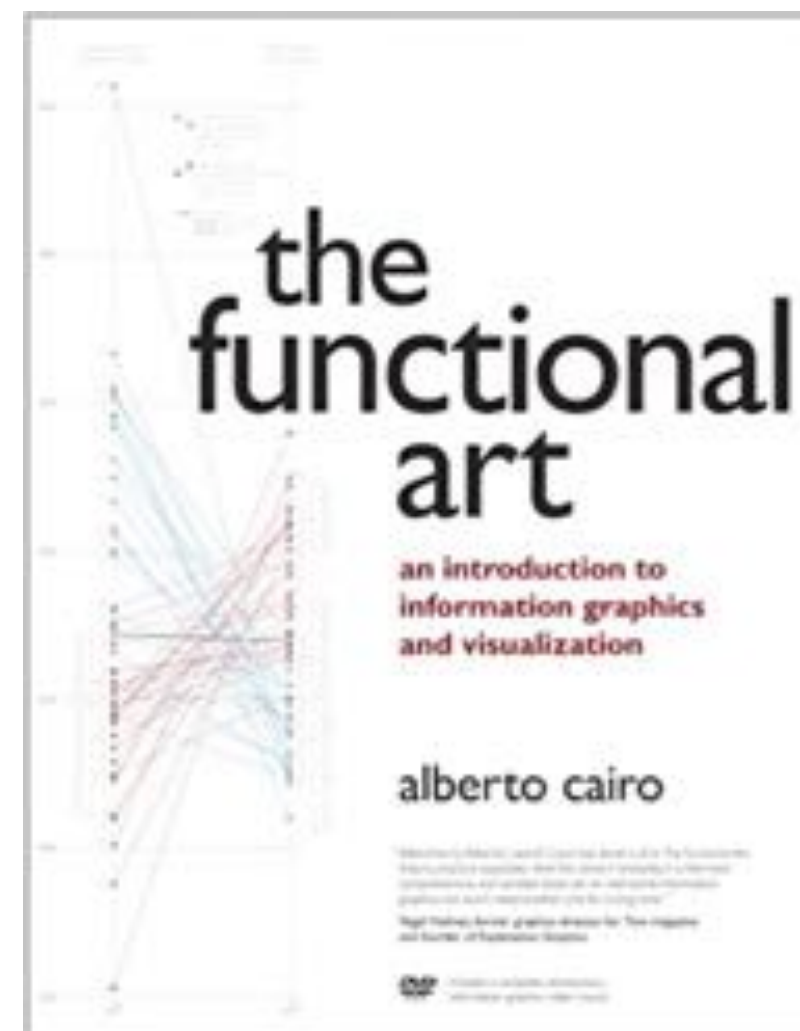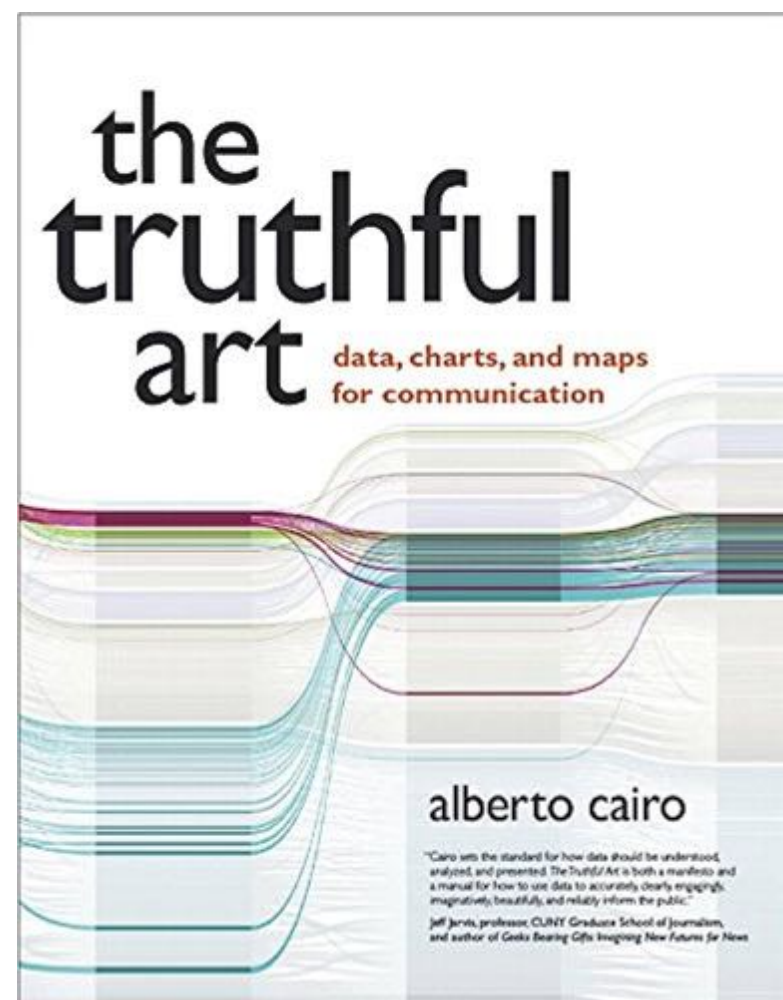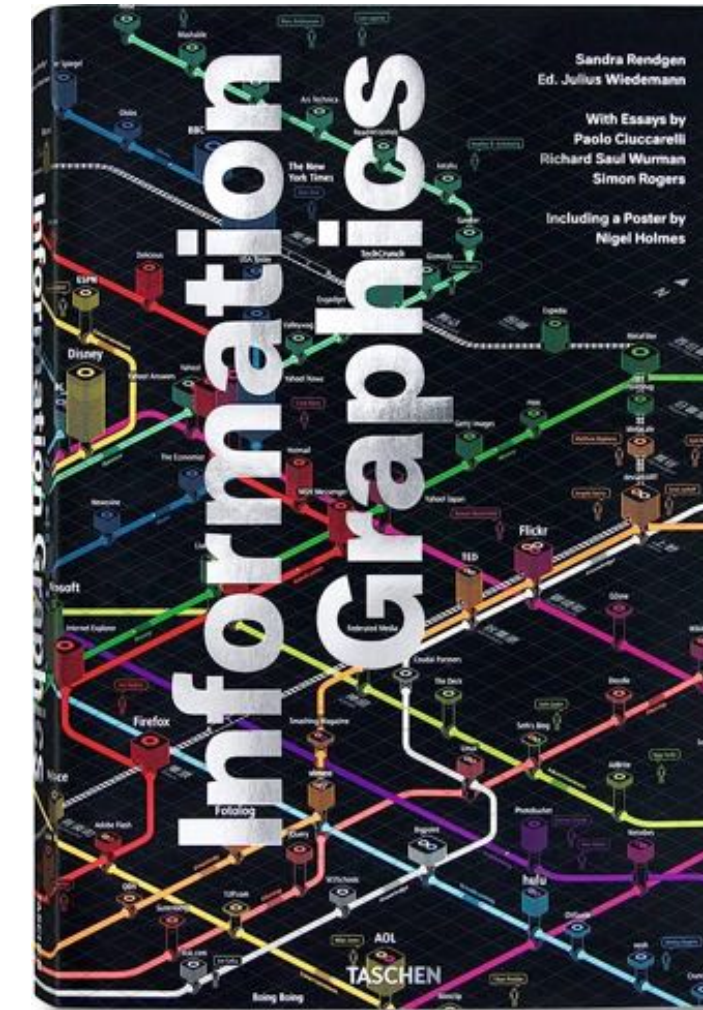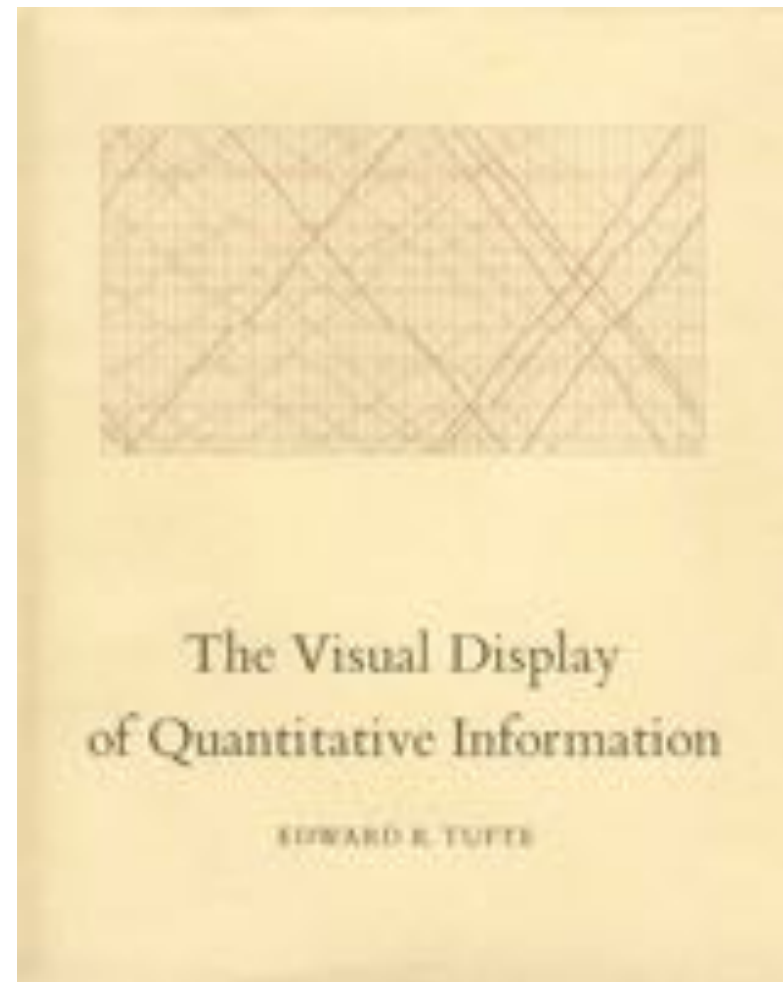# introduction to data visualization

in a scientific workflow

Bryan Scott - Data Science Fellowship Program Session 17 - Texas A&M - February 27 2023

Background

classic (and not so classic) resources for design and data visualization
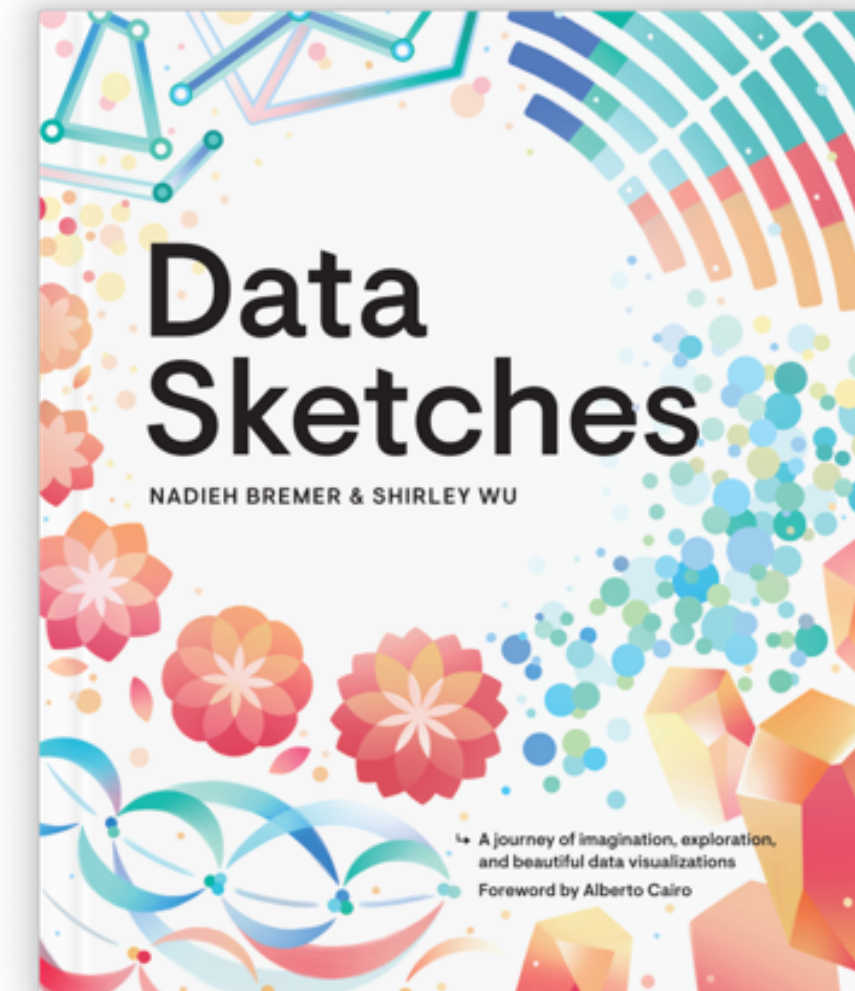
Tufte's rules for Graphical Excellence

Guiding principles of visualization

Show the data

Induce the viewer to think about substance

Avoid distracting data

Make large data coherent

Encourage comparison

Have a clear purpose

Integrate with statistical and verbal descriptions

Graphics Reveal Data
Anscombe's quartet

These four datasets have the same (first order) statistical description:

N=11

Mean of X = 9

Mean of Y = 7.5

Best fit line: Y = 3 + 0.5X

Standard Error of slope: 0.118

…

Graphics Reveal Data
Anscombe's quartet

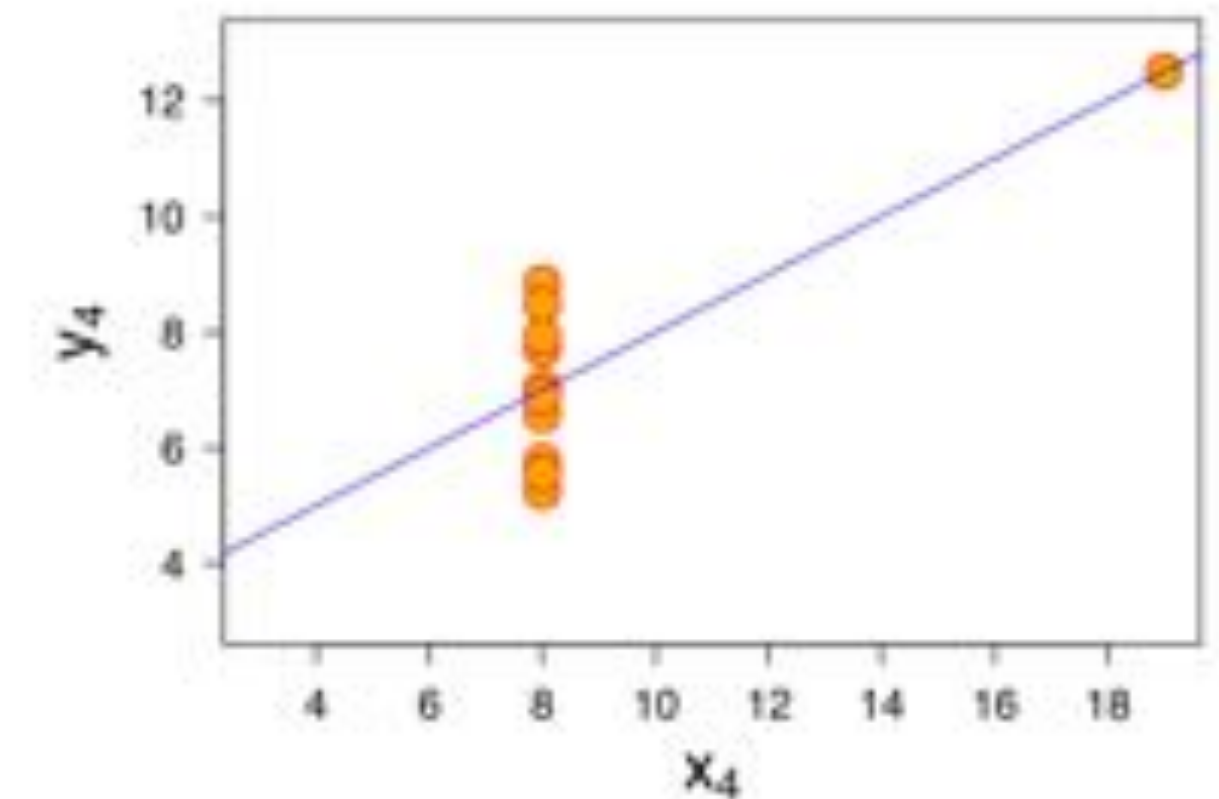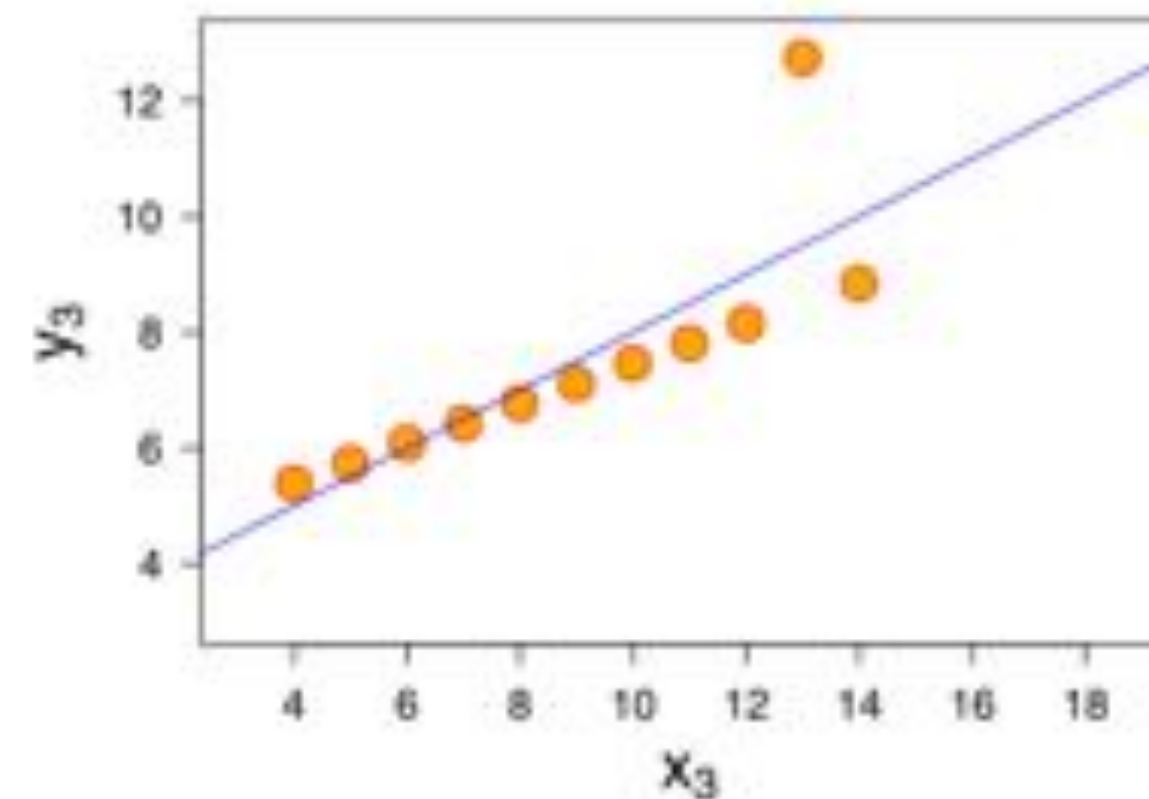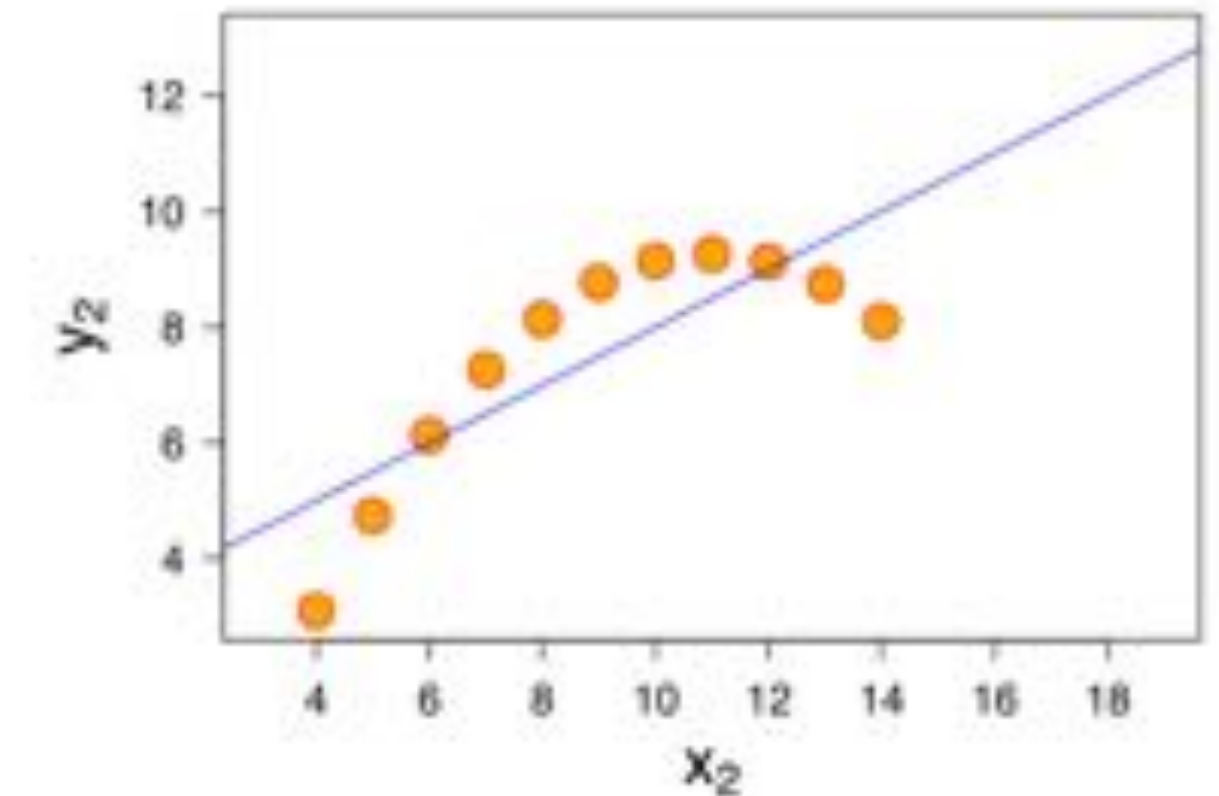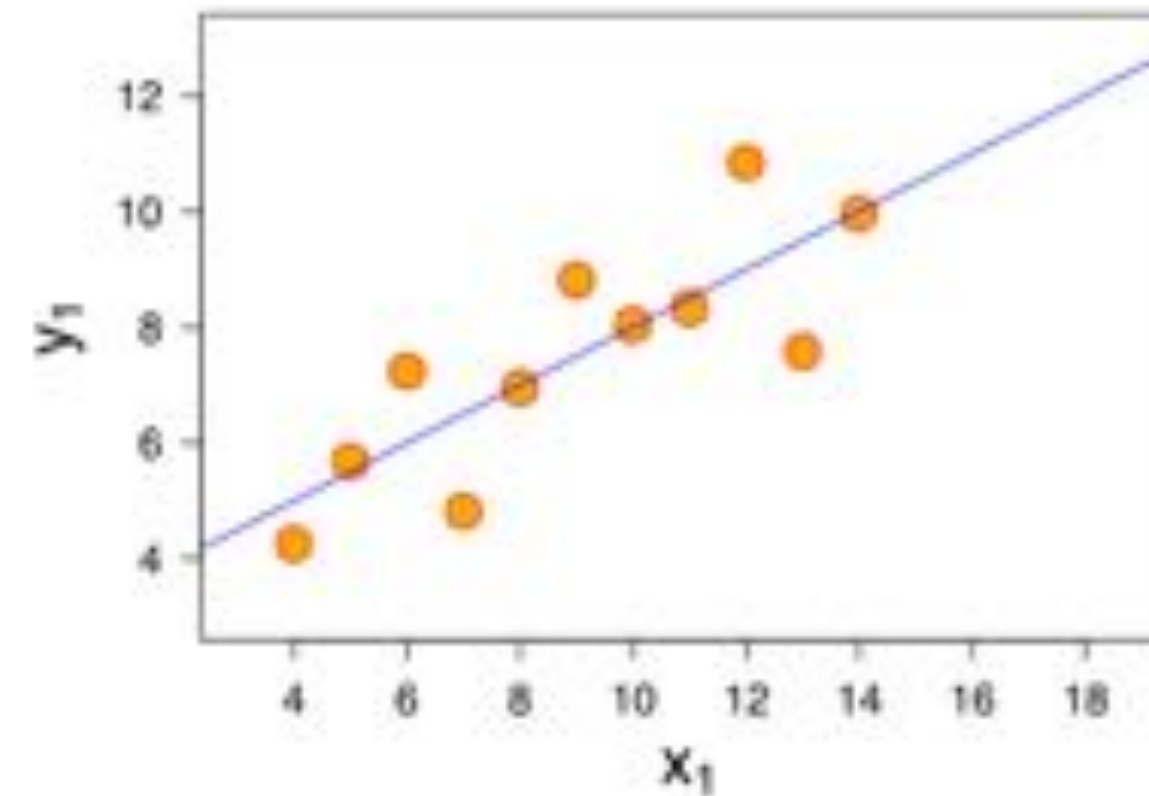These four datasets have the same (first order) statistical description:

