Visualization for Data Science

Channels

Administrivia – Office Hours

TA Office Hours (online)

Monday 11:00am – 12pm

Wednesday 10:00 – 11am

Instructor Office Hours

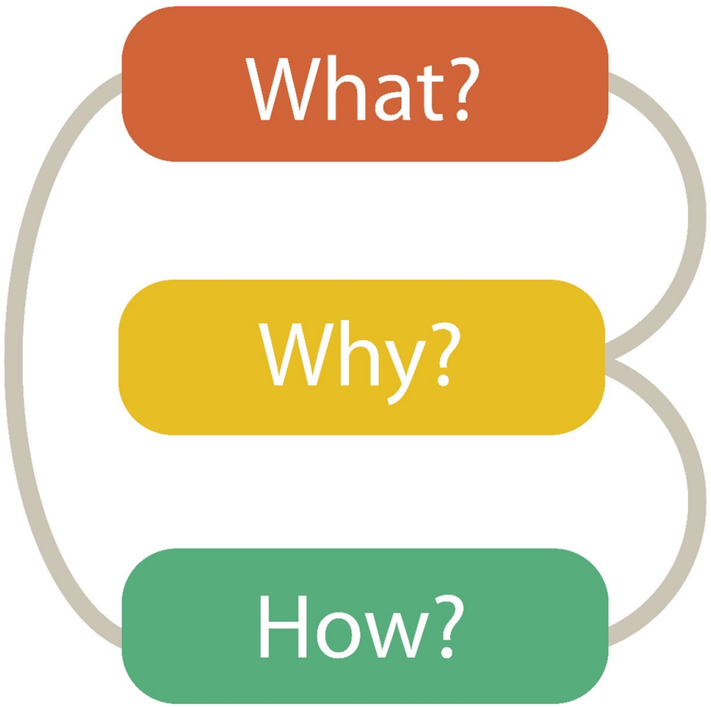
Monday and Wednesday 5 – 6pm (currently in this room)

Tuesdays 2 – 3pm ICCS 227

Learning Outcomes

- **Determine** the cardinality of a dataset attribute
- **Explain how cardinality impacts visual encoding choices**
- **Identify** the five characteristics of a visual channel
- **Explore** how channel characteristics influence visualization choices
- **Describe** how a channel is used based on its characteristics
- **Critique** a visualization using the principles of effectiveness and expressiveness

Data Visualization Ecosystem



What data is shown?

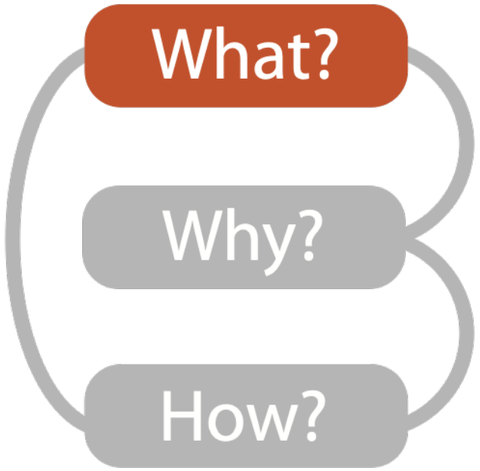
DATA ABSTRACTION

Why is the user analyzing / viewing it?

TASK ABSTRACTION

How is the data presented?

VISUAL AND INTERACTION ENCODING



What?

Datasets

Data Types

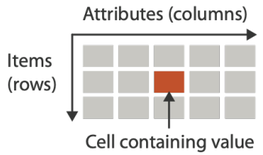
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

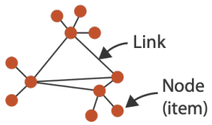
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

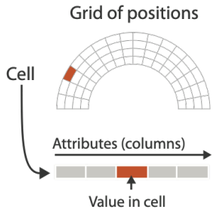
→ Tables



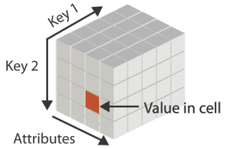
→ Networks



→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*

→ Geometry (Spatial)

Dataset Availability

→ Static

→ Dynamic

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

→ *Quantitative*

Ordering Direction

→ Sequential

→ Diverging

→ Cyclic

Data Abstraction

Moved from domain-specific language (e.g. births, cell growth in mm, social media platform) to generic language (e.g. quantitative, nominal).

This allows us to have a structure that transcends domain, so regardless of the data that you are working with, we can abstract away the details to focus on what is common across all datasets (its attribute type)

Data Abstraction

Column	Description	Attribute Type
country	Country name	nominal
year	Year of observation	temporal
population	Population in the country at each year	quantitative
region	Continent the country belongs to	nominal
sub_region	Sub-region the country belongs to	nominal
income_group	Income group	ordinal
life_expectancy	The mean number of years a newborn would live if mortality patterns remained constant	quantitative
income	GDP per capita (in USD) *adjusted for differences in purchasing power*	quantitative
children_per_woman	Average number of children born per woman	quantitative
child_mortality	Deaths of children under 5 years of age per 1000 live births	quantitative
pop_density	Average number of people per km^2	quantitative
co2_per_capita	CO2 emissions from fossil fuels (tonnes per capita)	quantitative
years_in_school_men	Mean number of years in primary, secondary, and tertiary school for 25-36 years old men	quantitative
years_in_school_women	Mean number of years in primary, secondary, and tertiary school for 25-36 years old women	quantitative

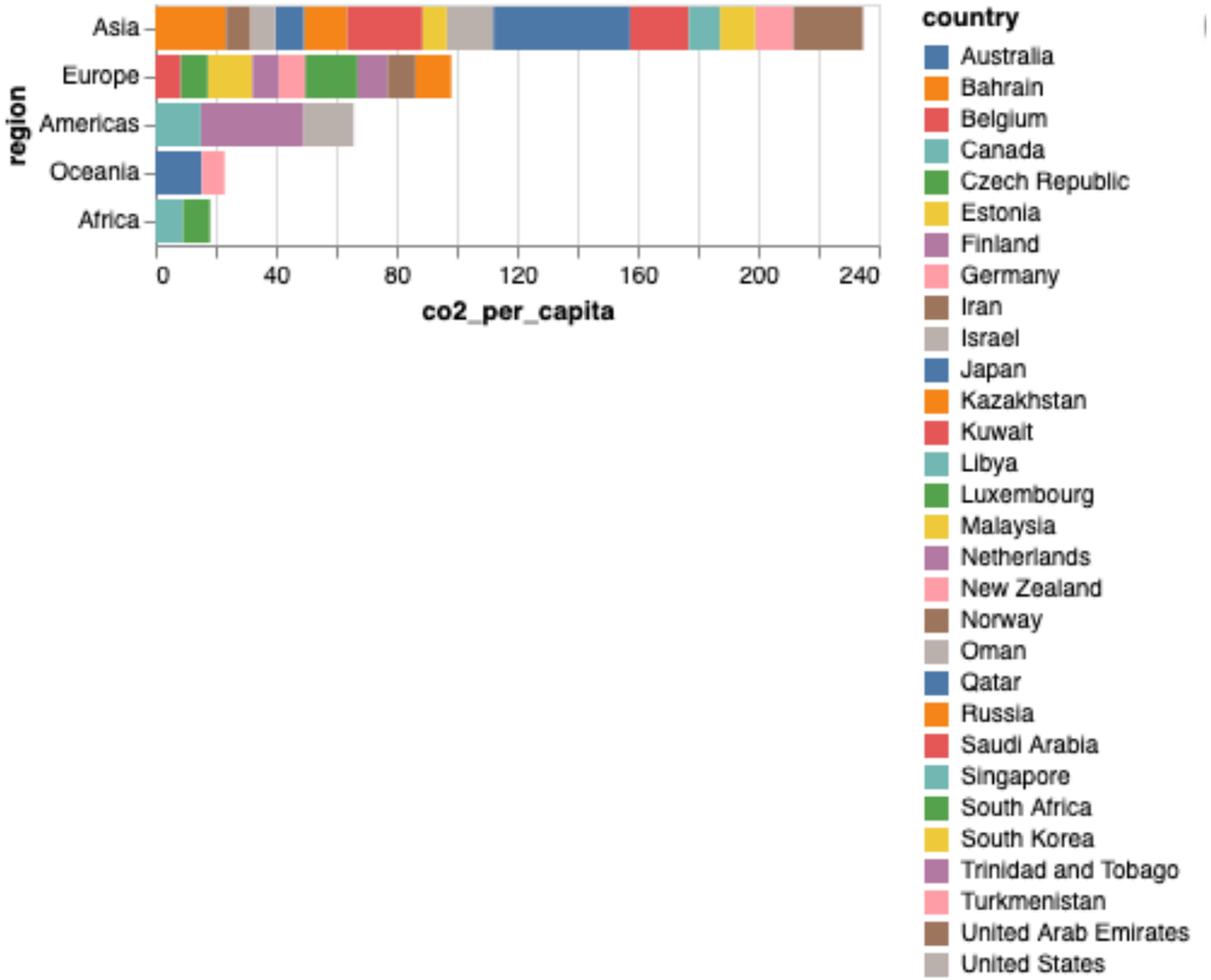
Task Abstraction - Why is the user viewing it

How are CO_2 emissions per capita distributed across different regions of the world, and what patterns emerge when comparing countries with the highest levels of emissions?

I choose to viz this by putting a categorical attribute on the y channel, another categorical attribute on the color channel, and CO2 emissions per capita on the x channel

Visual Encoding

I choose to viz this by putting a categorical attribute on the y channel, another categorical attribute on the color channel, and CO2 emissions per capita on the x channel.



