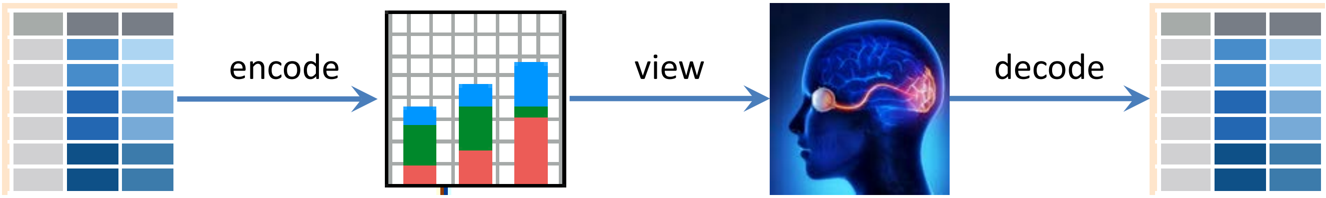


Graphical Perception

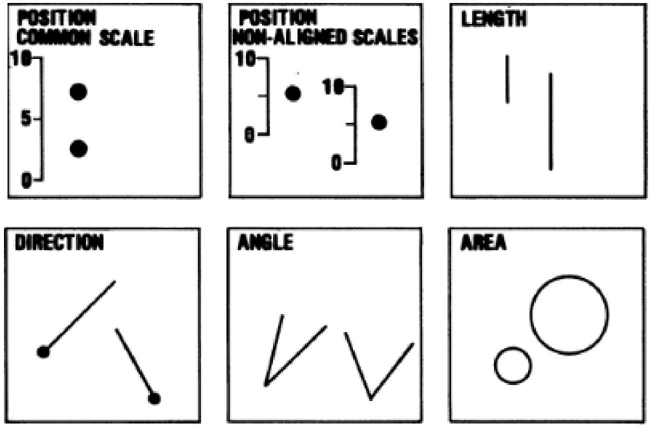
Michael Friendly

Psych 6135

http://euclid.psych.yorku.ca/www/psy6135/

Graphical Perception

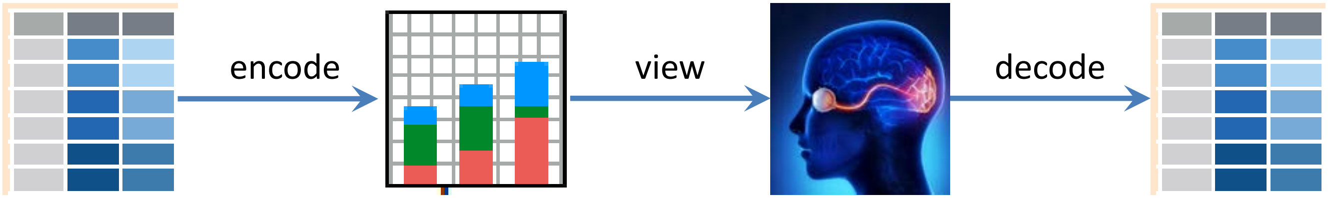
- In constructing a graph, quantitative and categorical information is encoded by visual attributes:
  - Length
  - Position along axis
  - Angle
  - Area
  - Color, shape, line style



- What determines the ability of graph viewers to:
  - Make comparisons (which is larger?)
  - Estimate a magnitude?
  - See patterns, trends, unusual features?

Encoding & decoding

- When we construct a graph, we encode a numerical or categorical variable as a graphical attribute
- When we view a graph, the goal is to decode the graphical attributes and extract information about the data that was encoded



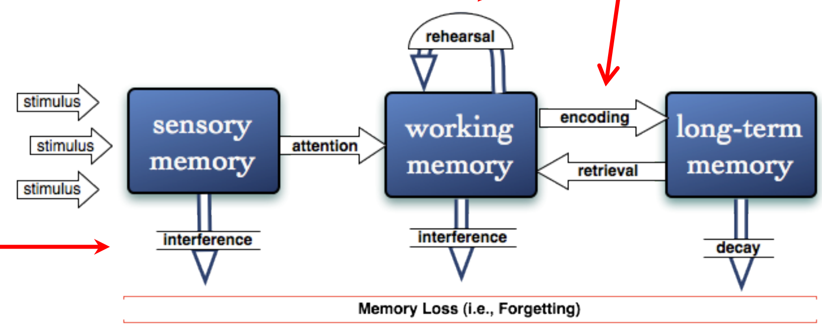
- Encoding should rely on features that can easily be decoded
- Often, easier said than done! The devil is in the details

Visual & cognitive systems

- A simplified model: Three stages

How to preserve info?

How is info lost?