N=11

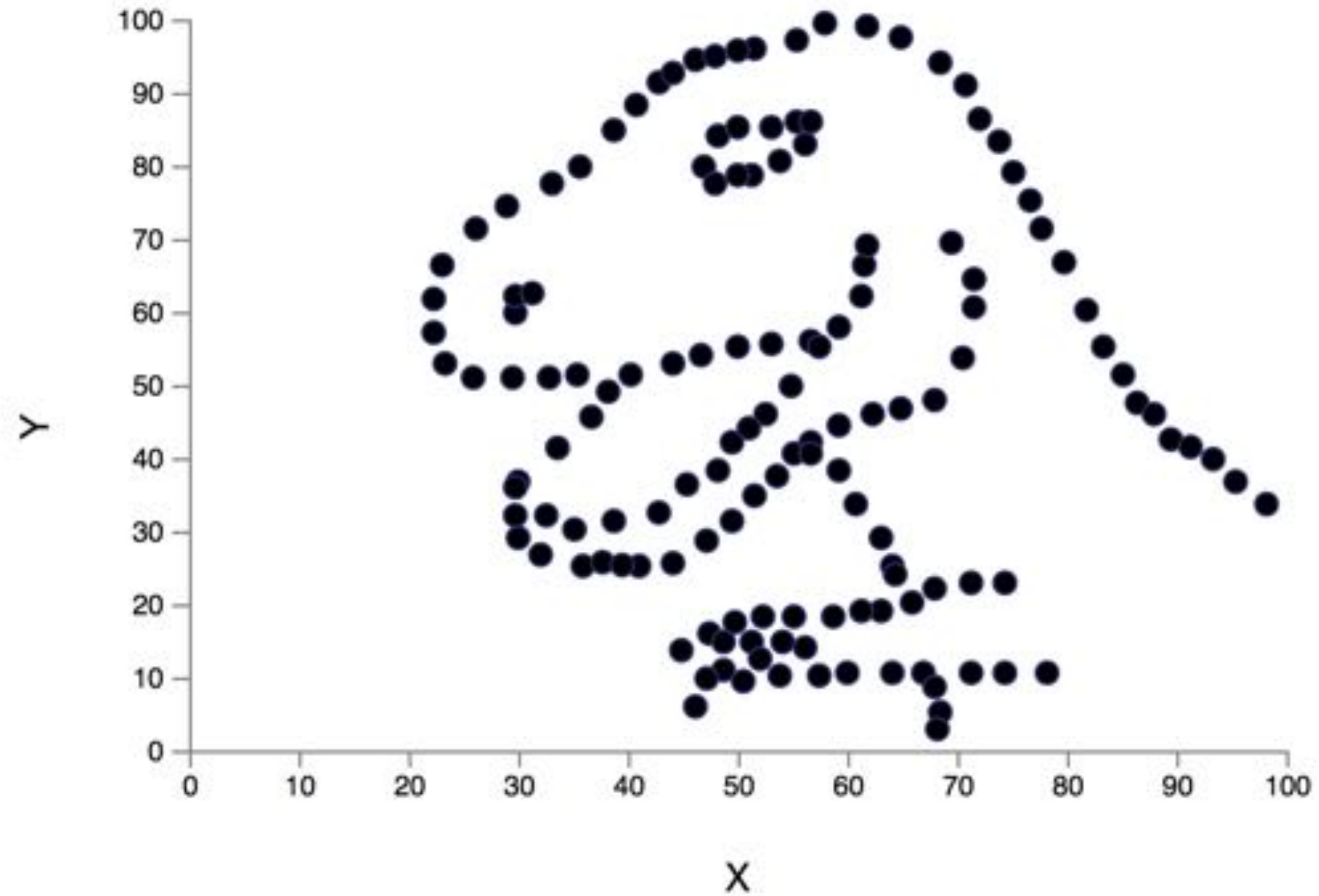Mean of X = 9

Mean of Y = 7.5

Best fit line: Y = 3 + 0.5X

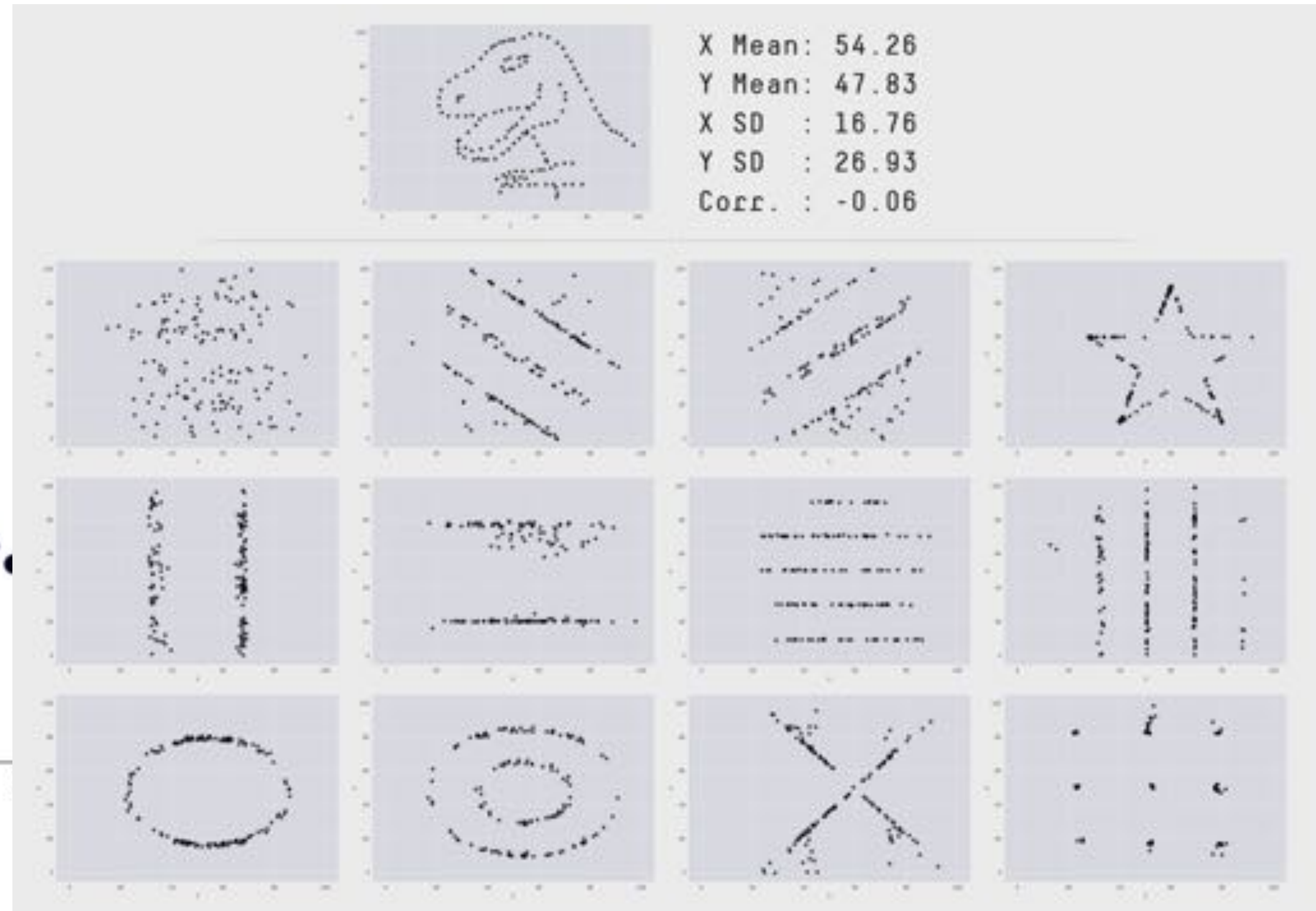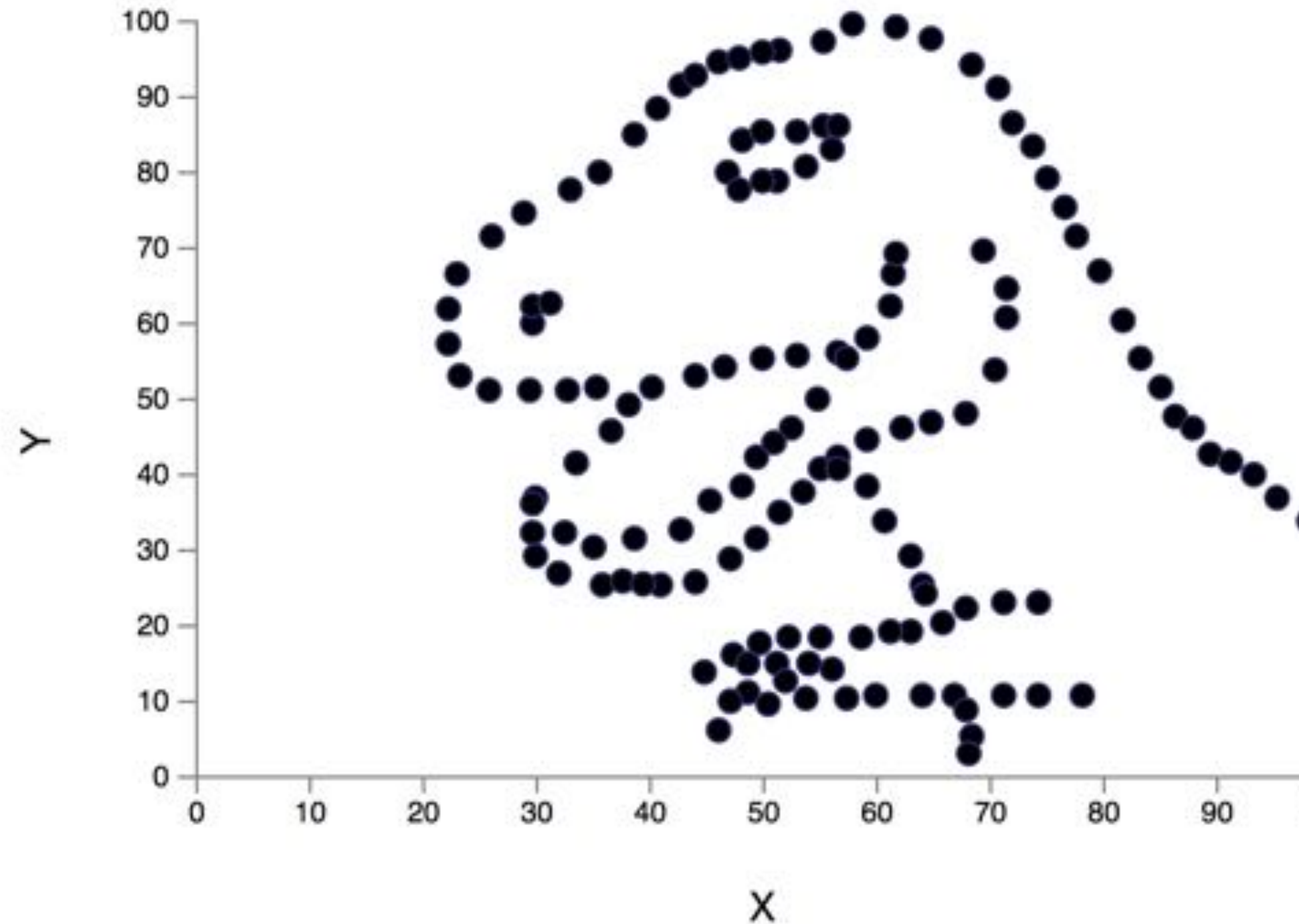Standard Error of slope: 0.118

…

ALWAYS visualize your data
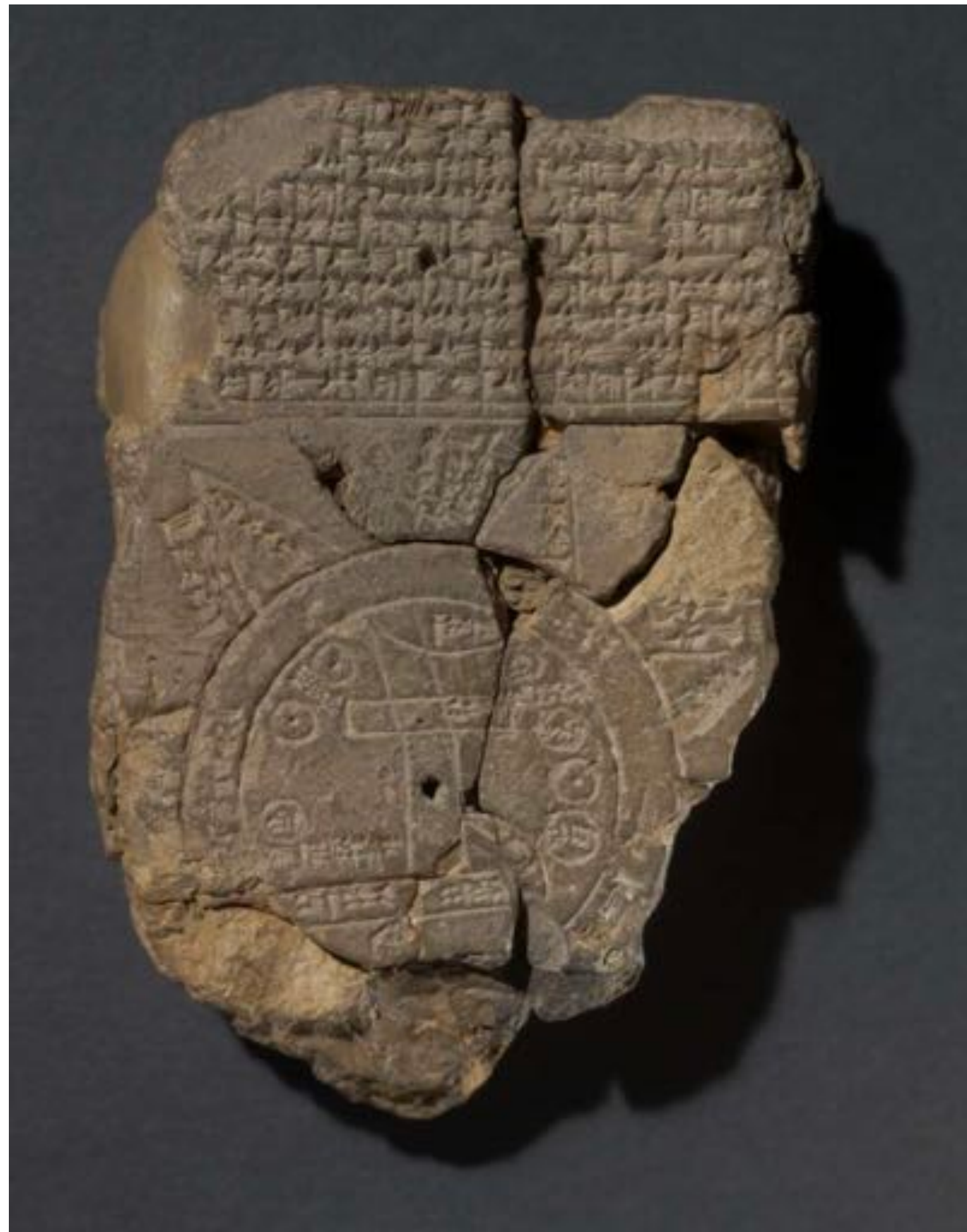**Summary Statistics can be misleading…**

ALWAYS visualize your data
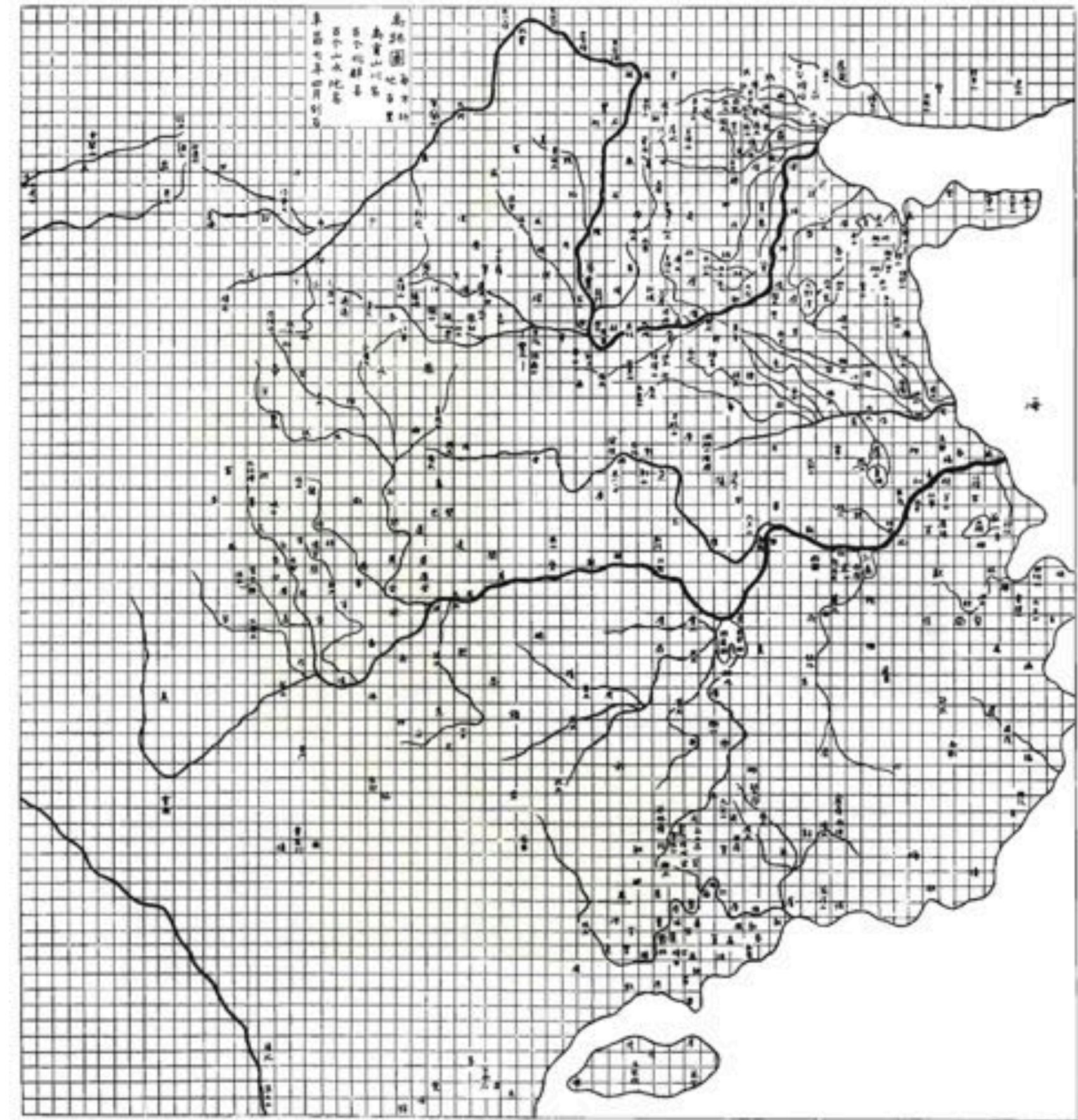## Summary Statistics can be misleading...

Click plot for reference and datasaurus dozen code.