Why is this viz problematic

Data Abstraction: Cardinality

Cardinality = number of unique values in an attribute

For **nominal and ordinal** data:

- Cardinality = number of unique categories

For **quantitative** data:

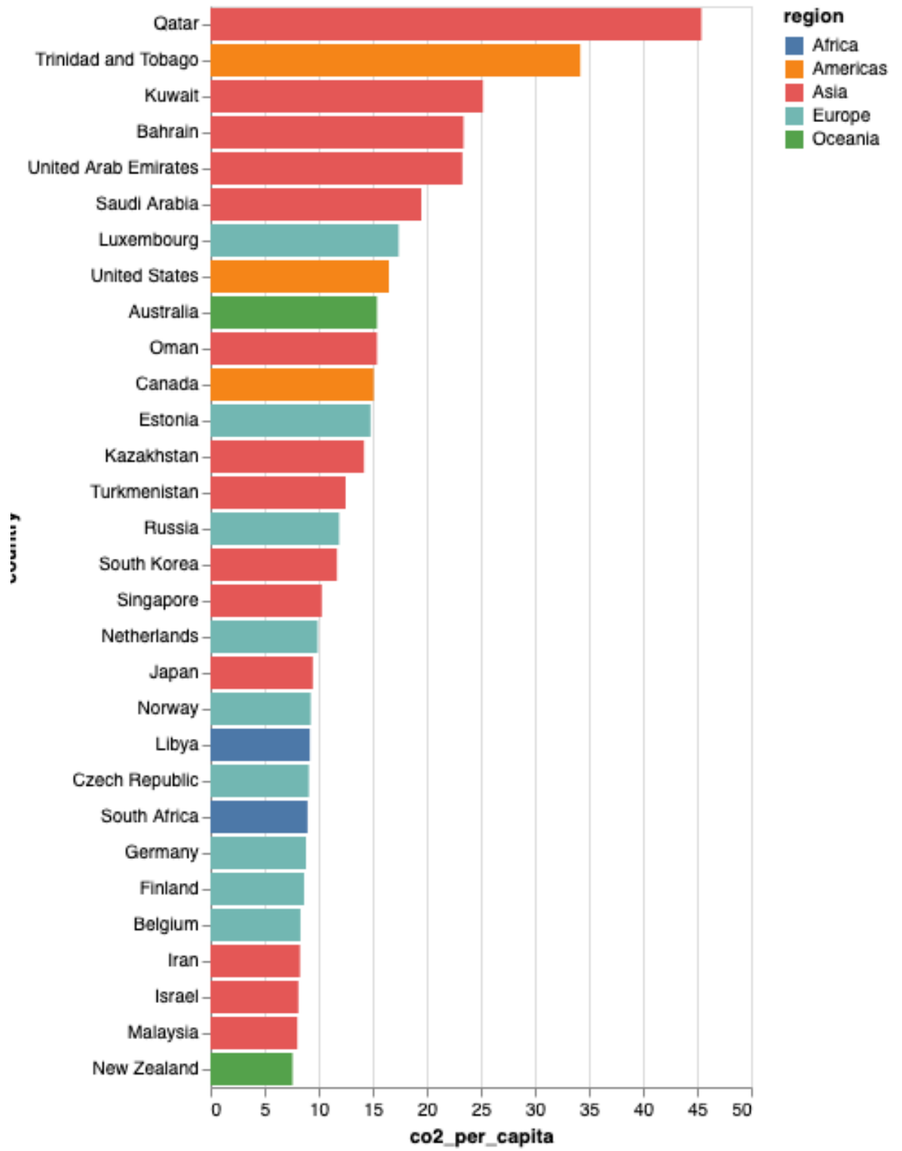
- Cardinality = range of values [min → max]
- Also consider whether the values are **discrete** or **continuous**

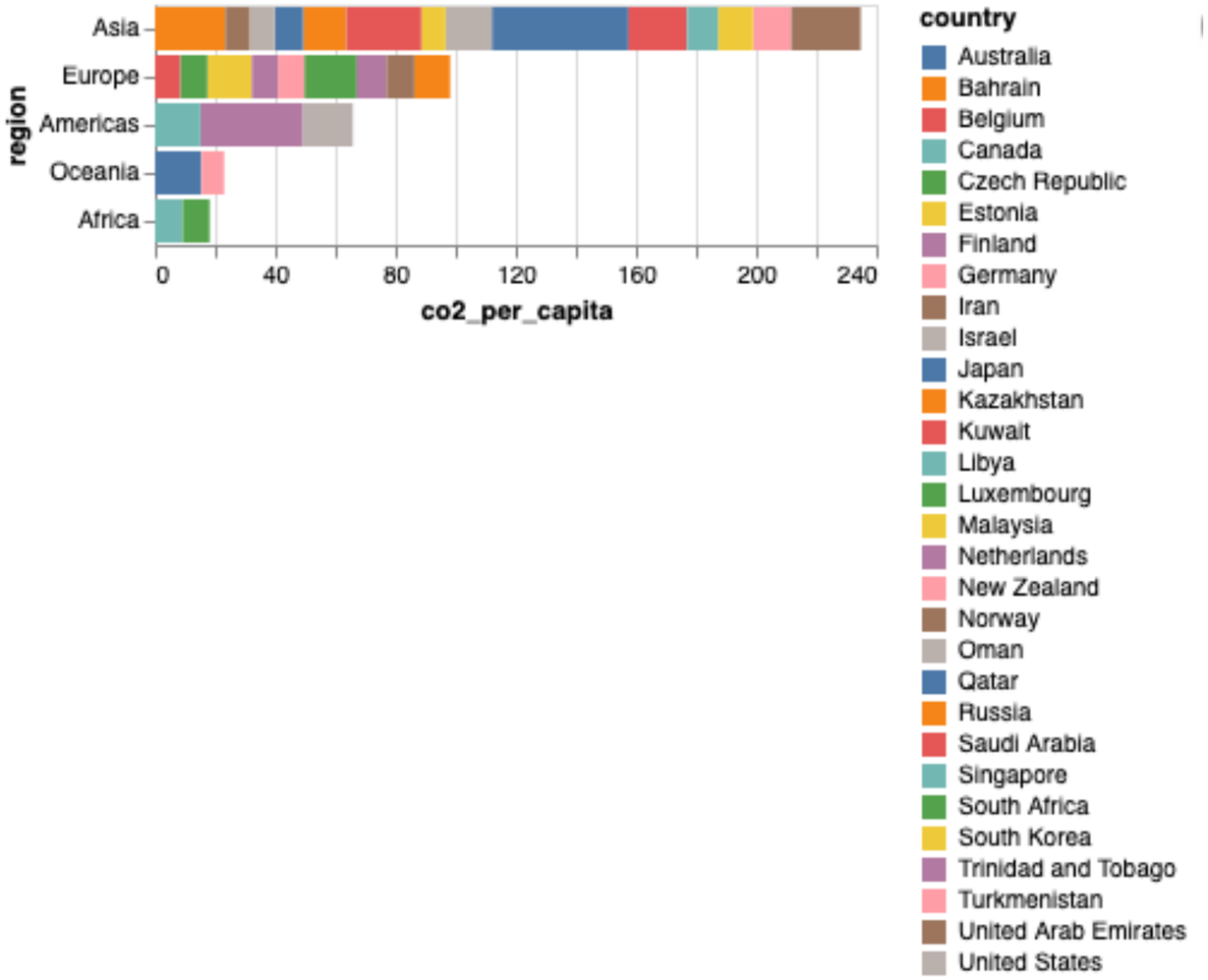
Cardinality is important because it affects which visual channels (e.g., color, shape, position) are appropriate

Cardinality – the missing link

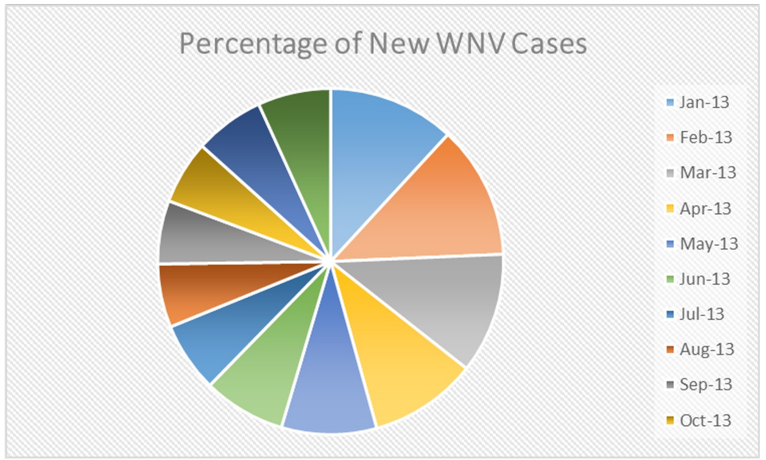
	column	type	cardinality
0	country	categorical	178 unique values
1	region	categorical	5 unique values
2	sub_region	categorical	17 unique values
3	income_group	categorical	4 unique values
4	population	quantitative	min = 12500, max = 1420000000
5	life_expectancy	quantitative	min = 1.0, max = 84.2
6	income	quantitative	min = 247, max = 178000
7	children_per_woman	quantitative	min = 1.12, max = 8.87
8	child_mortality	quantitative	min = 1.95, max = 756.0
9	pop_density	quantitative	min = 0.502, max = 8270.0
10	co2_per_capita	quantitative	min = 0.0, max = 101.0
11	years_in_school_men	quantitative	min = 0.9, max = 15.3
12	years_in_school_women	quantitative	min = 0.21, max = 15.7
13	year	temporal	min = 1800-01-01, max = 2018-01-01

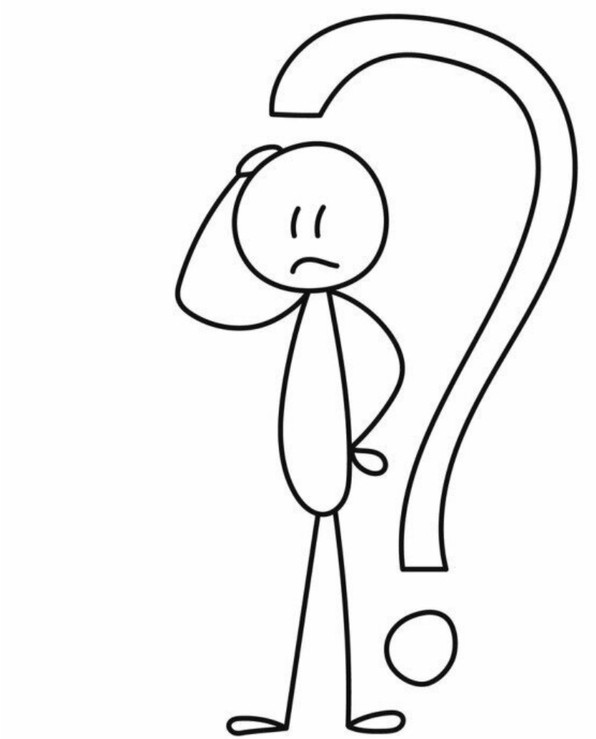
Encoding region on y channel as opposed to color

