Visualization for Data Science
Exploratory Data Analysis I

Why These Characteristics Matter: The Science Behind Good Design

Accuracy - *How precisely can we see quantitative differences?*

Foundation: Stevens' Power Law & Cleveland/McGill experiments
Principle: Position is most accurate, area systematically fools us
Design Guideline: Use position/length for precise comparisons, avoid area when accuracy matters

Discriminability - *How many levels can we actually distinguish?*

Foundation: Weber's Law (just noticeable differences)
Principle: We have built-in perceptual limits - only ~6 color lightness levels, ~10 area sizes
Design Guideline: Don't create more categories than people can actually see

Pop-out - *What grabs attention immediately?*

Foundation: Treisman's Feature Integration Theory
Principle: Color and motion jump out instantly, position requires focused attention
Design Guideline: Use color hue to highlight important categories, not for precise values

Separability - *Do channels interfere with each other?*

Foundation: Garner's Integral vs Separable Dimensions
Principle: Color hue + lightness interfere; position + shape work independently
Design Guideline: Be careful combining integral channels, use separable ones for multiple variables

Grouping - *How do we naturally organize/group visual information?*

Foundation: Gestalt Principles + Bertin's Visual Variables
Principle: Similar colors/shapes automatically create groups in our minds
Design Guideline: Use consistent encoding to show categories, vary it to show differences

Channels: Rankings

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)

Same

Same

Best

Effectiveness

Least

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

Shape

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

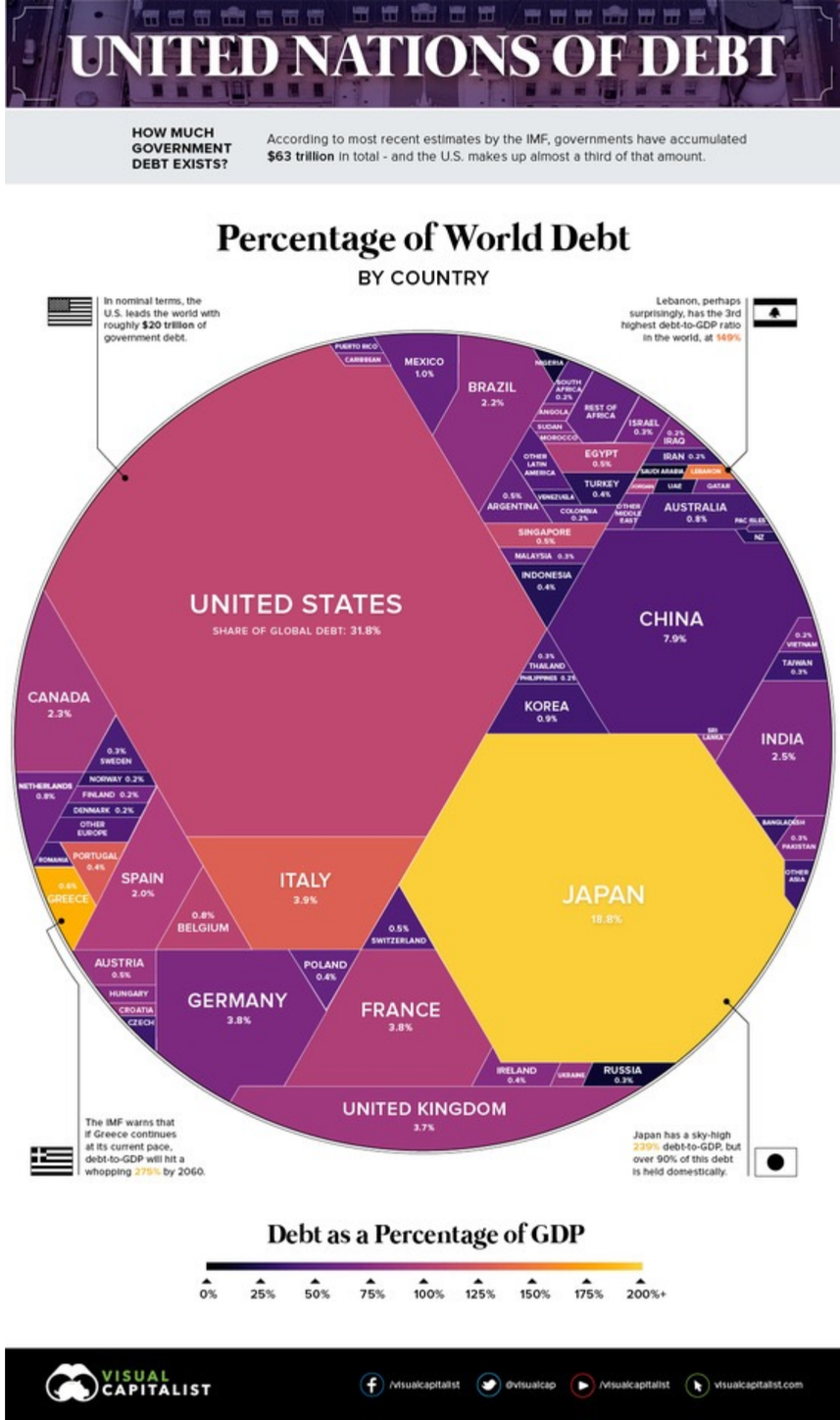
→ *Quantitative*

- expressiveness
  - match channel and data characteristics
    - magnitude for ordered
      - how much? which rank?
    - identity for categorical
      - what?

Table Talk

Deconstruct the viz and have a conversation about the following

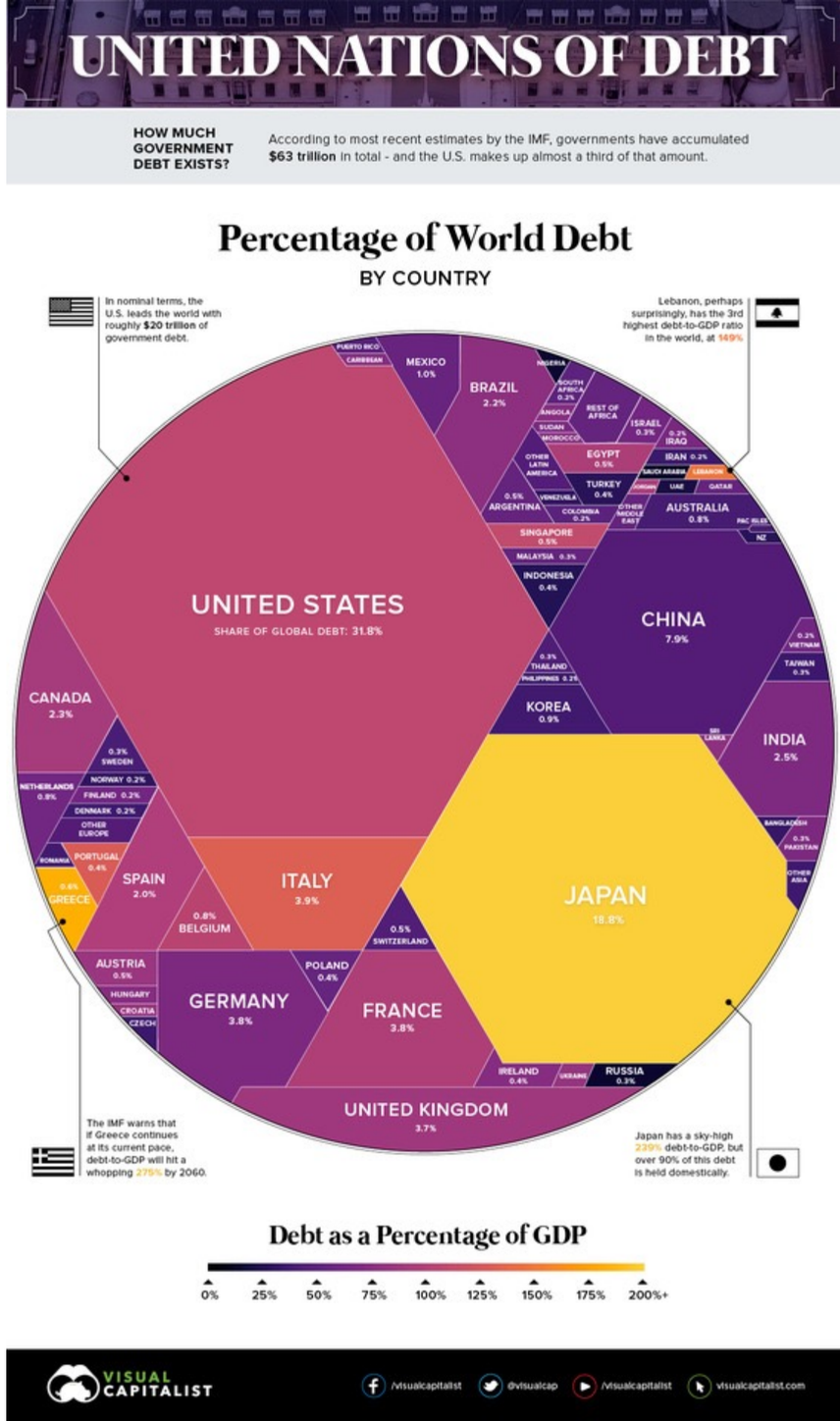
- Discriminability: how many unique steps can we perceive?
- Separability: is our ability to use this channel affected by another one?
- Popout: can things jump out using this channel?
- Grouping: can a channel show perceptual grouping of items?
- Accuracy: how precisely can we tell the difference between encoded items?



Clicker Question – Select ALL that apply

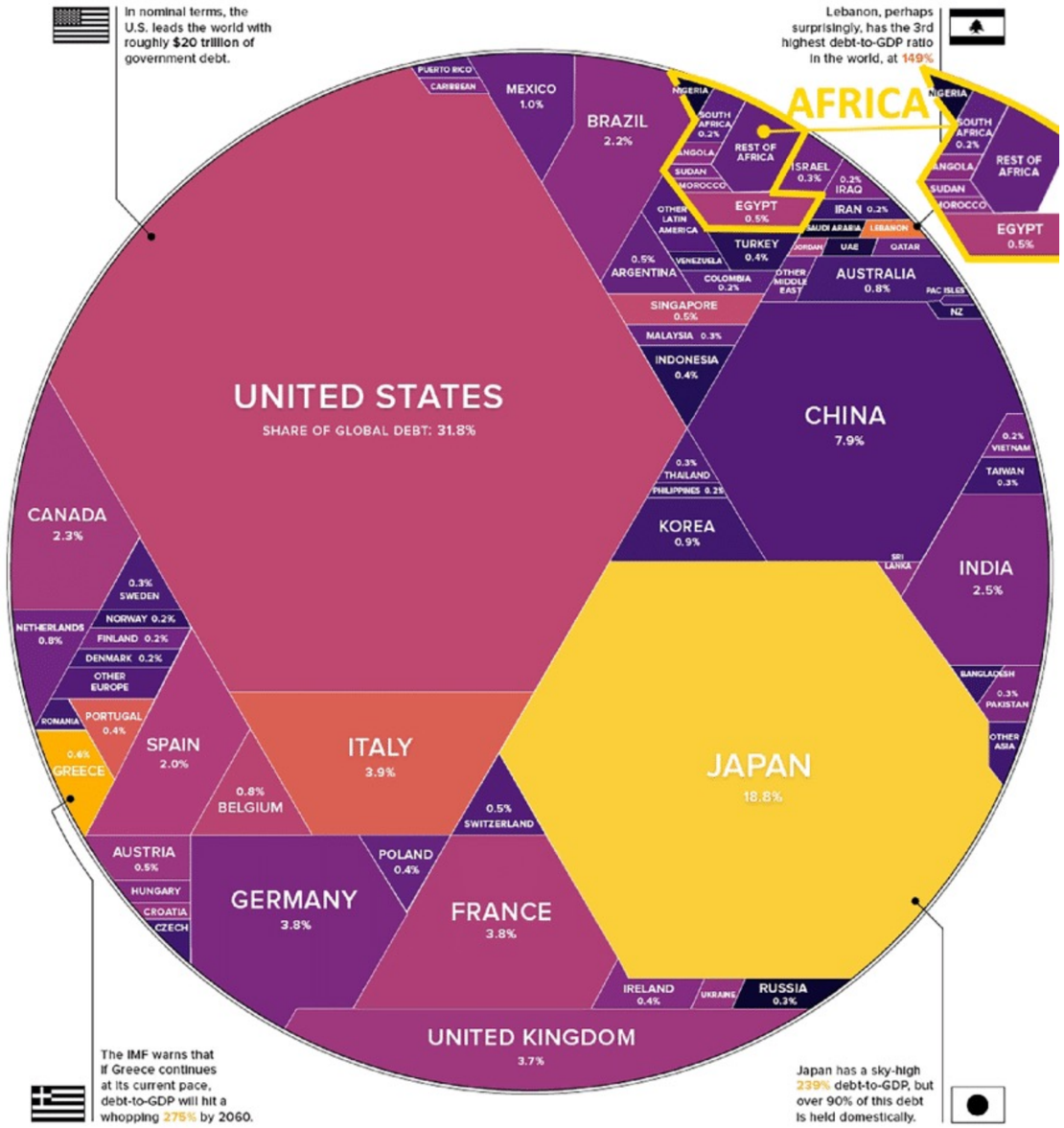
Looking at the viz, identify the encoding channels being used

A. Color Hue ✓
B. Area ✓
C. Length
D. Position ✓
E. Shape



Percentage of World Debt

BY COUNTRY



Debt as a Percentage of GDP

0% 25% 50% 75% 100% 125% 150% 175% 200%+

Clicker Question – Select ALL

This visualization uses area to show each country's percentage of global debt. What are the primary concern with this encoding choice?