A brief history of data visualization

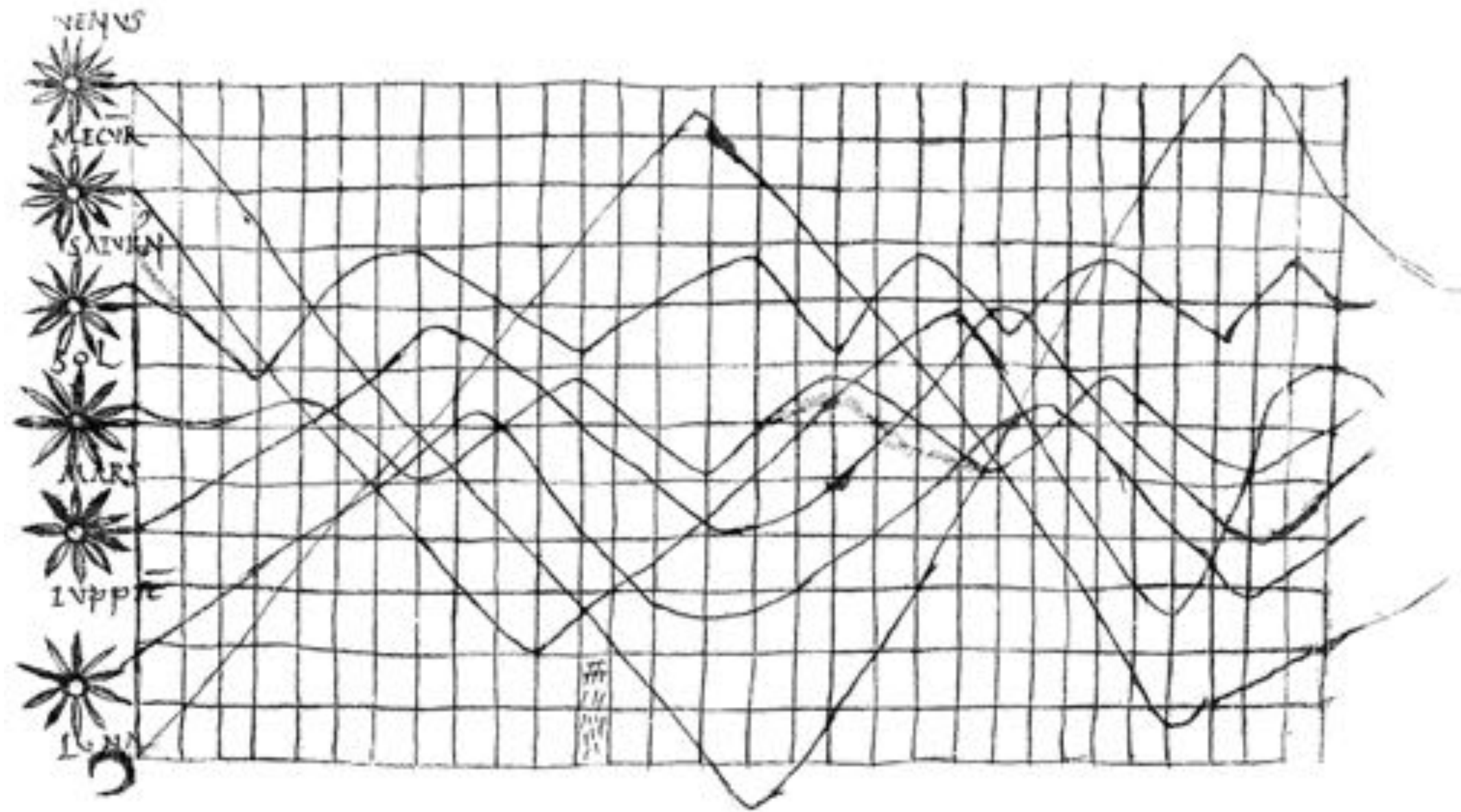*Around 6th Century BCE*

Around 1137 CE

A brief history of data visualization

Around 1686

Basic Visualization Types II
Time Series

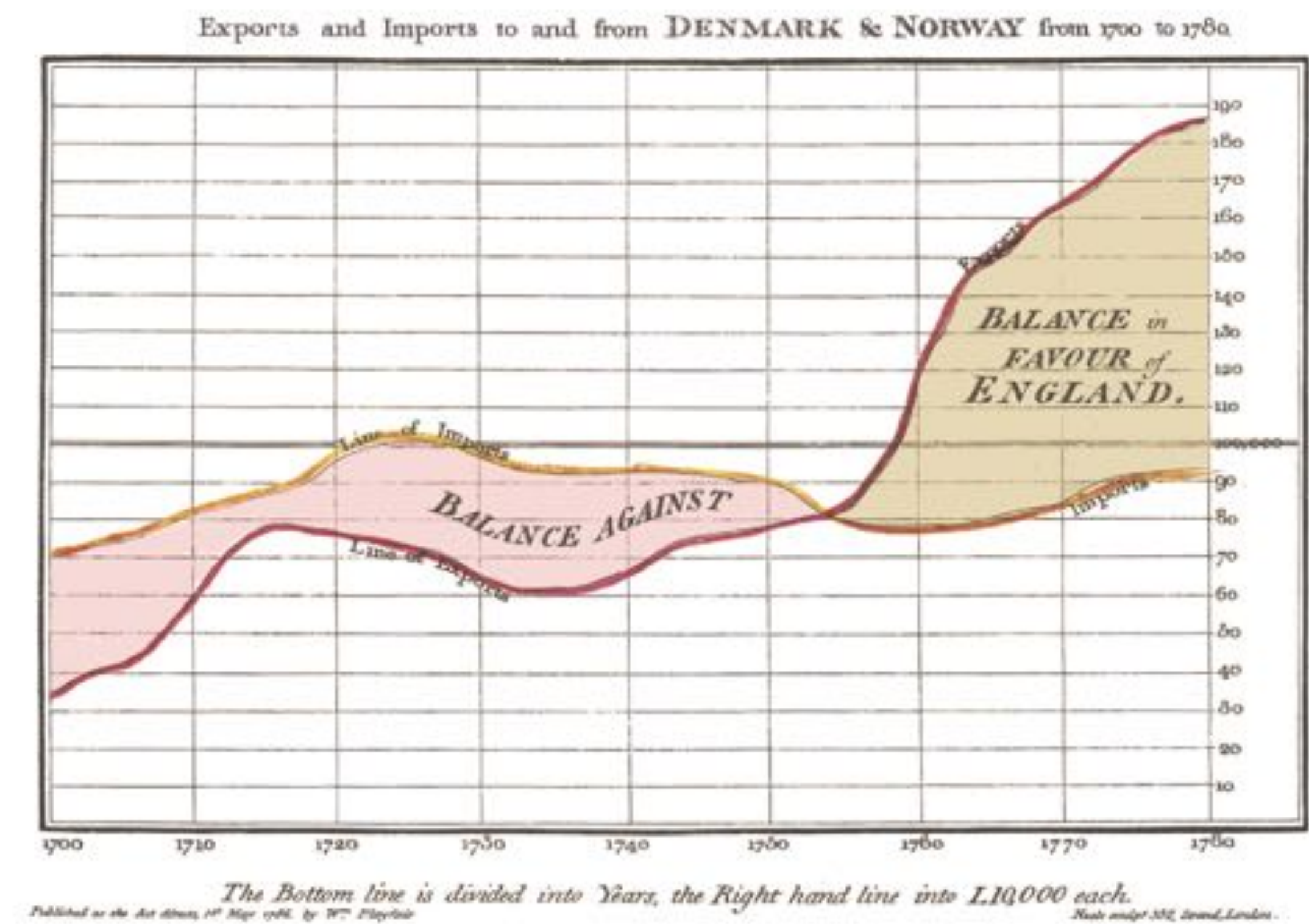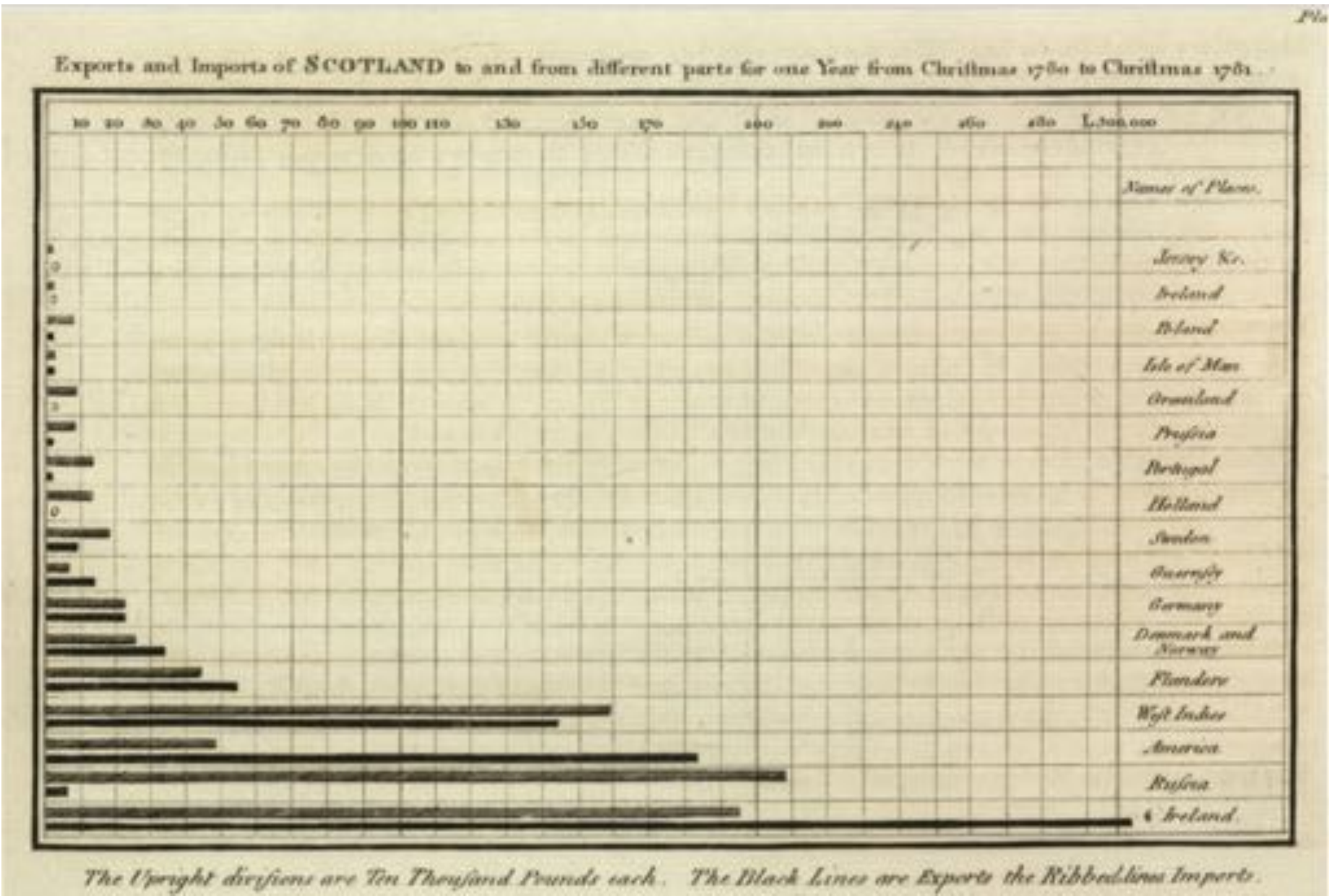*Around 10th century CE*

Nakota People, Early 1800s

Basic Visualization Types II
Time Series

William Playfair, 1786

William Playfair, 1786

Basic Visualization Types II
Time Series

EJ Marey 1885
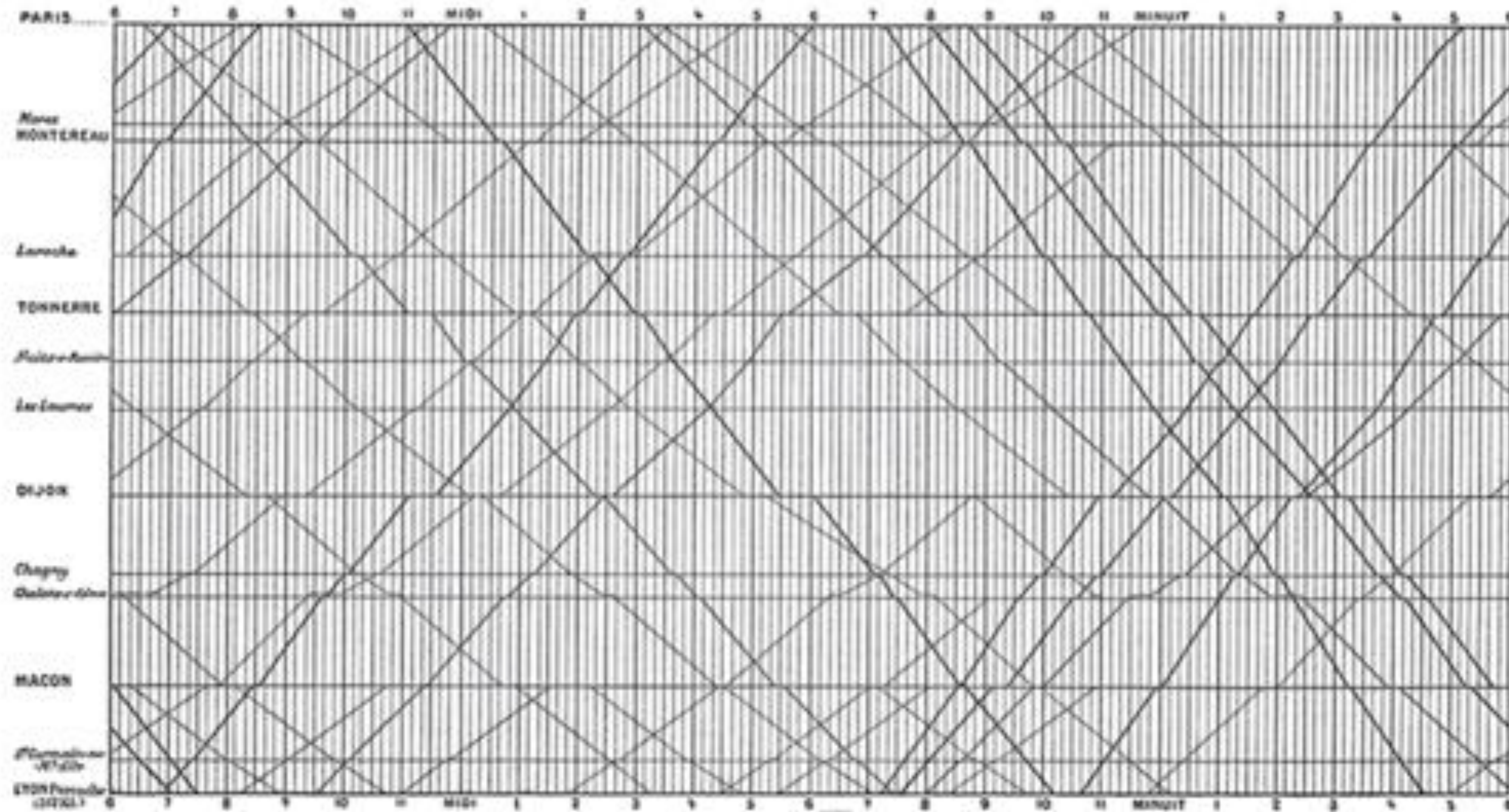
A brief history of data visualization
Spatio-temporal Visualization

Minard, 1869

A brief history of data visualization
Spatio-temporal Visualization

Sloan Digital Sky Survey
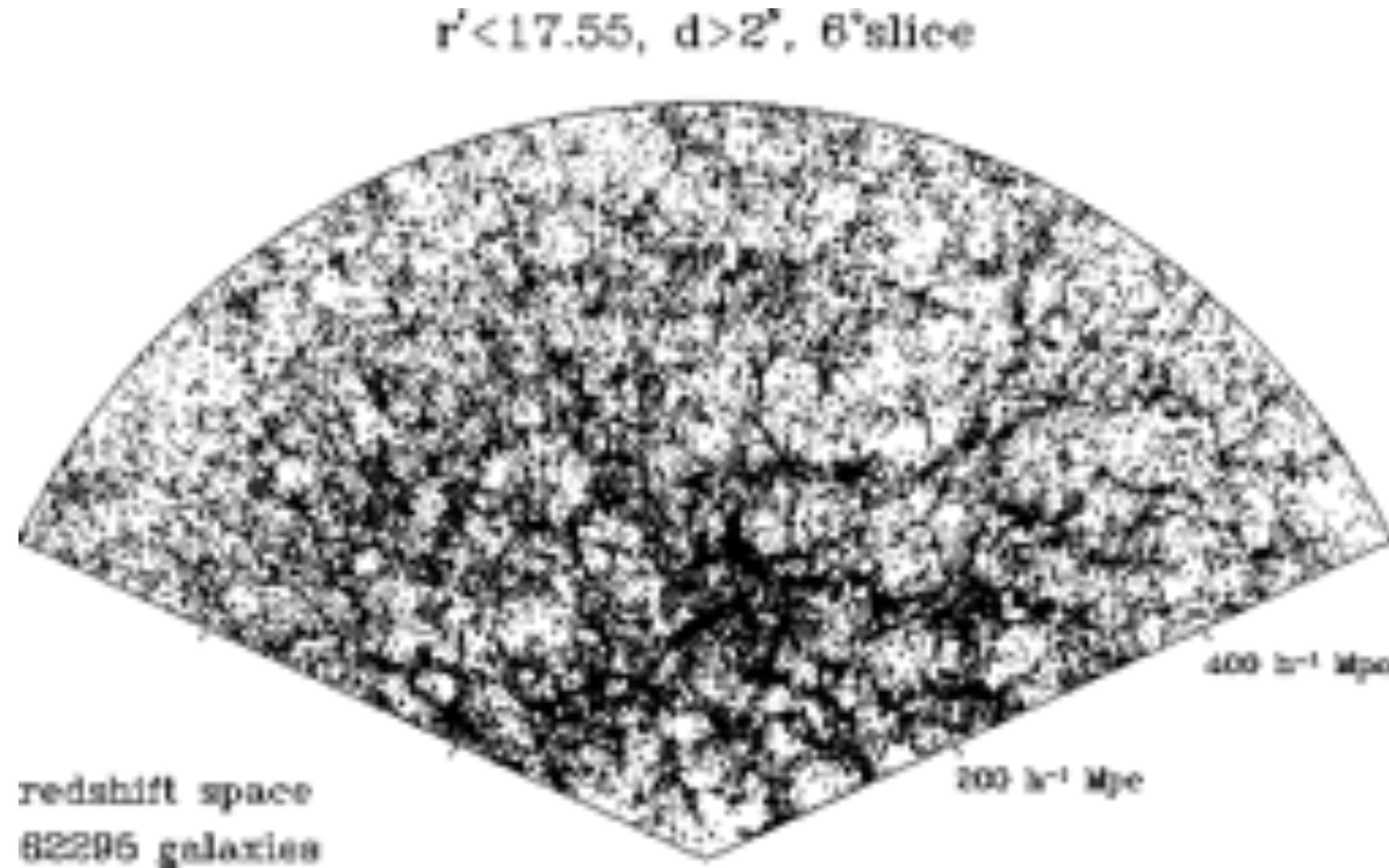
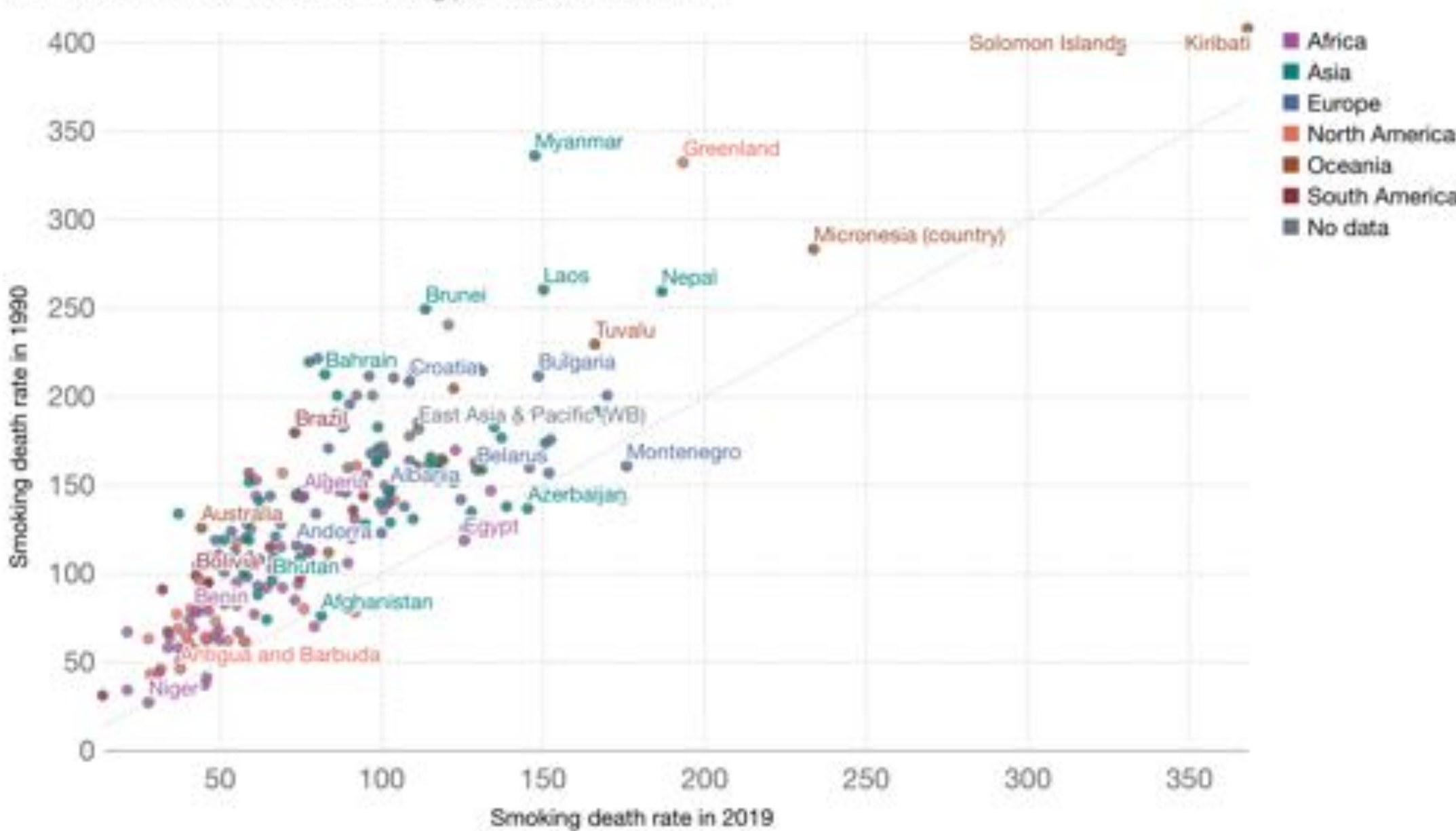# Graphical Sophistication
Relational Graphs

The Statistical Breviary
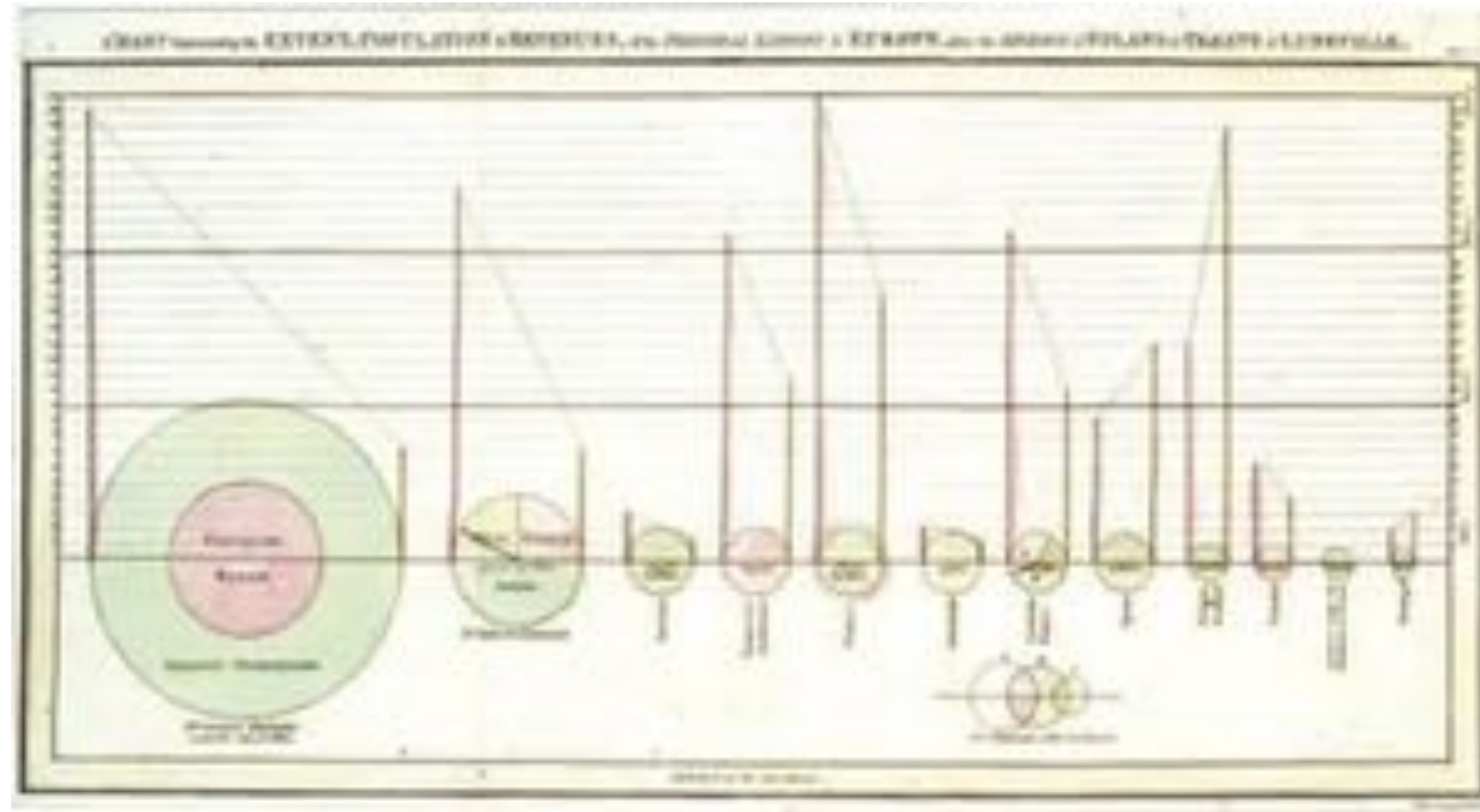
Relational Graphs

Smoking death rate in 1990 vs. 2019
Premature deaths attributed to smoking per 100,000 individuals.

Number of deaths from smoking in 1990 vs. 2019
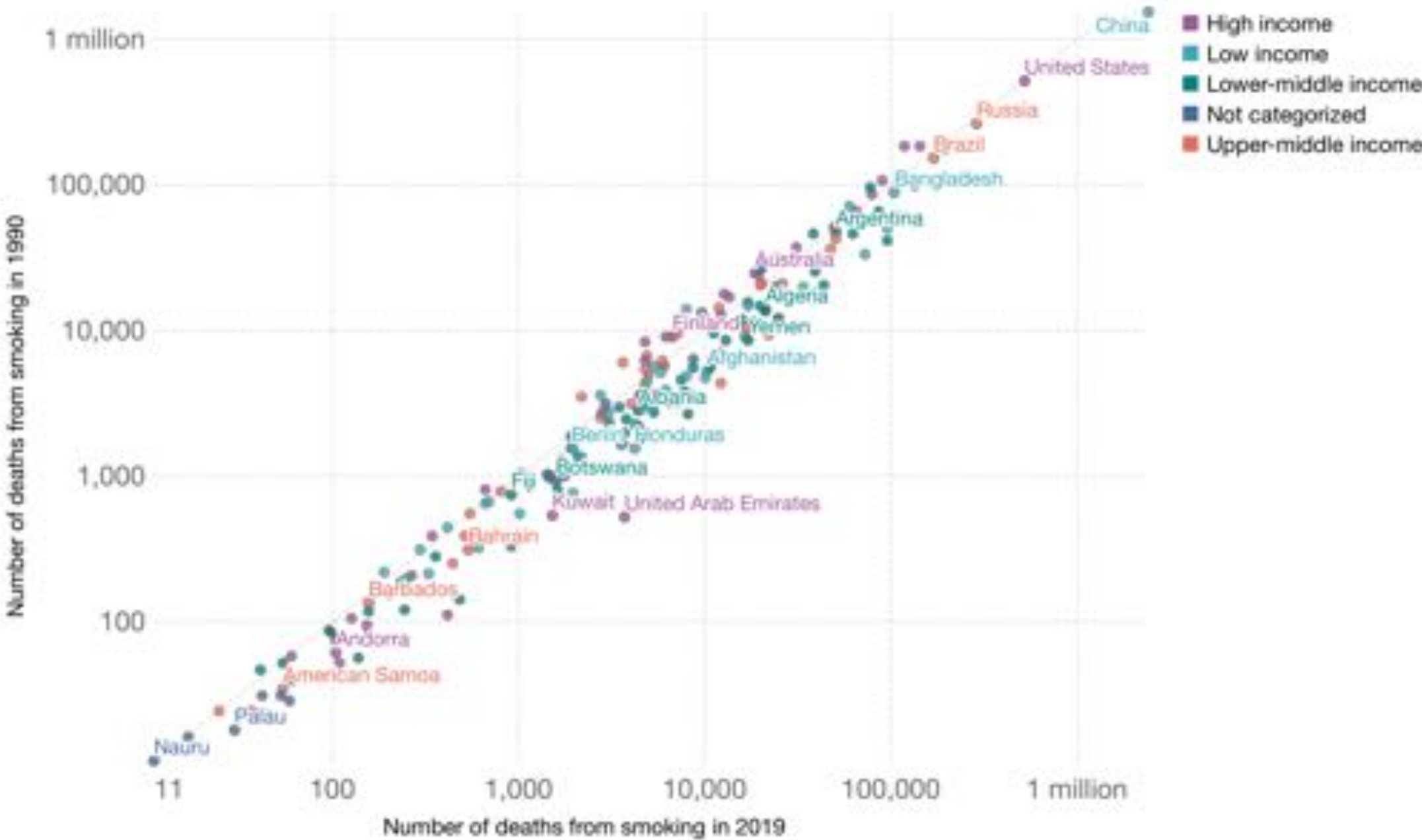
Measuring Graphical Sophistication

Quantitative Rules

$$Data-ink-ratio = \frac{Data-ink}{Ink-to-print-graphic} \approx 1$$

$$Data-density = \frac{area-devoted-to-data}{area-devoted-to-not-data} \approx 1$$

Complexity is approximated by the % of graphs that are relational.

Theory of Data Graphics

"Above all else: show the data."

Maximize the data-ink ratio and data-density

Erase non-data ink

Erase redundant-ink

Be cautious of unintentional optical art

"Graphics are reasoning tools, not art."

