Welcome (Again!) to

MATH 4100/COMP 5360– Introduction to Data Science

(Some) Principles of Data Visualization

February 20, 2024

Based on prior lectures from

Kate Isaacs, Alex Lex, and Bei Wang Phillips

+ others where noted

Announcements / Reminders / Schedule

HW 5 – Due this Friday!
HW 6 – Due Friday March 1!

Last Thursday: Visualization using Matplotlib, Seaborn and Altair

This class: basics of data visualization principles

This Thursday: Web-scraping

The purpose of computing is insight, not numbers.

- Richard Wesley Hamming

The purpose of visualization is insight, not pictures.

- Card, Mackinley, & Shneiderman

What is visualization?

One definition:

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

Good Data Visualization…

…makes data accessible

…combines the strengths of humans and computers

…enables insight

…communicates

Why Visualization?

To inform humans about complex situations:

Communication

What is the state of the election polls?

Why Visualization?

To inform humans about complex situations:

Communication

What is the state of the election polls?

When questions are not well defined:

Exploration

What is the structure of the scammer network?

Which drug can help my patient?

Is this data even right?

Why Visualization?

To inform humans about complex situations:

Communication

(when you know the data and want to share with others)

What is the state of the election polls?

When questions are not well defined:

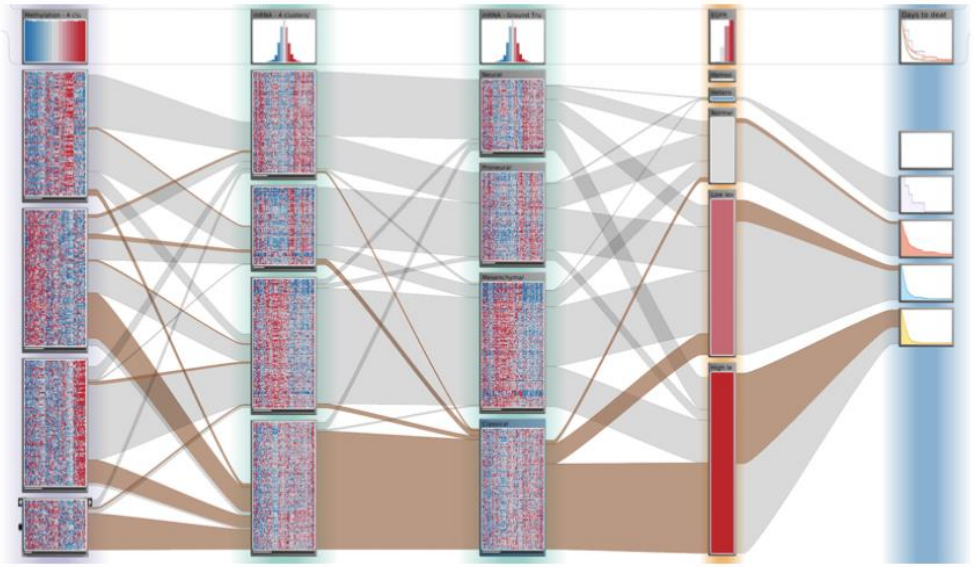
Exploration

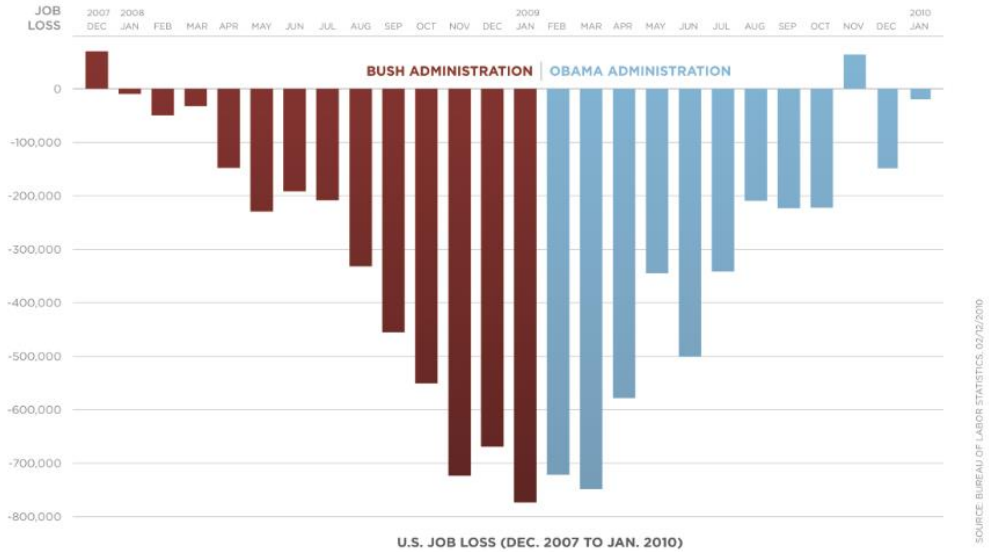
(when results and questions are unknown)

What is the structure of the scammer network?

Which drug can help my patient?

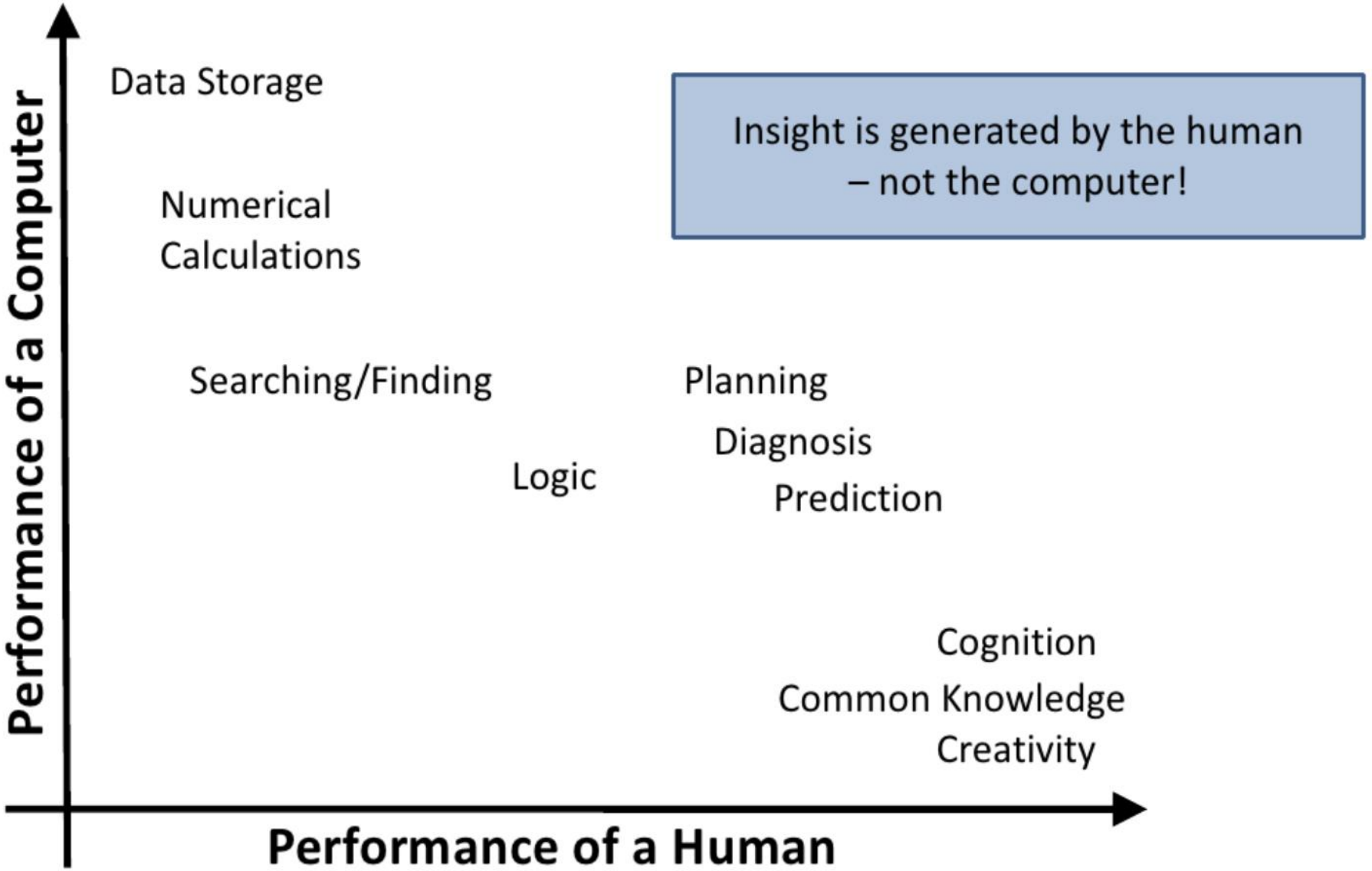
Is this data even right?



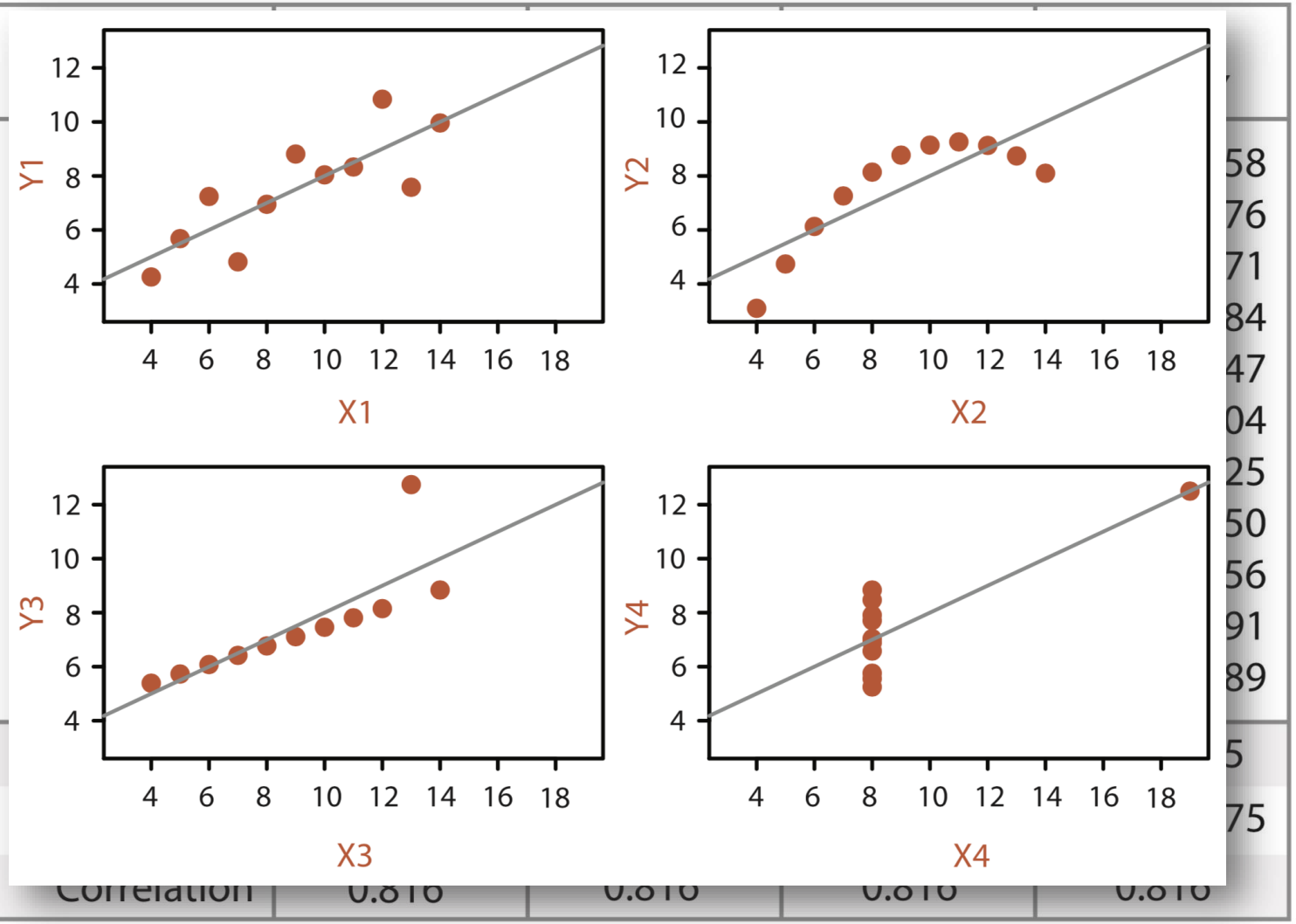


Open Exploration — Confirmation — Communication

Ability Matrix



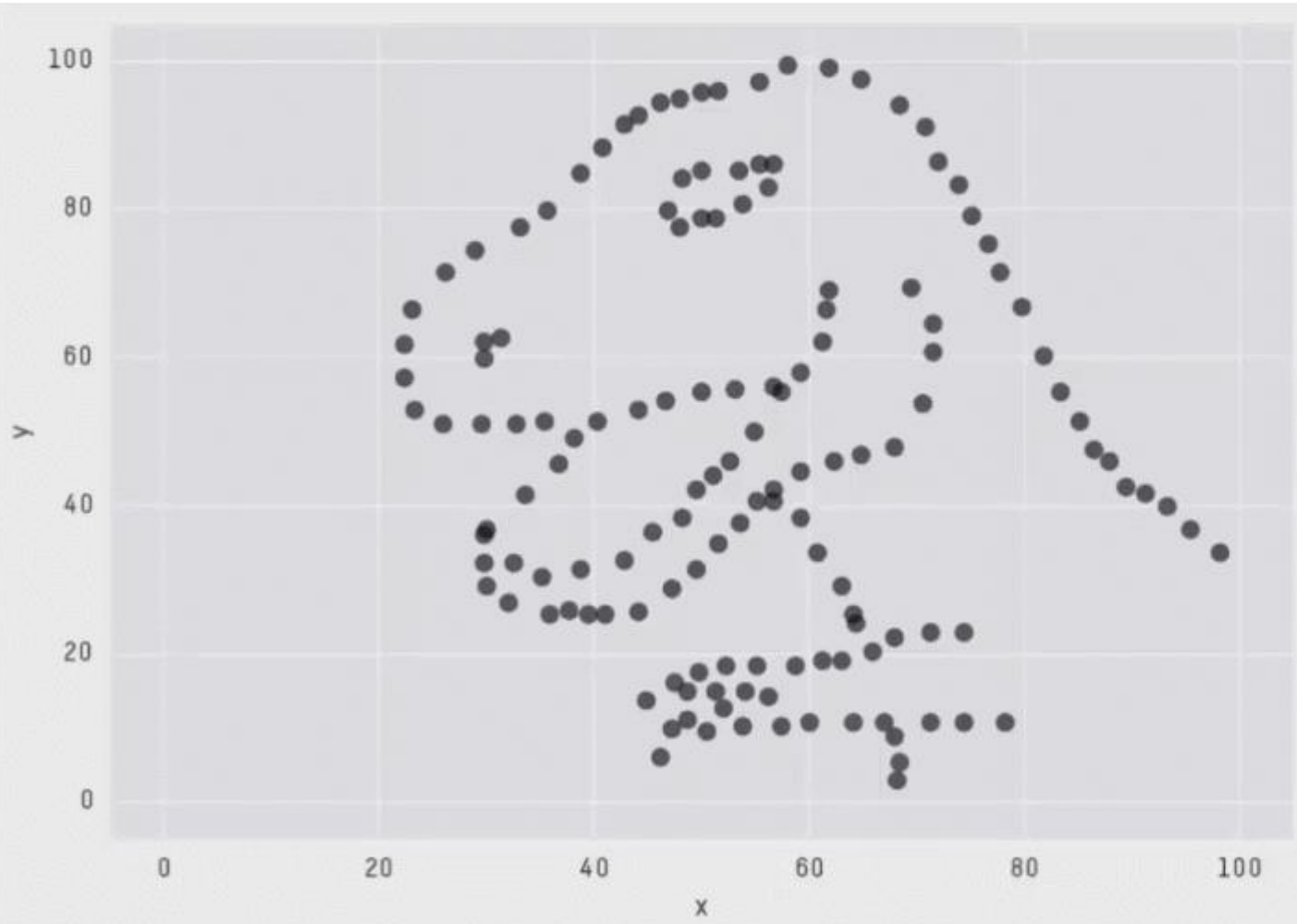
Why Visualization?



Sometimes summary statistics don't tell the full story.

Anscombe's Quartet figure from ***Visualization Analysis and Design***, Tamara Munzner

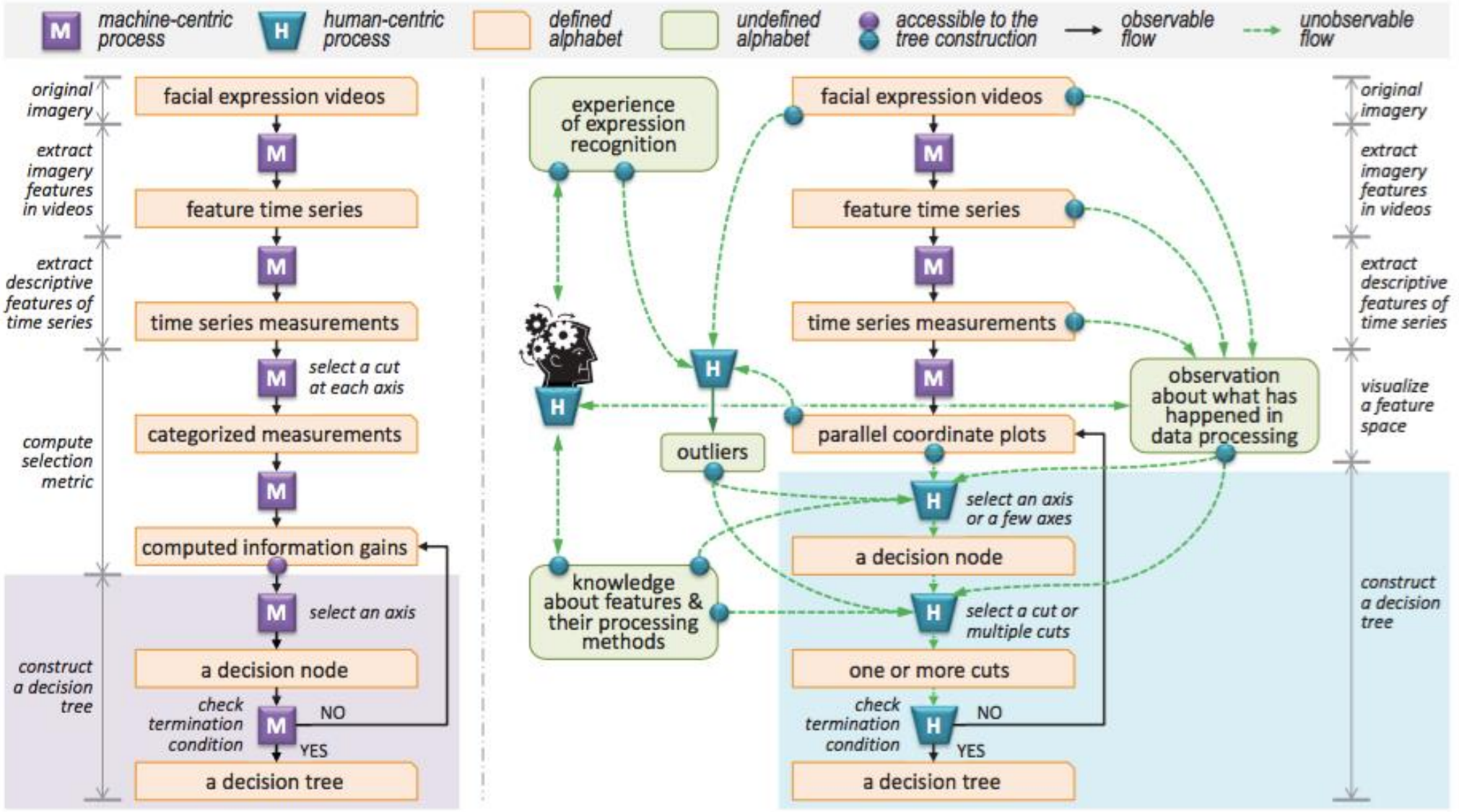
Why Visualization?



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

The Datasaurus Dozen of "**Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing**", Justin Matejka, George Fitzmaurice

Why Visualization?



Humans have access to information and context that pure automated methods don't.

From ***An Analysis of Machine- and Human-Analytics in Classification***, Lam et al. 2017

What can visualization do?

- Answer vague questions
- Answer multiple questions simultaneously
- Help generate new questions
- Help generate hypotheses
- Help find patterns
- Act as external memory
- Communicate to others
- Help “debug” your data
- Explain to others
- Please and delight

Visualization is…

Human Data Interaction

Visualization in the Data Science Process

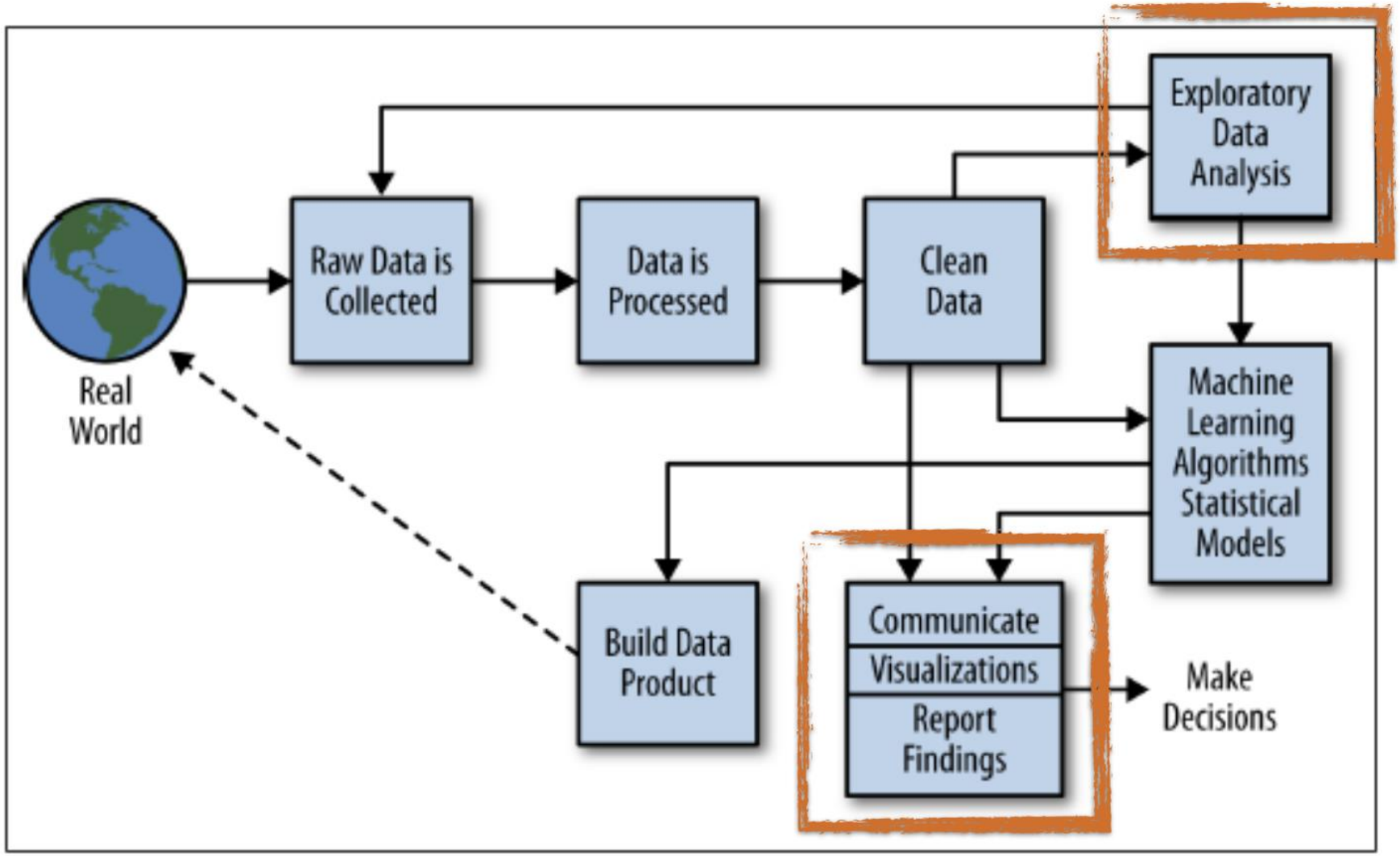
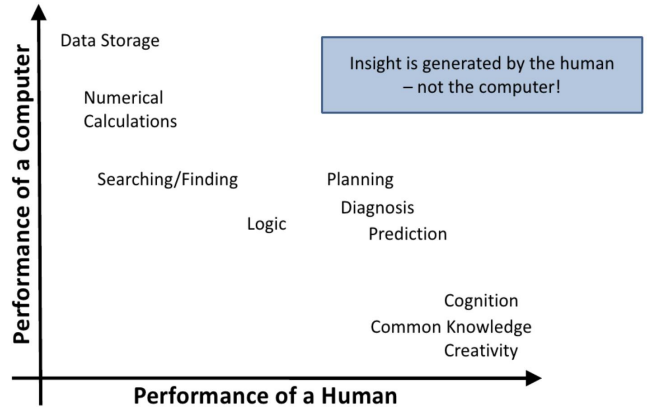


Figure 2-2. The data science process

Why Humans?



Leverage human capabilities:

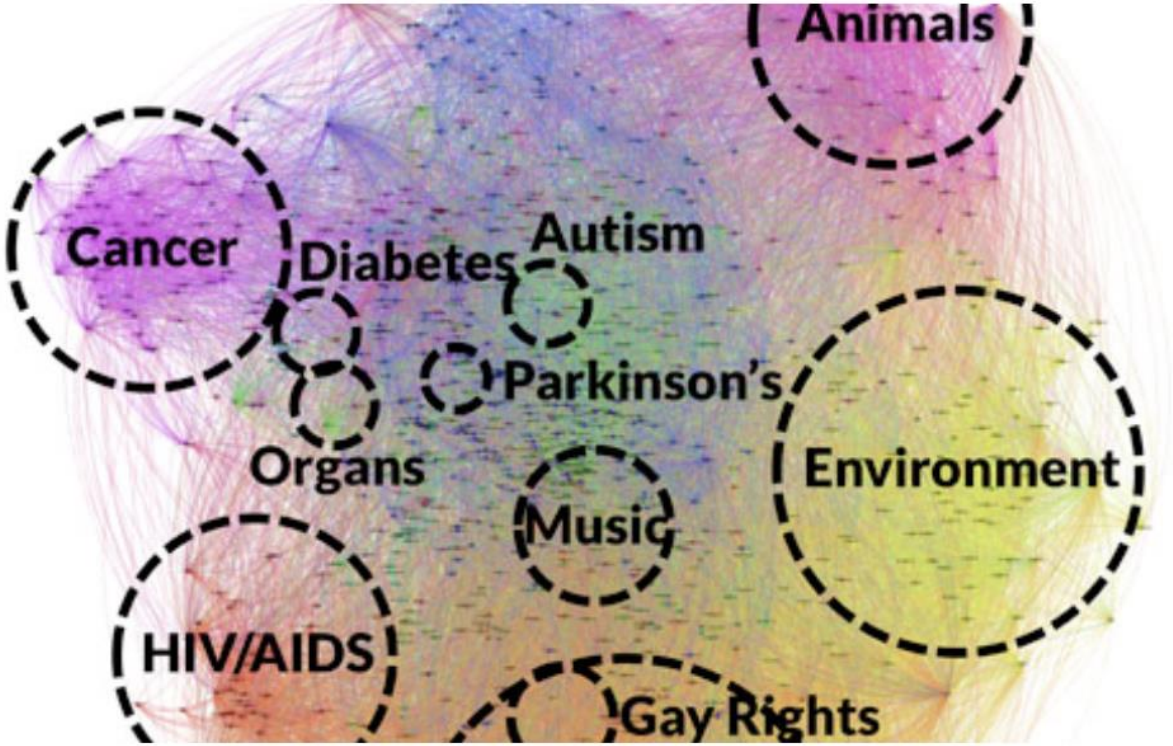
Pattern discovery: clusters, outliers, trends

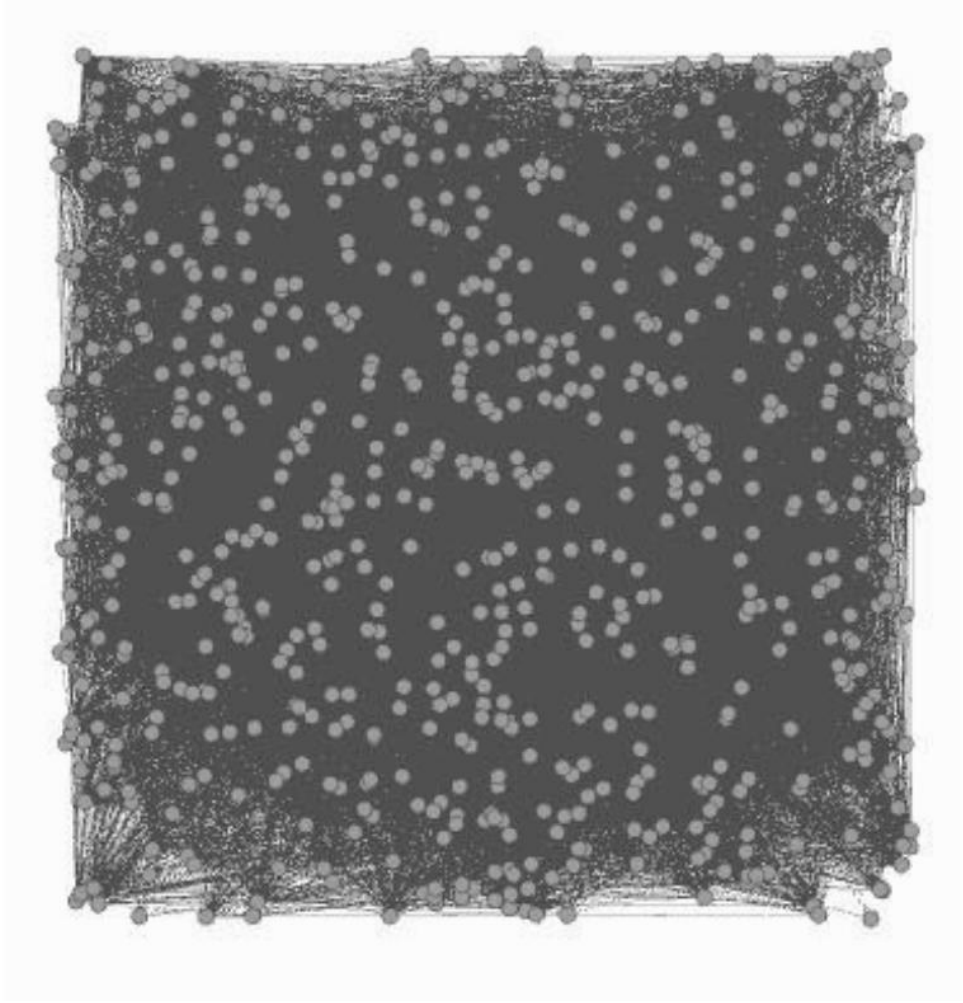
Contextual knowledge: expectations, explanations of patterns

Action: humans learn and take action

Because humans, we also have to ***design for humans their limitations***.

Just because we can draw it doesn't mean a human can read it.





What insight do we get from these?

