

Welcome (Again!) to **MATH 4100/COMP 5360– Introduction to Data Science**

(Some) Principles of Data Visualization

February 21, 2023

*Based on prior lectures from
Alex Lex and Bei Wang Phillips
+ others where noted*



Announcements / Reminders / Schedule

HW 5 – Due this Friday!

HW 6 – Due Friday March 3!

This class: basics of data visualization principles

Thursday: web-scraping

Next Tuesday: finishing up data visualization module

Am I recording this lecture? Remind me if I'm not yet.

**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:

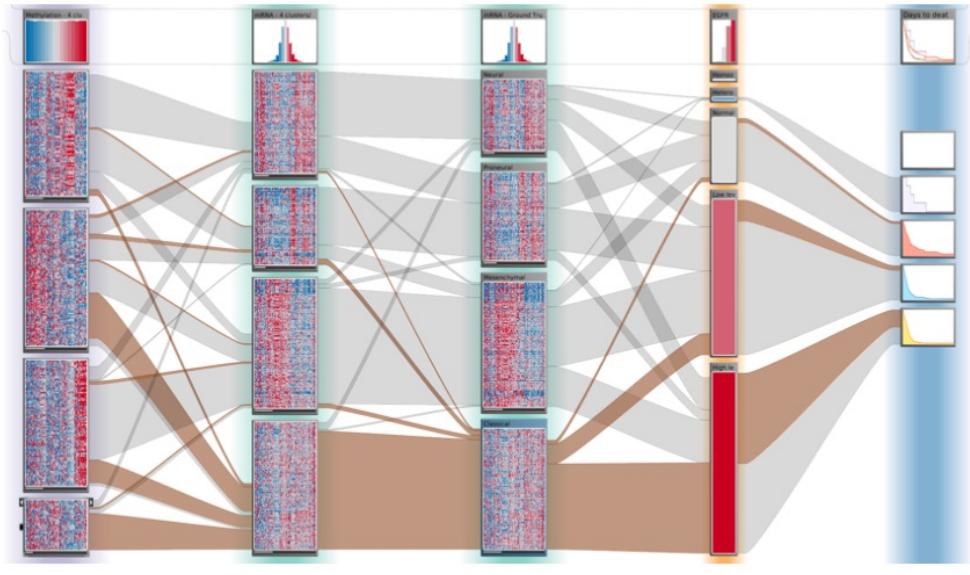
Exploration

(when results and questions are unknown)

What is the structure of the scammer network?

Which drug can help my patient?

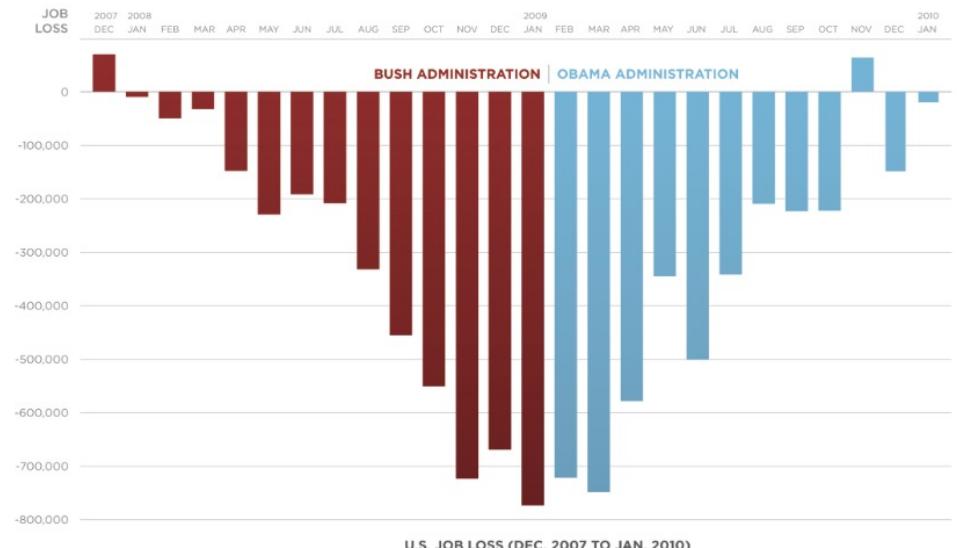
Is this data even right?



Open Exploration

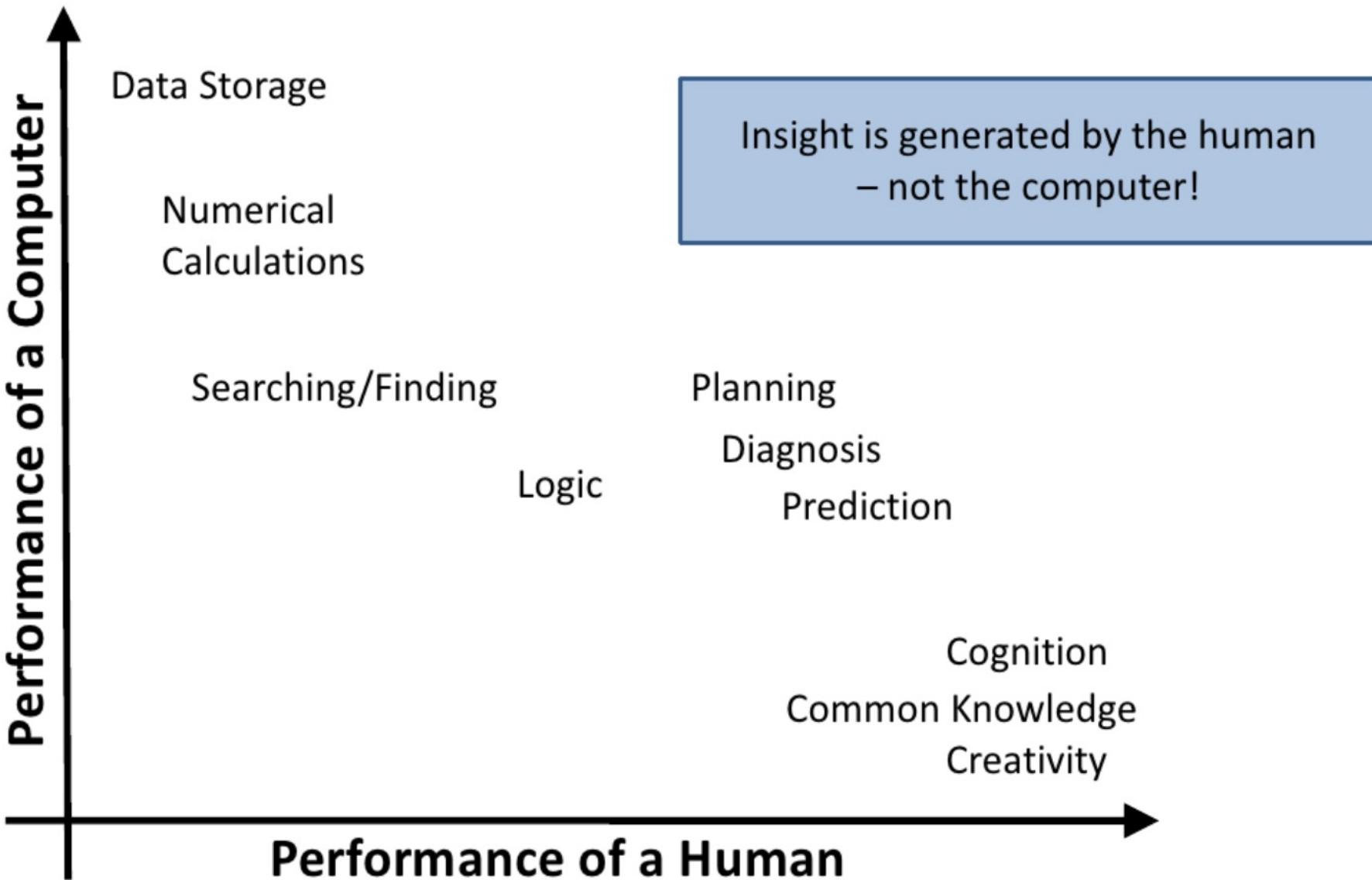
Confirmation

Communication

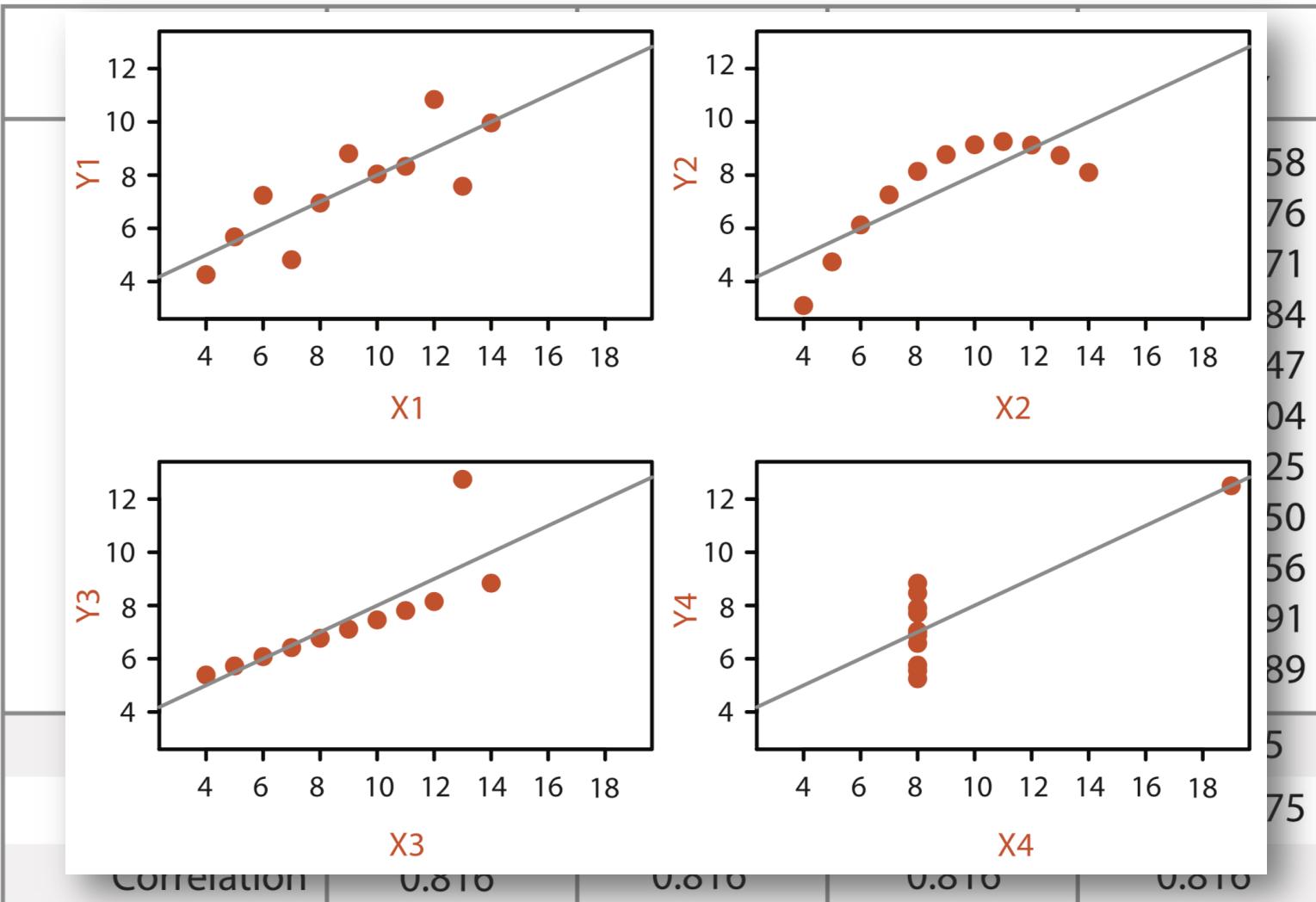


SOURCE: BUREAU OF LABOR STATISTICS, 02/02/2010

Ability Matrix

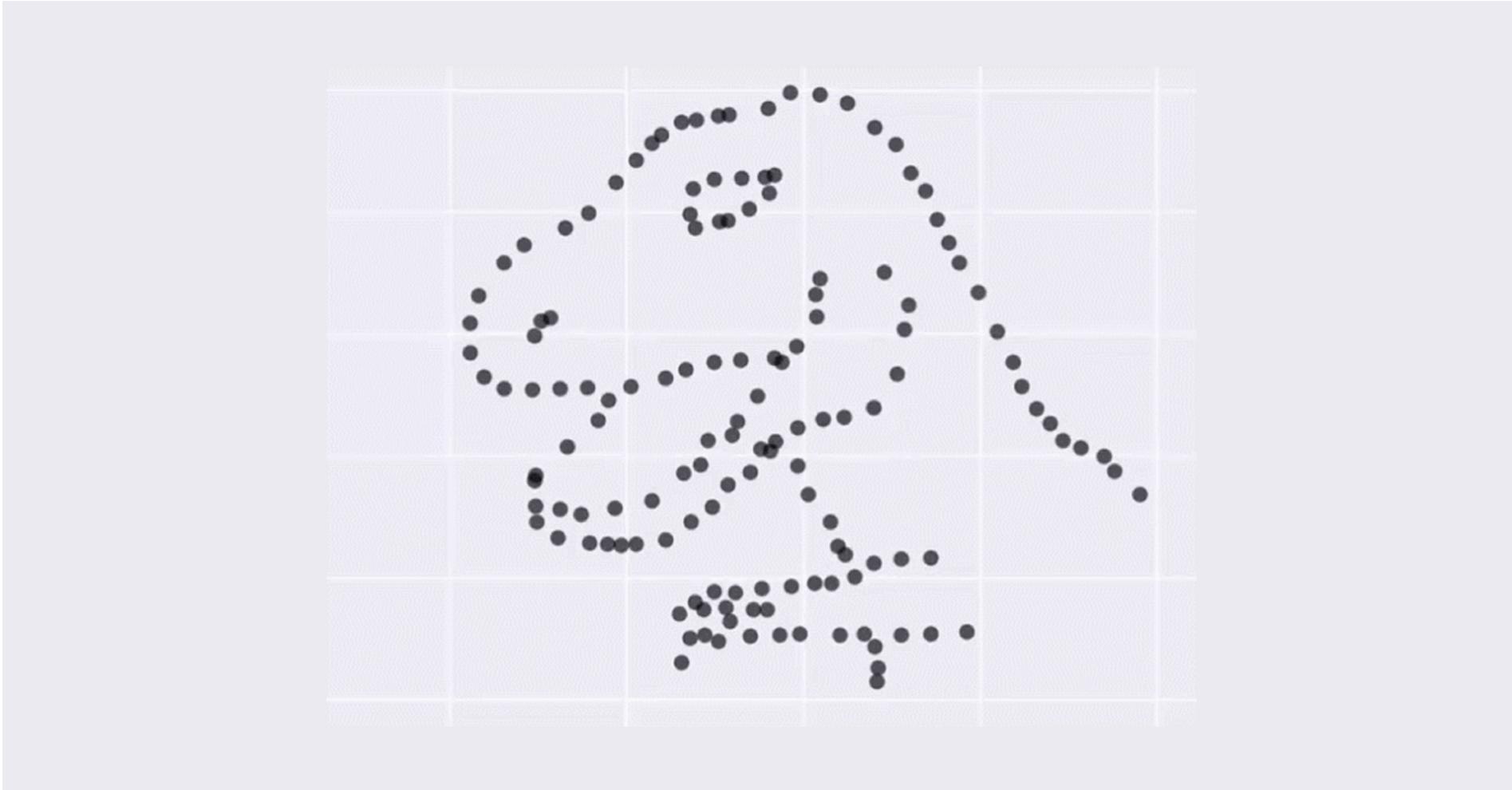


Why Visualization?



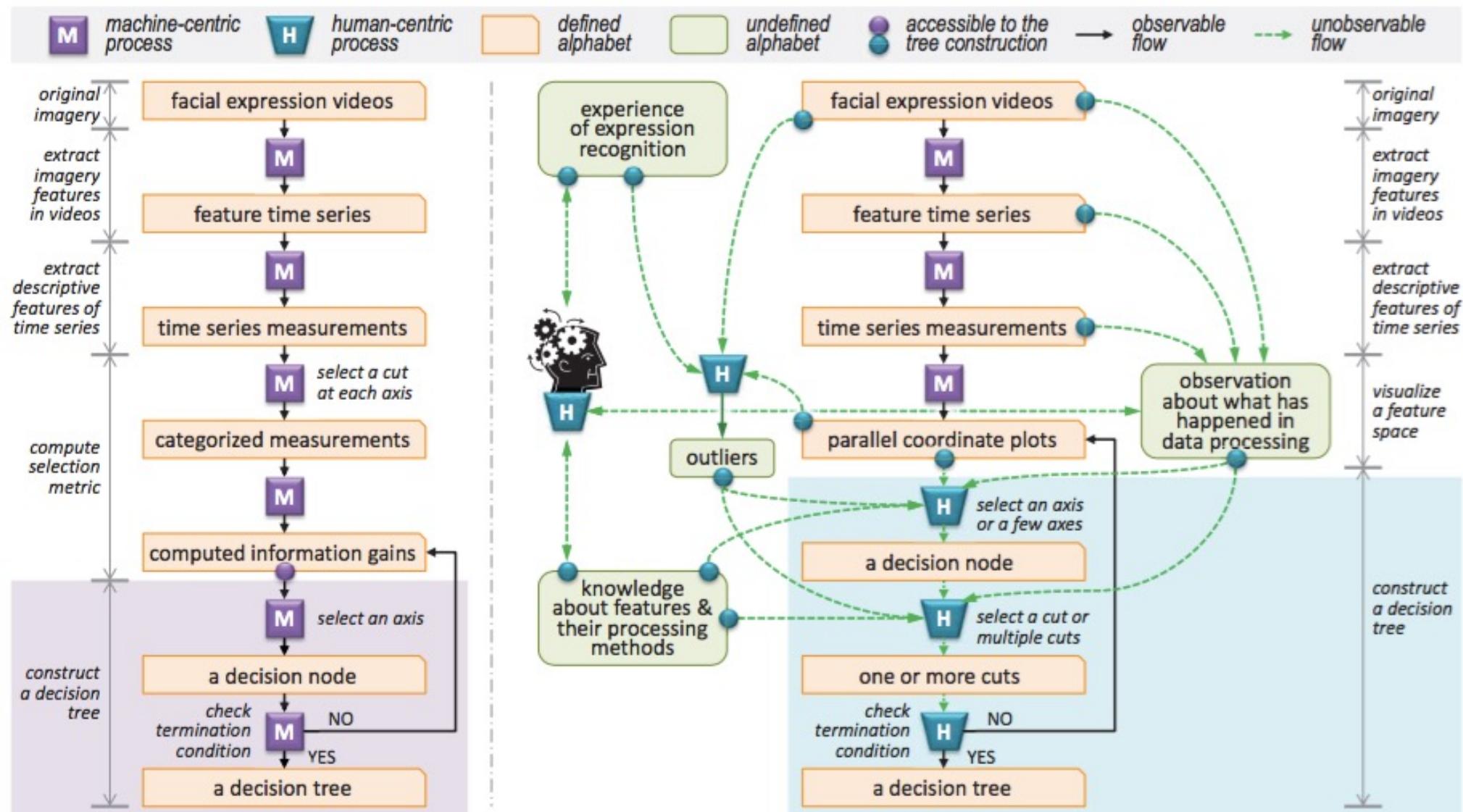
Sometimes summary statistics don't tell the full story.

Why Visualization?



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

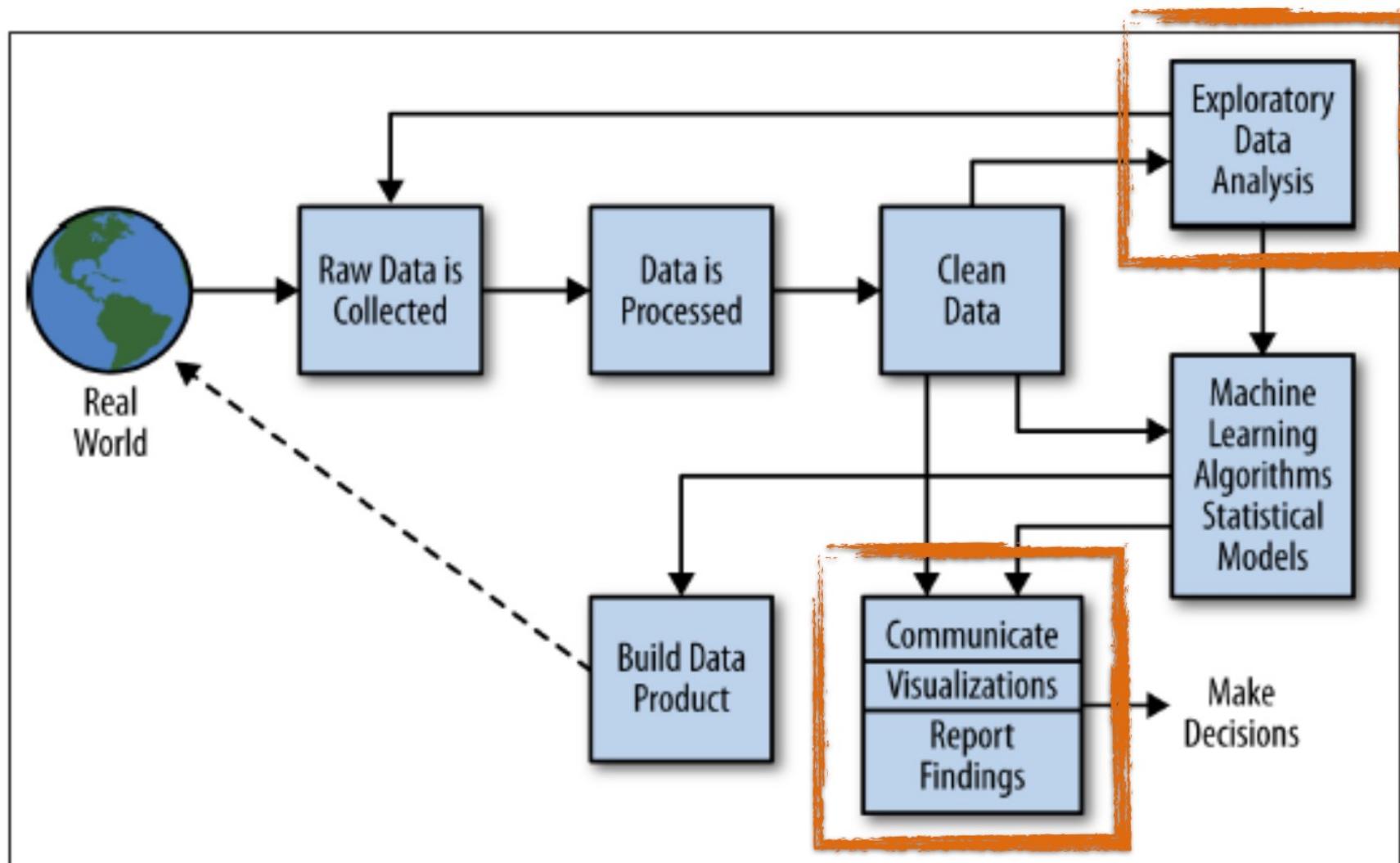


Figure 2-2. The data science process

Why Humans?

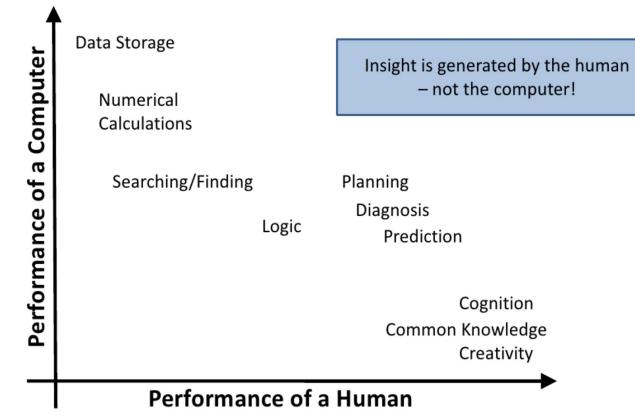
Leverage human capabilities:

Pattern discovery: clusters, outliers

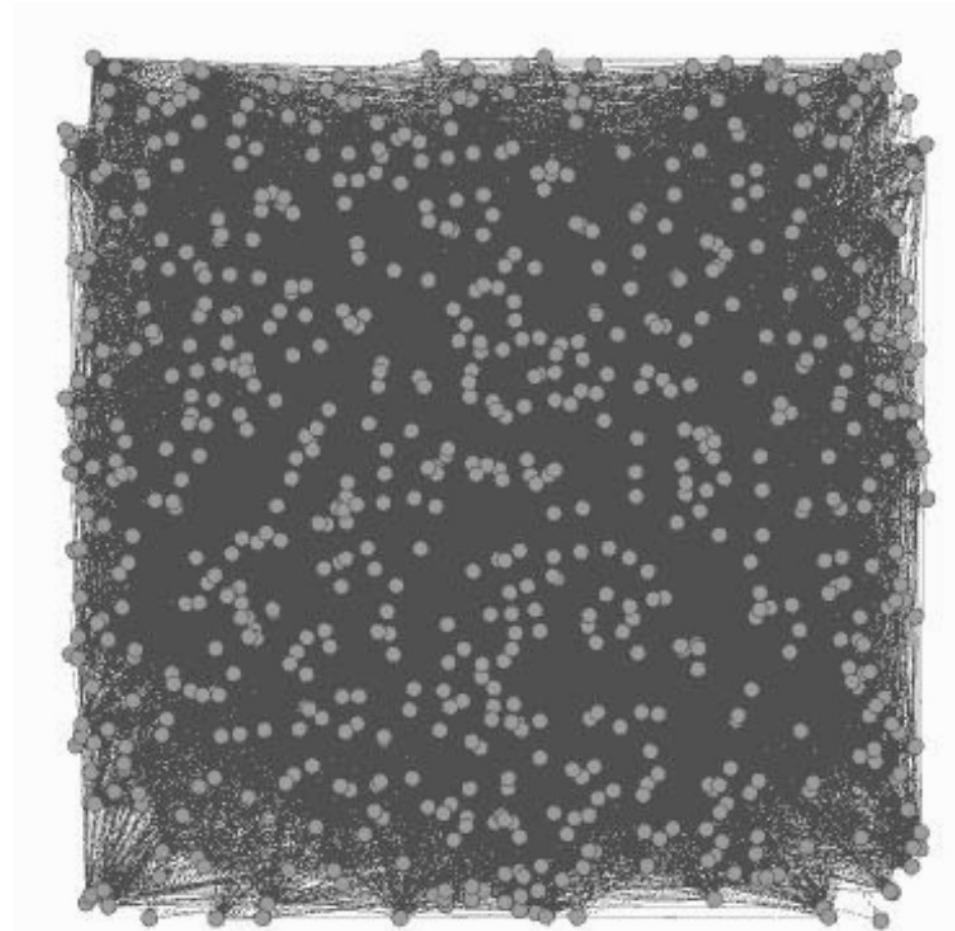
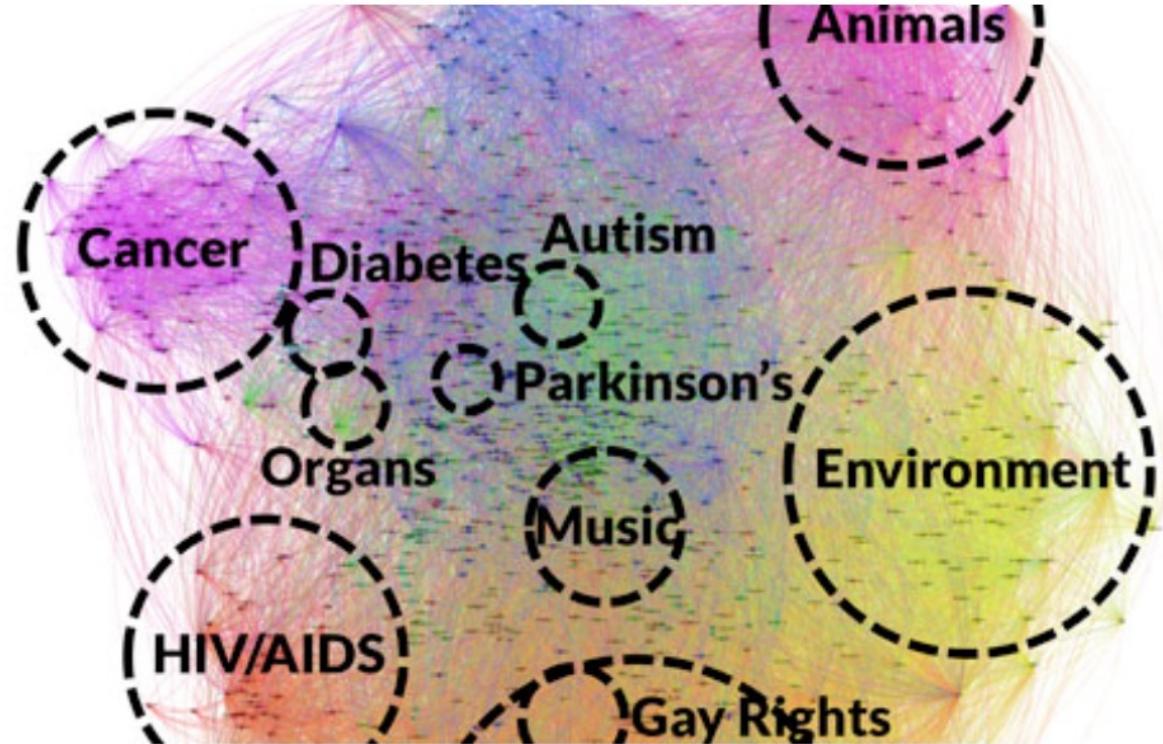
Contextual knowledge: expectations, explanations of patterns

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

Visual Vocabulary Deviation Correlation Ranking Distribution Change over Time Part-to-Whole Magnitude Spatial Flow

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationship is causal (one causes the other).