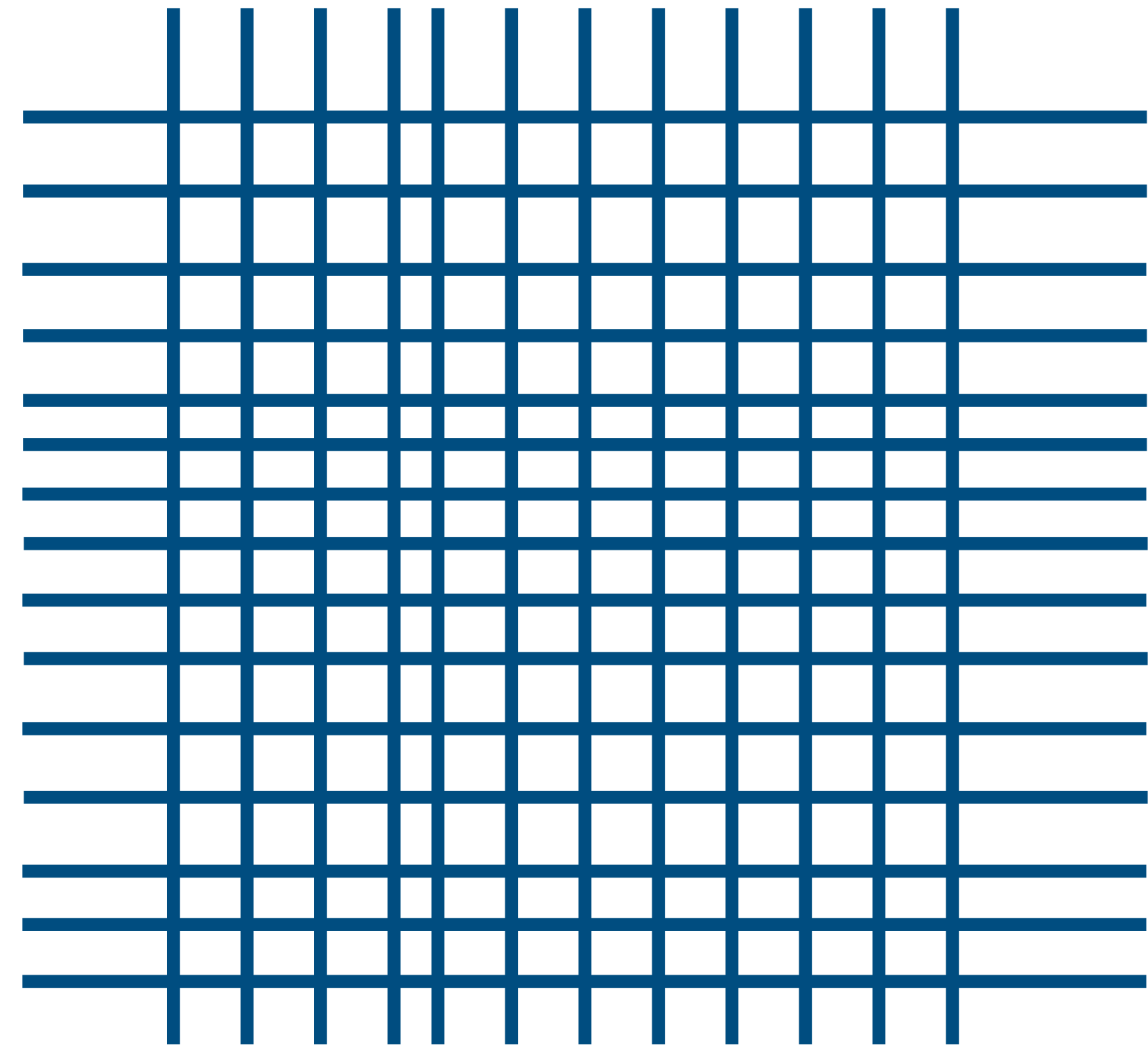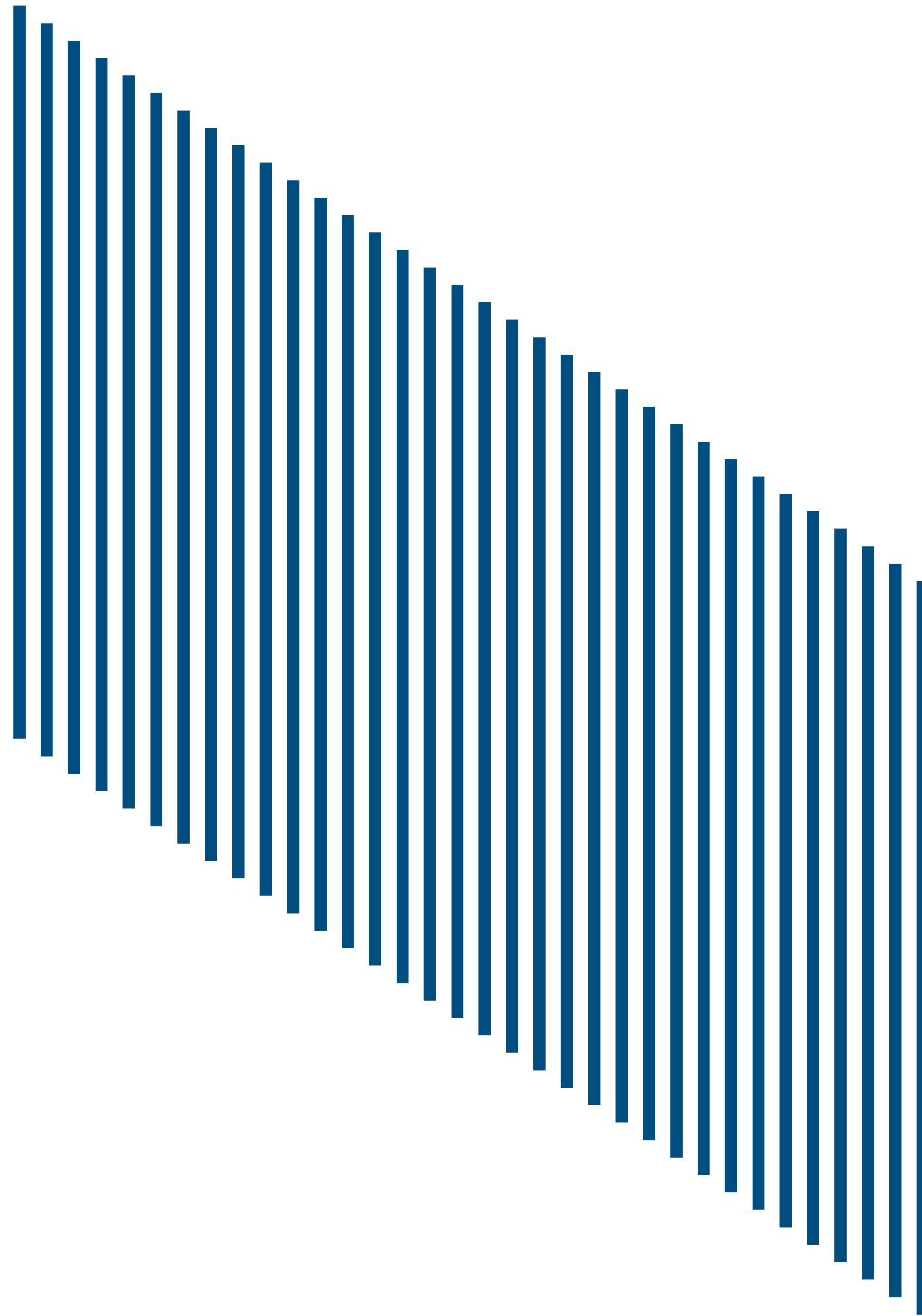# Unintentional Optical Art
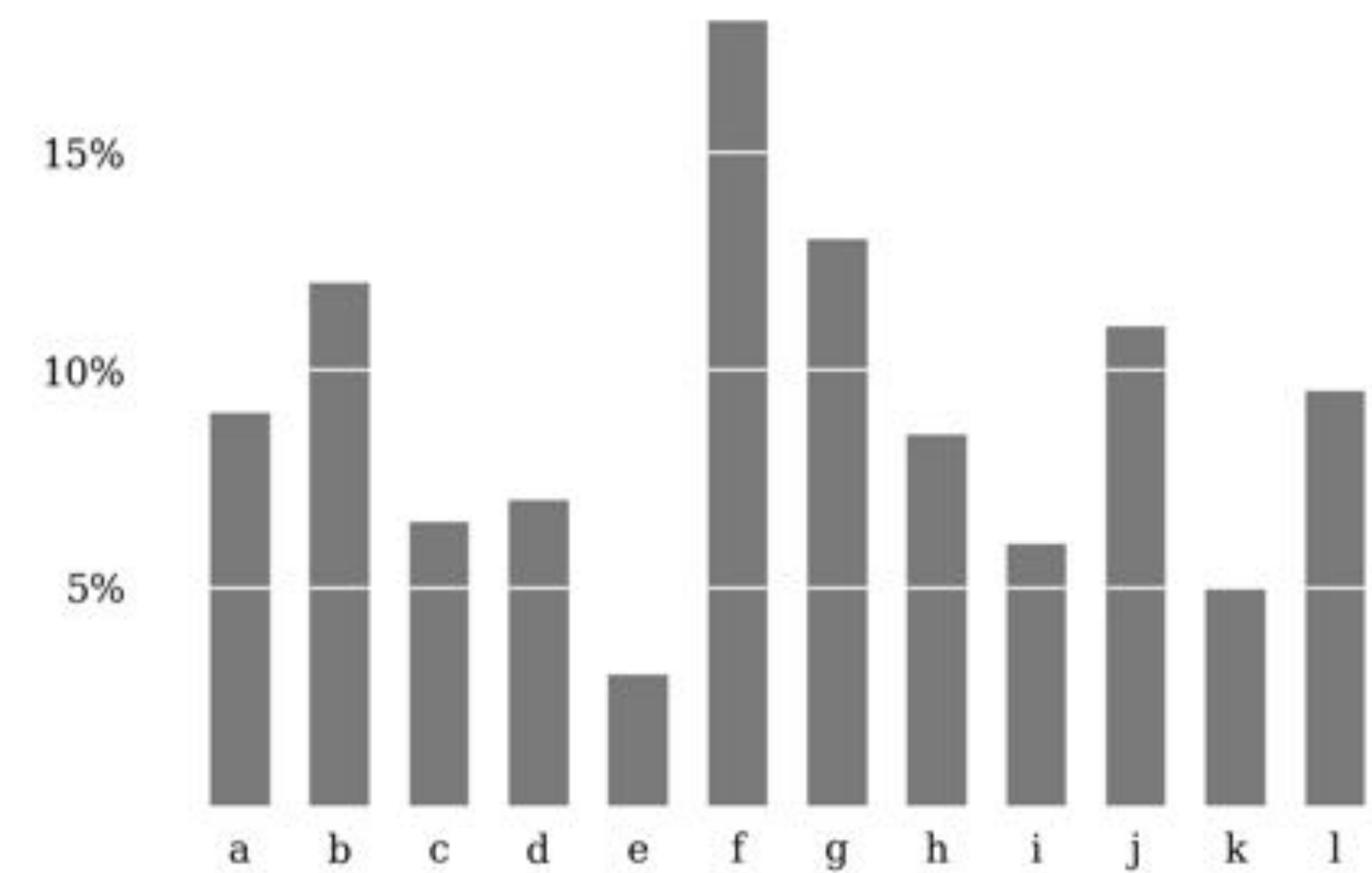
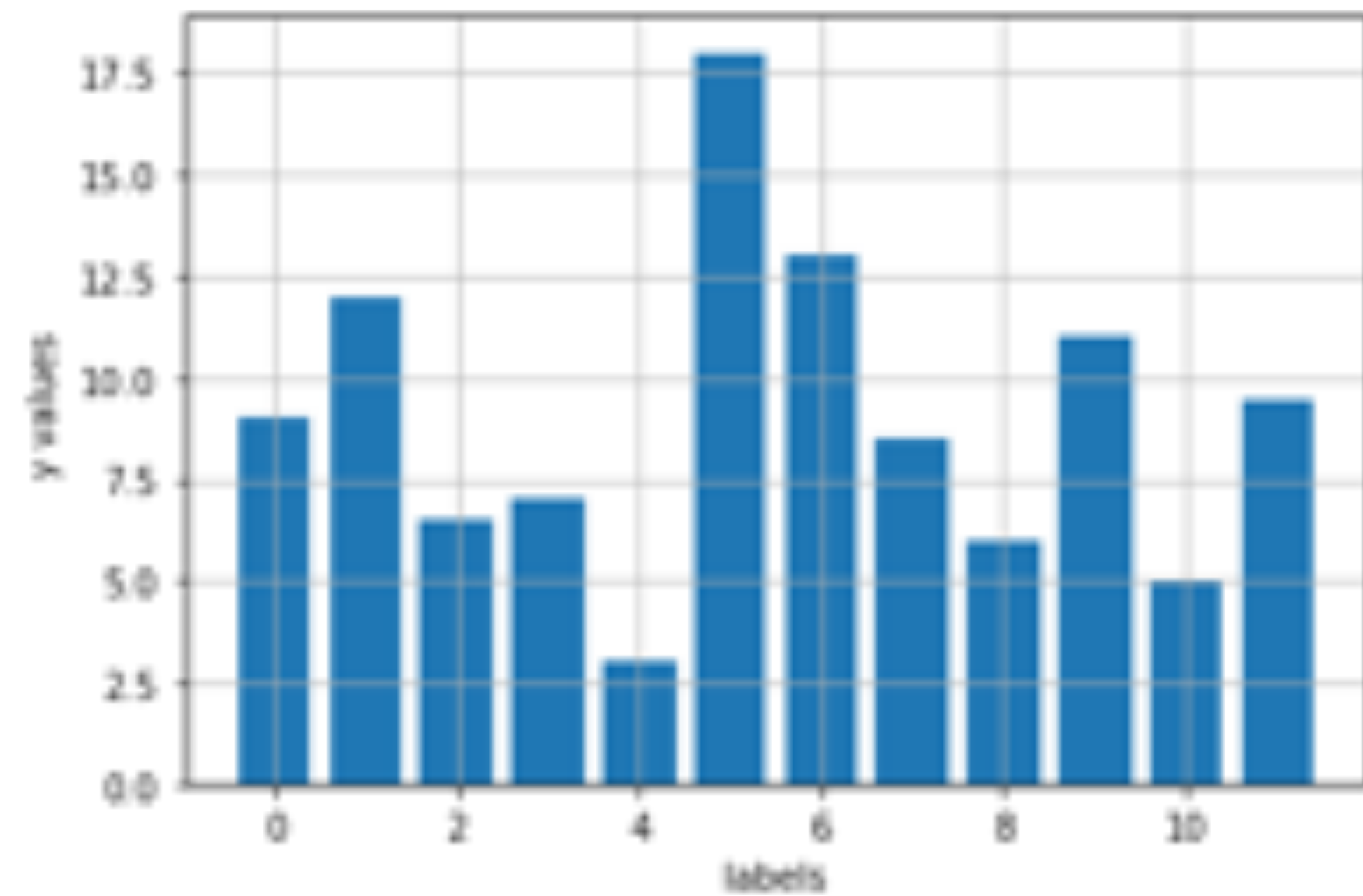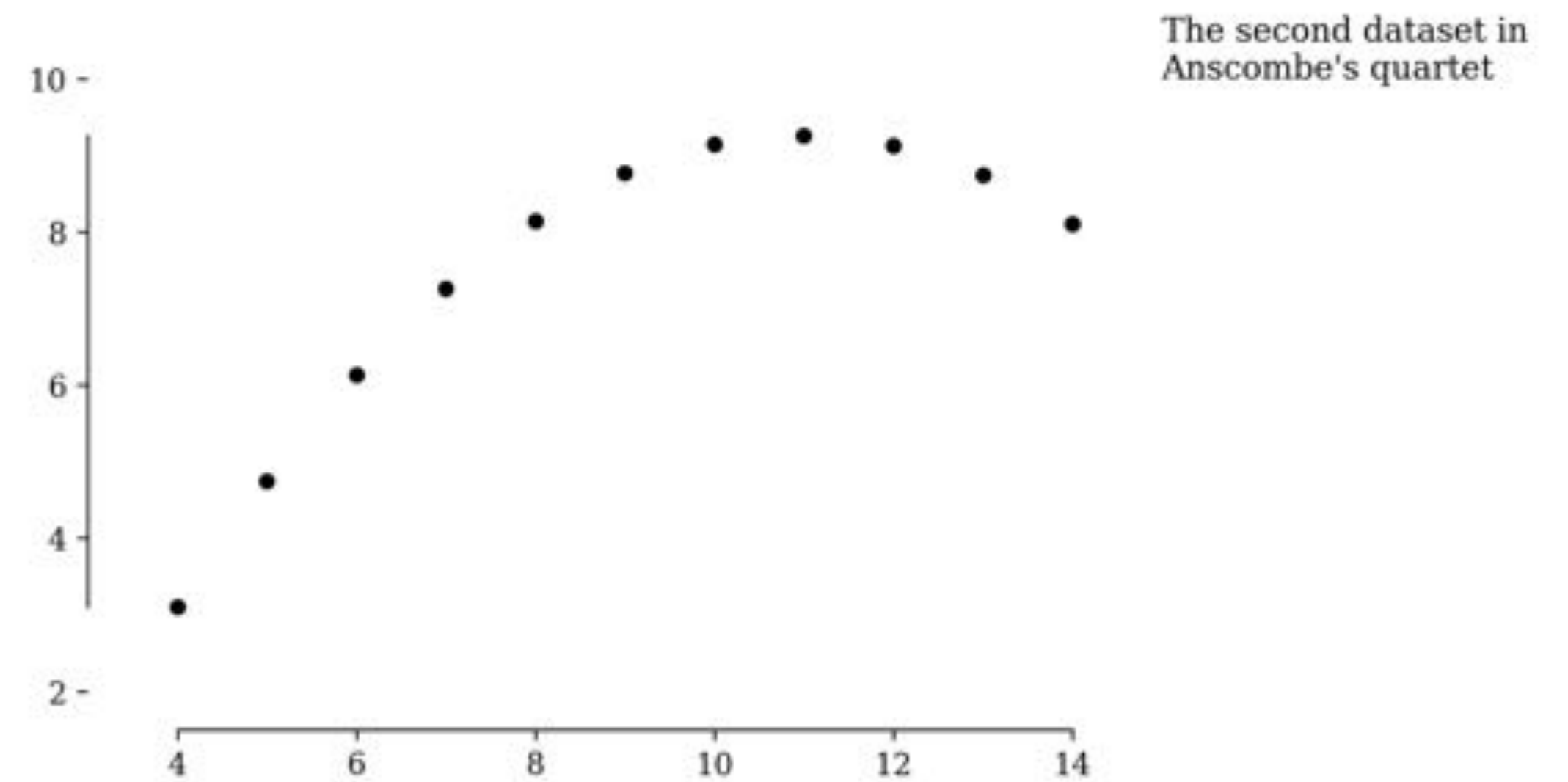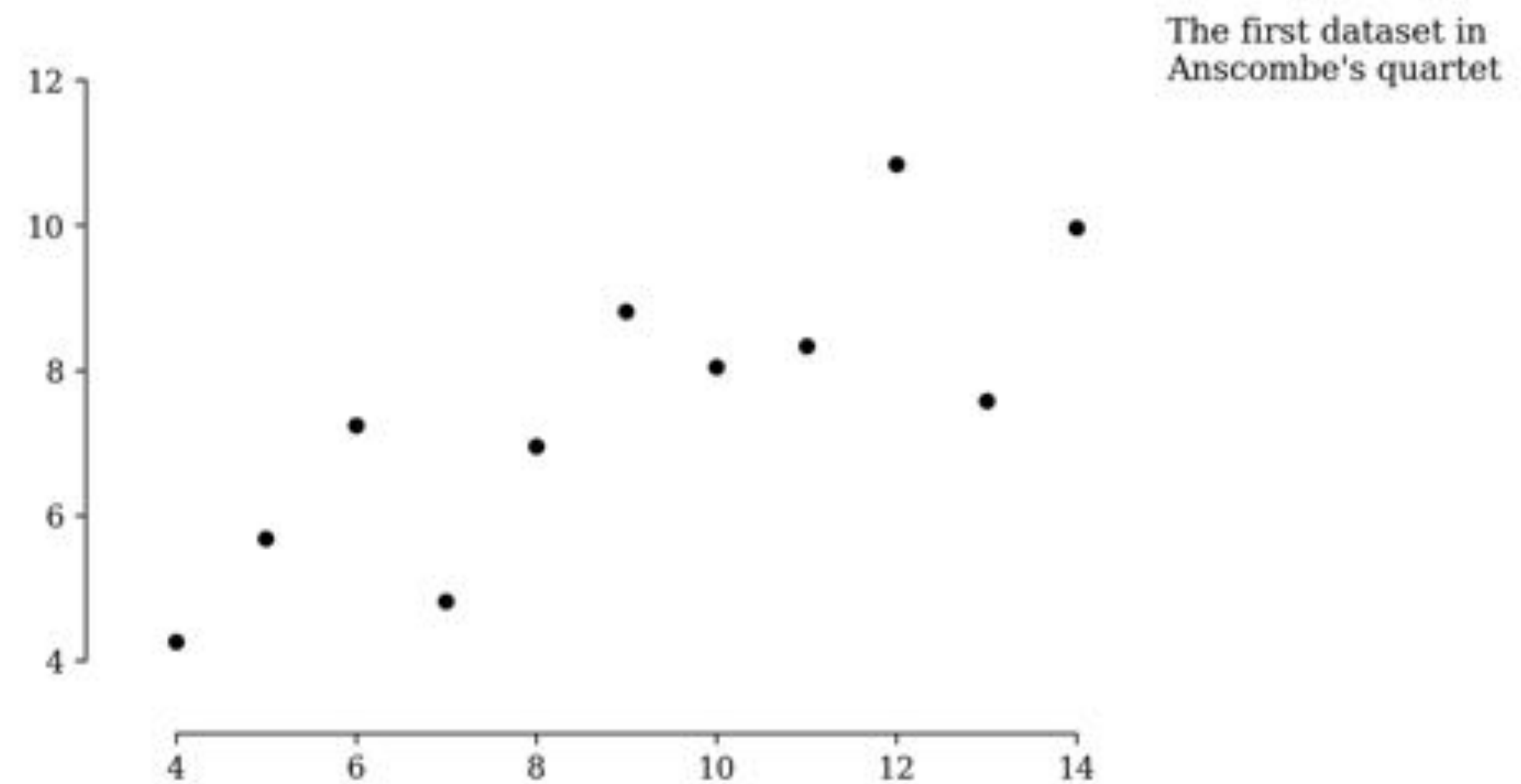**visual vibrations and optical illusions**

Chart-junk
Worked example

Click plot for reference and example matplotlib code.

Chart-junk
Improved Plots (according to Tufte)

The first dataset in
Anscombe's quartet

The second dataset in
Anscombe's quartet

Click plots for reference and example matplotlib code.

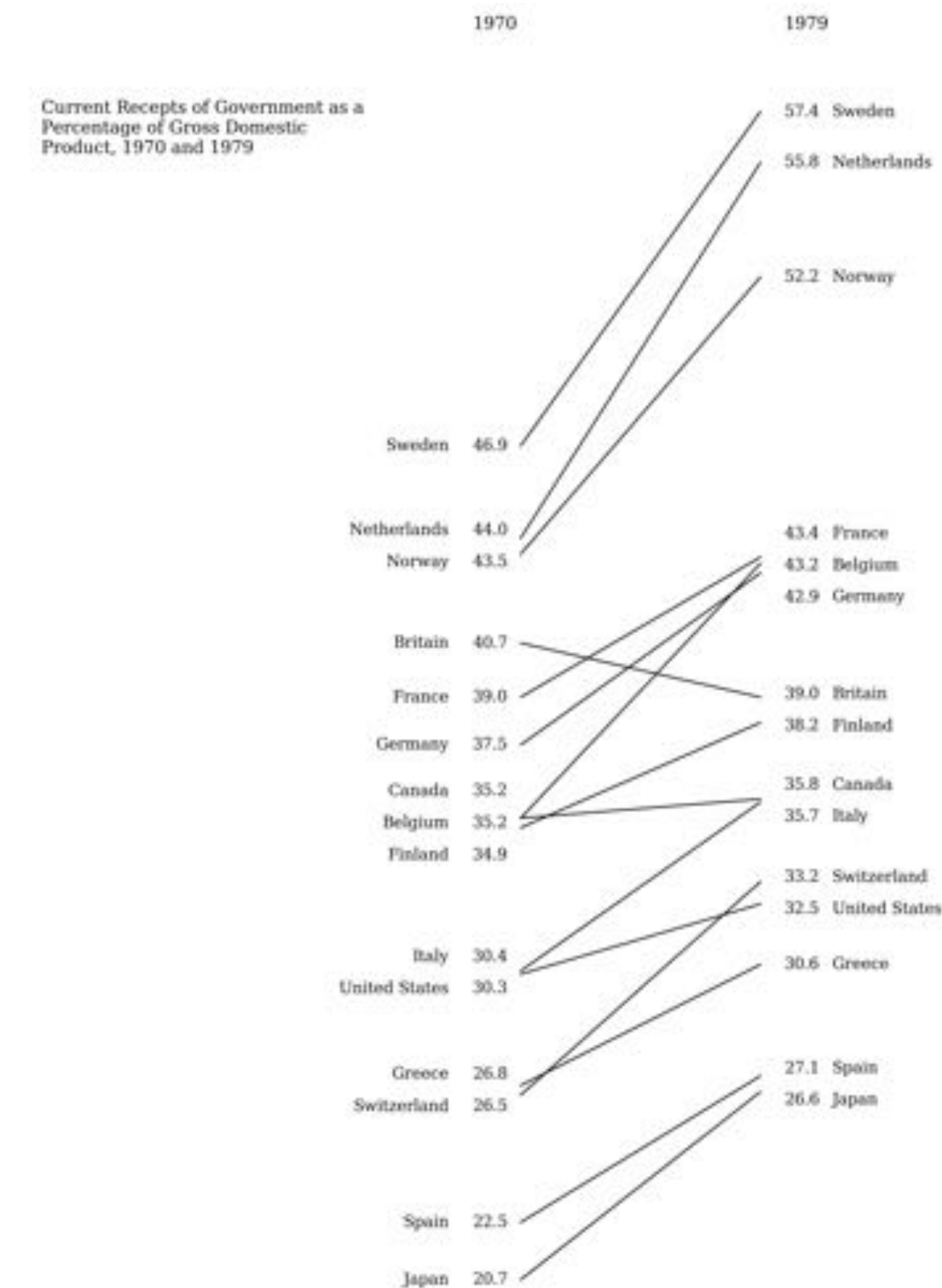# Multifunctional Graphical Elements

Using redundancy

Relative ordering of labels denotes rank in a distribution

Lines both connect over time and the slope carries meaningful information

Data is both integrated and clearly delineated.