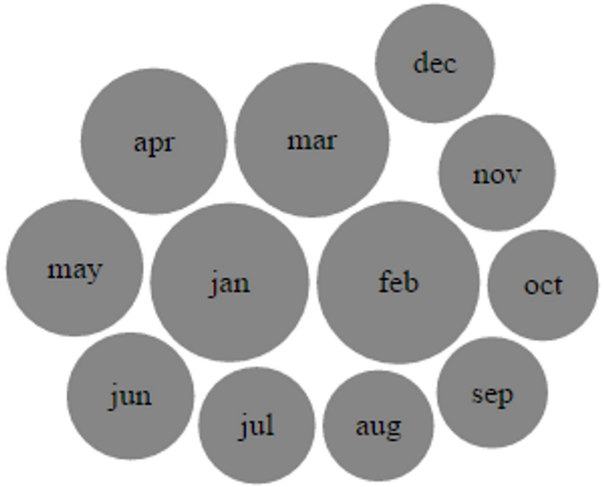


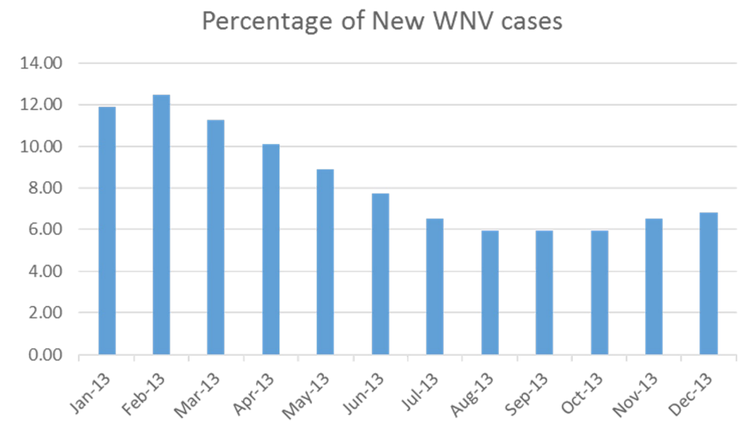


Representation Effect









How do I pick ***which*** marks or channels to use?

Channels

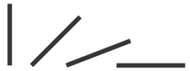
Magnitude Channels: Ordered Attributes

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

Spatial region

Color hue

Motion

Shape

Channel Selection

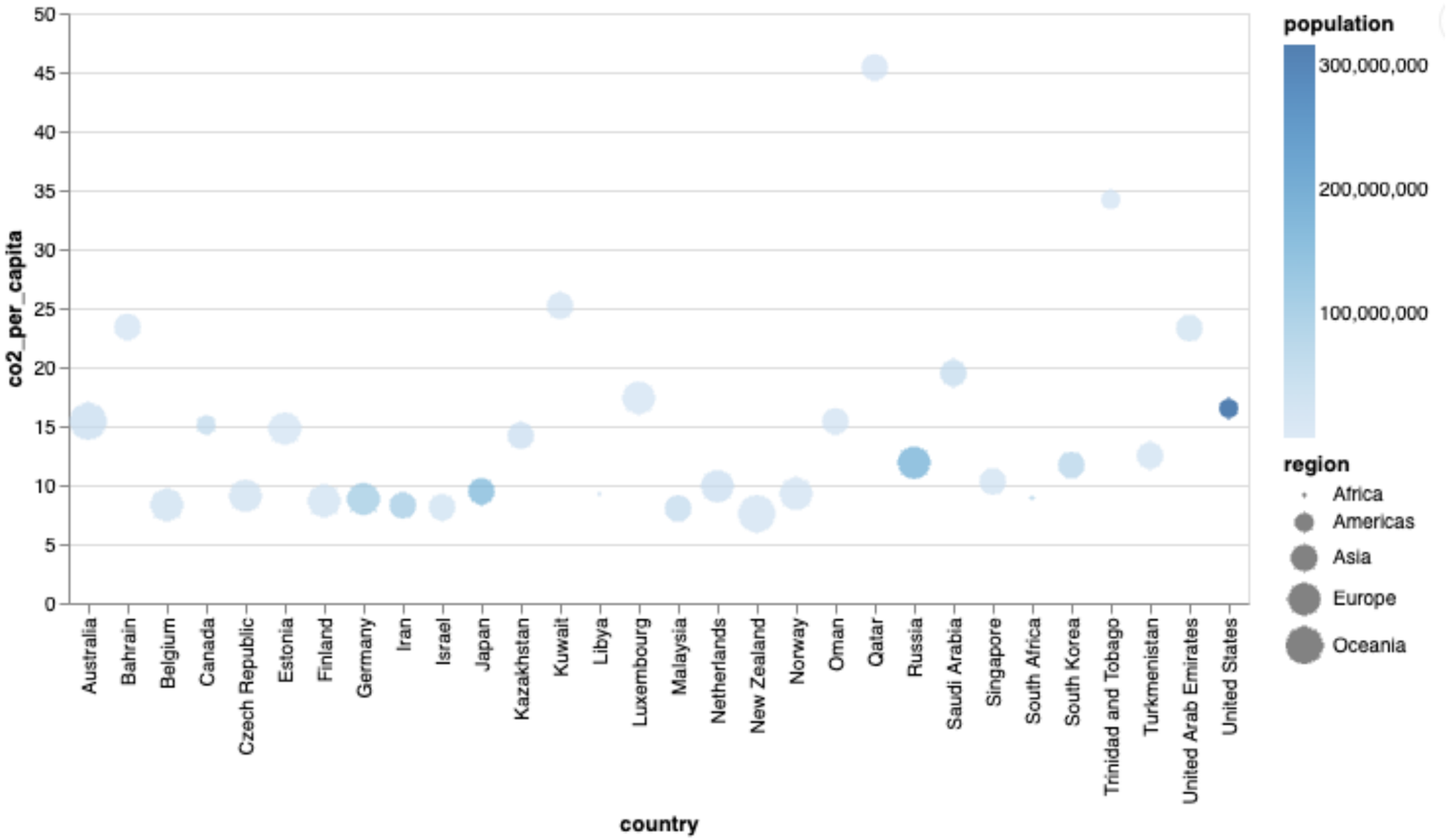
Before we can select channels we must understand how its various properties influence their use.

We must consider both

- **Expressiveness:** match the channel type to data characteristics: the visual encoding should express all of, **and only**, the information in the dataset attributes.

- **Effectiveness:** for a given task, some channels are **better** than others so it is important to select the most **effective** channel for the data

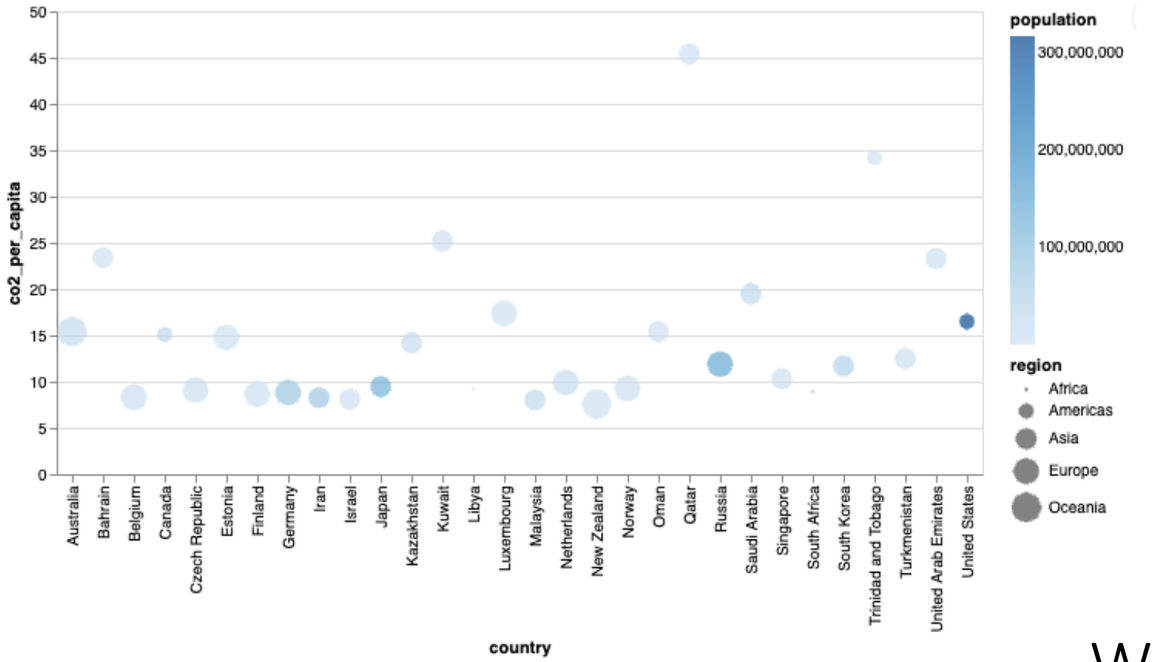
How are both principles broken?



Expressiveness: match the channel type to data characteristics: the visual encoding should express all of, **and only**, the information in the dataset attributes.

Effectiveness: for a given task, some channels are **better** than others so it is important to select the most **effective** channel for the data

How are both principles broken?



Which encoding mapping **violates expressiveness**?

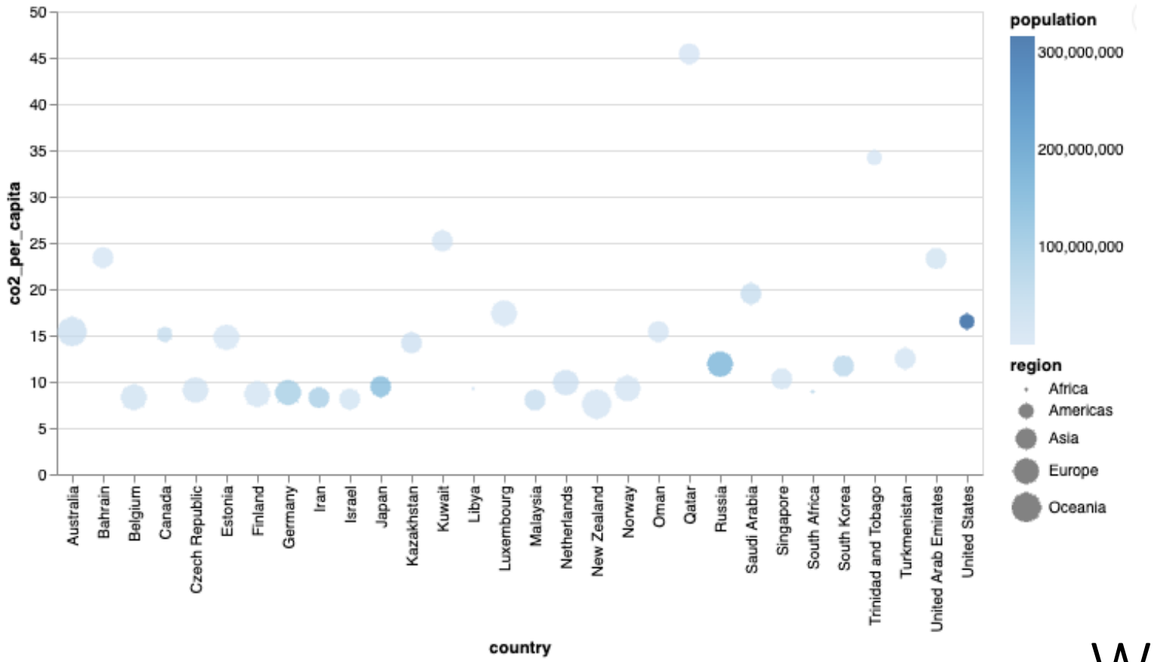
A. X-position

B. Y-position

C. Color

D. Size

How are both principles broken?



