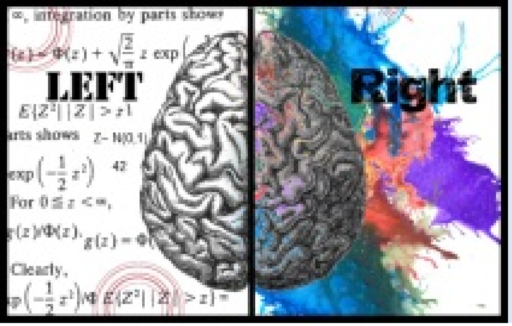


Visual Analytics: Where Art Meets Science

Tutorial on Data Visualization

Disclosures, Affiliations, and Acknowledgements

Contributors to the ideas presented today include:

- Abel Rodriguez, Professor in statistics at UCLA
- Tamara Munzner, Professor in information visualization at University of British Columbia
- Zhiheng Xu, reviewer at the FDA, CDRH

Disclosures

- I am a full time employee for Eli Lilly & Company.

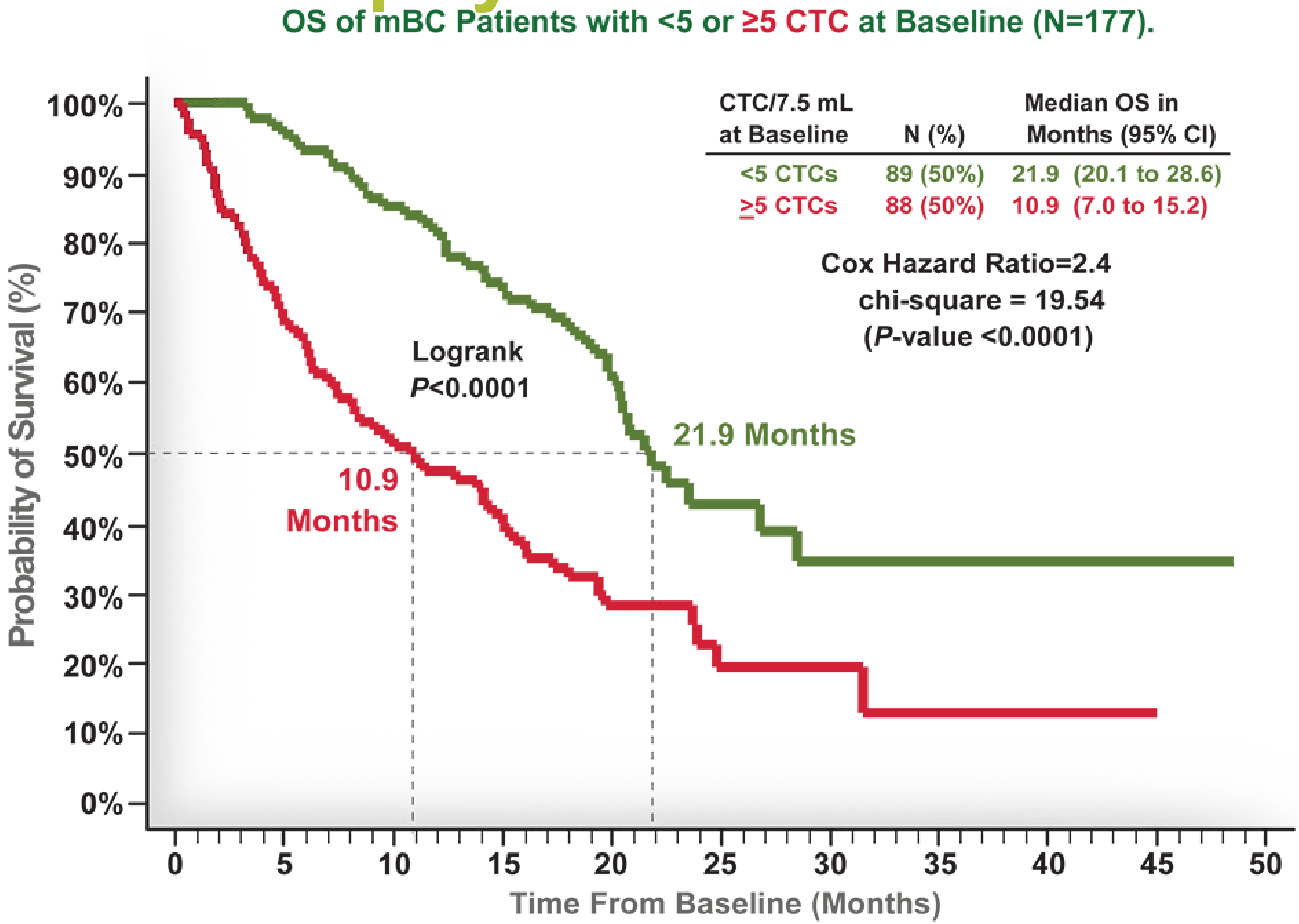
AADS
Advanced Analytics and Data Sciences

Educational Objectives

After completing this presentation you will be able to explain:

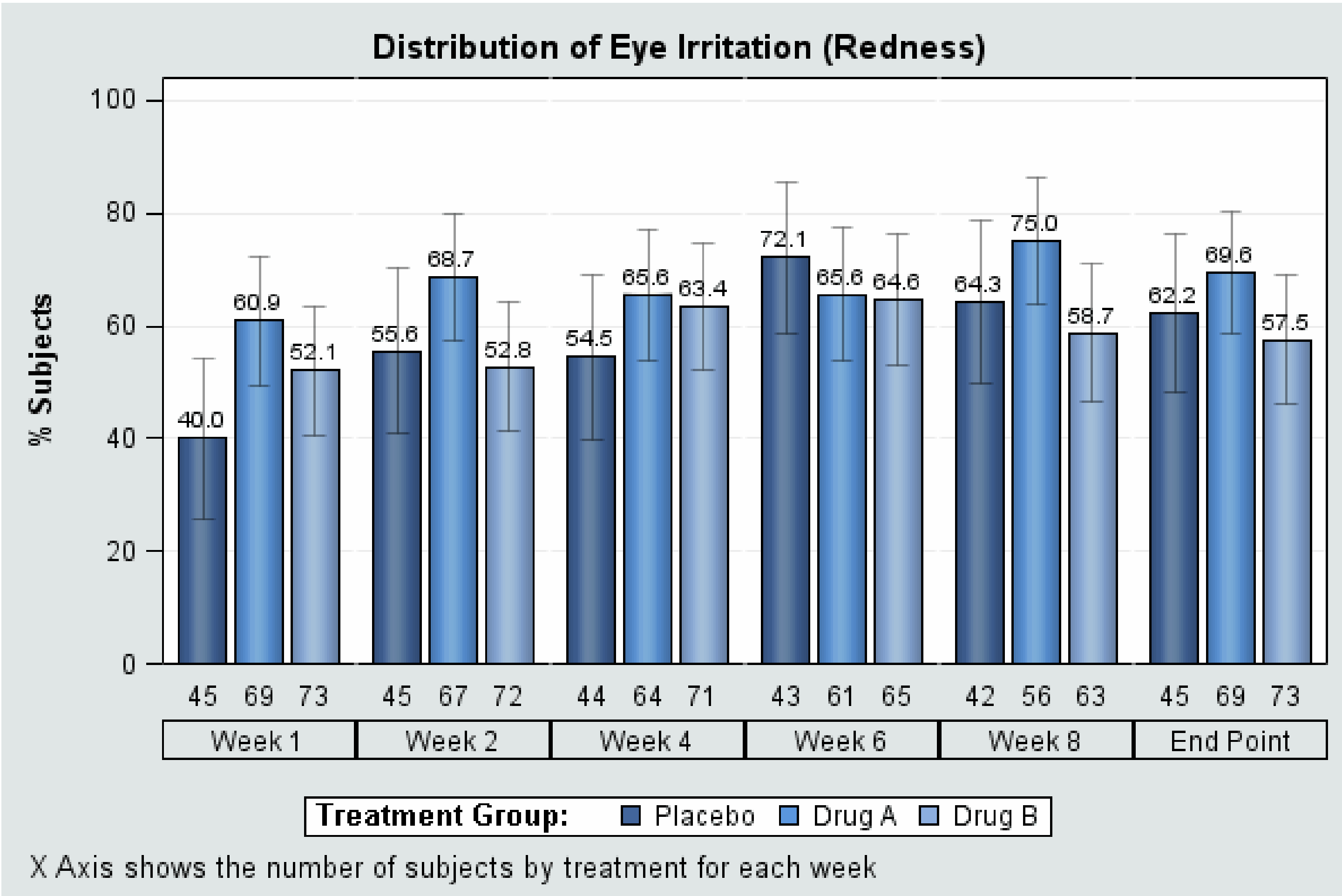
- Why is it important to visualize data?
- What does minimizing the ink to data ratio mean?
- What is Visual Analytics and how can it help in drug development?

Visual Display of Clinical Data



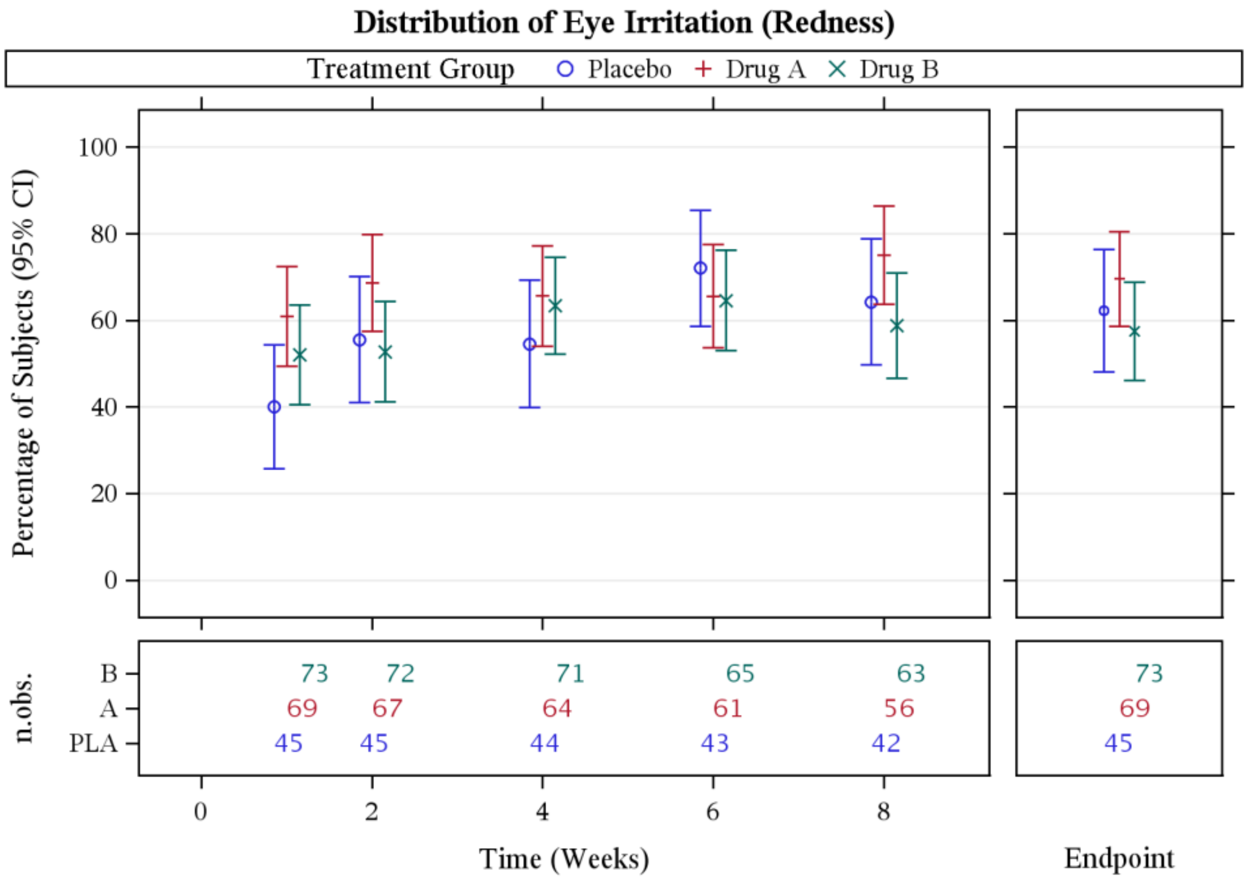
AADS
Advanced Analytics and Data Sciences

Poor Display of Clinical Data



AADS
Advanced Analytics and Data Sciences

Good Display of Clinical Data



n.obs = Number of Observations at Time point in Treatment group

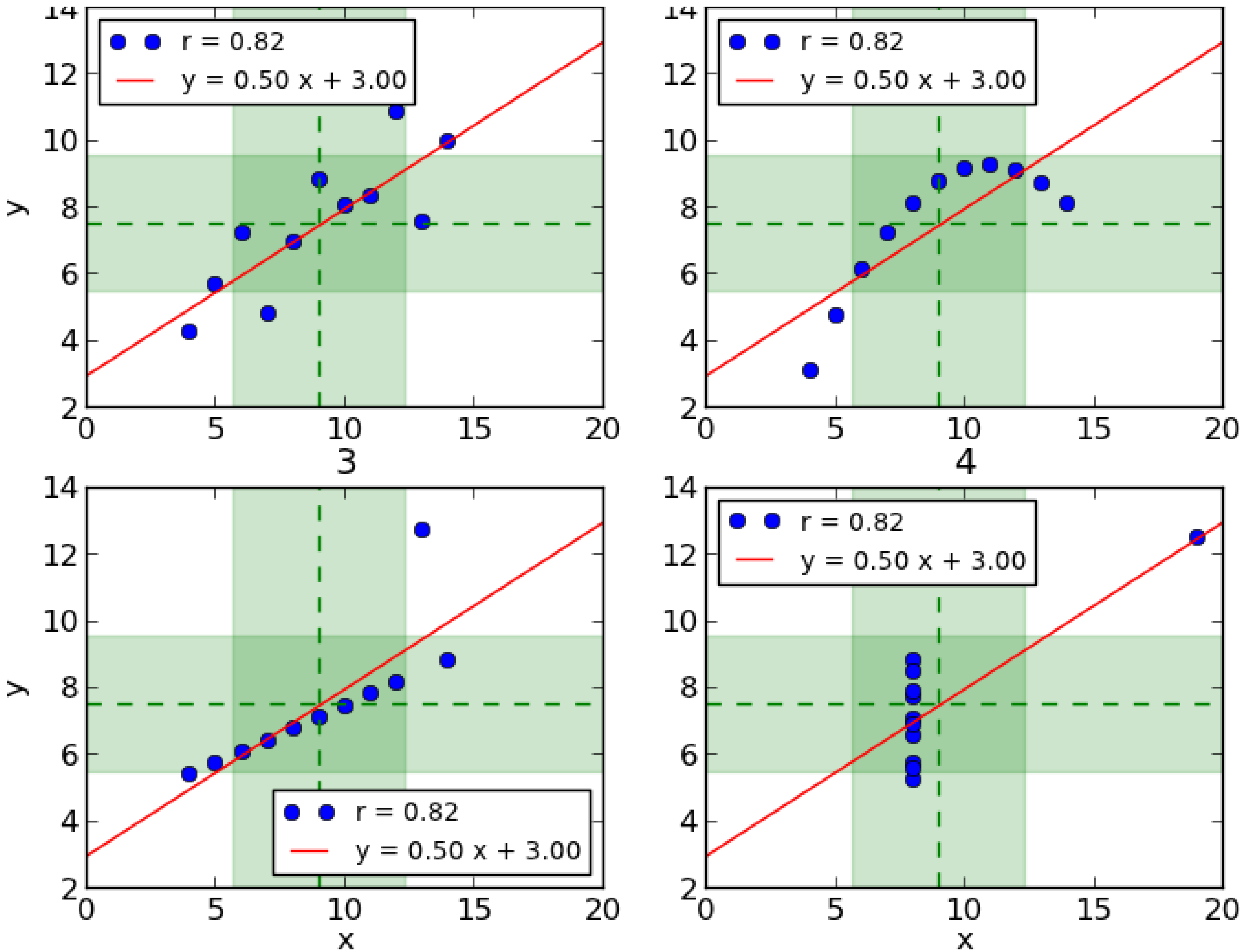
AADS
Advanced Analytics and Data Sciences

Data Visualization

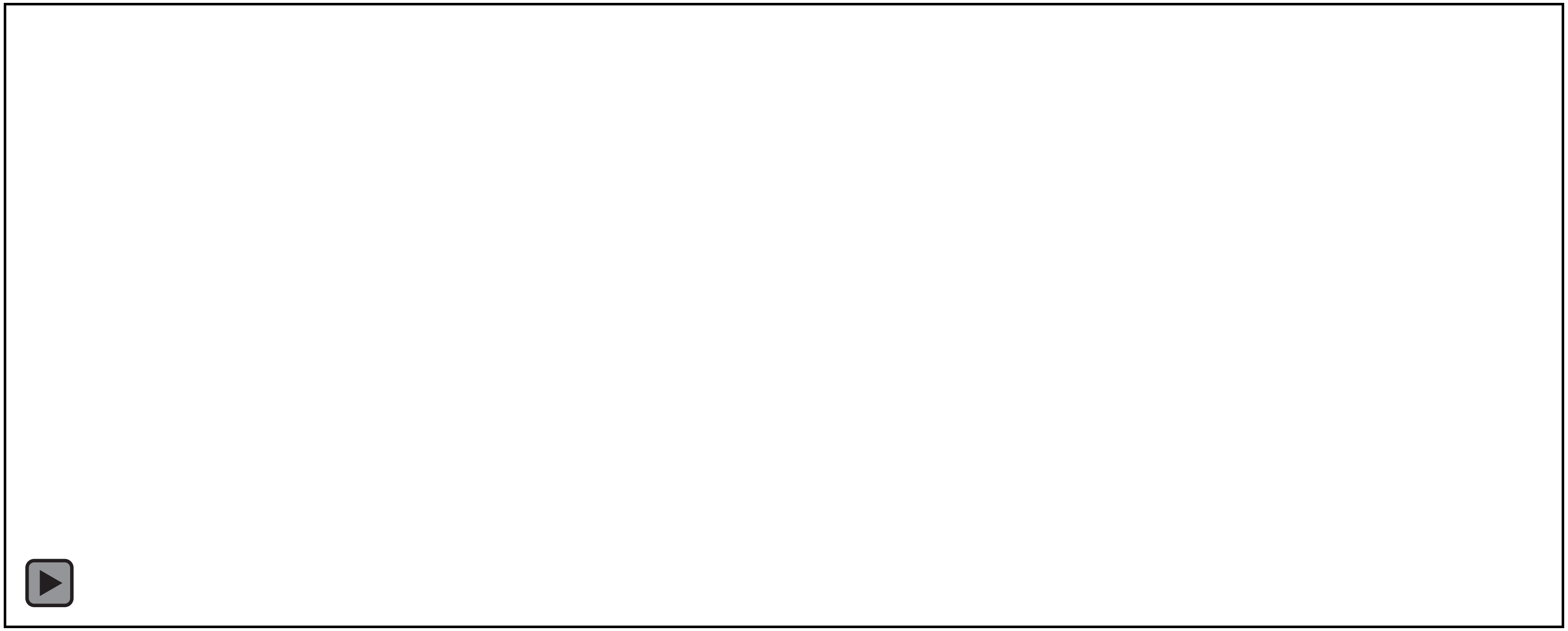
MOTIVATION

AADS
Advanced Analytics and Data Sciences

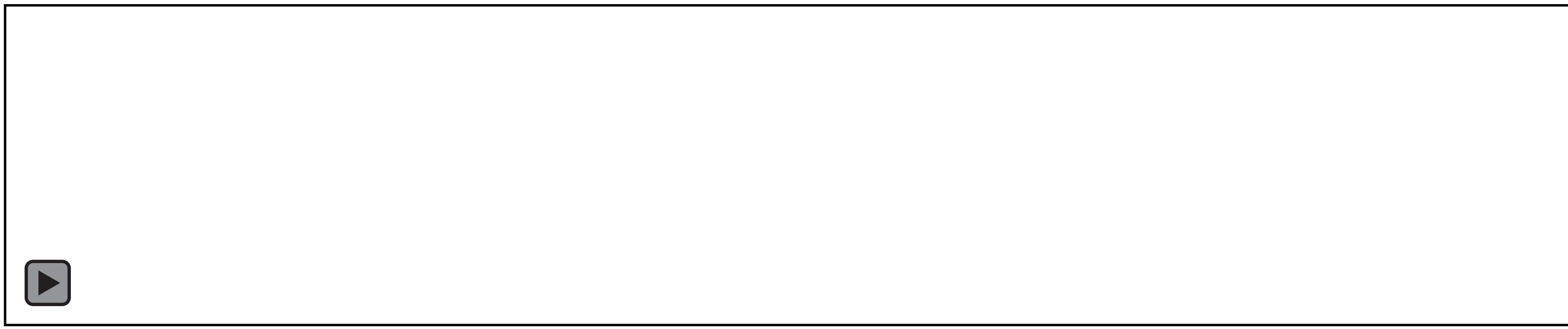
Anscombe's Quartet



2 dimensional vis vs table



1 dimensional: histograms



1 dimensional: violin plots

Are there situations where a table is better than a graph?

- Yes, but these are relative exceptions.
  - To convey a handful of numbers.
  - To report precise values for lookup.
  - To present many different types quantities (dimensions) for a small number of cases.
- Tables are usually a bad idea if comparison is important.
- I will not discuss principles of table design, but they are similar to the ones for visualization!

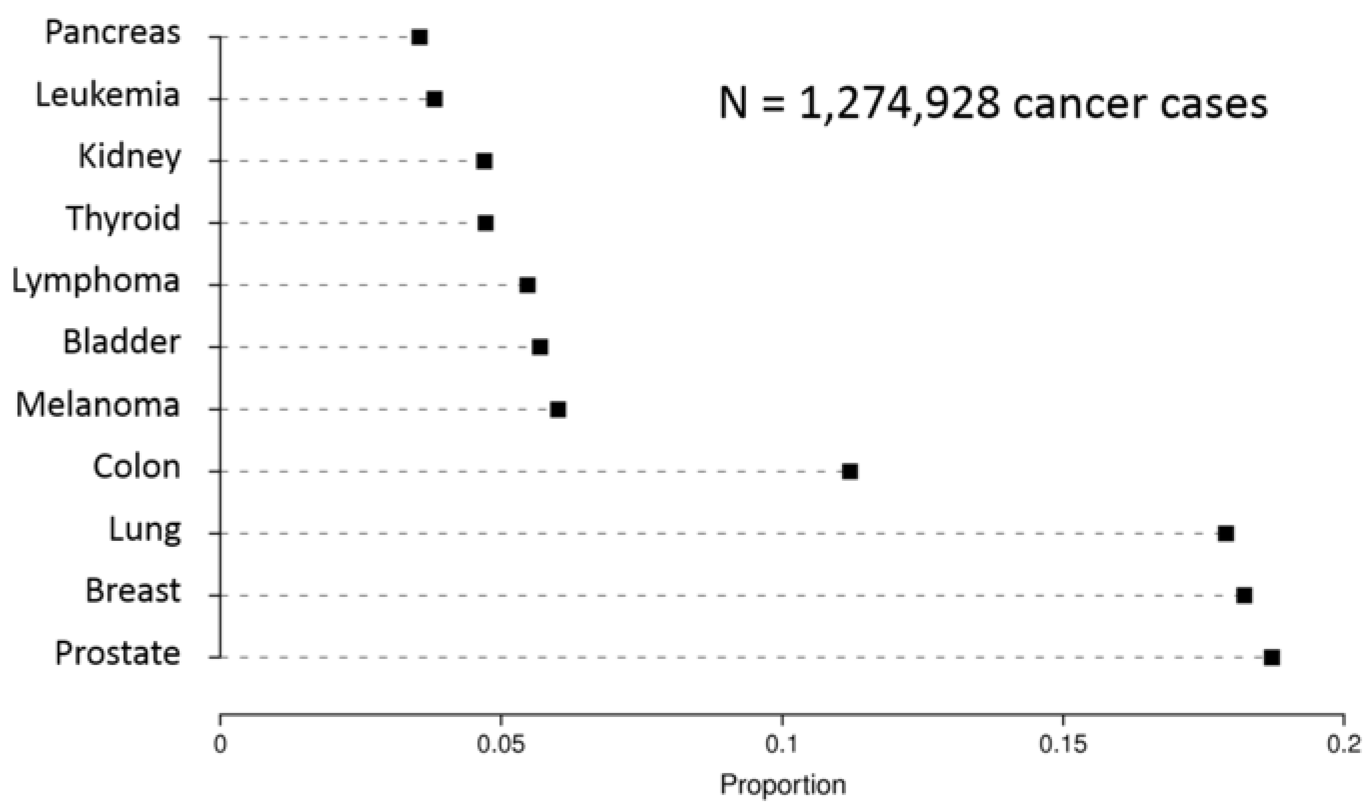
AADS
Advanced Analytics and Data Sciences

Table: Cancer Incidence by Type

Type	Incidence	Proportion
Prostate	238,590	18.7%
Breast	232,340	18.2%
Lung	228,190	17.9%
Colon	142,820	11.2%
Melanoma	76,690	6.0%
Bladder	72,570	5.7%
Lymphoma	69,740	5.5%
Thyroid	60,220	4.7%
Kidney	59,938	4.7%
Leukemia	48,610	3.8%
Pancreas	45,220	3.5%

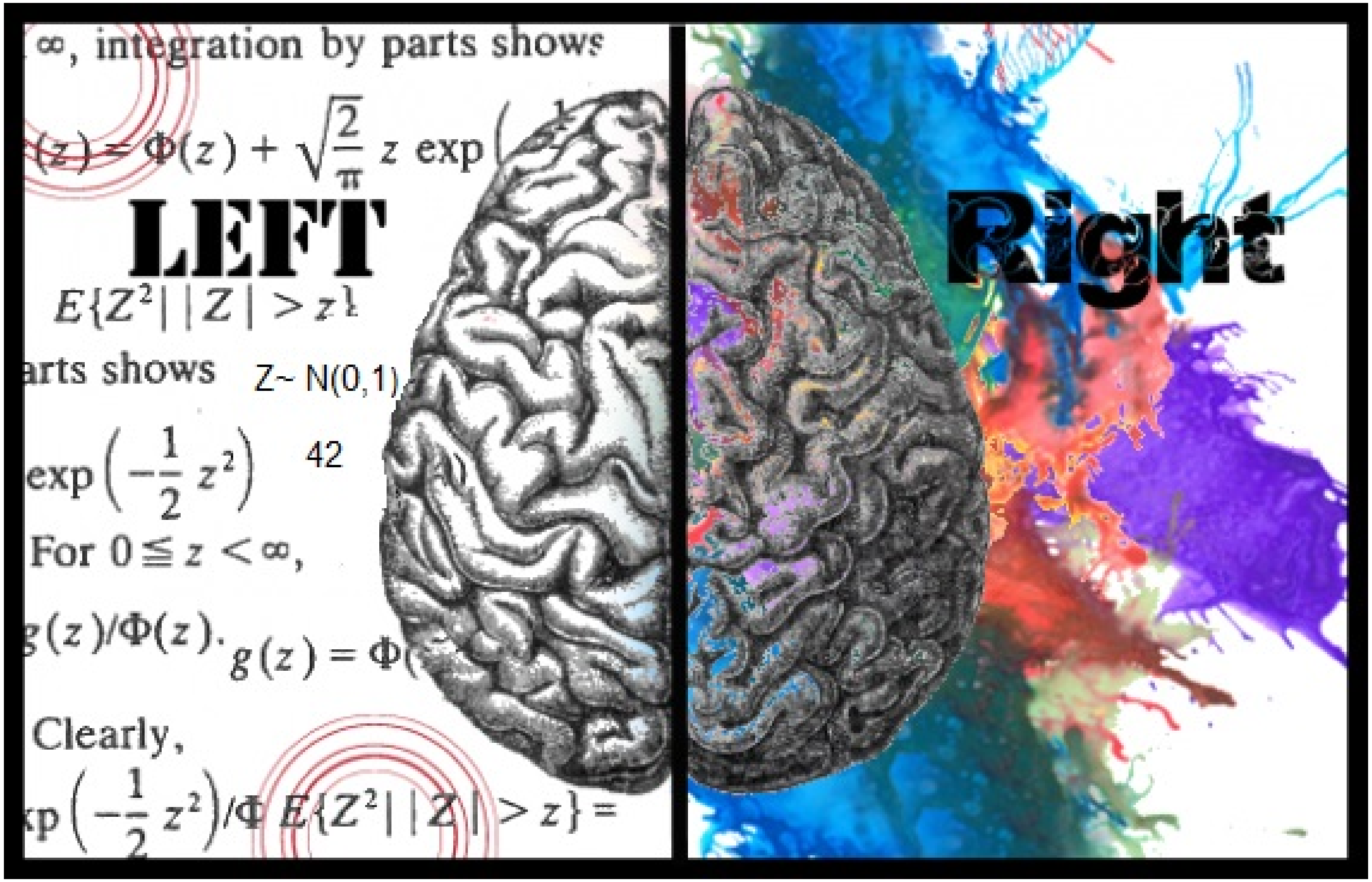
Data from http://www.cancer.gov/cancertopics/types/commoncancers

Graph: Cancer Incidence by Type



Data from http://www.cancer.gov/cancertopics/types/commoncancers

Visual Analytics: Where Art Meets Science



Exploration vs Explanation

vs

Insights from multiple disciplines….