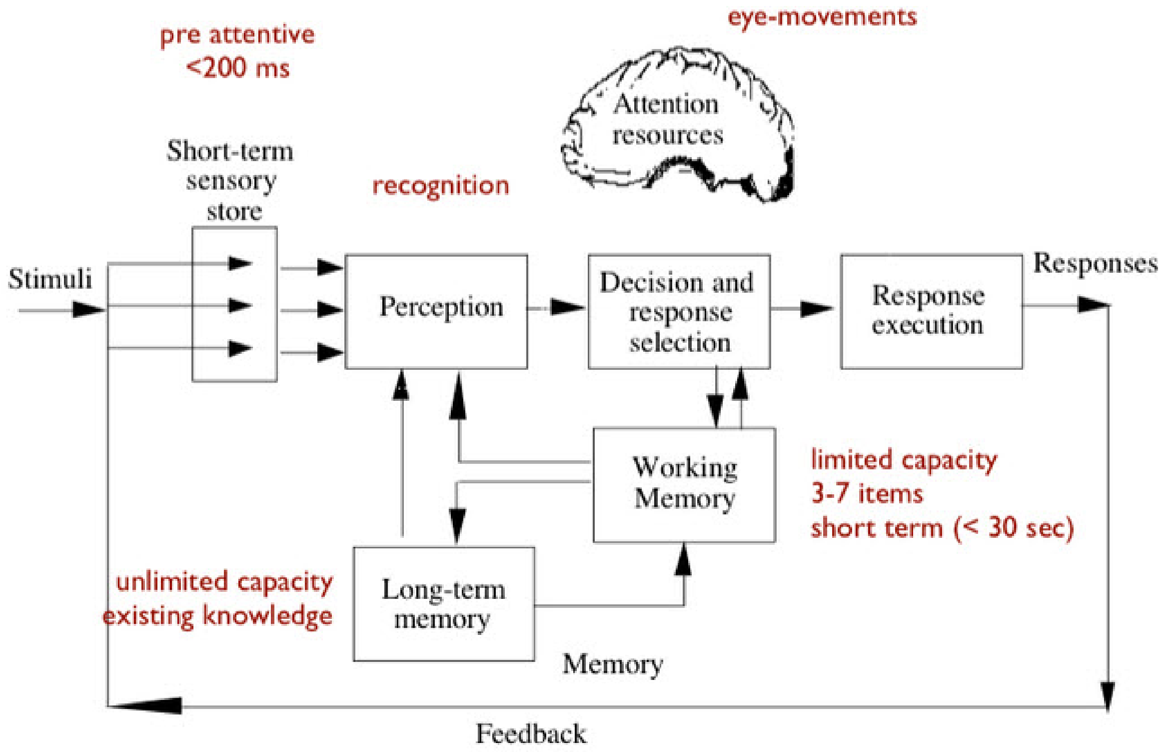


- Sensory (iconic) memory
  - pre-attentive, automatic, feature detection
  - massively parallel, short duration, easily fooled
- Working memory
  - requires attention, limited capacity (~ 4-6 "chunks")
- Long-term memory
  - real-world knowledge, unlimited capacity

Perception & cognition

Another coarse distinction:

- **Perception**: Processing of the signals coming in: what you "see"
- **Cognition**: How you understand and interpret what you see

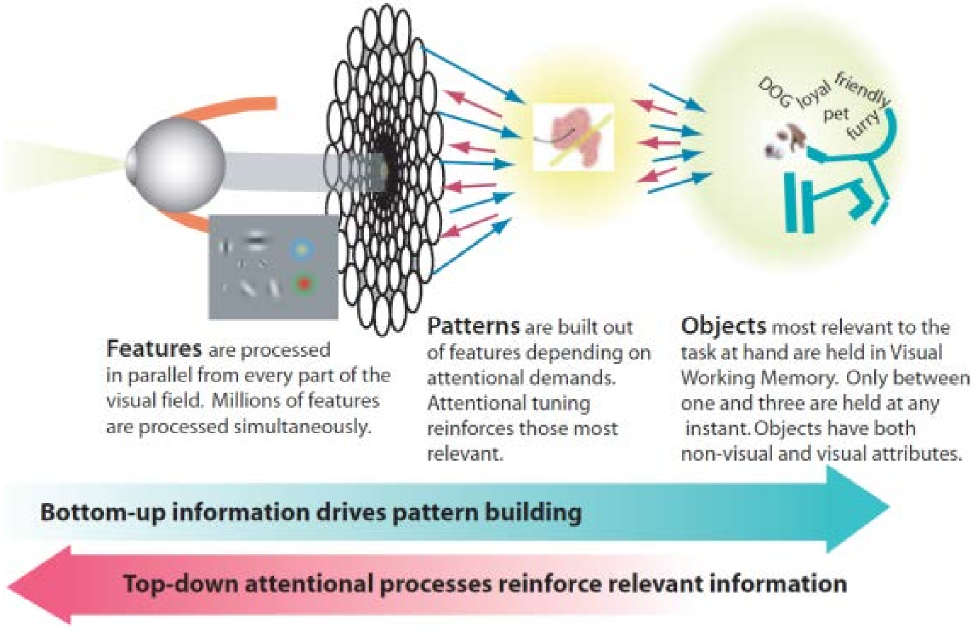


A nice scientific or textbook diagram
But where is cognition?