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

A scatterplot showing the relationship between the percentage of people with a BA degree (Y-axis) and the percentage of obese people (X-axis). The X-axis ranges from 20 to 36, and the Y-axis ranges from 10 to 50. A dashed regression line shows a negative correlation. Data points are labeled with state abbreviations. Two specific points are annotated: DC (high BA degree, very high obesity) and WV (low BA degree, low obesity). A vertical line at approximately 27.5% on the X-axis is labeled "US obesity average: 27.0%" and a horizontal line at approximately 27.2% on the Y-axis is labeled "People with a BA, US average: 27.2%".

Line + Column

A good way of showing the relationship between an amount (columns) and a rate (line).

A dual-axis chart showing Sales (left Y-axis, 0K to 400K) and Profit (right Y-axis, 0K to 80K) over time. The X-axis represents the Quarter of Order Date from 2015 Q3 to 2018 Q3. Sales are represented by brown bars, and Profit is represented by a black line. Both metrics show significant seasonal fluctuations, with a notable peak in Q3 2017.

Connected Scatter

Usually used to show relationships over time.

A line chart showing the top 0.01% of sales over time. The Y-axis represents the percentage of sales, ranging from 0% to 10%. The X-axis represents time, with labels at 15%, 10%, 5%, and the present. The data shows a general upward trend with some fluctuations.

Bubble

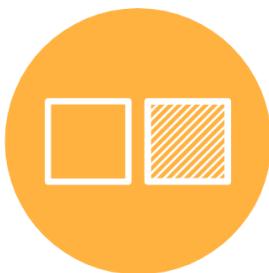
Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY Heatmap

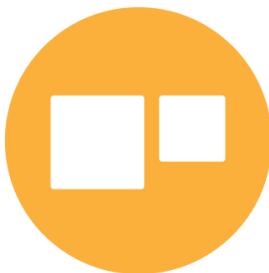
A good way of showing the patterns between 2 categories of data, less good at showing fine differences.

What do you want to show?

Here you can find a list of charts categorised by their data visualization functions or by what you want a chart to communicate to an audience. While the allocation of each chart into specific functions isn't a perfect system, it still works as a useful guide for selecting chart based on your analysis or communication needs.



Comparisons



Proportions



Relationships



Hierarchy



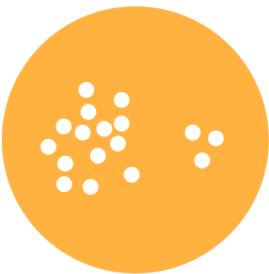
Concepts



Location



Part-to-a-whole



Distribution



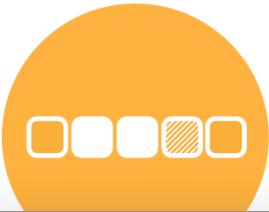
Data Sources



Tools



Flow



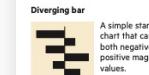
Progress

Visual Vocabulary

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it could be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change



A simple standard bar chart that can handle both negative and positive magnitude values.



Perfect for presenting survey results which involve sentiment (eg disagree/neural/agree).



Splits a single value into two contrasting components (eg male/female).

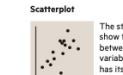


The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be careful with this as you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy



The standard way to show the relationship between two continuous variables, each of which has its own axis.



A good way of showing the relationship between an amount (columns) and a rate (line).



Usually used to show the relationship between 2 variables has changed over time.



Like a scatterplot but adds additional detail by sizing the circles according to a third variable.



A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.



Perfect for showing how ranks have changed over time or vary between categories.



Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.



Effective for showing changing rankings across multiple dates.

For large datasets, consider grouping lines using color.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results



Standard bar charts display the ranks of values much more easily when sorted into order.



See above.



Use when there are big differences between values and/or seeing fine differences between data is not so important.



Dot placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.



Like dot strip plot, good for displaying the data in a tabular form when best value highlighted individual values.



Summarise multiple distributions by showing the median (centre) and range (data)



Similar to a box plot, more effective with complex distributions (data that cannot be summarised with averages).



A standard way of showing the age breakdown of a population distribution, effectively back histograms.



A good way of showing how unequal a distribution is; it is always cumulative frequency, x axis is always a measure.



For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 data sets.



Use to emphasize individual points in a distribution. Points can be sized to an additional variable.

Distribution

Show values in a dataset and how they occur. The shape ('or skew') of distributions can be memorable when highlighting the lack of uniformity in equality in the data.

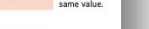
Example FT uses
Income distribution, population (age/gender) distribution, revealing inequality



The standard way to show a statistical distribution – key gaps between categories small to highlight 'shape' of the data.



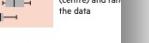
A simple way of showing the change (range/mean) of data across multiple categories.



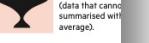
Good for showing individual values distribution, can be problem when too many dots have same value.



Like dot strip plot, good for displaying the data in a tabular form when best value highlighted individual values.



Summarise multiple distributions by showing the median (centre) and range (data)



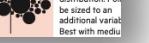
Similar to a box plot, more effective with complex distributions (data that cannot be summarised with averages).



A standard way of showing the age breakdown of a population distribution, effectively back histograms.



A good way of showing how unequal a distribution is; it is always cumulative frequency, x axis is always a measure.



Use to emphasize individual points in a distribution. Points can be sized to an additional variable.

What do you want to show?

Here you can find a list of charts categorised by their data visualization functions or by what you want a chart to communicate to an audience. While the allocation of each chart into specific functions isn't a perfect system, it still works as a useful guide for selecting chart based on your analysis or communication needs.

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

Comparisons

Proportions

Relationships

Hierarchy

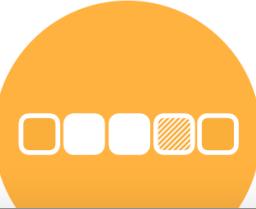
We'll go over some more chart types next week

Concepts

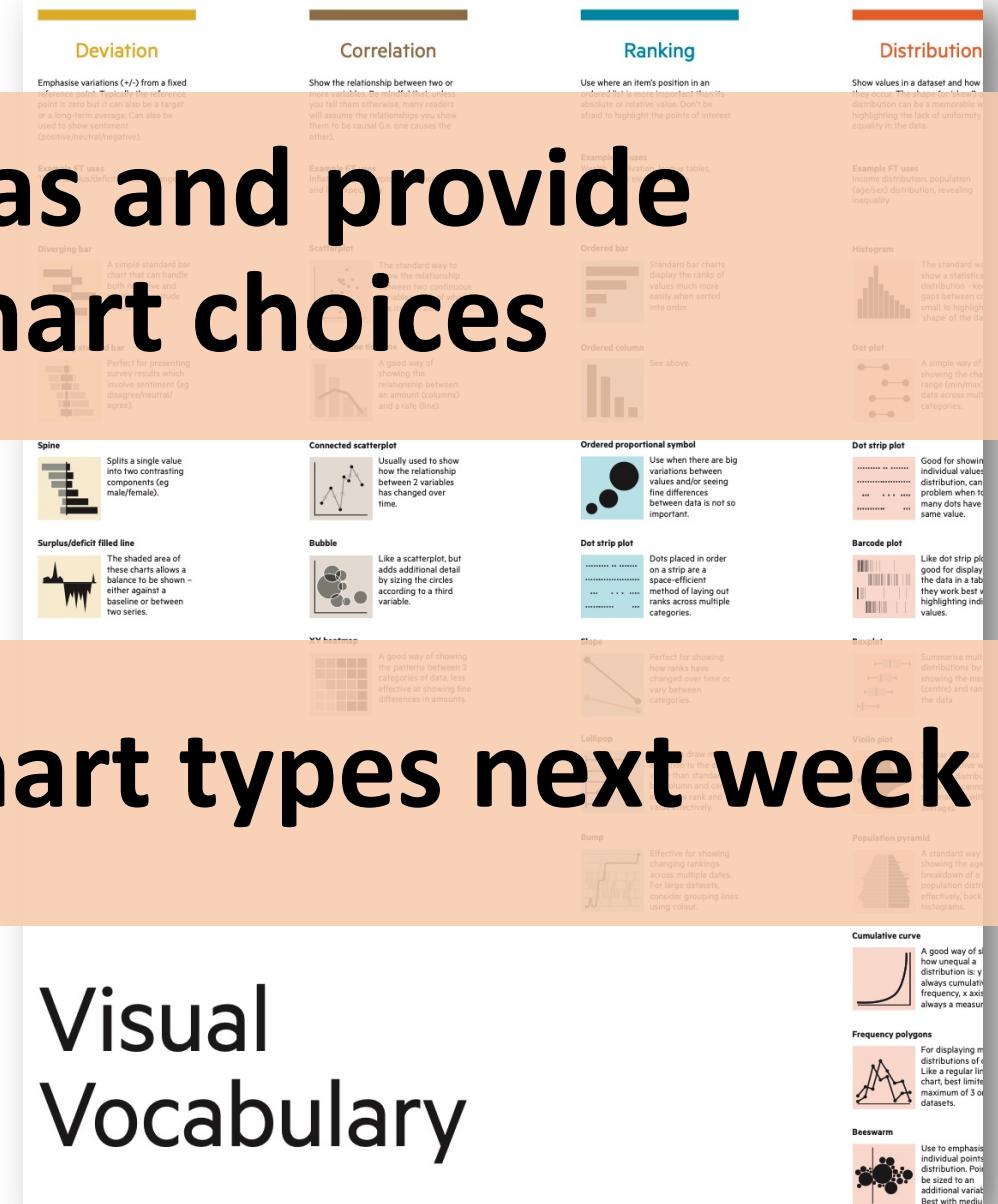
Location

Part-to-a-whole

Distribution



Visual Vocabulary



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 structure

Marks: Graphical primitives that encode data

Channels: Properties of mark appearance that map to data 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

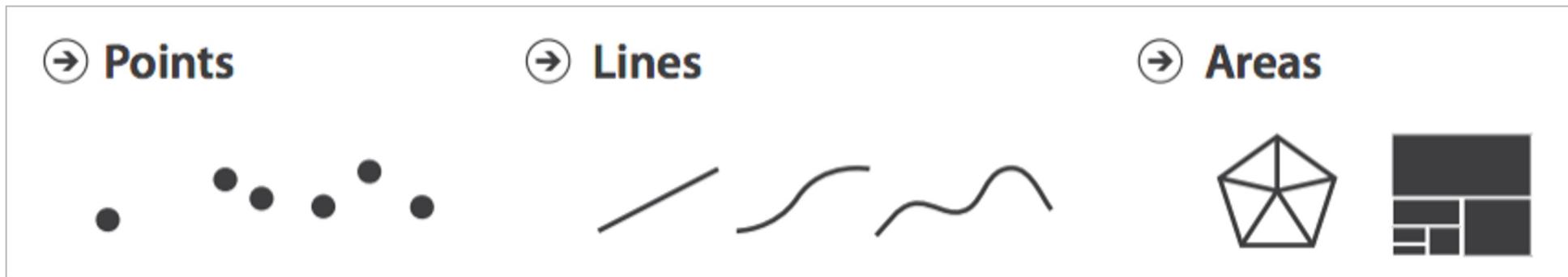
For a complete list of these encodings, see the [Encodings](#) section of the documentation.

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

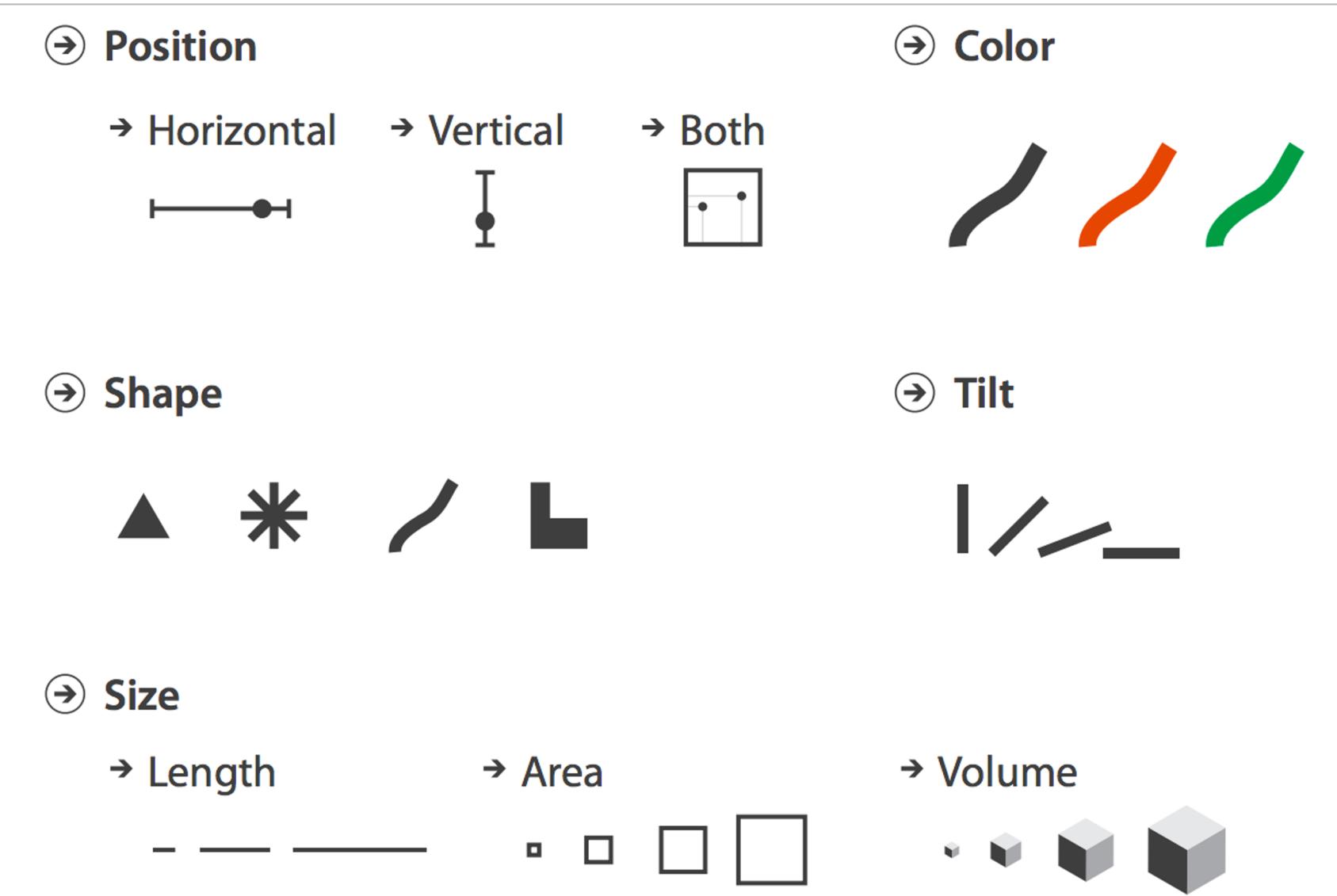
```
alt.Chart(movies_genre).mark_point().encode(
    y='AvgRating', size="Watches"
```

Marks, Channels, & Encoding

Marks: Graphical primitives that encode items / entities



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



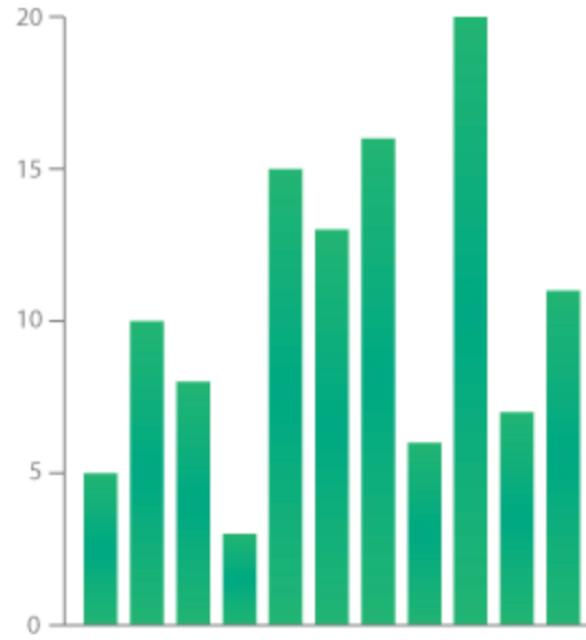
What are the marks & channels of...



Pie Chart

Marks: Wedge

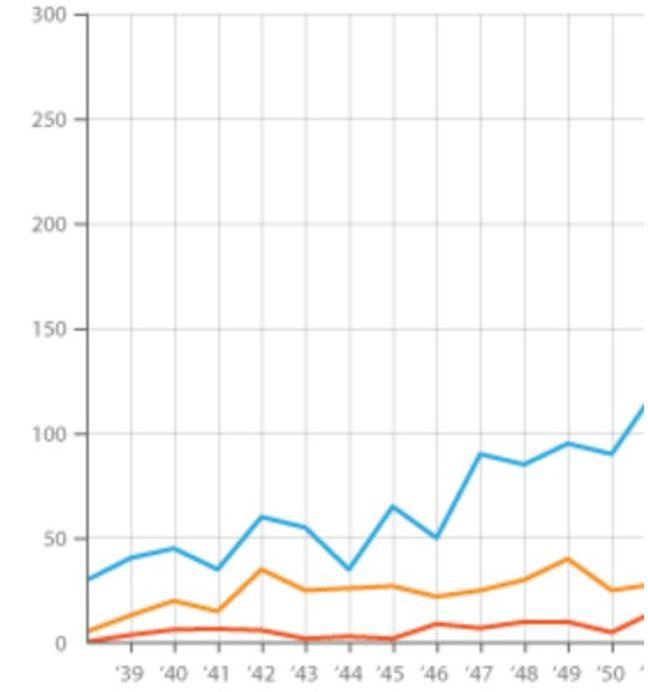
Channels: Color, Angle (not Area, not Size!)



Bar Chart

Marks: Rectangle

Channels: x-position, Length



Line Chart

Marks: Line (Path)

Channels: x,y-positions, Color

Expressiveness & Effectiveness

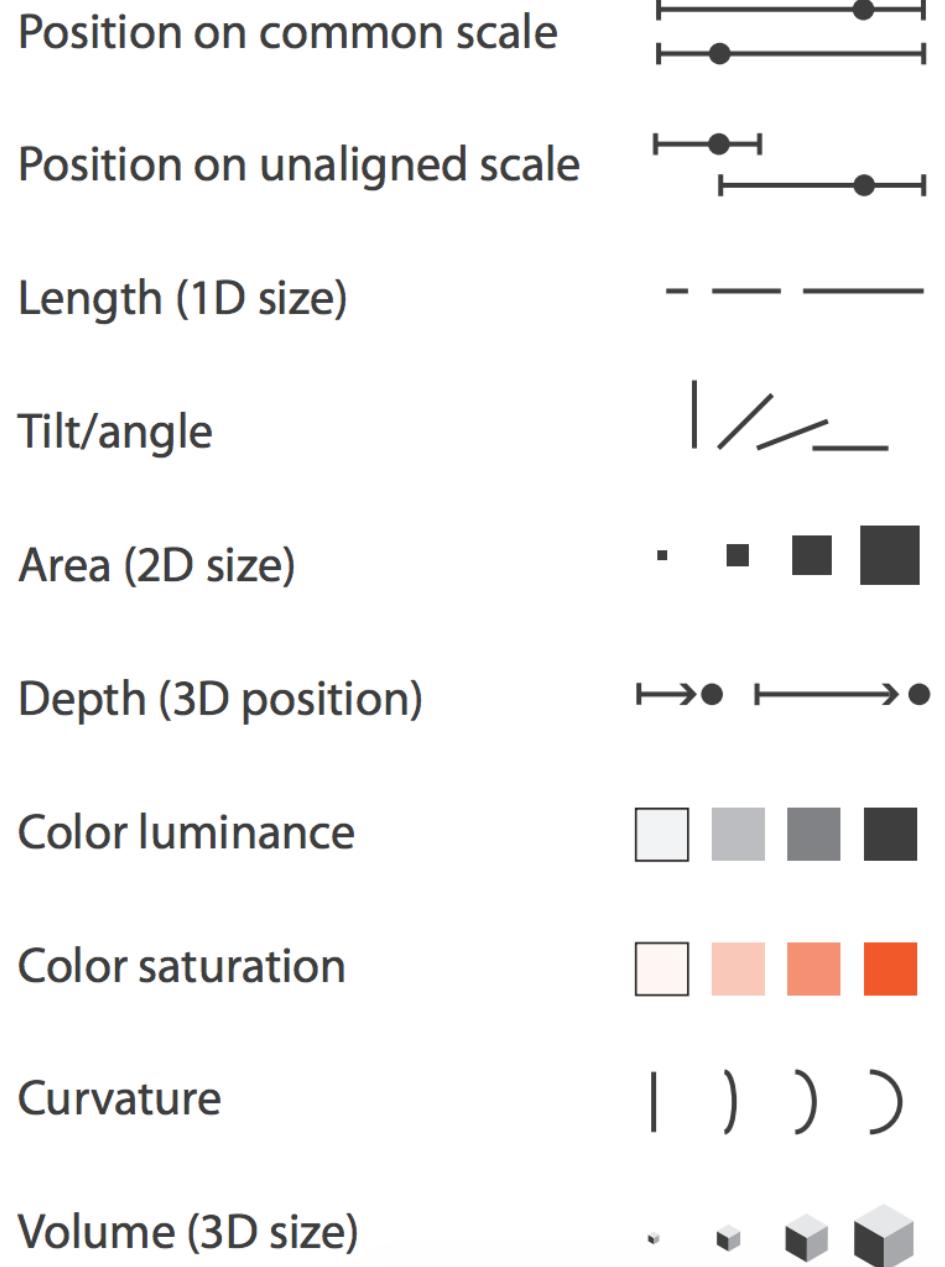
Expressiveness Principle: Encoding should express all of, and only, the information in the data

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

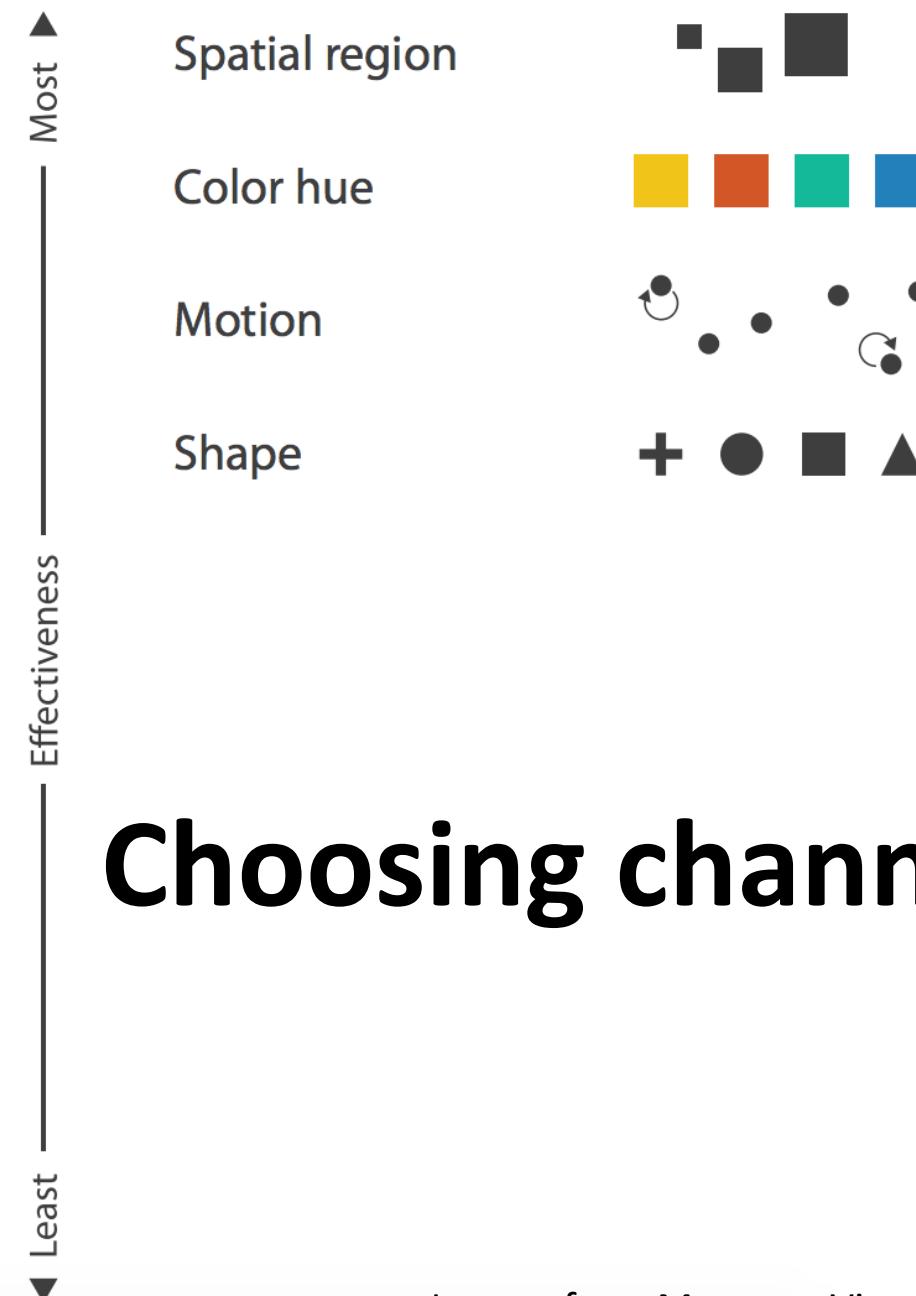
Effectiveness Principle: The more important the data/attribute, the more **salient** the encoding should be

- Important things should be noticeable

→ Magnitude Channels: Ordered Attributes



→ Identity Channels: Categorical Attributes



Choosing channels...

