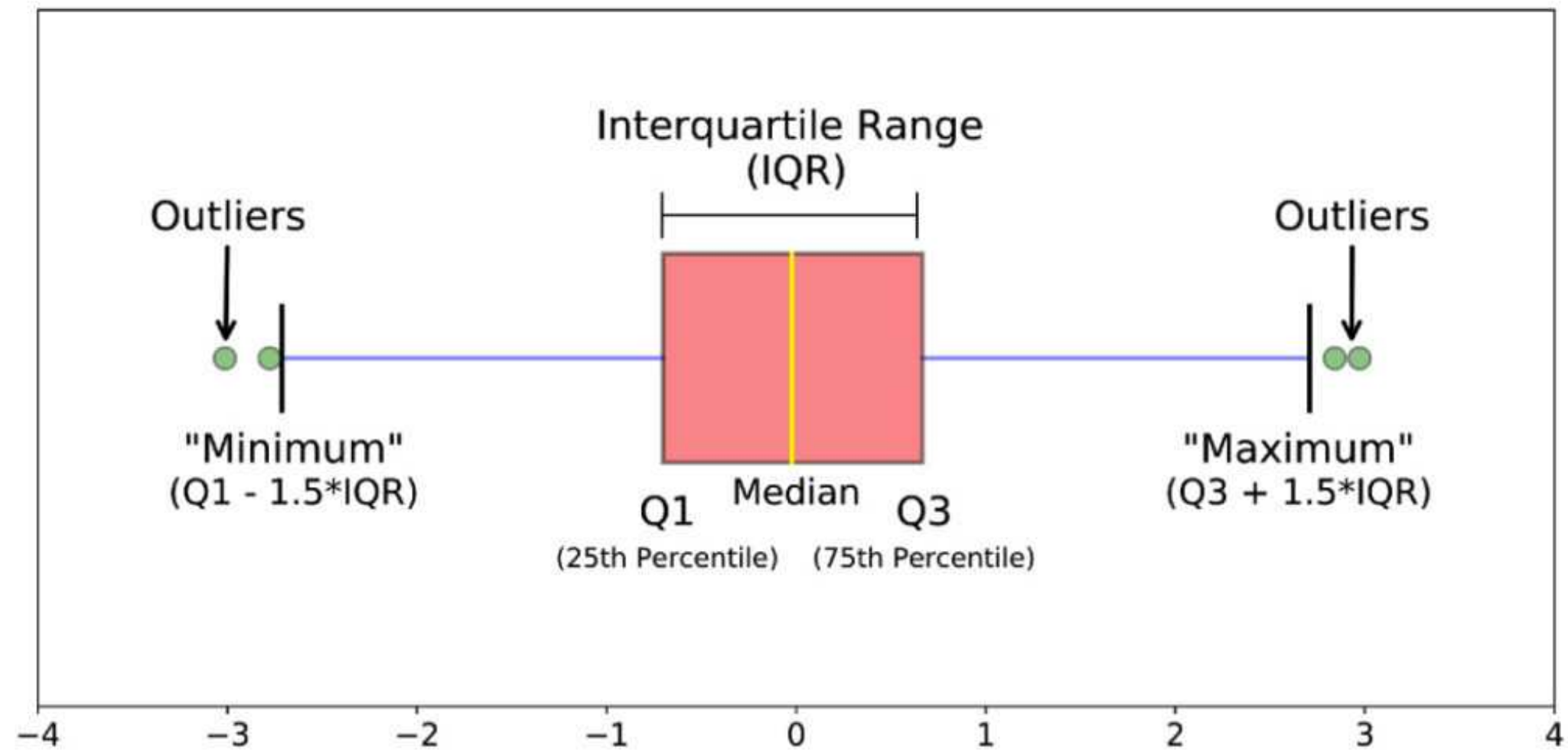


Cleveland & McGill (1984)
*Graphical Perception: Theory,
Experimentation, and Application
to the Development of Graphical
Methods*

One Variable: Box Plots



One Variable: Box Plots

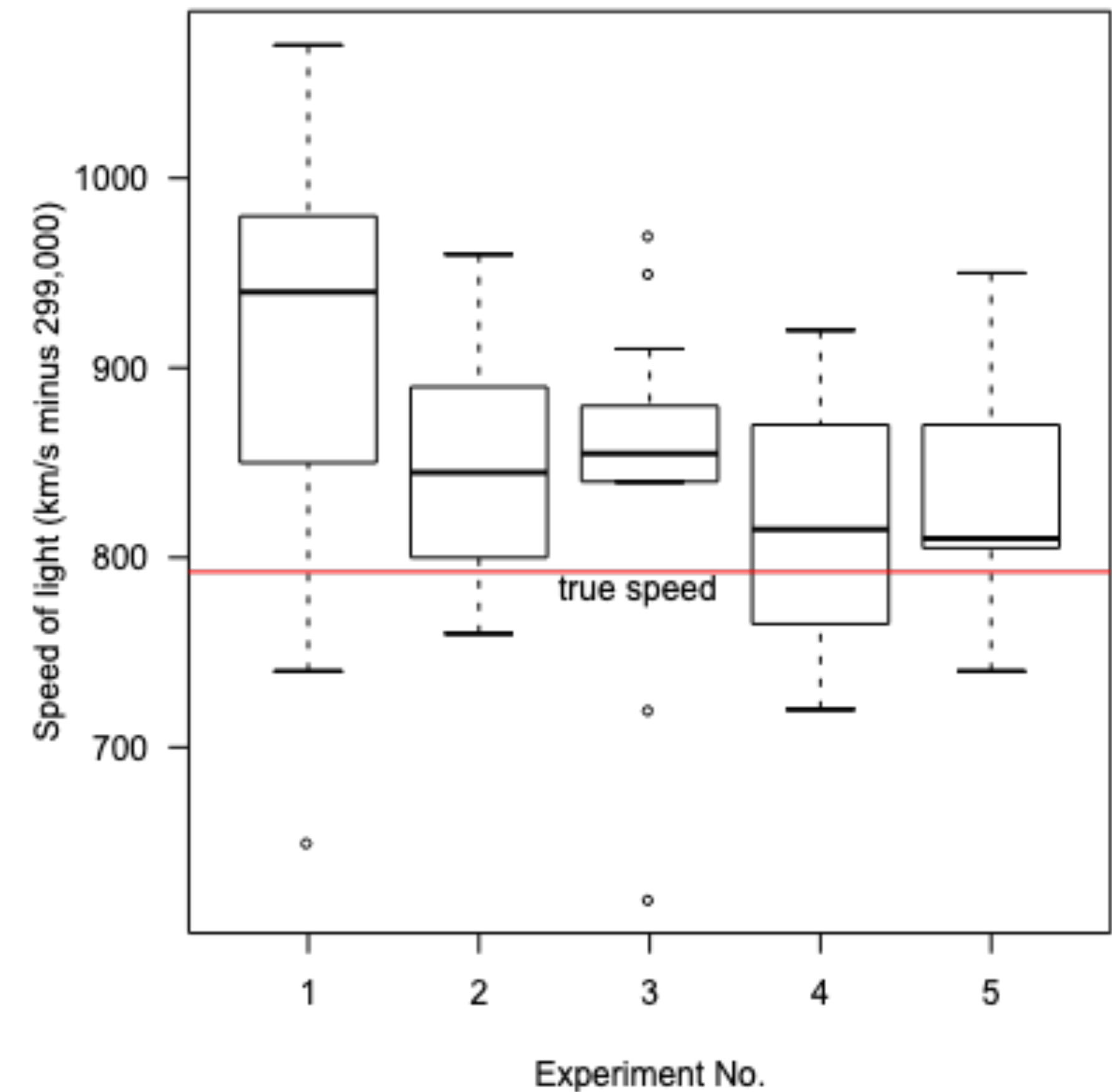
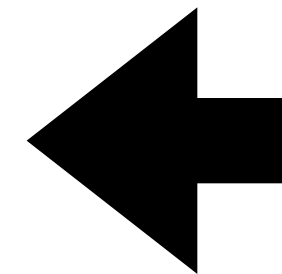


One Variable: Box Plots

In the first experiment, what % of the samples were below 850 km/s?

In the second experiment, what was the median speed of the samples?

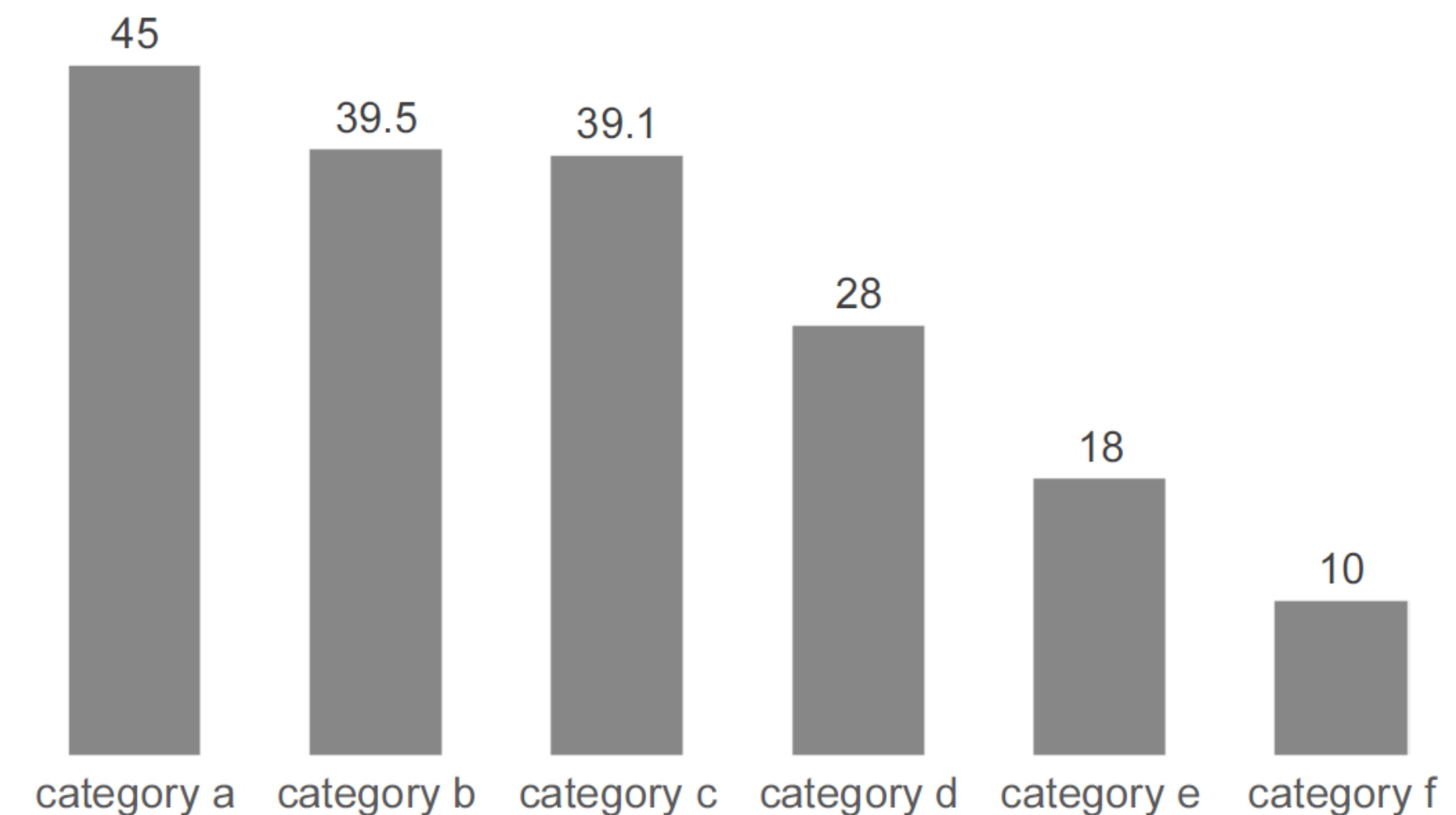
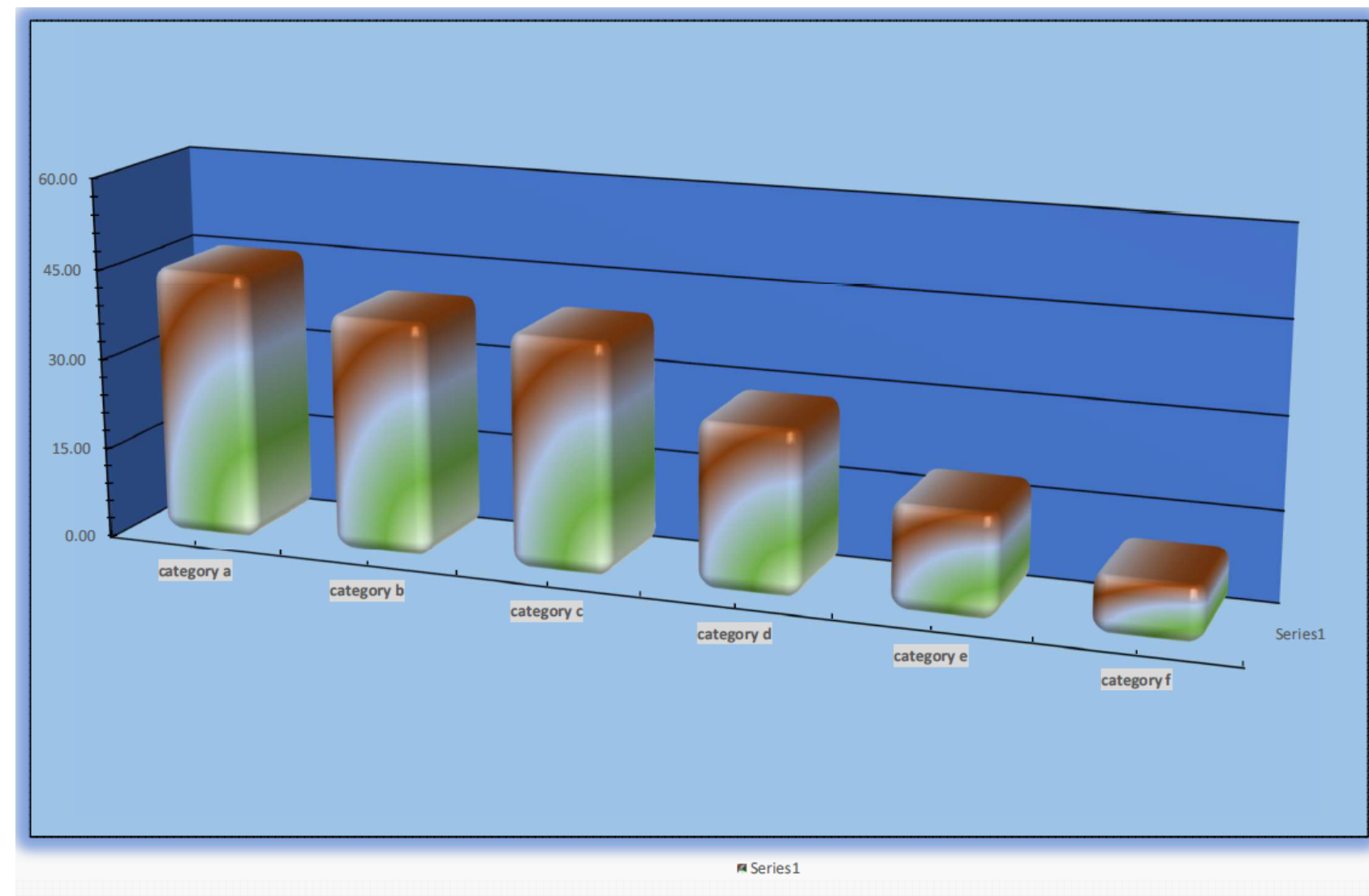
In the third experiment, what was the sample with the lowest recorded speed?



Box plots = high data-ink ratio

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$

= proportion of ink devoted to non-redundant stuff



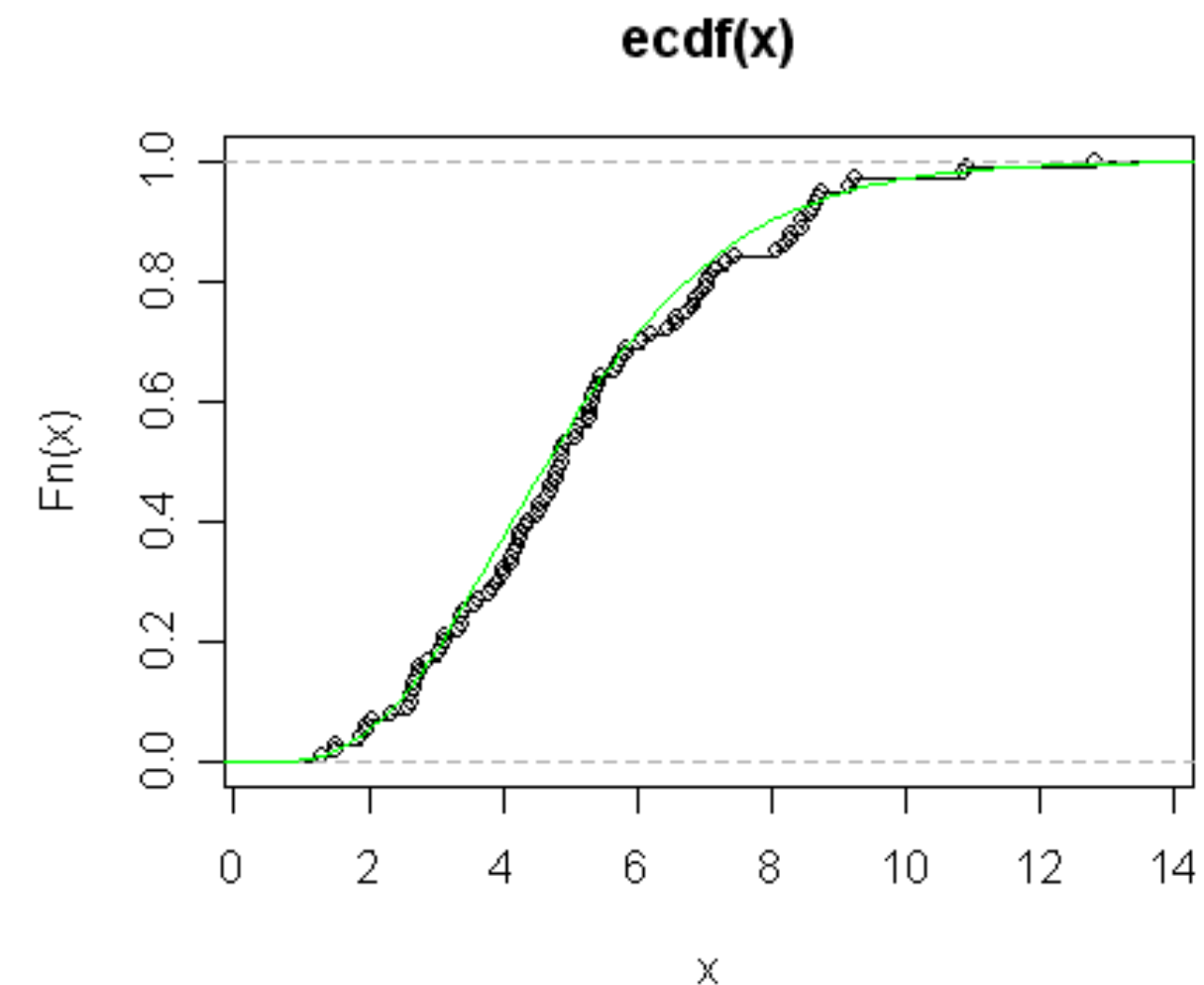
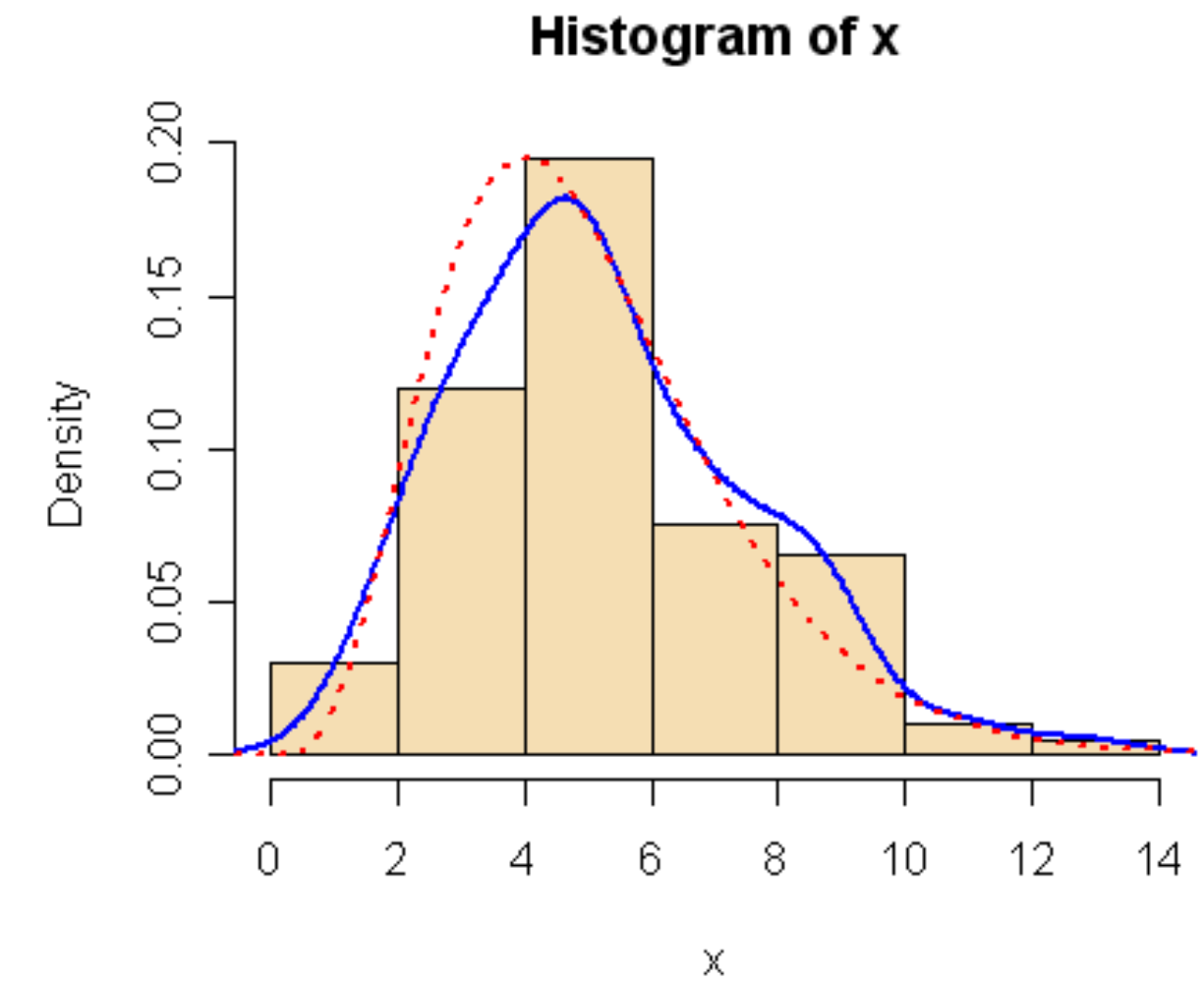
Empirical CDF Function