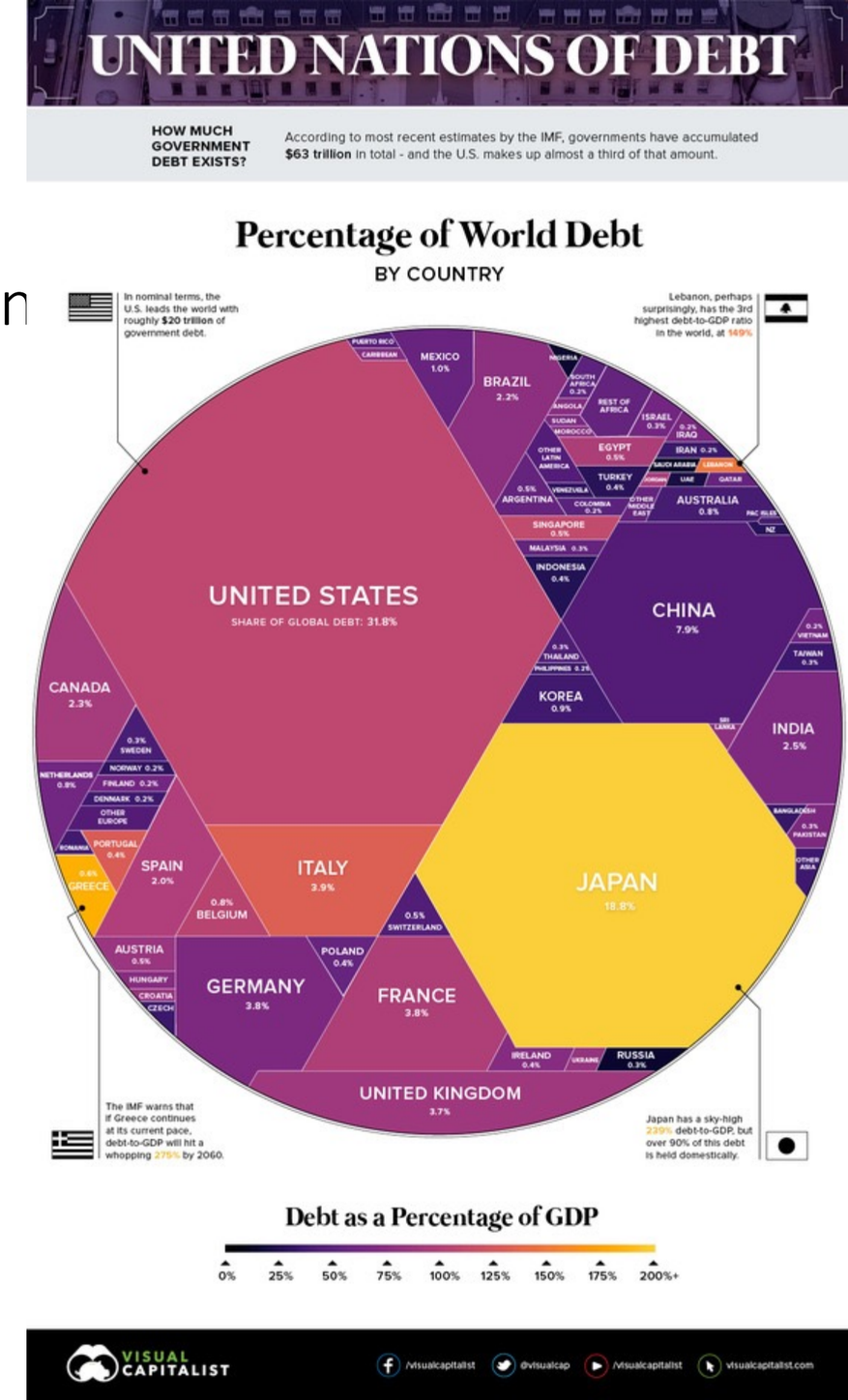
A. Poor discriminability

B. Underestimation of area differences

C. Users can't quickly identify outliers

D. Poor separability

E. Grouping is counterintuitive



Clicker Questions – Select ALL

If you wanted to improve the ACCURACY of comparing countries' share of global debt, which encoding change(s) would be most effective based on Cleveland & McGill's effectiveness hierarchy?

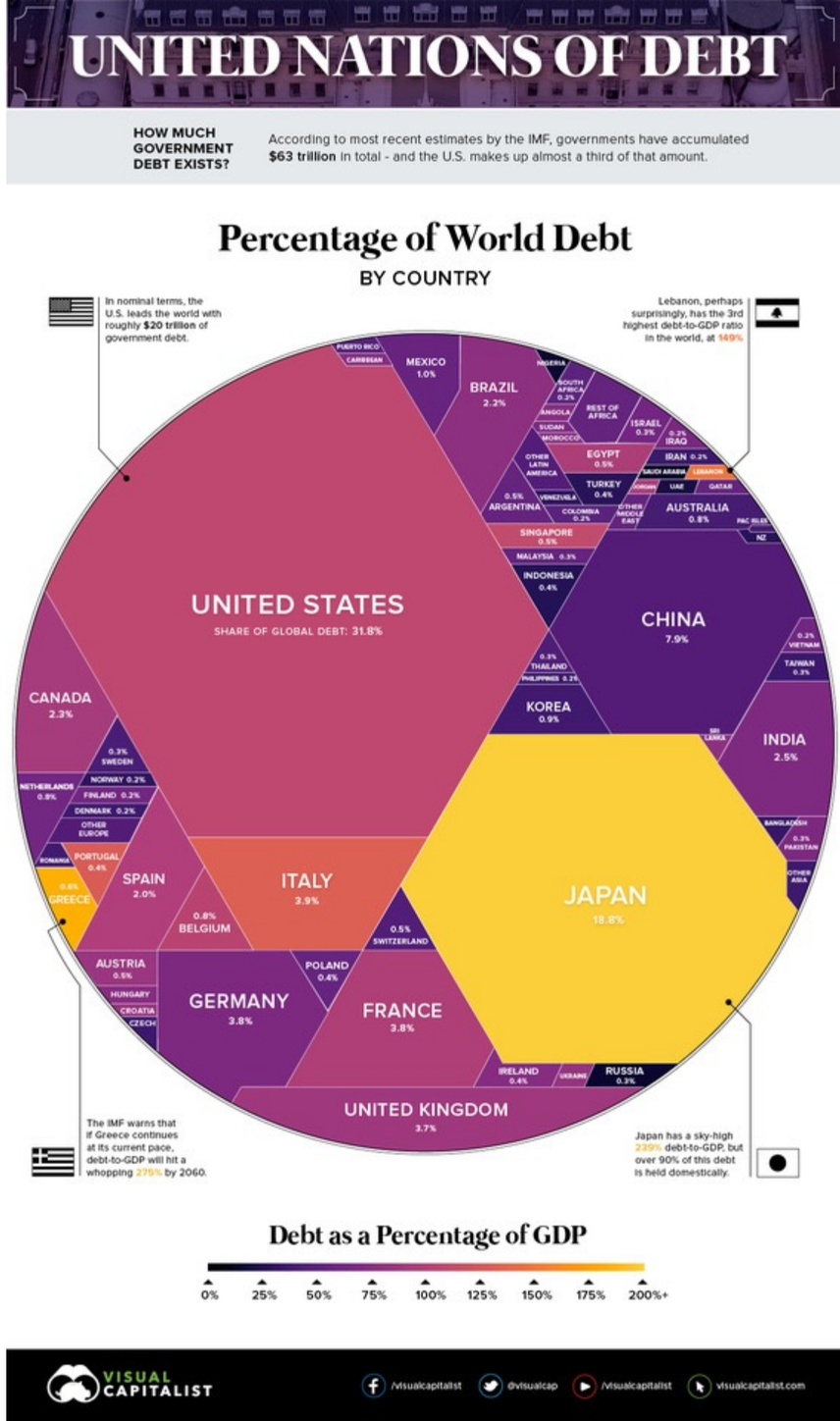
A. Switch from area encoding to length encoding for the debt percentages

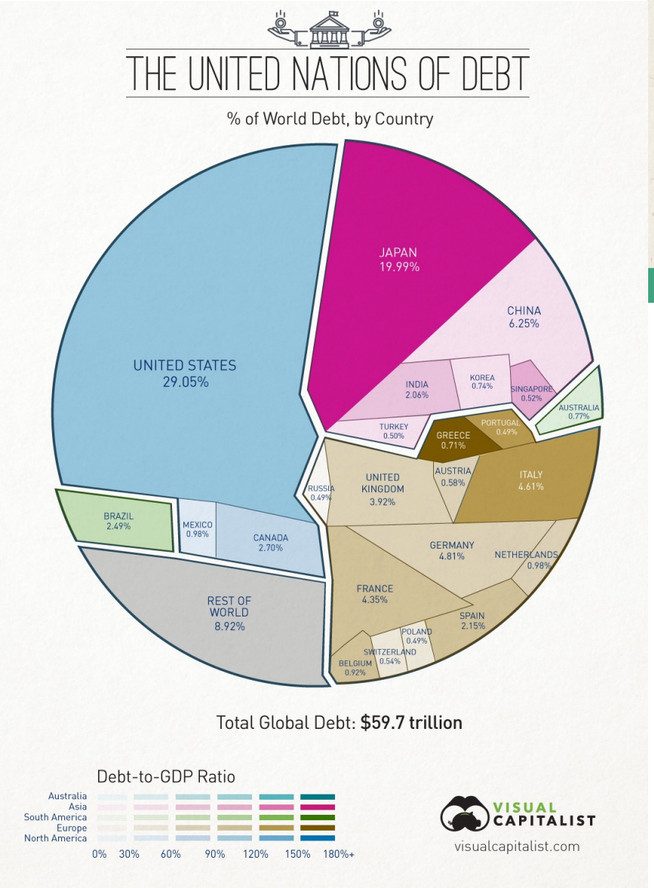
B. Switch from area encoding to position encoding along a shared scale

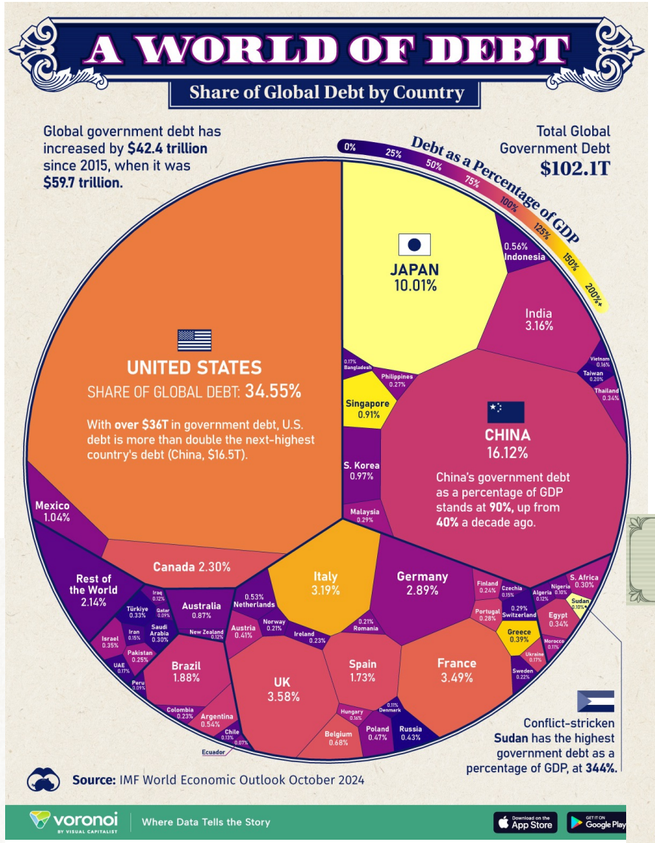
C. Keep area encoding but add color lightness to create redundant encoding

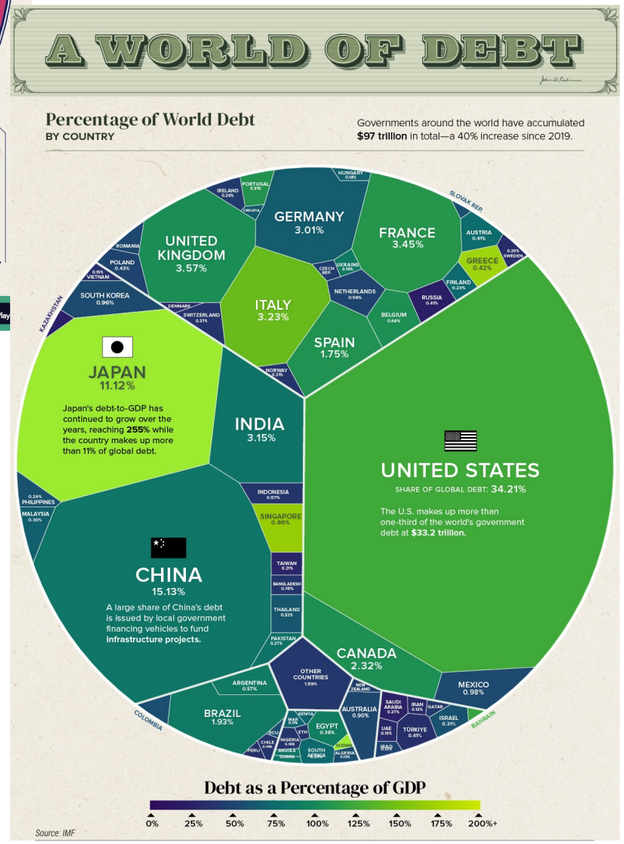
D. Keep area encoding but use rectangular shapes to reduce perceptual bias

E. Switch from area encoding to color hue encoding for categorical distinctions




THE UNITED NATIONS OF DEBT
% of World Debt, by Country
JAPAN 19.99%
CHINA 6.25%
UNITED STATES 29.05%
INDIA 2.06%
KOREA 0.74%
SINGAPORE 0.52%
AUSTRALIA 0.77%
TURKEY 0.50%
PORTUGAL 0.49%
GREECE 0.71%
UNITED KINGDOM 3.92%
AUSTRIA 0.58%
ITALY 4.61%
RUSSIA 0.49%
BRAZIL 2.49%
MEXICO 0.98%
CANADA 2.70%
GERMANY 4.81%
NETHERLANDS 0.98%
FRANCE 4.35%
SPAIN 2.15%
POLAND 0.49%
SWITZERLAND 0.54%
BELGIUM 0.92%
REST OF WORLD 8.92%
Total Global Debt: $59.7 trillion
Debt-to-GDP Ratio
Australia
Asia
South America
Europe
North America
0% 30% 60% 90% 120% 150% 180%+
VISUAL CAPITALIST
visualcapitalist.com


A WORLD OF DEBT
Share of Global Debt by Country
Global government debt has increased by $42.4 trillion since 2015, when it was $59.7 trillion.
Total Global Government Debt $102.1T
Debt as a Percentage of GDP
0% 25% 50% 75% 100% 125% 150% 200%+
UNITED STATES
SHARE OF GLOBAL DEBT: 34.55%
With over $36T in government debt, U.S. debt is more than double the next-highest country's debt (China, $16.5T).
JAPAN 10.01%
CHINA 16.12%
China's government debt as a percentage of GDP stands at 90%, up from 40% a decade ago.
Conflict-stricken Sudan has the highest government debt as a percentage of GDP, at 344%.
Source: IMF World Economic Outlook October 2024
voronoi BY VISUAL CAPITALIST | Where Data Tells the Story
Download on the App Store
GET IT ON Google Play