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



→ Identity Channels: Categorical Attributes

Spatial region



Motion



Shape



▲ Most
— Effectiveness
↓ Same

Which has order? Which doesn't?

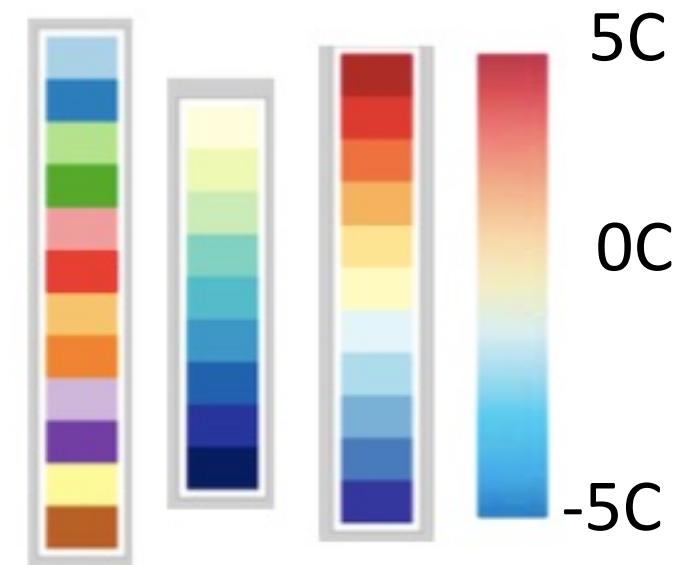


Vary the appropriate dimension depending on your data type

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!



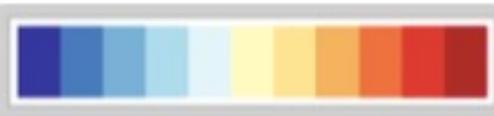
Color map axes: Design color map to match the attribute(s) you are encoding



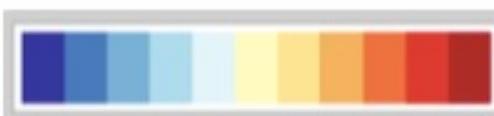
Categorical vs. Ordered



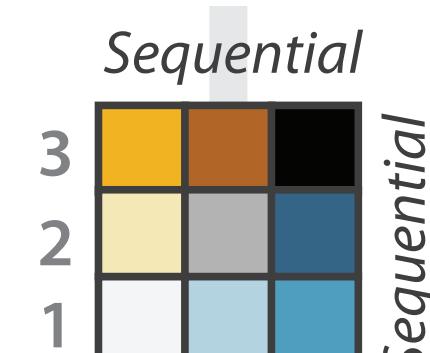
Diverging vs. Sequential



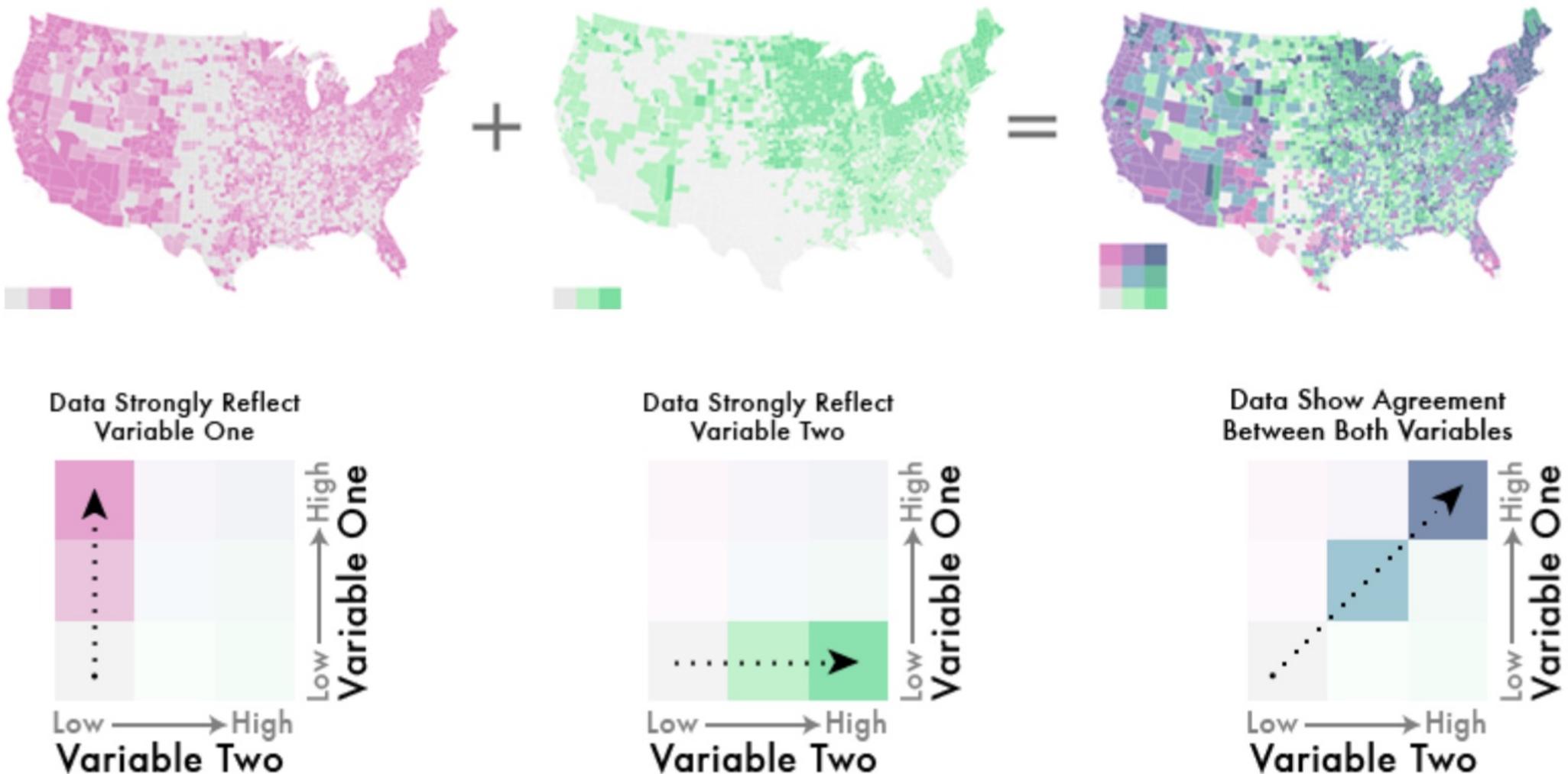
Segmented vs. Continuous



Univariate vs. Bivariate



Bivariate Example

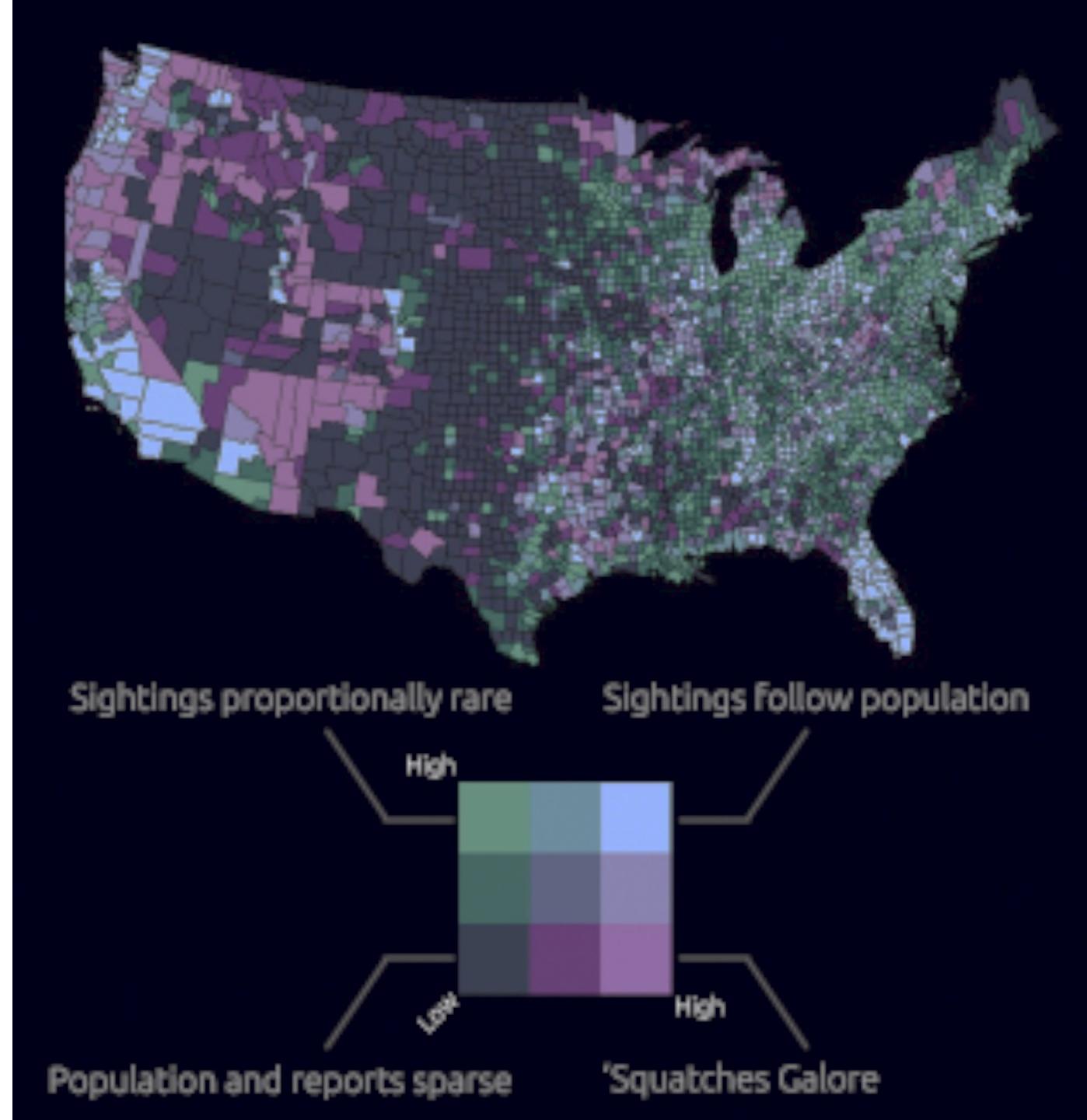


Bivariate Example

Population
&
Sasquatch (BigFoot) Sightings

...not so easy to read.

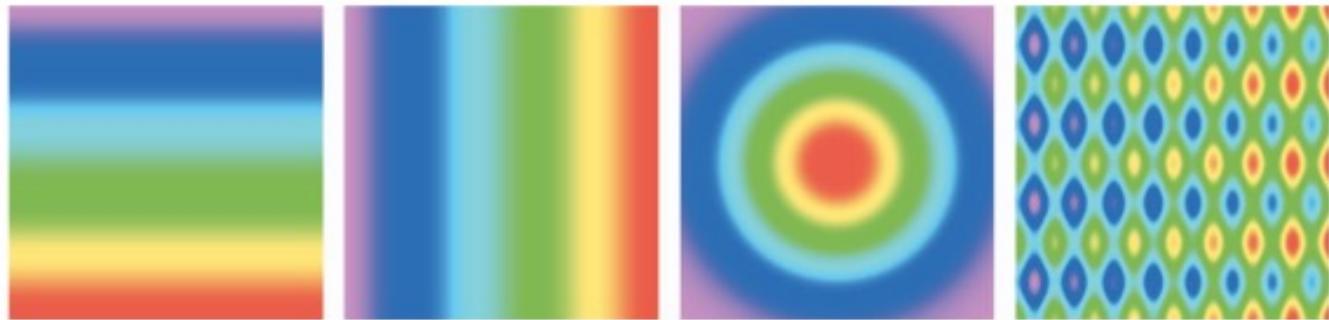
Consider also using
hue/lightness,
hue/thickness of border,
hue + mark&channel...



The Dangers of Rainbows

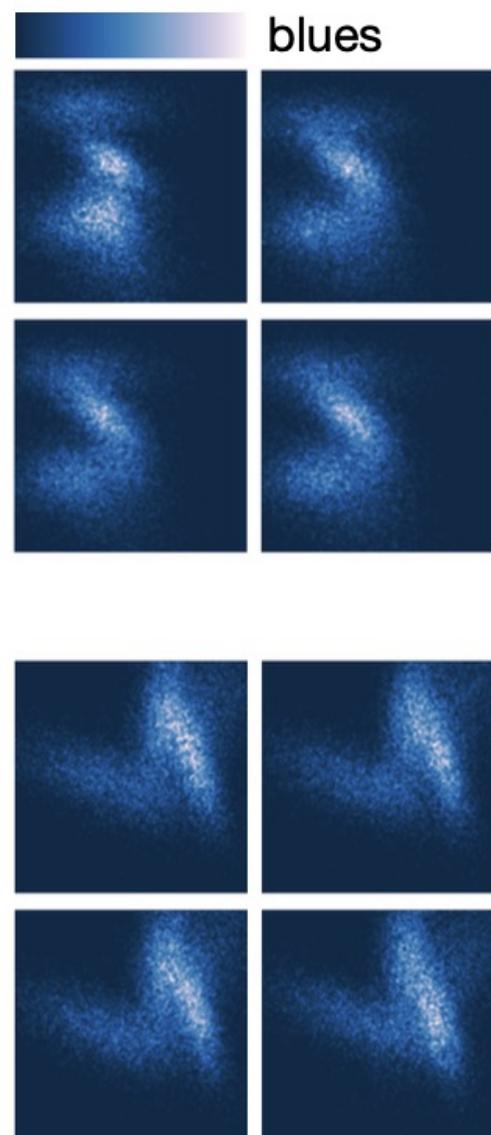
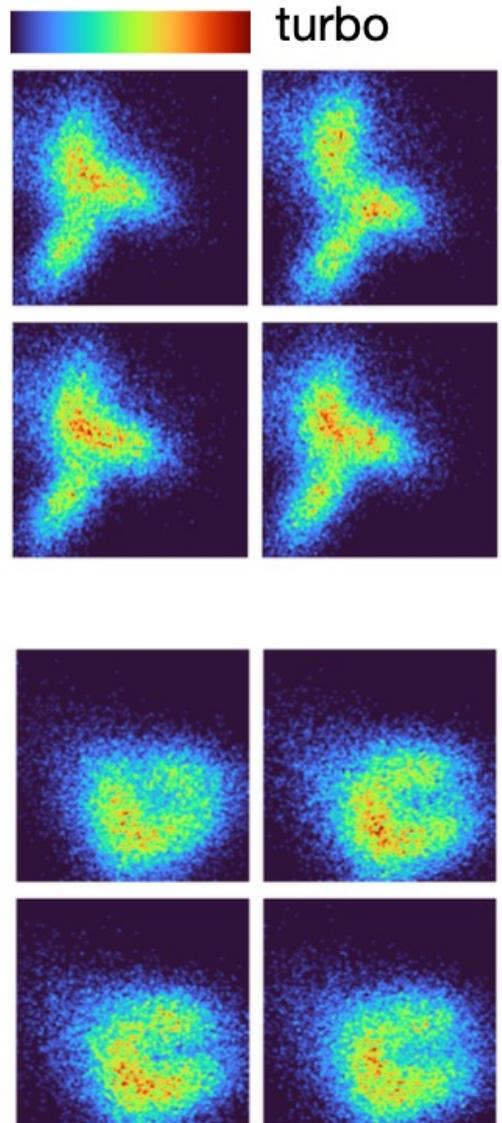
The common rainbow color map is often a poor choice due to:

- 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)

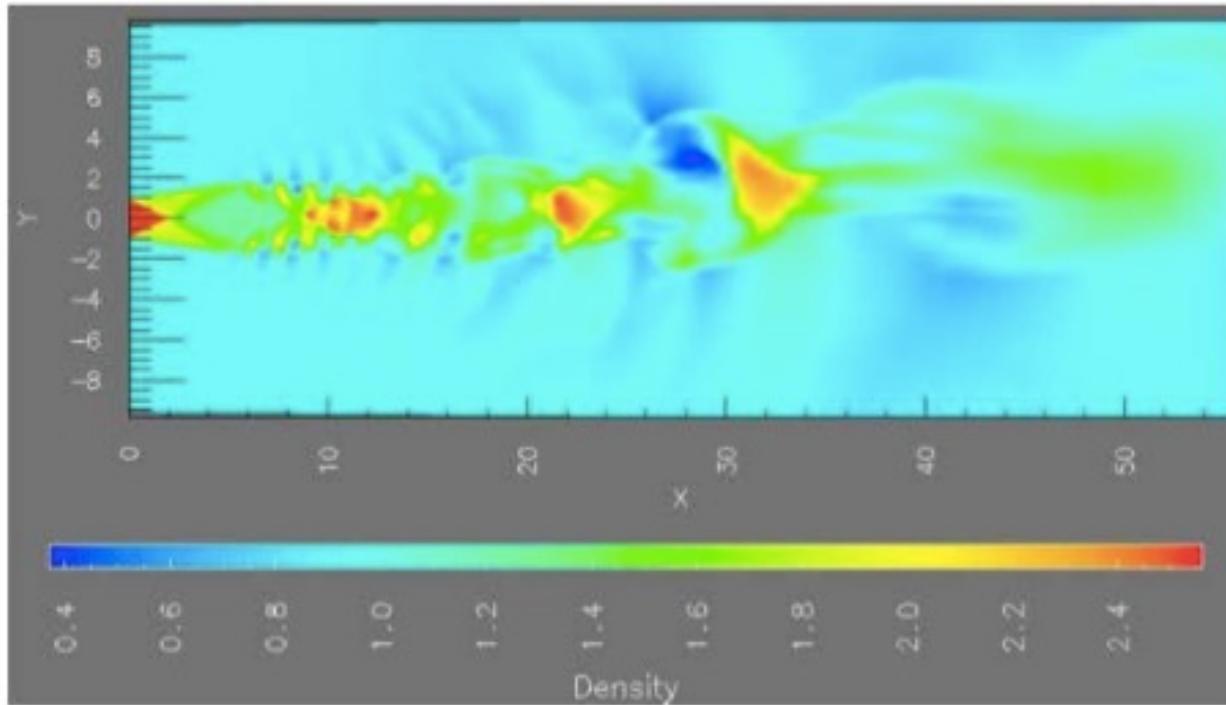
The Dangers of Dismissing Rainbows



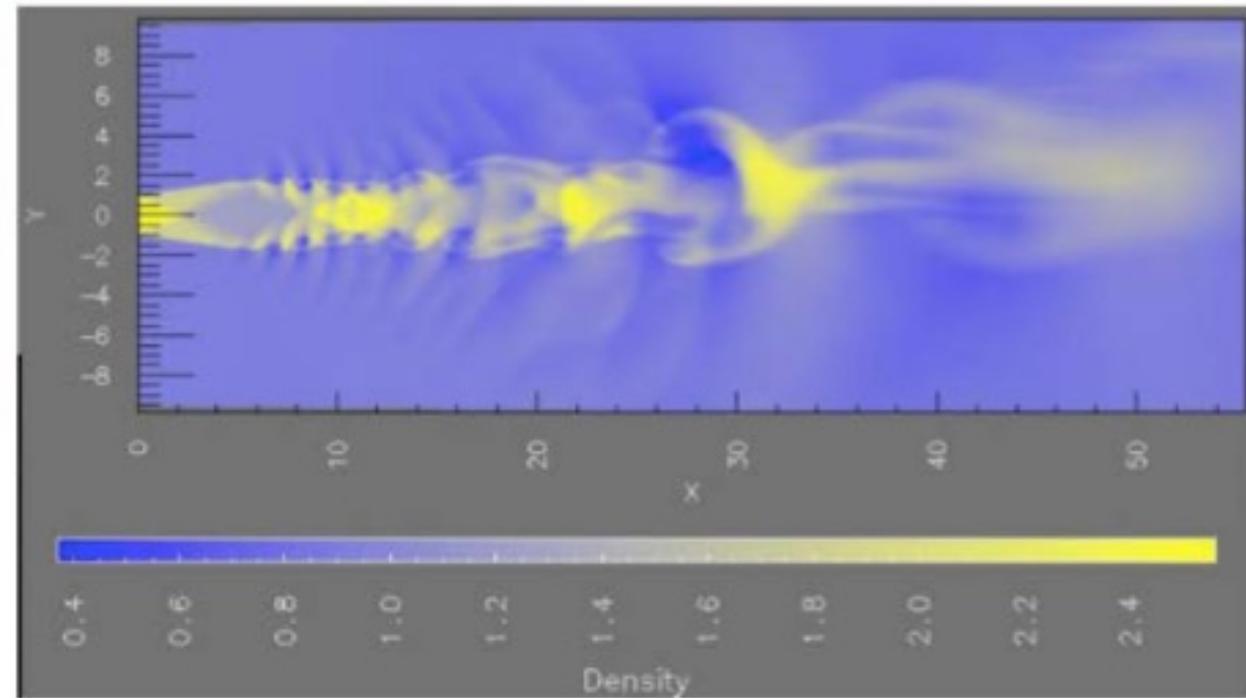
Some rainbow color maps may be helpful even in continuous tasks.

- Some tasks may benefit from the ability to name colors
- Detection of false features may not be as pervasive as we thought

Different color maps are good for different insights



(a)



(b)

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].

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

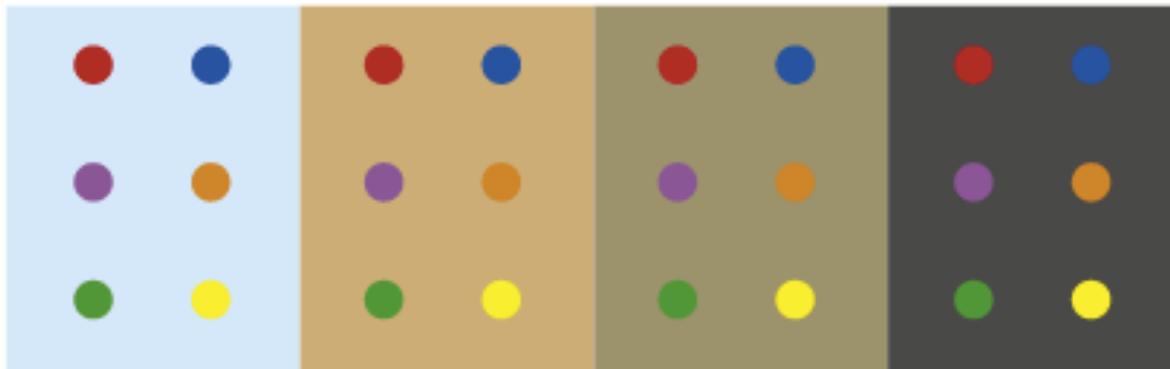
May assign colors hierarchically

- Pick colors we can differentiate by name, especially in collaborative contexts

e.g., dark blue & light blue, dark red & light red, etc.



Color Perception is Relative

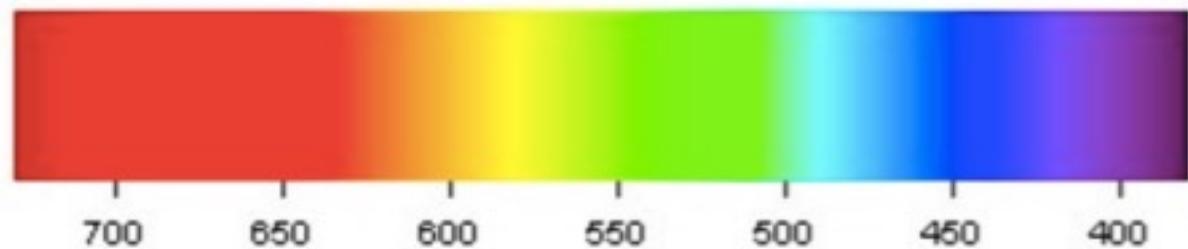


What happens when you take off tinted goggles?