Which encoding mapping **violates effectiveness**?

A. X-position

B. Y-position

C. Color

D. Size

Channel Selection

Before we can select channels, we must understand how their various properties influence their use.

Channel Characteristics

- 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?

Discriminability: How many usable steps?

- How many usable steps?

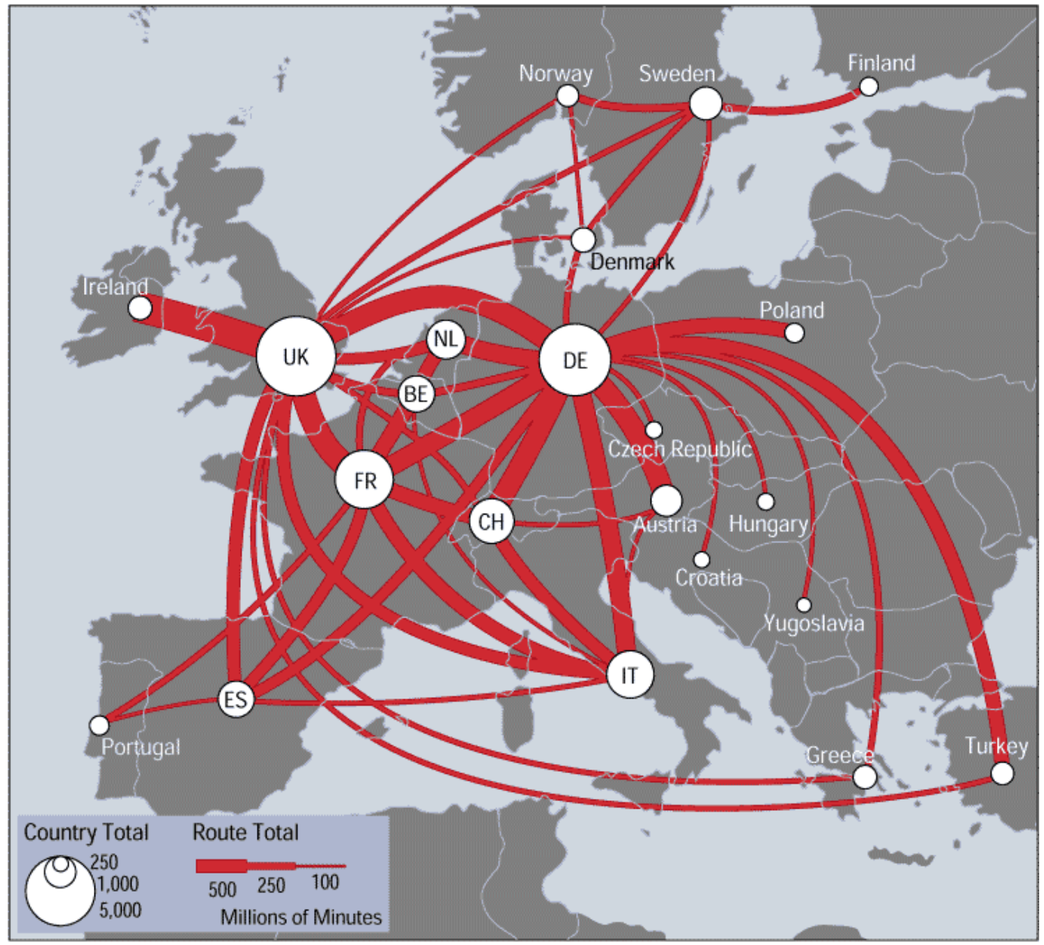
- How easily can differences between attribute levels be perceived?

Tips

The channel must be sufficient for number of attribute levels to show

- Linewidth – limited number of steps (maybe 4 at most)
- Color hue – max 5, using more isn't recommended
- Shapes – max 5, using more can create difficulties

ColorBrewer demo:
https://colorbrewer2.org/#type=qualitative&scheme=Paired&n=3



[mappa.mundi.net/maps/maps 014/telegeography.html]

READING, WRITING, AND EARNING MONEY

The latest data from the U.S. Census's American Community Survey paints a fascinating picture of the United States at the county level. We've looked at the educational achievement and the median income of the entire nation, to see where people are going to school, where they're earning money, and if there is any correlation.

① HIGH SCHOOL GRADUATES

② COLLEGE GRADUATES

③ MEDIAN HOUSEHOLD INCOME

The map at right is a product of overlaying the three sets of data. The variations in hue and value has been produced from the data shown above. In general, darker counties represent a more educated, better paid population while lighter areas represent communities with fewer graduates and lower incomes.

A collaboration between GOOD and Gregory Huhcenk
SOURCE: US Census

KING COUNTY, WA

EAGLE COUNTY, CO

HENNEPIN COUNTY, MN

NEW YORK COUNTY, NY

HOLMES COUNTY, OH

BREATHITT COUNTY, KY

ALACHUA COUNTY, FL

KOSTELA COUNTY, CO

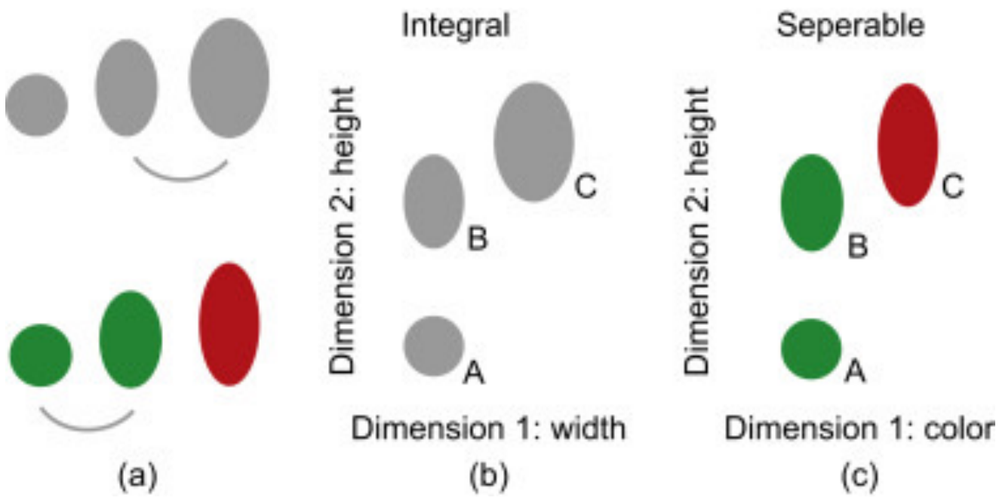
Separability vs. Integrality

Integral dimensions: two or more attributes are perceived holistically (not independently)

Separable dimensions: people tend to make separate judgements about each dimension (i.e., attribute)

Separability is our ability to use this channel affected by another one?

Can channels be used independently or is there interference from other channels being used?

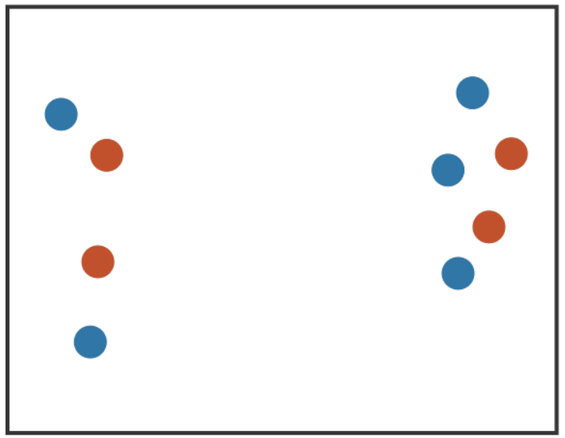


(a) The width and height of an ellipse are perceived integrally, so the ellipses are seen as more similar to each other (because they have the same shape) than the pair having the same width. The color and height of a shape are perceived separably, so the two green shapes are seen as most similar. (b, c) Space plots of the two examples.