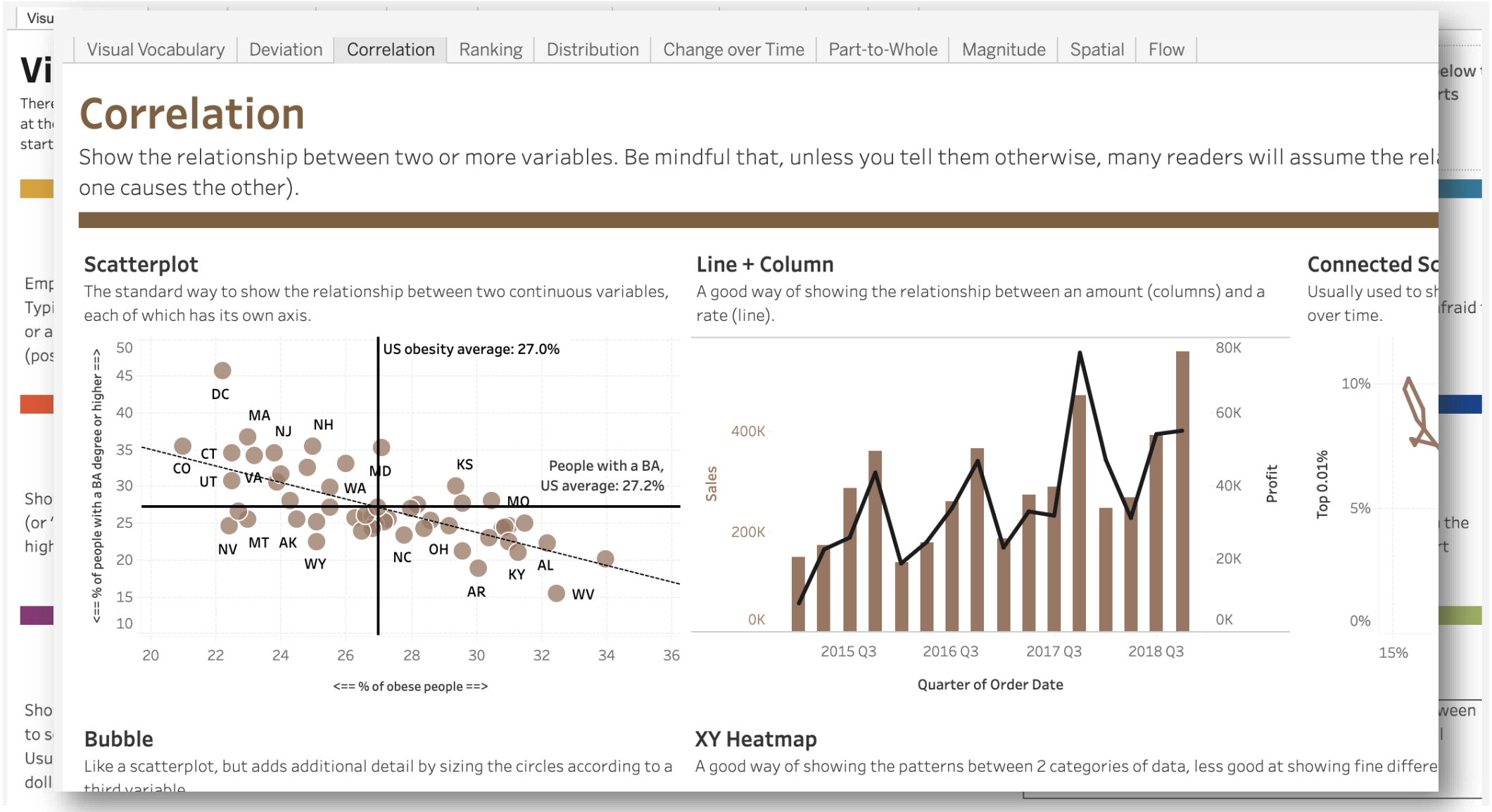
What chart should I use?

What are you trying to understand?

What is the nature of my data?

What is the nature of my task?

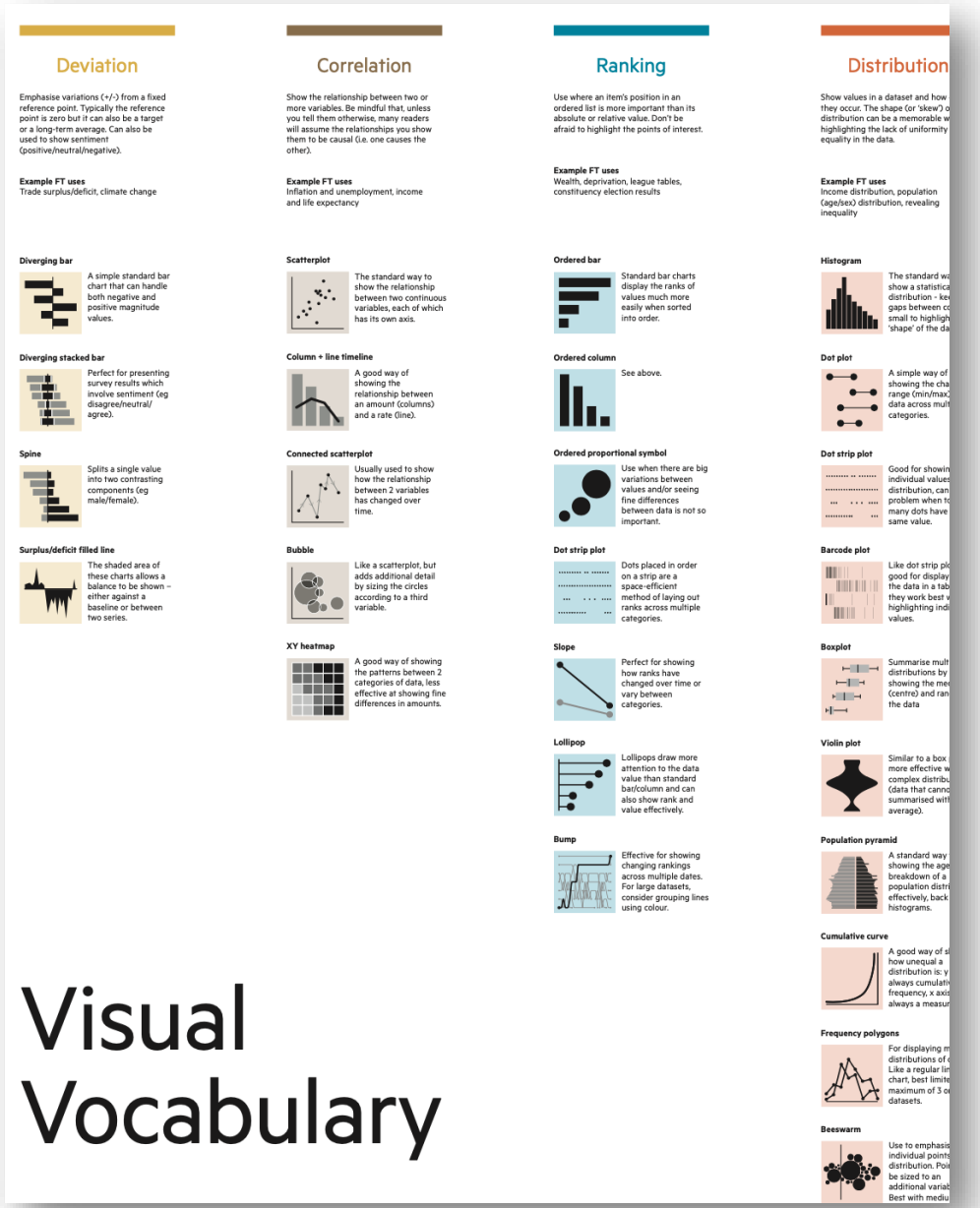
Some sources offer “visual vocabularies” for common abstract goals



https://public.tableau.com/en-us/gallery/visual-vocabulary

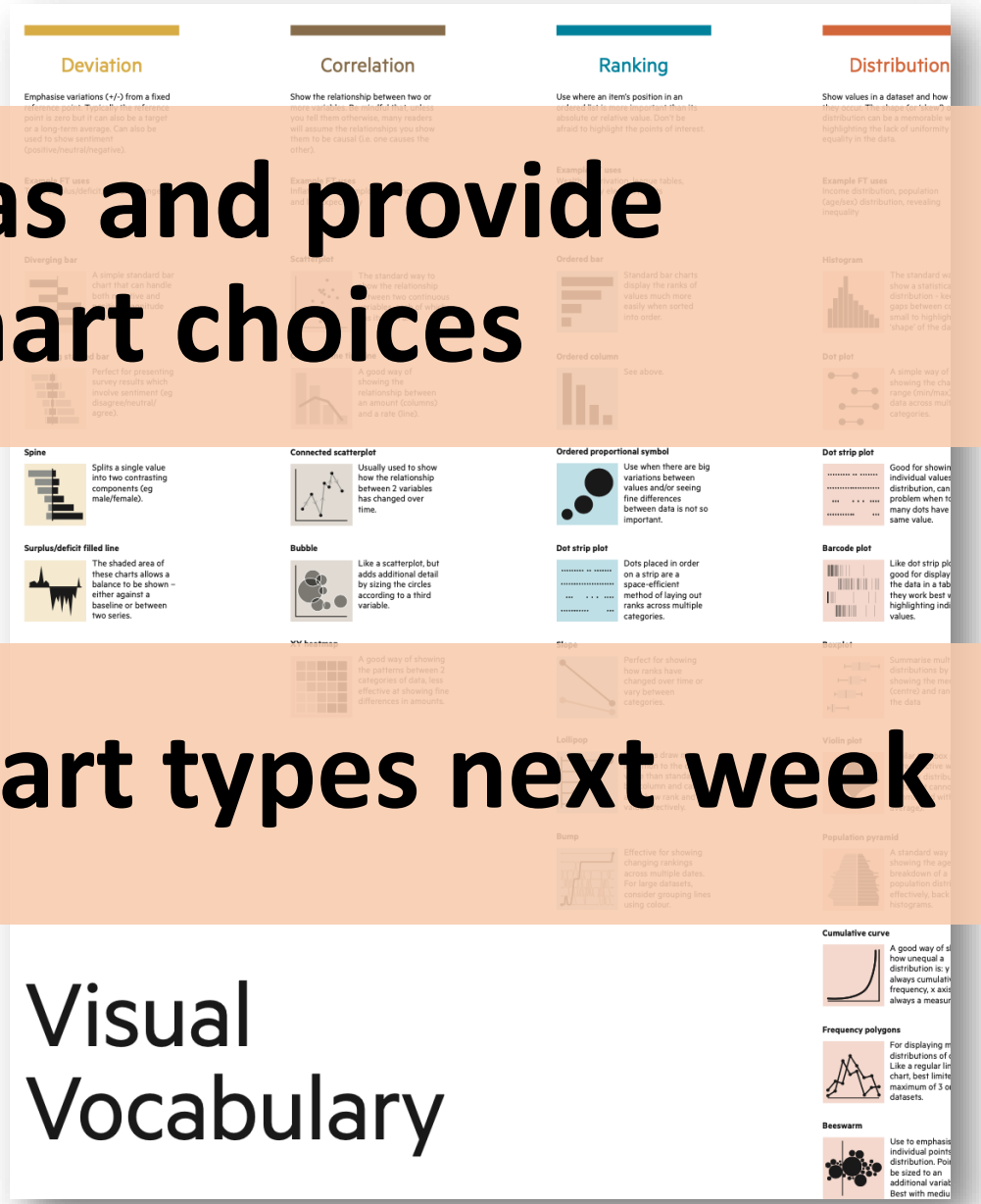
datavizcatalogue.com

Financial Times



dataviscatalogue.com

Financial Times



These can give you ideas and provide rationale for your chart choices

We'll go over some more chart types next week

Describing Visualizations

Marks & Channels

Marks, Channels, & Encoding

Encoding: Map data to visual structure

Marks: Graphical primitives that encode items / entities

Channels: Properties of mark appearance, often used to encode attributes or other information

Marks, Channels, & Encoding

Seem familiar?

Encoding: Map data to visual structu

Marks: Graphical primitives that enc

Channels: Properties of mark appea
attributes or other information

```
# Create a visualization
sns.relplot(
    data=tips,
    x="total_bill", y="tip",
    hue="smoker", size="size",
)
```

```
alt.Chart(movies_genre).mark_tick().encode(
  x='AvgRating'
)
```

Encodings

The next step is to add *visual encoding channels* (or *encodings* for short) to the chart. An encoding channel specifies how a given data column should be mapped onto the visual properties of the visualization. Some of the more frequently used visual encodings are listed here:

- `x` : x-axis value
- `y` : y-axis value
- `color` : color of the mark
- `opacity` : transparency/opacity of the mark
- `shape` : shape of the mark
- `size` : size of the mark
- `row` : row within a grid of facet plots
- `column` : column within a grid of facet plots

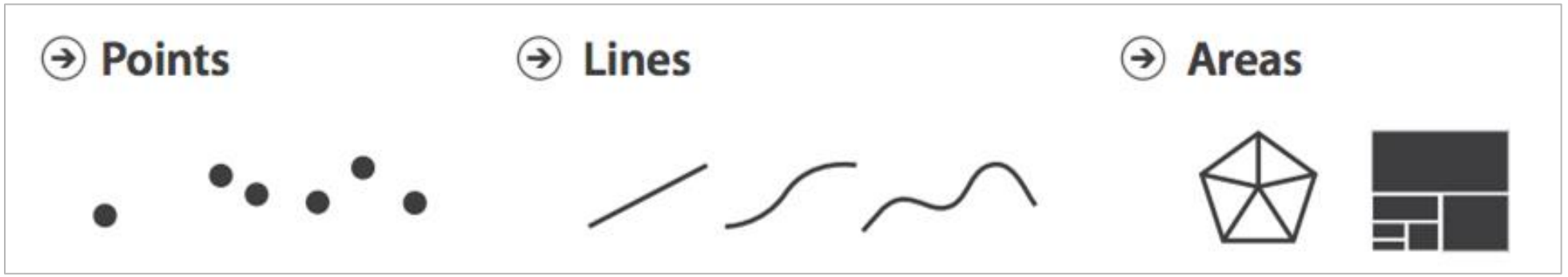
For a complete list of these encodings, see the Encodings section of the documentation.

Visual encodings can be created with the `encode()` method of the `Chart` object.

```
alt.Chart(movies_genre).mark_point().encode(
  y='AvgRating', size="Watches"
```

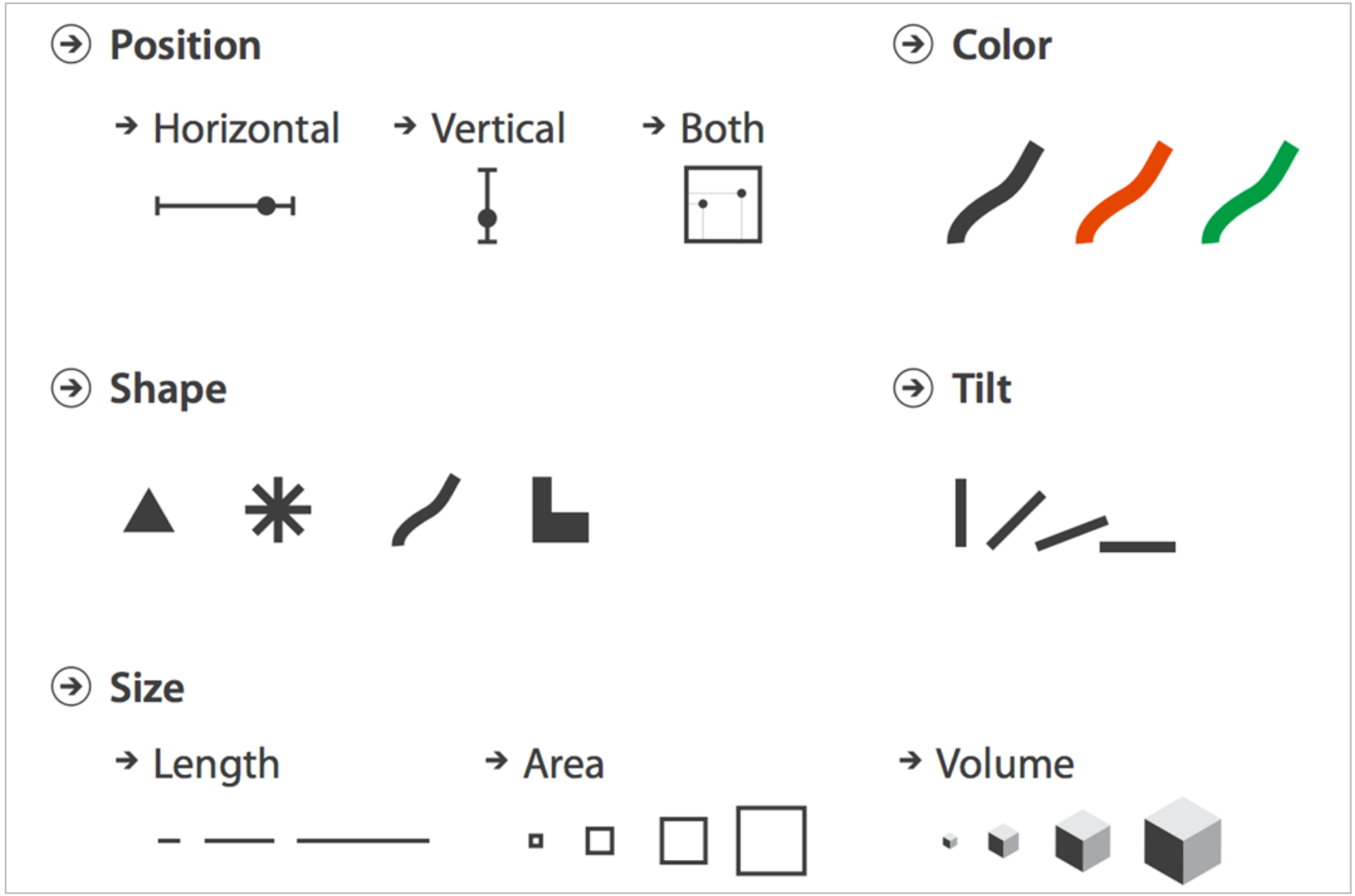
Marks, Channels, & Encoding

Marks: Graphical primitives that encode items / entities



Images from Munzner, Visualization Analysis and Design

Channels: Properties of mark appearance, often used to encode attributes or other information

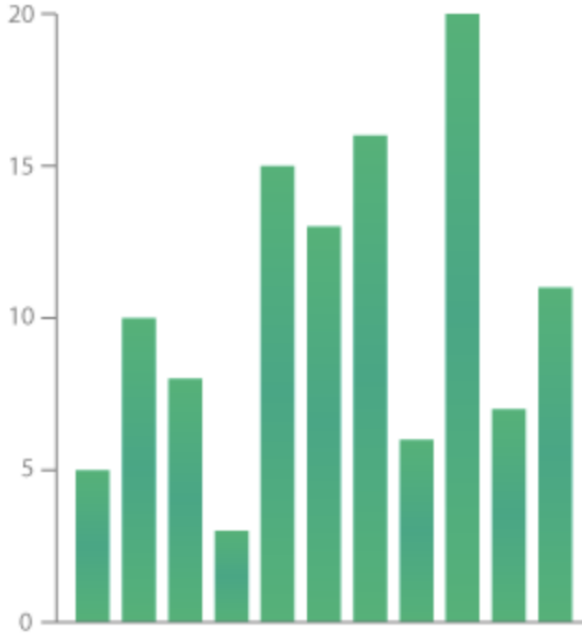


Images from Munzner, Visualization Analysis and Design

What are the marks & channels of…

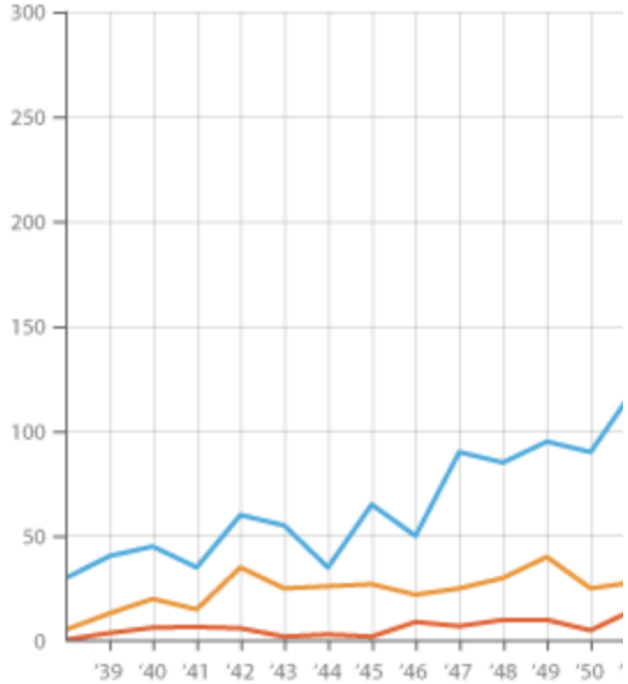
Pie Chart

Marks: Wedge
Channels: Color, Angle (not Area, not Size!)



Bar Chart

Marks: Rectangle
Channels: x-position, Length



Line Chart

Marks: Line (Path)
Channels: x,y-positions, Color

Images from datavizcatalogue.com

Expressiveness & Effectiveness

Expressiveness Principle: Encoding should express all of, and only, the information in the data

- Example: Don't imply order where this is not but imply order where there is

Effectiveness Principle: The more important the data/attribute, the more **salient** the encoding should be

- Important things should be noticeable

From Munzner, Visualization Analysis and Design

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

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

Shape

Choosing channels…

Images from Munzner, Visualization Analysis and Design

Loved & Dangerous – The Color Channel

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

Same

Most

Effectiveness

Identity Channels: Categorical Attributes

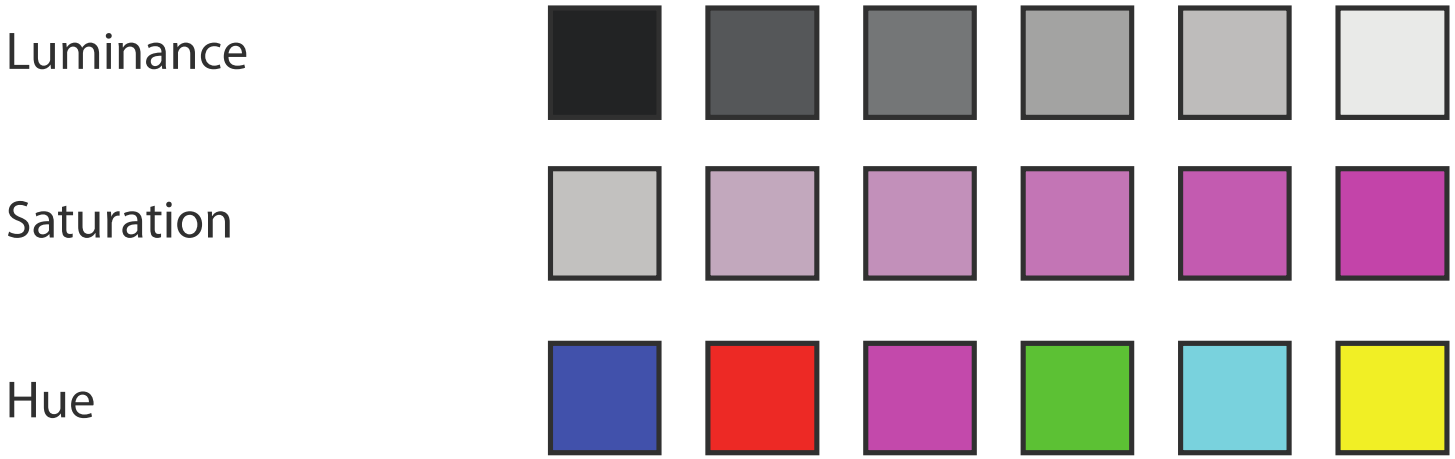
Spatial region

Color hue

Motion

Shape

Which has order? Which doesn't?



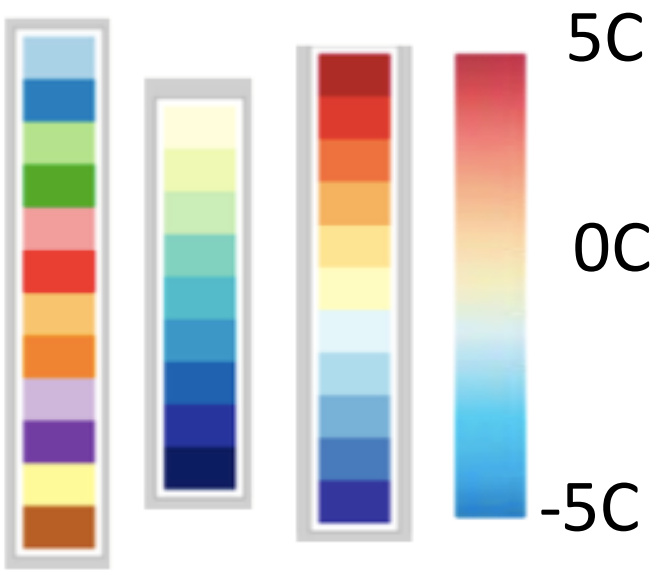
Vary the appropriate dimension depending on your data type

Ramps from Munzner, Visual Analysis and Design

Color maps specify a mapping between color and value

If you are using color to encode a value, you should include a representation of the color map.

Don’t forget to label their meaning!



Images from Josh Levine

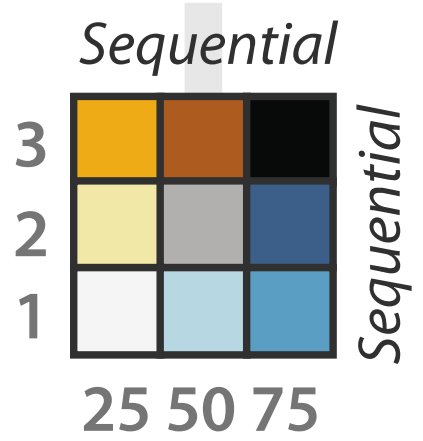
Color map axes: Design color map to match the attribute(s) you are encoding

Categorical vs. Ordered

Diverging vs. Sequential

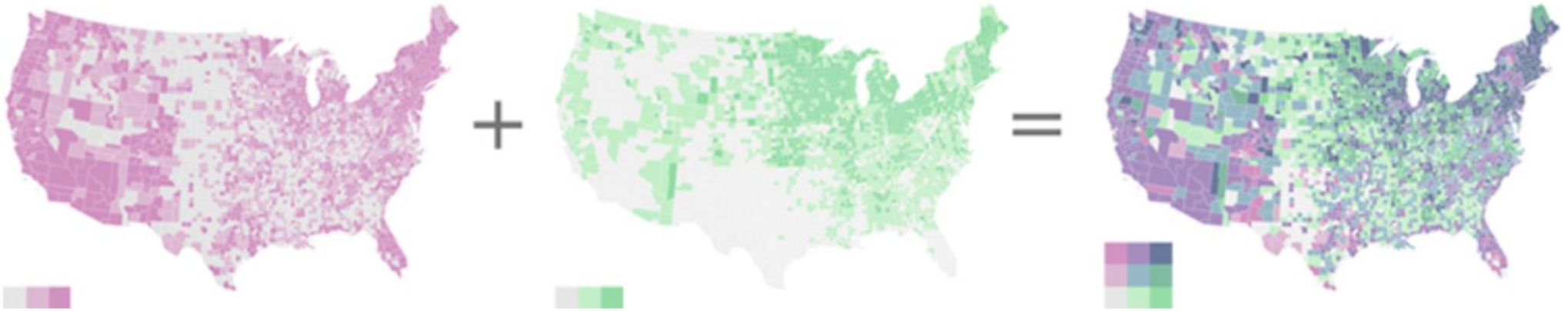
Segmented vs. Continuous

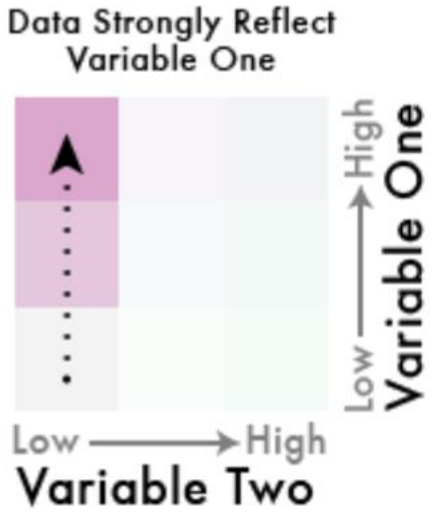
Univariate vs. Bivariate

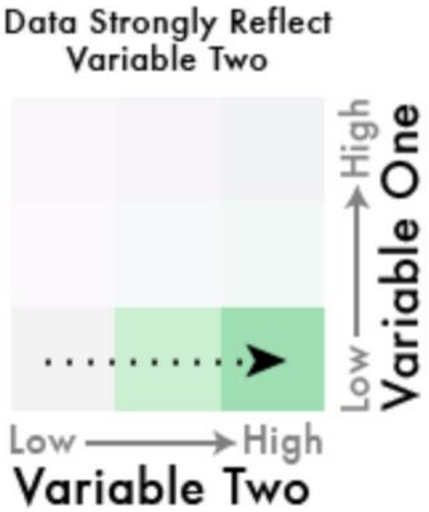


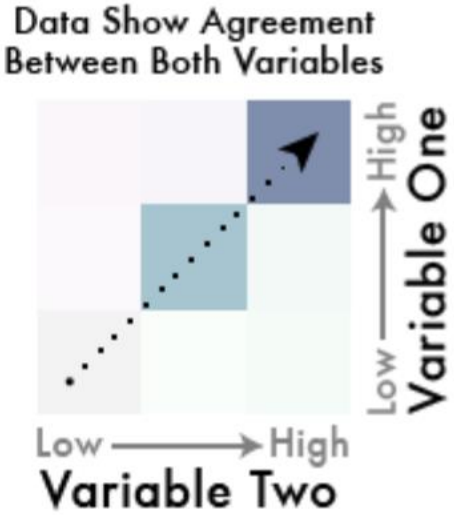
Adapted from Josh Levine

Bivariate Example





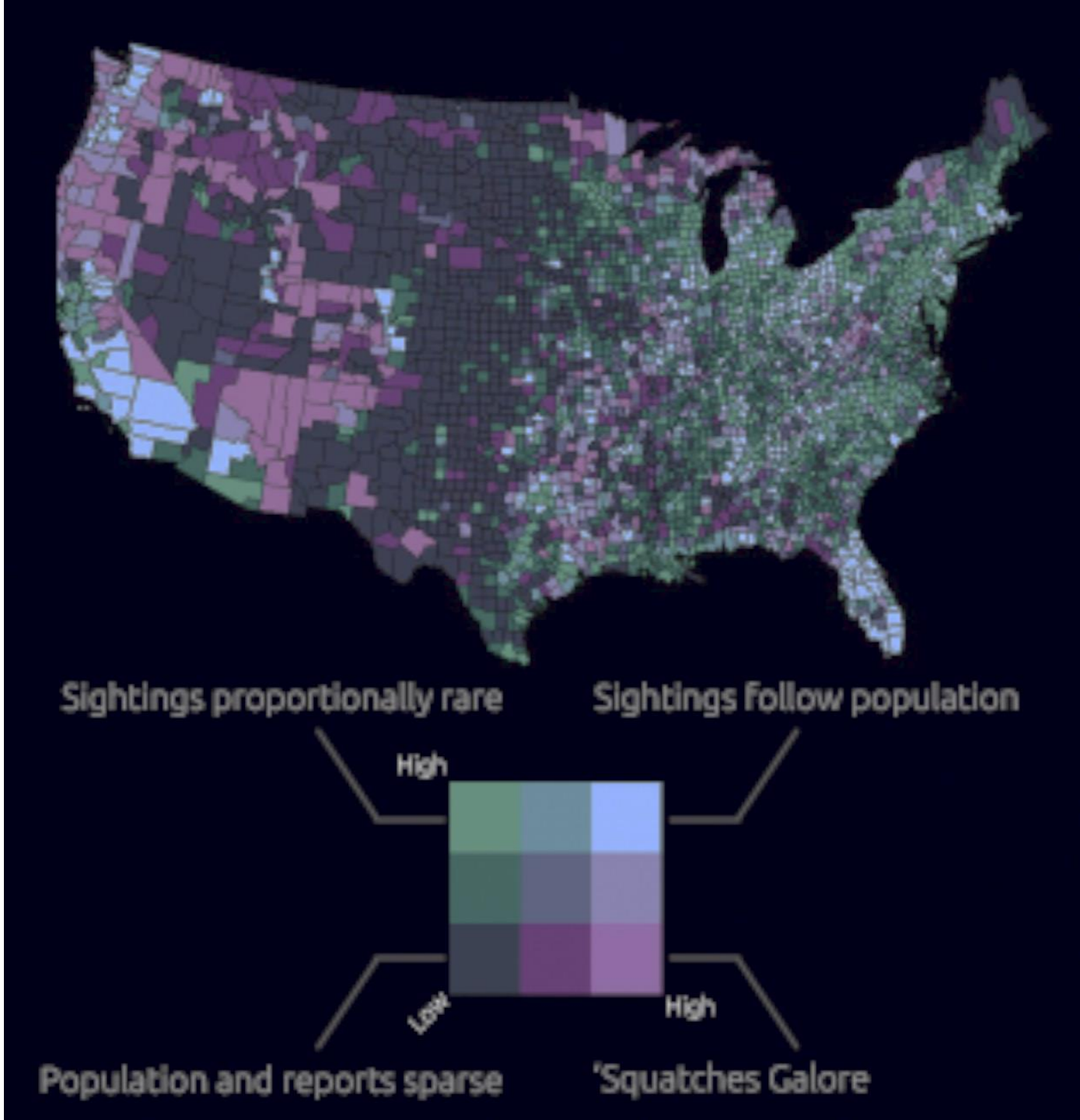




From josuhuastevens.net

Bivariate Example

Population
&
Sasquatch (BigFoot) Sightings



…not so easy to read.