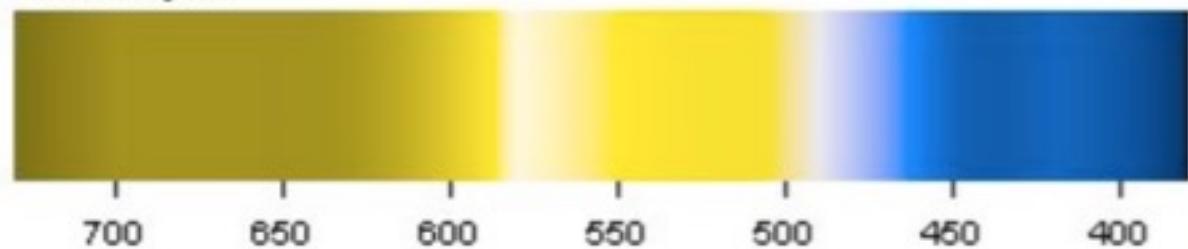
- To avoid contrast issues with background, surround points with white or black lines.
- Make sure colors do not have same luminance as background

Color Vision Deficiencies affect ~7% of the population

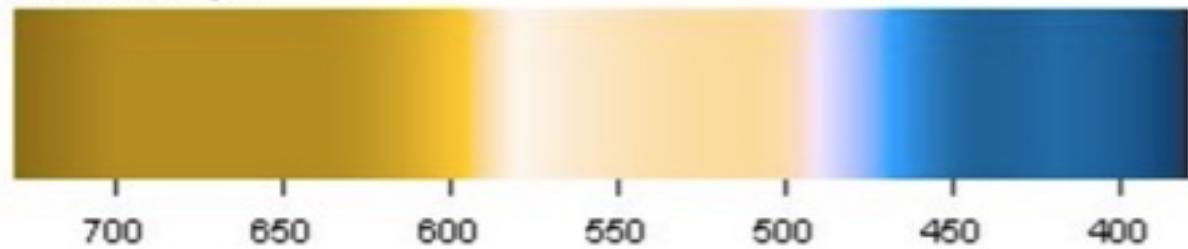
Normal



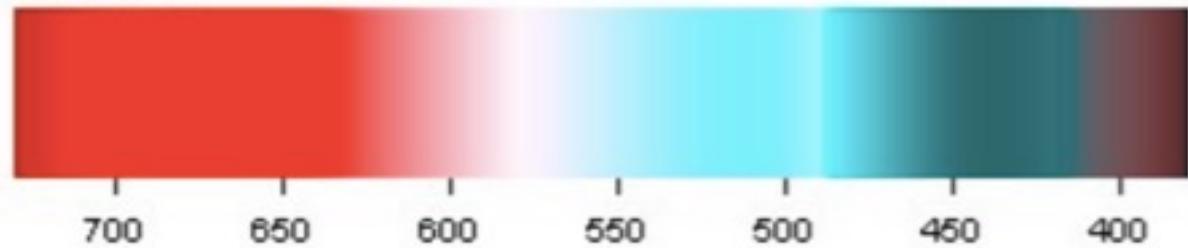
Protanopia



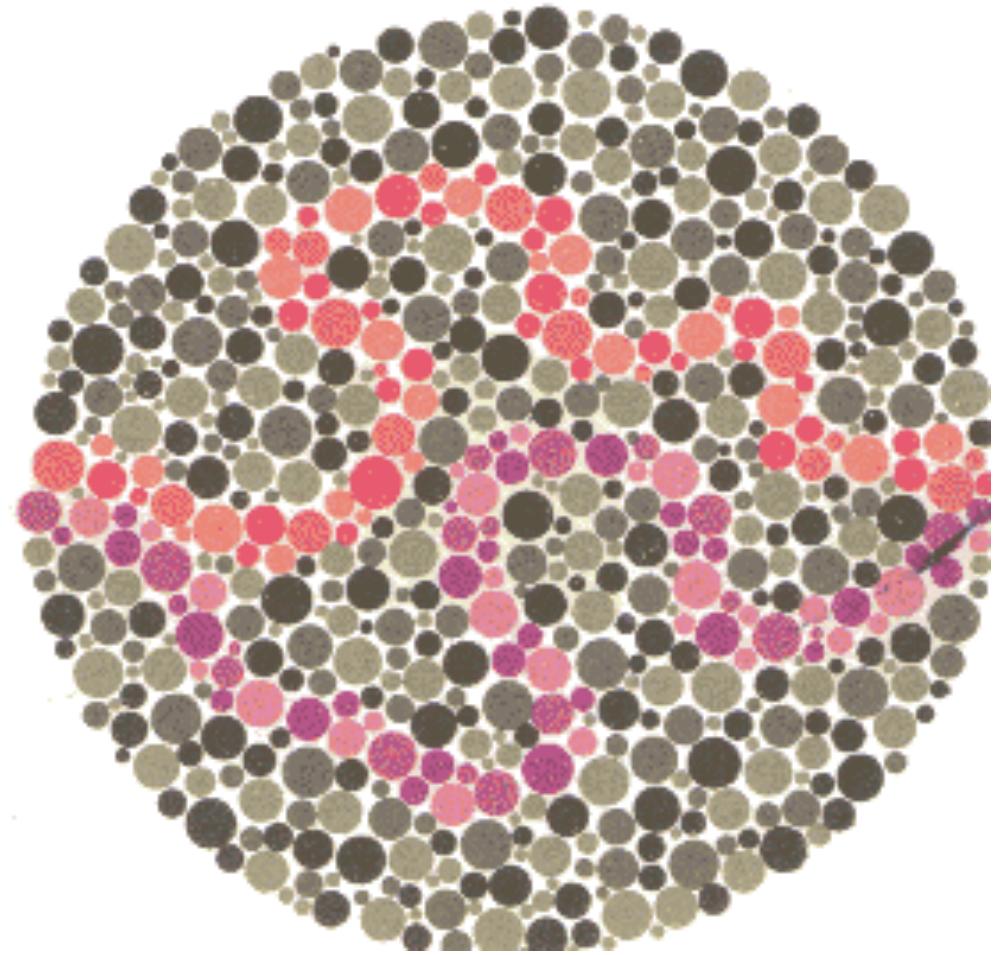
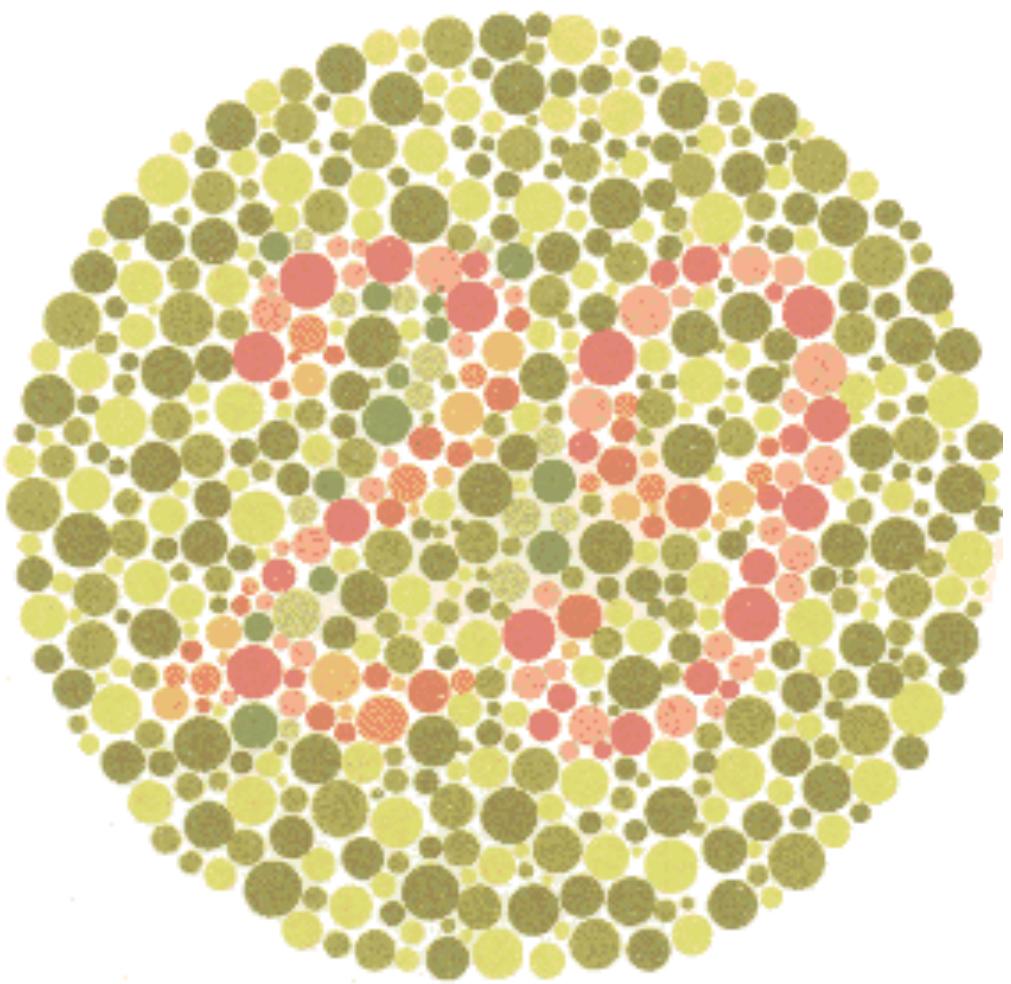
Deuteranopia



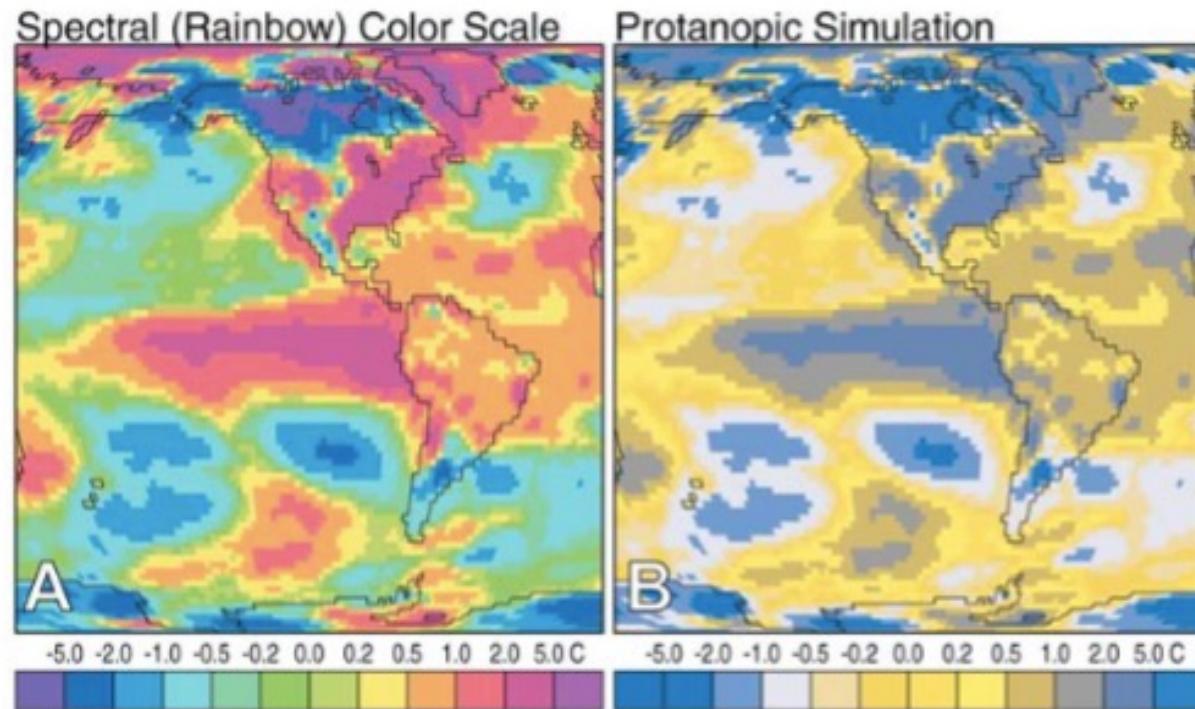
Tritanopia



Ishihara Plates are used to test for Color Vision Deficiencies



Rainbow Colormap & Color Vision Deficiency



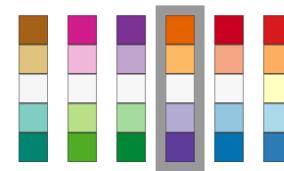
Guideline: “Get it right in black and white”

Number of data classes: 7

Nature of your data:

sequential diverging qualitative

Pick a color scheme:



Only show:

- colorblind safe
- print friendly
- photocopy safe

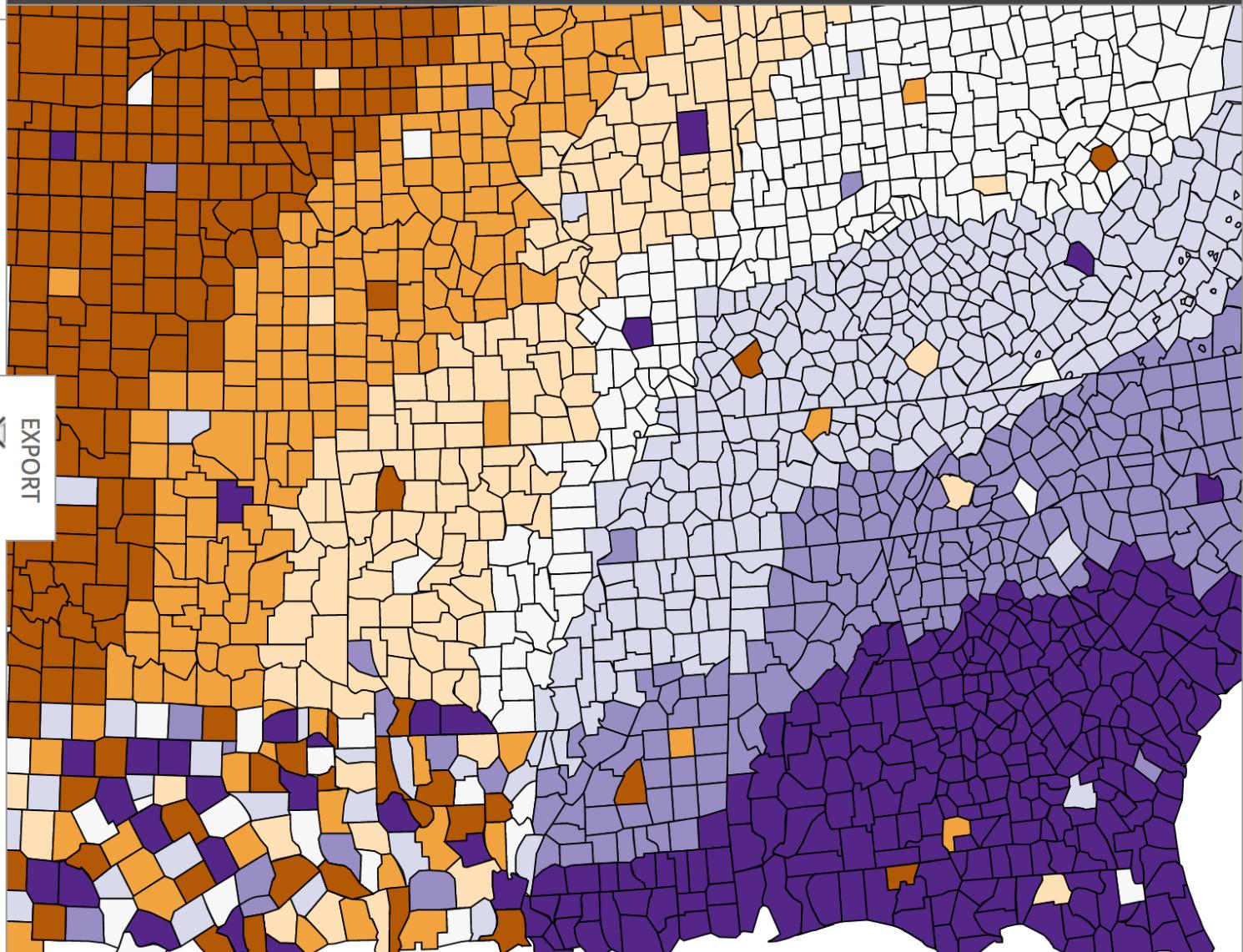
Context:

- roads
- cities
- borders

Background:

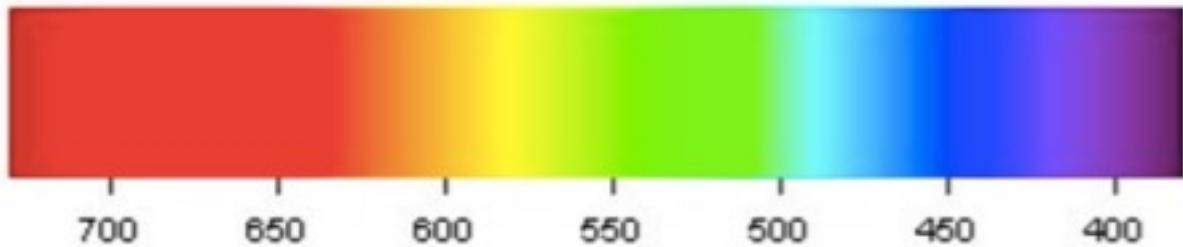
- solid color
- terrain

color transparency

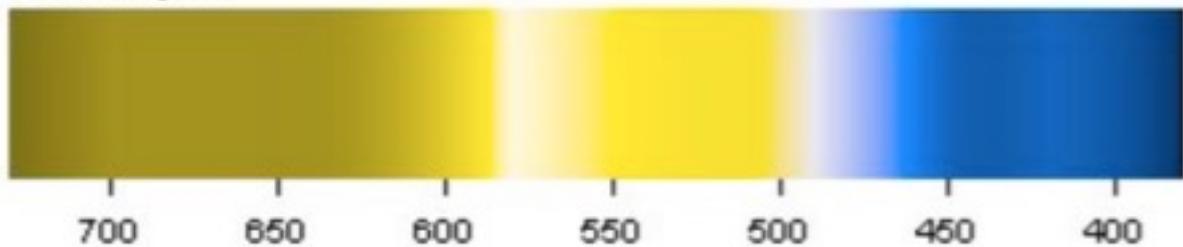


Color Vision Deficiencies affect ~7% of the population

Normal



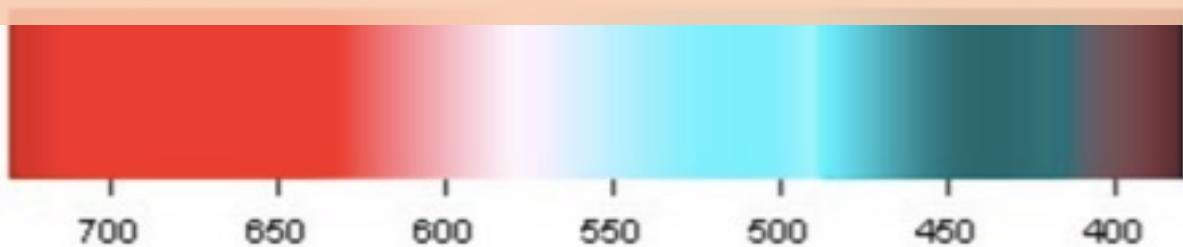
Protanopia



Deuteranopia



Tritanopia



...but which 7%?

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



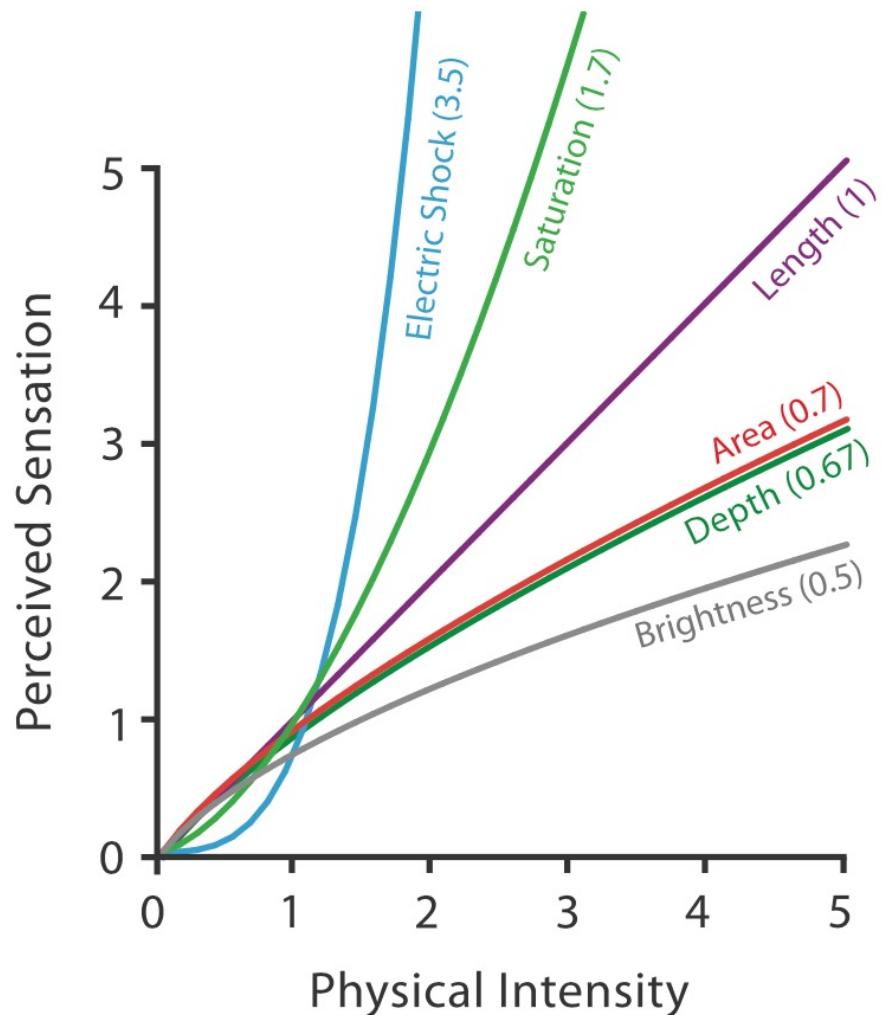
Shape



▲ Most Effective
— Effectiveness
— Same

Where do these rankings come from?

Stevens's Psychophysical Power Law: $S = I^n$



S = sensation
I = intensity

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

The general form of the law is

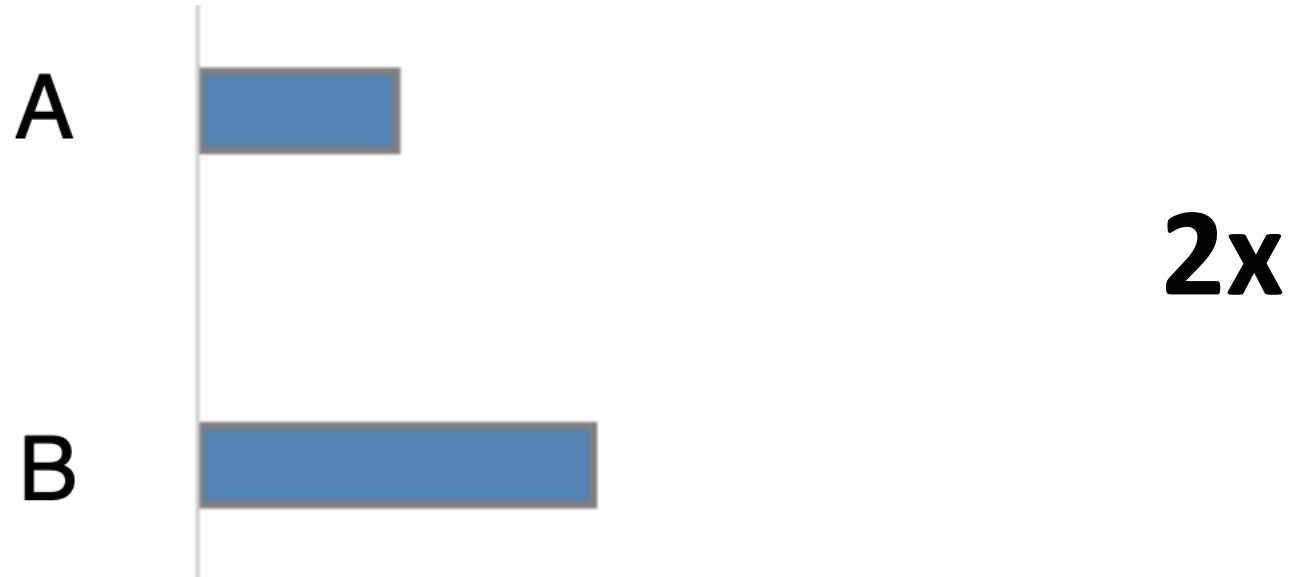
$$\psi(I) = kI^a,$$

where I is the intensity or strength of the stimulus in physical units (energy, weight, pressure, mixture proportions, etc.), $\psi(I)$ 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**

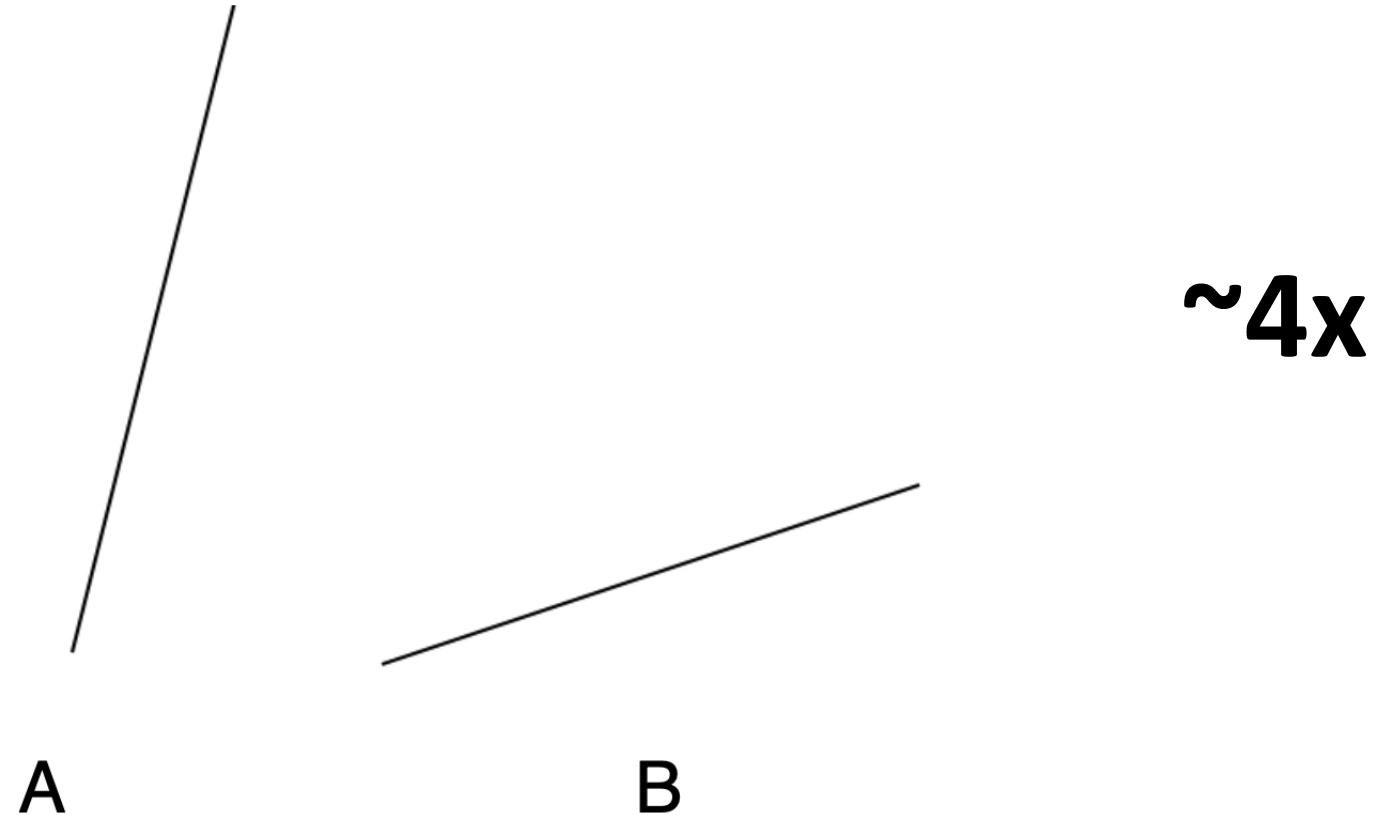
How Much Longer?



How Much Longer?



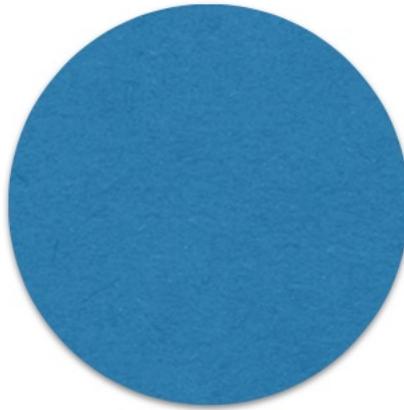
How Much Steeper?



How Much Larger?



A



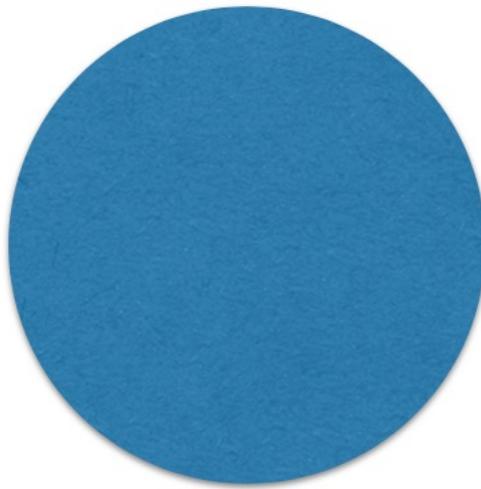
B

5x

How Much Larger?



A



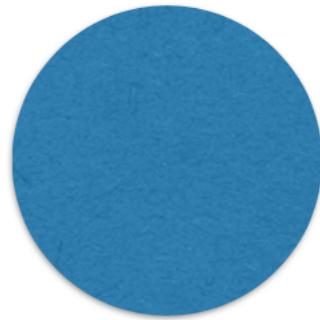
B

**4x area
2x diameter**

How Much Larger (by area)?



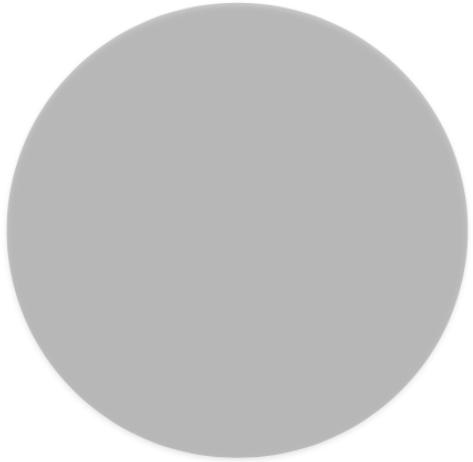
A



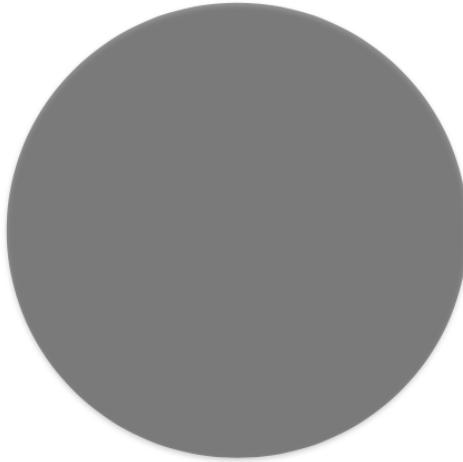
B

3x

How Much Darker?