- Another way of visualizing a distribution!

$$Fn(x) = \Pr(X \leq x)$$

- **Alternative:** Complementary CDF

$$1 - Fn(x)$$



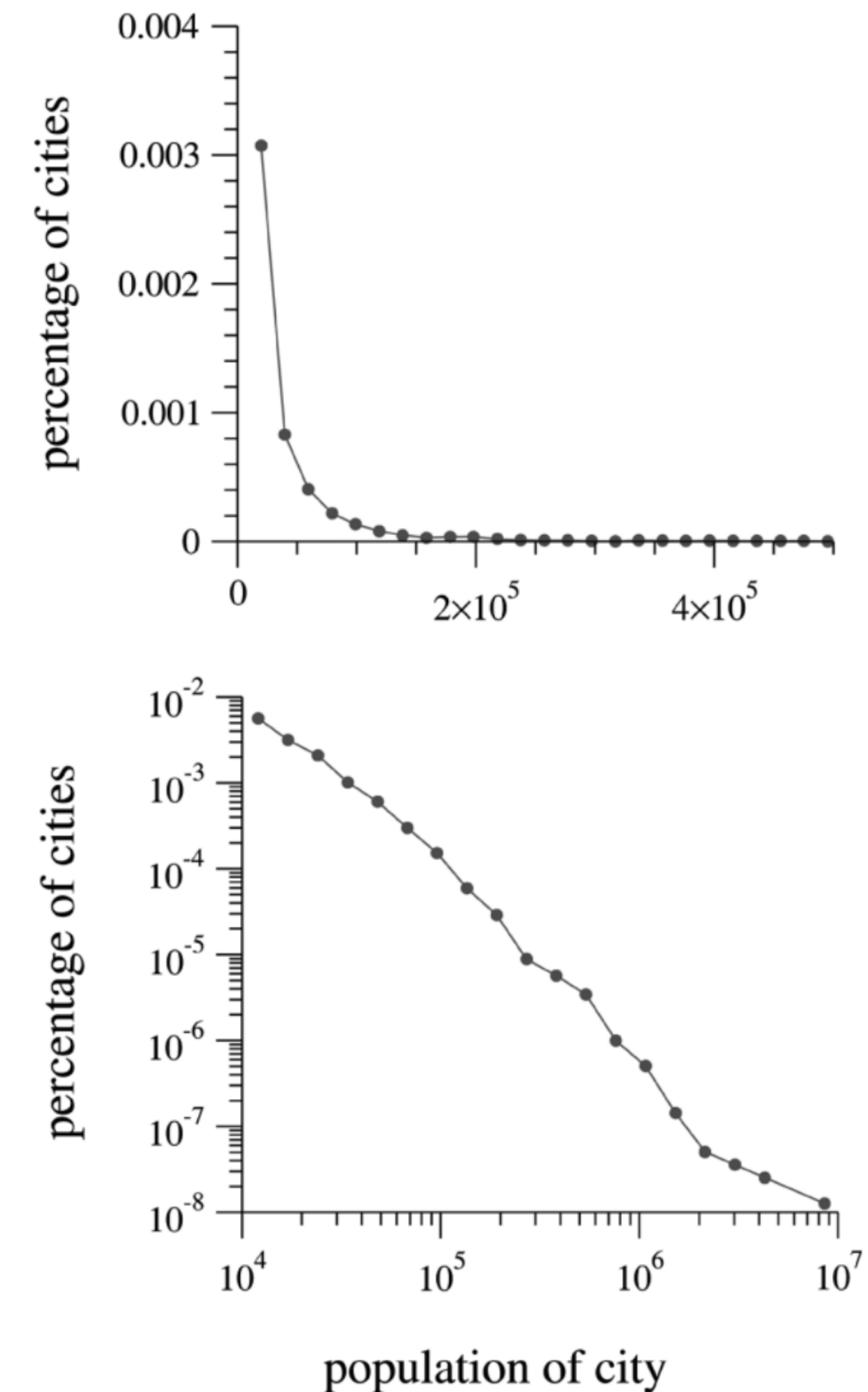
Scales and Power Laws

- Power law distribution

$$\Pr(x) = Cx^{-\alpha}$$

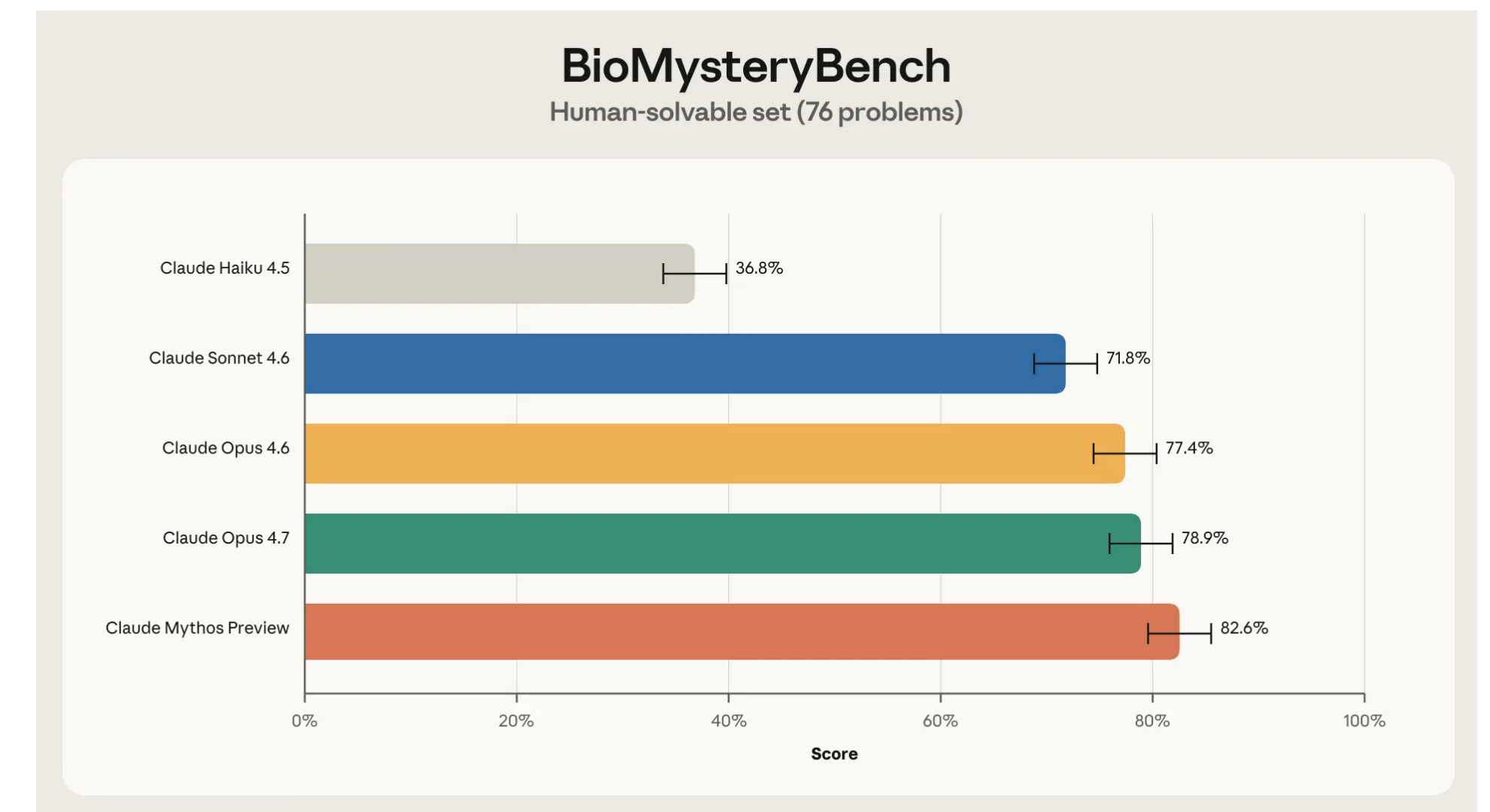
- Tricky to visualize and describe! For small α , mean, and variance are ∞
- Straight lines in log-log axes:

$$\log(\Pr(x)) = \log(C) - \alpha \log(x)$$



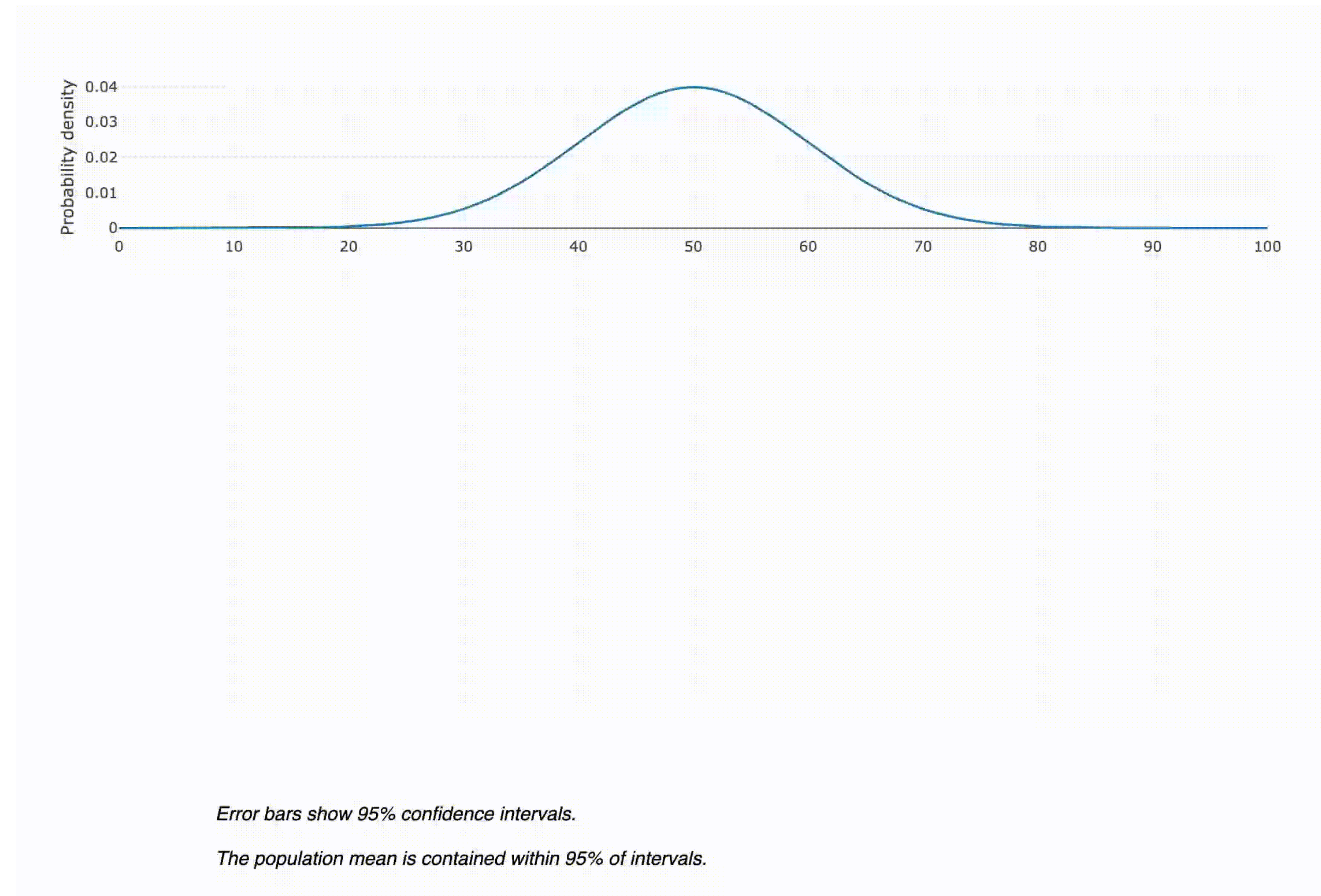
Visualizing Statistics

- Error bars are employed to denote the uncertainty of an estimate!
- Usually underspecified:
 - Standard error
 - Standard error of the mean
 - Prediction interval
 - Confidence intervals



(Link)

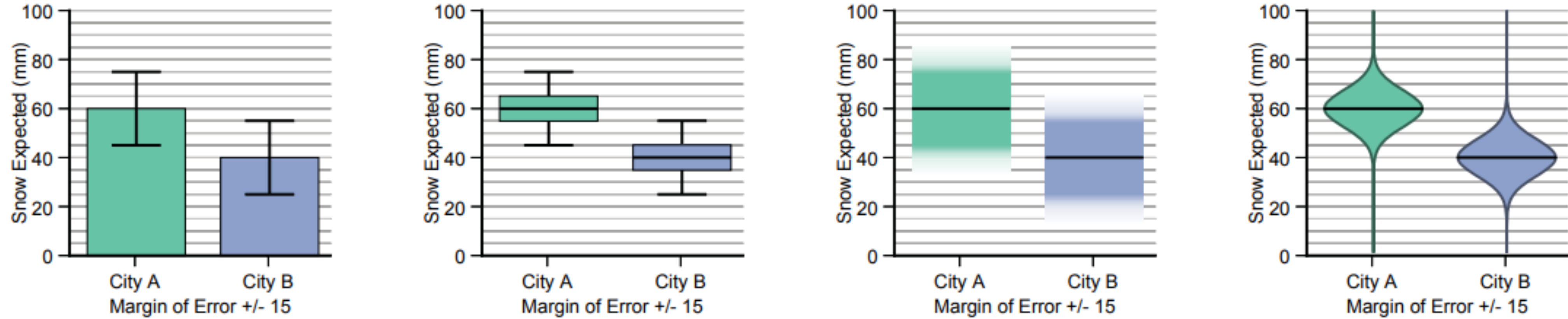
Visualizing Statistics



<https://logarithmic.net/2017/dance/>

Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error

Michael Correll *Student Member, IEEE*, and Michael Gleicher *Member, IEEE*



(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.

(b) **Modified box plot:** The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.

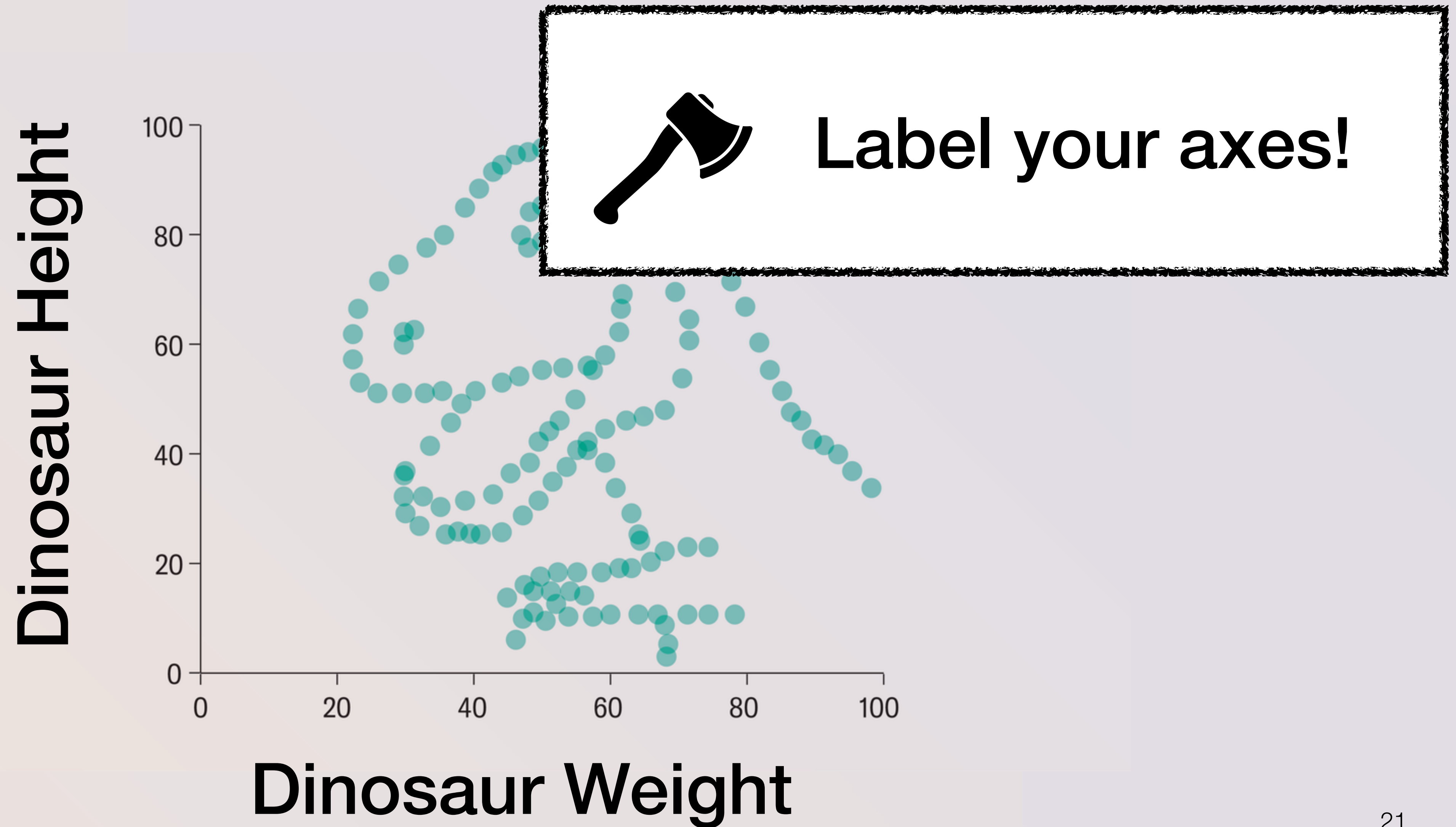
(c) **Gradient plot:** the transparency of the colored region corresponds to the cumulative density function of a t-distribution.

(d) **Violin plot:** the width of the colored region corresponds to the probability density function of a t-distribution.

Fig. 1. Four encodings for mean and error evaluated in this work. Each prioritizes a different aspect of mean and uncertainty, and results in different patterns of judgment and comprehension for tasks requiring statistical inferences.