A WORLD OF DEBT
Percentage of World Debt
BY COUNTRY
Governments around the world have accumulated $97 trillion in total—a 40% increase since 2019.
JAPAN 11.12%
Japan's debt-to-GDP has continued to grow over the years, reaching 255% while the country makes up more than 11% of global debt.
UNITED STATES
SHARE OF GLOBAL DEBT: 34.21%
The U.S. makes up more than one-third of the world's government debt at $33.2 trillion.
CHINA 15.13%
A large share of China's debt is issued by local government financing vehicles to fund infrastructure projects.
Debt as a Percentage of GDP
0% 25% 50% 75% 100% 125% 150% 175% 200%+
Source: IMF
VISUAL CAPITALIST
/visualcapitalist @visualcap visualcapitalist.com
COLLABORATORS RESEARCH + WRITING Omri Wallach, Dorothy Neufeld | ART DIRECTION + DESIGN Joyce Ma

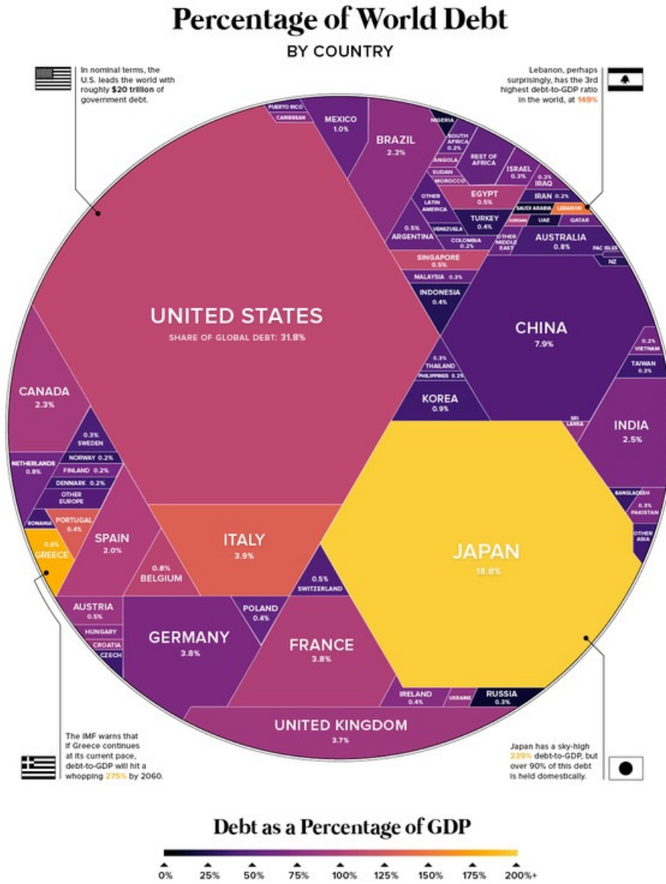

Percentage of World Debt
BY COUNTRY
In nominal terms, the U.S. leads the world with roughly $20 trillion of government debt.
Lebanon, perhaps surprisingly, has the 3rd highest debt-to-GDP ratio in the world, at 149%
UNITED STATES
SHARE OF GLOBAL DEBT: 31.8%
CHINA 7.9%
JAPAN 18.8%
GERMANY 3.8%
FRANCE 3.8%
UNITED KINGDOM 3.7%
ITALY 3.9%
SPAIN 2.0%
CANADA 2.3%
INDIA 2.5%
BRAZIL 2.2%
MEXICO 1.0%
AUSTRALIA 0.8%
KOREA 0.9%
The IMF warns that if Greece continues at its current pace, debt-to-GDP will hit a whopping 275% by 2060.
Japan has a sky-high 239% debt-to-GDP, but over 90% of this debt is held domestically.
Debt as a Percentage of GDP
0% 25% 50% 75% 100% 125% 150% 175% 200%+

Table Talk

Deconstruct the viz and have a conversation about the following

- Discriminability: how many unique steps can we perceive?
- Separability: is our ability to use this channel affected by another one?
- Popout: can things jump out using this channel?
- Grouping: can a channel show perceptual grouping of items?
- Accuracy: how precisely can we tell the difference between encoded items?

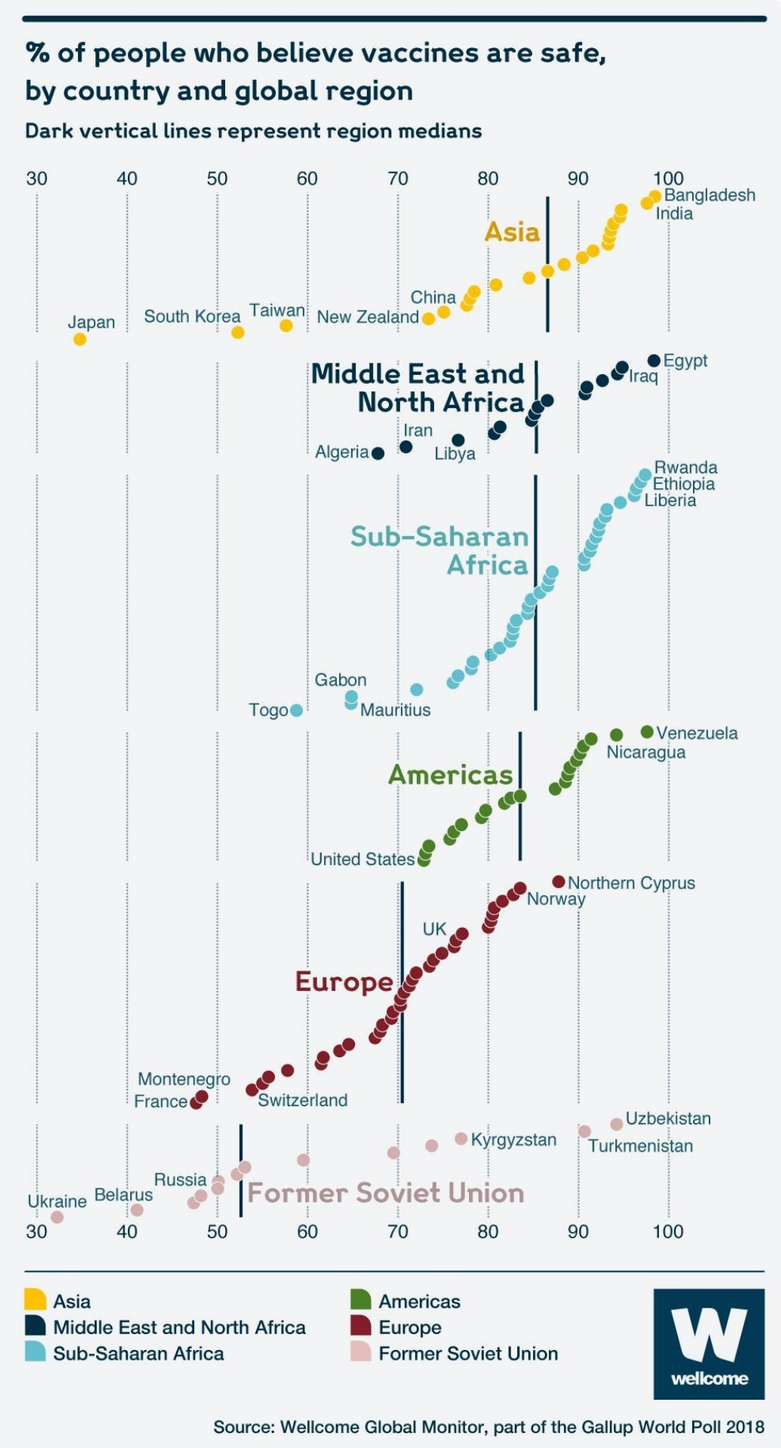


Table Talk

Looking at the viz, identify the encoding channels being used

What does length of the line represent?

What does color hue represent?

What does x position represent?

What does y position represent?

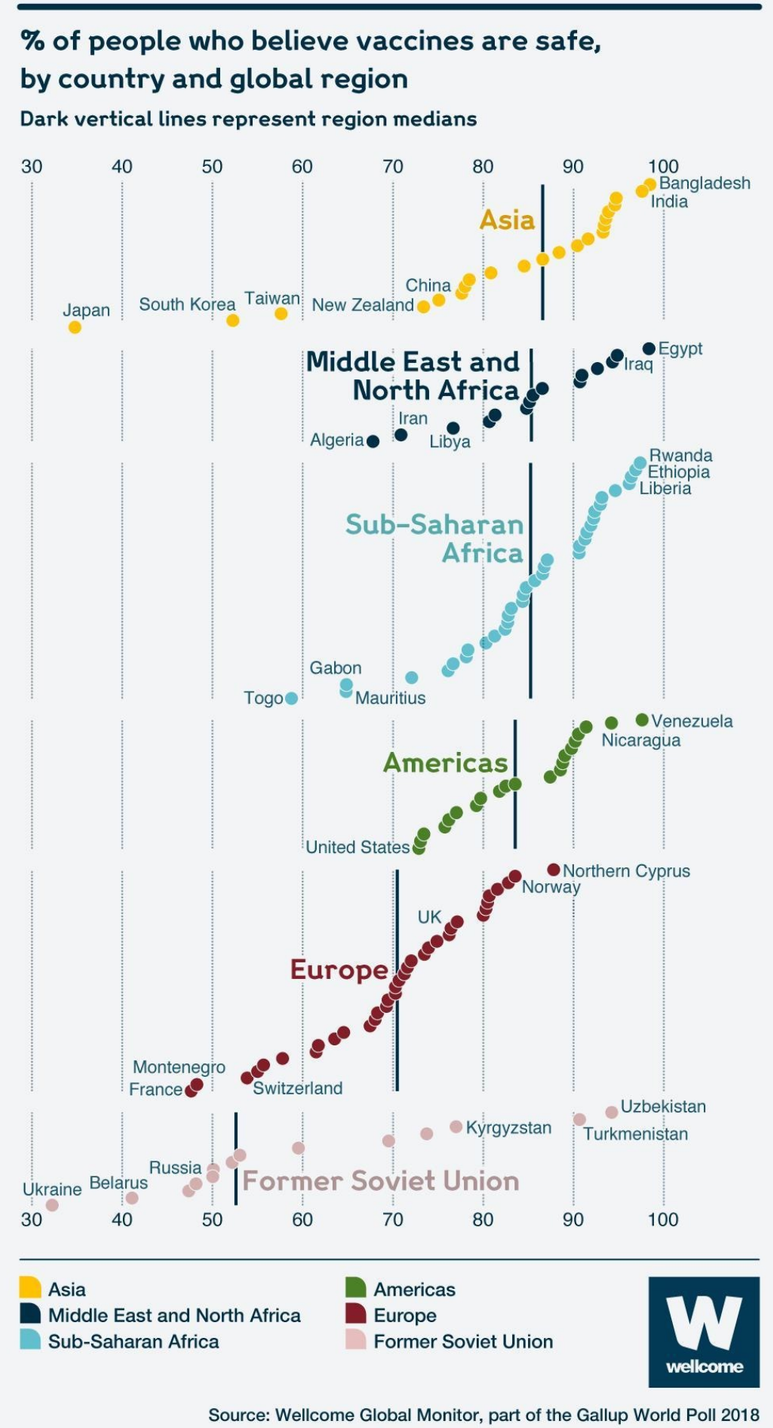
A. % of people who believe vaccines are safe

B. Regions of the world

C. Regional medians % of people who believe vaccines are safe

D. Nothing

E. Extent of people who believe vaccines are safe across the region



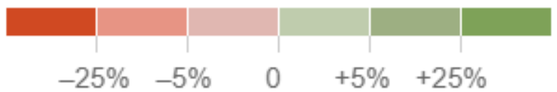
All Spending | Types of Spending | Changes | Department Totals

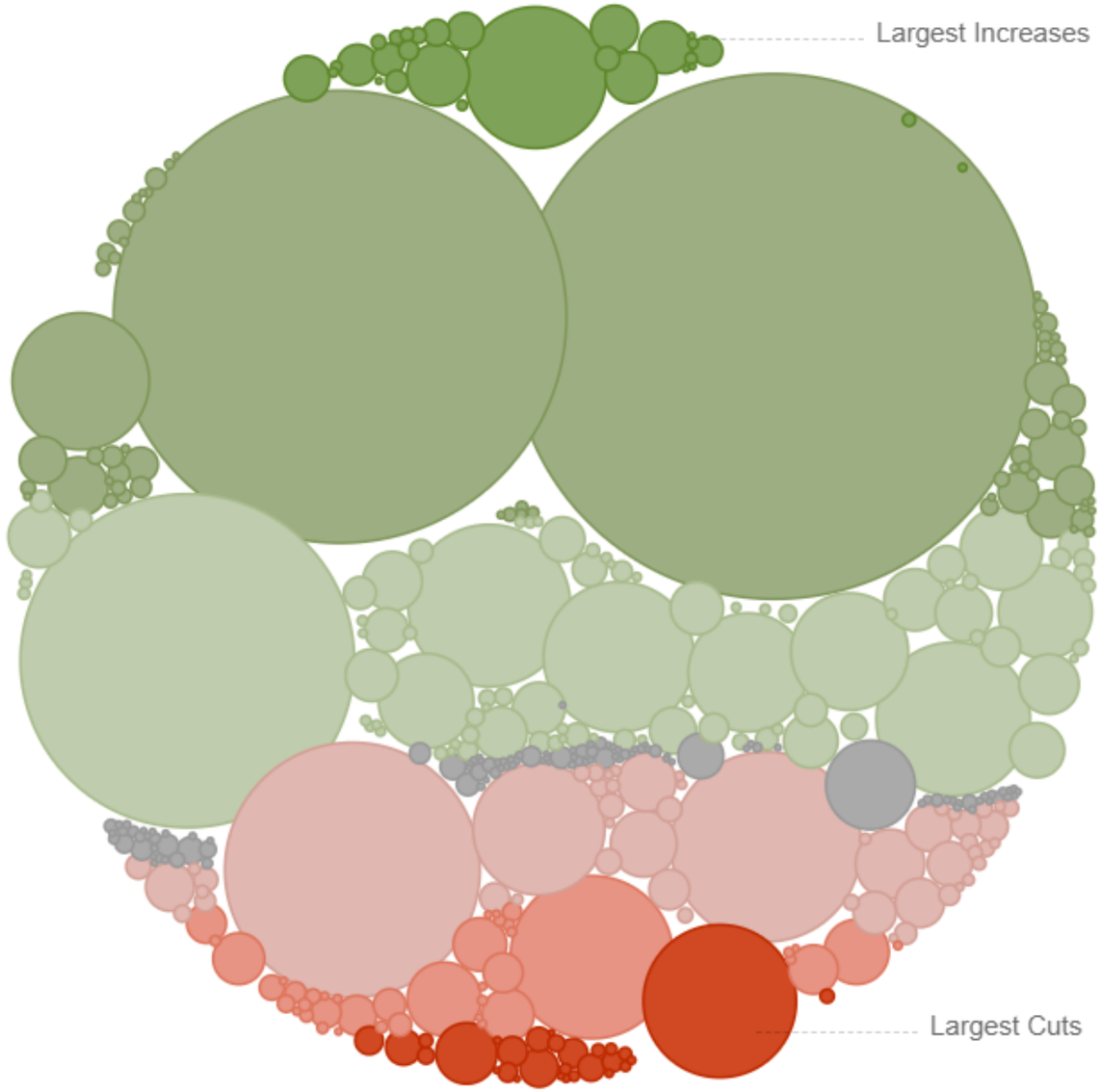
How $3.7 Trillion Is Spent

Mr. Obama's budget proposal includes $3.7 trillion in spending in 2013, and forecasts a $901 billion deficit.