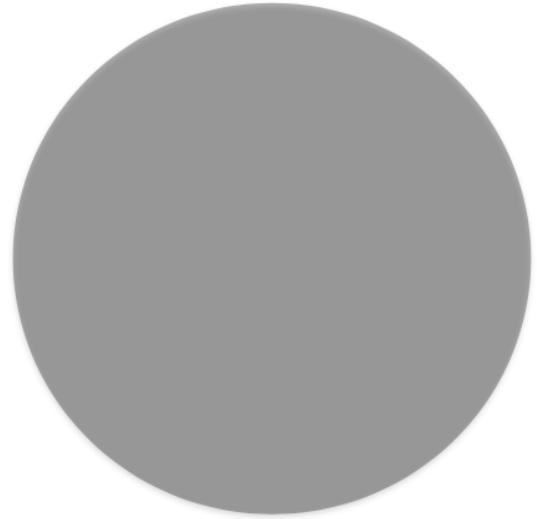
A



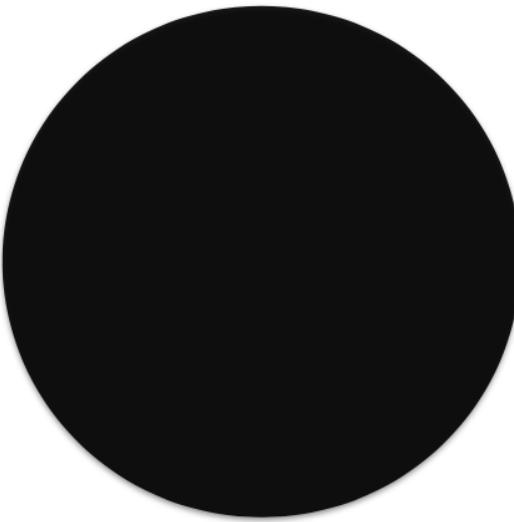
B

2x

How Much Darker?



A

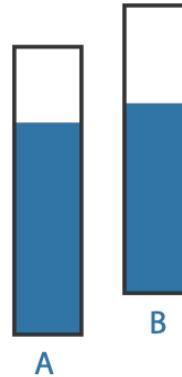


B

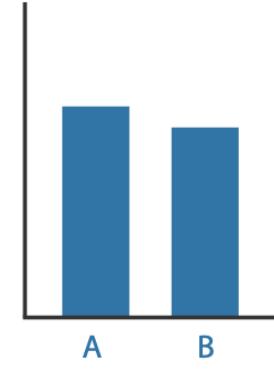
3x

Other factors affect accuracy too...

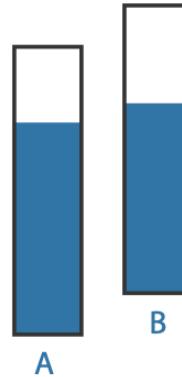
Alignment



Distractors

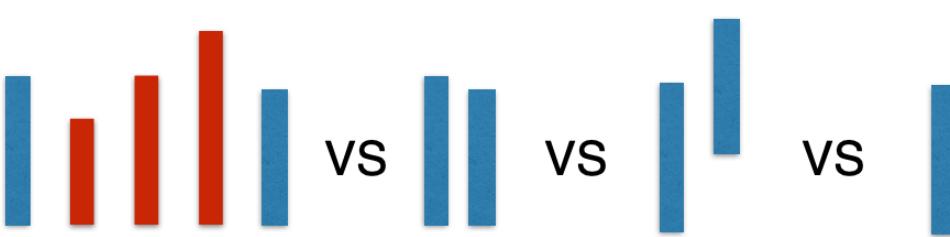


Distance



Common scale

...



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

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



▲ Most Effective
— Effectiveness
↓ Same

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

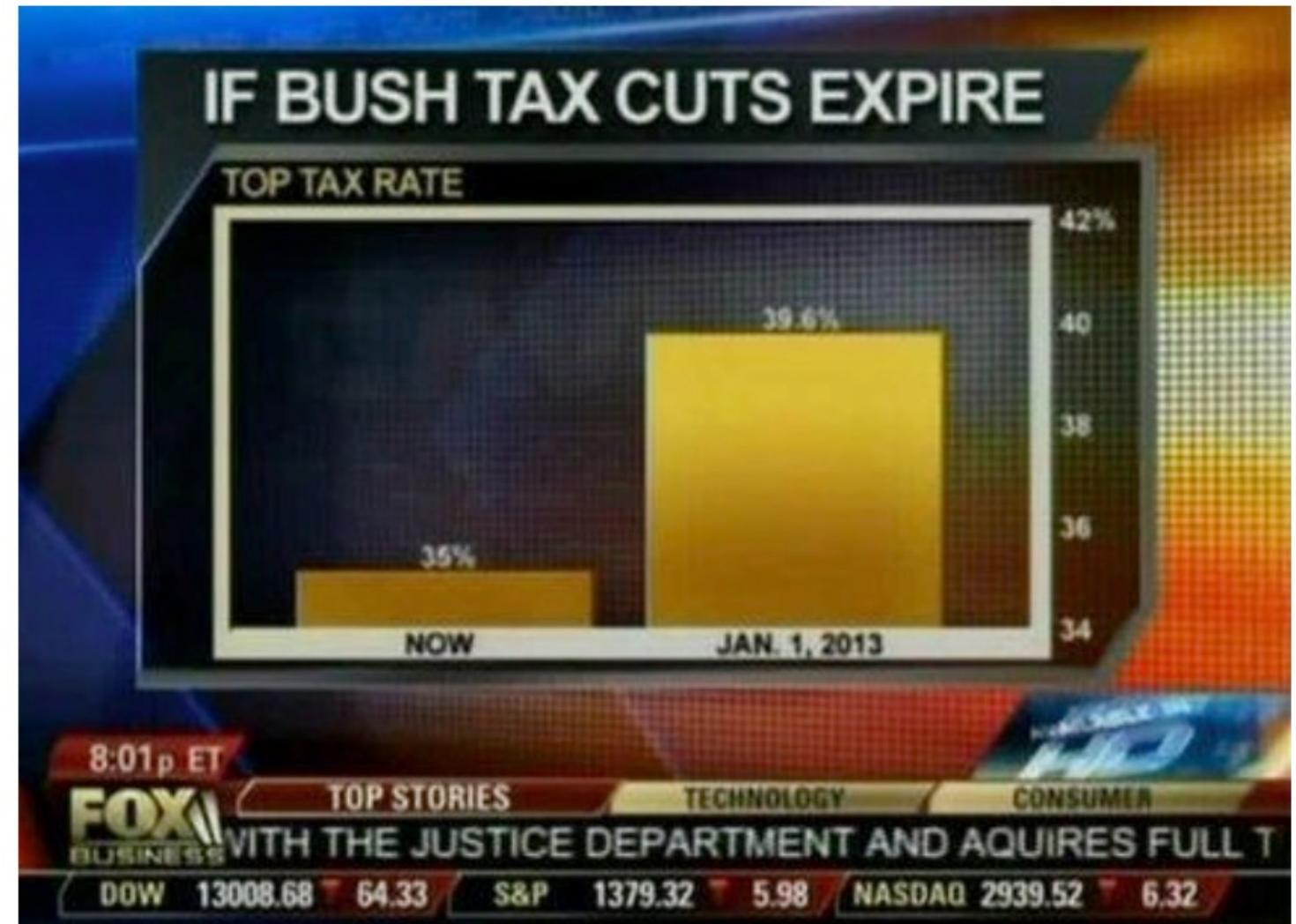
Size of the effect shown in graphic

Size of the effect in data

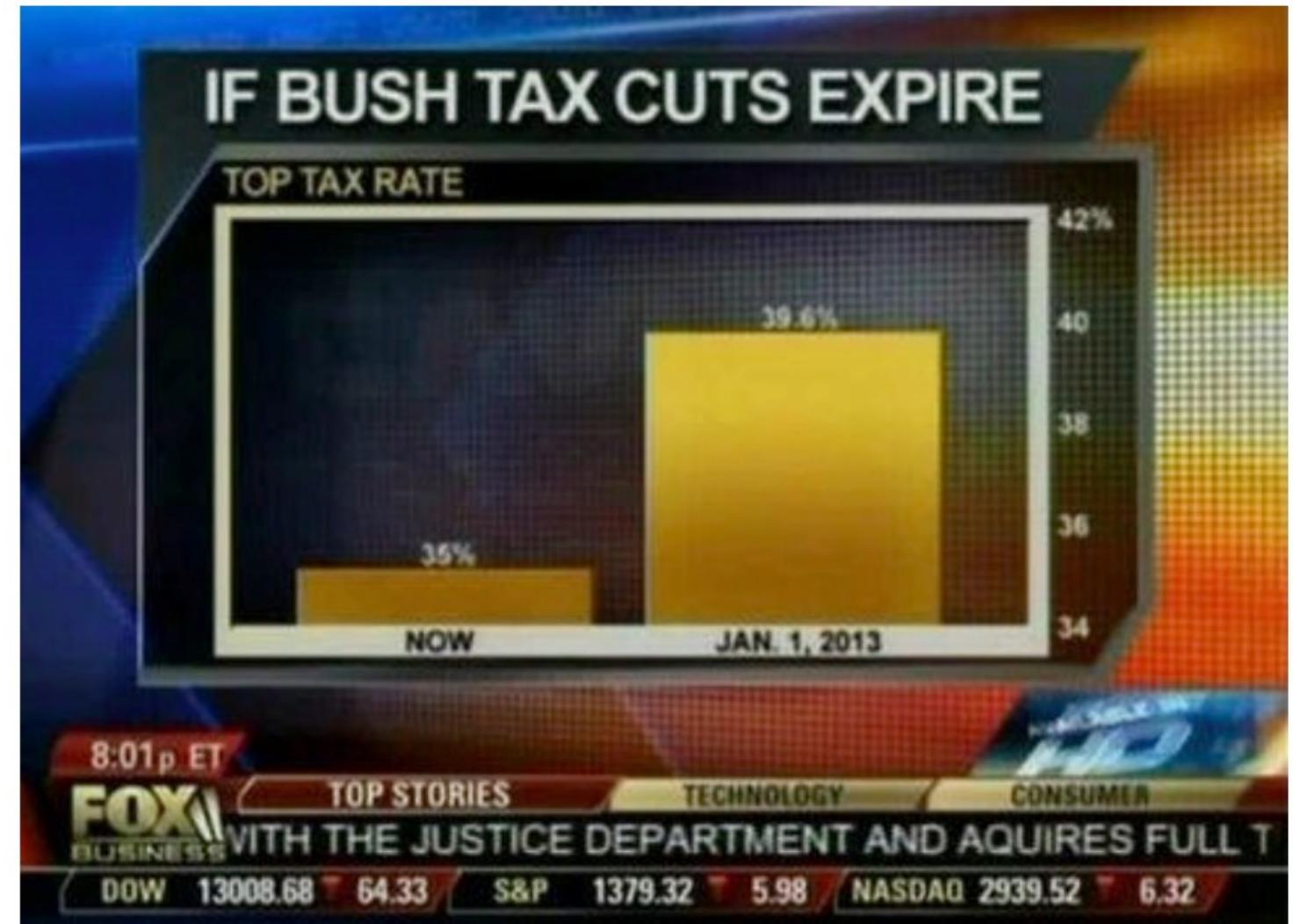
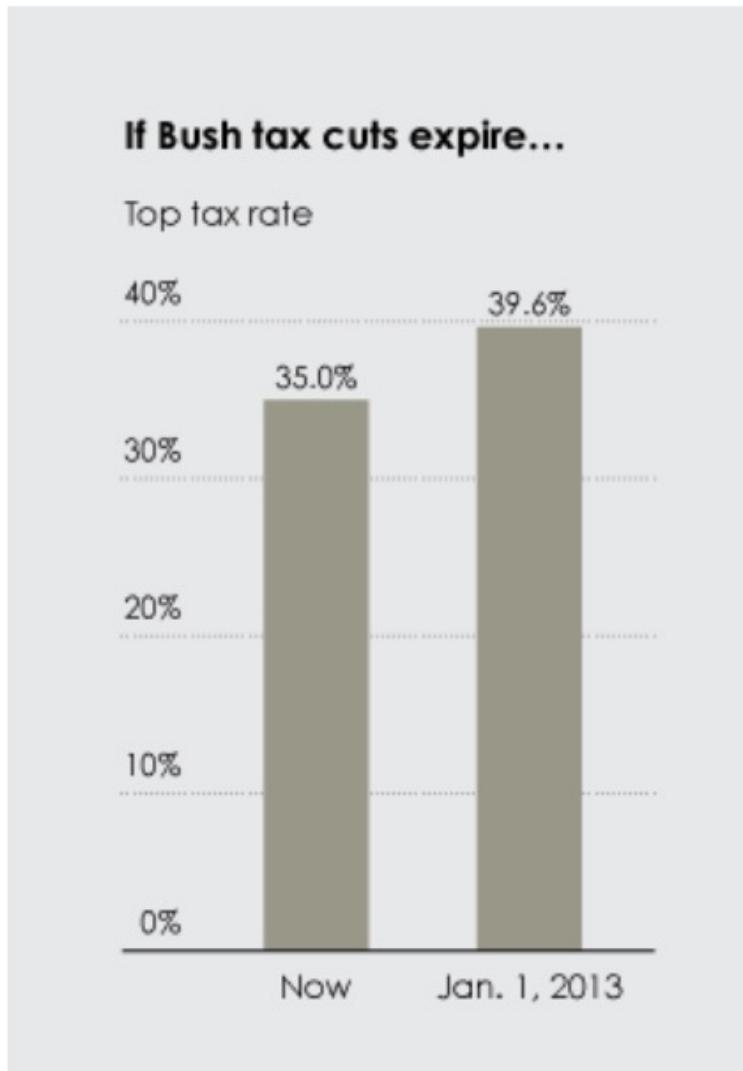
The Lie Factor – Graphical Integrity

Magnitude in
data must
correspond to
magnitude of
mark

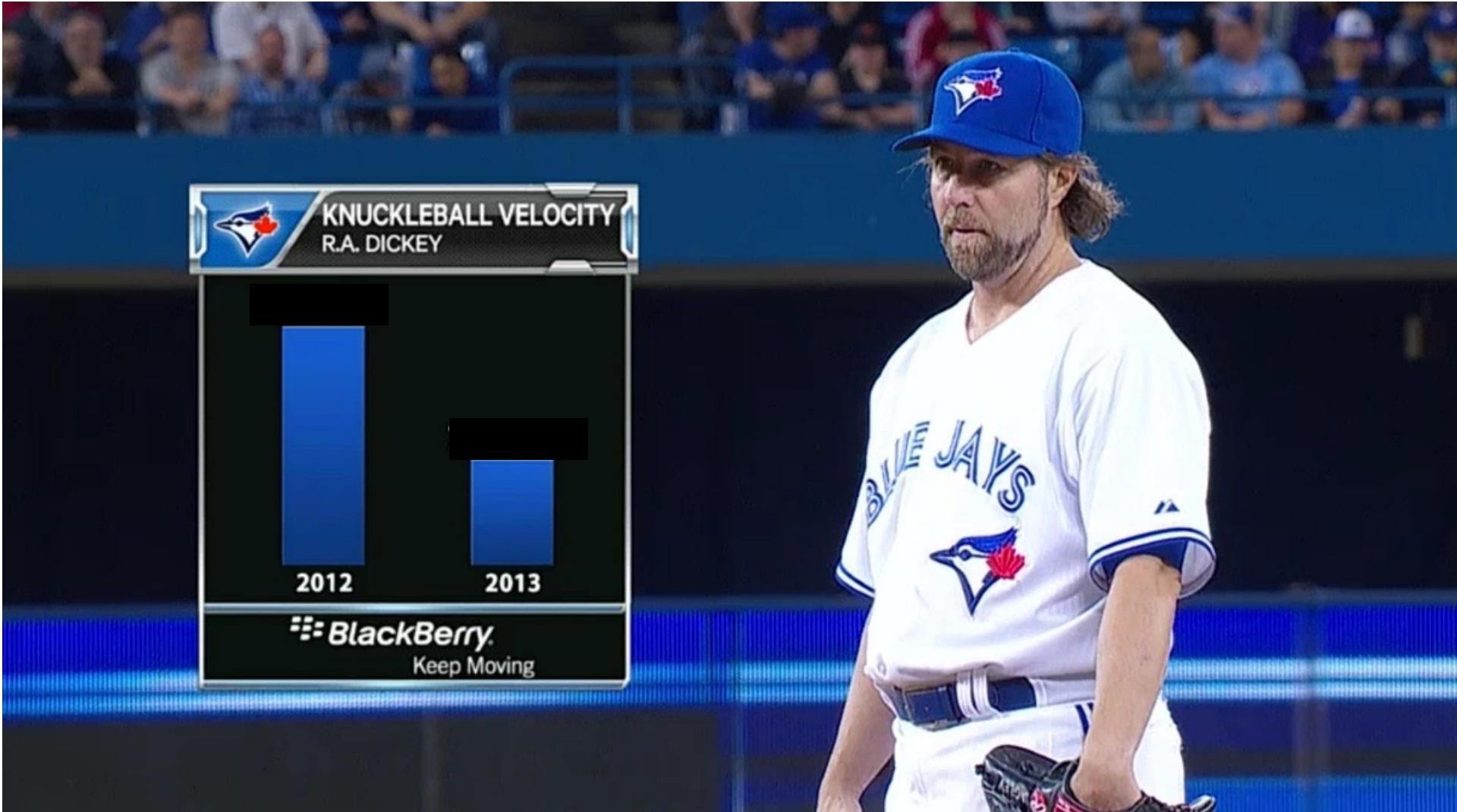
Effect in data: 1.14
Effect in graphic: 5
Lie factor: $5/1.14 = 4.38$



The Lie Factor – Graphical Integrity



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

Grafik
in echt



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

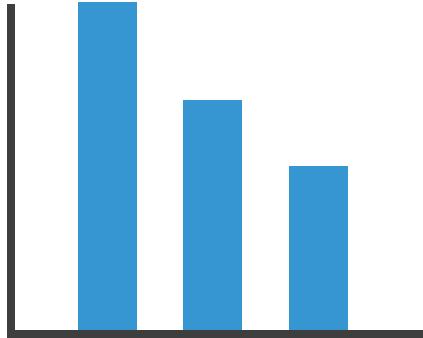


ACTUAL
ENROLLMENT

GOAL



Where should the y-axis start?

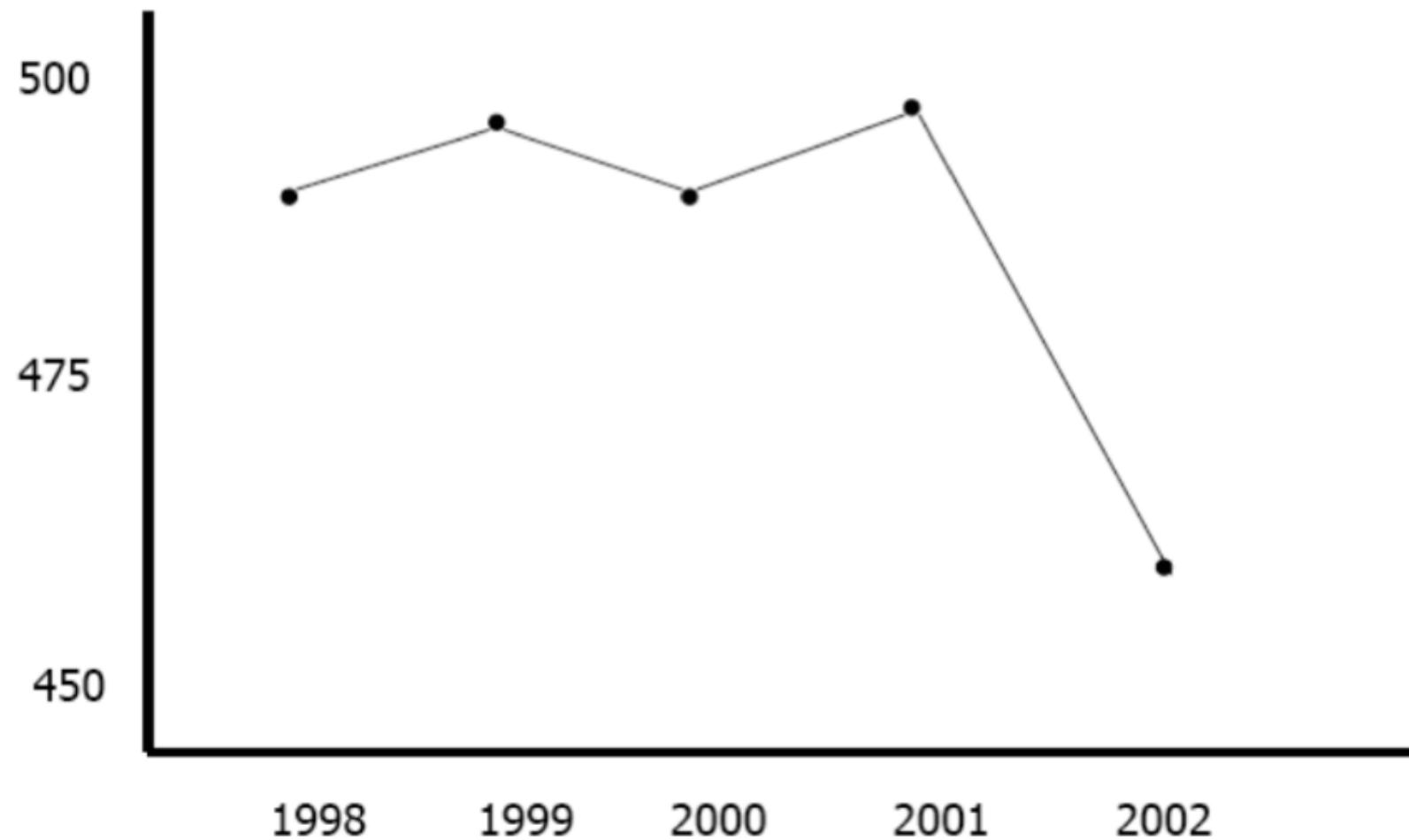


Length
(aligned)

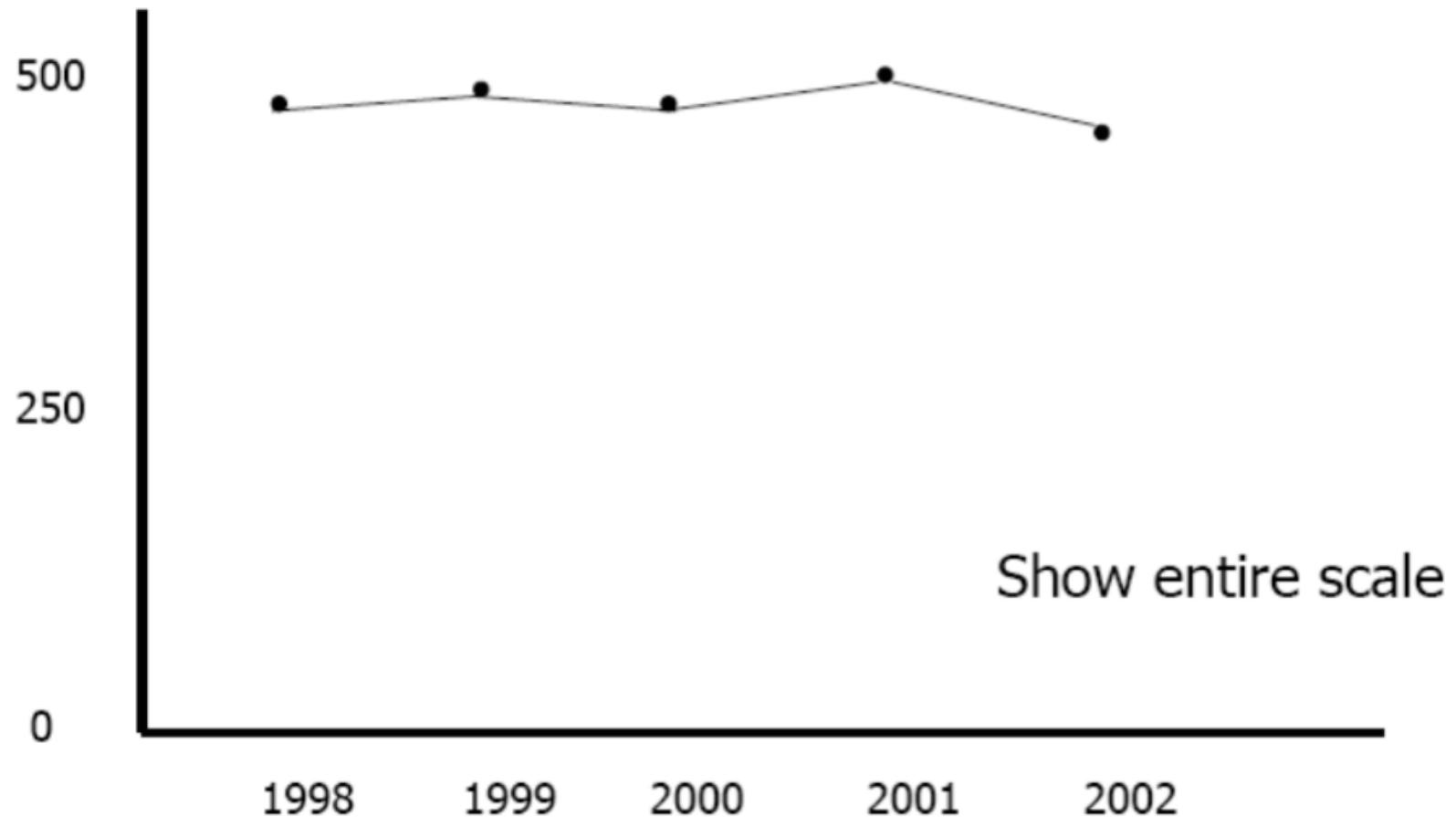
In **bar charts**, bar length is being compared. Therefore, starting the y-axis at an arbitrary position will work against the visualization task... often tricking the viewer.

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

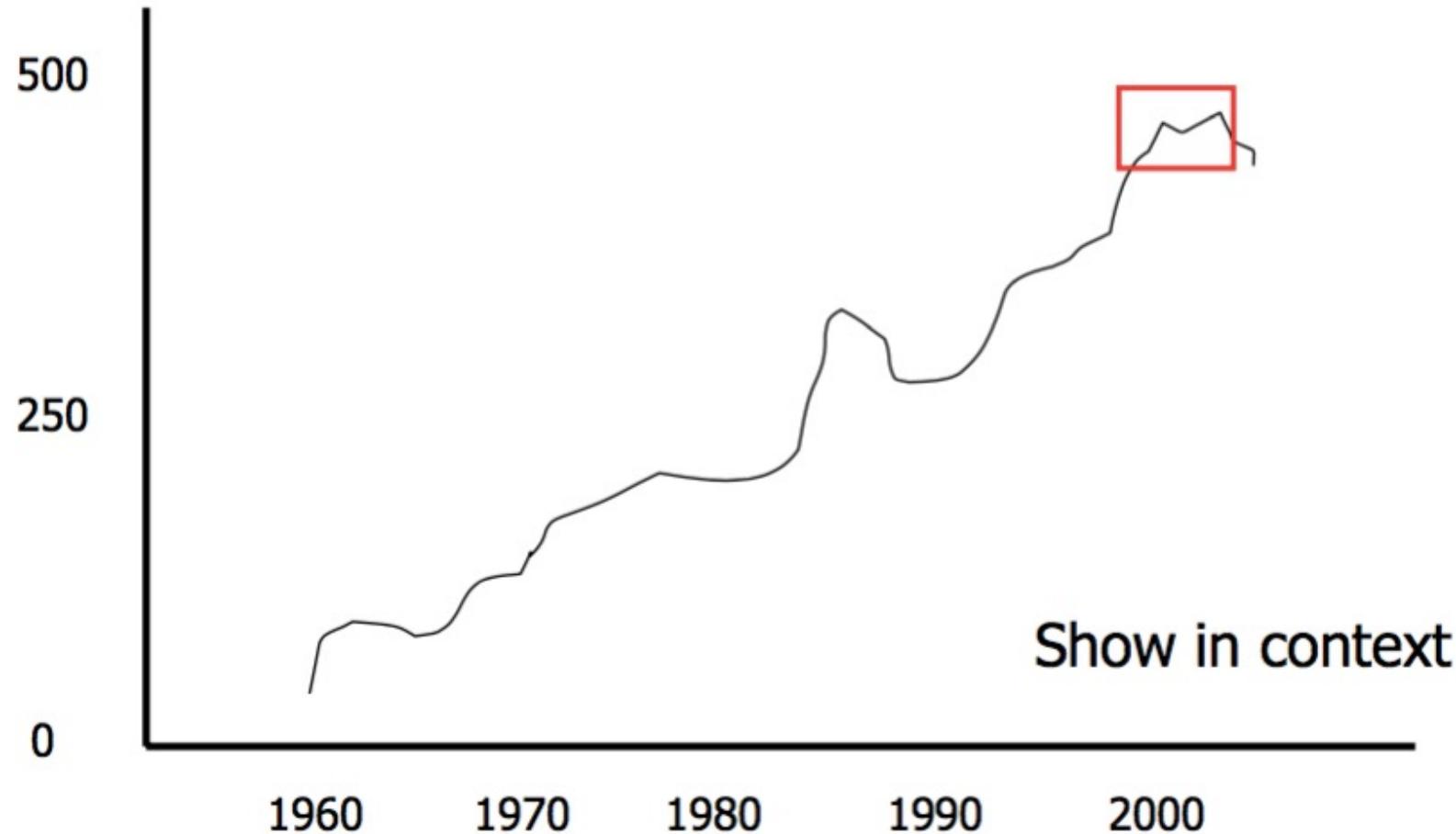
What's happening here?



