

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

## **(Some) Principles of Data Visualization + Project Overview**

February 3, 2026

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



**...let's talk about the Project first**

# Project

## *What you need to get started:*

- *A team of 3*
- *An idea*
- *A dataset (or multiple)*
  - *...that you can actually get!*

<https://datasciencecourse.net/2026/project/>

*Website has more details about milestones*

The screenshot shows a website header with the text "DS Introduction to Data Science" and the University of Utah logo. Below the header, there is a navigation bar with links: Home, Syllabus, Schedule, Project, Fame, and Resources. A large blue arrow points to the "Project" link in the navigation bar. The main content area contains several sections: "Projects", "Forms", "Introduction", "Project Team", and "Project Steps". Each section includes a brief description and some text below it.

**DS Introduction to Data Science**

Home Syllabus Schedule Project **Fame** Resources

**THE UNIVERSITY OF UTAH**

**Projects**

Submit Project Team and Topic (link forthcoming) by February 20, 2026

Submit Team Peer Feedback after each milestone (Proposal, Milestone, Final), link forthcoming

**Forms**

**Introduction**

For your final assignment in this course you will work on a data science project over several weeks. The goal of the project is to go through the complete data science process to answer questions about the dataset you choose. You will acquire the data, reshape and clean it, explore it, run analysis on it, visualize your results and present it to a reader.

**Project Team**

You will work closely with other classmates in a **3 person project team**. You can come up with your own teams and use our discussion forum to find prospective team members. If you can't find a partner we will team you up randomly. In general, we do not anticipate that the grades for each group member will be different. However, we reserve the right to assign different grades to each group member based on peer assessments (see below).

**Project Steps**

# Project Phases

- **Feb 15 (optional):** Request random team
- **Feb 20:** Submit your team & title
- **Mar 6:** Project Proposal due
- **Mar 19 (in-class):** peer feedback (required)
- **Mar 27:** Project Milestone due
- **Mar 30 (week of):** Staff feedback
- **Apr 17:** Final Report due
- **Apr 21:** Project Awards



DS Team Finder Survey - 2026

Please fill out this survey by February 15th to be randomly matched to one or more teammates.

DS Project Teams and Titles 2026

\* Indicates required question

Project Title \*

Your title should indicate the topic of your project.

Your answer

Team member information \*

Put one team member on each line:

[Name], [Email], [UID]. [Programming Experience on a Scale of 1 - 5, 1 being less comfortable, 5 being very comfortable]

*Projects phases must be done on time, assignment late policy does not apply – needed for feedback*

# Project Requirements

Scope as agreed upon with Staff

Project should include:

- Data acquisition (e.g., scraping, using an API)
- Data cleaning
- Exploratory analysis including visualization
- At least two different analysis methods (e.g., classification, regression, clustering, dimensionality reduction, NLP)
  - Evaluation of alternative approaches for each one (e.g., comparing classification methods)
- Ethical considerations

With the exception of ethical considerations, one inclusion may be minimal if made up for in other areas. For example, if your acquisition is a simple download, the analysis and cleaning may need to be more sophisticated.

# Ethical considerations

We will discuss ethics in a couple weeks, but some things to consider now:

- Where in your process are/were ethical decisions made? What were they?
- Who are the stakeholders of this project?
- If the project is successful and were to be released widely, who would be affected? Will some be more affected than others?
- Is your data biased in someway? If so, how?

“There are no ethical considerations” must be **strongly** defended. If the staff can identify any ethical considerations, you will get zero credit for that part of the project.

# Don'ts

**Don't** use a standard machine learning data (e.g., MNIST, common datasets from UCI ML Repository, many on Kaggle...)

- These are often pre-processed for specific analyses, and don't require other parts of the DS process like cleaning and exploratory analysis
- If you use from these repositories, must demonstrate in proposal how data is appropriate for the entire project process

**Don't** choose a dataset where it is too hard to extract structured data:

- e.g., requires advanced NLP, extracting from PDF, etc.
- ...or, if you do go this route, proposal should demonstrate you are prepared to extract this data. You must state specifically how you will do so, perhaps even giving an example.

# Proposal

Sections lists on website

Submit as Jupyter notebook

Group submits one together

Individual feedback forms  
allow you to tell staff about  
the division of work

## Proposal

You start your project by forming your groups and letting us know what topic you are interested in exploring by submitting a project data form. **Please submit only one form per team!** In addition to the form, you will create a proposal document in the form of a Jupyter Notebook, addressing the following points. Use these points as headers in your document.

- **Basic Info.** The project title, your names, e-mail addresses, UIDs.
- **Background and Motivation.** Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.
- **Project Objectives.** Provide the primary questions you are trying to answer in your project. What would you like to learn and accomplish? List the benefits.
  - This should include both questions about the data and any learning objectives you would like to fulfill. In other words, there are two kinds of benefits to address.
- **Data Description and Acquisition.** What format is your data in? How many items are there? What attributes do those items have? Are there special structures in it (e.g., networks, geographical)? From where and how are you collecting your data? If appropriate, provide a link to your data sources.
  - This part should be specific enough that the instructional staff is assured you have or will be able to obtain data.
    - If it's online through direct download, link to the specific page from which you will download it.
    - If you will scrape it from the web, link to the page from which you will scrape it and a statement regarding how you have confirmed you are permitted to scrape it.
    - If you will use an API to access it, link to the documentation of the API and explain how you have access to that API
      - If it requires an account, state you have one.
      - If it doesn't require an account, state that it does not require one.
    - If it is data you have access to through other means, describe in detail what the data is, how you have access to it, and why you have permission to use it.
- **Ethical Considerations.** Complete a stakeholder analysis for your project.
  - Who may be affected by your project and its outcomes? How could your project be used for harm?
  - "There are no ethical considerations" must be *strongly* defended. No one successfully done this before in this class.
- **Data Cleaning and Processing.** Do you expect to do substantial data cleanup or data extraction? What quantities do you plan to derive from your data? How will data processing be implemented?
- **Exploratory Analysis.** Which methods and visualizations are you planning to use to look at your dataset?
- **Analysis Methodology.** How are you planning to analyze your data?
  - What specific questions do you hope to calculate?
  - What methods (from class or otherwise) do you think you will use?
- **Project Schedule.** Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.

As a ballpark number: your proposal should be about 2-3 pages of text and figures. You can also include some preliminary data acquisition / analysis steps.

# Project Milestone

For your milestone, we expect you to have acquired, cleaned, and explored your dataset. You should also explain in more detail what will go into your final analysis. Explain deviations from your initial project plan. In other words, we expect an elaborated data description, acquisition, cleaning, exploratory analysis, and an updated project schedule that discusses changes in plans from your project proposal. The acquisition, cleaning, and exploratory analysis should include the code to accomplish these tasks. Please revise any other sections as necessary.

If you are uncertain about the scope, please contact the staff member responsible for your project.

The milestone should be submitted as a zip file containing a Jupyter notebook and any supporting documents. the Jupyter notebook should contain all the narrative and code. Do no submit a separate document with the write up from the Jupyter notebook. Note this zip file should also include your in-class feedback as a separate file with the name

`feedback_exercise`.

Like with the assignments, submit the Jupyter notebook with the output. Make a large note at the top if you are not able to include your data due to size.

Needs to contain narrative and code that demonstrates acquisition, cleaning, and exploratory analysis of your dataset.

Submission still single Jupyter notebook + individual feedback forms

# Final Project Submission

For your final project you must complete the analysis in your notebook and present your results in a compelling way. We recommend you include revised versions of all the sections from your proposal (except for the Project Schedule) *along with* the results of your analysis, the limitations of your analysis, and your conclusions from your data analysis.

Like the previous milestones, you should submit as a zip file containing a Jupyter notebook and any supporting documents. the Jupyter notebook should contain all the narrative and code. Do no submit a separate document with the write up from the Jupyter notebook.

Like with the assignments, submit the Jupyter notebook with the output. Make a large note at the top if you are not able to include your data due to size.

## Project Screen-Cast

You must include a **three minute video including audio walking us through your project**. Each team will create a **three minute screen-cast with narration** showing a demo of your project and/or some slides. You can use any screencast tool of your choice. Please make sure that the sound quality of your video is good. Upload the video to an online video-platform such as YouTube or Vimeo and link to it from your notebook.

Present your analysis questions and your main contributions, but also explain your methods and justify your choices. What do you feel is the best part of your project? What insights did you gain? What is the single most important thing you would like your audience to take away?

We will strictly enforce the three minute time limit for the video, so please make sure you are not running longer. Use principles of good storytelling and presentations to get your key points across. Focus the majority of your screencast on your main contributions rather than on technical details. What do you feel is the best part of your project? What insights did you gain? What is the single most important thing you would like your audience to take away? Make sure it is front and center rather than at the end.

Full narrative of the project, including your analyses, your conclusions, and a narrative of the whole project (within the notebook & the movie)

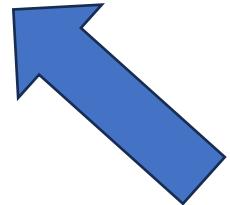
Submission is the Jupyter notebook, individual feedback, + a video & screencast

# Example Projects on Course Website



Home Syllabus Schedule Project Fame Resources

## Hall of Fame



## Best Projects 2022

### Winners

#### Deja Brew

Ja-Rey Corcuera, Brianna Mendoza

[Project Video](#)

#### Analyzing Induced Microseismicity at Utah FORGE

Faris Khan, Patrick Bradshaw, Barrett Kilroy

[Project Video](#)

#### Optical Spectroscopic Object Classification

Alexander Millar, Roanna Rague

[Project Video](#)

*\*Some groups smaller than three because we had more Staff back then.*

*Please scale up your project accordingly.*

# Group Work Expectations

## Be responsive to your teammates

- Communicate regularly, status updates help:
  - “I can’t make it today, but I have done X, Y, and Z and put it in the shared Drive.”
  - “I have work at that time, but I’ll get you A and B before the meeting.”
  - “Sorry I’m behind but I will have it to you by DATETIME.”
    - ...and if that doesn’t work for your team, they can move on without you.
- If a teammate has not responded in **48 hours**, you may move on without them. Please let the course staff know.

## Recognize that “glue work” takes time

- Your teammates can’t add your work at the last minute. It takes time to integrate it. If they don’t have time to integrate it, you can’t get credit for it.
- Glue work is work. It is a contribution to the project.

# Handling Group Work Issues

If problems arise, let the course staff know early.

- We can't go back in time and change assignment grades
- We don't always adjust grades when something is reported
- If someone didn't contribute at all, they will get a zero
- We can't share with you any individual student's grade but your own

Worried about hard feelings? Communicate directly rather than staff.

- “I don’t feel work was evenly distributed in the proposal. [Details]. How can we divide it better in the next milestone?”
- “I scrambled to integrate your work at the last minute but I can’t do that in the next milestone. I need it by ABC or it will not get submitted.”

Worried your work isn't recognized? Use version control (git, etc)

- Version control through git or other platforms can keep a history of the work you did. We will review if there are discrepancies.