Circles are sized according to the proposed spending.

Color shows amount of cut or increase from 2012.



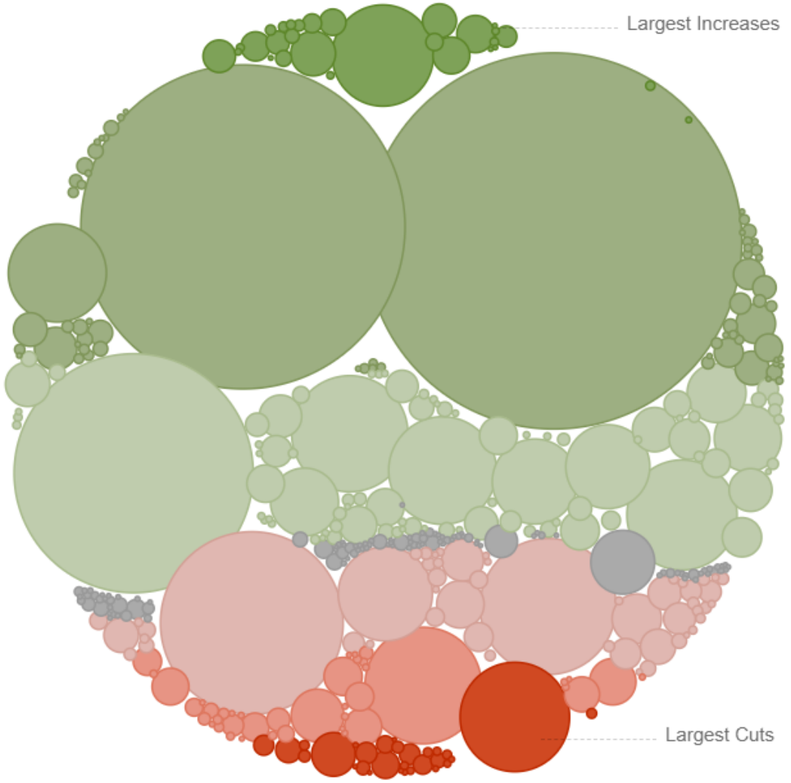


https://archive.nytimes.com/www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html

Table Talk

Deconstruct the viz and have a conversation about the following

- Discriminability: how many unique steps can we perceive?
- Separability: is our ability to use this channel affected by another one?
- Popout: can things jump out using this channel?
- Grouping: can a channel show perceptual grouping of items?
- Accuracy: how precisely can we tell the difference between encoded items?



Clicker Question - Select ALL

Looking at the viz, identify the main encoding channels being used

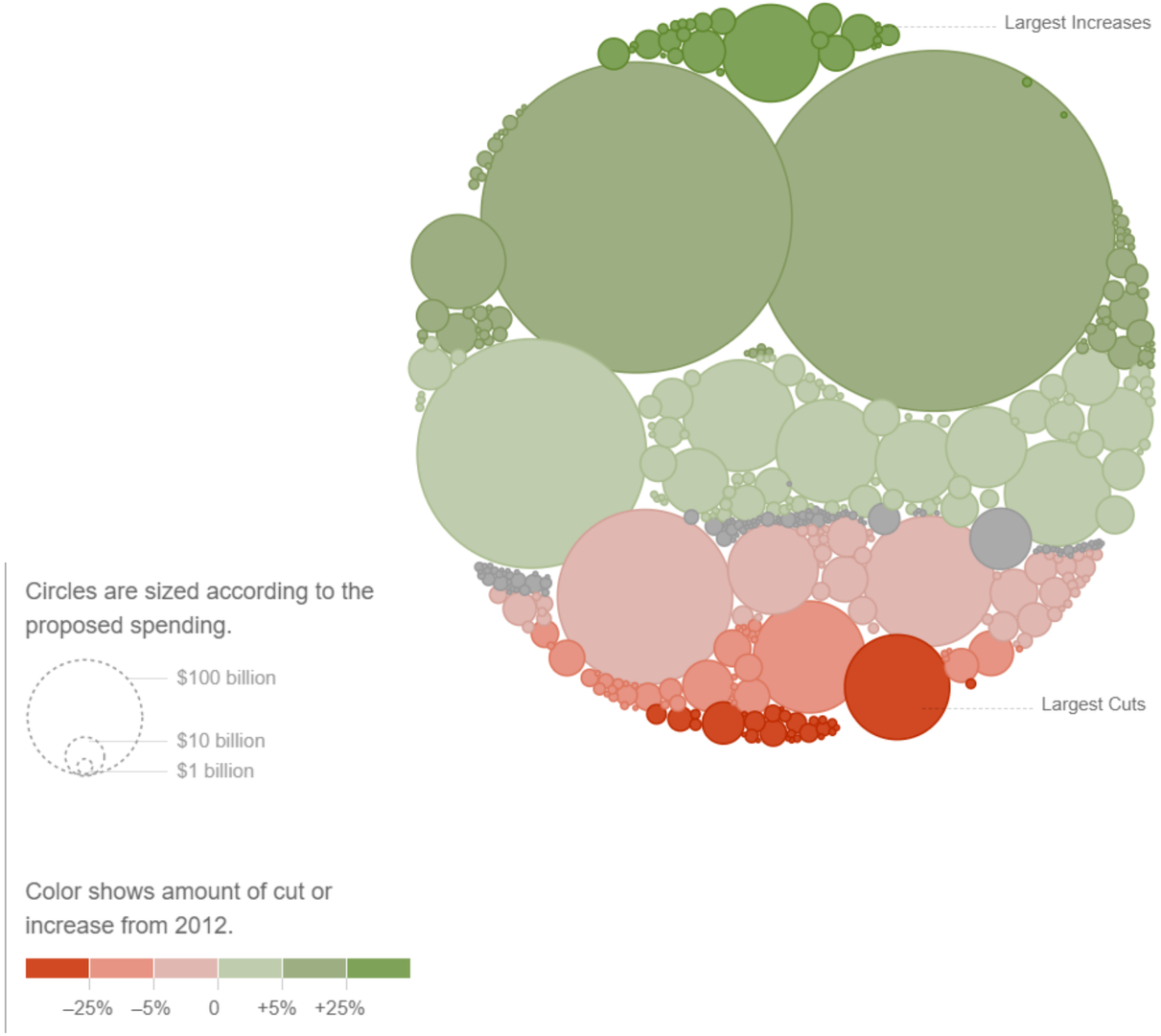
A. Area

B. Color hue

C. Color lightness

D. Position

E. Stroke width



Clicker Question

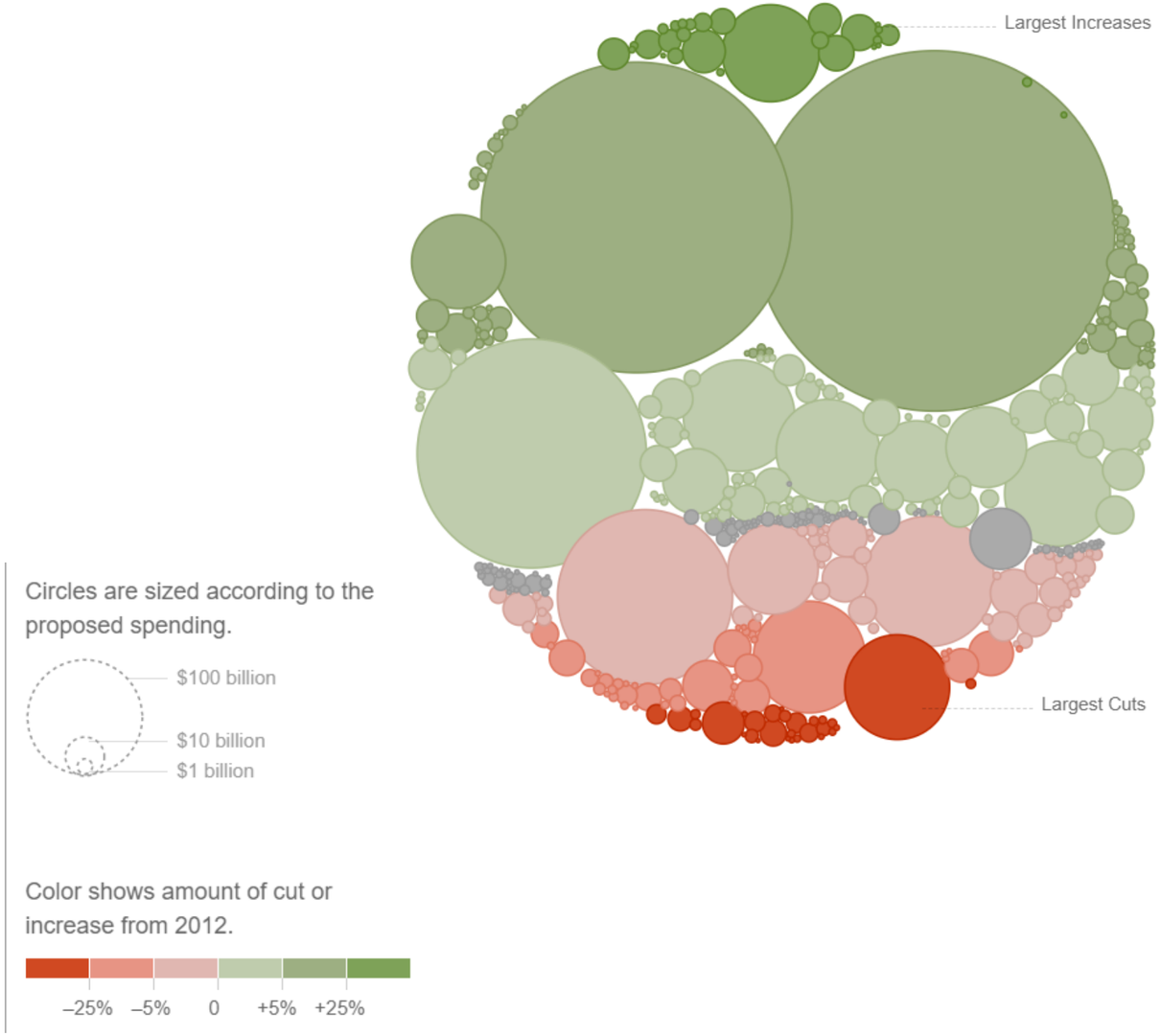
What kind of attribute is encoded on y channel

A. Quantitative

B. Nominal

C. Ordinal

D. Temporal



Clicker Question - Select ALL

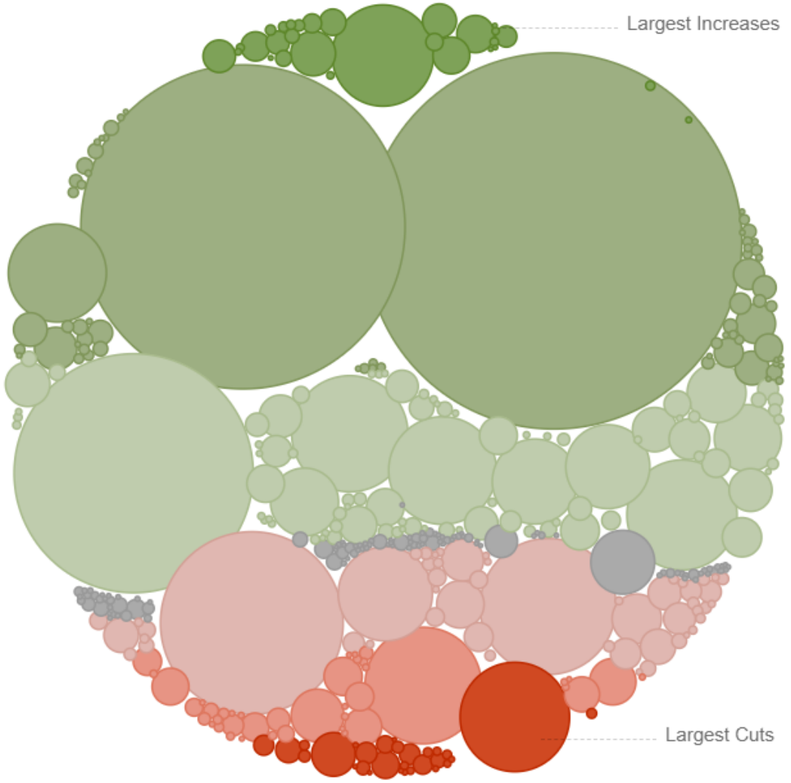
The vis uses AREA (bubble size) to represent the magnitude of budget changes. What are the two main perceptual problems with this encoding choice?