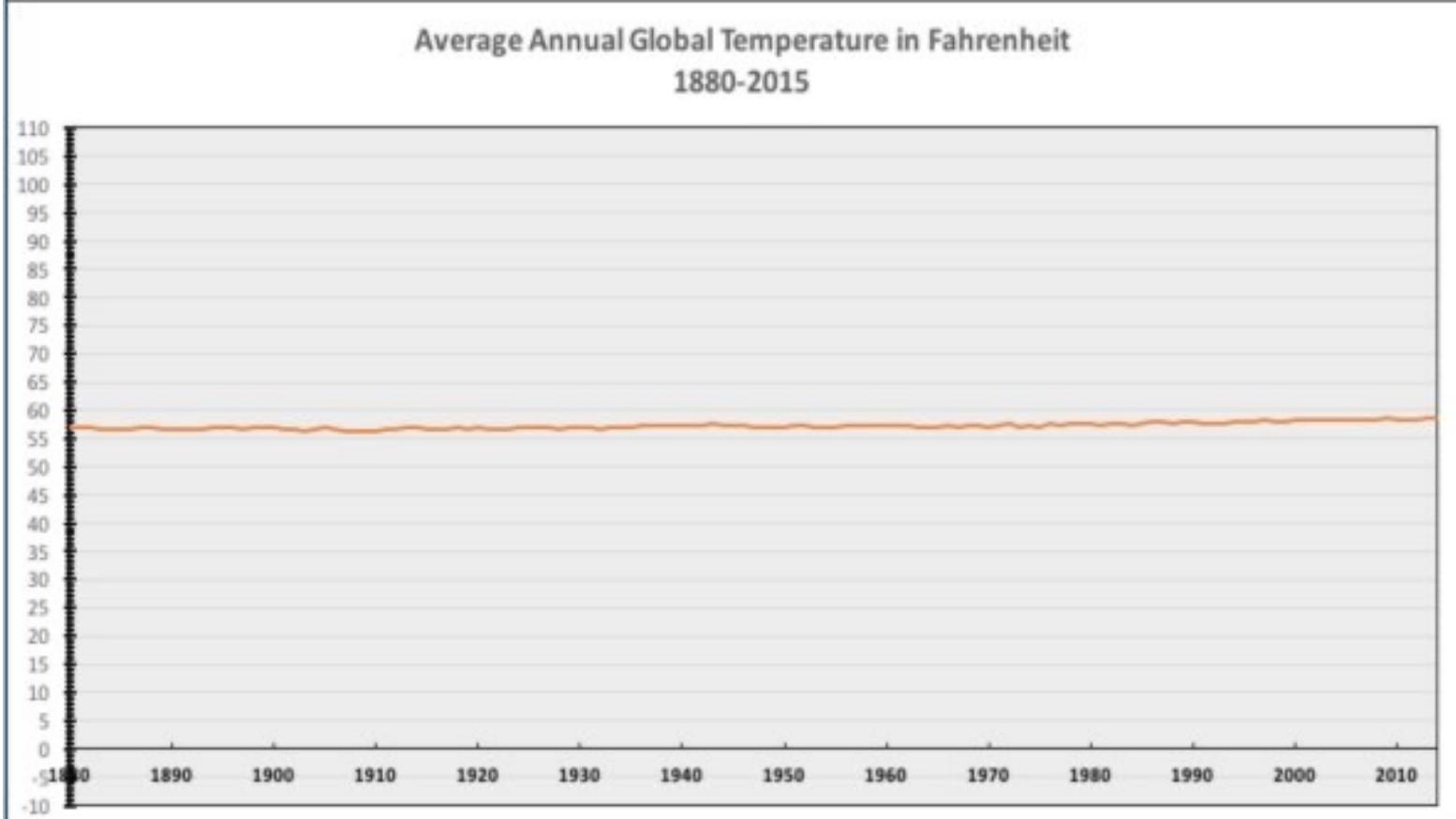
What's happening here?



What's happening here?



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!



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 deviation, 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 pack-

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).

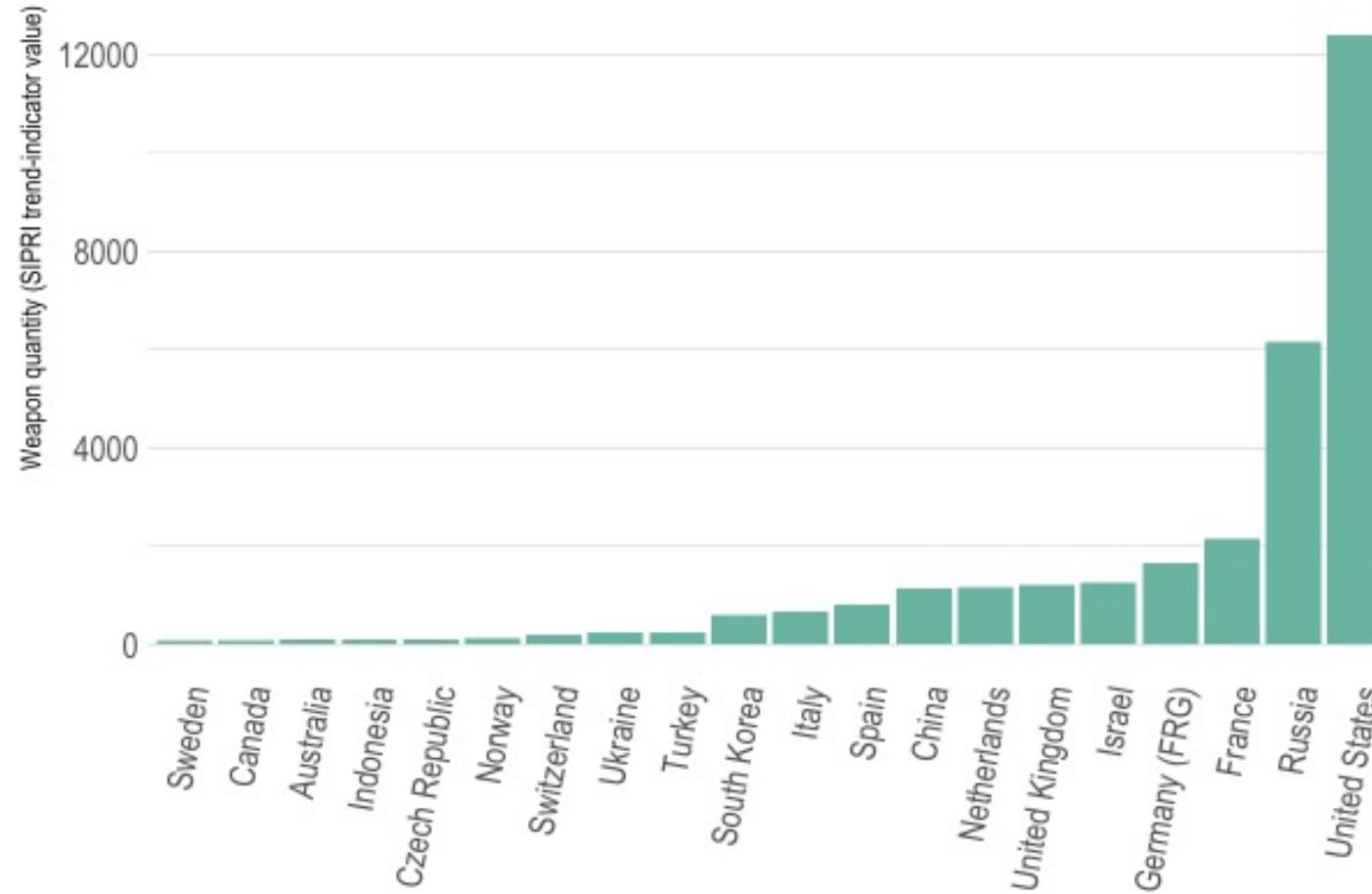
What range should I use?

Data range?
Start from zero?

Like all things, depends on the task.

Guideline: Rotate for Readability

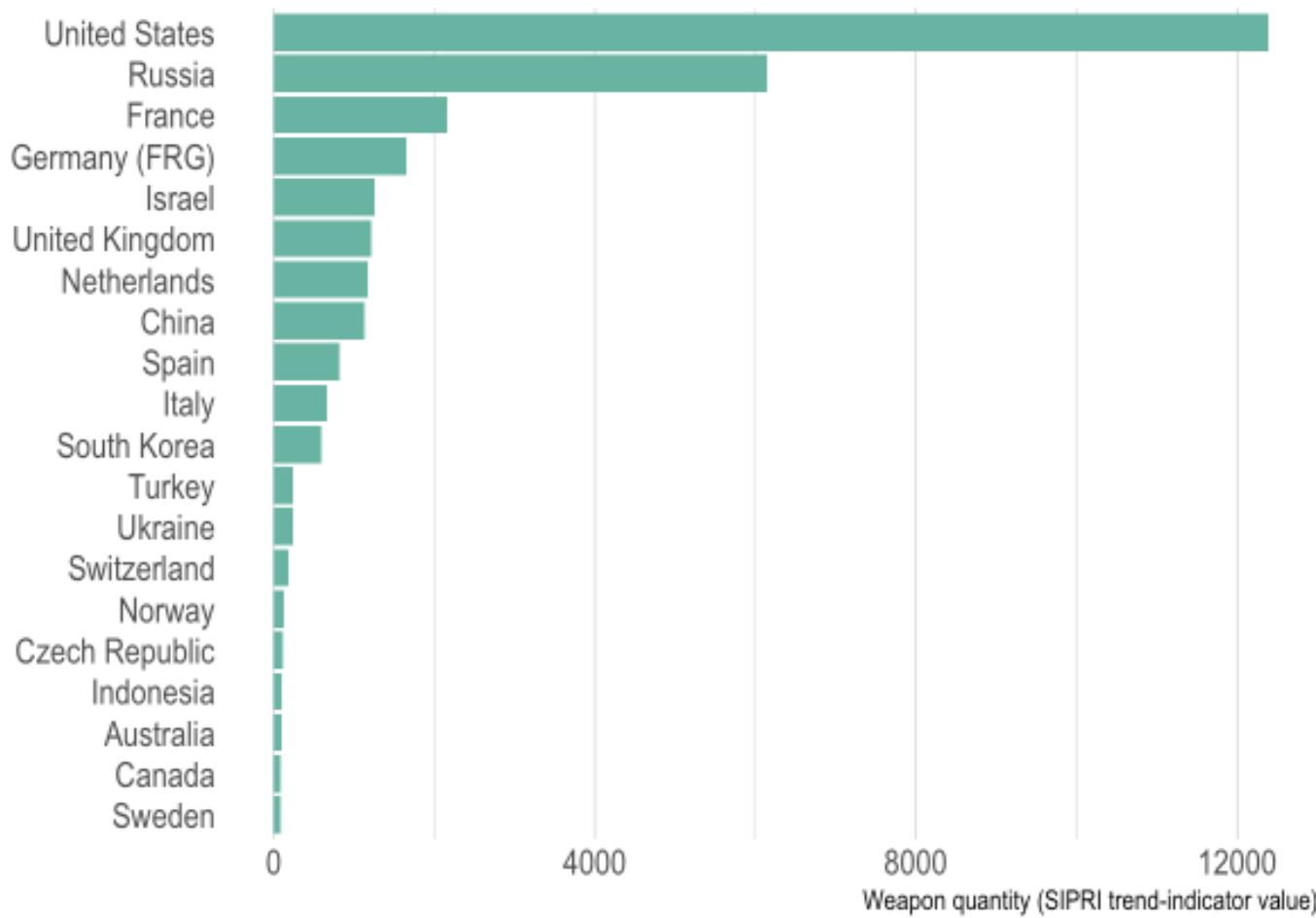
Be Aware of Text Angle



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

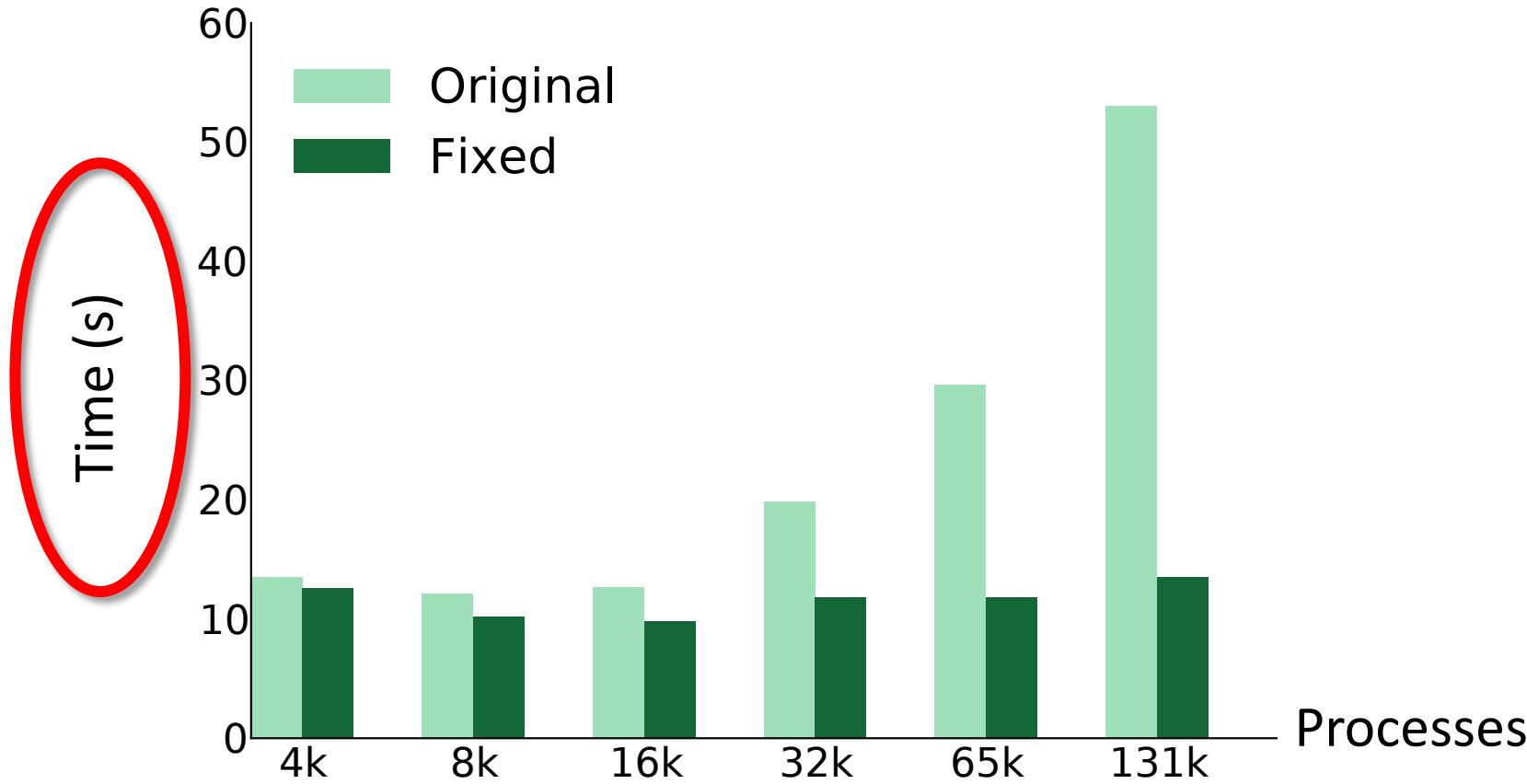
Furthermore, while we can rotate fonts at any angle, they distort and become jagged.

Consider rotating a bar chart

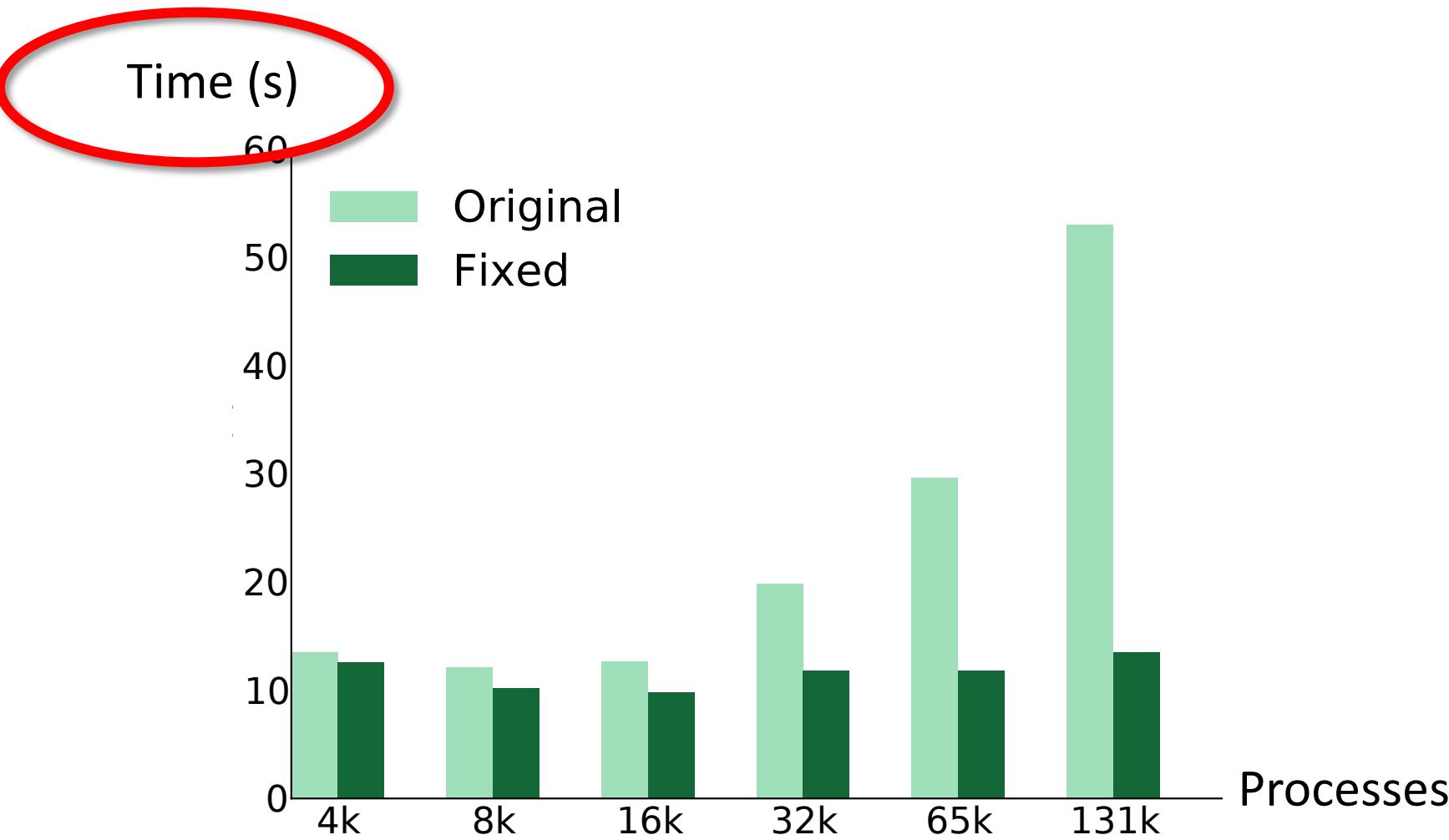


*Note however that some axes have a strong association with the x-axis (e.g., time). In that case, the design trade off may leave tilted text.

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

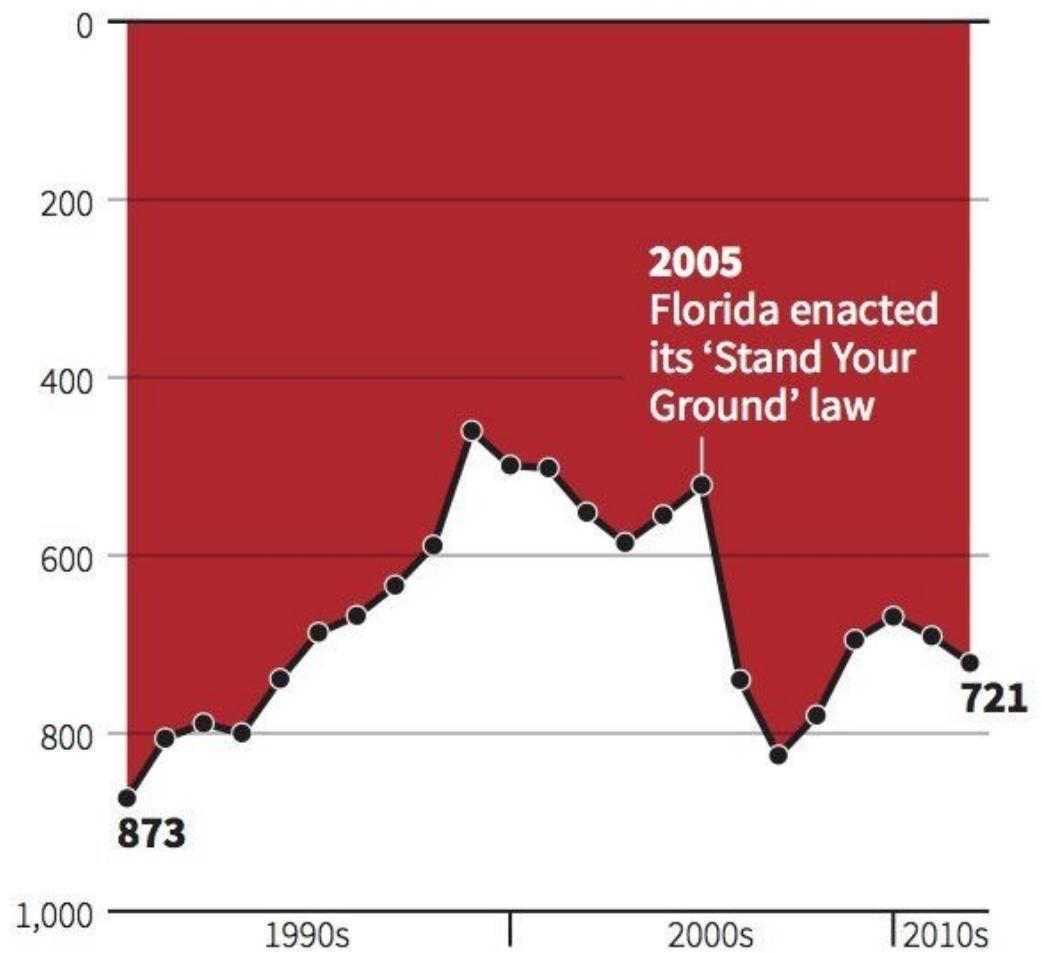
What is your initial reaction?

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

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

Gun deaths in Florida

Number of murders committed using firearms



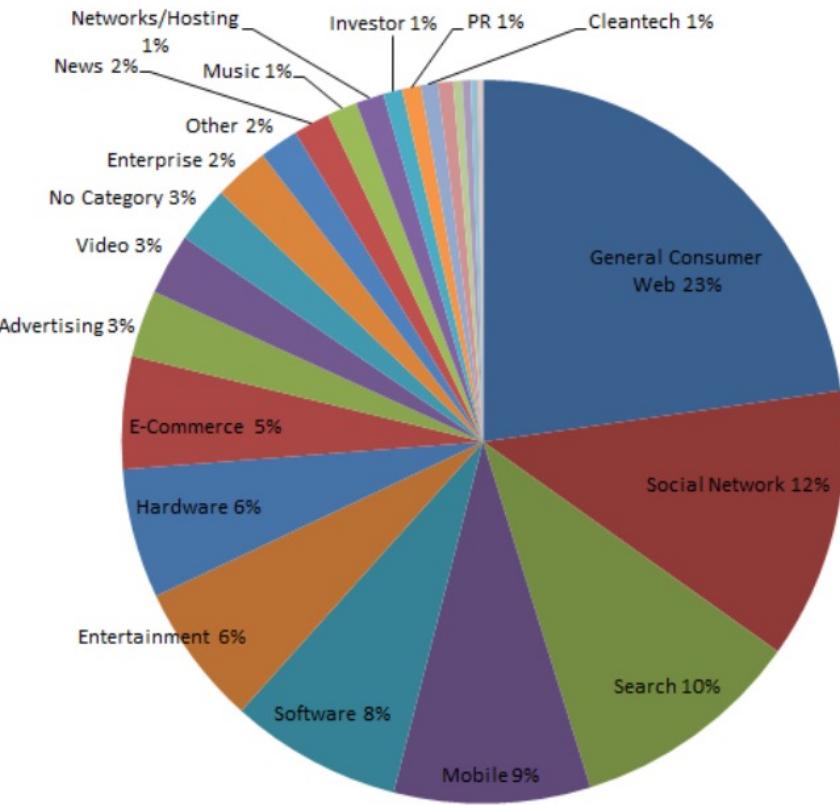
Source: Florida Department of Law Enforcement

C. Chan 16/02/2014



Guideline: Pie with Care

Pie Charts... easy to get wrong

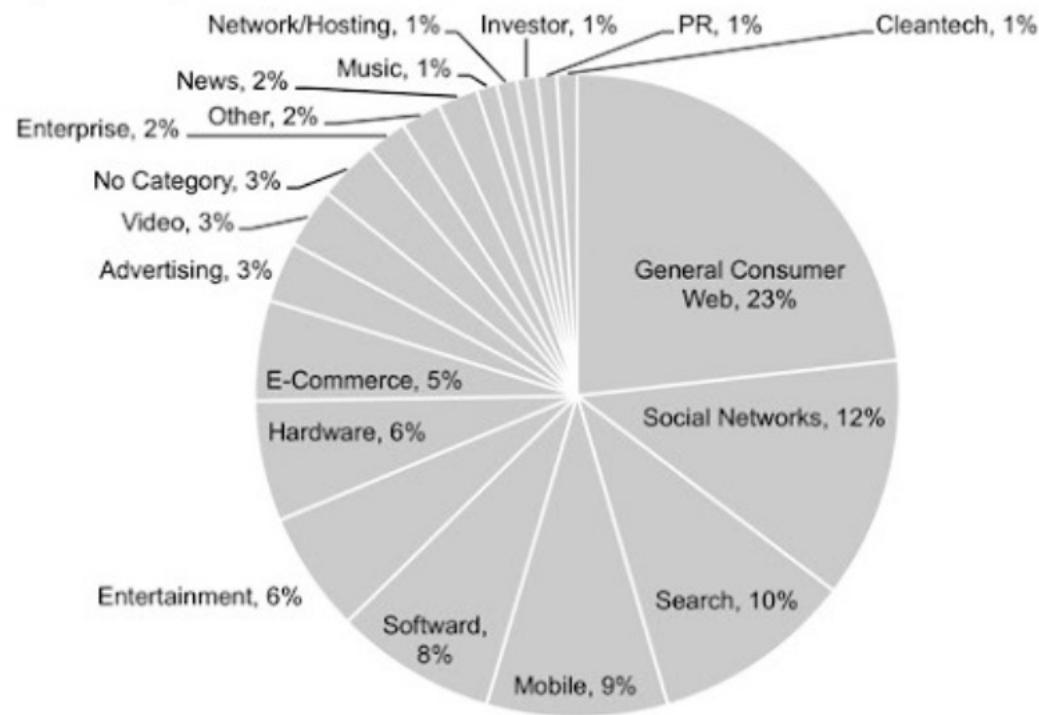


Share of coverage
on TechCrunch

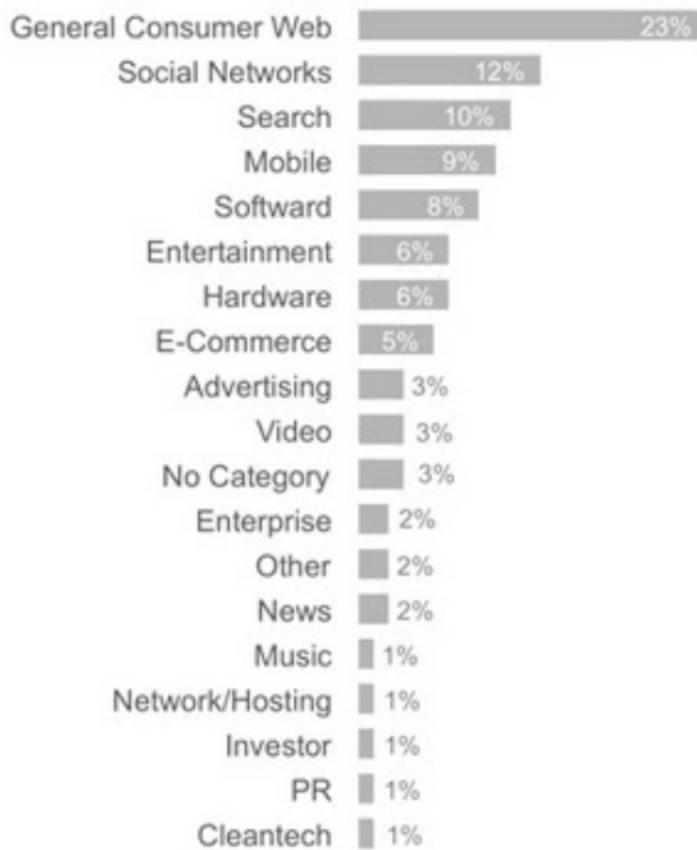
“I hate pie charts.
I mean, really hate them.”

