COMP 5360 / MATH 4100
Introduction to Data Science
Visualization

Bei Wang Phillips

THE UNIVERSITY OF UTAH

Lecture 15: (Some) Principles of Data Visualization

Slides modified from those done by Alex Lex

The purpose of computing is insight, not numbers.

- Richard Wesley Hamming

The purpose of visualization is insight, not pictures.

- Card, Mackinlay, Shneiderman

Visualization Definition

Visualization is the process that **transform**s
(abstract) **data** into
interactive graphical representations for the purpose of
exploration, confirmation, or presentation.

Good Data Visualization

… makes data accessible

… combines strengths of humans and computers

… enables insight

… communicates

Why Visualize?

To inform humans about complex situations:
Communication

How is ahead in the election polls?

When questions are not well defined:
Exploration

What is the structure of a terrorist network?

Which drug can help patient X?

Why Visualize?

Communication

Know about the data: communicate to others

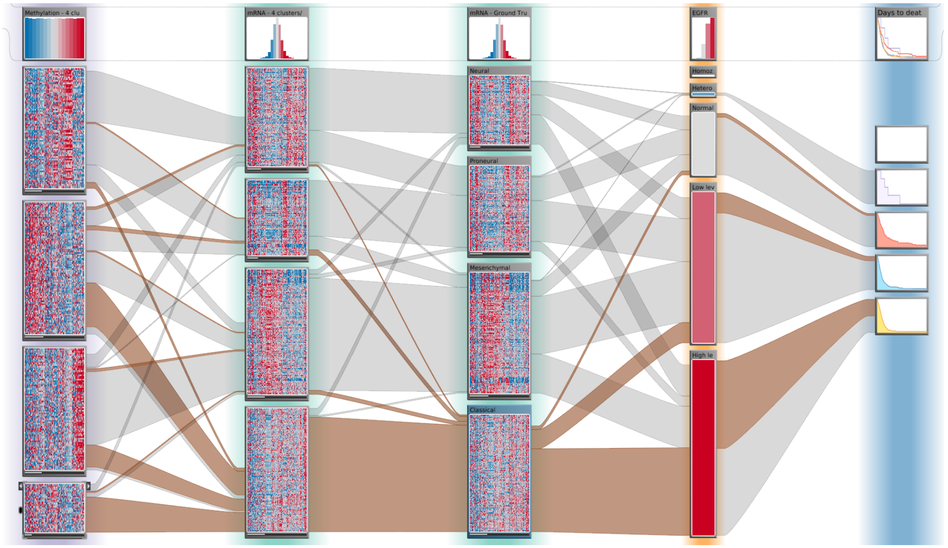
Exploration

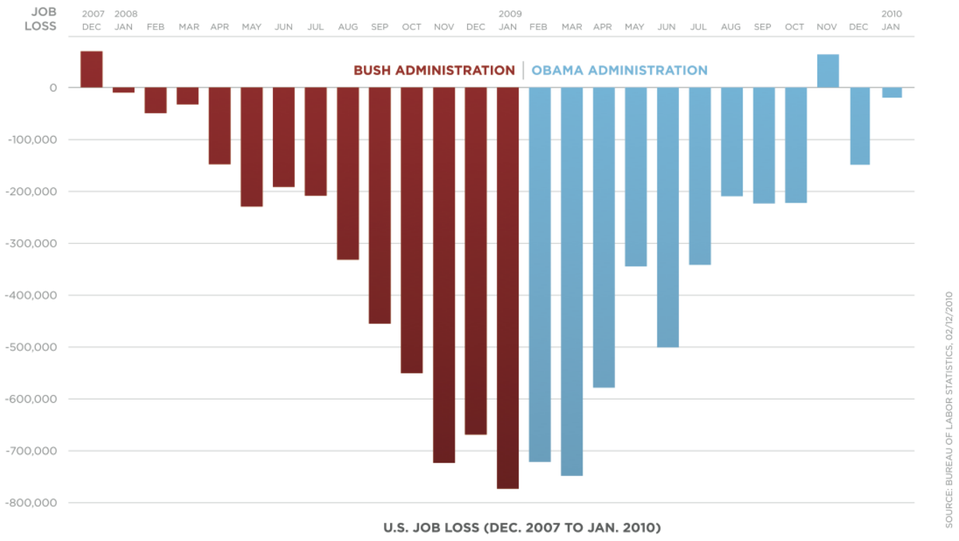
Results unknown

Questions unknown

Purpose of Visualization

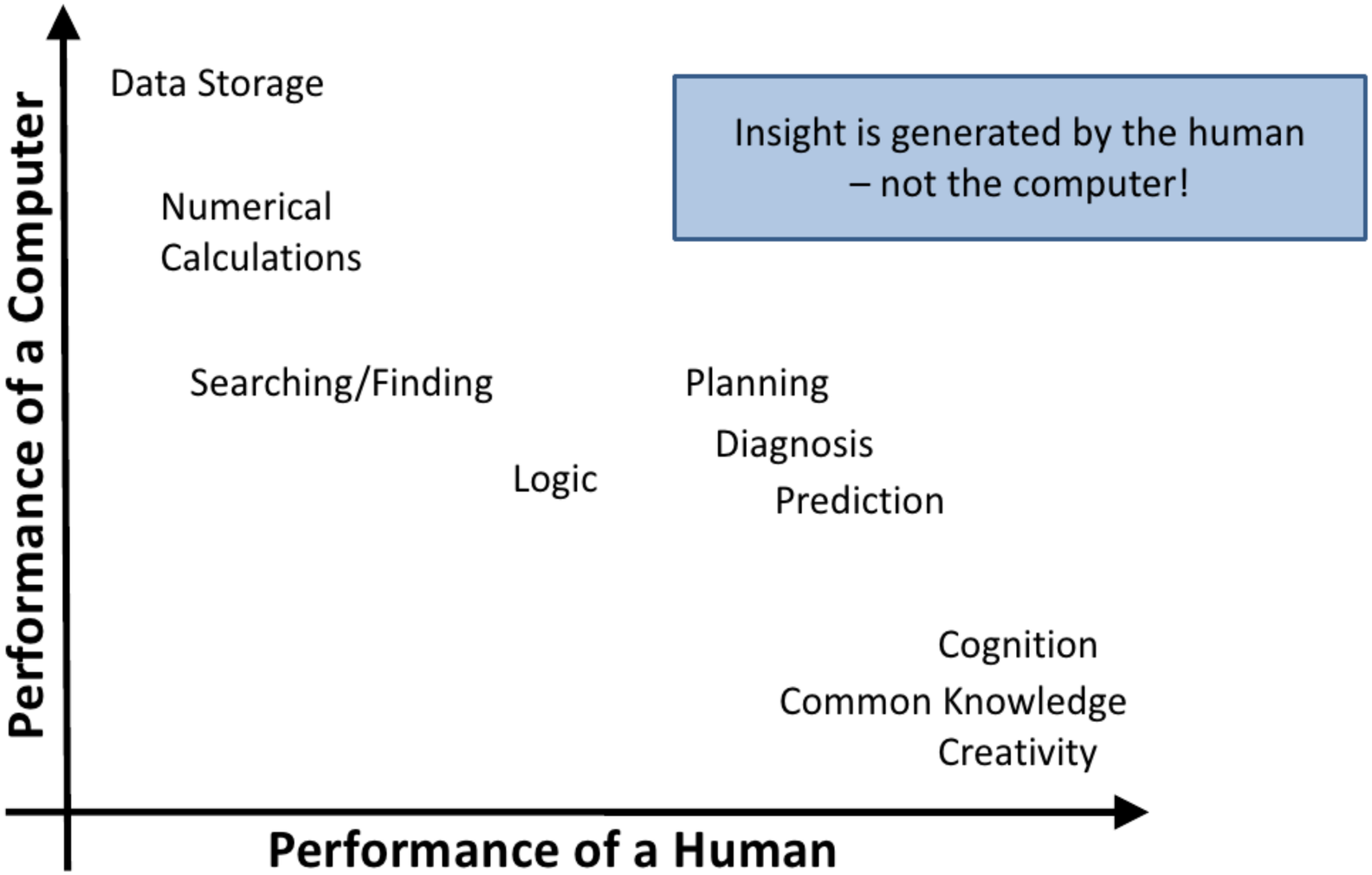
[Obama Administration]





Open Exploration ⟷ Confirmation ⟷ Communication

The Ability Matrix



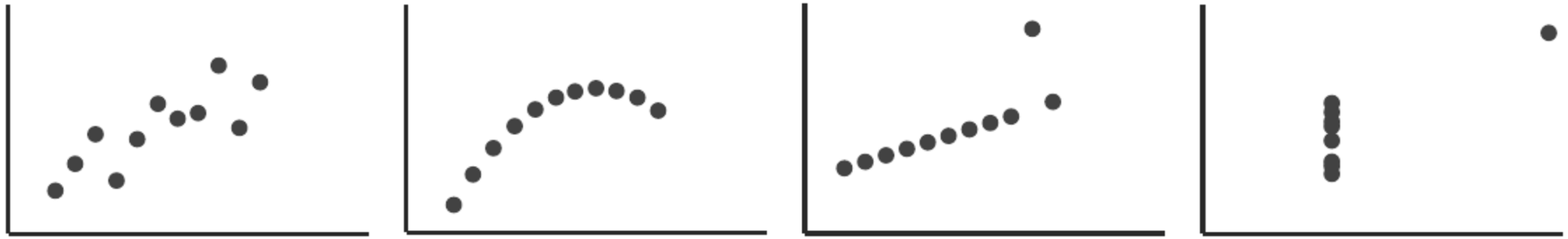
Insight is generated by the human – not the computer!

Why not just use Statistics?

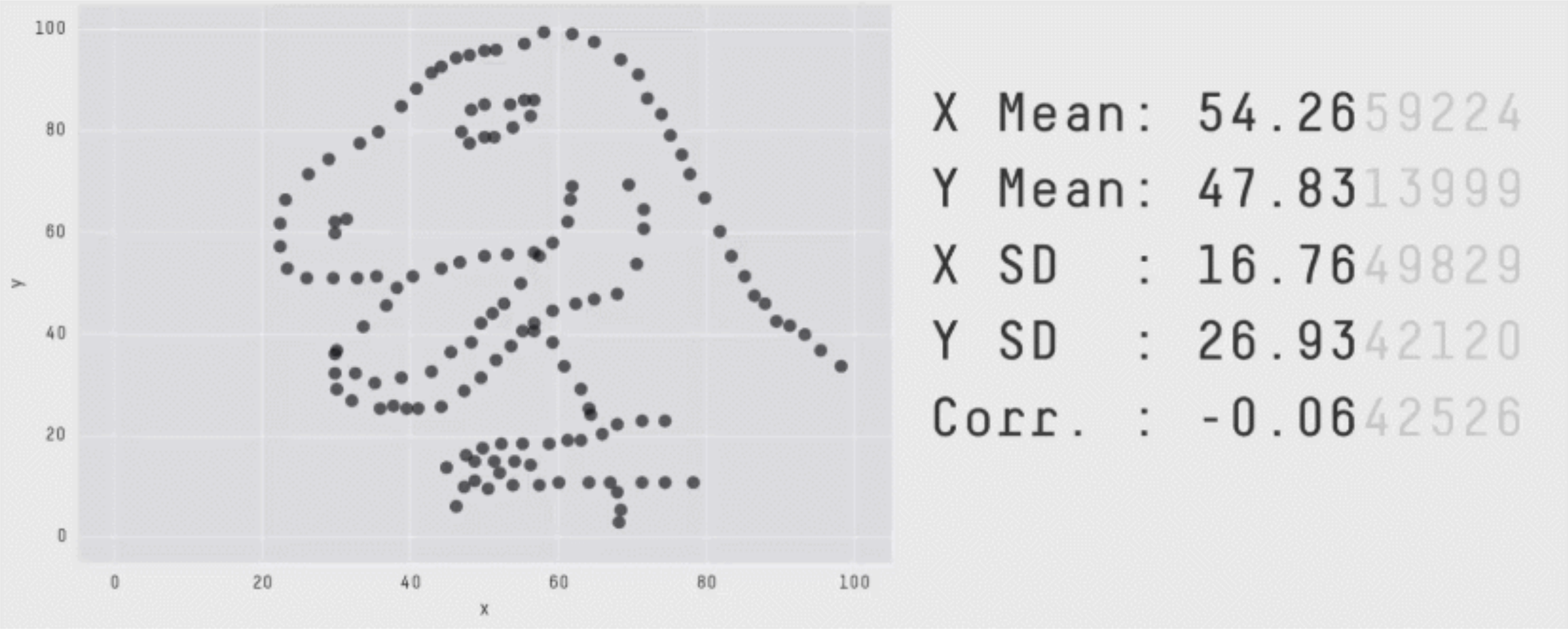
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.0	10	9.1	10	7.4	8	6.5
8	6.9	8	8.1	8	6.7	8	5.7
13	7.5	13	8.7	13	12.	8	7.7
9	8.8	9	8.7	9	7.1	8	8.8
11	8.3	11	9.2	11	7.8	8	8.4
14	9.9	14	8.1	14	8.8	8	7.0
6	7.2	6	6.1	6	6.0	8	5.2
4	4.2	4	3.1	4	5.3	19	12.
12	10.	12	9.1	12	8.1	8	5.5
7	4.8	7	7.2	7	6.4	8	7.9
5	5.						6.8

Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 + 0.500x

Anscombe’s Quartett



Mean x: 9 y: 7.50
Variance x: 11 y: 4.122
Correlation x – y: 0.816
Linear regression: y = 3.00 +



Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, CHI 2017, Justin Matejka, George Fitzmaurice

Visualization =

Human Data Interaction

Visualization in the Data Science Process

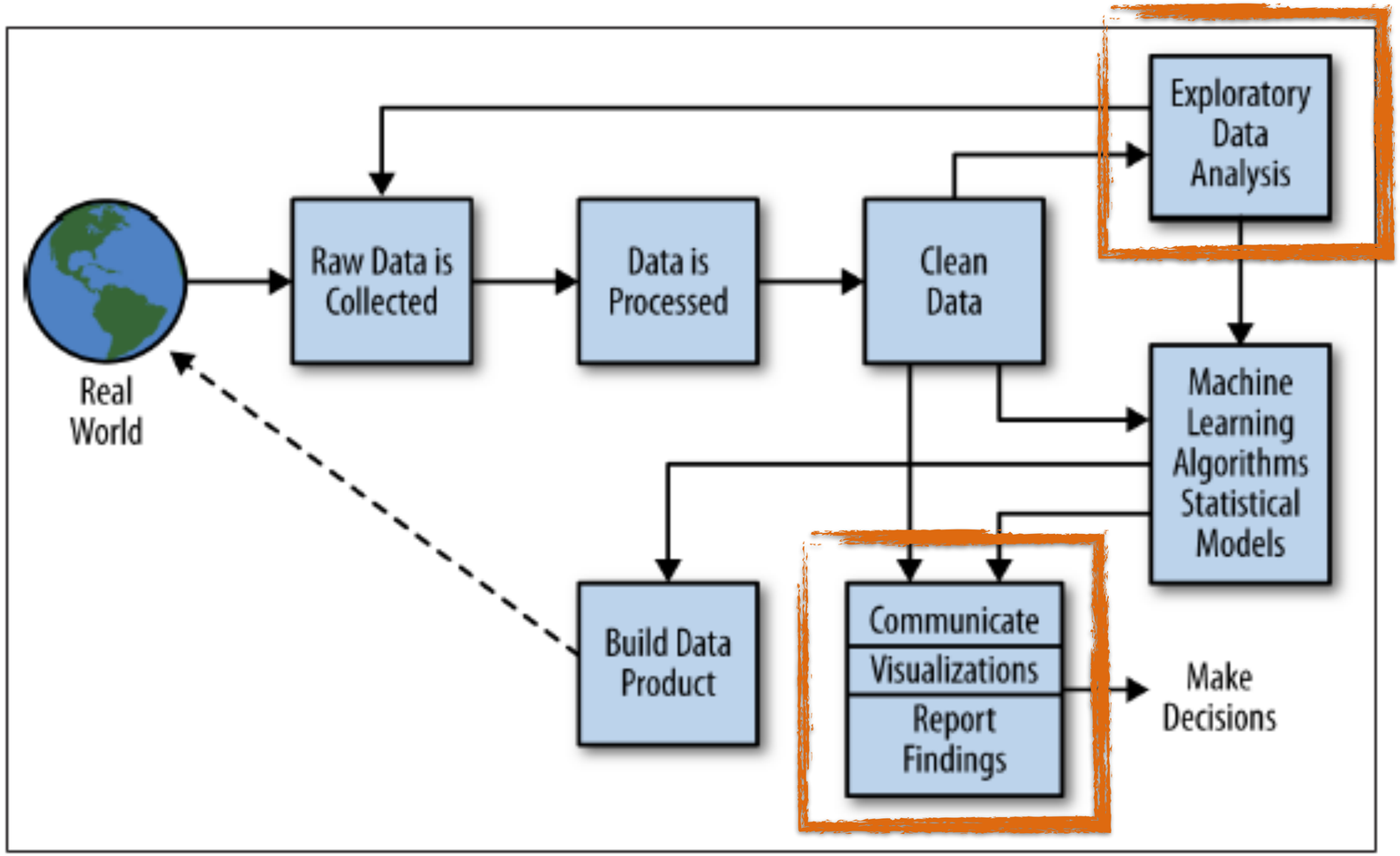


Figure 2-2. The data science process

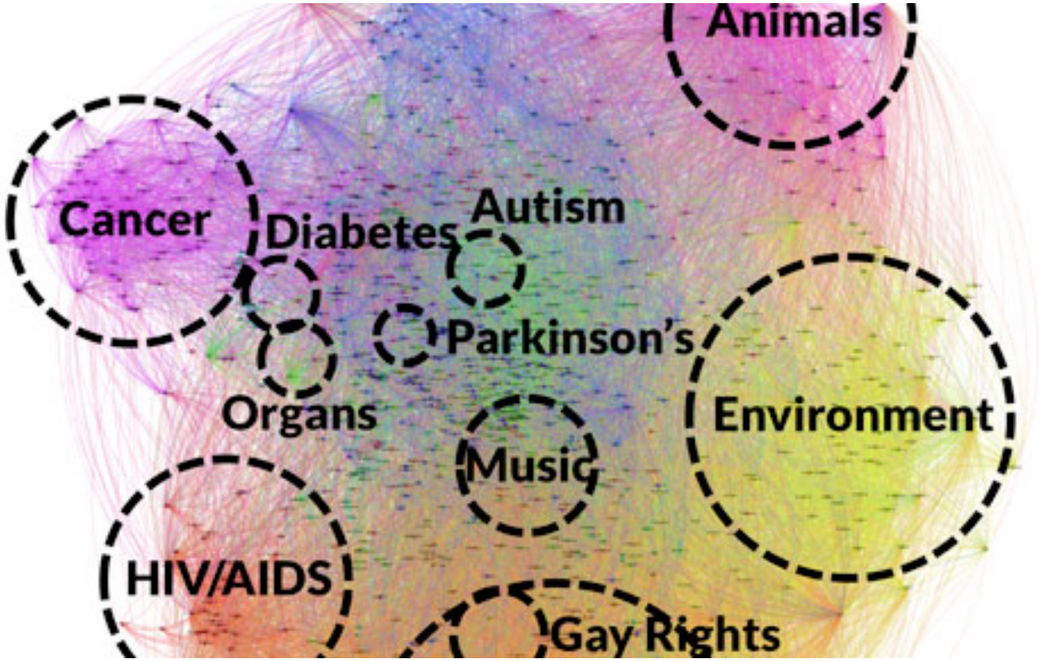
Why Humans?