- Graphic design: Emphasizes aesthetics.
- Computer science: Emphasizes algorithms.
- Cognitive psychology: Provides insights into the most effective tools.
- Journalism: Emphasizes storytelling.
- Statistics: Emphasizes quantification of information.

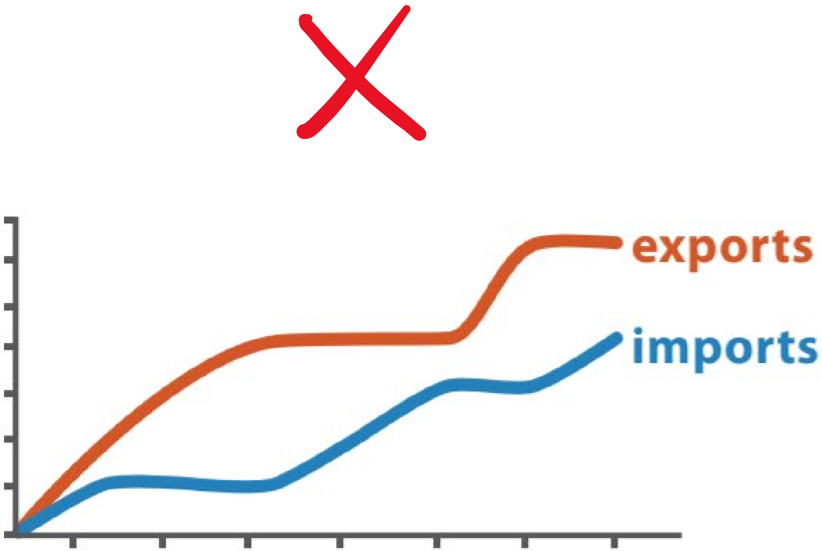
A (Very) Short History of Visualization

- 1637 – Descartes first uses 2D grids to visually encode numbers.
- 1786 – William Playfair's "The Commercial and Political Atlas".
- 1855 – John Snow uses maps to link the 1854 London cholera epidemic to contaminated drinking water.
- 1857 – Florence Nightingale uses stacked bar and pie charts to persuade Queen Victoria to improve conditions on British military hospitals.
- 1954 – Darrel Huff's "How to Lie with Statistics".
- 1977 – John Tukey introduces boxplots.
- 1983 – Edward Tufte's "Visual Displays of Quantitative Information".
- 1994 – William Cleveland's "The Elements of Graphing Data".
- 2004 – Stephen Few "Show me the Numbers".
- Nowadays dominated by computer scientists (on the technical side) and business analytics (on the more applied side).

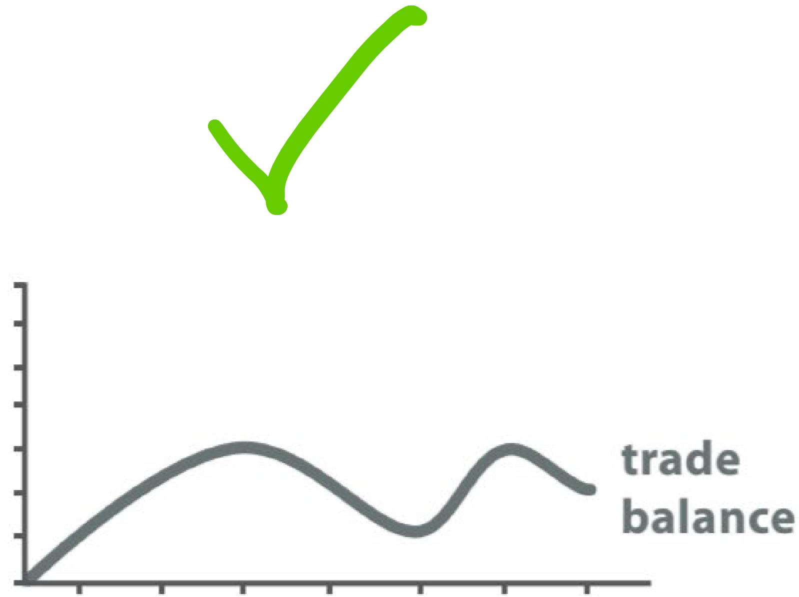
Data Visualization

BASIC PRINCIPLES

Objective Drives the Plot



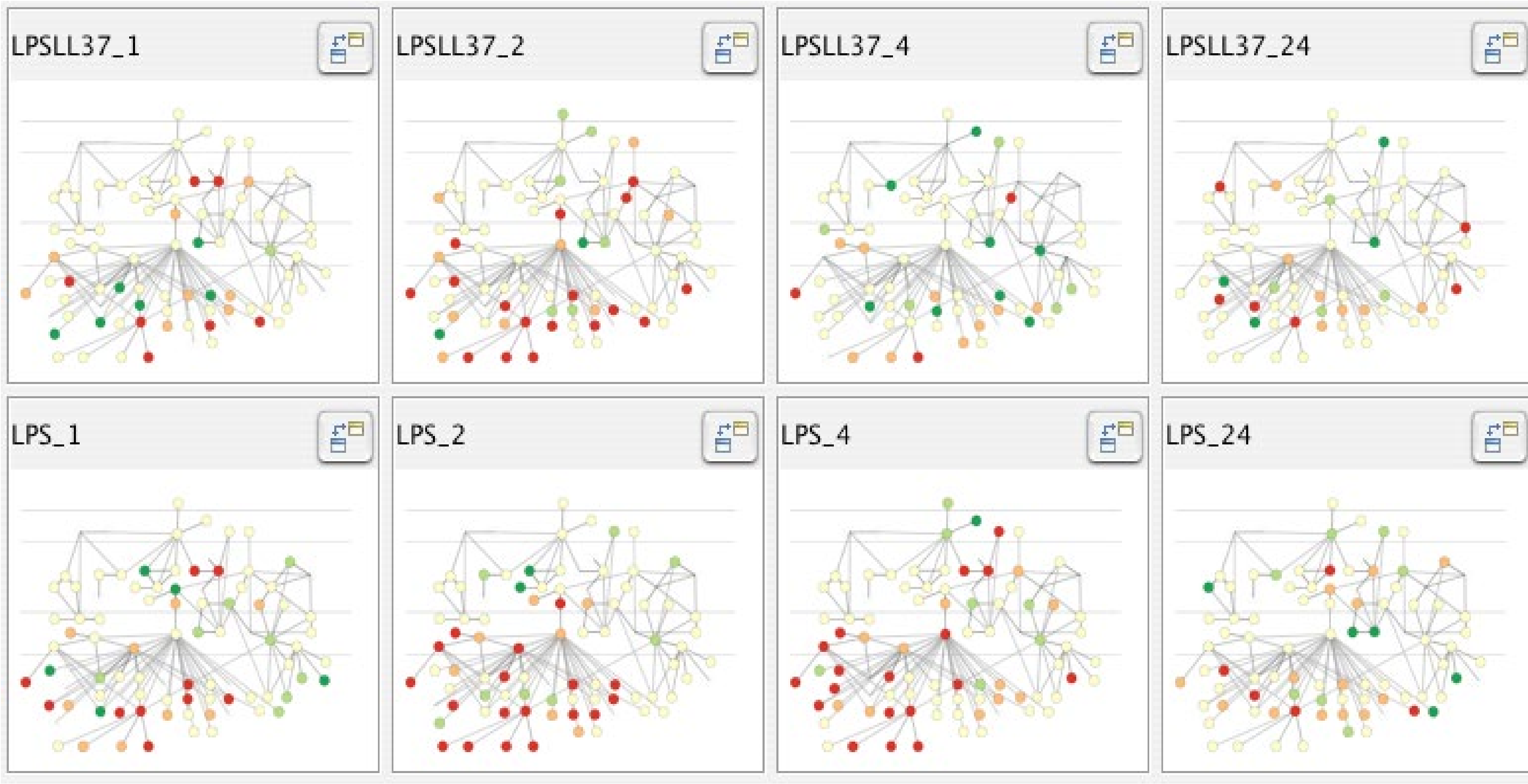
Original Data



trade balance = *exports* − *imports*

Derived Data

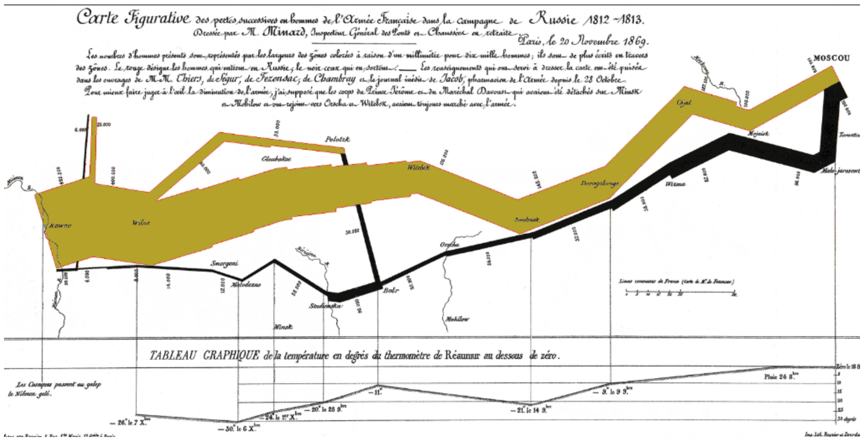
Small multiples: Eyes Beat Memory

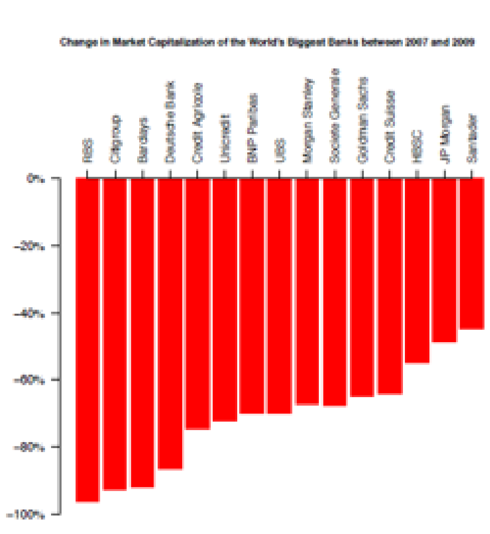


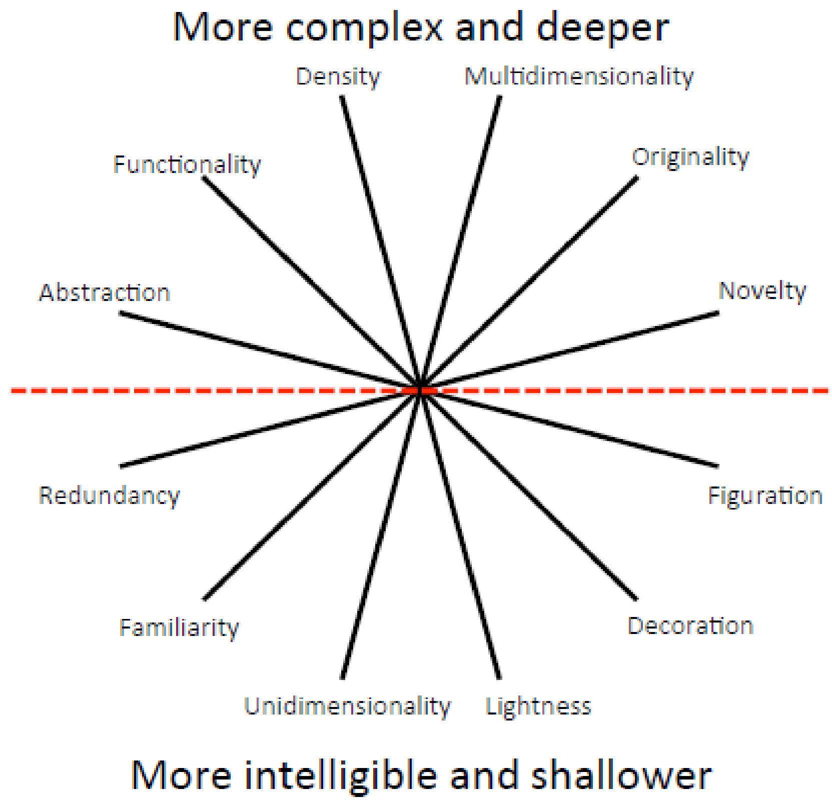
AADS

[Cerebral: Visualizing Multiple Experimental Conditions on a Graph with Biological Context. Barsky
Munzner, Gardy, and Kincaid. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis

Complexity vs comprehension



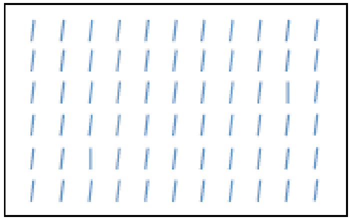


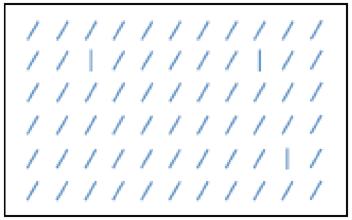


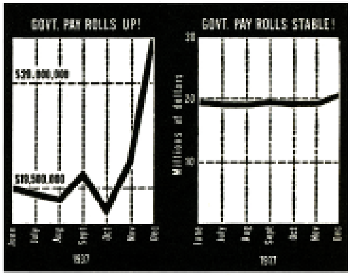
AADS

A few lessons from cognitive psychology…

- Attention is drawn to large perceptible differences: humans think in terms of differences.
- People expect changes in properties to carry information.
- Form and meaning need to be compatible.
- People can only hold in mind up to four groups of information at once.
- People automatically group elements into units.
- Try to maximize data/ink ratio.
- When possible, interactivity is your friend.





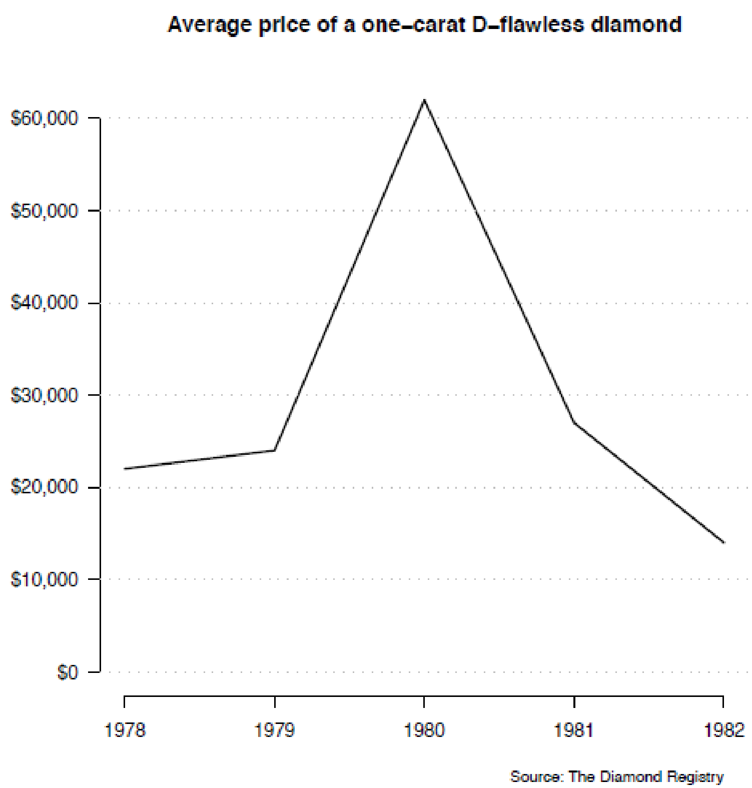


Blue   Red

[_]   _][

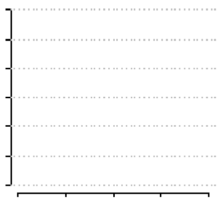
XXOO   XO   XK

The structure of visualizations

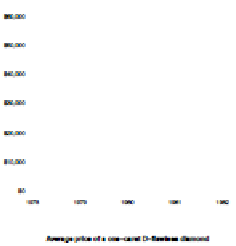


Visual cues

Coordinate system

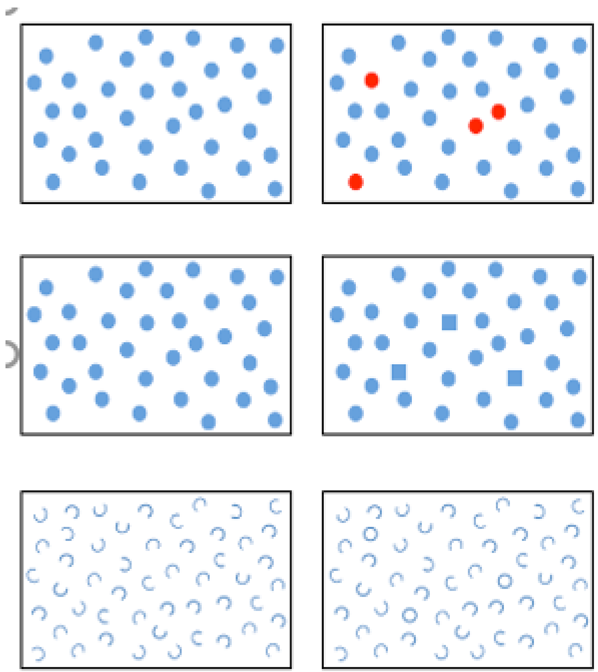


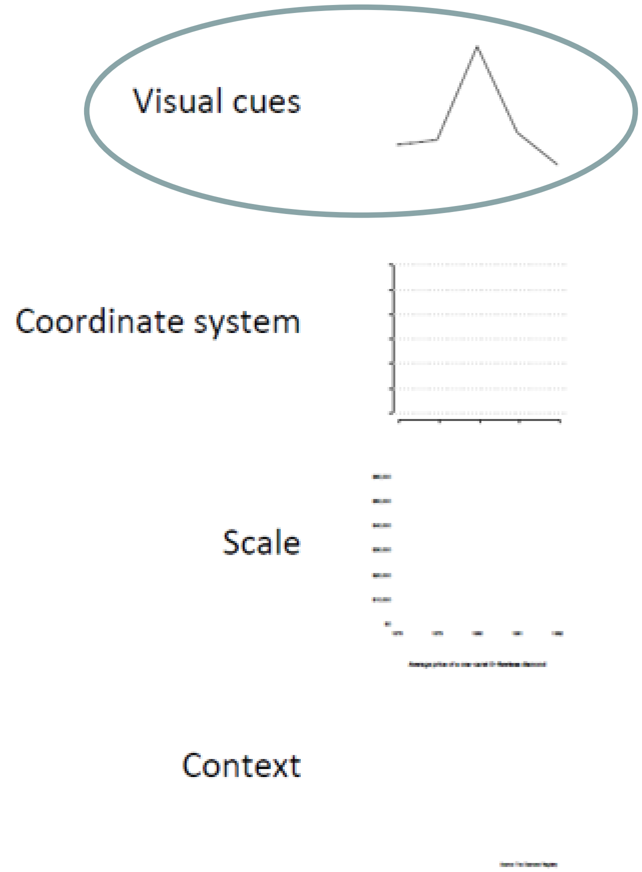
Scale



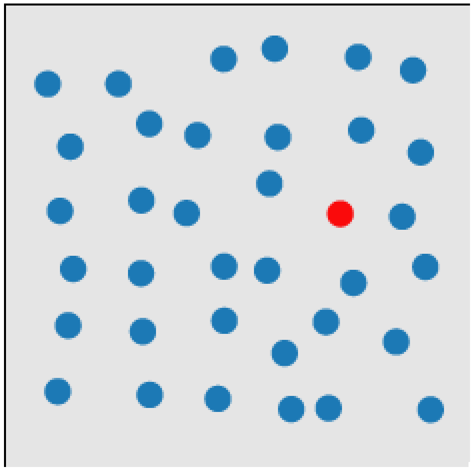
Context

Visual Cues & Pre-attentive processing

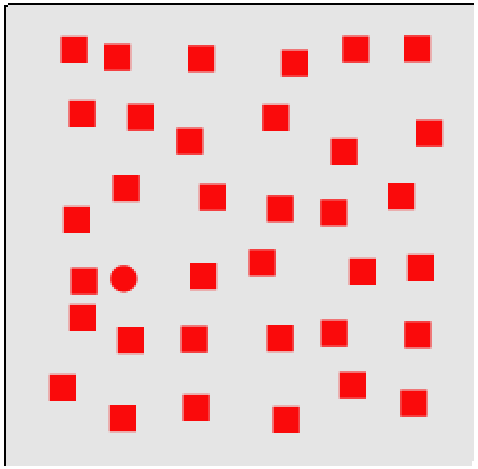




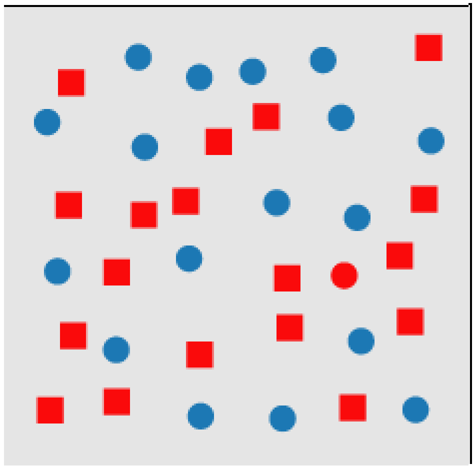
Target Selection Visual Cue: color



Target Selection Visual Cue: shape



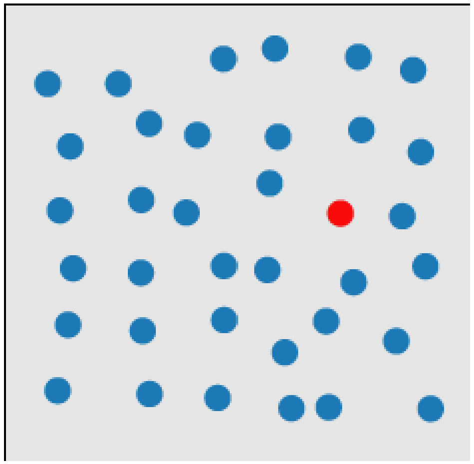
Target Selection Visual Cue: conjunction



Pre-attentive processing

EXPERIMENT INSTRUCTIONS

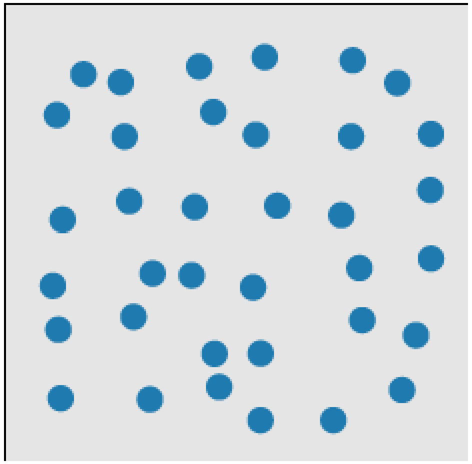
Target Selection Visual Cue: color



a) Anomaly present

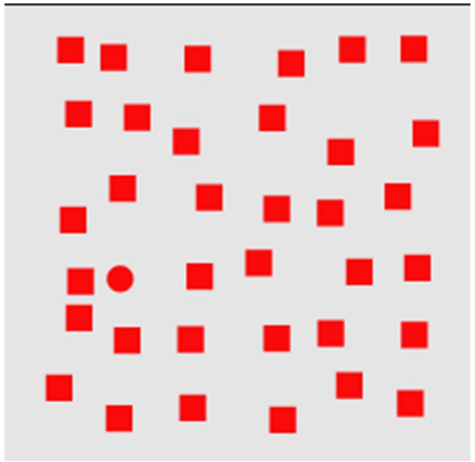
b) Anomaly absent

Target Selection Visual Cue: color



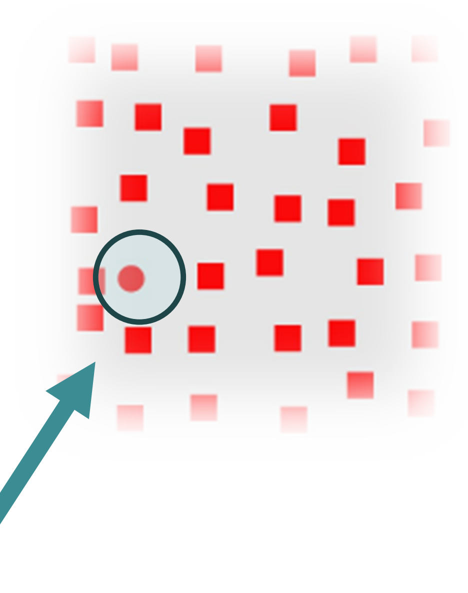
a) Anomaly present
b) Anomaly absent

Target Selection Visual Cue: shape



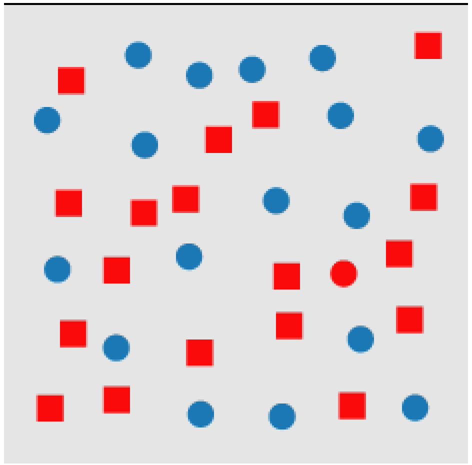
a) Anomaly present
b) Anomaly absent

Target Selection Visual Cue: shape



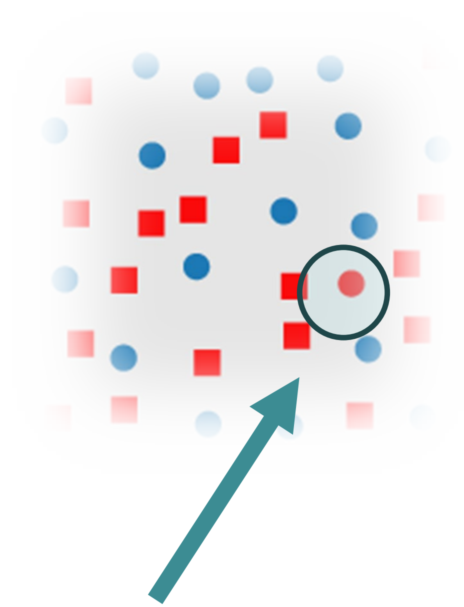
a) Anomaly present
b) Anomaly absent

Target Selection Visual Cue: conjunction



a) Anomaly present
b) Anomaly absent

Target Selection Visual Cue: conjunction



a) Anomaly present

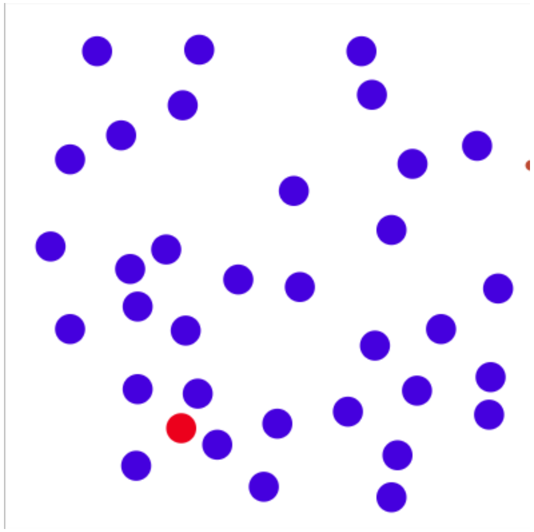
b) Anomaly absent

AADS

Pre-attentive processing

EXPERIMENT BEGINS

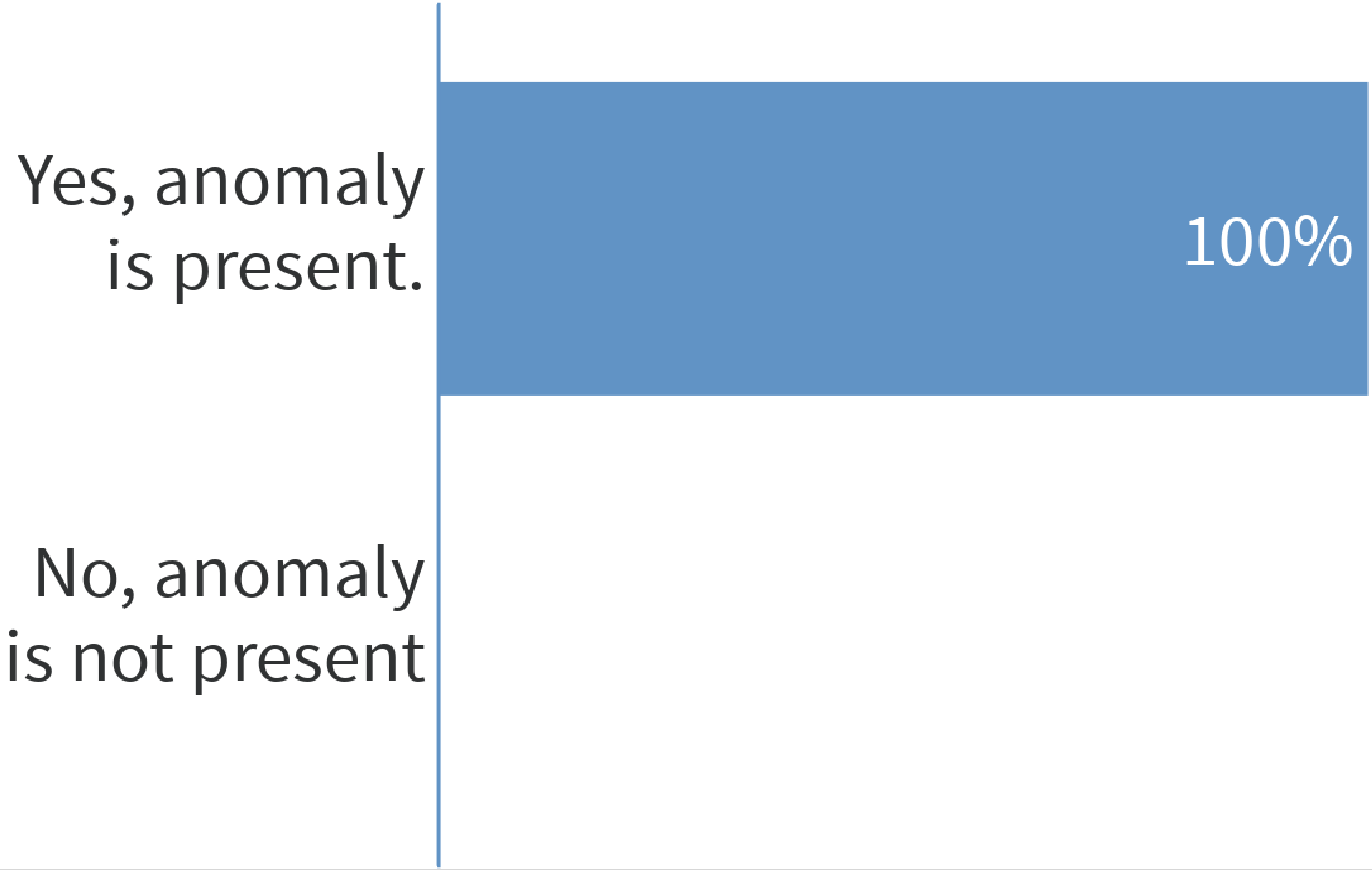
Experiment #1: color



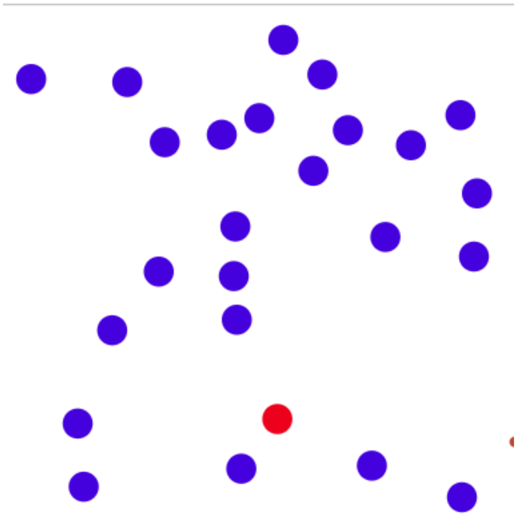
a) Anomaly present
b) Anomaly absent

AADS

Is an anomaly present?