Information Visualization by Colin Ware. Ch. 5.Fig. 20 Visual Salience: Finding and Reading Data Glyphs

Separability vs. Integrality

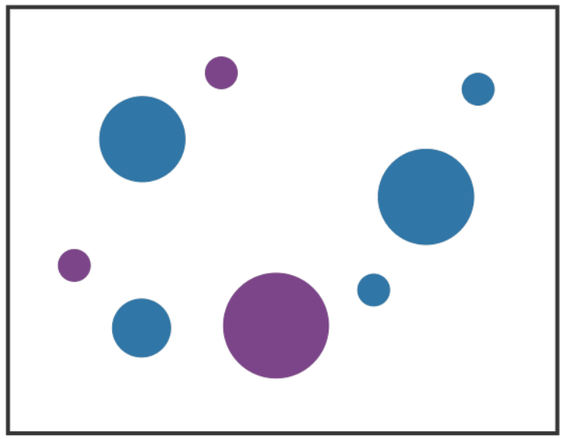
Position
+ Hue (Color)



Fully separable

2 groups each

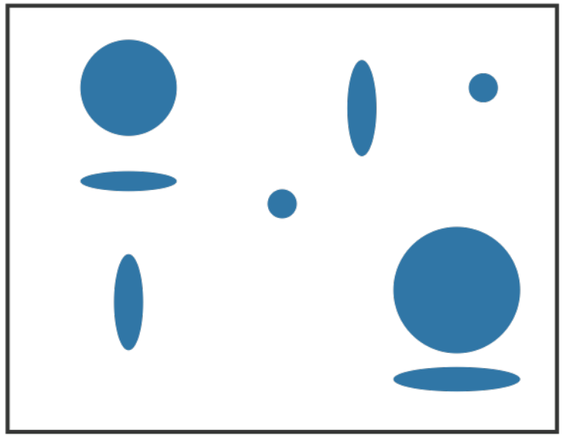
Size
+ Hue (Color)



Some interference

2 groups each

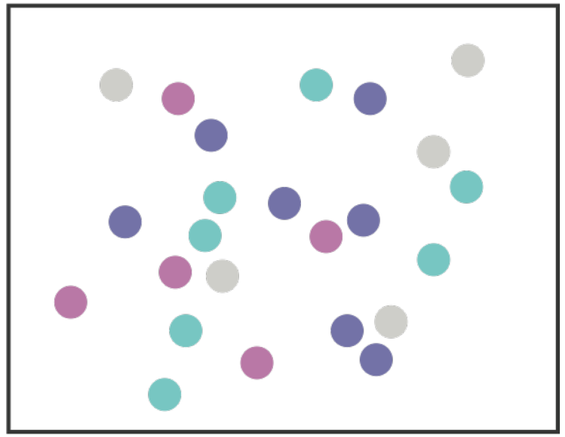
Width
+ Height



Some/significant interference

3 groups total: integral area

Red
+ Green



Major interference

4 groups total: integral hue

Design Considerations

Use Integral Dimensions: Opt for this when aiming for a holistic impact. This approach encourages viewers to perceive multiple elements (2 or 3) as a single cohesive unit, ensuring they are seen as one simple entity.

Use Separable Dimensions: This is suitable when the objective is for viewers to concentrate on individual elements sequentially. It allows them to distinguish between different channels one by one.

Information Visualization by Colin Ware. Ch. 5.Fig. 20 Visual Salience: Finding and Reading Data Glyphs

Popout/Ease of recognition

Can things jump out using this channel?

Can a channel provide popout where a difference is perceived preattentively?

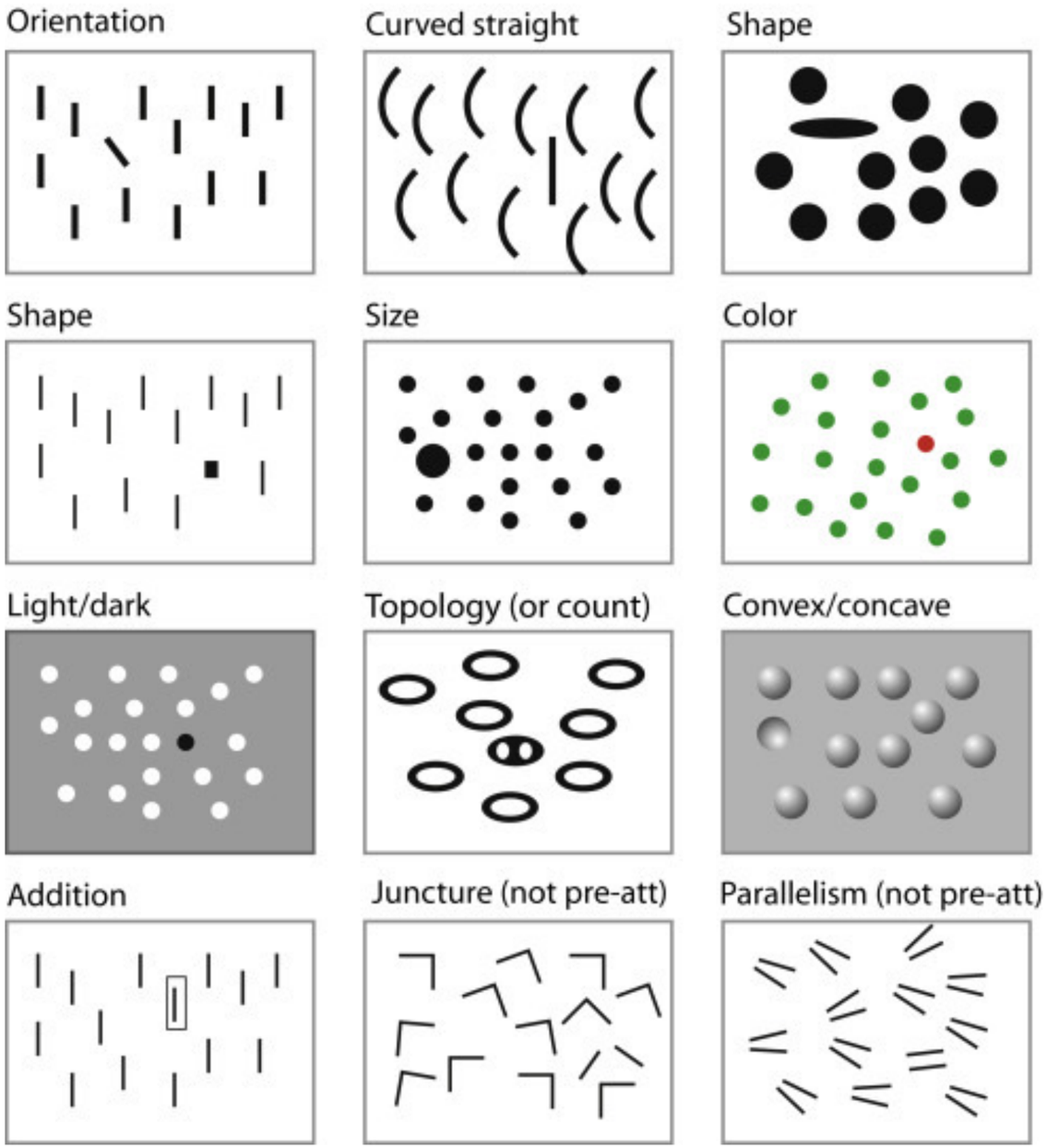
- Properties detected by the low-level visual system
- very rapid - 200-250 ms
- very accurate
- processed in parallel
- happens before focused attention -> preattentive

Design Guideline: use whatever channel is used least in other parts of the design

Popout

- find the red dot
  - how long does it take?
- parallel processing on many individual channels
  - speed independent of distractor count
  - speed depends on channel and amount of difference from distractors
- serial search for (almost all) combinations
  - speed depends on number of distractors

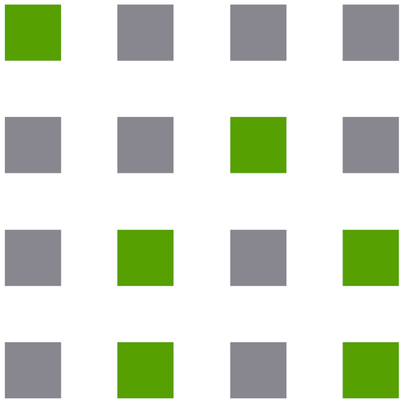
Popout

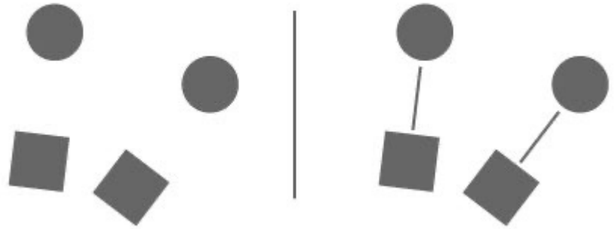


Information Visualization by Colin Ware. Ch. 5. Figure 12 Visual Salience: Finding and Reading Data Glyphs

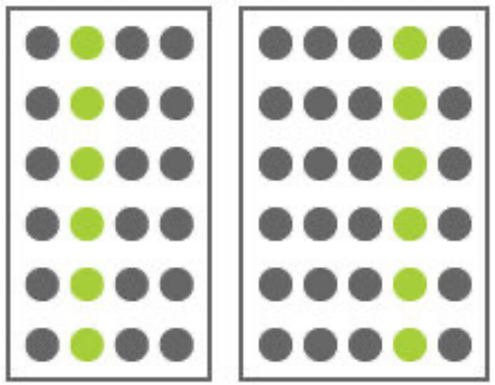
Grouping

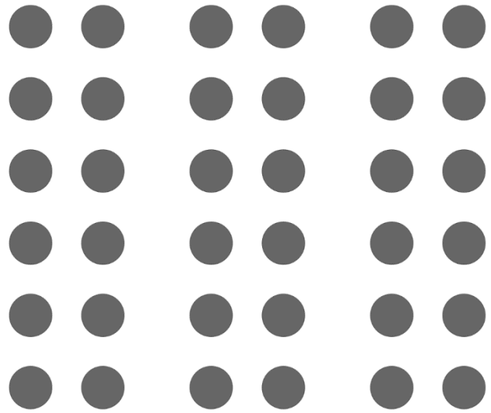
can channel show perceptual grouping of items?





A. containment
B. connection
C. proximity
D. similarity





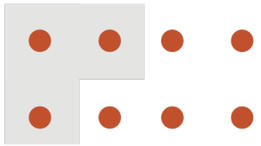
https://graybox.co/knowledge/blog/gestalt-principles-applied-to-design

Grouping *(more when we talk about Gestalt Properties)*

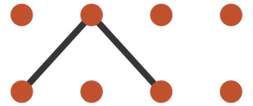
can channel show perceptual grouping of items?

- containment
- connection
- proximity
- similarity

Containment



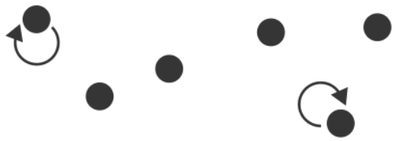
Connection



Spatial region

Color hue

Motion



Shape

Designing a visualization experiment

Imagine that you are visualization researchers.

Work with 3-4 people around you to design an experiment - involving human subjects - that would help visualization researchers answer the question:

Which channel is most accurate for encoding a quantitative attribute?