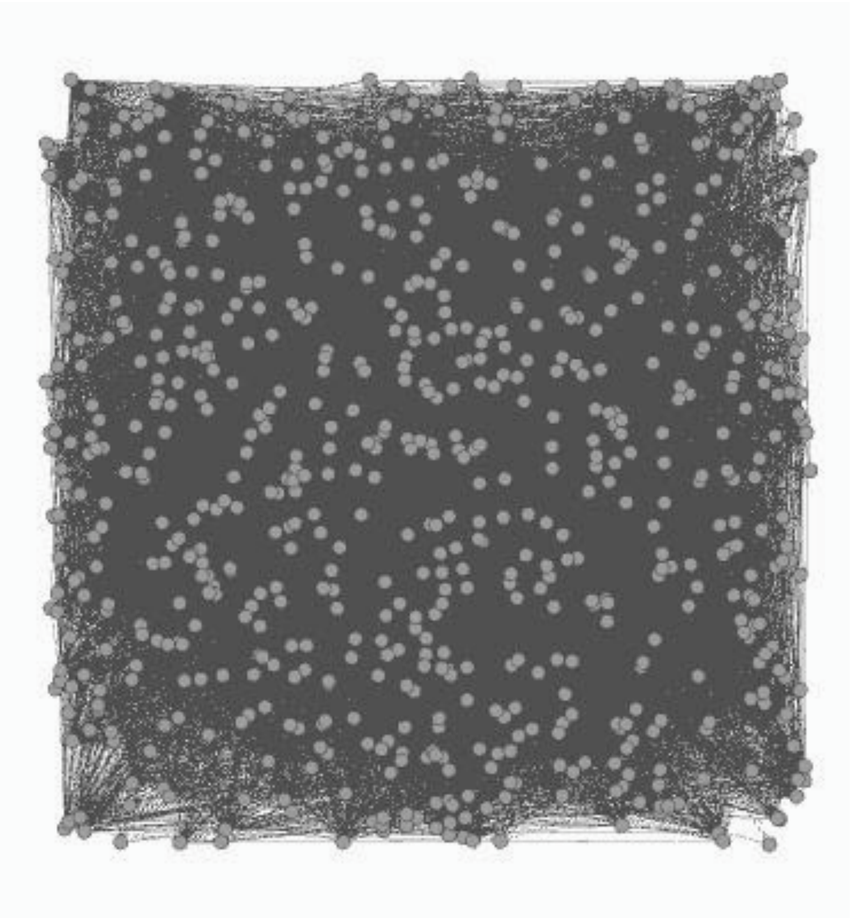
Leveraging human capabilities

- **Pattern Discovery**: clusters, outliers, trends
- **Contextual Knowledge**: expectations for dataset, explanations for patterns
- **Action**: humans learn and take action

But: we also have to **design for Humans and their limitations**

Not everything that can be drawn can be read!





Limits of Cognition: change blindness

A phenomenon of visual perception that occurs when a stimulus undergoes a change without this being noticed by its observer.

Daniel J. Simons and Daniel T. Levin, Failure to detect changes to people during a real world interaction, 1998

History: John Snow

1854 Broad Street cholera outbreak in London
Cause for cholera was thought to be "bad air". Didn’t know about germs

A physician, John Snow, had theory that it’s in the water.
Studied patterns of outbreak and identified a water pump.
Persuaded the authorities to disable the well pump by removing its handle.

E. Tufte, Visual Explanations, 1997

John Snow: disease map

John Snow, 1854

Snow's map is one of the earliest and most powerful uses of data **visualization**.

Each black bar represents one death. When there are multiple deaths at the same address, the bars corresponding to those deaths are stacked on top of each other. The black discs mark the locations of water pumps. The map displays a striking revelation—the deaths are roughly clustered around the Broad Street pump.

E. Tufte, Visual Explanations, 1997

Marks & Channels

Marks & Channels

Marks

Encode existence of an item (a row in a table).

Point, bar, line, etc.

Channels

Encode magnitude or category of item.

Position, size, height, color, etc.

Types of Channels

Magnitude Channels

How much?

Position

Length

Saturation …

Identity Channels

What? Where?

Shape

Color (hue)

Spatial region …

Ordinal & Quantitative Data

Categorical Data

Channels: Expressiveness Types and Effectiveness Ranks

Magnitude Channels: Ordered Attributes

Position on common scale

Position on unaligned scale

Length (1D size)

Tilt/angle

Area (2D size)

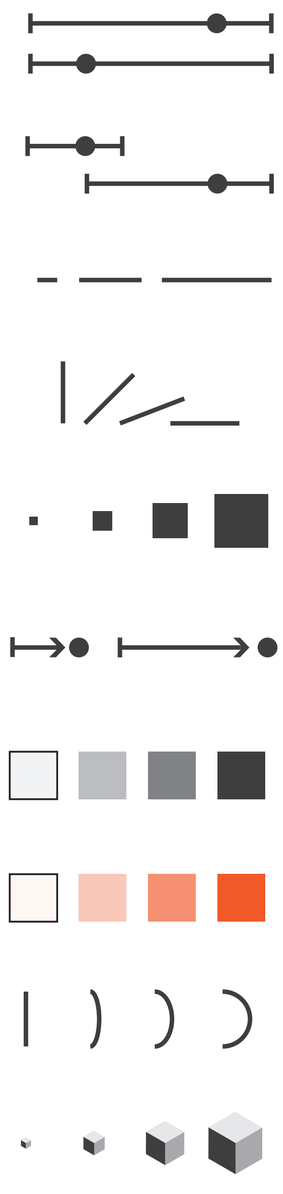
Depth (3D position)

Color luminance

Color saturation

Curvature

Volume (3D size)



Most

Effectiveness

Least

Same

Same

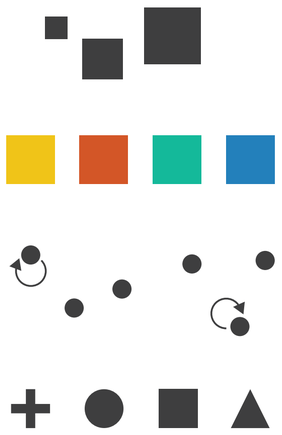
Identity Channels: Categorical Attributes

Spatial region

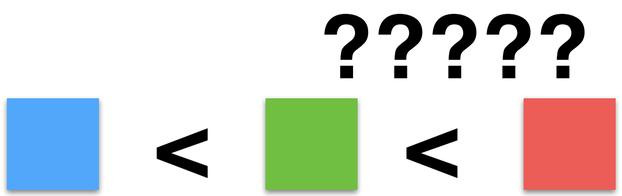
Color hue

Motion

Shape



Color



Good for qualitative data (identity channel)

Limited number of classes/length (~7-10!)

Does not work for quantitative data!

Lots of pitfalls! Be careful!

My rule:

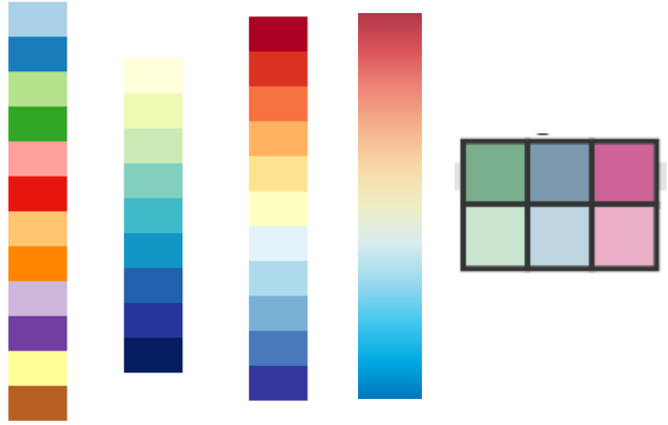
- minimize color use for encoding data
- use for brushing

What is a colormap?

[0,8] →

specifies a mapping between color and values

- categorical vs ordered
- sequential vs diverging
- segmented vs continuous
- univariate vs bivariate



expressiveness: match colormap to attribute characteristics!

http://colorbrewer2.org/

Quantitative Data Vis

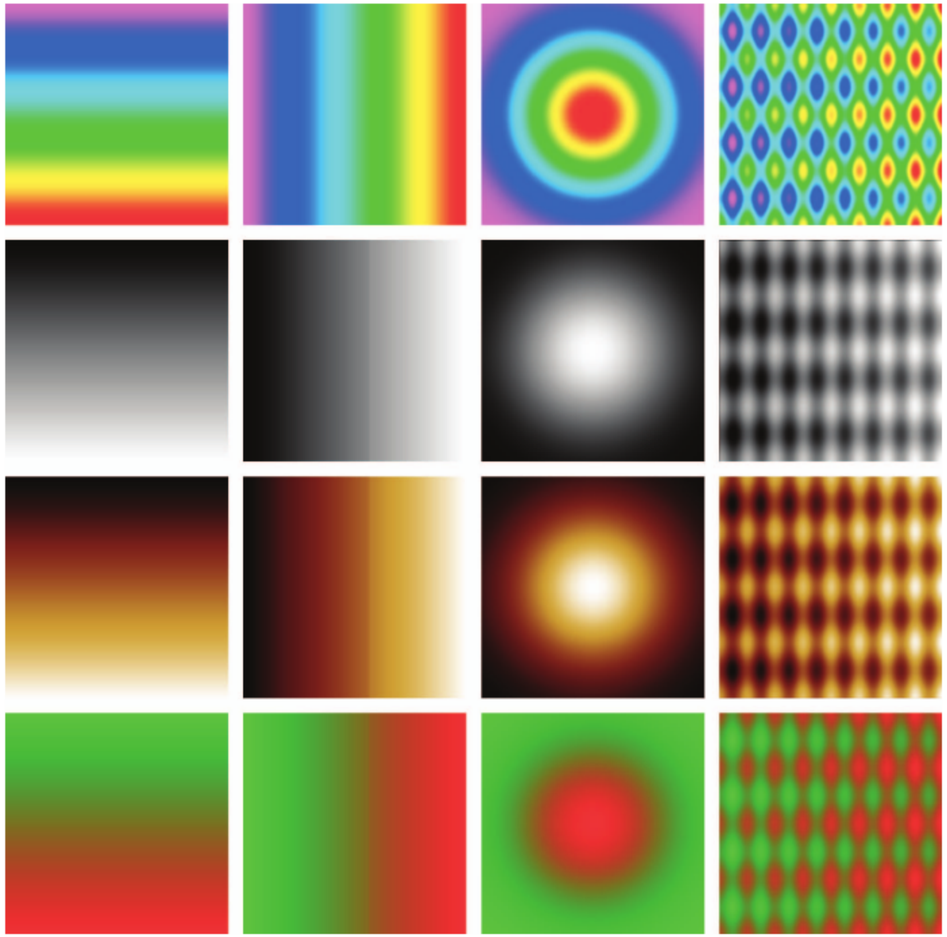
use value

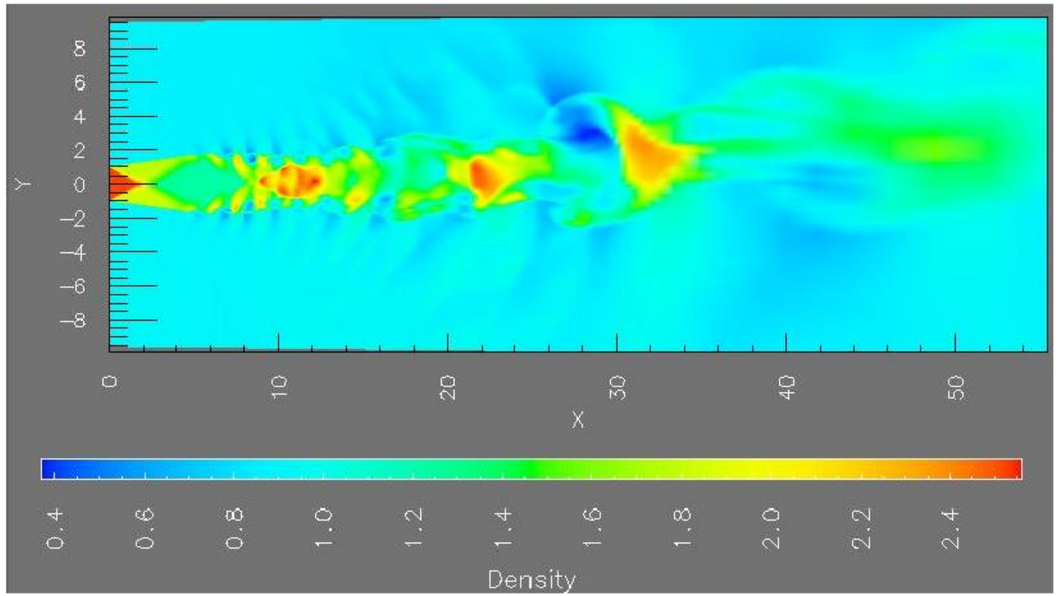
saturation works but not as good

don‘t use hue!

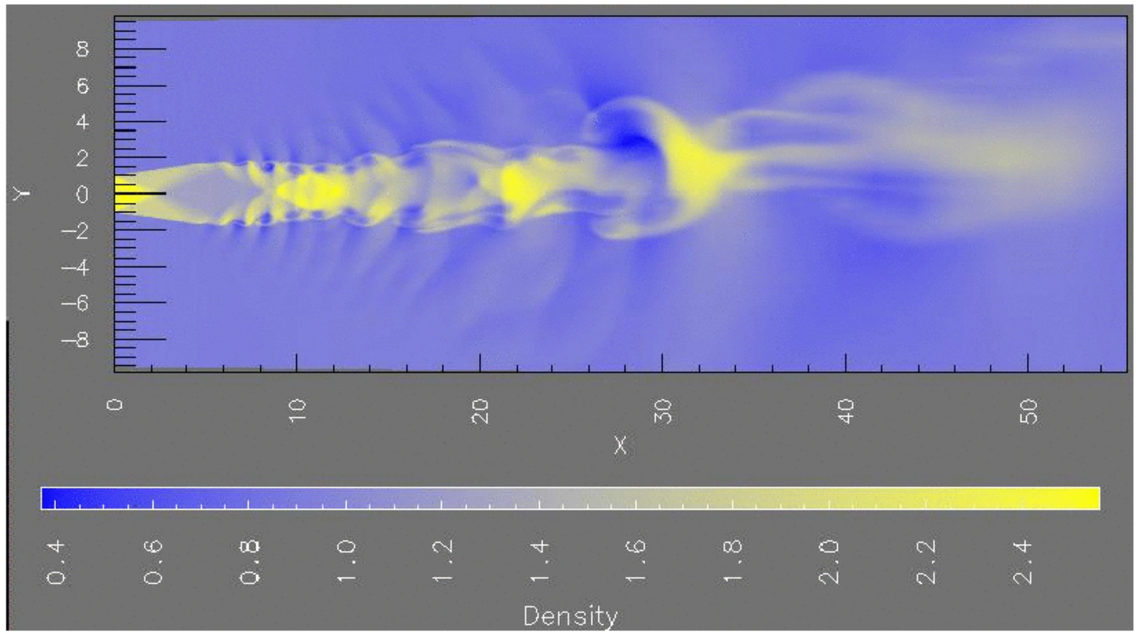
Danger: rainbow color map

Rainbow color map: confuses viewers through its lack of perceptual ordering, obscures data through its uncontrolled luminance variation, and actively misleads interpretation through the introduction of non-data-dependent gradients. (Rainbow Color Map (Still) Considered Harmful, Borland and Taylor II **10.1109/MCG.2007.46**)





[Rogowitz and Treinish, Why Should Engineers and Scientists Be Worried About Color? http://www.research.ibm.com/people/l/lloydt/color/color.HTM]



[Rogowitz and Treinish, How NOT to Lie with Visualization, www.research.ibm.com/dx/proceedings/pravda/truevis.htm]

Color Blindness

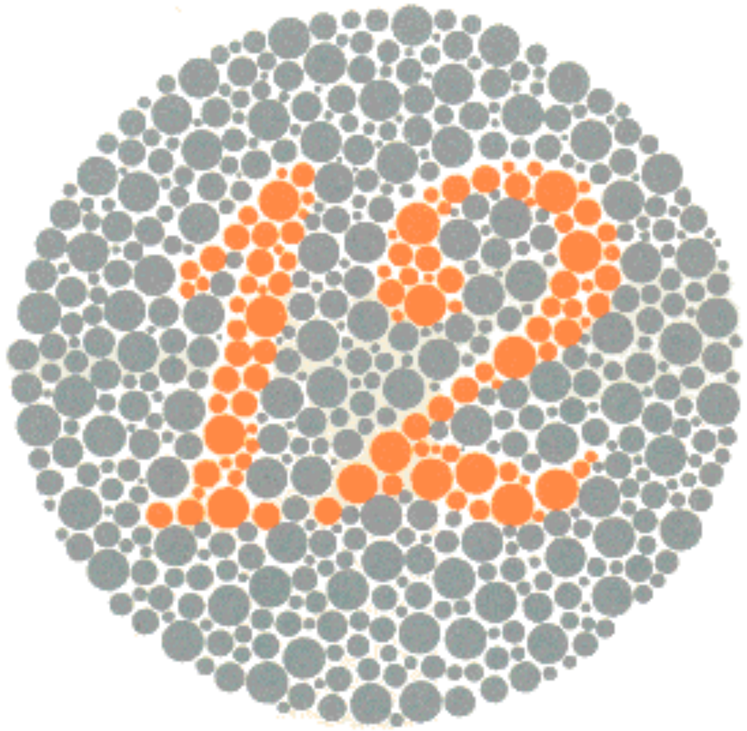
10% of males, 1% of females (probably due to X-chromosomal recessive inheritance)

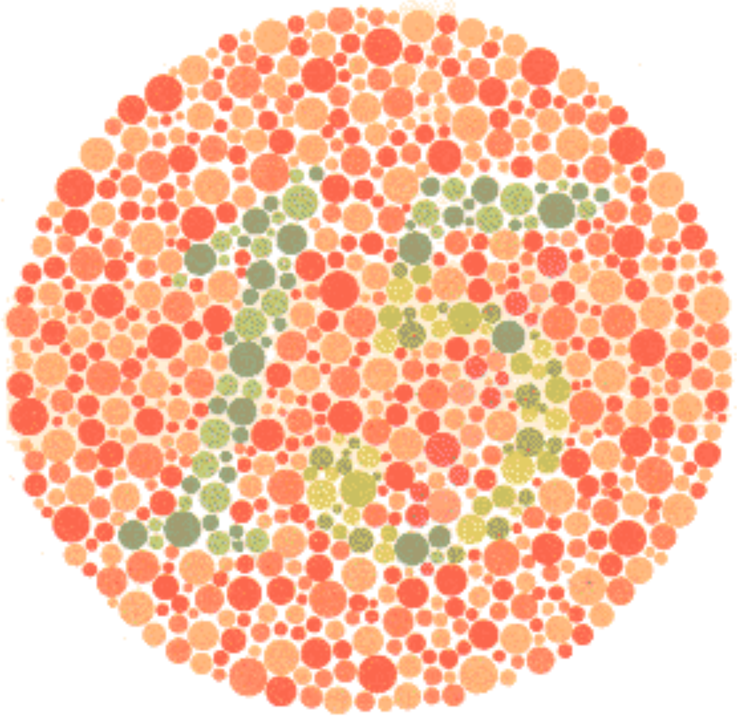
Most common: red-green weakness / blindness