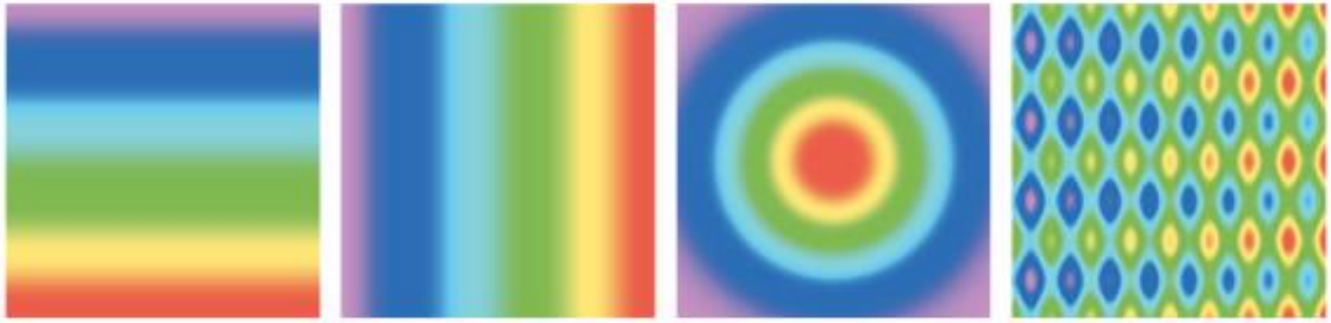
Consider also using
hue/lightness,
hue/thickness of border,
hue + mark&channel…

From josuhuastevens.net & Ware, Information Visualization

The Dangers of Rainbows

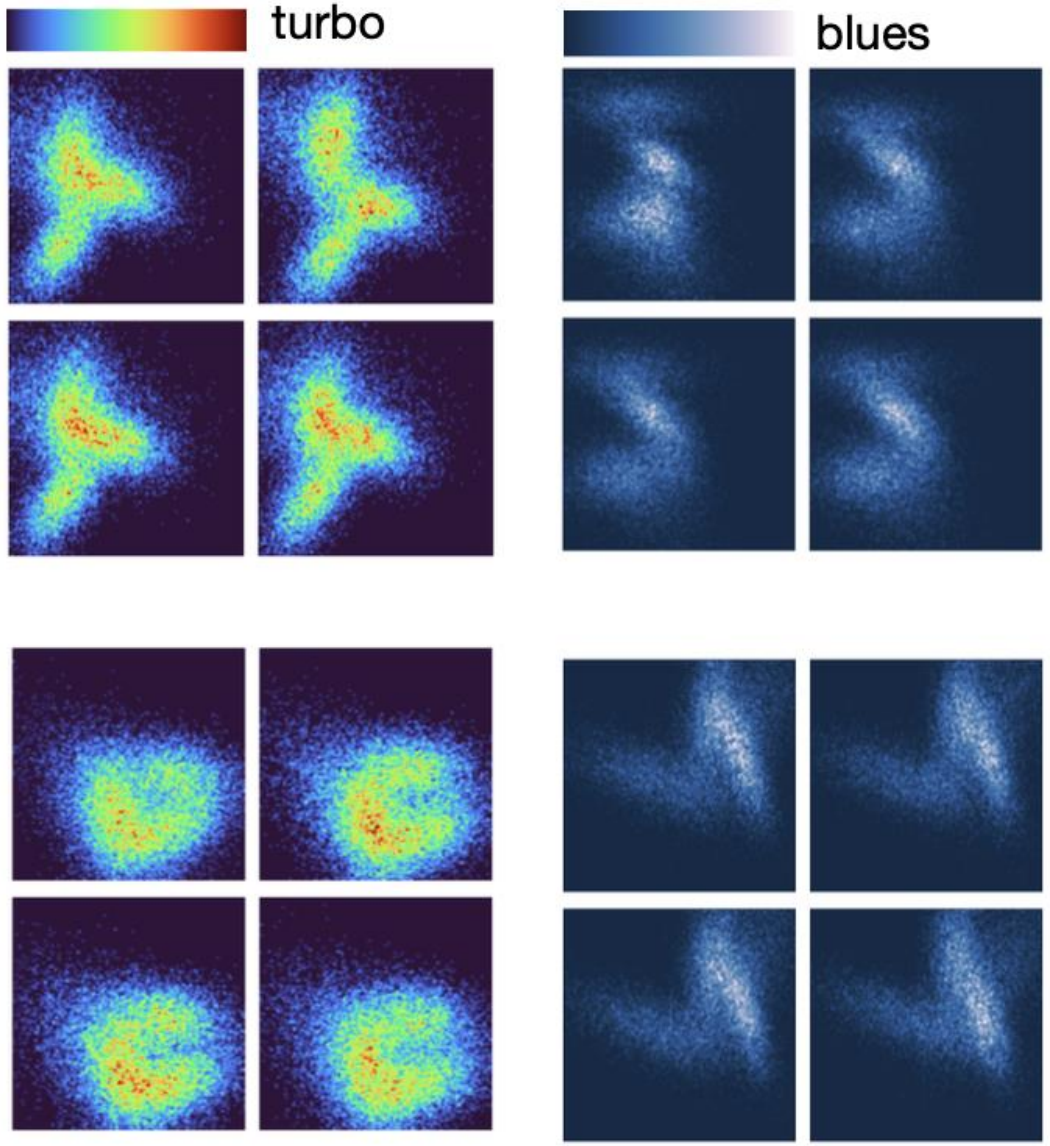
The common rainbow color map is often a poor choice due to:

- Lack of perceptual linearity
- Using hue for ordering
- Using hue for fine-grained detail



Munzner, Visualization Analysis and Design, with images from slides of Josh Levine (left) from "Rainbow Color Map (Still) Considered Harmful" (right)

The Dangers of Dismissing Rainbows



Some rainbow color maps may be helpful even in continuous tasks.

- Some tasks may benefit from the ability to name colors

- Detection of false features may not be as pervasive as we thought

From “Rainbows Revisited: Modeling Effective Colormap Design for Graphical Inference” Reda and Szafir, 2021

Different color maps are good for different insights

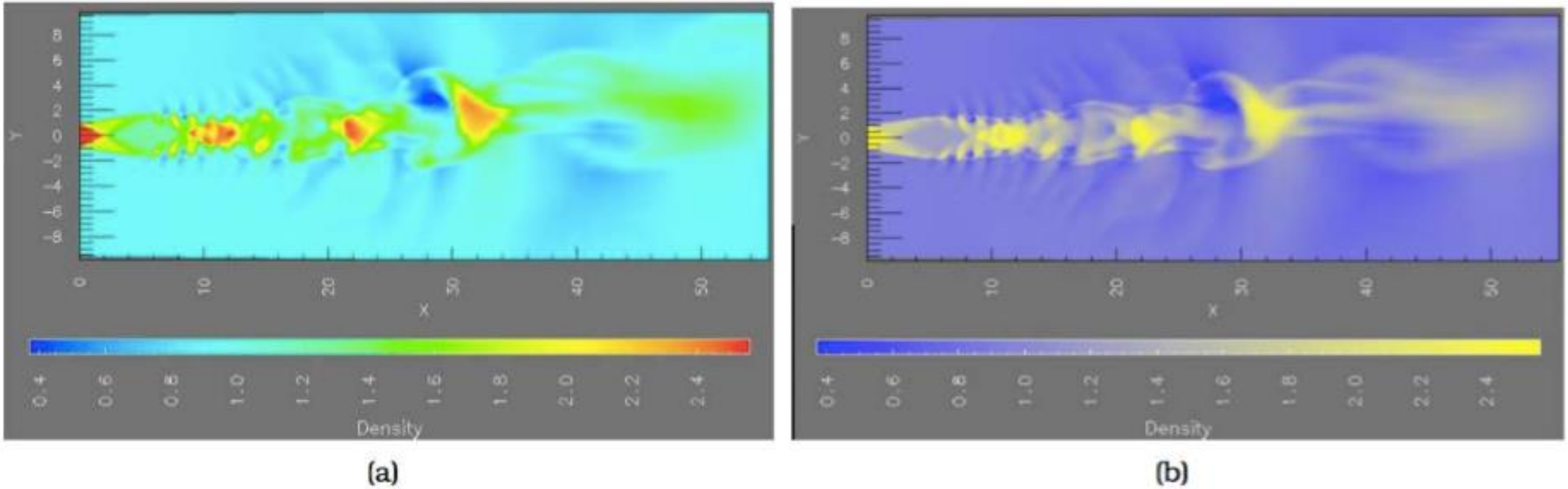


Figure 10.11. Rainbow versus two-hue continuous colormap. (a) Using many hues, as in this rainbow colormap, emphasizes mid-scale structure. (b) Using only two hues, the blue–yellow colormap emphasizes large-scale structure. From [Bergman et al. 95, Figures 1 and 2].

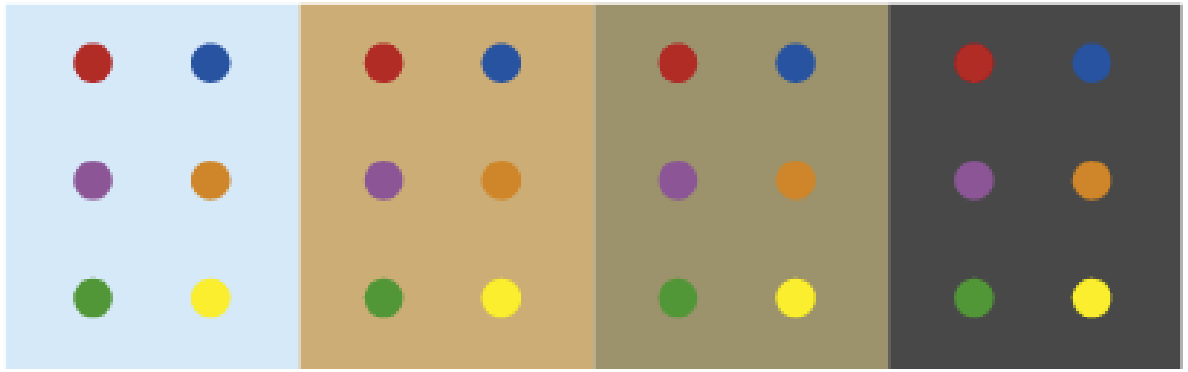
Figure/caption from Munzner, Visualization Analysis and Design

Categorical Data: Color Categories & Naming

- We can only quickly differentiate 5-10 colors. Limit categorical use of color.
  - With work we can do more, but best not to rely on it
  - May assign colors hierarchically
- Pick colors we can differentiate by name, especially in collaborative contexts
  - e.g., dark blue & light blue, dark red & light red, etc.

Ware, Information Visualization: Perception for Design

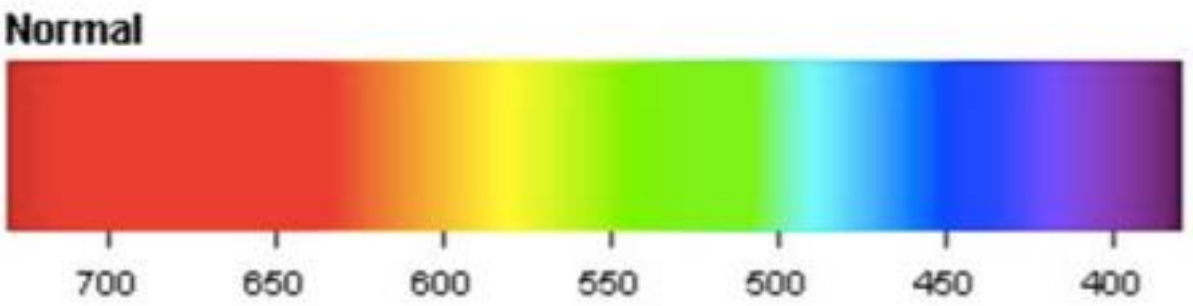
Color Perception is Relative

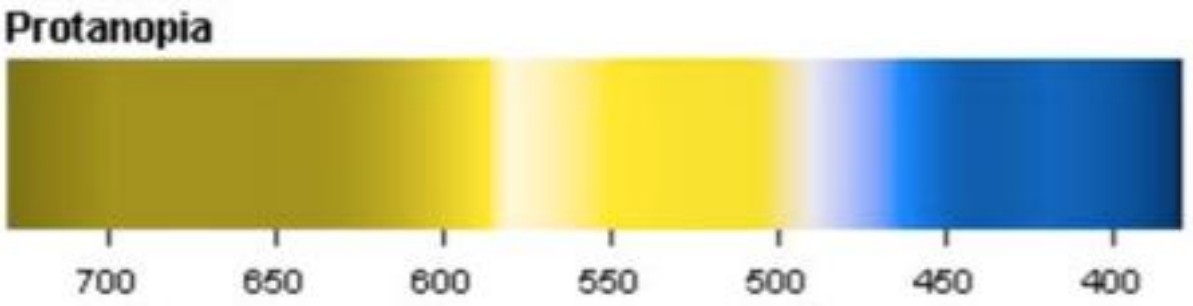


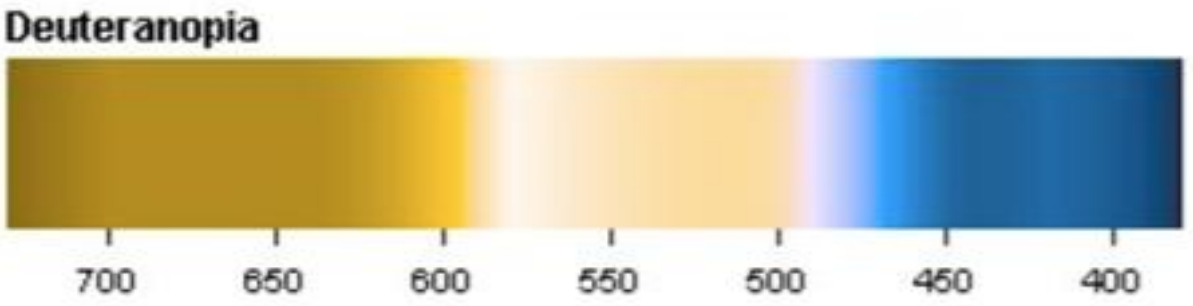
What happens when you take off tinted goggles?

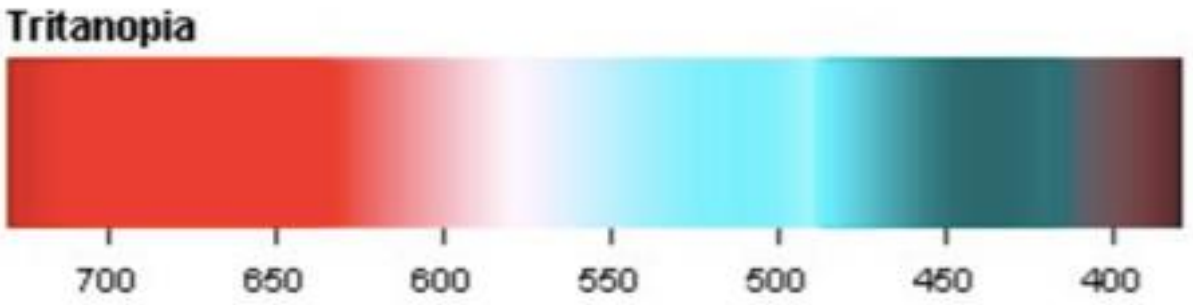
- To avoid contrast issues with background, surround points with white or black lines.
- Make sure colors do not have same luminance as background

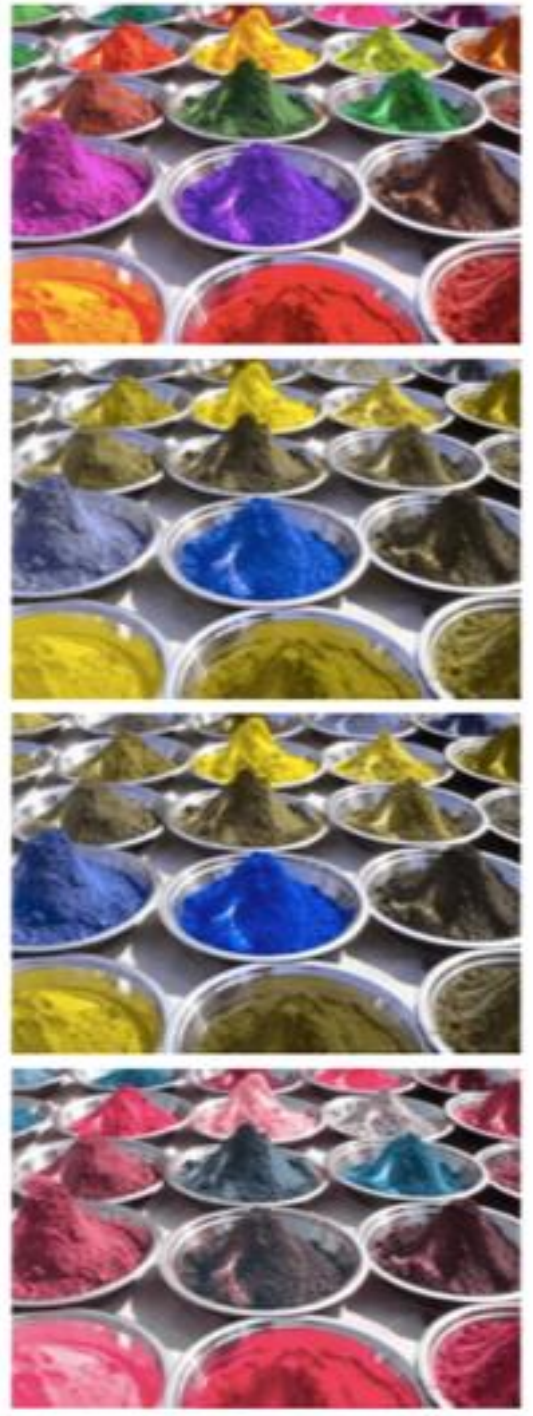
Color Vision Deficiencies affect ~7% of the population





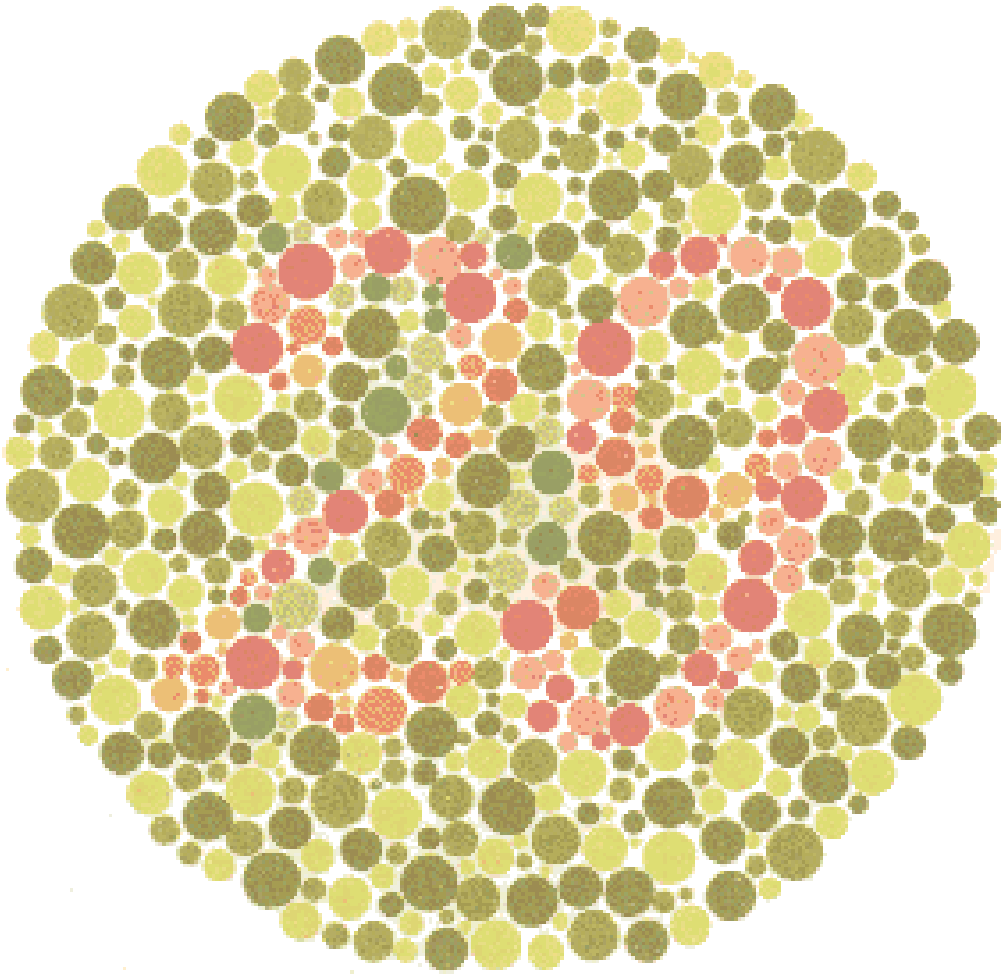


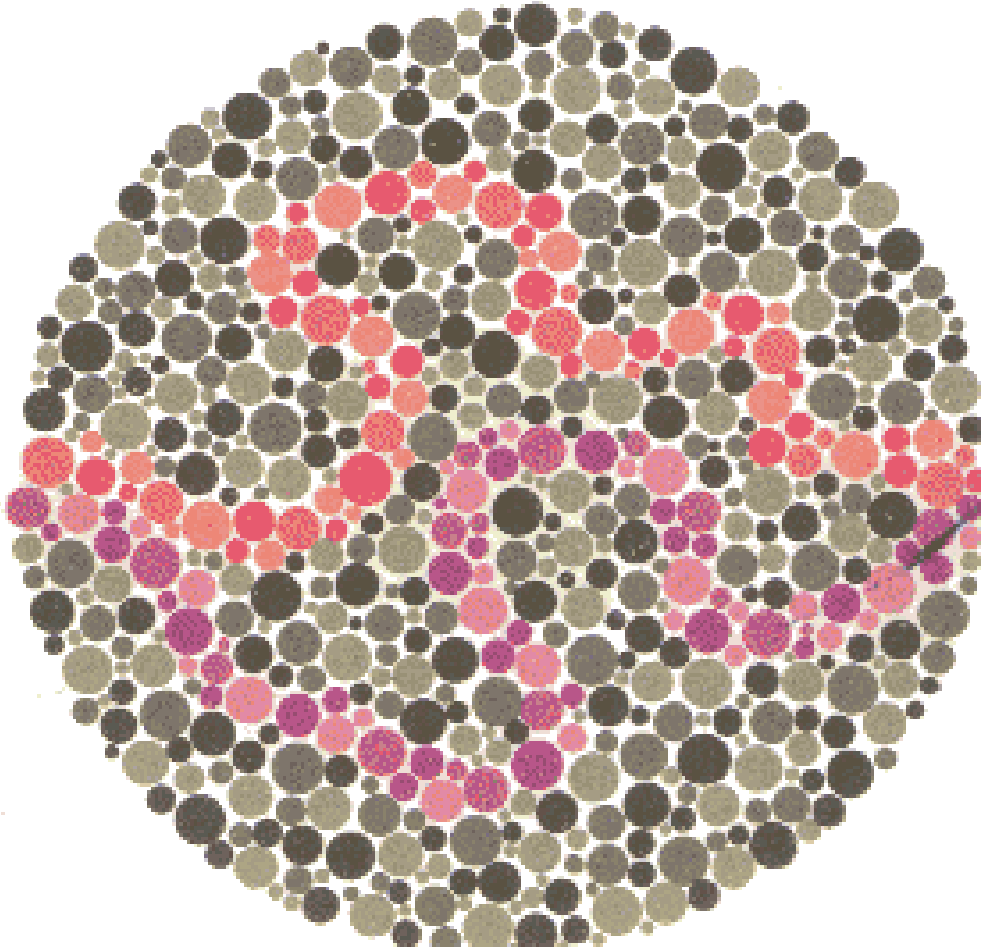




Slide adapted from Carlos Scheidegger

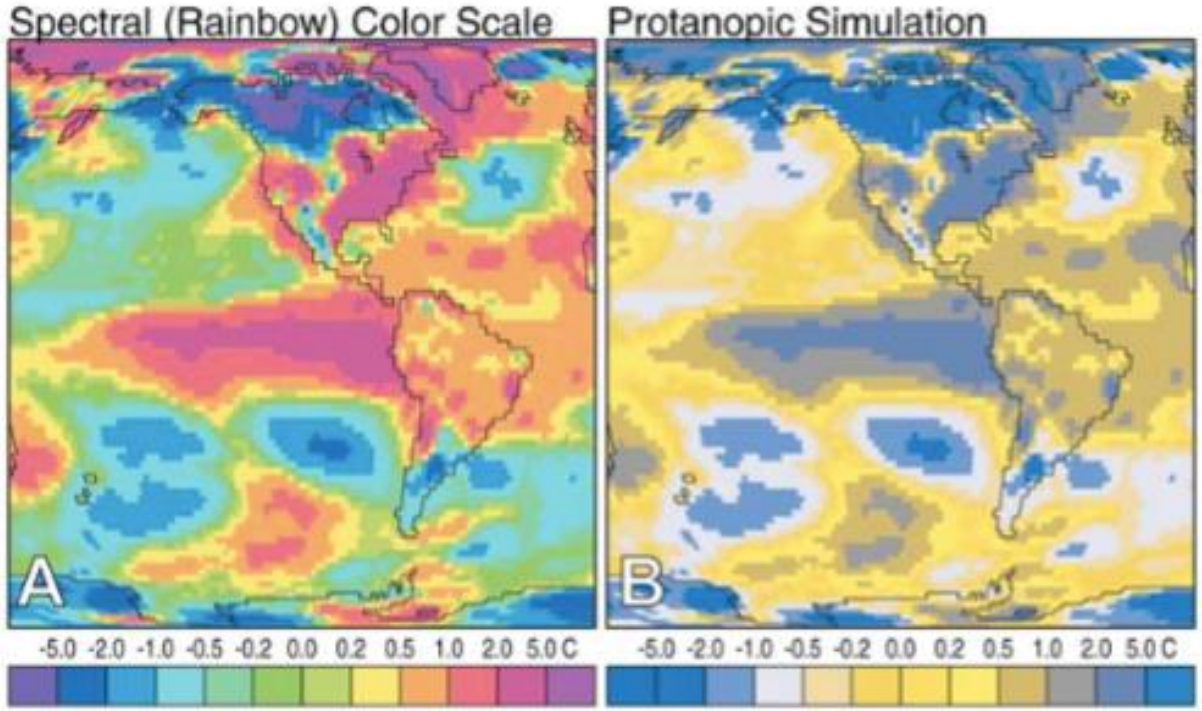
Ishihara Plates are used to test for Color Vision Deficiencies





Slide adapted from Carlos Scheidegger

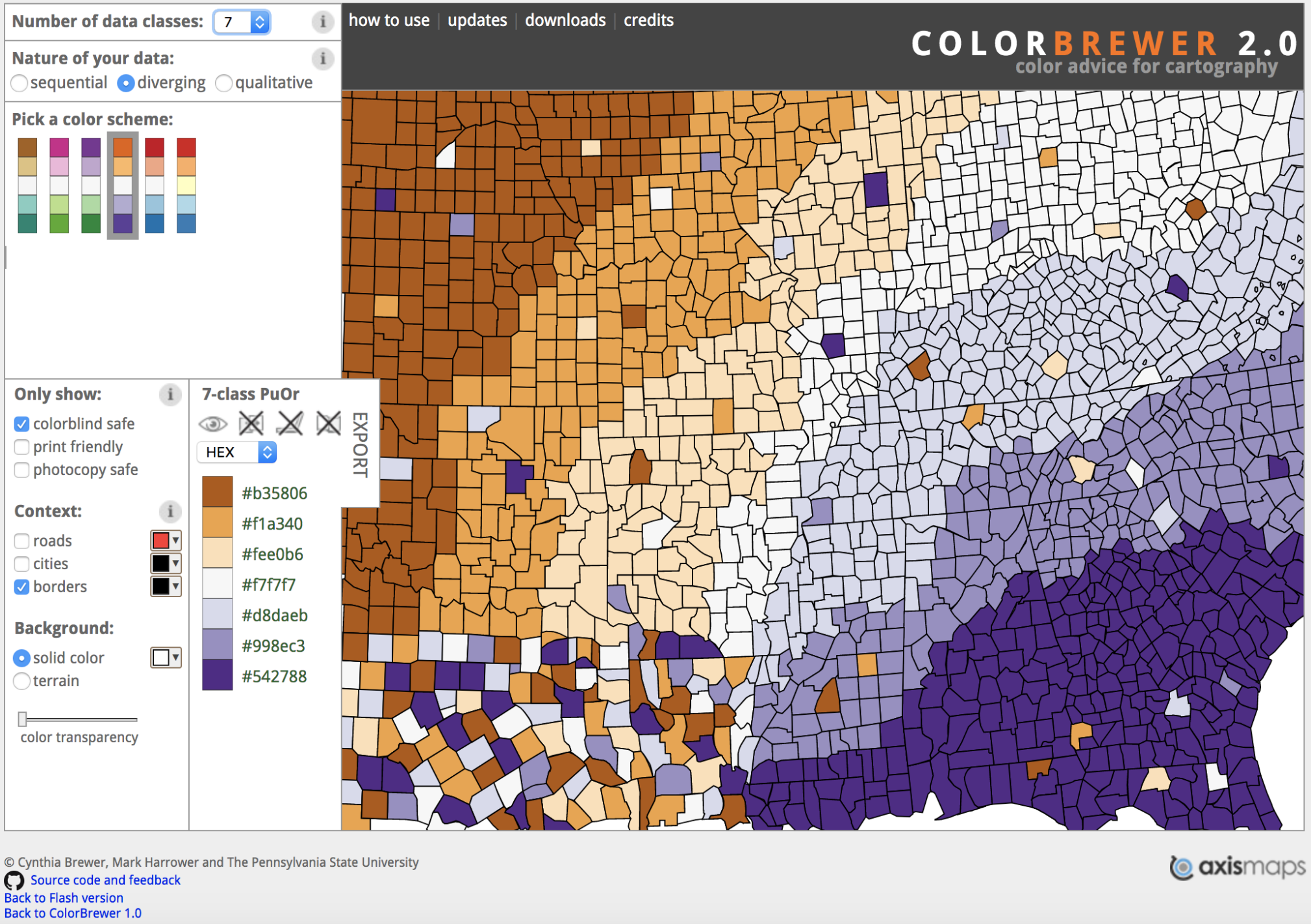
Rainbow Colormap & Color Vision Deficiency