Based on slide from G. Grinstein

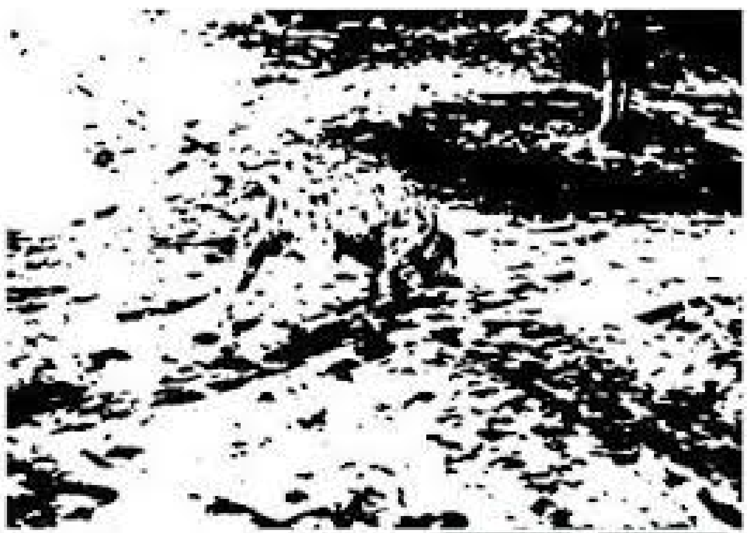
Perception: Bottom-up & Top-down

- Bottom-up processing
  - Low level: features $\rightarrow$ pattern $\rightarrow$ object
  - Detect edges, contours, color, motion
- Top-down processing
  - Driven by goals, expectations
  - Uses prior knowledge, experience, filters what we "see"

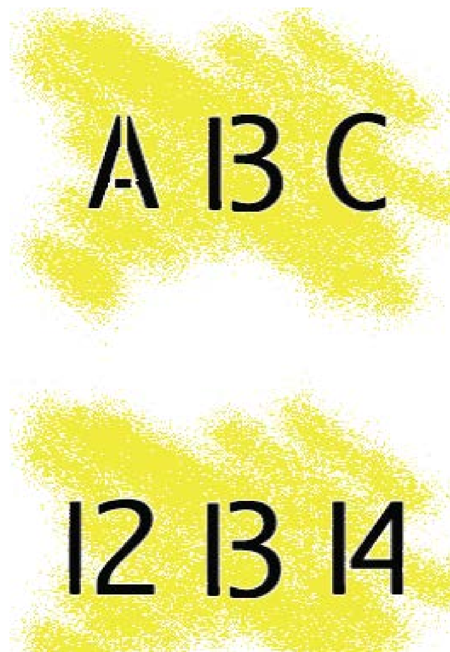


Perception: Top-down

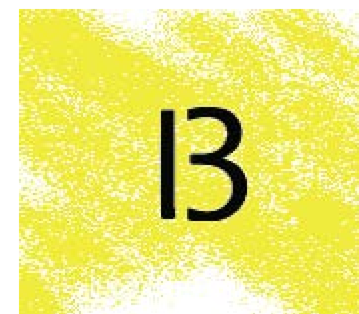
What is in this scene?



What is the middle character?



What here?



An ambiguous figure!

What is the middle letter in each word?

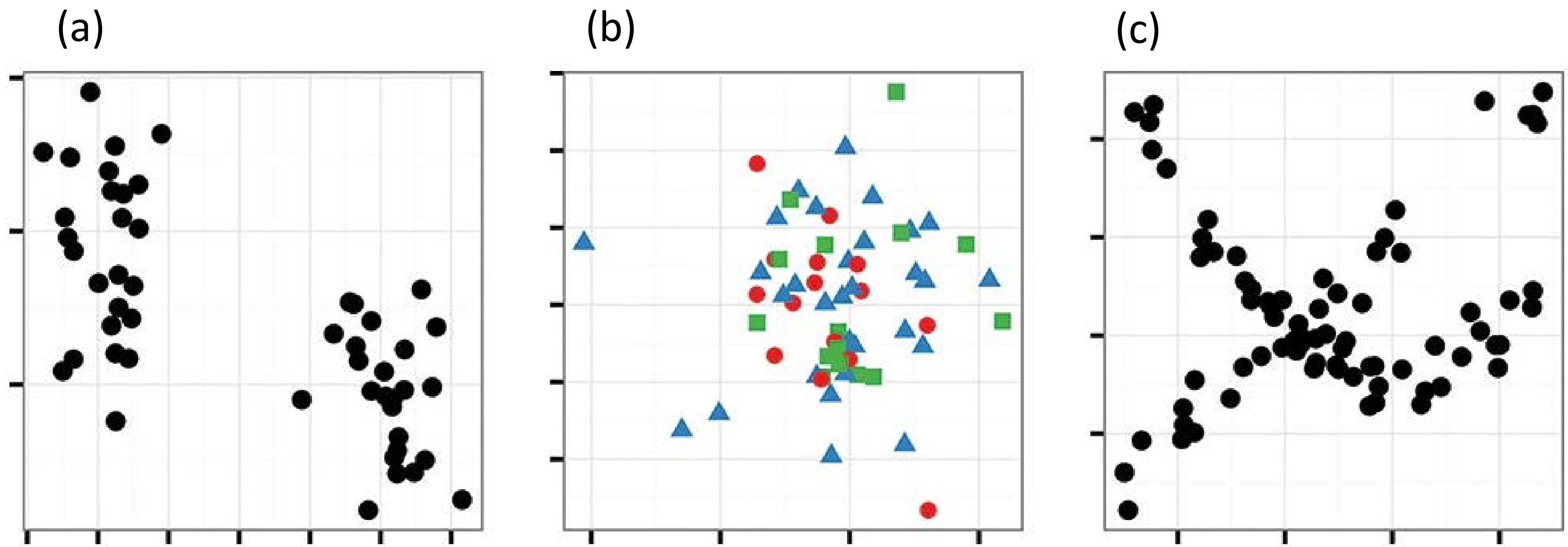
All of these are demonstrations of the role of expectations (top-down) in determining what we “see”

