

# Data Visualization

## Lecture 9

# Data visualization

---

Data visualization is the process of creating graphical representations of information.

- **Represent** data and information **graphically**.
  - Translate data into a visual context using charts, plots, animations, infographics, etc.
- **Discover** the trends and patterns of data.
- **Summarize** the information



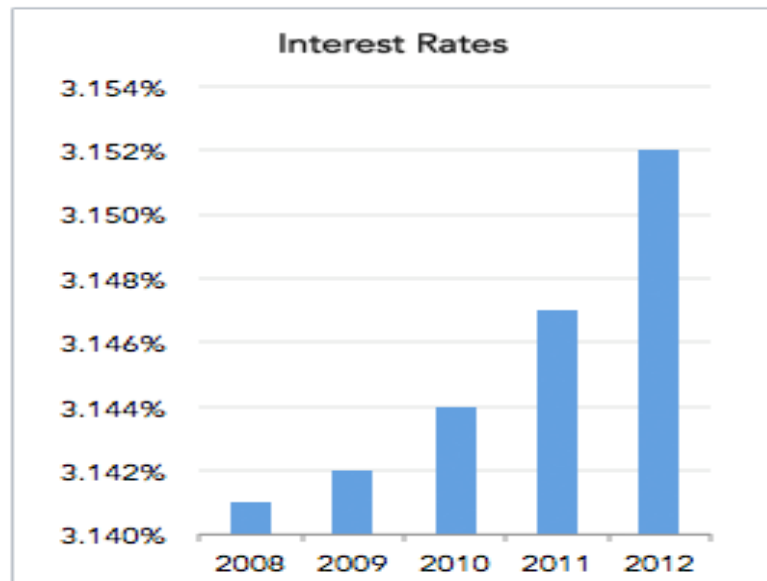
# The Bad:



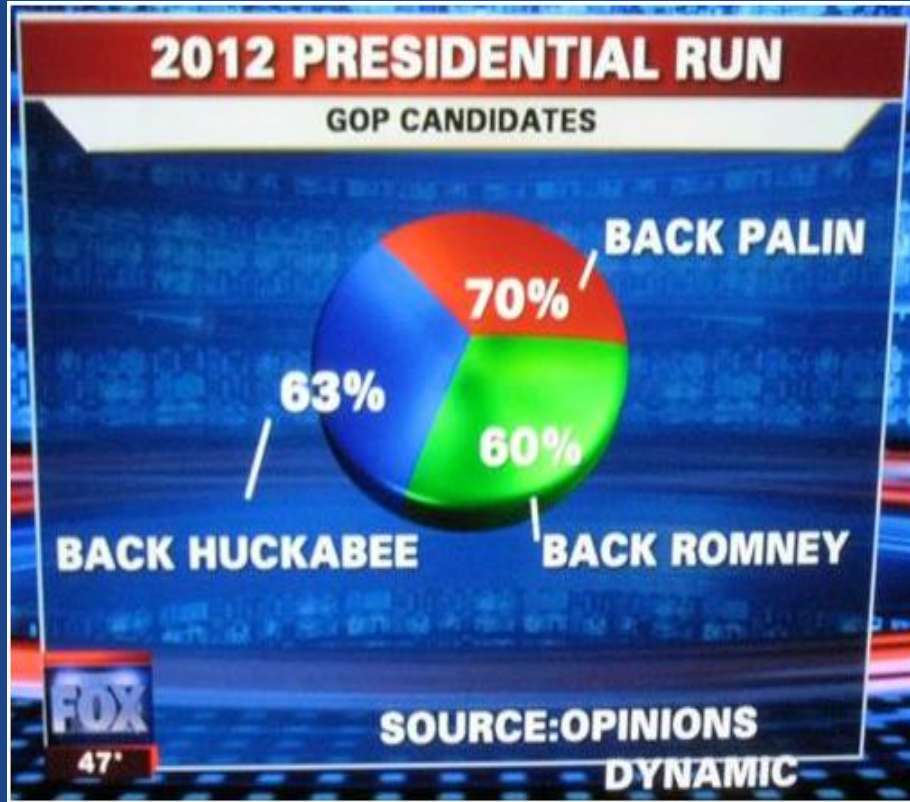
*Sometimes, visualizations can be misleading*



## Same Data, Different Y-Axis



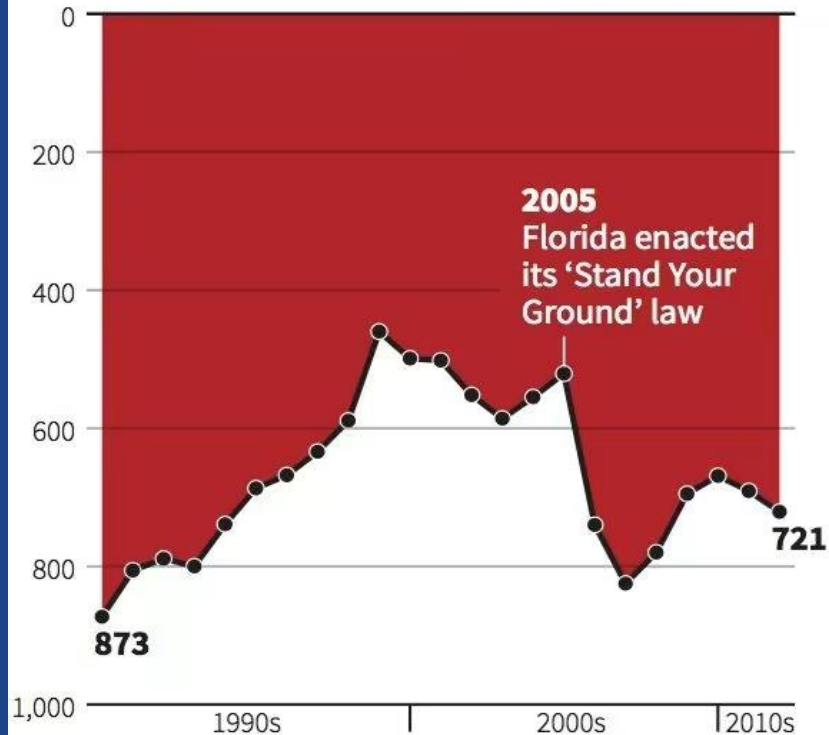
Think about, is it same data or different data?



Think about, what's wrong with this graph?

# Gun deaths in Florida

Number of murders committed using firearms

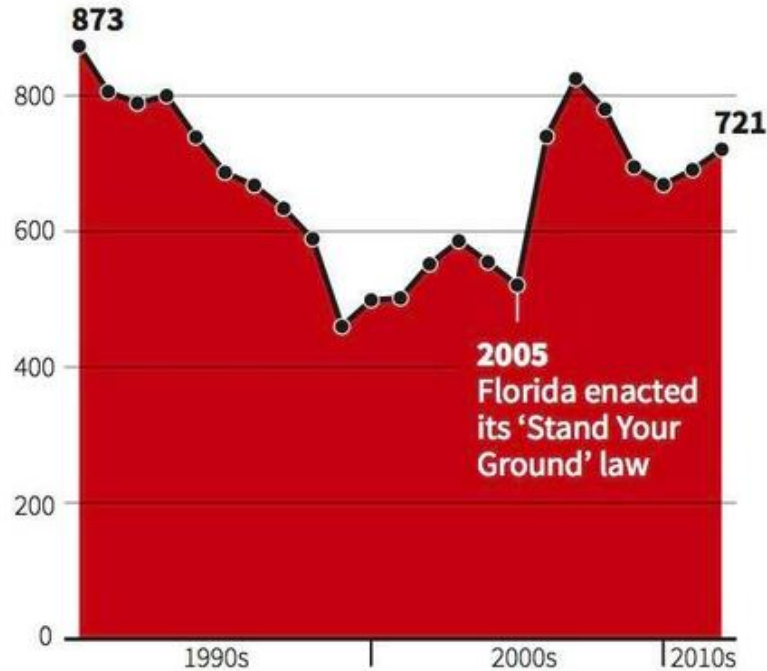


Source: Florida Department of Law Enforcement

Think about, what's wrong with this graph?

# Gun deaths in Florida

Number of murders committed using firearms



Source: Florida Department of Law Enforcement

P.A. Fedewa and Reuters

A reader of Business Insider, P.A. Fedewa, was kind enough to **revise** this graph, **using all of the same numbers, with the y-axis starting at the normalized bottom left, 0.**

## UNEMPLOYMENT RATE UNDER PRESIDENT OBAMA



2011

SOURCE: BUREAU OF LABOR STATISTICS

WAR AND A TROOP WITHDRAWAL AT THE E NAS FUT 2,292.50

Think about, what's wrong with this graph?



## UNEMPLOYMENT RATE UNDER PRESIDENT OBAMA



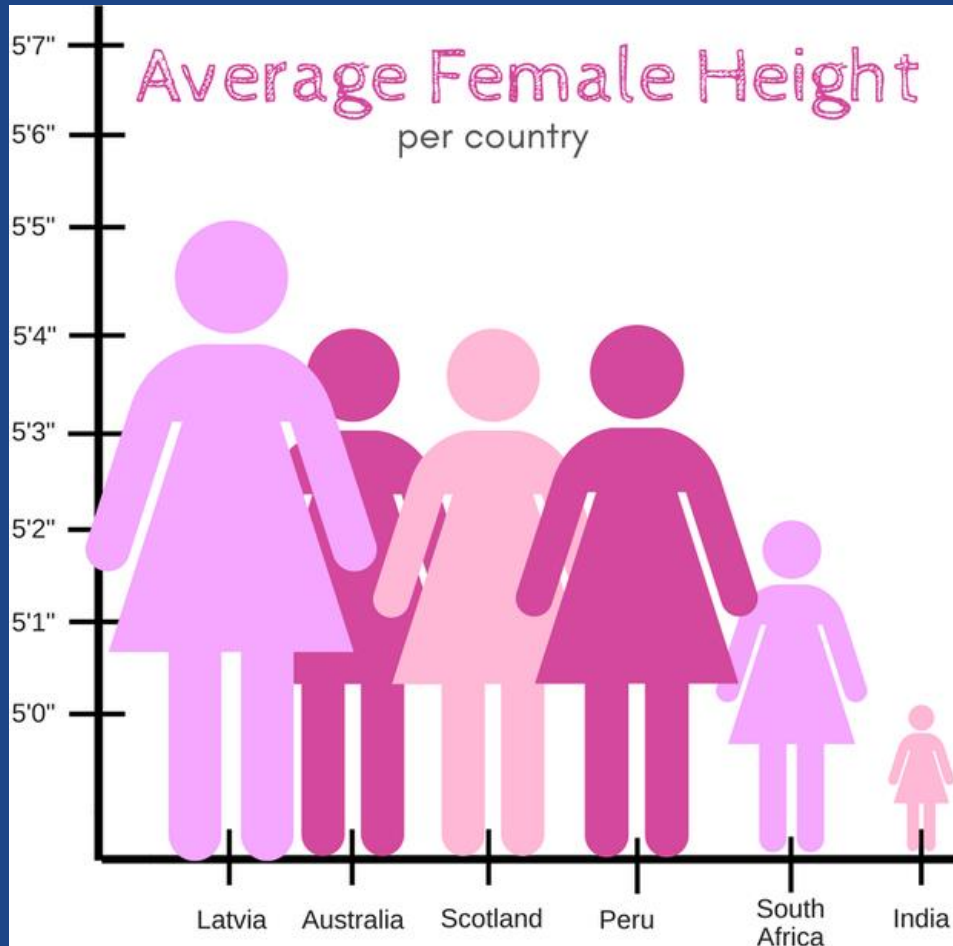
2011

SOURCE: BUREAU OF LABOR STATISTICS



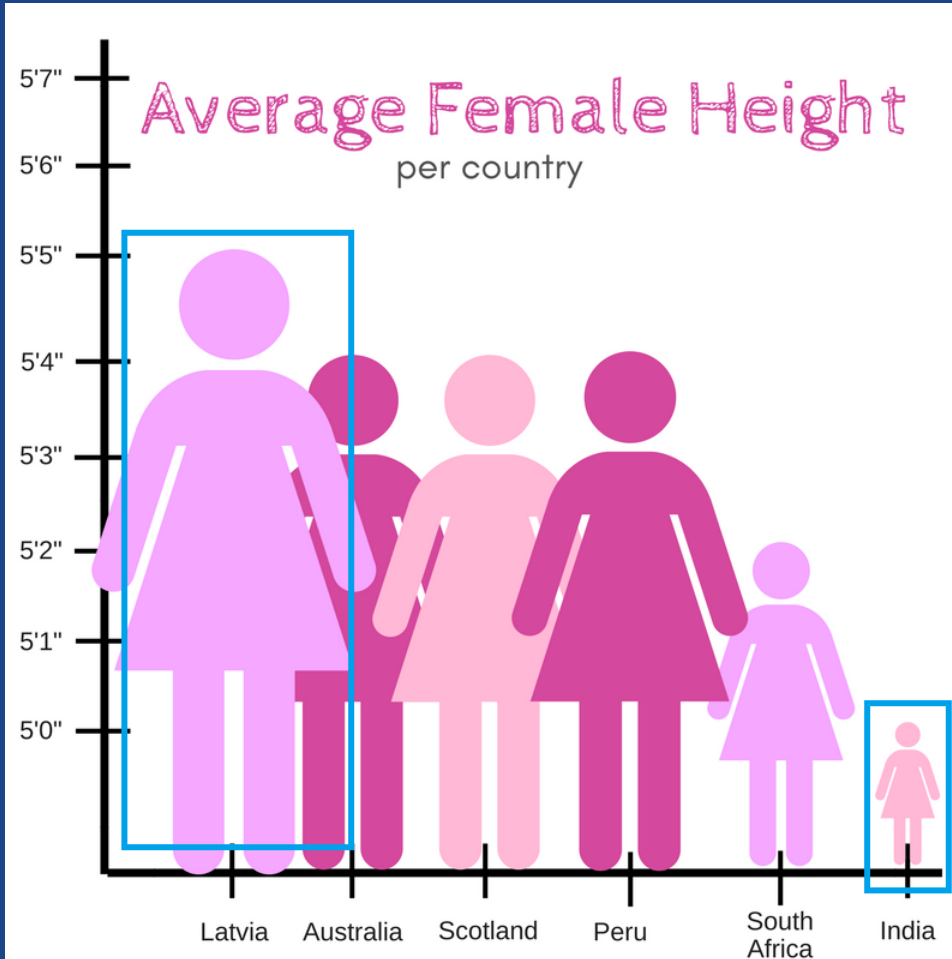
EAR WAR AND A TROOP WITHDRAWAL AT THE E NAS FUT 2,292.50

Think about, what's wrong with this graph?