Guideline: "Get it right in black and white"

Adapted from Josh Levine, from Light & Bartlein 2004

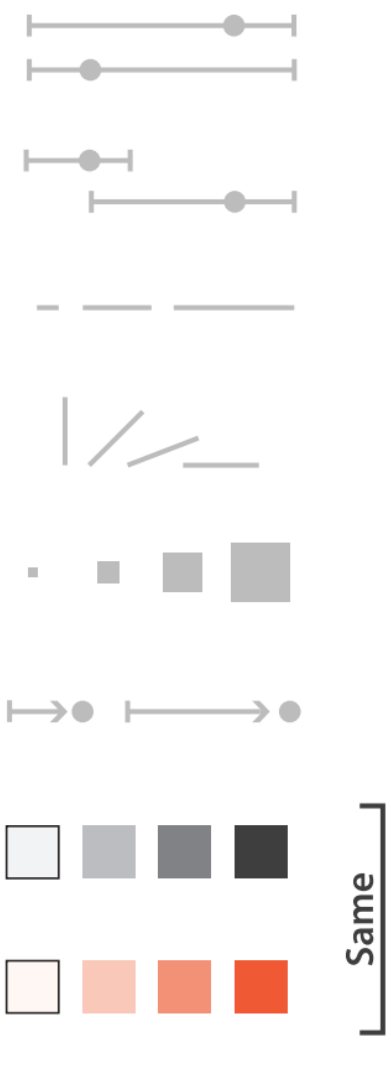
Many online resources for CVD-friendly color maps



Guideline: Use color deliberately & sparingly

Magnitude Channels: Ordered Attributes

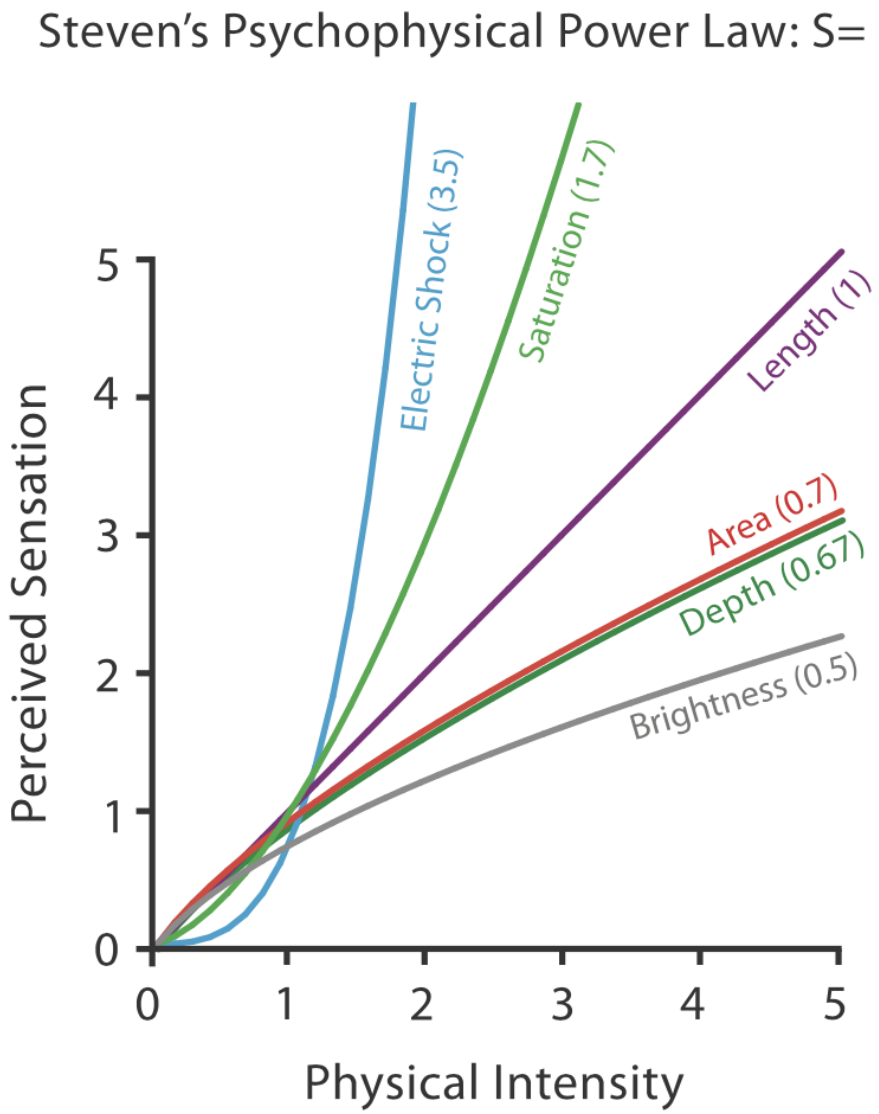
- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation



Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape

Where do these rankings come from?



S = sensation
I = intensity

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

The general form of the law is

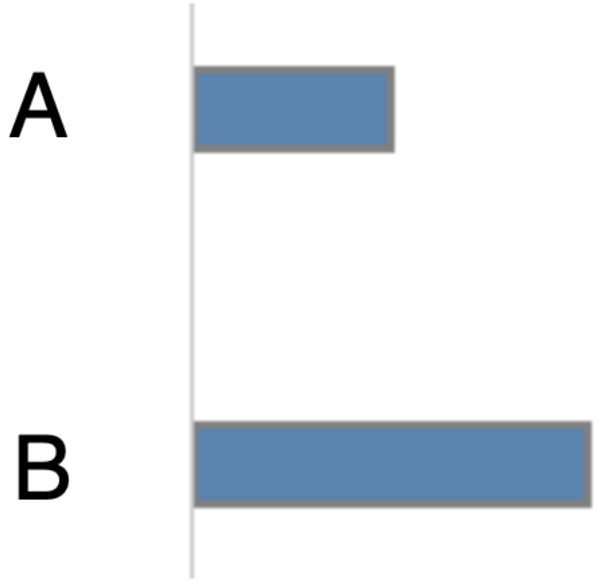
$$\psi(I) = kI^a,$$

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

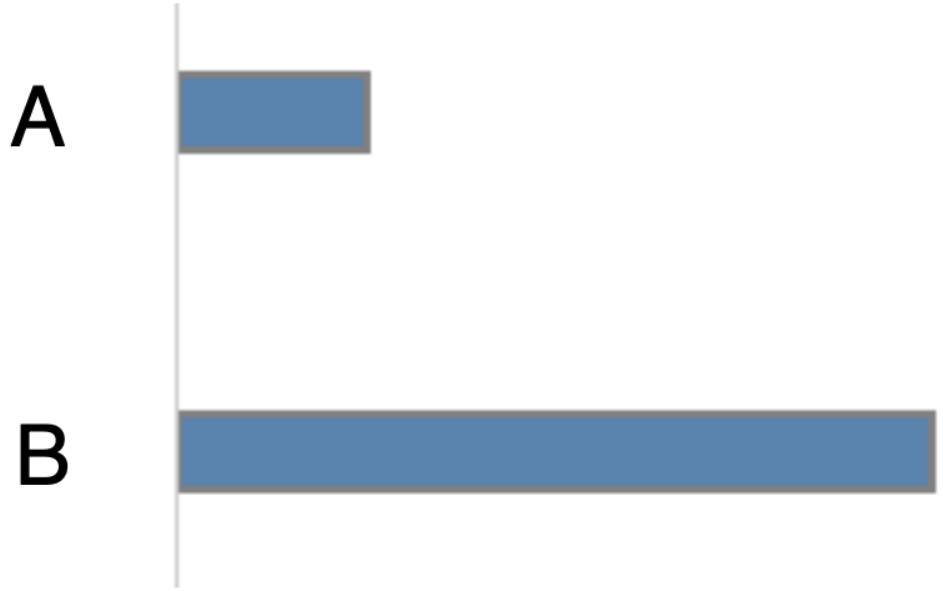
***In part**

How Much Longer?



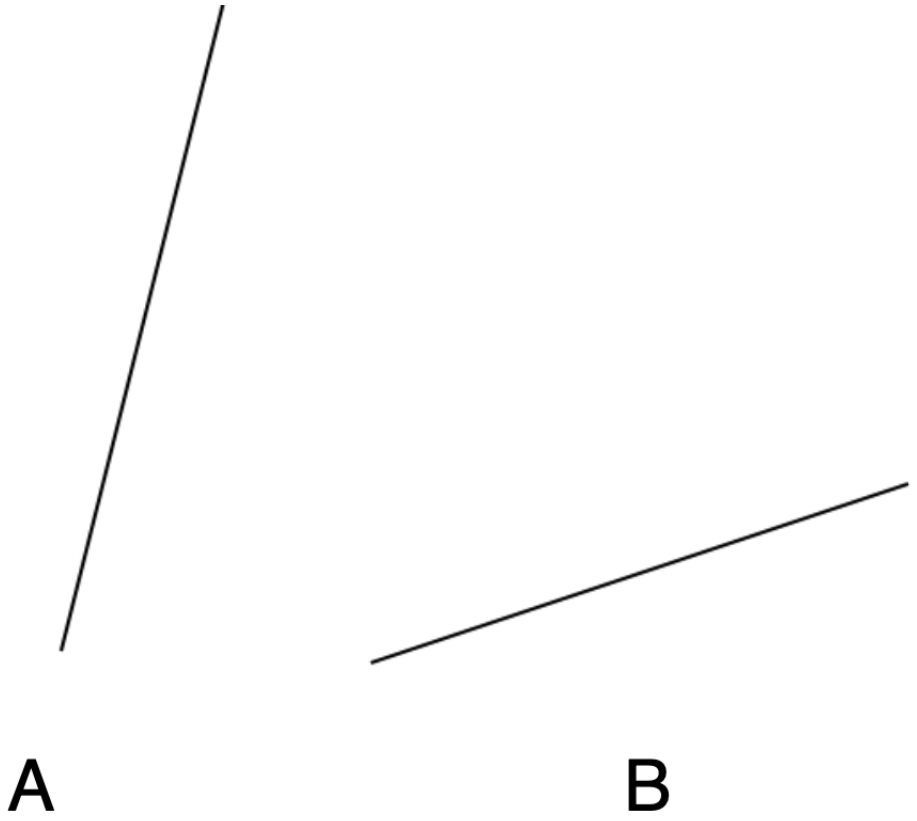
2x

How Much Longer?



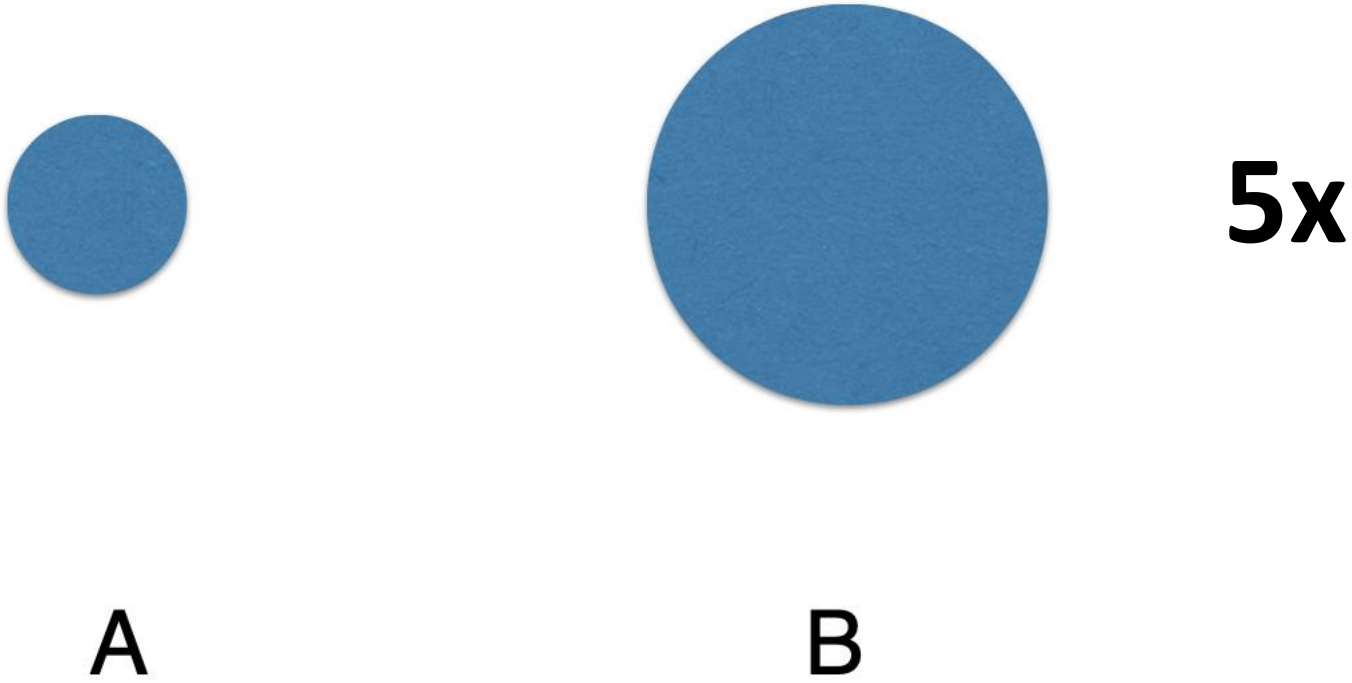
4x

How Much Steeper?



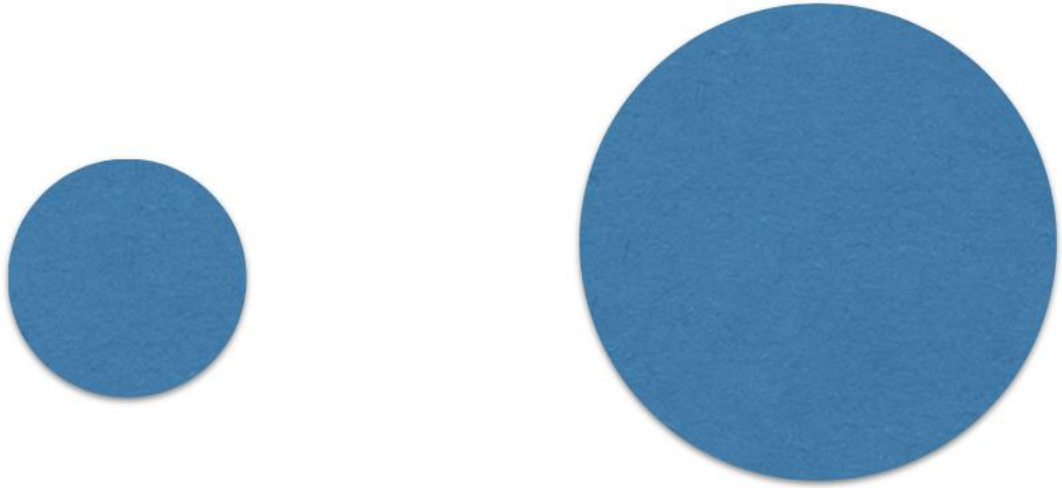
~4x

How Much Larger?



5x

How Much Larger?



A

B

4x area
2x diameter

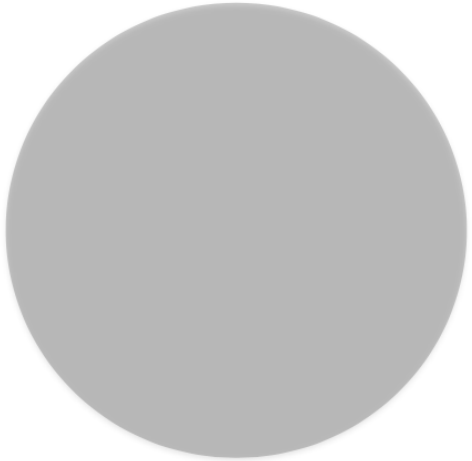
How Much Larger (by area)?

3x

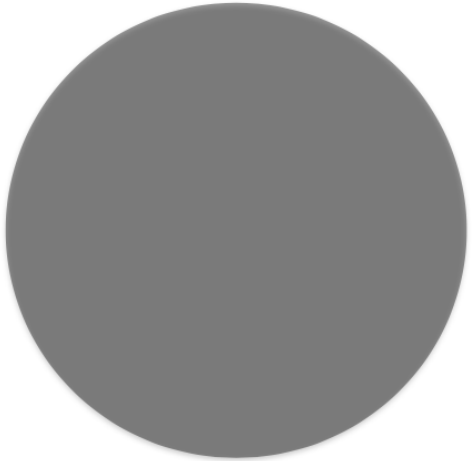
A

B

How Much Darker?



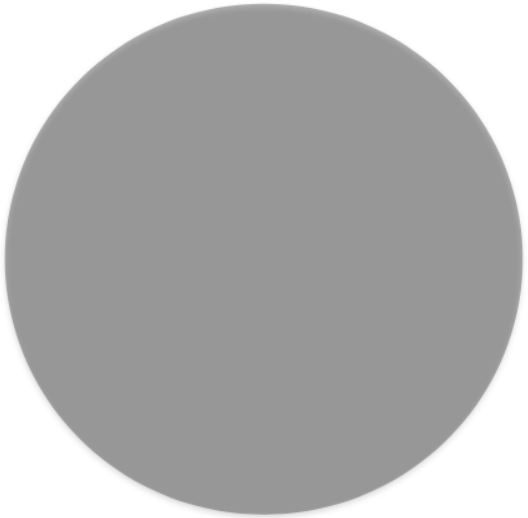
A



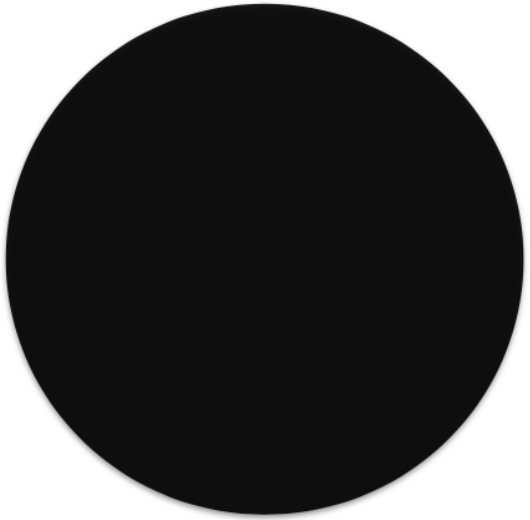
B

2x

How Much Darker?



A



B

3x

Other factors affect accuracy too…

Alignment

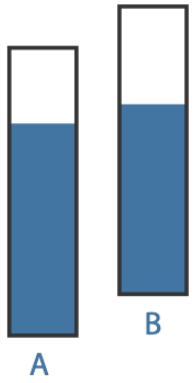
Distractors

Distance

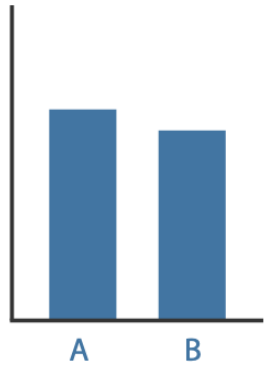
Common scale

…

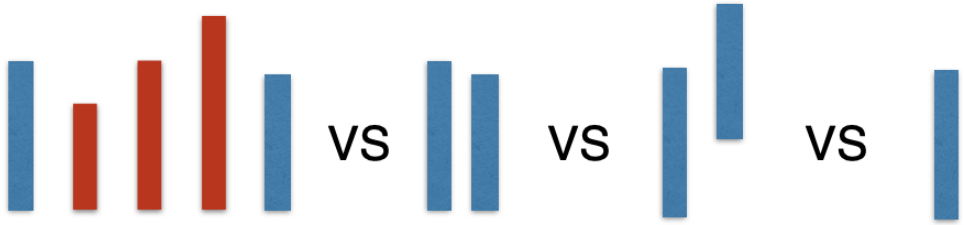
Unframed
Unaligned



Framed
Unaligned



Unframed
Aligned



Some more guidelines…

Expressiveness & Effectiveness

Expressiveness Principle: Encoding should express all of, and only, the information in the data

- Example: Don't imply order where this is not but imply order where there is

Effectiveness Principle: The more important the data/attribute, the more **salient** the encoding should be

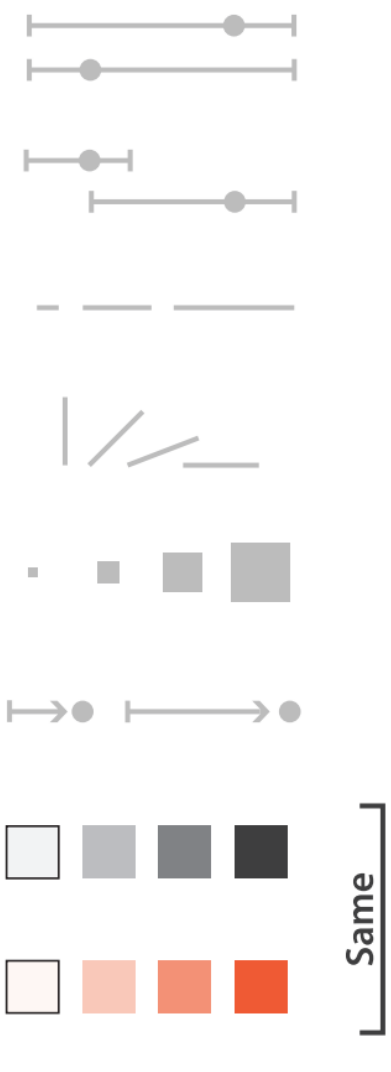
- Important things should be noticeable

From Munzner, Visualization Analysis and Design

Guideline: Use color deliberately & sparingly

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



Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape

Tufte's Integrity Principles

Show **data variation**, not design variation

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

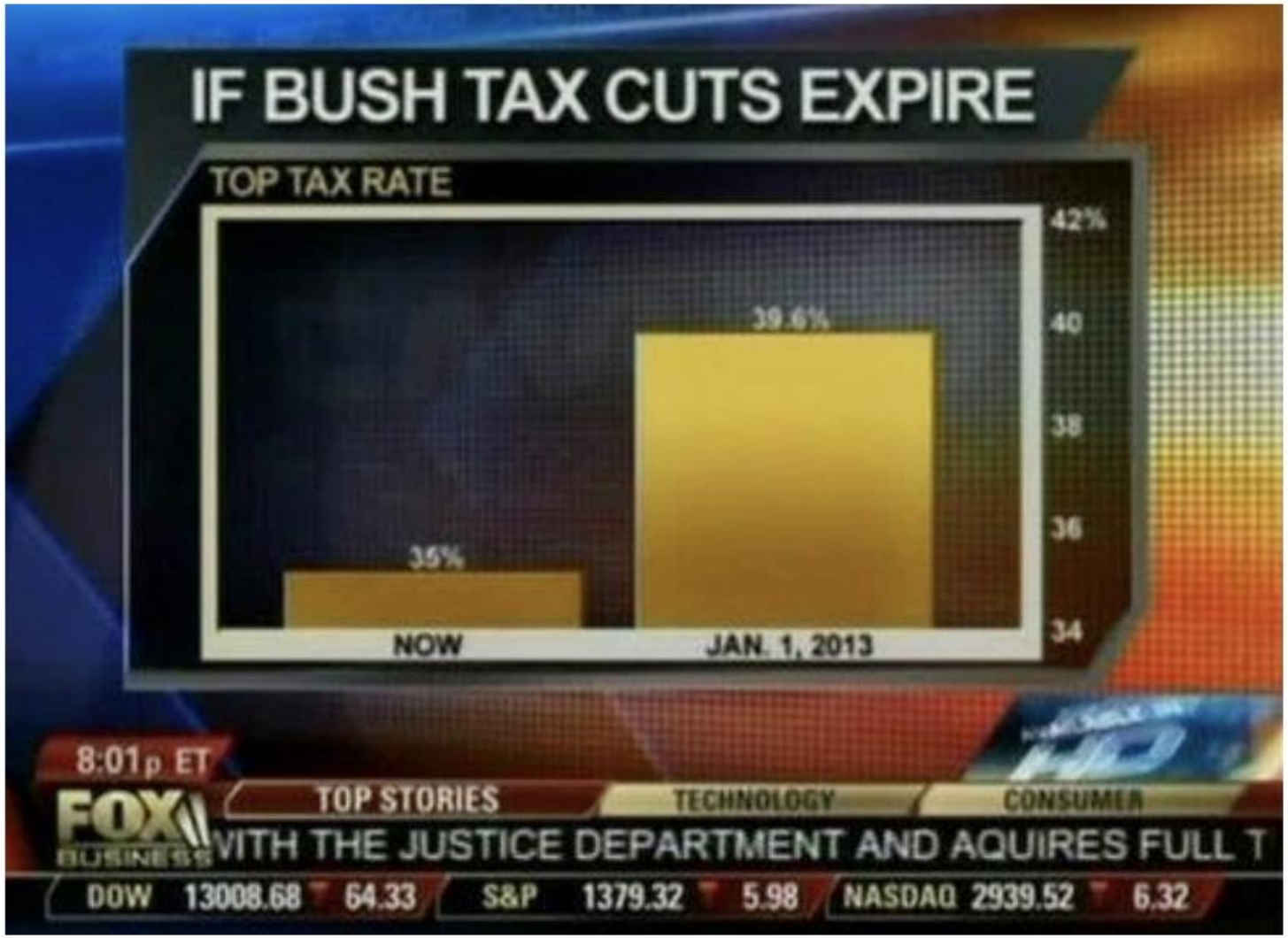
Lie Factor: Size of the **graphic effect** should be **directly proportional to the numerical quantities.**

The Lie Factor

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

The Lie Factor – Graphical Integrity

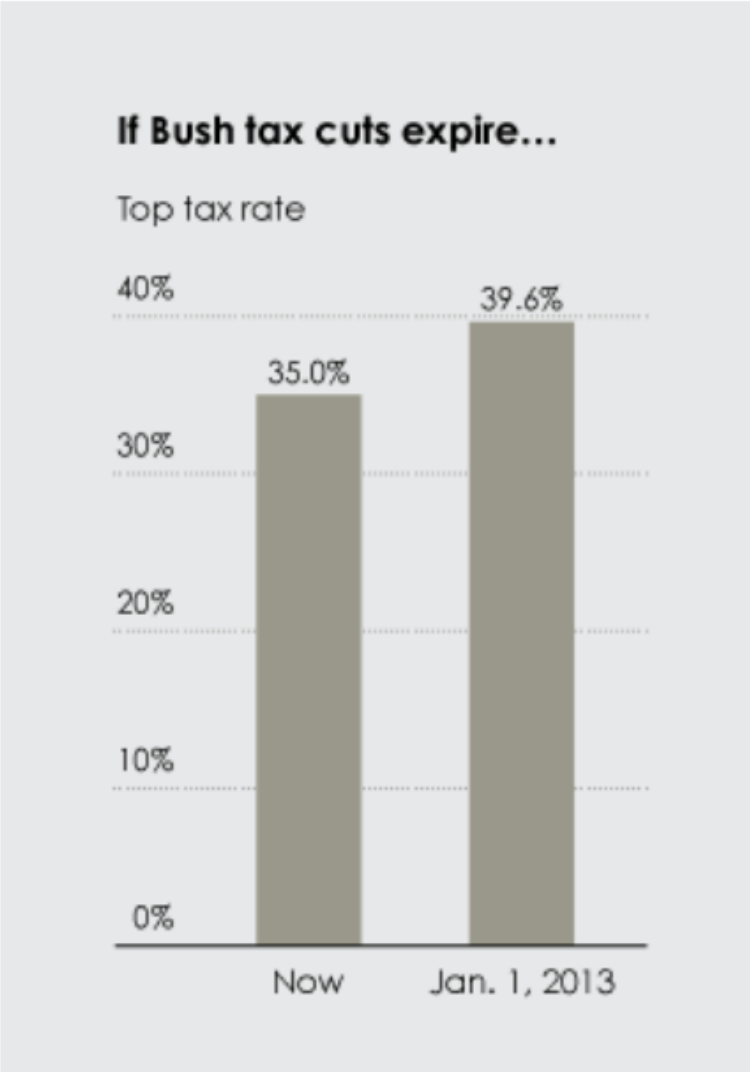
Magnitude in data must correspond to magnitude of mark