**...back to Visualization**

**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:

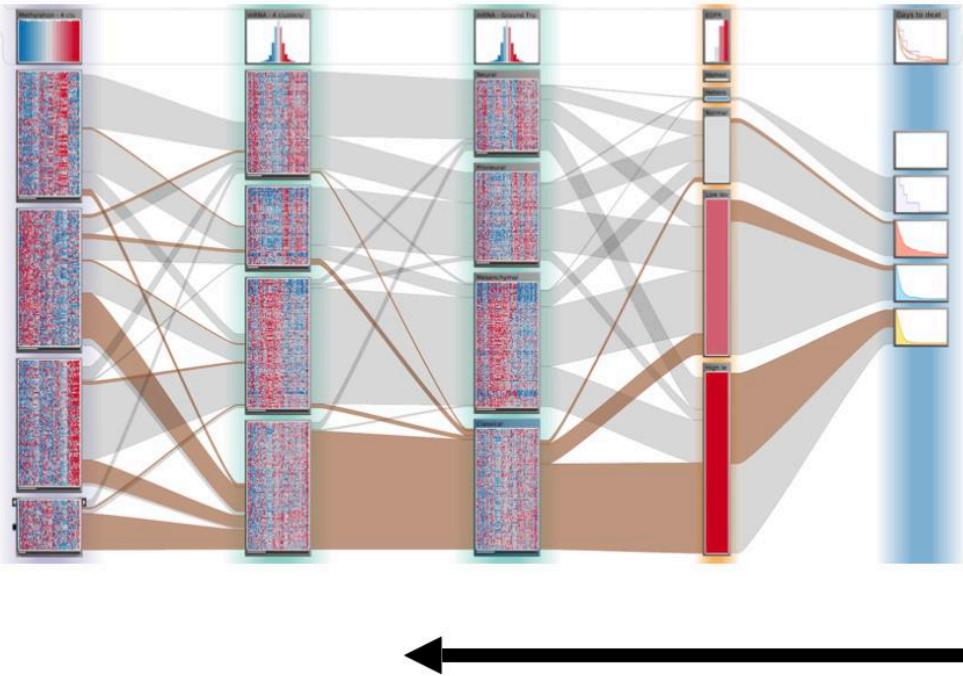
## Exploration

**(when results and questions are unknown)**

*What is the structure of the scammer network?*

*Which drug can help my patient?*

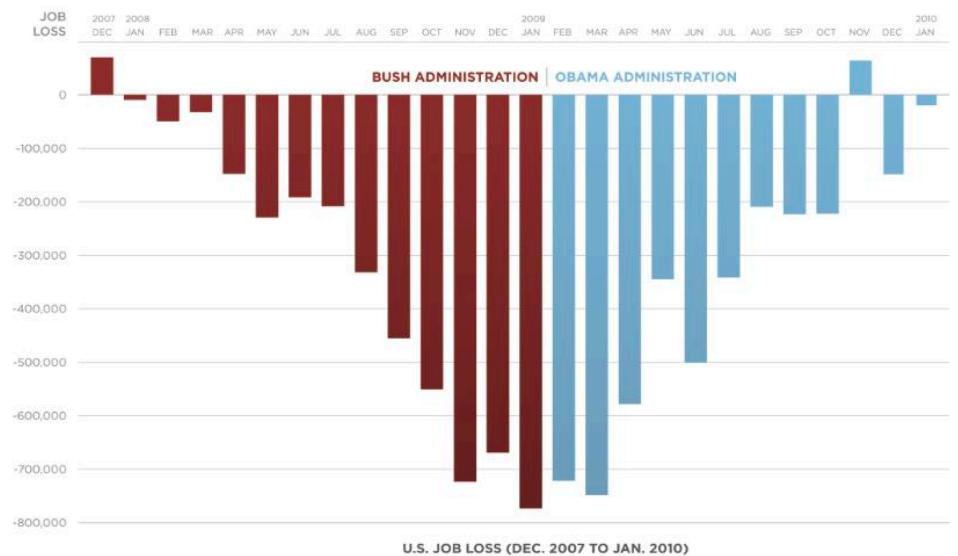
*Is this data even right?*



Open Exploration

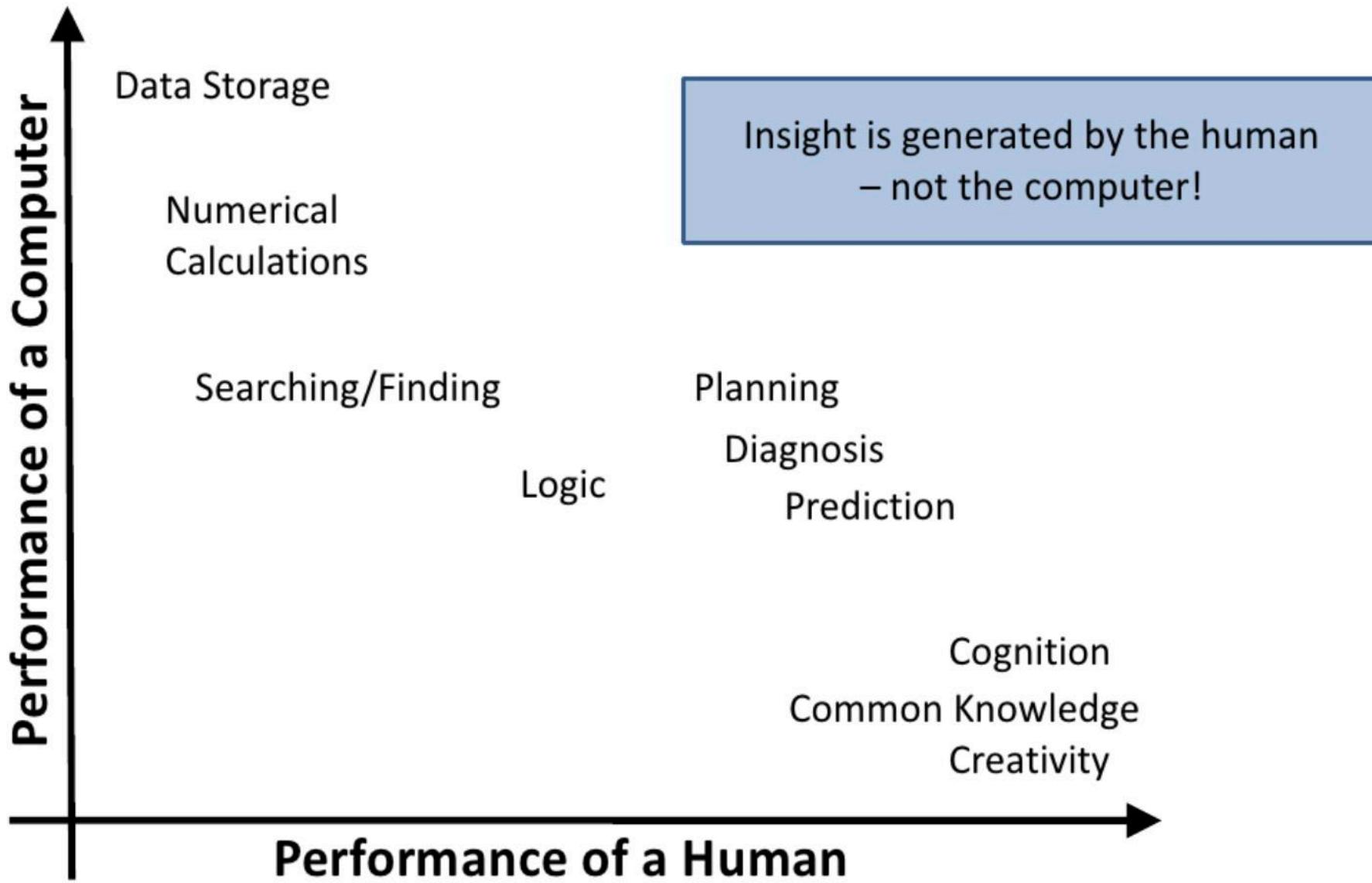
Confirmation

Communication

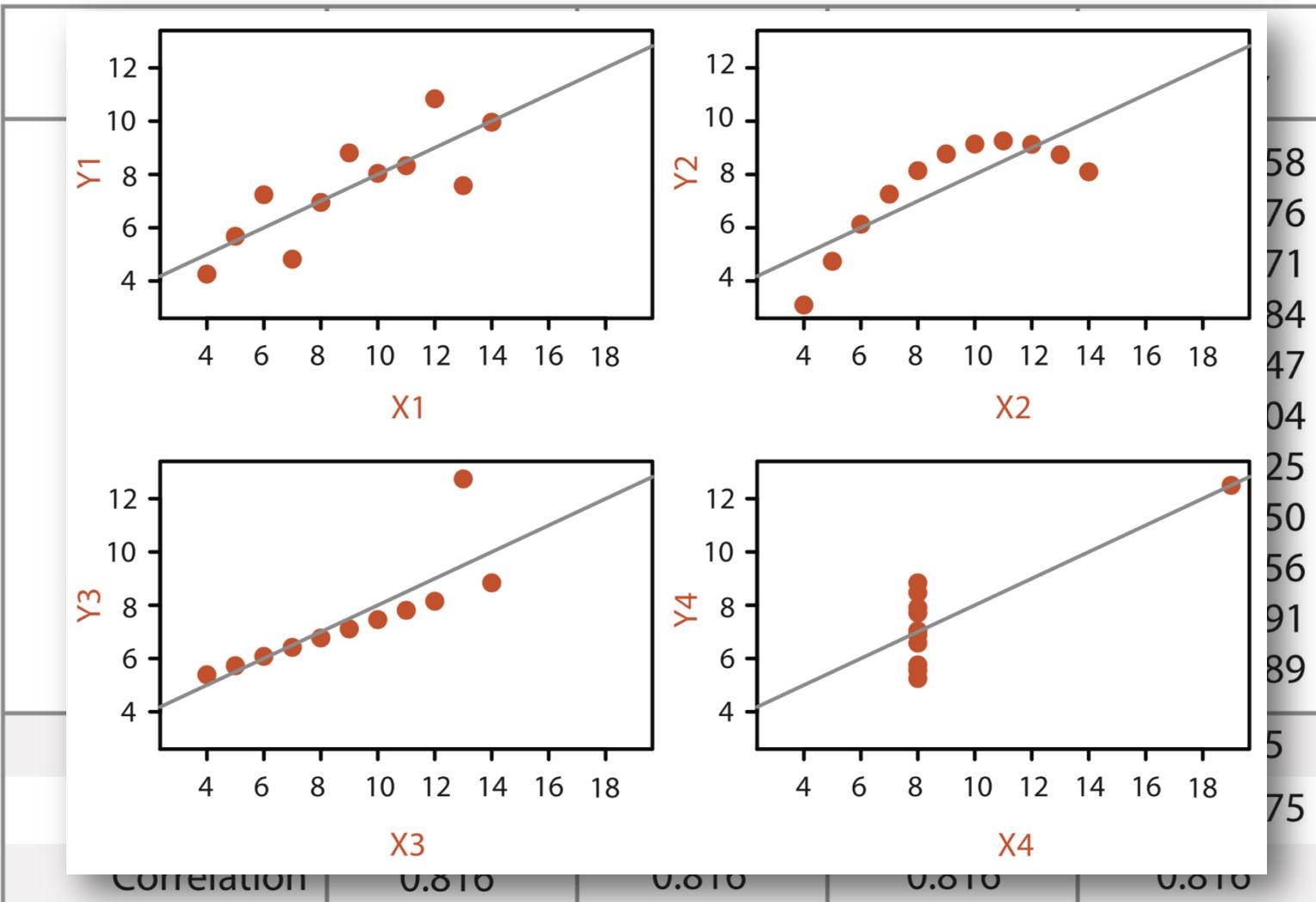


SOURCE: BUREAU OF LABOR STATISTICS, 02/22/2010

# Ability Matrix

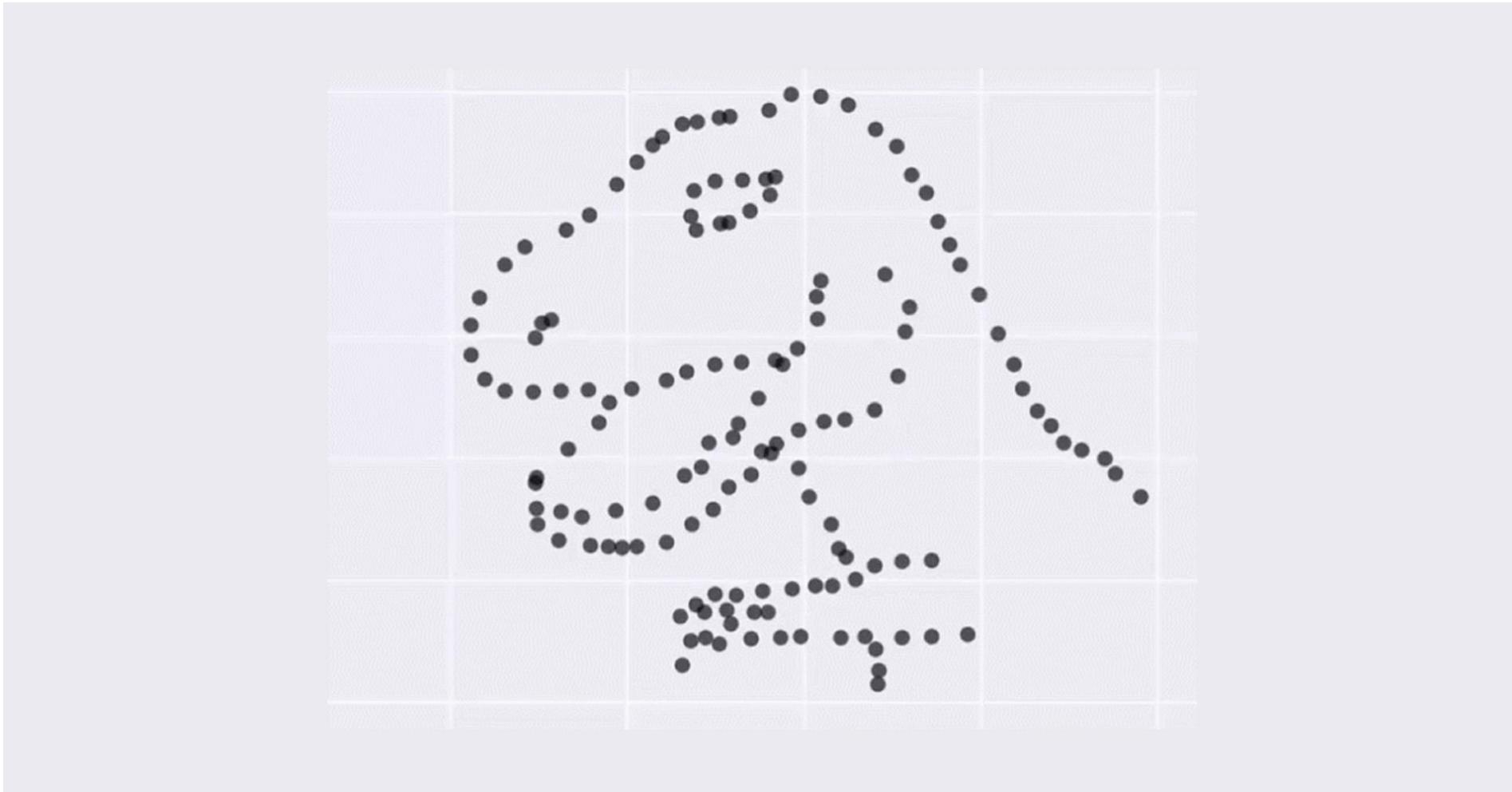


# Why Visualization?



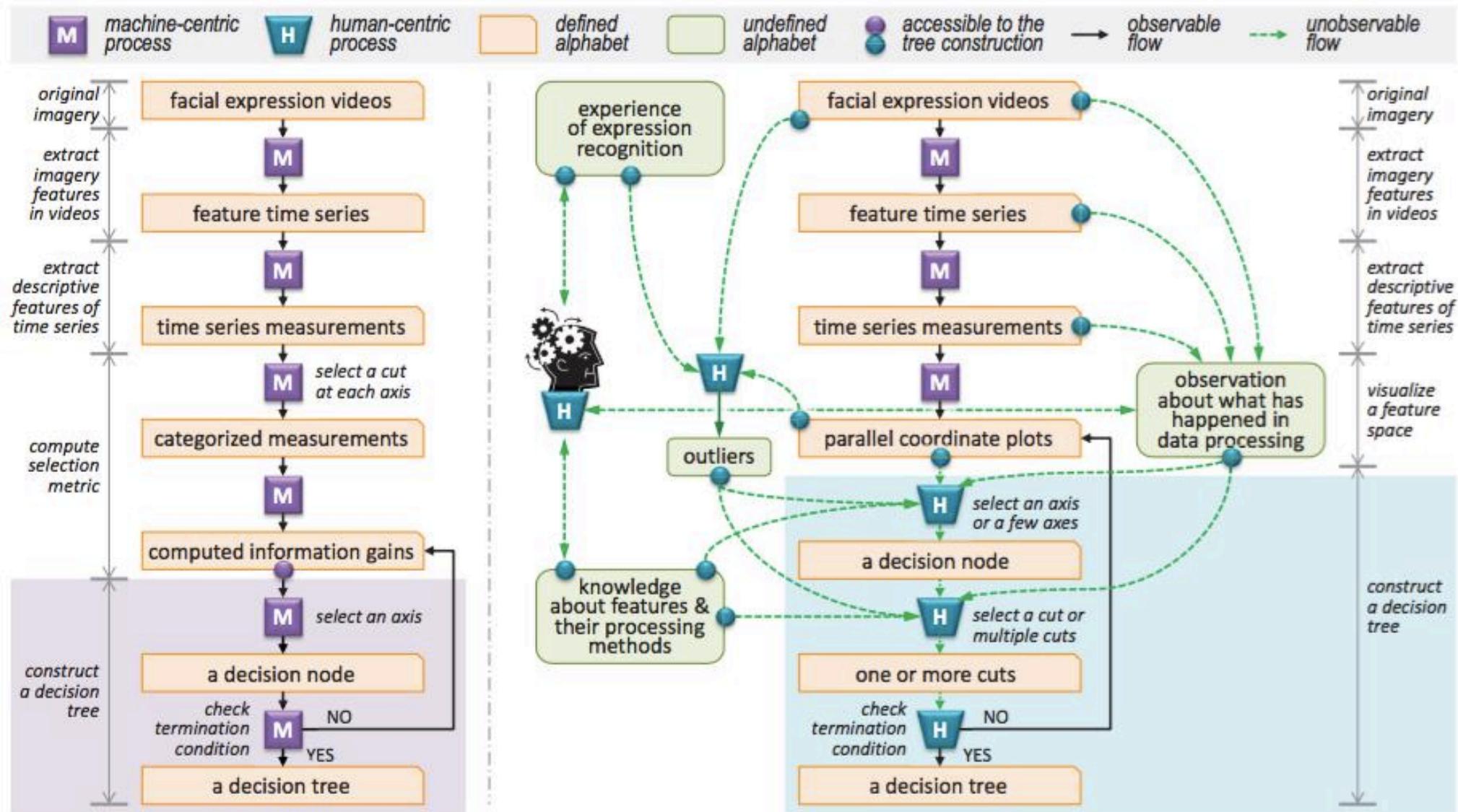
Sometimes summary statistics don't tell the full story.

# 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

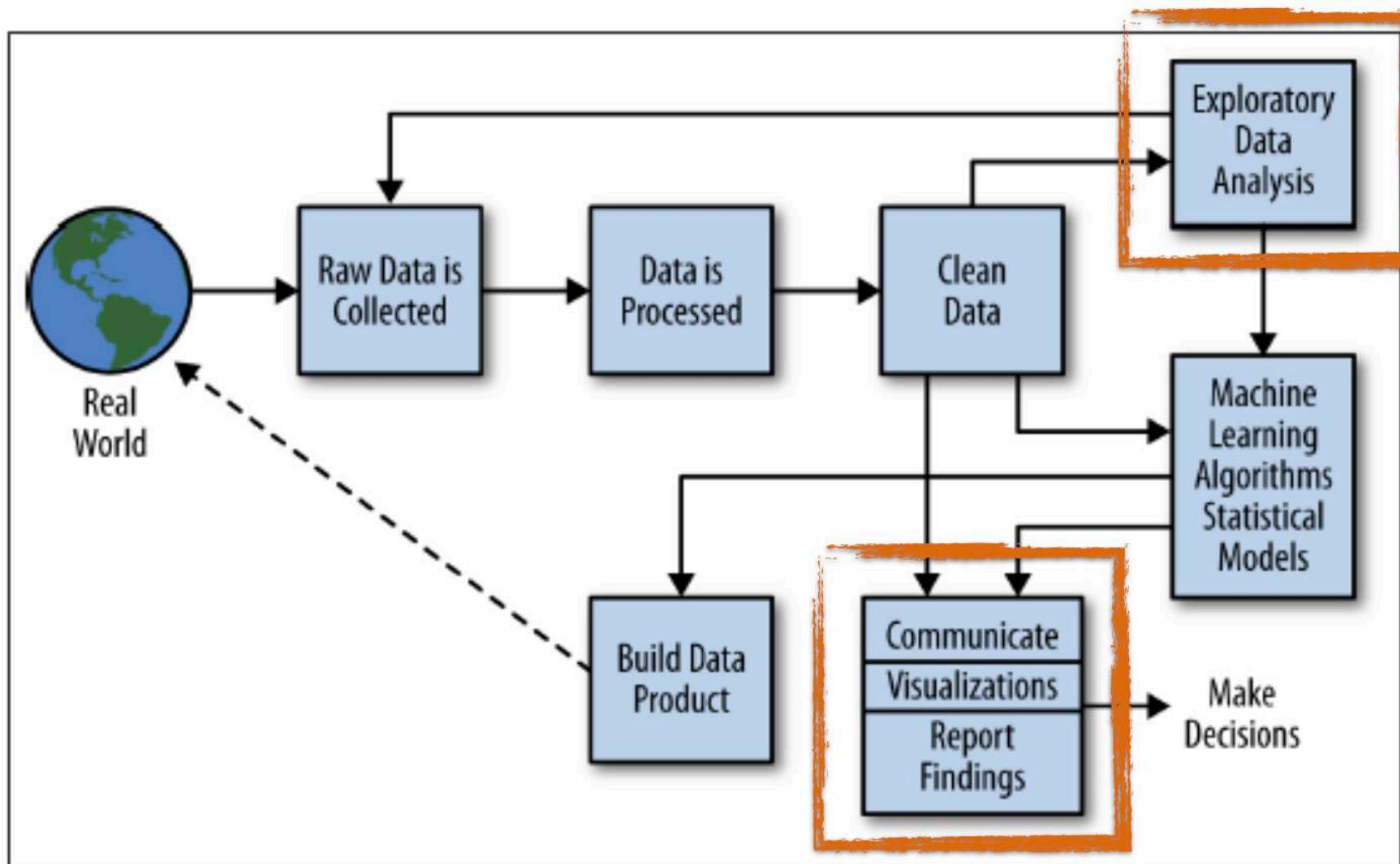


Figure 2-2. The data science process

# Why Humans?

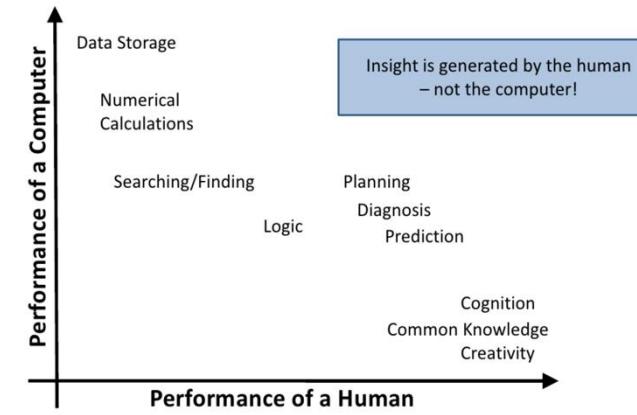
Leverage human capabilities:

**Pattern discovery:** clusters, outliers

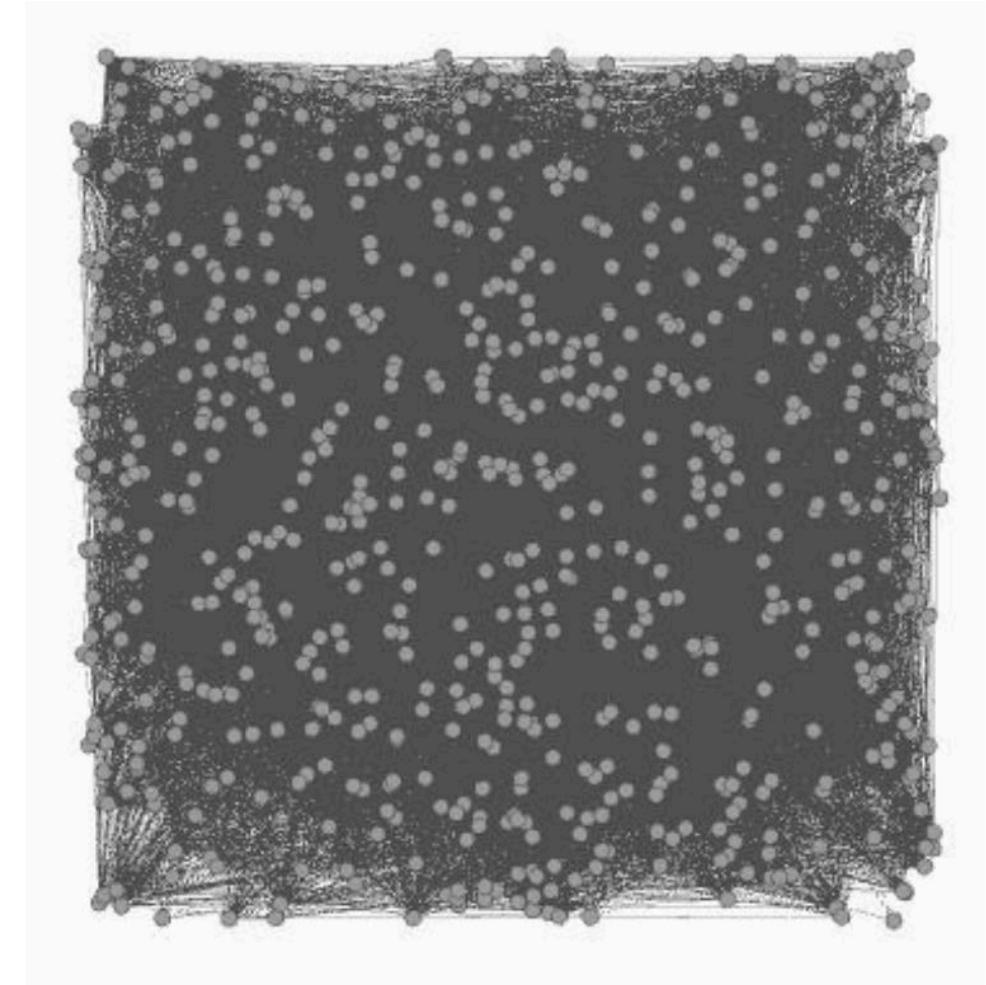
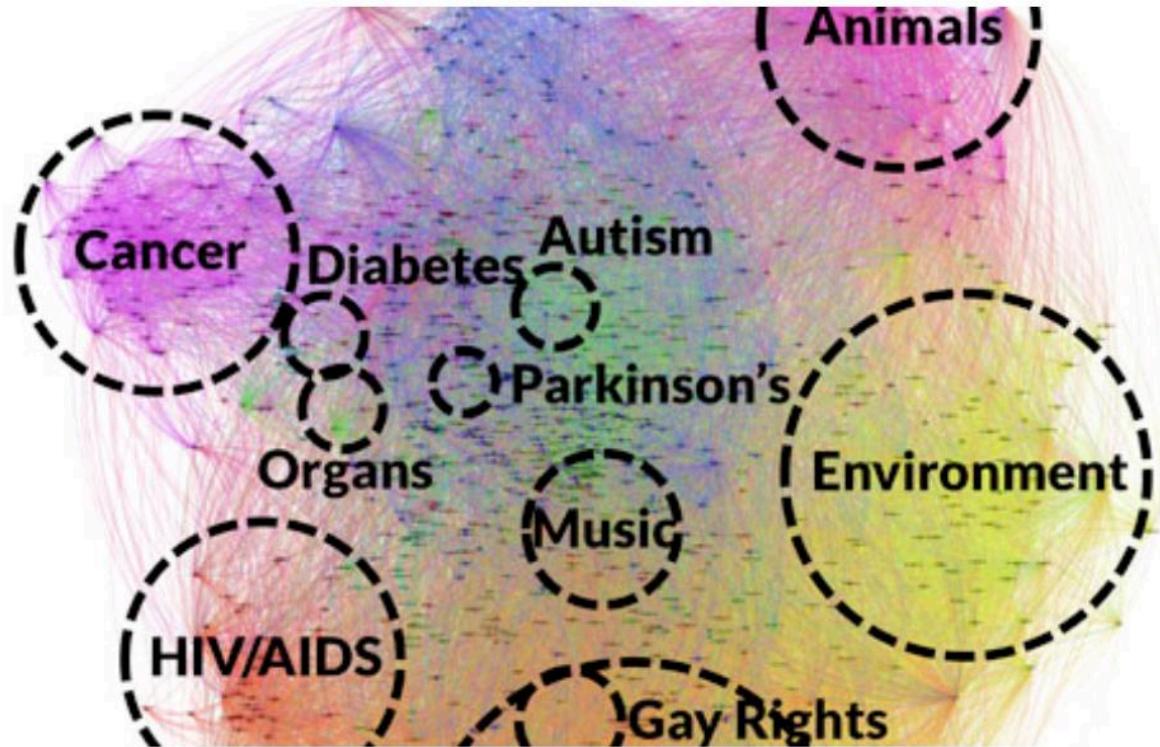
**Contextual knowledge:** expectations, explanations of patterns

**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?*

# 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 appear based on 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

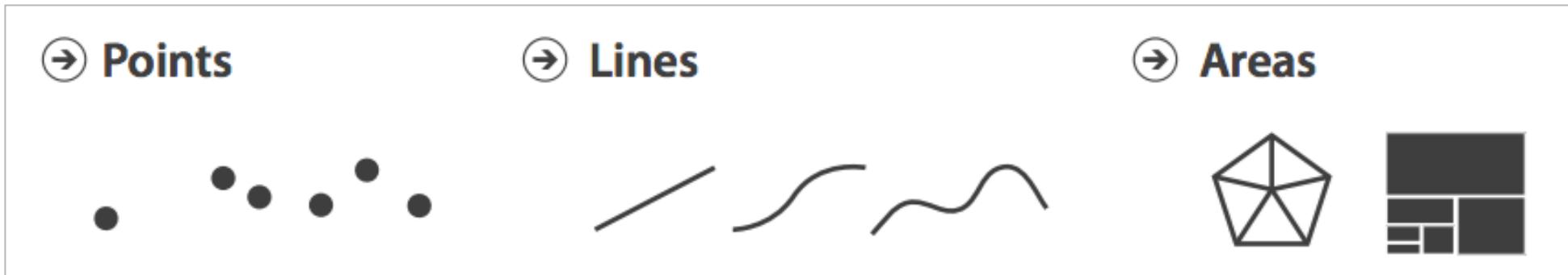
For a complete list of these encodings, see the [Encodings](#) section of the documentation.

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

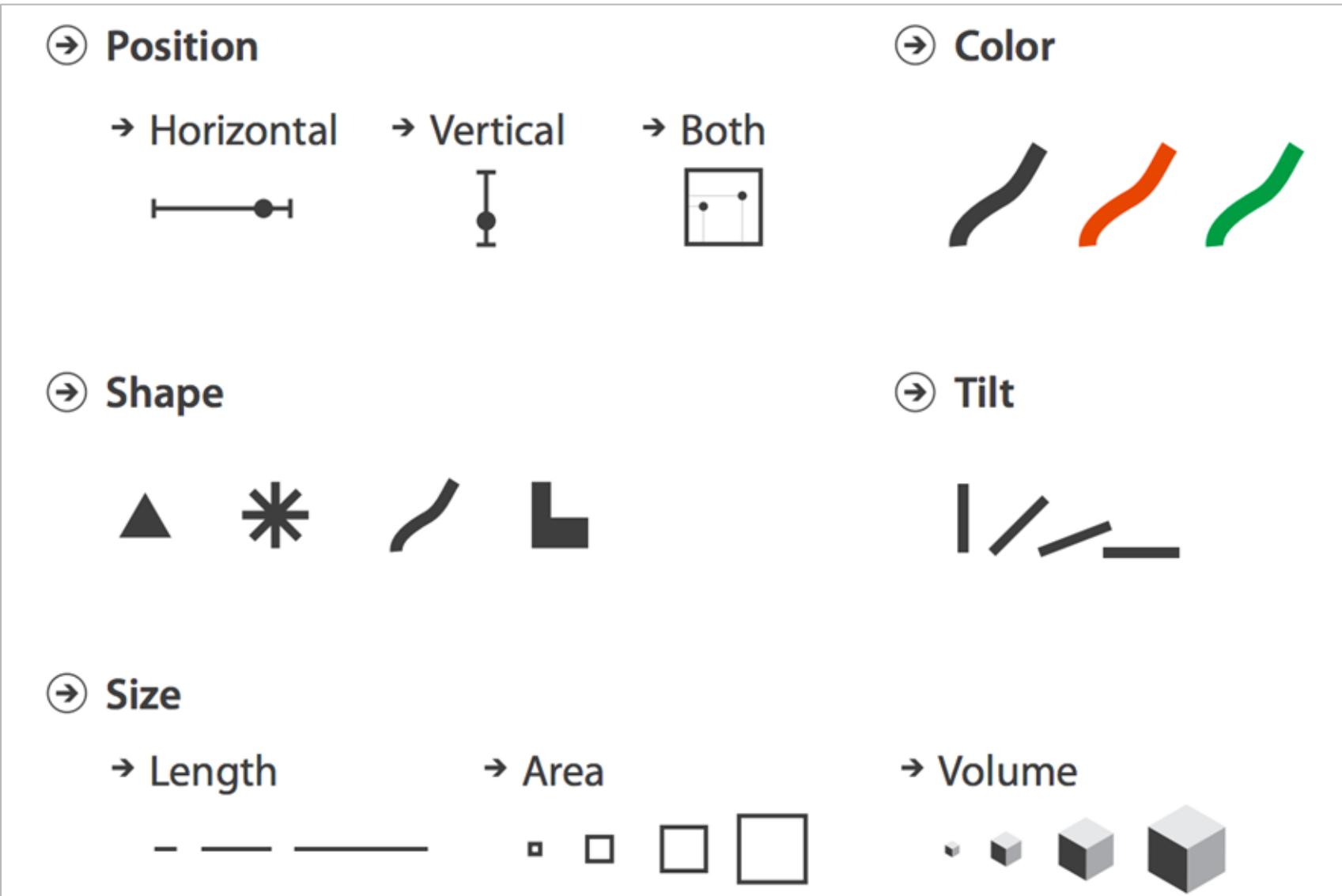
```
alt.Chart(movies_genre).mark_point().encode(
    y='AvgRating', size="Watches"
```

# Marks, Channels, & Encoding

**Marks:** Graphical primitives that encode items / entities



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



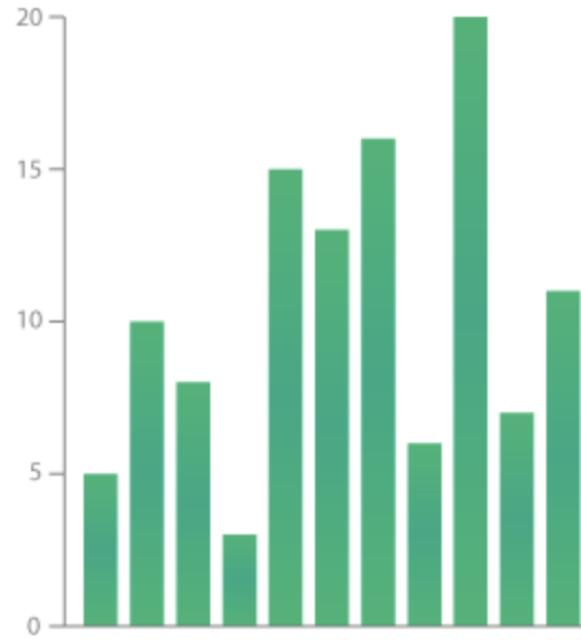
# What are the marks & channels of...



**Pie Chart**

**Marks:** Wedge

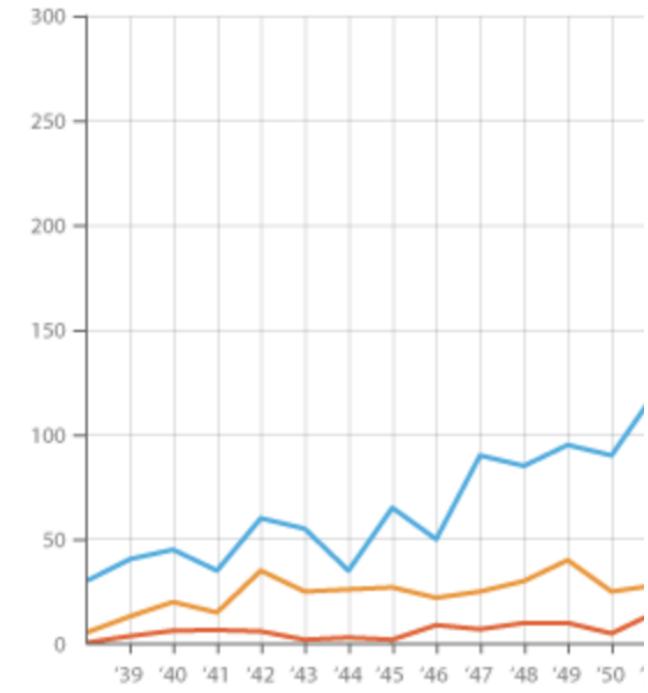
**Channels:** Color, Angle (not Area, not Size!)