High measures of data-density and graphical sophistication

Click plot for reference and example matplotlib code.

visualizations to communicate and discover

Foundations

Alberto Cairo defines:

Visualization - Any kind of visual representation designed to enable communication, analysis, discovery.

Infographic - A multi-section visual representation of information intended to communicate one or more specific messages.

Data Visualization - A display of data designed to enable analysis, exploration and discovery.

Infographics are about *communication* while data visualization is about *discovery*.

Insightful visualizations
Clearing the path to valuable discoveries

When designing a visualization, think about what you want to **enable**.

Spontaneous insights are best enabled by *static* visualizations. The *EUREKA!* moment.

Complicated and in-depth discoveries that involve a slow **accumulation** of knowledge require *interactive* visualizations and high sophistication.

Visualizations are **models**. We ultimately still have to make decisions about what features of our data we think are important or that will be highlighted.

Basic Visualization Workflow
Mapping data onto visual properties

Think about the task you want to enable

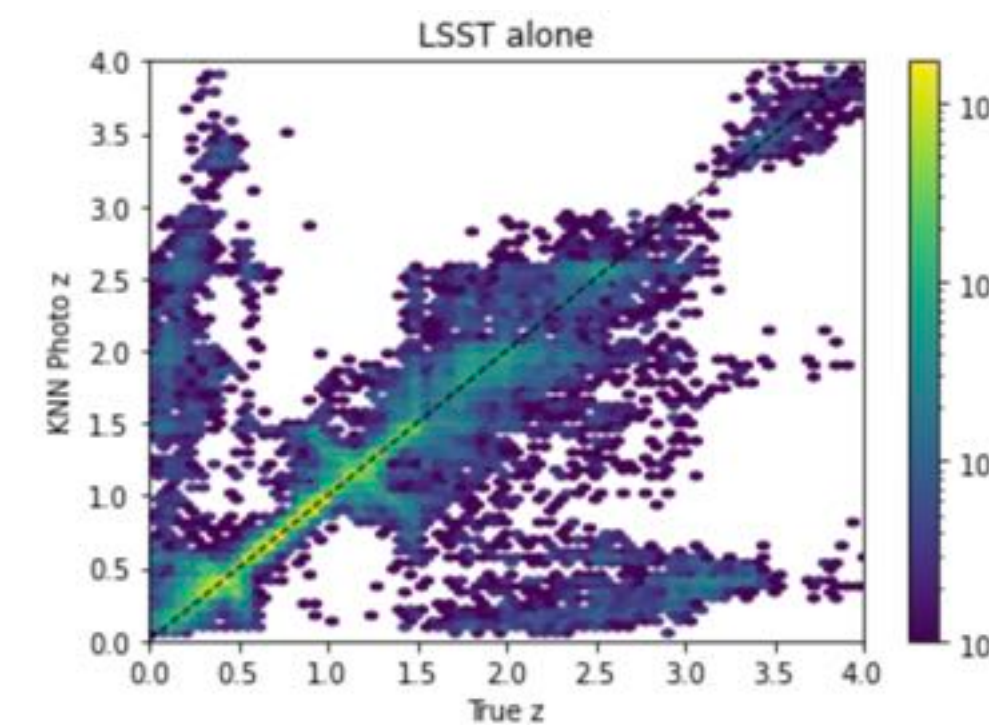Try different graphical forms
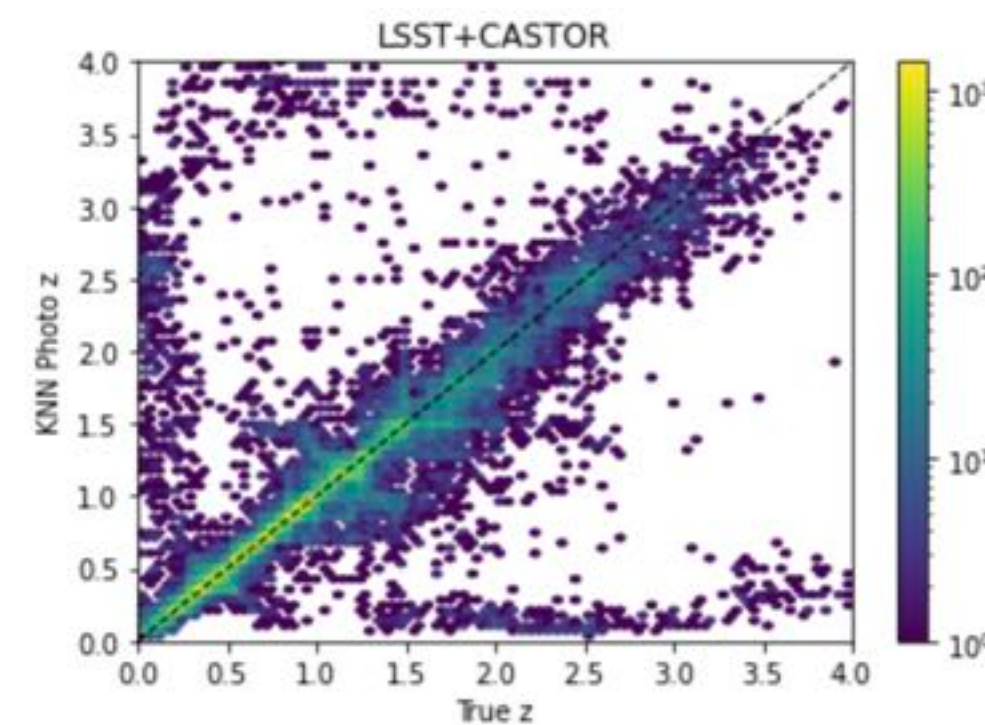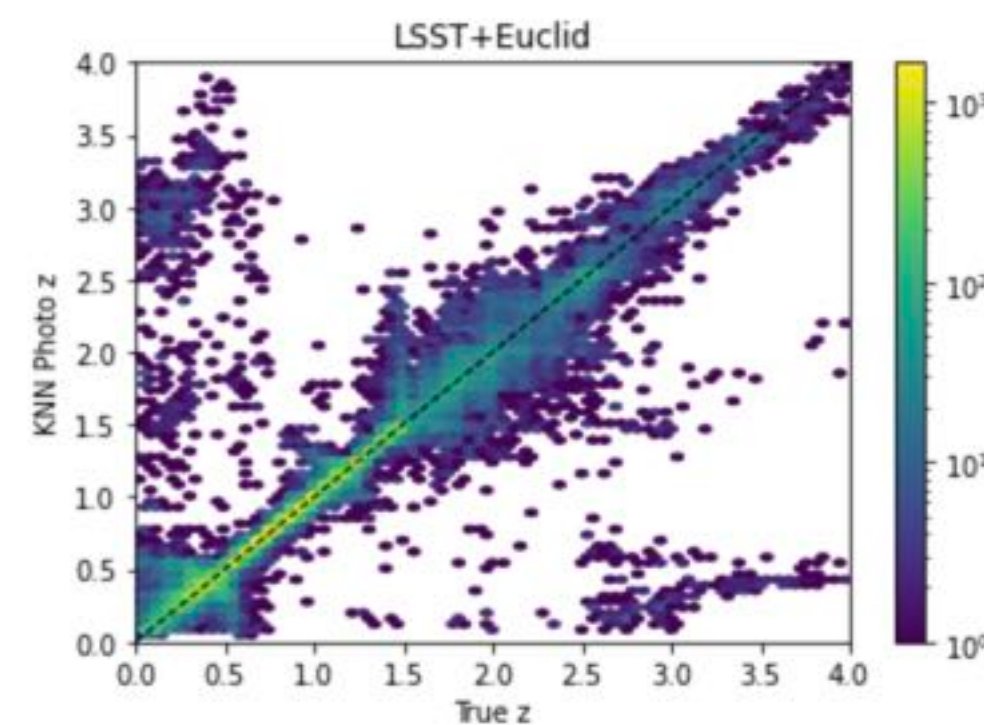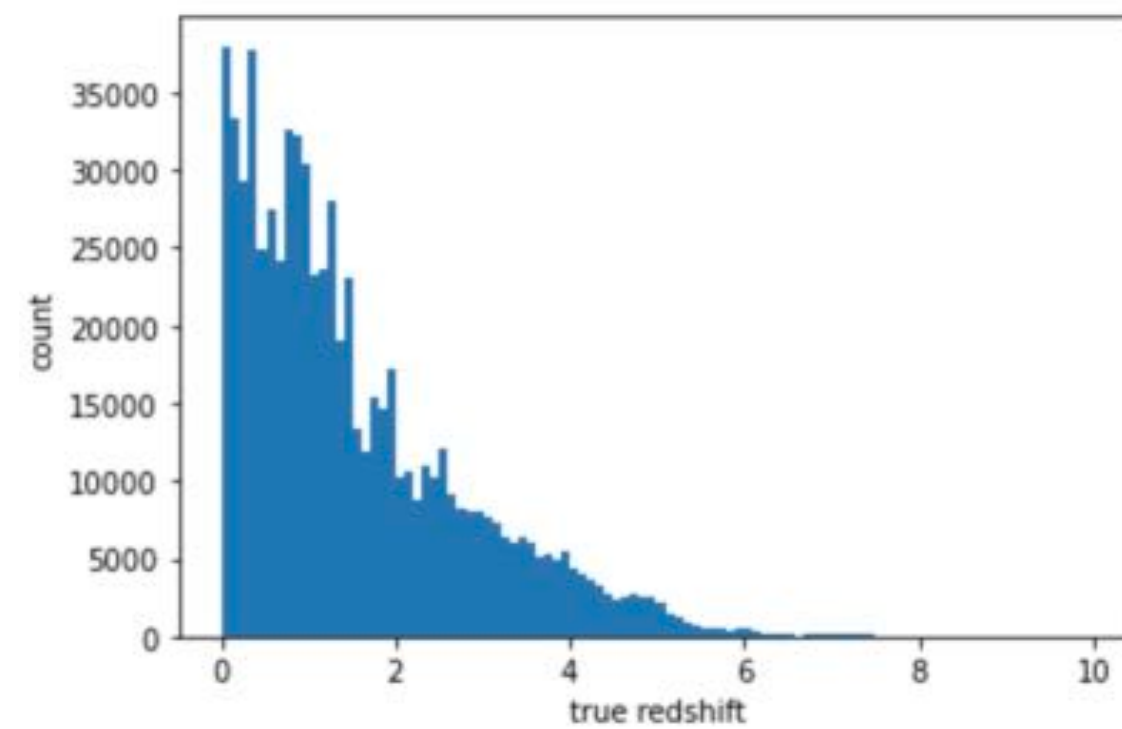
Arrange components of the graphic

Test alternatives

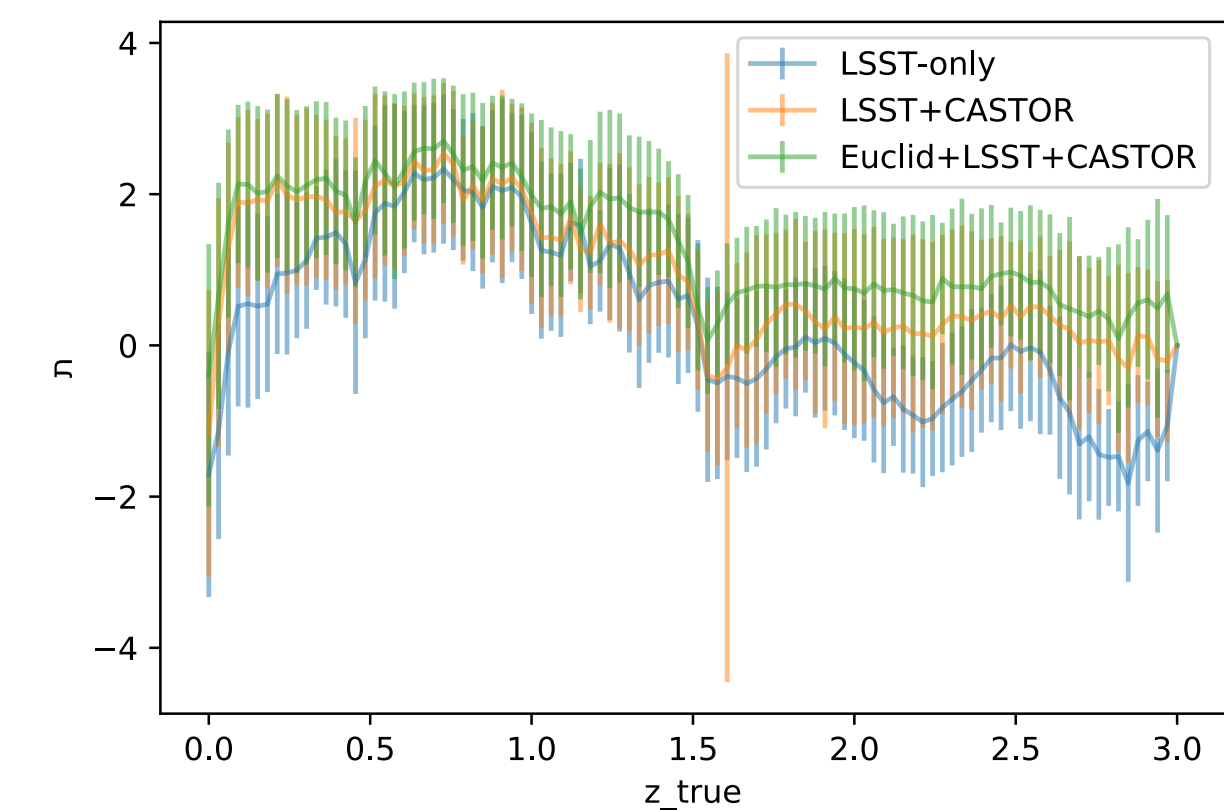   Data visualization catalog

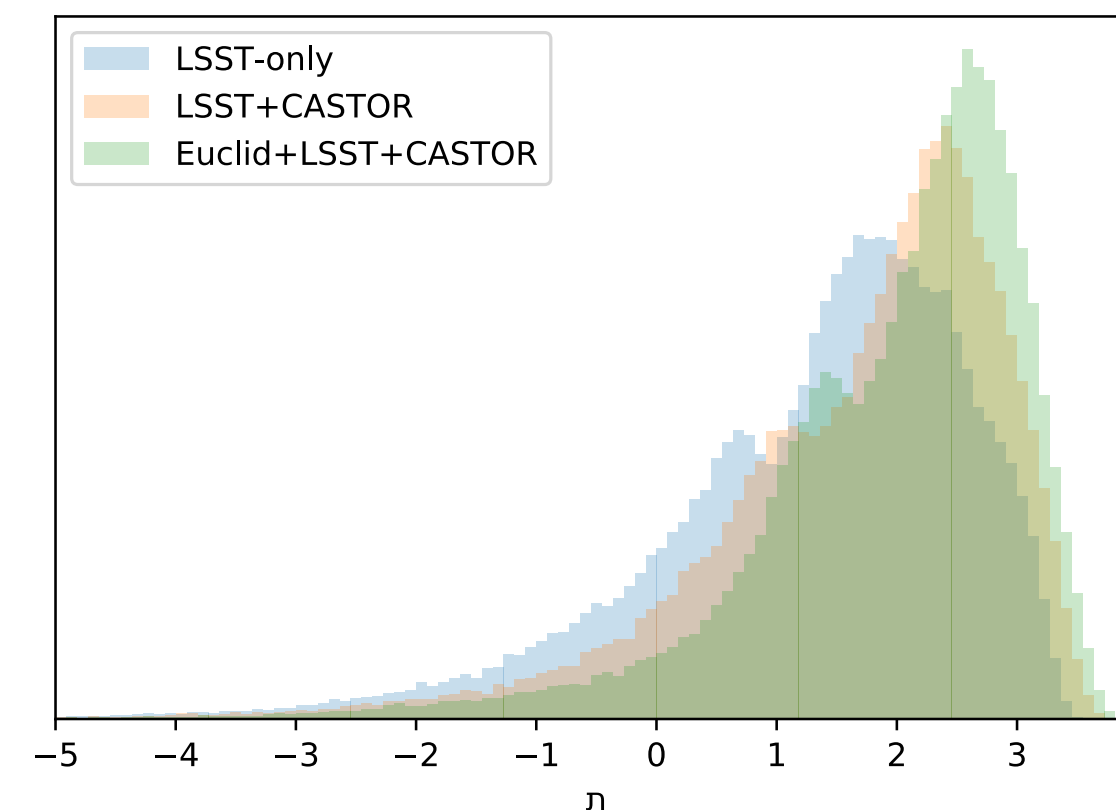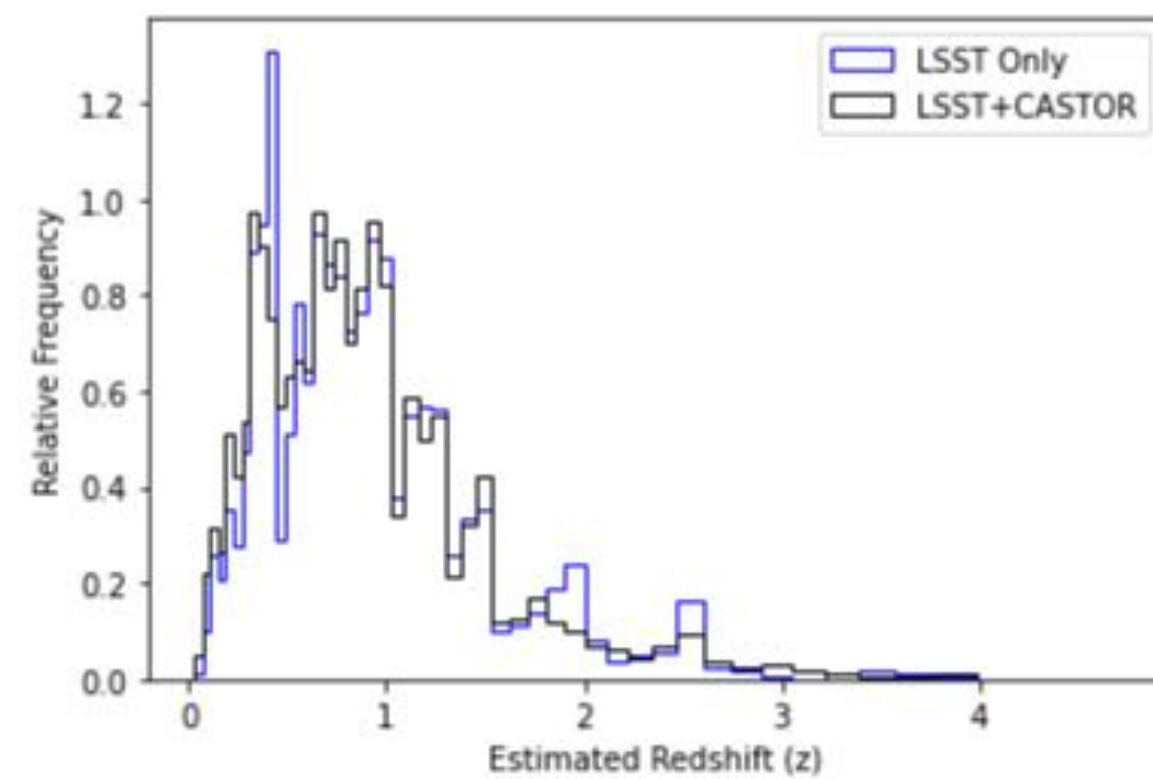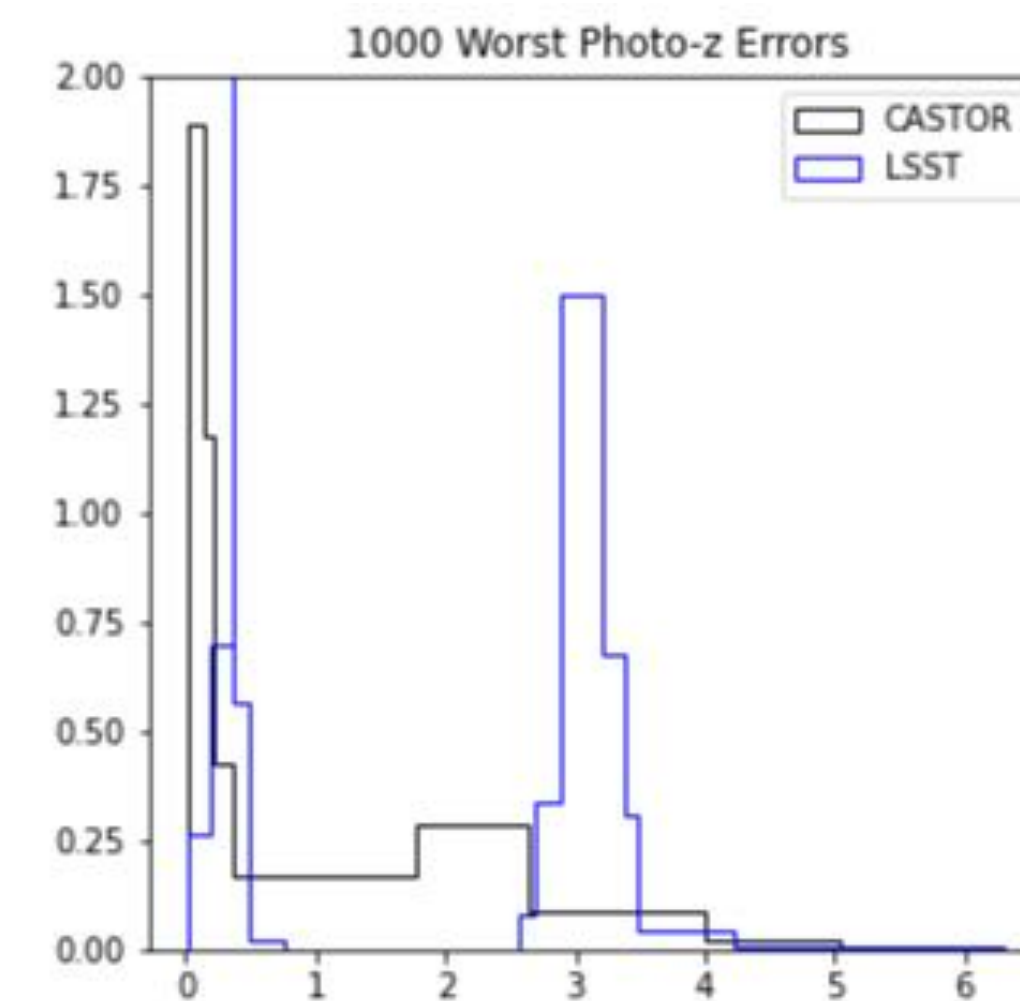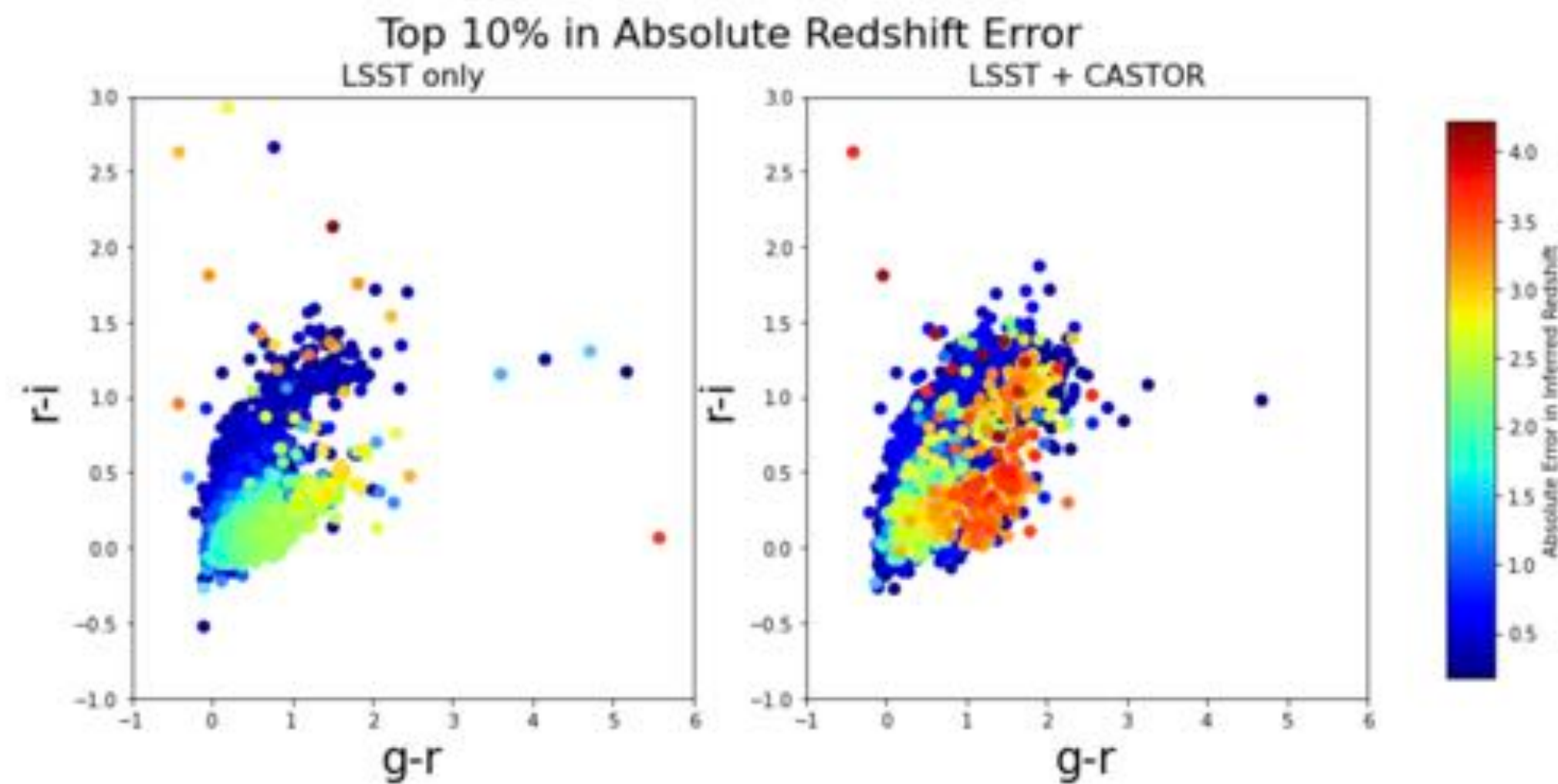   Scale of elementary perceptual tasks