[\(Link\)](#)

Scatter plot



All of the following 13 graphs have the same summary statistics:

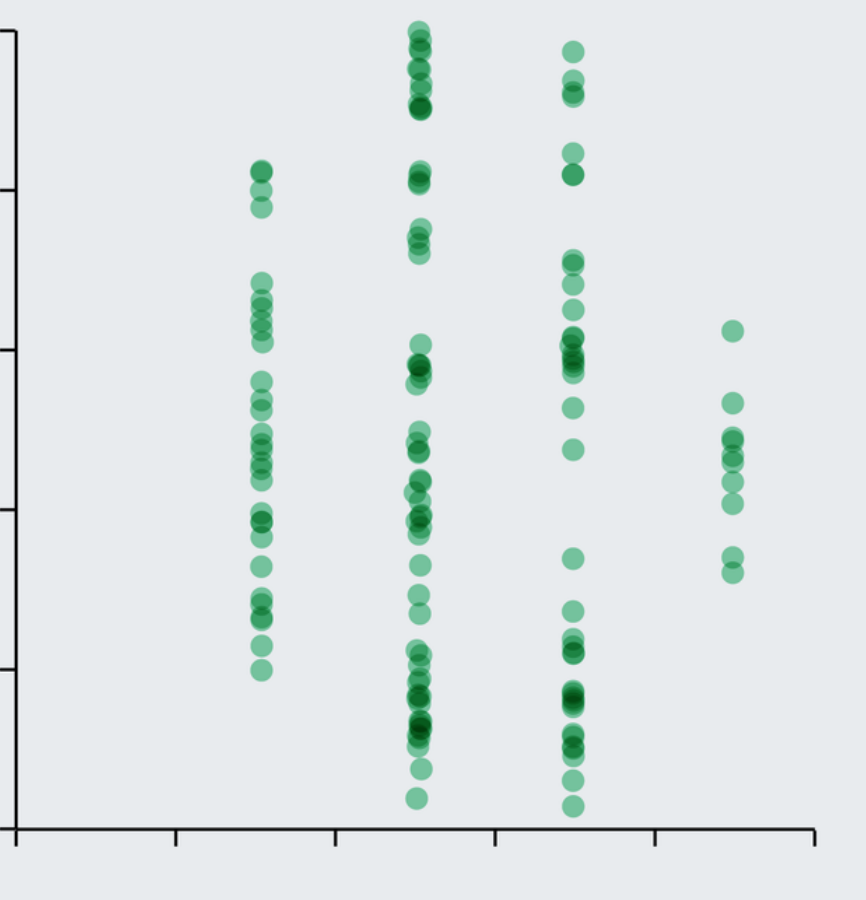
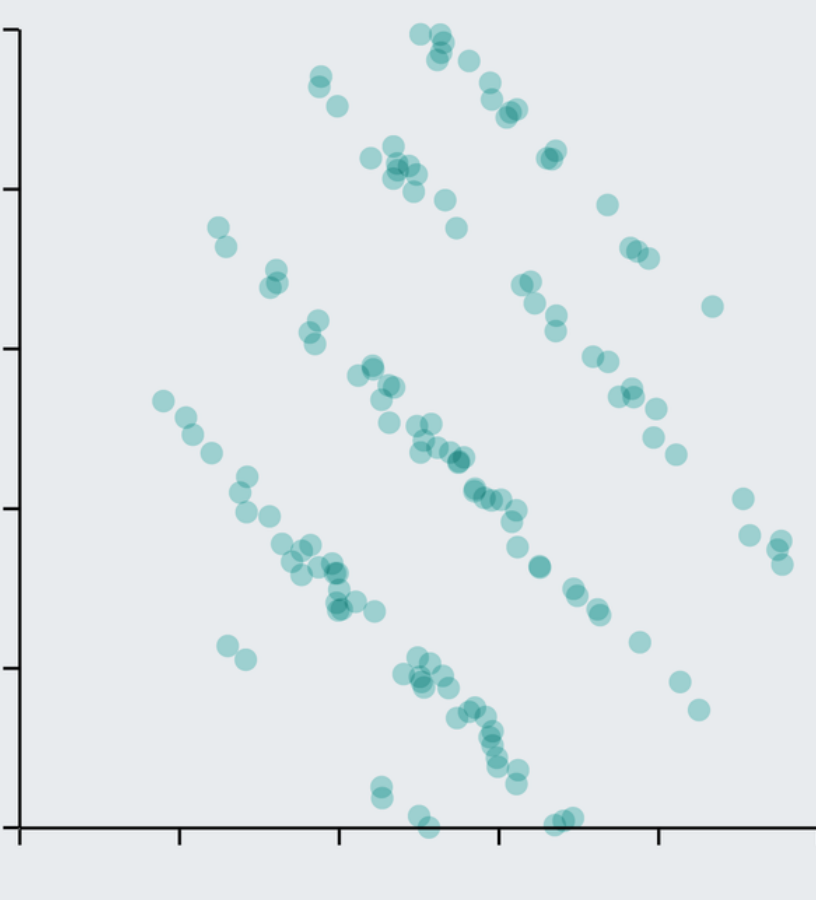
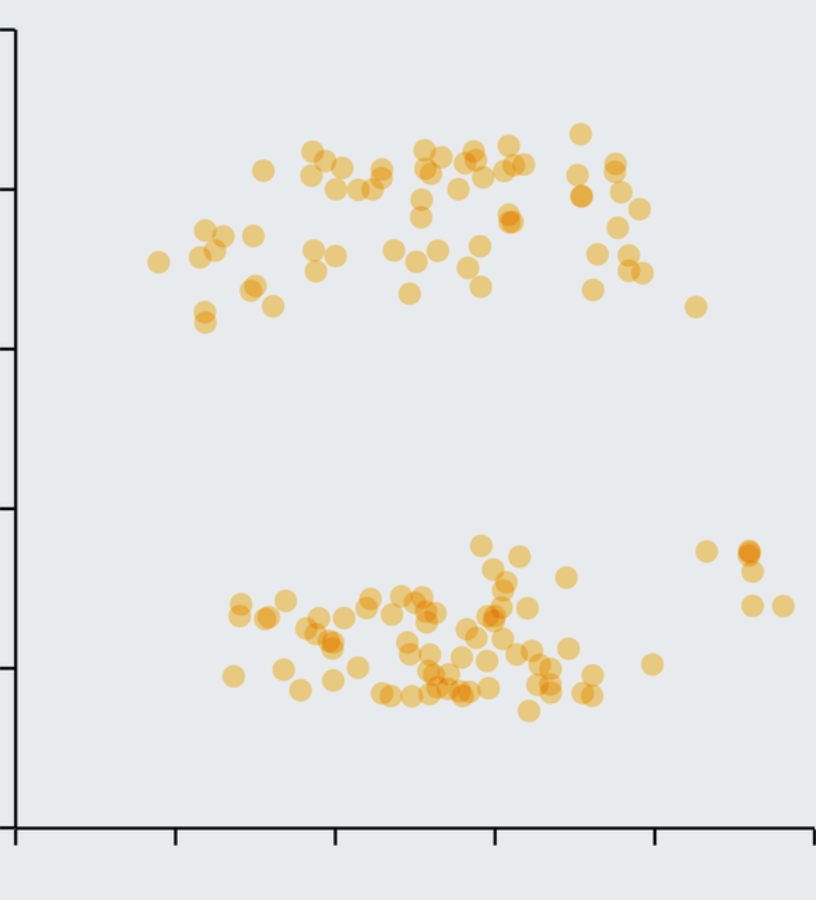
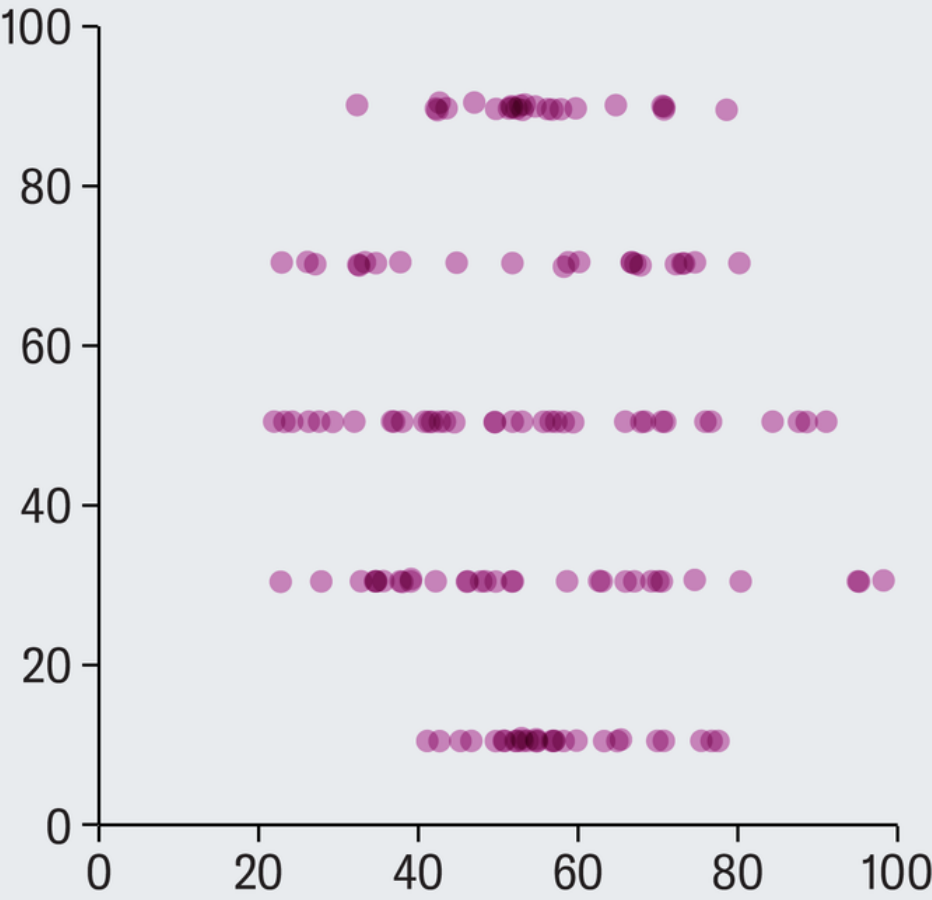
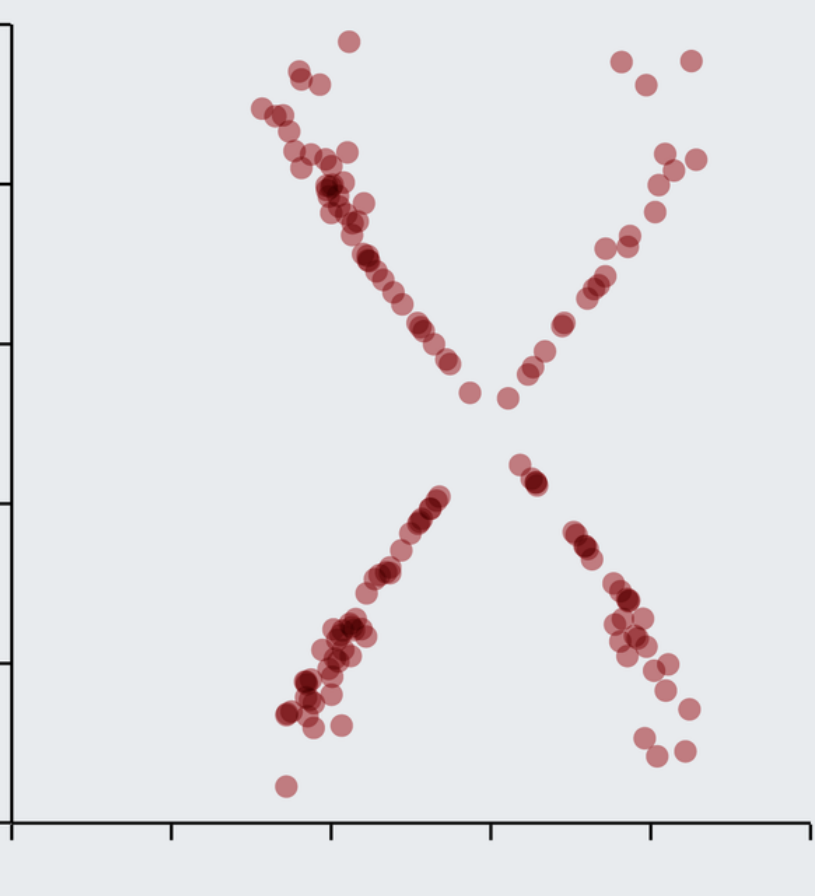
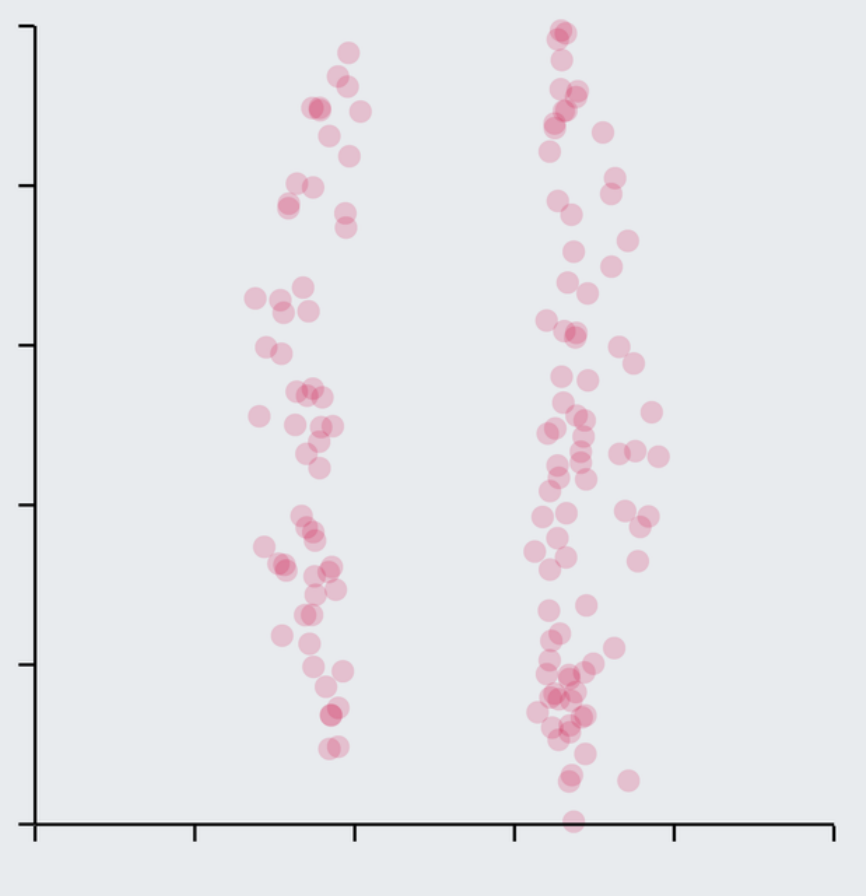
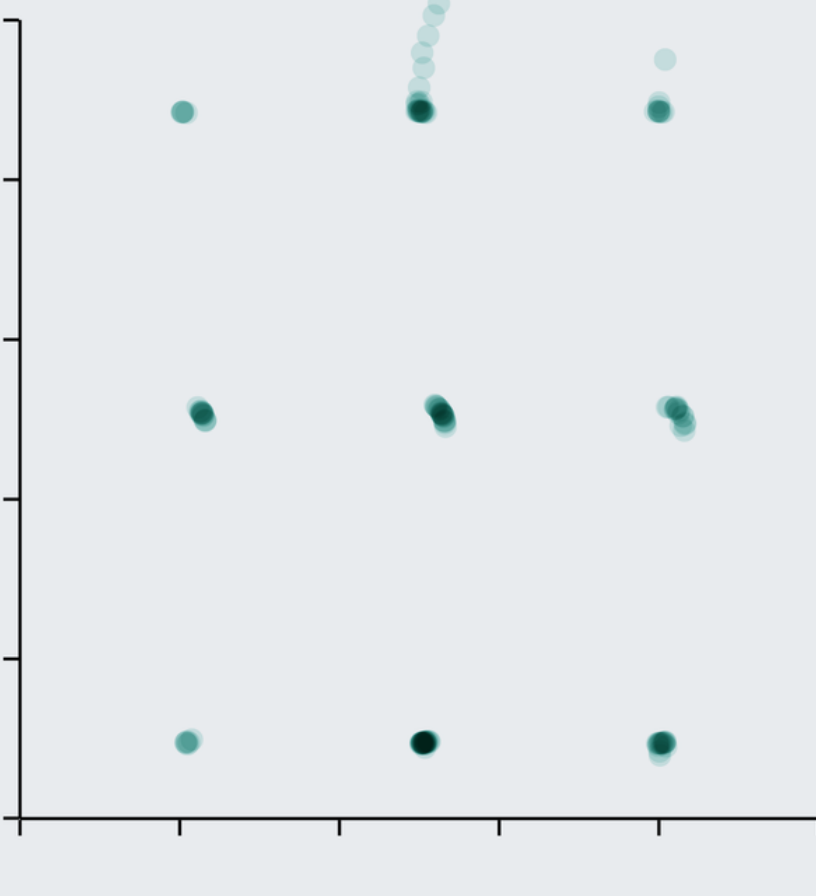
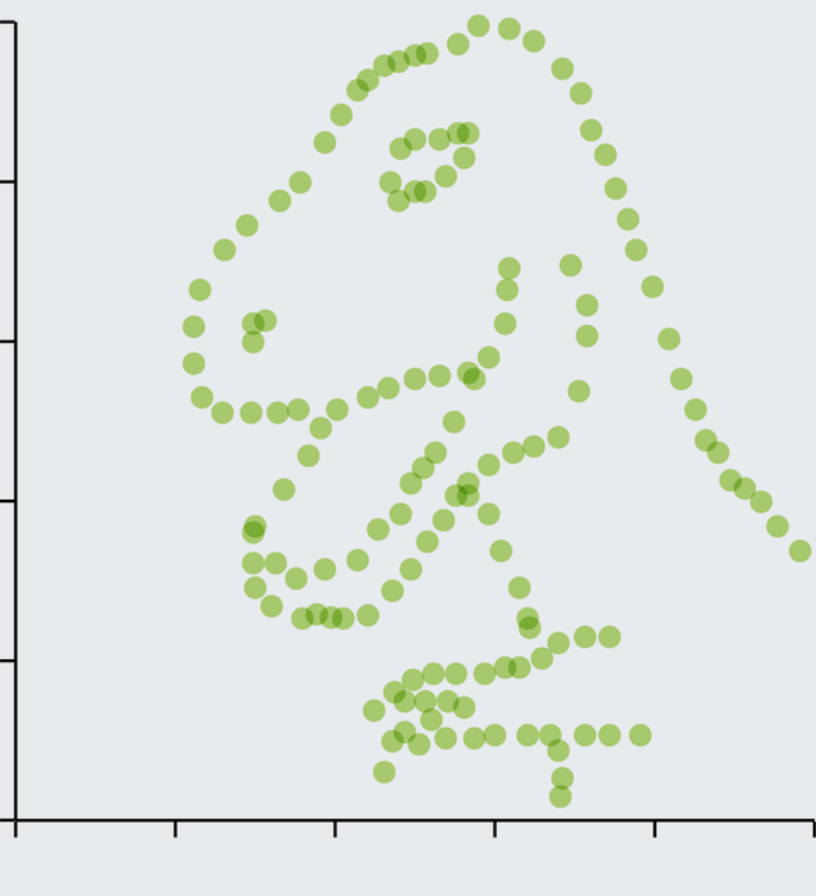
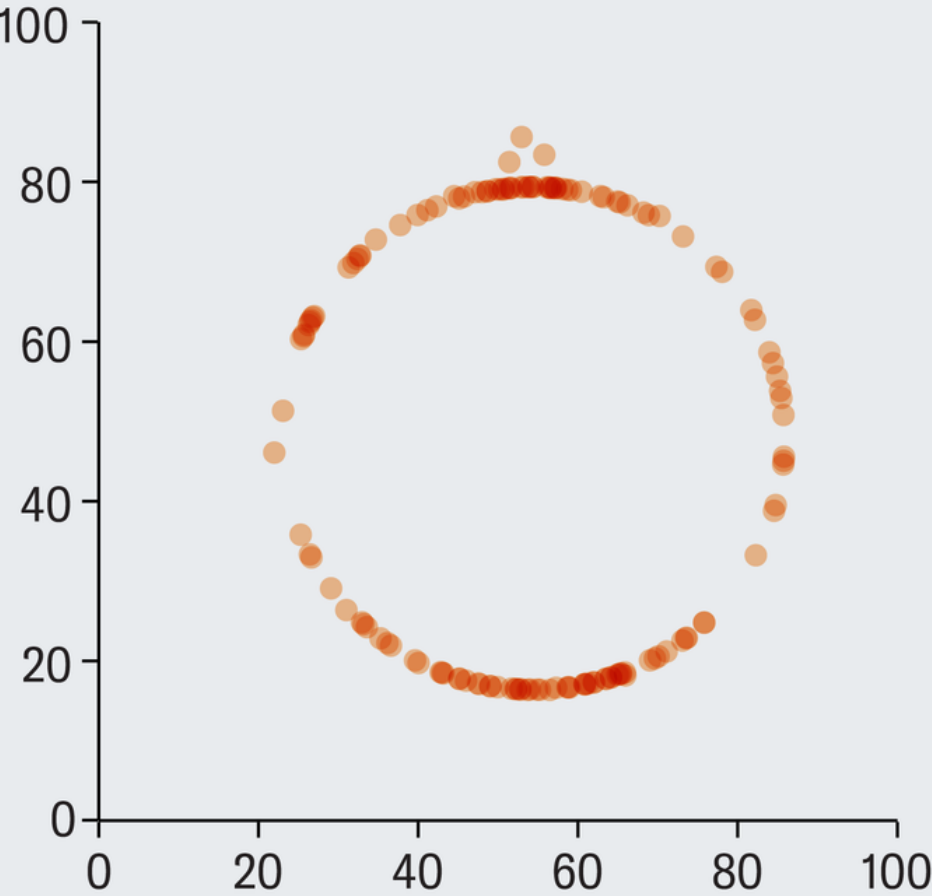
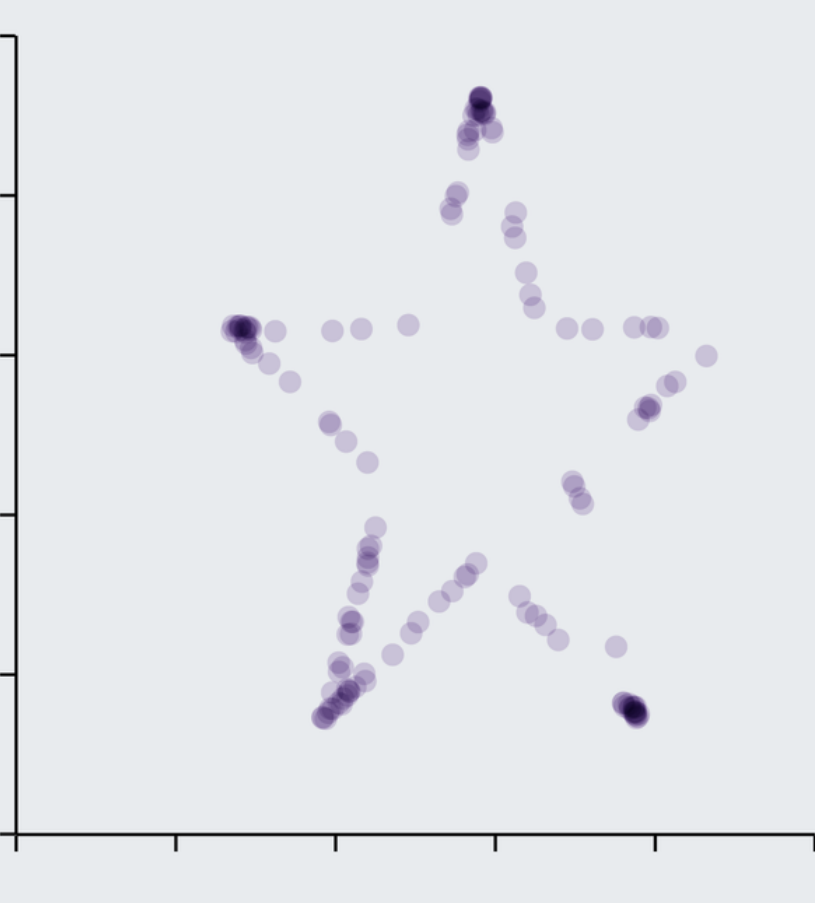
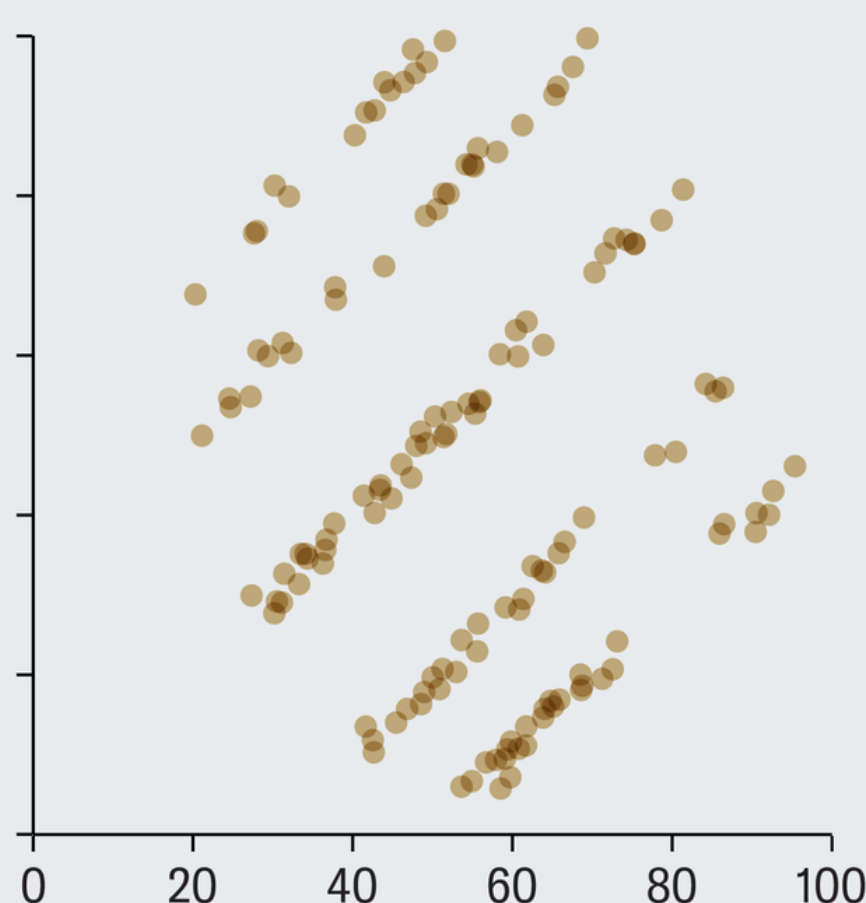
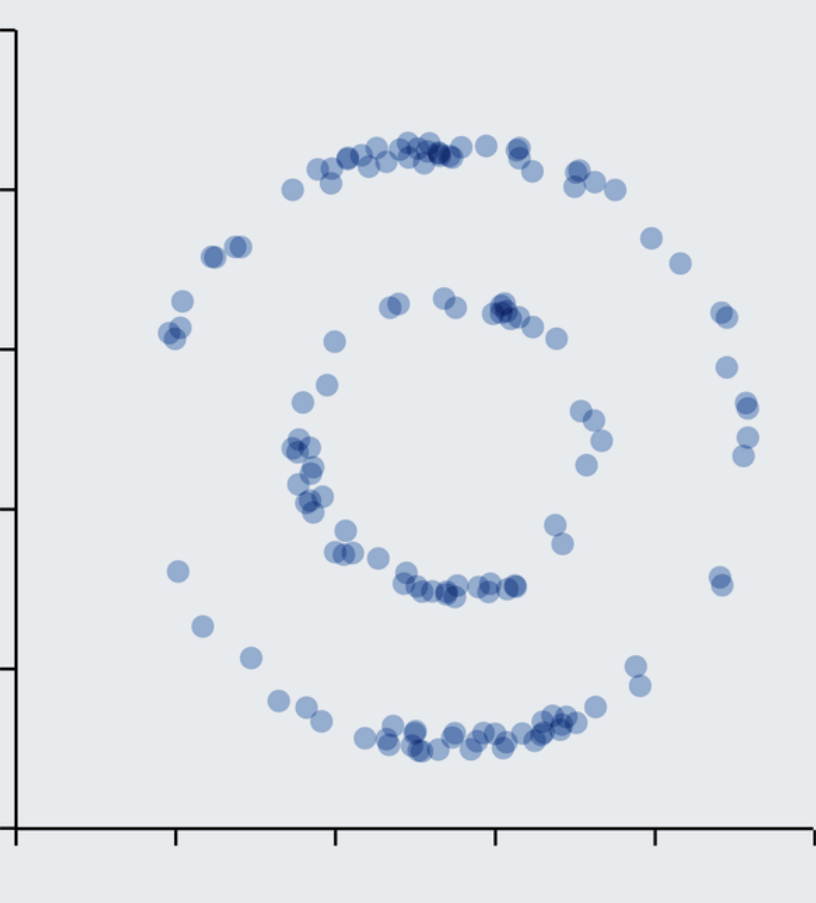
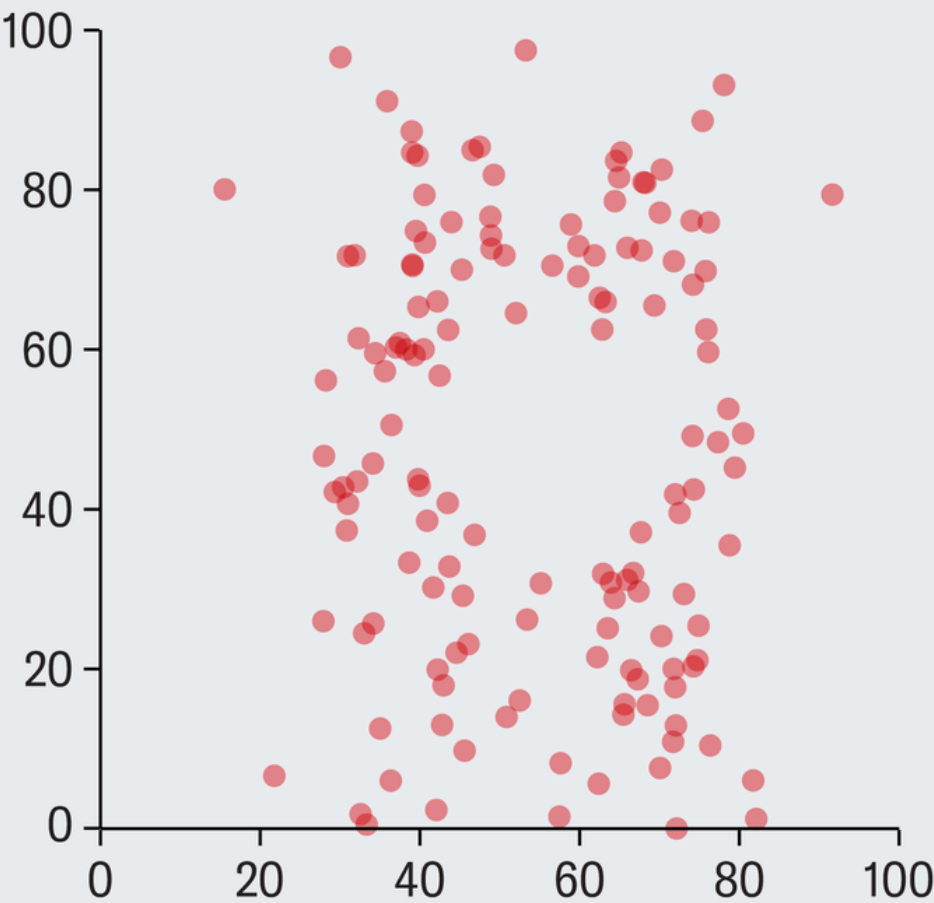
Mean, x-axis = 54.26