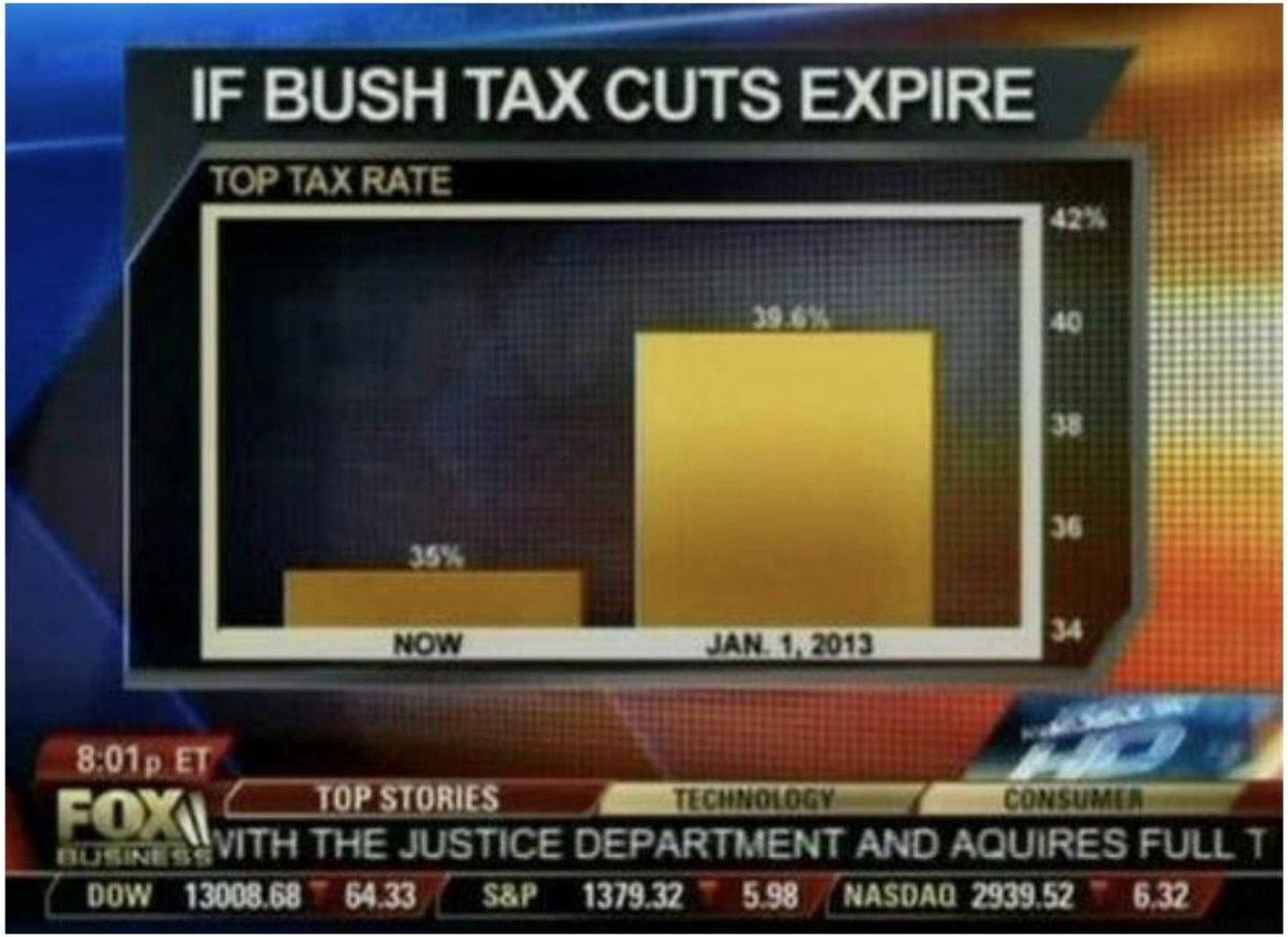


Effect in data: 1.14
Effect in graphic: 5
Lie factor: 5/1.14 = 4.38

From Flowing Data

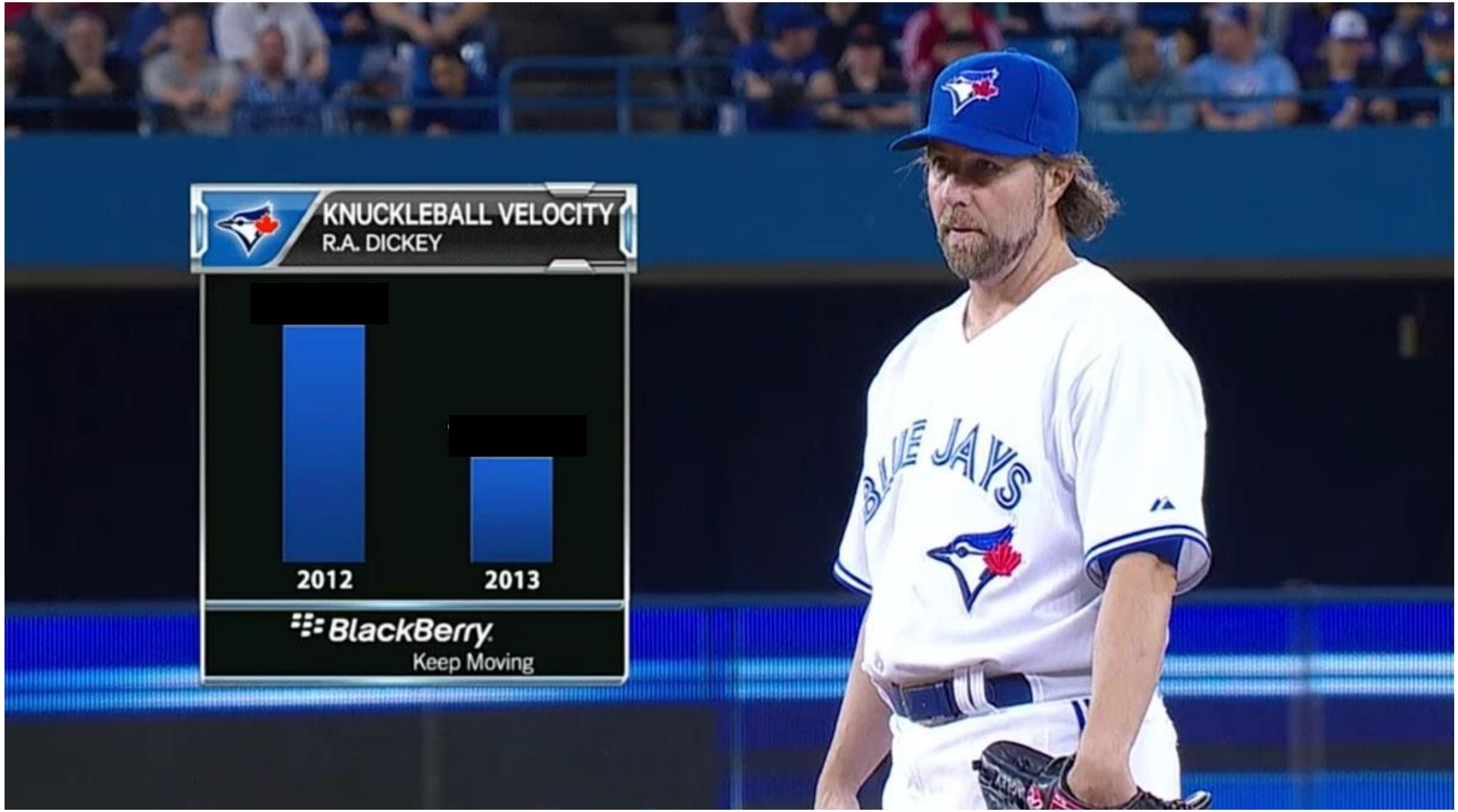
The Lie Factor – Graphical Integrity





From Flowing Data

How much has Dickey's knuckleball slowed?



Images https://www.huffingtonpost.com/raviparikh/lie-with-data-visualization_b_5169715.html

What’s wrong?

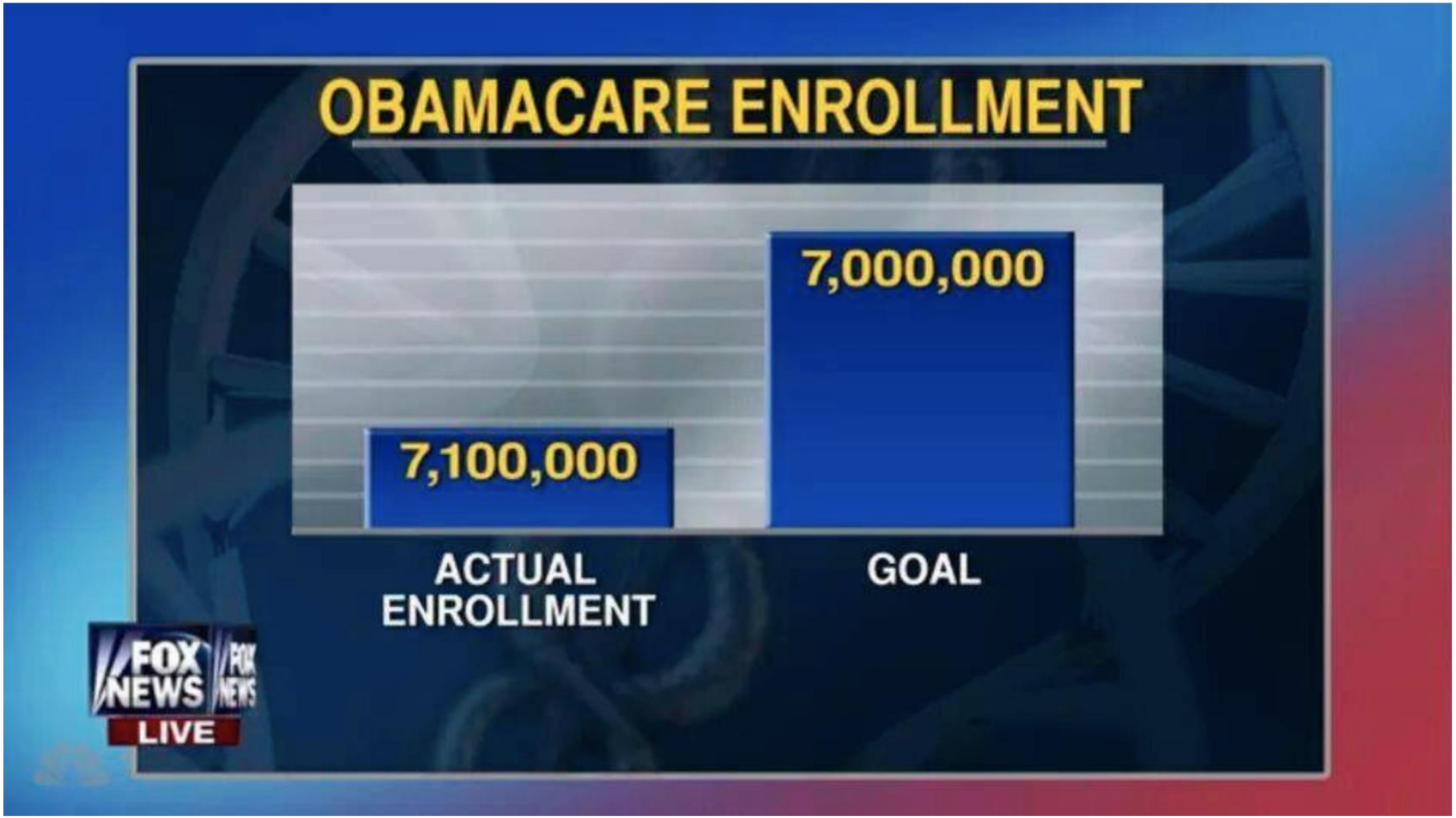
Viele Bezieher mit "ungeklärter Staatsbürgerschaft"

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

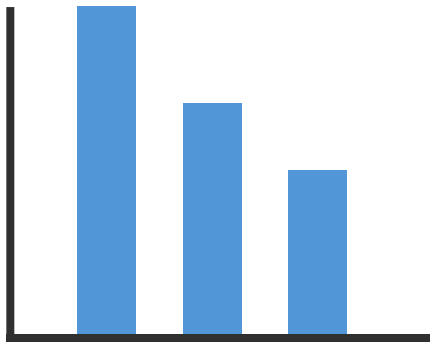
What’s wrong?

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


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

Where should the y-axis start?



Length
(aligned)

In **bar charts**, bar length is being compared. Therefore, starting the y-axis at an arbitrary position will work against the visualization task… often tricking the viewer.

This is referred to as "*truncating the y-axis*." Be careful when you start at something other than zero.

Images from Munzner, Visual Analysis and Design

What's happening here?

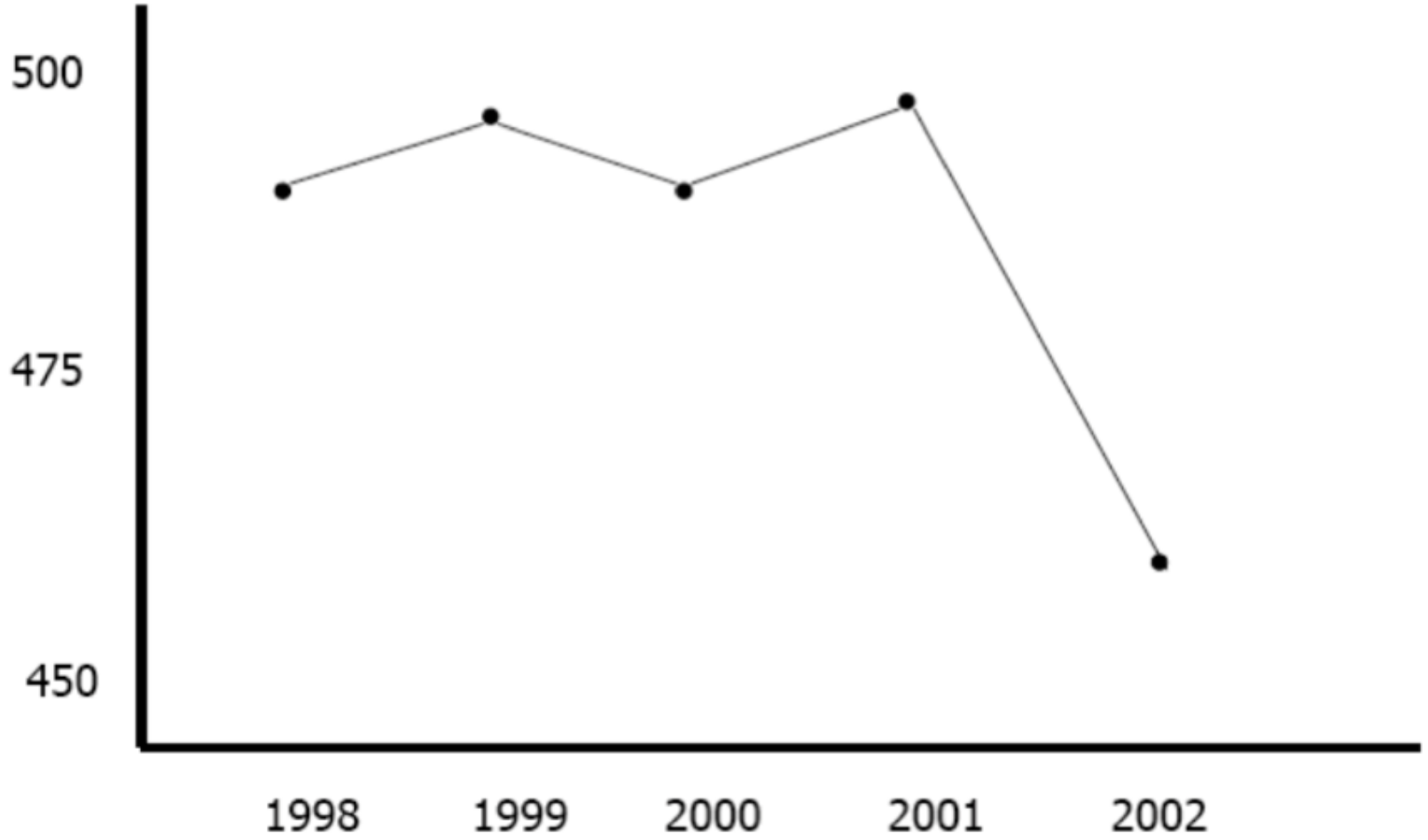


Image from slides of Josh Levine, adapted from John Stasko

What's happening here?

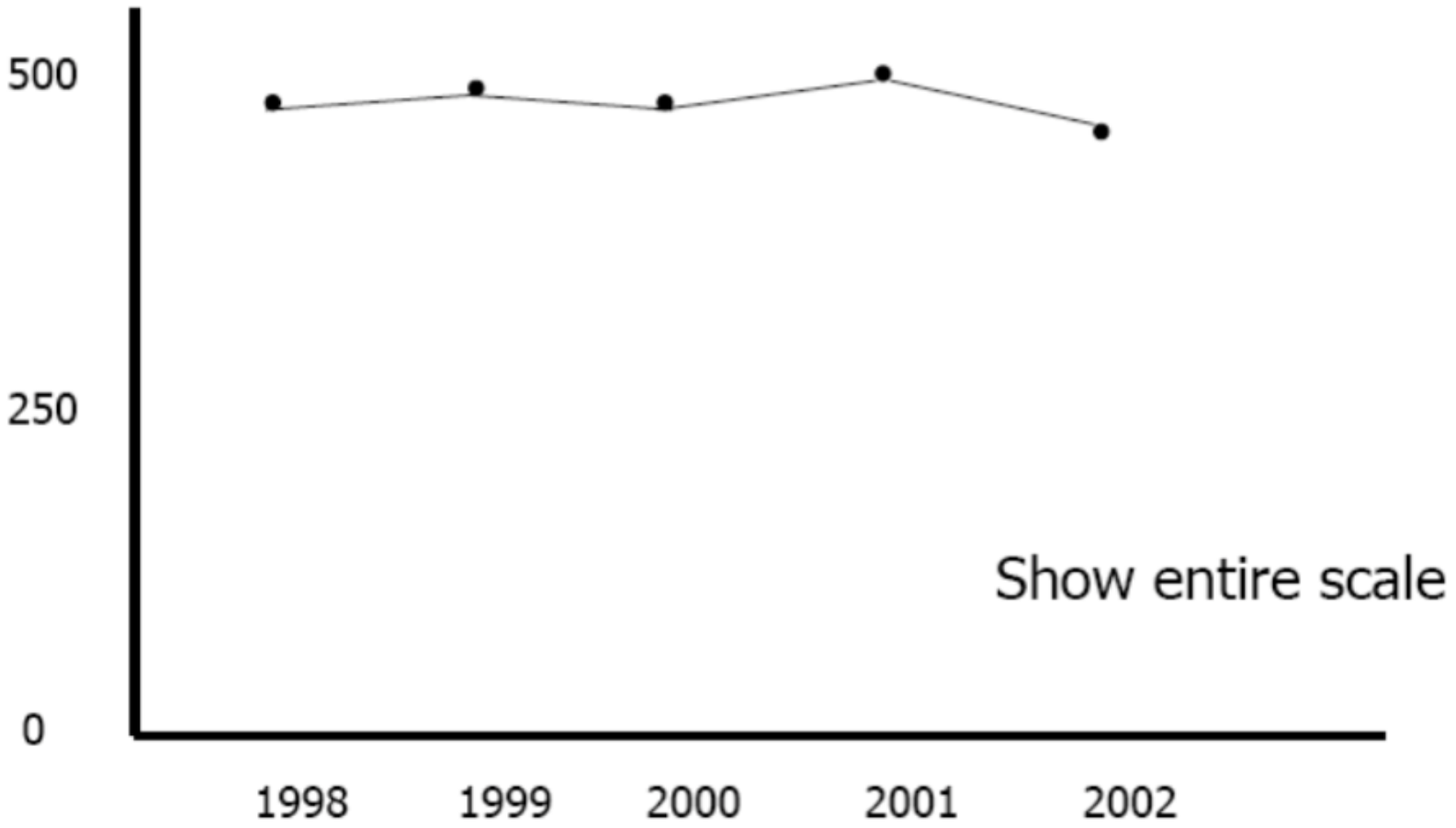


Image from slides of Josh Levine, adapted from John Stasko

What's happening here?

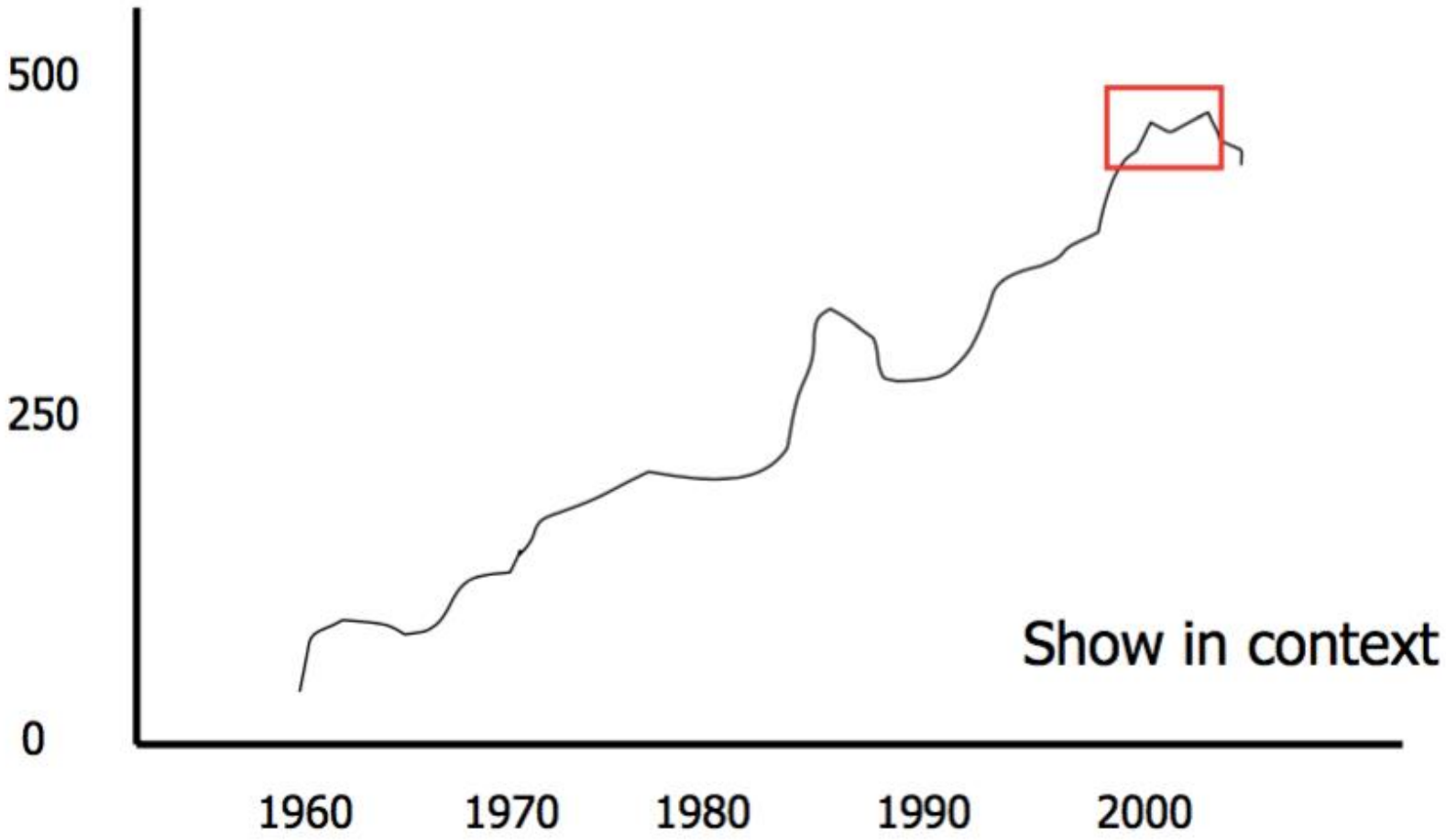
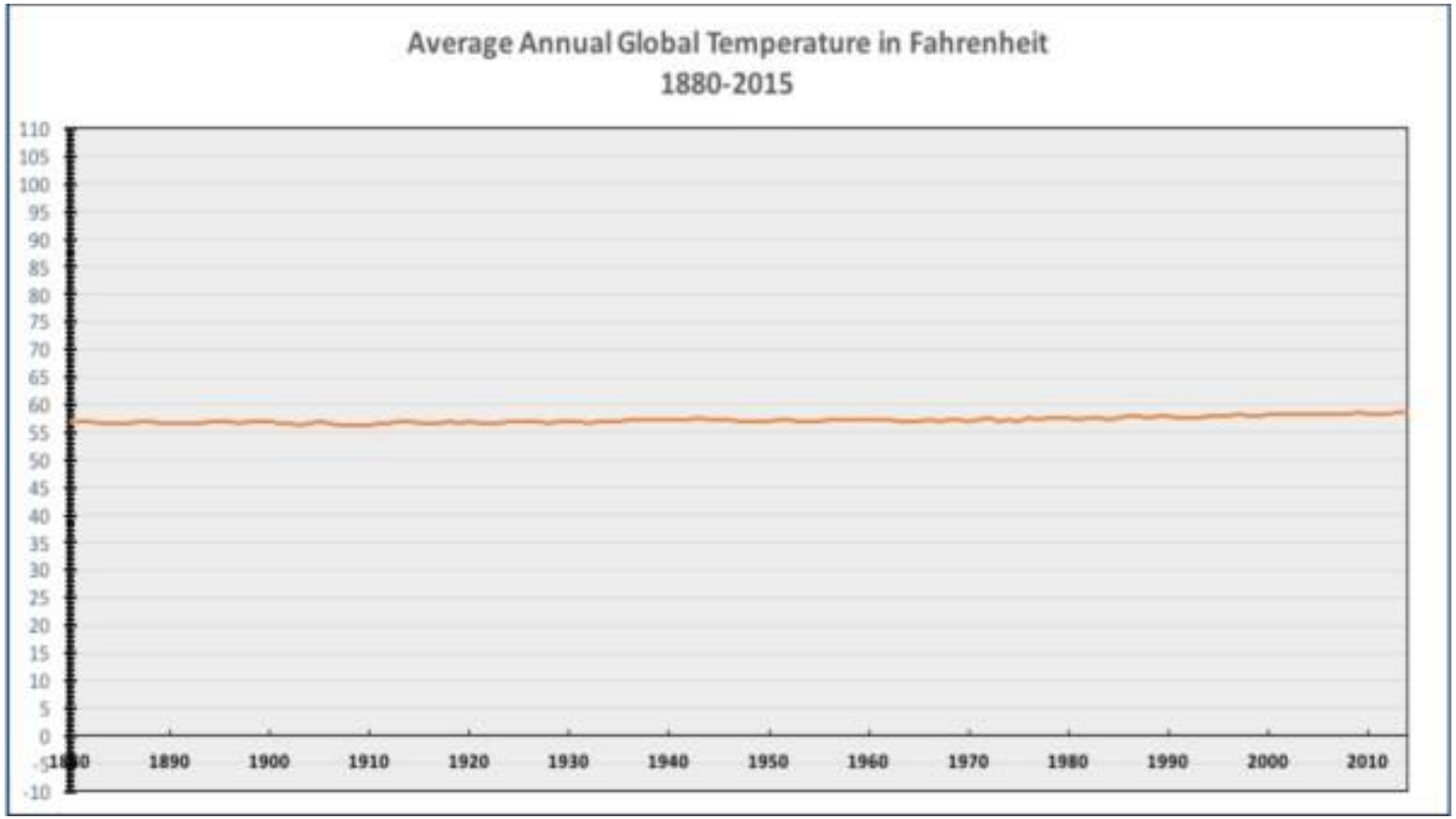


Image from slides of Josh Levine, adapted from John Stasko

Where should the y-axis start?



There are several possible "zero" points. Which one is the most natural?

Line graphs are generally used to analyze *change* in a range rather than absolute. The analysis task matters!

Image from the National Review

Meta-Psychology, 2019, vol 3, MP.2018.895,
https://doi.org/10.5626/MP.2018.895
Article type: Original Article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Felix Schönbrodt
Reviewed by: Julia Rohrer, Gordon Feld, Steve Haroz
Analysis reproduced by: Tobias Mühlmeister
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/HXK2U

Graph Construction: An Empirical Investigation on Setting the Range of the Y-Axis

Jessica K. Witt
Colorado State University

Graphs are an effective and compelling way to present scientific results. With few rigid guidelines, researchers have many degrees-of-freedom regarding graph construction. One such choice is the range of the y-axis. A range set just beyond the data will bias readers to see all effects as big. Conversely, a range set to the full range of options will bias readers to see all effects as small. Researchers should maximize congruence between visual size of an effect and the actual size of the effect. In the experiments presented here, participants viewed graphs with the y-axis set to the minimum range required for all the data to be visible, the full range from 0 to 100, and a range of approximately 1.5 standard deviations. The results showed that participants' sensitivity to the effect depicted in the graph was better when the y-axis range was between one to two standard deviations than with either the minimum range or the full range. In addition, bias was also smaller with the standardized axis range than the minimum or full axis ranges. To achieve congruency in scientific fields for which effects are standardized, the y-axis range should be no less than 1 standard deviations, and aim to be at least 1.5 standard deviations.

Keywords: Graph Design, Effect size, Sensitivity, Bias

One way to lie with statistics is to set the range of the y-axis to form a misleading impression of the data. A range set too narrow will exaggerate a small effect and can even make a non-significant trend appear to be a substantial effect (Pandey, Rall, Satterthwaite, Nov, & Bertini, 2015). Yet the default setting of many statistical and graphing software package range set too wide also creates a misleading impression of the data by making effects seem smaller than they are. Here, I argue that for scientific fields that use standardized effect sizes and adopt Cohen's convention that an effect of d = 0.8 is big, the range of the y-axis should be approximately 1.5 standard deviations (SDs).

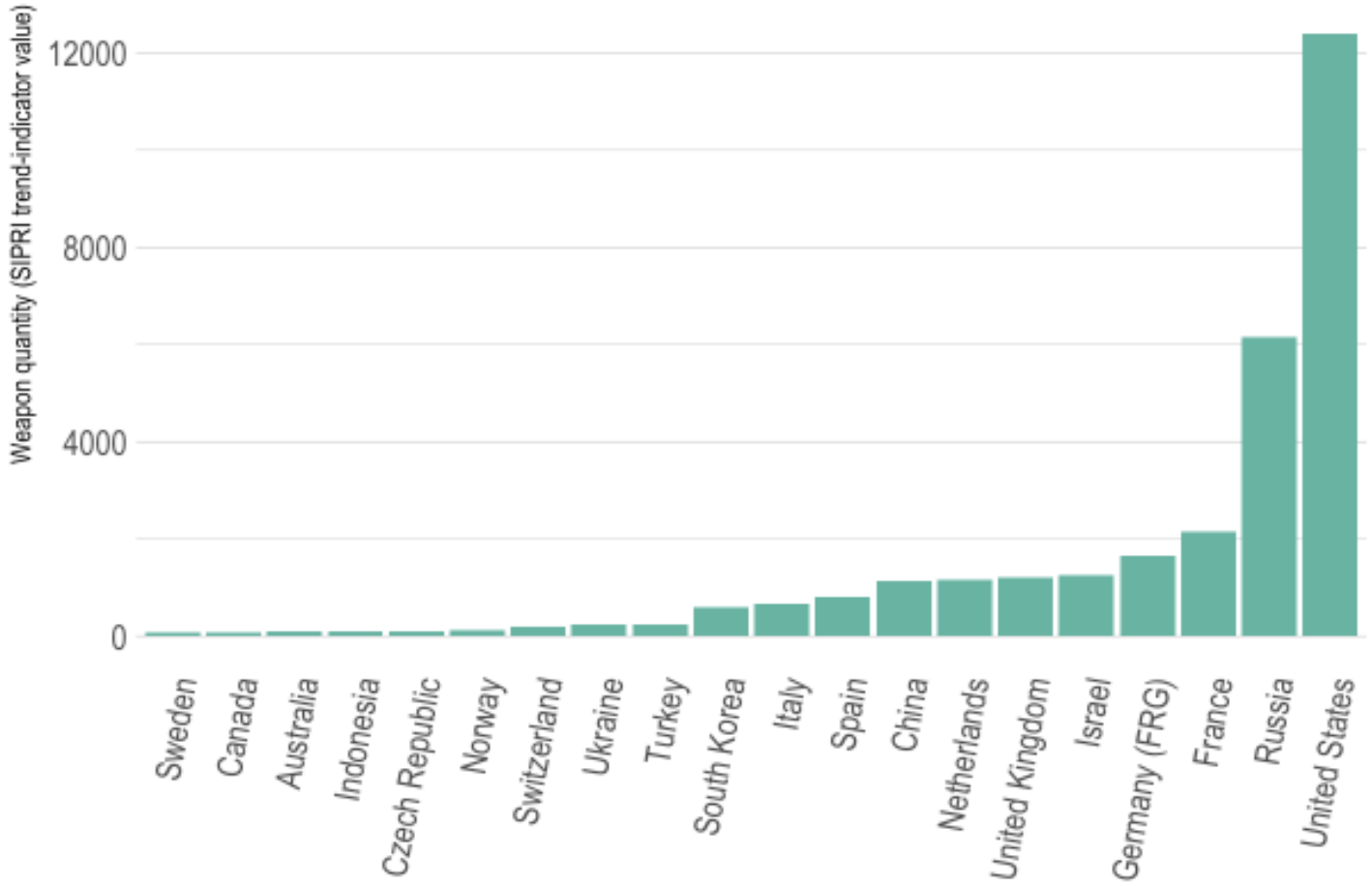
What range should I use?