**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

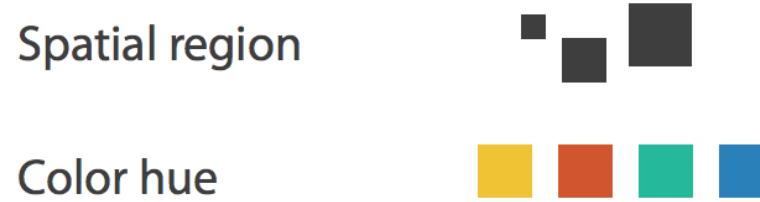
**Effectiveness Principle:** The more important the data/attribute, the more **salient** the encoding should be

- Important things should be noticeable

## → Magnitude Channels: Ordered Attributes

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)	

## → 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



Color hue



Motion



Shape



▲ Most Effective  
— Effectiveness  
— Same

# Which has order? Which doesn't?

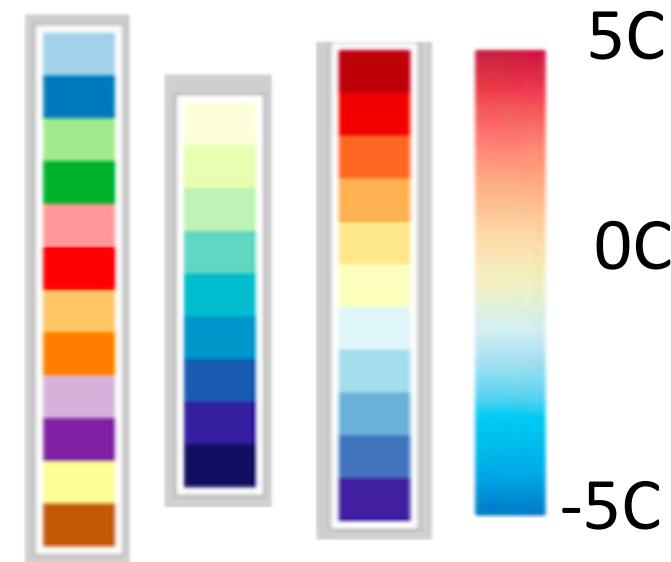


Vary the appropriate dimension depending on your data type

# 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



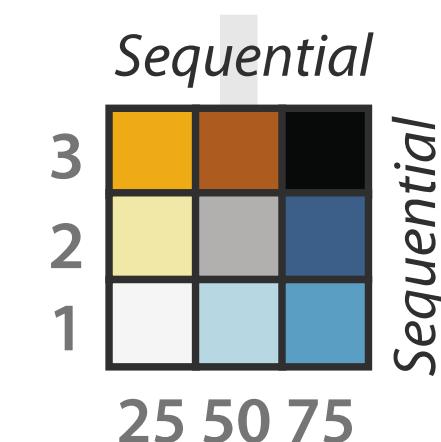
Diverging vs. Sequential



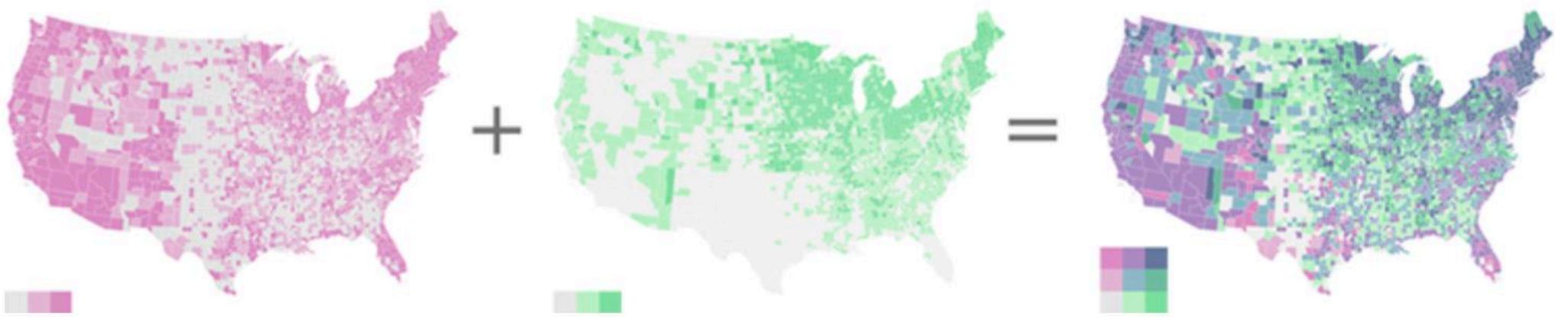
Segmented vs. Continuous



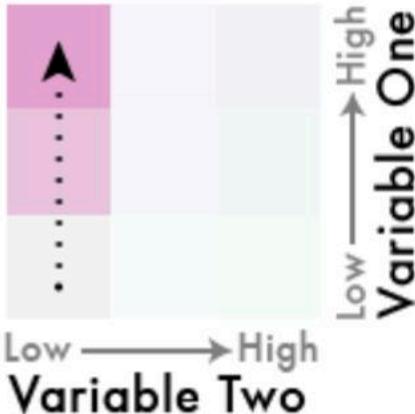
Univariate vs. Bivariate



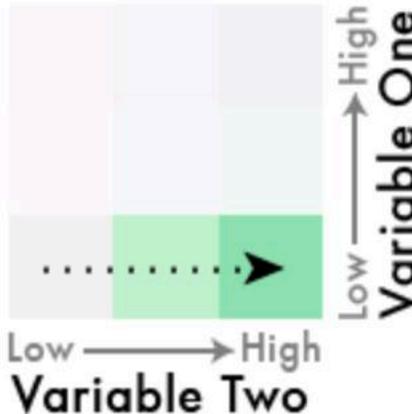
# Bivariate Example



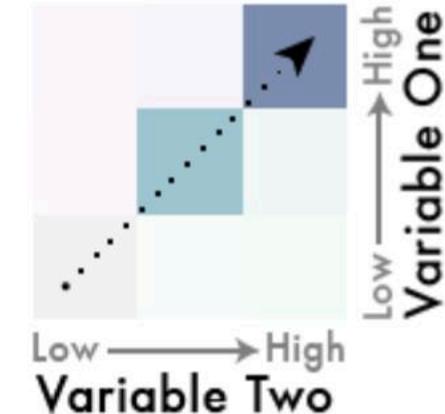
Data Strongly Reflect  
Variable One



Data Strongly Reflect  
Variable Two



Data Show Agreement  
Between Both Variables

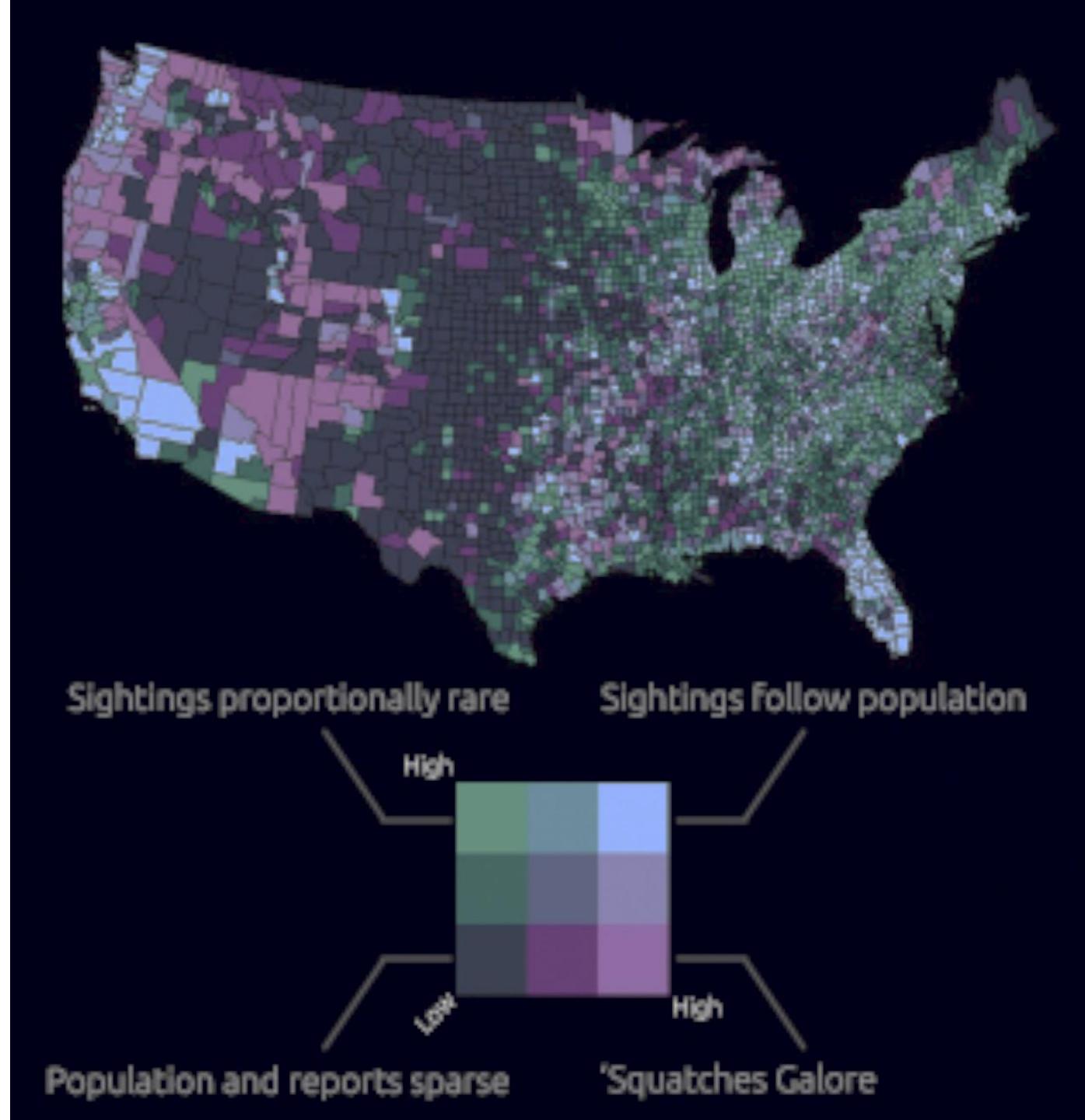


# Bivariate Example

Population  
&  
Sasquatch (BigFoot) Sightings

...not so easy to read.

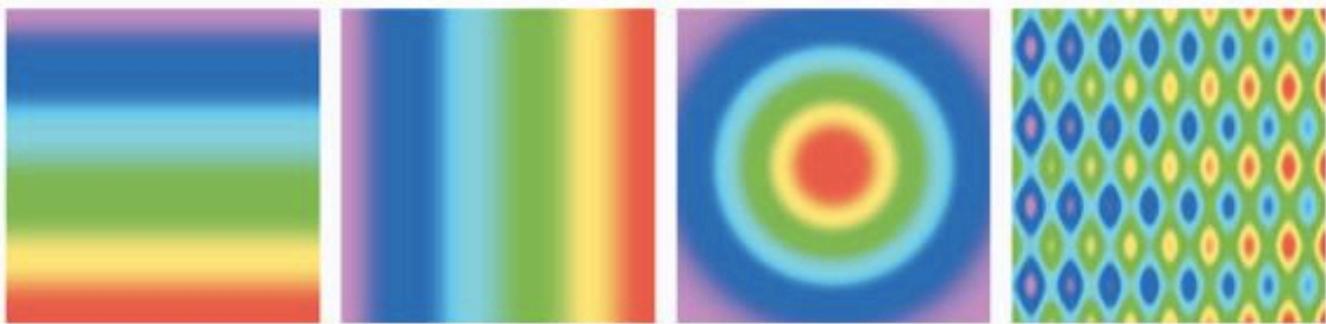
Consider also using  
hue/lightness,  
hue/thickness of border,  
hue + mark&channel...



# The Dangers of Rainbows

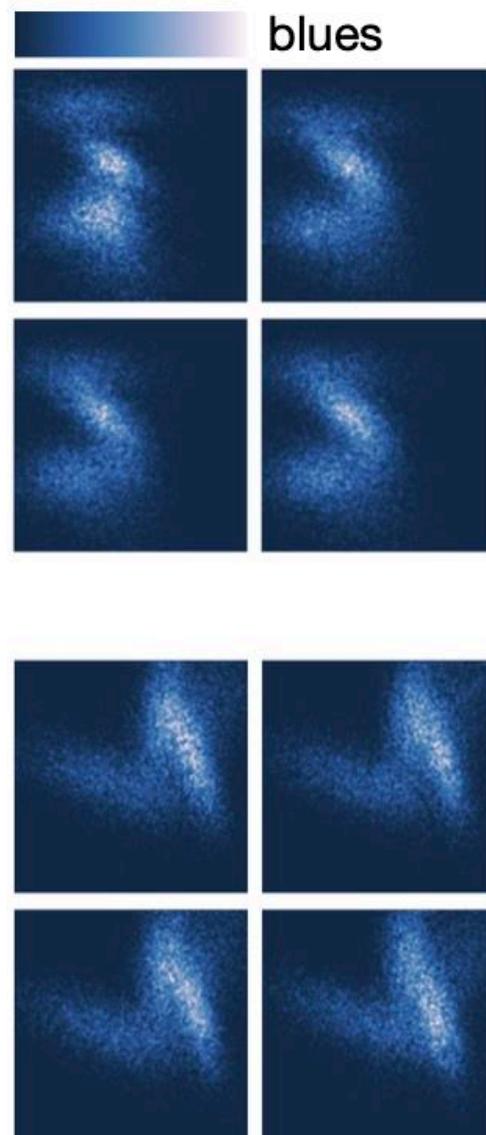
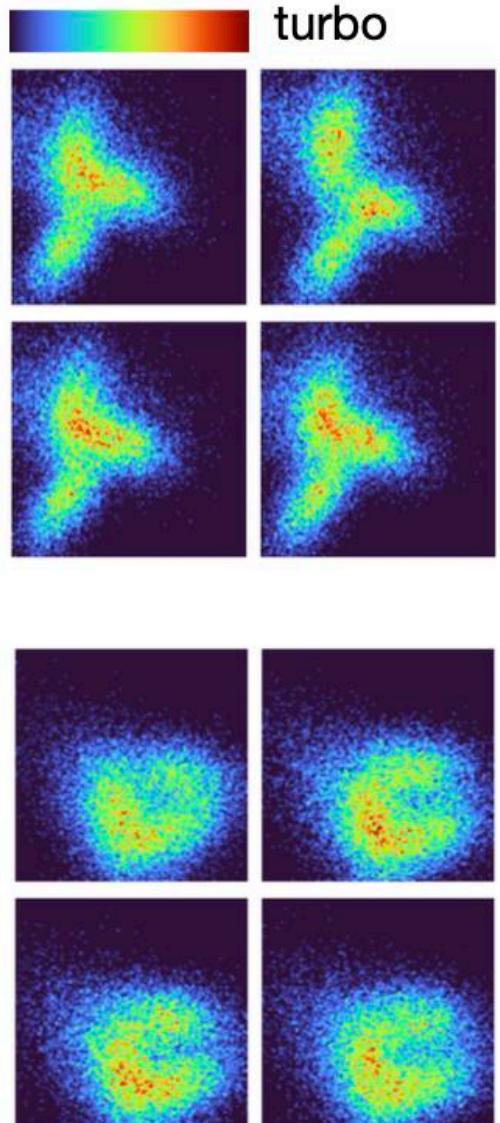
The common rainbow color map is often a poor choice due to:

- 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)

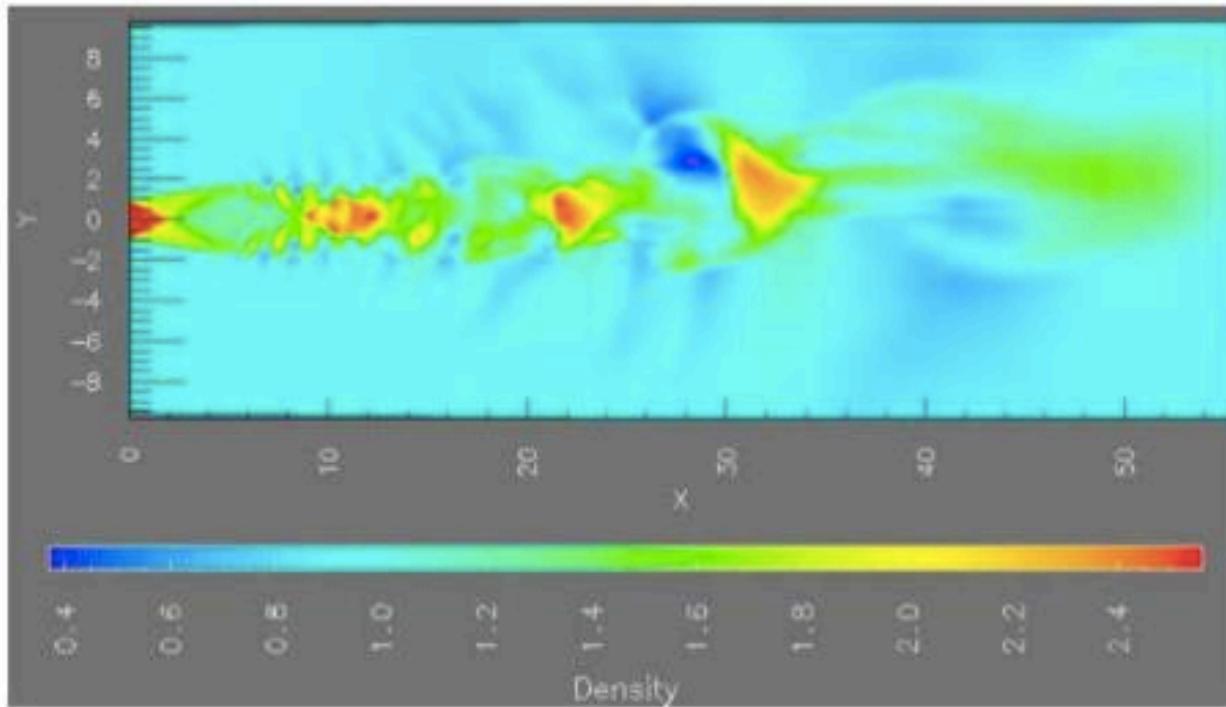
# The Dangers of Dismissing Rainbows



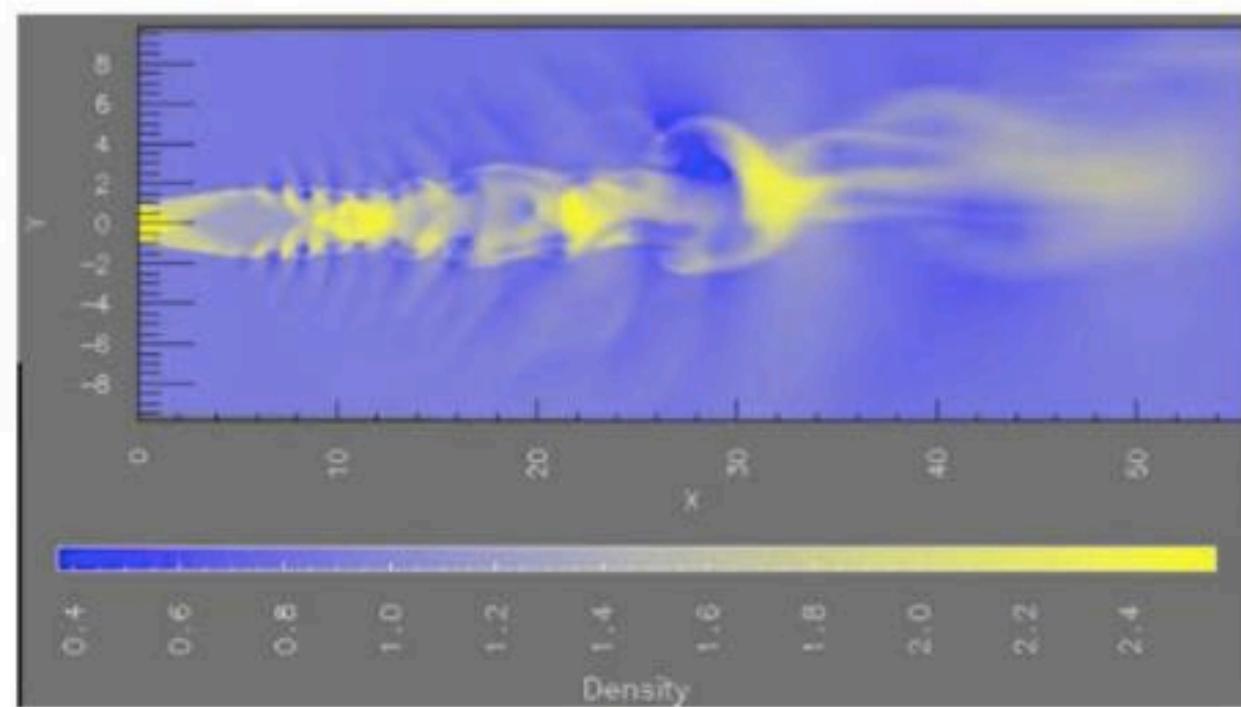
Some rainbow color maps may be helpful even in continuous tasks.

- 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

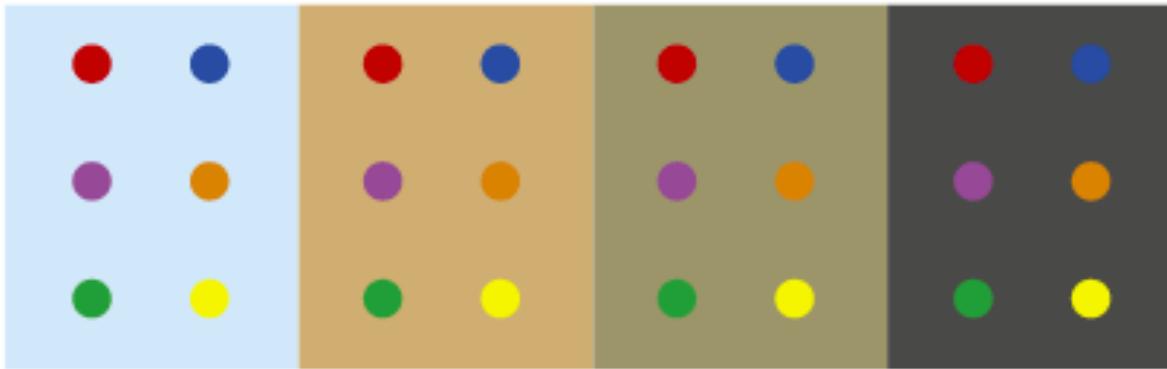
May assign colors hierarchically

- 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

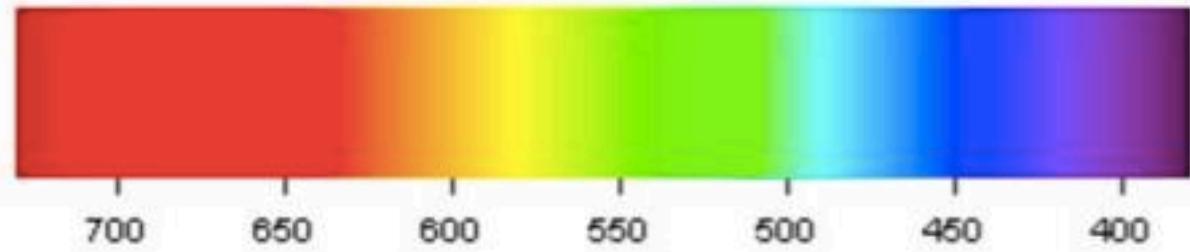


What happens when you take off tinted goggles?