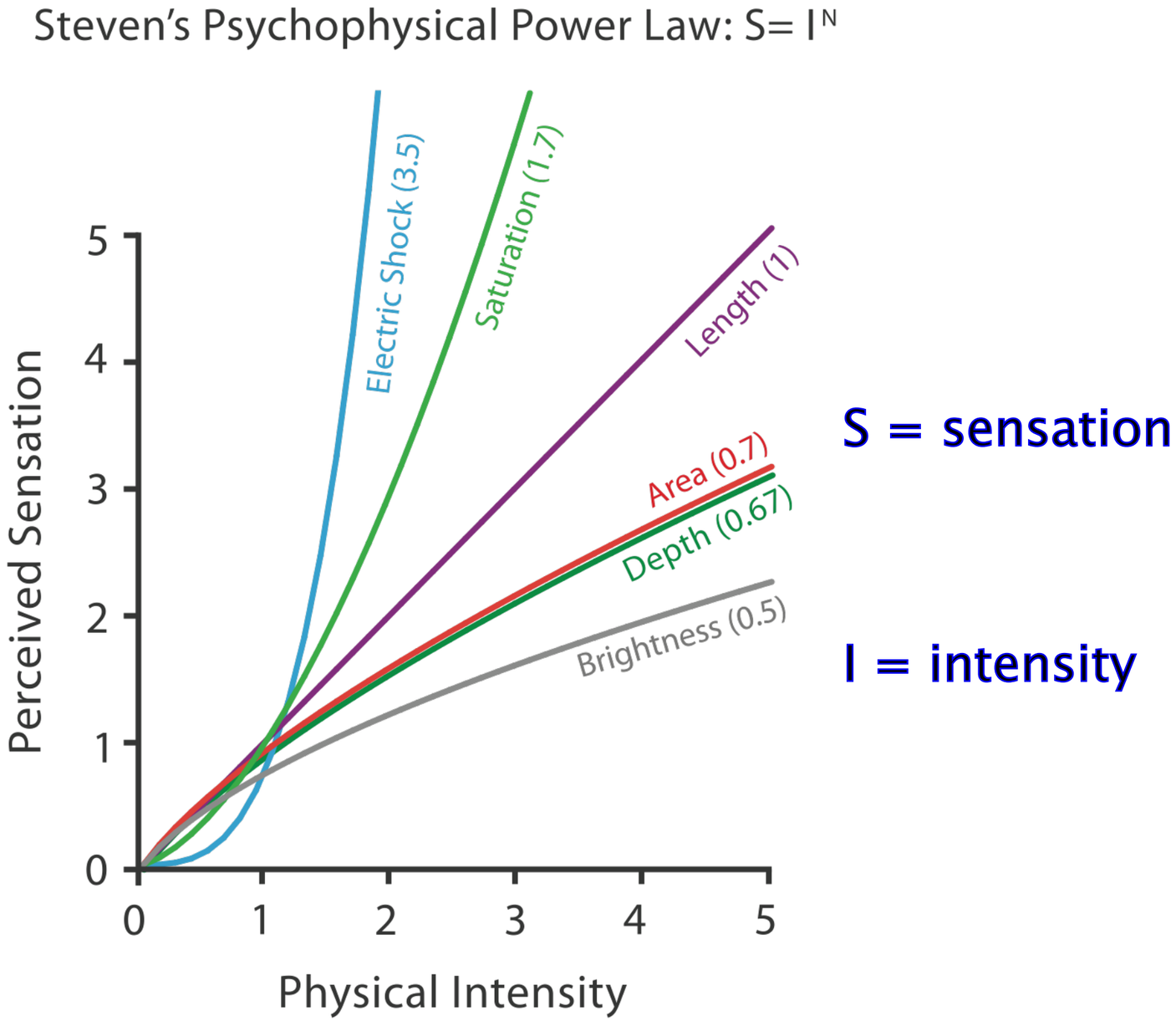
Accuracy: Fundamental theory

How precisely can we tell the difference between encoded items?

length is accurate: linear

others magnified or compressed

– exponent characterizes



S = sensation

I = intensity

Accuracy: User studies

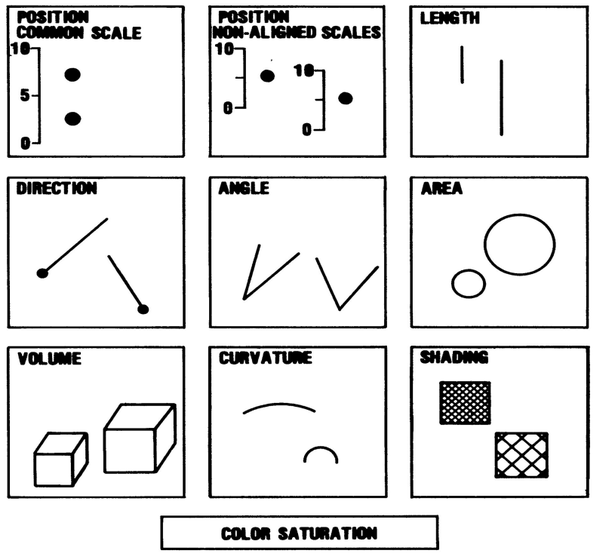
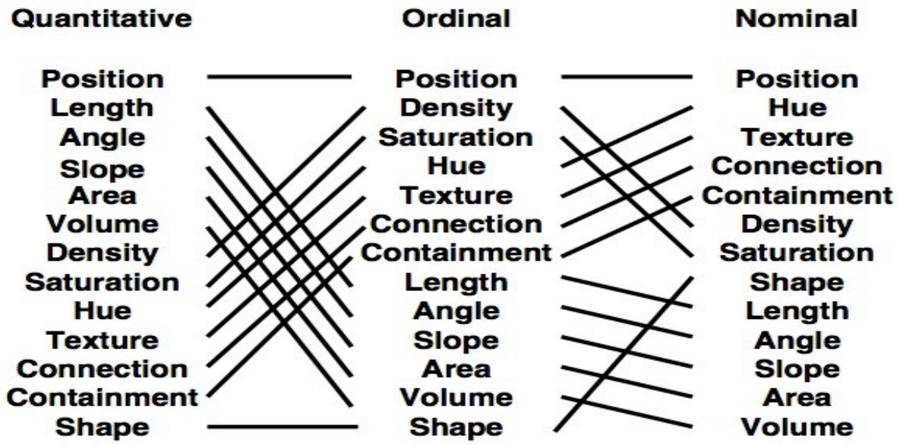


Figure 1. Elementary perceptual tasks.



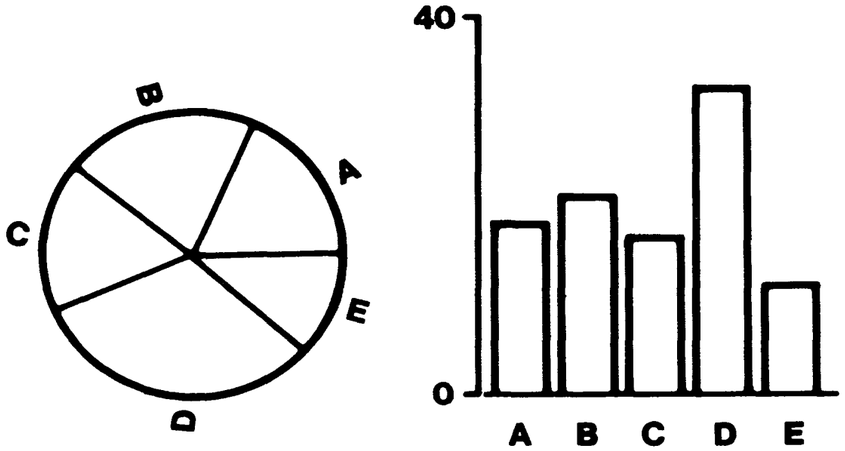


Figure 3. Graphs from position–angle experiment.

[Cleveland & McGill, 1984]

[Mackinlay, 1986]