A. Stevens' Power Law bias (underestimation) + limited discriminability for comparing similar-sized bubbles

B. No natural ordering + circular confusion with bubble shapes

C. Poor pop-out effects + integral relationship with color channels

D. Weak grouping properties + accessibility issues for color-blind users



Clicker Question - Select ALL

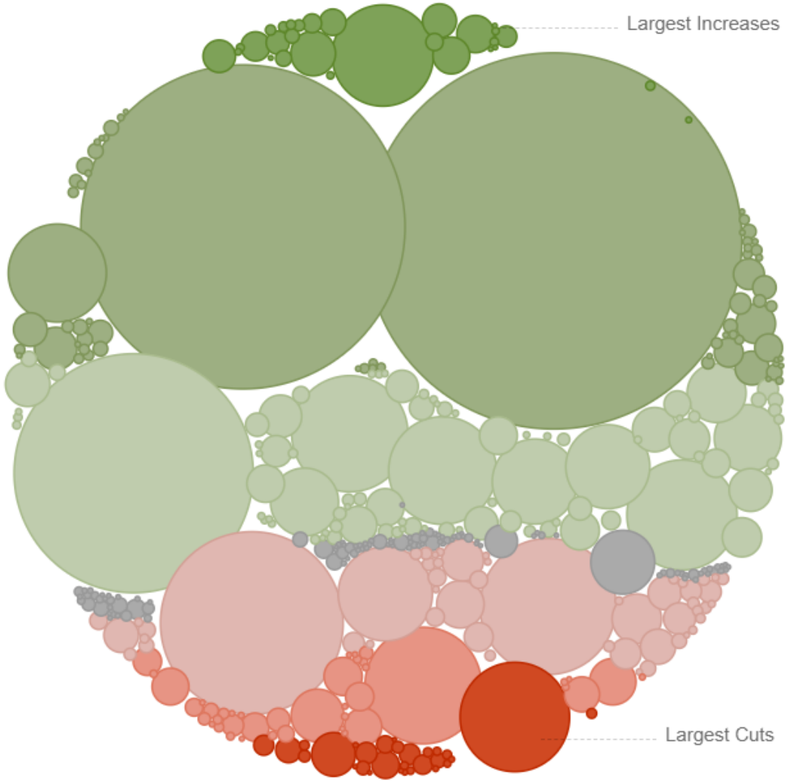
This design uses BOTH color (green/red/grey) and spatial position (top/bottom) to show increases vs cuts. Evaluate this redundant encoding strategy:

A. Unnecessary - creates visual clutter without adding information

B. Effective - provides accessibility and reinforces the categorical distinction

C. Problematic - color and position are integral dimensions that interfere

D. Confusing - creates competing organizational systems

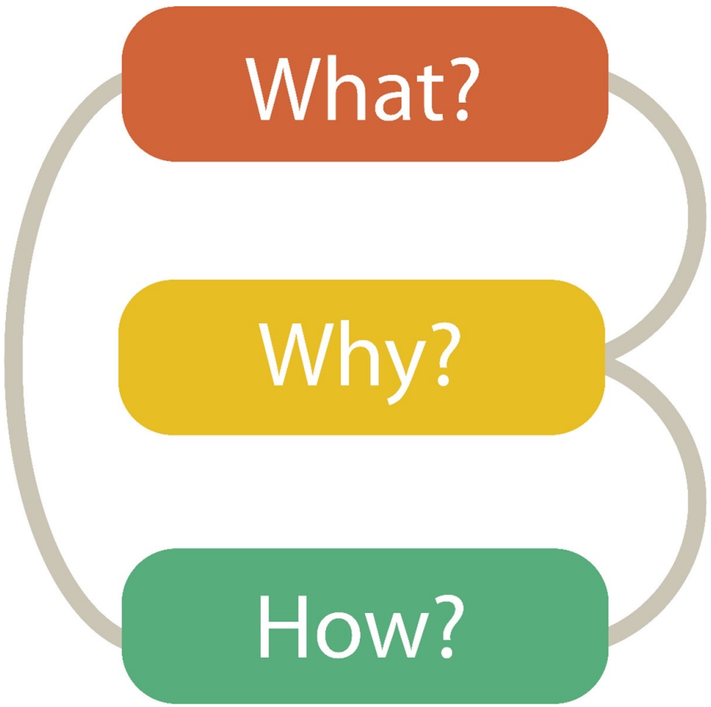


Questions We Still Need to Answer

We’ve only scratched the surface — there’s more to explore in the coming weeks:

- ~~Which channels are best suited for which types of data attributes?~~
- How can we combine multiple channels effectively without causing conflicts?
- **Which channels are most effective for specific analytical tasks?**

Data Visualization Ecosystem



What data is shown?

DATA ABSTRACTION

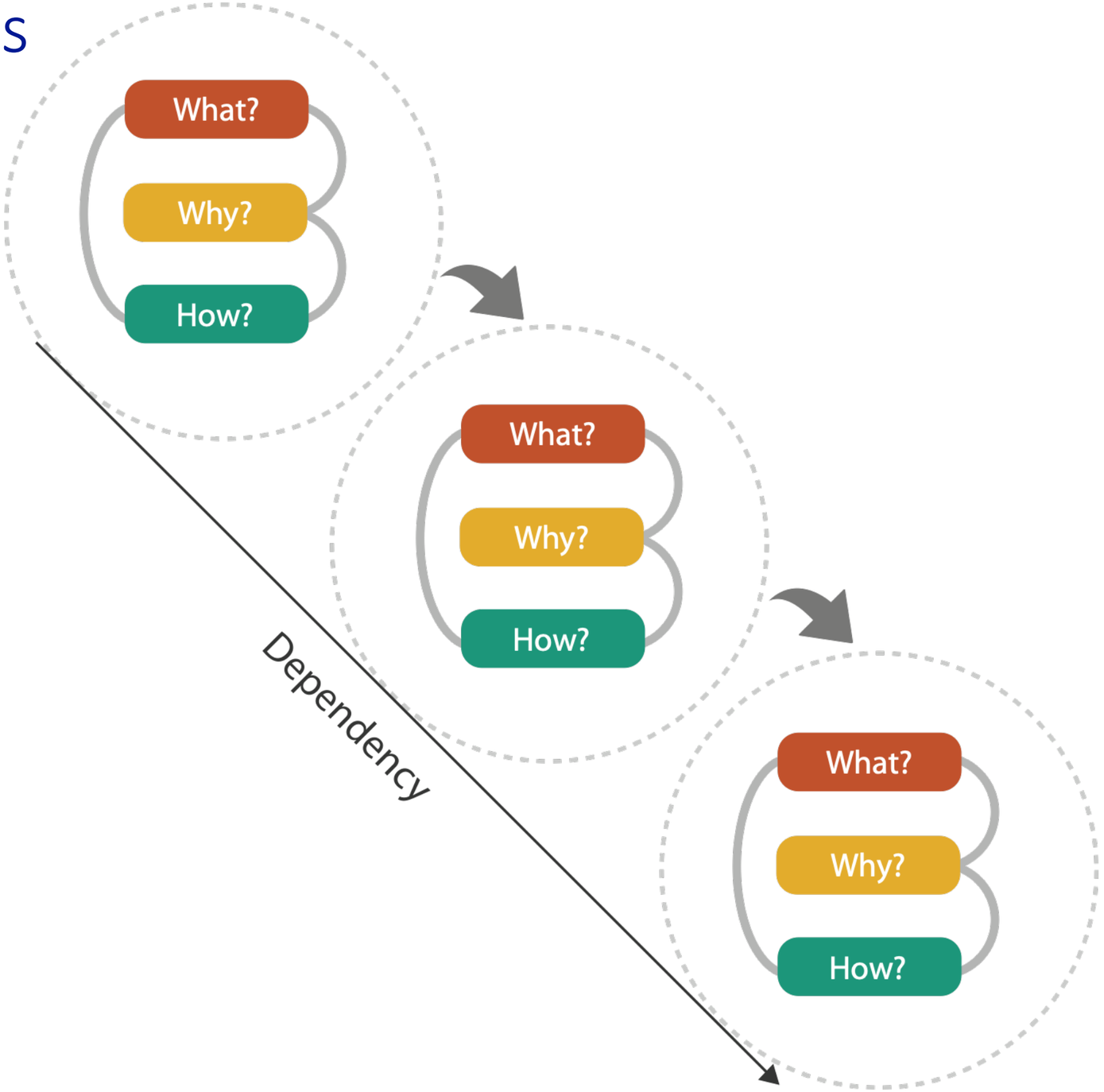
Why is the user analyzing / viewing it?

TASK ABSTRACTION

How is the data presented?

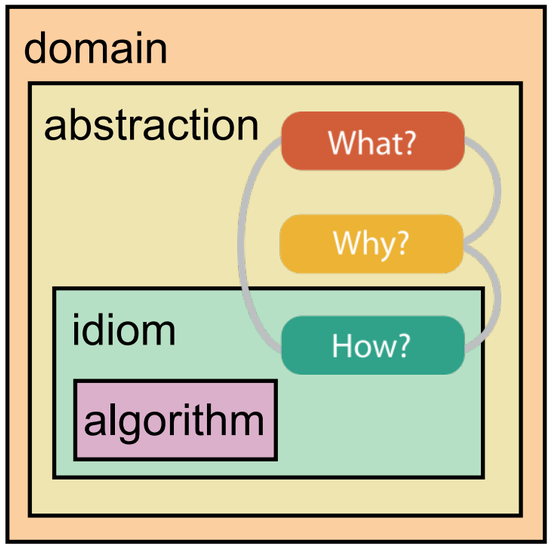
VISUAL AND INTERACTION ENCODING

Means and ends



Analysis framework: Four levels, three questions

- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - what is shown? **data** abstraction
    - why is the user looking at it? **task** abstraction
- *idiom*
  - how is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm*
  - efficient computation

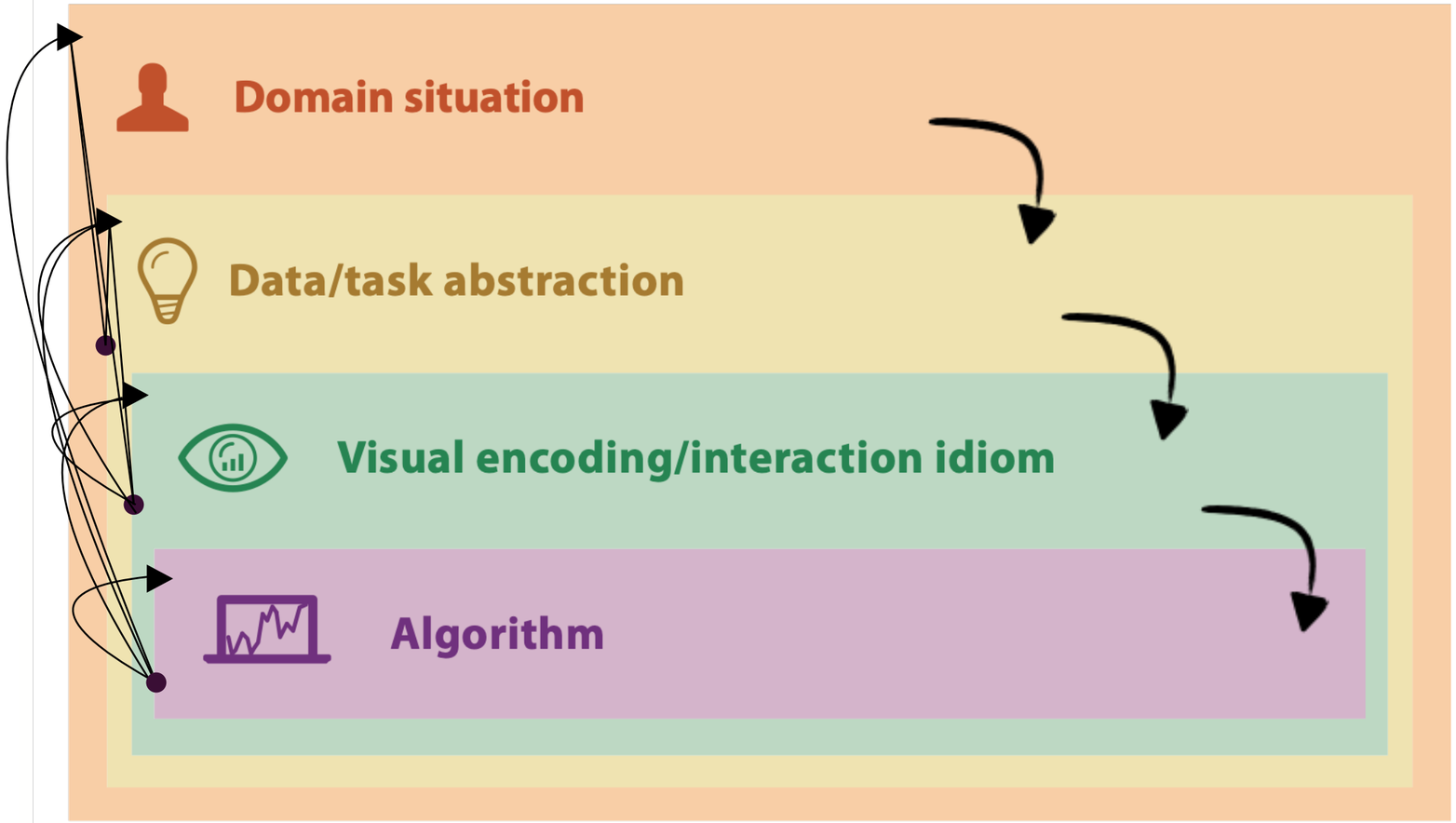


[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Nested model

- downstream: cascading effects
- upstream: iterative refinement



[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Whose Task: Designer vs. User

From Tasks to Tools – A continuum

- Specific tools
  - Narrow use: typically designed for a specific context
  - Designer has customized the tool in such a way that the user cannot change it
  - Limited design choices
  - High stakes
- General tools
  - Users have a lot of power

Task-Driven Visualization Design

- What is the overarching purpose of your visualization?
- What questions do you want to answer?
- What insights should the user be able to reach?
- What tasks should the user be able to accomplish?
  - More on this when we talk about Interaction

Task Abstraction – First Look

What's in there?

Overview and Individual Values – looking at all the data through specific attributes

- Tasks: Retrieve Value, determine range, characterize the distribution

– Vizzes: Histograms, density plots, box plots, overview displays

Where are the interesting cases?

Finding and Filtering – searching and locating actions

– Tasks: Filter, sort, find anomalies

– Vizzes: bubble plots, sorted bars, interactive views

How do things relate?

Relationships and patterns

– Tasks: correlation, clustering, computing derived values

– Vizzes: Scatter plot, correlation matrices, grouped/faceted displays

September 30th: National Day for Truth and Reconciliation

Administrivia – Office Hours (THIS WEEK ONLY)

TA Office Hours (online)

Monday 11:00am – 12pm

Wednesday 10:00 – 11am

Pop up in-person office hours Thursday 3 – 6 in LAB ROOM.

Instructor Office Hours

Monday and Wednesday 5 – 6pm (currently in this room)

~~Tuesdays 2 – 3pm ICCS 227~~

Participation is Clickers, not PL (unless otherwise stated)

Get Stepping

- Office Hours this week are different – see earlier slides
- Quiz this week is Wednesday – Thursday
  - This Pandas Cheet Sheet from Pydata has been added as a resource for the quiz
  - https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf
  - On course website – there are some programming exercises,
  - Run your code, no red errors or that is a 0
- General Notes
  - No late work accepted, you are nearing graduation

Activity – Motivating EDA

(Submit on PrairieLearn)

Cholera in 1854 London

Cholera is an infectious disease that affects the small intestine.
In 1854, a deadly cholera outbreak swept through the Soho district in London. In the first week alone, more than 150 people died. It's 1854 (no computers!) and the government has hired your team to investigate the cause of the outbreak.