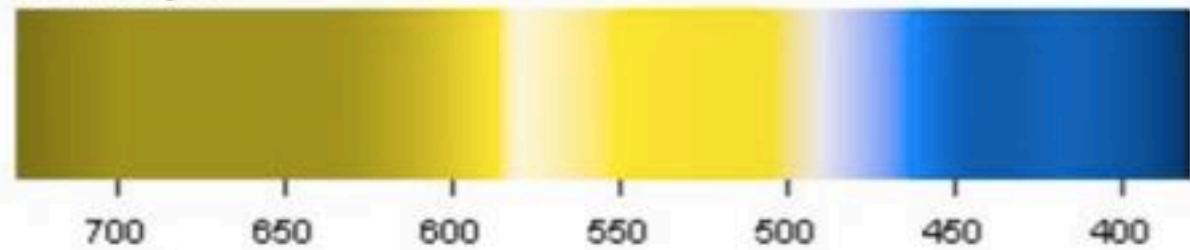
- To avoid contrast issues with background, surround points with white or black lines.
- Make sure colors do not have same luminance as background

**Color Vision Deficiencies affect ~7% of the population**

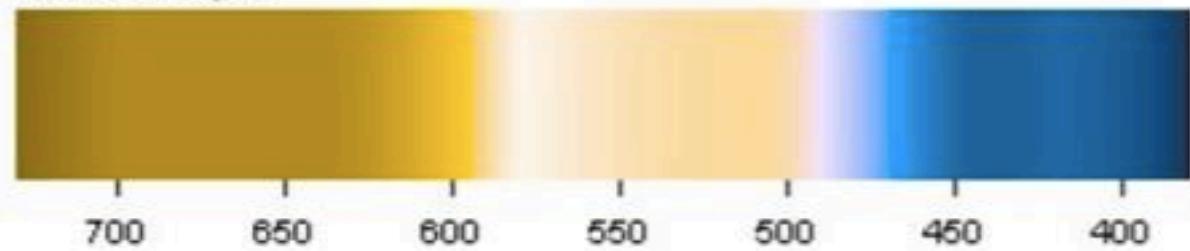
**Normal**



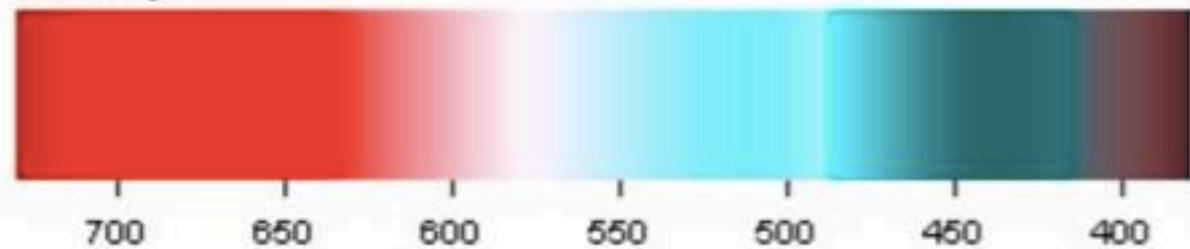
**Protanopia**



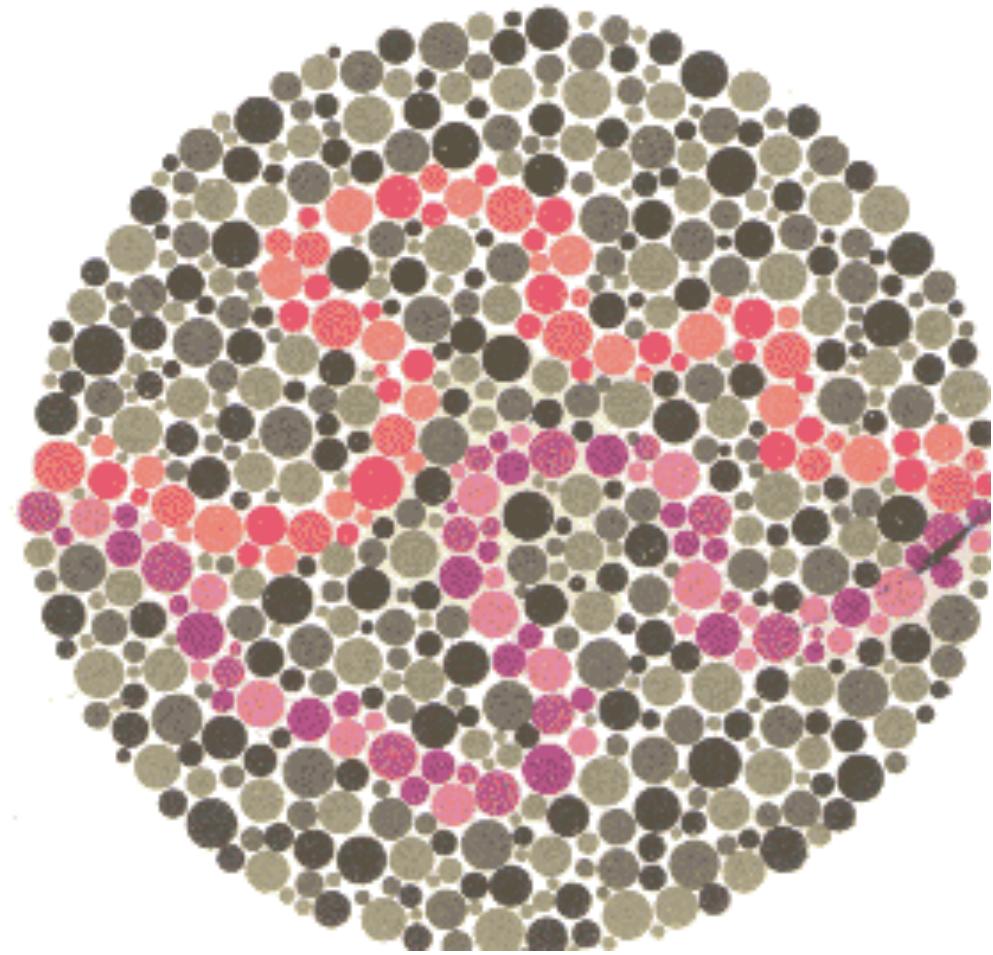
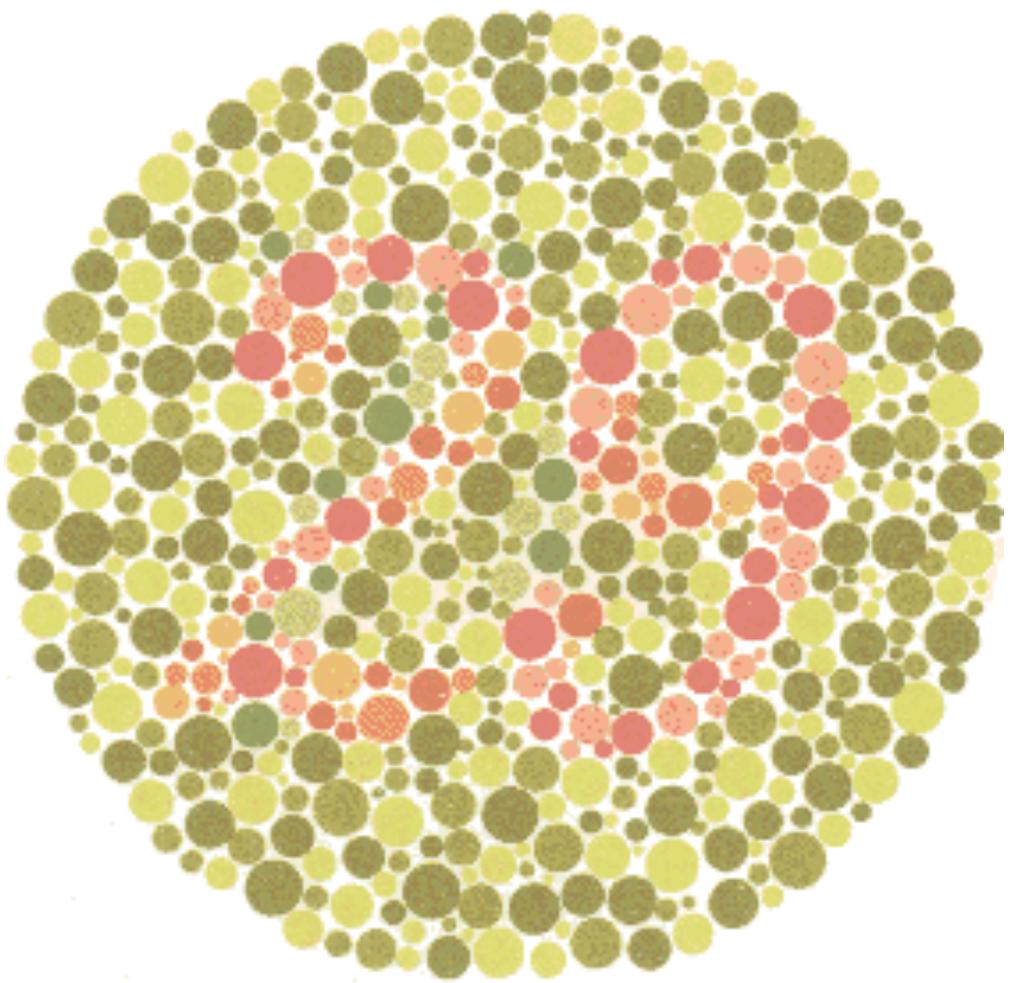
**Deuteranopia**



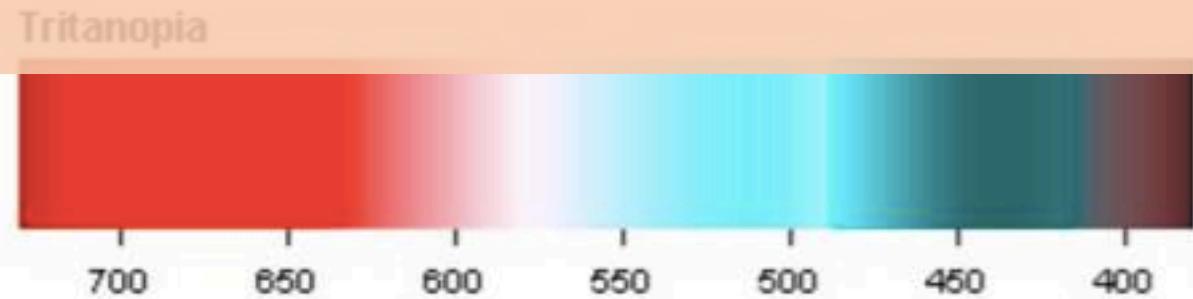
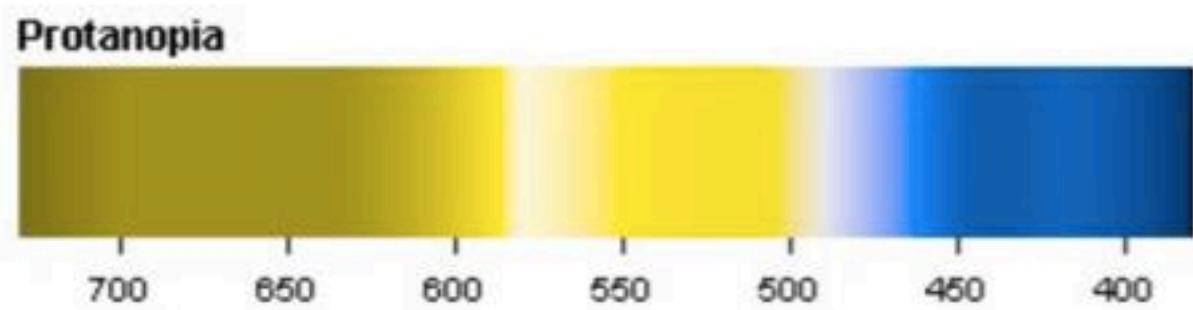
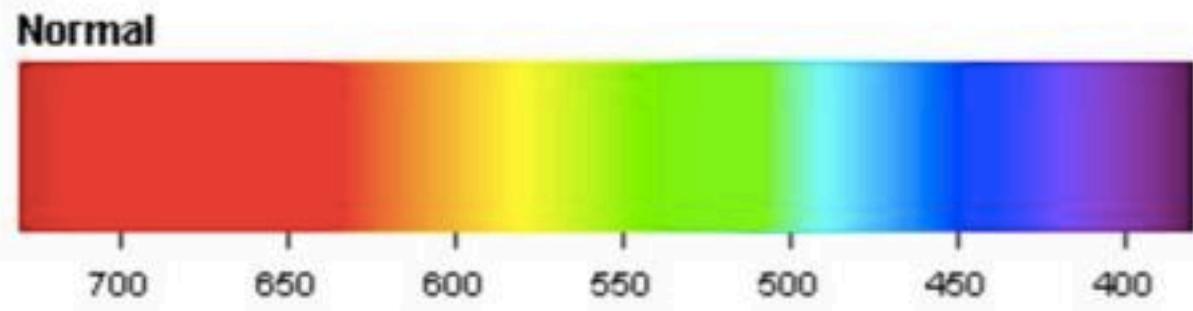
**Tritanopia**



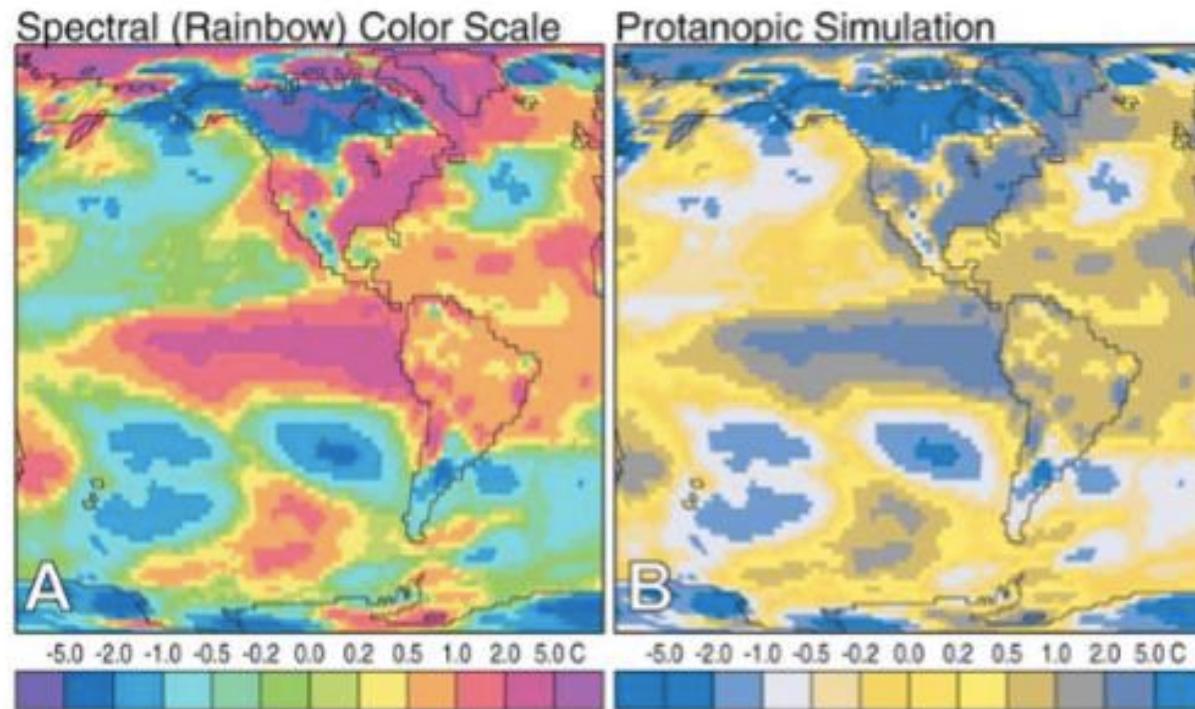
# Ishihara Plates are used to test for Color Vision Deficiencies



**Color Vision Deficiencies affect ~7% of the population**



# Rainbow Colormap & Color Vision Deficiency



***Guideline: “Get it right in black and white”***

Number of data classes: 7

how to use | updates | downloads | credits

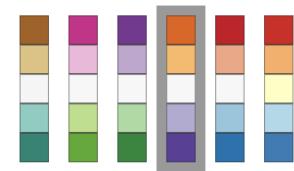
COLORBREWER 2.0

color advice for cartography

Nature of your data:

sequential  diverging  qualitative

Pick a color scheme:



Only show:

- colorblind safe
- print friendly
- photocopy safe

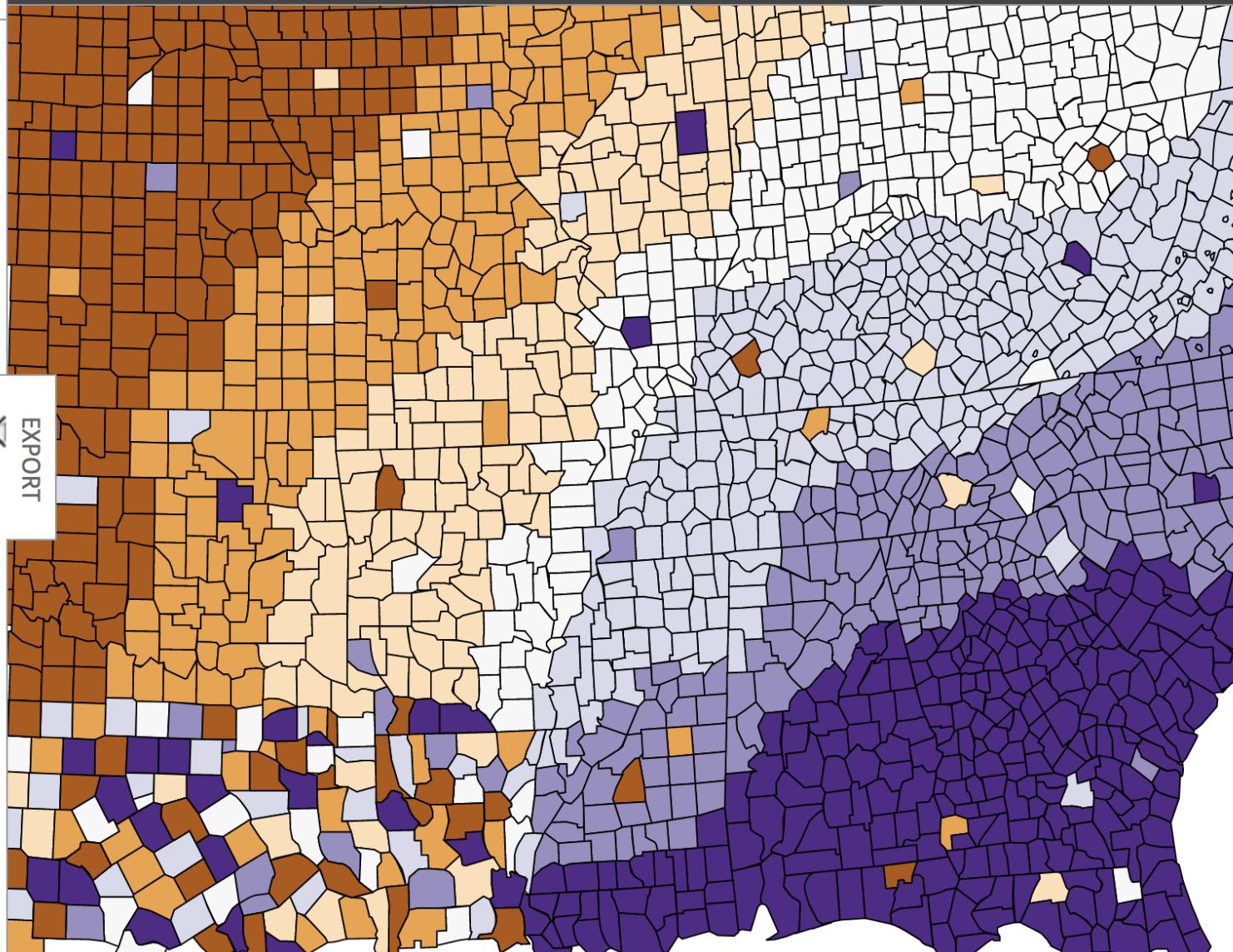
Context:

- roads
- cities
- borders

Background:

- solid color
- terrain

color transparency



7-class PuOr

EXPORT

HEX

▼

- #b35806
- #f1a340
- #fee0b6
- #f7f7f7
- #d8daeb
- #998ec3
- #542788

© Cynthia Brewer, Mark Harrower and The Pennsylvania State University

[Source code and feedback](#)

[Back to Flash version](#)

[Back to ColorBrewer 1.0](#)

axismaps

# *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



Motion



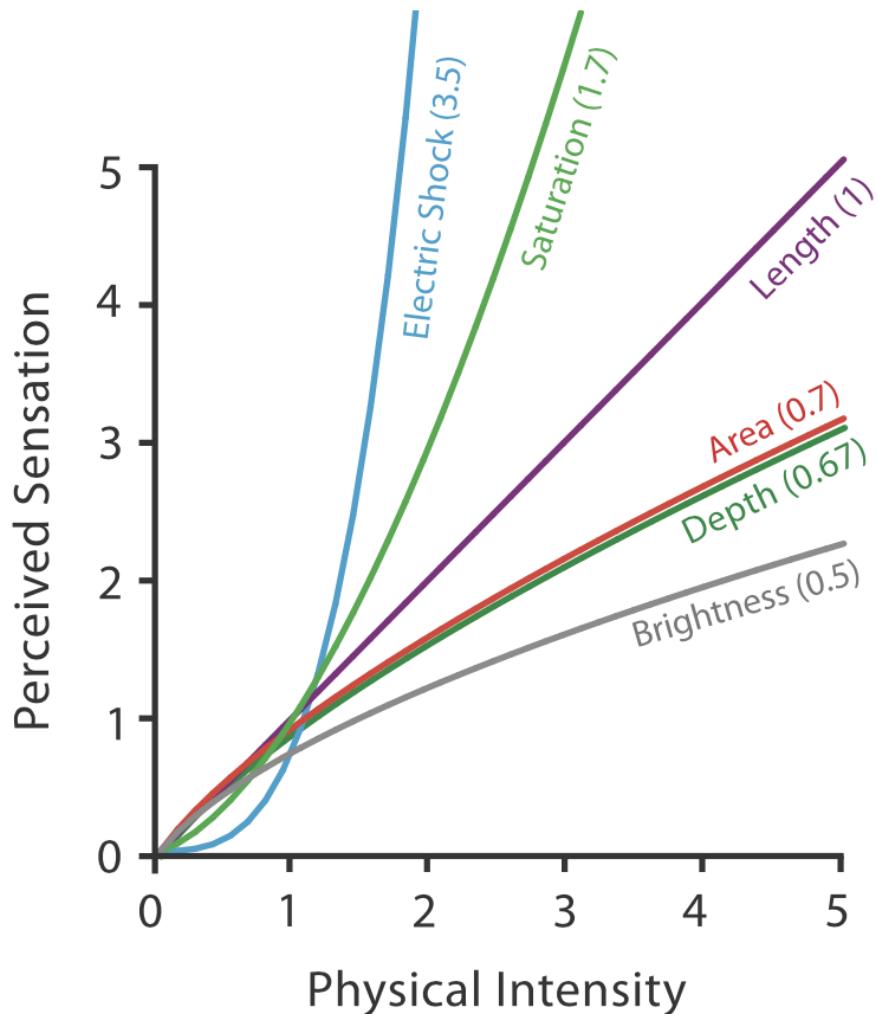
Shape



▲ Most  
— Effectiveness  
↓ Same

# Where do these rankings come from?

Steven'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](https://en.wikipedia.org/wiki/Stevens%27s_power_law)

**\*In part**

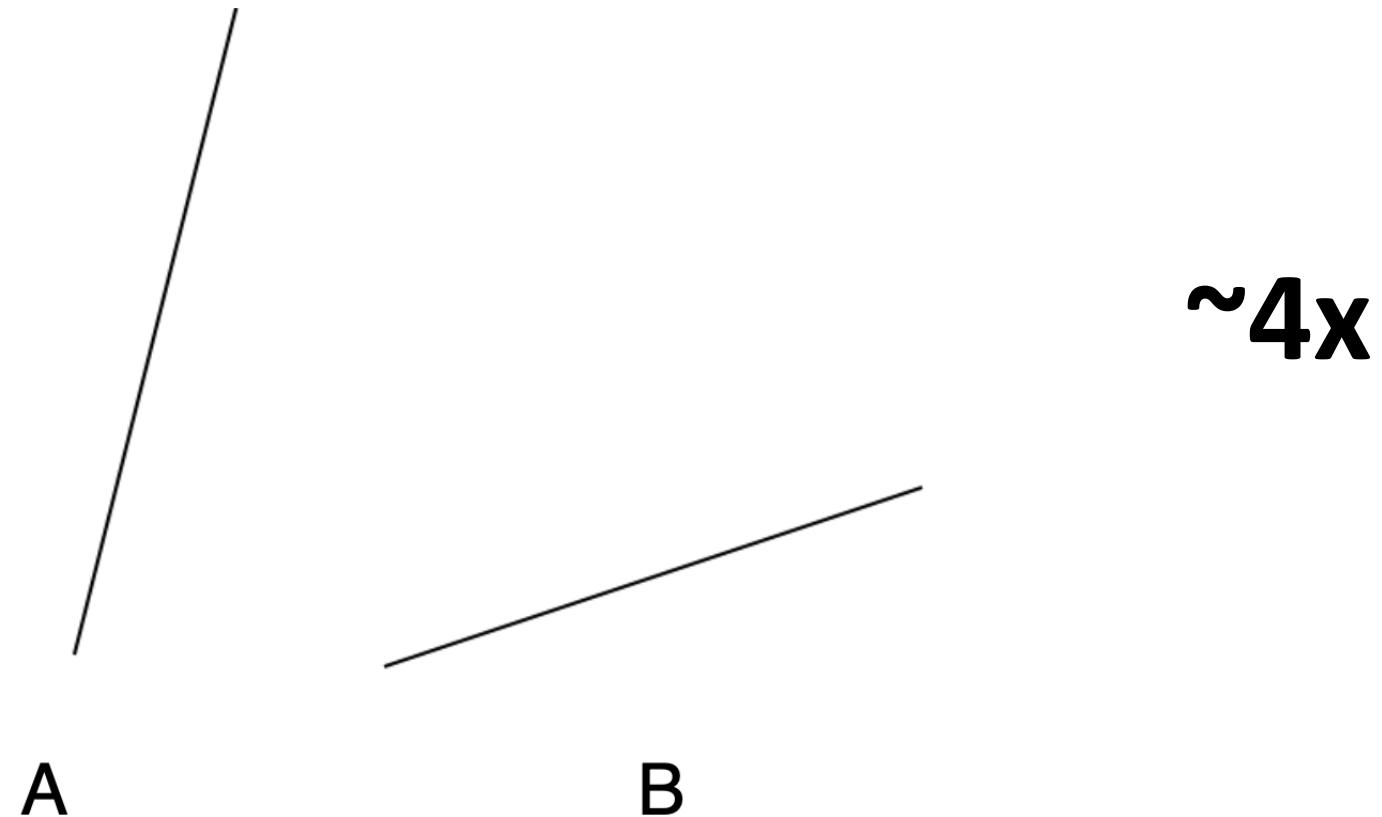
# How Much Longer?



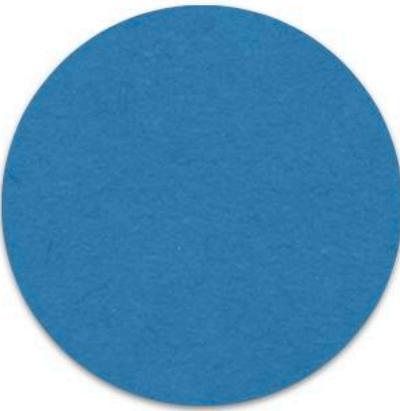
# How Much Longer?