Accuracy: User studies

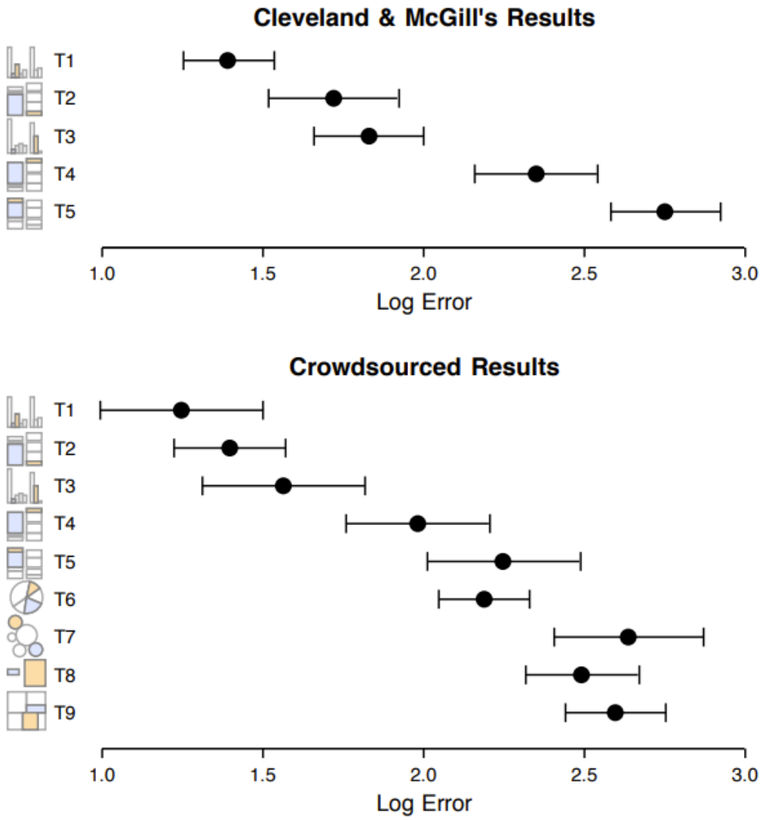
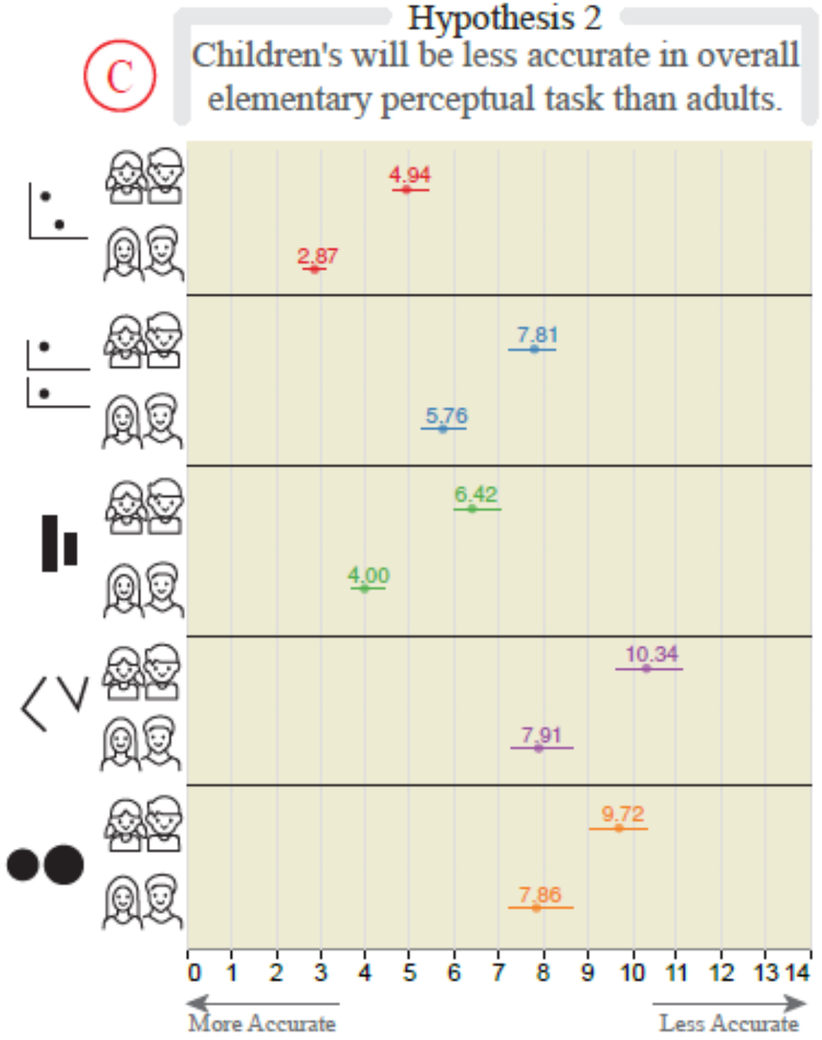
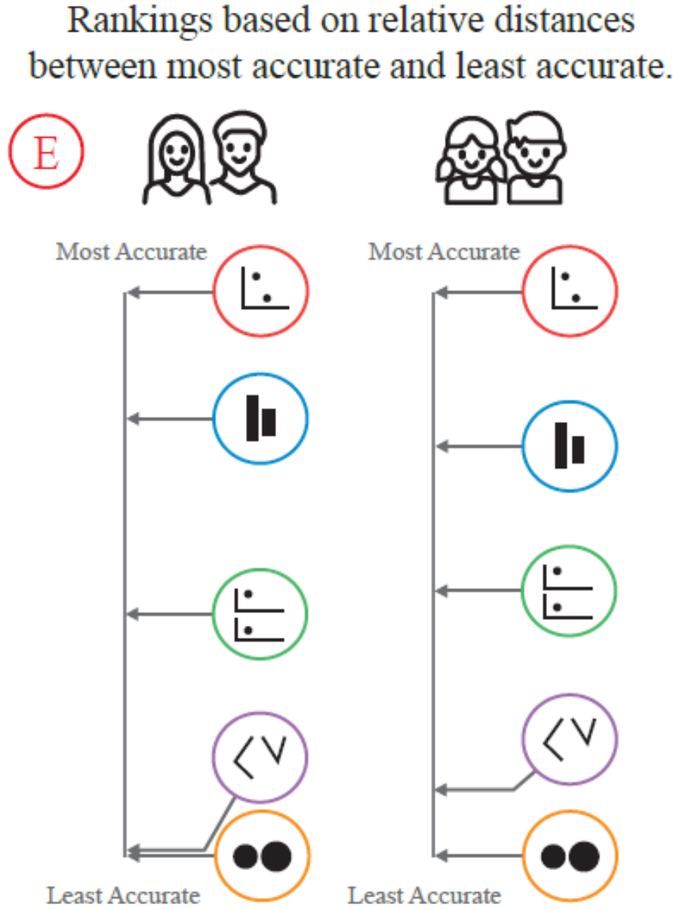


Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.





[Cleveland & McGill, 1984]
[Mackinlay, 1986]
[Heer & Bostock, 2010]
[Panavas et al., 2022]

Factors affecting accuracy

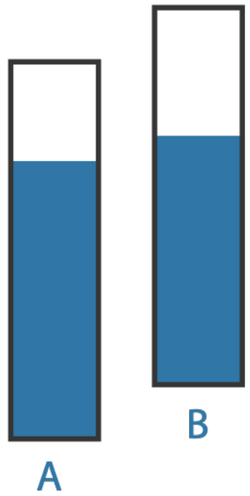
- alignment
- distractors
- distance
- common scale / alignment

Relative vs. absolute judgements

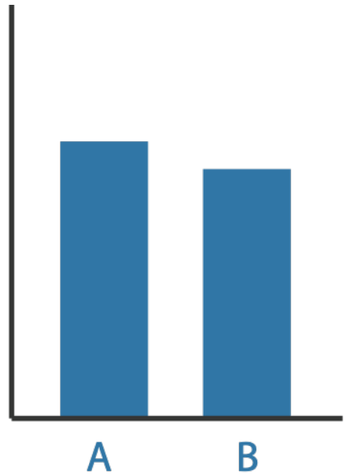
Perceptual system mostly operates with relative judgements, not absolute

- that's why accuracy increases with common frame/scale and alignment
- Weber's Law: ratio of increment to background is constant
  - filled rectangles differ in length by 1:9, difficult judgement
  - white rectangles differ in length by 1:2, easy judgement

length

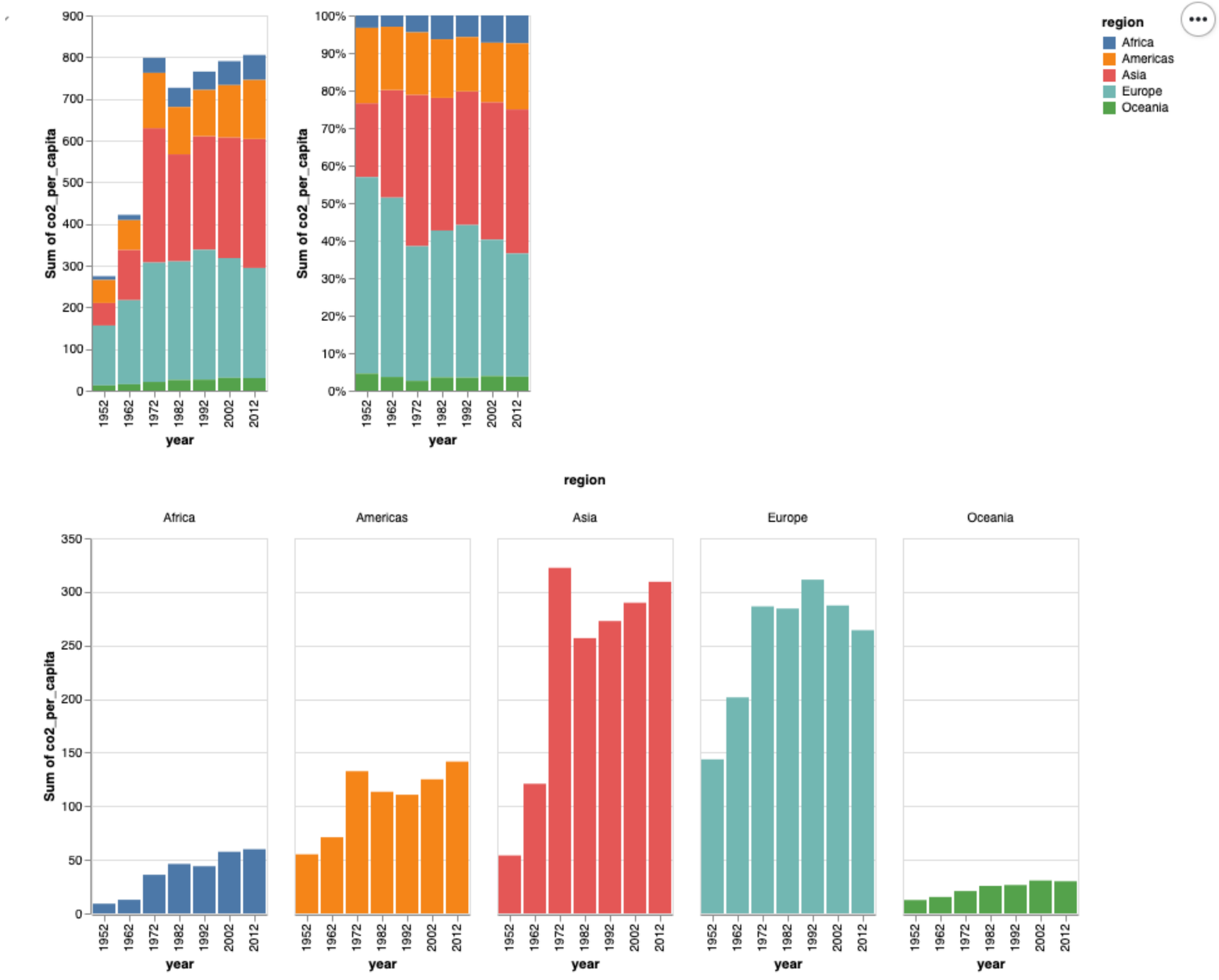


position along unaligned common scale

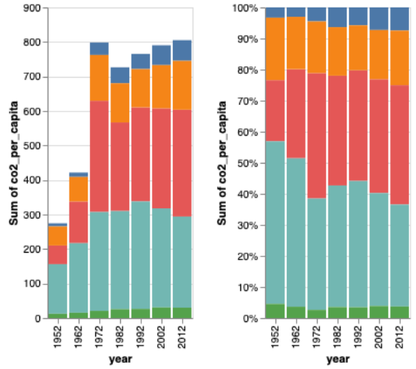


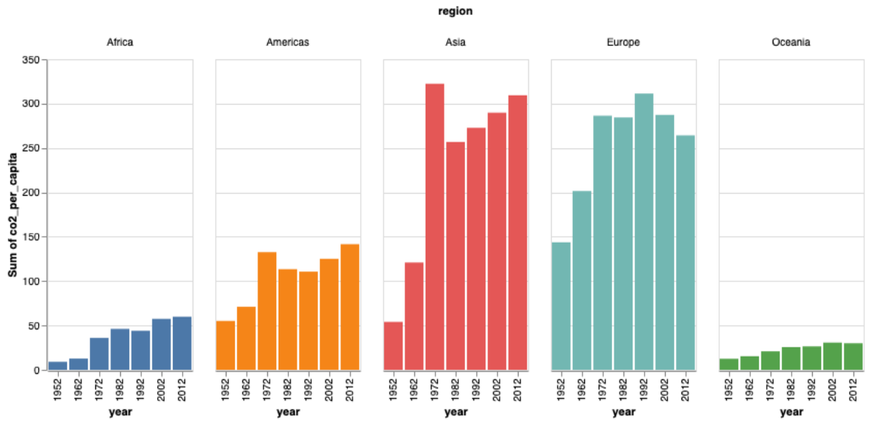
position along aligned scale

after [Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Cleveland and McGill. Journ. American Statistical Association 79:387


region
Africa
Americas
Asia
Europe
Oceania
Sum of co2_per_capita
year
region
Africa
Americas
Asia
Europe
Oceania

Comparing Bar Chart Variants





- **Regular Bar Chart**
  - Single variable across categories
  - Good for: comparing magnitudes between categories
  - Judgement: **length and position along aligned scale** (i.e., a common baseline)
- **Grouped (Clustered) Bar Chart**
  - Multiple variables shown side by side for each category
  - Good for: comparing within and between groups
  - Judgement: **length and position along aligned scale**
- **Stacked Bar Chart**
  - Segments stacked to show total and contribution of parts
  - Good for: seeing overall total + approximate part contributions
  - Judgement: **length**, but harder for internal segments (position along unaligned common scale)
- **Normalized Stacked Bar Chart (100%)**
  - Segments scaled to a common total of 100%
  - Good for: comparing **proportions** across categories
  - Judgement: **position/length**, but only relative proportions are visible, not totals

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

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?