Mean, y-axis = 47.83

Standard deviation, x-axis = 16.76

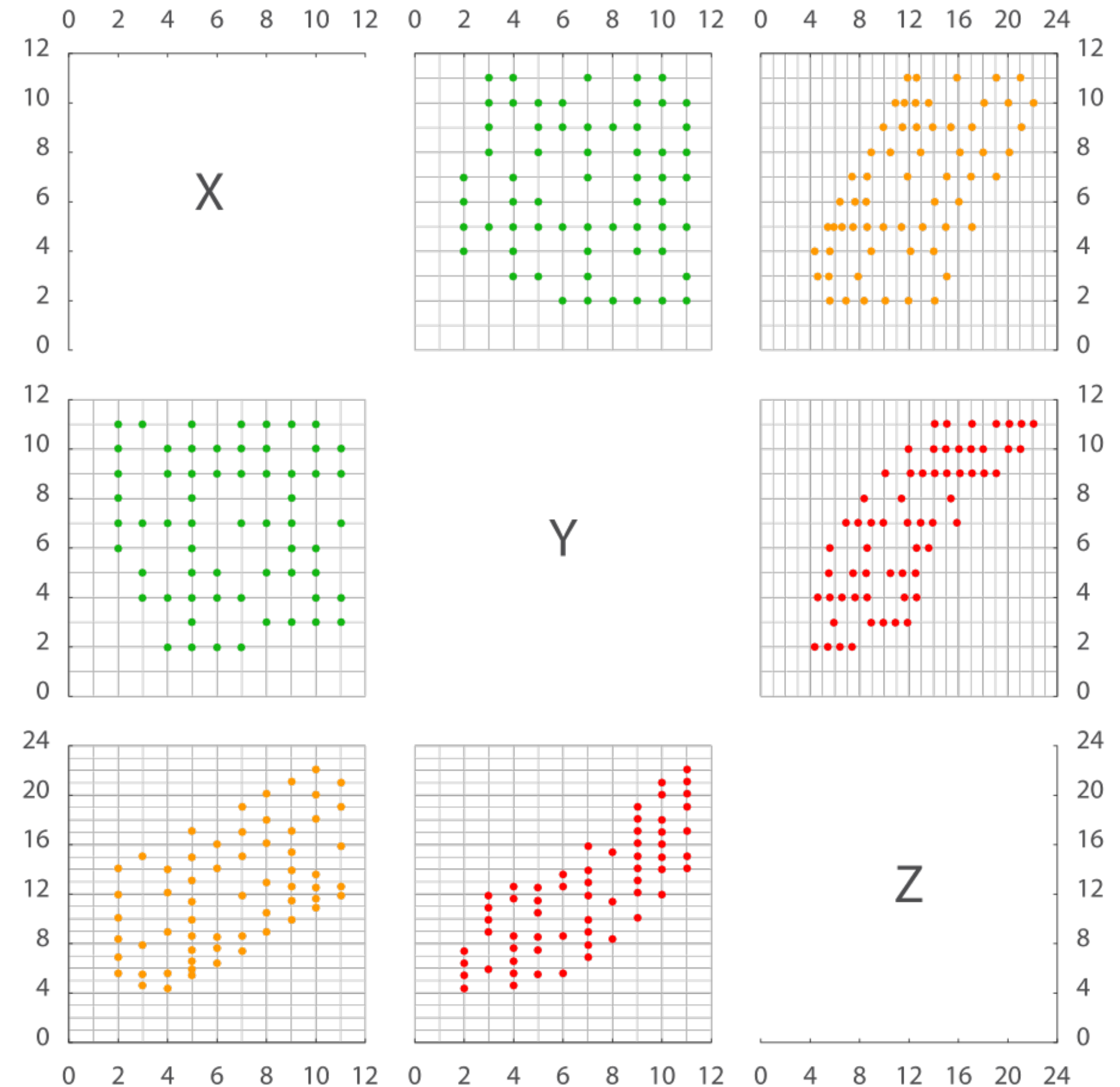
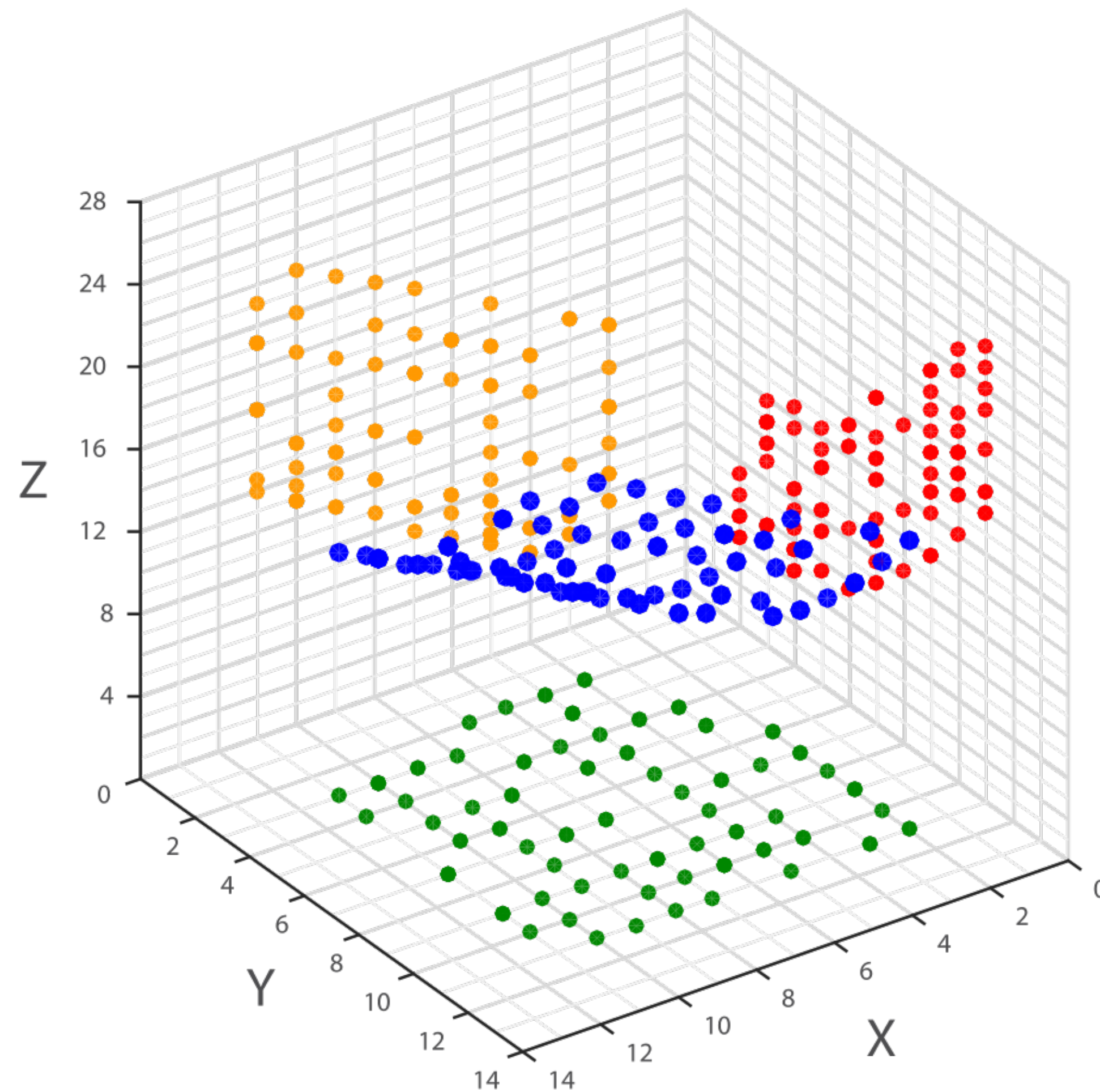
Standard deviation, y-axis = 26.93

Correlation = -0.06



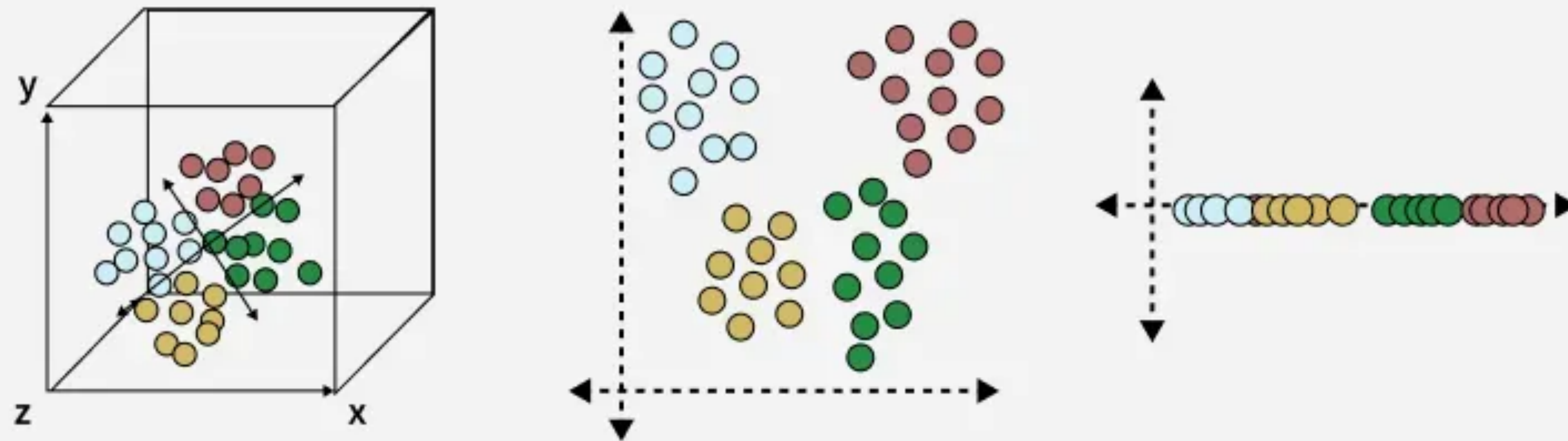
(Link)