What's wrong with this graph?

*Visualizations should never prioritize style over substance.*



What's wrong with this graph?

Actual difference – 0.5 feet. Visual Difference – 10 feet.

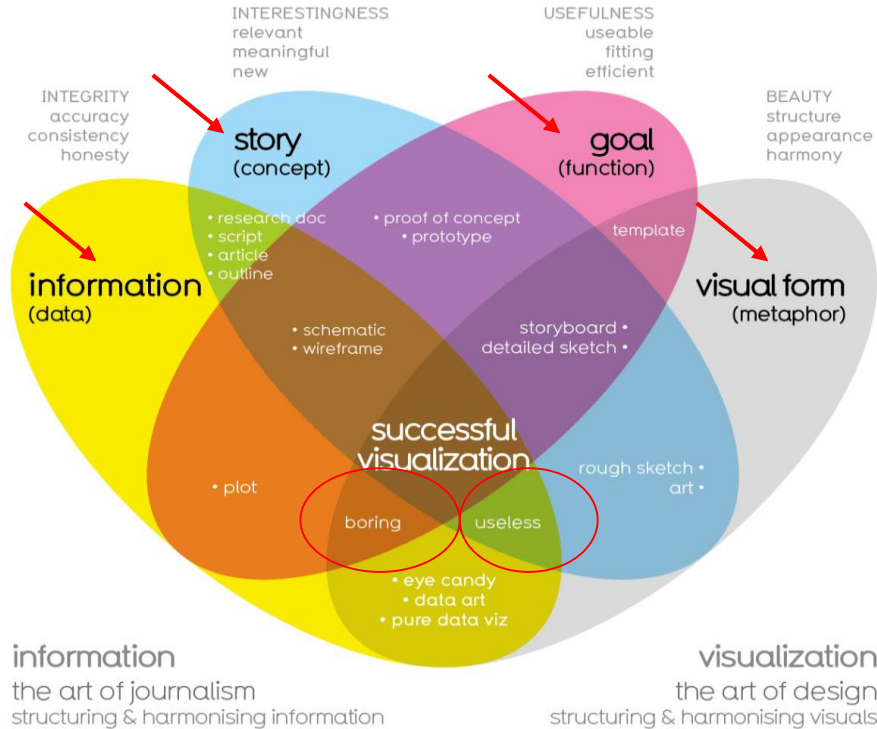
*A good visualization needs to be arranged carefully and in a logical manner*

# Takeaway

---

**Be careful** when designing visualizations, and **be extra careful** when interpreting graphs created by others.

# What Makes a Good Visualization?



David McCandless  
InformationIsBeautiful.net

find out more  
[bit.ly/KIB\\_Books](http://bit.ly/KIB_Books)

## Four Elements:

1. **Information (data):** collect the “correct and required” data/information.
2. **Story (Concept):** provide context, explaining why data matters and suggesting necessary actions. They bridge the gap between data and the audience.
3. **Goal (Function):** Define your data **visualization purpose and approach** upfront, and choose suitable tools, charts, and techniques aligned with your goal for effective data analysis and decision-making.
4. **Visual Form (Metaphor)** presentation of your visualization

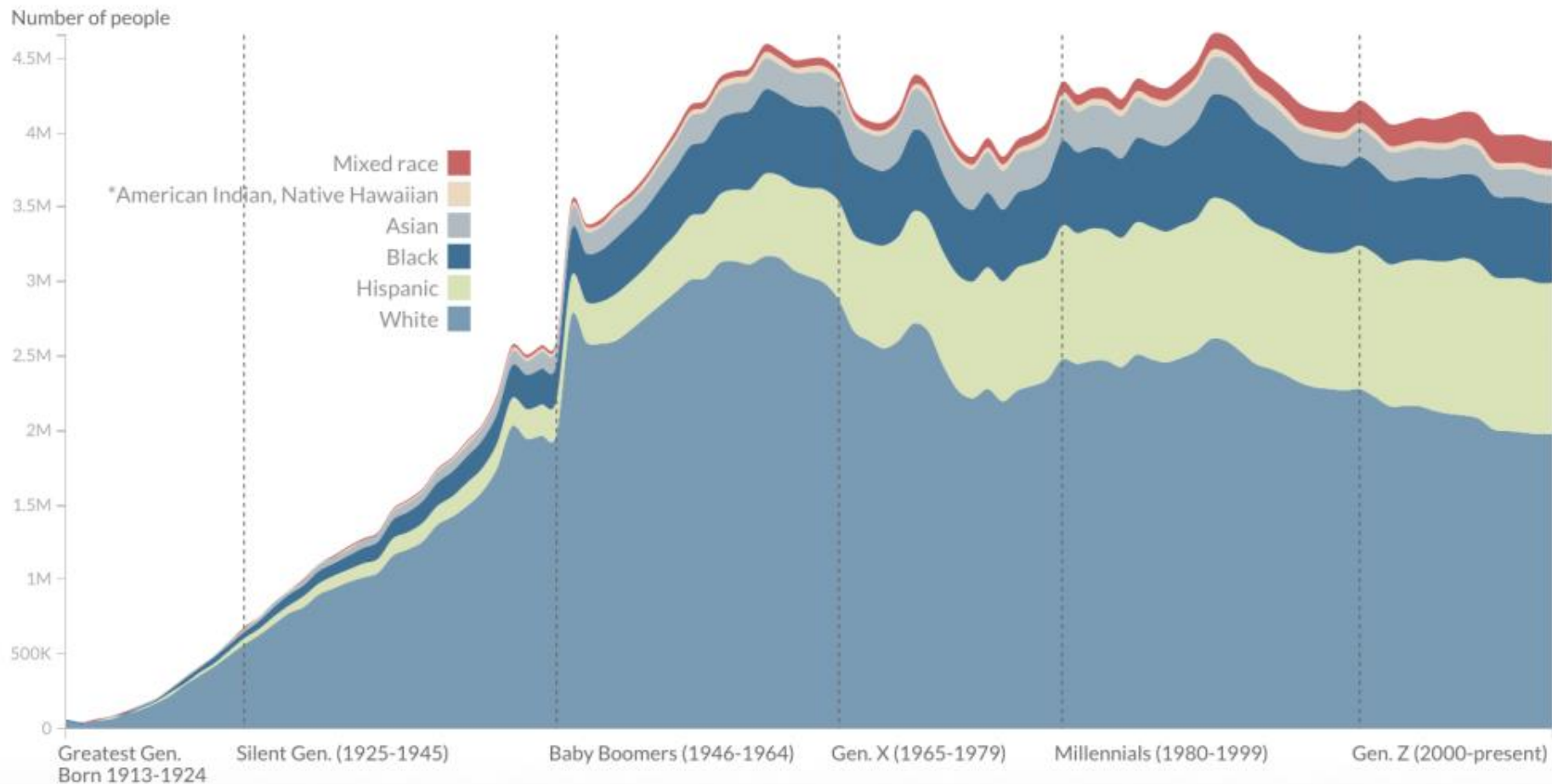


# The Good:



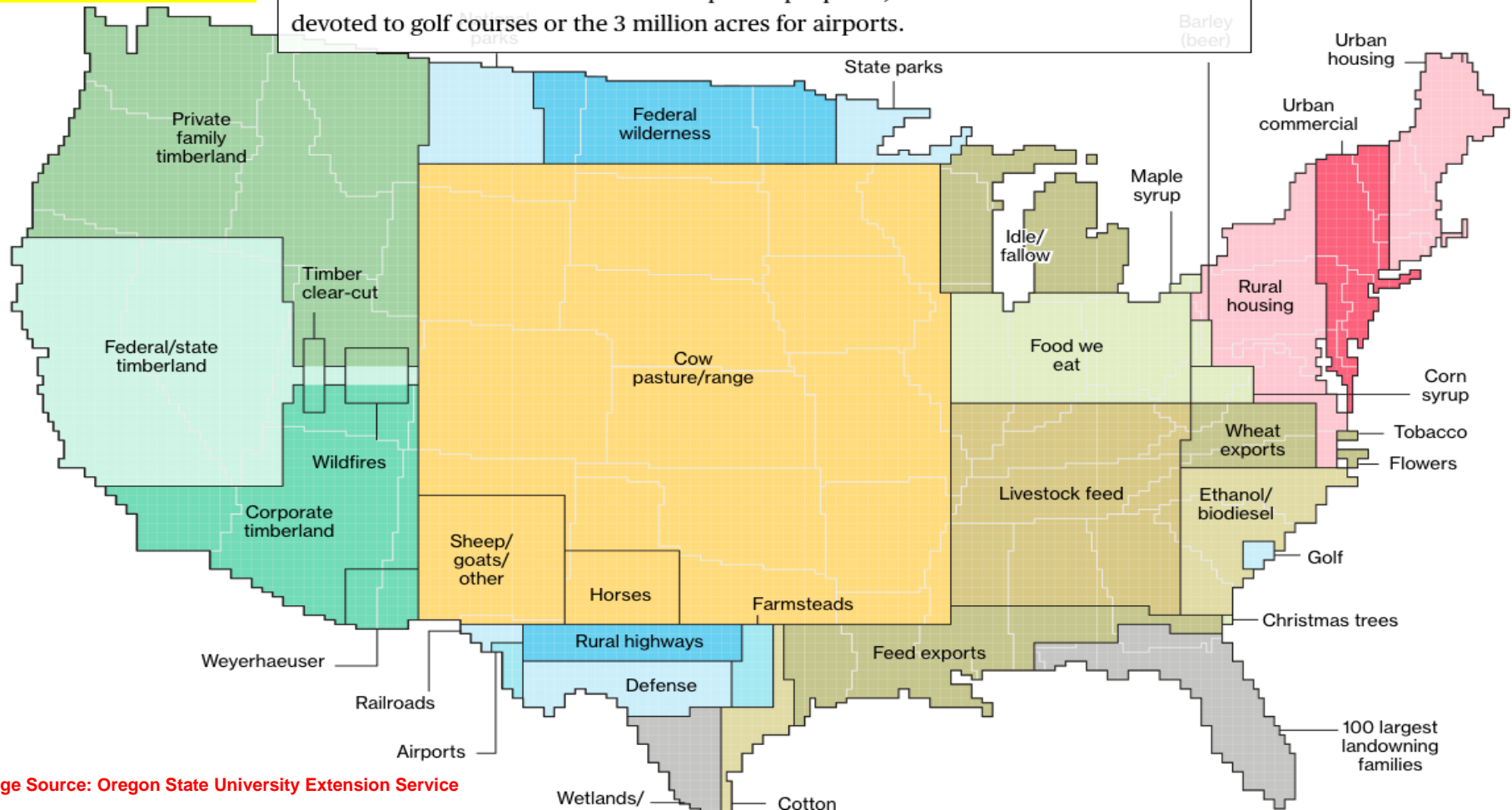
See how diverse your generation is by rolling over the chart below.