Your Task (work in small groups, 15 minutes):

- **Form a Hypothesis:** What do you suspect might be causing the outbreak?
- **Collect Evidence:** What kinds of information would you want to gather from the community to test your suspicion? (Think about *people, places, and timing*.)
- **Show the Evidence**
  - How could you present this information so patterns become clear to others?
  - What would your specific visualization be, describe the mark, channels used,
  - If you have time, go a step further and visualize it

Clicker Question – ONE WORD

Hypothesis:

In one word or phrase, what do you think caused the outbreak?

Clicker Question – Select ALL

EVIDENCE:

Which types of information would be useful to collect to investigate the cholera outbreak?

- A. Home addresses of cholera victims
- B. Dates when symptoms began
- C. Number of people in each household
- D. Locations of nearby water pumps
- E. Reports of deaths from other neighborhoods

Clicker Question – Select ONE

VISUALIZATION

Which visualization would make it easiest to spot the source of the outbreak?

A. A line graph showing deaths over time
B. A pie chart of total deaths by gender
C. A map showing deaths by household location
D. A bar chart of deaths by day of the week
E. A scatter plot showing the mortality rate relative to size of household

Learning Outcomes

- Describe what is exploratory data analysis and why it is important
- Explore a dataset by visualizing various one-dimensional and 2D distributions across multiple categories.
- Inspect and describe relationships between variables.
- Detect missing values and suspicious observations in a dataset.

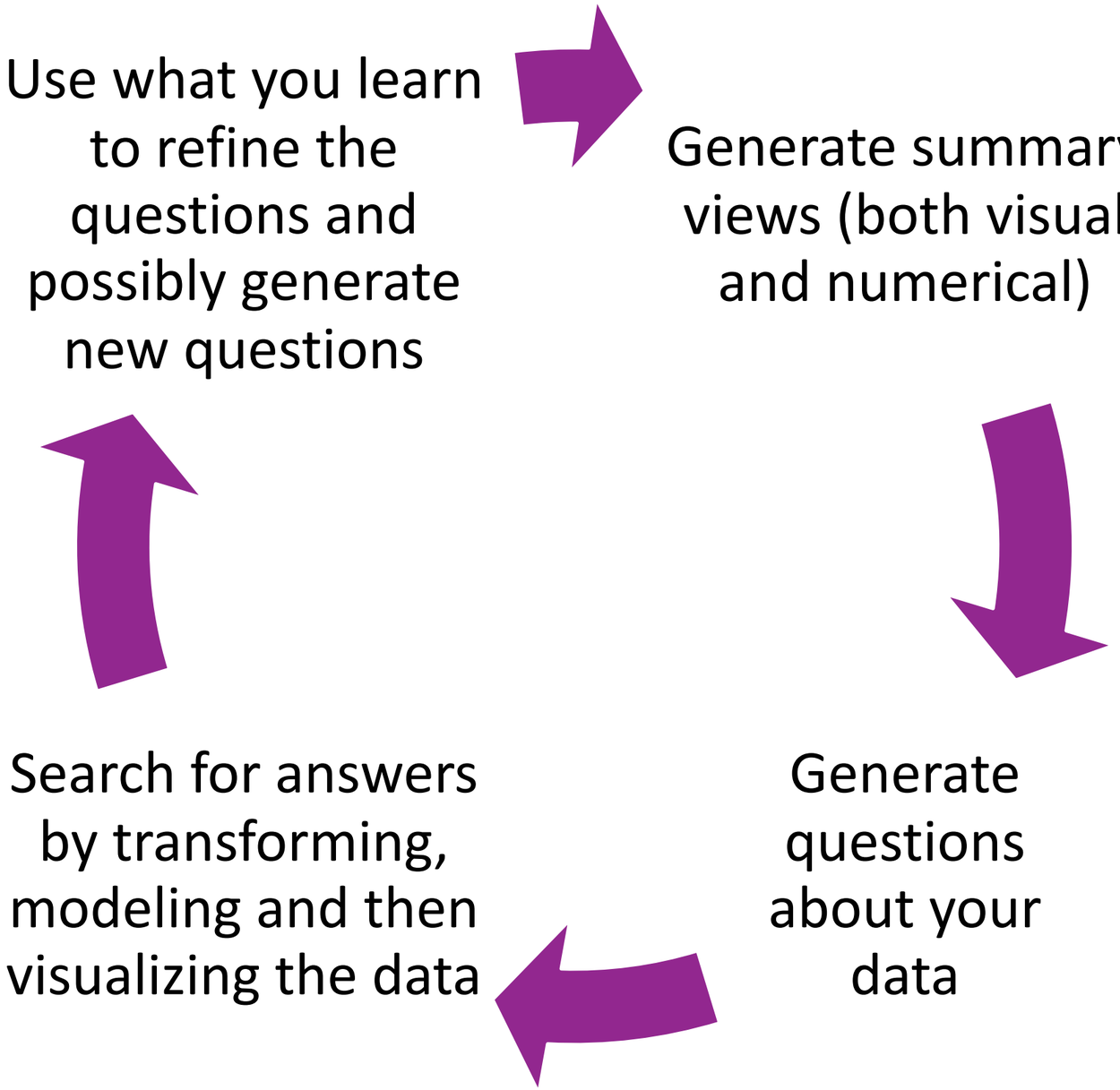
Exploratory Data Analysis (EDA)

Is a process that includes

- detection of mistakes
- checking of assumptions
- preliminary selection of appropriate models
- determining relationships among the explanatory variables, and
- assessing the direction and rough size of relationships between explanatory and outcome variables.

Experimental Design and Analysis. Howard J. Seltman

Exploratory Data Analysis Iterative Cycle



https://r4ds.had.co.nz/exploratory-data-analysis.html

Making the case for EDA – John Tukey (Father of EDA)

- Between 1900 – 1970s the status quo was
  - Formal theories of statistics
  - Rapid advancement in computers (e.g. Moore's law, applications, storage)
  - Data collection increases (based on measurements and growth of the scholar class)
  - Quantification vs. qualitative
- "Exposure" is critical
  - Numbers are great, formal processes are good
  - But the **flexibility** of an **informed human mind** is **irreplaceable**"
- We need strategies and techniques that
  - Facilitate the exploration of data by humans
  - Externalize representations of data
  - Support manipulation of data models

Exploratory Data Analysis (EDA)

10.1 Introduction

This chapter will show you how to use visualization and transformation to explore your data in a systematic way, a task that statisticians call exploratory data analysis, or EDA for short. EDA is an iterative cycle. You:

1. Generate questions about your data.

2. Search for answers by visualizing, transforming, and modelling your data.

3. Use what you learn to refine your questions and/or generate new questions.

EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends. As your exploration continues, you will home in on a few particularly productive insights that you'll eventually write up and communicate to others.

EDA is an important part of any data analysis, even if the primary research questions are handed to you on a platter, because you always need to investigate the quality of your data. Data cleaning is just one application of EDA: you ask questions about whether your data meets your expectations or not. To do data cleaning, you'll need to deploy all the tools of EDA: visualization, transformation, and modelling.

R for Data Science by Hadley Wickham, Mine Cetinkaya-Rundel, Garrett Grolemund

Categorizing Exploratory Data Analysis

Medium

- Numerical Summaries
- Visual Data Analysis

Attribute

- Univariate - one column at a time
- Multivariate – two or more variables at a time, looking for relationships.
- Attribute Type – categorical or quantitative

Role

- Outcome
- Explanatory

Categories of EDA

- Univariate Numerical Summaries
- Univariate Visual Idioms
- Multivariate Numerical Summaries
- Multivariate Visual Idioms

Read through the next **twelve** slides. Make sure you understand the different visualizations and when they will be used.

Use the references (links in the lower-right corner) to aid your understanding.

We will NOT spend time in class describing each standard visualization so please do this work.

Univariate Numerical Summaries

- Categorical Variable
  - Range of values
  - Frequency of each value (proportion)
- Quantitative Variable
  - Distribution: center, spread, modality, shape and outliers
  - Central Tendency: mean, mode, median
  - Spread: variance, standard deviations, interquartile range

Tukey advocates for focusing on max, min, median and quartiles

Univariate Visual Idioms

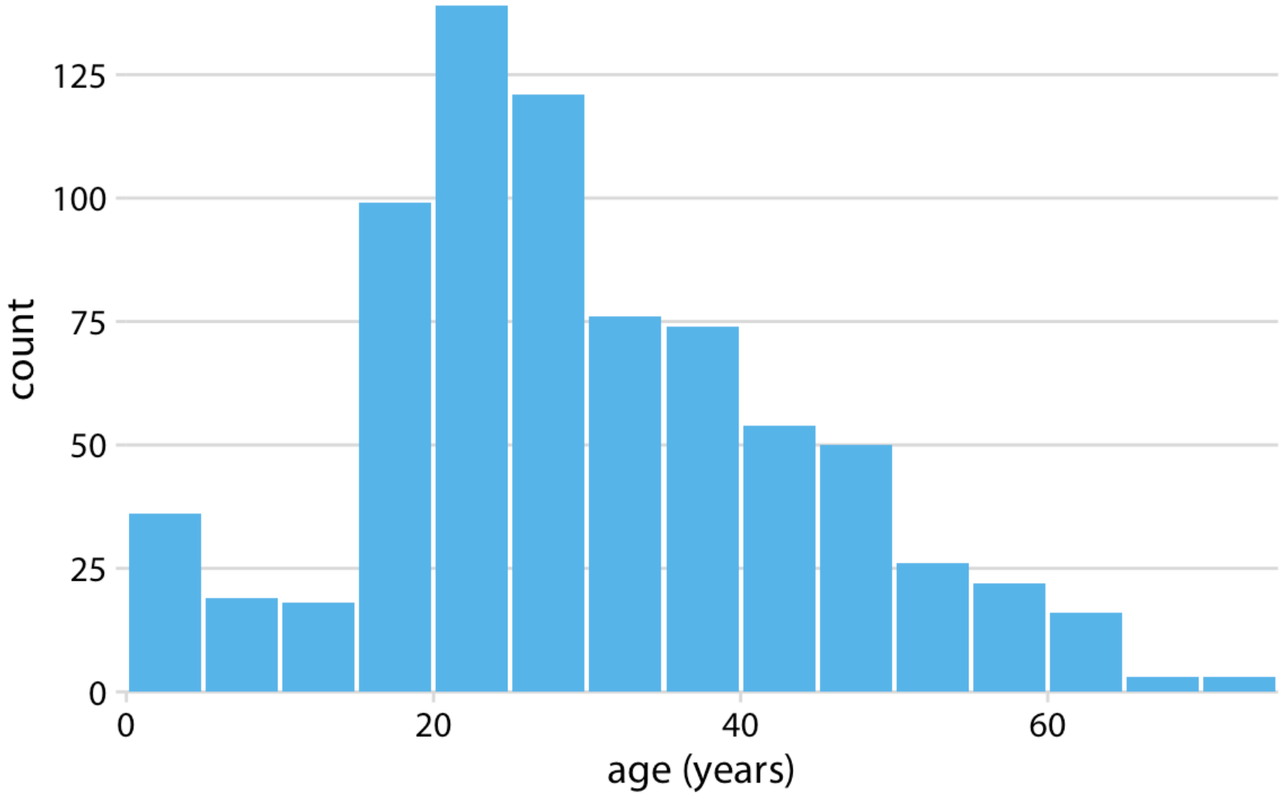
- Distribution of Categorical Variable
  - Bar Charts
- Distribution of Quantitative Variable
  - Histogram
  - Density Plots

Univariate Visual Idiom: Histograms

Histograms are used to visualize the shape, center, range and variation for a continuous variable.

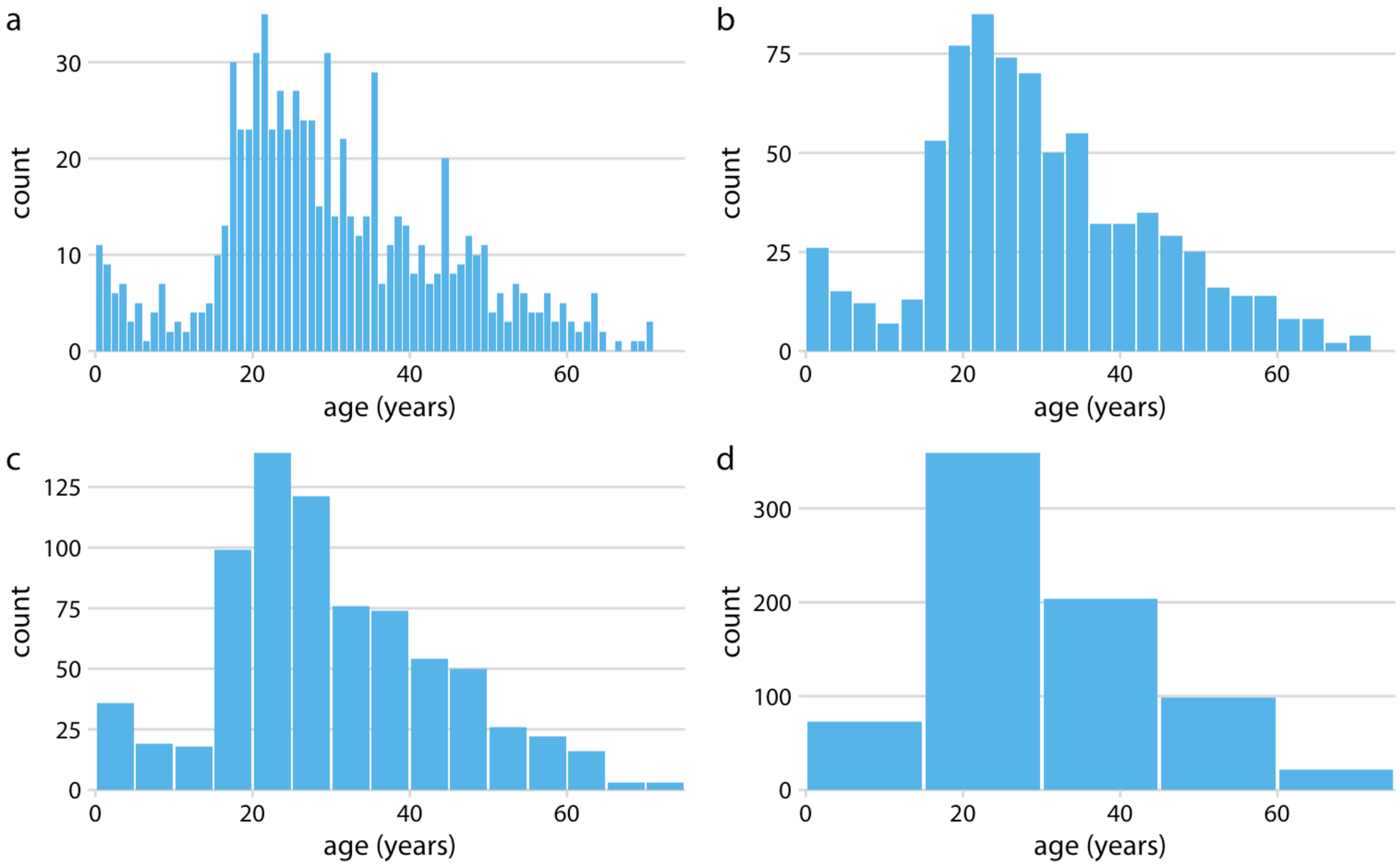
The size of the bin is extremely important.