# How Much Steeper?



# How Much Larger?



**5x**

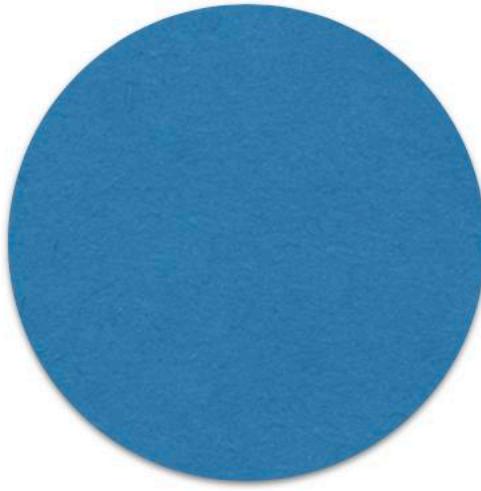
**A**

**B**

# How Much Larger?



A



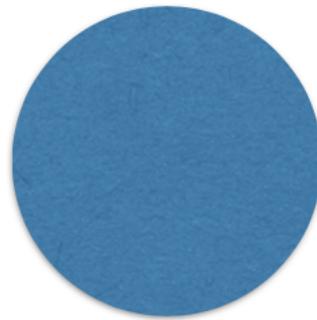
B

**4x area  
2x diameter**

# How Much Larger (by area)?



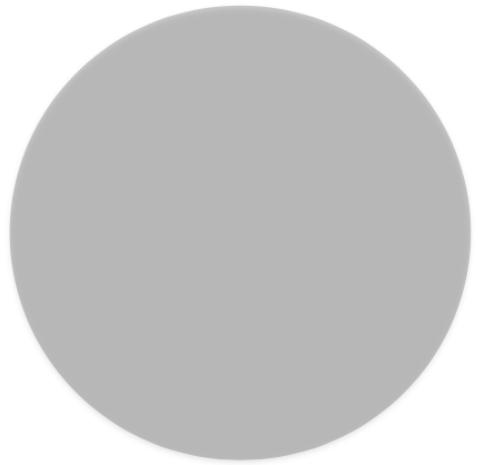
A



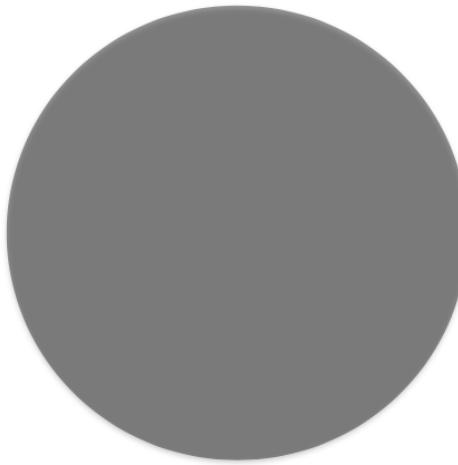
B

3x

# How Much Darker?



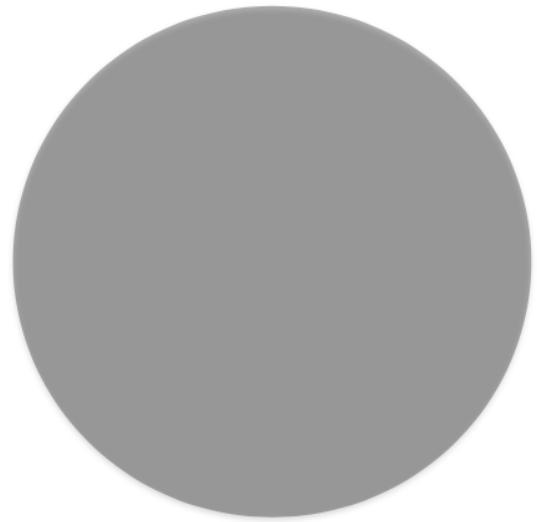
A



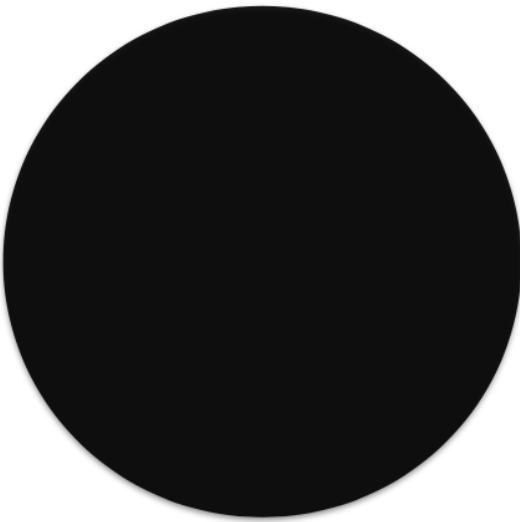
B

**2x**

# How Much Darker?



A



B

3x

# Other factors affect accuracy too...

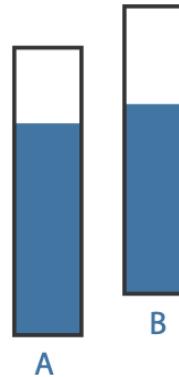
Alignment

Distractors

Distance

Common scale

...



# 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  
— Effectiveness  
↓ Same

# What **chart** should I use? ...it depends on what you are trying to show! **Visualization vocabularies can help!**

## Financial Times Visualization Vocabulary

Deviation Correlation Change v Time Ranking Distribution Part to whole Magnitude Spatial Flow

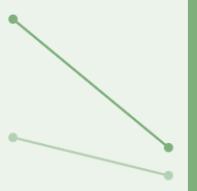
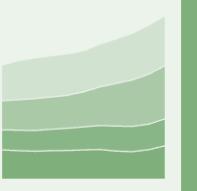
### Change v Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader

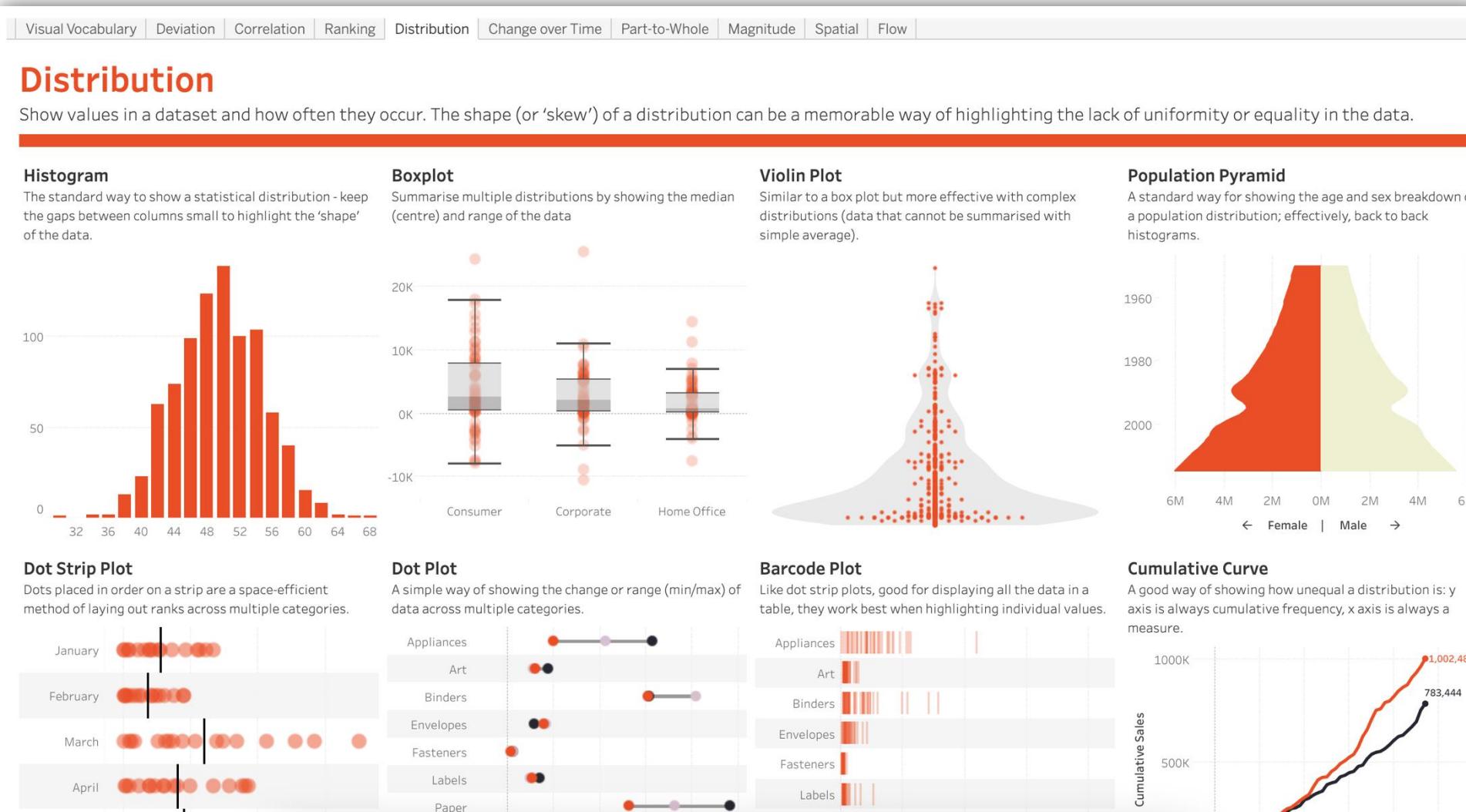
#### Examples of use

Share price movements, economic time series

#### Chart types

line	column-timeline	column-line-timeline	stock-price	slope	area
					
The standard way to show a changing time series. If data are irregular,	Columns work well for showing change over time - but usually best with	A good way of showing the relationship over time between an	Usually focused on day-to-day activity, these charts show opening/closing and	Good for showing changing data as long as the data can be simplified	Use with care. These are good at showing changes to total, but seeing

# What **chart** should I use? ...it depends on what you are trying to show! **Visualization vocabularies** can help!



## Tableau Visualization Vocabulary

# 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

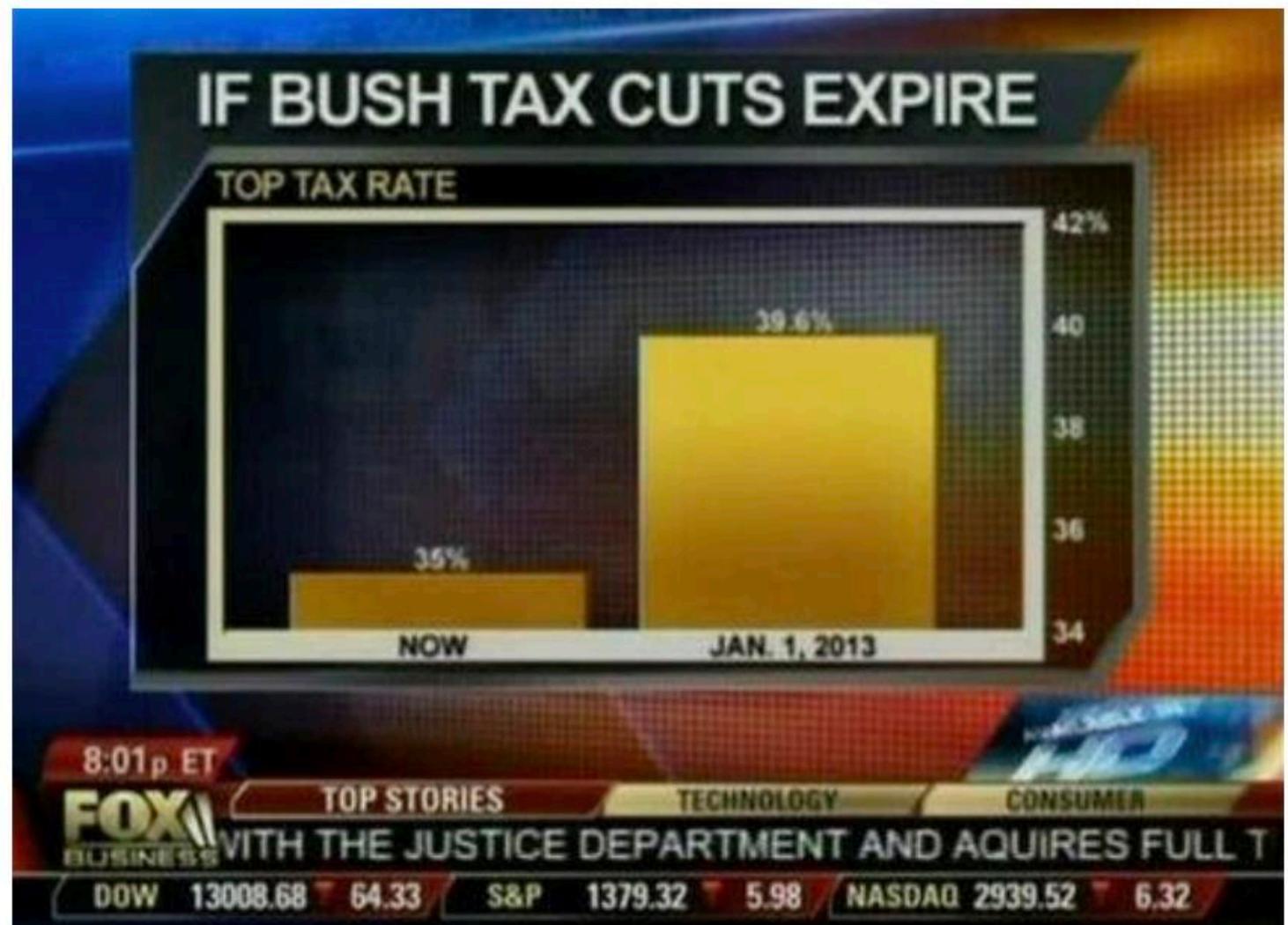
---

Size of the effect in data

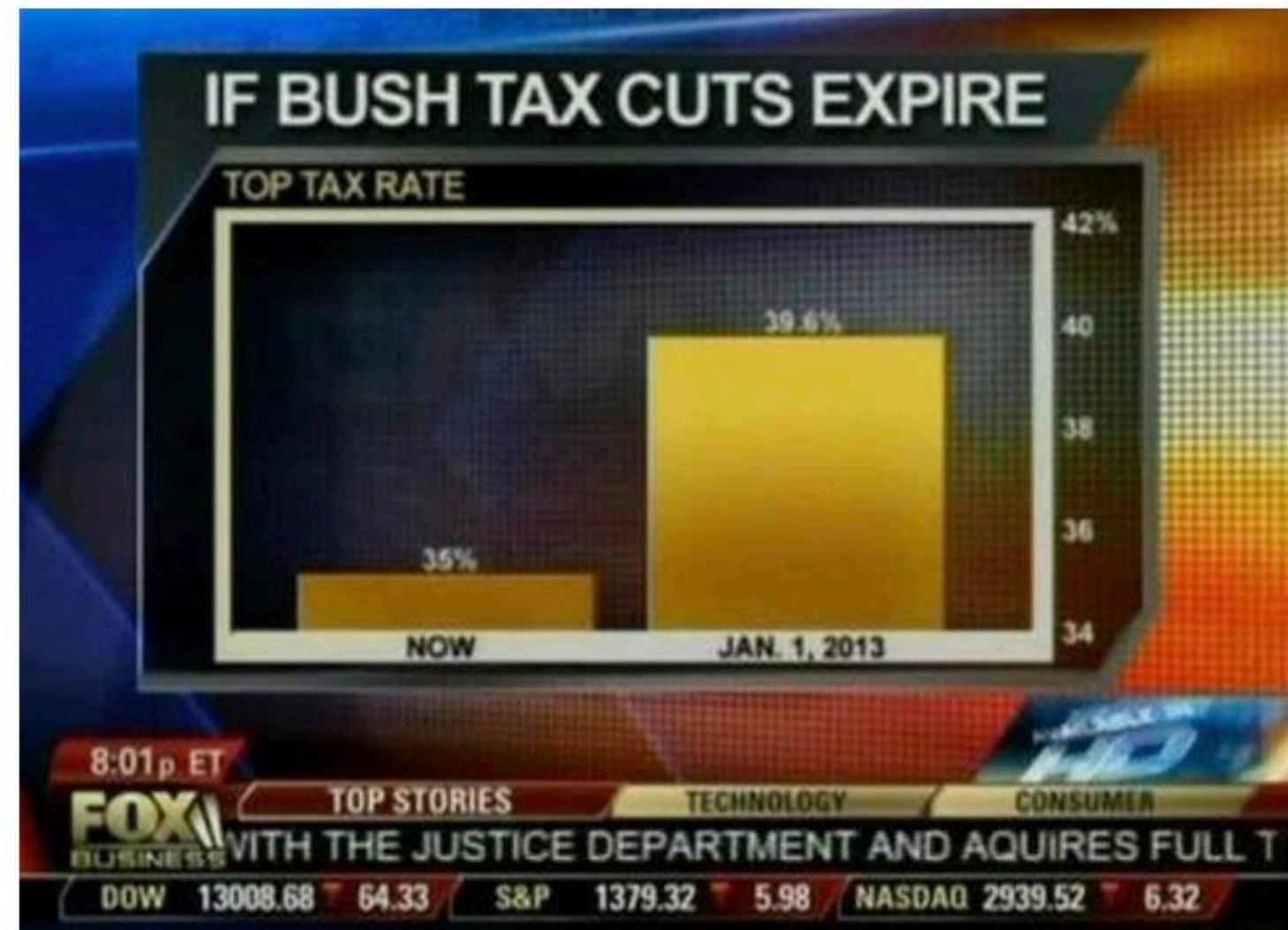
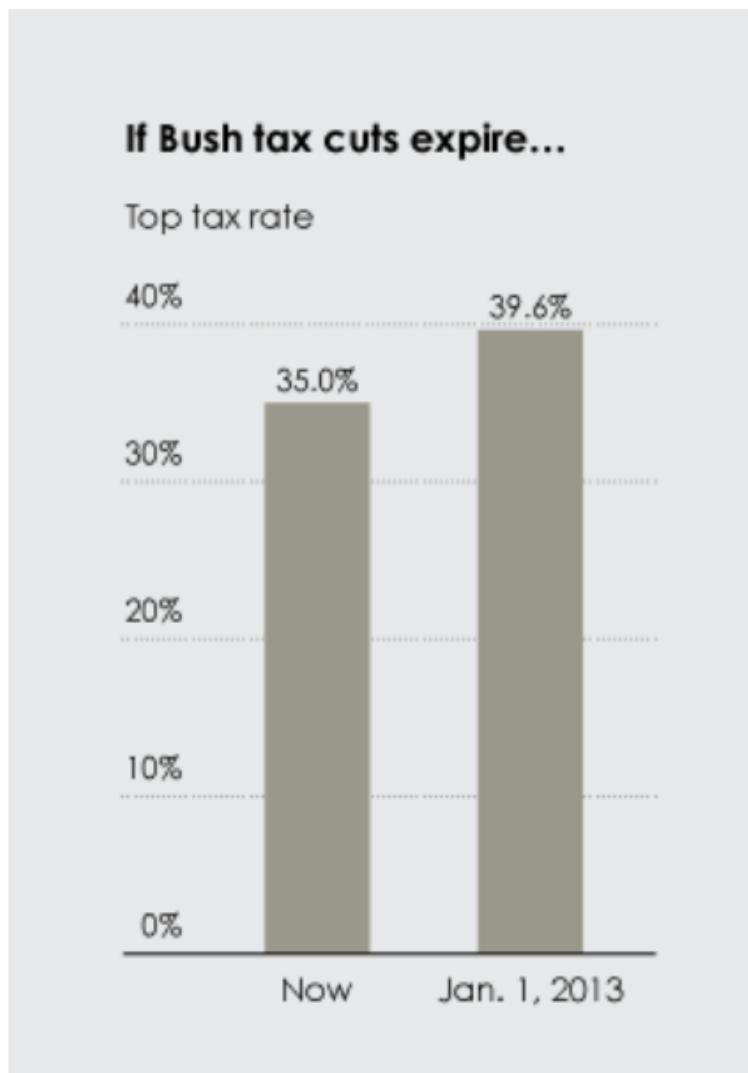
# Graphical Integrity

Magnitude in  
data must  
correspond to  
magnitude of  
mark

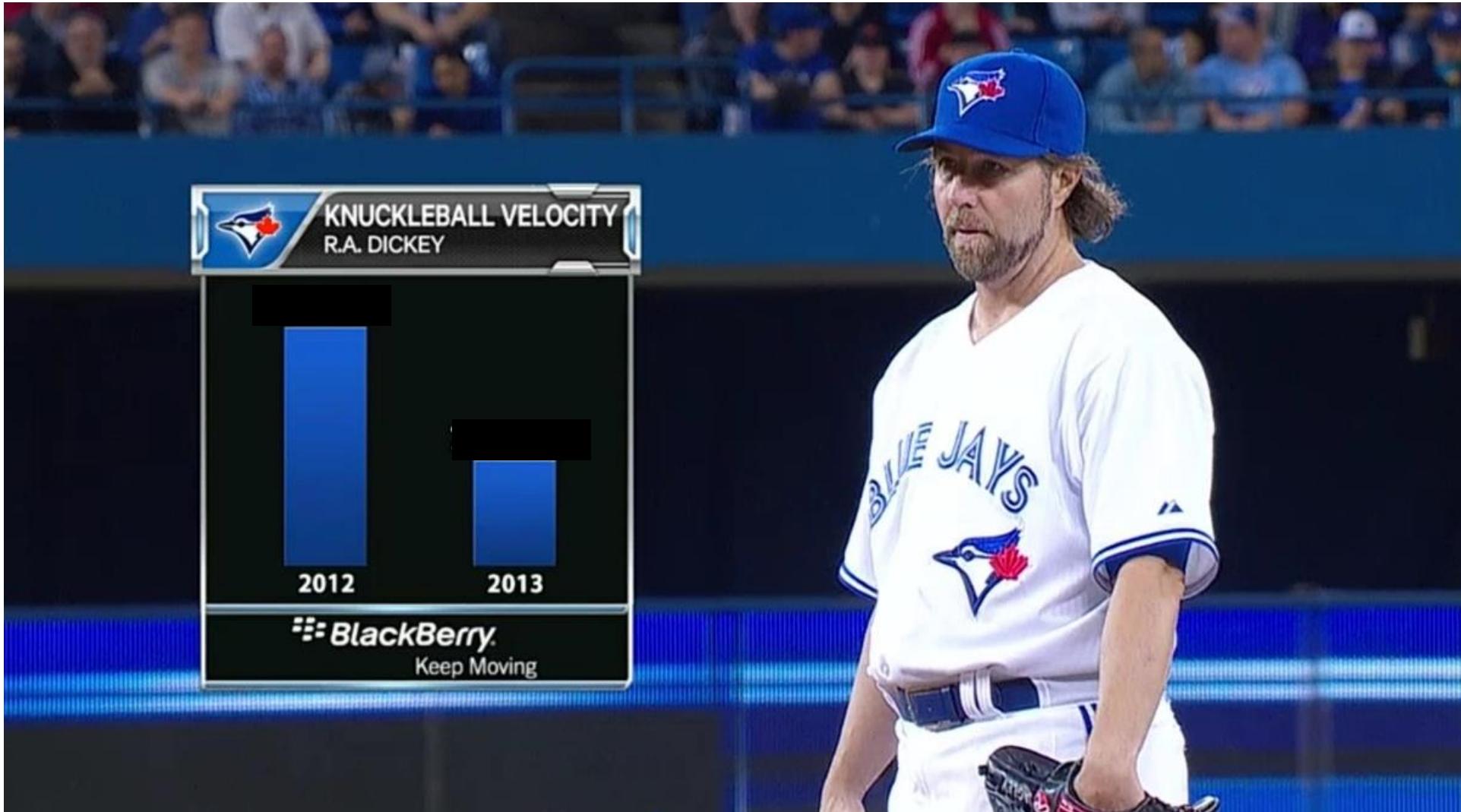
Effect in data: 1.14  
Effect in graphic: 5  
Ratio:  $5/1.14 = 4.38$