Data range?
Start from zero?

Like all things, depends on the task.

Guideline: Rotate for Readability

Be Aware of Text Angle

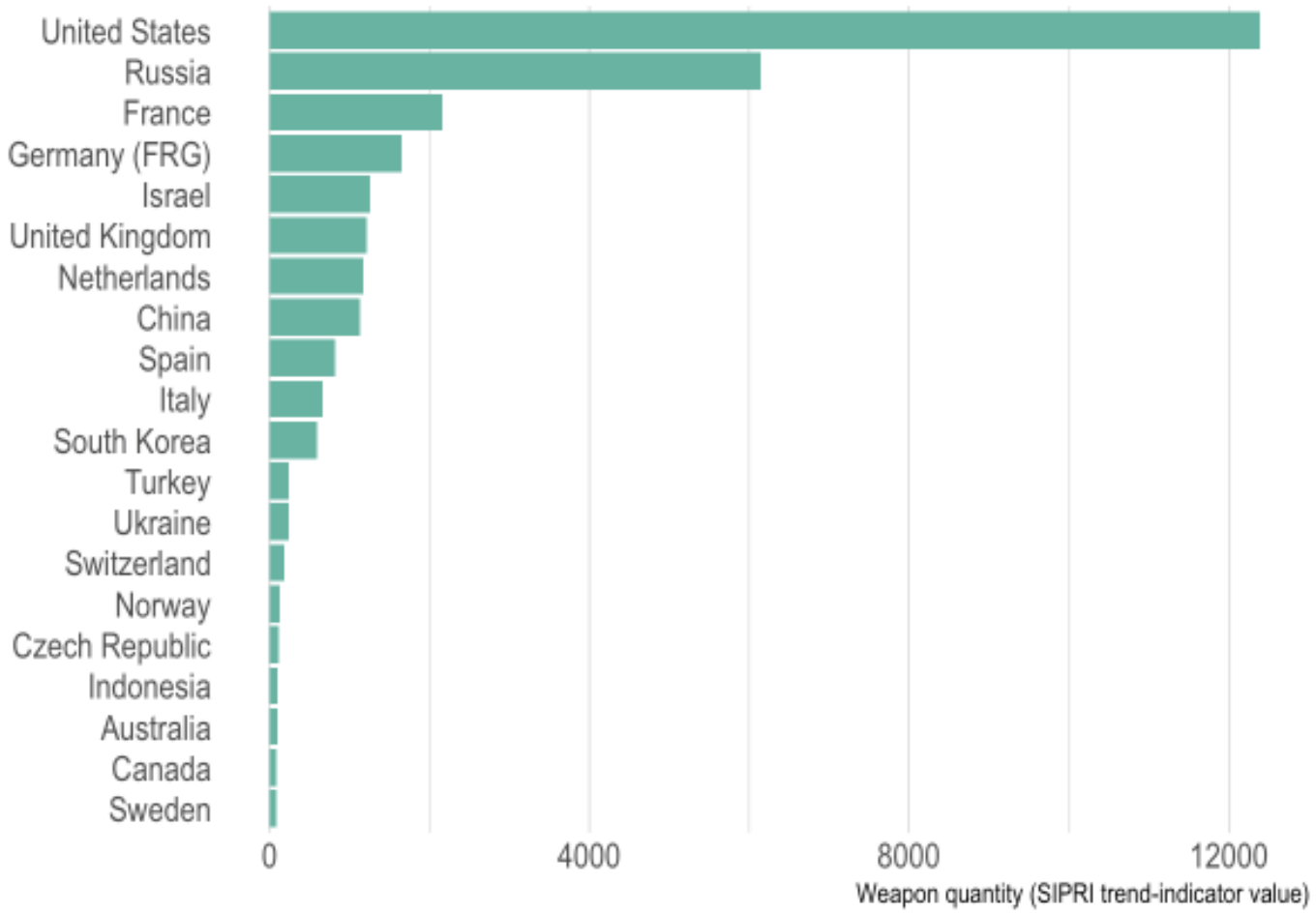


Users shouldn't feel the need to tilt their head to read labels.

Furthermore, while we can rotate fonts at any angle, they distort and become jagged.

Image from https://www.data-to-viz.com/caveat/hard_label.html

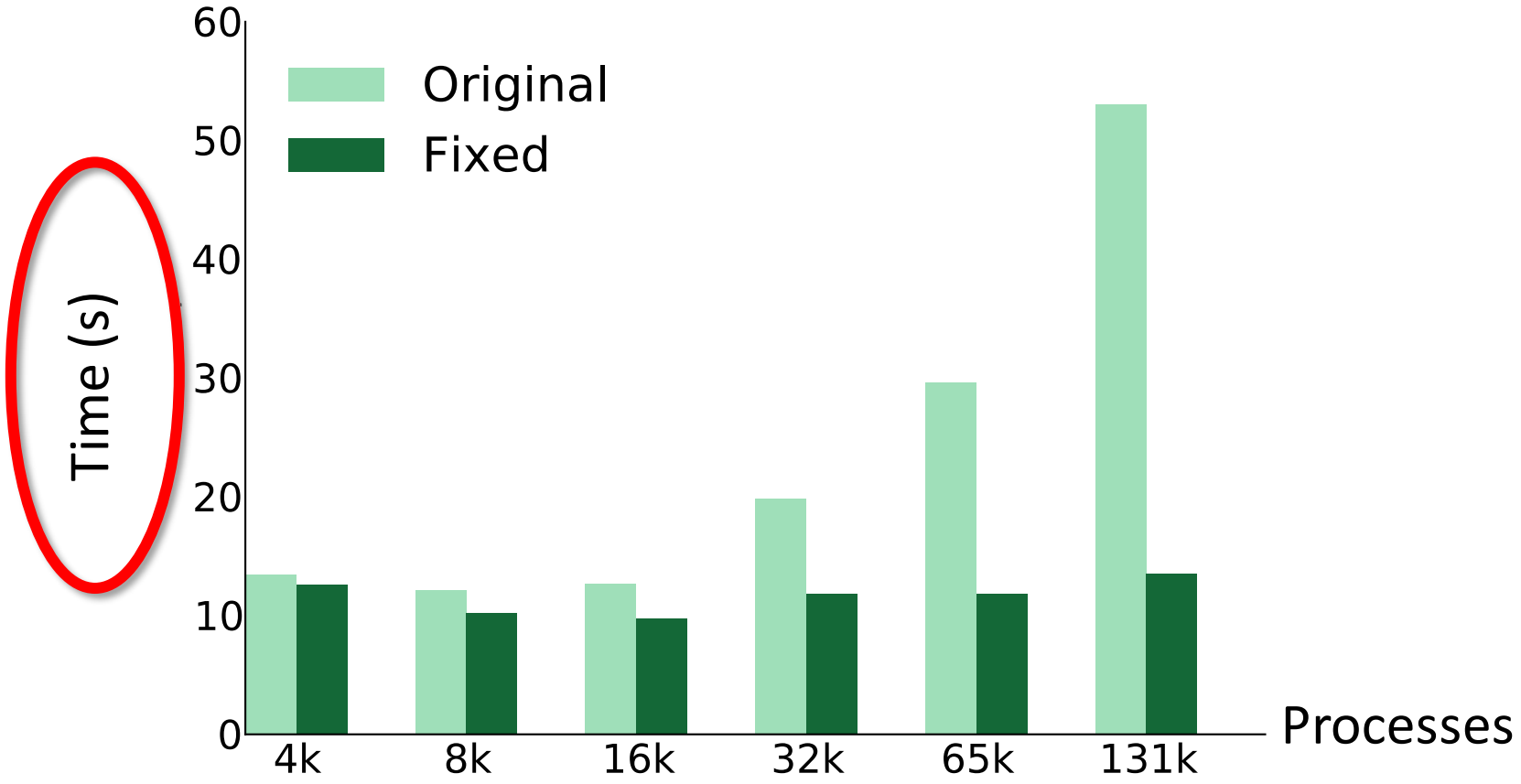
Consider rotating a bar chart



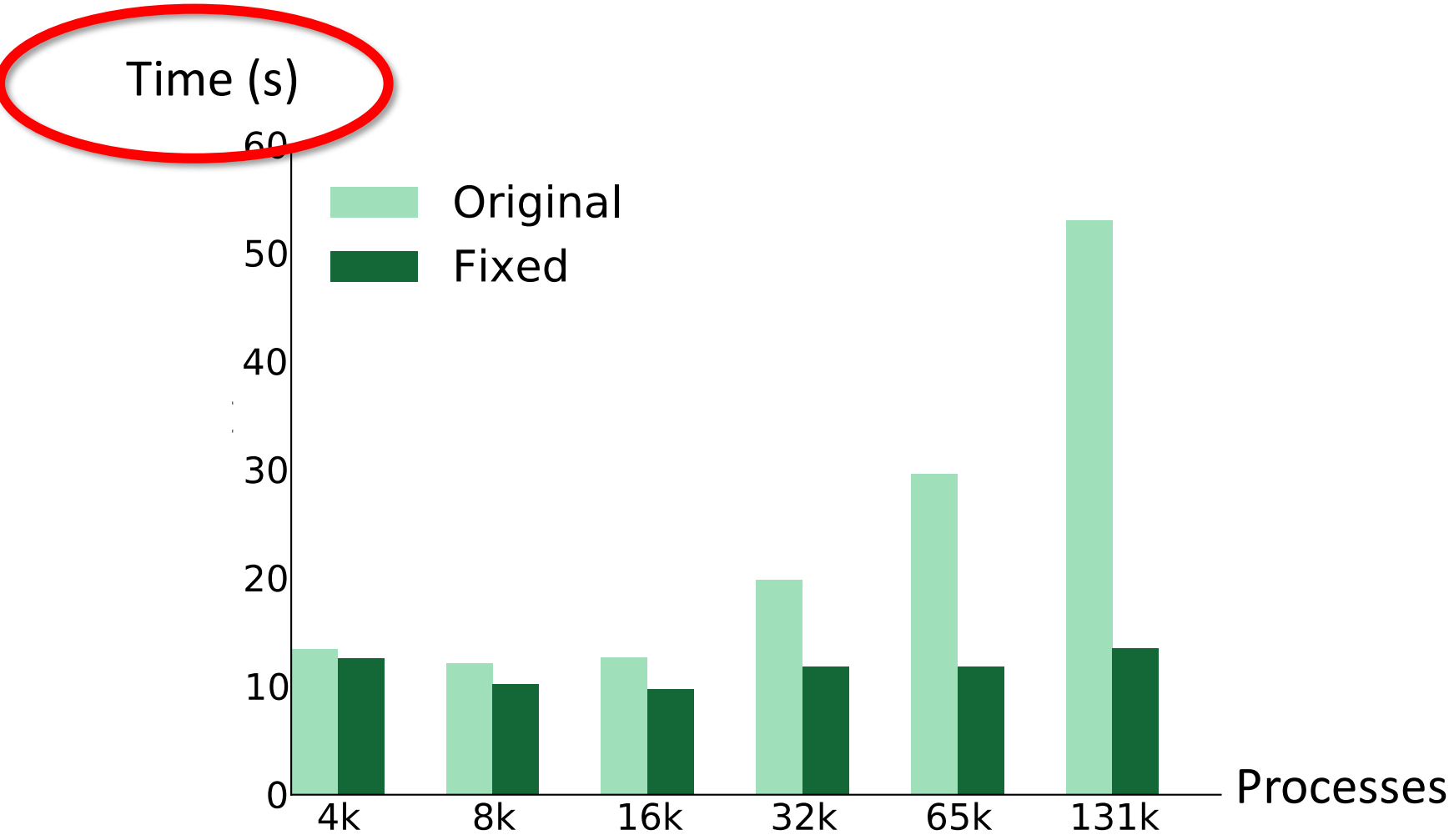
*Note however that some axes have a strong association with the x-axis (e.g., time). In that case, the design trade off may leave tilted text.

Image from https://www.data-to-viz.com/caveat/hard_label.html

Labels on the y-axis need not be vertical



Labels on the y-axis need not be vertical



When the rotation bucks convention, it may be misinterpreted

What is your initial reaction?

The designer's desire was to evoke blood running down a wall.

Takeaway: any design counter to well known conventions must be strongly justified.

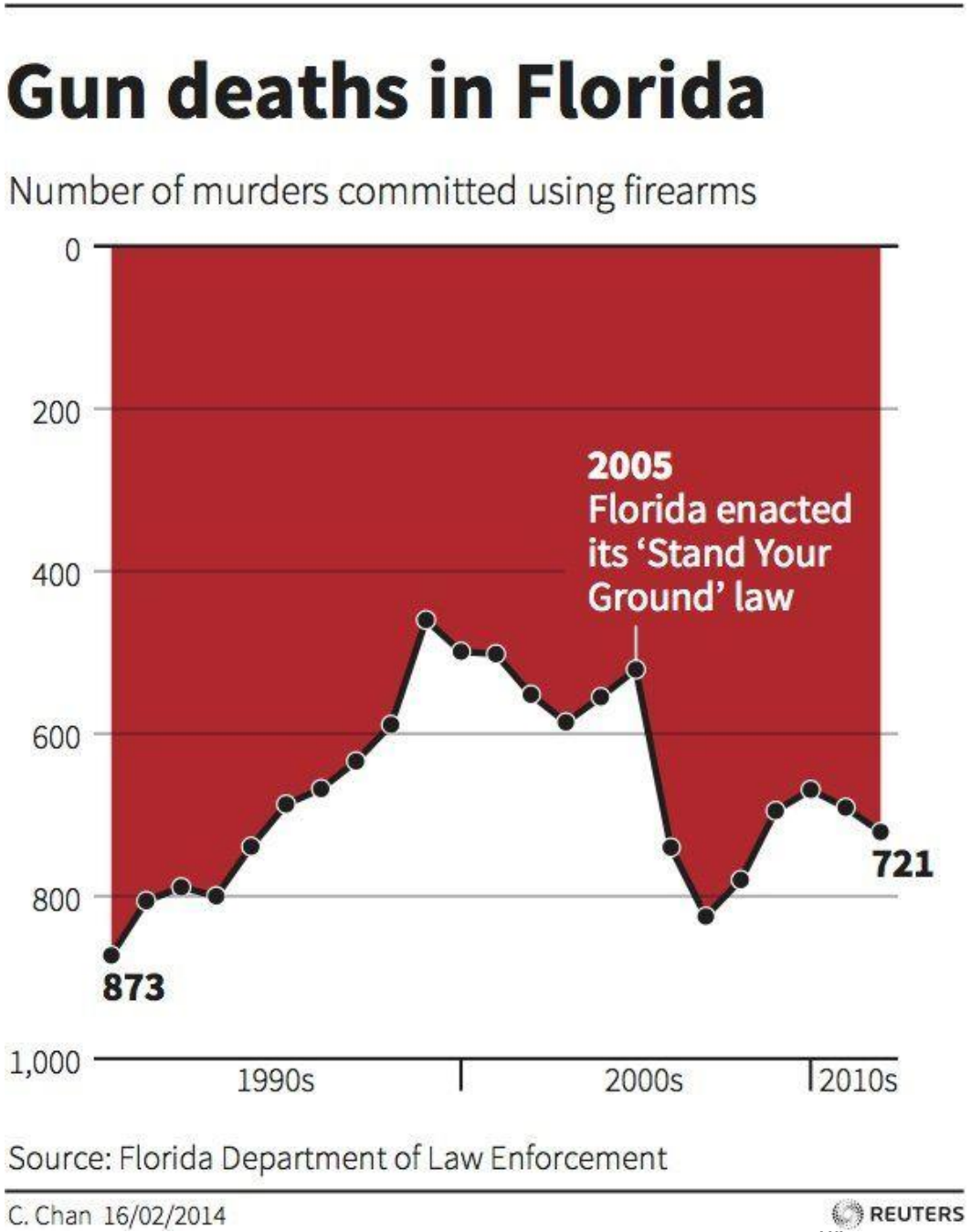
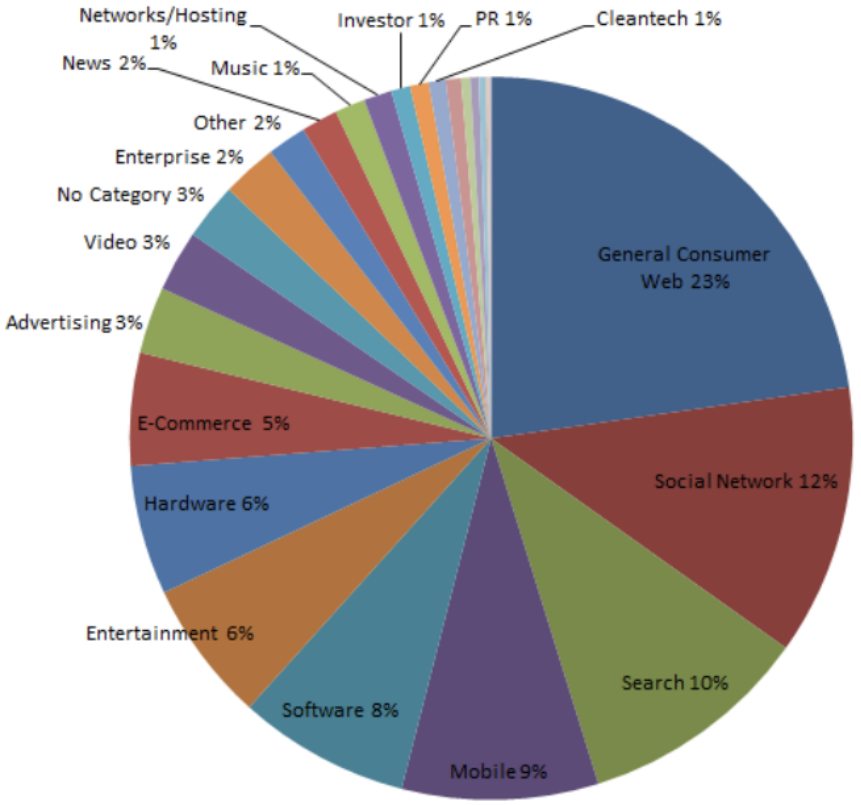


Image retrieved from https://www.livescience.com/45083-misleading-gun-death-chart.html

Guideline: Pie with Care

Pie Charts… easy to get wrong

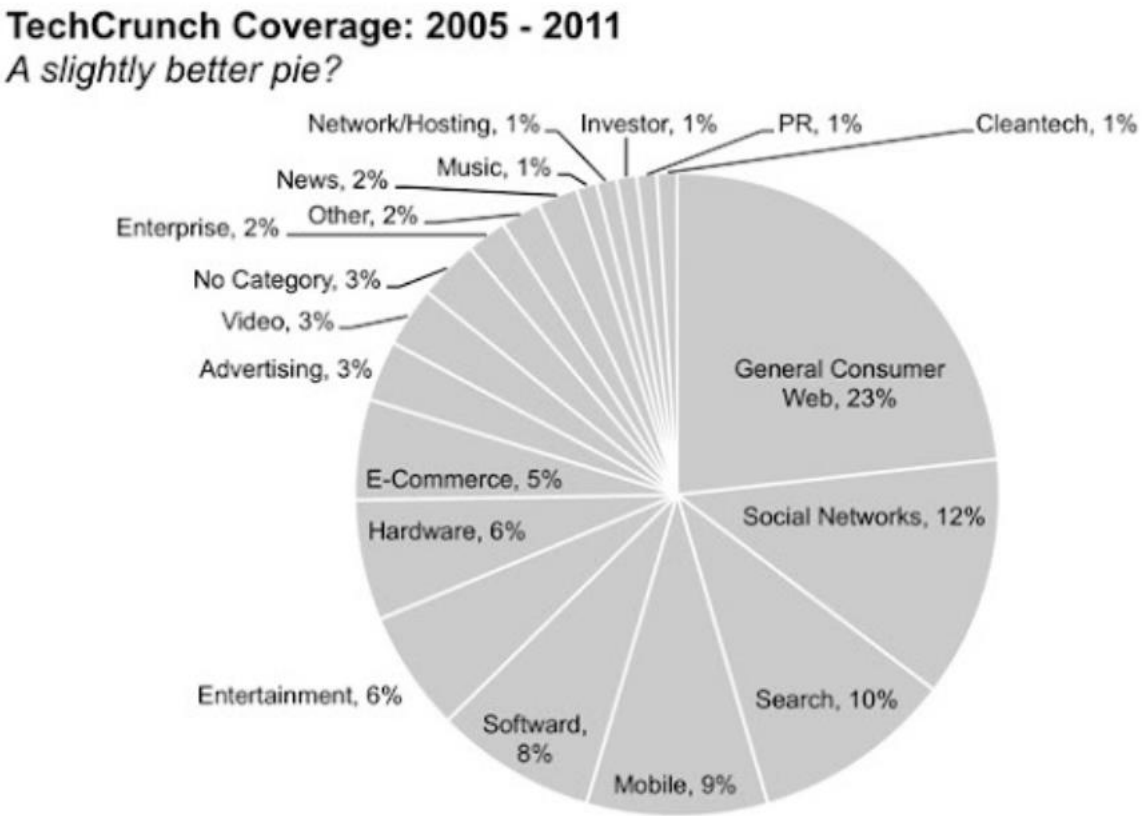


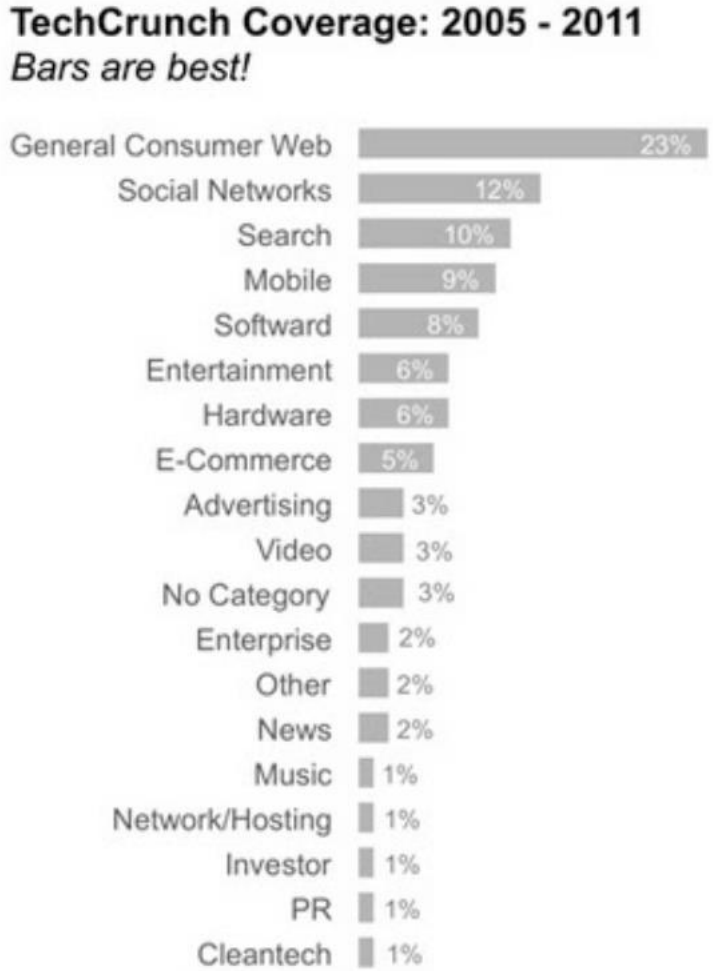
Share of coverage on TechCrunch

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

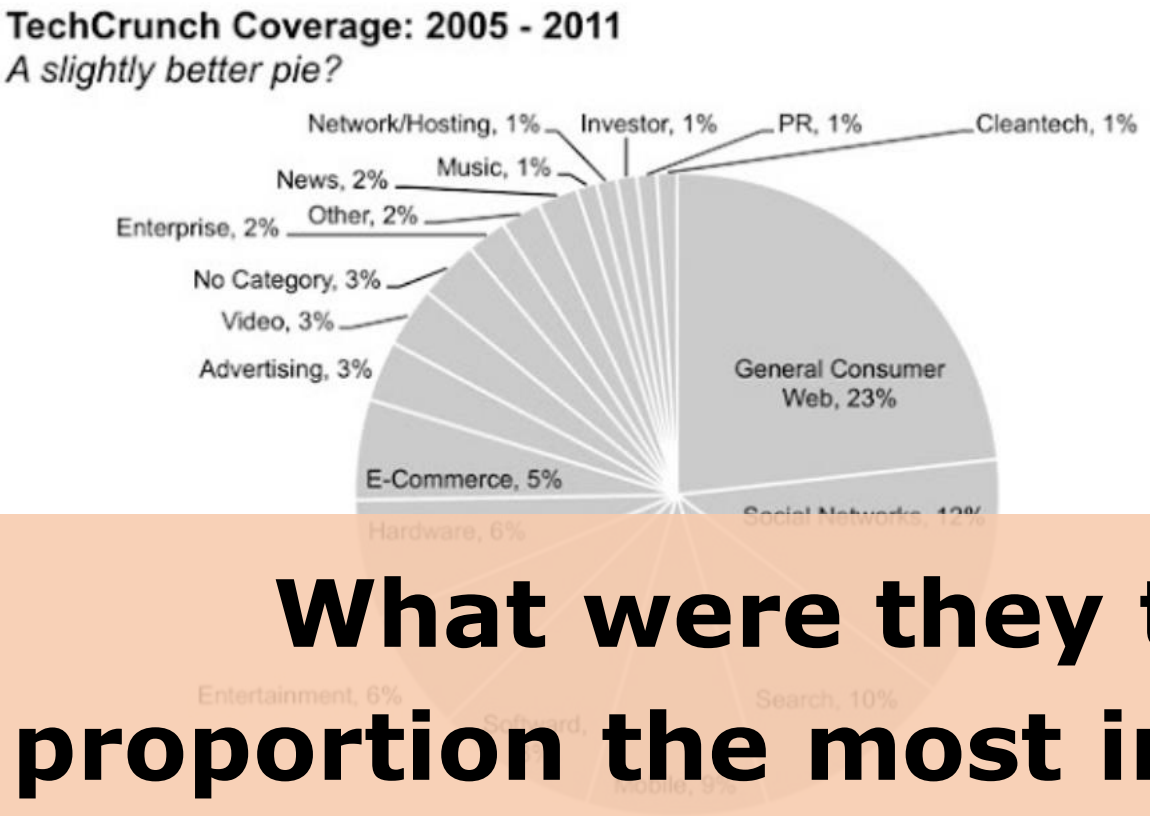
www.storytellingwithdata.com/2011/07/death-to-pie-charts.html

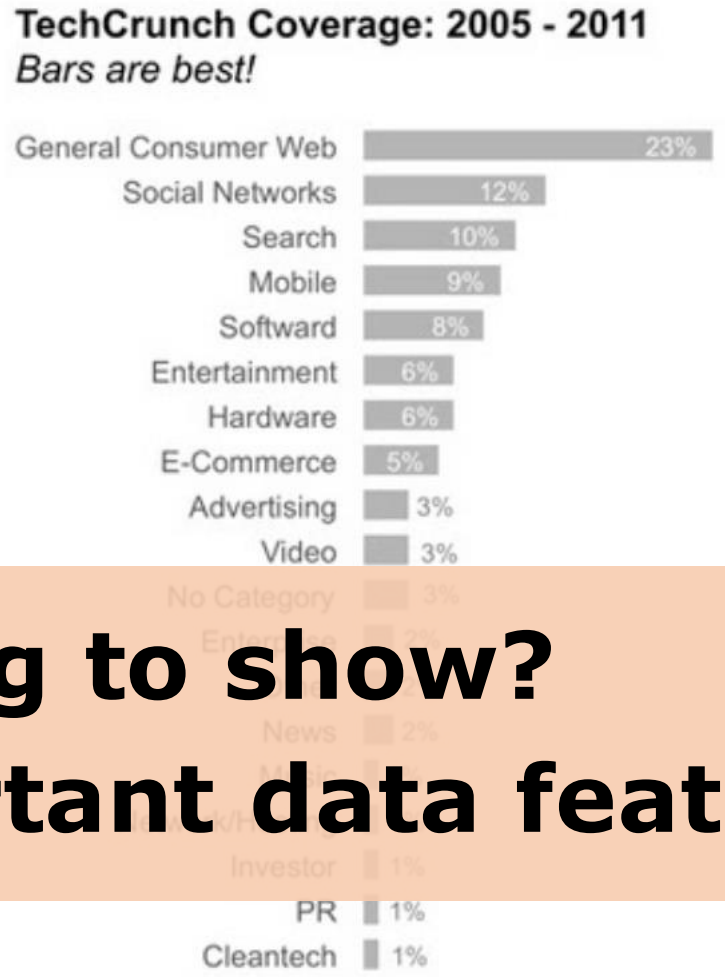
Redesign


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%
Search, 10%
Entertainment, 6%
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%

Redesign

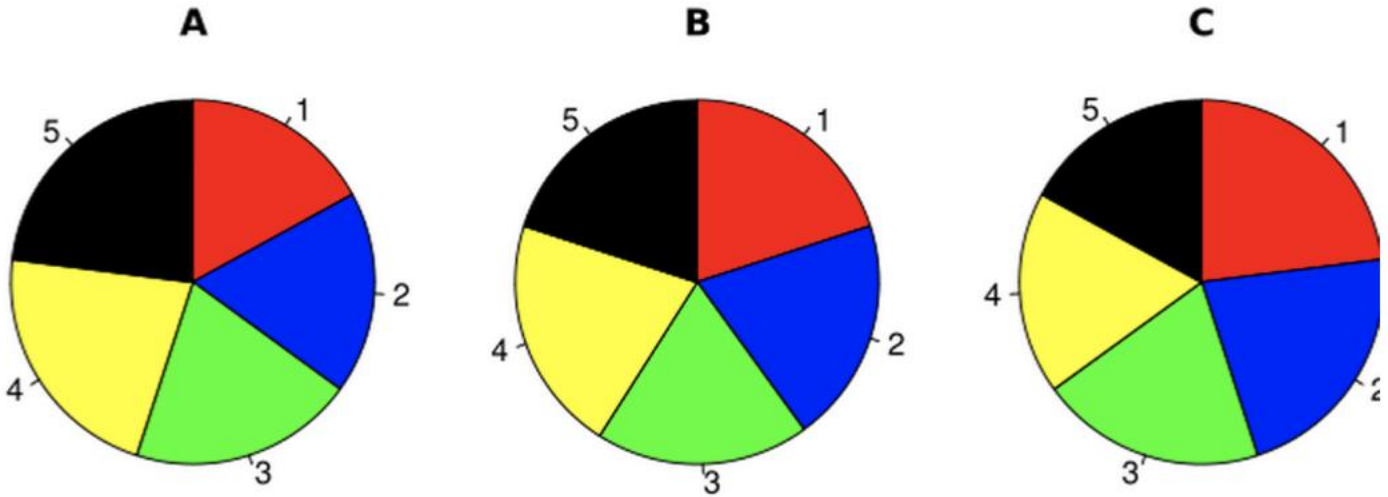




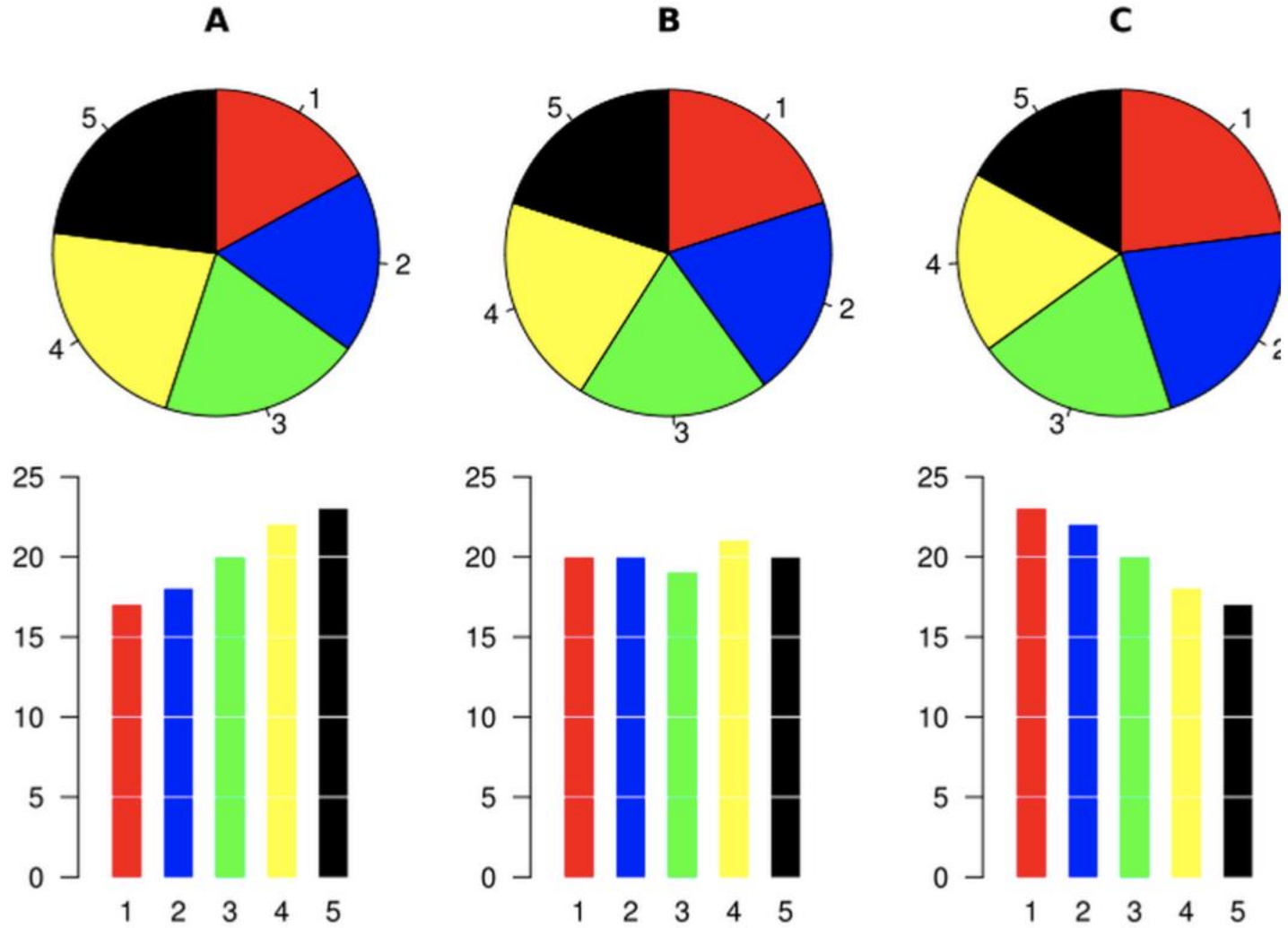
What were they trying to show?