Design Tip: Vary bin size during EDA



https://clauswilke.com/dataviz/histograms-density-plots.html

Univariate Visual Idiom: Histograms

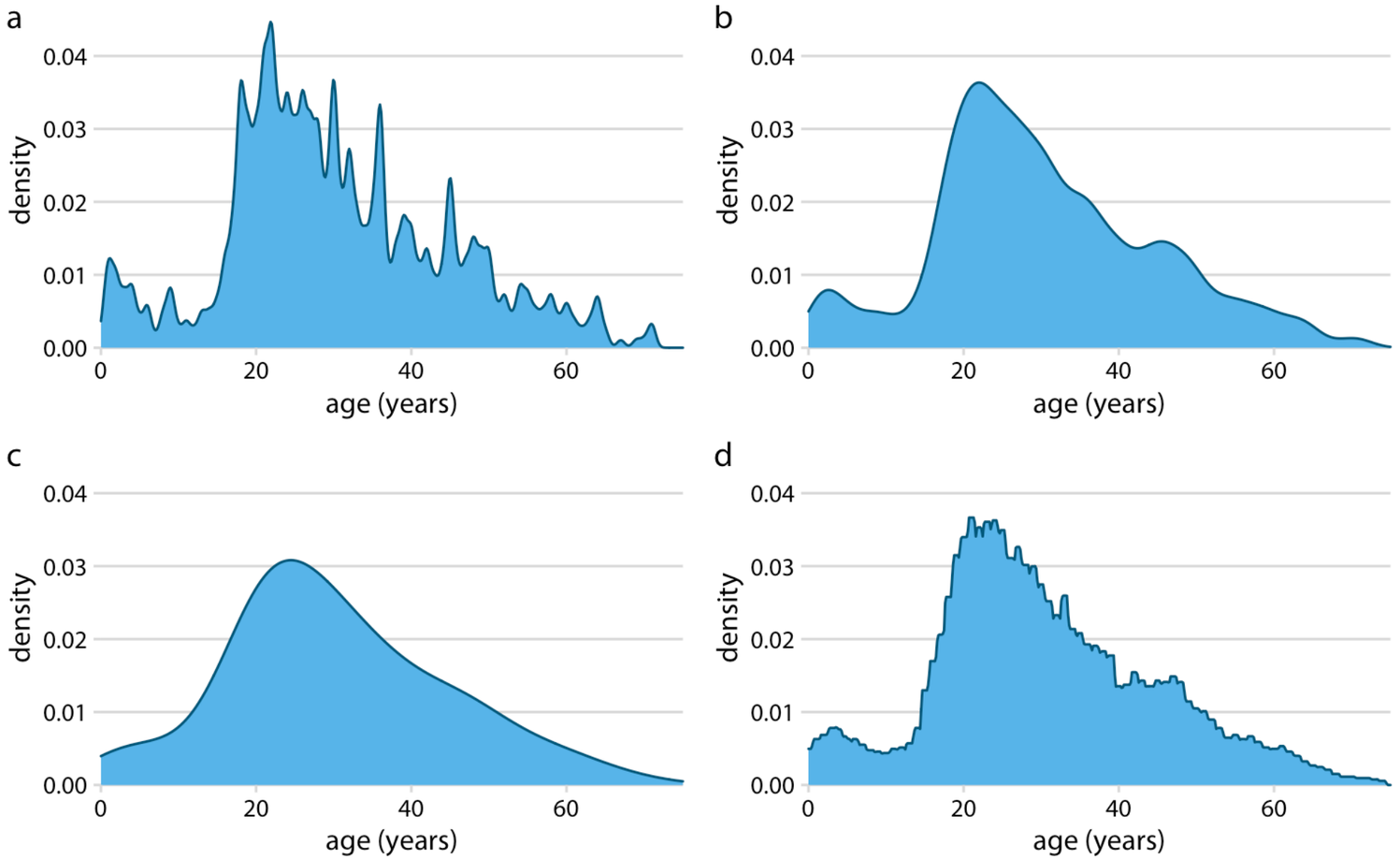


Histograms depend on the chosen bin width. Here, the same age distribution of Titanic passengers is shown with four different bin widths: (a) one year; (b) three years; (c) five years; (d) fifteen years.

https://clauswilke.com/dataviz/histograms-density-plots.html

Univariate Visual Idioms: Density Plot

A density plot is a representation of the distribution of a numeric variable. It uses the kernel density estimate to show the probability density function of a variable. It is basically a smoothed out version of the histogram.



Kernel density estimates depend on the chosen kernel and bandwidth. Here, the same age distribution of Titanic passengers is shown for four different combinations of these parameters: (a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2; (c) Gaussian kernel, bandwidth = 5; (d) Rectangular kernel, bandwidth = 2.

https://clauswilke.com/dataviz/histograms-density-plots.html

Multivariate Numerical Summaries

Categorical Variable

- cross-tabulation
- Univariate statistics by category

	Cake	Ice	Donut	Total
Female	4	3	6	**13**
Male	5	7	9	**21**
Total	**9**	**10**	**15**	**34**

Quantitative Variable

- Correlation & Covariance: a measure of how much (and in what direction) should we expect one variable to change when the other changes.
- Correlation & Covariance Matrix for >2 variables

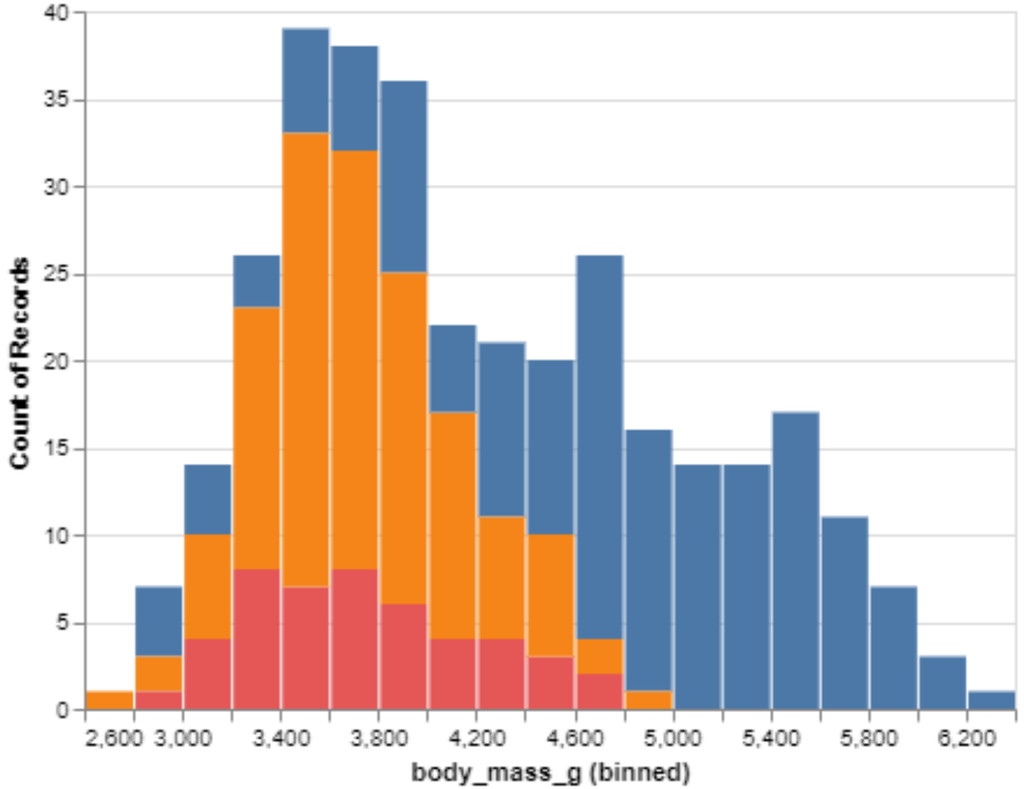
Experimental Design and Analysis. Howard J. Seltman

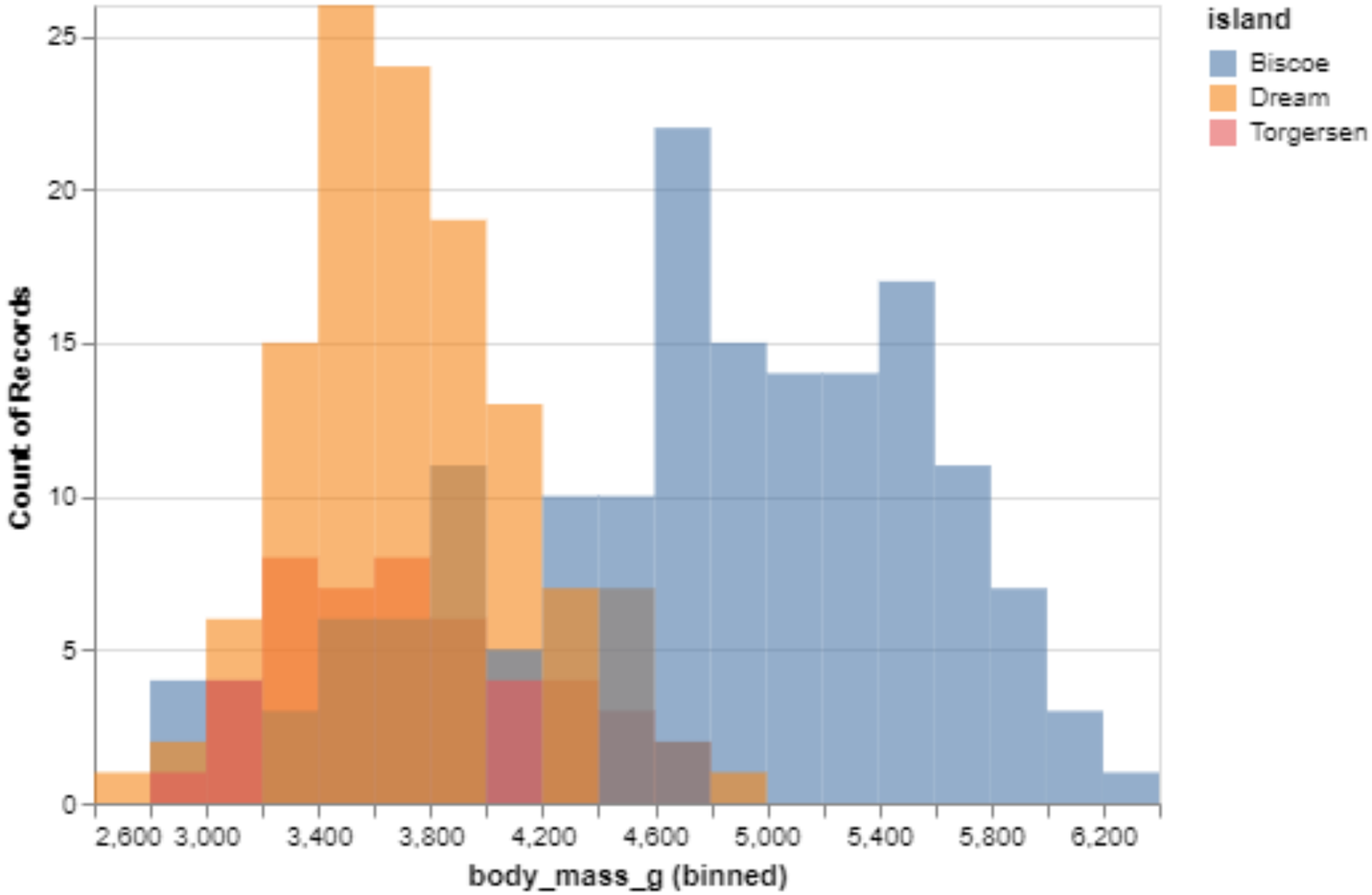
https://datatab.net/tutorial/cross-table

Multivariate Visual Idioms

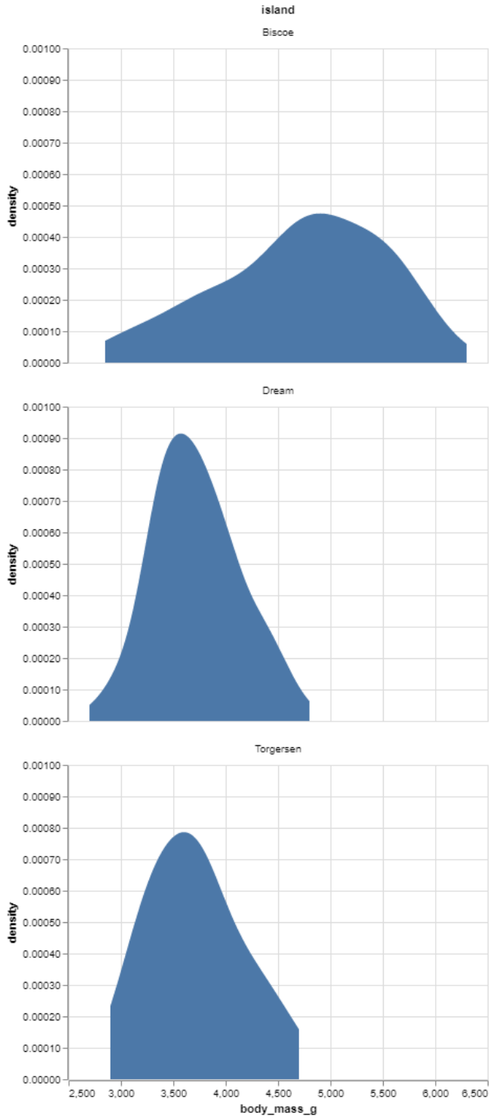
- Categorical Data
  - Stacked Bar Charts
- Quantitative Data
  - Overlapping Density Plots
  - Scatterplots
  - Box-plots & Violin Plots (see T8)
  - Univariate graphs by category – typically used when we have one explanatory variable (categorical) and one outcome (quantitative) variable