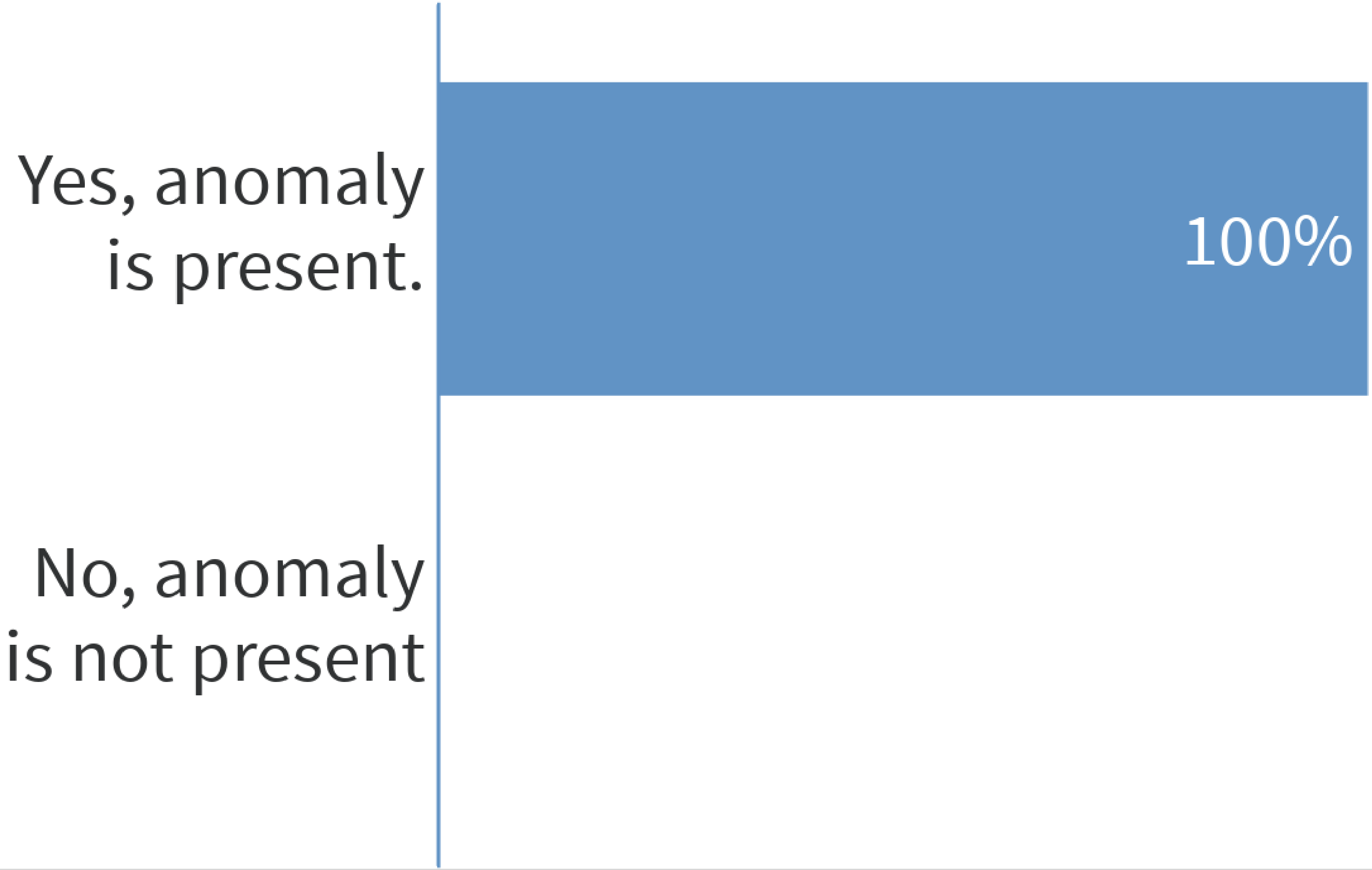


Experiment #2: color

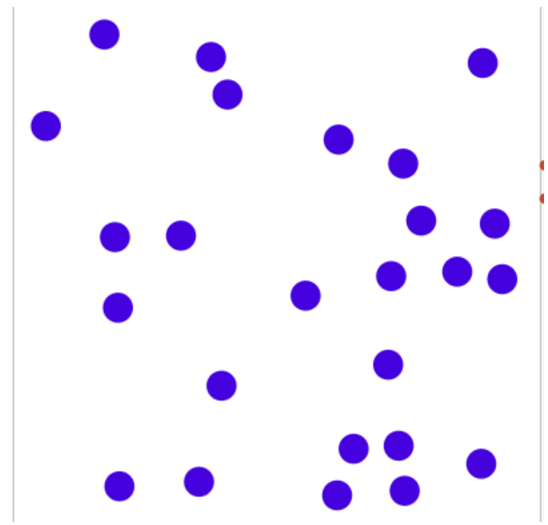


a) Anomaly present
b) Anomaly absent

Is an anomaly present?

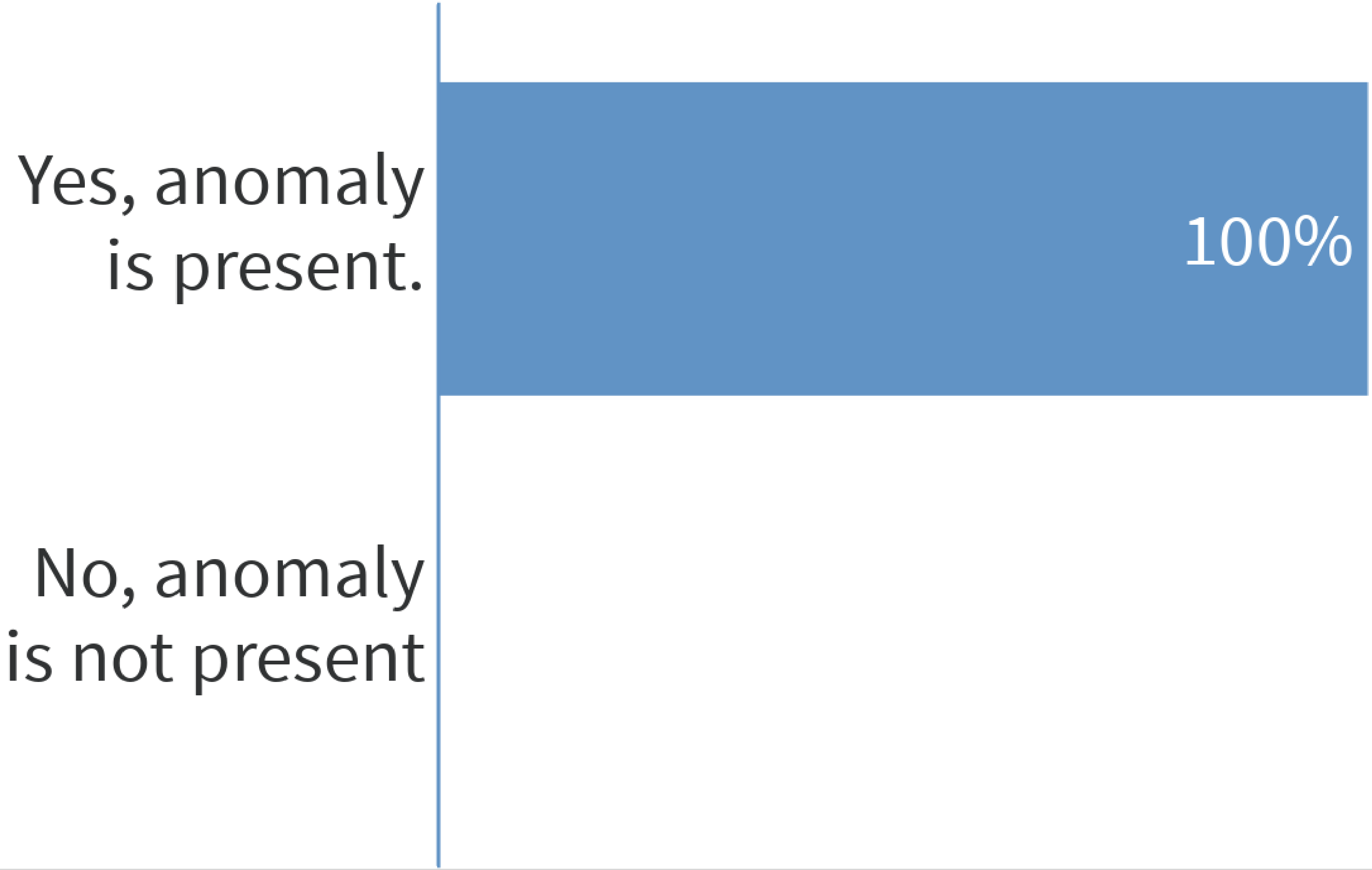


Experiment #3: color

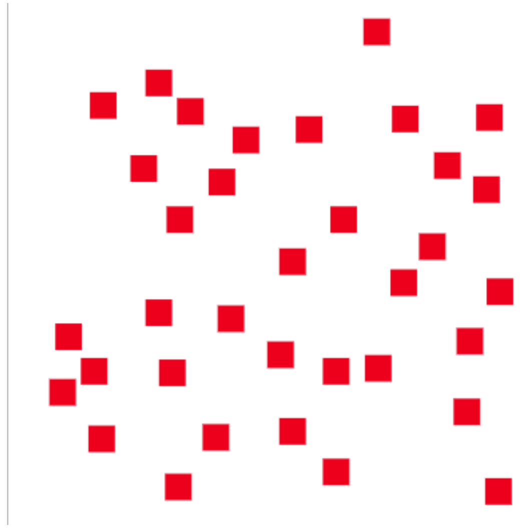


a) Anomaly present
b) Anomaly absent

Is an anomaly present?

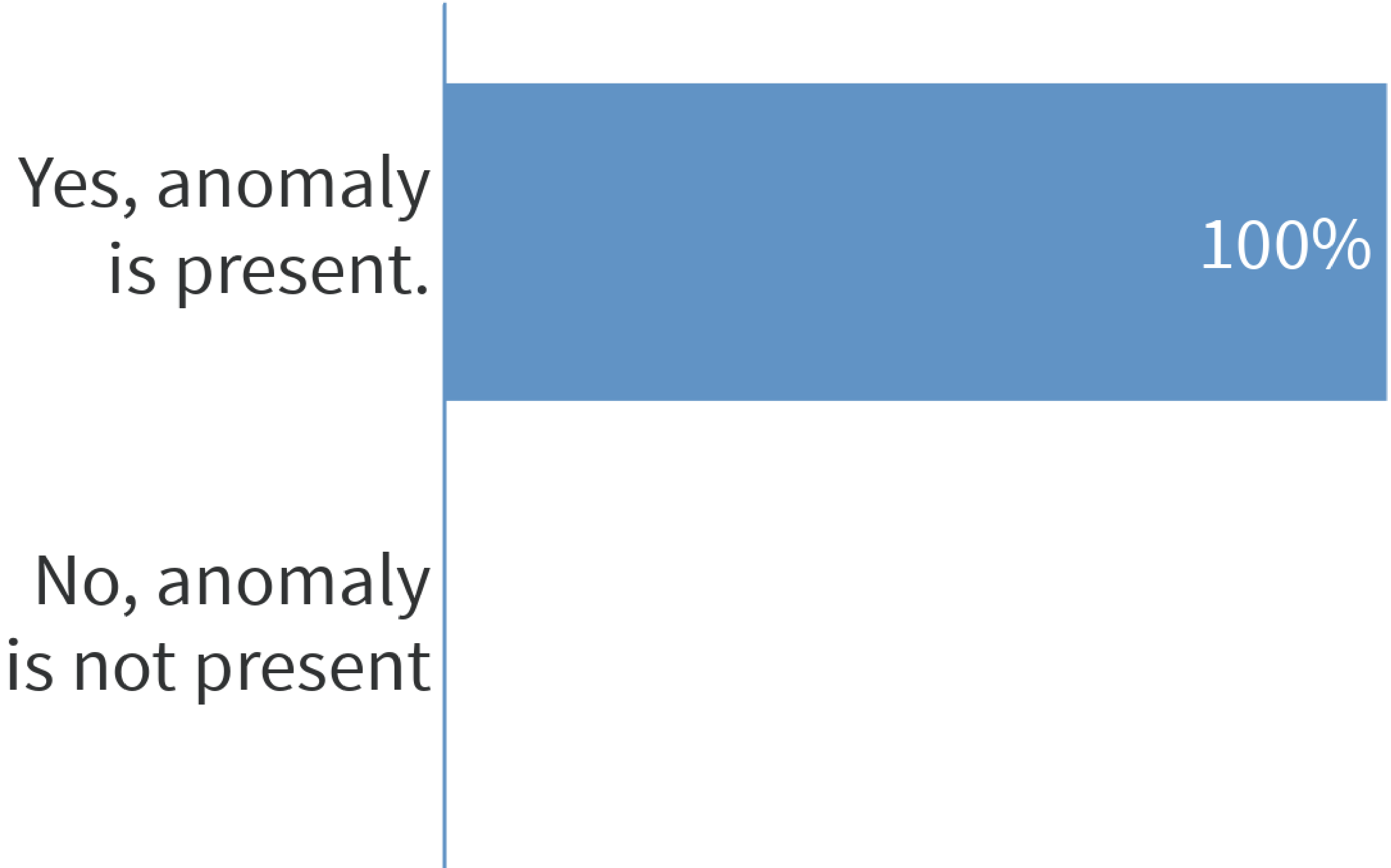


Experiment #4: shape

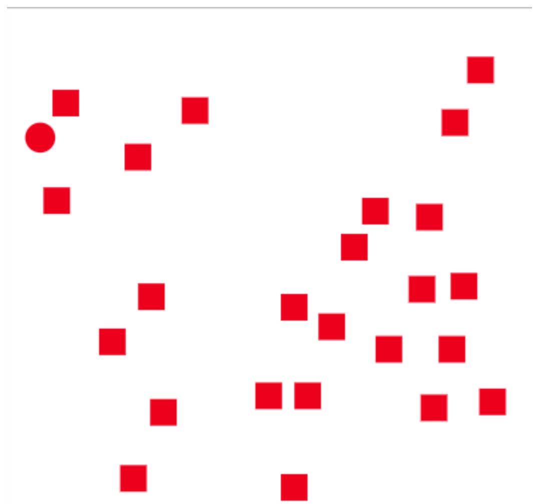


a) Anomaly present
b) Anomaly absent

Is an anomaly present?

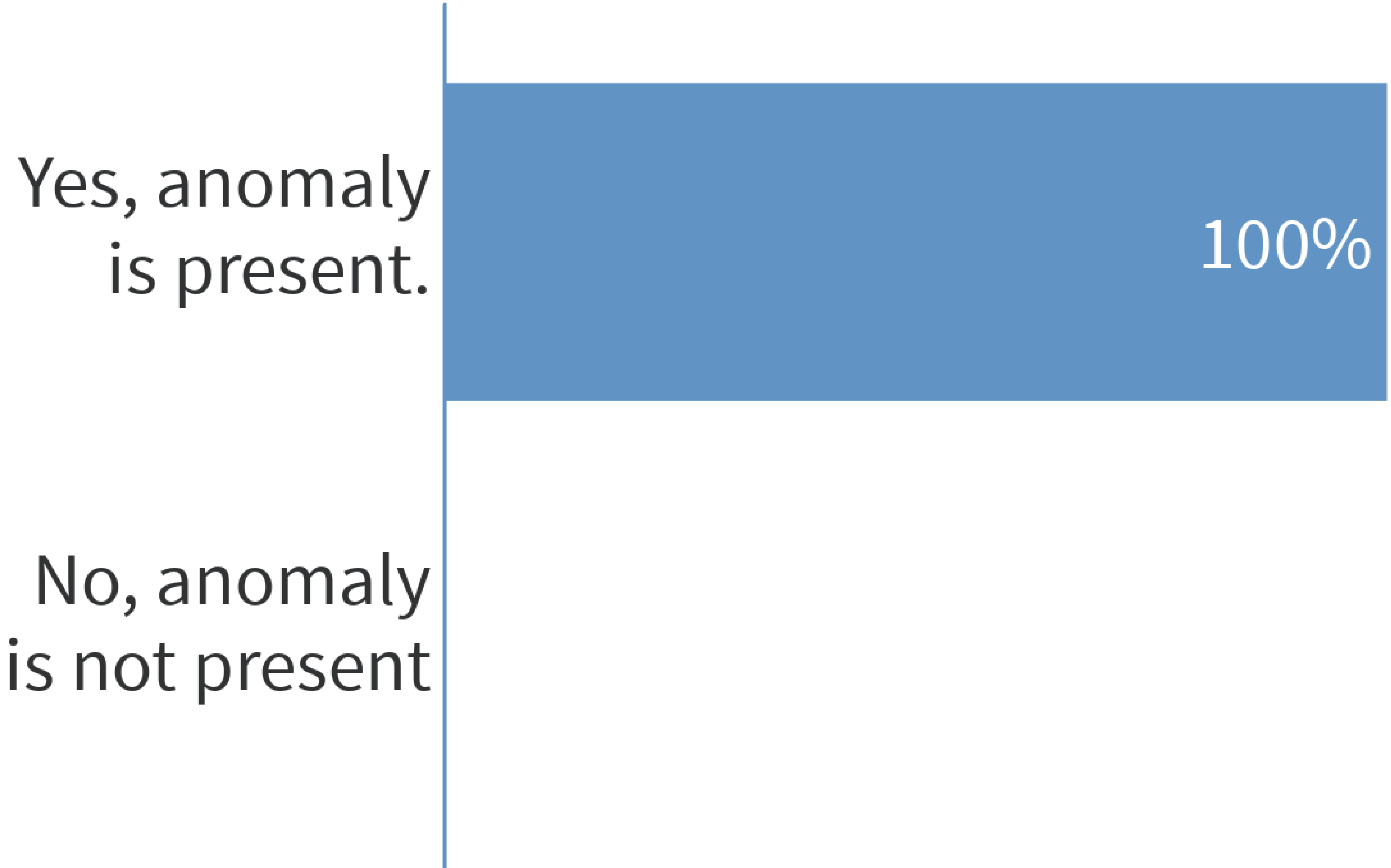


Experiment #5: shape

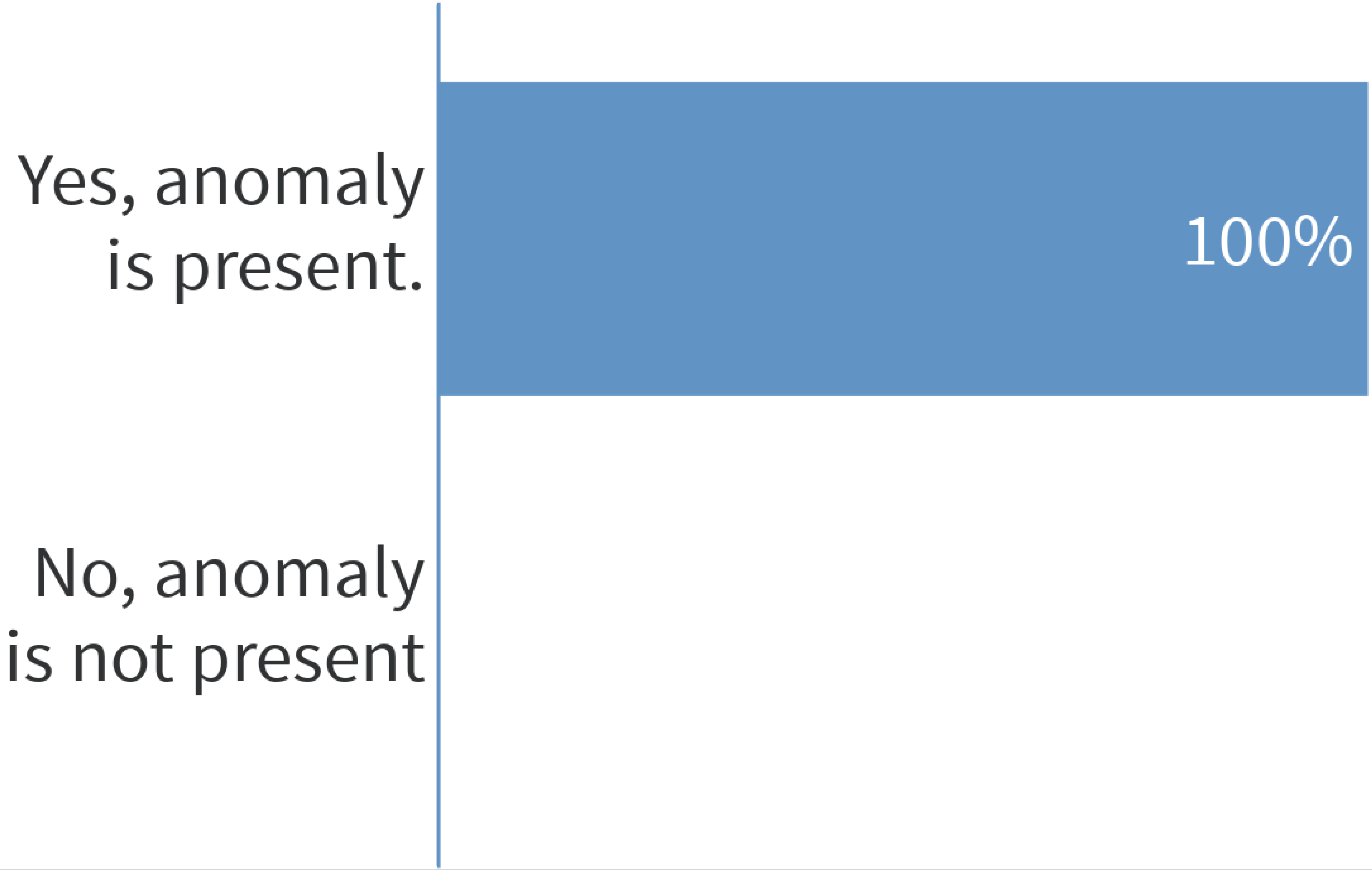


a) Anomaly present
b) Anomaly absent

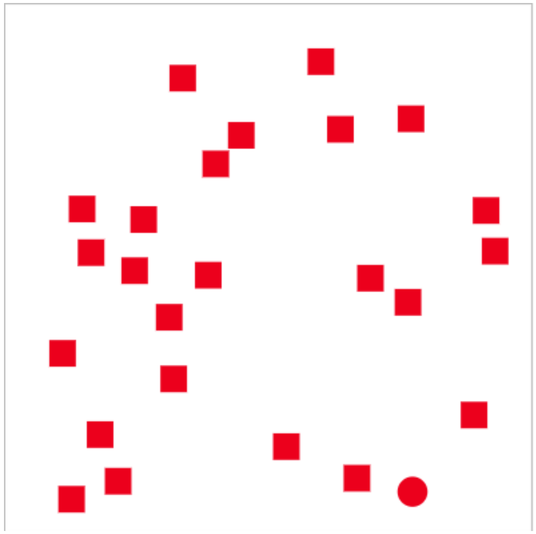
Is an anomaly present?



Is an anomaly present?



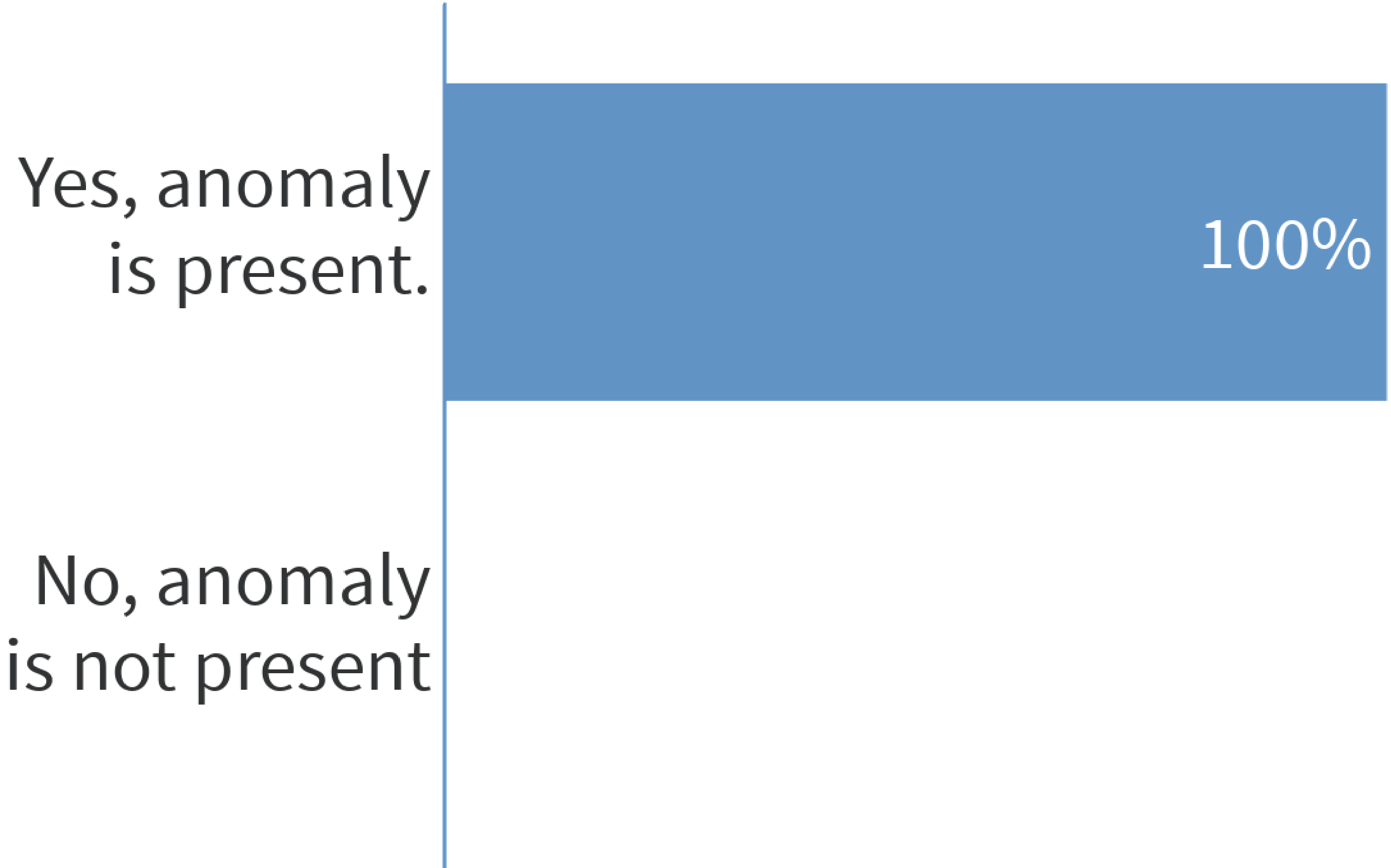
Experiment #6: shape



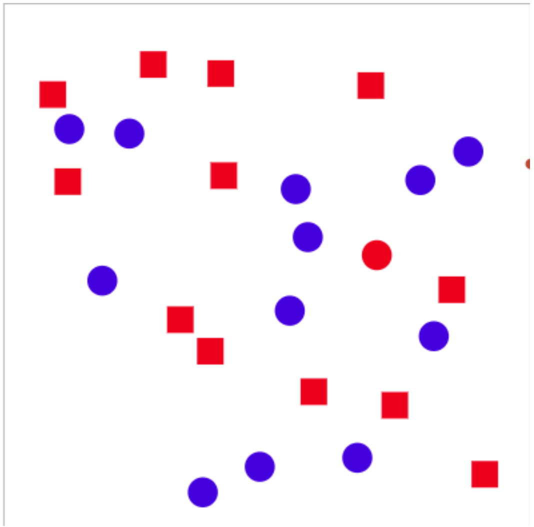
a) Anomaly present
b) Anomaly absent

AADS

Is an anomaly present?