*Choose the visualization that tells the story*

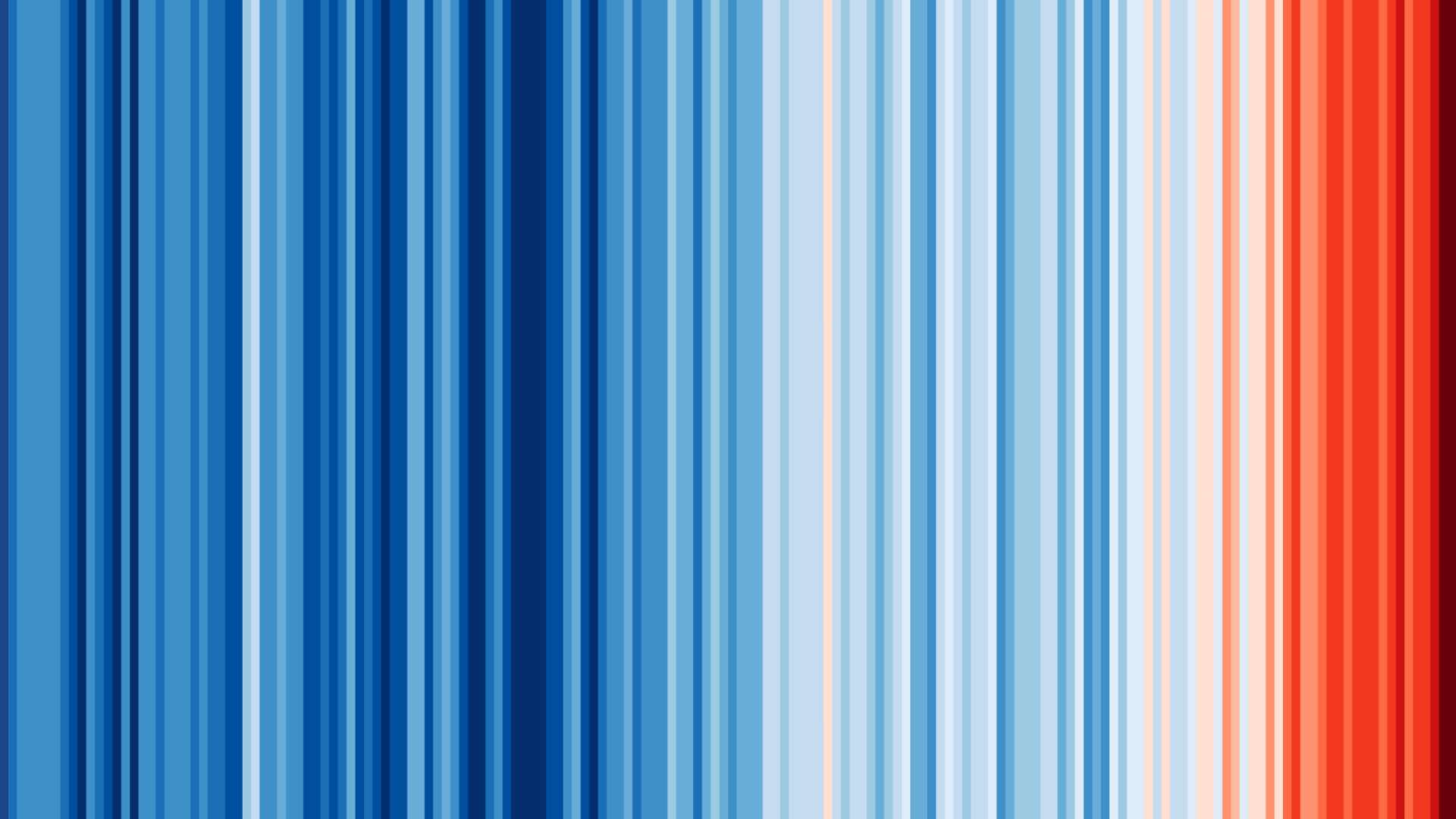


How US uses its land to create wealth

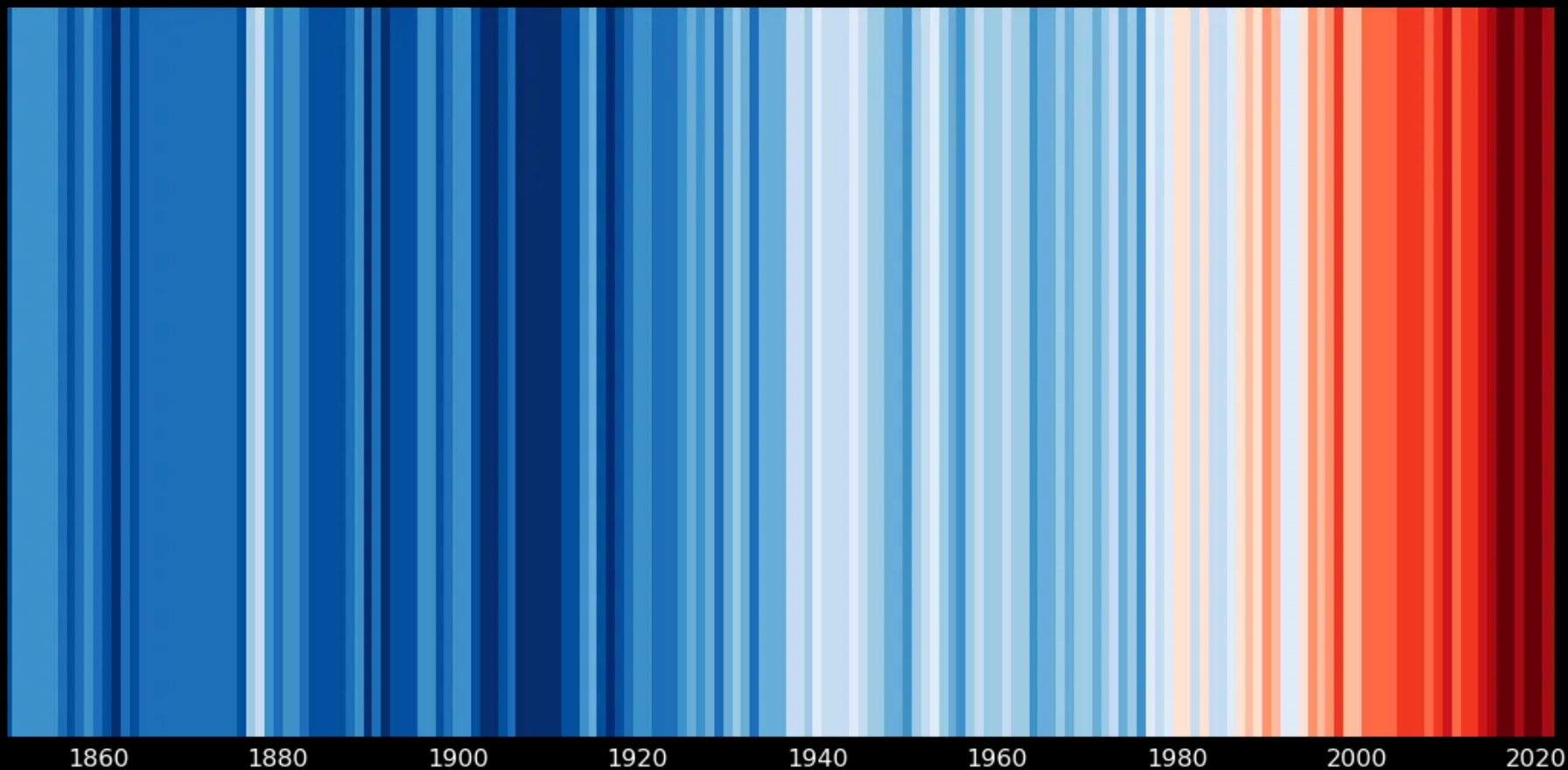
Putting all those pieces together, this map gives you a rough sense of all the ways U.S. land is used. Much of U.S. land serves specific purposes, such as the 2 million acres devoted to golf courses or the 3 million acres for airports.







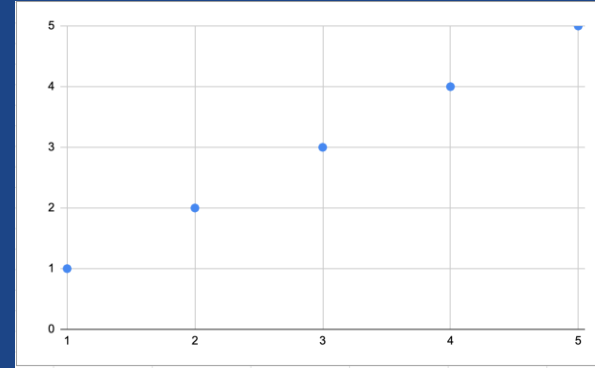
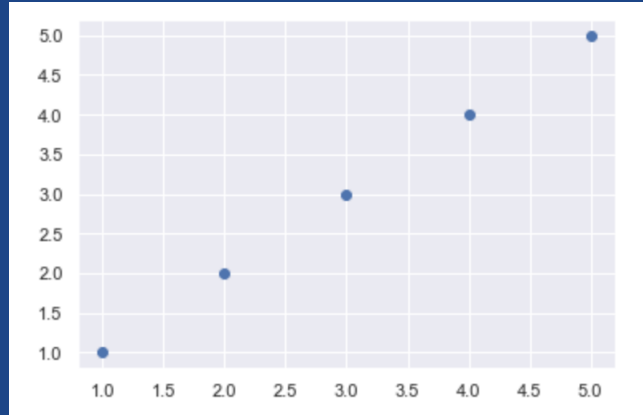
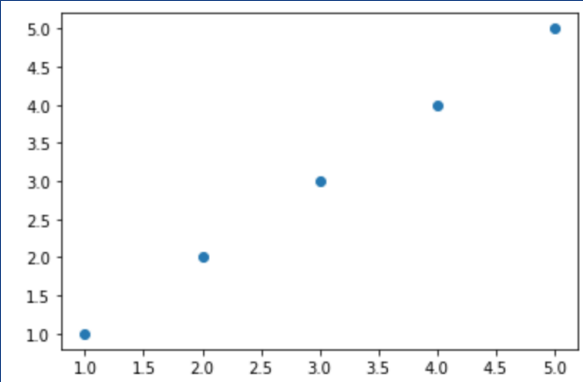
## Global Temperature Change (1850 - 2022)



Source: Met Office - <https://www.metoffice.gov.uk/hadobs/hadcrut5/index.html> / Original idea - Ed Hawkins

# Different Styles of Same Plot (can use Matplotlib)

---



# Importance of Data Visualizations

**Good data visualizations allow us to reason and think effectively about our data.**

By presenting information visually, it allows us offload internal cognition to the perceptual system, making it easier to understand and process.

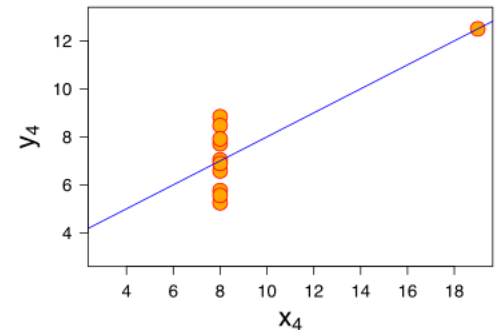
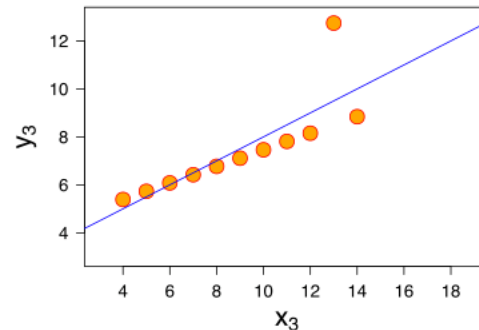
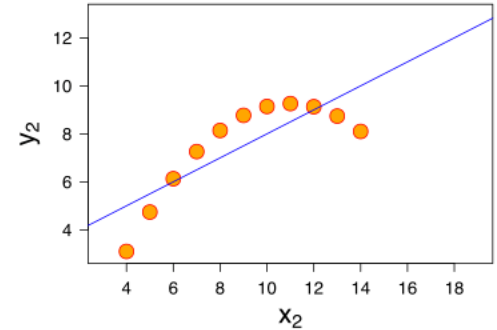
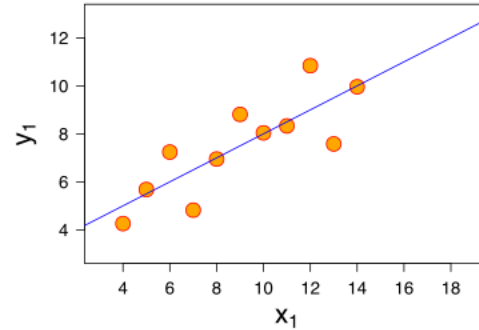
I		II		III		IV	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

**"Anscombe's quartet" (Francis Anscombe, 1973):** four datasets that share the same descriptive statistics, including mean, variance, and correlation.

# Importance of Data Visualizations

When graphically represent them, the patterns are clearly different when plotted on a graph.

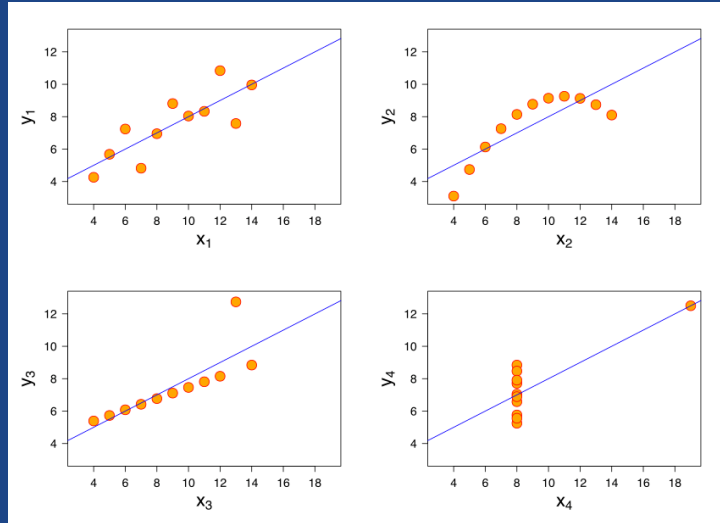
This is the power of effective data visualization: it allows us to bypass cognition by communicating directly with our perceptual system (understand things quickly without having to think too hard).



# Why do visualizations?

Think about how effortlessly your visual cortex understood the complex mathematical relationships in these data sets.