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

Effectiveness

Best

Least

Identity Channels: Categorical Attributes

Spatial region

Color hue

Motion

Shape

- expressiveness
  - match channel and data characteristics
- effectiveness
  - channels differ in accuracy of perception
  - spatial position ranks high for both

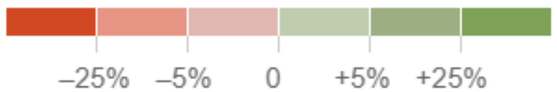
All Spending | Types of Spending | Changes | Department Totals

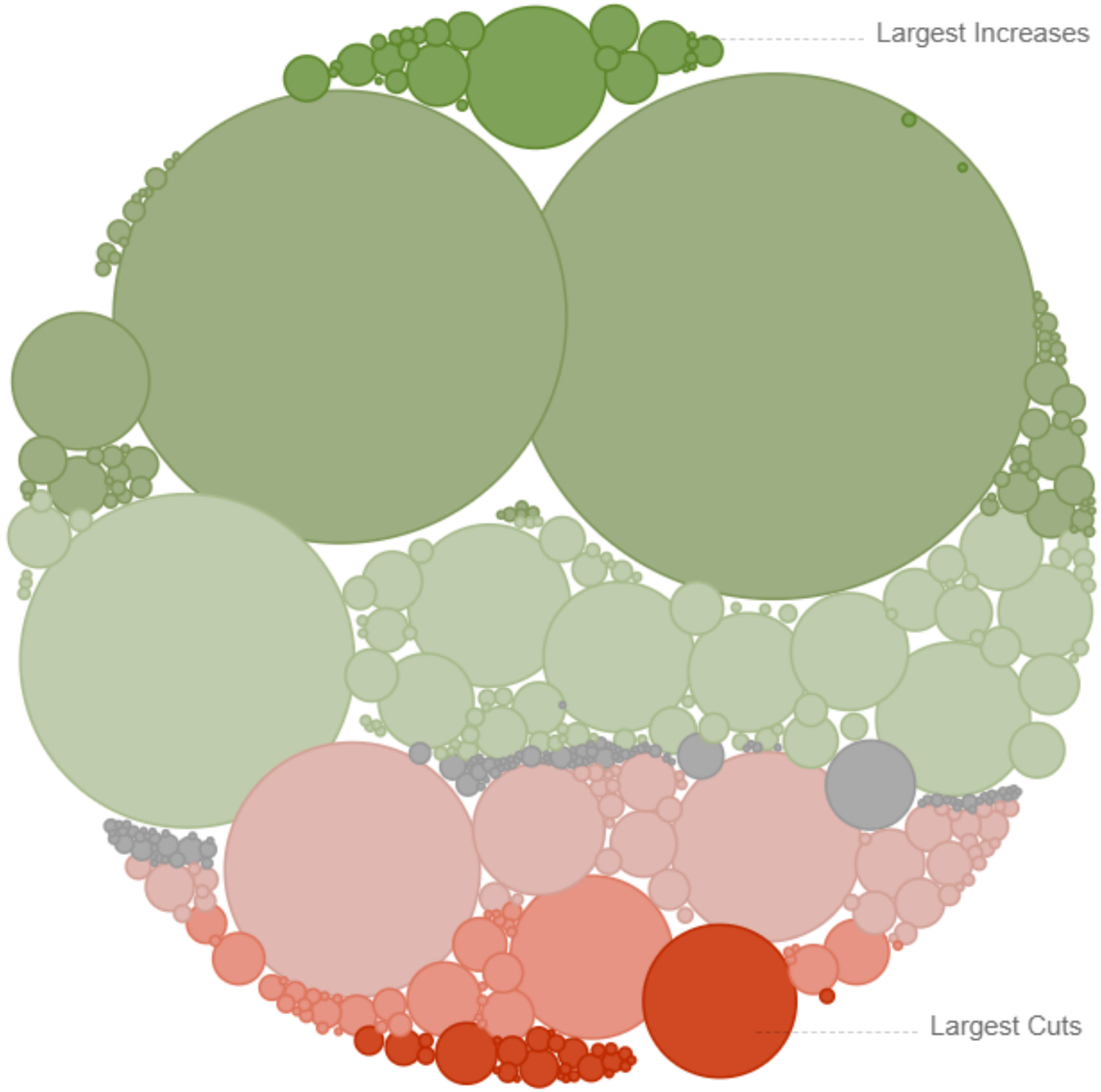
How $3.7 Trillion Is Spent

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.

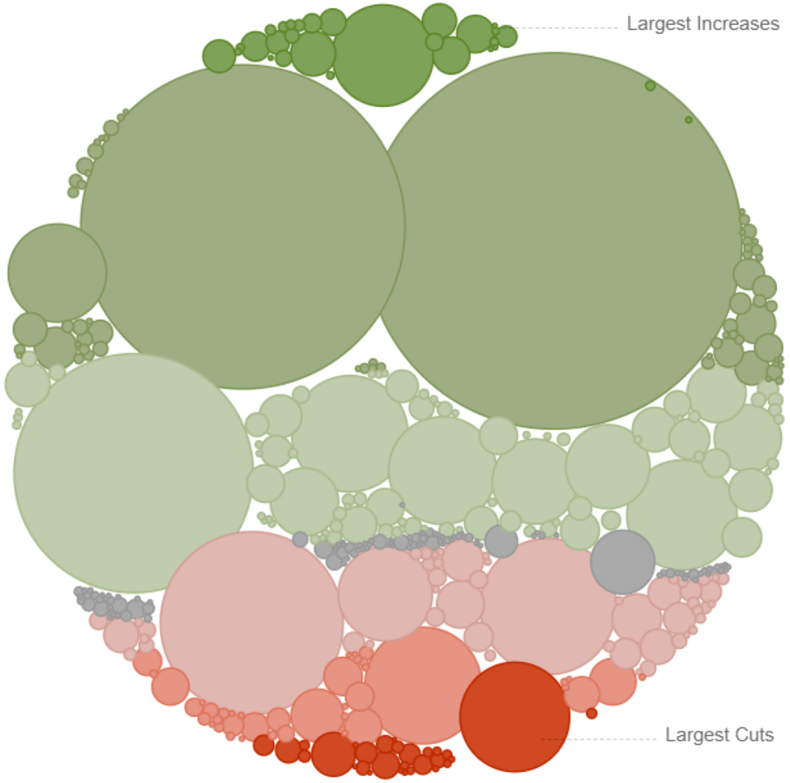




https://archive.nytimes.com/www.nytimes.com/interactive/2012/02/13/us/politics/2013-budget-proposal-graphic.html

For the color channel in the viz discuss the following

- Does it violate the expressiveness or effectiveness principle?
- 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

While reviewing a visualization, you notice that the most critical data attribute is encoded using texture variations. Given the importance of this attribute and the principles of effectiveness and expressiveness, a more suitable encoding might be

A. Color intensity, as it might provide slight differences in perception.
B. Position along a common scale, for clear and distinct differentiation.
C. Shape variations, as all shapes are perceived with equal priority.
D. Size variations of arbitrary symbols, relying on viewer interpretation.

Clicker

A data visualization represents years of experience (an ordered attribute) using different shapes (an identity channel). Why might this design choice be criticized?

A. Shapes are always more engaging than other channels and adhere to the expressiveness principle

B. Shapes, as identity channels, do not inherently convey an order, violating the expressiveness principle when used for ordered data.

C. Using shapes ensures that the visualization is universally understood and keeps with the expressiveness principle

D. Years of experience should always be represented by color hue to capitalize on the expressiveness principle

Clicker

In a chart displaying different brands (a categorical attribute), the designer decided to use length (a magnitude channel) to differentiate them. What potential issue might arise from this design choice?

A. Length is too effective and will overshadow other elements in the visualization.
B. Brands, being categorical, do not have an intrinsic order, and using length might mislead by implying an unwarranted hierarchy or order.
C. Length is universally recognized, and therefore the best choice for any data type.
D. Brands are best represented in 3D visualizations only.

Learning Outcomes

- **Determine** the cardinality of a dataset attribute
- **Explain how cardinality impacts visual encoding choices**
- **Identify** the five characteristics of a visual channel
- **Explore** how channel characteristics influence visualization choices
- **Describe** how a channel is used based on its characteristics
- **Critique** a visualization using the principles of effectiveness and expressiveness

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?

Get Stepping

- Lab this week is on design, come with your sketching paper and colored pens/pencils
- Read the lab overview page on the course website
- Work through Tutorial 3 before class on Wednesday
- Complete Quiz 2 in CBTF which is on Altair Basics
- Take a walk for it is starting to smell rain.

Flying: Above and Beyond

Accuracy study Papers
You do NOT need to read these papers, but if you are interested in the studies

- Heer & Bostock 2010
- Panavas et al. 2022
- Cleveland & McGill 1984

A paper that delves deeper into the theories around perception and visual working memory

https://www.csc2.ncsu.edu/faculty/healey/download/tvcg.12a.pdf