Practice: Truth, breadth, depth, and context
information theory metrics for future wide field surveys

Practice: Truth, breadth, depth, and context
information theory metrics for future wide field surveys
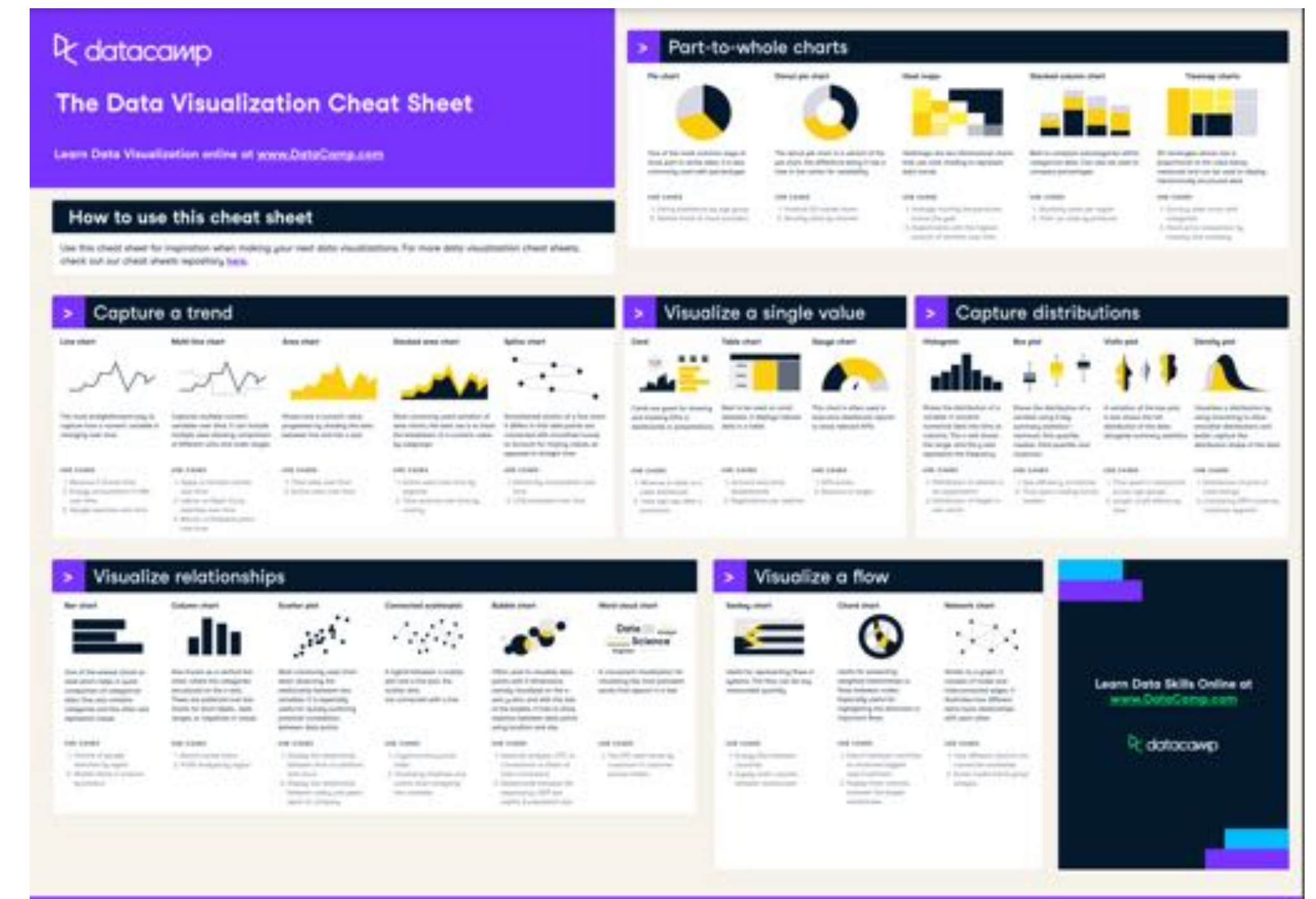
Types of Visualizations

What's in your toolbox.

Part to whole - heat maps and pie charts

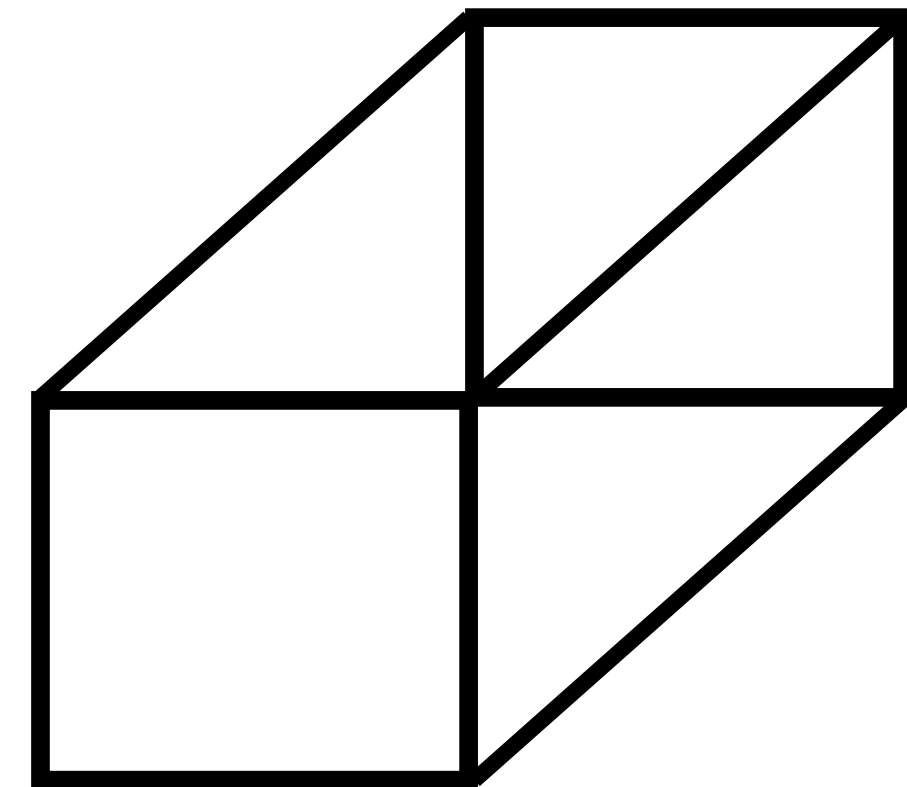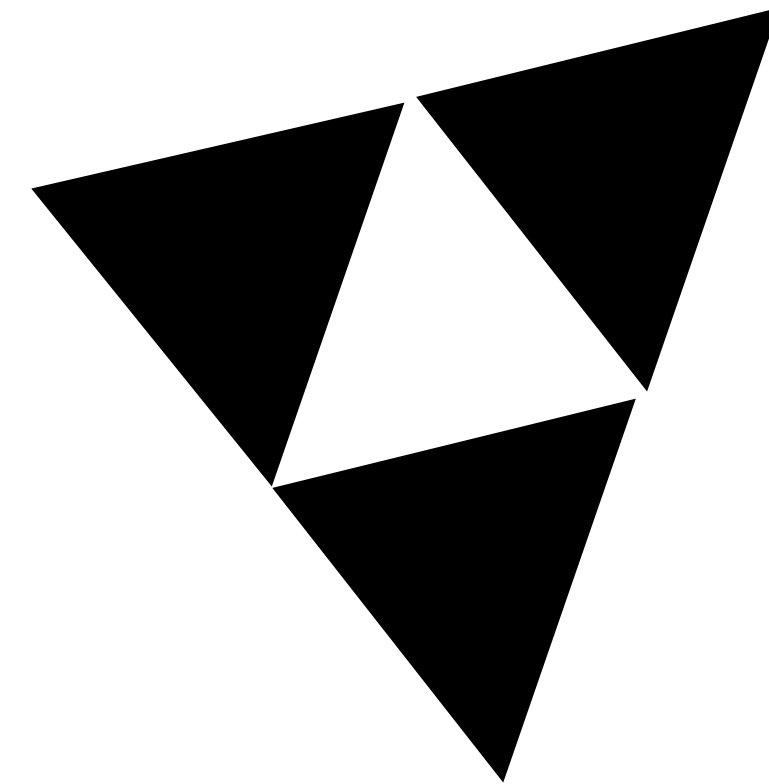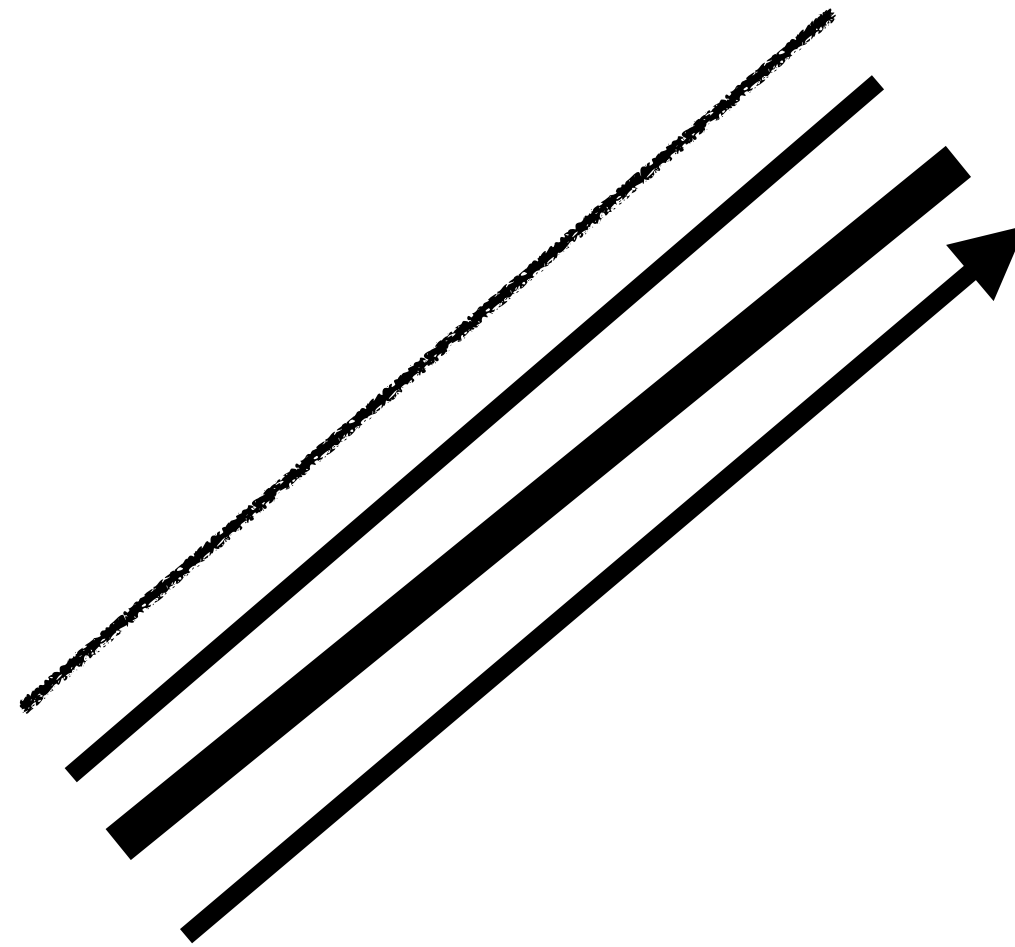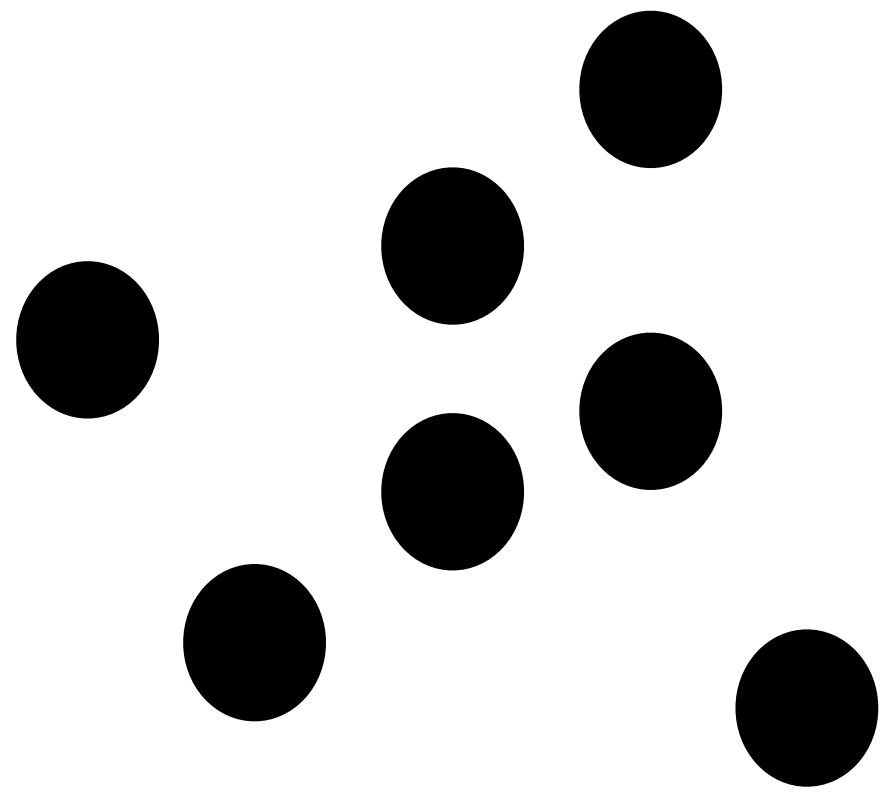Trends and relationships - scatter plots

Distributions - histograms and box plots

Flows and hierarchies - Graphs

Inventory of Graphical Elements

for customized visualizations

Inventory of Symbols and Encodings

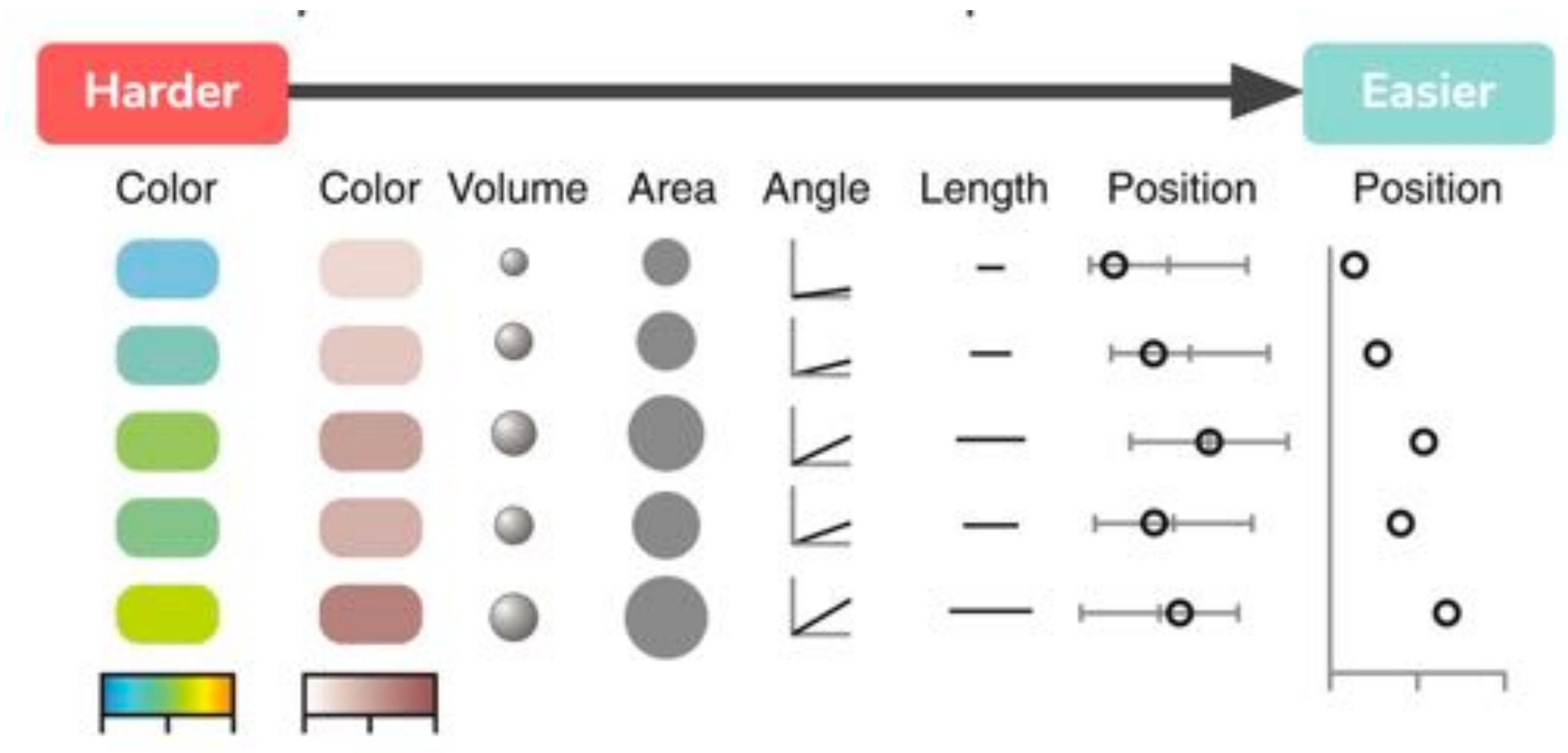**Pick symbols that are distinct when they overlap…**