# 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](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

**7,100,000**

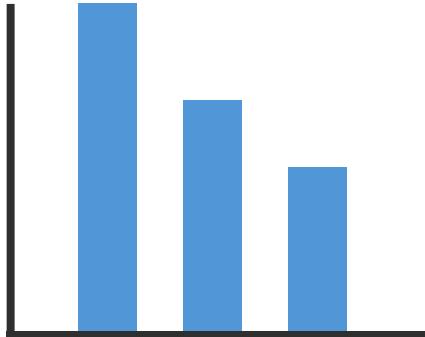
**7,000,000**

ACTUAL  
ENROLLMENT

GOAL



# Where should the y-axis start?

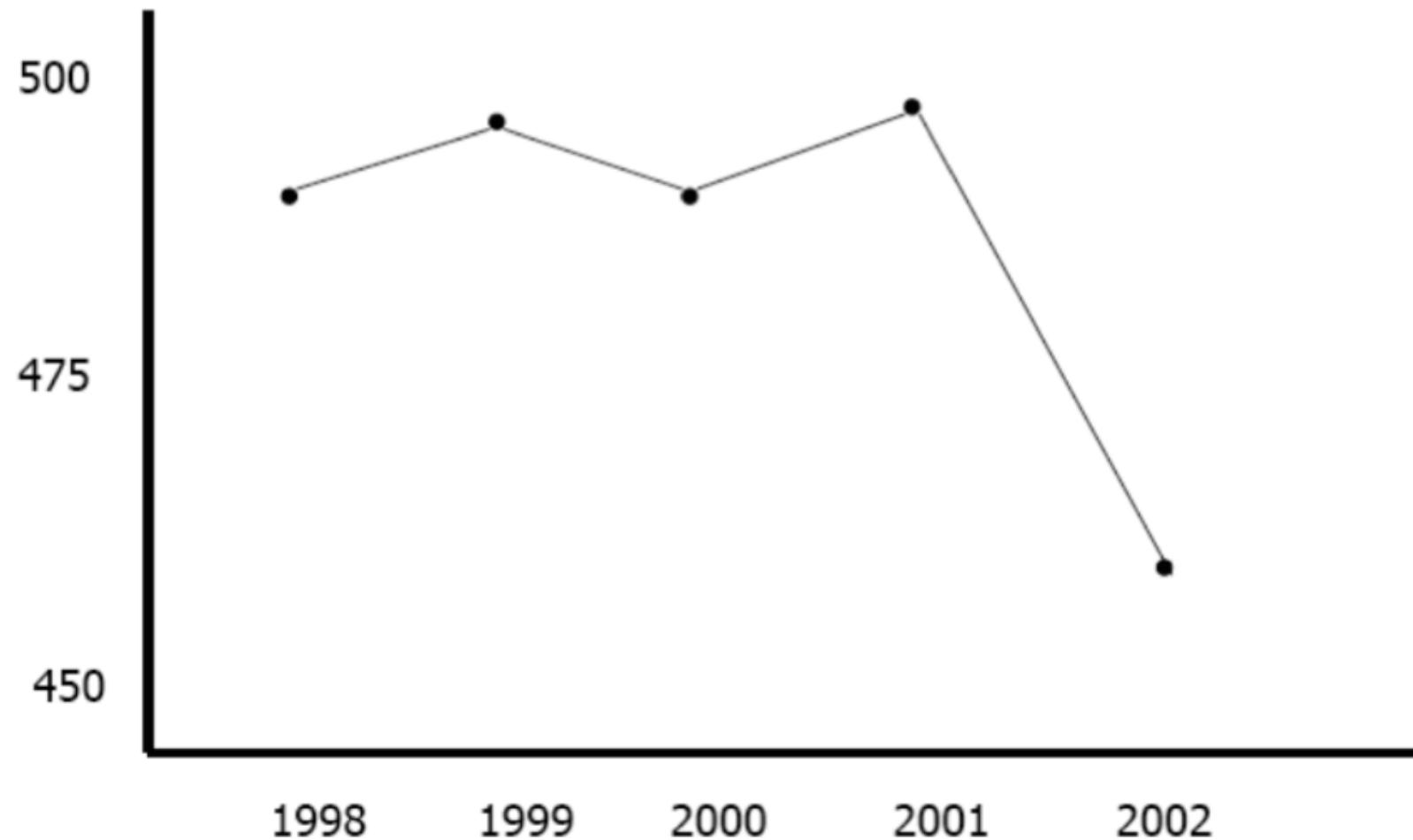


Length  
(aligned)

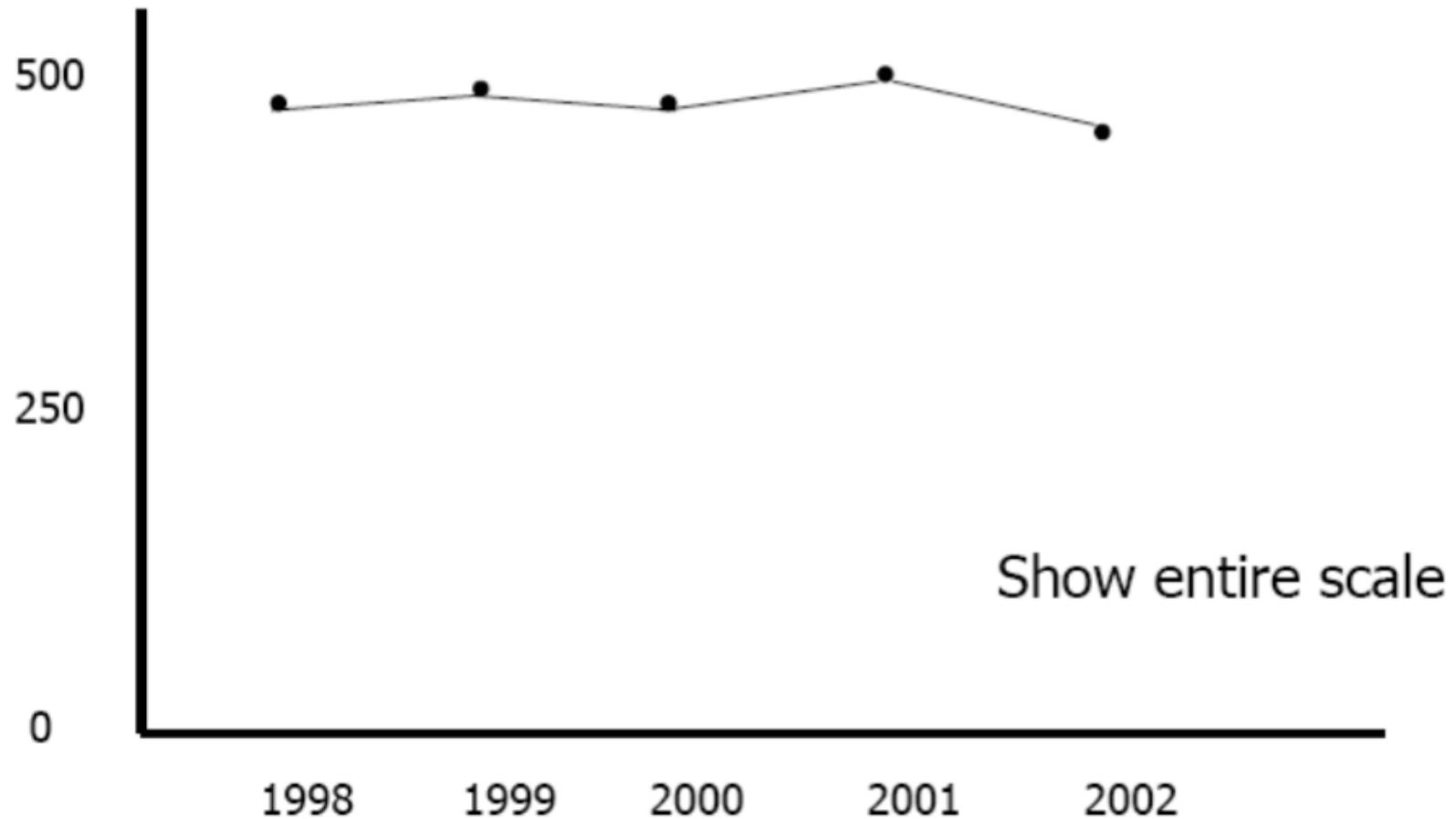
In bar charts, bar length is being compared. Therefore, starting the y-axis at an arbitrary position will work against the visualization task... often tricking the viewer.

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

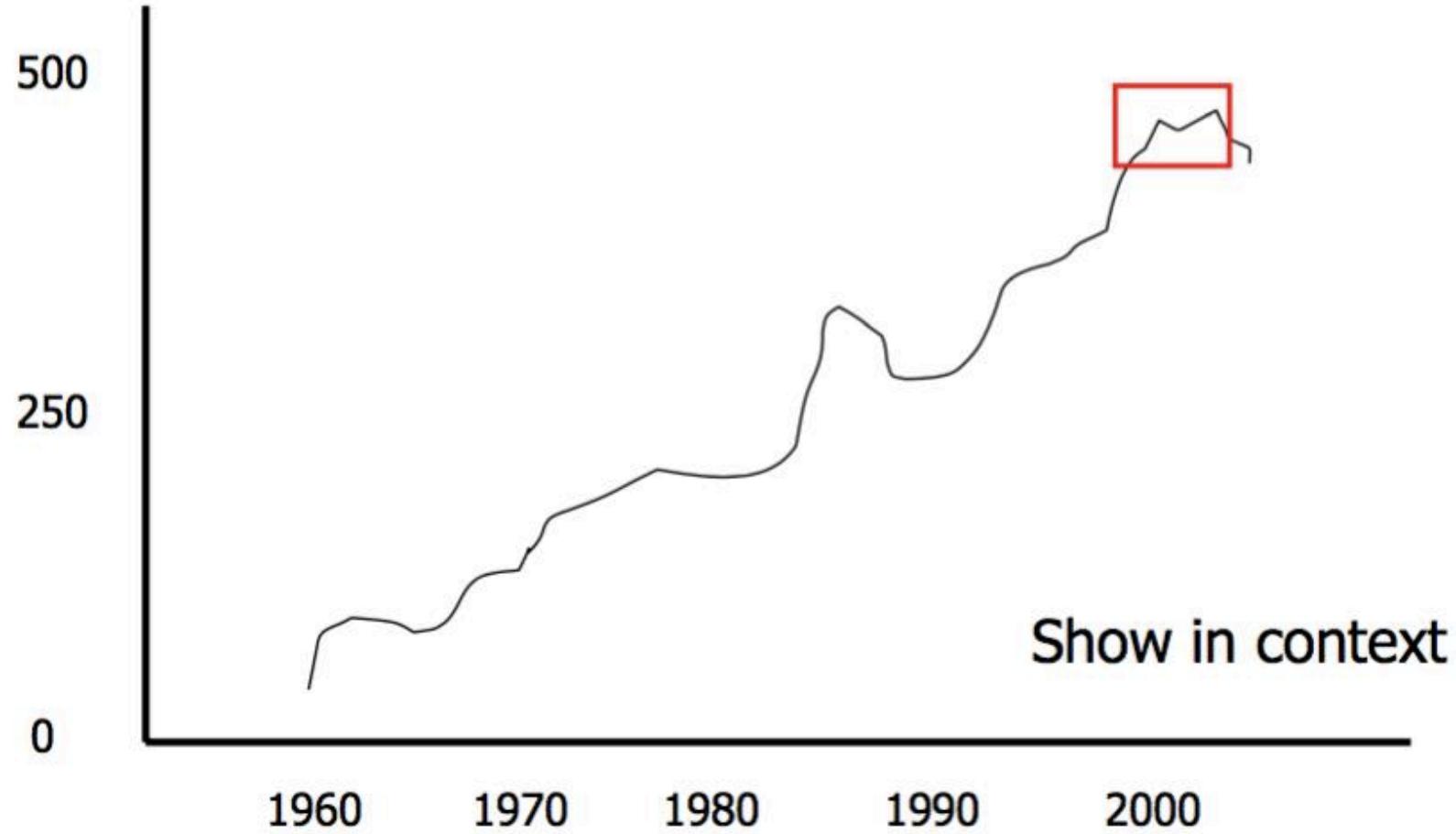
# What's happening here?



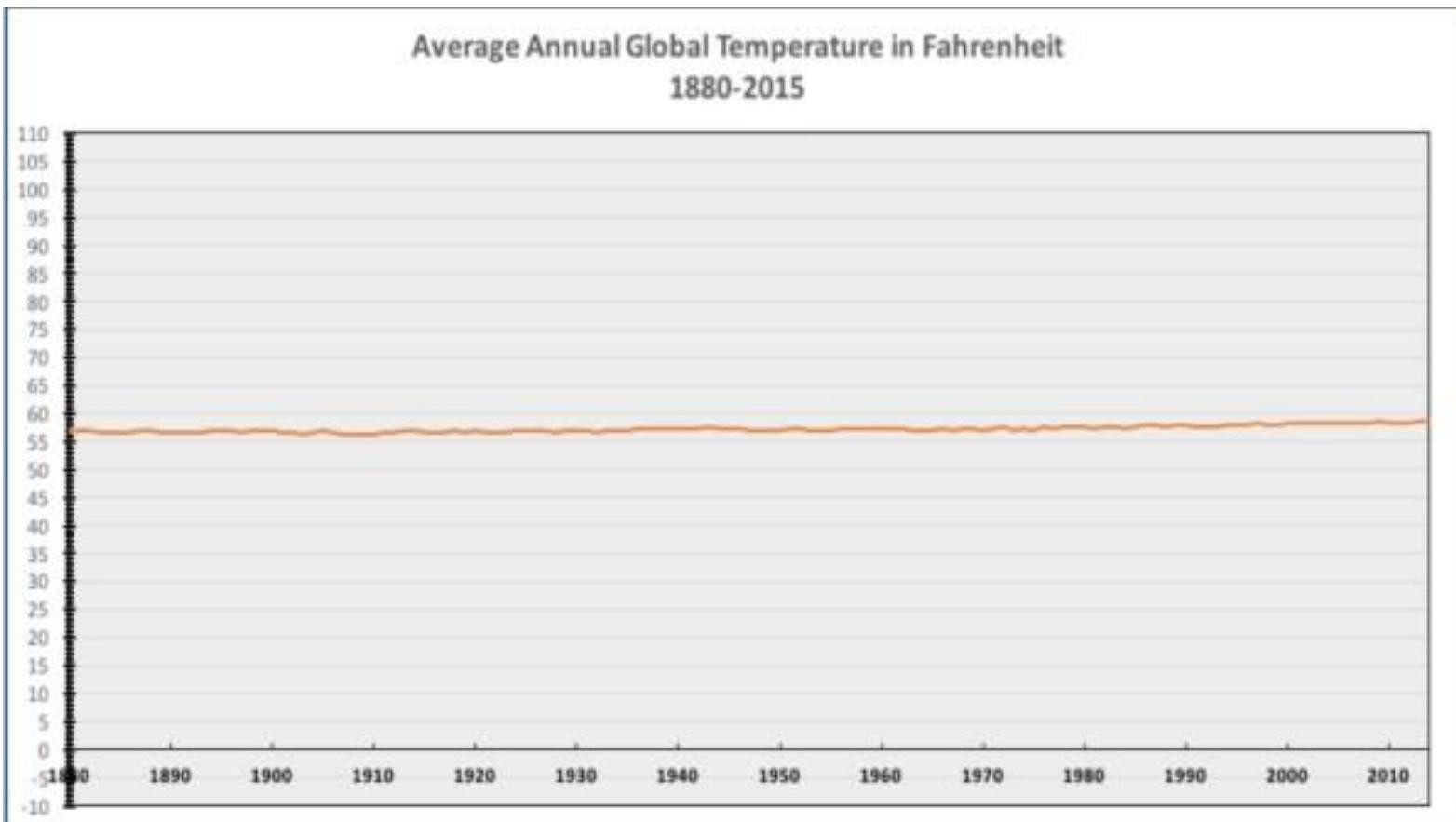
# What's happening here?



# 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).

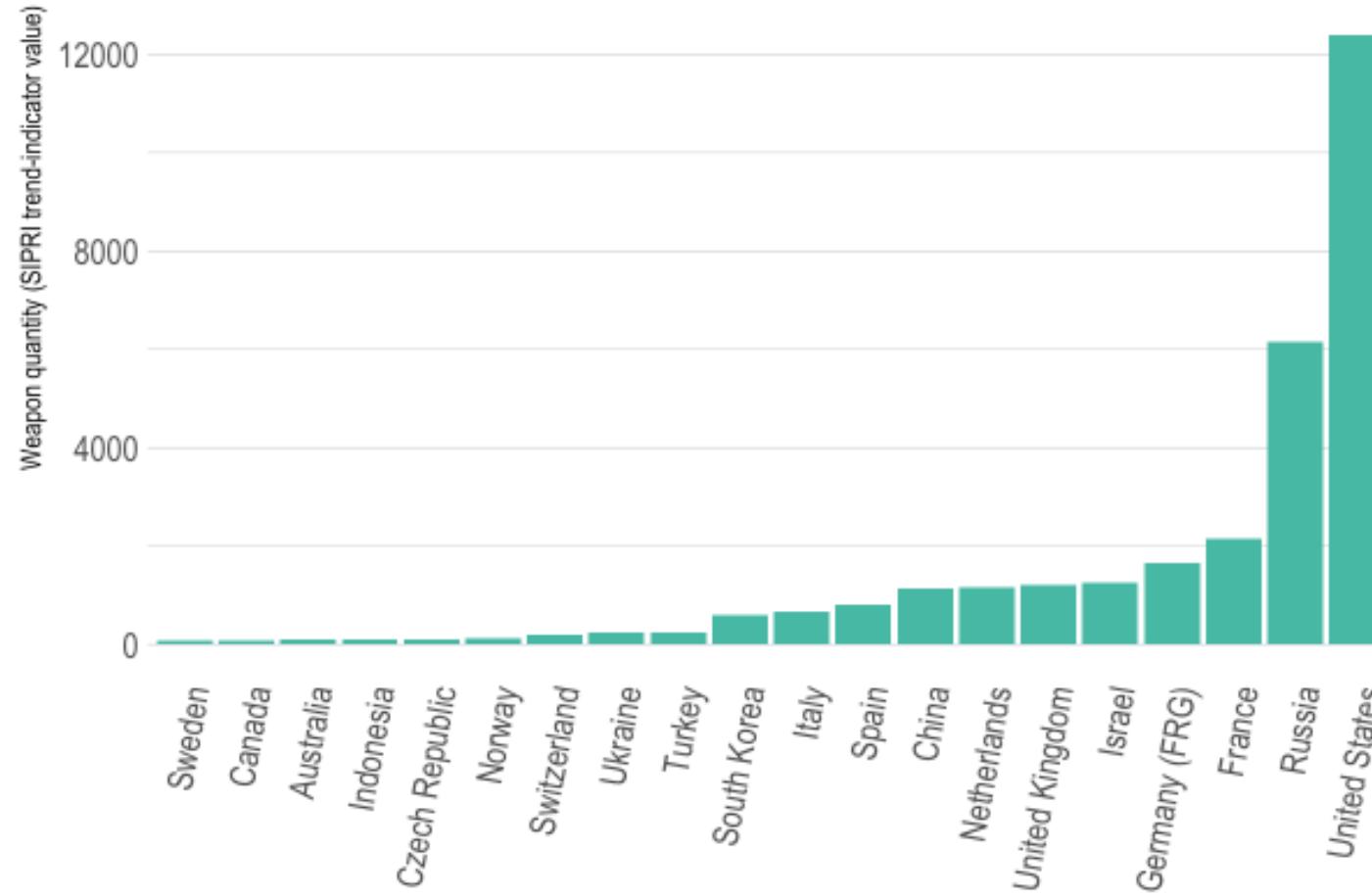
# What range should I use?

Data range?  
Start from zero?

Like all things, depends on the task.

# **Guideline: Rotate for Readability**

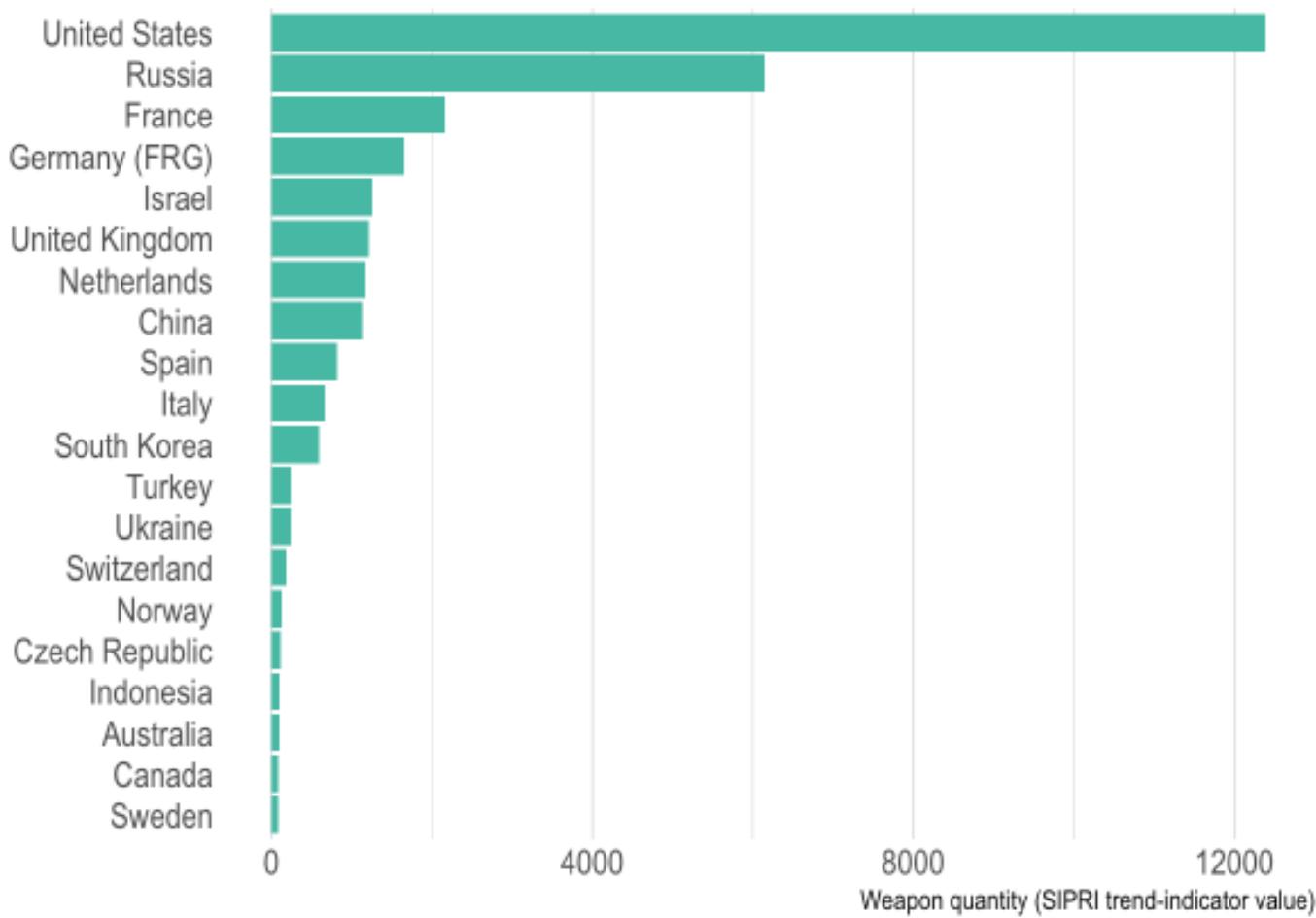
# Be Aware of Text Angle



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

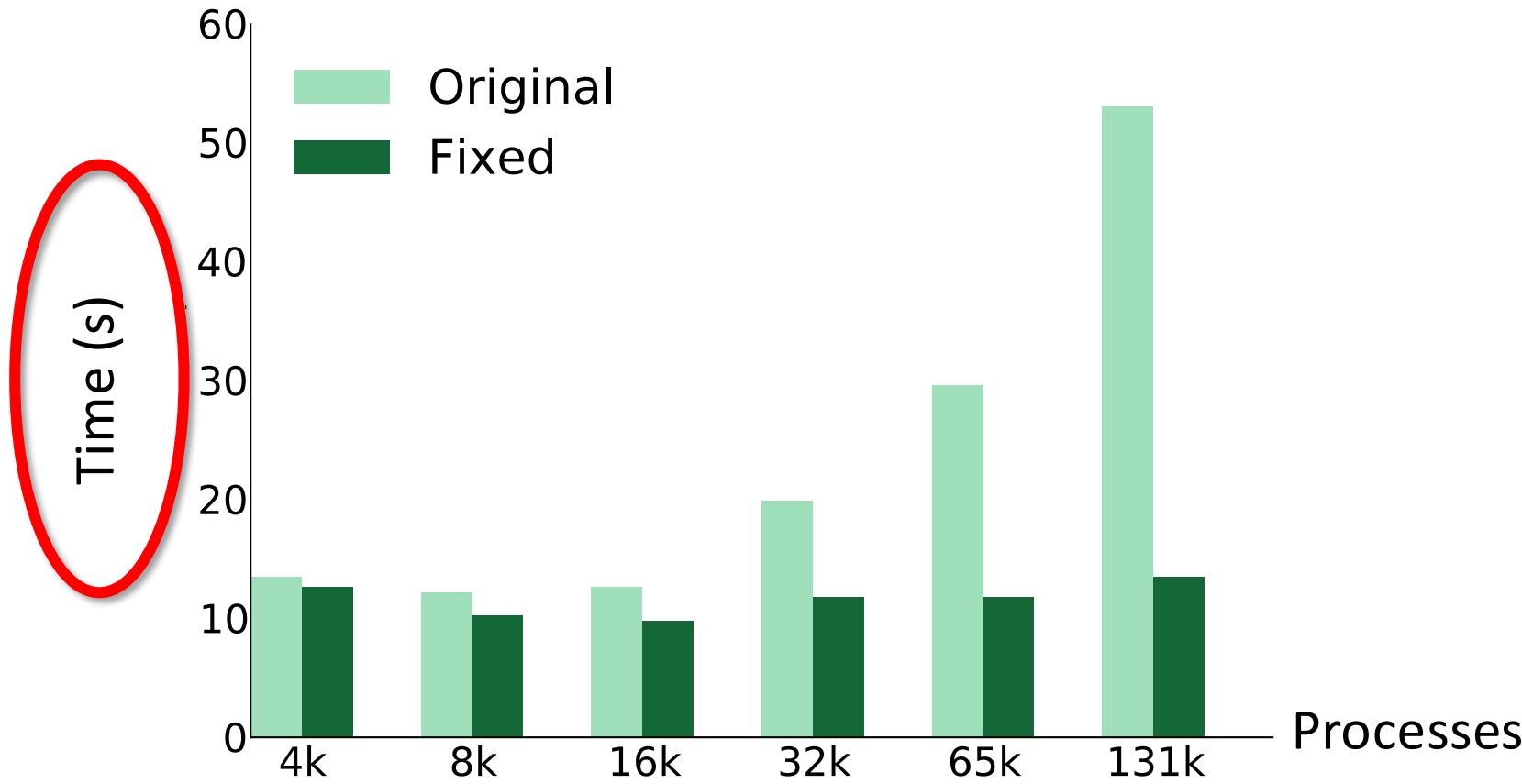
Furthermore, while we can rotate fonts at any angle, they distort and become jagged.