Redesign

TechCrunch Coverage: 2005 - 2011
A slightly better pie?

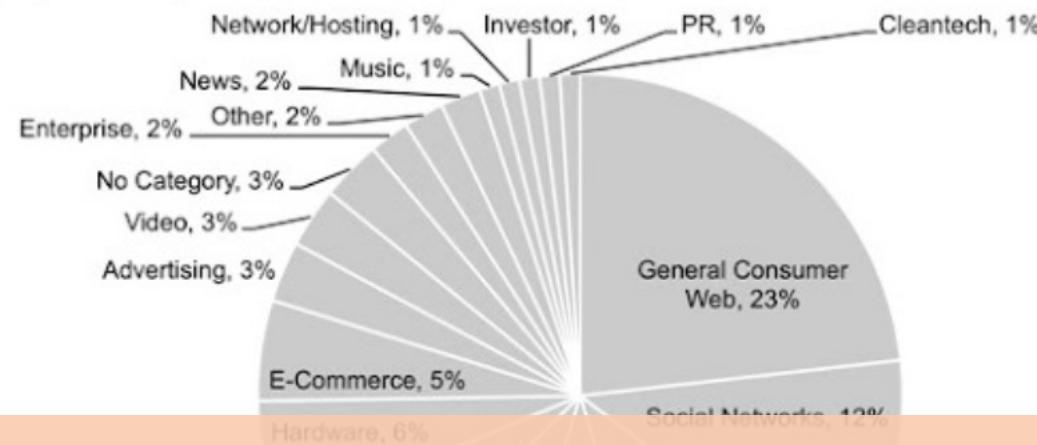


TechCrunch Coverage: 2005 - 2011
Bars are best!

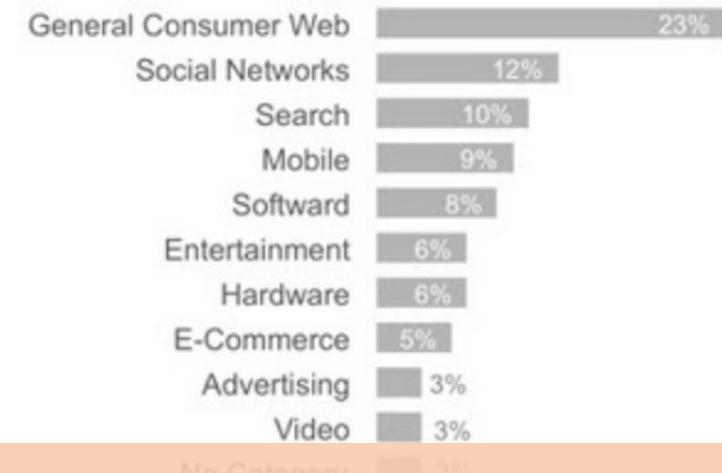


Redesign

TechCrunch Coverage: 2005 - 2011
A slightly better pie?



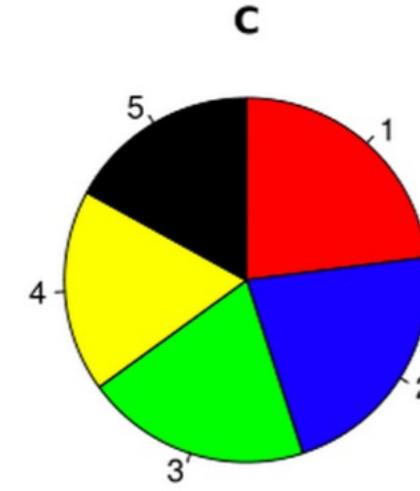
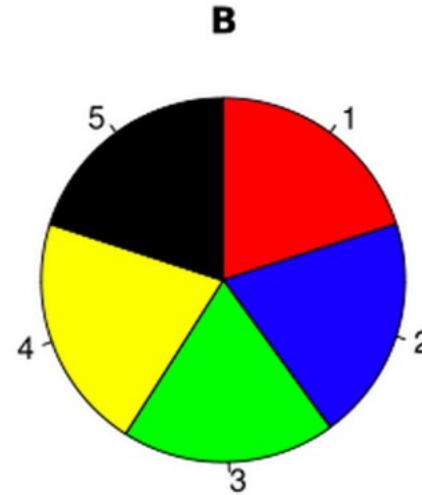
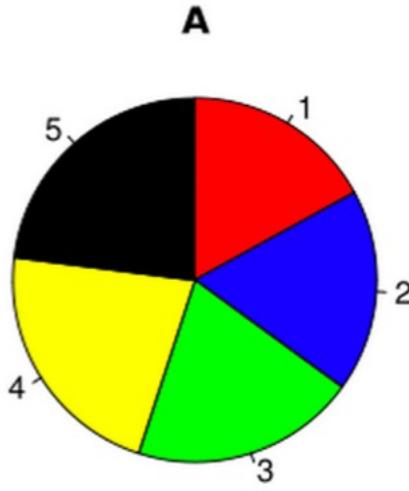
TechCrunch Coverage: 2005 - 2011
Bars are best!



What were they trying to show?

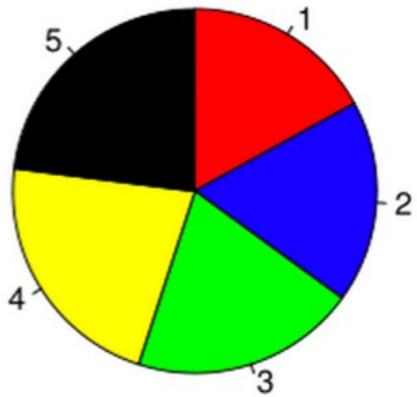
Is proportion the most important data feature?

Can you spot the differences?

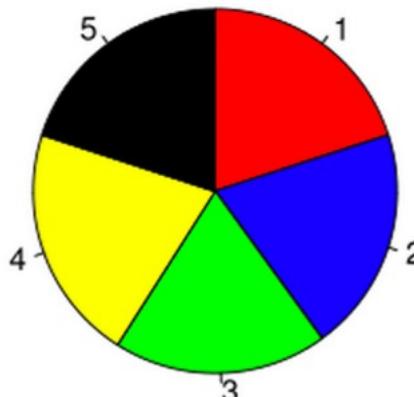


Can you spot the differences?

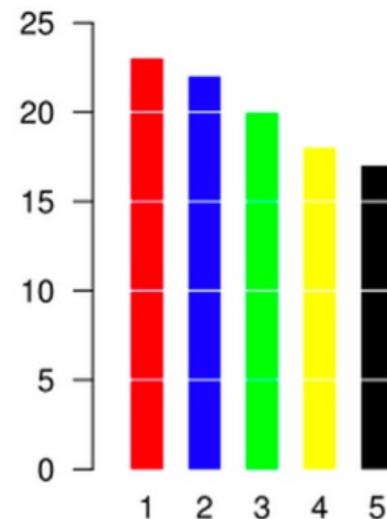
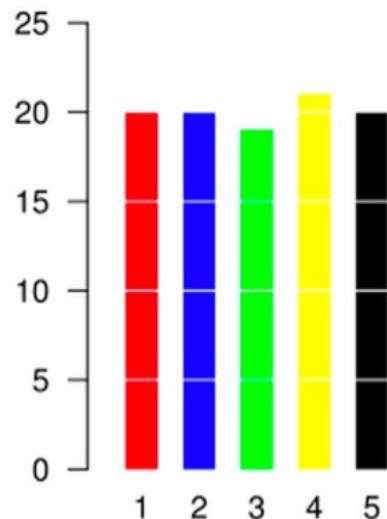
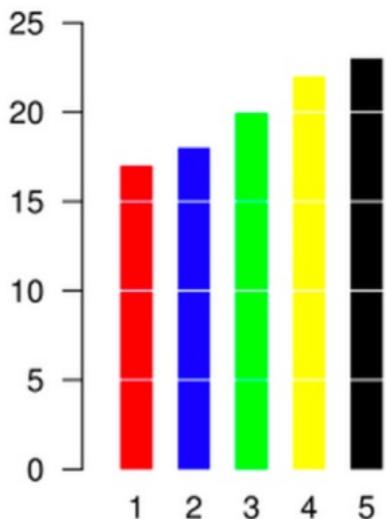
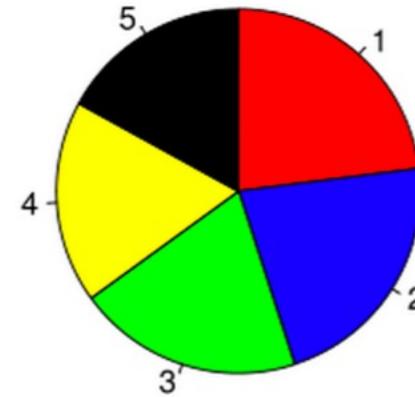
A



B

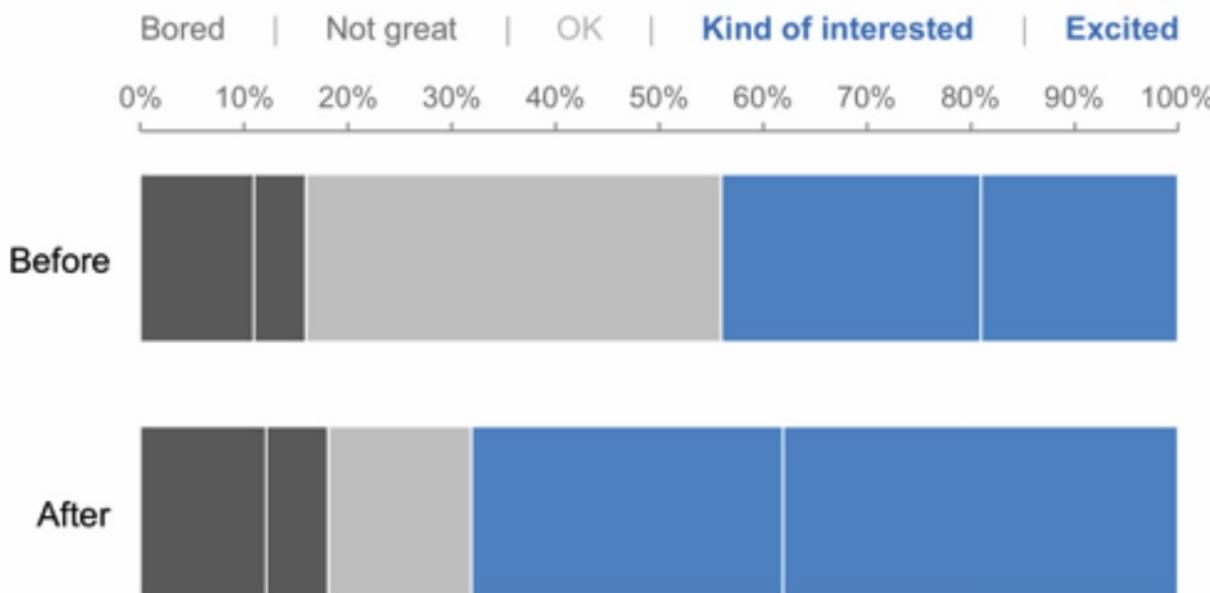


C



Pie Alternatives: Stacked Bar Charts

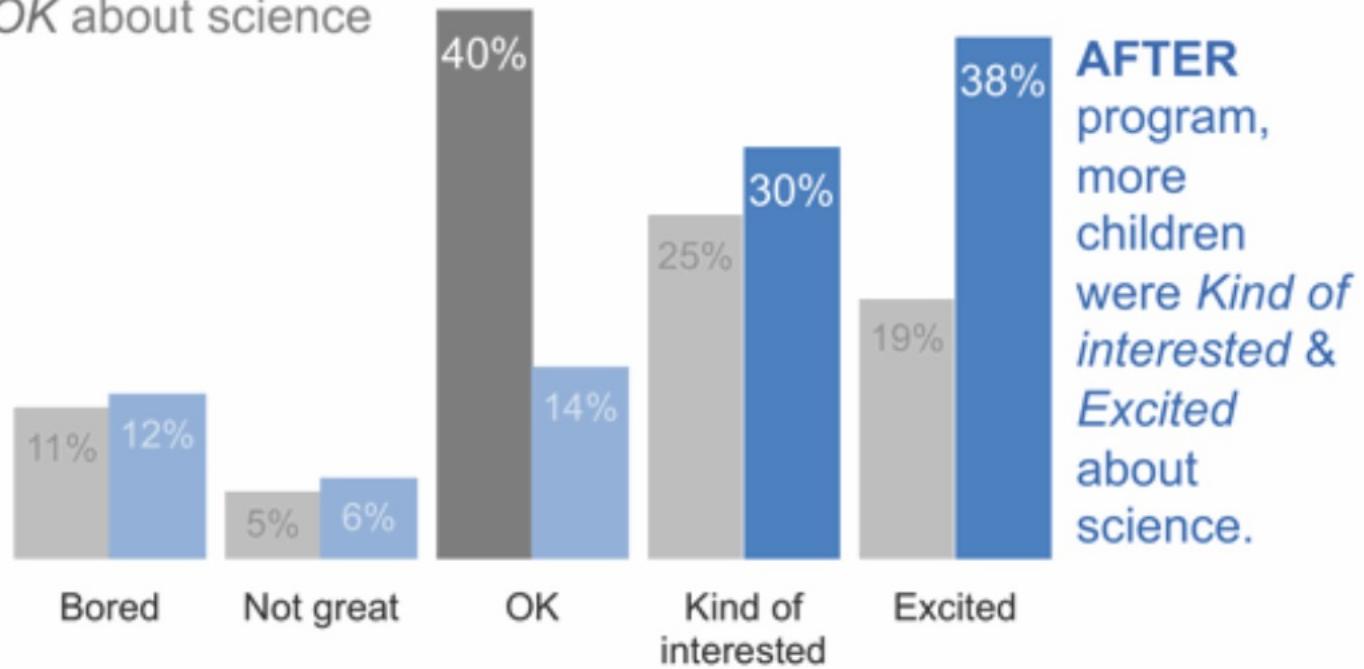
How do you feel about science?



Pie Alternatives: Bar charts

How do you feel about science?

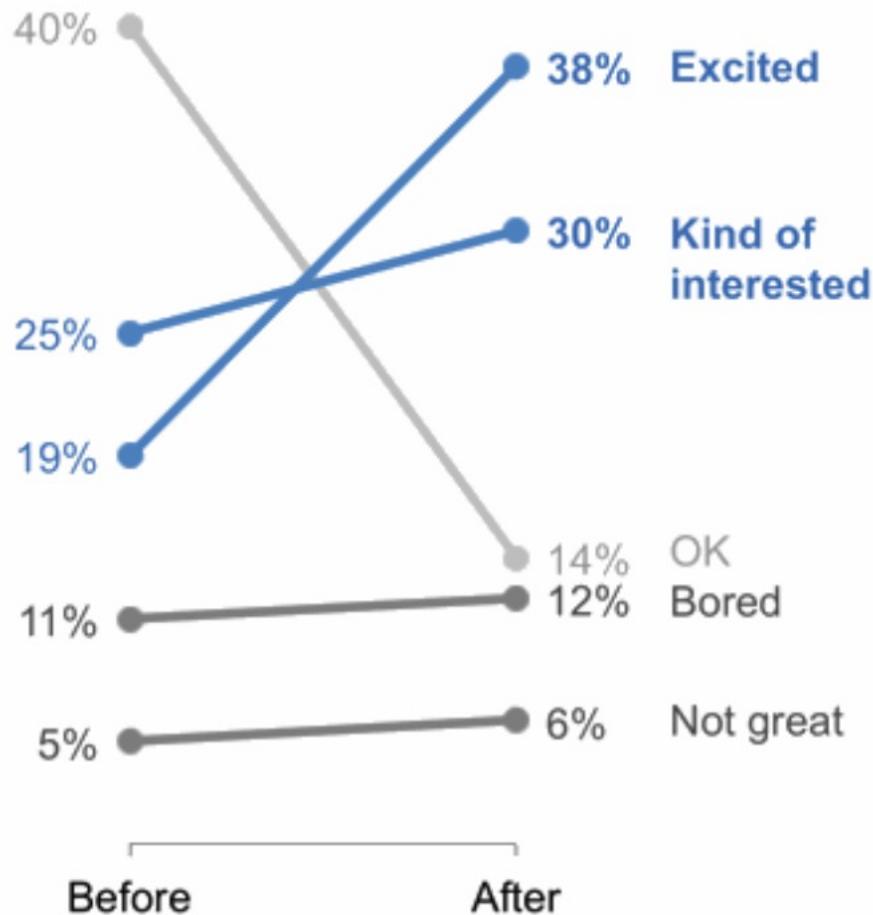
BEFORE program, the majority of children felt just OK about science



AFTER program,
more
children
were *Kind of
interested &
Excited*
about
science.

Pie Alternatives: Slope graphs

How do you feel about science?



Pie Alternatives: Just show the numbers

After the pilot program,

68%

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

Pie Alternatives: Just show the numbers

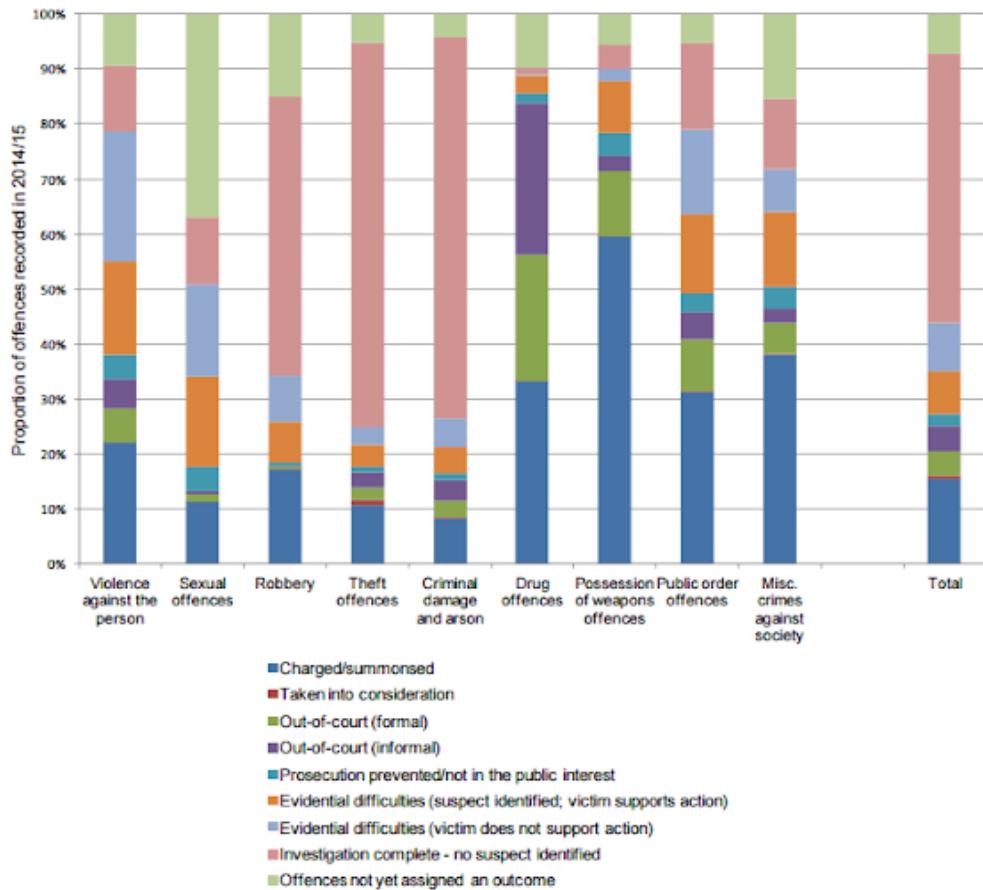
After the pilot program,

68%

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

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

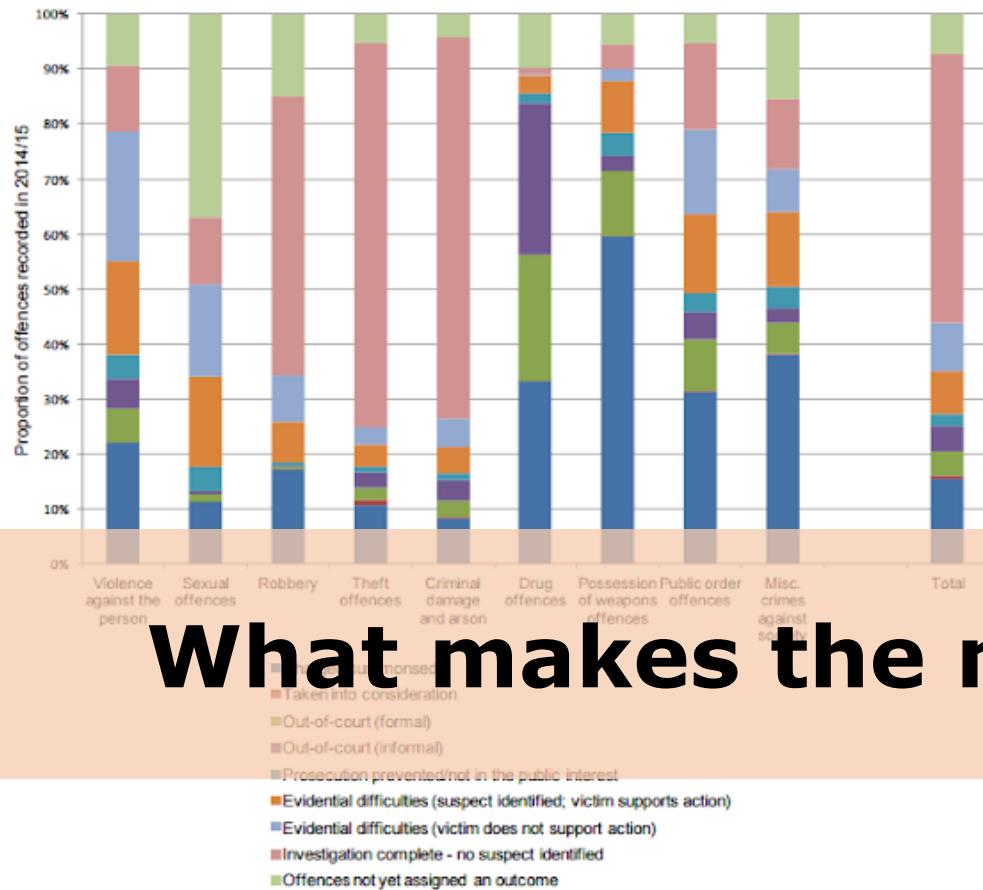
1. Based on 38 forces that supplied data as referenced in Table 2.1.

2. The numbers behind this chart are in Table 2.3

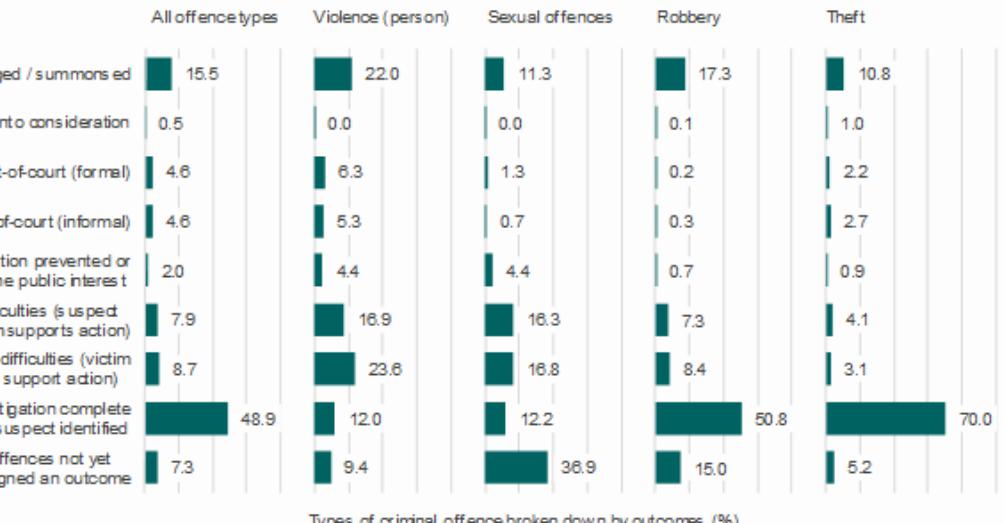


Stacked Bar Charts vs. Small Multiples

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



What makes the most sense for your task?

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

Google search results for "bad pie charts":

Search bar: bad pie charts

Filter: Images

Other filters: All, Videos, News, Shopping, More, Settings, Tools, SafeSearch

Autocomplete suggestions: wrong, media, example, data visualization, male female, economy florida, 2016 presidential election, attractive, advanced, 2...