Gestalt principles

- Perception as top-down process governed by holistic principles. “Gestalt” = “form”
  - proximity: elements close together likely to belong to the same unit
  - similarity: more common visual elements increases belonging together
  - good continuation: elements that blend together are likely in the same unit
  - common region: elements in the same region likely belong together

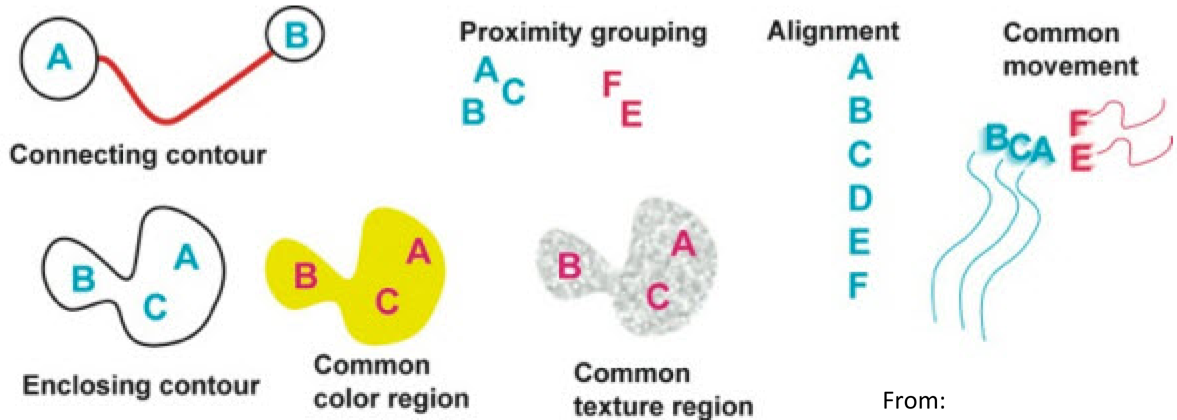
Gestalt principles

- (a) proximity creates impression of 2 groups
- (b) similarity: 3 groups via color & shape
- (c) good continuation gives impression of 2 groups