# Consider rotating a bar chart

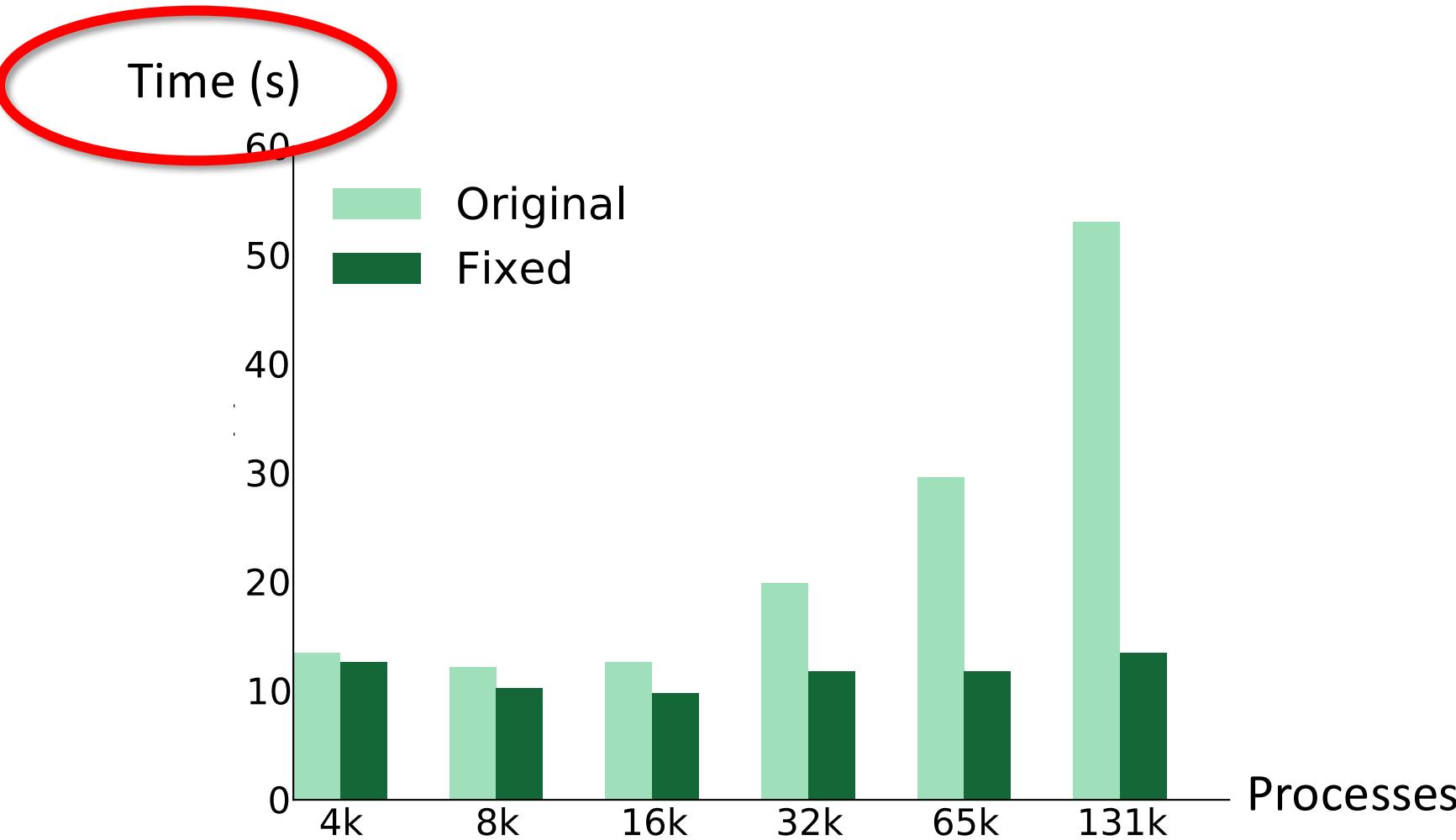


\*Note however that some axes have a strong association with the x-axis (e.g., time). In that case, the design trade off may leave tilted text.

# 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

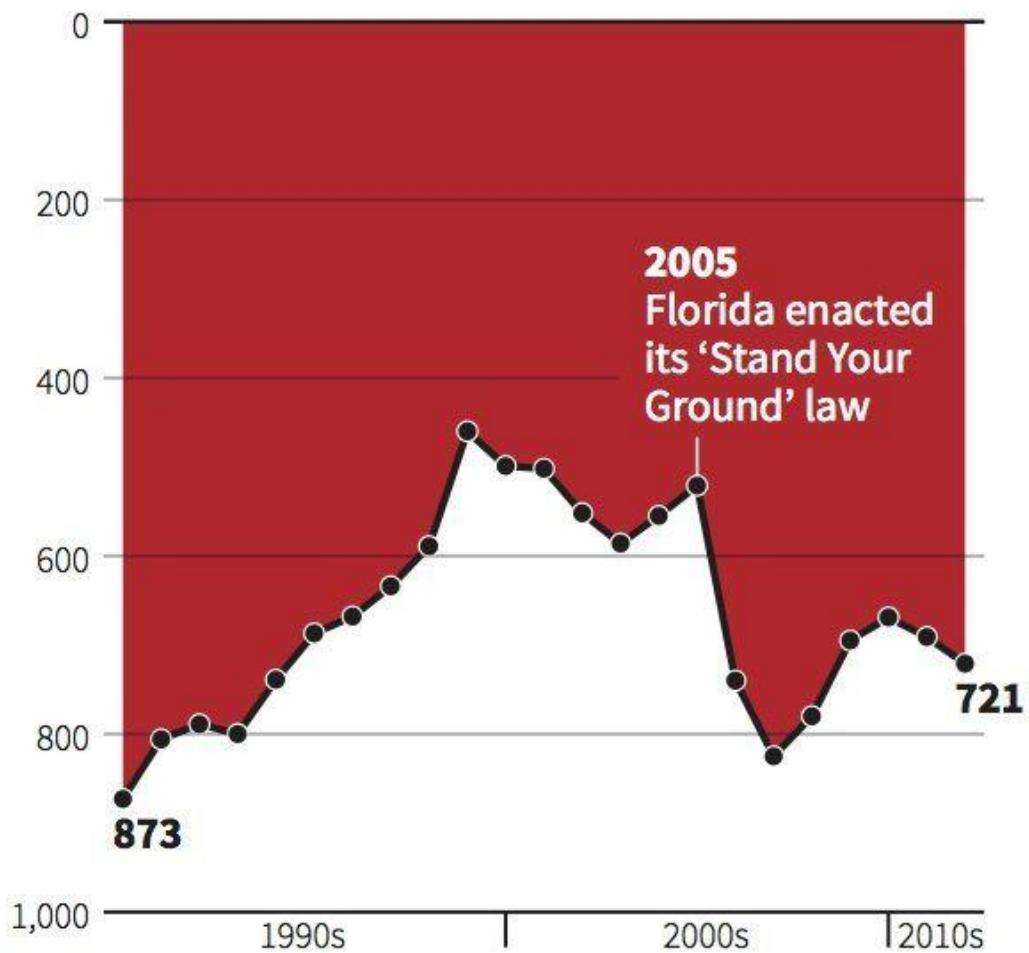
What is your initial reaction?

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

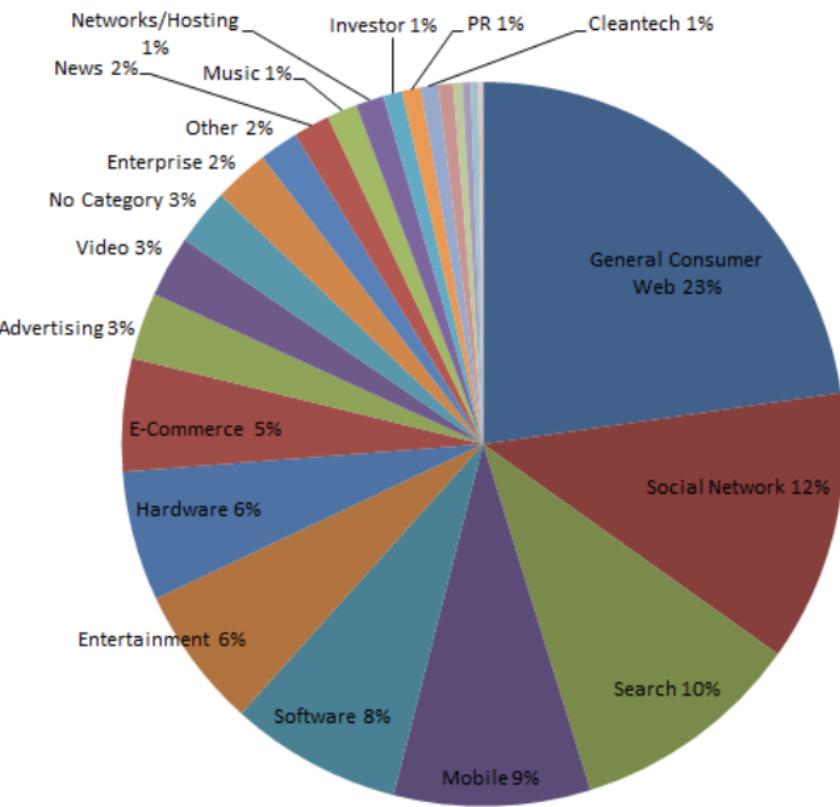


Source: Florida Department of Law Enforcement

C. Chan 16/02/2014

# **Guideline: Pie with Care**

# Pie Charts... easy to get wrong

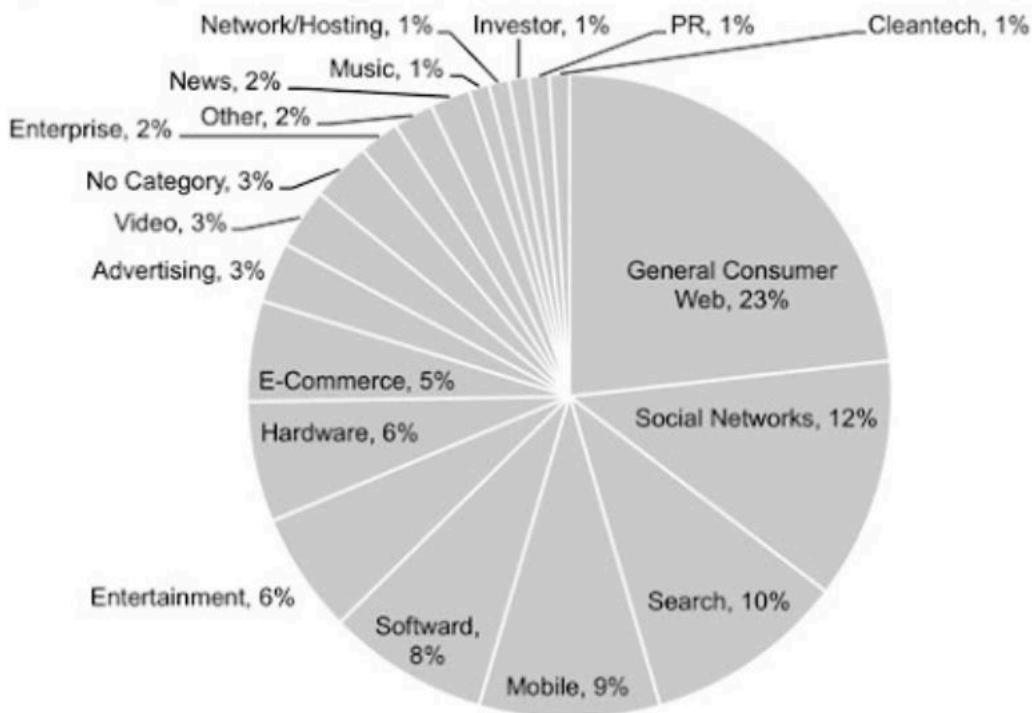


Share of coverage  
on TechCrunch

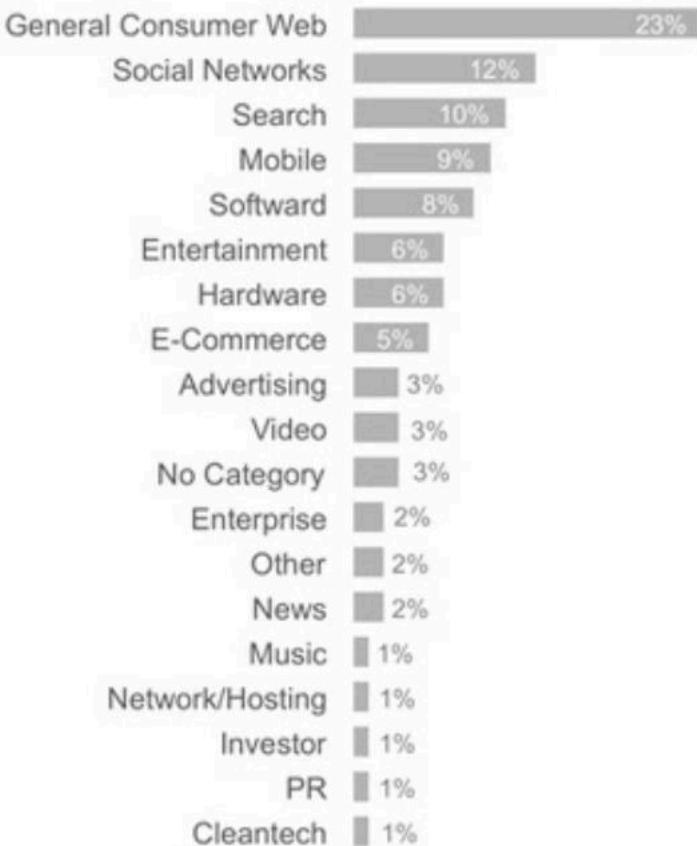
“I hate pie charts.  
I mean, really hate them.”

# Redesign

**TechCrunch Coverage: 2005 - 2011**  
*A slightly better pie?*

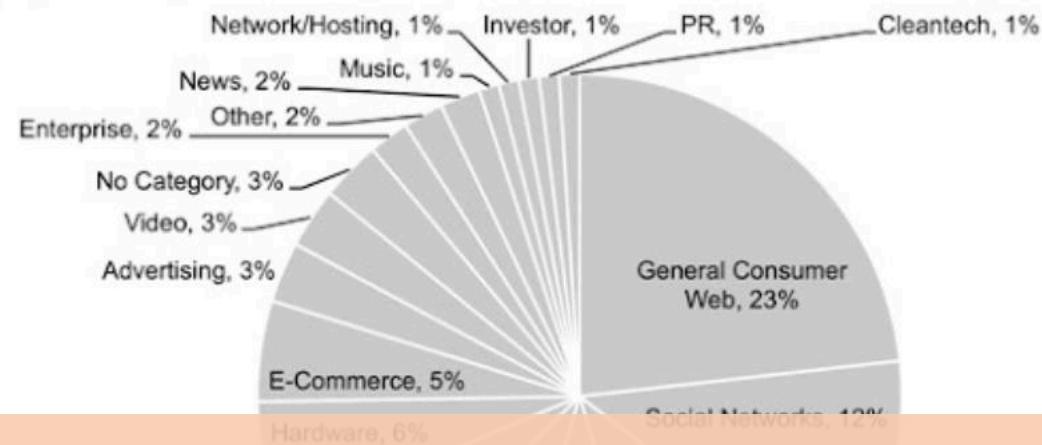


**TechCrunch Coverage: 2005 - 2011**  
*Bars are best!*

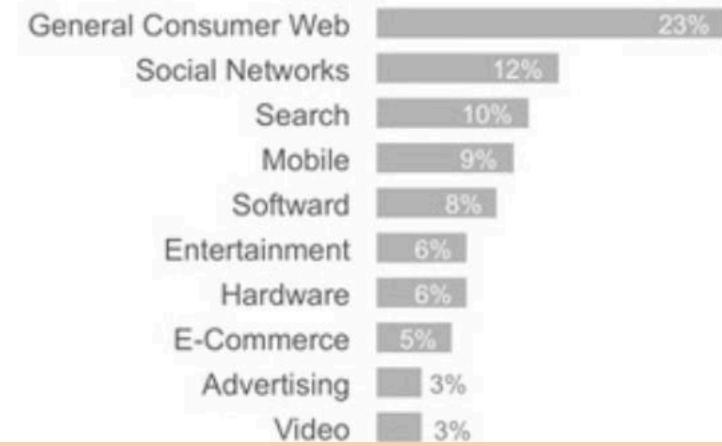


# Redesign

TechCrunch Coverage: 2005 - 2011  
A slightly better pie?



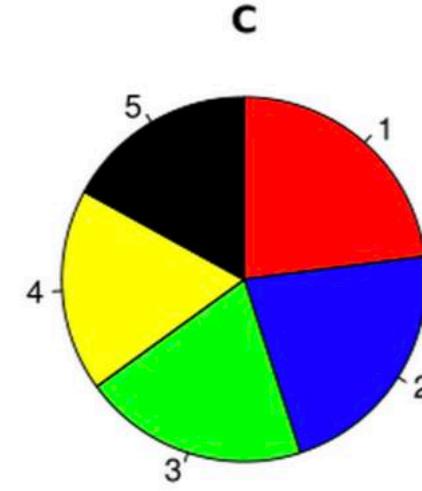
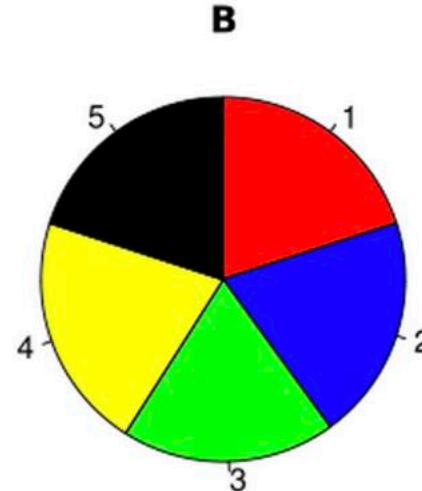
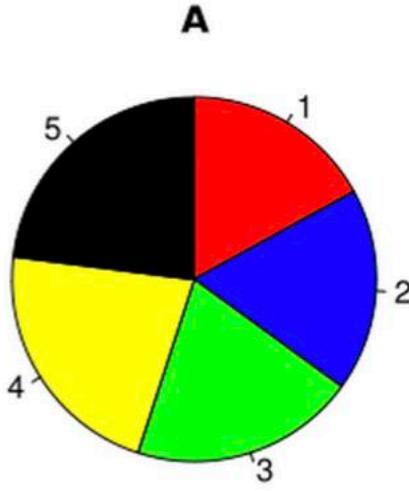
TechCrunch Coverage: 2005 - 2011  
Bars are best!



What were they trying to show?

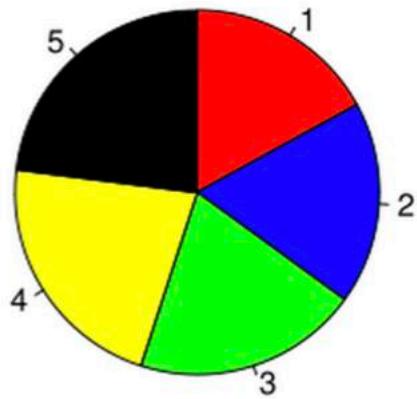
Is proportion the most important data feature?

# Can you spot the differences?

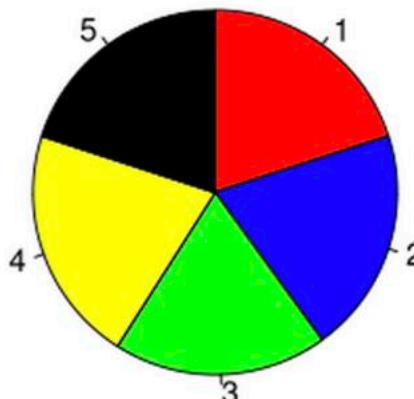


# Can you spot the differences?

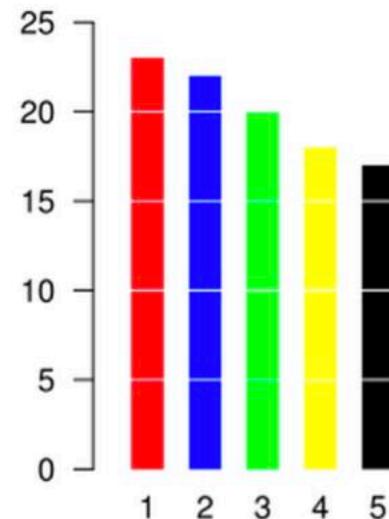
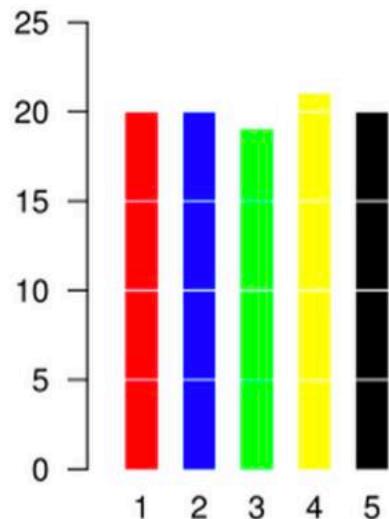
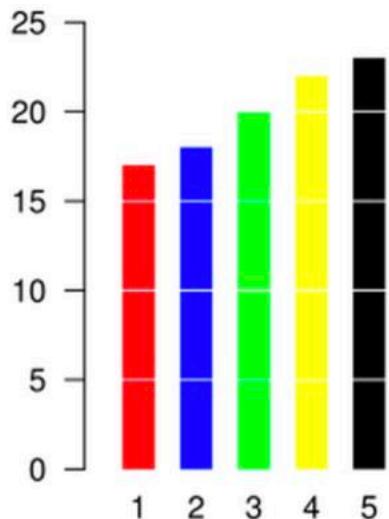
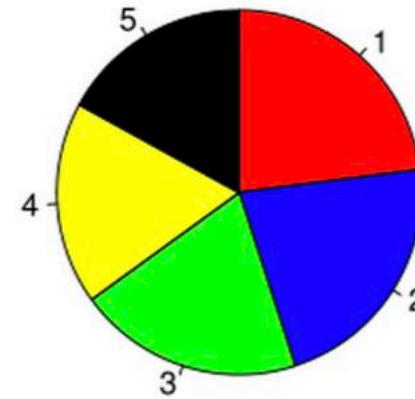
A