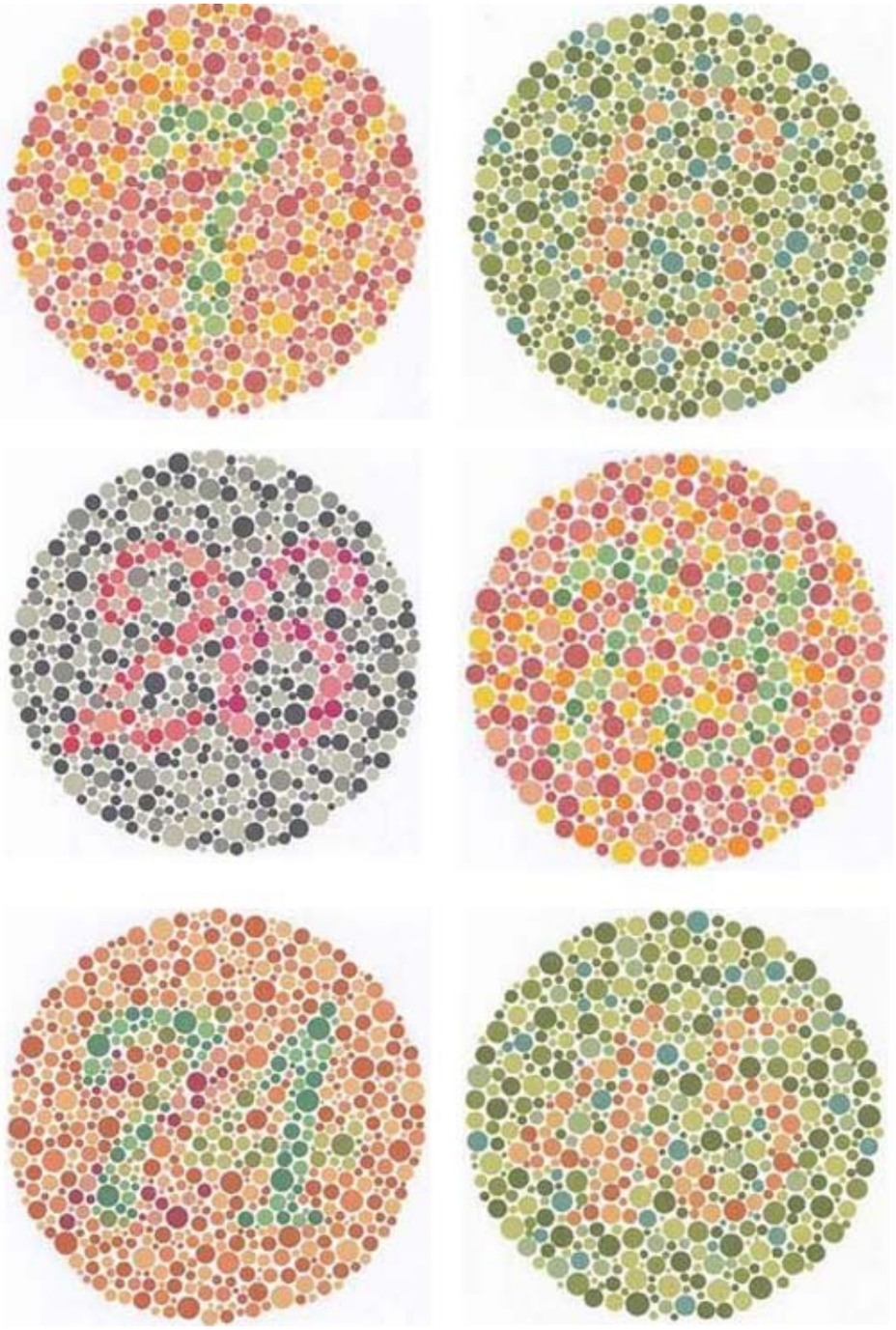
Reason: lack of medium or long wavelength receptors, or altered spectral sensitivity (most common: green shift)

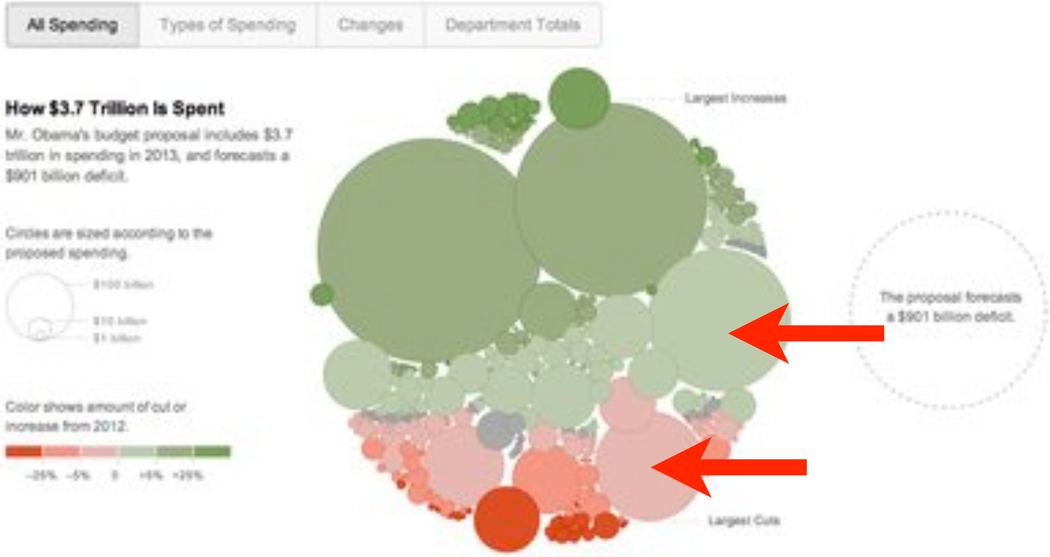
Normal Color Perception | Deuteranopia (no green receptors) | Protanopia (no red receptors)

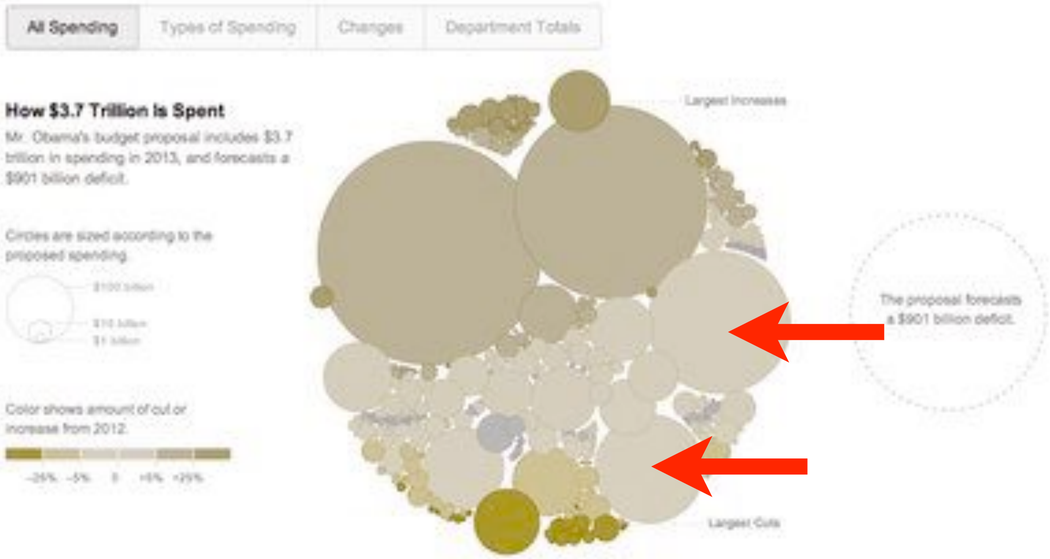
Color Blindness Tests





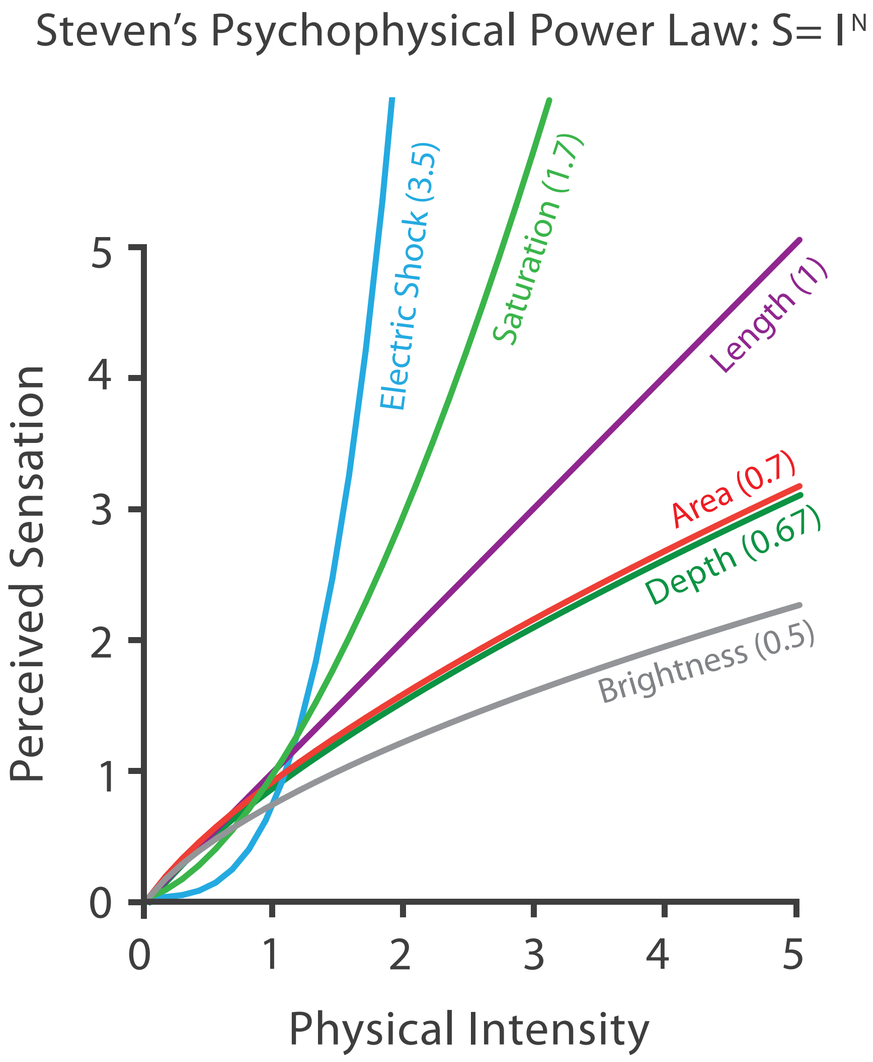






NYT, Feb 12, 2012

Why are quantitative channels different?



S = sensation
I = intensity

Psychological intensity ("Sensation") increases as the nth power of stimulus intensity

The general form of the law is

$$\psi(I) = kI^a,$$

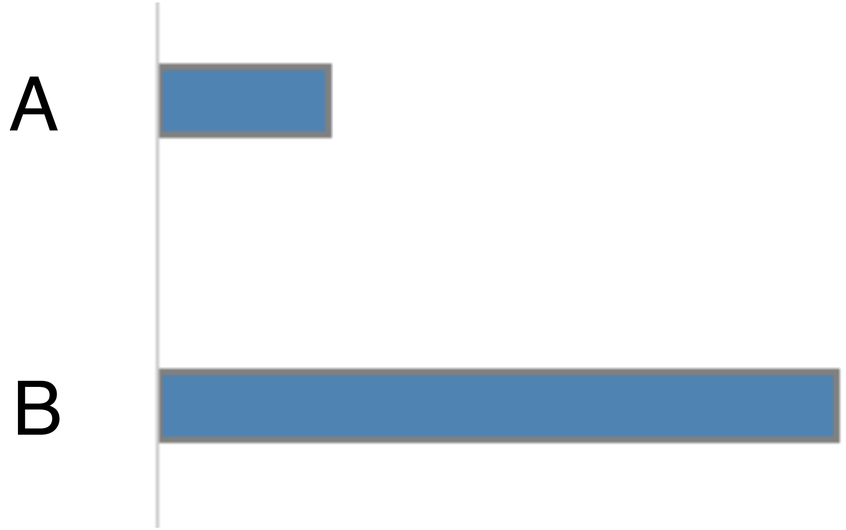
where *I* is the intensity or strength of the stimulus in physical units (energy, weight, pressure, mixture proportions, etc.), ψ(*I*) is the magnitude of the sensation evoked by the stimulus, *a* is an exponent that depends on the type of stimulation or sensory modality, and *k* is a proportionality constant that depends on the units used.

https://en.wikipedia.org/wiki/Stevens%27s_power_law

How much longer?

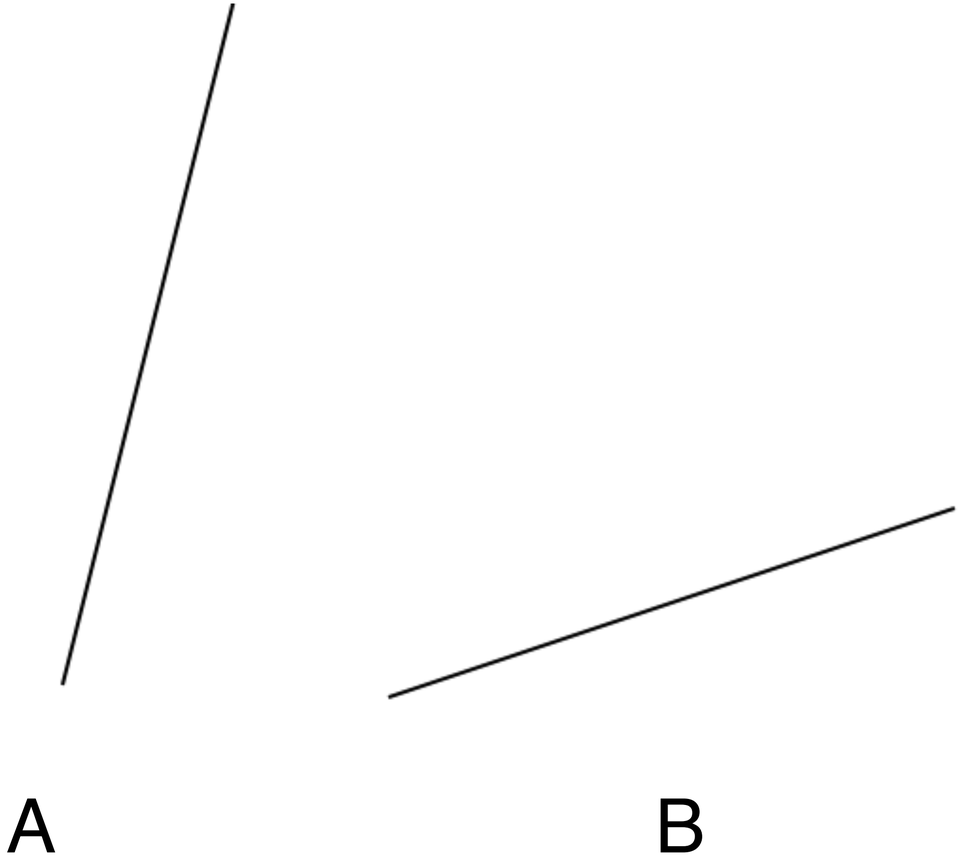
2x

How much longer?



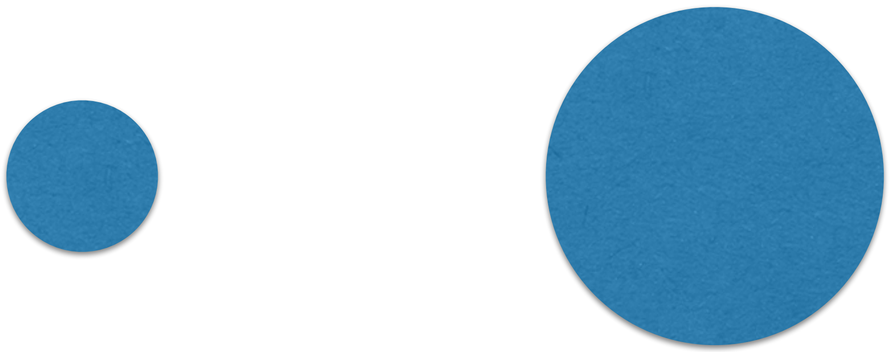
4x

How much steeper?



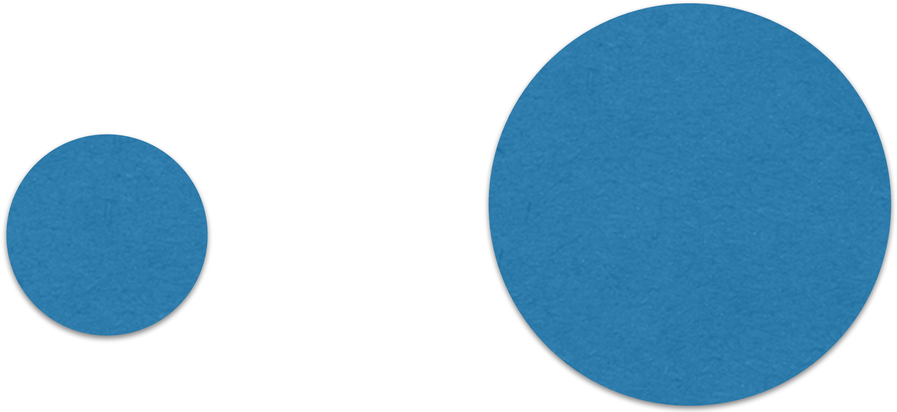
~4x

How much larger?



5x

How much larger?



A B

2x diameter

4x area

area is proportional to diameter squared

How much larger (area)?

A B

3x

How much darker?

A B

2x

How much darker?