Is proportion the most important data feature?

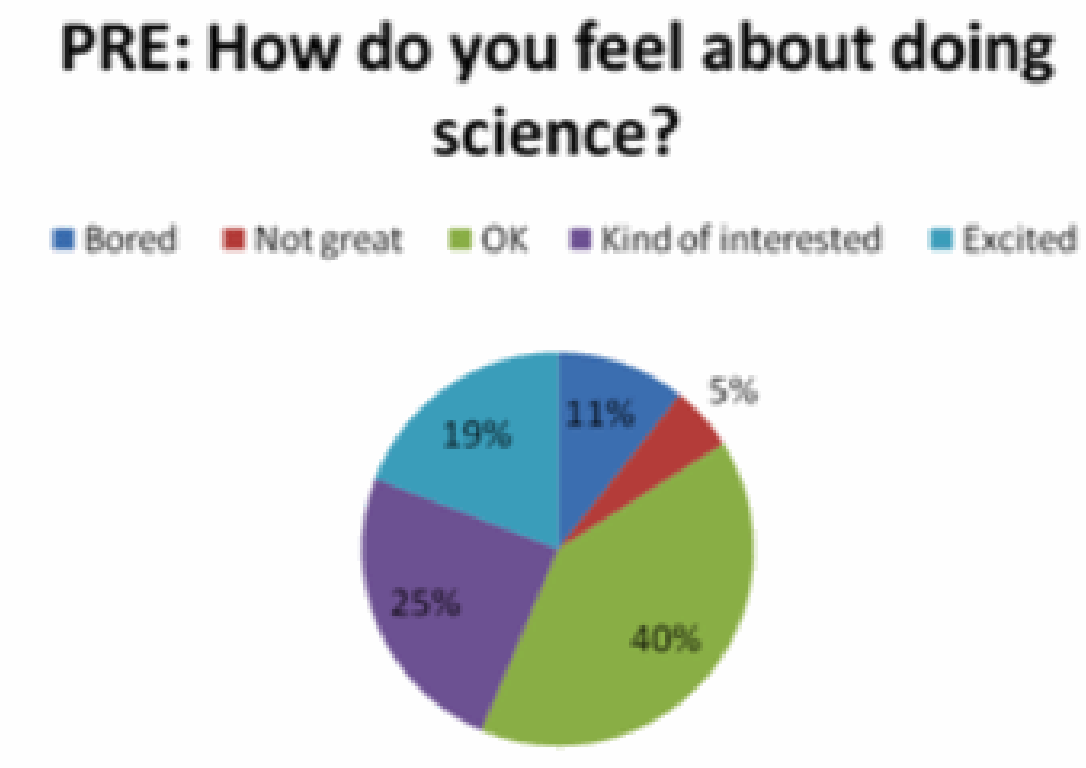
Can you spot the differences?



Can you spot the differences?

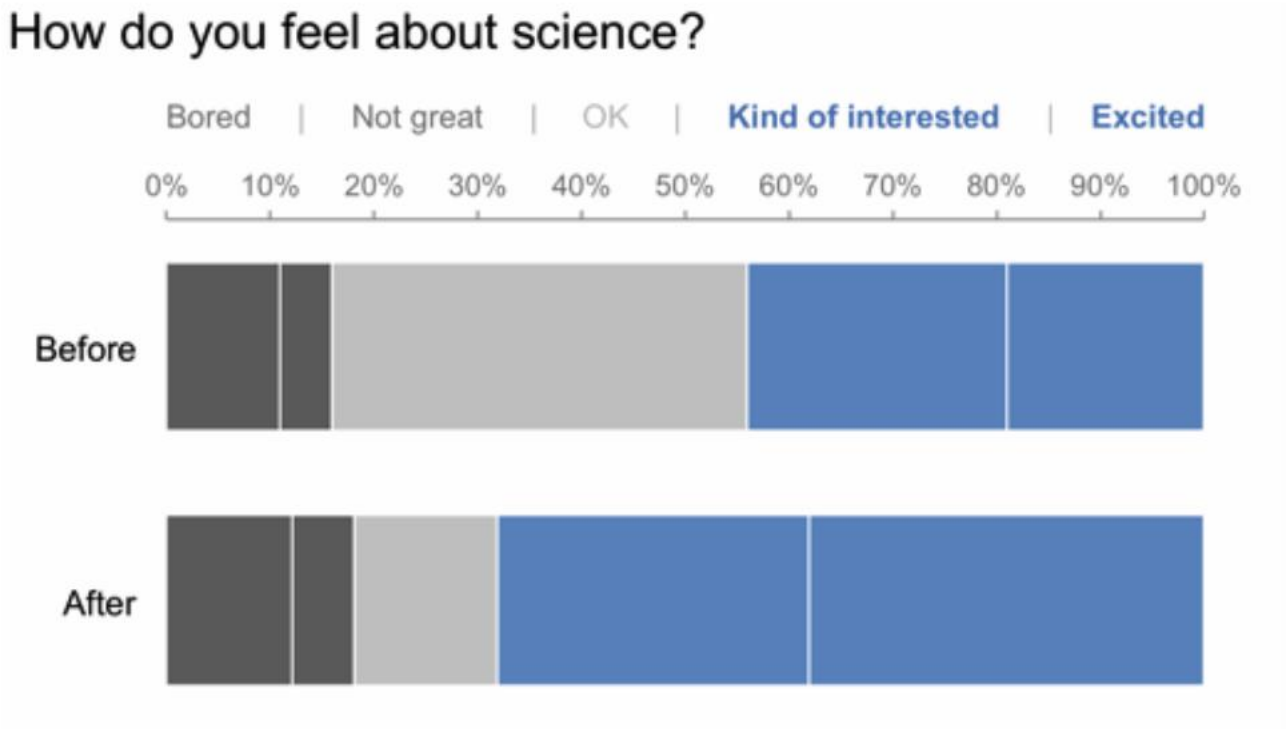


So, what to use instead?

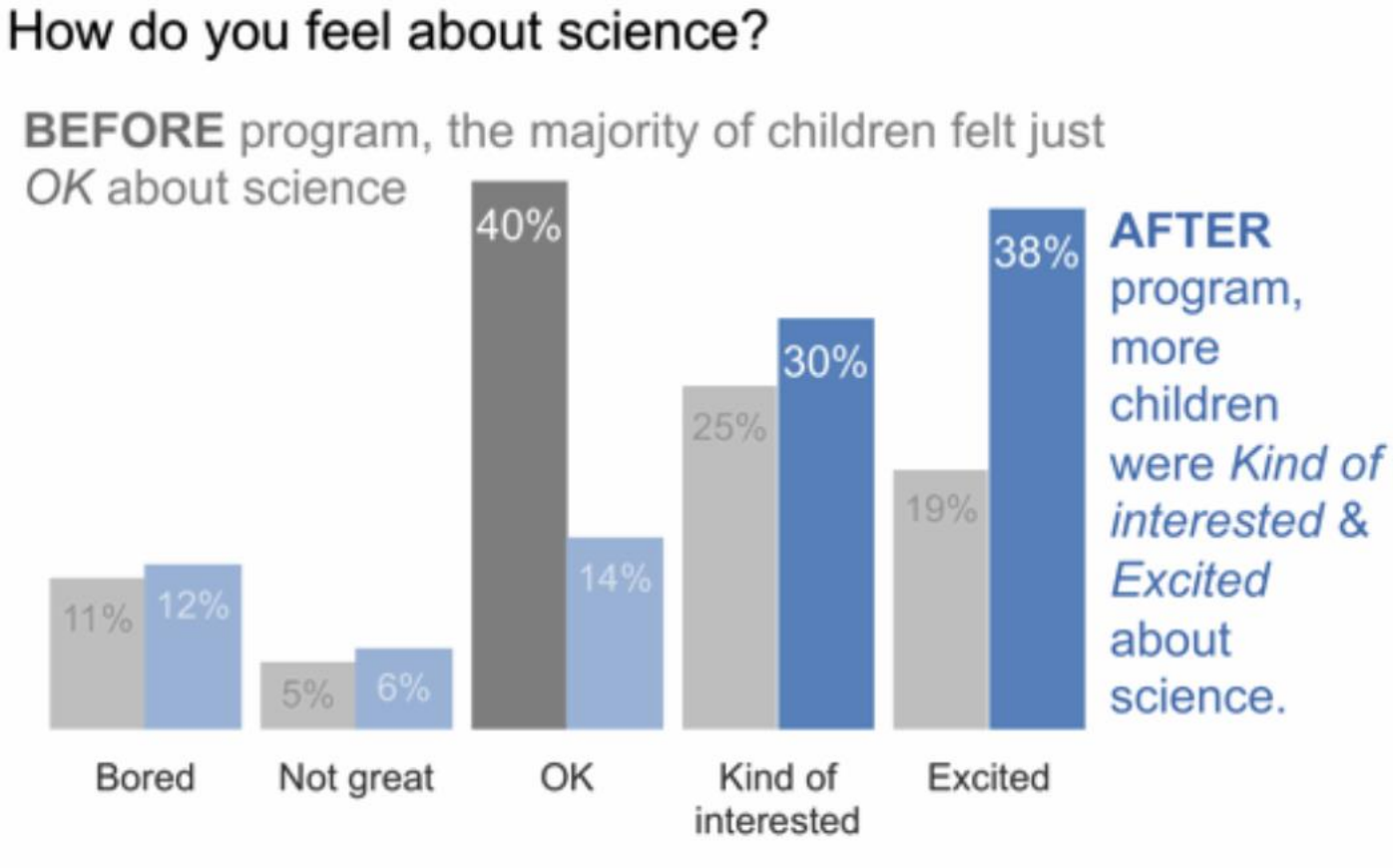


https://www.storytellingwithdata.com/blog/2014/06/alternatives-to-pies

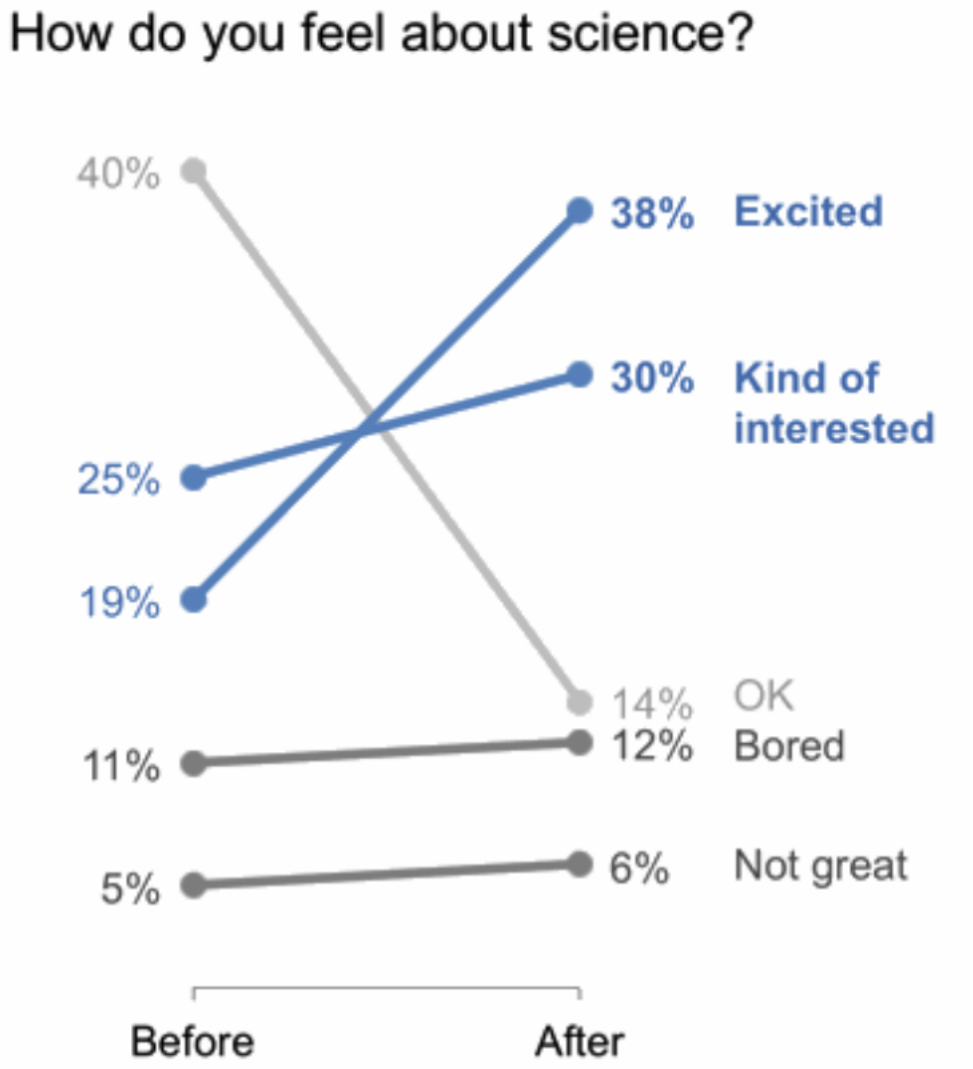
Pie Alternatives: Stacked Bar Charts



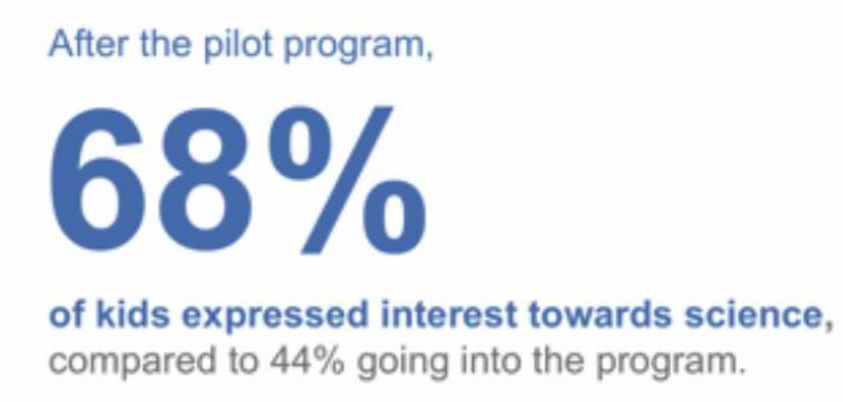
Pie Alternatives: Bar charts



Pie Alternatives: Slope graphs

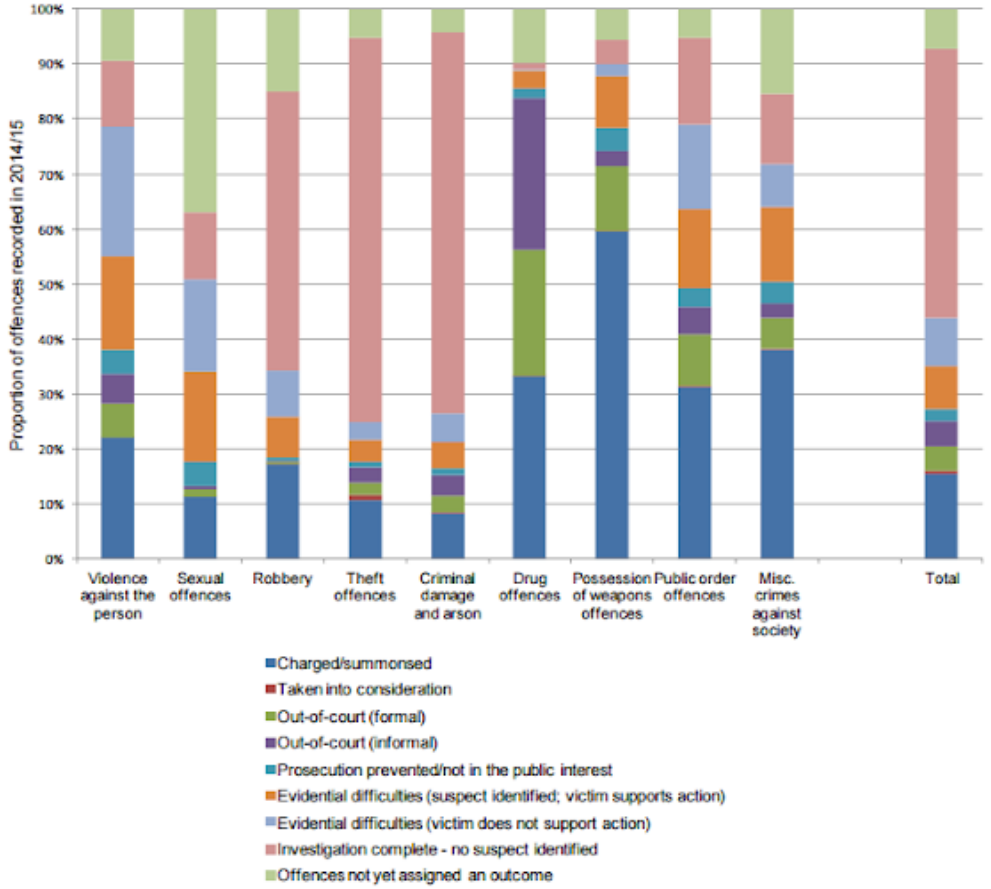


Pie Alternatives: Just show the numbers

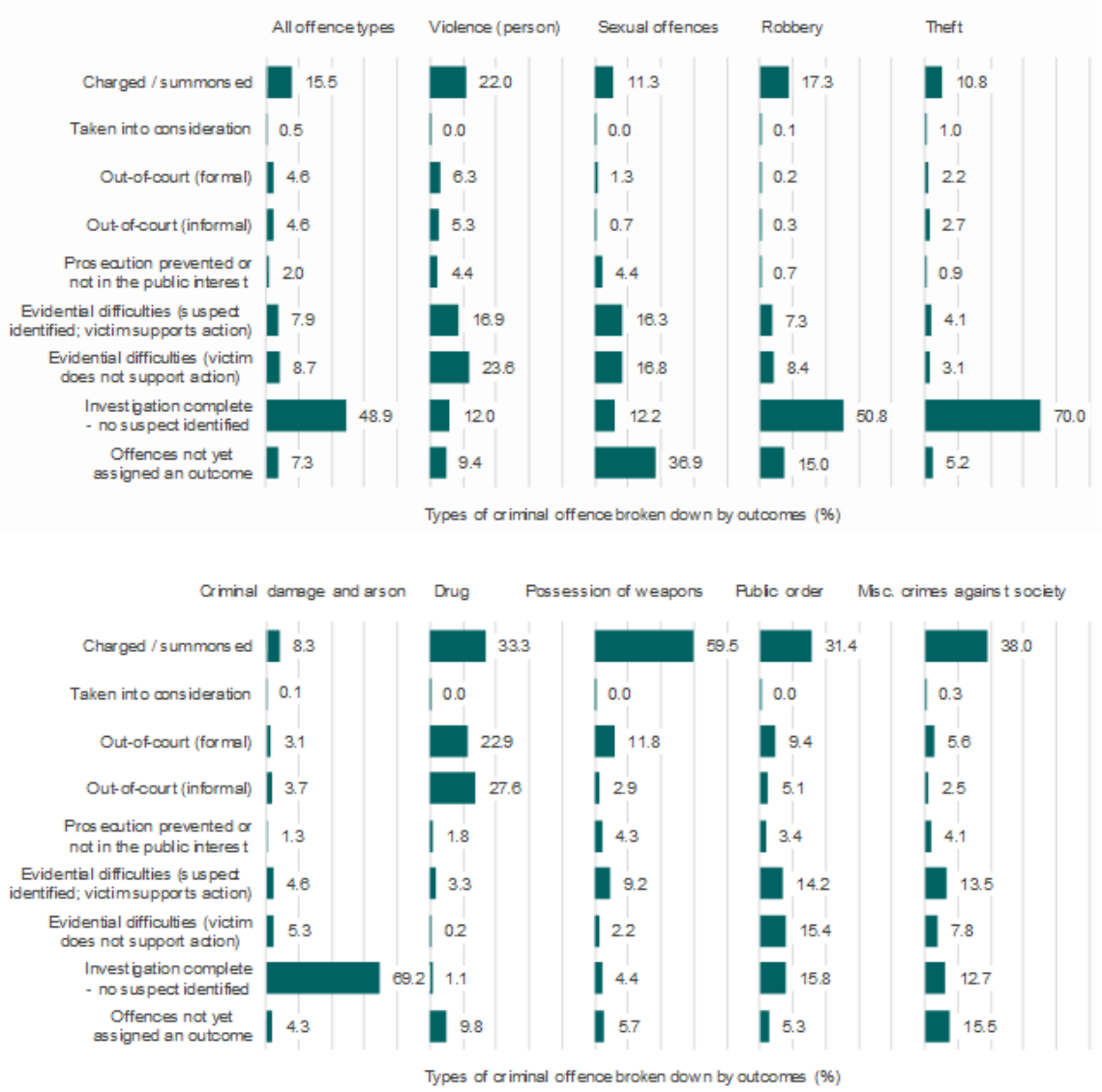


Stacked Bar Charts vs. Small Multiples

Figure 2.1: Outcomes assigned to offences recorded in 2014/15, by outcome group and offence group

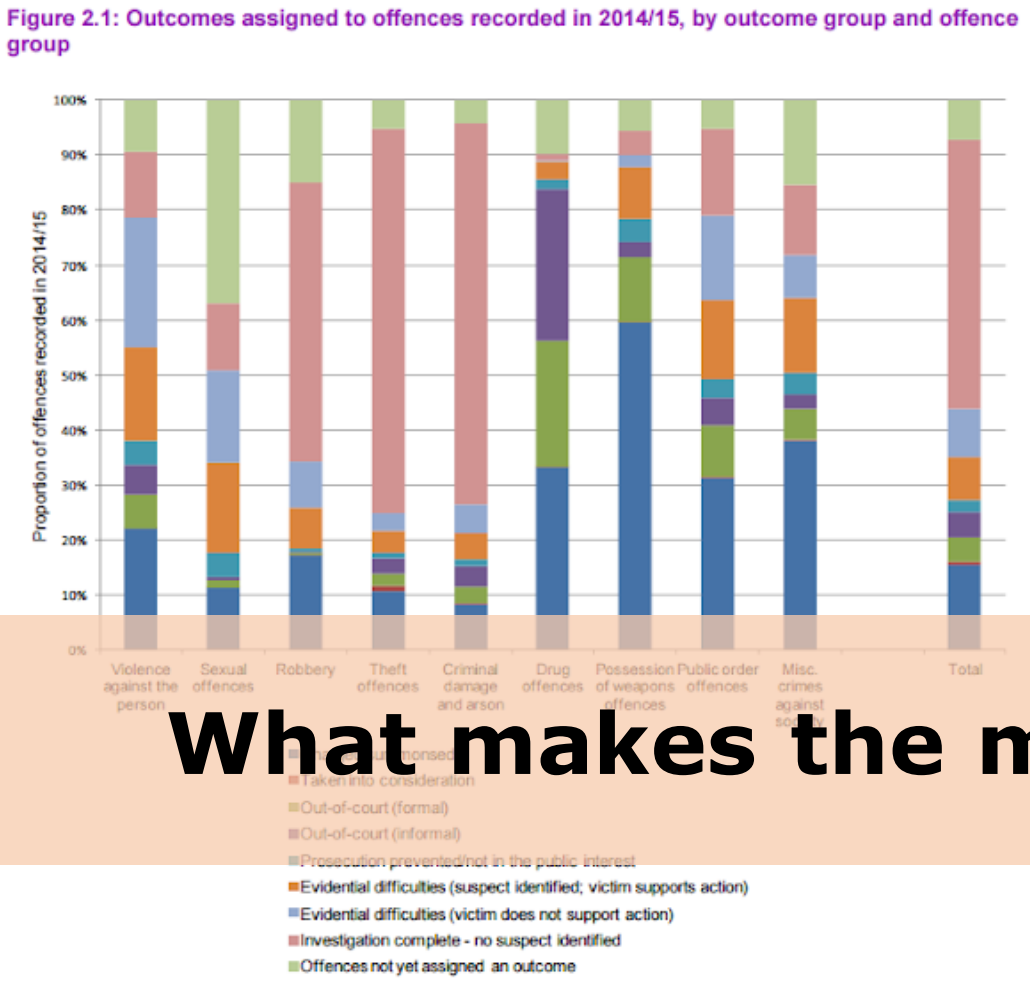


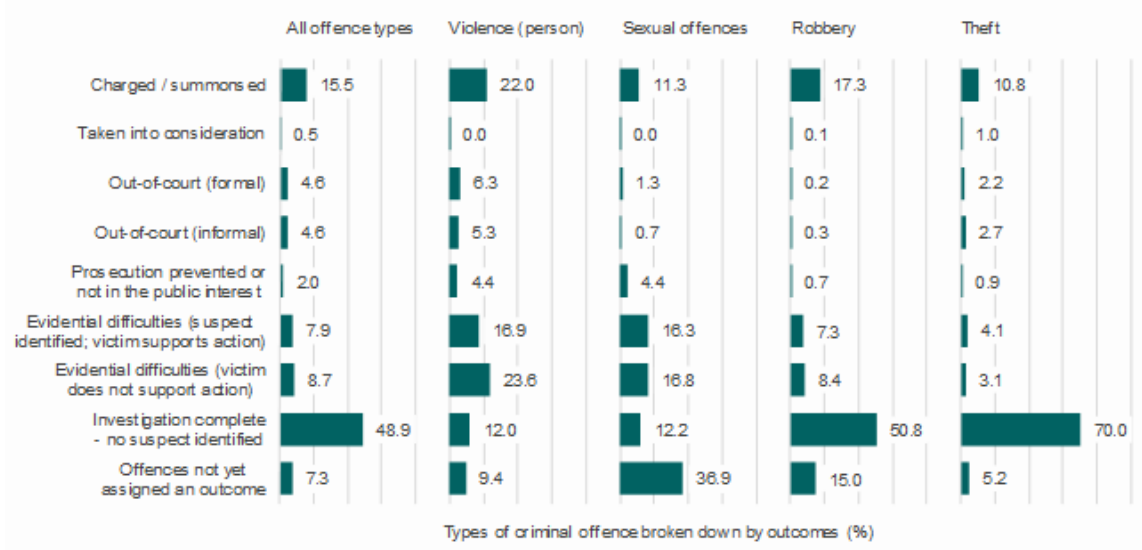
Source: Home Office Data Hub and voluntary spreadsheet return
1. Based on 38 forces that supplied data as referenced in Table 2.1.
2. The numbers behind this chart are in Table 2.3

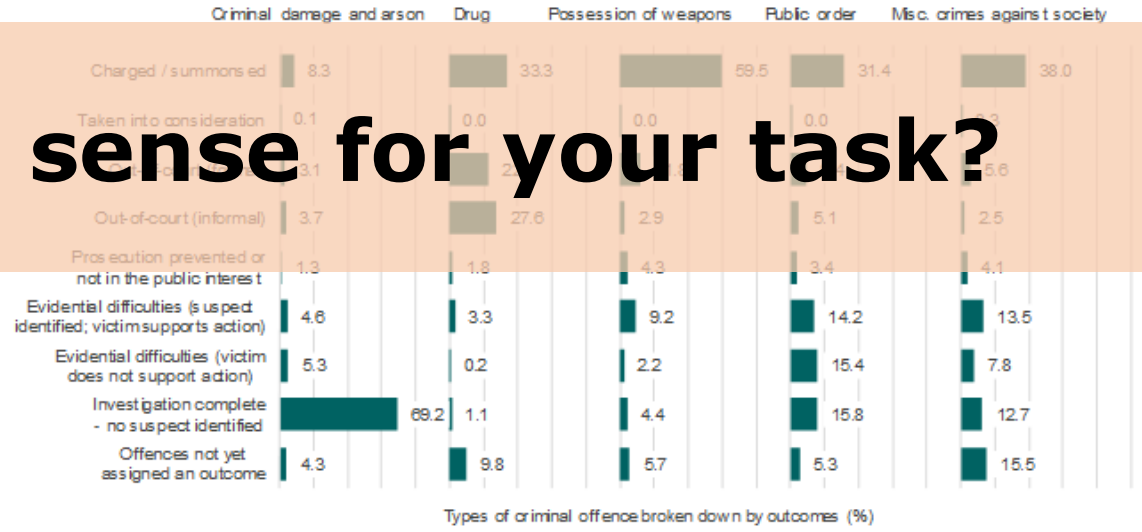


Images from http://chartartistry.blogspot.com/2016/02/small-multiples-when-stack-bars-just.html

Stacked Bar Charts vs. Small Multiples



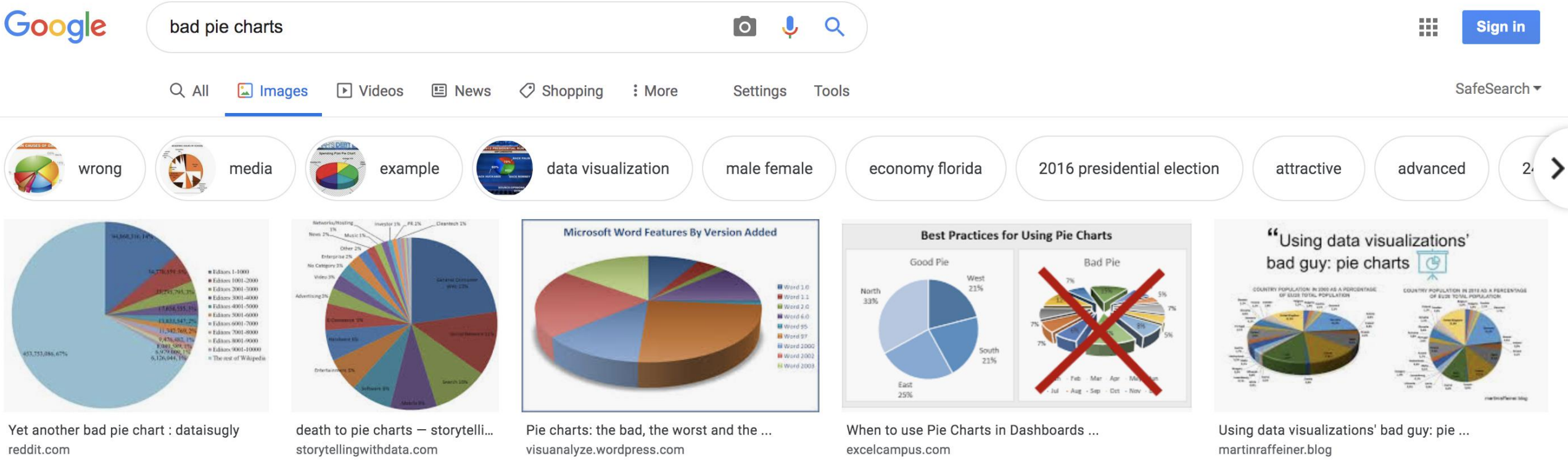




What makes the most sense for your task?