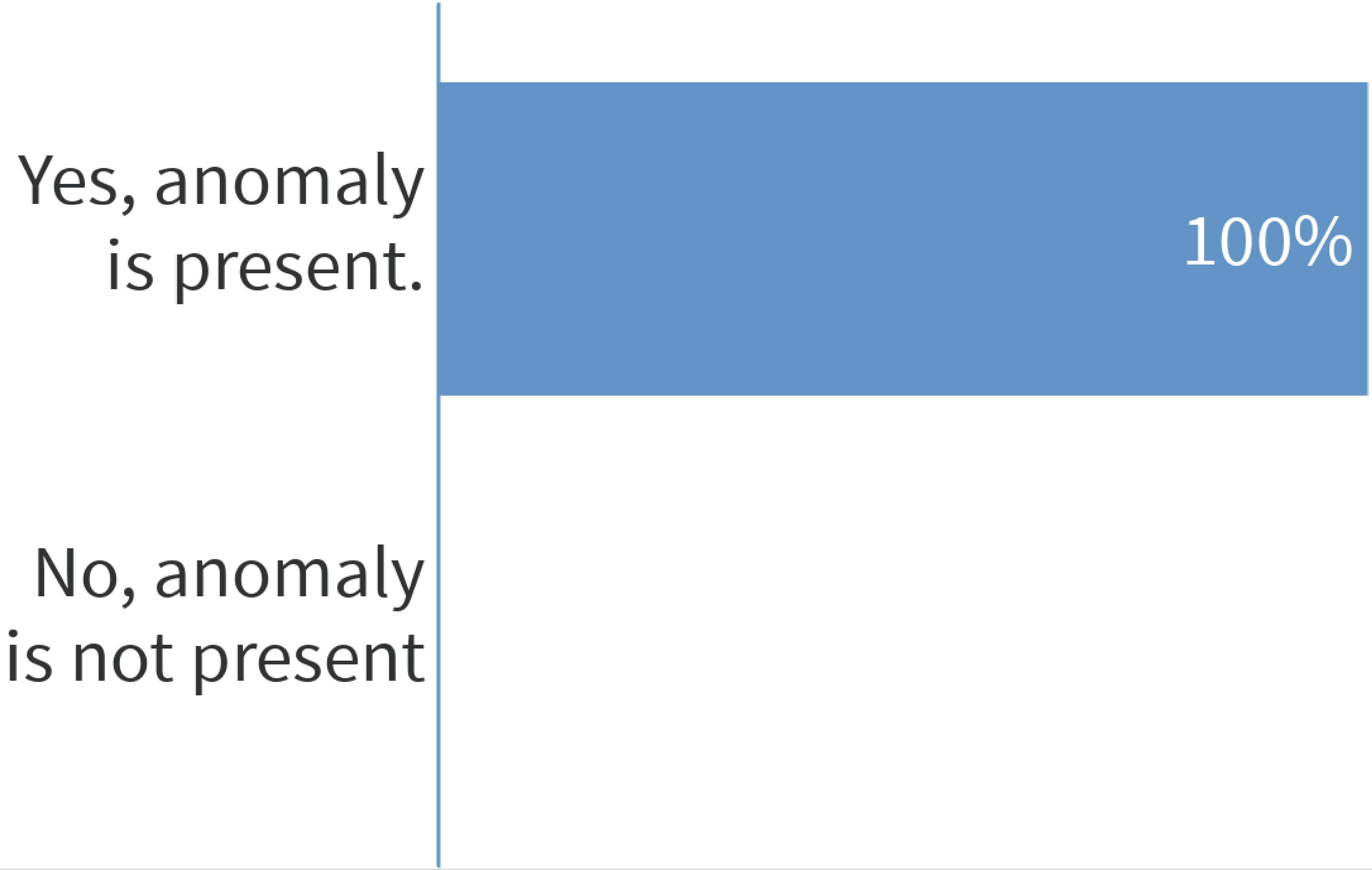


Experiment #7: conjuction

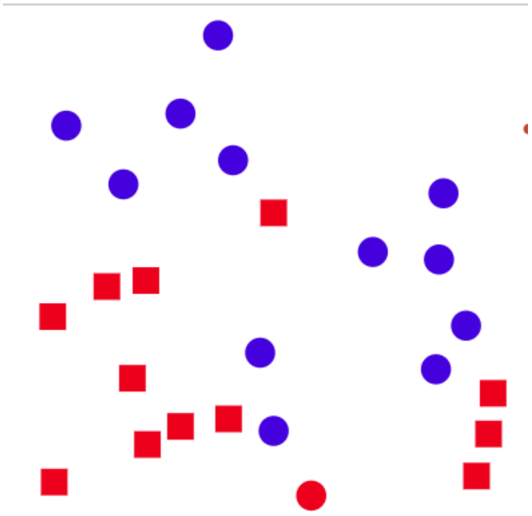


a) Anomaly present
b) Anomaly absent

Is an anomaly present?



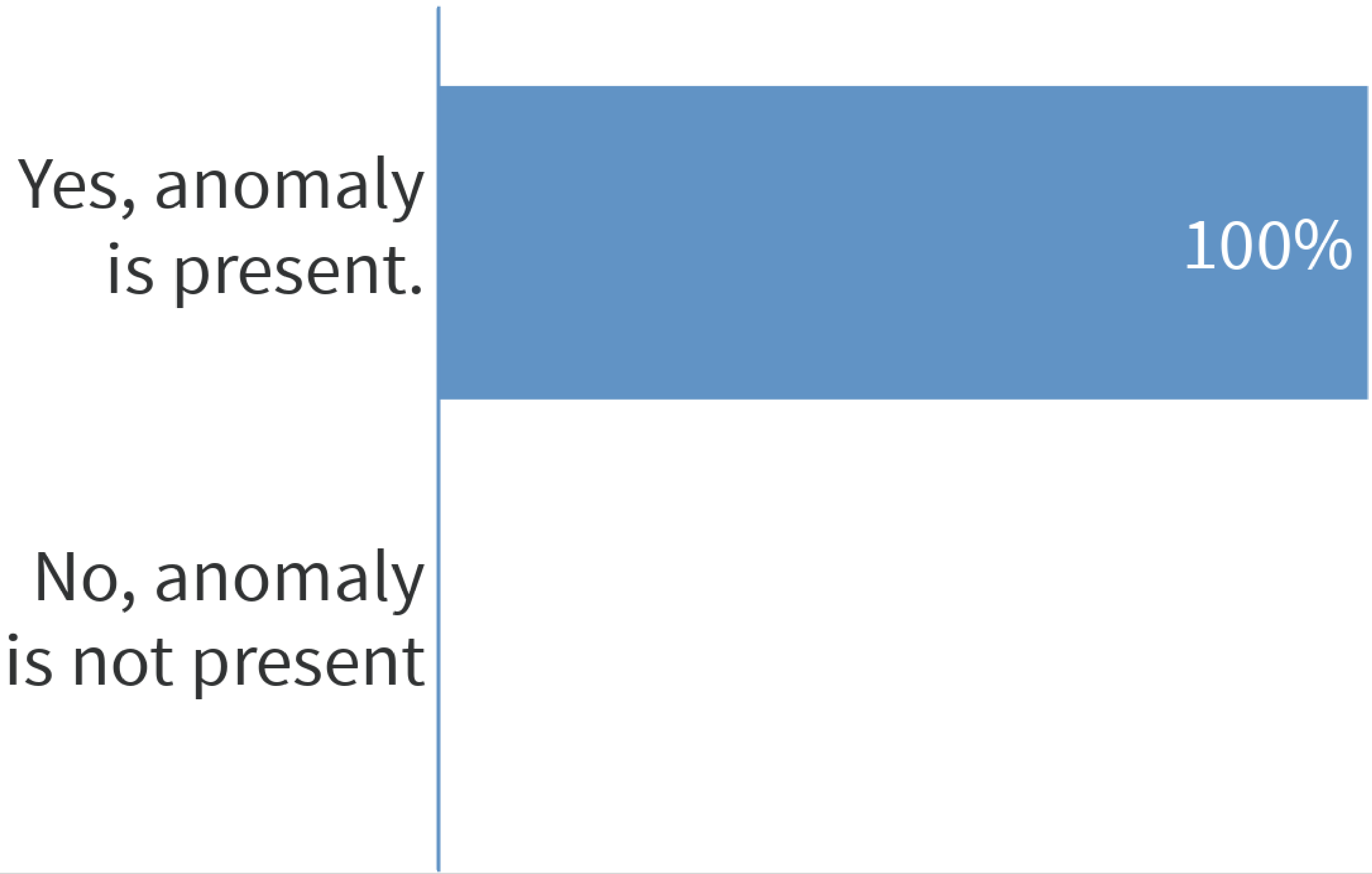
Experiment #8: conjuction



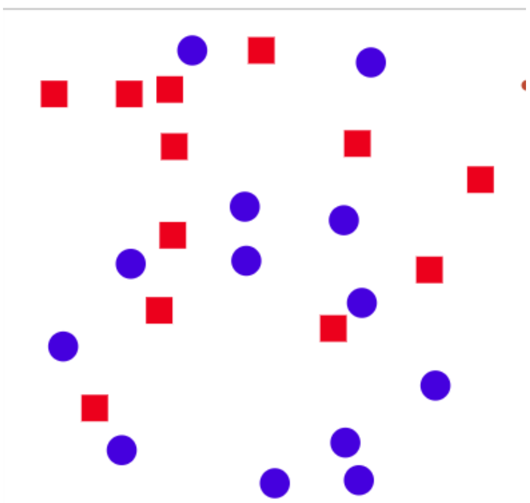
a) Anomaly present
b) Anomaly absent

AADS

Is an anomaly present?

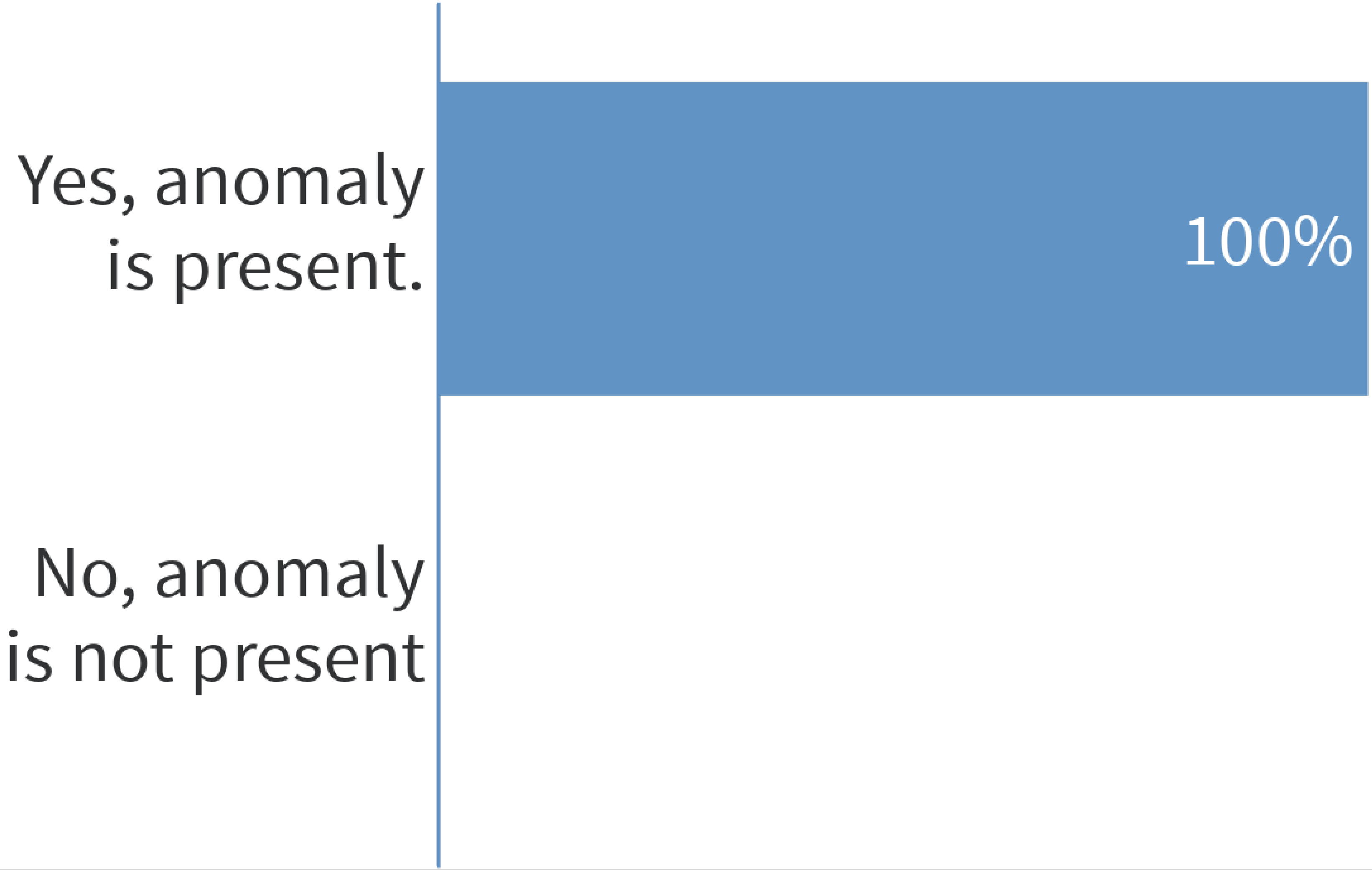


Experiment #9: conjuction

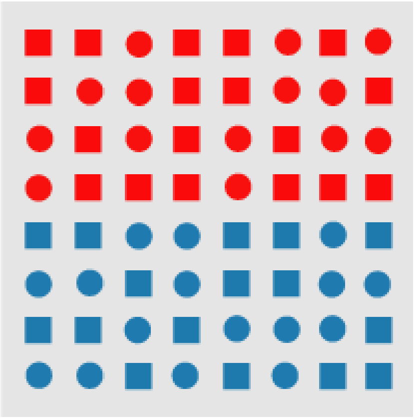


a) Anomaly present
b) Anomaly absent

Is an anomaly present?



Boundary Detection



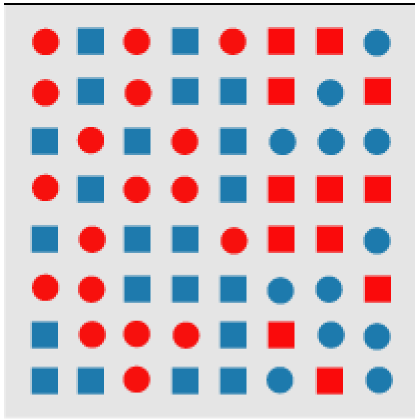
a) Boundary present
b) Boundary absent

Is boundary present?

Yes, boundary is present

No, boundary is not present

Boundary Detection



a) Boundary present
b) Boundary absent

Is boundary present?

Yes, boundary is present

No, boundary is not present

Cognitive scale of visual cues for qualitative variables

Color Hue

Orientation

Shape

Color Intensity

Size

Curvature

Added marks

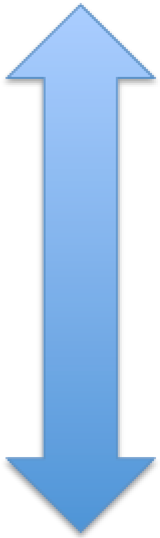
Closure

More accurate comparisons

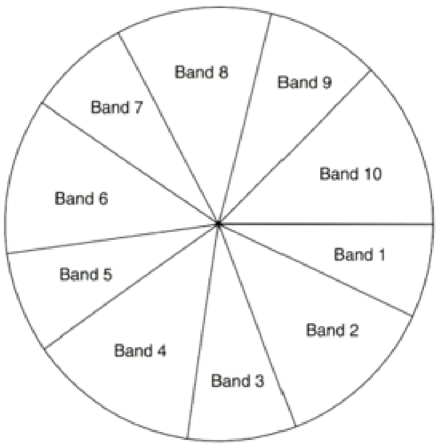
Less accurate comparisons

Cognitive scale of visual cues for quantitative variables

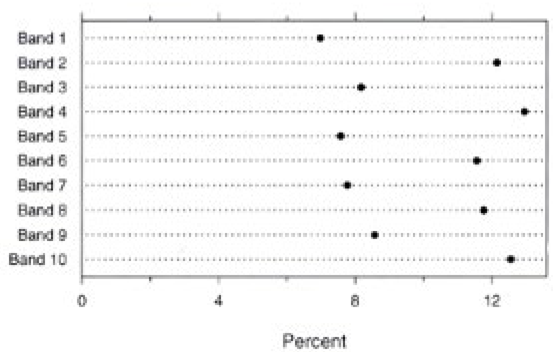
Visual cue	Chart types	Accuracy
Position	Dot charts, Bar plots, Scatterplots	More accurate comparisons
Length		
Angle		
Direction		
Shape		
Area	Piecharts	
Volume		
Color saturation	Heatmaps	
Color Hue		Less accurate comparisons



The accuracy of visual cues

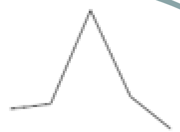


From "The Elements of Graphing Data", by Cleveland.

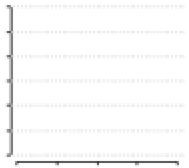


From "The Elements of Graphing Data", by Cleveland.

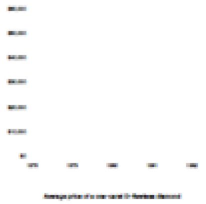
Visual cues



Coordinate system



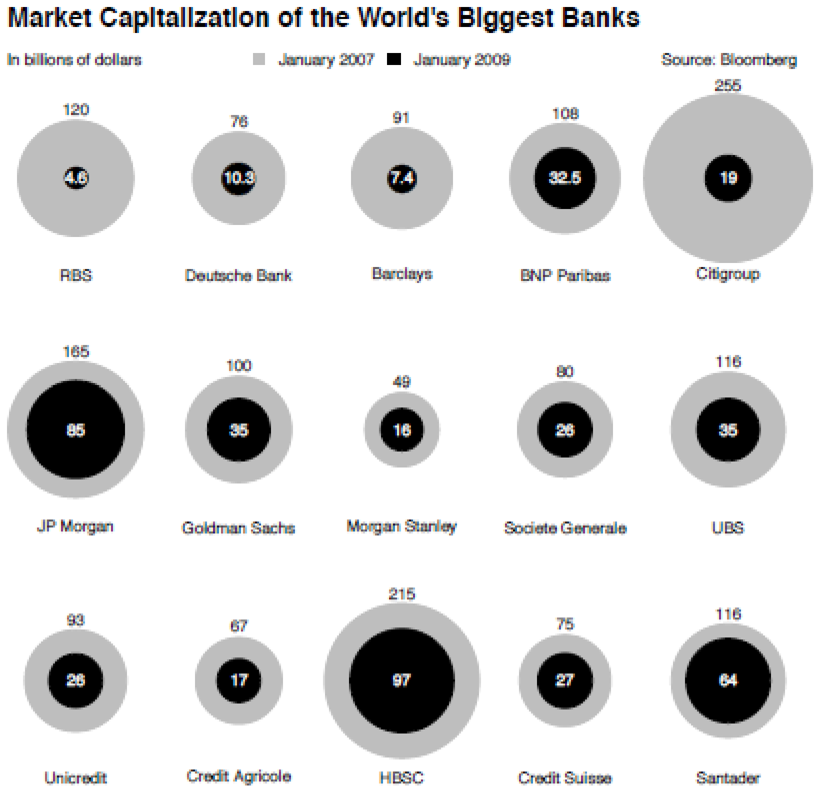
Scale



Context

The accuracy of visual cues

- This graph shows the market capitalization of the worlds biggest banks in January 2007 and January 2009.
- The original version was published by J.P. Morgan. This is a reinterpretation of the original graph (which we will see later).



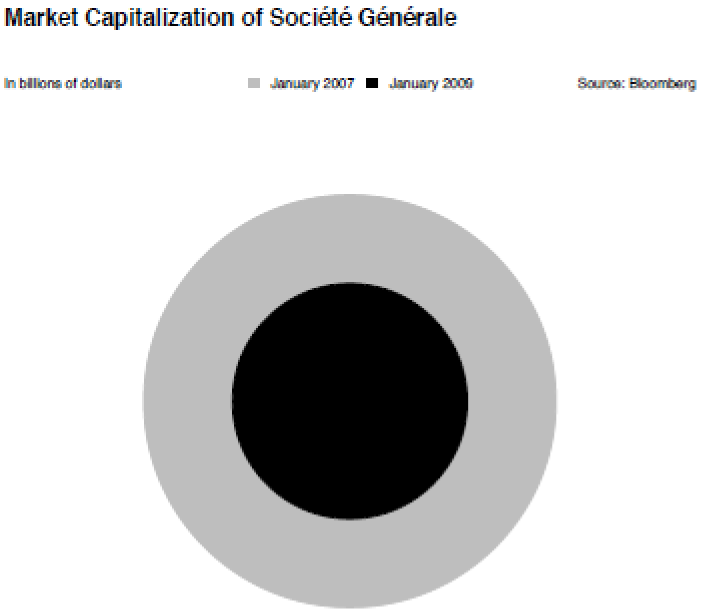
The accuracy of visual cues

- Let’s focus on one specific bank.
- If the largest bubble represents $80 billion, how much money does the second bubble represents?

a) Slightly less than than $40 billion?

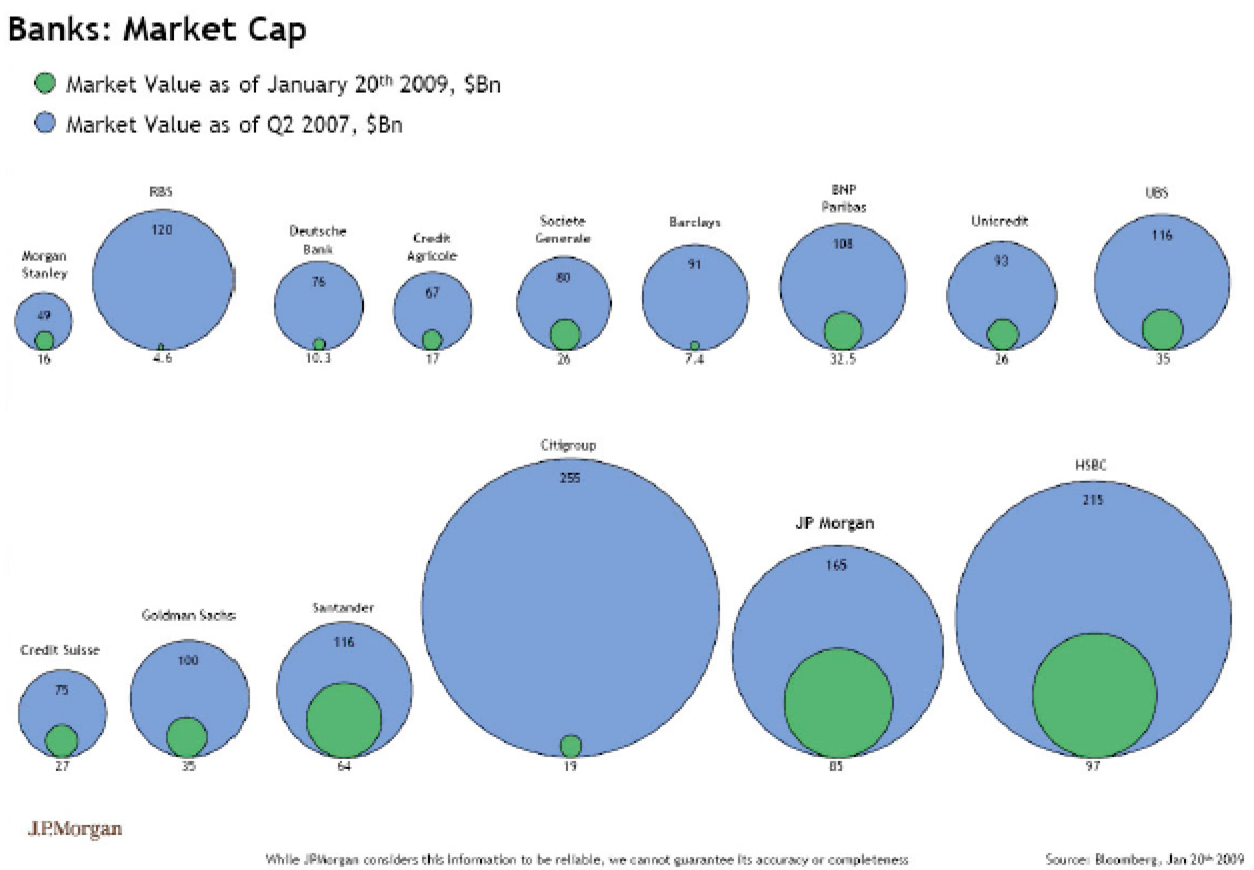
b) Slightly more than $25 billion?

c) Slightly more than $50 billions?



AADS

Other issues with areas

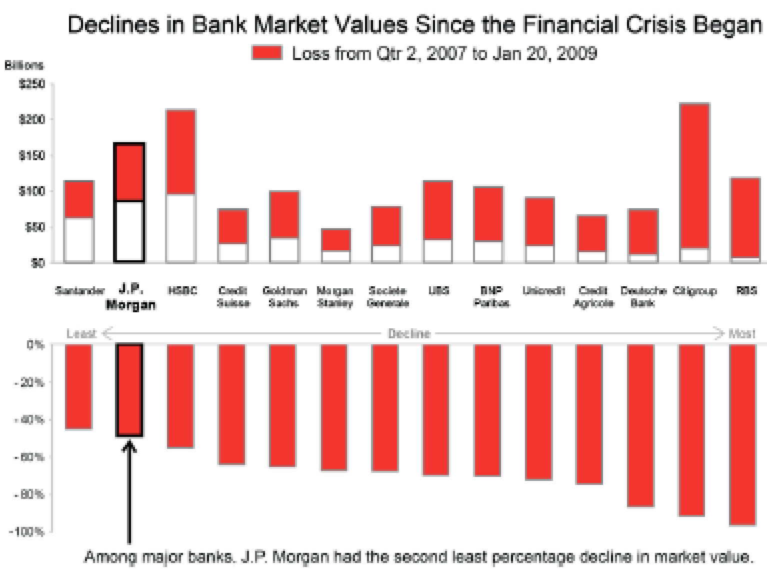


Original visualization created by J.P. Morgan Bank. Taken from http://www.perceptualedge.com/example18.php

Other issues with areas

- Why does the graphs appear to give the wrong message?
  - Values are encoded through the **diameters**, but the eye is might be trying compare the areas! This is a common problem with bubble plots.
  - This “trick” helps in making the differences across banks appear bigger than they really are.
  - What makes it worse is that there is no hint in the graph that suggests that you should be looking at the diameters rather than the areas.

Stephen Few's visualization for the Bank's market capitalization



- Stephen Few suggests using two sets of bar plots, one showing absolute values, and a second one showing relative declines.
- Much easier to make comparisons because cue is higher in the cognitive scale.
- Additional text highlights J.P. Morgan!

Taken from http://www.perceptualedge.com/example18.php

Take home message....

- Try to avoid bubbles/areas if accurate comparisons are important.
- There is almost always a better alternative than pie charts!
- However, areas and bubbles can still be a helpful tool if accurate comparisons are not key and either:
  - You want to build in redundancy in your visualization.
  - Other visual cues have been used for other variables.