A B

3x

Other Factors Affecting Accuracy

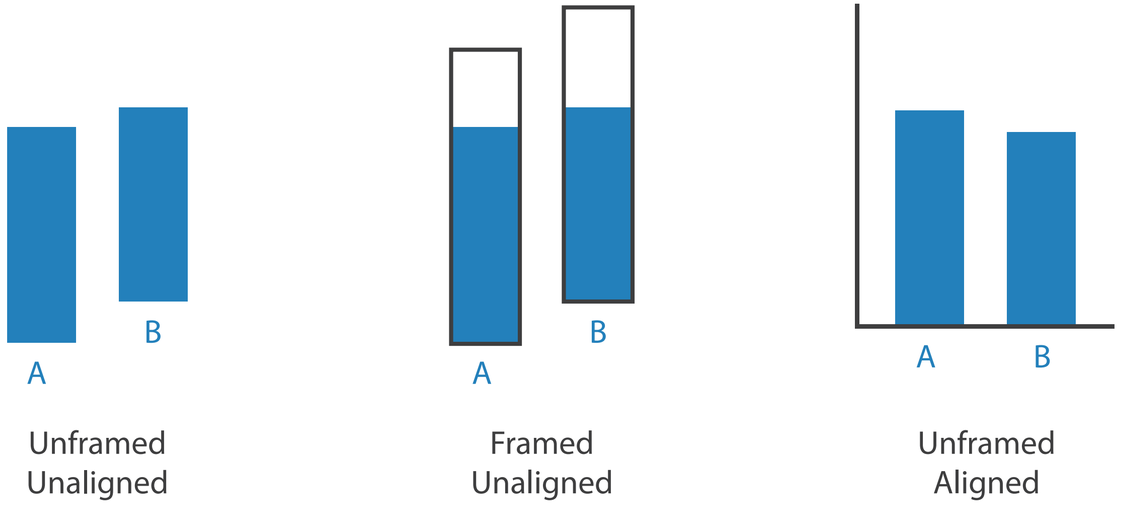
Alignment

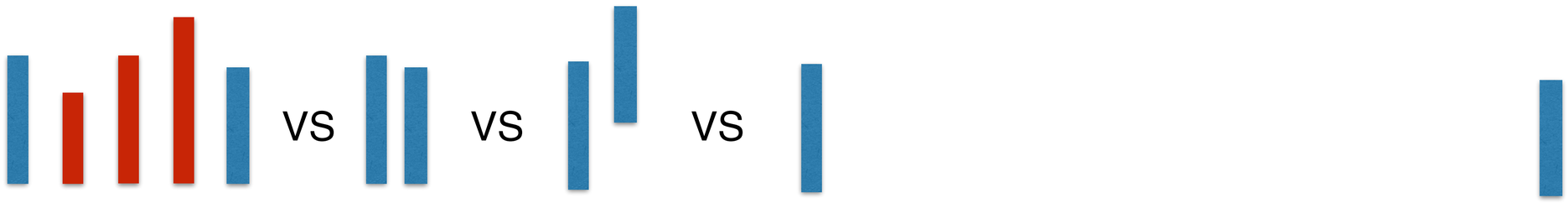
Distractors

Distance

Common scale

…





Design Guidelines

Rule #1: Use the Best Visual Channel Available for the Most Important Aspect of your Data

Rule #2: The visualization should show all of the data, and only the data

Tufte’s Integrity Principles

Show **data variation**, not design variation

Clear, detailed, and thorough **labeling** and **appropriate scales**

Size of the **graphic effect** should be **directly proportional to the numerical quantities** (“lie factor”)

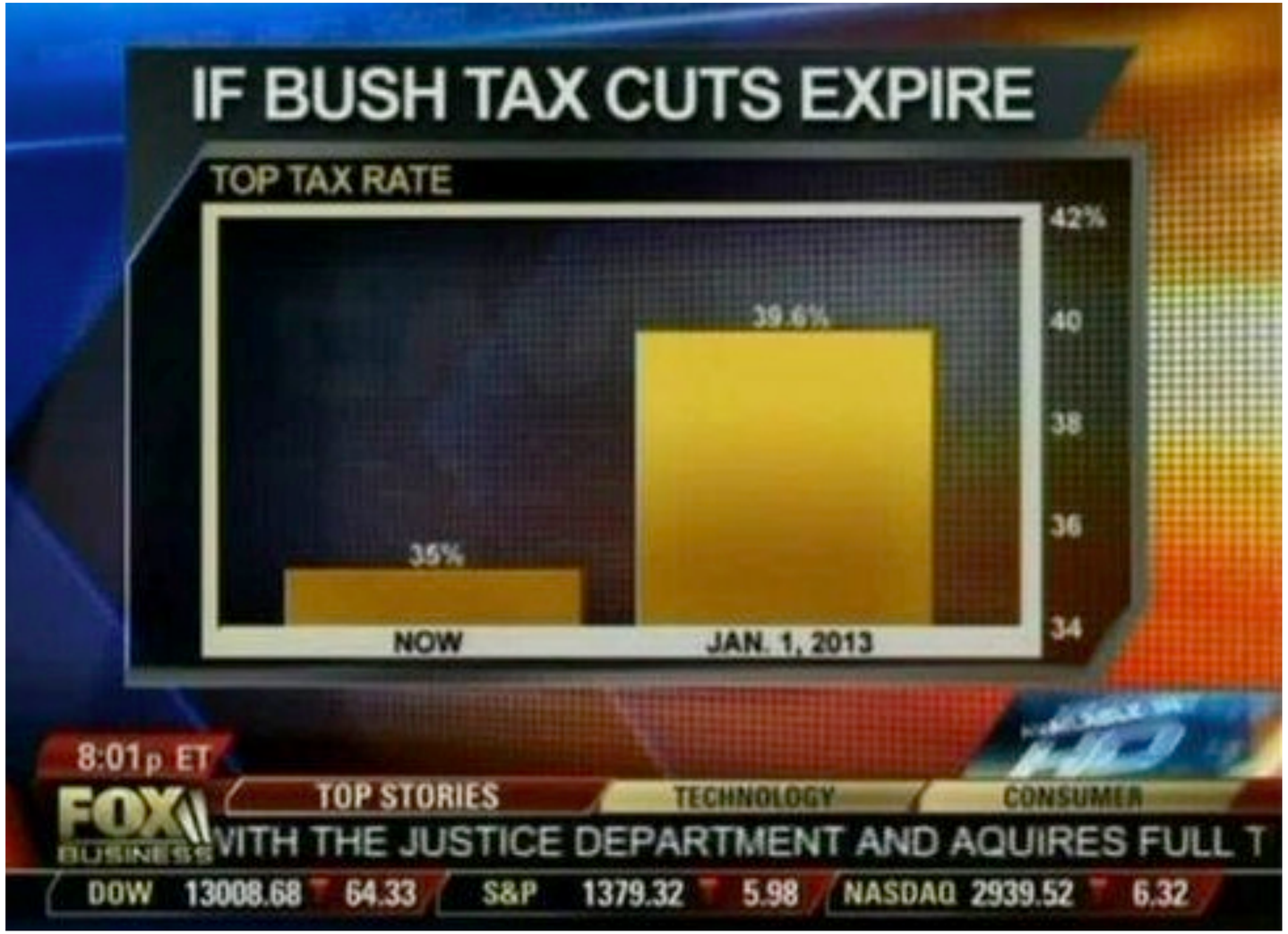
The Lie Factor

$$\frac{\text{Size of effect shown in graphic}}{\text{Size of effect in data}}$$

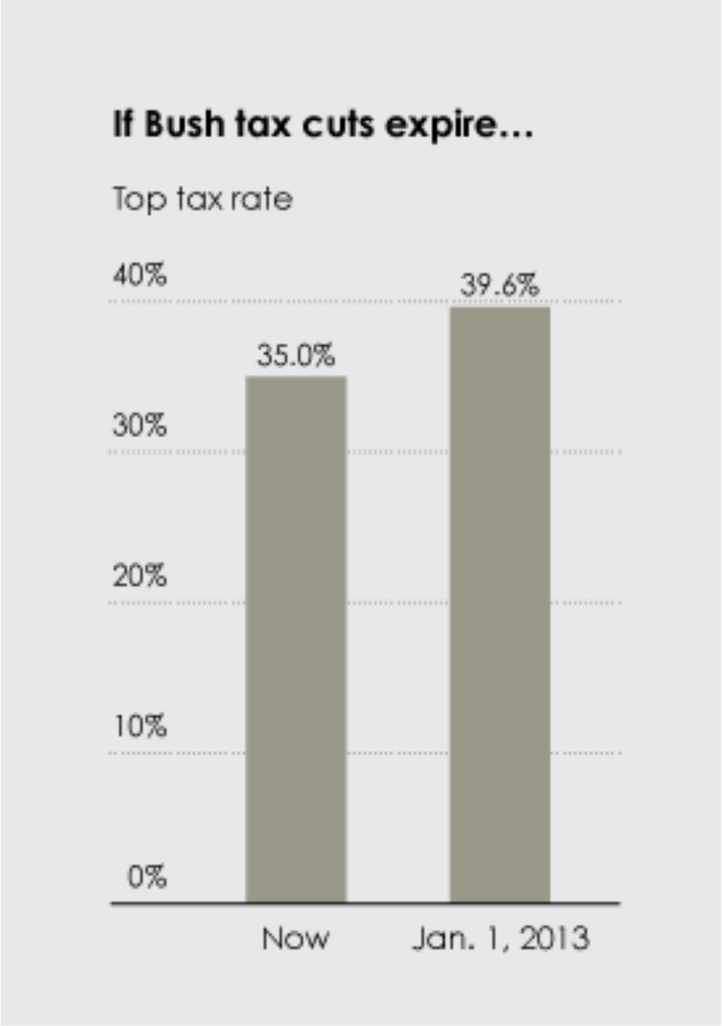
Lie Factor - Graphical Integrity

Magnitude in data must correspond to magnitude of mark

Effect in Data: factor 1.14
Effect in Graphic: factor 5
Lie Factor: 5/1.14 = 4.38



Scale Distortions



Flowing Data

What’s wrong?

Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

What’s wrong?

Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

1
2
3

Asyl
Opf
frei
"Kr
Räd
geh
Zwe
Mit
Wol
Am
Stre
Mes
Abe

Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

Die größte Gruppe in der Liste der Mindestsicherungsbezieher ist aber jene der "ungeklärten Staatsbürgerschaft". Dass es sich bei den 16.712 Personen um

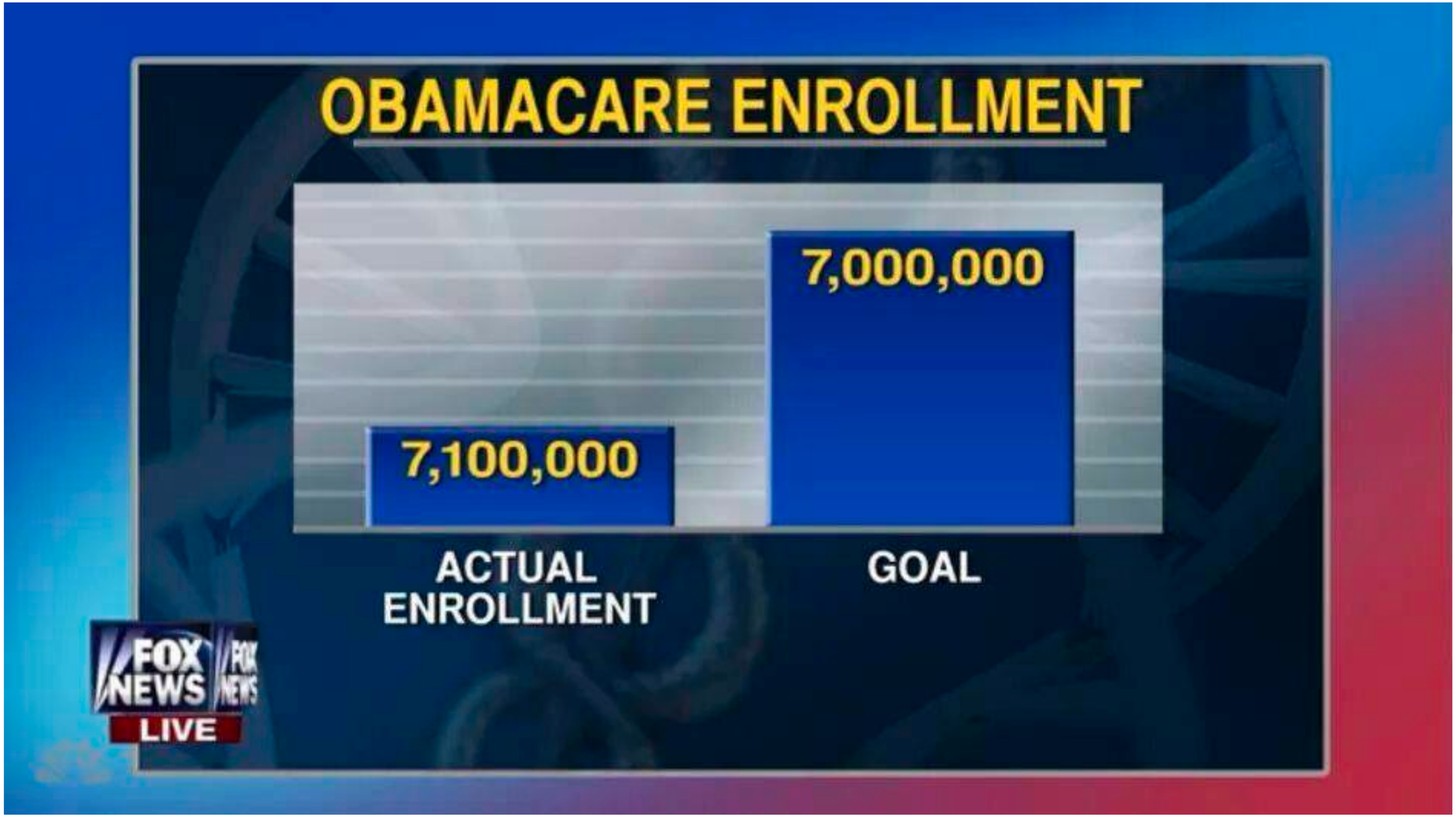
What’s wrong?

graphic of crown newspaper

Grafik der Kronen-zeitung

Grafik in echt

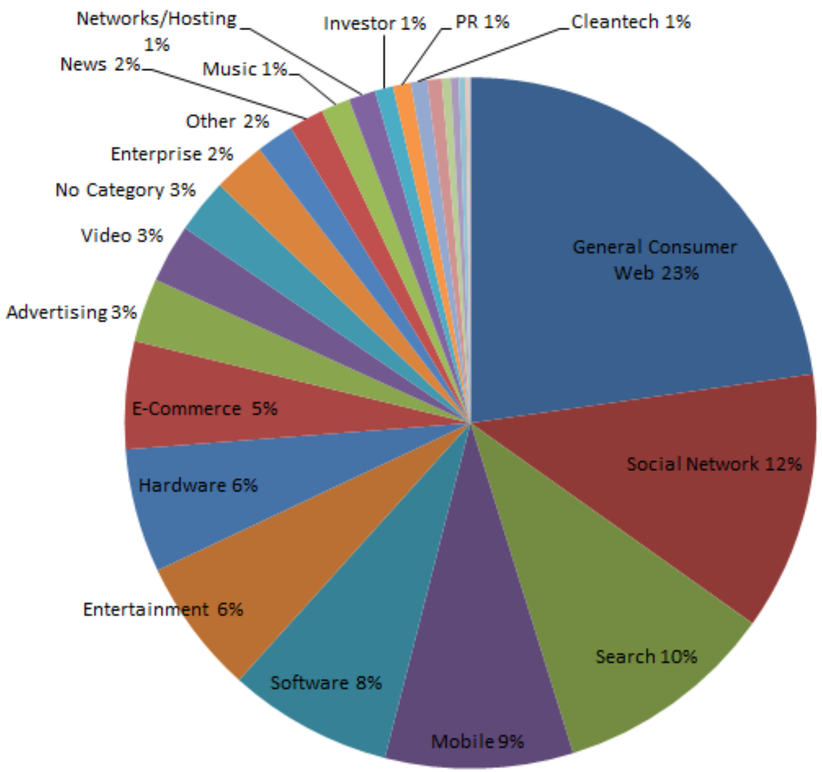
graphic in real


OBAMACARE ENROLLMENT
7,000,000
7,100,000
ACTUAL ENROLLMENT
GOAL
FOX NEWS
LIVE

Scales at 0

Use a baseline that shows the data, not the zero-point.

Death to Pie Charts



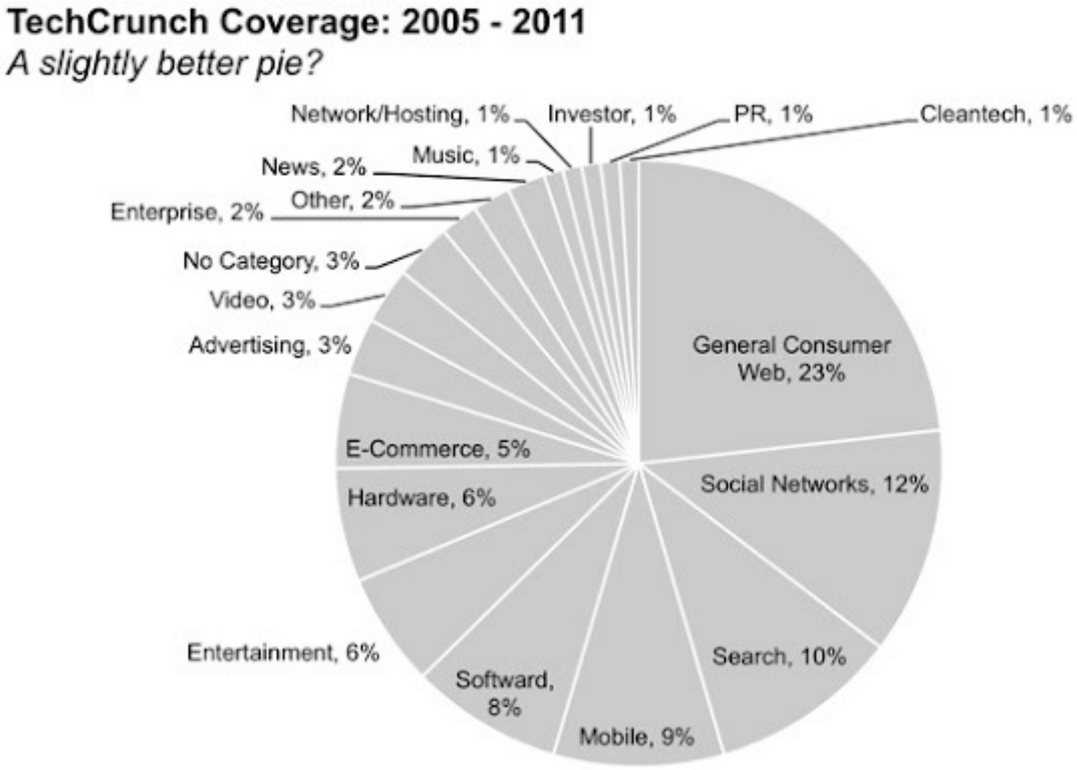
Share of coverage on TechCrunch

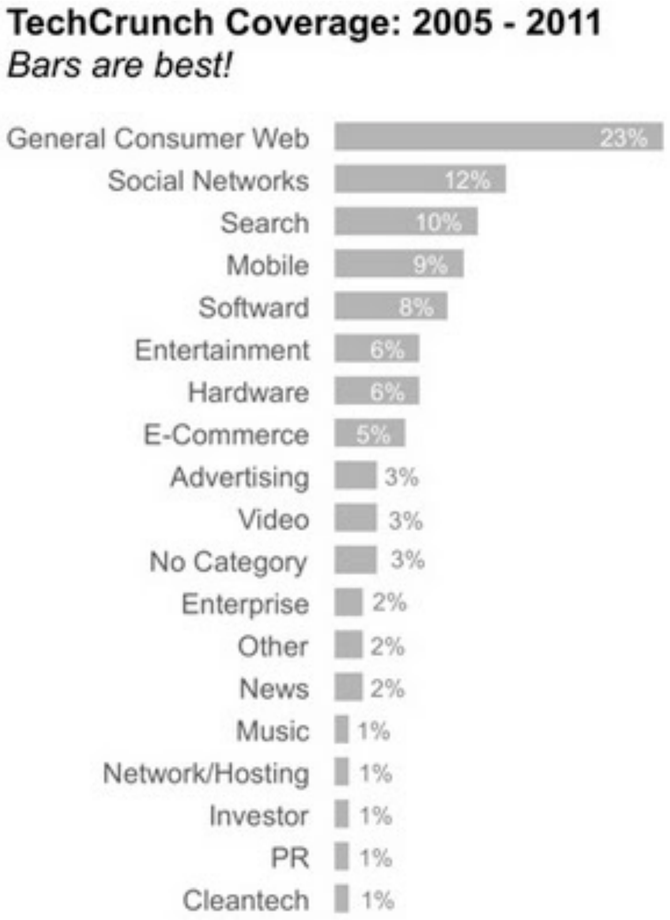
"I hate pie charts.
I mean, really hate them."