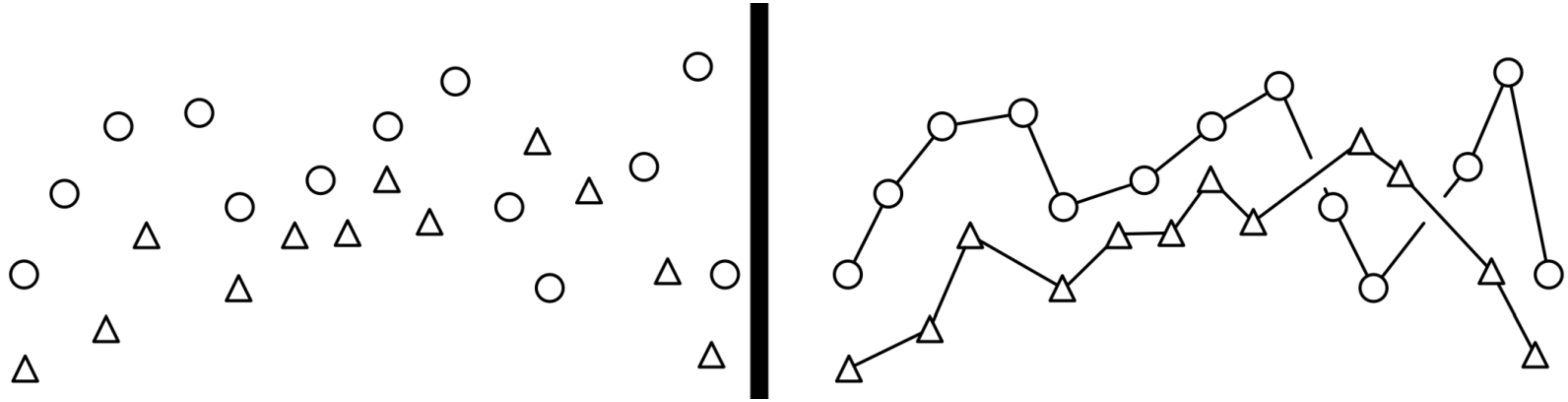
Gestalt principles

More gestalt ideas



From:
http://blog.yhwu.me/notes/visualizations/cs171.html

Why lines are good in time series graphs



Perception: Bottom-up

How many 5s in this display?

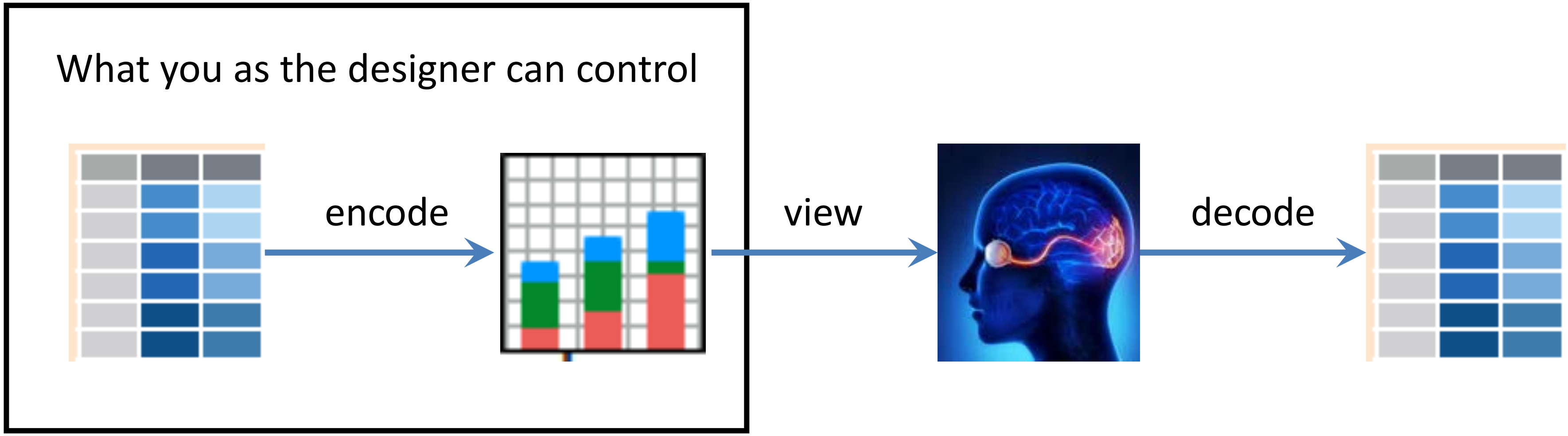
1561321203658413076510374627
4173127527327592732990709742
1703707774179527931749270973
4019743217909370945179279417

How many 5s in this display?

1561321203658413076510374627
4173127527327592732990709742
1703707774179527931749270973
4019743217909370945179279417

Numerals differ only in shape, and are high-level symbols
You have to literally scan them **all** & count the 5s.
The distinction of color is immediate & **pre-attentive**
You only have to scan & count the 5s.

This is why color is an important visual attribute for a categorical variable in graphs


What you as the designer can control
encode
view
decode

Encoding: All data visualizations map data values into quantifiable features (variables) of the resulting graphic. We refer to these features (variables) as ***aesthetics*** (sometimes called *channels*).