The occipital lobe is highly optimized, especially compared to newer inventions like language and the frontal lobe (which does traditional “math”)



# Who do we create visualizations for?

---

- For ourselves:
  - We want to learn about and explore our data
  - We want to investigate a specific hunch

# An Aside On Communication and Persuasiveness

---

We live in a social world, as much as some of us might not want to.



# Why do visualizations?

---

- For ourselves:
  - We want to learn about and explore our data
  - We want to investigate a specific hunch
- For others:
  - **We want to communicate something to other people**

# Why Data Visualization

---

**Data can be difficult to interpret:** when data is presented in a solely textual manner (Excel docs, CSVs)

*Using graphical, visual representations of information, data visualization presents complex data in easy-to-understand ways, allowing users to quickly discover essential patterns, trends, and anomalies.*

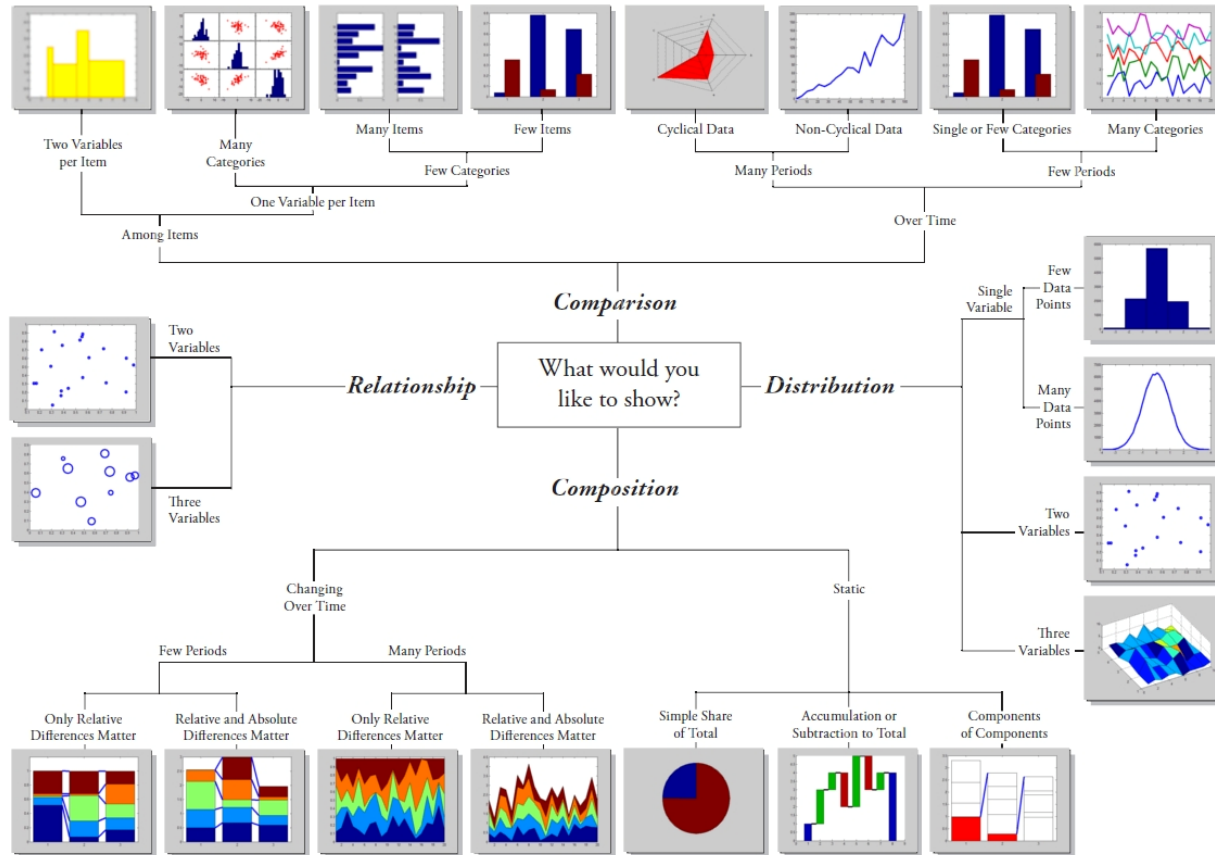


# What Type of Visualization to Use?



# (Data Visualization Cheat Sheet)

## Chart Suggestions—A Thought-Starter



# Why are we spending the lectures on data visualization?

---

- When visualizations are good, they're very good
- When they're bad, they are VERY bad

# Exploration

---

- What sort of things might we want to know about our data?



# Types of Visualization (And when to use them)



Colab: <https://colab.research.google.com/drive/1B0hCLPsGWi2aepKHxtG-PohXZ-piHi39?usp=sharing>

Dataset: [https://drive.google.com/file/d/1IEJVdHjDDHZTfx\\_J7R-lfoLmAWjAIVbM/view?usp=sharing](https://drive.google.com/file/d/1IEJVdHjDDHZTfx_J7R-lfoLmAWjAIVbM/view?usp=sharing)



# Main Types

---

- Data Over Time
- Comparison
- Correlation
- Part-To-Whole
- Distribution



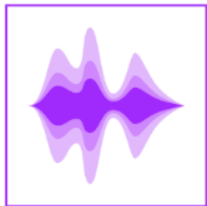
# DATA OVER TIME (TEMPORAL)



AREA CHART



STACKED AREA CHART



STREAM GRAPH



BUMP CHART



BUMP AREA CHART



LINE CHART



SPLINE CHART



STEP LINE CHART



CANDLESTICK CHART



GANTT CHART



BARCODE CHART

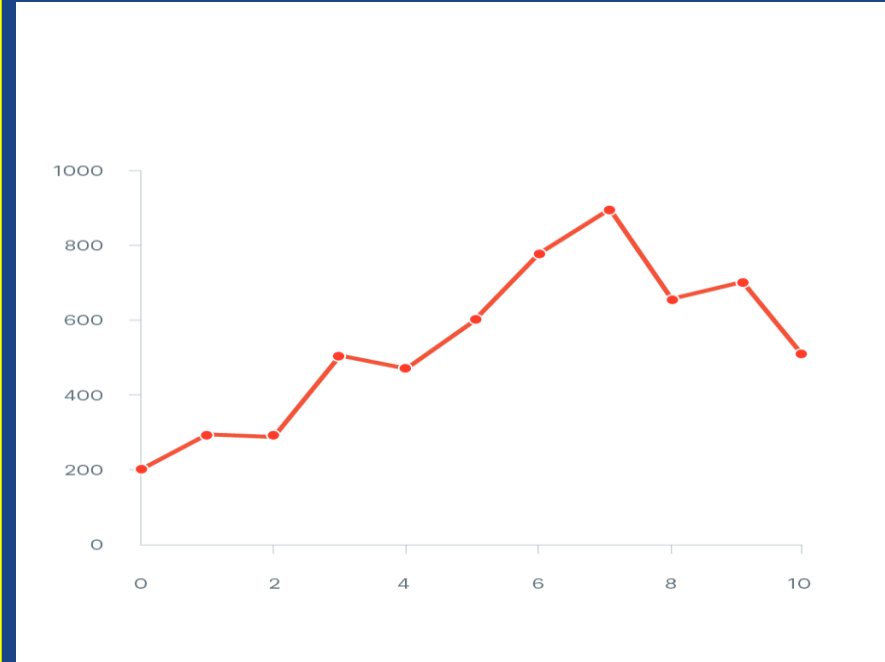


OHLC CHART

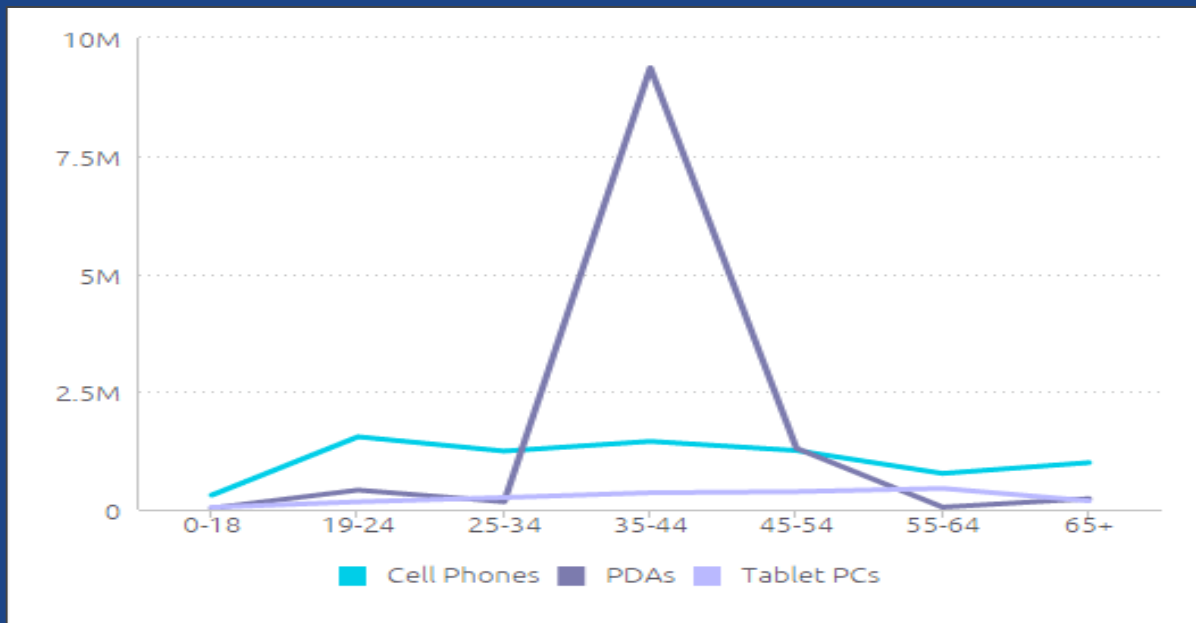
# Line charts - display trends

A line graph reveals **trends or progress over time**.

- Show trends across categories for easy comparison.
- should use it when we chart a continuous data set.



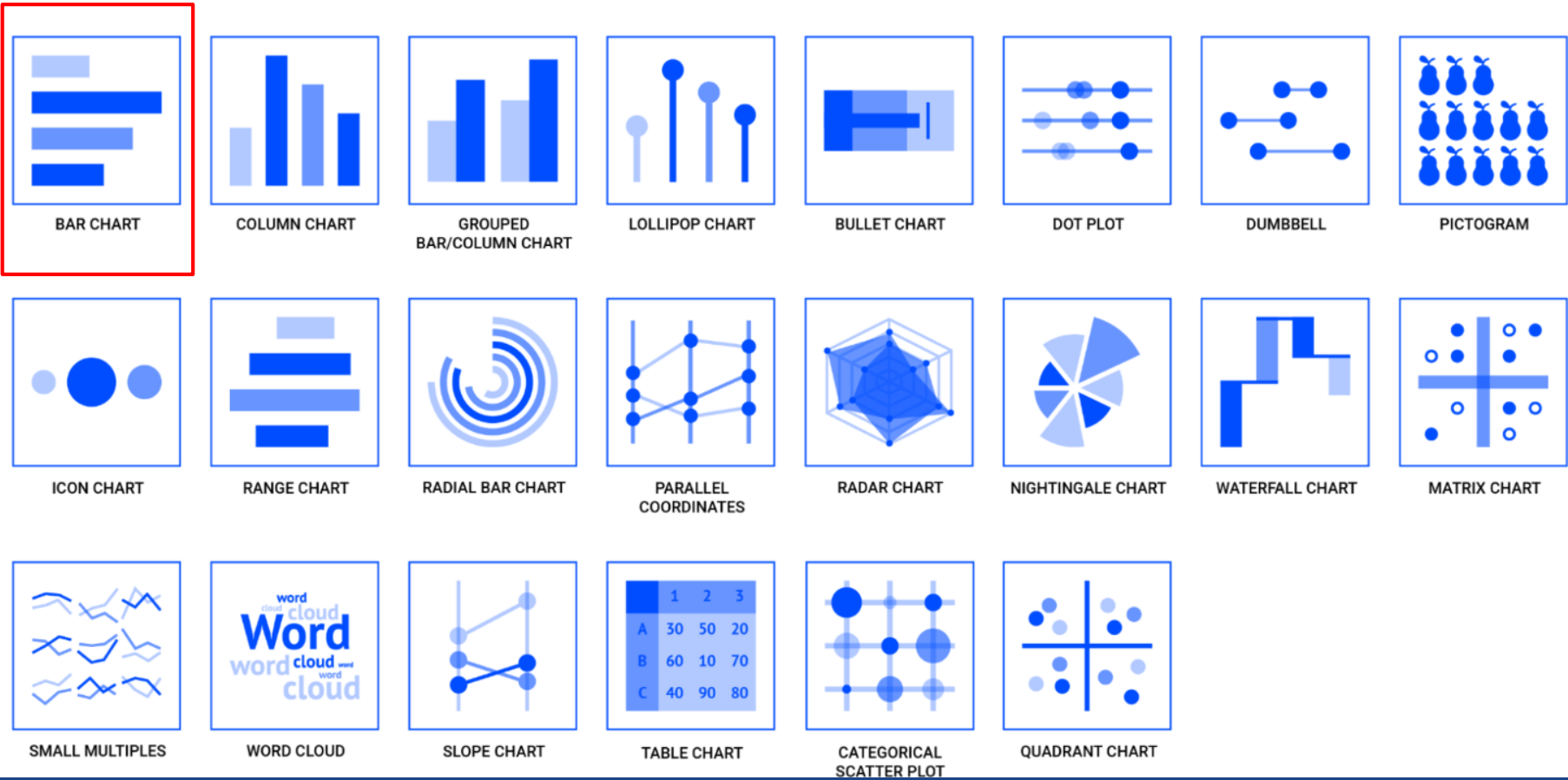
# Example: Line charts



**Observe:** biggest customers are 34-45 year old buyers of PDAs, followed by 19-24 year old buyers of cell phones.

Ex: Sales figures by age group for three different product lines

# COMPARISON



# Bar charts - break things down, simply

Bar graphs can help **comparing data between different groups** or track changes over time.