Visual Cue: Color

Qualitative

Color Hue
Orientation
Shape
Color Intensity
Size
Curvature
Added marks
Closure

More accurate comparisons

Less accurate comparisons

Quantitative

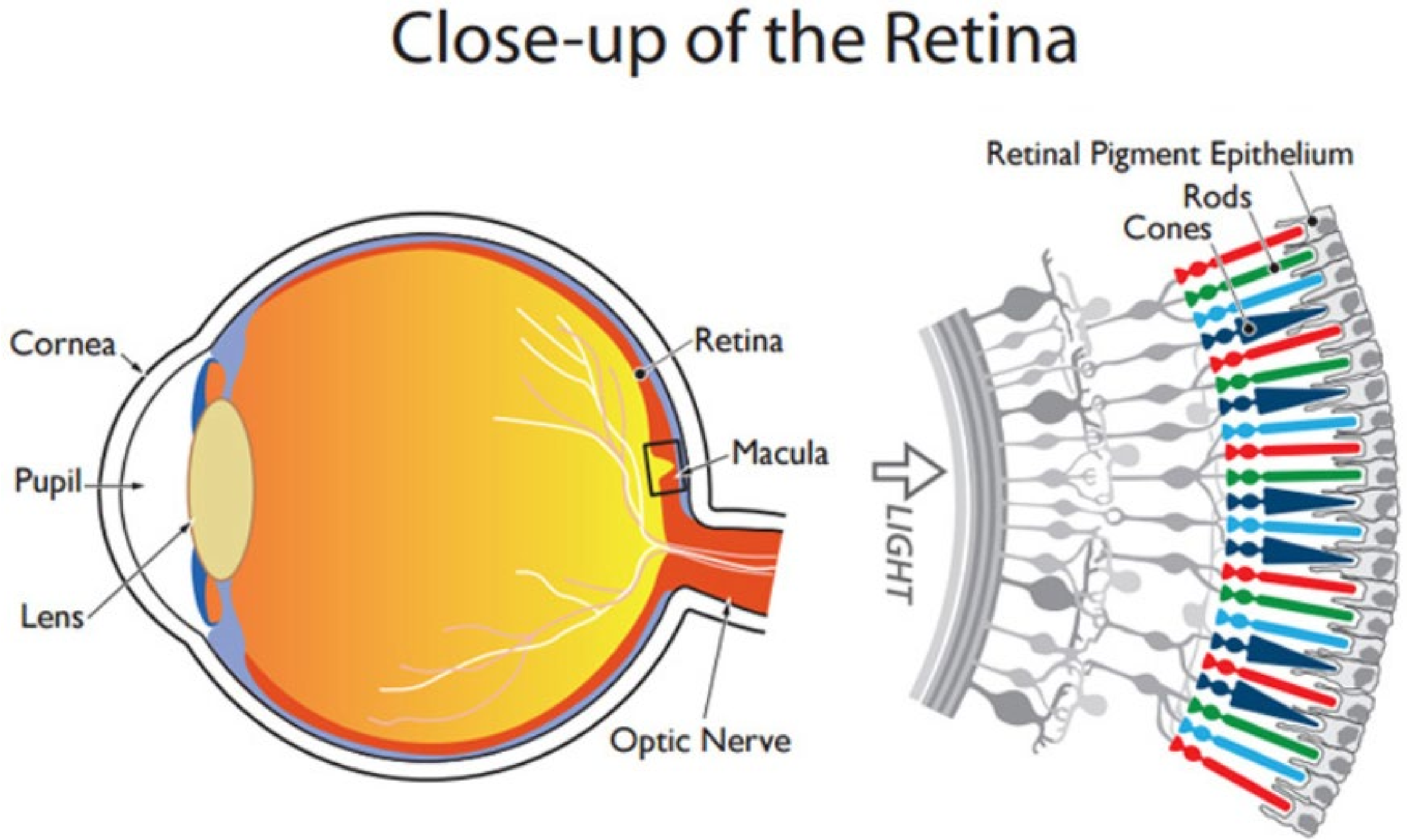
Position
Length
Angle
Direction
Shape
Area
Volume
Color saturation
Color Hue

Dot charts
Bar plots
Scatterplots
Piecharts
Heatmaps

More accurate comparisons

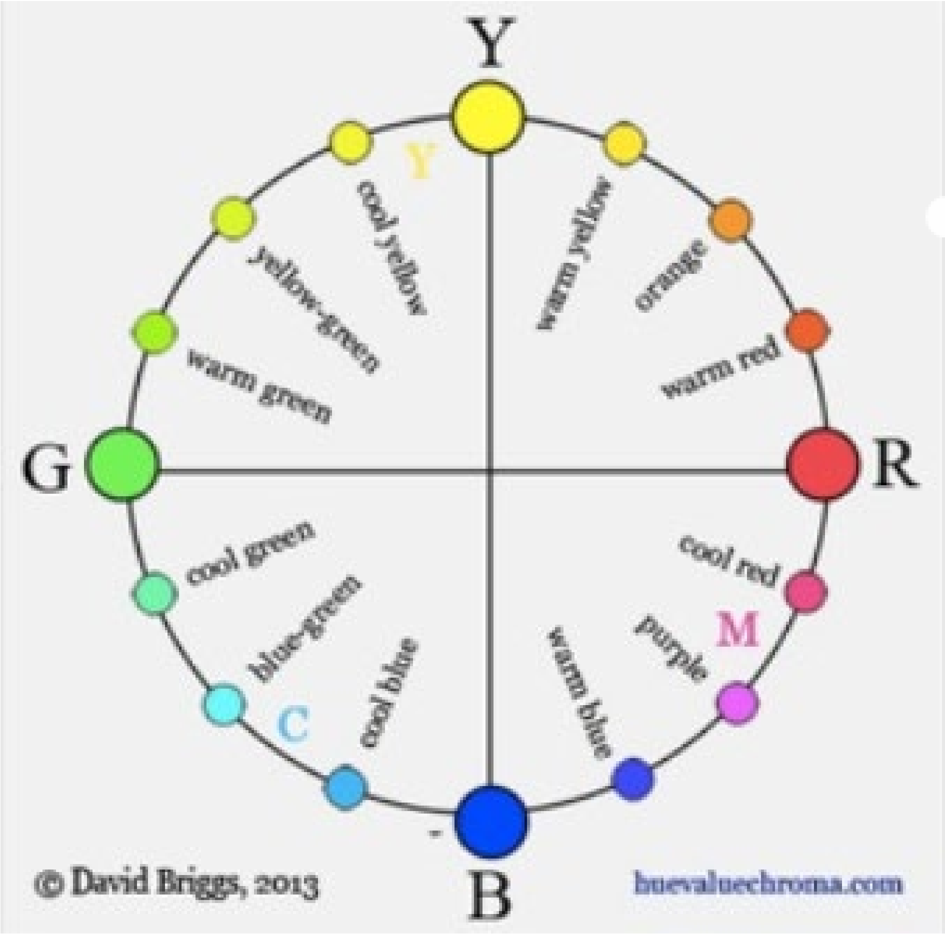
Less accurate comparisons

How eyes perceive color



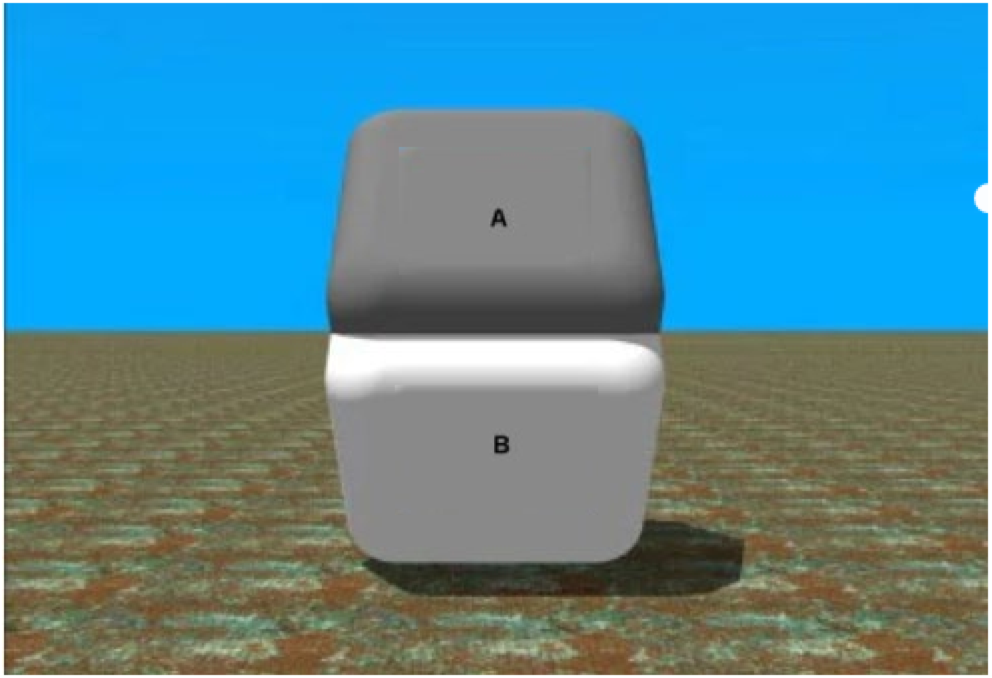
AADS https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

Antagonistic color theory



AADS

Antagonistic color theory



Representing Color

Main color schemes

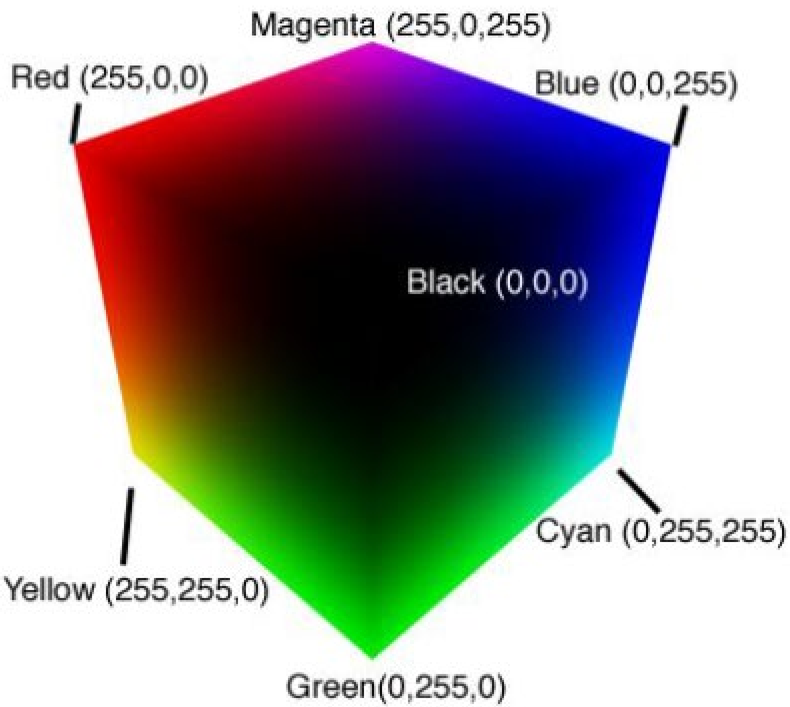
1. Computer scales
2. Perceptual scales

Decomposition of color:

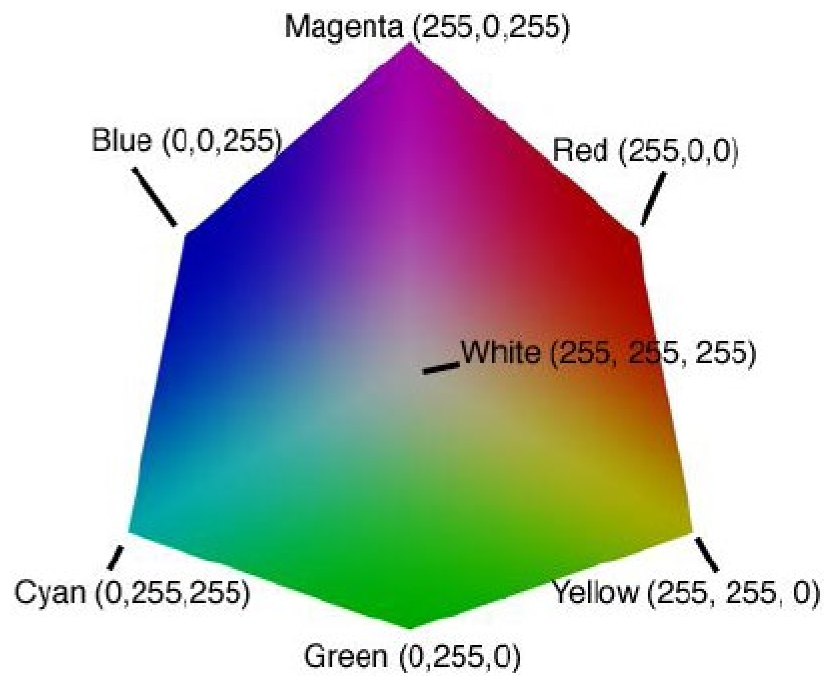
1. Hue
2. Saturation
3. Value

The RGB scale

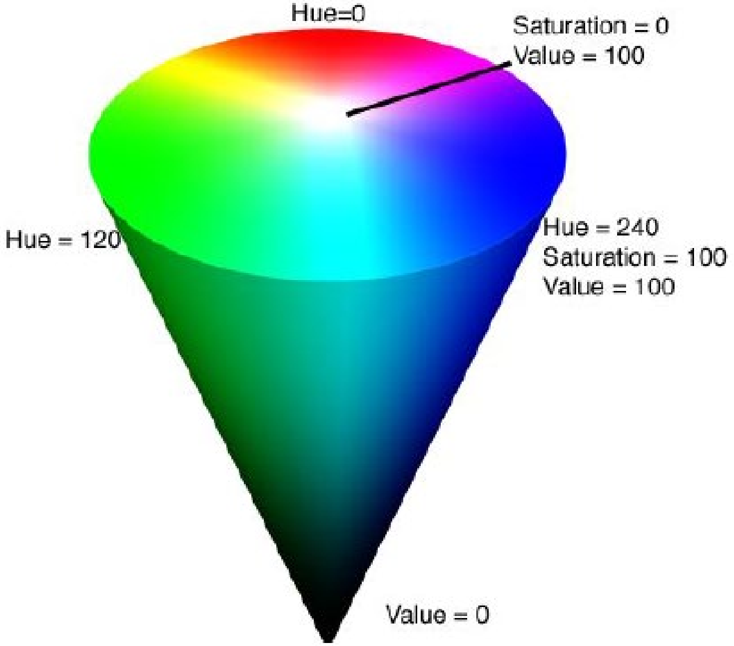
The black corner of the RGB cube

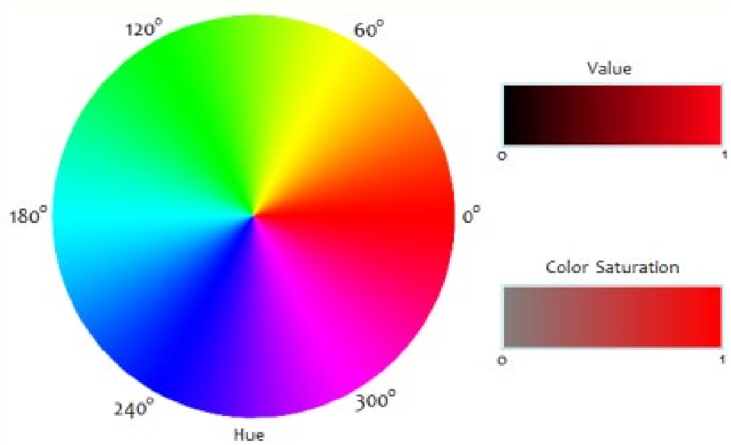


The white corner of the RGB cube

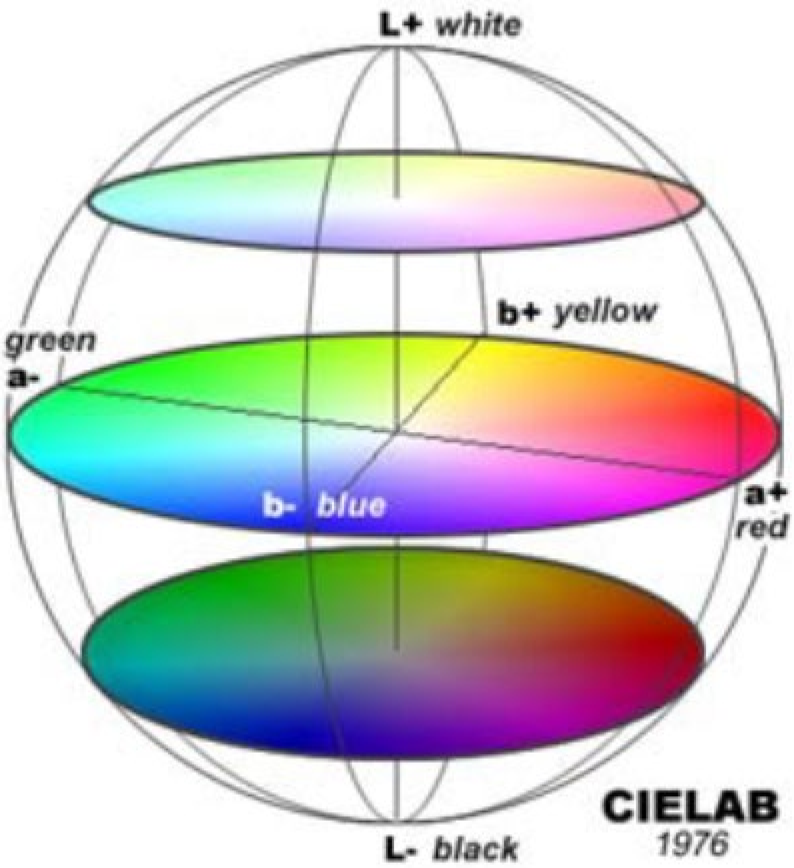


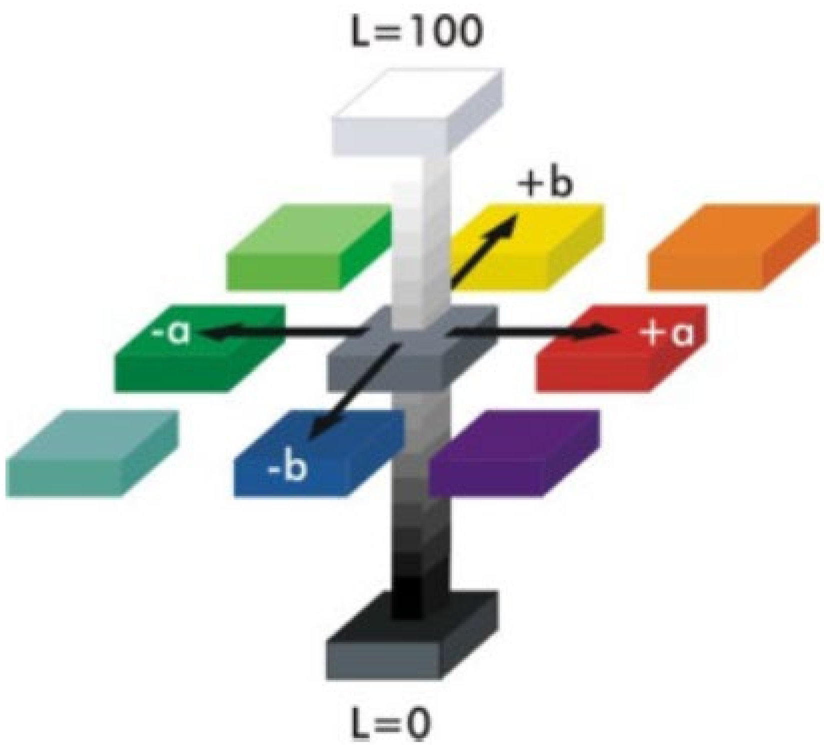
The HSV scale





The CIElab scale



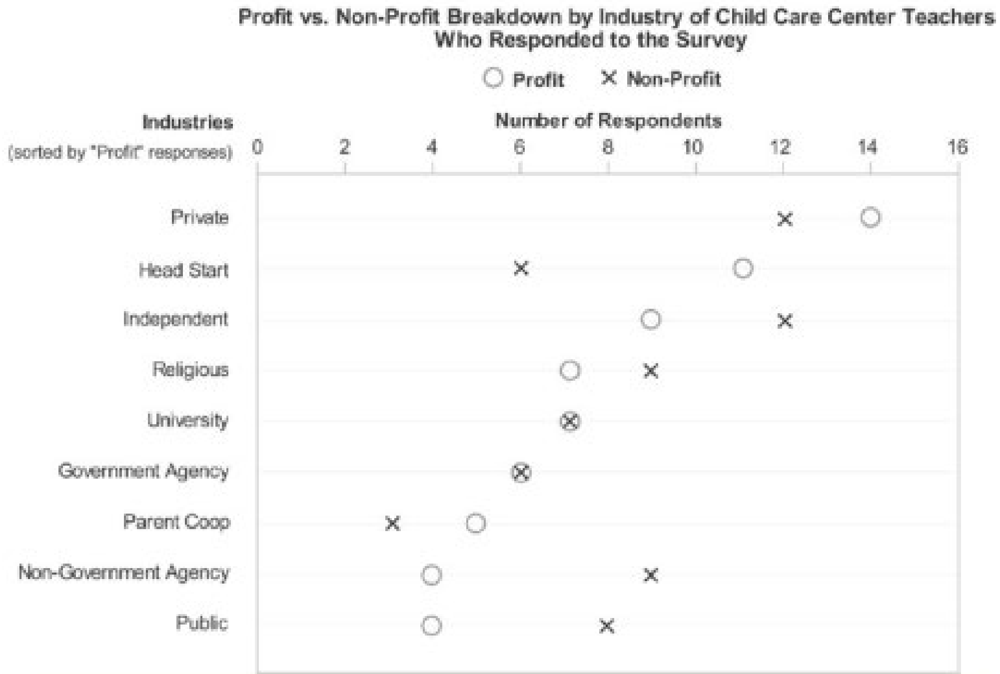


Examples of CIElab palettes

Qualitative | Quantitative (divergent) | Quantitative (sequential)

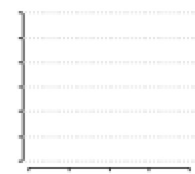
Color Brewer palettes:
http://www.personal.psu.edu/cab38/ColorBrewer/ColorBrewer_intro.html
Recommended palette: https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

Visual Cue: Shape

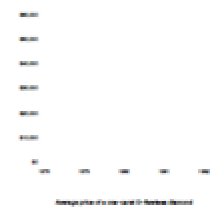


Visual cues

Coordinate system



Scale



Context

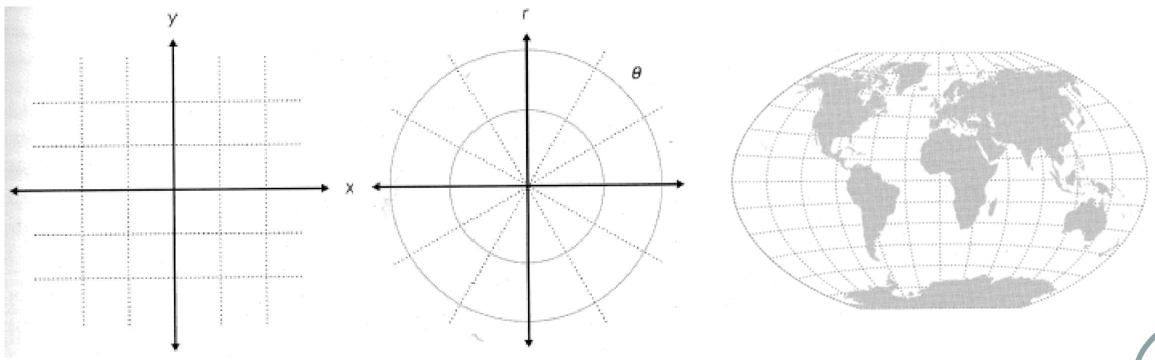
Visual Cue: Shape

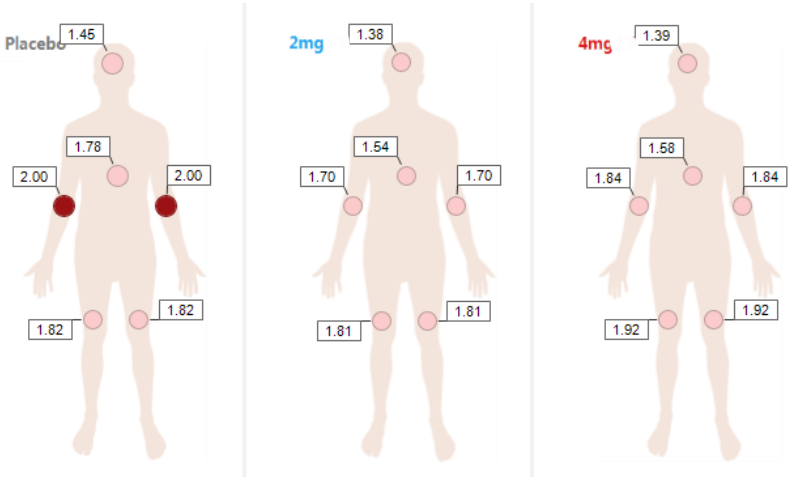
"Pre-attentive symbols become less distinct as the variety of distracters increases. It is easy to spot a single hawk in a sky full of pigeons, but if the sky contains a greater vaiety of birds, the hawks will be more difficult to see. A number of studies have shown that the immediacy of any pre-attentive cue declines as the variety of alternative patterns increases, even if all the distracting patterns are individually distinct from the target."

Colin Ware (2000) "Information Visualization: Perception and Design".