Inventory of Symbols and Encodings
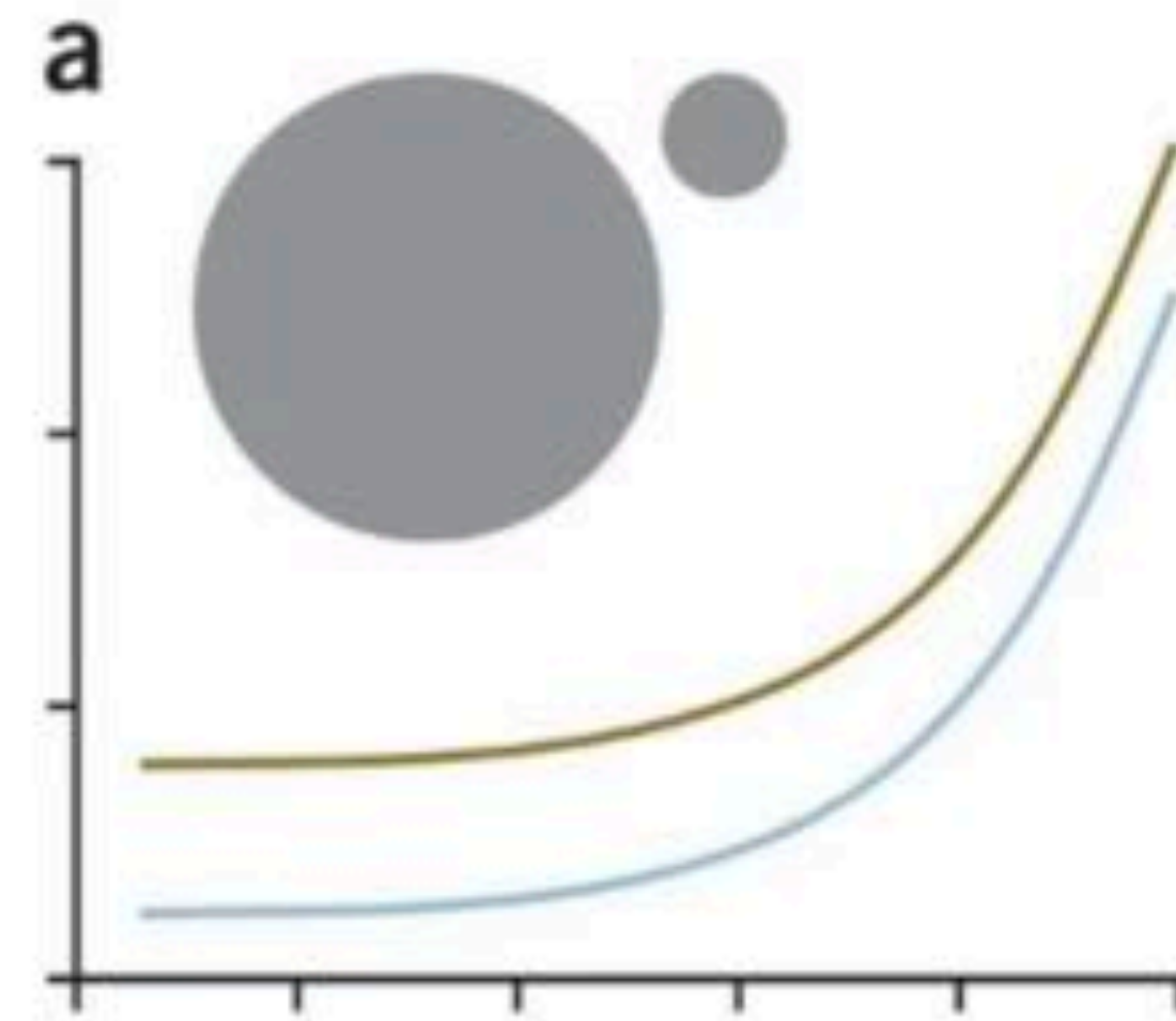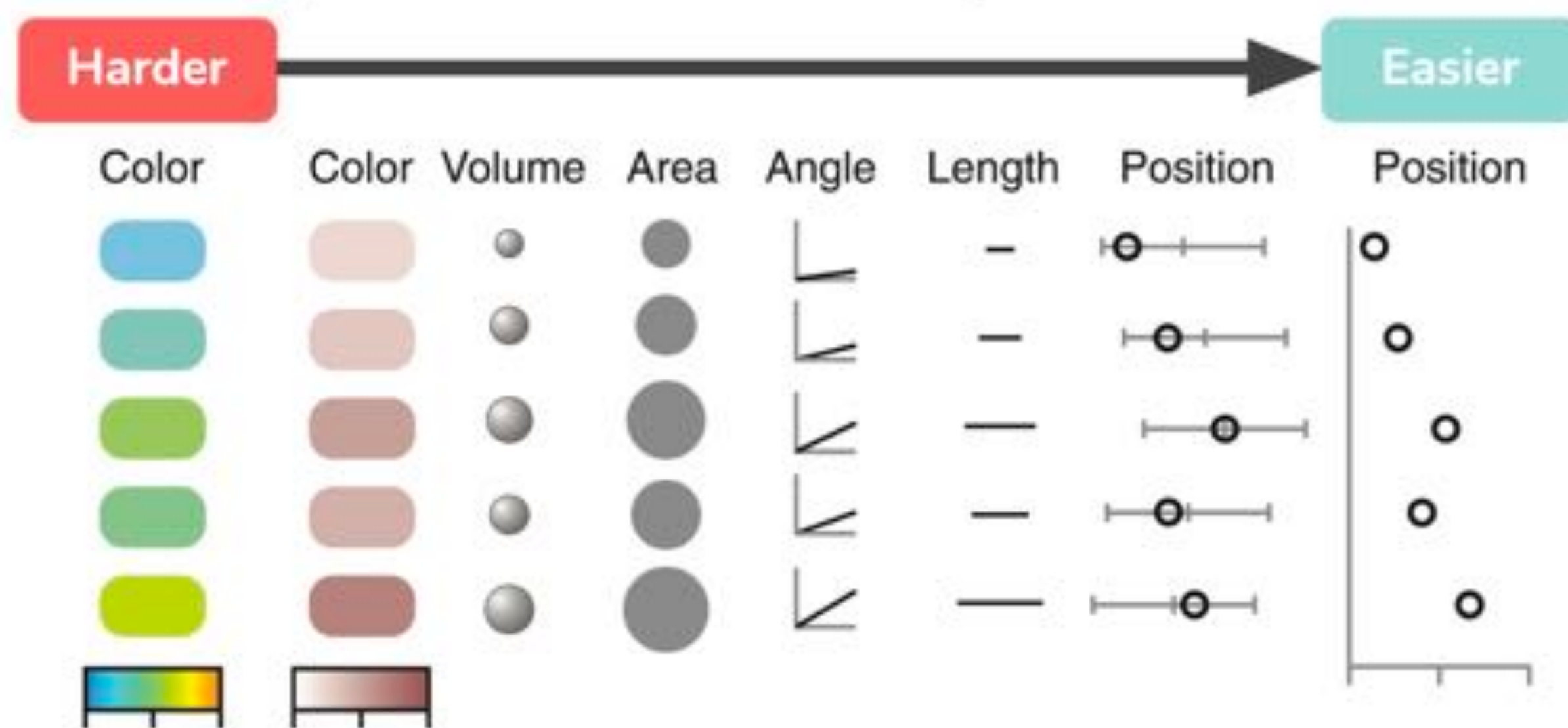## … and that form natural hierarchies.
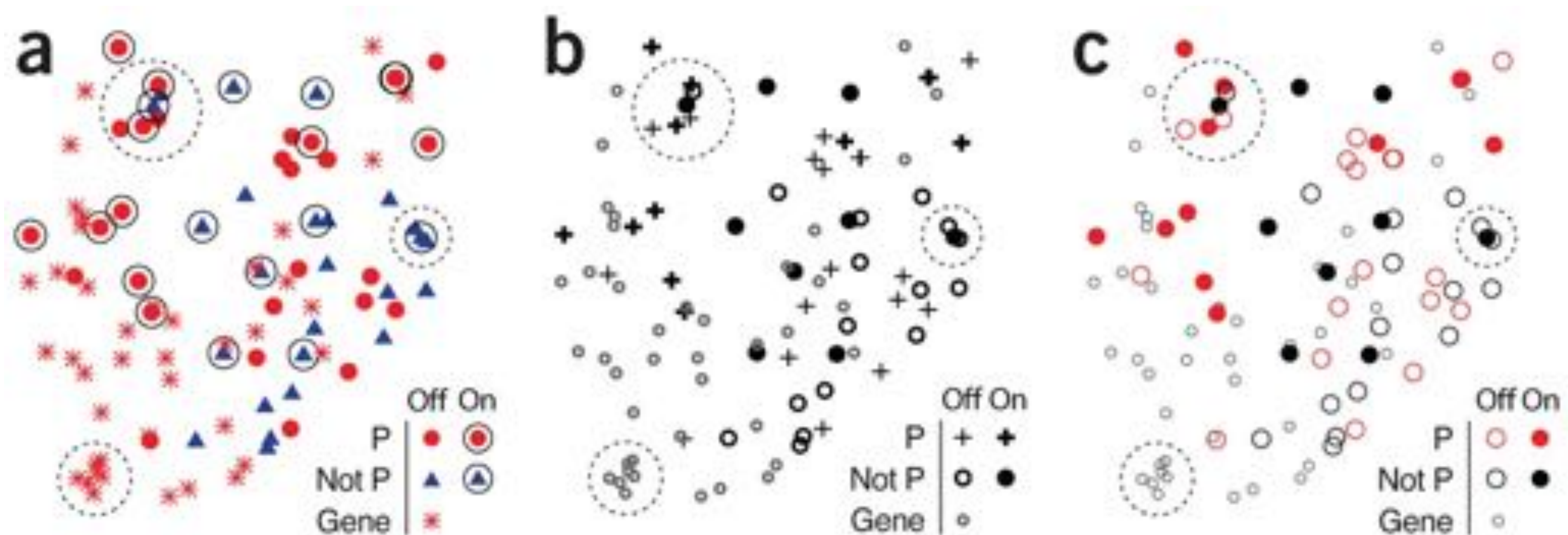
Scale of Elementary Perceptual Tasks
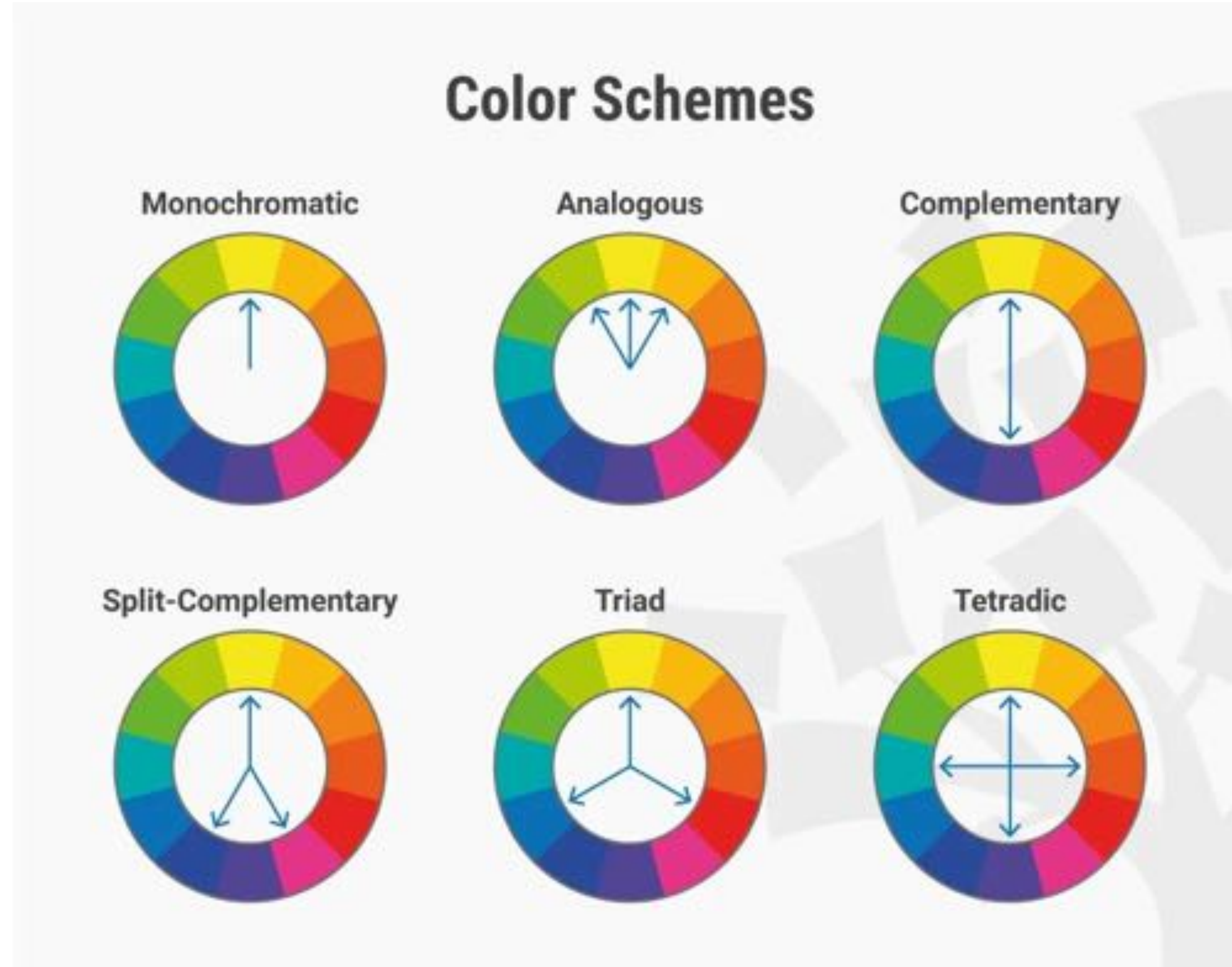## Enabling accurate and truthful insight

Scale of Elementary Perceptual Tasks
Enabling accurate and truthful insight

Color Theory
The color circle

Color Schemes

Monochromatic

Analogous

Complementary

Split-Complementary

Triad

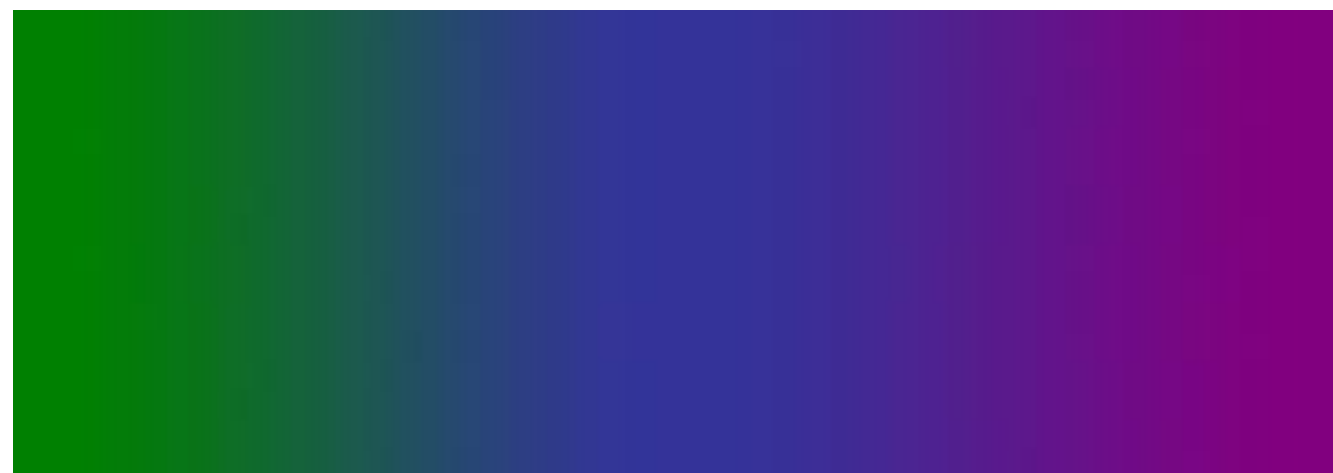Tetradic

Color Cultural Associations

Color spectra

Warm colors are energizing, passionate, and positive.

West: Importance and danger/anger, Asia: Prosperity

Cool colors are calming, relaxing, and somewhat reserved.

West: Nature, renewal, peace, religious connotations

East: Political

Neutral colors are often accents or backgrounds

West: purity, cleanliness, and virtue, as well as medicine.

East: death and mourning.
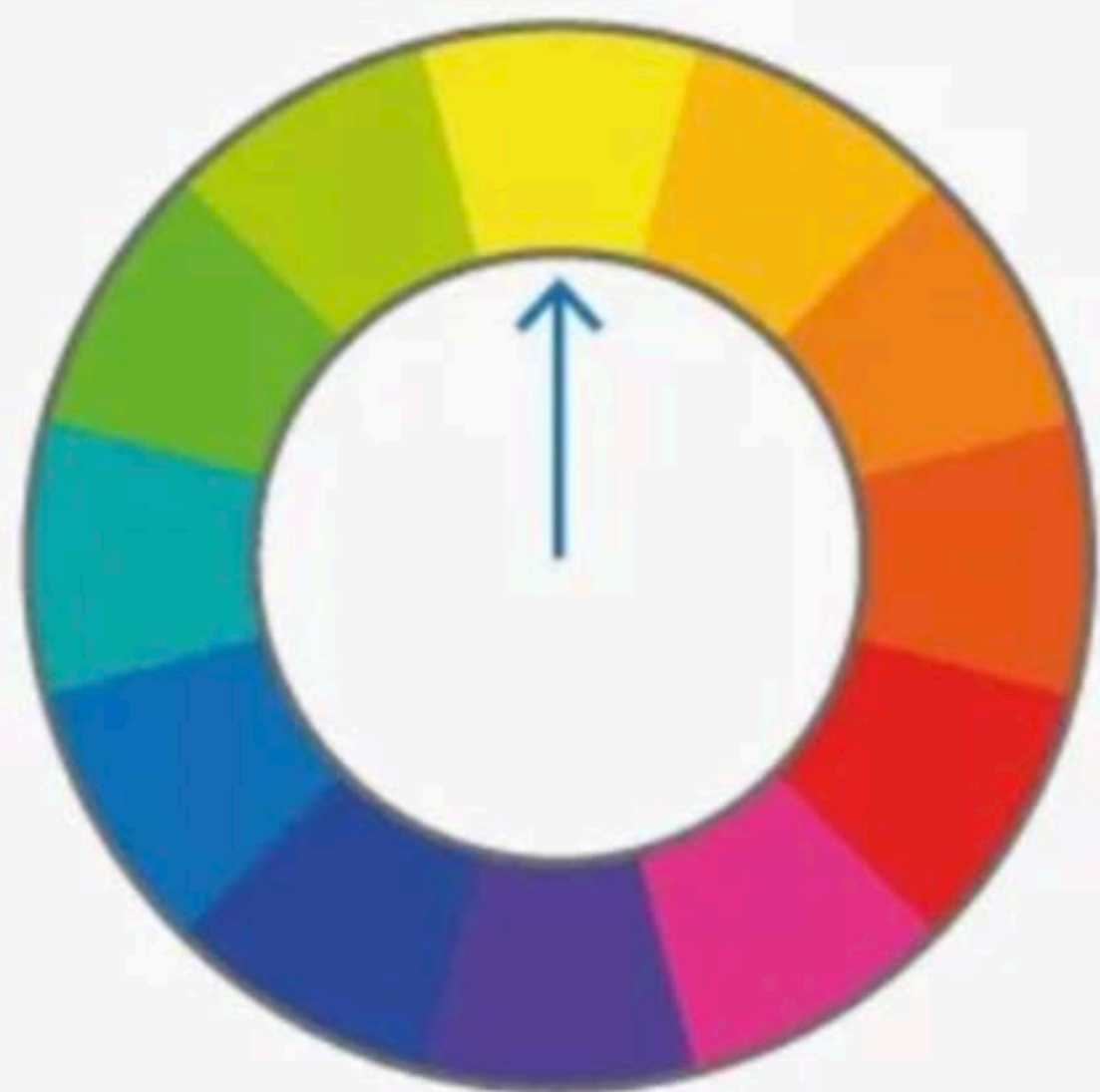
Color Cultural Associations
Ex: Red & Green

Color Cultural Associations
Ex: Red & Green
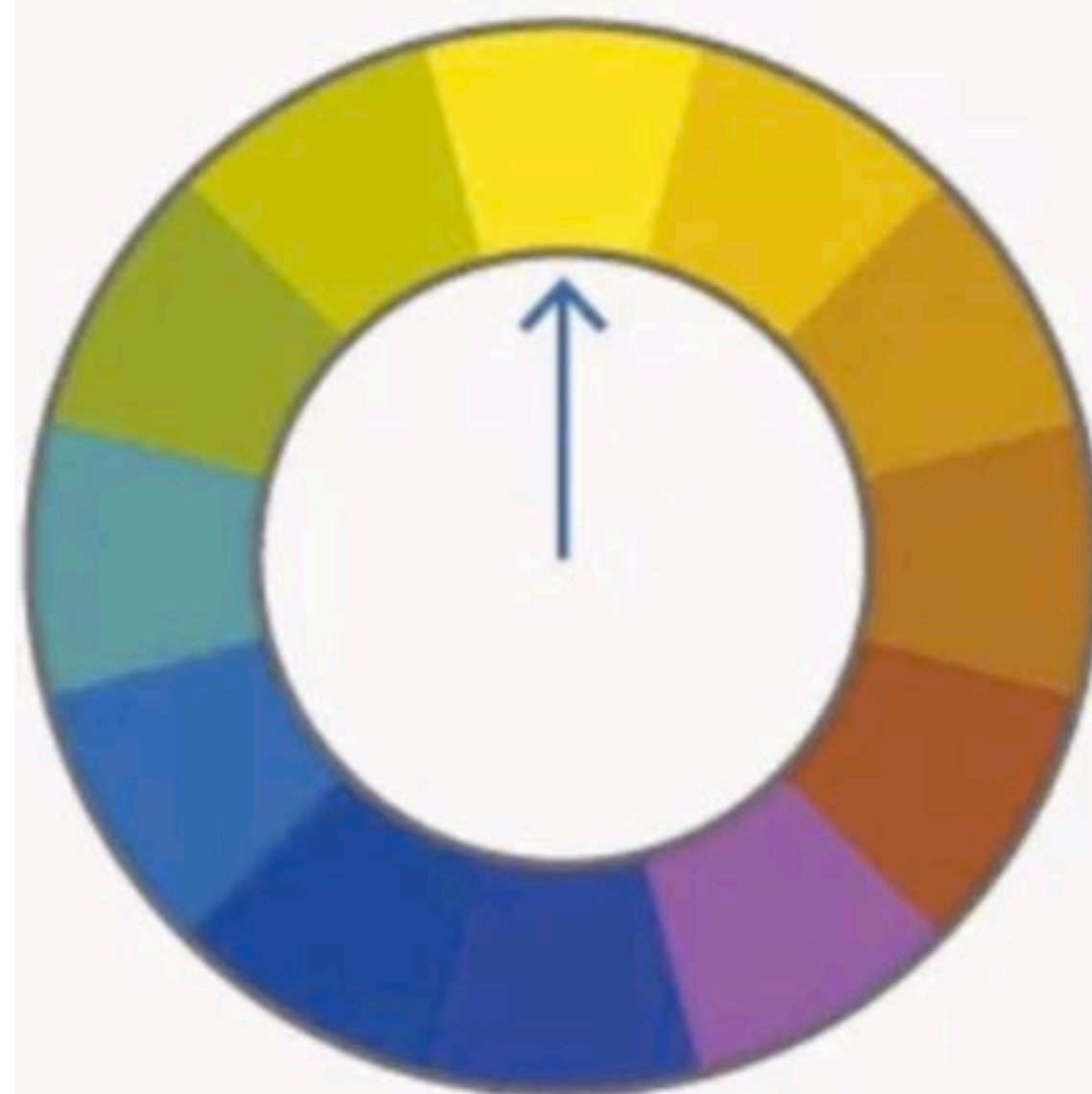
Designing for the Visually Impaired
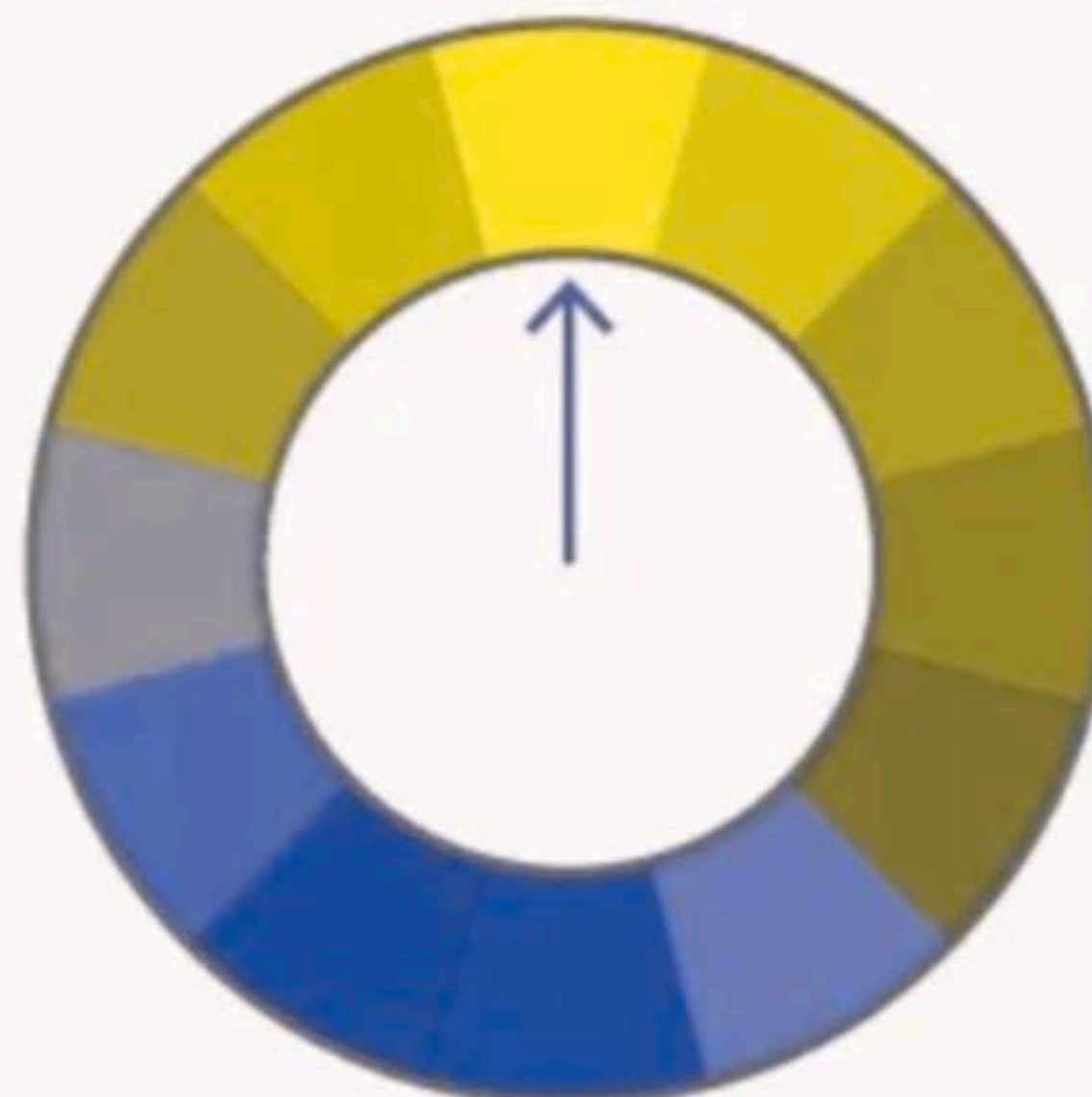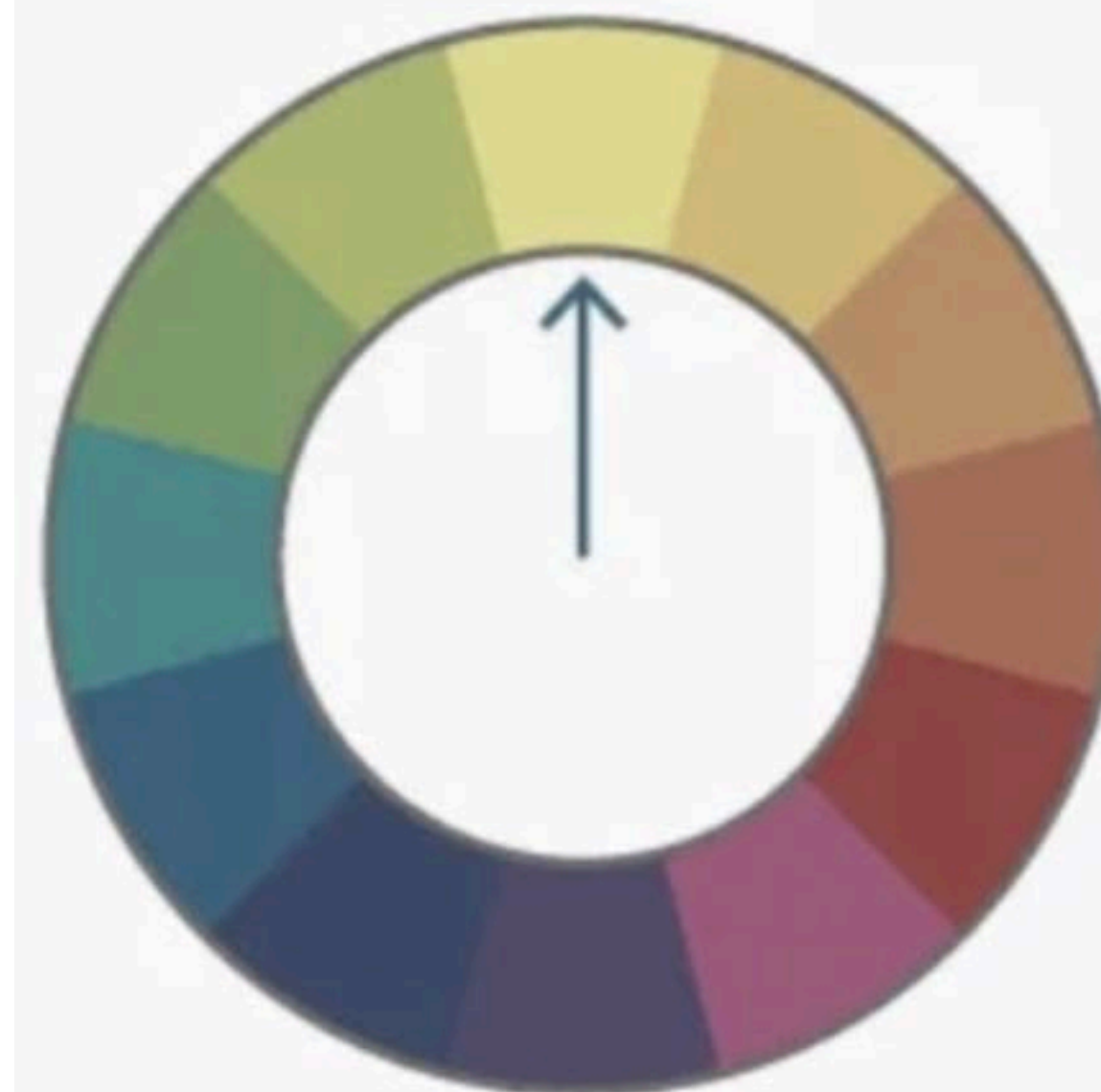
contemporary perspectives: interactive visualization paradigms
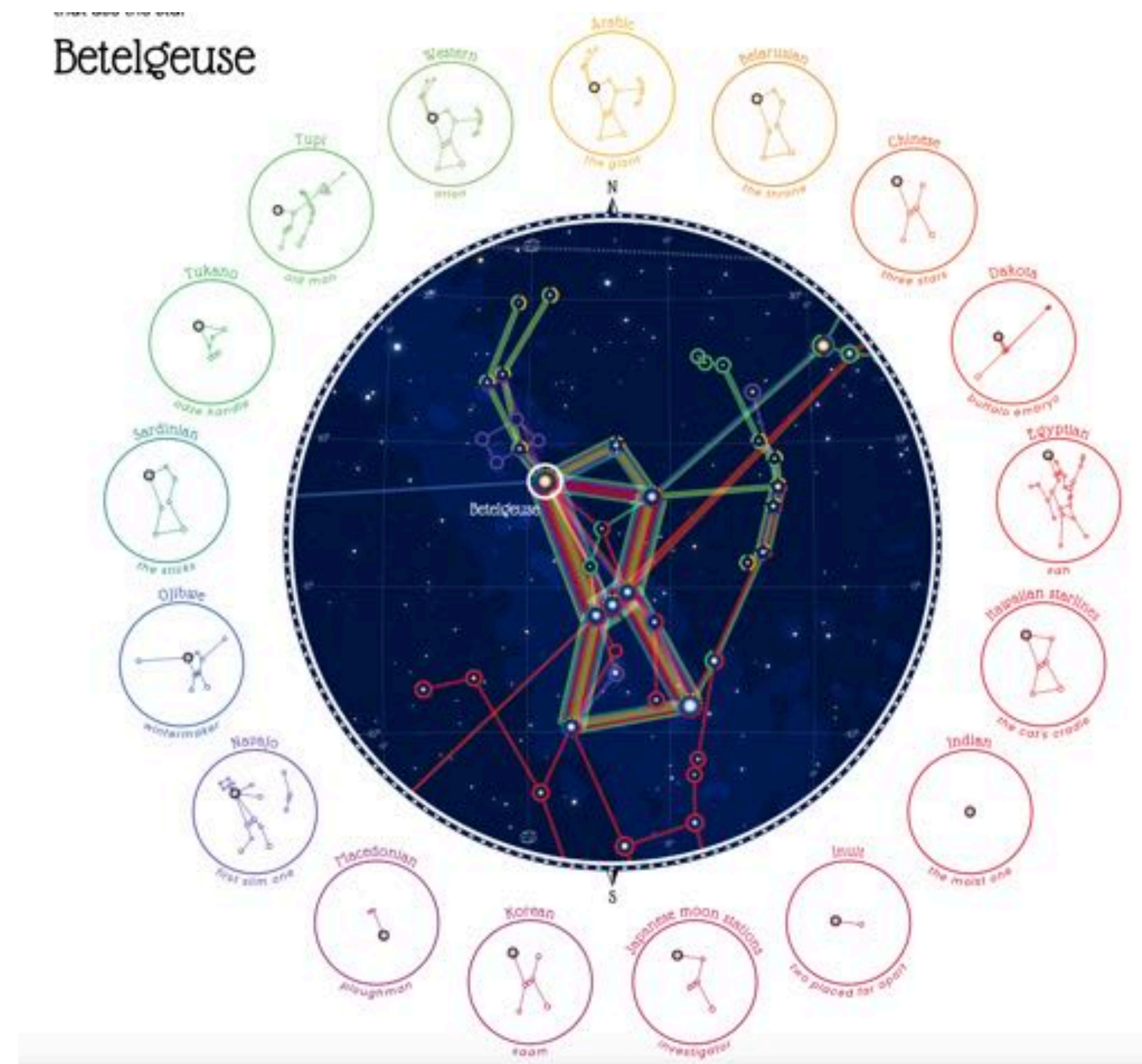
Form and Function
Whoever said pleasure wasn't functional? - Charles Eames
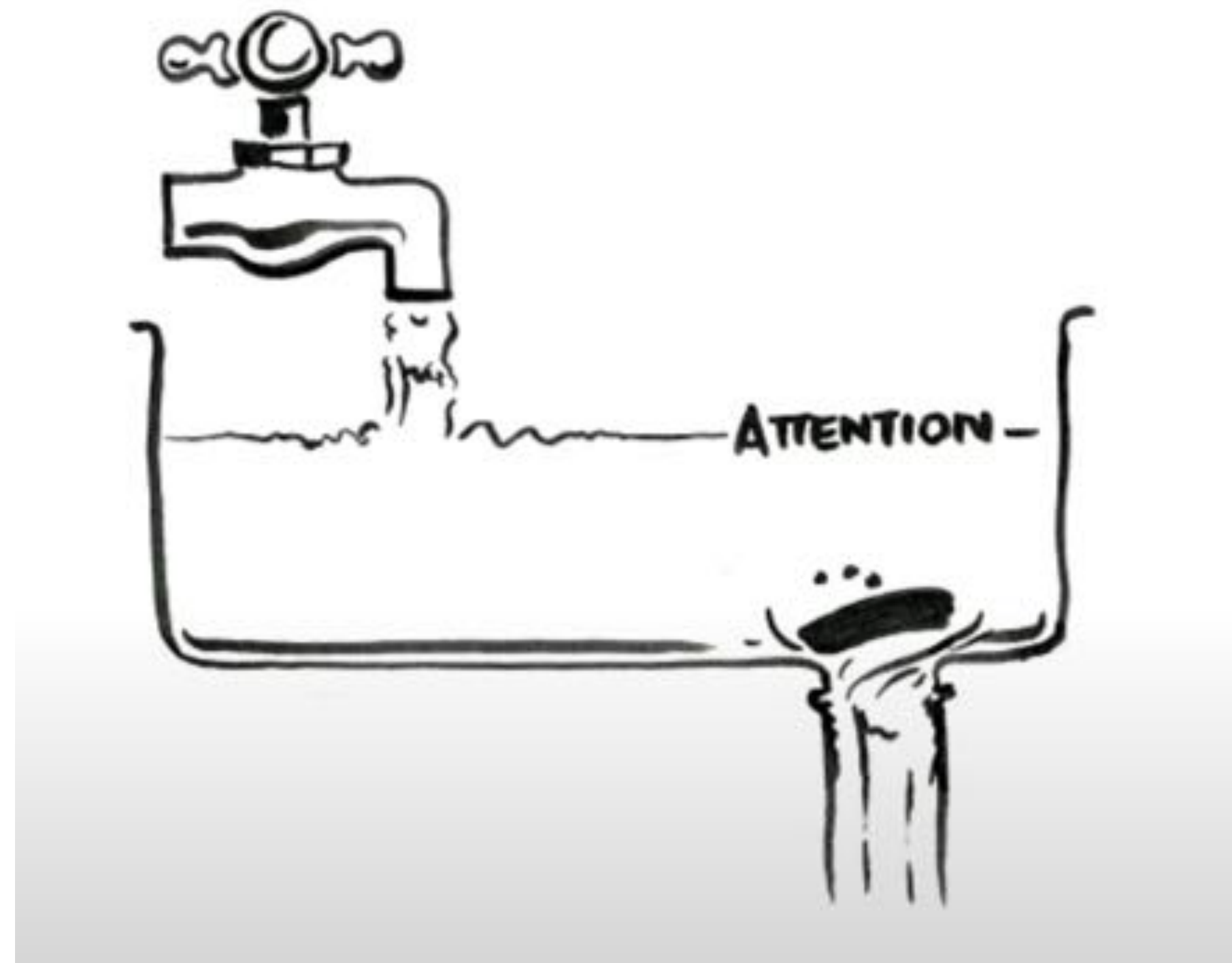