Scatter plot matrix



Dimensionality Reduction

We are very bad at visualizing data in three dimensions (let alone four)



All of the following 13 graphs have the same summary statistics:

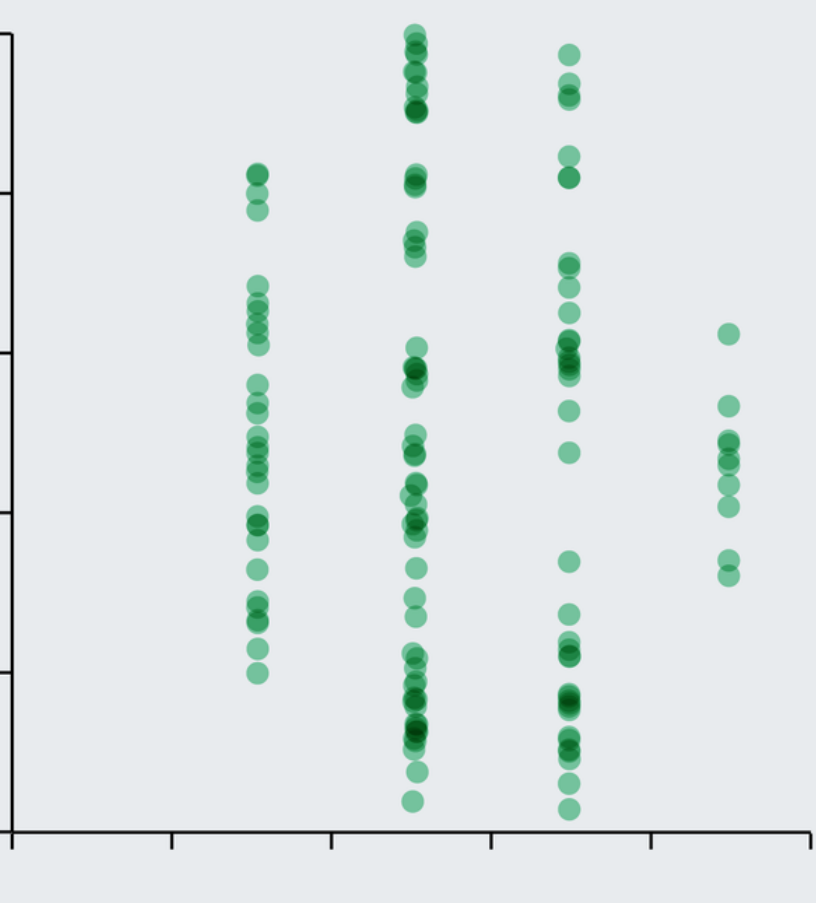
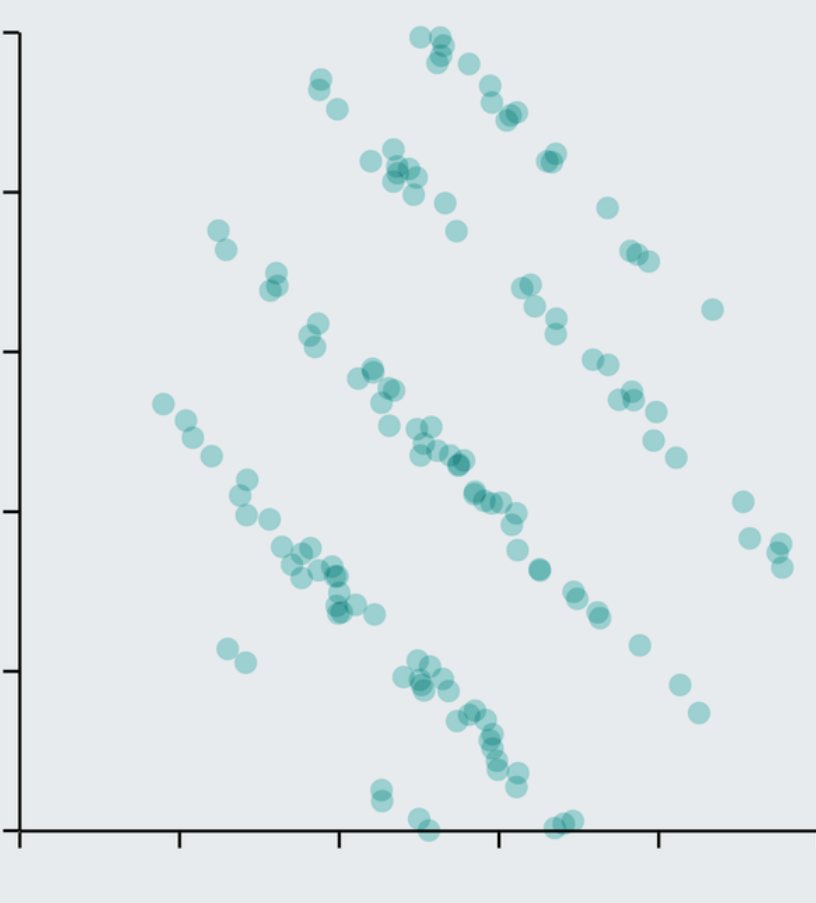
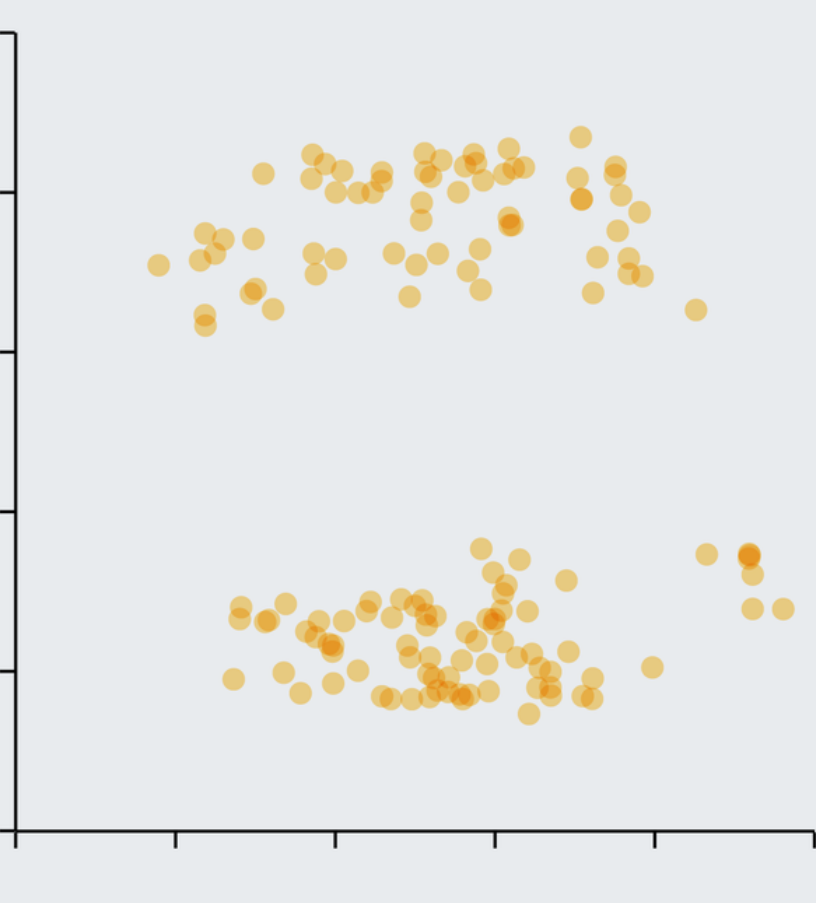
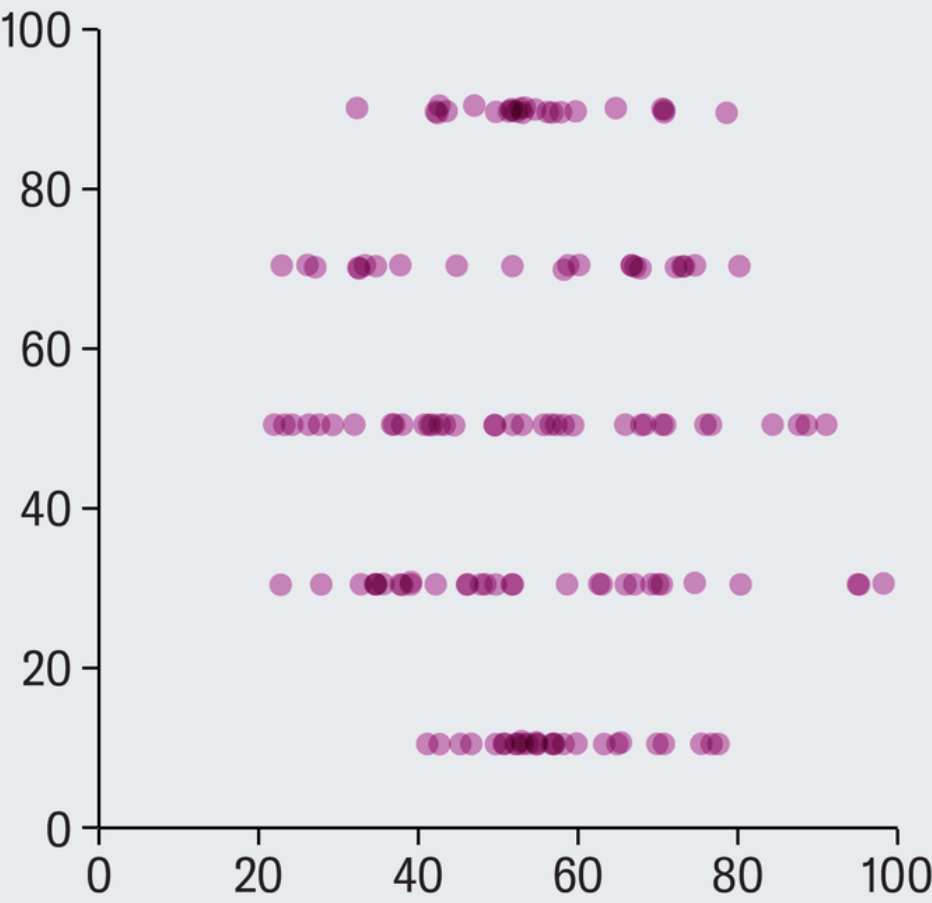
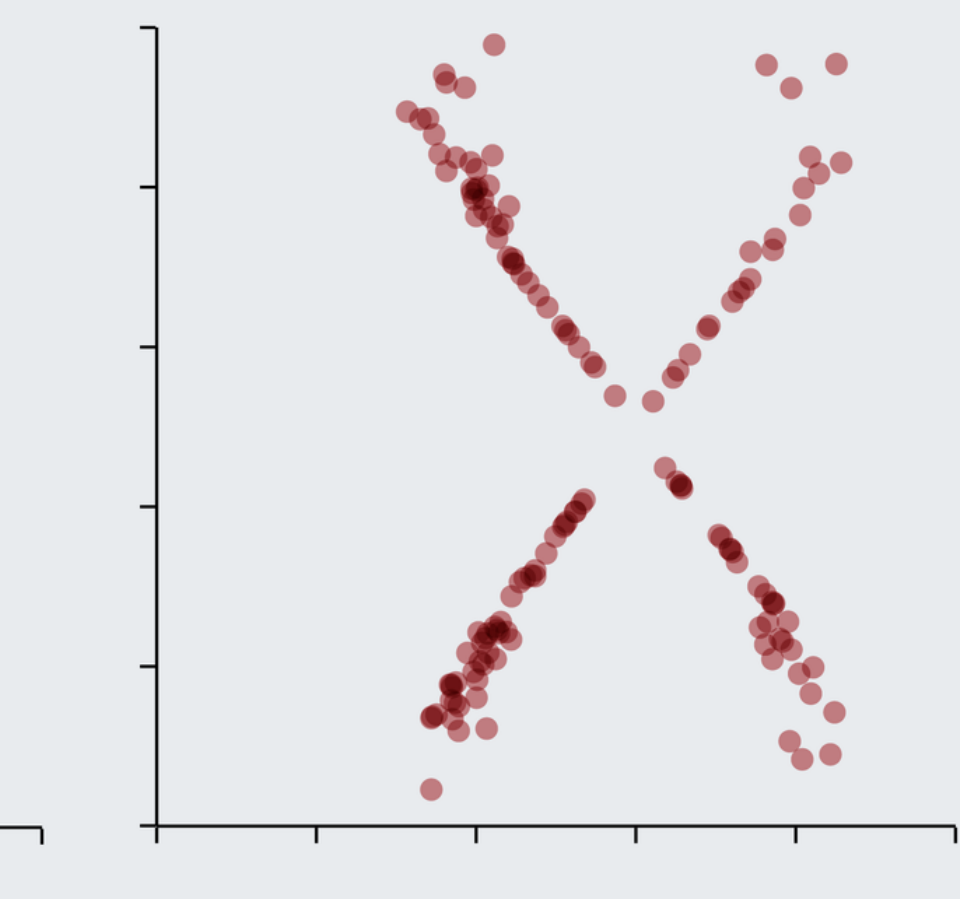
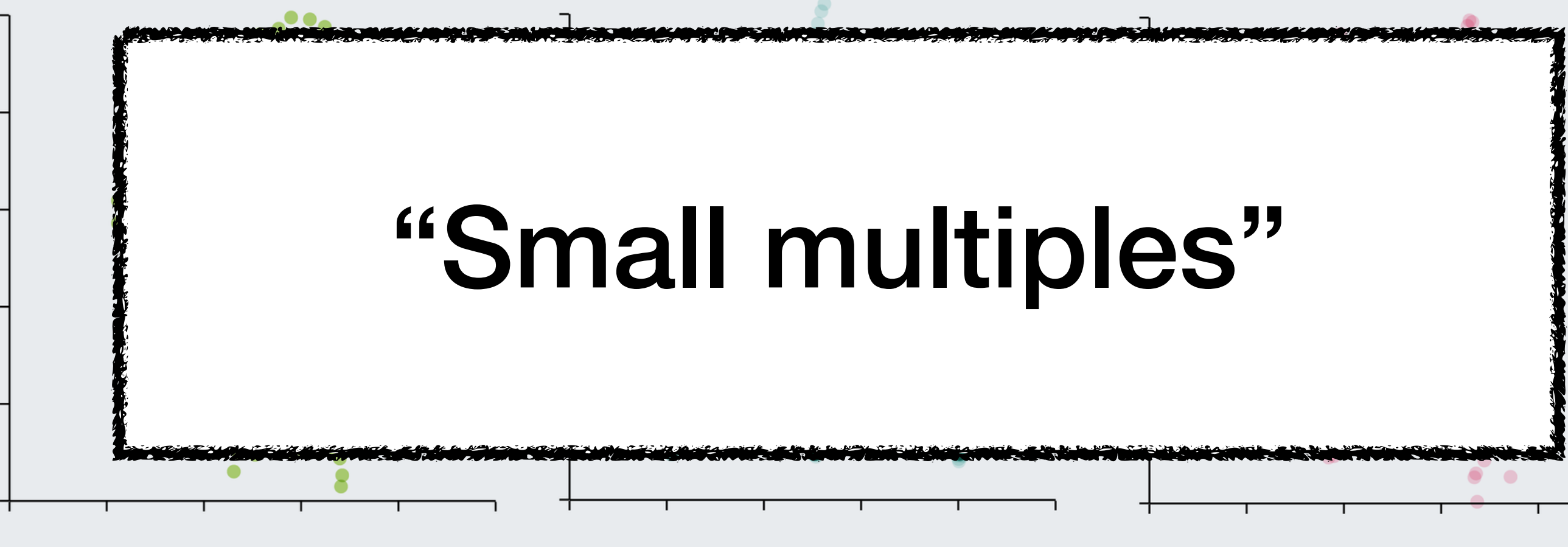
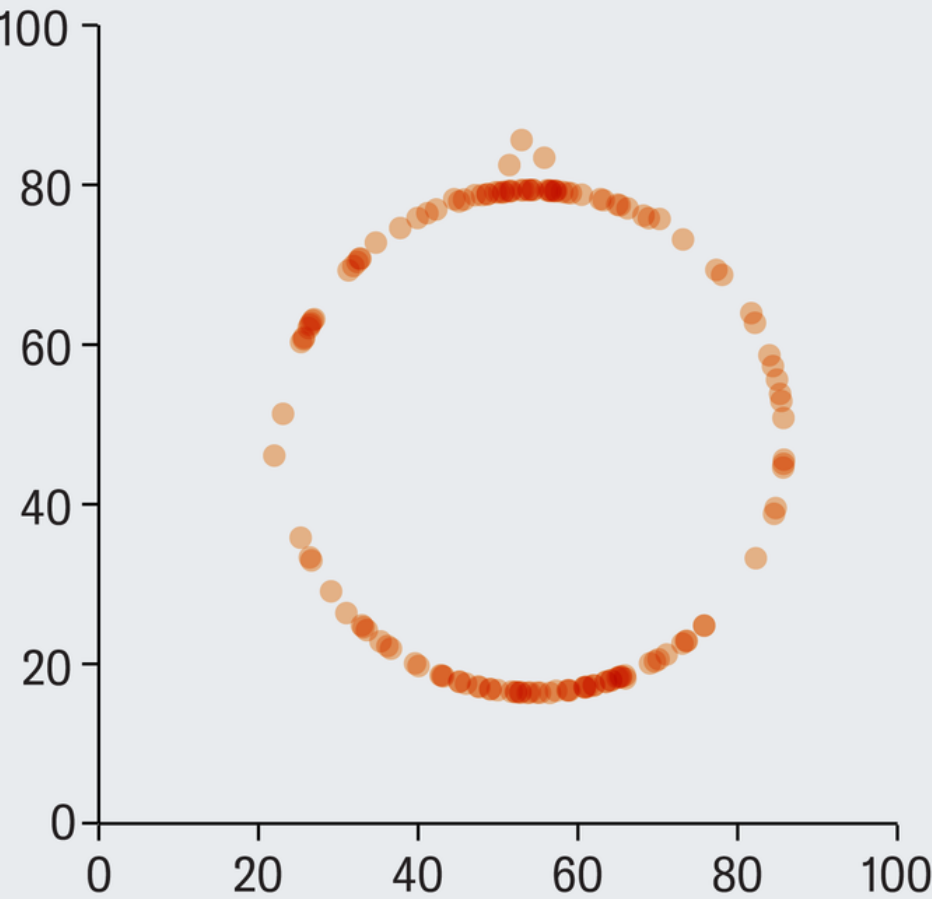
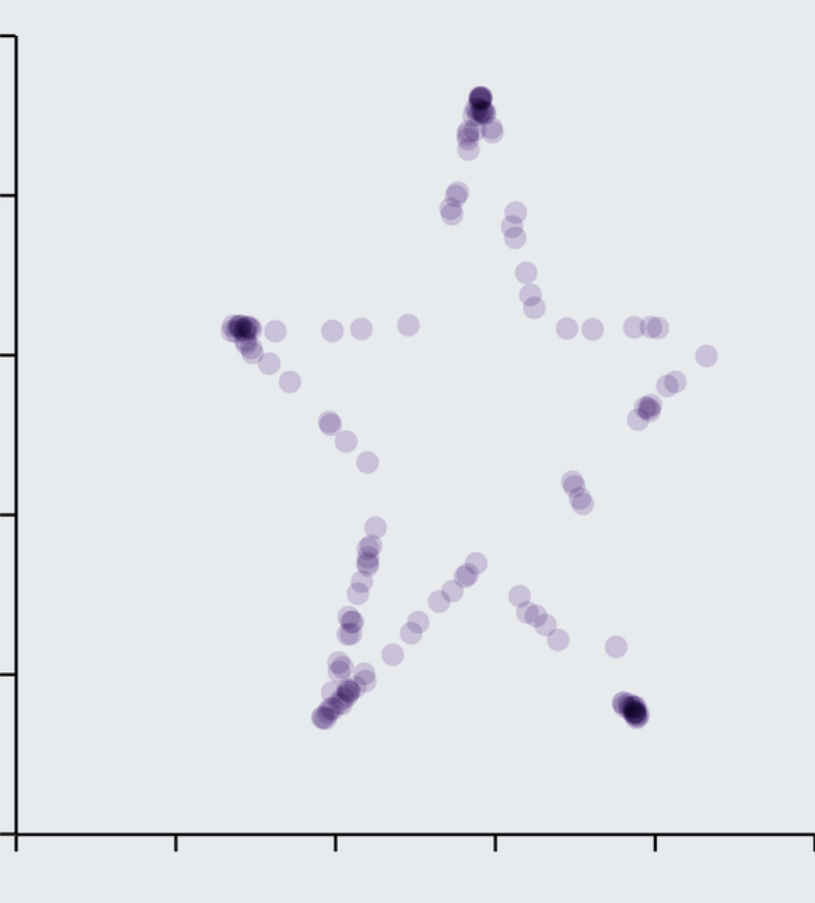
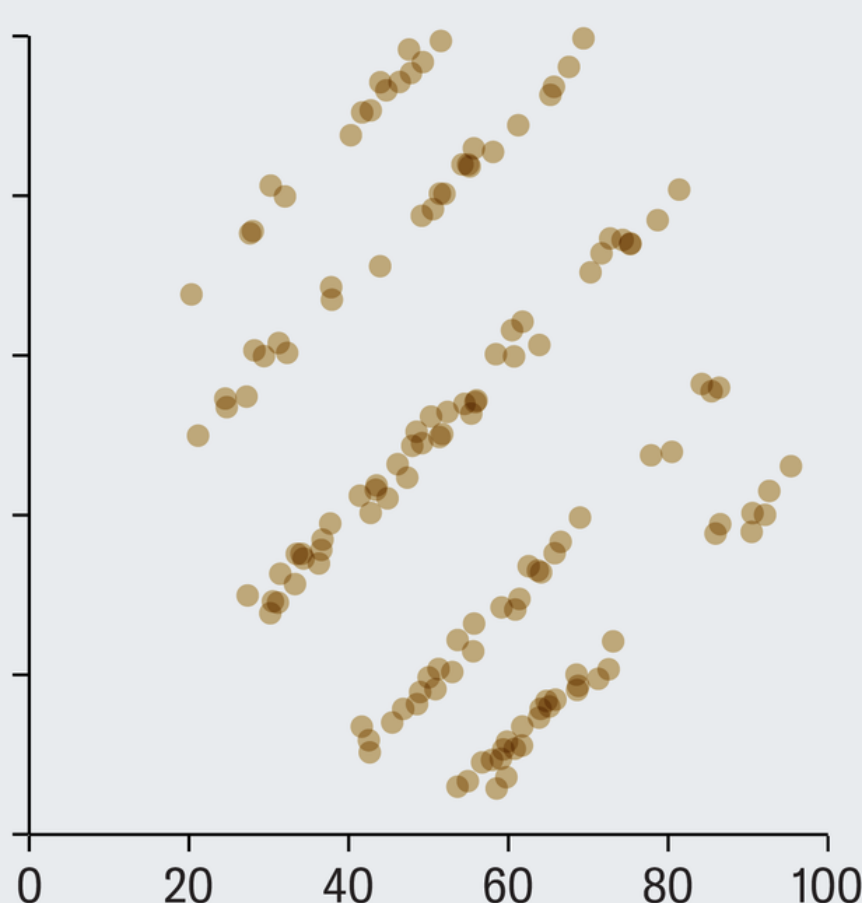
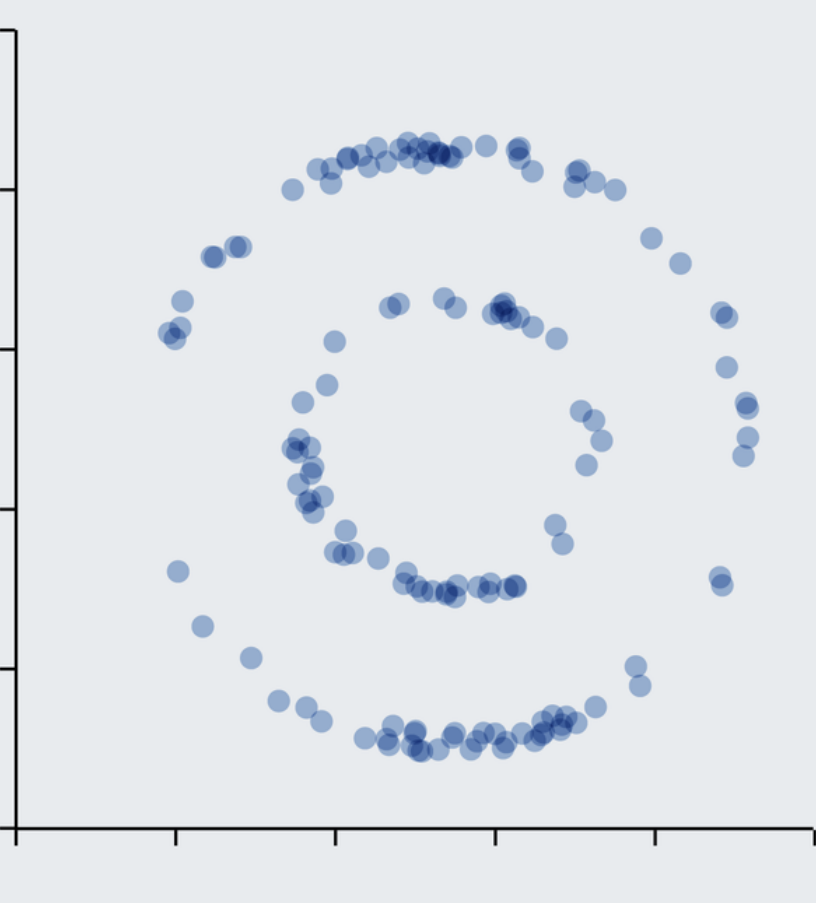
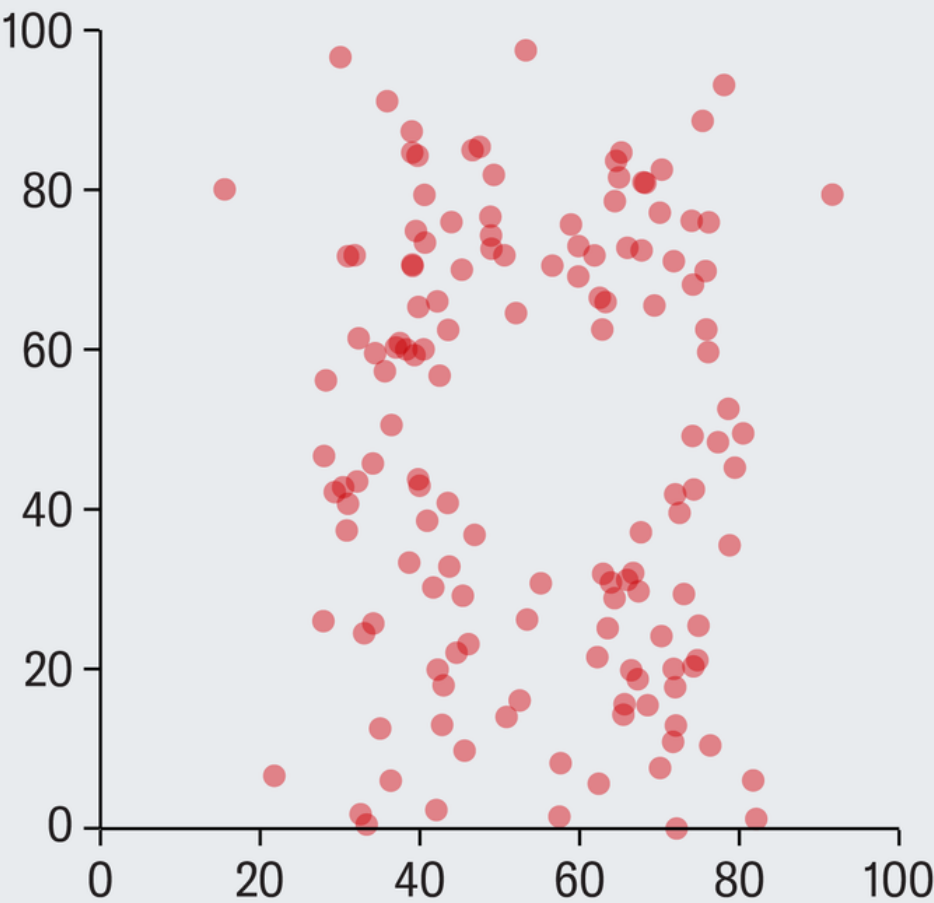
Mean, x-axis = 54.26