AADS

Coordinate System



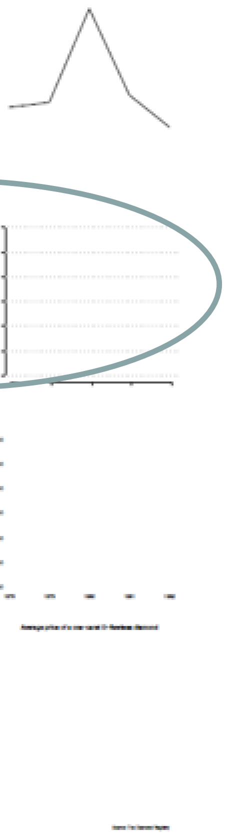


Visual cues

Coordinate system

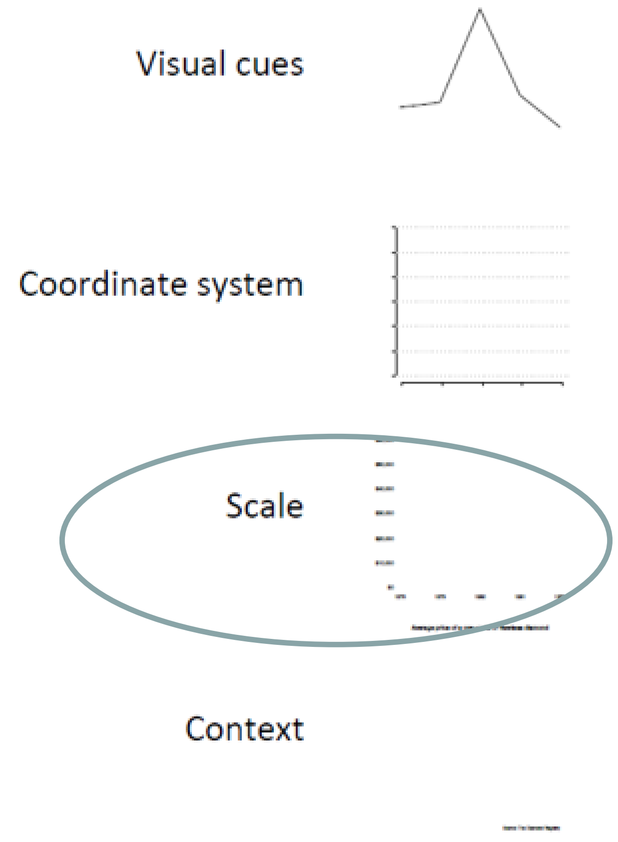
Scale

Context



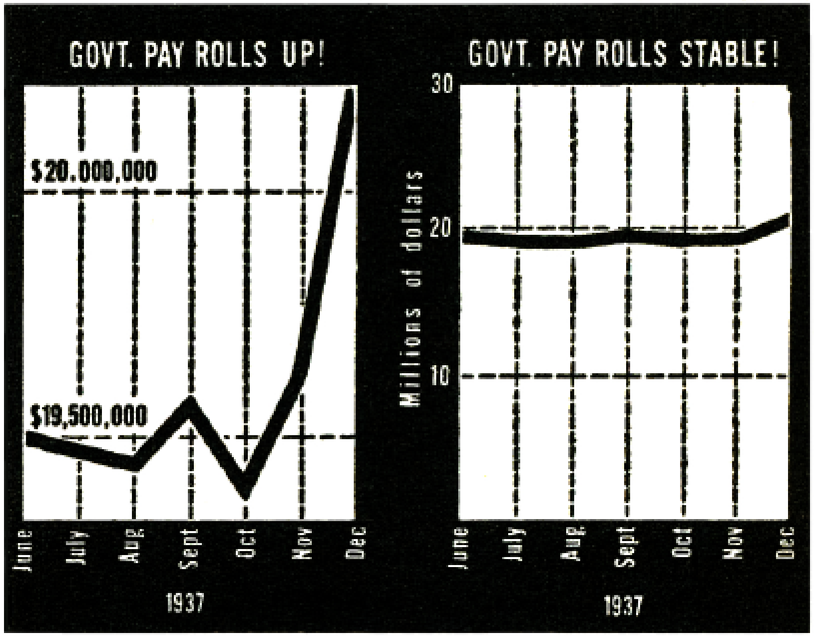
Scale

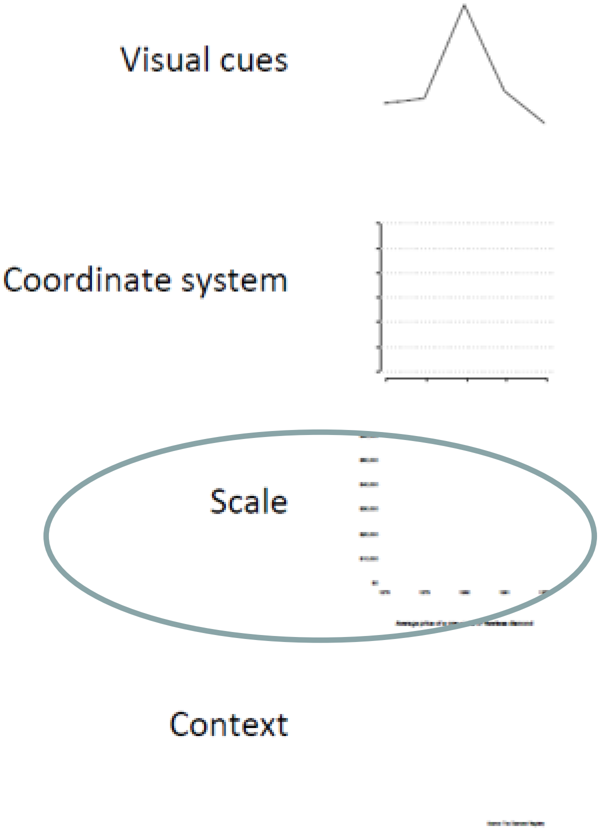
1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…



Scale

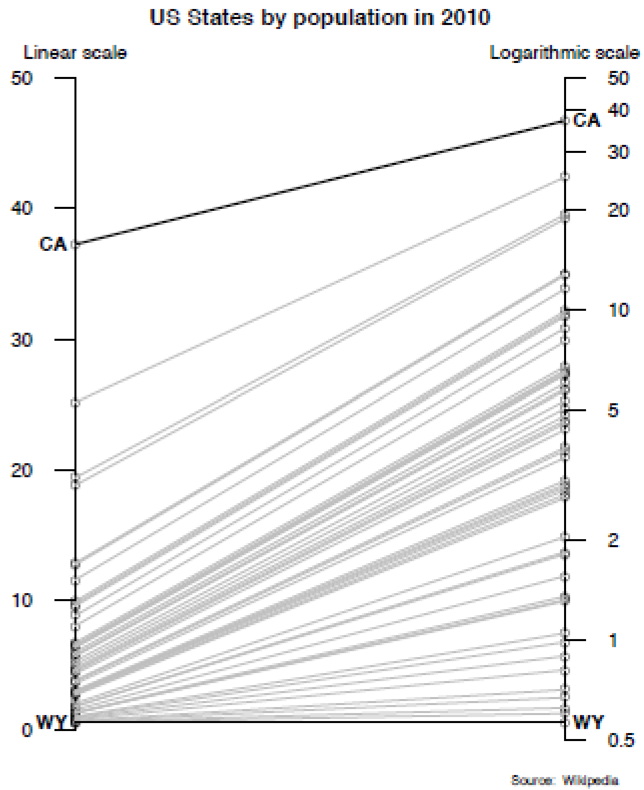
1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

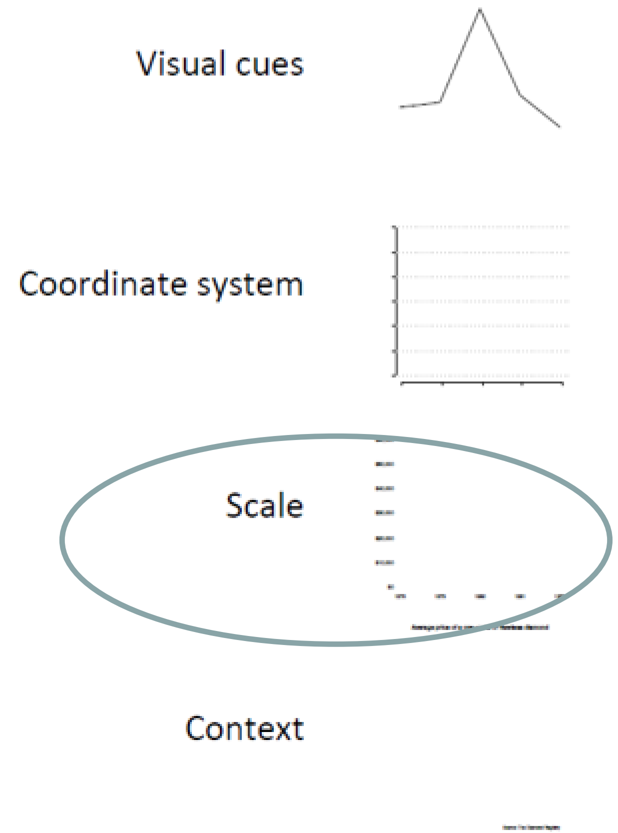




Scale

1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

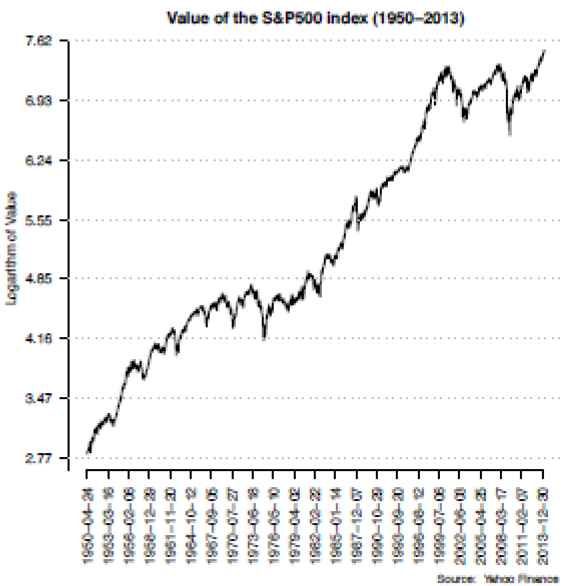




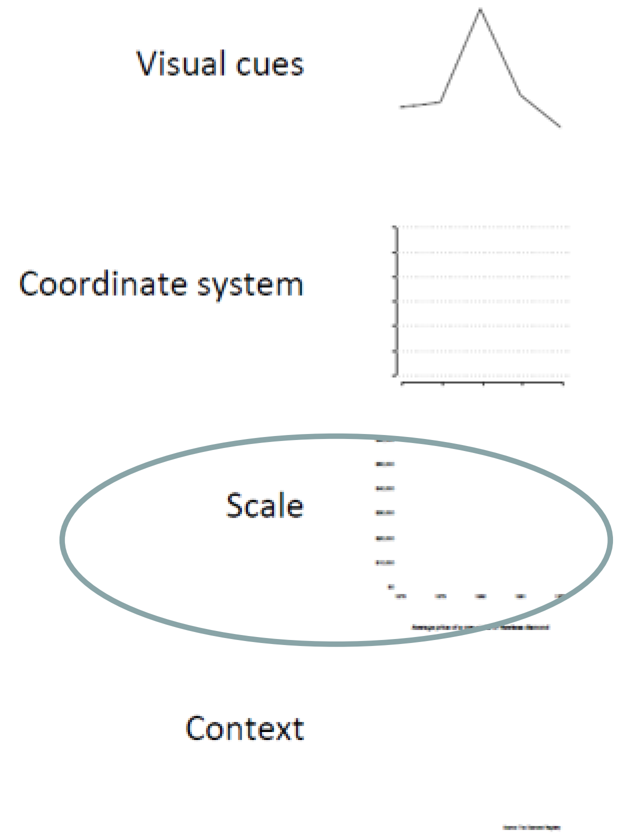
AADS

Scale

1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

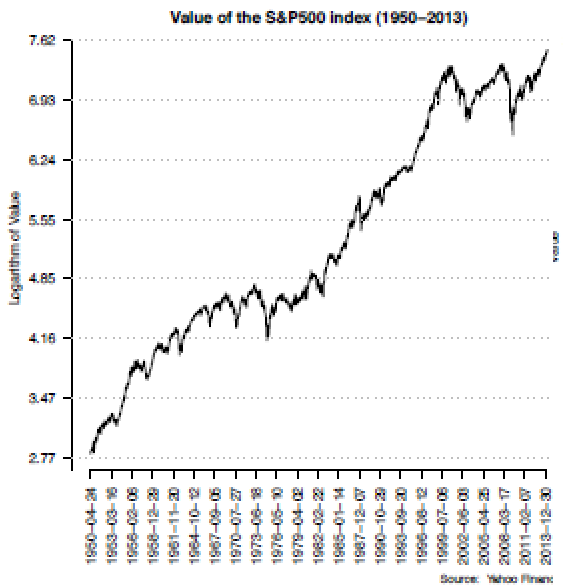


NO!

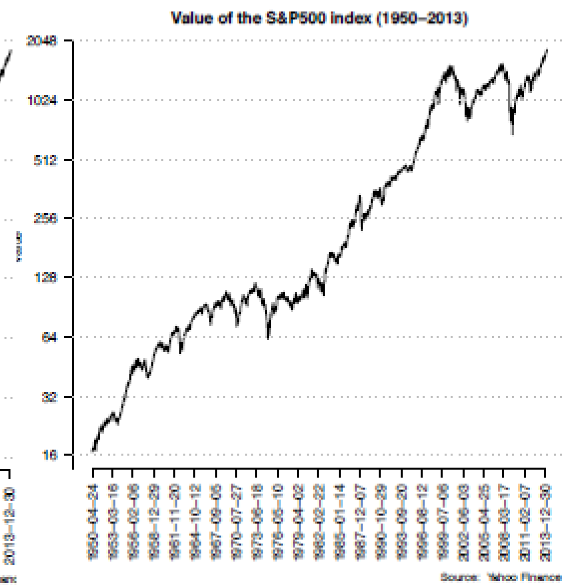


Scale

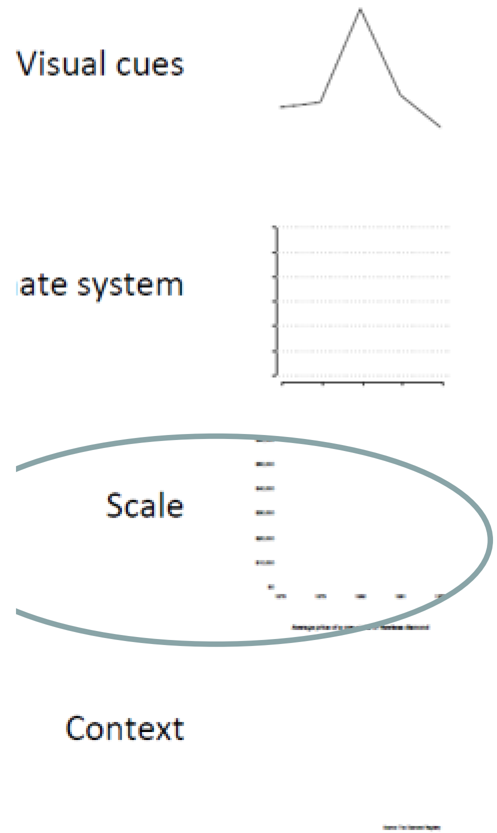
1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…



NO!

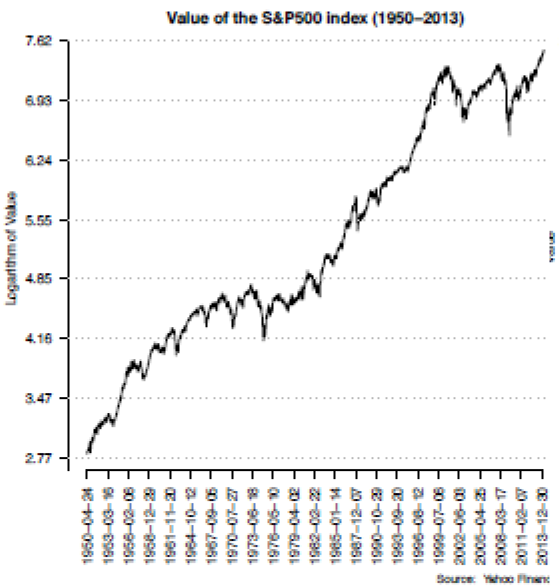


Most common

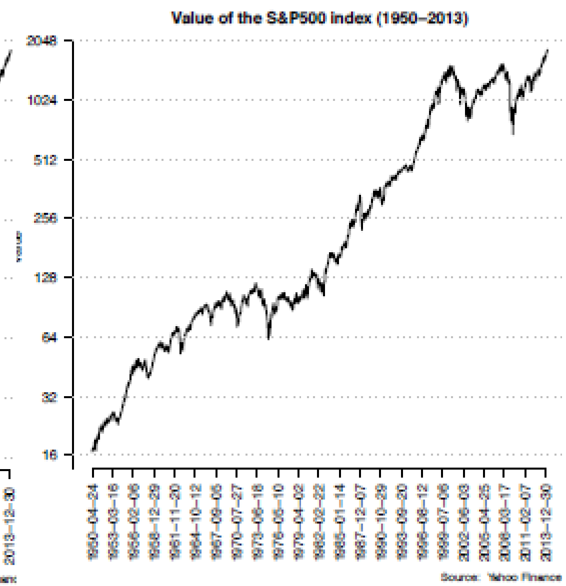


Scale

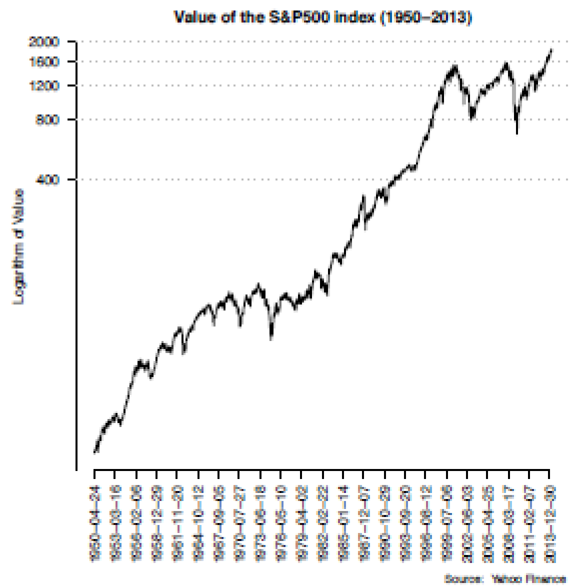
1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…



NO!



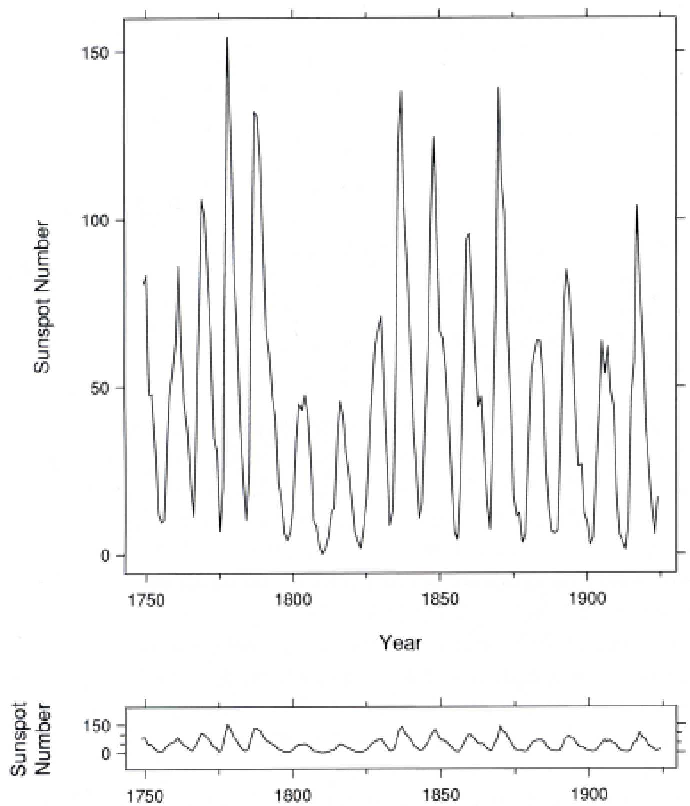
Most common

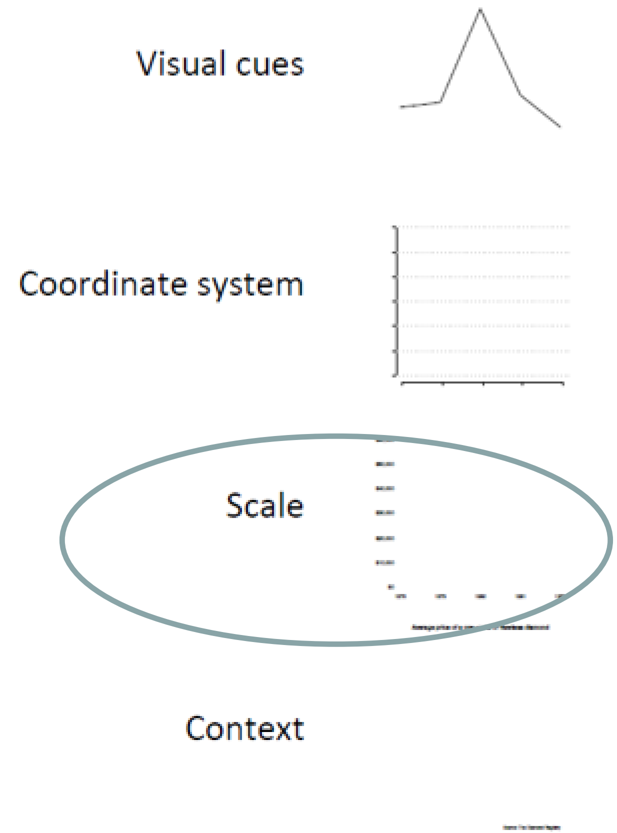


An alternative

Scale & Aspect Ratio

1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

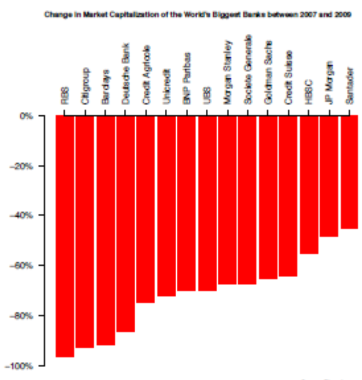




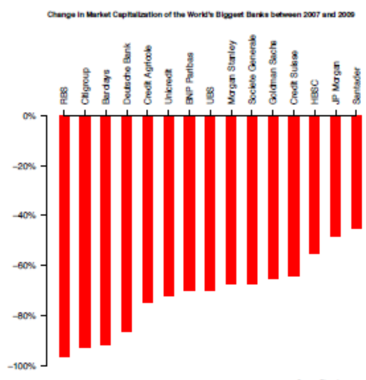
Scale

1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

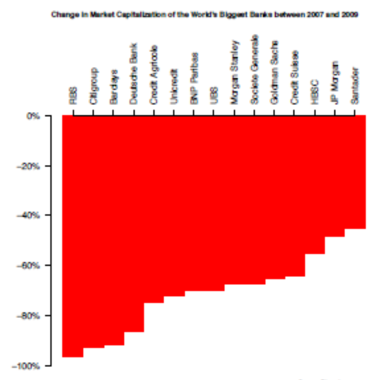
About the right amount of space between bars



Too much space between bars

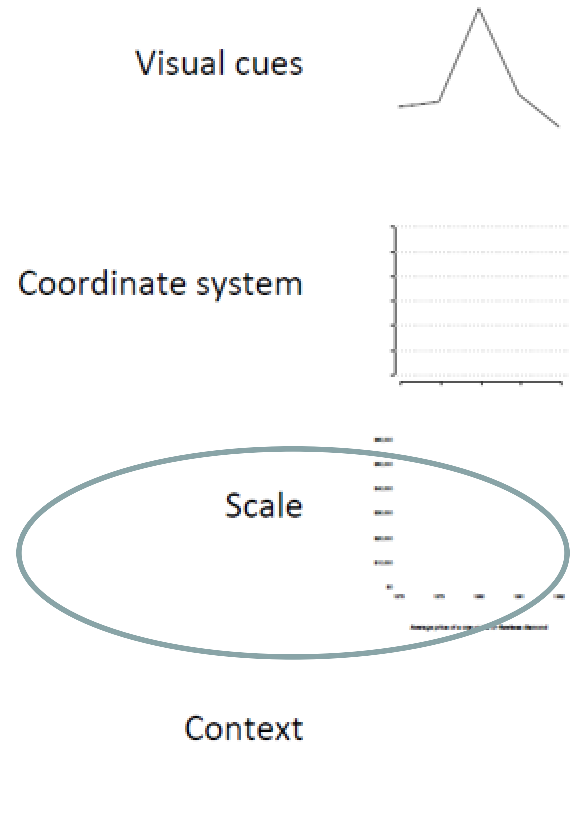


No space between bars



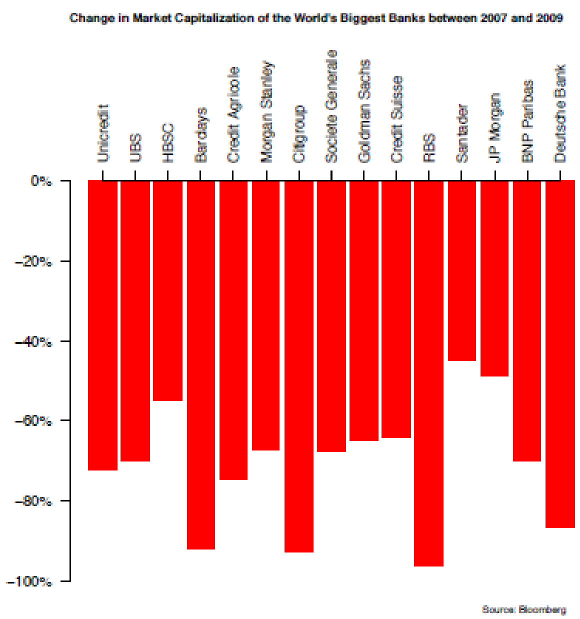
Space in between bars is about 20% of the width of the bar

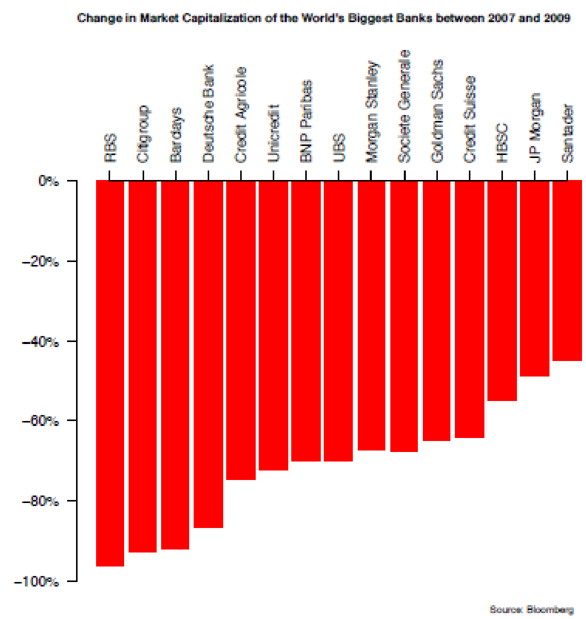
Arguably, less visually appealing

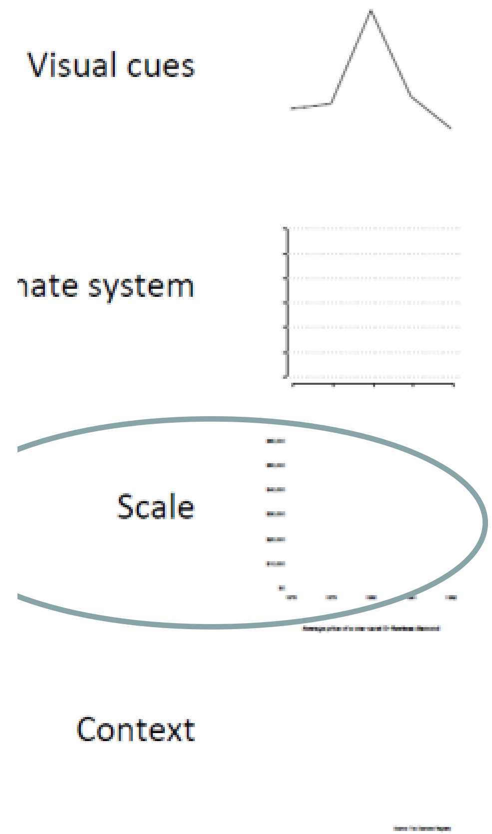


Scale: sorting

1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

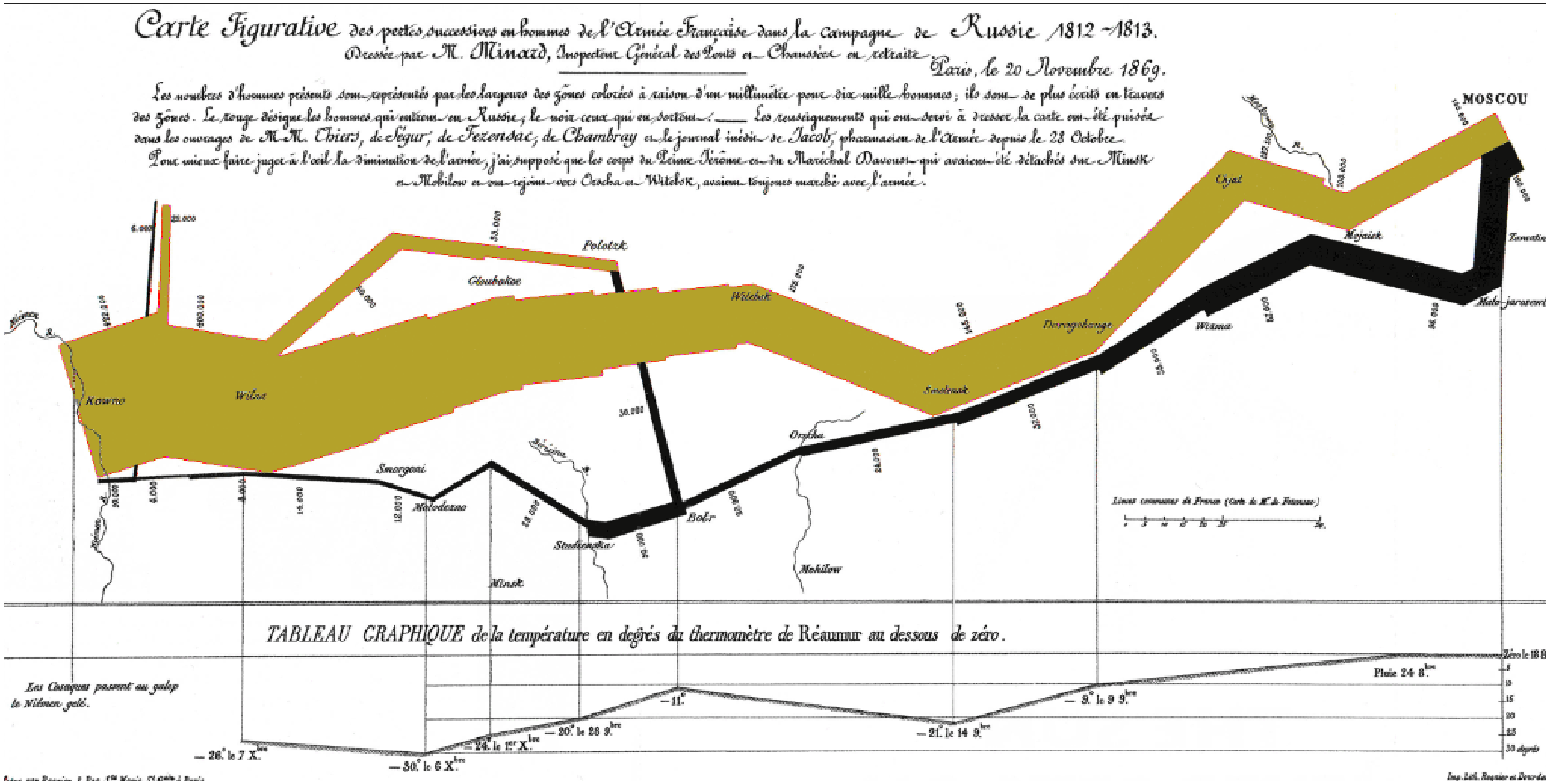






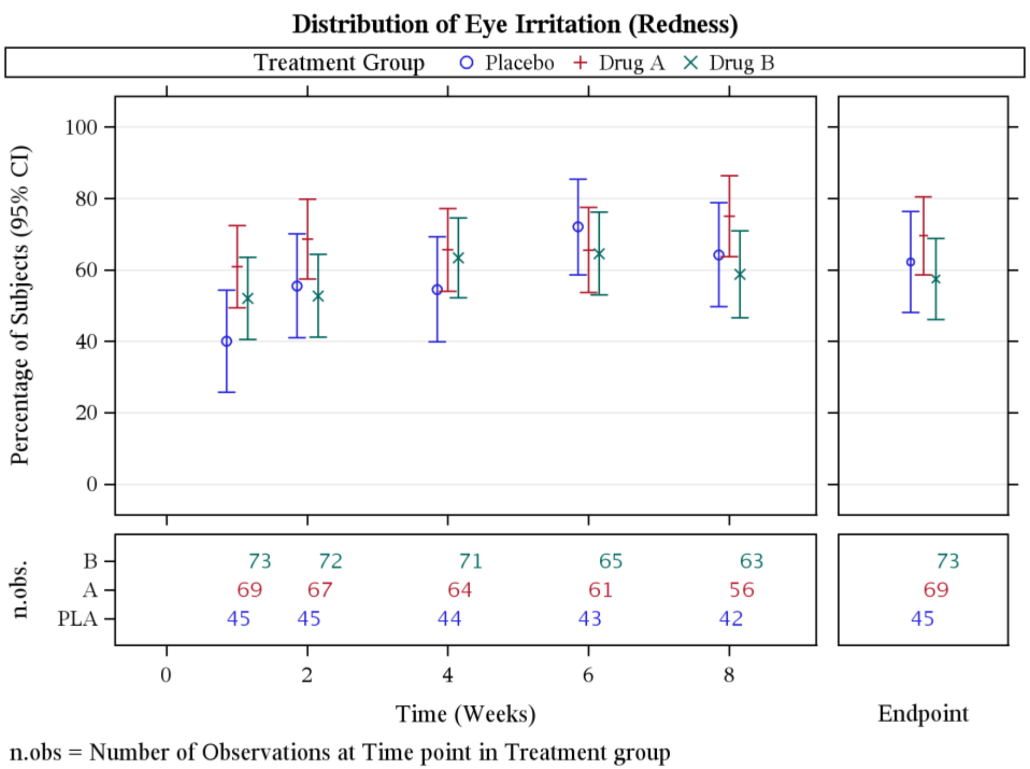
Scale

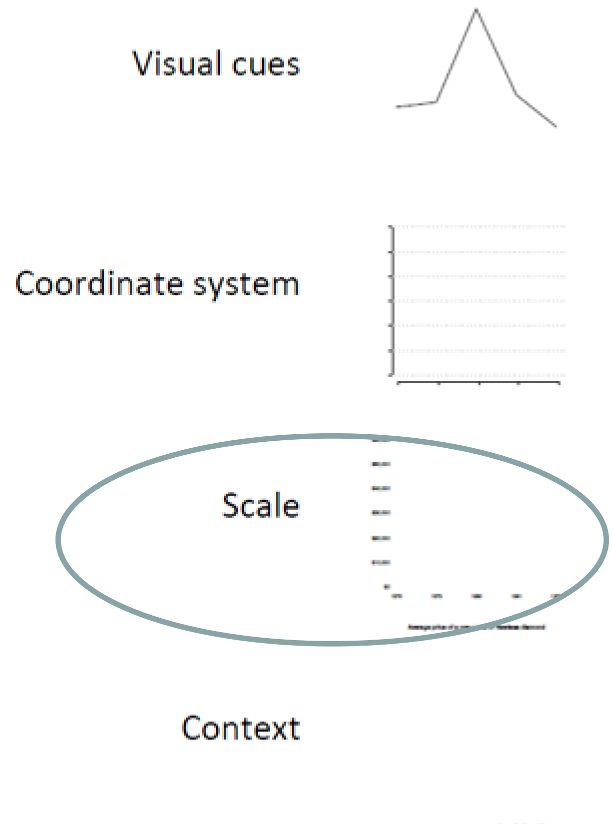
1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…