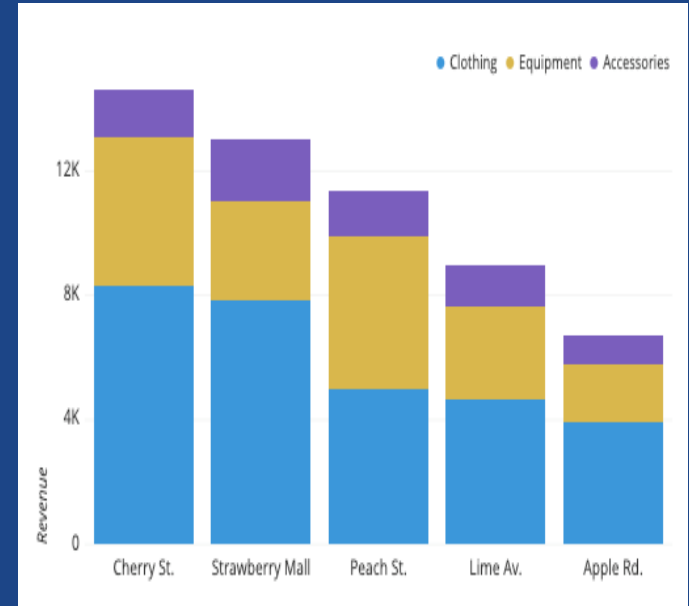
- most useful when there are **big changes** or **when comparing one group against another**.



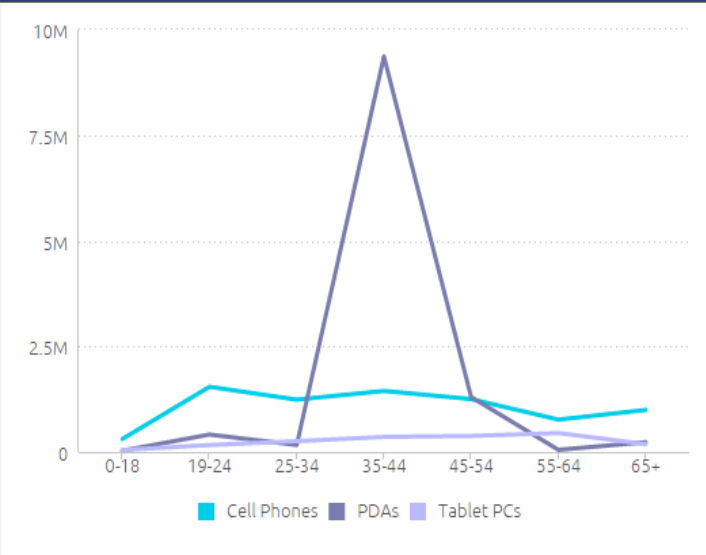
# Bar charts - break things down, simply

Bar graphs can help **comparing data between different groups** or track changes over time.

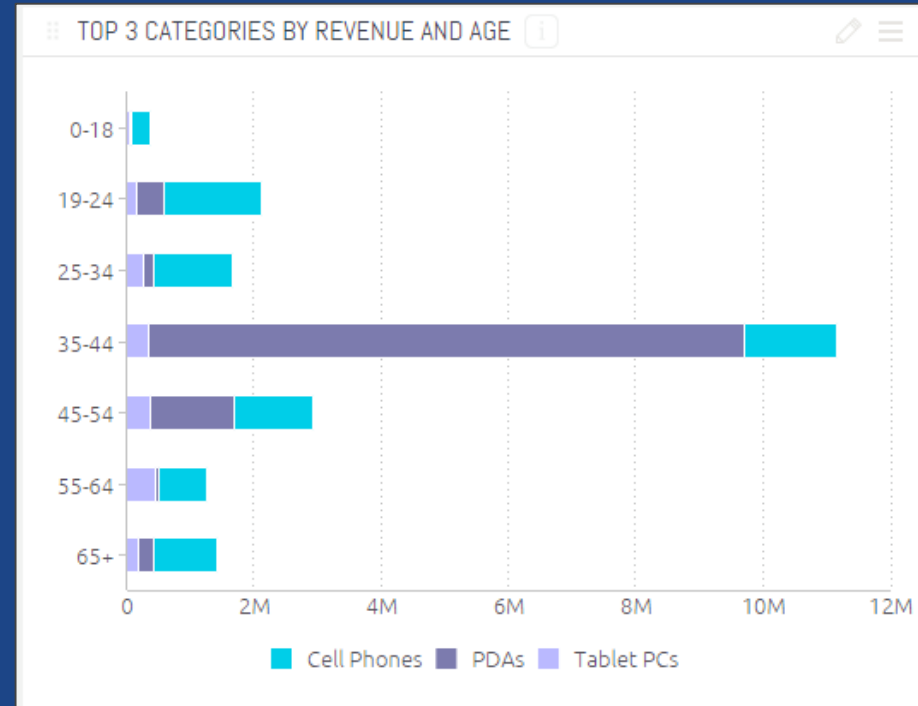
- most useful when there are **big changes** or **when comparing one group against another**.
- **Great for comparing several different values, especially when some of these are broken into color-coded categories.**



# Example: Bar charts



Revisualize  
previous  
chart as a  
bar chart



- Products are group by age
- Explore detailed sales differences within each age category.
- Quickly identify the most valuable age groups for business

# COMPARISON



BAR CHART



COLUMN CHART



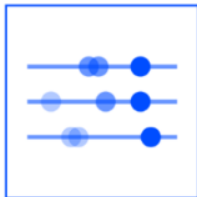
GROUPED  
BAR/COLUMN CHART



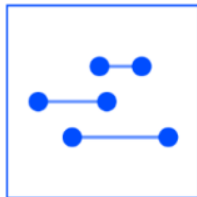
LOLLIPOP CHART



BULLET CHART



DOT PLOT



DUMBBELL



PICTOGRAM



ICON CHART



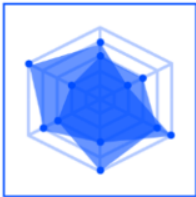
RANGE CHART



RADIAL BAR CHART



PARALLEL  
COORDINATES



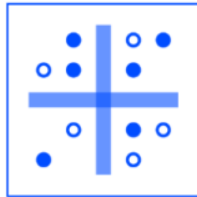
RADAR CHART



NIGHTINGALE CHART



WATERFALL CHART



MATRIX CHART



SMALL MULTIPLES



WORD CLOUD



SLOPE CHART

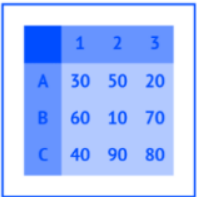
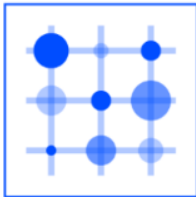


TABLE CHART



CATEGORICAL  
SCATTER PLOT



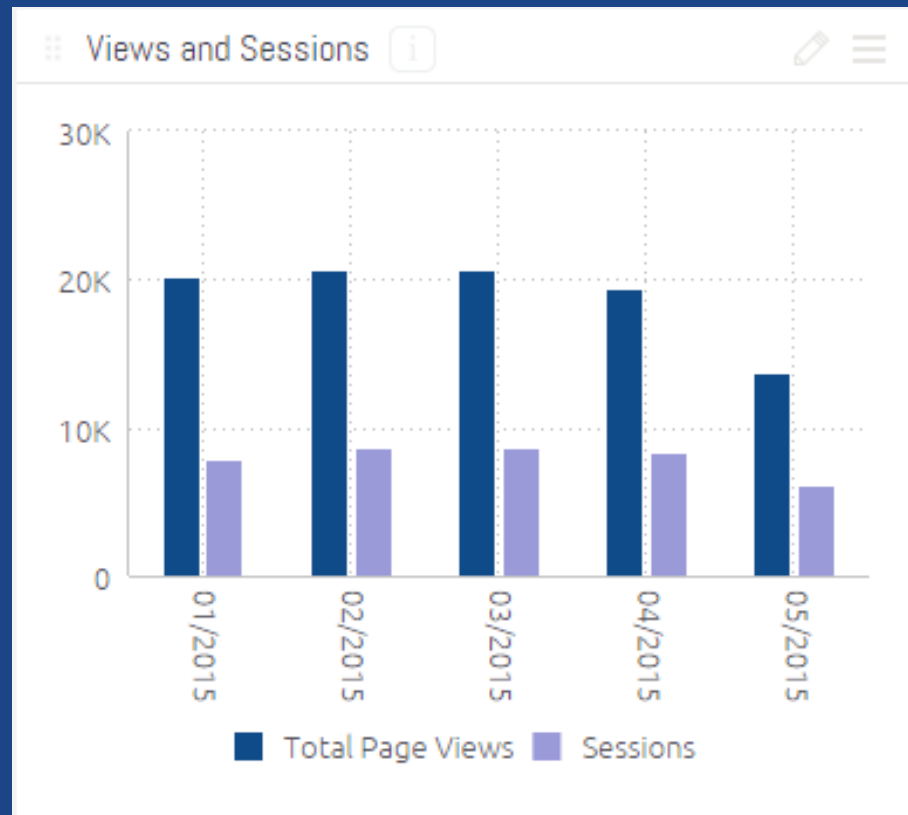
QUADRANT CHART



# Column charts - compare values side-by-side

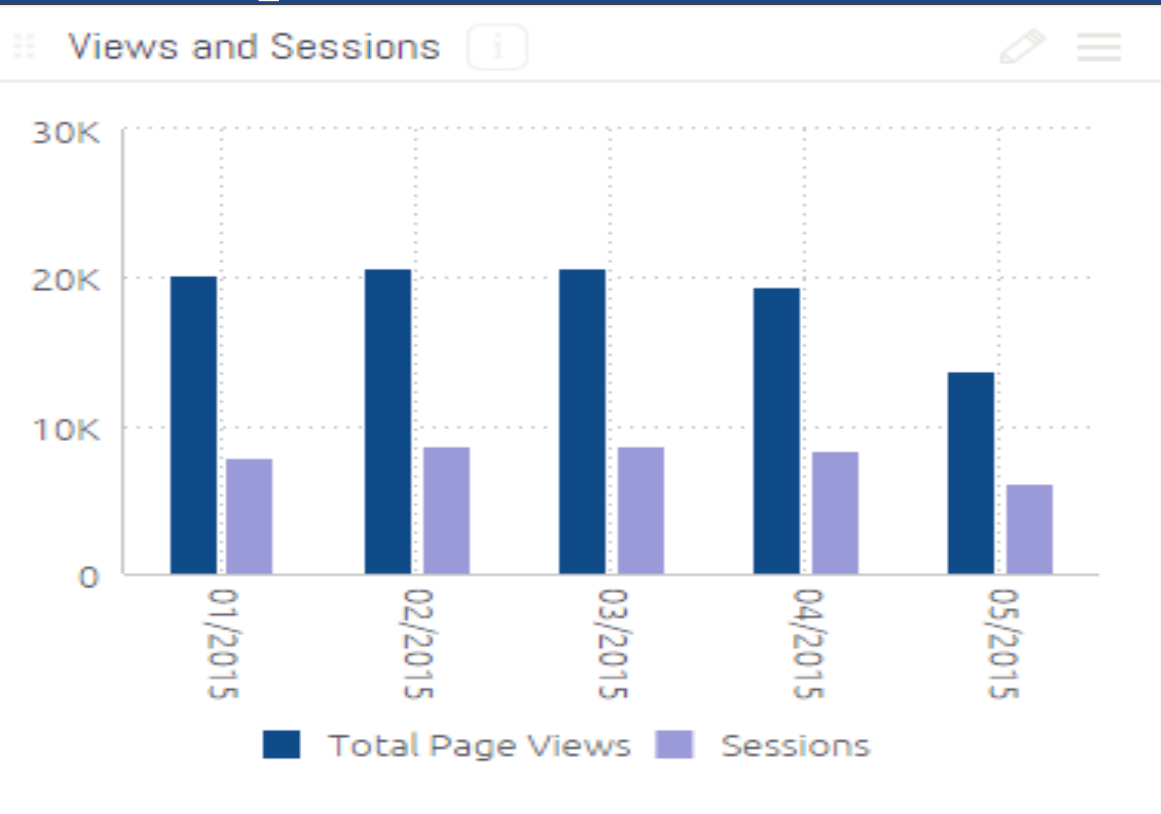
Use for **side-by-side comparisons of different values.**

- Show changes over time
- Ideal for scenarios **where daily changes are minimal.**



Total website page views vs. sessions on various dates.

# Example: Column charts



Use for side-by-side comparisons of different values.

- Highlights concrete numbers, such as the daily number of website visitors.

Plot: Total website page views vs. sessions on various dates.