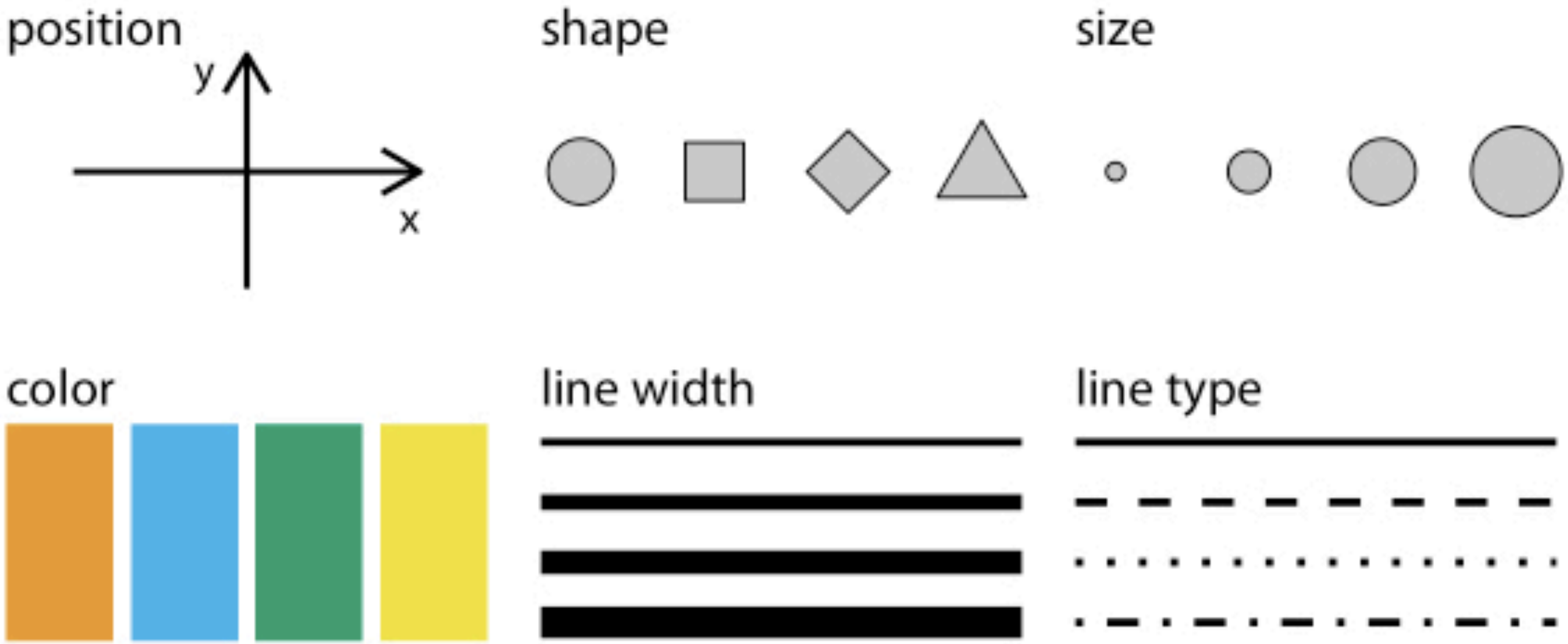


Figure 2.1: Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can usually only represent discrete data (shape, line type).

Wilke, Chapter 2

Table 2.1: Types of variables encountered in typical data visualization scenarios.

Type of variable	Examples	Appropriate scale	Description
quantitative/numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}	continuous	Arbitrary numerical values. These can be integers, rational numbers, or real numbers.
quantitative/numerical discrete	1, 2, 3, 4	discrete	Numbers in discrete units. These are most commonly but not necessarily integers. For example, the numbers 0.5, 1.0, 1.5 could also be treated as discrete if intermediate values cannot exist in the given dataset.
qualitative/categorical unordered	dog, cat, fish	discrete	Categories without order. These are discrete and unique categories that have no inherent order. These variables are also called *factors*.
qualitative/categorical ordered	good, fair, poor	discrete	Categories with order. These are discrete and unique categories with an order. For example, "fair" always lies between "good" and "poor". These variables are also called *ordered factors*.
date or time	Jan. 5 2018, 8:03am	continuous or discrete	Specific days and/or times. Also generic dates, such as July 4 or Dec. 25 (without year).
text	The quick brown fox jumps over the lazy dog.	none, or discrete	Free-form text. Can be treated as categorical if needed.

Wilke, Chapter 2

Match variable type with each dataset variable.

Type of variable	Examples
quantitative/numerical continuous	1.3, 5.7, 83, $1.5x10^{-2}$
quantitative/numerical discrete	1, 2, 3, 4
qualitative/categorical unordered	dog, cat, fish
qualitative/categorical ordered	good, fair, poor
date or time	Jan. 5 2018, 8:03am
text	The quick brown fox jumps over the lazy dog.

Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2
Jan	3	Chicago	USW00014819	25.3
Jan	3	San Diego	USW00093107	55.3
Jan	3	Death Valley	USC00042319	51.3
Jan	3	Houston	USW00012918	53.8

?	?	?	?	?

Wilke, Chapter 2

Match variable type with each dataset variable.

Type of variable	Examples
quantitative/numerical continuous	1.3, 5.7, 83, 1.5×10^{-2}
quantitative/numerical discrete	1, 2, 3, 4
qualitative/categorical unordered	dog, cat, fish
qualitative/categorical ordered	good, fair, poor
date or time	Jan. 5 2018, 8:03am
text	The quick brown fox jumps over the lazy dog.

Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2
Jan	3	Chicago	USW00014819	25.3
Jan	3	San Diego	USW00093107	55.3
Jan	3	Death Valley	USC00042319	51.3
Jan	3	Houston	USW00012918	53.8

Categorical Ordered	Numerical Discrete	Categorical Unordered	Categorical Unordered	Numerical Continuous

Wilke, Chapter 2

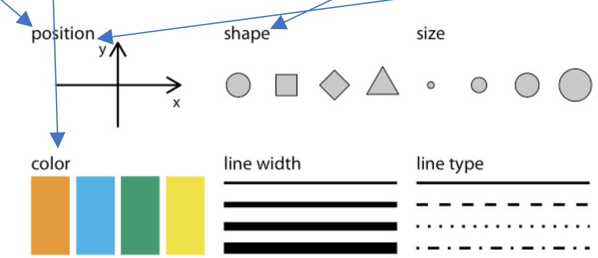
What is **encoding**? **Mapping** data variables onto aesthetics

Table 2.2: First 12 rows of a dataset listing daily temperature normals for four weather stations. Data source: NOAA.

Month	Day	Location	Station ID	Temperature
Jan	1	Chicago	USW00014819	25.6
Jan	1	San Diego	USW00093107	55.2
Jan	1	Houston	USW00012918	53.9
Jan	1	Death Valley	USC00042319	51.0
Jan	2	Chicago	USW00014819	25.5
Jan	2	San Diego	USW00093107	55.3
Jan	2	Houston	USW00012918	53.8
Jan	2	Death Valley	USC00042319	51.2
Jan	3	Chicago	USW00014819	25.3
Jan	3	San Diego	USW00093107	55.3
Jan	3	Death Valley	USC00042319	51.3
Jan	3	Houston	USW00012918	53.8

Categorical Ordered	Numerical Discrete	Categorical Unordered	Categorical Unordered	Numerical Continuous

After defining variables, the next goal is "**mapping**" variables onto aesthetics.

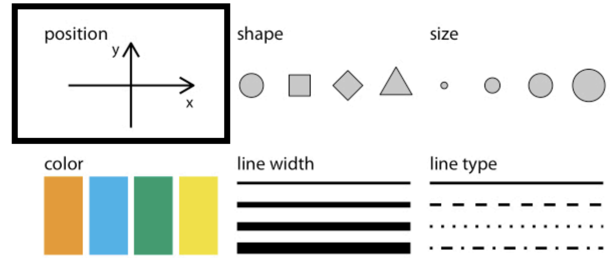


This is your job as a Data Vis **designer**.

Wilke, Chapter 2

“Position” aesthetics are critical as they’re (typically) required for any graph.

Position is encoded in **coordinate systems**.



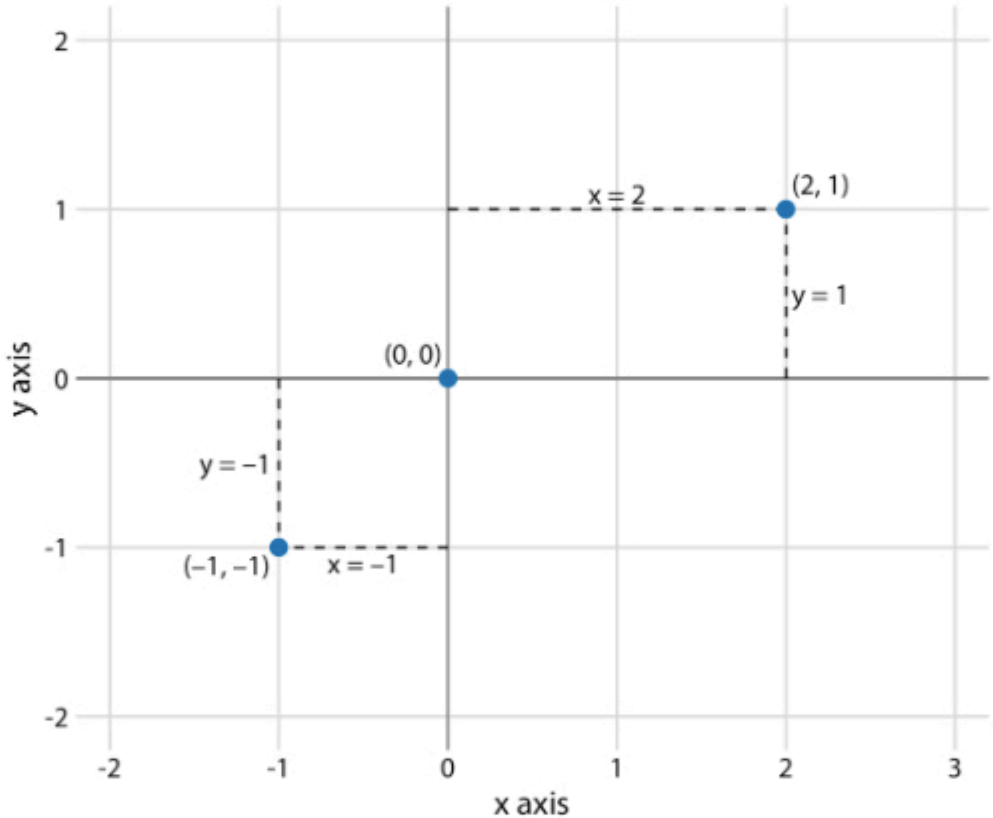
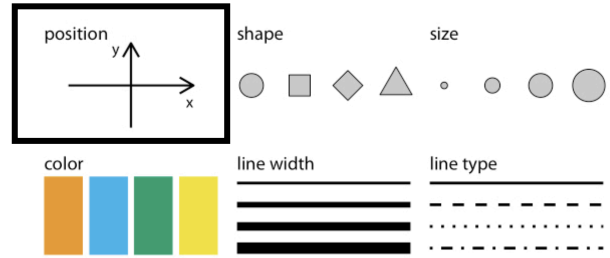