Images from http://chartartistry.blogspot.com/2016/02/small-multiples-when-stack-bars-just.html

Pie charts not inherently bad. Maybe the biggest problem with pie charts is that they have been so often done poorly…

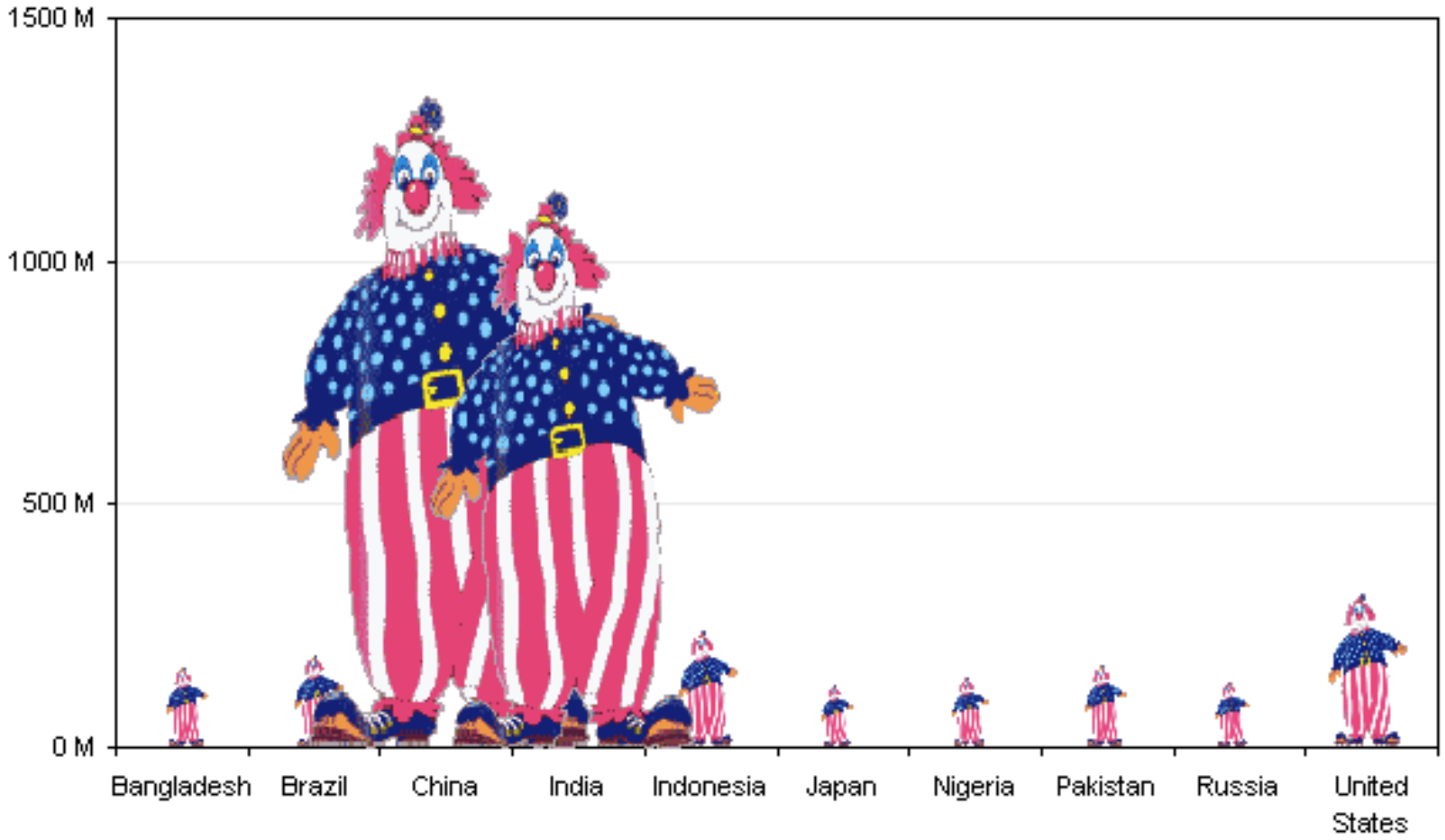


Guideline: Area-as-Quantity with Care

How many streams are there in November compared to December?

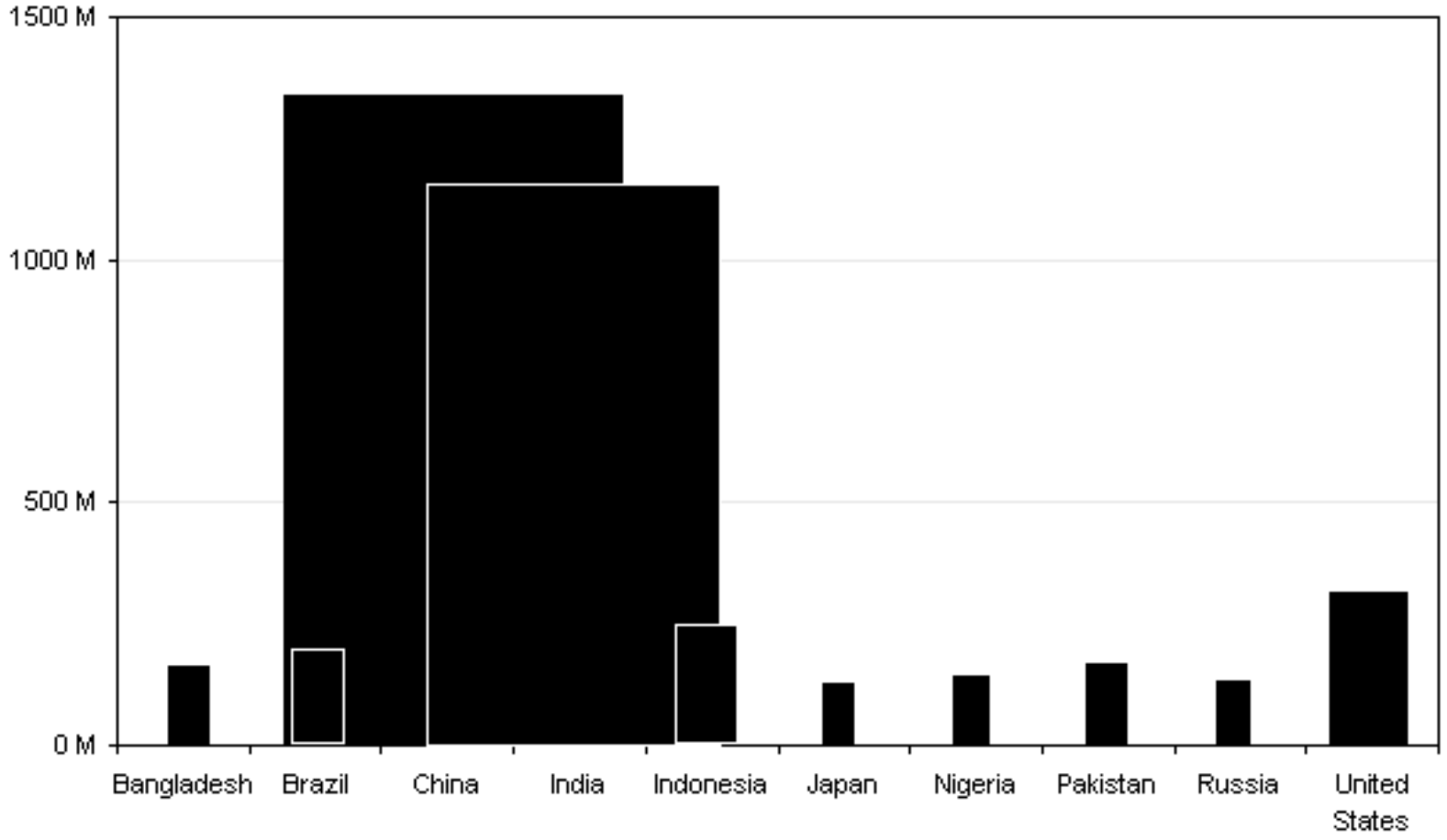
7.5 times as many streams!

Be careful of length vs. area for other marks



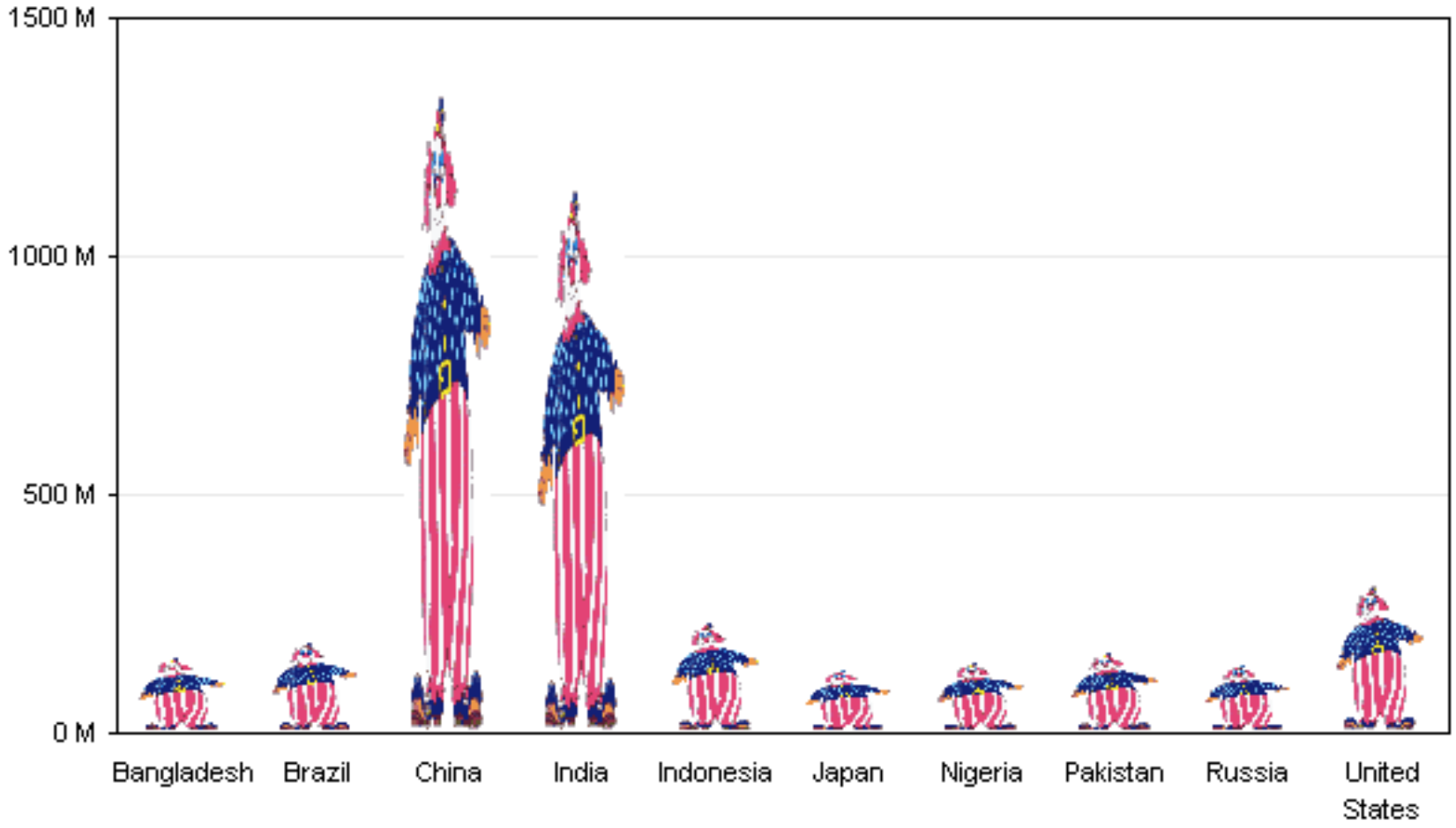
https://peltiertech.com/bad-bar-chart-practices-or-send-in-the-clowns/

What is being perceived?



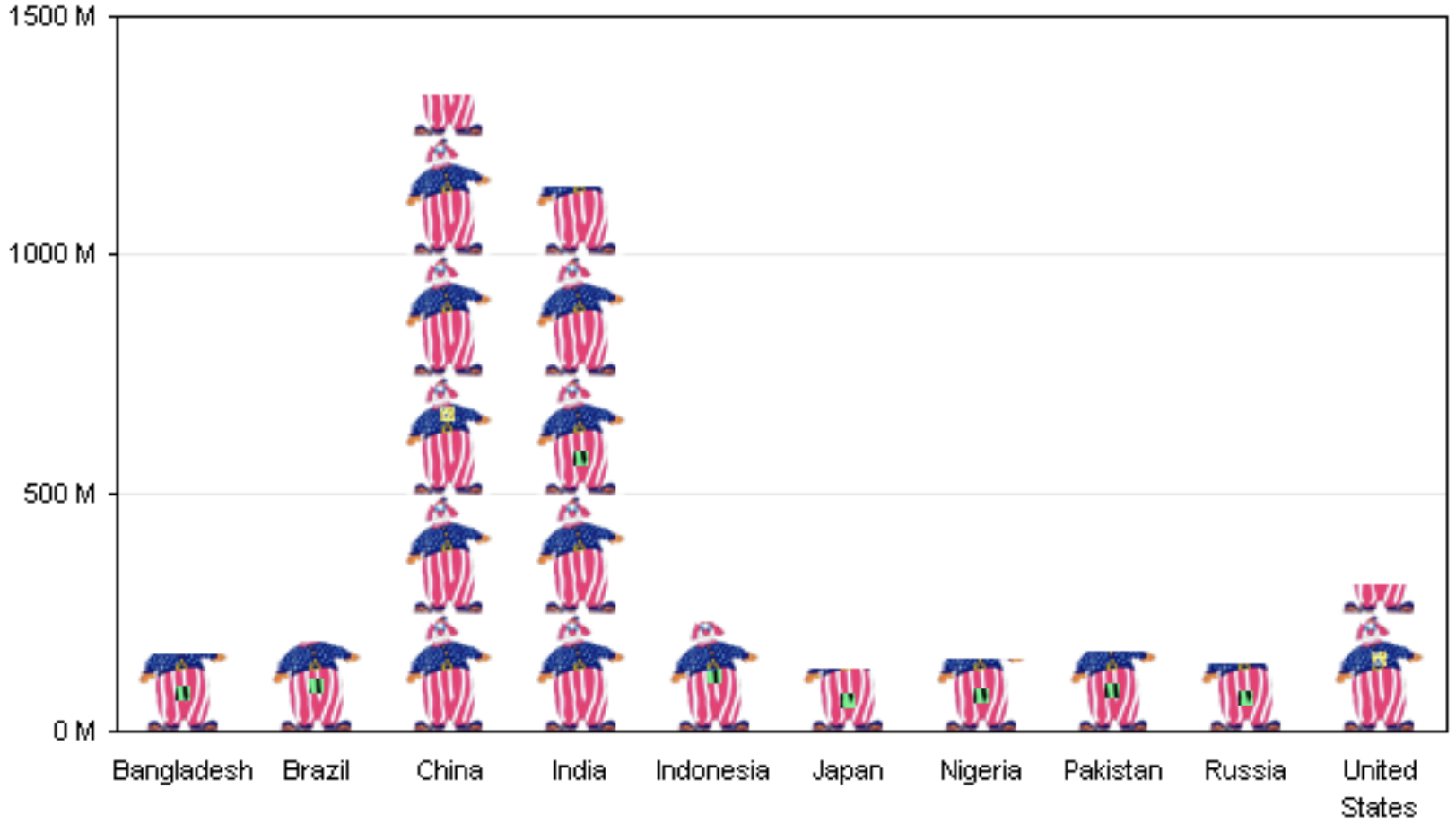
https://peltiertech.com/bad-bar-chart-practices-or-send-in-the-clowns/

Fixing the width



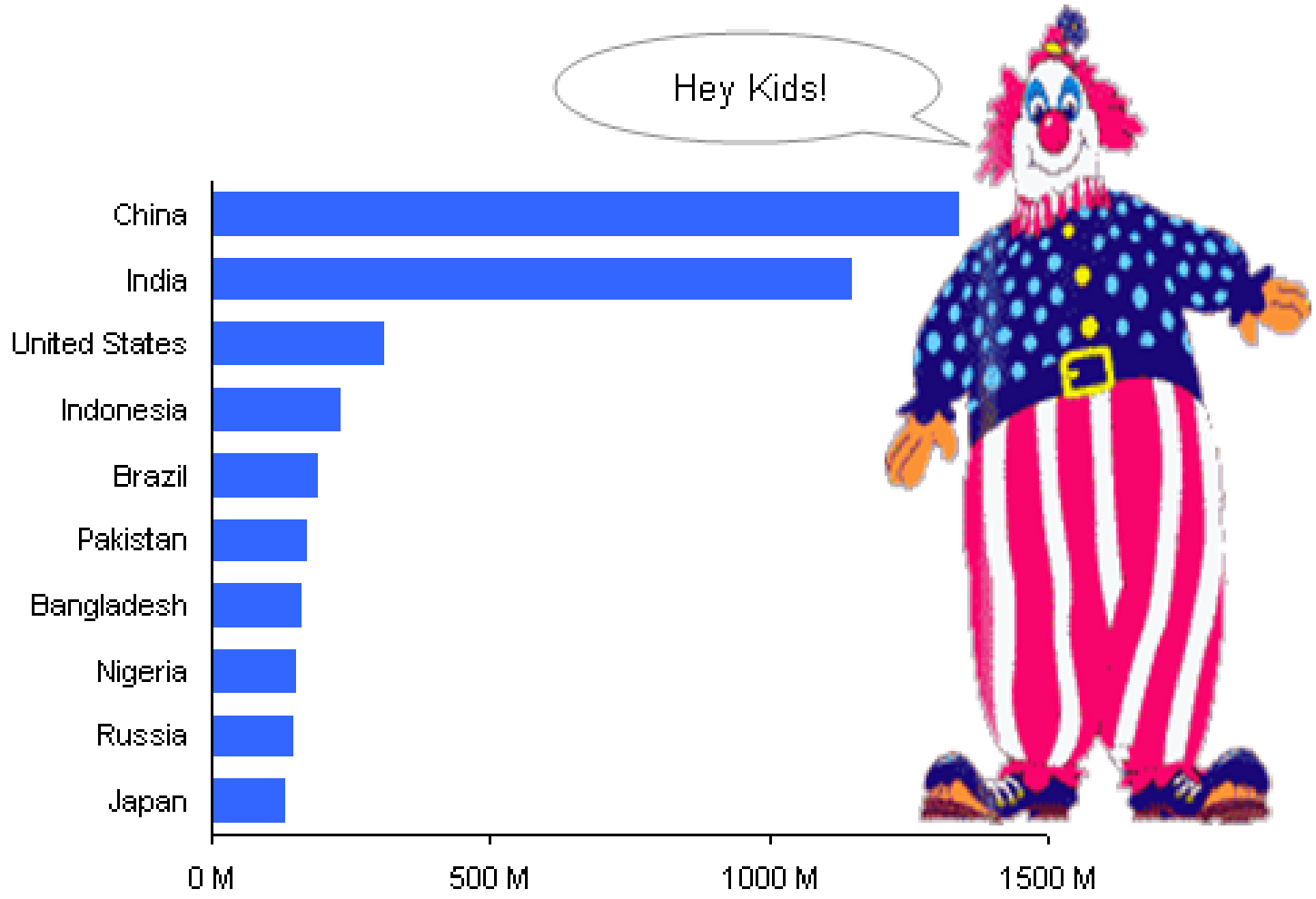
https://peltiertech.com/bad-bar-chart-practices-or-send-in-the-clowns/

Consider using an Isotope Chart



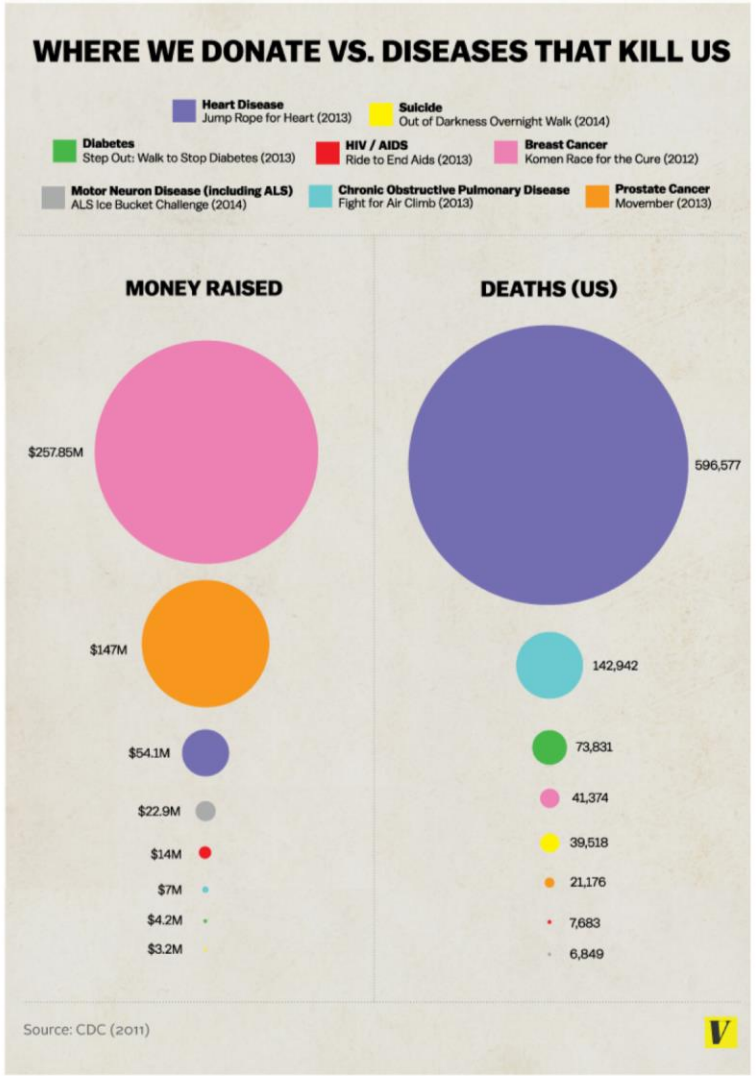
https://peltiertech.com/bad-bar-chart-practices-or-send-in-the-clowns/

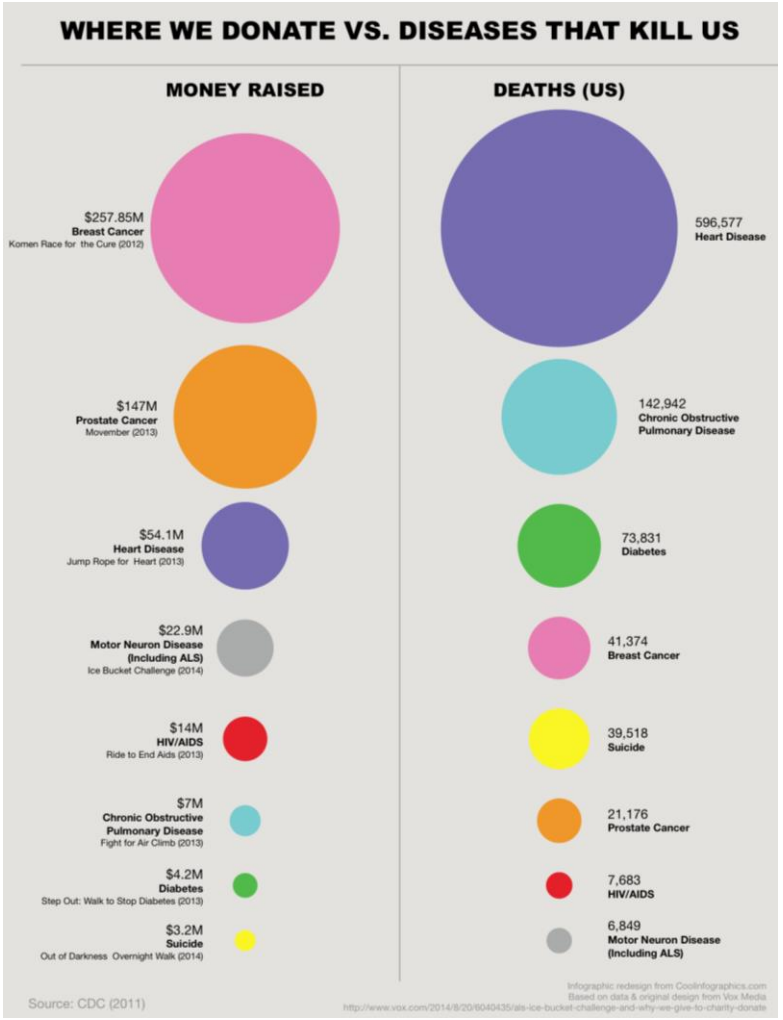
Consider using Horizontal Bar Chart

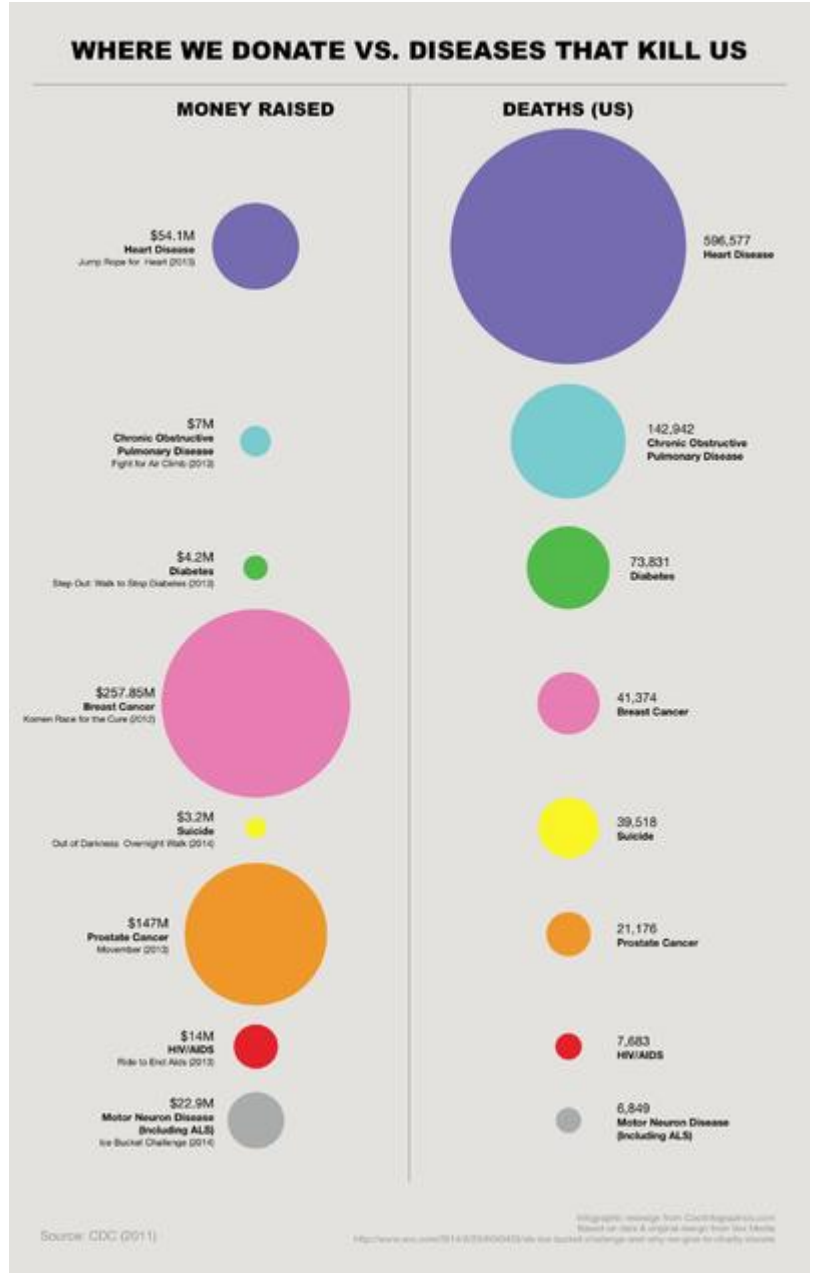


https://peltiertech.com/bad-bar-chart-practices-or-send-in-the-clowns/

Circles: Encode by Area not Radius







Images from Vox and http://coolinfographics.com/blog/2014/8/29/false-visualizations-sizing-circles-in-infographics.html