Scale

1. Quantitative: linear, logarithmic
2. Categorical: nominal, ordinal
3. Temporal: time, date, week, year, etc…

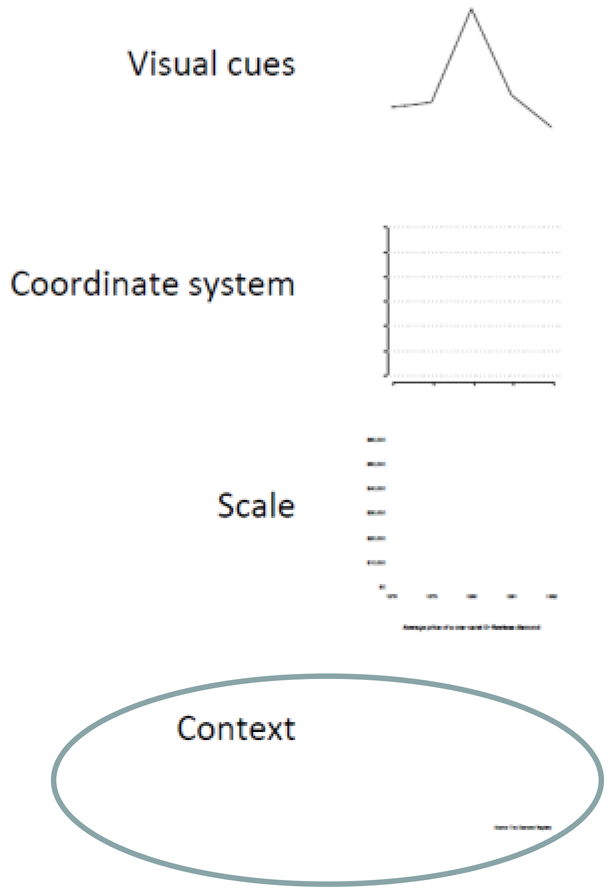




AADS

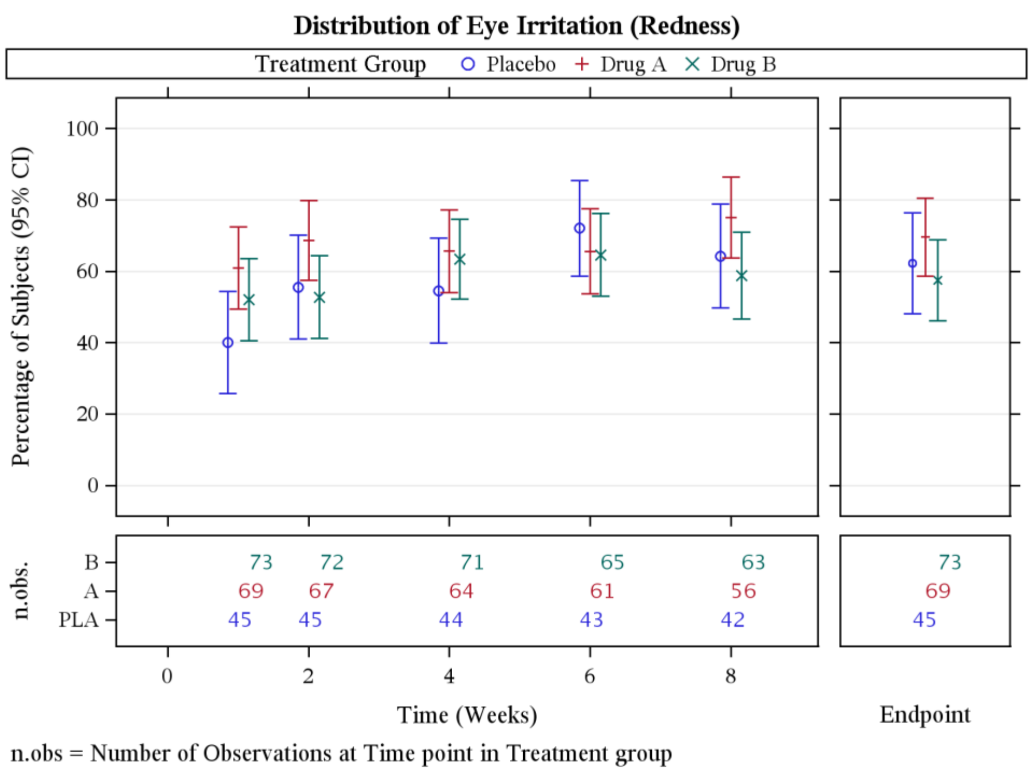
Context

- Title/subtitle
- Legend
- Grid lines
- Highlights
- Trend lines
- Reference lines/areas
- Other important components



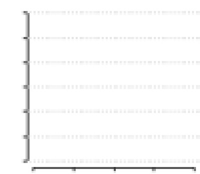
AADS

Context

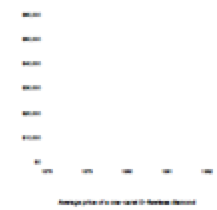


Visual cues

Coordinate system



Scale



Context

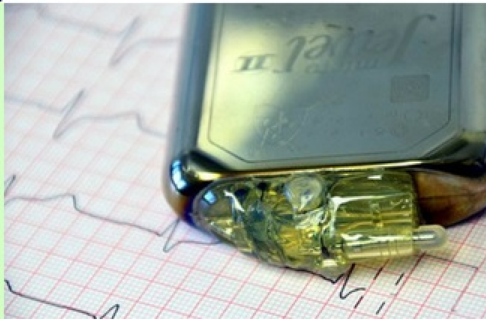
Putting it all together

- Group the data: visually (or explicitly!) segment the data into meaningful subsets.
- Plot the raw data whenever possible.
- Prioritize the data: rank the data by importance
- Sequence the data: give direction for the order in which the data should be read. Storytelling!
- Vertical and horizontal alignment of figures and/or text is important for clear visual flow and to facilitate comparisons (particularly across multiple graphs)
- Use the same scale for similar variables on different graphs to **facilitate comparisons**.

Examples Based on Clinical Data

APPLYING THE PRINCIPLES

Data Visualization In the Regulation of Medical Devices



Zhiheng Xu, Ph.D.
FDA/CDRH

AADS

“The simple graph has brought more information to the data analyst’s mind than any other device.”

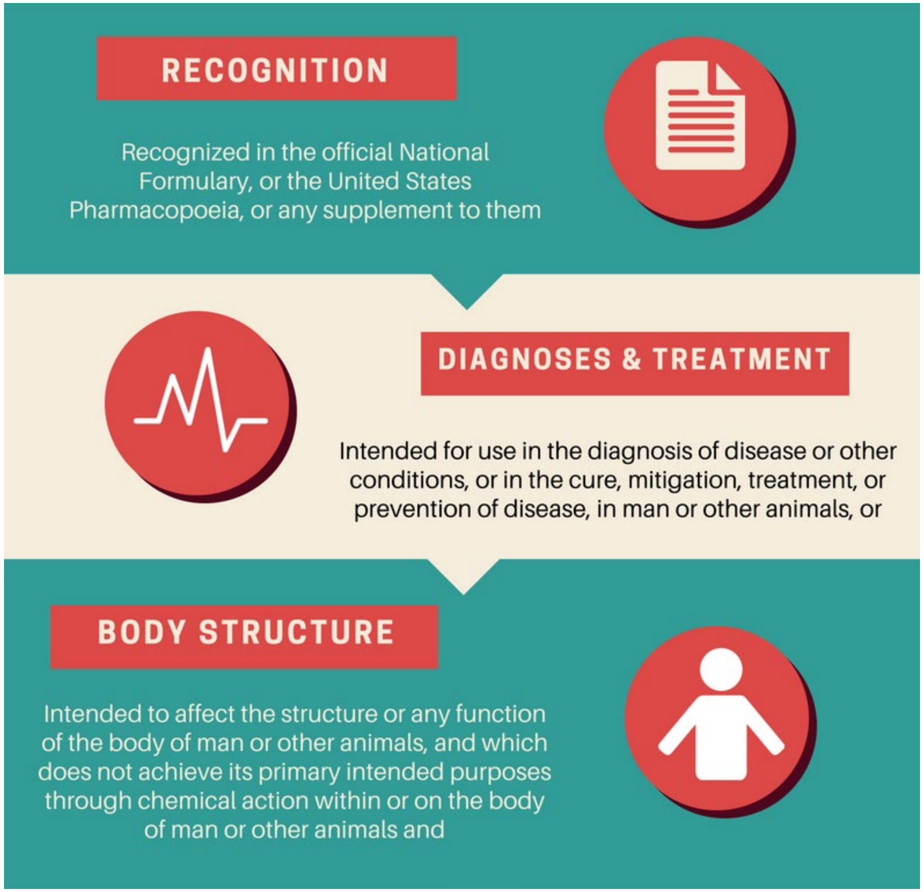
– John Turkey

Outline

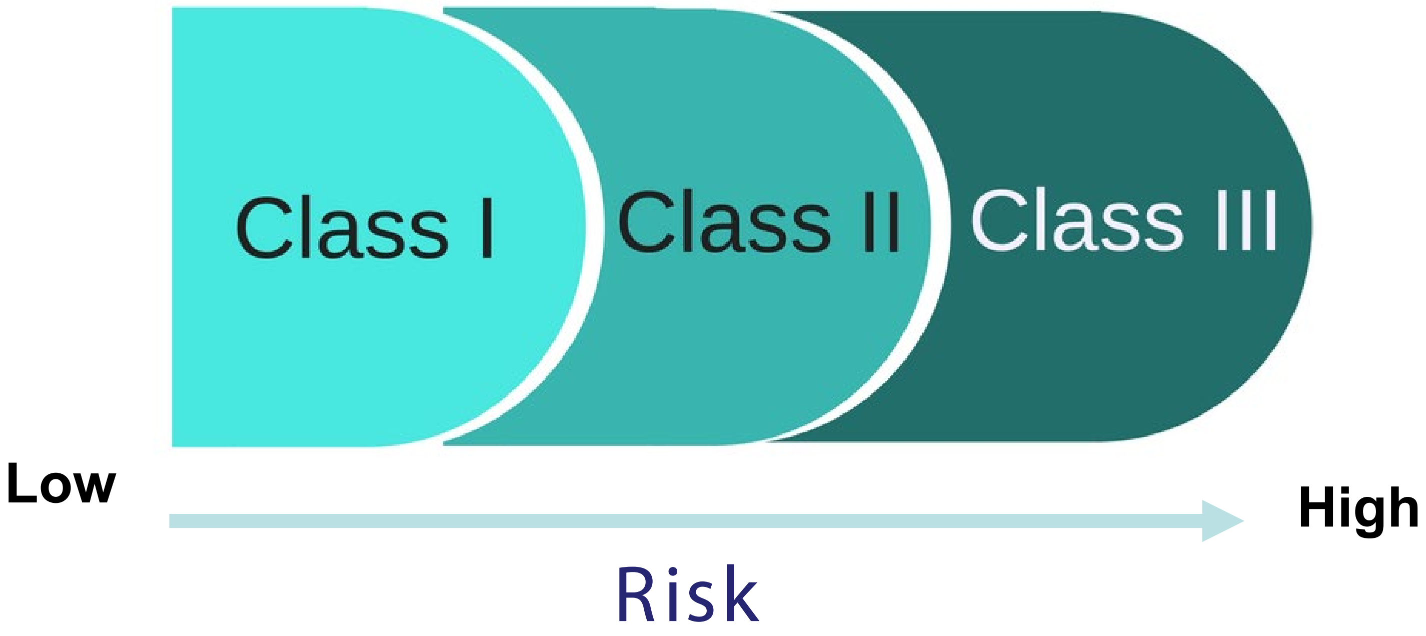
- Medical Device Regulations
- Simple Graphs
  - Scatter Plot
  - Bland-Altman Plot
  - Box plot
  - Venn Diagram

What is Medical Device?

Any instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article.

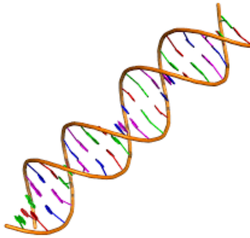


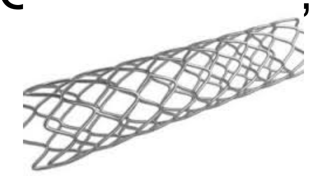
Device Classification

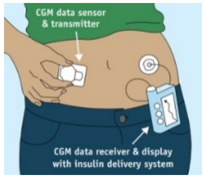


Medical Devices

- Therapeutical Devices
- Diagnostic Devices
  - In-Vivo Diagnostics
  - In-Vitro Diagnostics (IVD)
    - Analytical Validation: Precision, Linearity, etc.
    - Clinical Validation: Accuracy, Method C







METHOD COMPARISON

Test
Device

Predicate
Device
(previously
cleared)

Do they
agree?

Substantial Equivalence (SE)

Regression Analysis

AADS

REGRESSION ANALYSIS

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Test Device

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

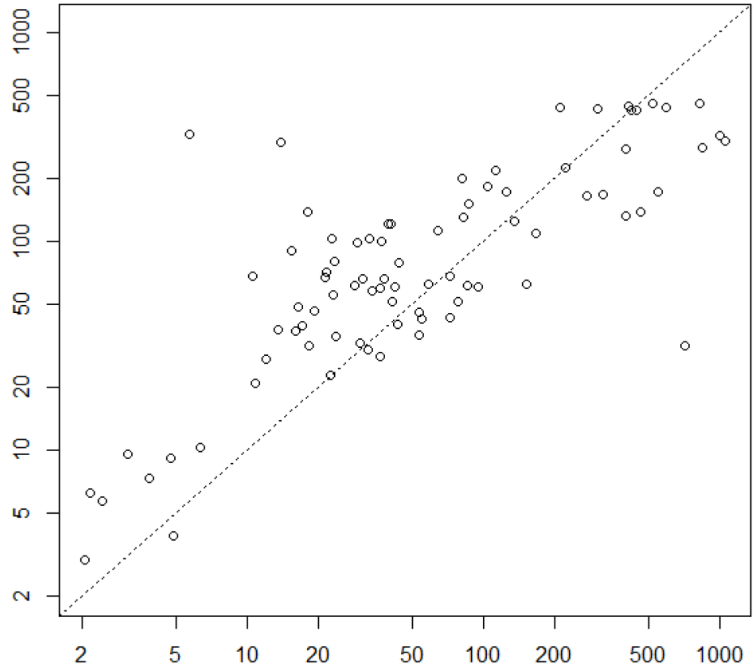
Predicate Device (previously cleared)

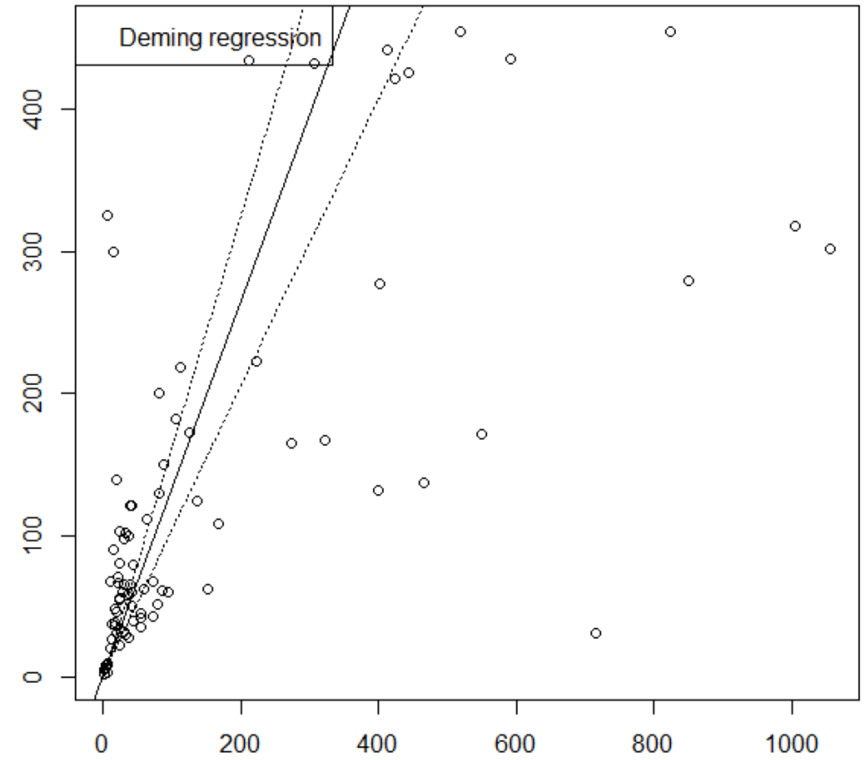
$$H_0: \ |\beta_0| > \delta, \quad |\beta_1 - 1| > \lambda$$

$$H_a: \ |\beta_0| \le \delta, \quad |\beta_1 - 1| \le \lambda$$

- Ordinary Least Squares (OLS)
- Deming Regression
- Passing-Bablok Regression

Scatter Plot





β_0=1.81 (95% CI: -0.91-4.66)

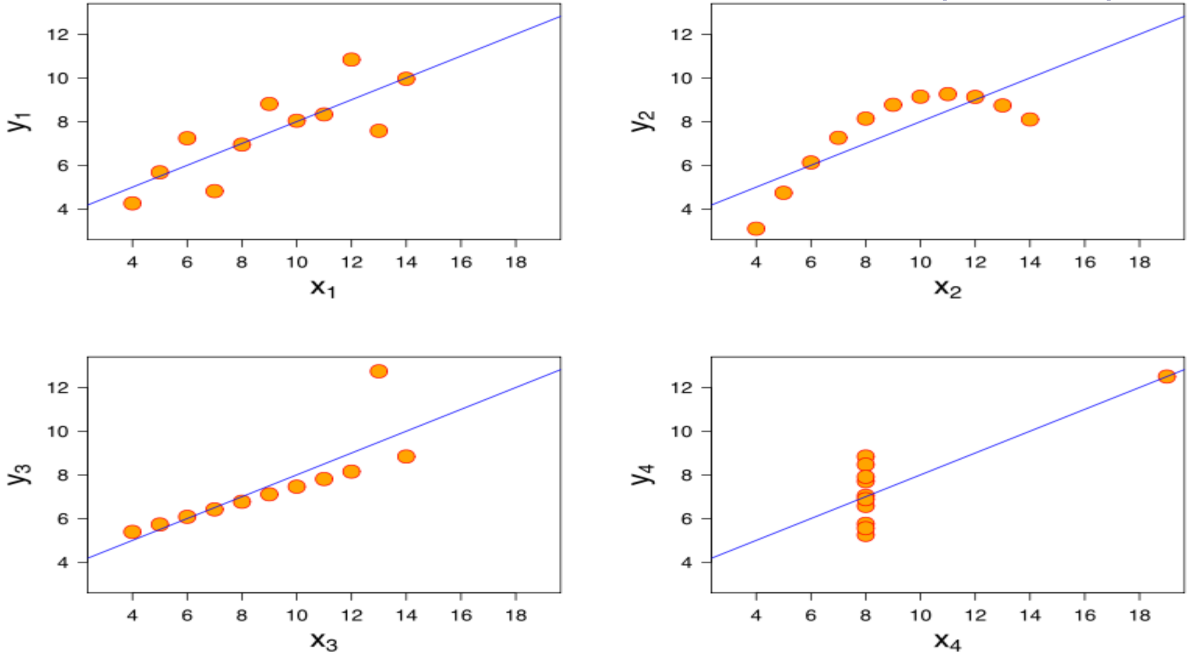
β_1=1.31 (95% CI: 1.01-1.57)

AADS

Visualization Tips

Use the same scales on both X and Y axis

Four datasets with same correlation (0.816)

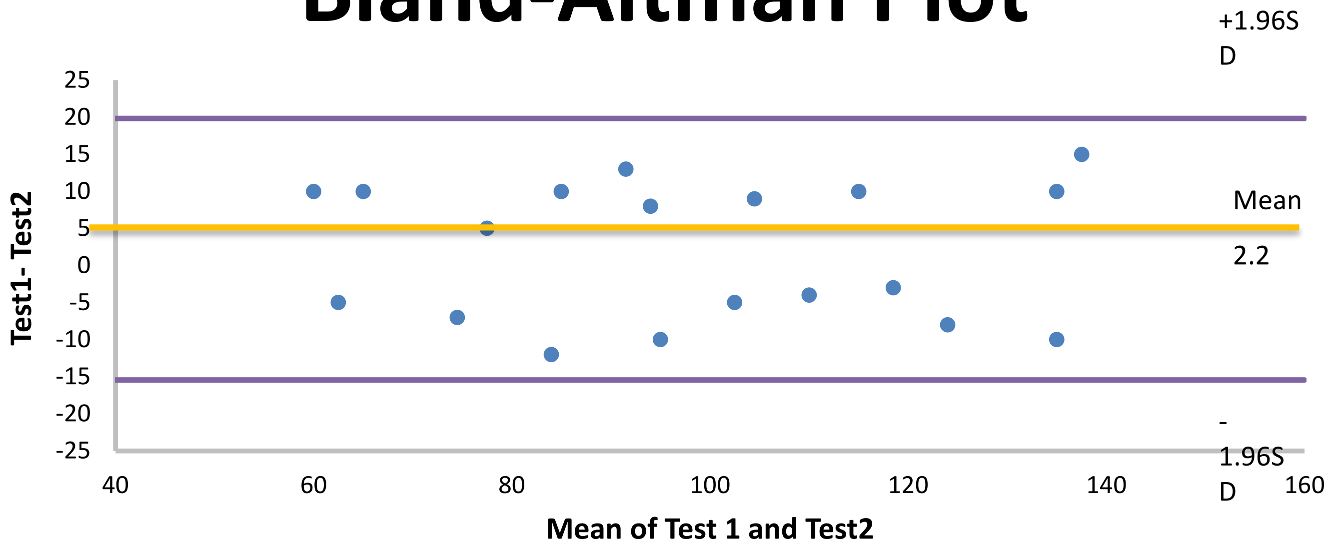


From: http://en.wikipedia.org/wiki/Correlation_and_dependence

CORRELATION IS NOT AGREEMENT

AADS

Bland-Altman Plot



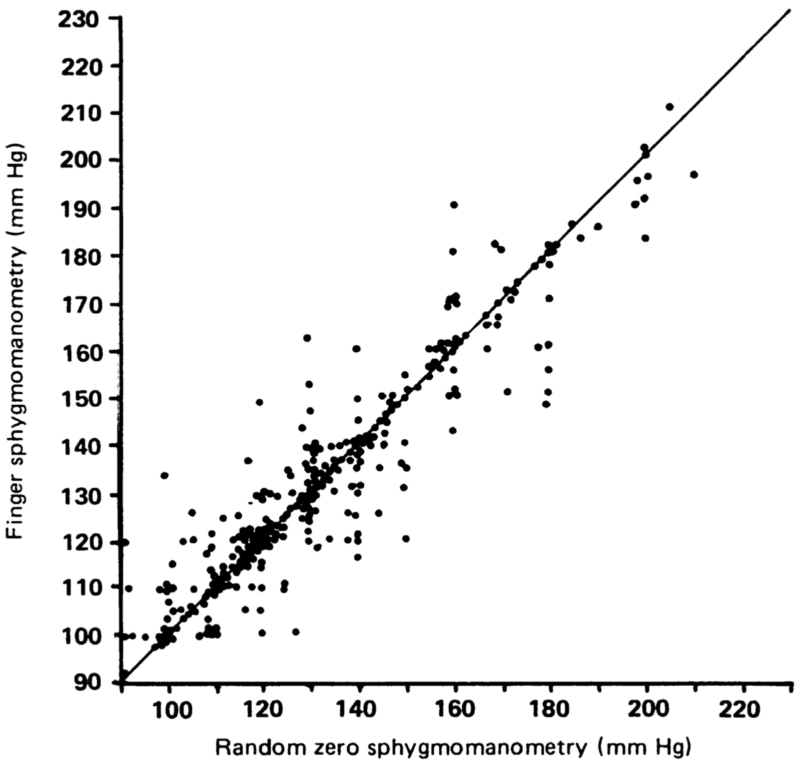
Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1(8476): 307-10.

TOP 3

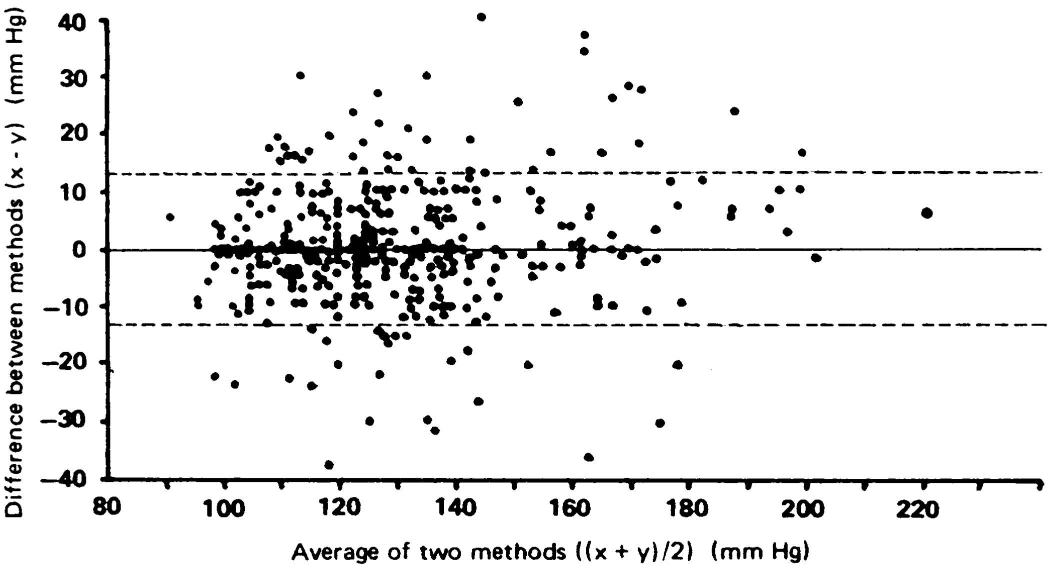
Most Cited Stat. Paper

Bland-Altman Plot

Why not plot Test 1 vs. Test 2 directly?