Attention and invitation to explore, beauty and fun

the importance of attention
Modern perspectives emphasize attention and structure

Headline and Visual Hierarchy

And you will read this last

You will read this first

And you will read this

Then this one

Invitations
Invite the viewer to explore the data

Point and Click
Adventures with data
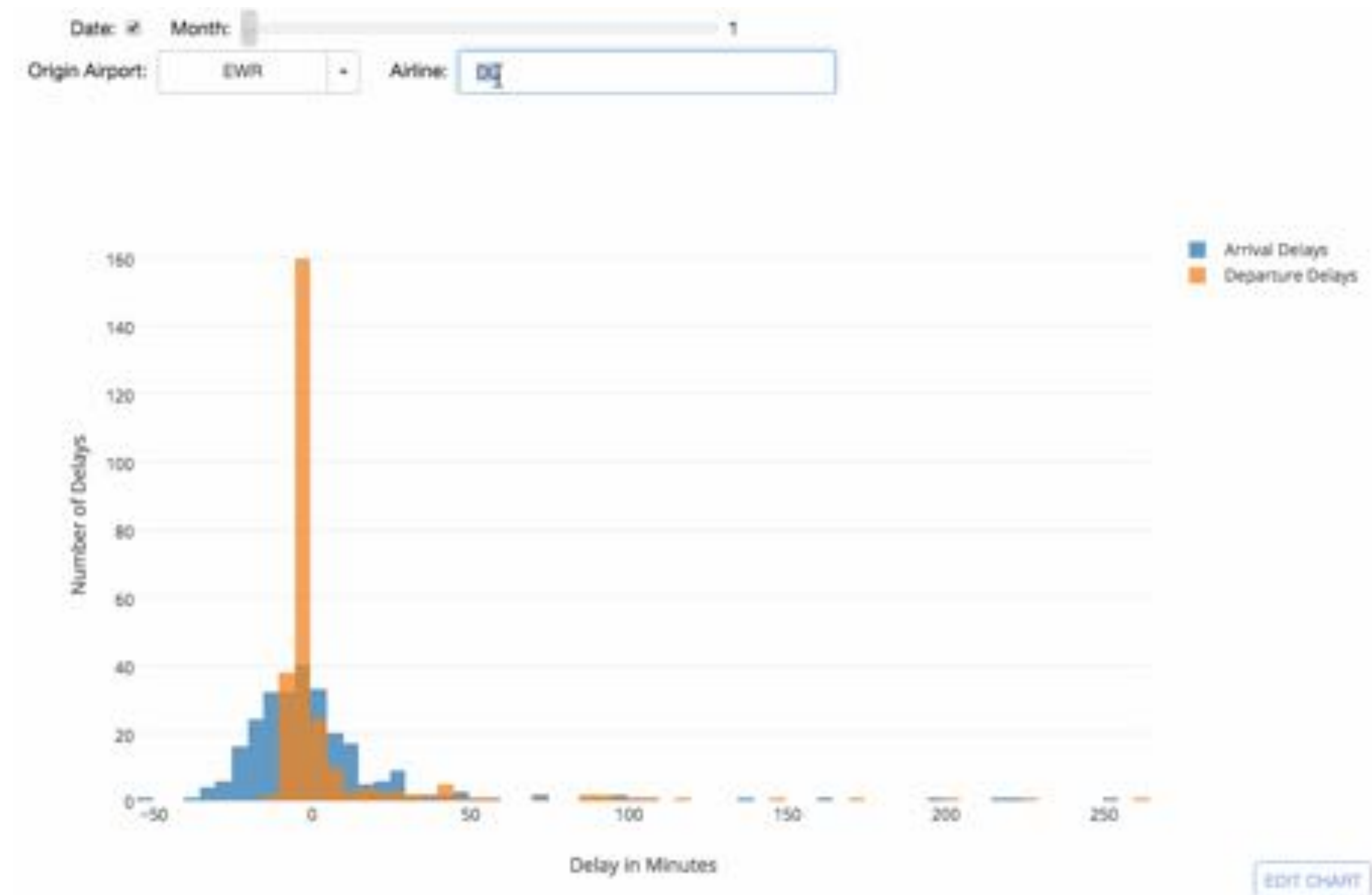
Drag to select
Linking two different views of the data

Variant: Drag and drop
for example, if I want to interactively create new plot windows.
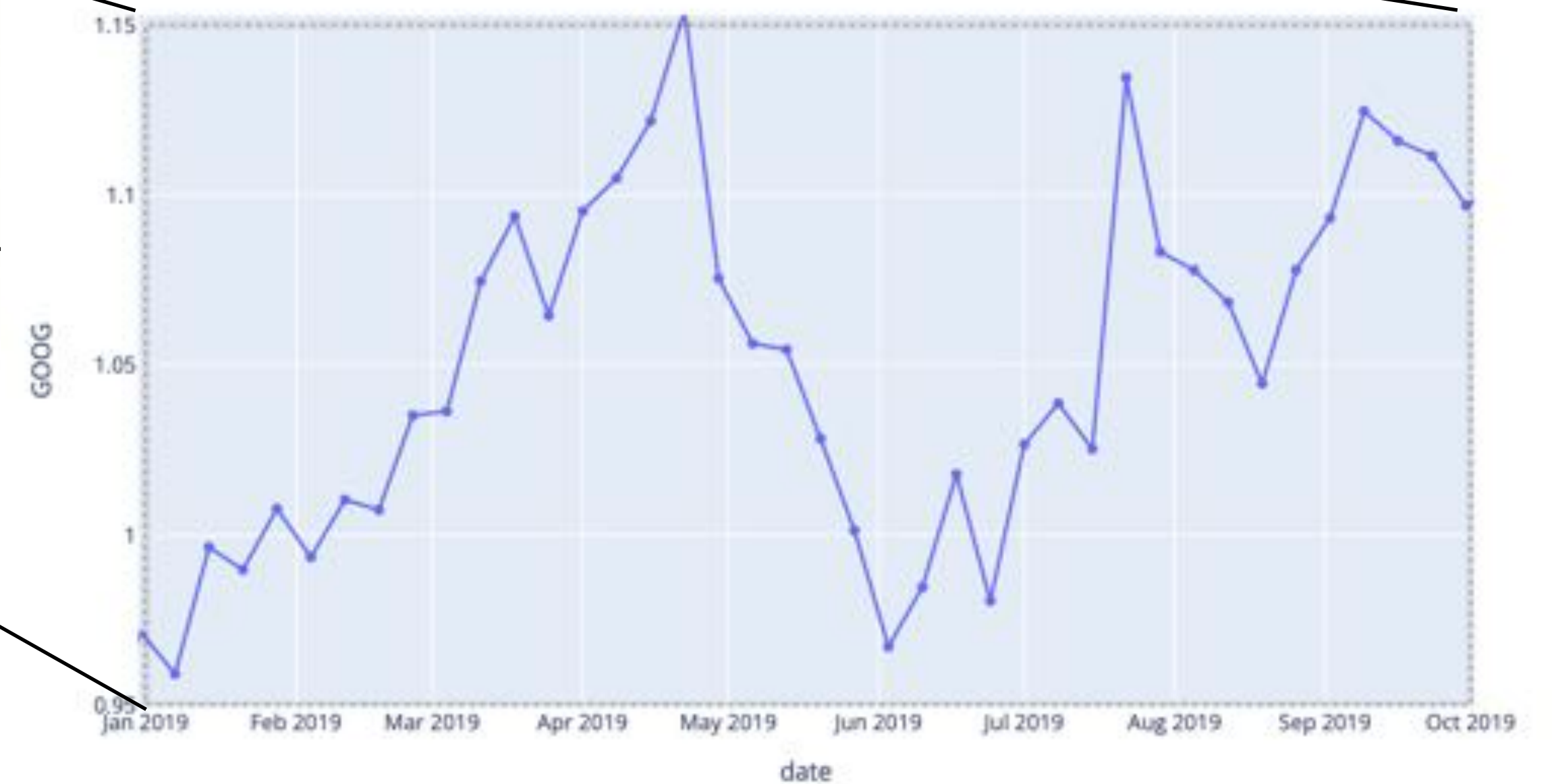
Animation
Tony Chu's Visual Introduction to Machine Learning

Smooth Transitions

Real Time Animation

Extra Visual Channels
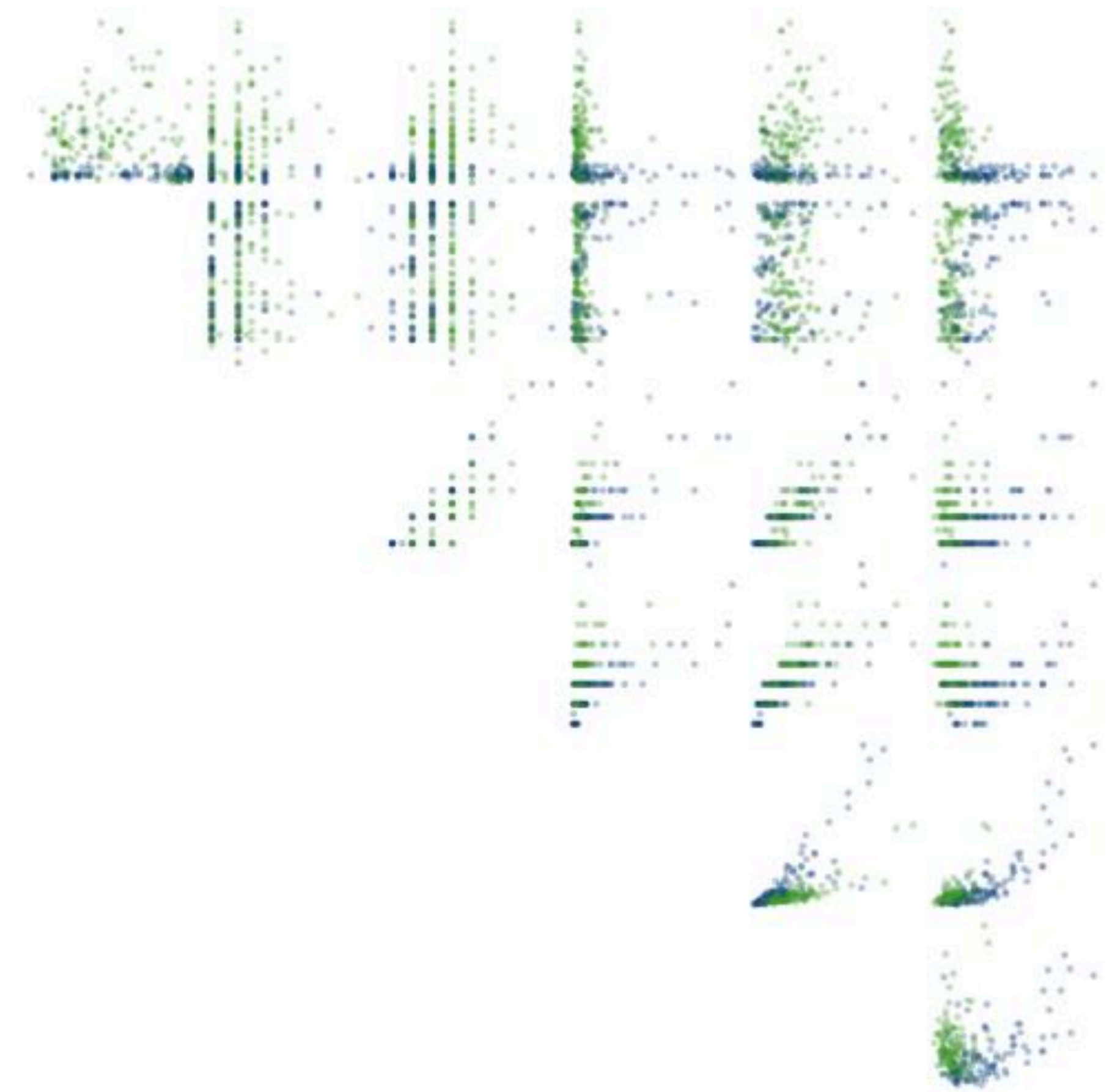
Q: Is this Chart Junk?

Q: Is this Chart Junk?

This doesn't violate the 'chart junk' rules because it increases data density and adds context.

For example, by showcasing change by including comparison plots - time series, histograms, etc and working with our knowledge of human attention and concentration.

Data Sketches Exercise

"If you can't make your design work on paper, you're never going to make it work on a computer with actual data."
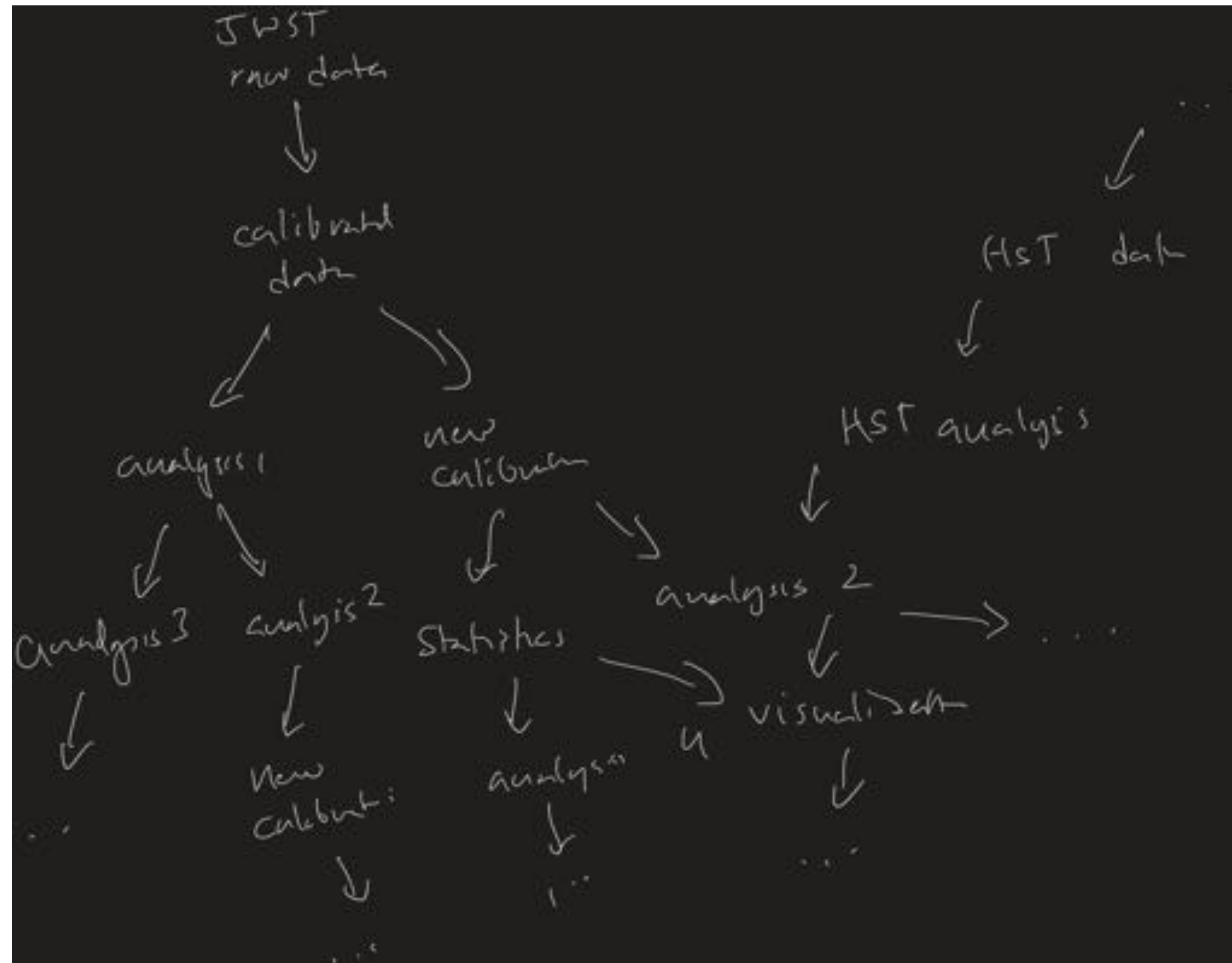
Exploratory Data Analysis:

Questions to explore creates focus when working with huge datasets generate

Interactive visualizations and test hypothesis

Sketches help especially with multipart designs

Hierarchical visualization for JWST
A data sketch example

Hierarchical visualization for JWST

A data sketch example

In the Session 17/Day 1 repository, find the dsfp_session17_data_sketches.ipynb tutorial notebook.

This will walk through producing a data sketch. This notebook also links to some example datasets that may be inspiration but feel free to sketch anything you'd like.

As you work through the notebook, think about how to employ these principles to enable informative, beautiful, and *insightful* visualizations in your research.