www.storytellingwithdata.com/2011/07/death-to-pie-charts.html

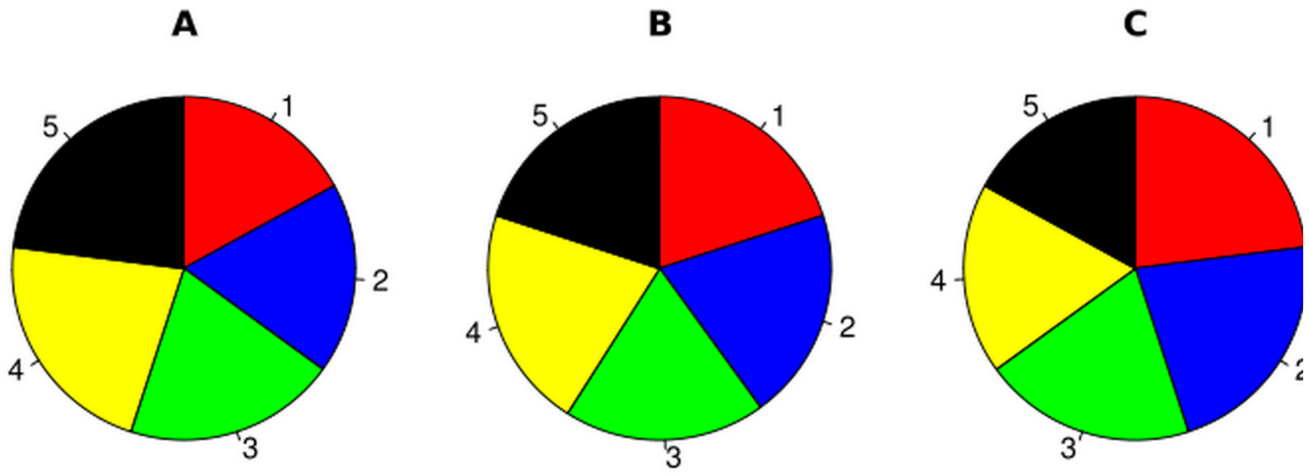
Cole Nussbaumer

Redesign

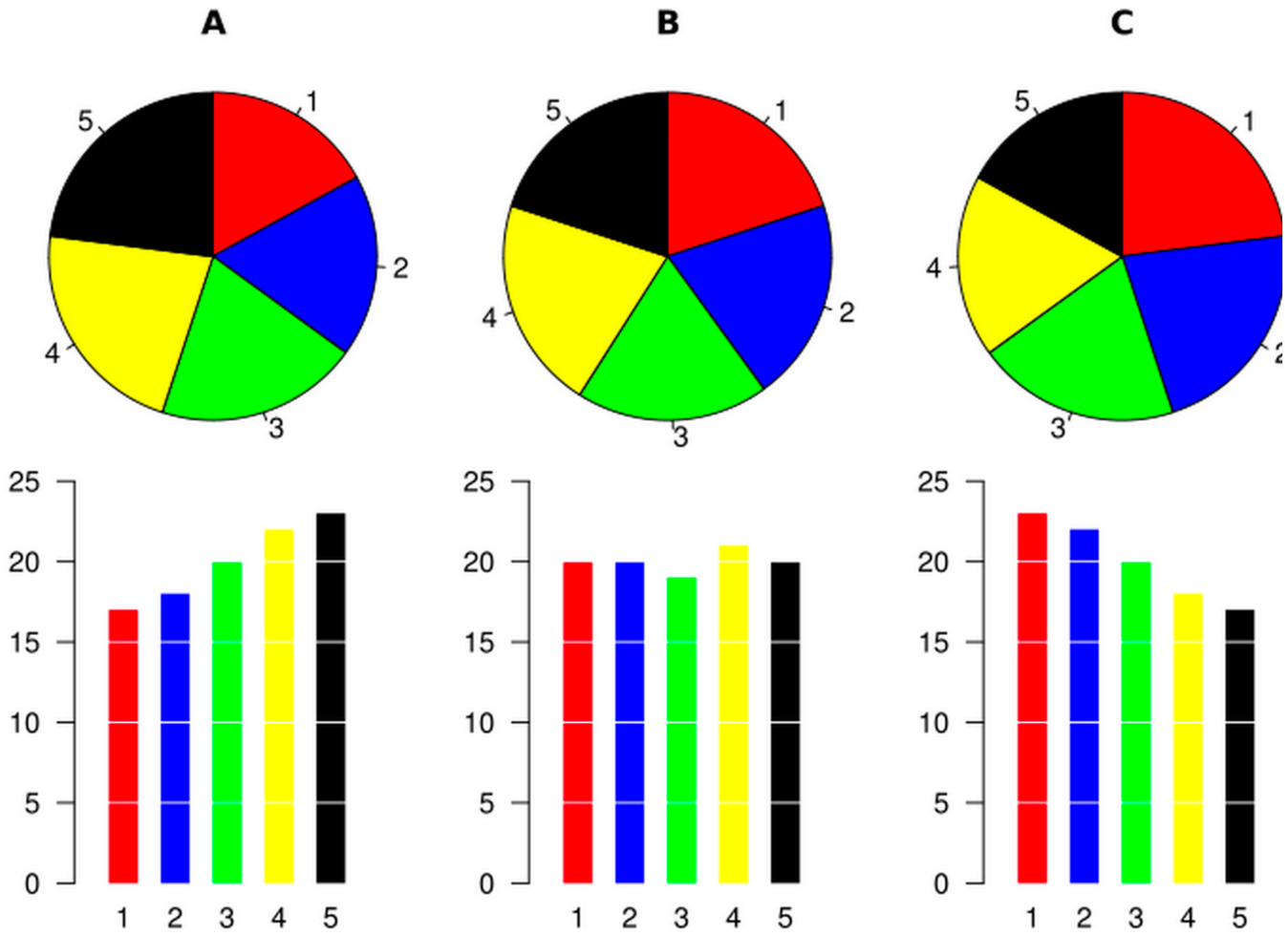

TechCrunch Coverage: 2005 - 2011
A slightly better pie?
Network/Hosting, 1%
Investor, 1%
PR, 1%
Cleantech, 1%
Music, 1%
News, 2%
Other, 2%
Enterprise, 2%
No Category, 3%
Video, 3%
Advertising, 3%
General Consumer Web, 23%
E-Commerce, 5%
Social Networks, 12%
Hardware, 6%
Entertainment, 6%
Search, 10%
Softward, 8%
Mobile, 9%


TechCrunch Coverage: 2005 - 2011
Bars are best!
General Consumer Web 23%
Social Networks 12%
Search 10%
Mobile 9%
Softward 8%
Entertainment 6%
Hardware 6%
E-Commerce 5%
Advertising 3%
Video 3%
No Category 3%
Enterprise 2%
Other 2%
News 2%
Music 1%
Network/Hosting 1%
Investor 1%
PR 1%
Cleantech 1%

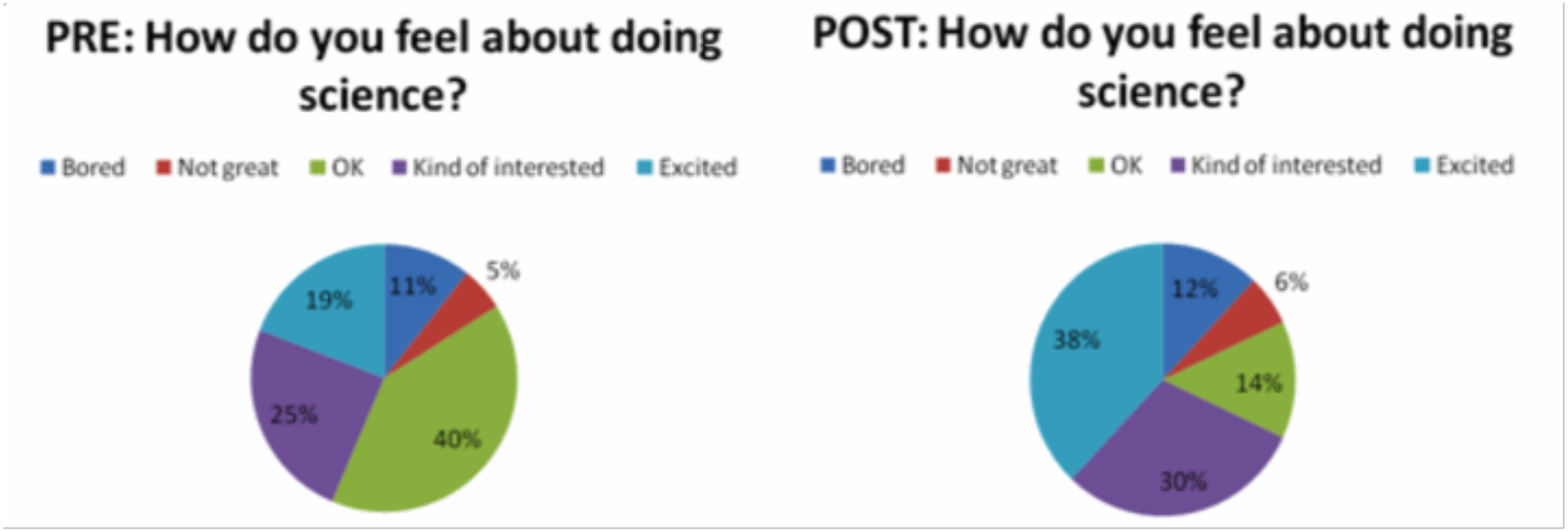
Can you spot the differences?



Can you spot the differences?



So, what to use instead?



imagine you just completed a pilot summer learning program on science aimed at improving perceptions of the field among 2nd and 3rd grade elementary children

http://www.storytellingwithdata.com/blog/2014/06/alternatives-to-pies

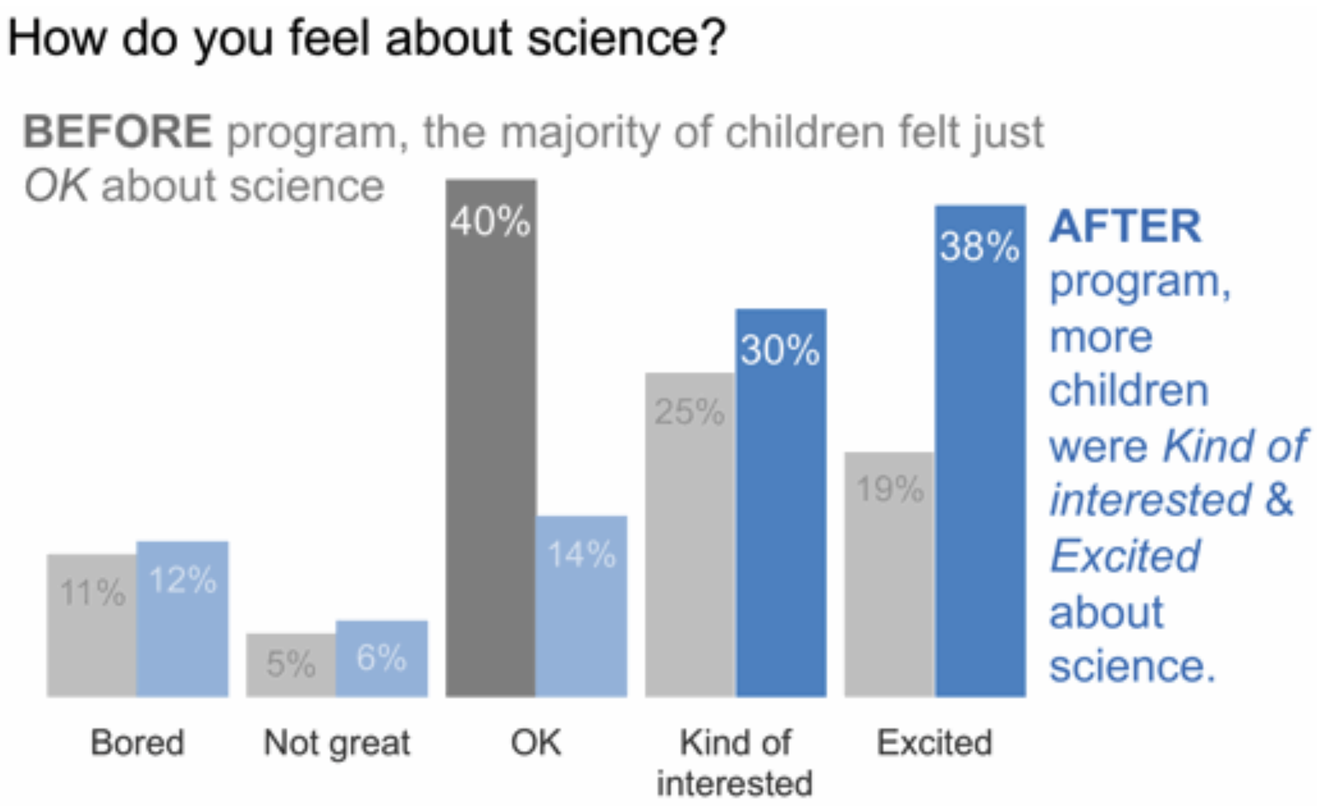
Alternative #1: Show the Number(s) Directly

After the pilot program,

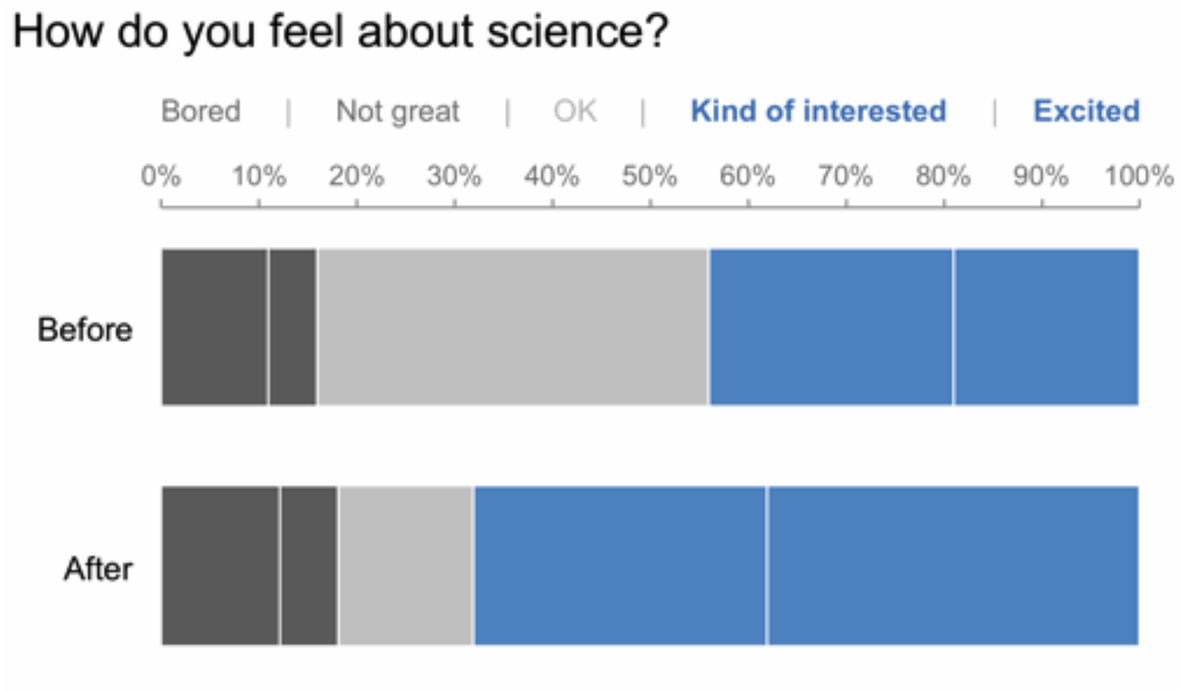
68%

of kids expressed interest towards science,
compared to 44% going into the program.

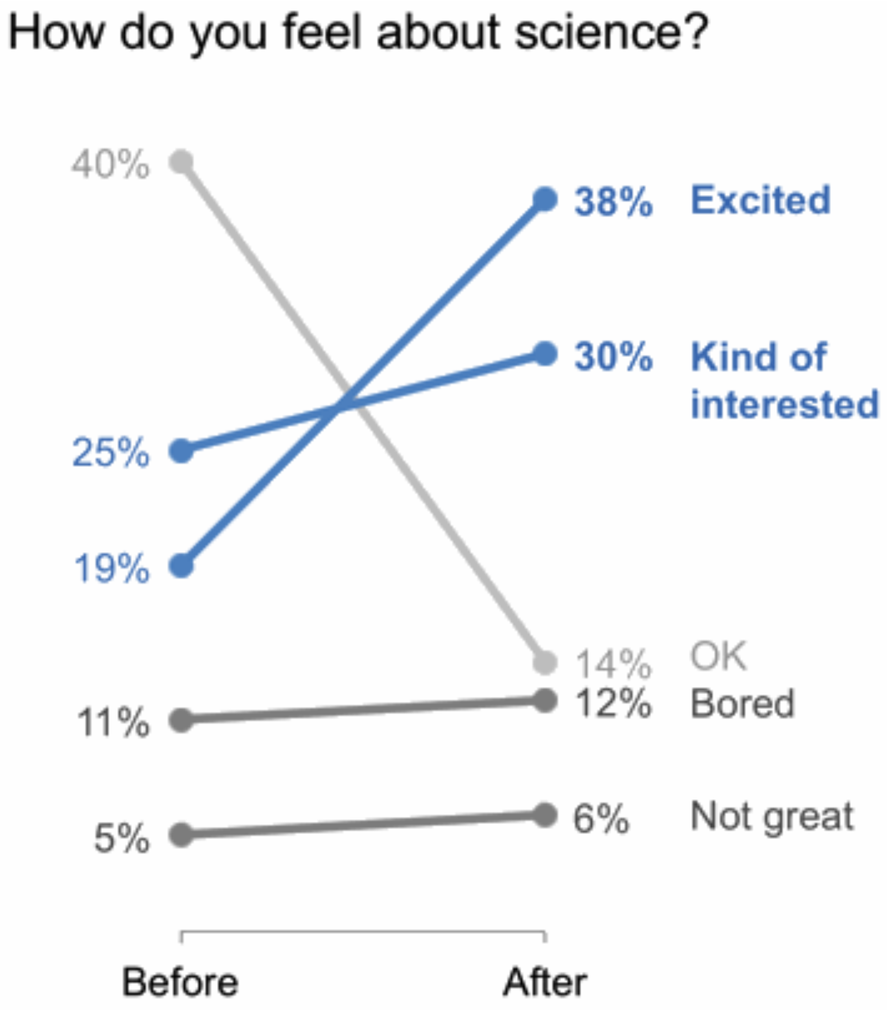
Alternative #2: Simple Bar Graph



Alternative #3: 100% Stacked Horizontal Bar Graph



Alternative #4: Slopegraph

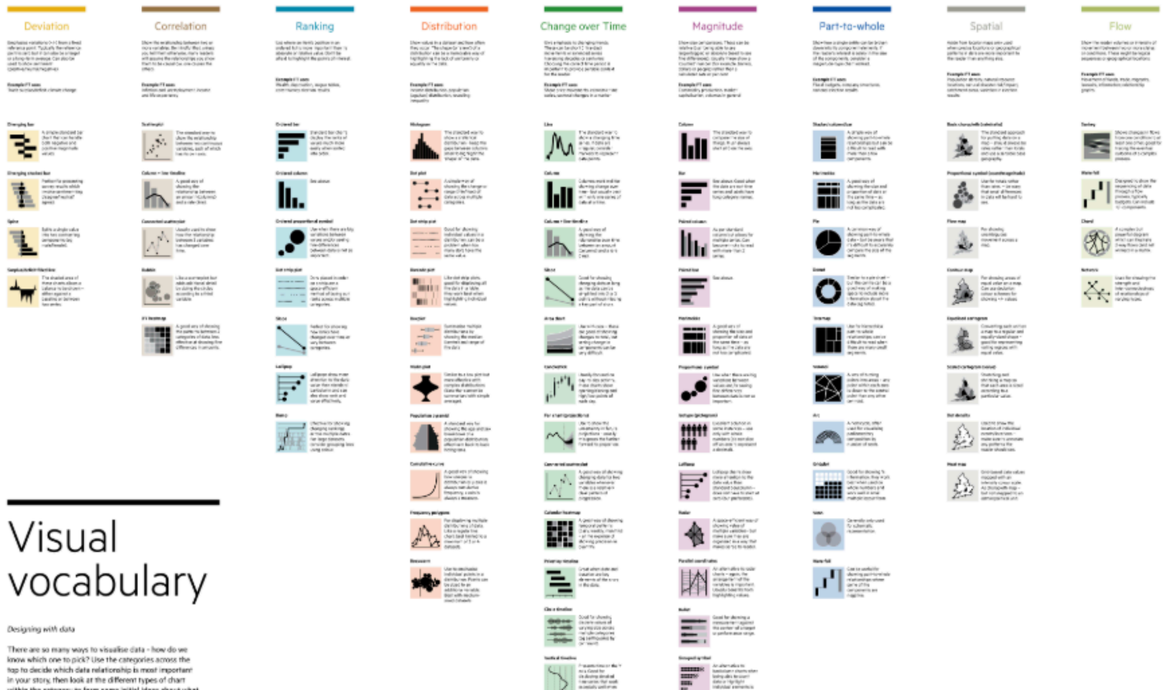


Tabular Data

Techniques and Tasks

Magnitude Distribution Deviation Correlation

Ranking Part to whole Change over Time



https://github.com/ft-interactive/chart-doctor/tree/master/visual-vocabulary
https://gramener.github.io/visual-vocabulary-vega/#/Magnitude/