Close A, Hamilton G, Muriss S. Finger systolic pressure: its use in screening for hypertension and monitoring. *Brit Med J* 1986; **293**: 775-778.



Close A, Hamilton G, Muriss S. Finger systolic pressure: its use in screening for hypertension and monitoring. *Brit Med J* 1986; **293**: 775-778.

Bland-Altman Plot

Why plot difference against average, not standard?

S: standard measurement
T: test measurement

$Var(S) = \sigma_S^2 \quad , Var(T) = \sigma_T^2 \quad , \mathrm{Corr}(\mathrm{S}, \mathrm{T}) = \rho$

AADS

Difference vs. Standard

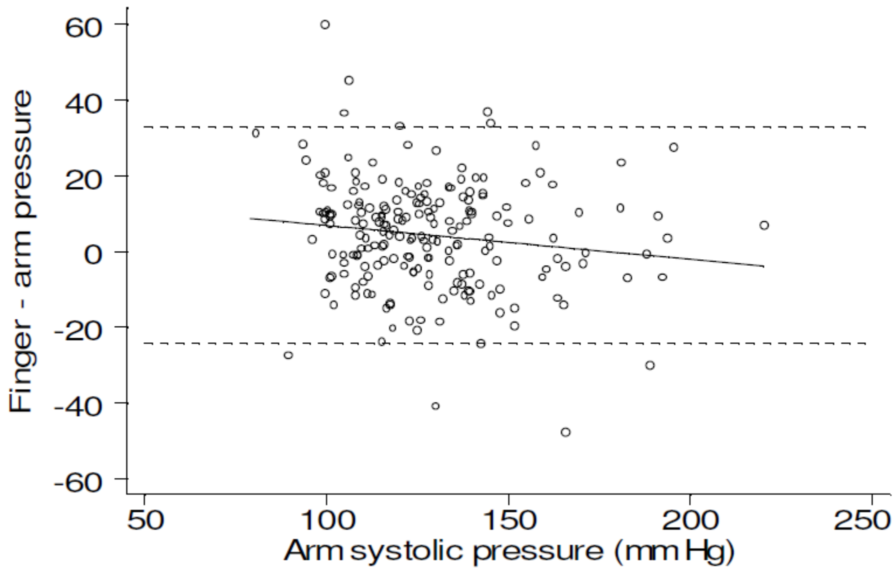


Figure 3. Difference against standard measurement, with 95% limits of agreement (broken lines) and regression line

$$Corr(T - S, S) = \frac{\rho\sigma_T - \sigma_S}{\sqrt{\sigma_S{}^2 + \sigma_T{}^2 - 2\rho\sigma_T\sigma_S}}$$

Bland, JM, Altman, DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1996; 346: 1085-87.

AADS

Difference vs. Test

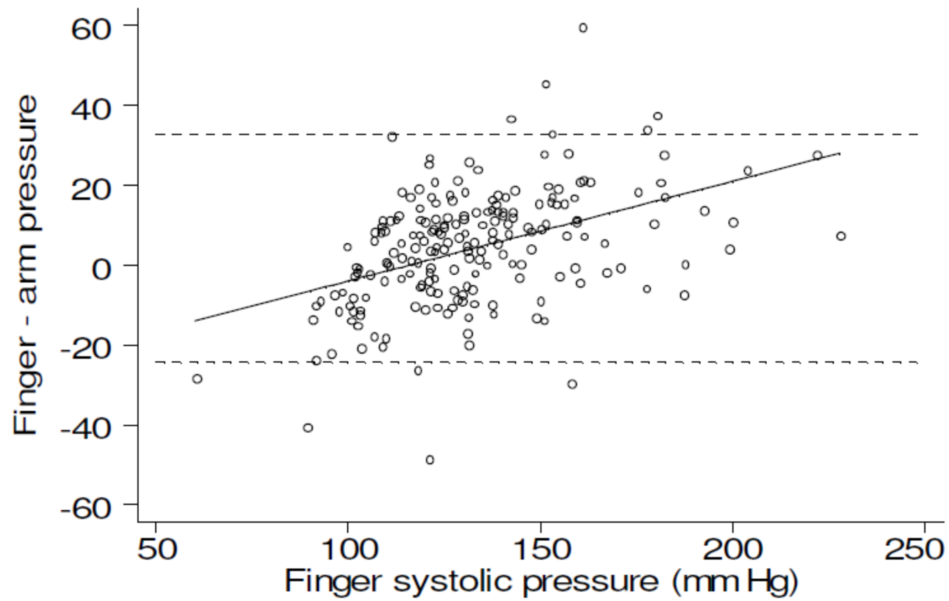
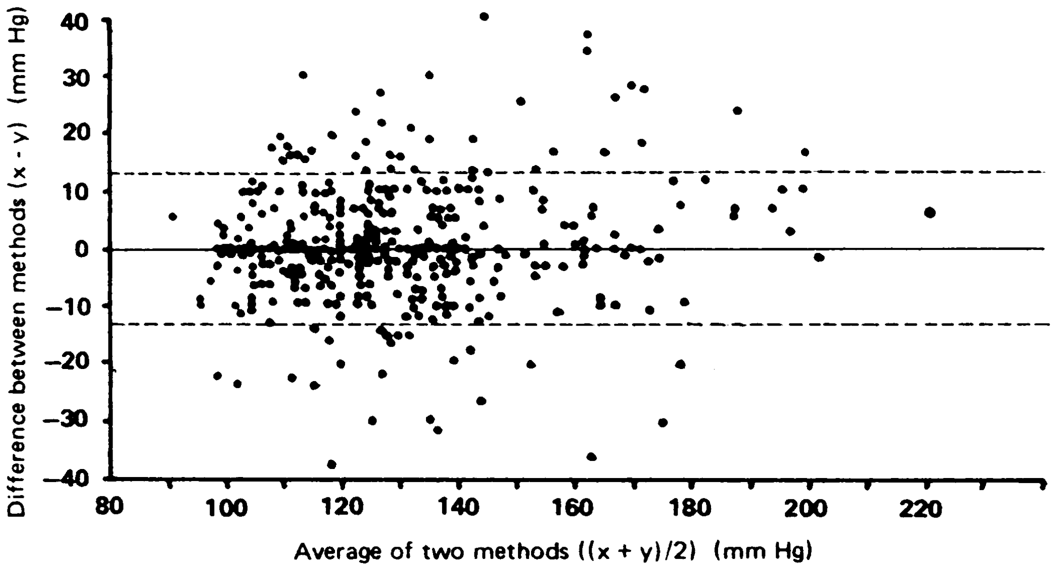


Figure 4. Difference against test measurement, with 95% limits of agreement (broken lines) and regression line

$$Corr(T-S,T) = \frac{\sigma_T - \rho\sigma_S}{\sqrt{\sigma_S{}^2 + \sigma_T{}^2 - 2\rho}}$$

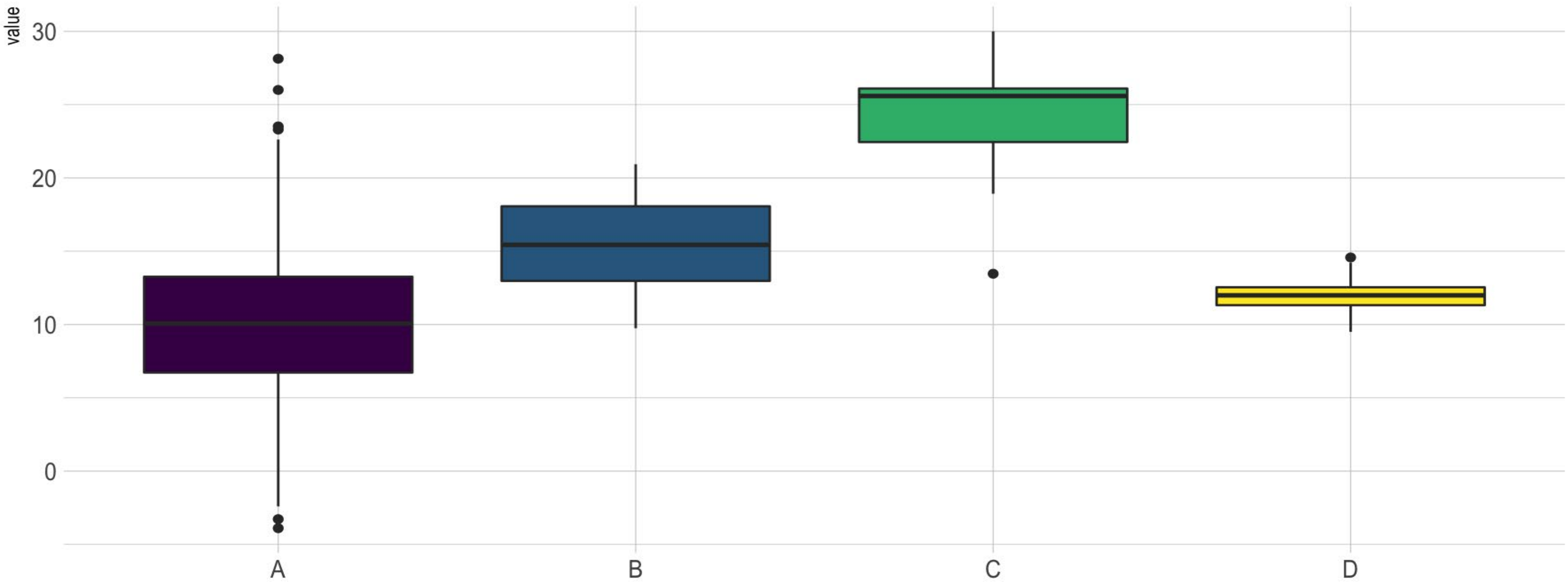
Bland, JM, Altman, DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1996; 346: 1085-87.

Difference vs. Average



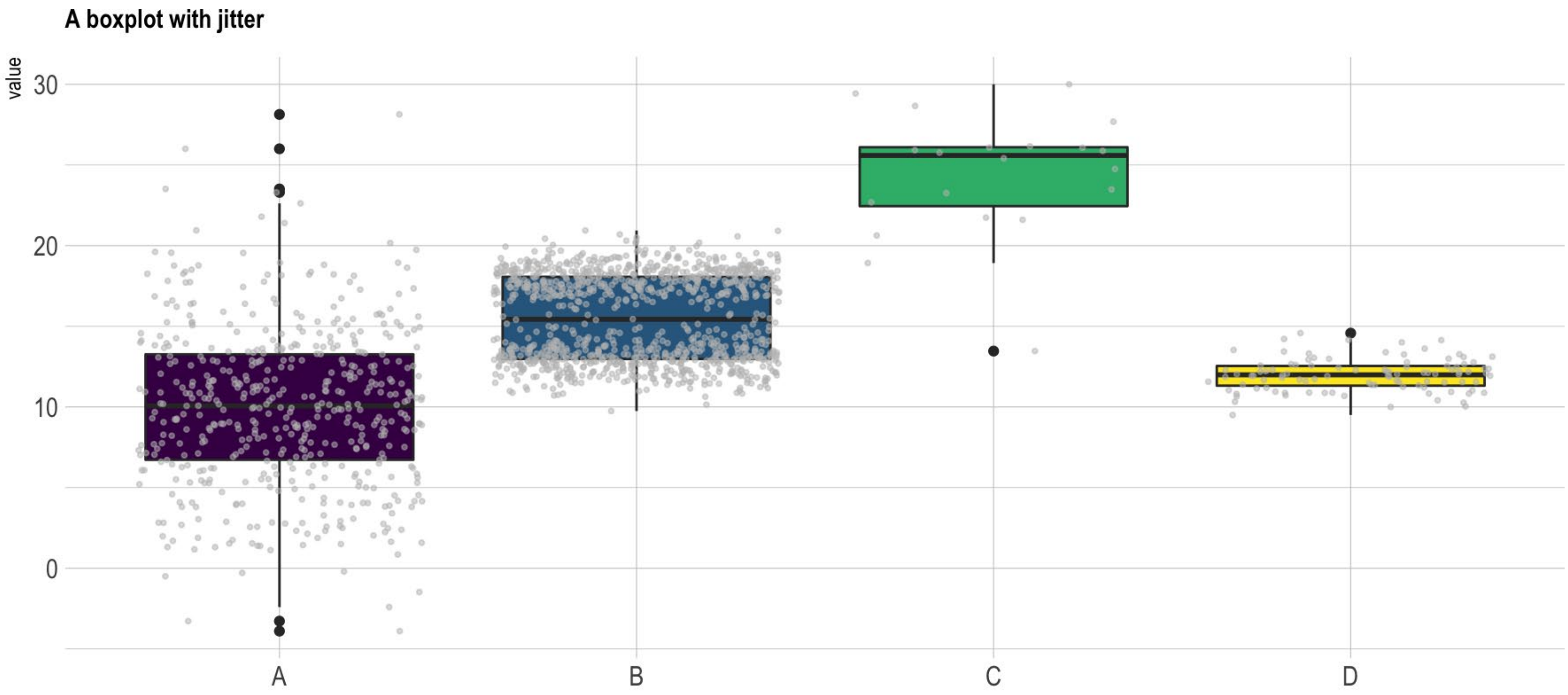
Close A, Hamilton G, Muriss S. Finger systolic pressure: its use in screening for hypertension and monitoring. *Brit Med J* 1986; **293**: 775-778.

Box Plot



https://www.data-to-viz.com/caveat/boxplot.html

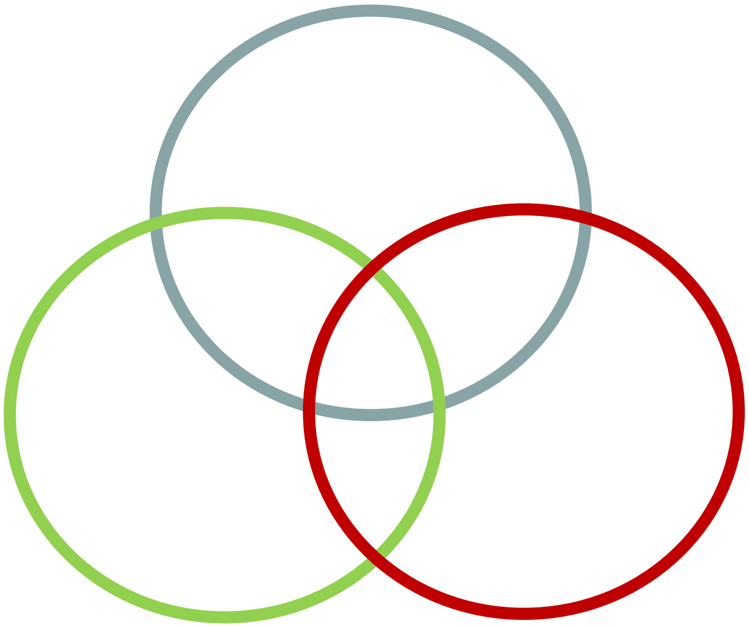
AADS



https://www.data-to-viz.com/caveat/boxplot.html

AADS

Venn Diagram



Human Genetic Tests

- 3 DNA extraction methods
  - RES (resection)
  - CNB (core needle biopsy)
  - FNA (fine needle aspiration)
- Equivalency

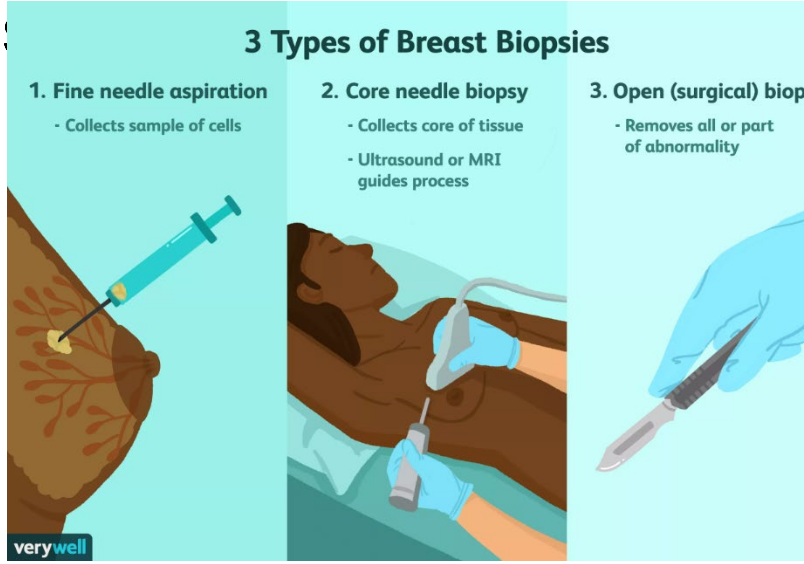
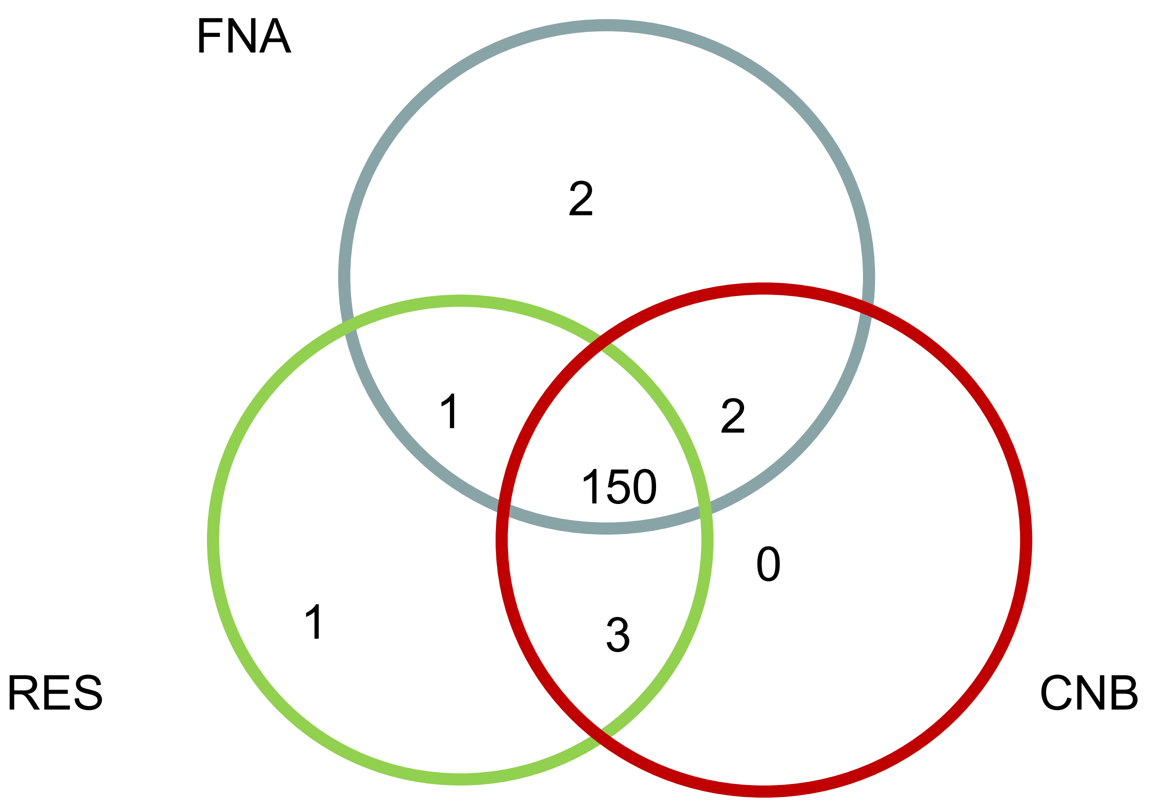


Image from https://www.verywellhealth.com/open-surgical-breast-biopsy-429949

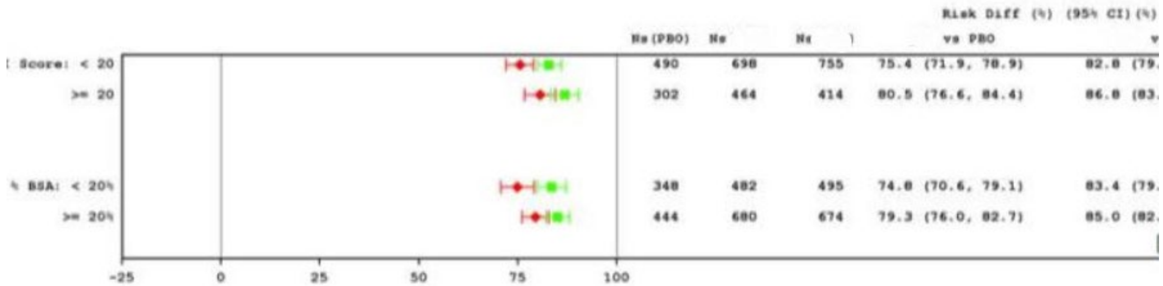
AADS

Comparison of RES, CNB and FNA valid calls

Acquisition Methods	Agreement	Freq	Percentage
CNB vs. FNA	OA	152/155	98.1%
	PPA	30/32	93.8%
	NPA	122/123	99.2%
CNB vs. RES	OA	153/155	98.7%
	PPA	31/32	96.9%
	NPA	122/123	99.2%
FNA vs. RES	OA	151/155	97.4%
	PPA	30/33	90.9%
	NPA	121/122	99.2%


FNA
2
1
2
150
0
1
3
RES
CNB

Forest plot
of and Response Rates at Week 16 (NRI) by Subgroups
reat Population
- Blinded Treatment Dosing Period



ns: PBO = Placebo; mg (;
atients in each subgroup; CI = confidence interval.
ntage of response is calculated by n/Ns*100%
dence intervals are constructed using the simple asymptotic method, without continuity correction (that is, norm
on to the binomial distribution).

taset location/dataset name.sas7bdat
ogram location/program name.sas
put location/output name.rtf

Notes:
s include all the subgroups mentioned in demographic, geographic region, baseline disease severity, and other pa
tic subgroup analyses. Include ADA40Q2W in above display.
Risk Diff(%) (95%CI)(%)" to "Difference (95% CI vs. PBO(%)".

Forest Plot

Subgroup Analysis

AADS

Forest Plot

Interactive Subgroup Screening Tool

AADS

Summary

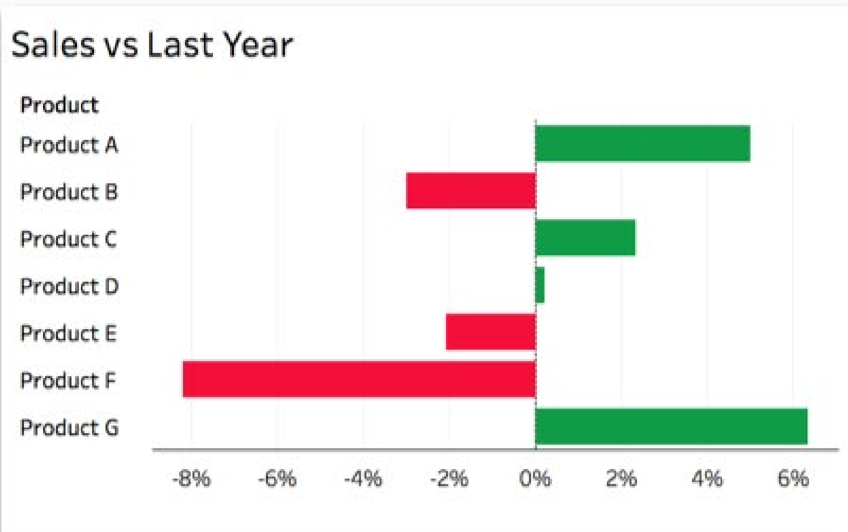
Data Visualization

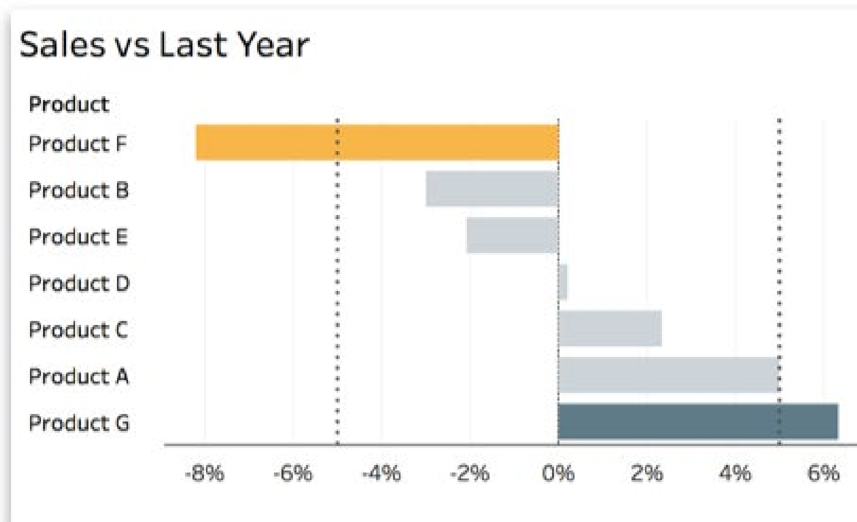
- Effective tool
- Understand statistics behind graphs
- Avoid graph pitfalls

A picture is worth a thousand words!

Telling the Story

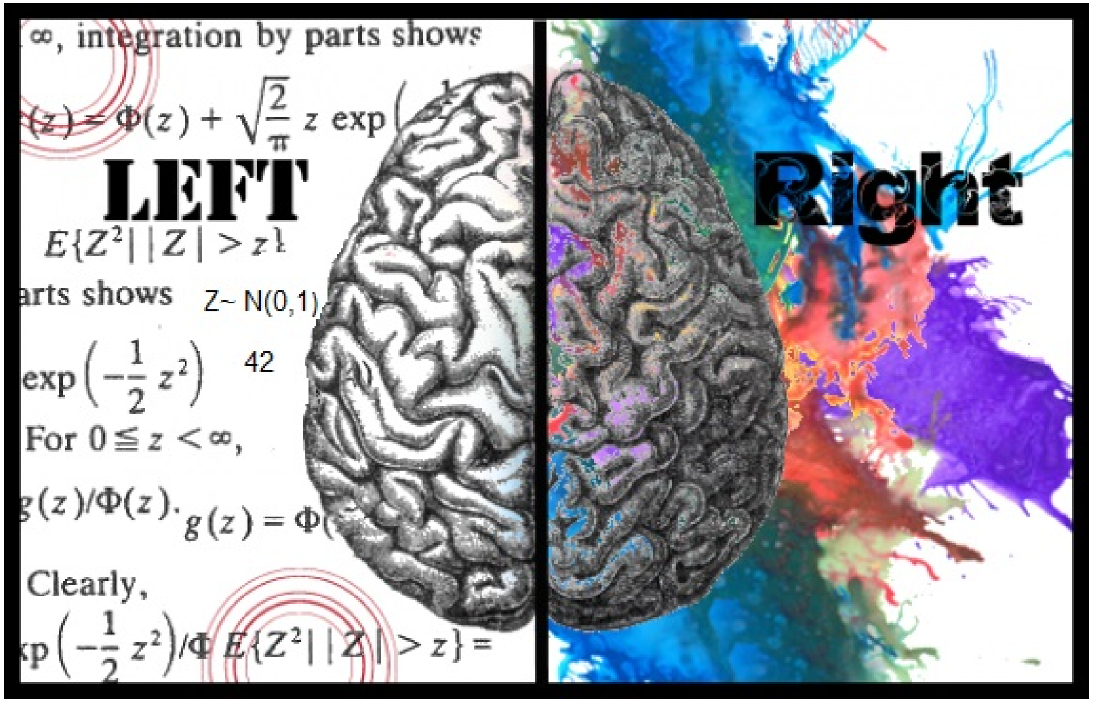
[Scrollytelling applied to clinical data]





Visual Analytics: Where Art Meets Science

- 0.1 Second
- 1 Second
- 10 Seconds



Summary

Suggested additional reading

- **Visualizing Data** by William, Cleveland. 1993
- **Visualization Analysis & Design** by Tamara Munzner
- Susan Duke, Fabrice Bancken, Brenda Crowe, Mat Soukup, Taxiarchis Botsis an Richard Forshee, *Seeing is believing: Good graphic design principles for medical research*, **Statistics in Medicine, Special Edition Accepted** May 2015**.**
- **The Grammar of Graphics (Statistics and Computing) 2nd Edition** by Leland Wilkinson (Author), D. Wills (Contributor), D. Rope (Contributor), A. Norton (Contributor), R. Dubbs (Contributor). 2005
- http://www.perceptualedge.com/
- http://flowingdata.com/