Results:

- Yet another bad pie chart : dataisugly reddit.com**: A pie chart showing the distribution of Wikipedia editors by version added, with many small slices and illegible labels.
- death to pie charts – storytellingwithdata.com**: A complex sunburst chart illustrating the 100 most active tweeters, showing a hierarchical breakdown of users by location and activity level.
- Pie charts: the bad, the worst and the ... visuanalyze.wordpress.com**: A comparison of three pie charts: a good one (balanced segments), a bad one (irregular segments), and a worst one (tiny segments).
- When to use Pie Charts in Dashboards ... excelcampus.com**: A guide on when to use pie charts in dashboards, including a section on best practices.
- Using data visualizations' bad guy: pie ... martinraffaeiner.blog**: Two pie charts showing country population in 2010 as a percentage of total population, with one labeled as a "bad guy".
- Understanding Pie Charts eagereyes.org**: A detailed explanation of pie charts, including their strengths and weaknesses.
- Pie charts: the bad, the worst an... visuanalyze.wordpress.com**: Another comparison of pie charts, highlighting their不足 (bad, worst) and potential for improvement.
- Remake: Pie-in-a-Donut Chart - Policy Viz policyviz.com**: A remake of a pie-in-a-donut chart, showing a more effective way to represent nested data.
- Pin on Chartjunk Data Visualization pinterest.com**: A collection of various bad pie charts from Pinterest.
- Pie Charts Are The Worst - Business Insider businessinsider.com**: A critique of pie charts, emphasizing their limitations and suggesting better alternatives like stacked bars or treemaps.

Guideline: Area-as-Quantity with Care

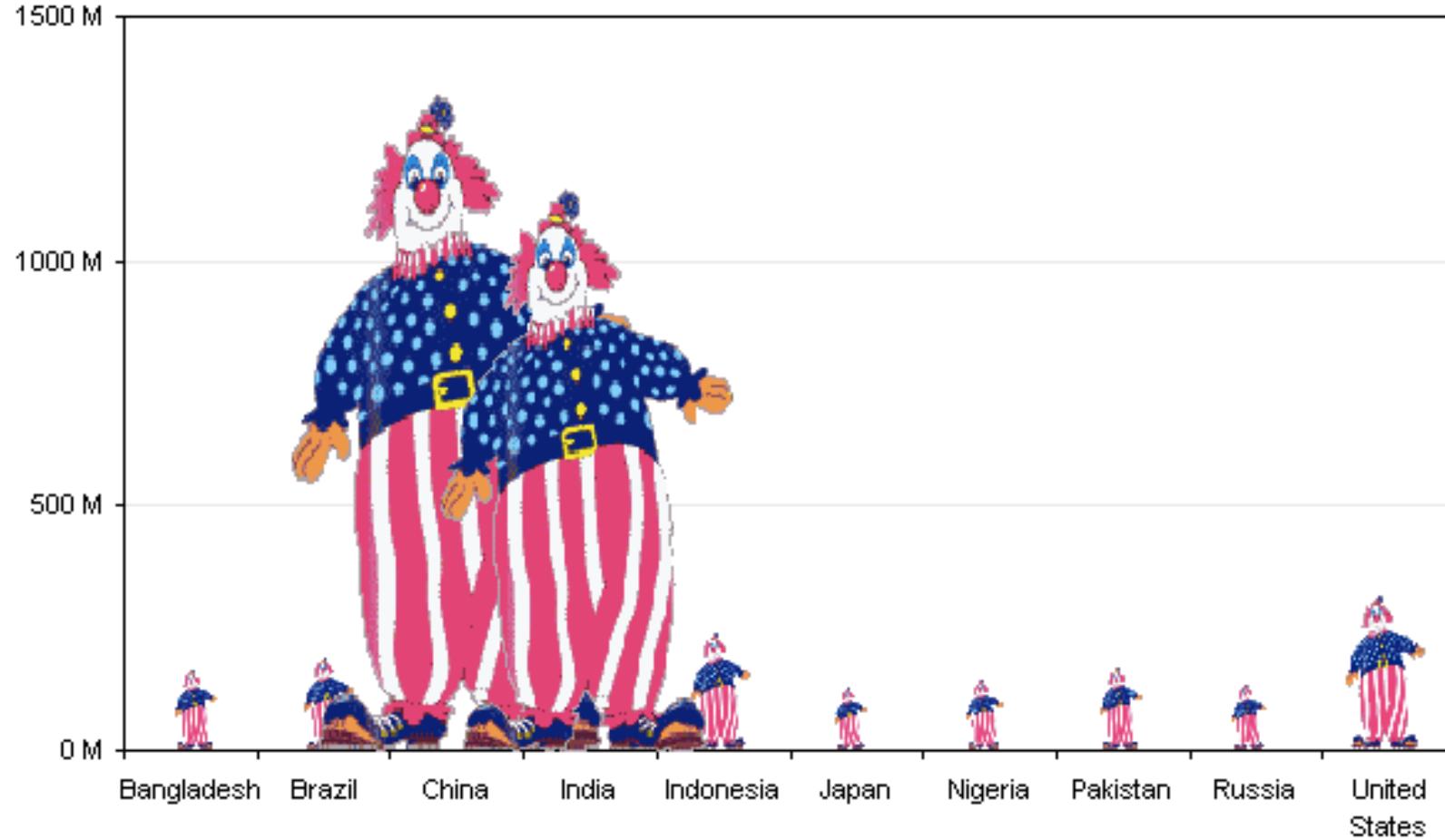
How many streams are there in November compared to December?



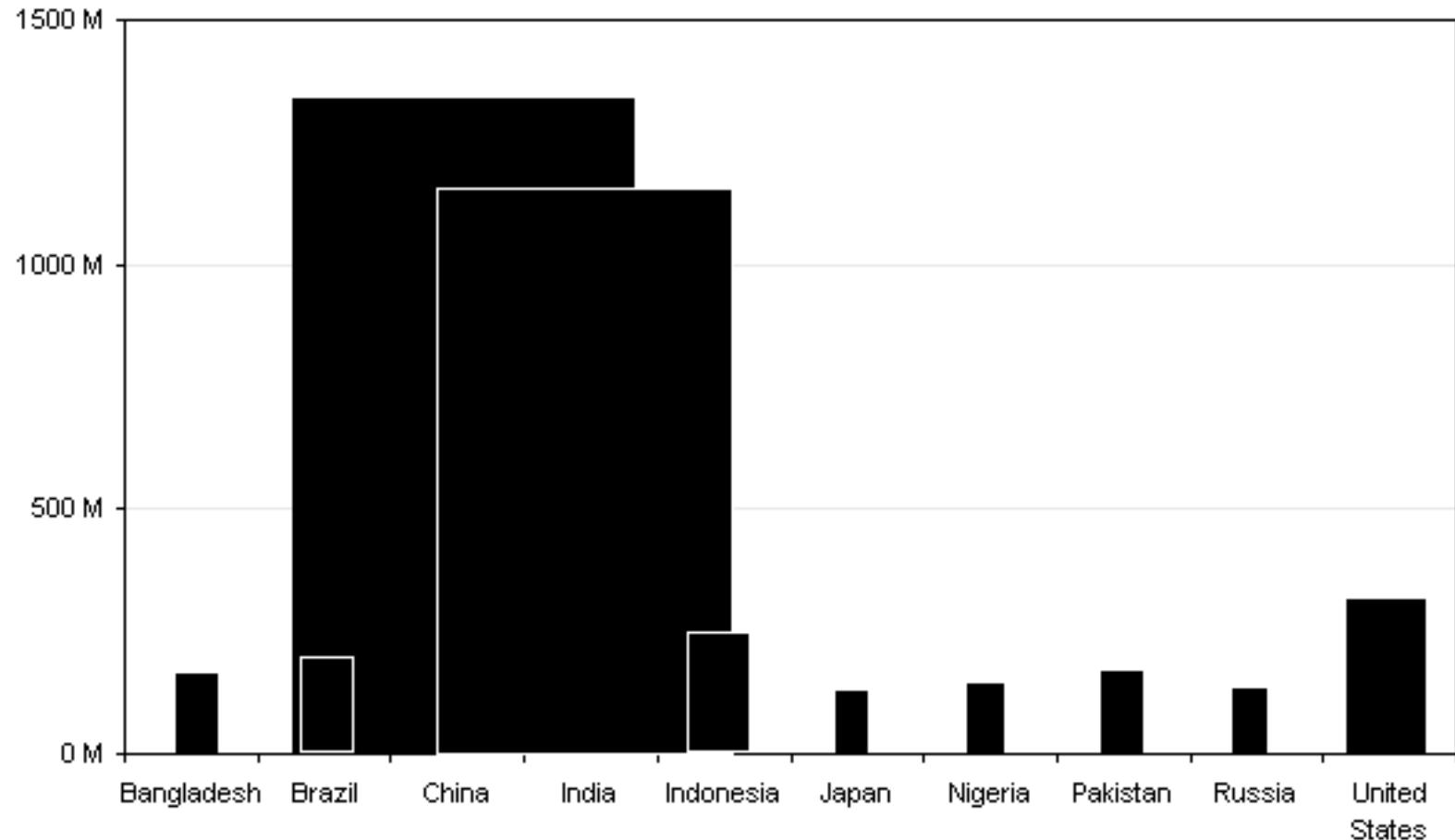
7.5 times as many streams!



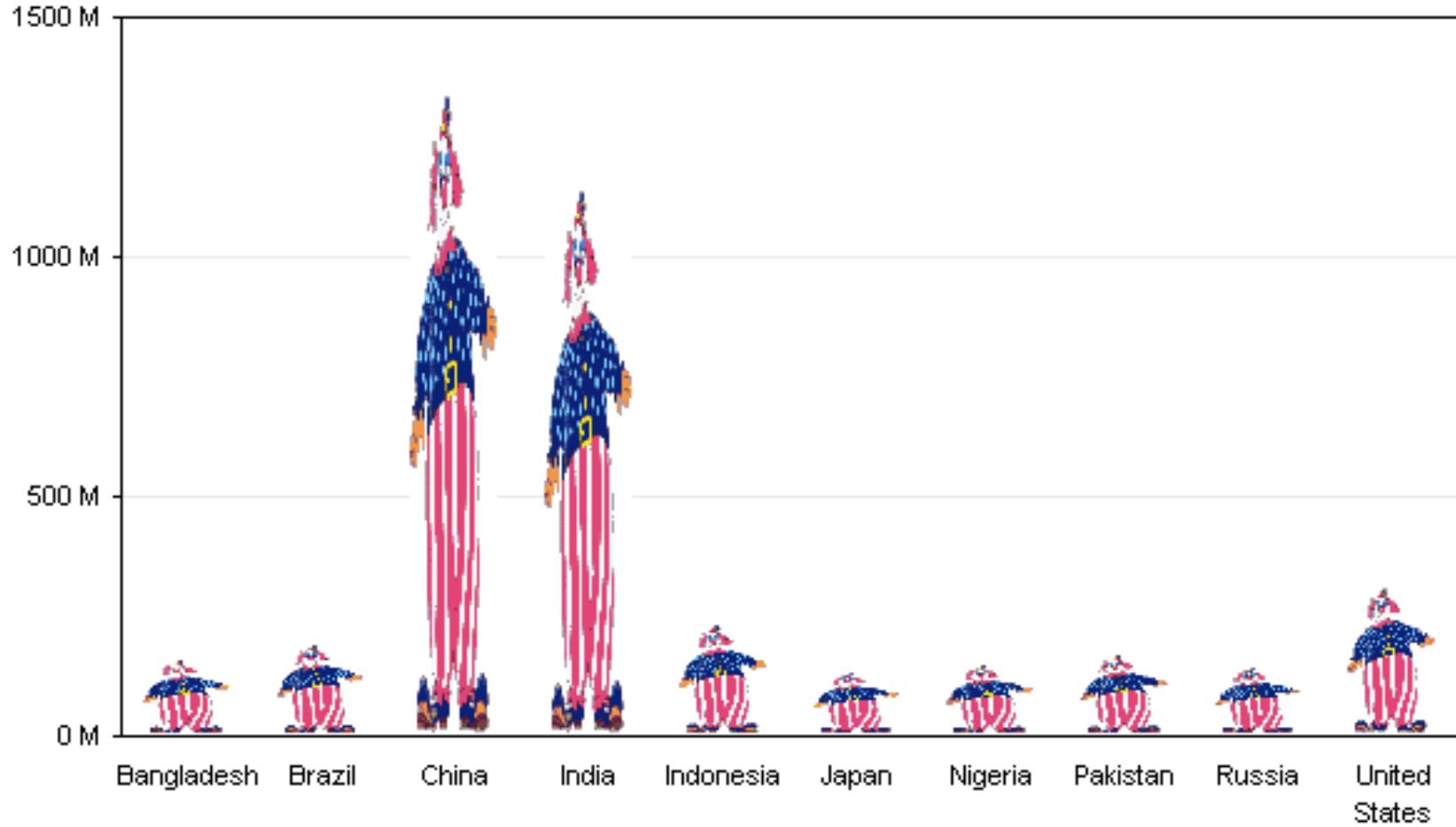
Be careful of length vs. area for other marks



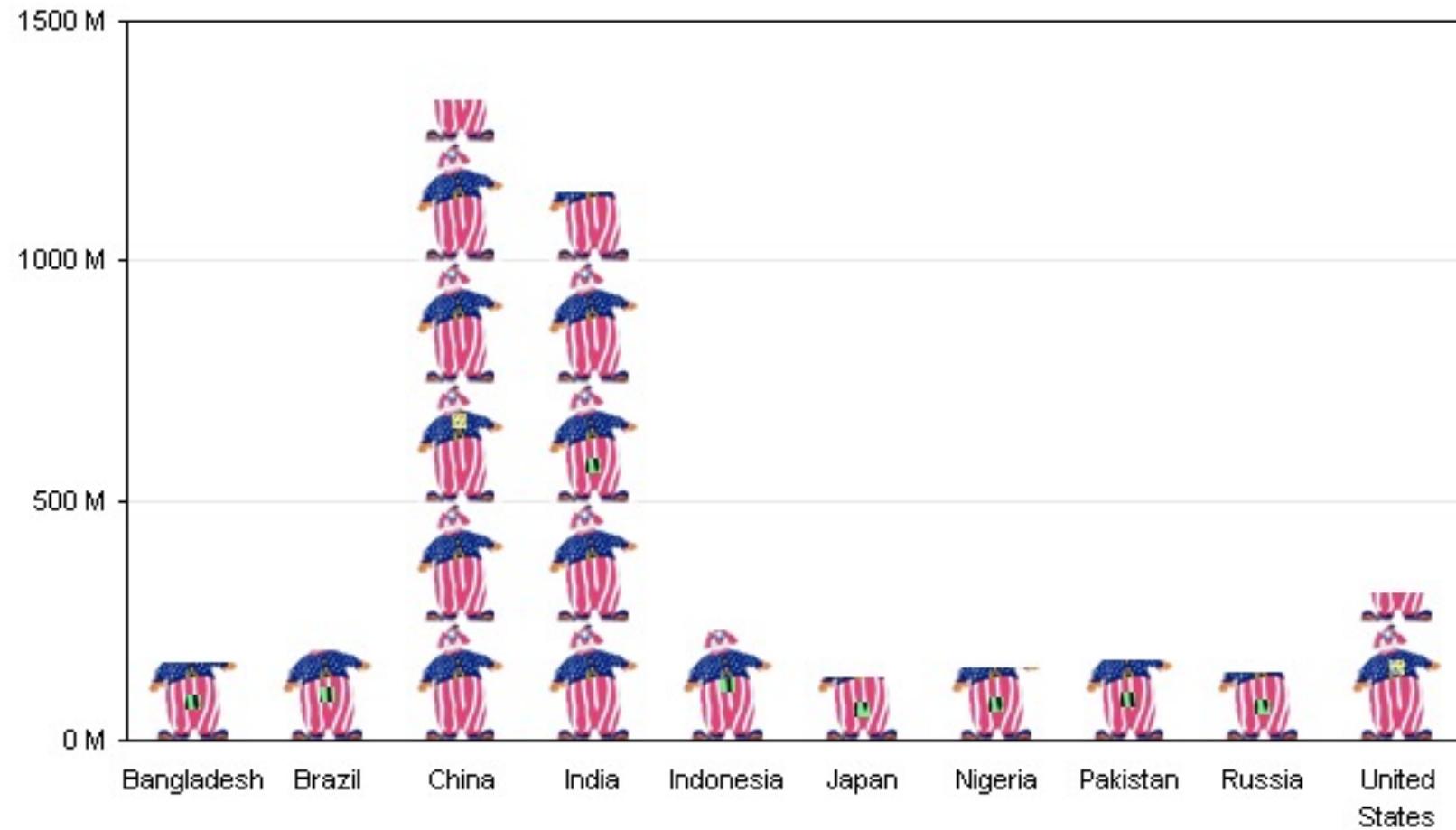
What is being perceived?



Fixing the width



Consider using an Isotope Chart



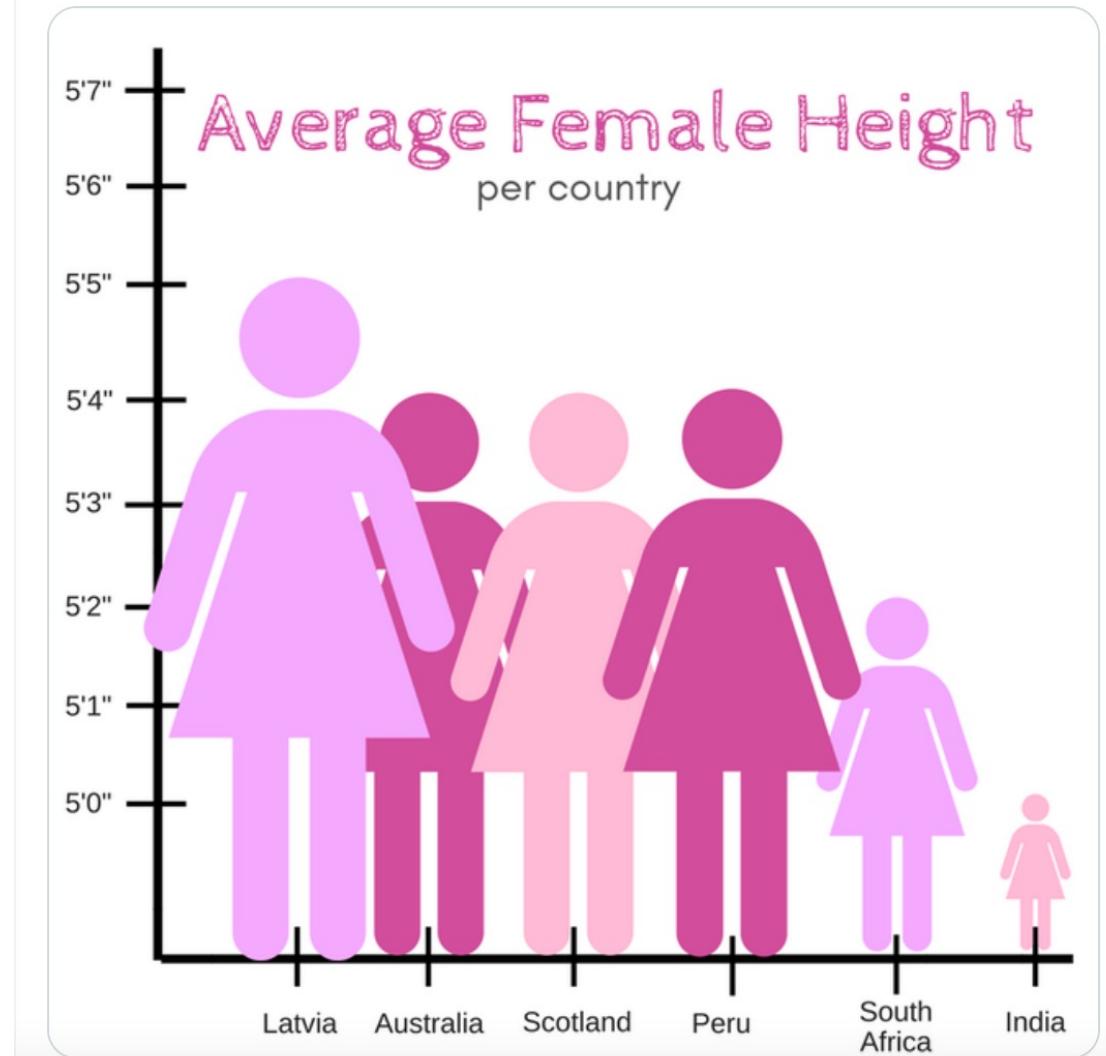
Now we're just breaking multiple guidelines at once



Sabah Ibrahim
@reina_sabah

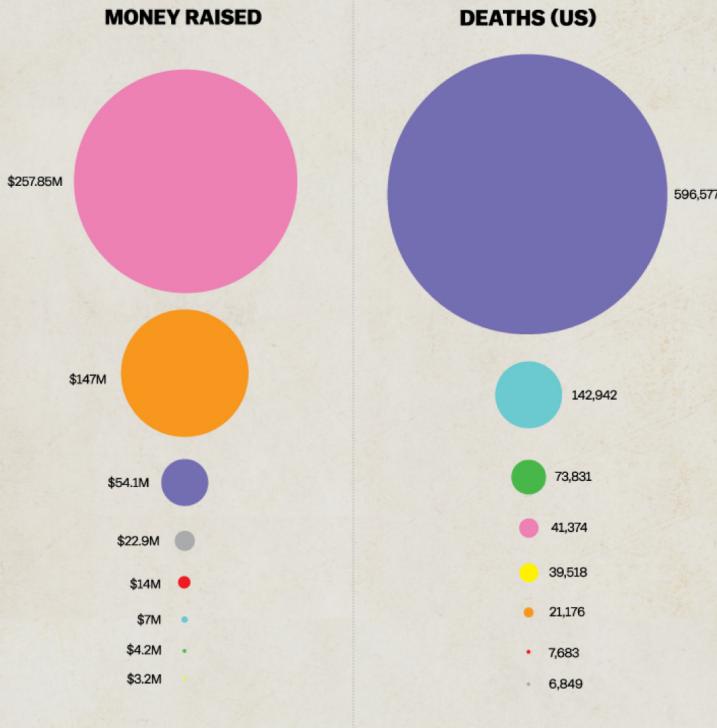
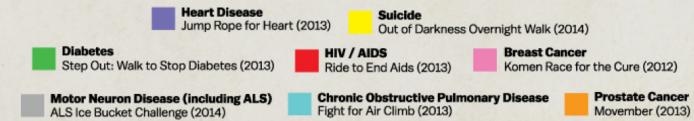
...

As an Indian woman, I can confirm that too much of my time is spent hiding behind a rock praying the terrifying gang of international giant ladies and their Latvian general don't find me



Circles: Encode by Area not Radius

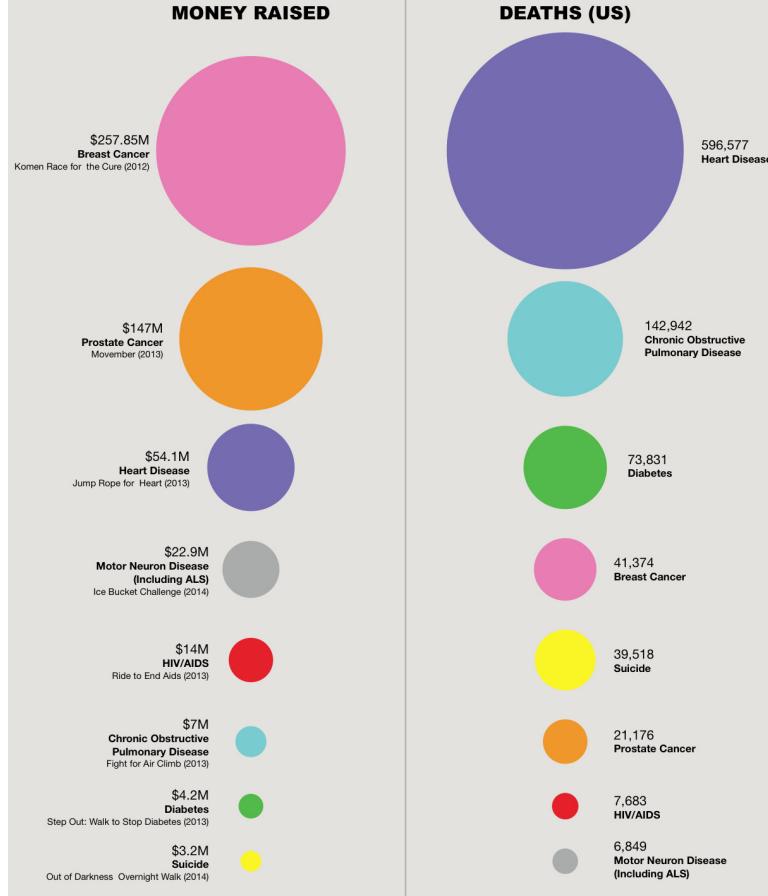
WHERE WE DONATE VS. DISEASES THAT KILL US



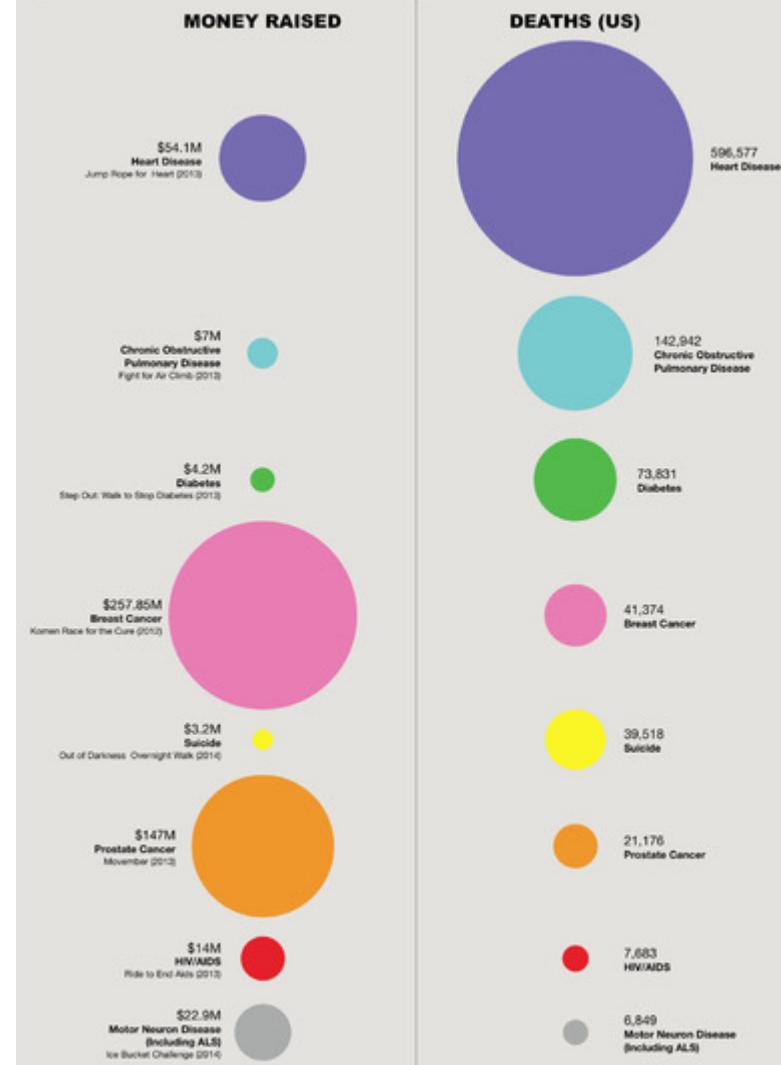
Source: CDC (2011)

V

WHERE WE DONATE VS. DISEASES THAT KILL US



WHERE WE DONATE VS. DISEASES THAT KILL US



Images from Vox and

<http://coolinfographics.com/blog/2014/8/29/false-visualizations-sizing-circles-in-infographics.html>