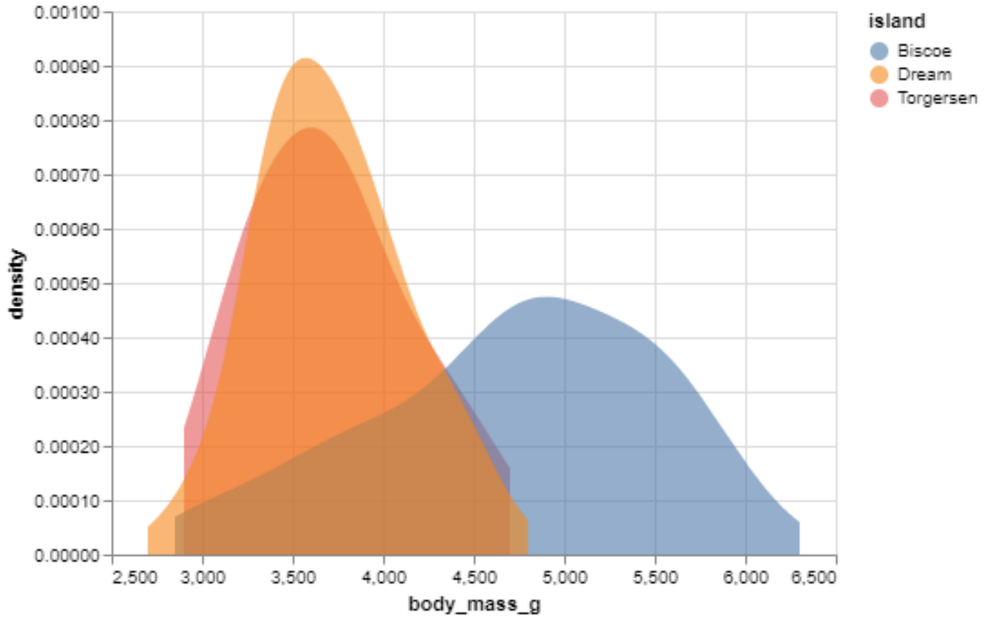
Bivariate Visual Idioms: Stacked Bar Chart

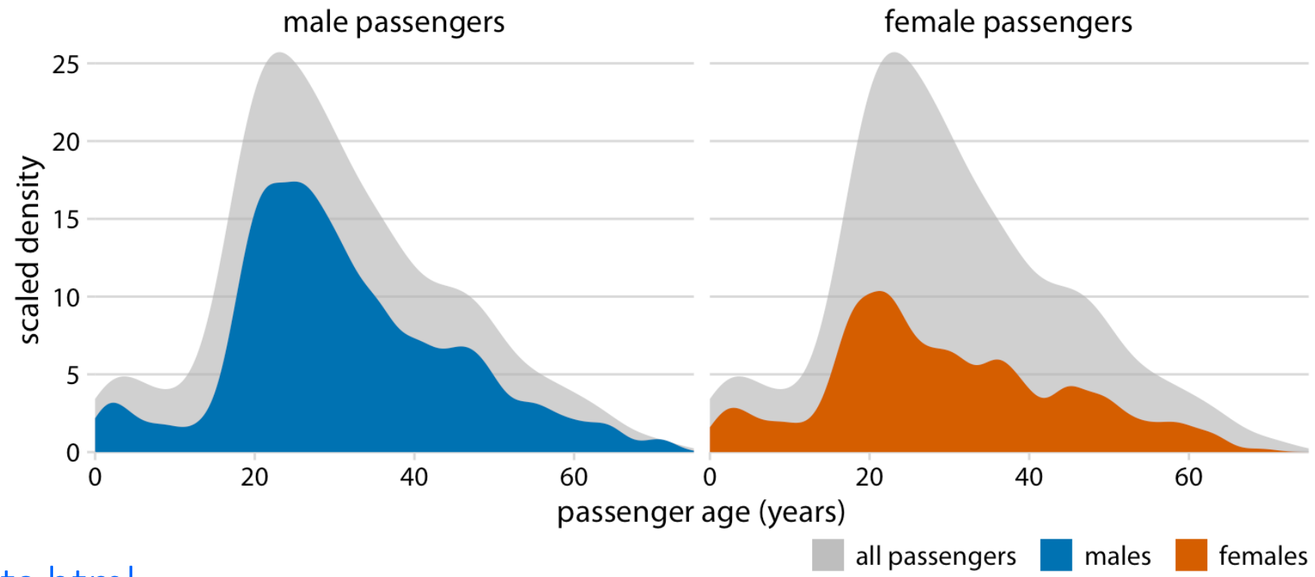




Multivariate Visual Idioms: Overlapping & Faceted Density Plots

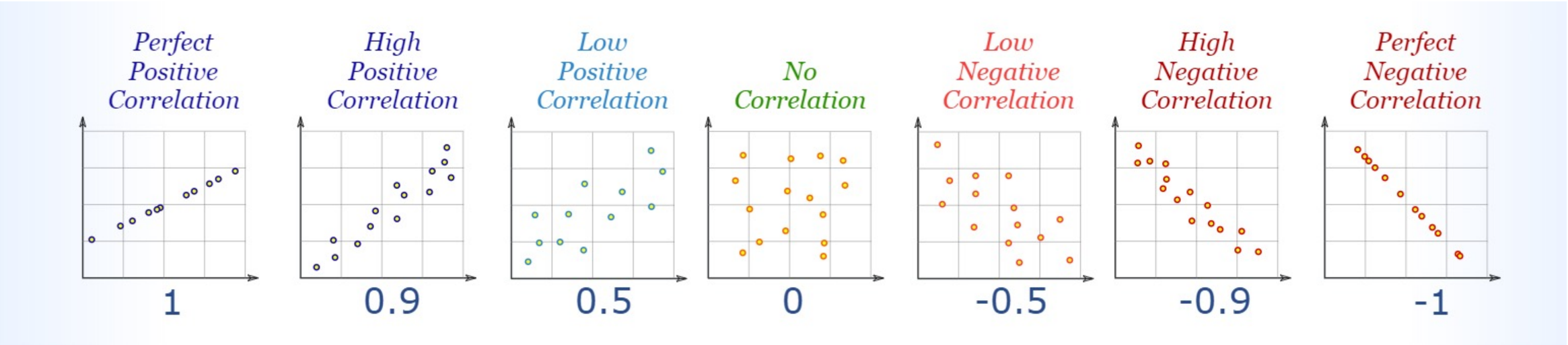






https://clauswilke.com/dataviz/histograms-density-plots.html

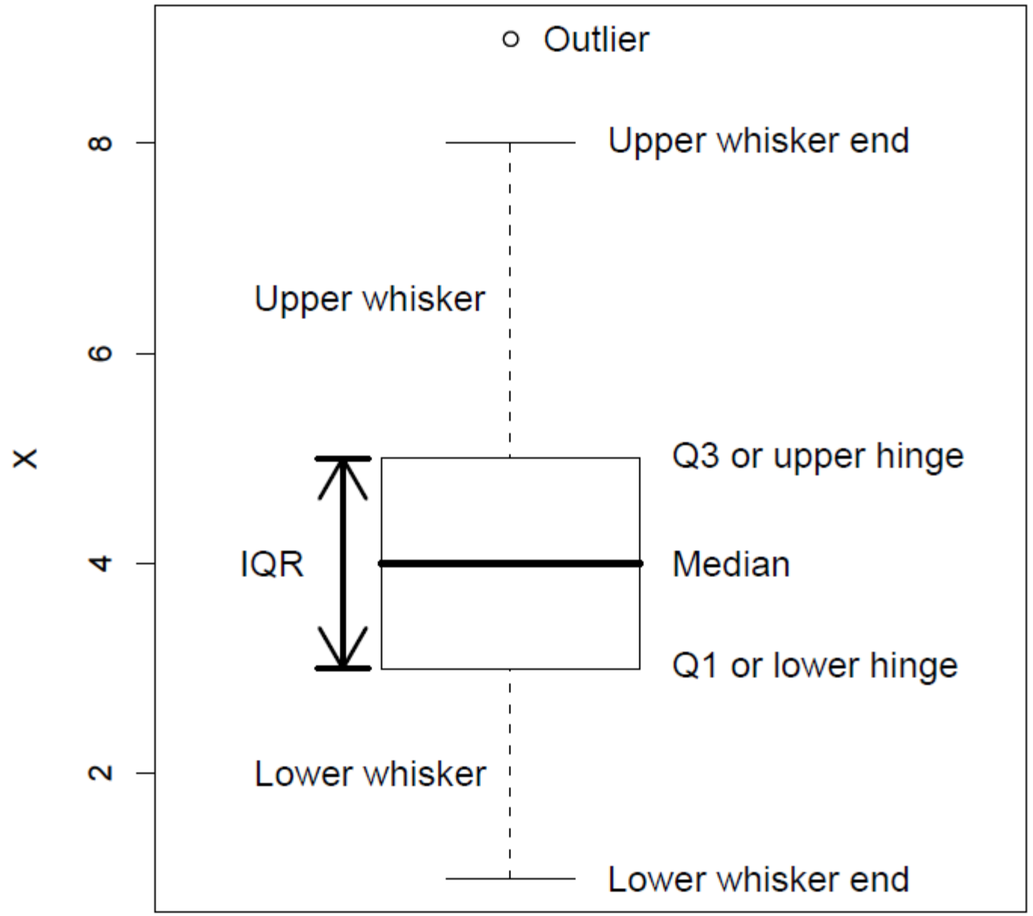
Bivariate Visual Idioms: Scatterplot



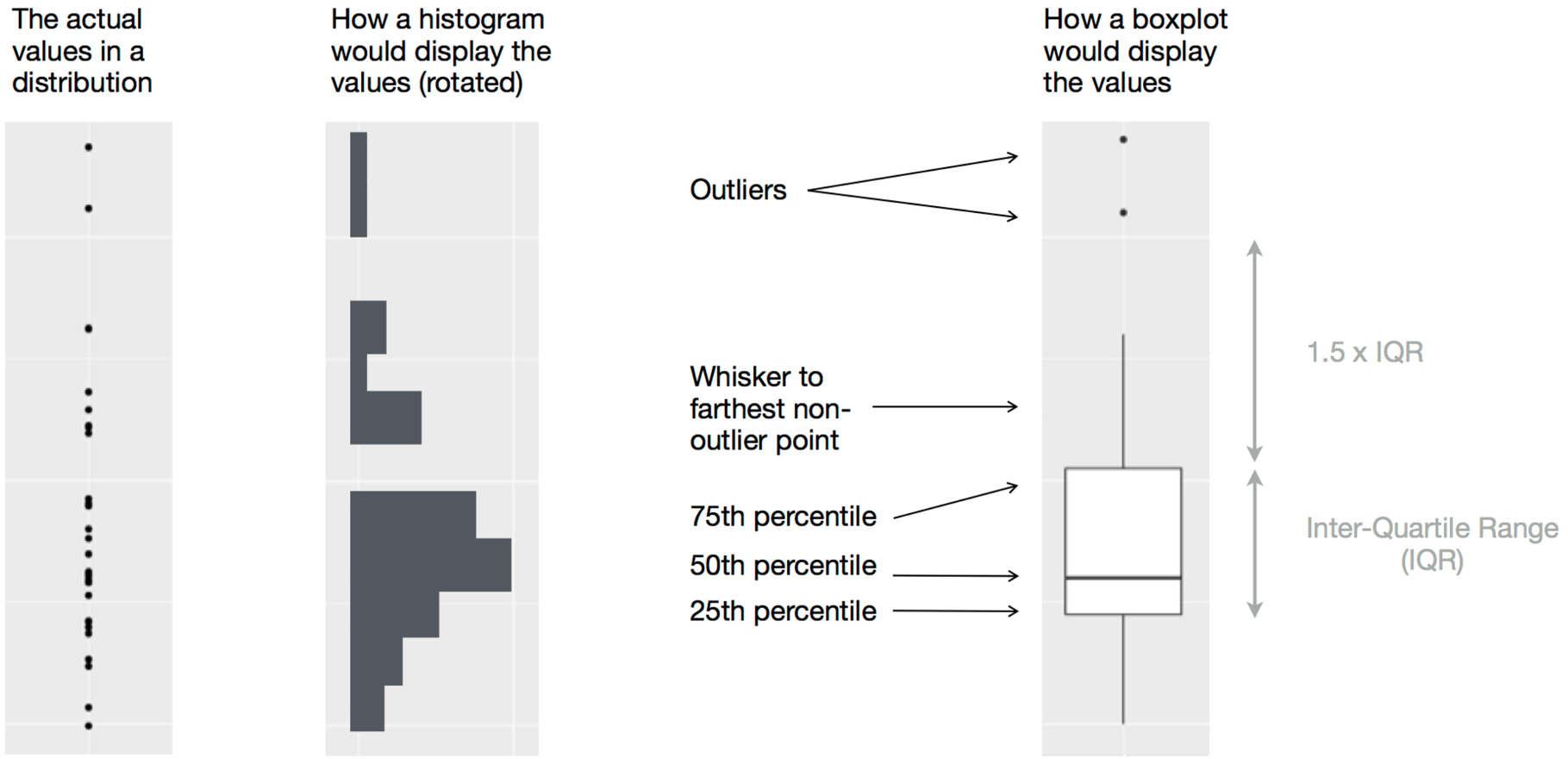
https://www.mathsisfun.com/data/scatter-xy-plots.html

Multivariate Visual Idioms: Boxplots

Very good at representing data related to the central tendency, symmetry and skew.



Multivariate Visual Idioms: Boxplots



https://r4ds.had.co.nz/exploratory-data-analysis.html

“In a nutshell: You should always perform appropriate EDA before further analysis of your data. Perform whatever steps are necessary to become more familiar with your data, check for obvious mistakes, learn about variable distributions, and learn about relationships between variables. EDA is not an exact science ***{ it is a very important art!}”***

- Howard J. Seltman

Get Stepping

- Office Hours this week are different – see earlier slides
- Quiz this week is Wednesday – Thursday
  - This Pandas Cheet Sheet from Pydata has been added as a resource for the quiz
  - https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf
  - On course website – there are some programming exercises, if stuck visit office hours or post on Ed.
  - Run your code, no red errors or that is a 0
- General Notes
  - No late work accepted, you are nearing graduation