Mean, y-axis = 47.83

Standard deviation, x-axis = 16.76

Standard deviation, y-axis = 26.93

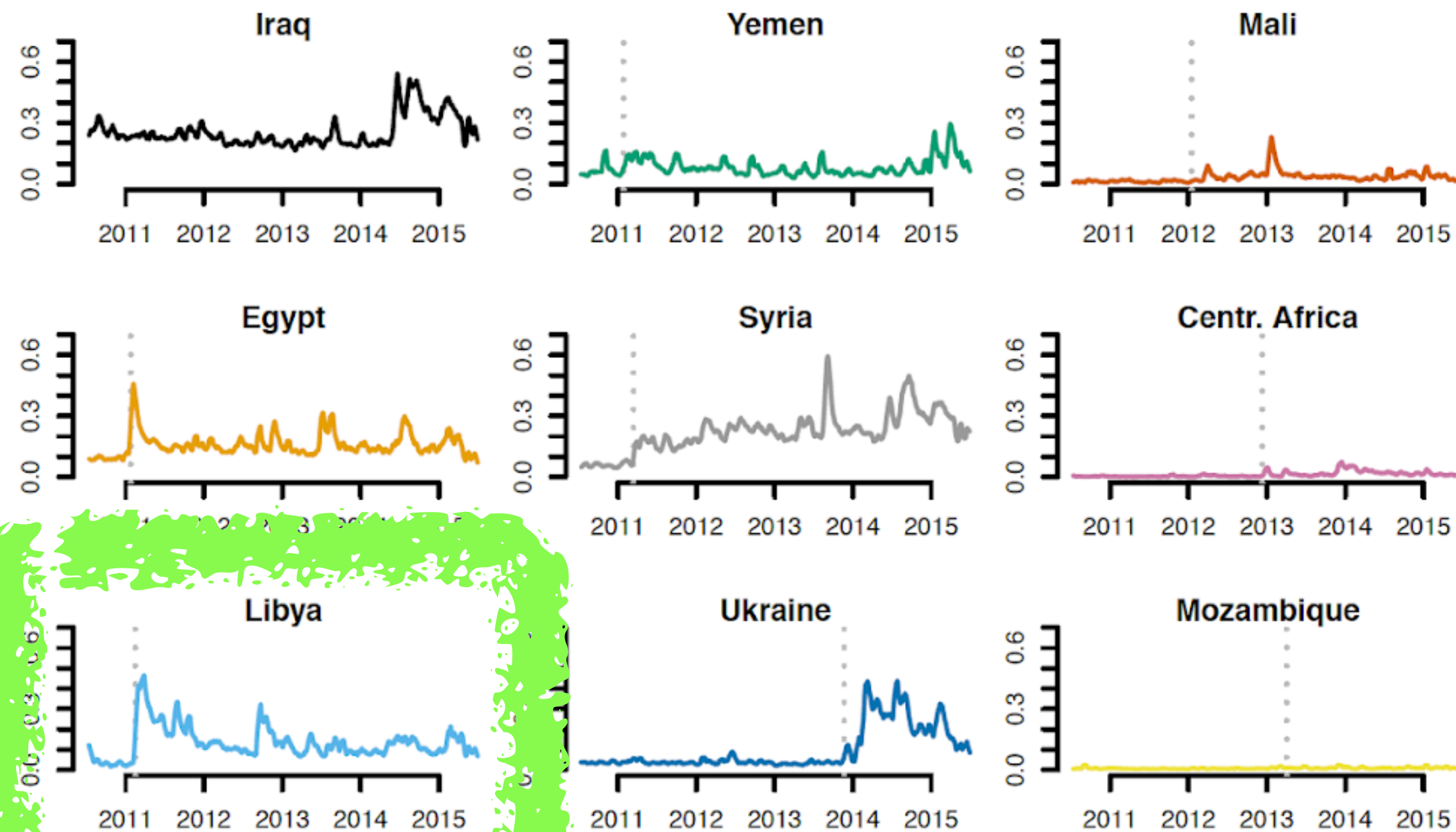
Correlation = -0.06



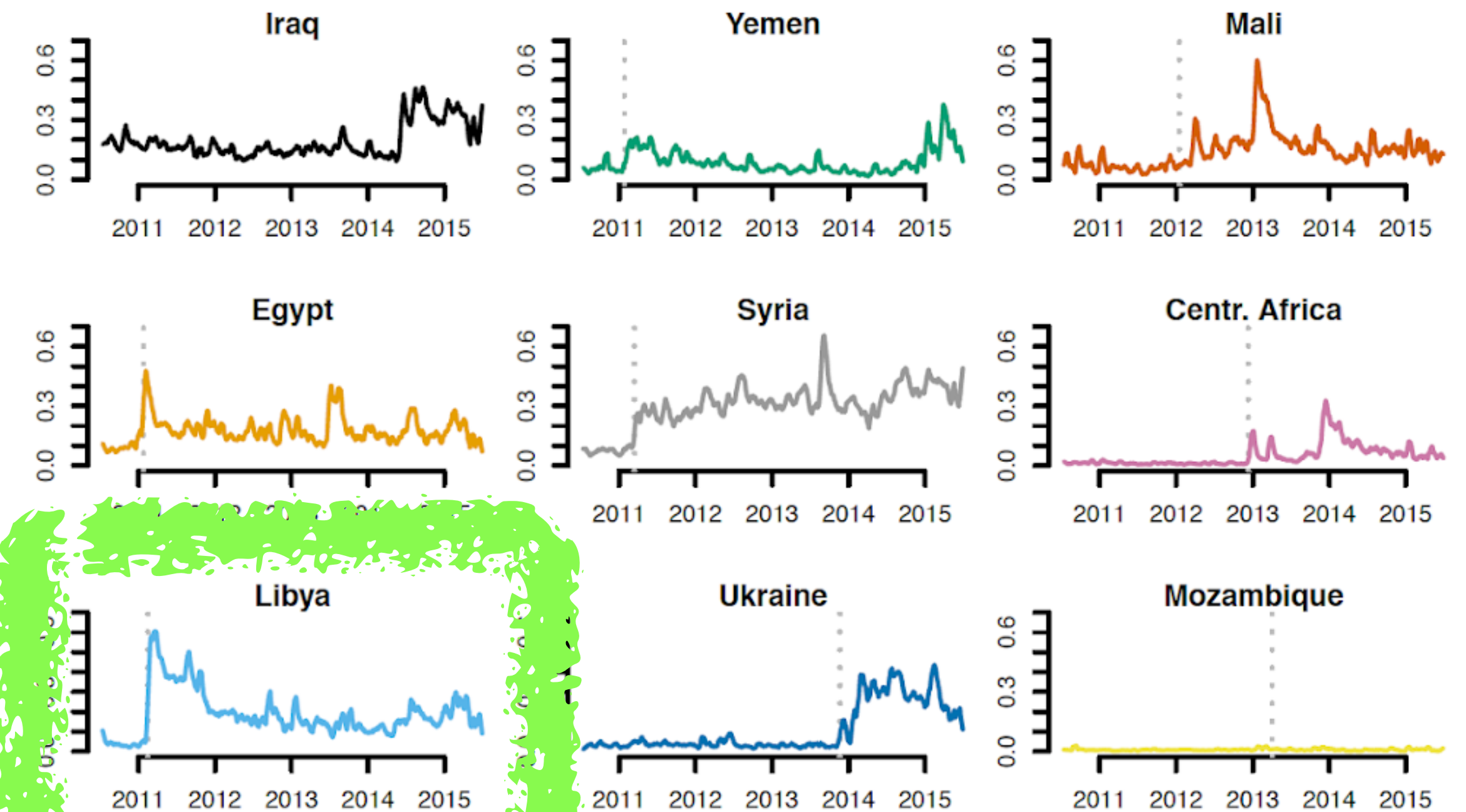
(Link)

Consider using small multiples

Media Attention



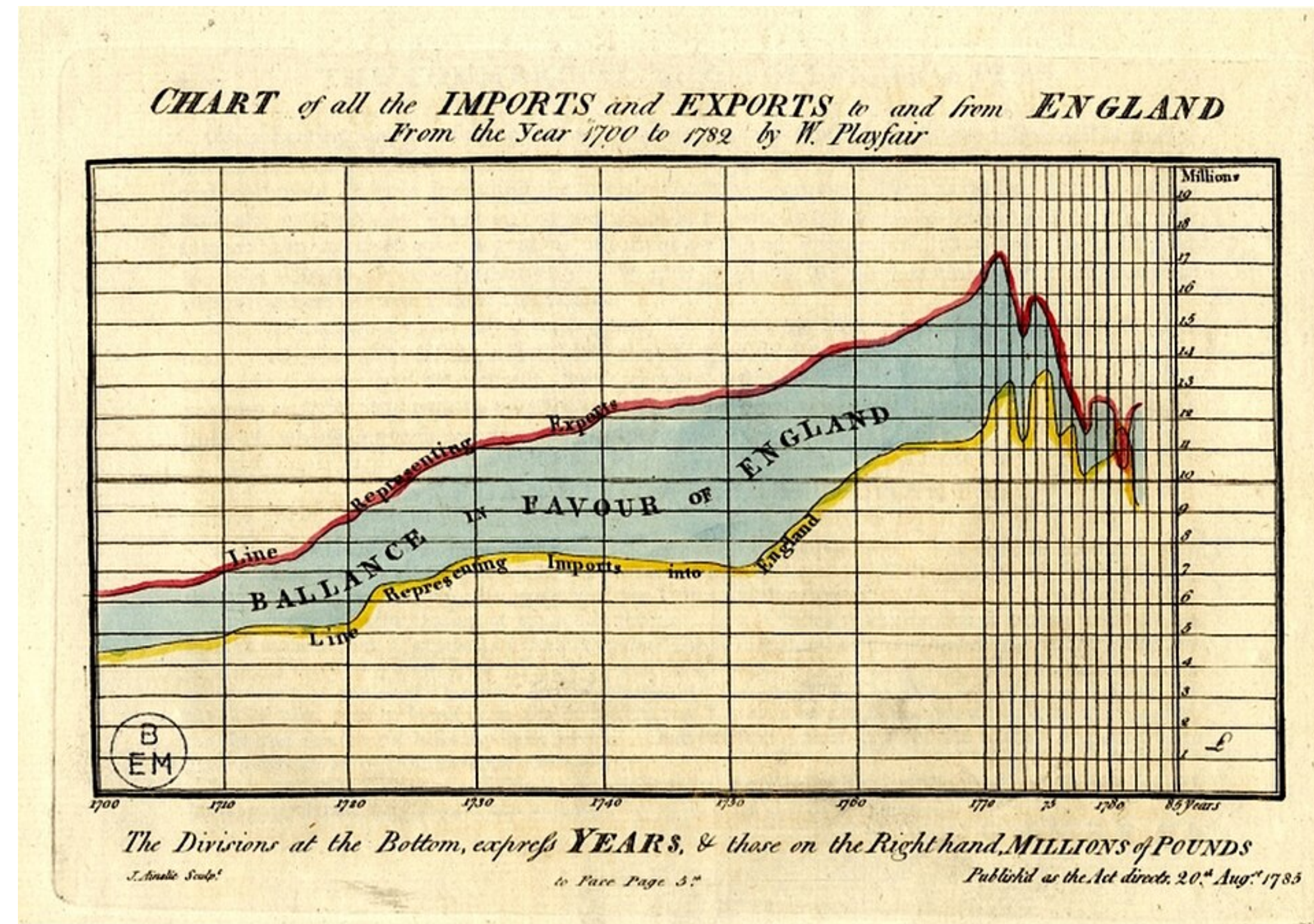
(a) English



(b) French

Be consistent with color!