# CORRELATION



HEATMAP



BUBBLE CHART



SCATTER PLOT



CONNECTED  
SCATTER PLOT

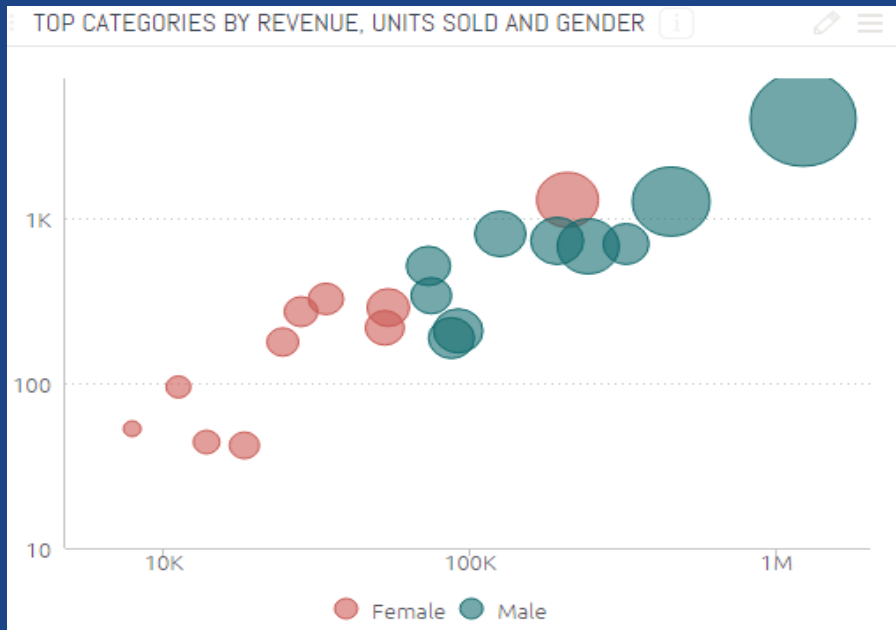


HEXAGONAL BINNING



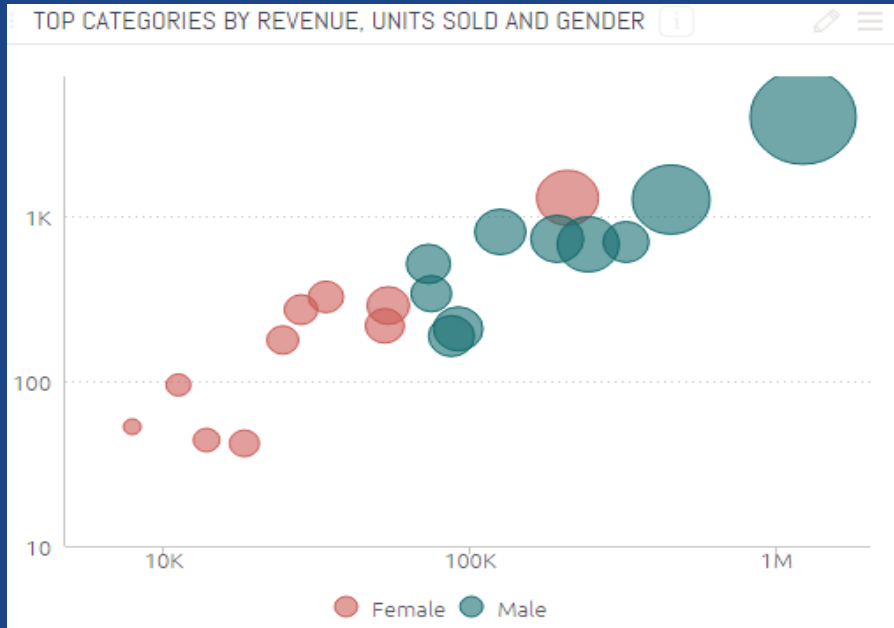
CONTOUR PLOT

# Scatter charts - relationships



- Displays the relationship **between two continuous variables**.
- It involves **plotting individual data points as dots on a graph**, with one variable on the x-axis and the other variable on the y-axis.
  - Each dot represents an observation in the dataset, and the pattern formed by the dots can reveal the nature and strength of the relationship between the variables.

# Scatter charts - relationships



- The chart **visualizes each product line by the number of units sold (x-axis) and the revenue this brings in (y-axis), representing the value in physical size.**
- It also **breaks this down by gender** (hovering over the circles would reveal the name of the product in the original).

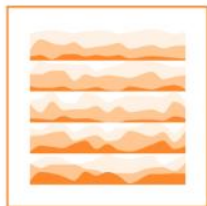
# DISTRIBUTION



DENSITY PLOT



RIDGELINE PLOT



HORIZON CHART



HISTOGRAM



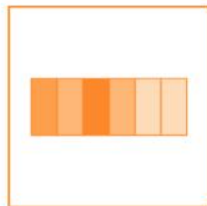
RADIAL HISTOGRAM



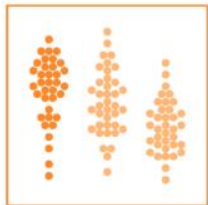
STRIP PLOT



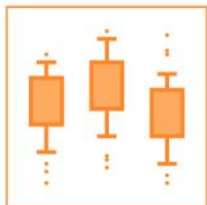
JITTER PLOT



ONE-DIMENSIONAL  
HEATMAP



BEESWARM CHART



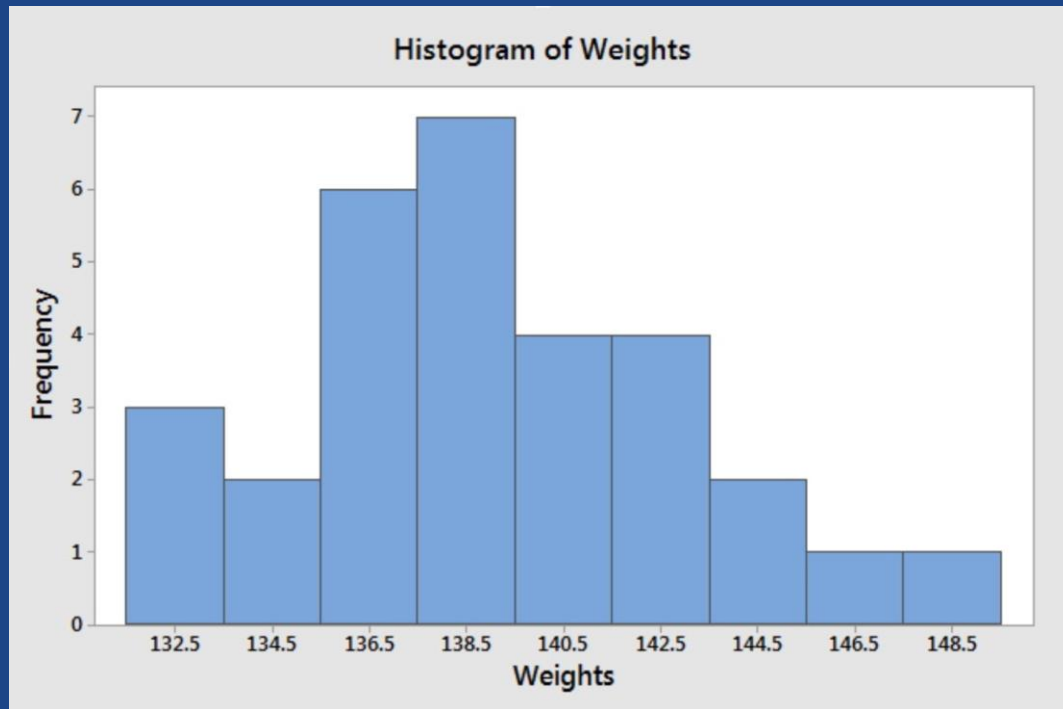
BOX CHART



VIOLIN PLOT

# Histogram

A histogram is a graph that shows the **frequency of numerical data using rectangles**.



# PART-TO-WHOLE & HIERARCHICAL



STACKED  
BAR/COLUMN CHART



DIVERGING  
BAR CHART



POPULATION PYRAMID



ICON ARRAY



WAFFLE CHART



PIE CHART



DONUT CHART



SEMI-CIRCLE  
DONUT CHART



MARIMEKKO CHART



TREEMAP



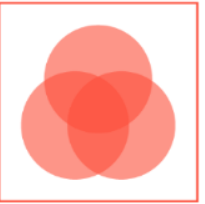
CIRCULAR TREEMAP



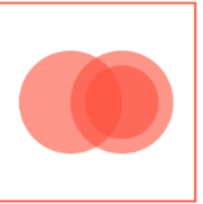
CONVEX TREEMAP



DENDROGRAM



VENN DIAGRAM



EULER DIAGRAM



CIRCULAR GAUGE



SUNBURST CHART



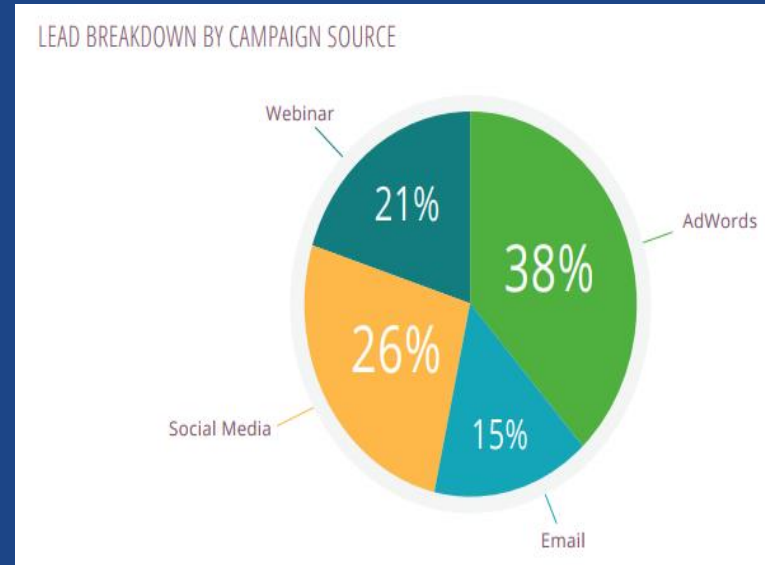
FUNNEL & PYRAMID  
CHART



# Pie charts clearly show proportions

Pie charts easily show the **share each value contributes to the whole**.

- More intuitive than simply listing percentages that **add up to 100%**.

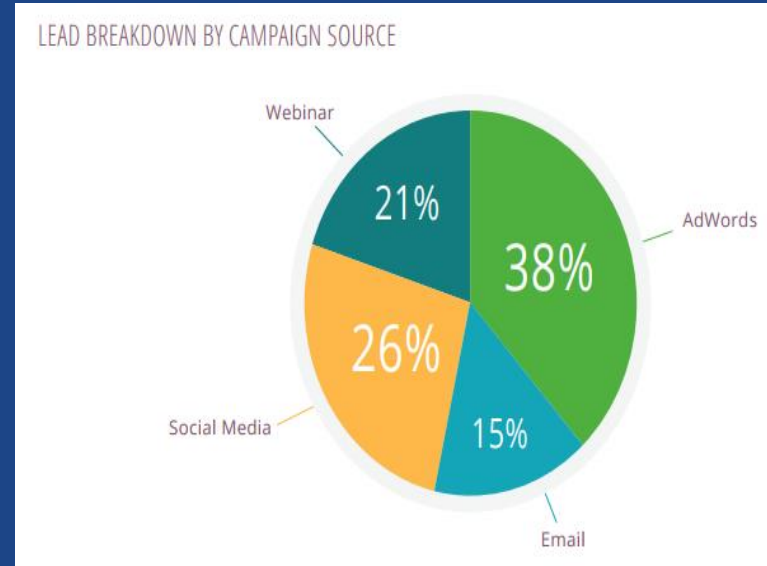


Example: This pie chart illustrates the **effectiveness of different marketing campaigns** in generating leads

# Pie charts clearly show proportions

**Example:** Clearly shows that **AdWords is the most effective**, followed by social media and webinar signups.

- An instant insight would illuminate to the marketing team what's working best, helping them to rapidly reassign resources or refocus their efforts to maximize lead generation.



Example: This pie chart illustrates the **effectiveness of different marketing campaigns** in generating leads

# GEOSPATIAL & OTHER



GEOGRAPHIC  
HEATMAP



CHOROPLETH MAP



TILE MAP



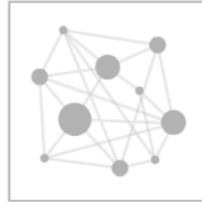
CHORD DIAGRAM



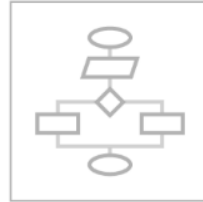
ARC DIAGRAM



SANKEY



NETWORK DIAGRAM



FLOWCHART

# Visualization using Programming

---

## Python

- matplotlib
- seaborn
- plotly
- pylab

## R

- graphics
- ggplot2

# Common Tools for Data Visualization

---

- Tableau
- Microsoft Power BI
- Infogram
- ChartBlocks
- D3.js
- Google Charts
- Fusion Charts
- Chart.js



# Storytelling



# Why Tell Stories with Data?

---

We've been using **stories to teach since there was language**. We are much more likely to remember facts and internalize an idea when presented as a story.

Psychologists argue that its **especially difficult for us to make sense of statistics without narrative**.