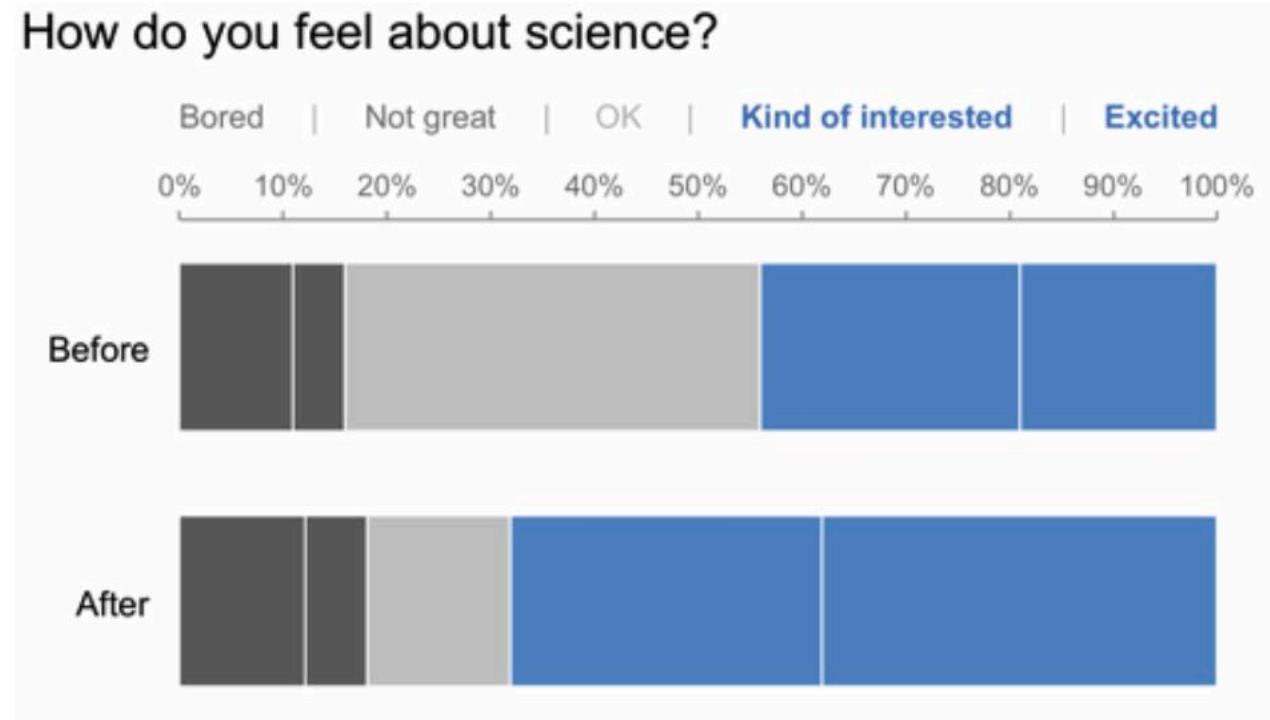
B



C



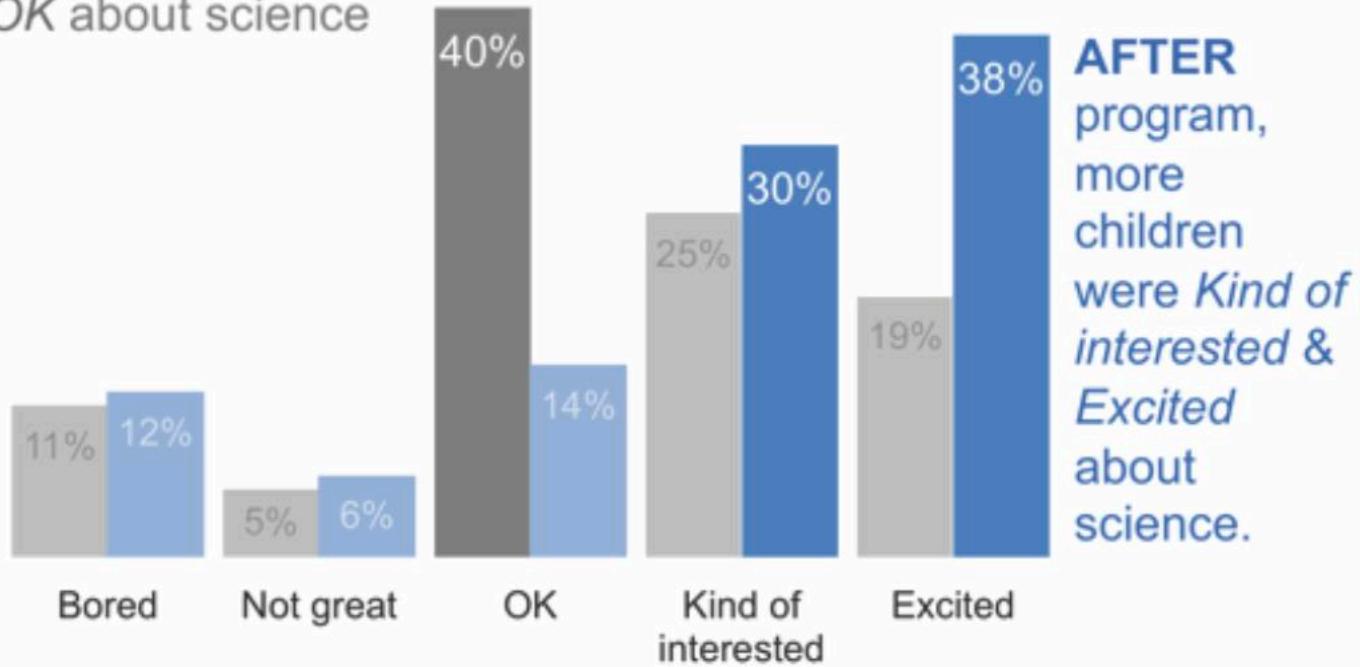
# Pie Alternatives: Stacked Bar Charts



# Pie Alternatives: Bar charts

## How do you feel about science?

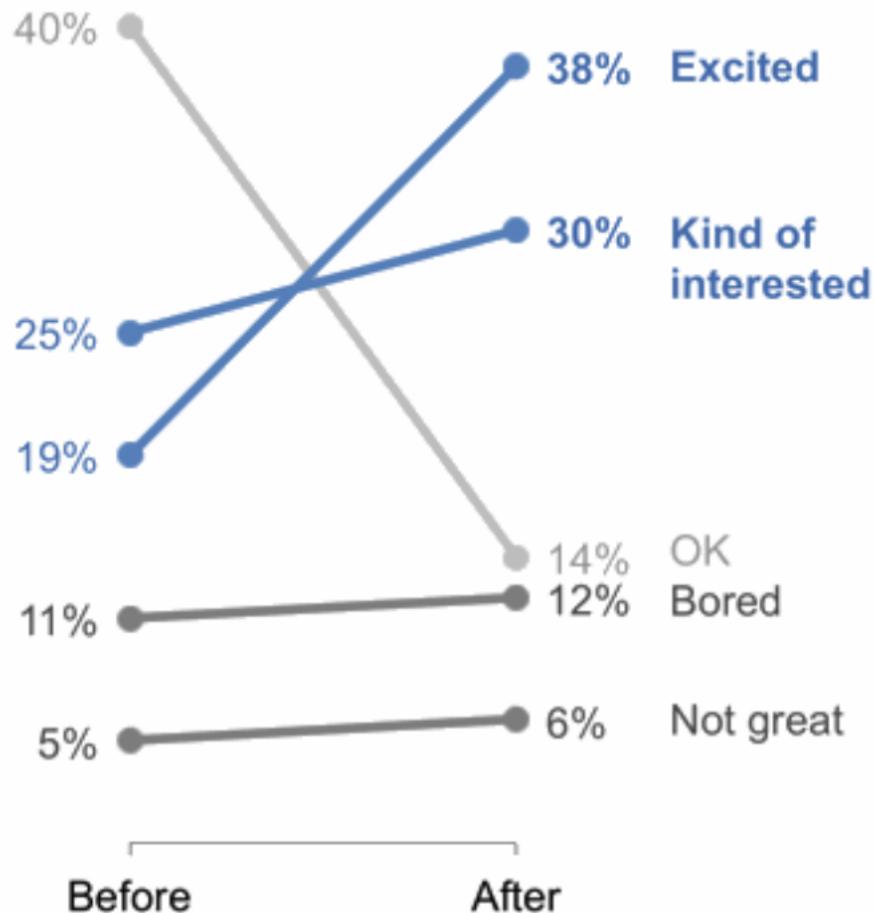
**BEFORE** program, the majority of children felt just OK about science



**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

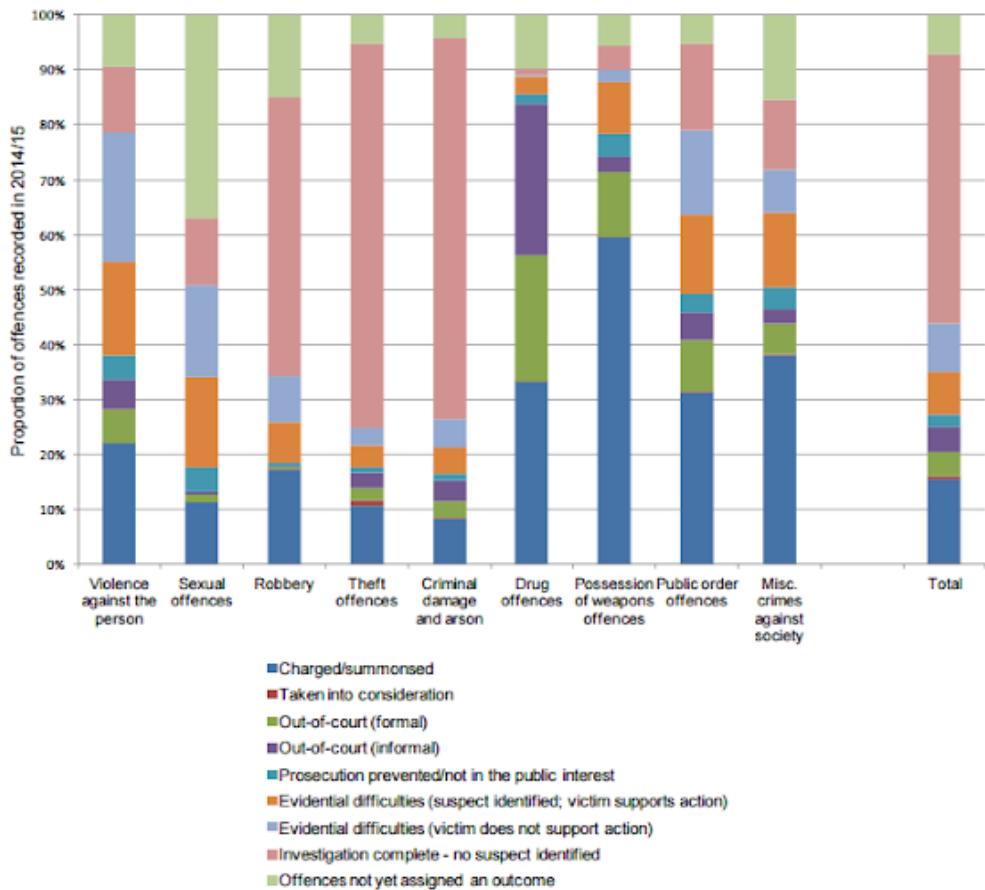
After the pilot program,

**68%**

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

# Stacked Bar Charts vs. Small Multiples

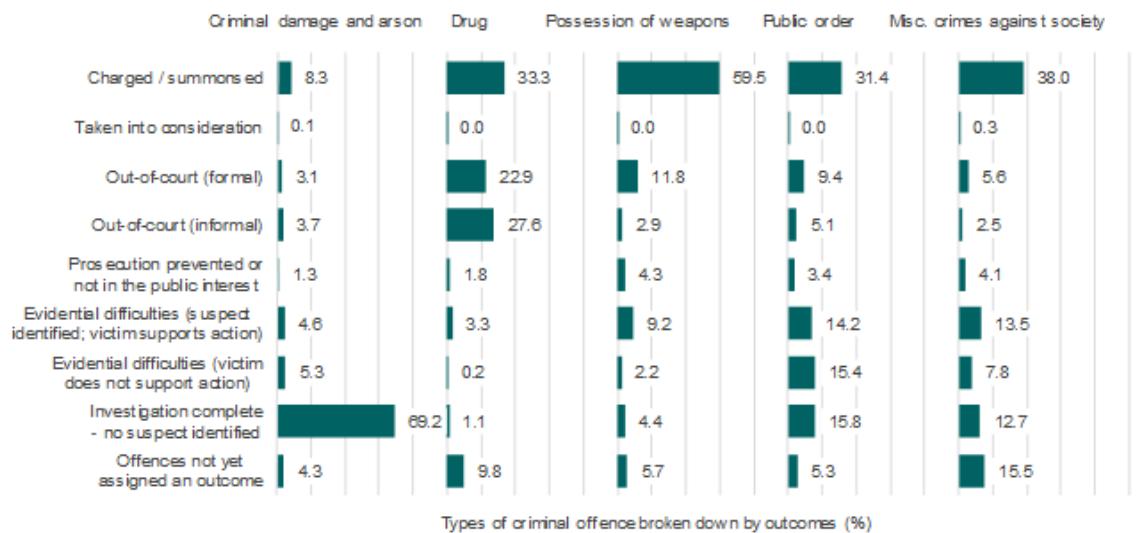
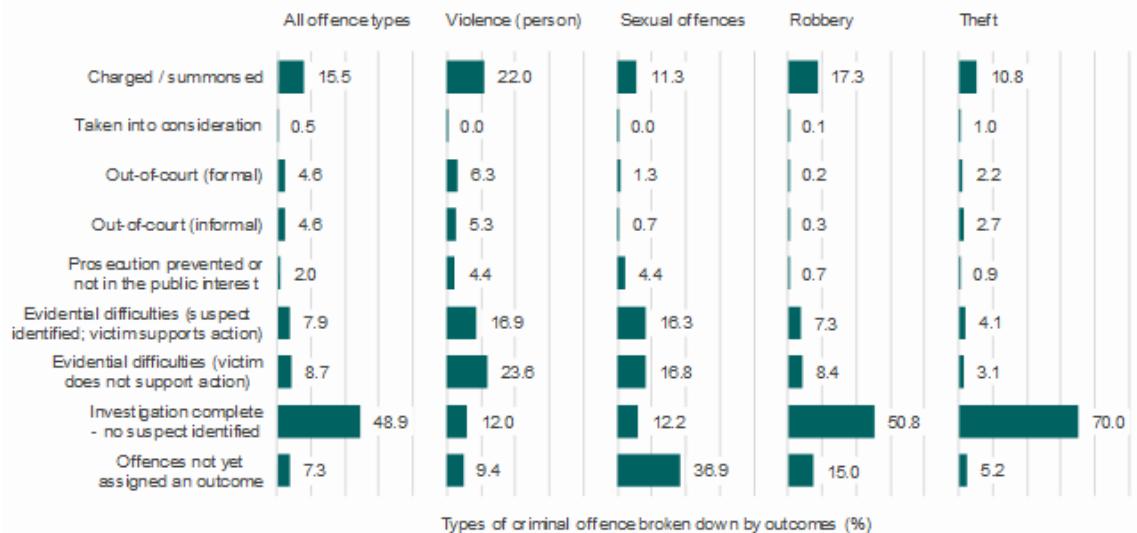
Figure 2.1: Outcomes assigned to offences recorded in 2014/15, by outcome group and offence group



Source: Home Office Data Hub and voluntary spreadsheet return

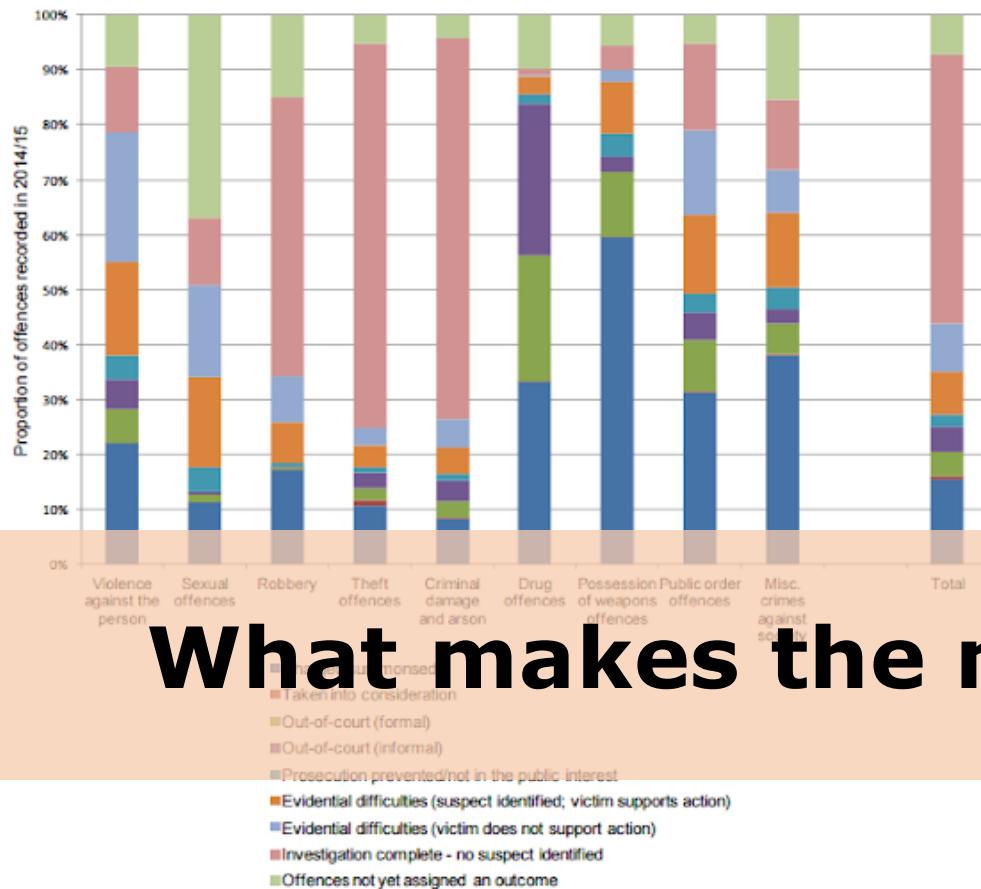
1. Based on 38 forces that supplied data as referenced in Table 2.1.

2. The numbers behind this chart are in Table 2.3



# Stacked Bar Charts vs. Small Multiples

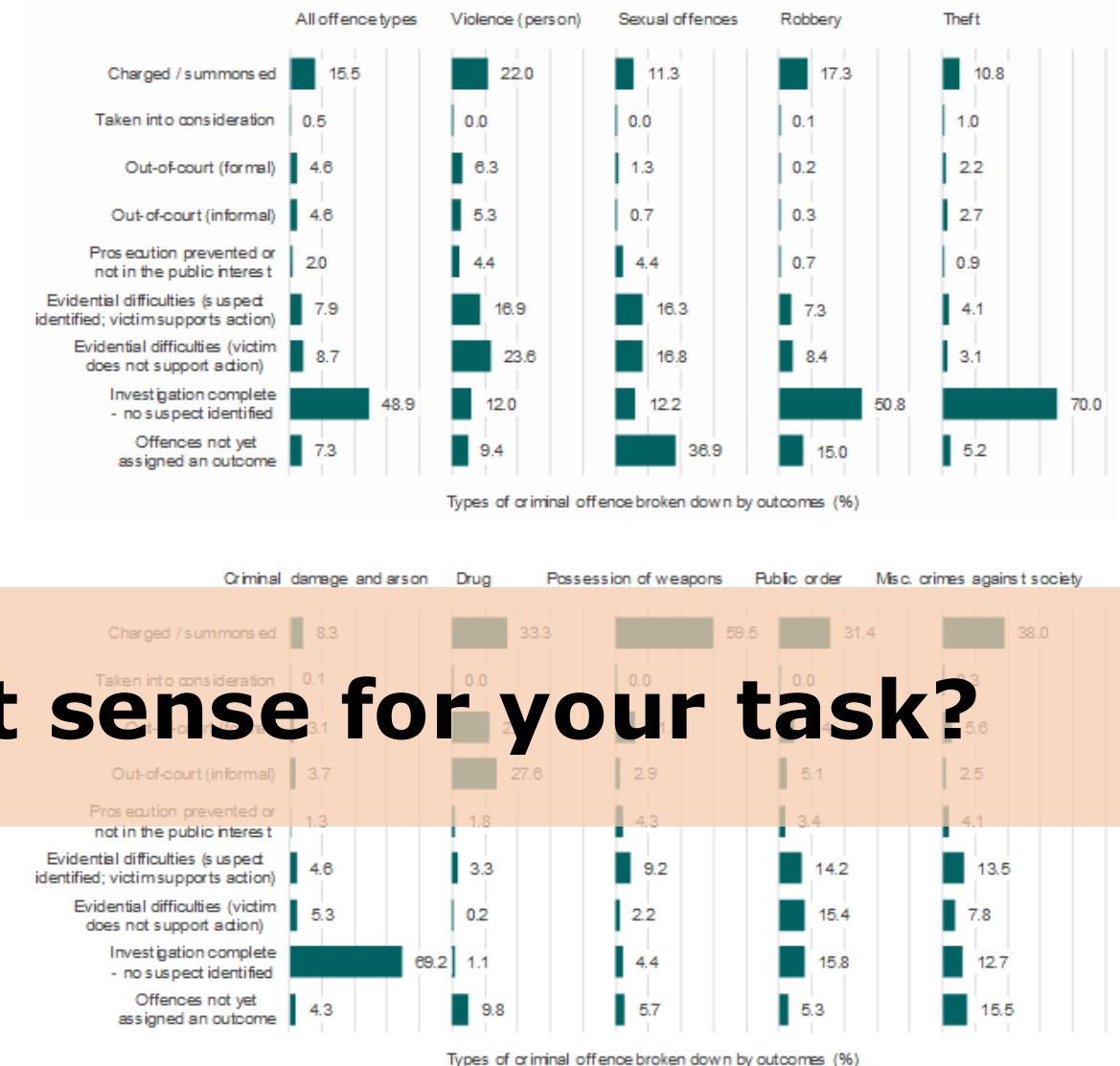
Figure 2.1: Outcomes assigned to offences recorded in 2014/15, by outcome group and offence group



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?

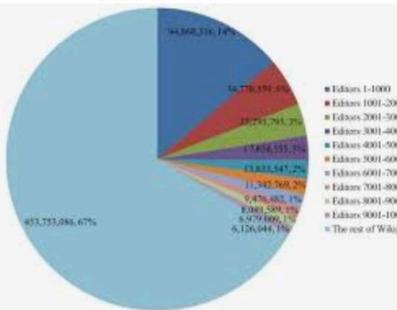
# Pie charts **not** inherently bad. Maybe the biggest problem with pie charts is that they have been so often done poorly...

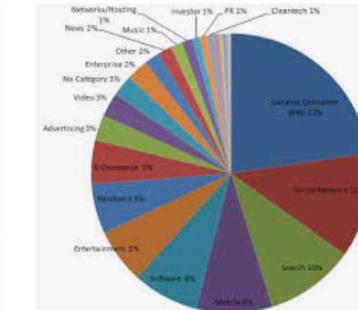
Google search results for "bad pie charts":

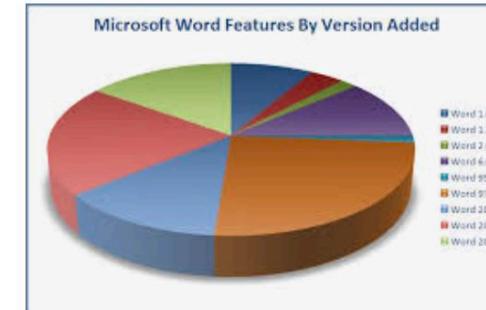
- wrong
- media
- example
- data visualization
- male female
- economy florida
- 2016 presidential election
- attractive
- advanced

Search filters: All, Images (selected), Videos, News, Shopping, More, Settings, Tools, SafeSearch.

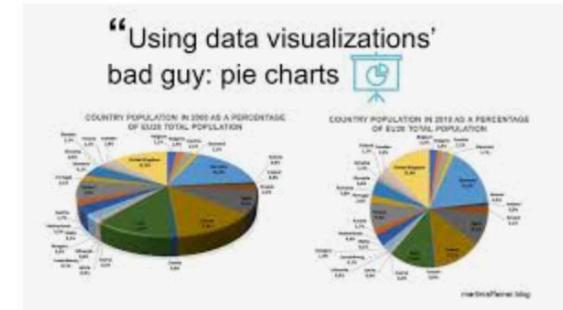
Results:

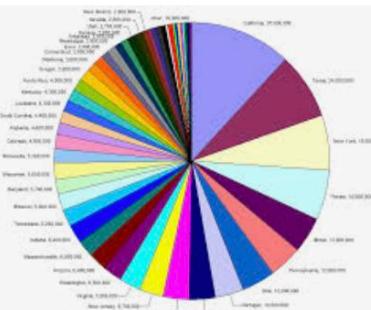
- 

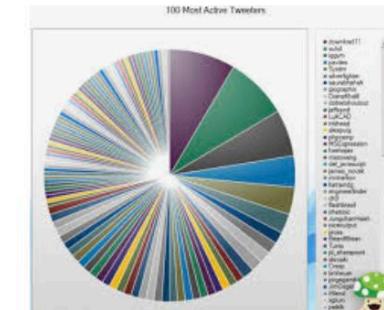
Yet another bad pie chart : dataisugly reddit.com
- 

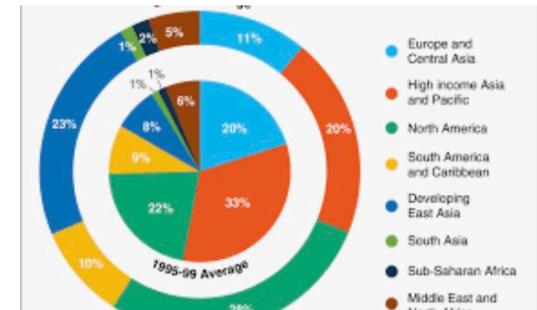
death to pie charts – storytellingwithdata.com
- 

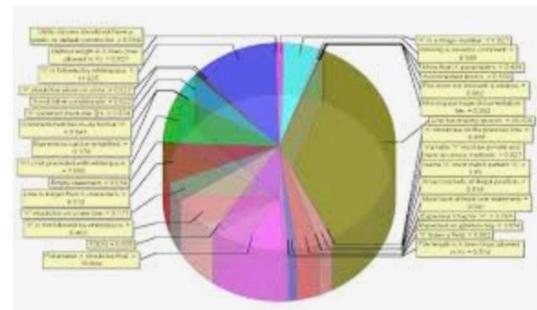
Pie charts: the bad, the worst and the ... visuanalyze.wordpress.com
- 

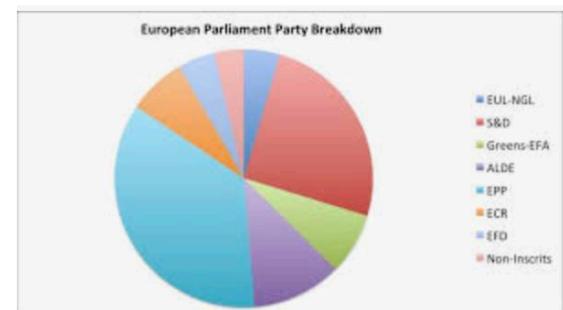
When to use Pie Charts in Dashboards ... excelcampus.com
- 

Using data visualizations' bad guy: pie ... martinraffineer.blog
- 

Understanding Pie Charts eagereyes.org
- 

Pie charts: the bad, the worst an... visuanalyze.wordpress.com
- 

Remake: Pie-in-a-Donut Chart - Policy Viz policyviz.com
- 

Pin on Chartjunk Data Visualization pinterest.com
- 

Pie Charts Are The Worst - Business Insider businessinsider.com