Line plots

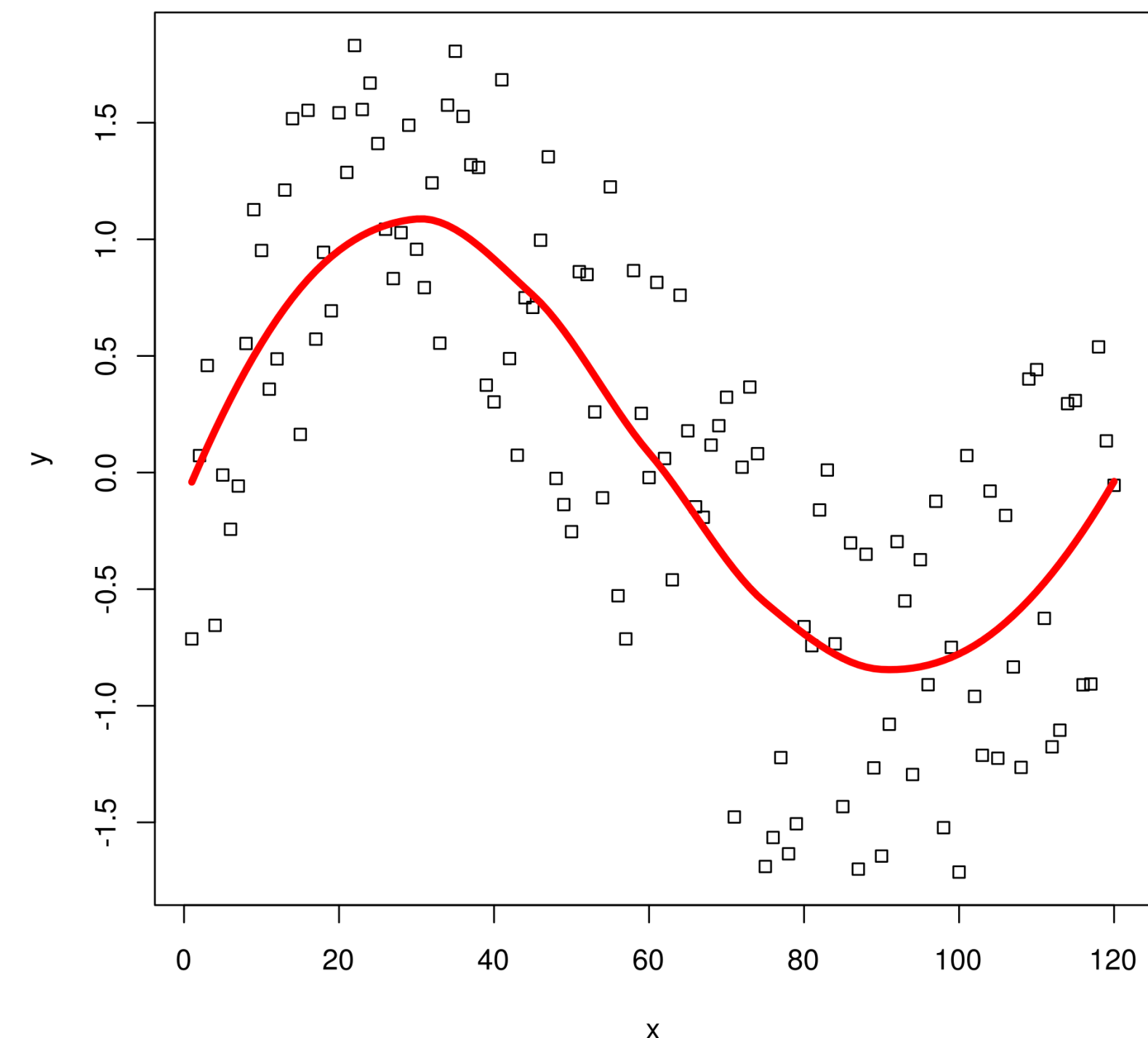


Moving Averages

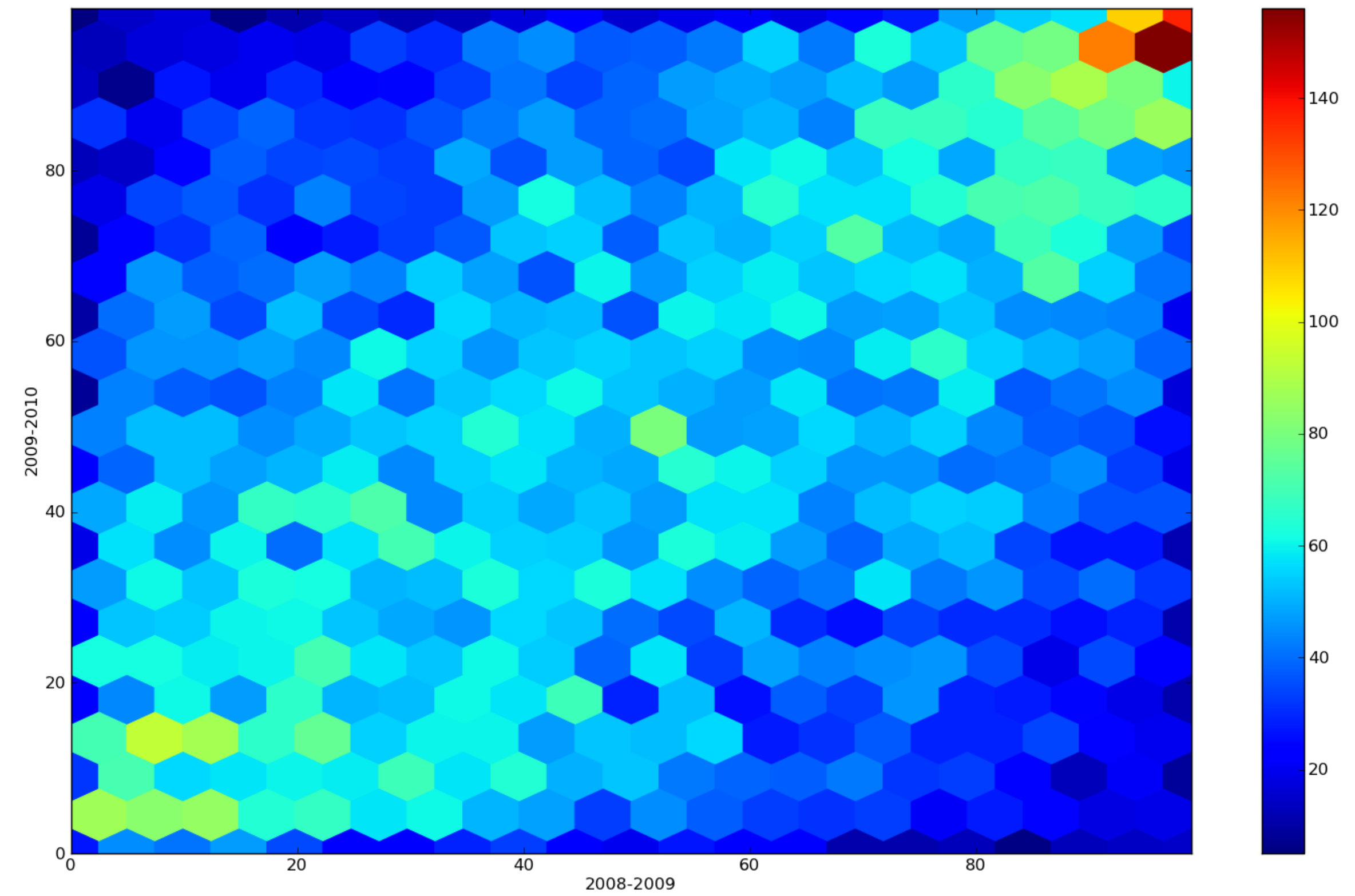
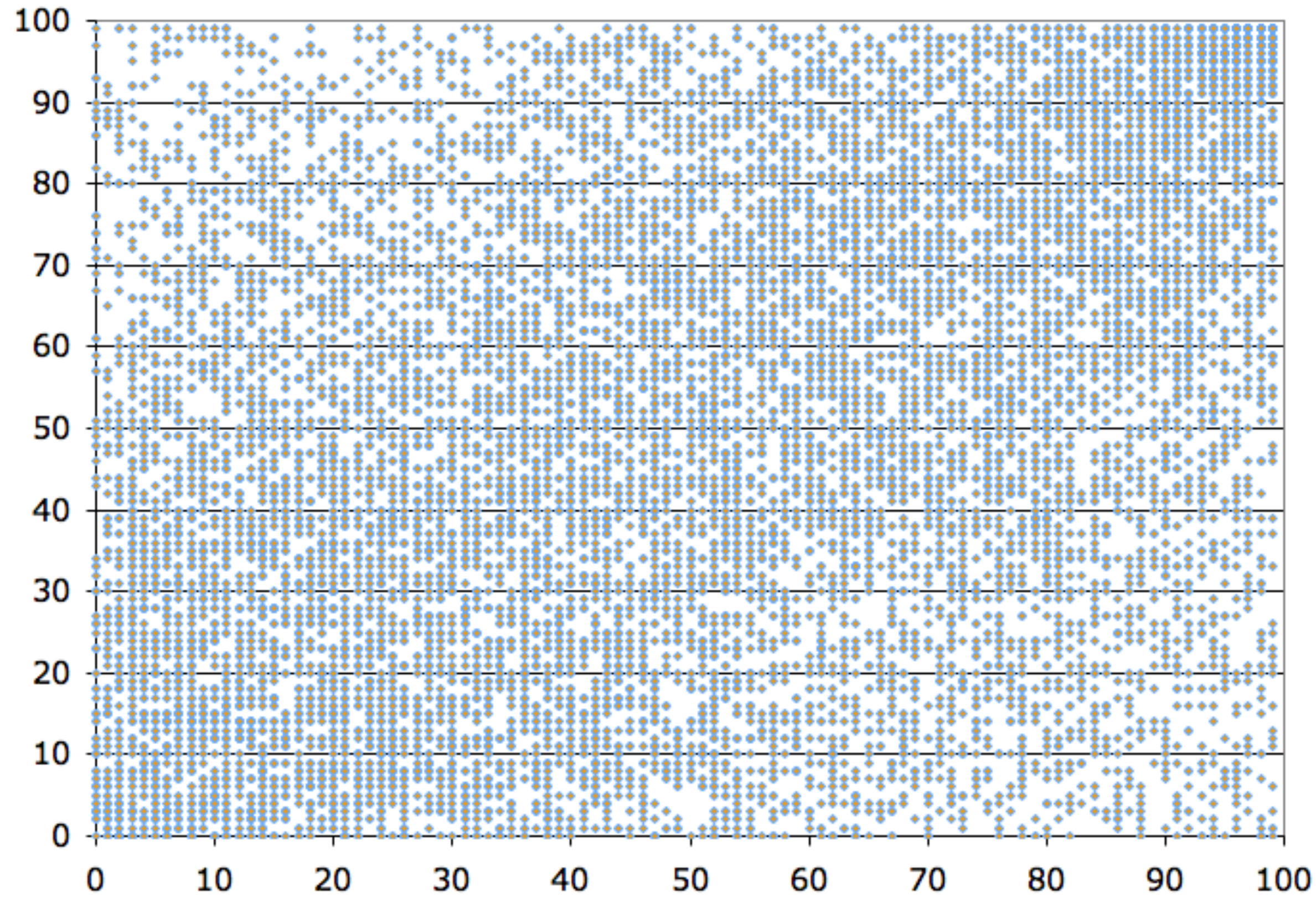


LOWESS

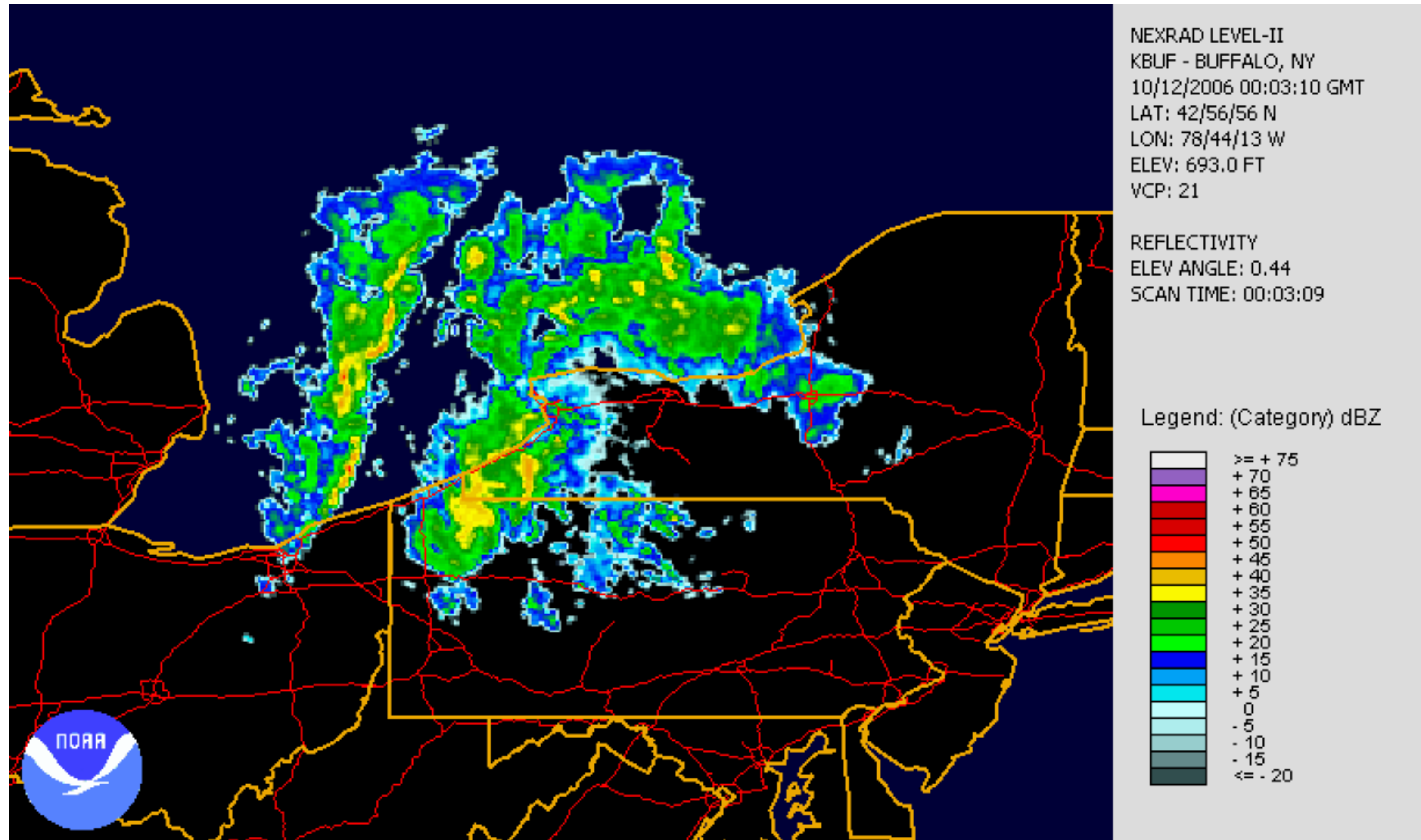
- Pick a specific point p on the x -axis
- Look at the nearby data points within a window.
- Calculate their average value $\hat{y}_p$, weighted by their distance to p .
- Plot a dot on $\langle p, \hat{y}_p \rangle$



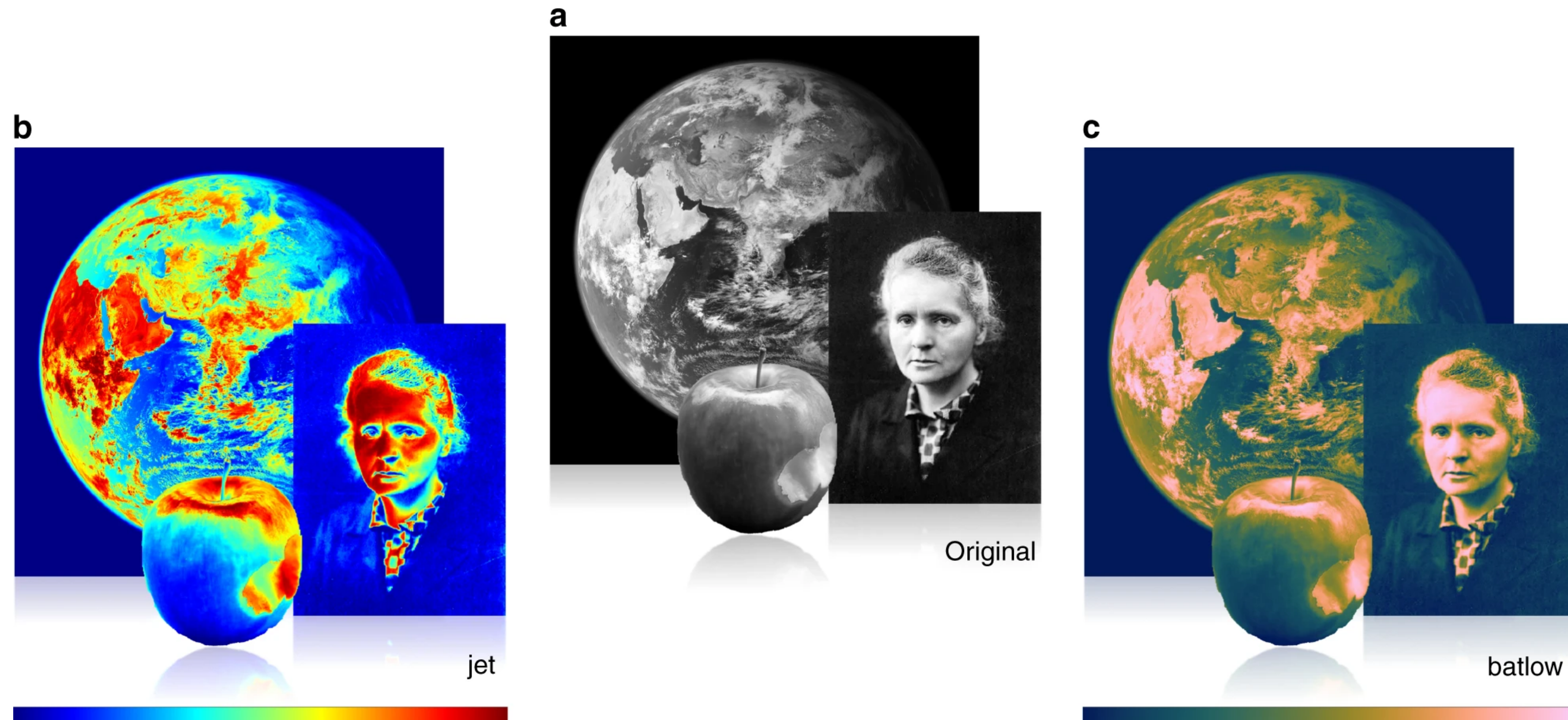
Heatmaps



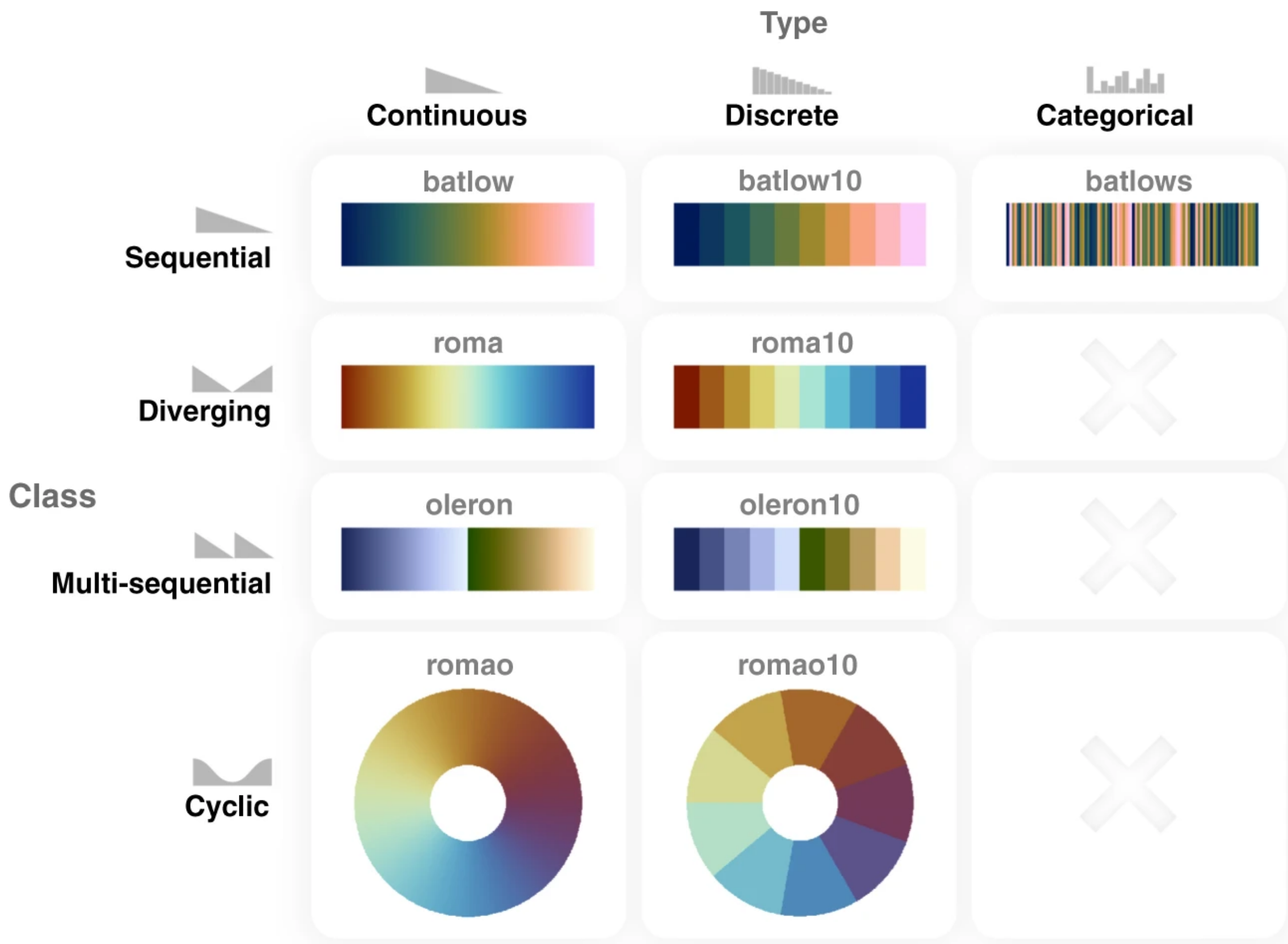
They are everywhere!