# Let the data tell the story

---

The data **cannot tell** you what is important.

- ❑ If you **don't give** the audience **a story**, they will have to do the work of creating one (or worse, they won't bother to engage at all).
- ❑ With this understanding, **let the data speak for itself** — but present it in a way that makes it easy for the audience to build the correct story.
- ❑ When creating a narrative structure, the most basic question is, “**what do you want the audience to know, and when?**” ... this is also the basic question of data storytelling.



# What's in a Story?

---

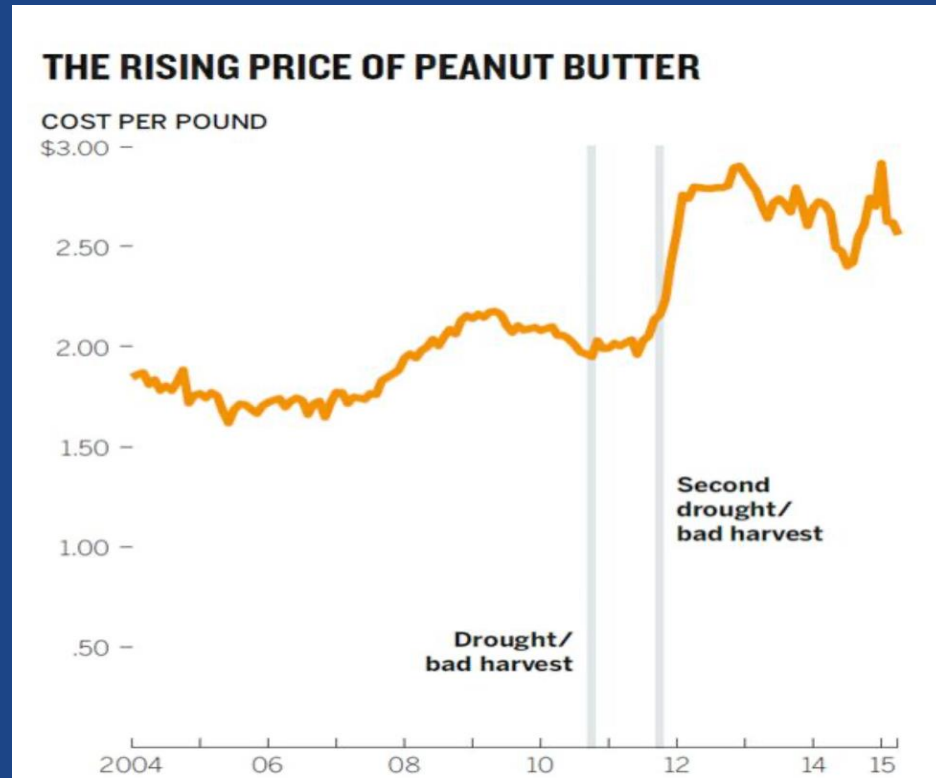
- Sequence, change over time
- Characters and setting
- Conflict and resolution

# Example: A Story

---

Throughout the 1990s and into the early 2000s, the price of copper was stable, and historically low at about 65 cents per pound. But then in 2003 a landslide at a mine sent prices over \$1 per pound. After a strike at a mine in Chile in 2004, prices passed \$2 per pound. Because of these events and continued high demand, production fell below consumption, which caused prices to reach nearly \$4 per pound by 2006.

# A Data Vis Story

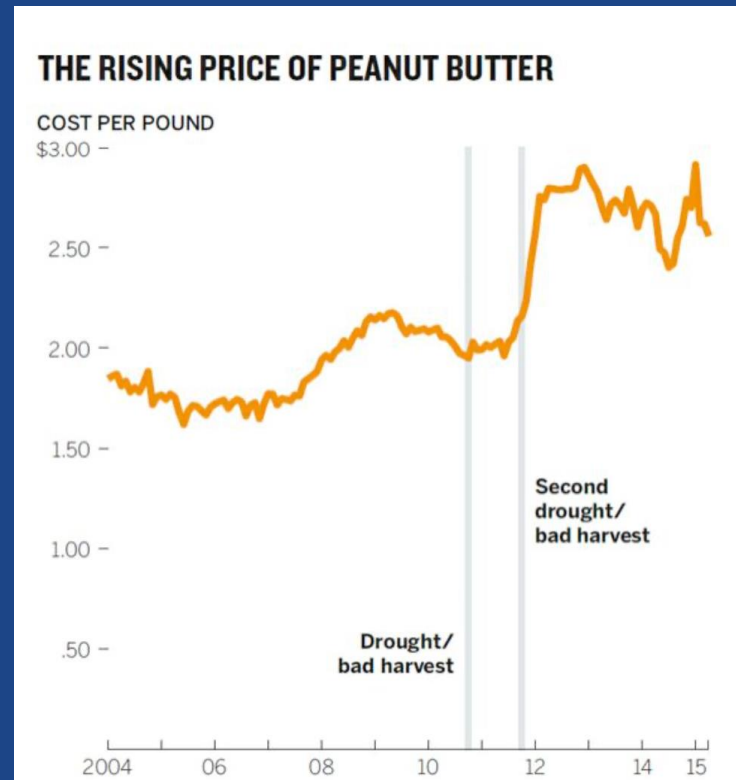


# Data Viz Stories

**Attention theory** predicts that

- you will **first see the spike in price**,
- then **look for events** related to it,
- then **read the title** of the chart.

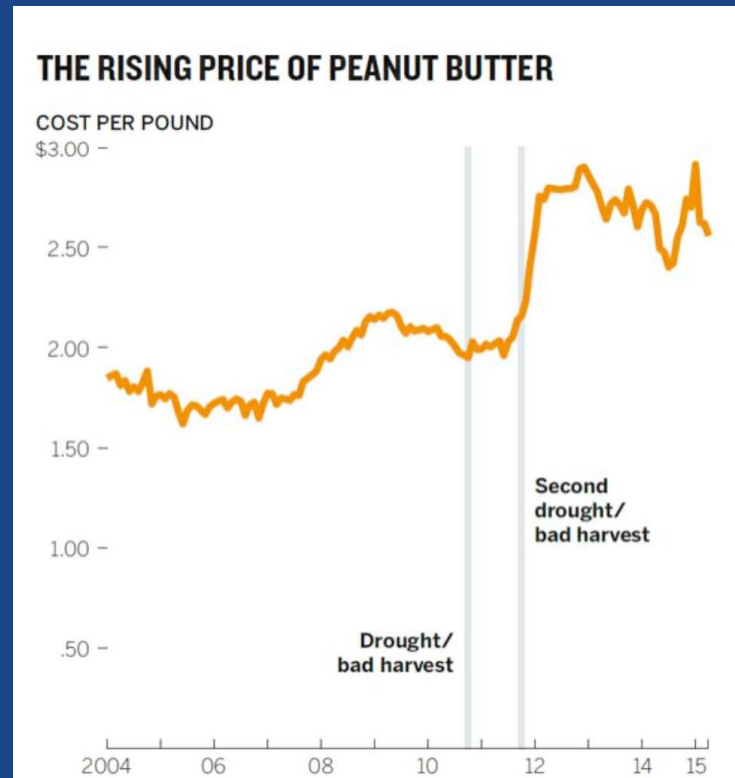
**Preattentive features** govern the sequence of the narrative (instead of the text “contract”).



In simpler terms, your eyes are **drawn to visual cues** like changes in price before you read the actual words on the chart.

# But Why Data Viz Stories?

- **Transfer of information is quick and effortless** with the **picture** when compared to text.
- We **need narrative to make sense of statistics**, and **narratives emerge most quickly** when they are visual.



# What's in a Story: Sequence

---

Many ways to create sequence and control pacing in a data story...

- **static visual** that guides the viewer's eyes in a specific sequence via visual salience.
- **interactive data visual** with a “click-through” feature
- **sequence of slides** in a presentation

# Slow Reveal

---

## WHAT CAUSED THE HIGHER PRICE OF PEANUT BUTTER?

PRICE PER POUND

\$3 -

2 -

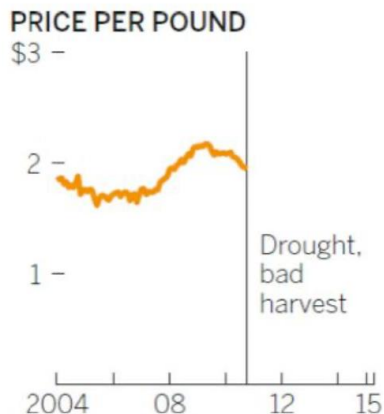
1 -

2004 08 12 15

[Pause 5 beats] For nearly a decade, prices of peanut butter were smooth. Modest increases during the recession mapped to rises in most food prices.

# Slow Reveal

## WHAT CAUSED THE HIGHER PRICE OF PEANUT BUTTER?



But then there was a drought and a bad peanut harvest. What happened? Where did prices go from here? [Pause, invite speculation].



# Slow Reveal

---

## WHAT CAUSED THE HIGHER PRICE OF PEANUT BUTTER?



For most of the following year, prices were actually flat, until the next harvest approached, when prices started to rise as growers feared another drought.

# Slow Reveal

## WHAT CAUSED THE HIGHER PRICE OF PEANUT BUTTER?



And that's what happened: another bad harvest.

Prices spiked almost 50%. And since then . . .

[Pause]

# Storytelling

---

**Controlling the chronology of your story** is important, but it takes more to make the story impactful ...

**Characters:** stories resonate because they happen to specific people with whom we can develop empathy. Make sure your audience knows exactly who is affected by your data story.

**Setting:** the background information, the context in which the data exists; without which interpretation would be impossible.

# Three Act Structure

The Three-Act Structure is a commonly used storytelling framework which divides a story into three main parts:

**Act 1 - Setup:** here is some reality.

Peanut butter costs X

**Act 2 - Conflict:** here is new information that complicates or changes reality.

A drought!

**Act 3 - Resolution:** here is the new reality.

Peanut butter costs Y

*For more details, check additional slides*



# Focusing Attention



# Visual Perception

---

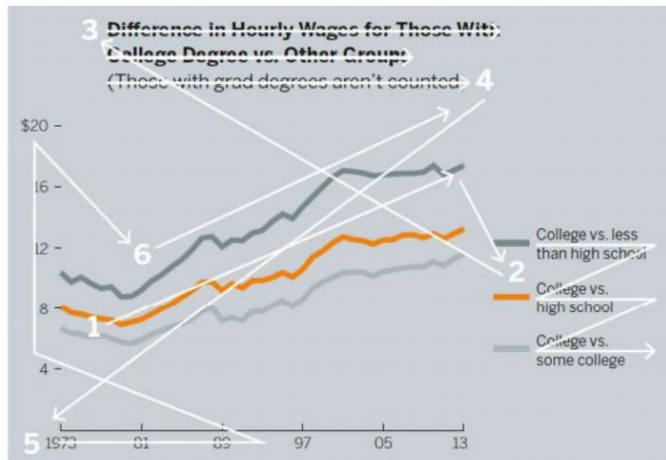
Visual processing happens in its own part of the brain, based on its own rules!

Presentators can leverage research helping us understand these rules ...

- Visual Attention Theory
- Visual Salience
- Preattentive features

●

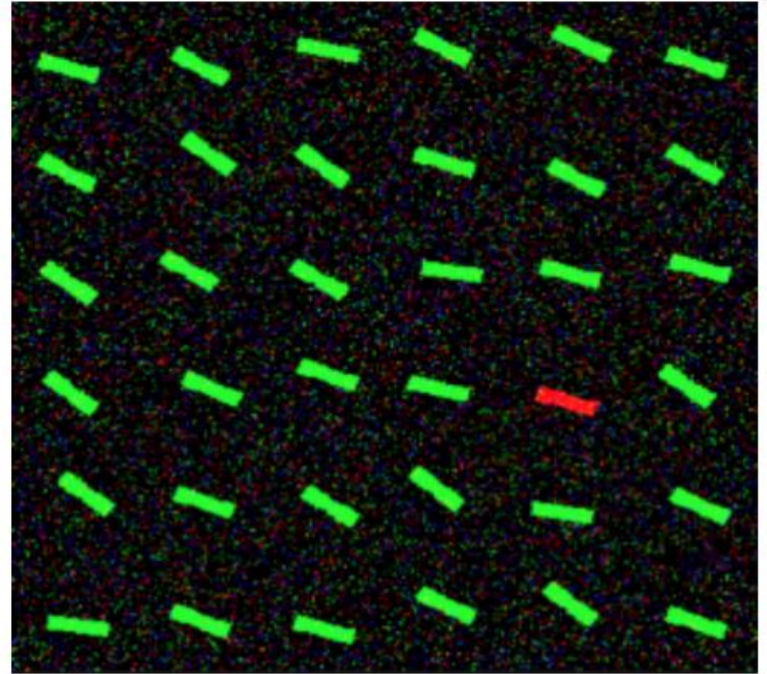
**19.1**



# Visual Attention Theory

---

Track where your attention goes!