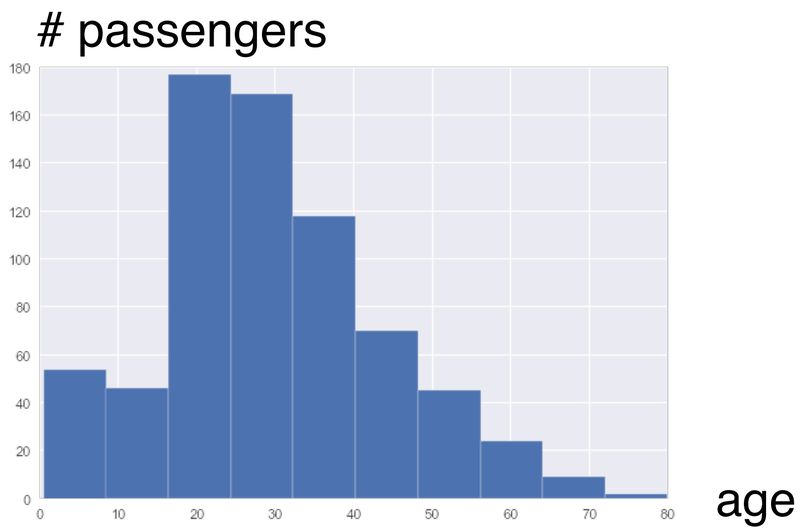
Histogram

Good #bins hard to predict

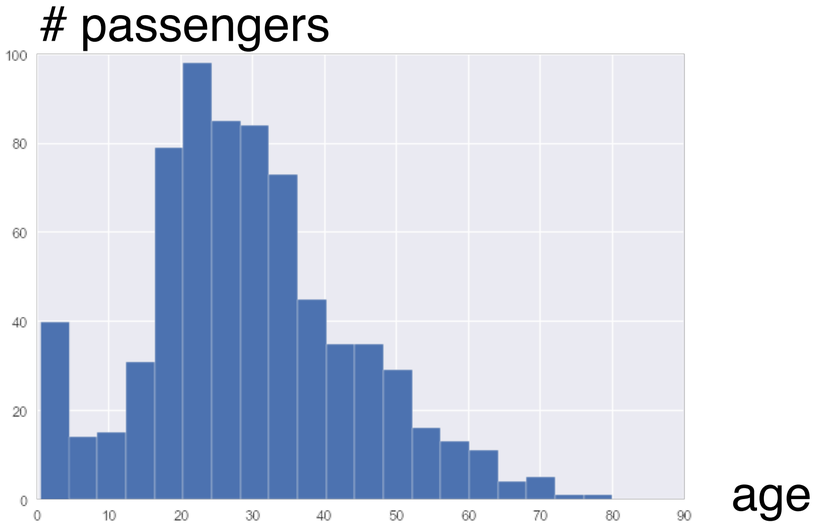
make interactive!

rules of thumb:

- #bins = sqrt(n)
- #bins = log2(n) + 1



10 Bins



20 Bins

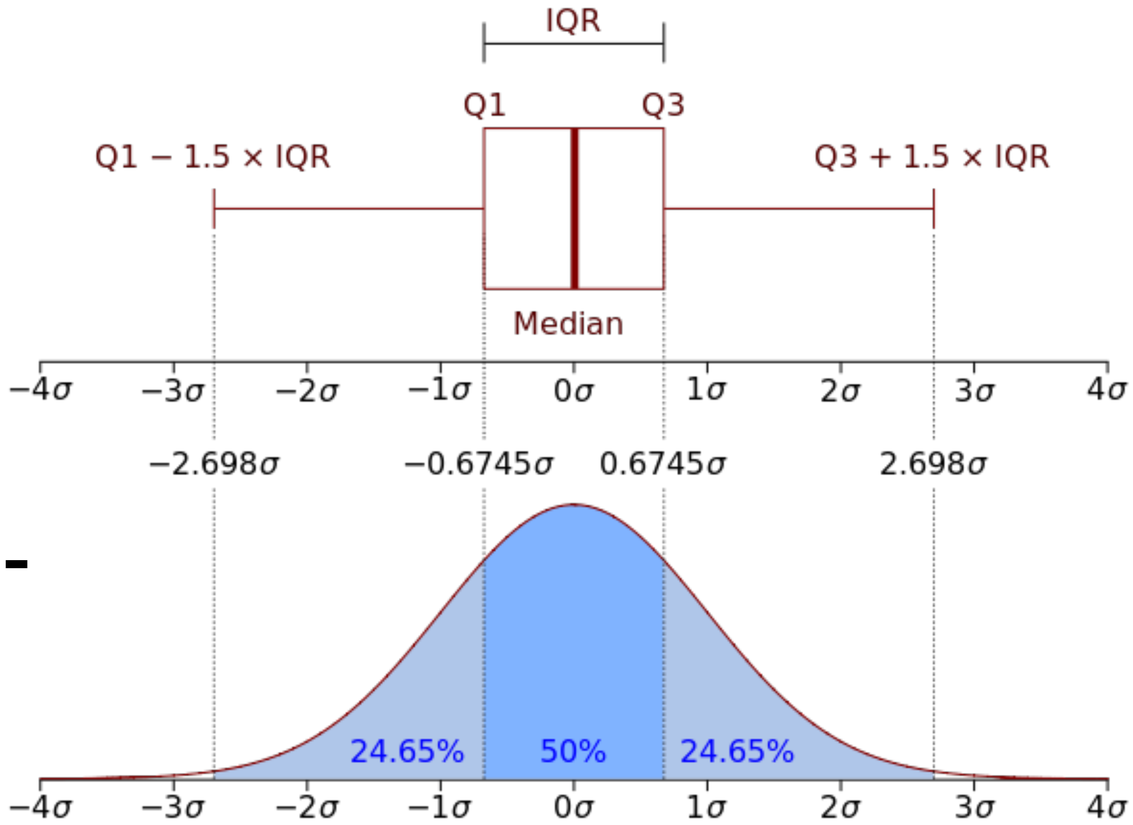
Box Plots

aka Box-and-Whisker Plot

Show outliers as points!

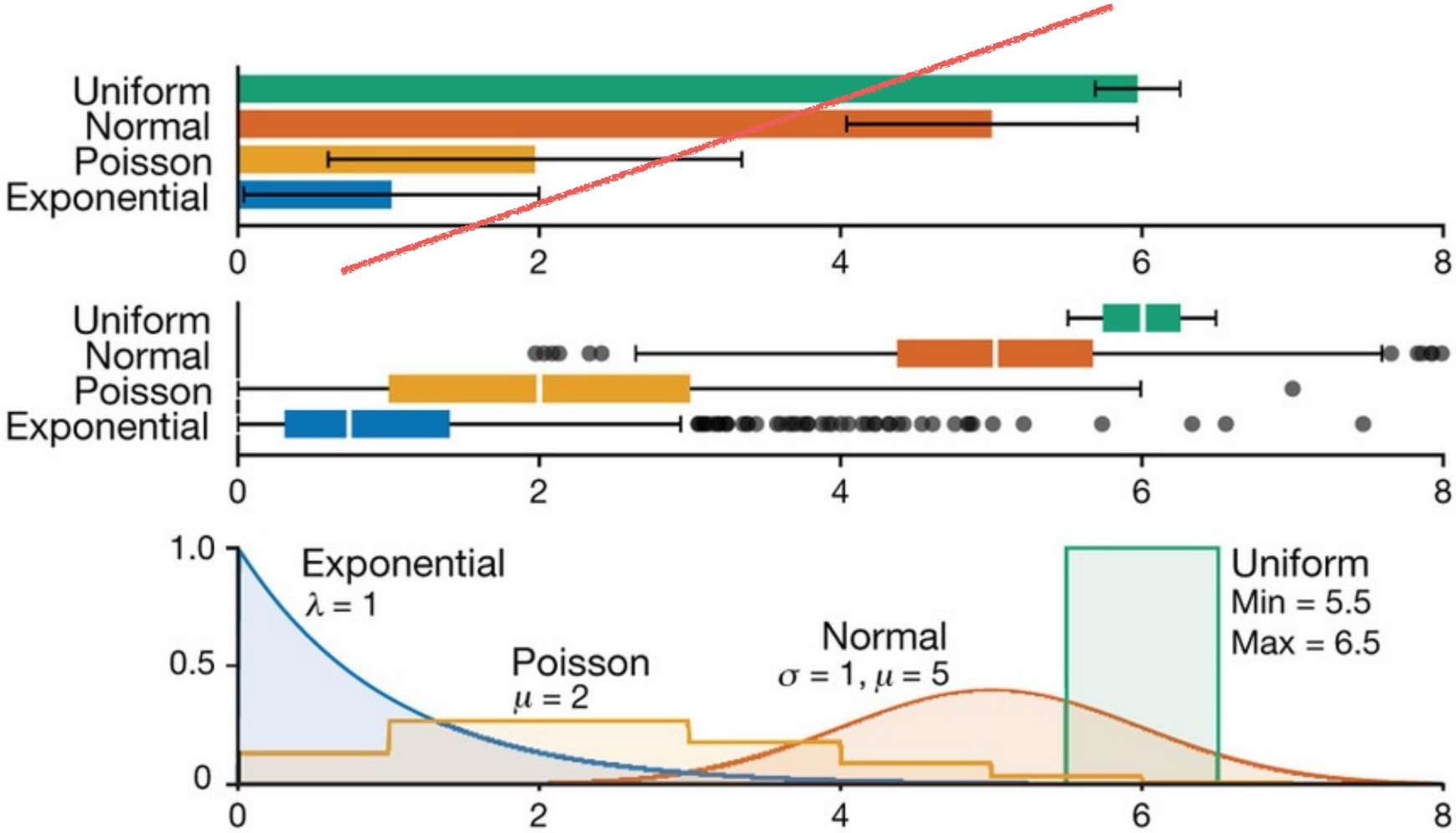
Not so great for non-normal distributed data

Especially bad for bi- or multi-modal distributions



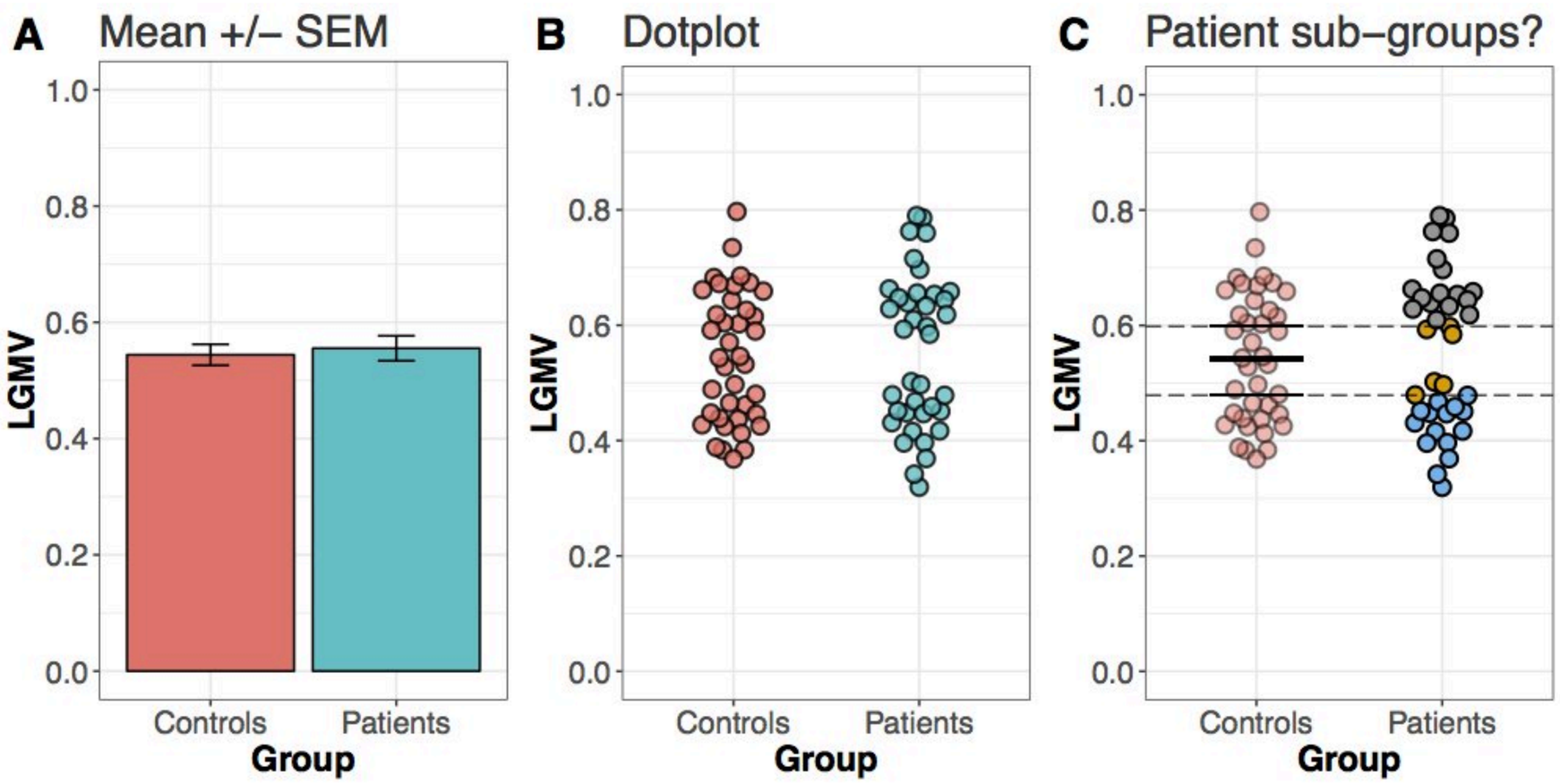
Wikipedia

Comparison



Streit & Gehlenborg, PoV, Nature Methods, 2014

Bar Charts vs Dot Plots

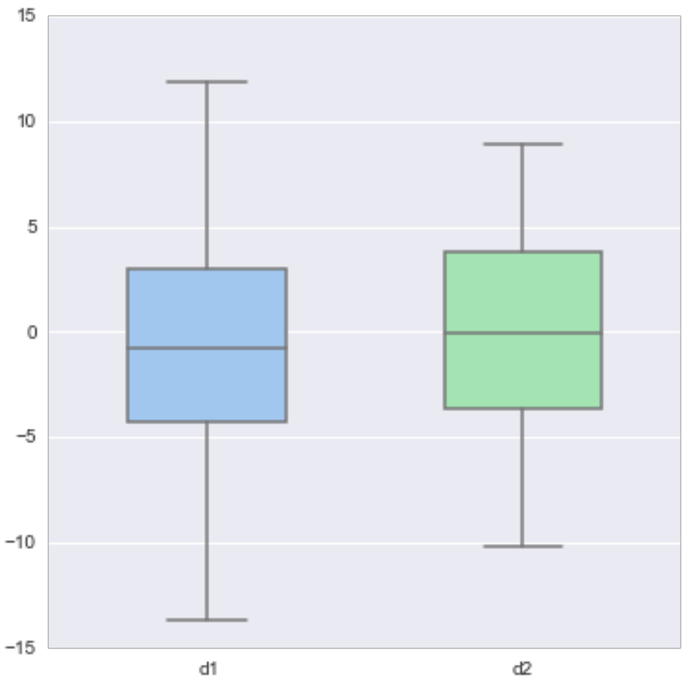


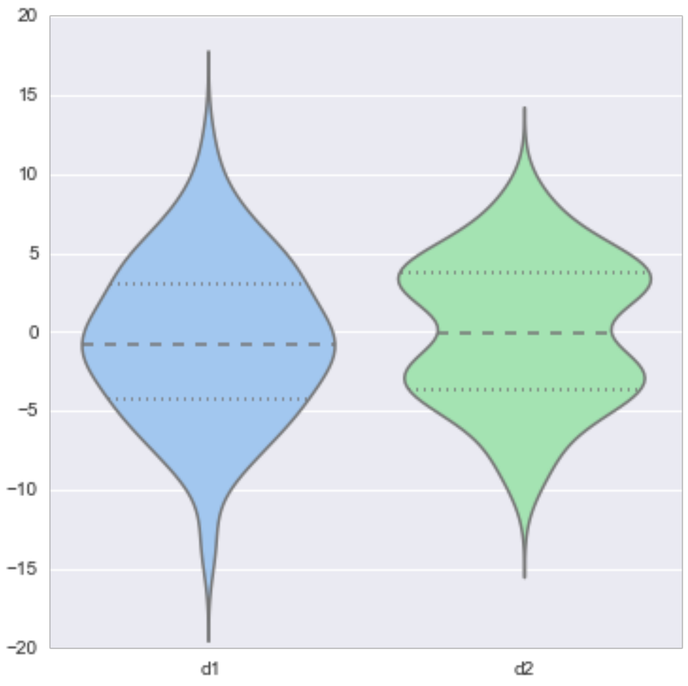
Data Source https://bmcneurosci.biomedcentral.com/articles/10.1186/1471-2202-10-67

https://twitter.com/robustgar/status/859318971920769024

Violin Plot

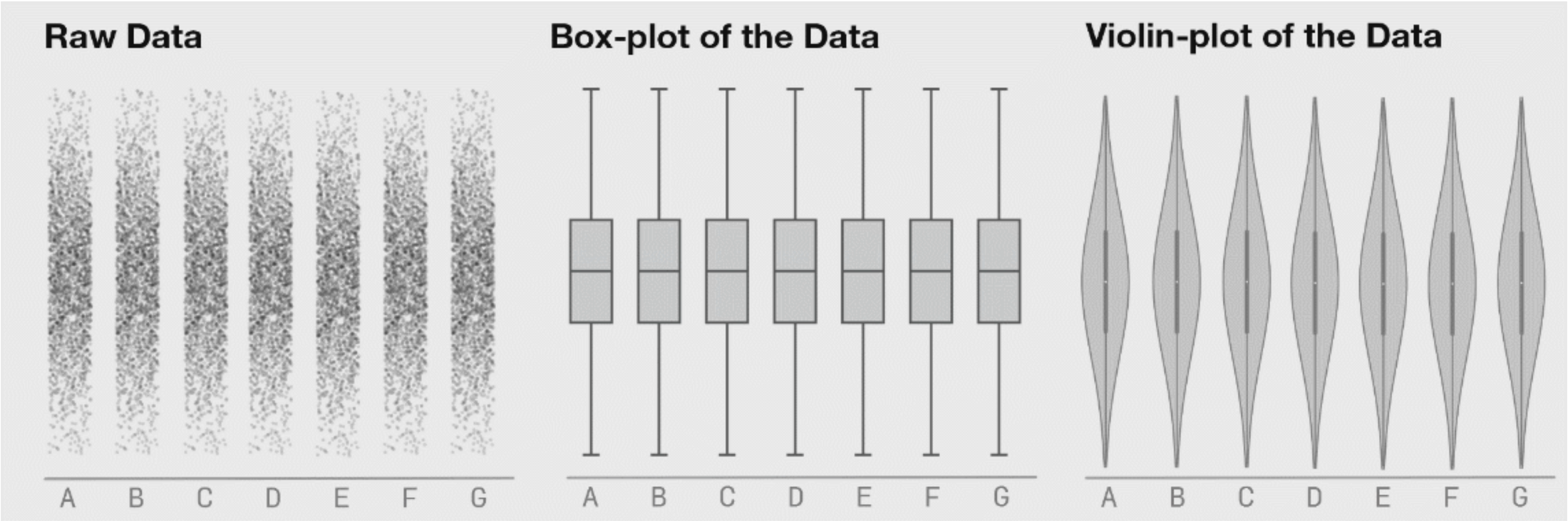
= Box Plot + Probability Density Function