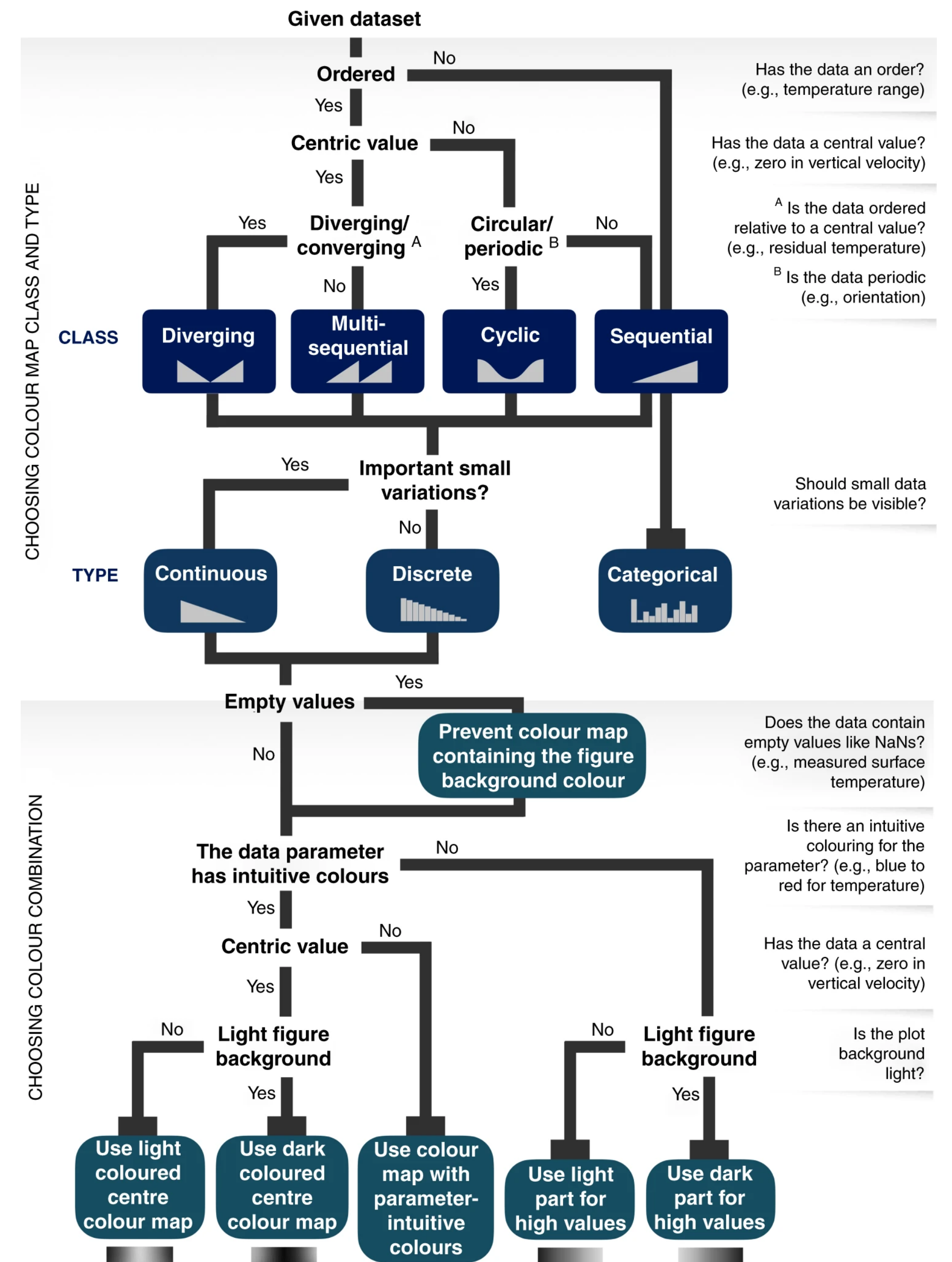
Color matters!



(Link)



[\(Link\)](#)

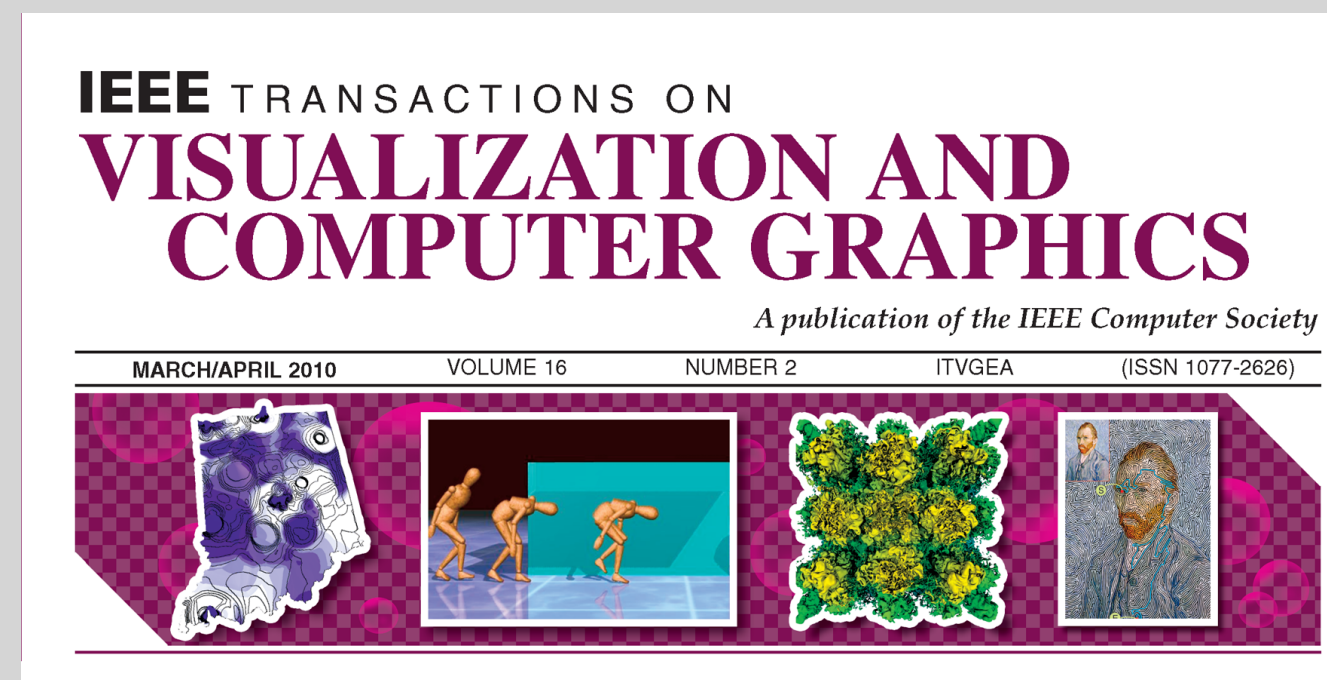


Why visualize data?

Analysis

Communication

Decision making



There is a whole research community dedicated to DataVis, can be a course on its own!

Decision making



HomeMembersJoin UsPublicationsTeachingPress



Spring 2026 · Fall 2025 · Spring 2025

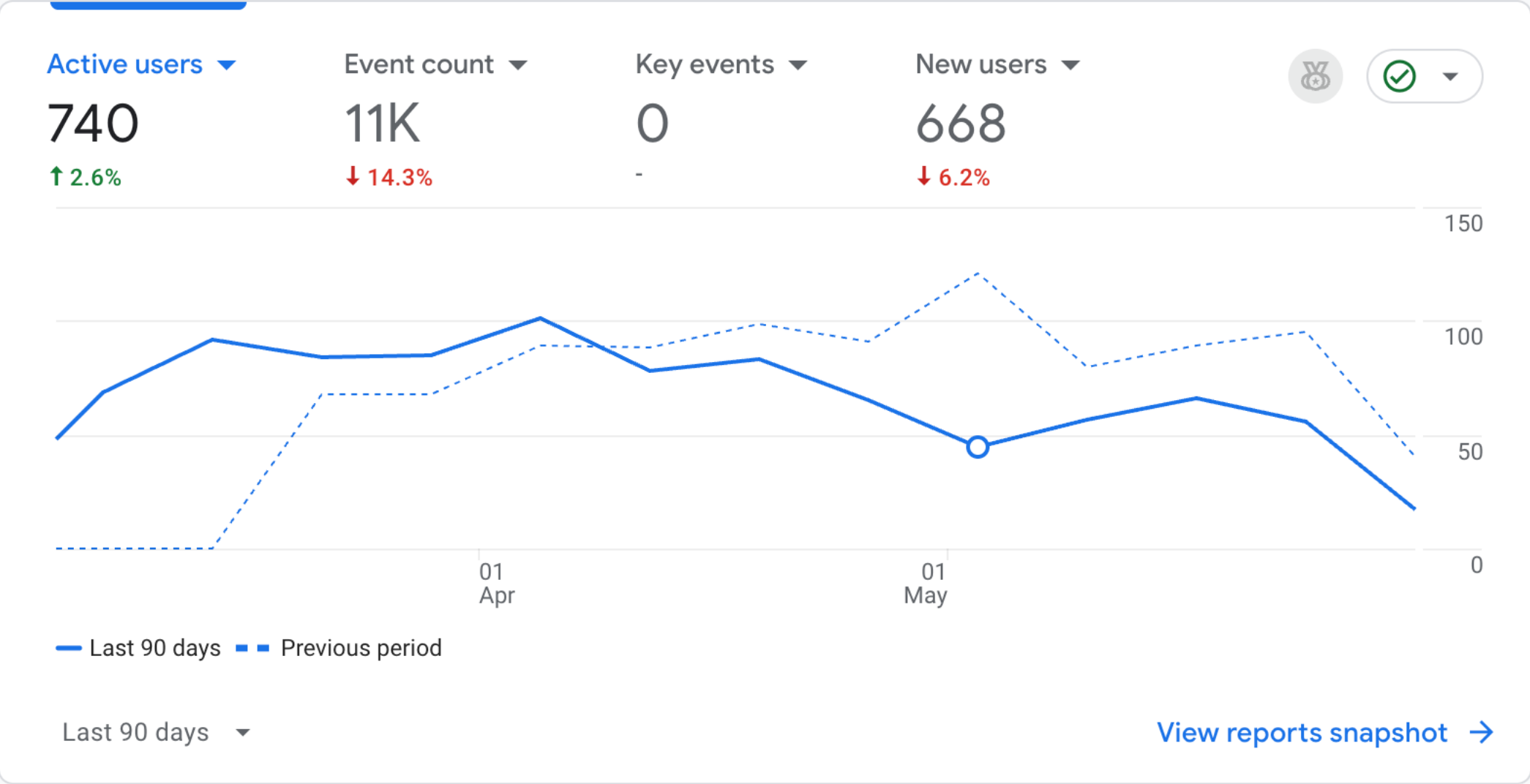
The Humans and Machines Lab is a research group within the [Princeton CS department](#) led by [Manoel Horta Ribeiro](#). We are affiliated with the [Center for Information Technology Policy](#) and with [Princeton HCI](#). We do research at the intersection of Computational Social Science, Human-Computer Interaction, and Artificial Intelligence.

Our research seeks to understand, design, and improve sociotechnical systems in the age of generative AI.

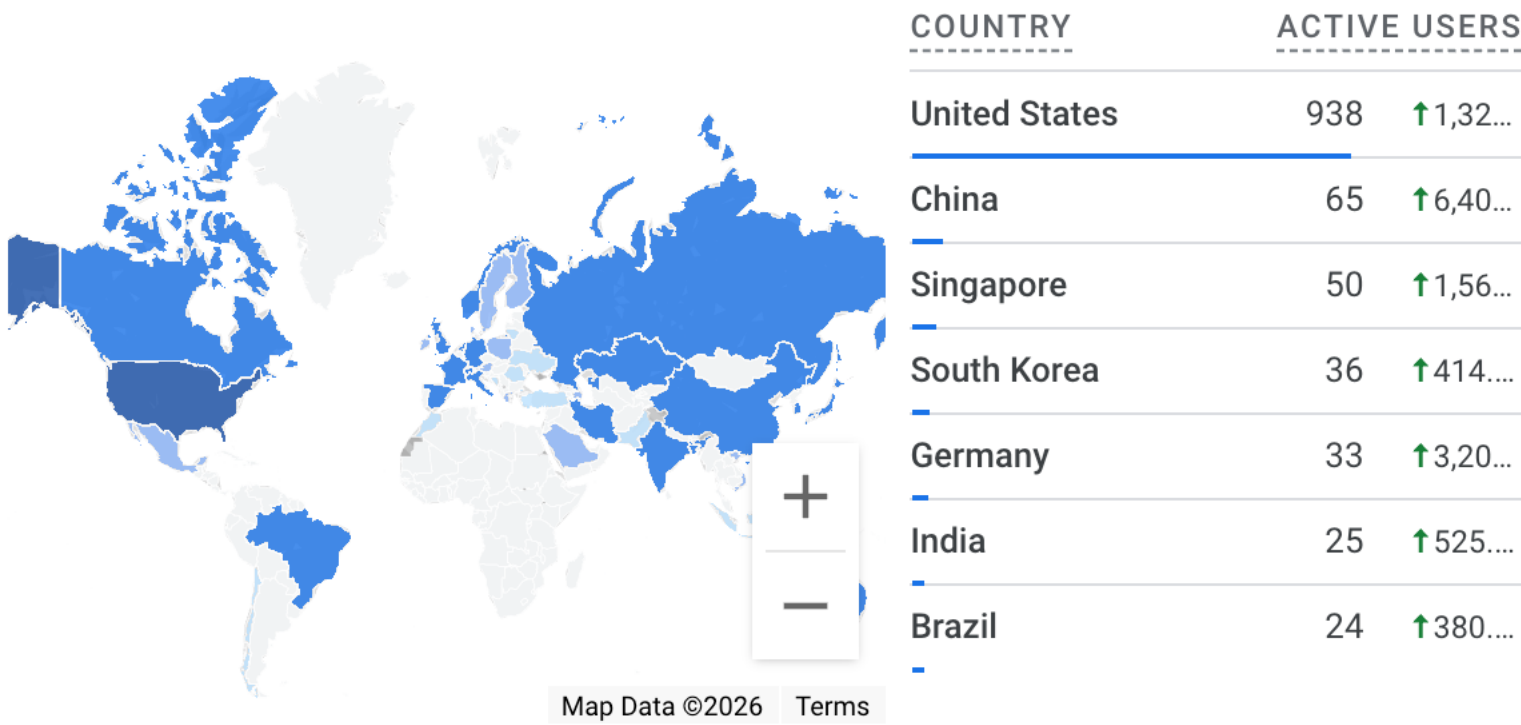
Bulletin Board

PreviousNext

- 2026-04-04 📄 Paper accepted at Findings of the ACL 2026: "The Illusion of Insight in Reasoning Models" [\(link\)](#)
- 2026-03-15 📄 Paper accepted at ICWSM 2026: "Short-form video platforms drive mobile overuse" [\(link\)](#)
- 2026-01-17 📄 Paper accepted at CHI 2026: "Bonsai: Intentional and Personalized Social Media Feeds" [\(link\)](#)
- 2025-12-04 📄 Paper accepted at USENIX 2026: "Chameleon Channels: Measuring YouTube Accounts Repurposed for Deception and Profit" [\(link\)](#)
- 2025-11-10 🏆 Manoel was commended for his teaching at Princeton! [\(link\)](#)
- 2025-10-18 🎉 The very first cohort of Postdocs, PhD students, and MSc students joined the lab!



Active users ▾ by Country ID ▾



This year (Jan - Today) ▾

View countries →

Views by Page title and scree...

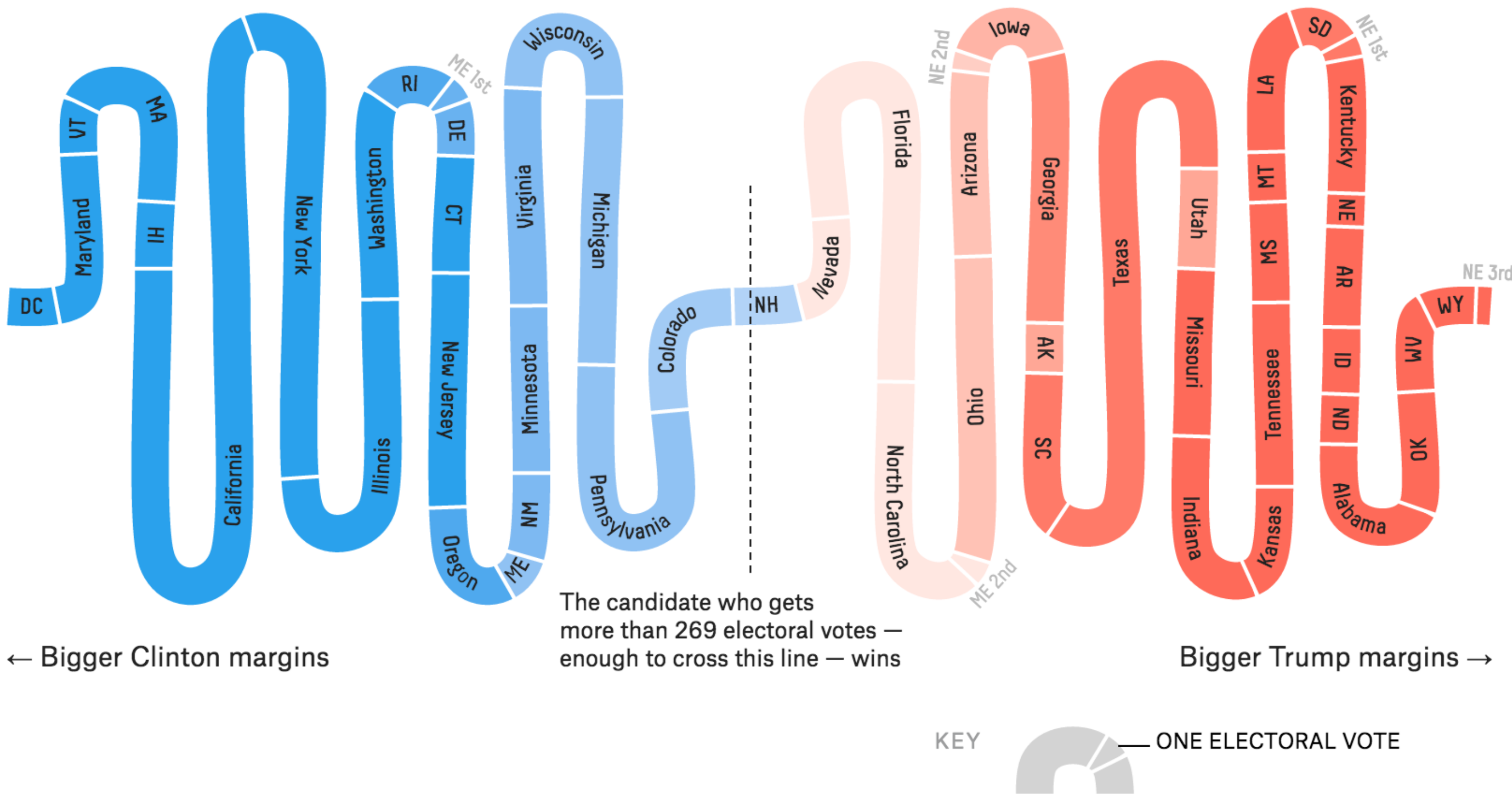
PAGE TITLE AND S...	VIEWS
Humans and Machin...	168 -
Publications	121 -
People	119 -
Join Us	55 -
Teaching	24 -
Spring 2025 – COS 5...	13 -
Press	11 -

Last calendar year View pages and s... →

FiveThirtyEight

The winding path to 270 electoral votes

A candidate needs at least 270 electoral votes to clinch the White House. Here's where the race stands, with the states ordered by the projected margin between the candidates — Clinton's strongest states are farthest left, Trump's farthest right — and sized by the number of electoral votes they will award.





Give us some feedback!



[https://tinyurl.com/
reasoningwithdata](https://tinyurl.com/reasoningwithdata)