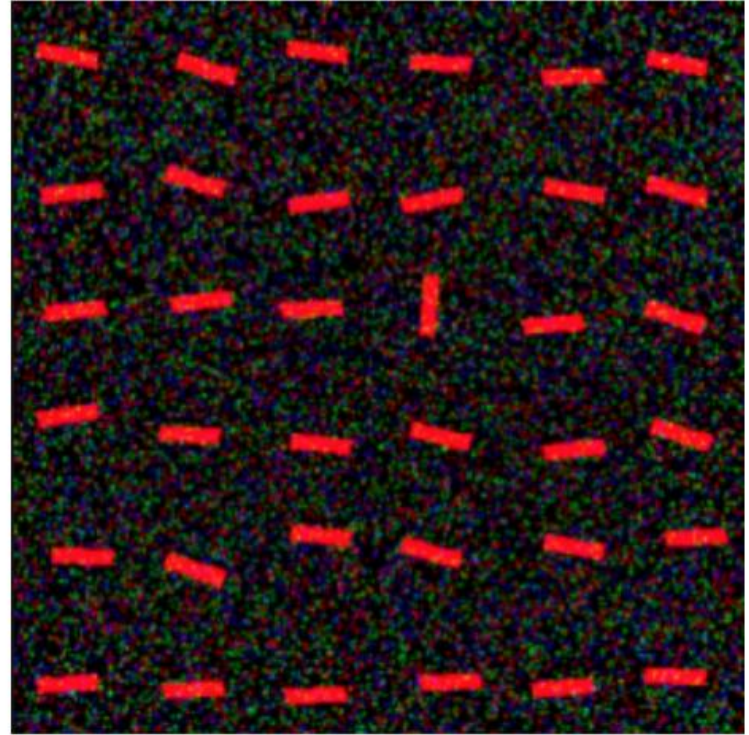


# Visual Attention Theory

The horizontal red bar was salient in the first image ... but visual properties alone do not determine salience; context is crucial.

Orientation and color are examples of preattentive features.

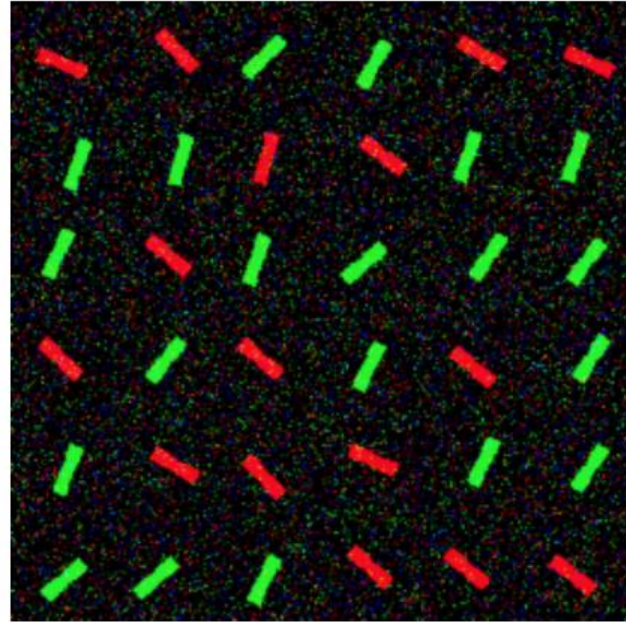
Preattentive processing is performed automatically and instantly\* on the entire visual field, but only for a limited set of object features.



# Visual Attention Theory

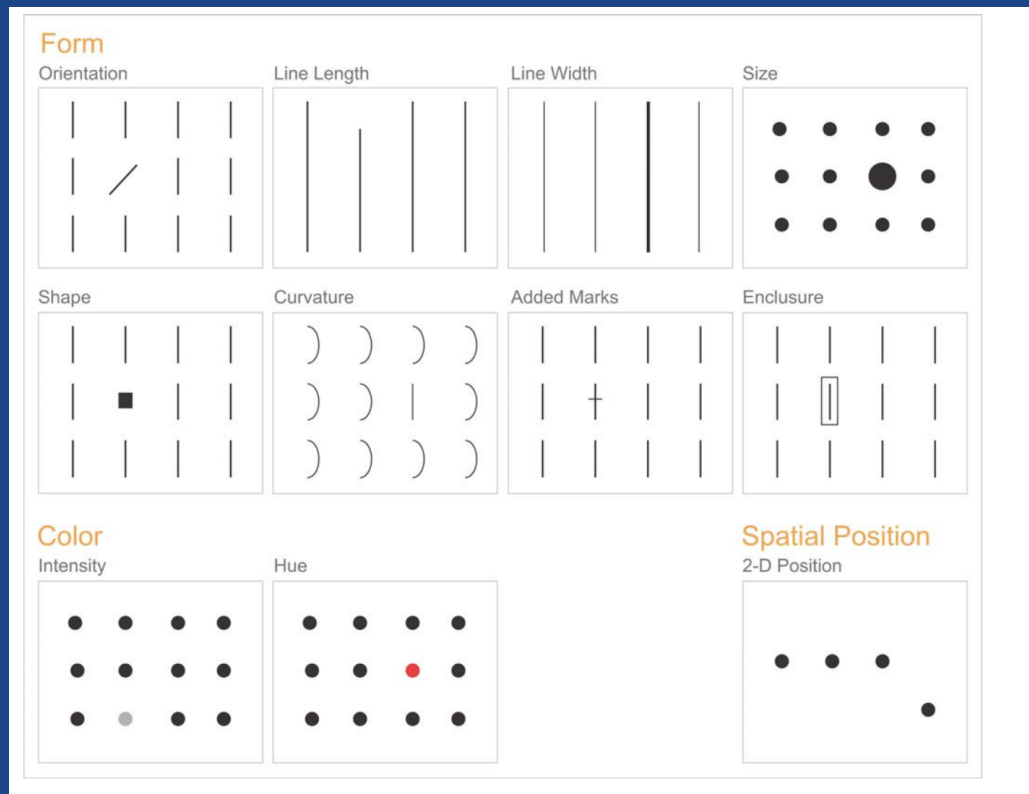
---

Hard to parse both at once!



# Preattentive Features

Basic features such as **color**, **size**, or **tilt** (known as channels) are extracted from the display effortlessly — without the need to focus attention



List of the preattentive attributes that are of particular use in visual displays of data

# Using Color

Don't overuse your most powerful tool!

"If you have more than one color (other than grayscale), people will look for meaning in the colors."

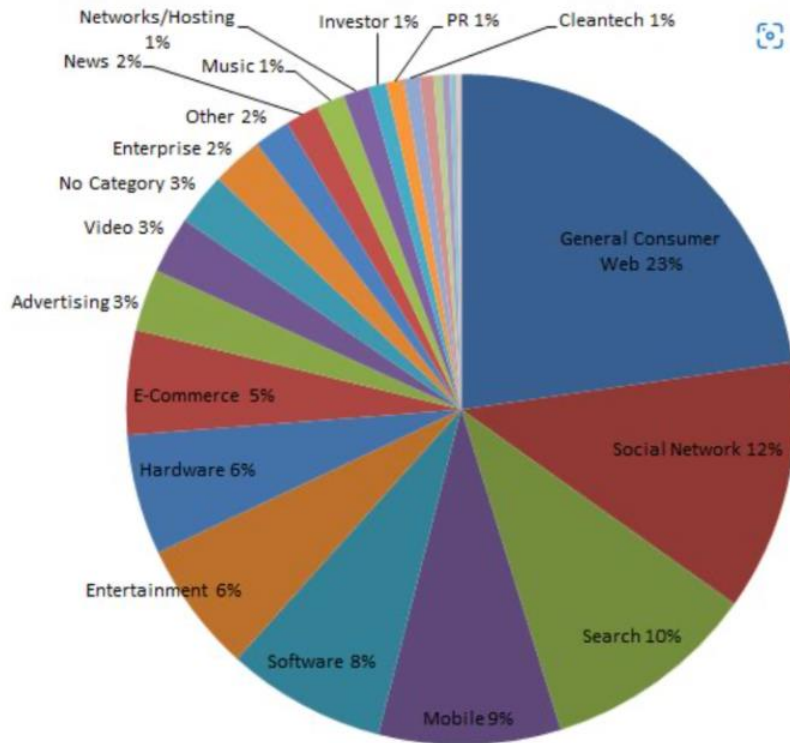
*Edward Tufte*

"Bright colors often serve no purpose, except for being disruptive."

*Edward Tufte*

"Use color sparingly. Learn to love grayscale."

*Enrico Bertini*



# Design Principles

---

- Each visualization **should be designed** to **communicate a specific insight or two**.
  - **Don't ever include** a graph just because you feel like you should!
  - **Know what** each graph is telling your viewers
- **Maximize information** and minimize ink
  - **Avoid** anything extraneous
- **Organize** hierarchically
  - **Start** broad and zoom in
- **Dazzle**
  - A good visualization is worth a lot



# Additional Reading Slides



# The Good:



# Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Léger, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davout qui avaient été détachés sur Minsk et Mohilew et qui rejoignent Oescha et Witebsk, avaient toujours marché avec l'armée.

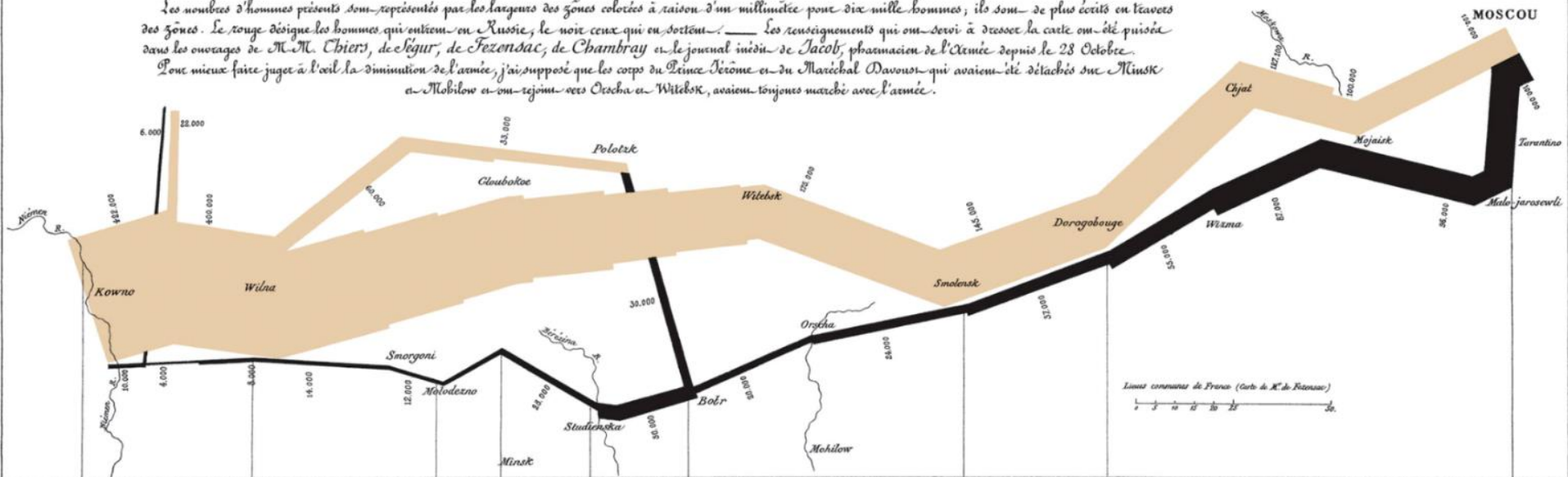
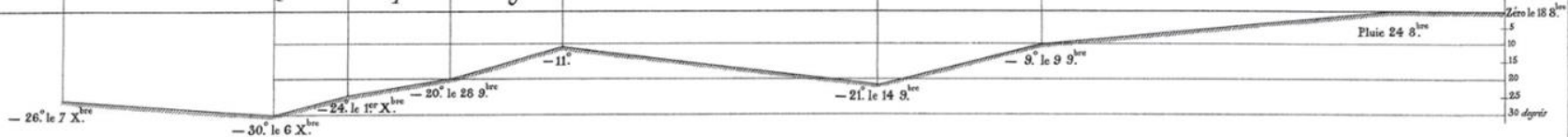


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Les Cosaques passent au galop le Niemen gelé.



# Storytelling

---

# Act One: Setup

---

“The first act of a film serves to introduce the audience to a protagonist, establish the setting, provide information into how the characters’ world works, and introduce an obstacle that sets the rest of the story in motion.”

In traditional dashboards and reports, this information is often missing and leads to users not knowing where to start.

If your audience is going to go on a data adventure with you, they should start off by caring about the situation that exists.

Data stories should start with a high-level summary that then lets users progressively and logically drill into more complex details and context.

# Act Two: Conflict

---

Pixar states that the second act of a story is “a series of progressive complications.”

When it comes to data stories, the conflict can often take the form of an event that needs explanation — but even more tangible problems can be framed as “the mystery of how my problem can be solved.”

Act two is the back-and-forth exploration of the problem; slicing-and-dicing, changing the way the data is viewed and trying out interpretations.

Tension is created by highlighting a mismatch between expectation and reality.

# Act Two: Conflict

Use reference charts. Companion visuals that show “ideal” or “average” users can add context and make your chart easier to understand.

With Radar/Spider graph, one shape combines many measures. But the shape is meaningless without context to aid in interpretation.



# Act Three: Resolution

---

In traditional storytelling, the third act is the part of the story where the main character learns what she truly needs — as opposed to what she thought she wanted. The character has gone on a transformation along the course of the story, and that is evidenced in the final act.

In a data story that is intended to inspire the audience to act, the audience should be considered the protagonist. Your presentation will help them to resolve the conflict between their expectations and reality, to answer their questions, to learn something new about the world.

If you've also helped them relate to the characters impacted by the story, they will be motivated to take action.