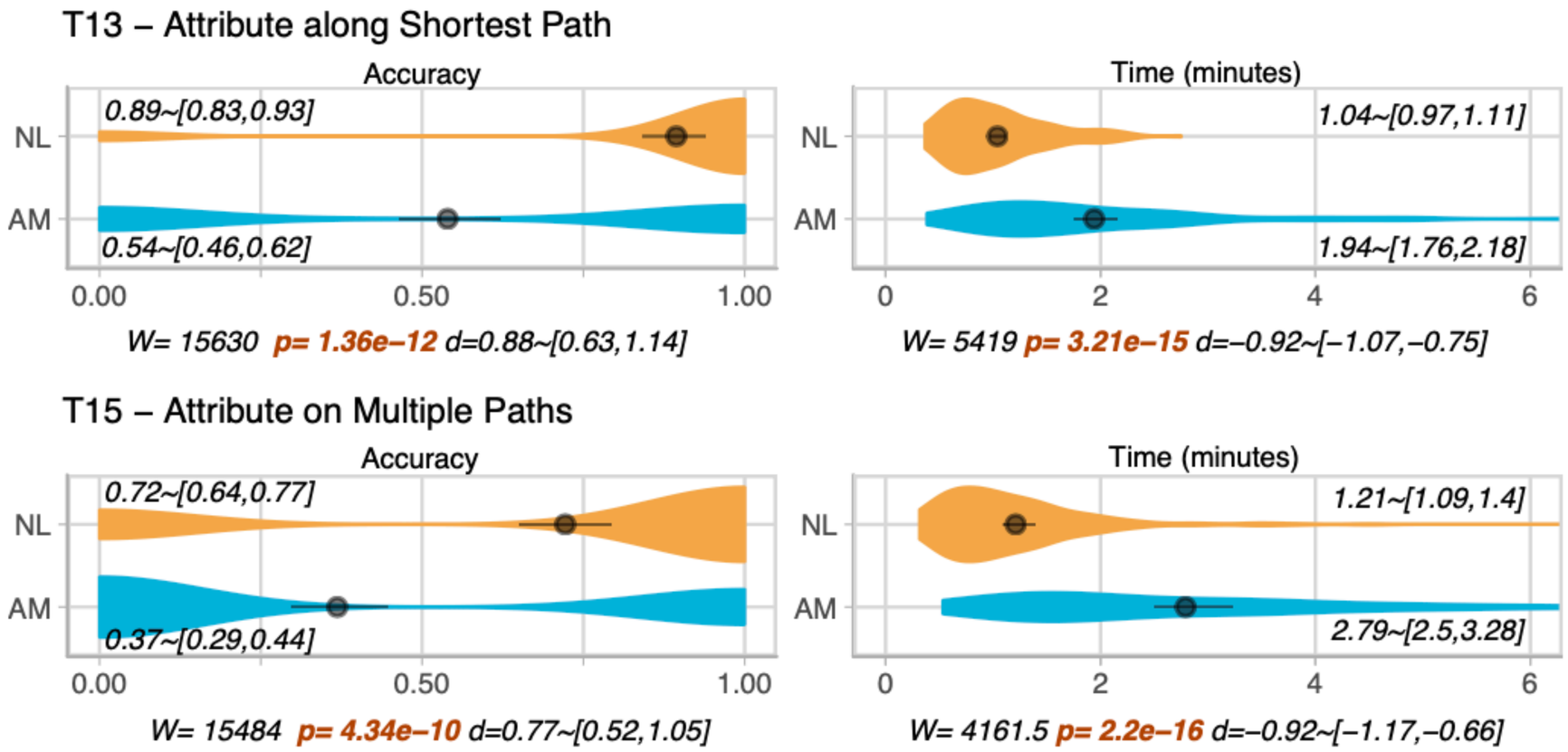
http://web.stanford.edu/~mwaskom/software/seaborn/tutorial/plotting_distributions.html

Different Distributions



https://blog.bioturing.com/2018/05/16/5-reasons-you-should-use-a-violin-graph/

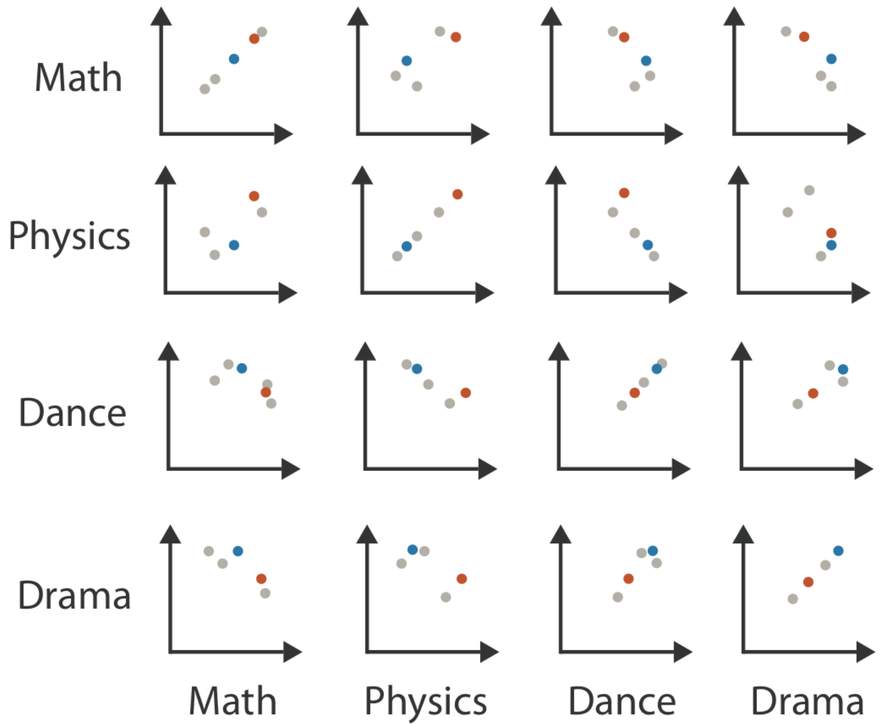
How to Visually Report Distributions for Experiments



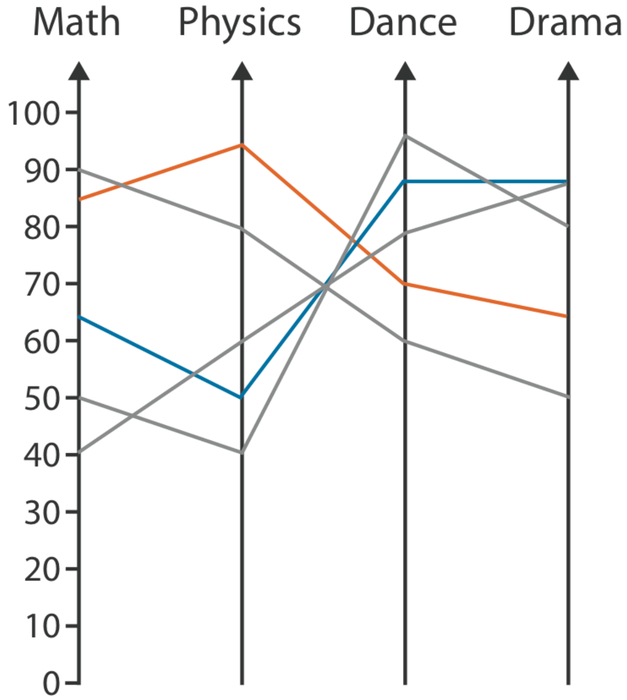
[Nobre 2019]

Scatterplot Matrix & PCs

Scatterplot Matrix



Parallel Coordinates



Table

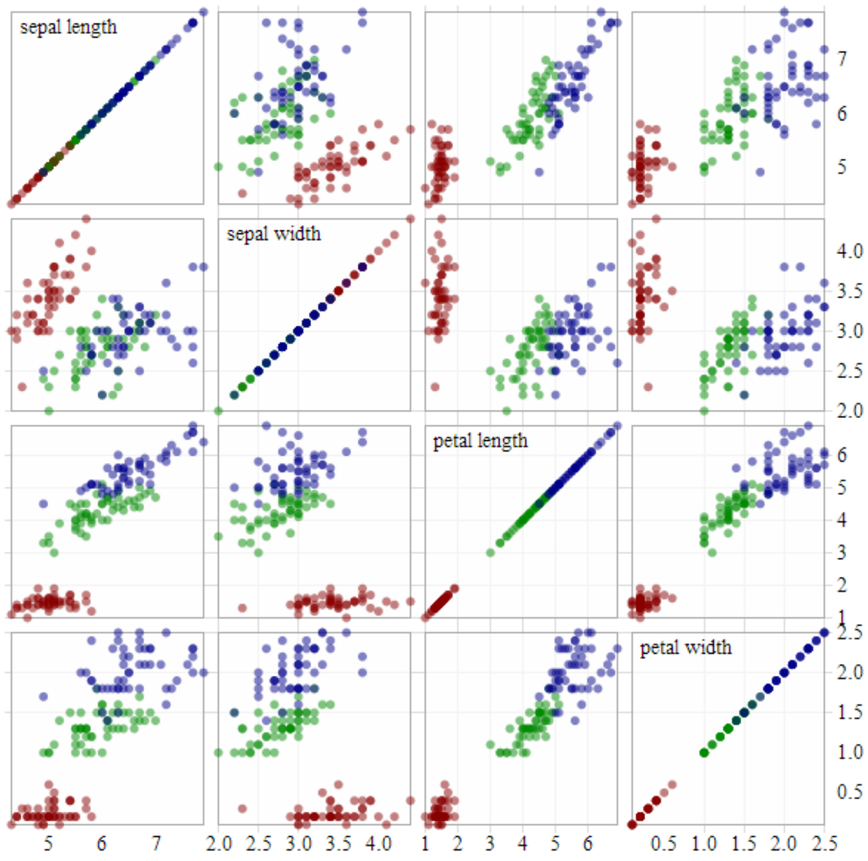
Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrices (SPLOM)

Matrix of size d*d

Each row/column is one dimension

Each cell plots a scatterplot of two dimensions

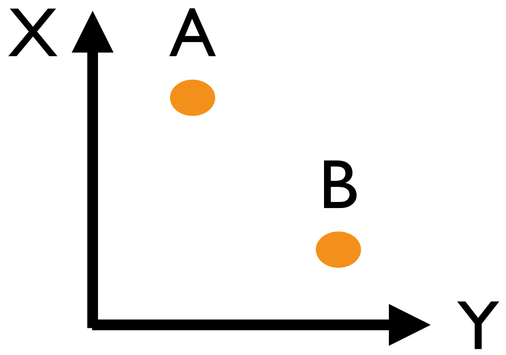


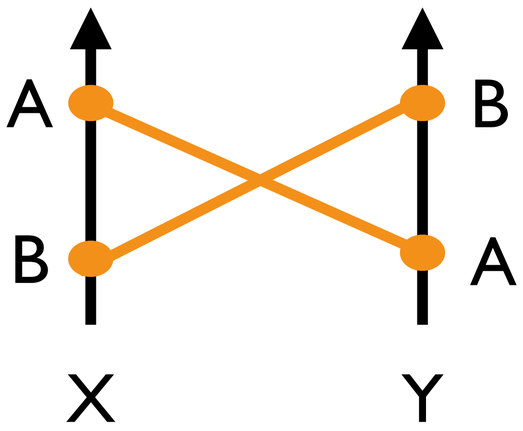
Parallel Coordinates (PC)

Inselberg 1985

Axes represent attributes

Lines connecting axes represent items





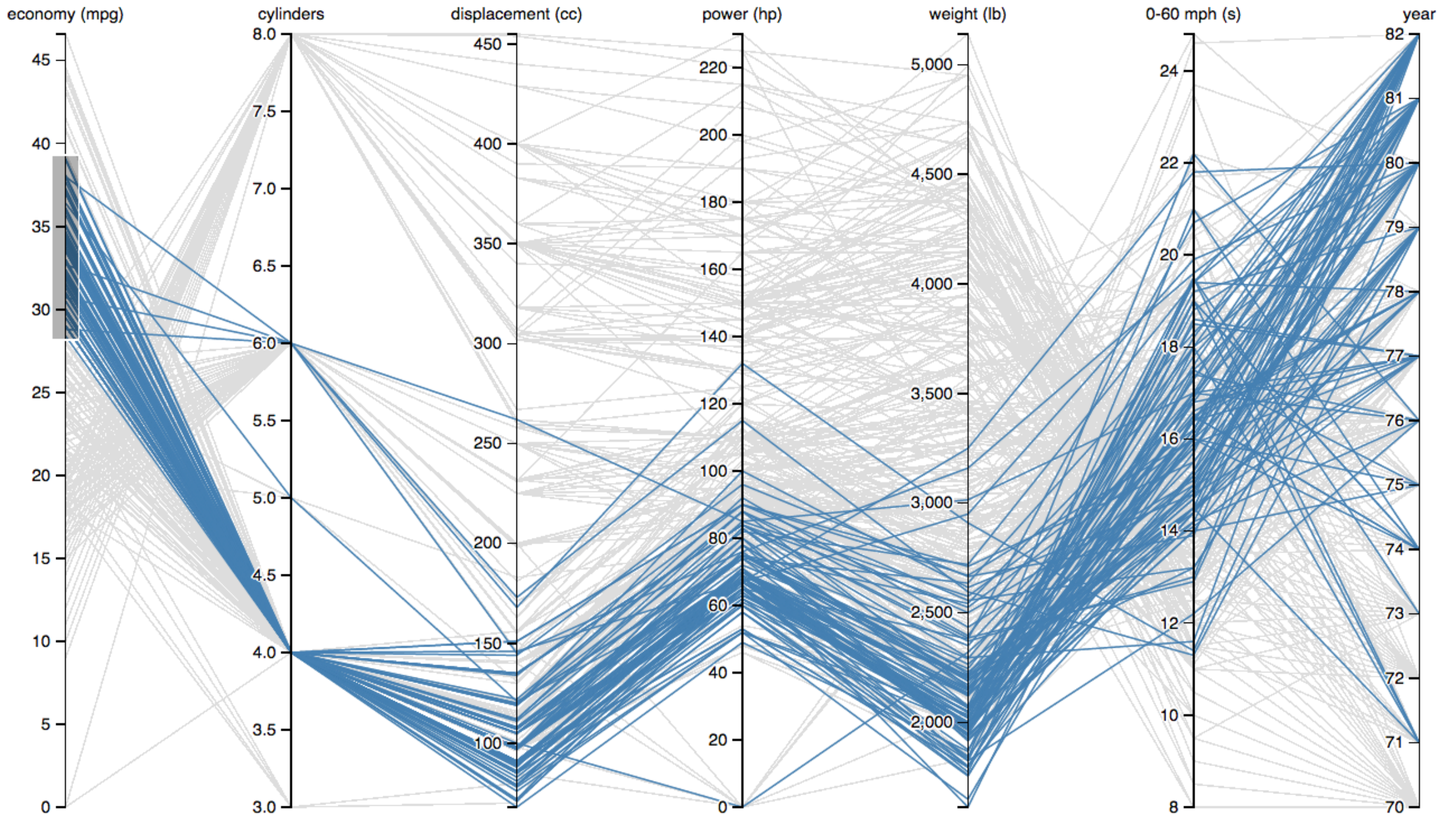
Parallel Coordinates

Each axis represents dimension

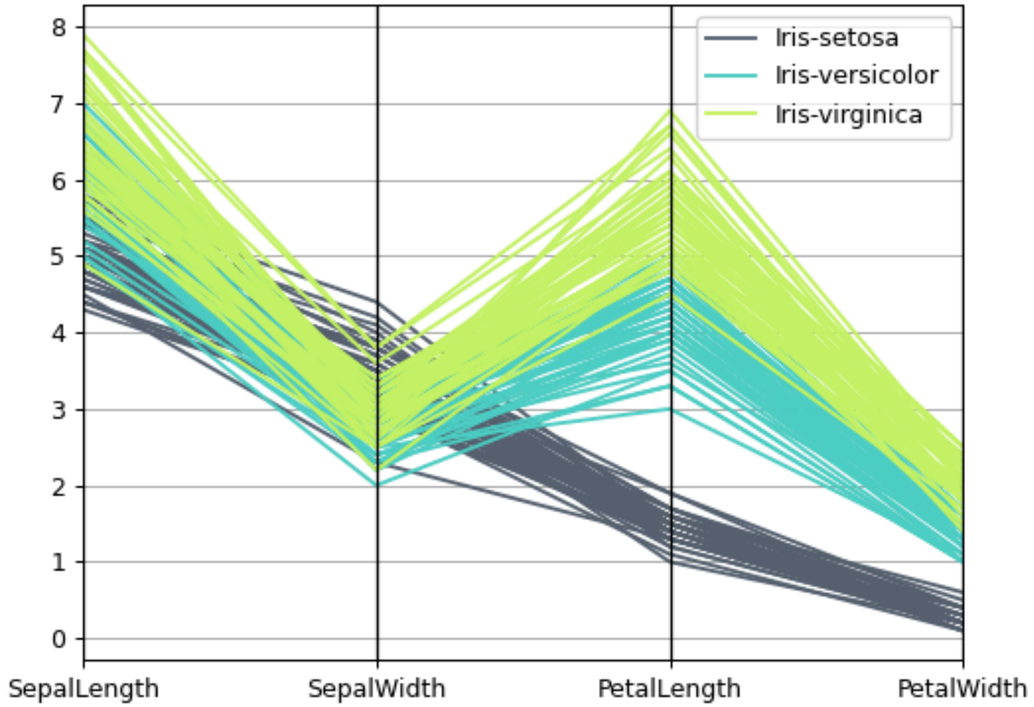
Lines connecting axis represent records

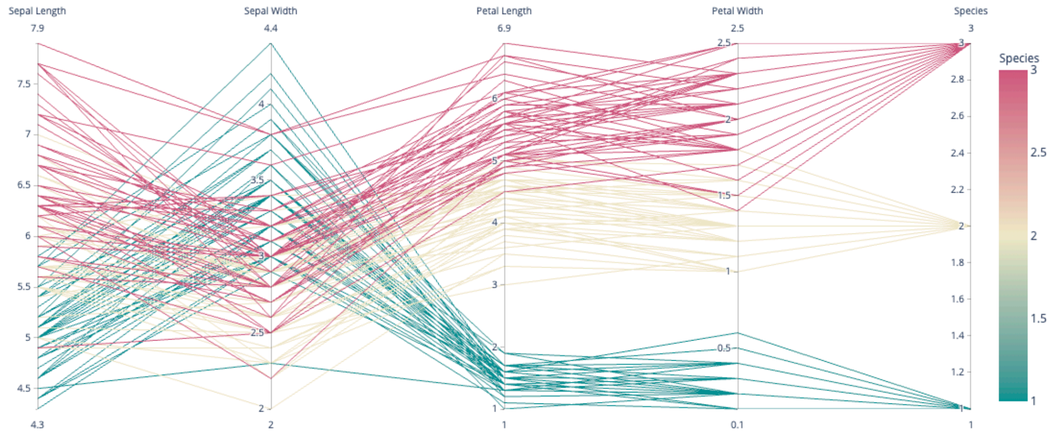
Suitable for

- all tabular data types
- heterogeneous data



Pandas and Plotly PCs





https://pandas.pydata.org/docs/reference/api/pandas.plotting.parallel_coordinates.html

https://plotly.com/python/parallel-coordinates-plot/

Pixel Based Displays

Each cell is a “pixel”, value encoded in color / value

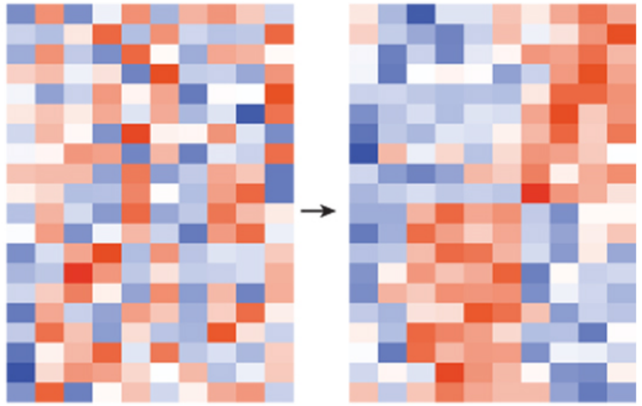
Ordering critical for interpretation

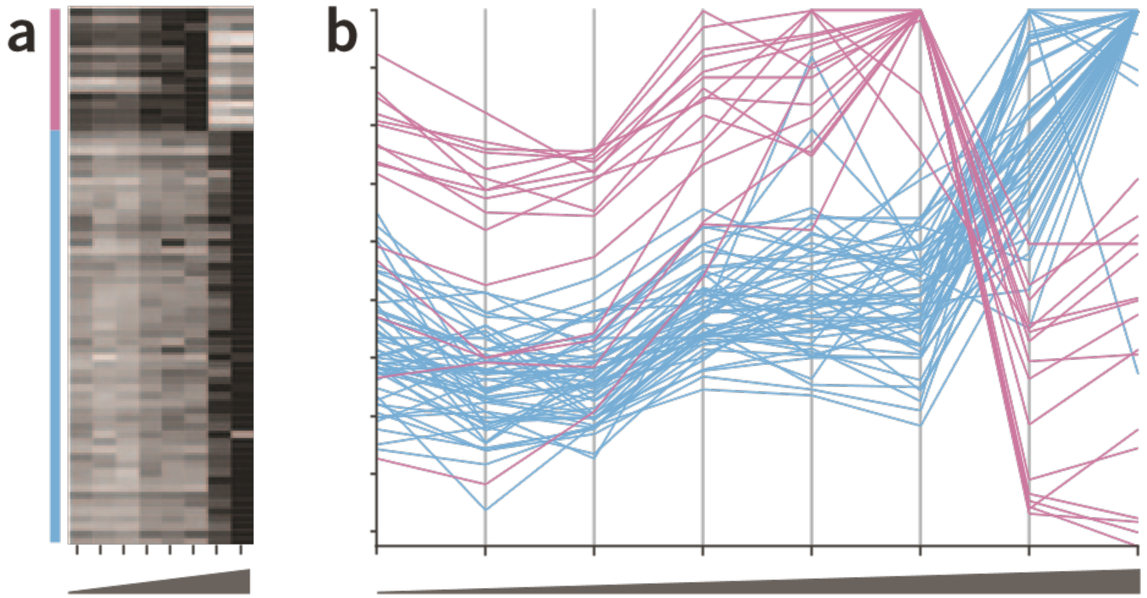
If no ordering inherent, clustering is used

Scalable – 1 px per item

Good for homogeneous data

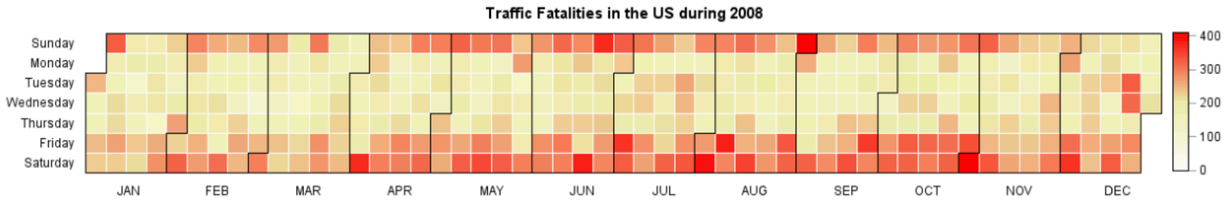
same scale & type





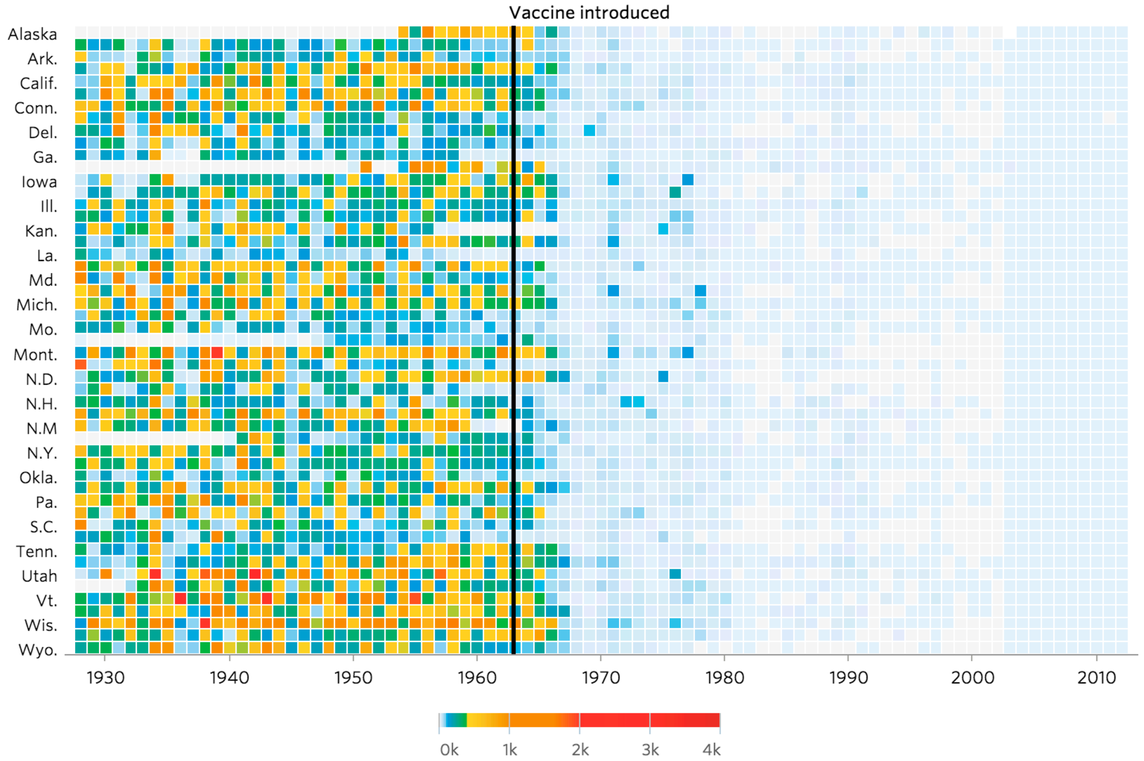
[Gehlenborg & Wong 2012]

Heat Map and Calendar Heat Map



The heat maps below show number of cases per 100,000 people.

Measles

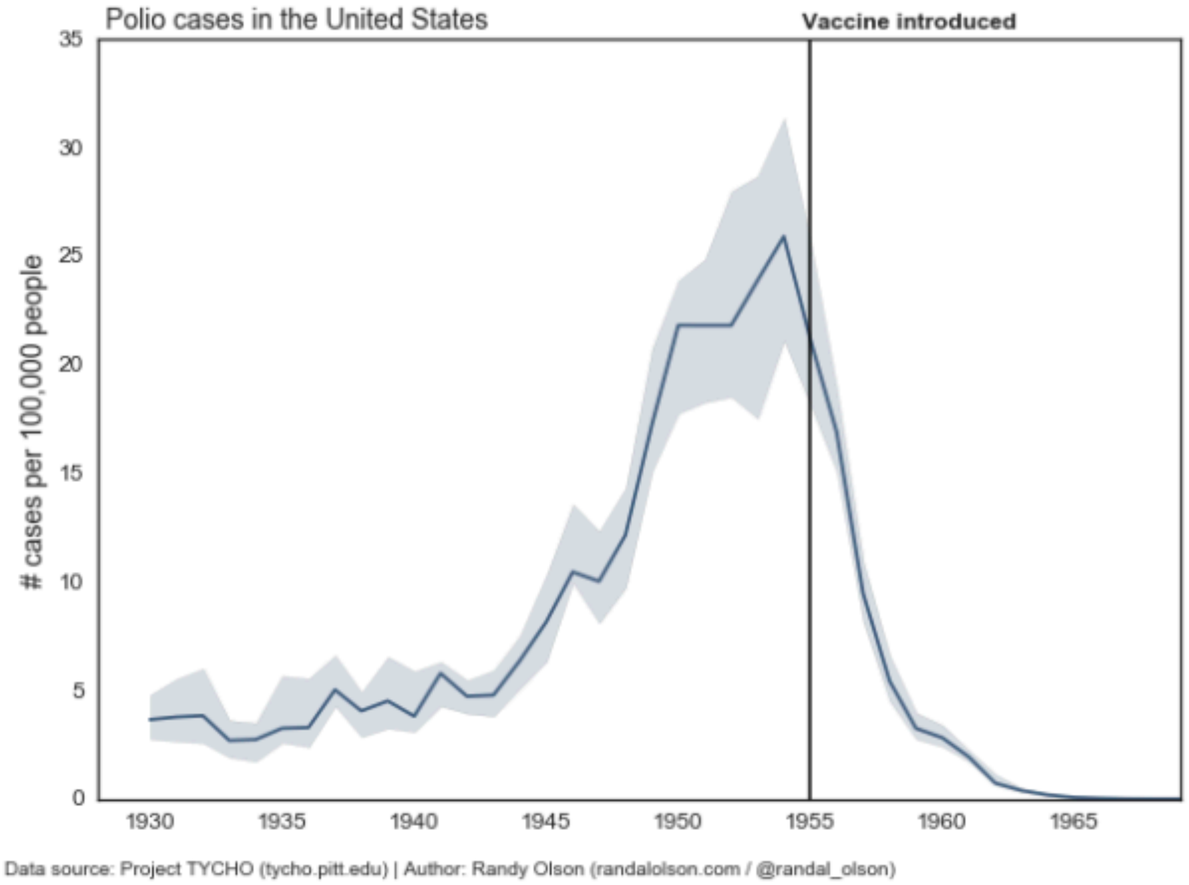


Note: CDC data from 2003-2012 comes from its Summary of Notifiable Diseases, which publishes yearly rather than weekly and counts confirmed cases as opposed to provisional ones.

https://blogs.sas.com/content/graphicallyspeaking/2011/12/08/calendar-heatmaps-in-gtl/

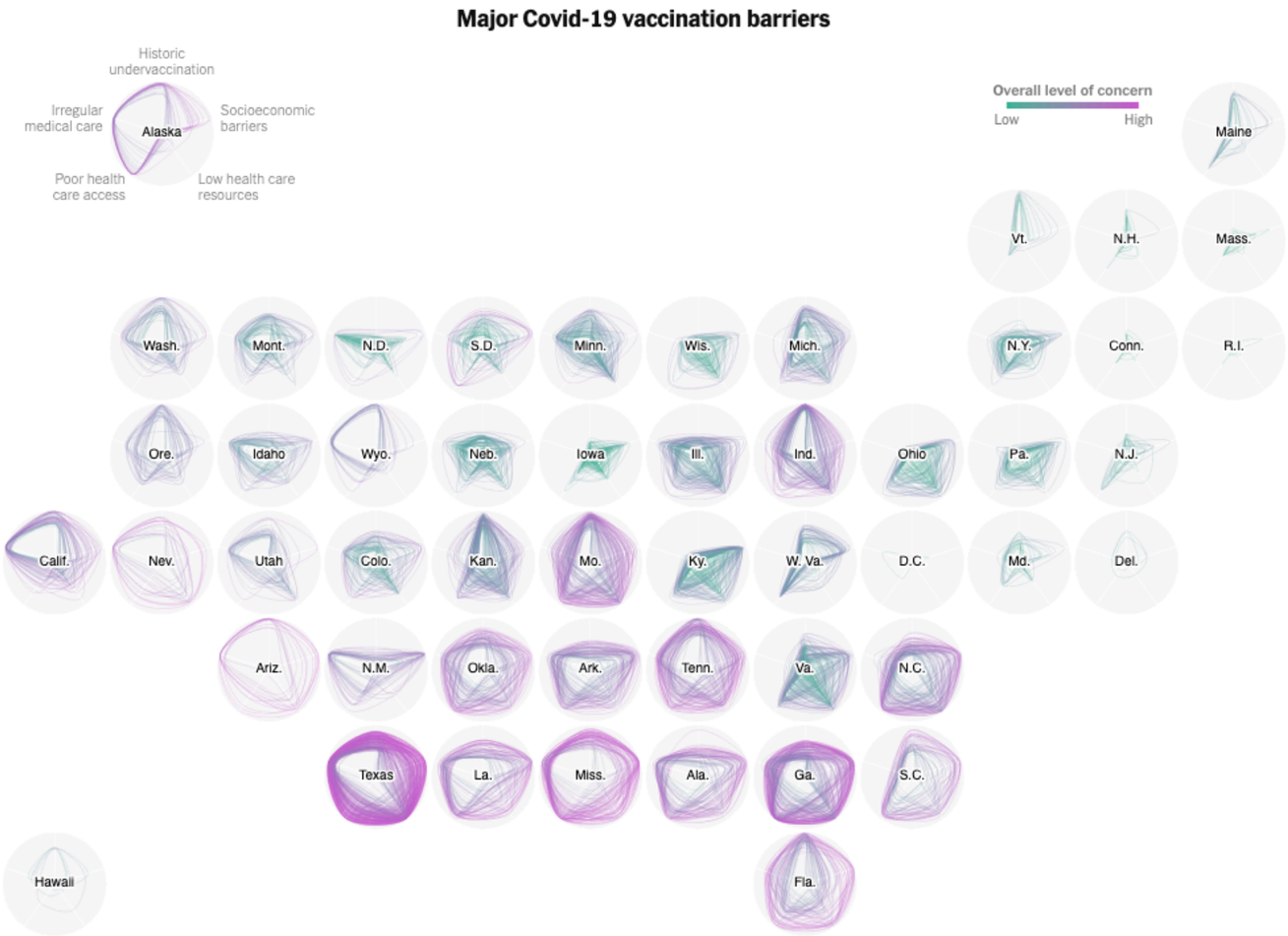
https://www.informationisbeautifulawards.com/showcase/660-vaccines-and-infectious-diseases

Sometimes you can Show Too Much Data



http://www.randalolson.com/2016/03/04/revisiting-the-vaccine-visualizations/

Critique: Vaccine Roadblocks



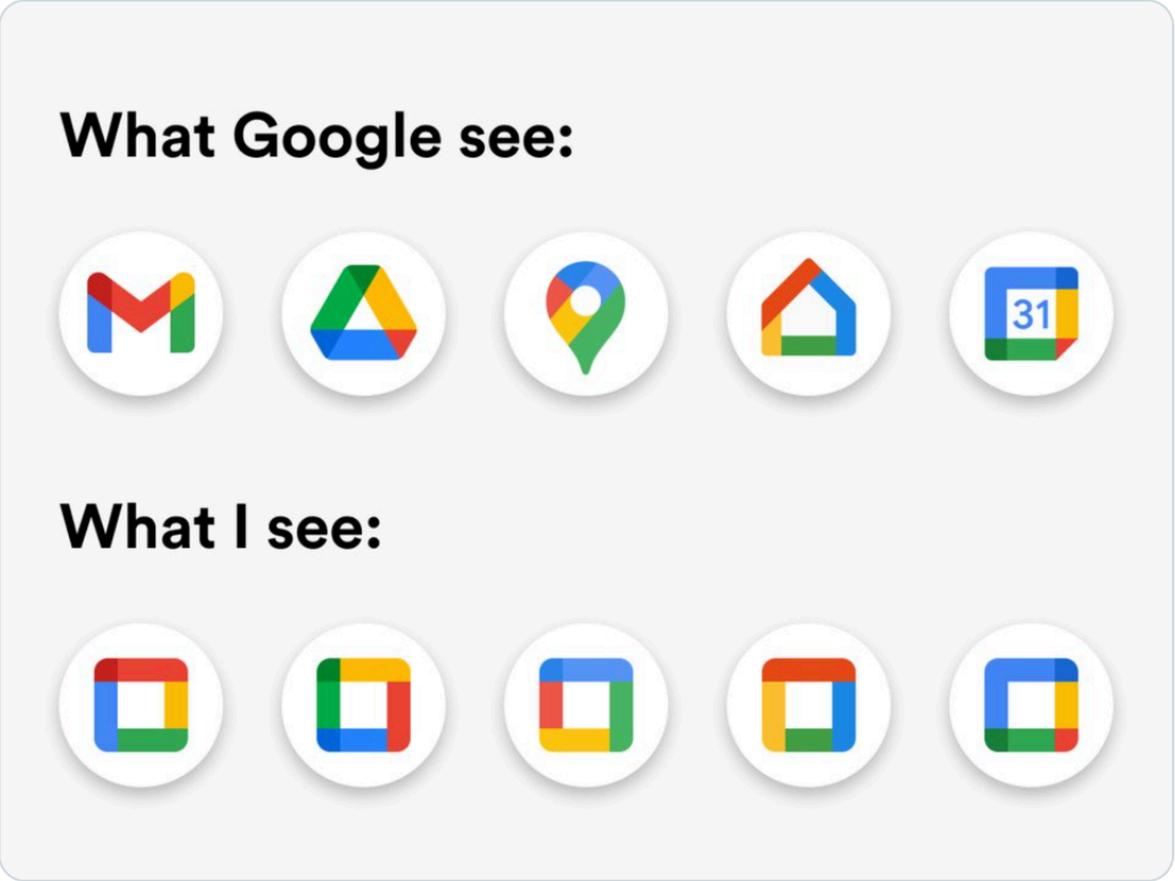
https://www.nytimes.com/interactive/2021/02/25/opinion/covid-vaccination-barriers.html

Image Credits: Google

https://techcrunch.com/2020/10/06/googles-new-logos-are-bad/

Frederic Lardinois
@fredericl

This is correct. (via reddit.com/r/google/comme...)



10:30 AM · Oct 26, 2020

19.4K Reply Copy link

Read 134 replies

Critique

1. Color: having distinct color is important
2. Shapes: having distinguishable shape is important

https://techcrunch.com/2020/10/06/googles-new-logos-are-bad/