Figure 3.1: Standard cartesian coordinate system. The horizontal axis is conventionally called *x* and the vertical axis *y*. The two axes form a grid with equidistant spacing. Here, both the *x* and the *y* grid lines are separated by units of one. The point (2, 1) is located two *x* units to the right and one *y* unit above the origin (0, 0). The point (-1, -1) is located one *x* unit to the left and one *y* unit below the origin.

Wilke, Chapter 3

"Position" aesthetics are critical as they're (typically) required for any graph.



Another example is **polar coordinates**.

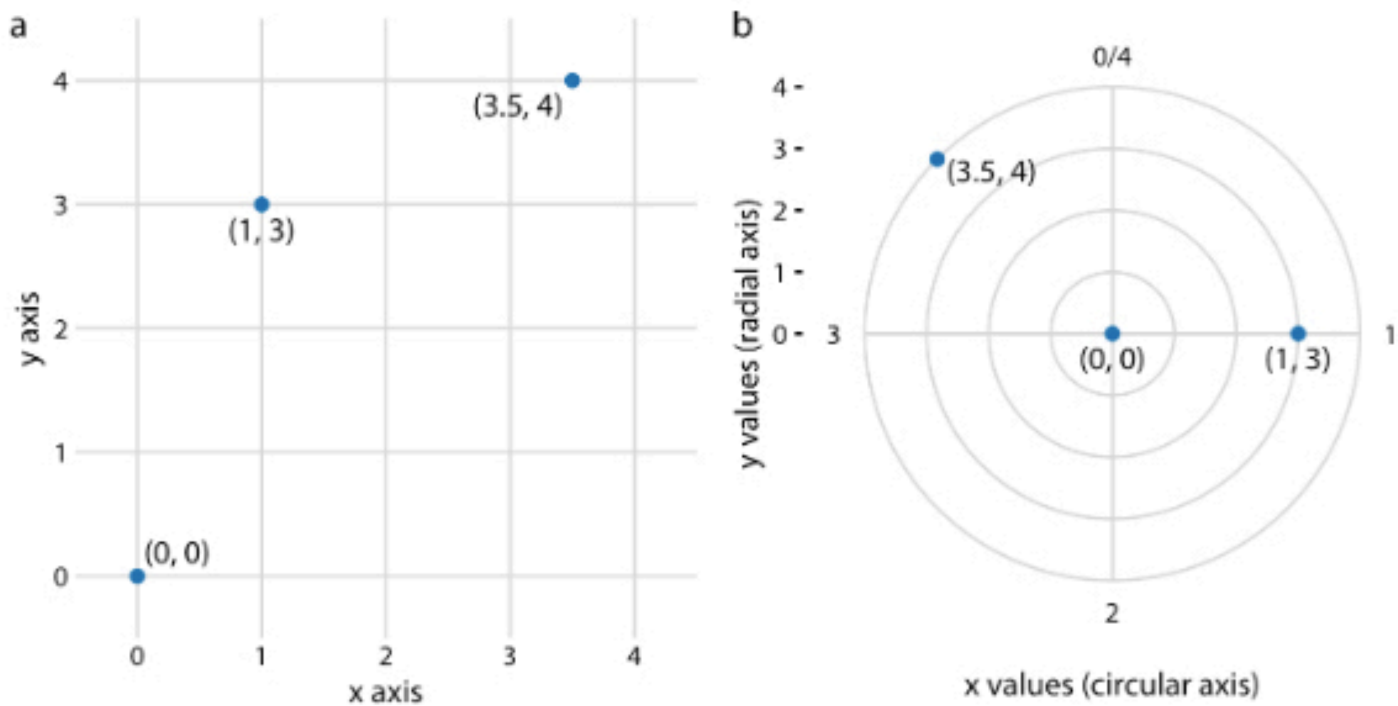


Figure 3.9: Relationship between Cartesian and polar coordinates. (a) Three data points shown in a Cartesian coordinate system. (b) The same three data points shown in a polar coordinate system. We have taken the *x* coordinates from part (a) and used them as angular coordinates and the *y* coordinates from part (a) and used them as radial coordinates. The circular axis runs from 0 to 4 in this example, and therefore $x = 0$ and $x = 4$ are the same locations in this coordinate system.

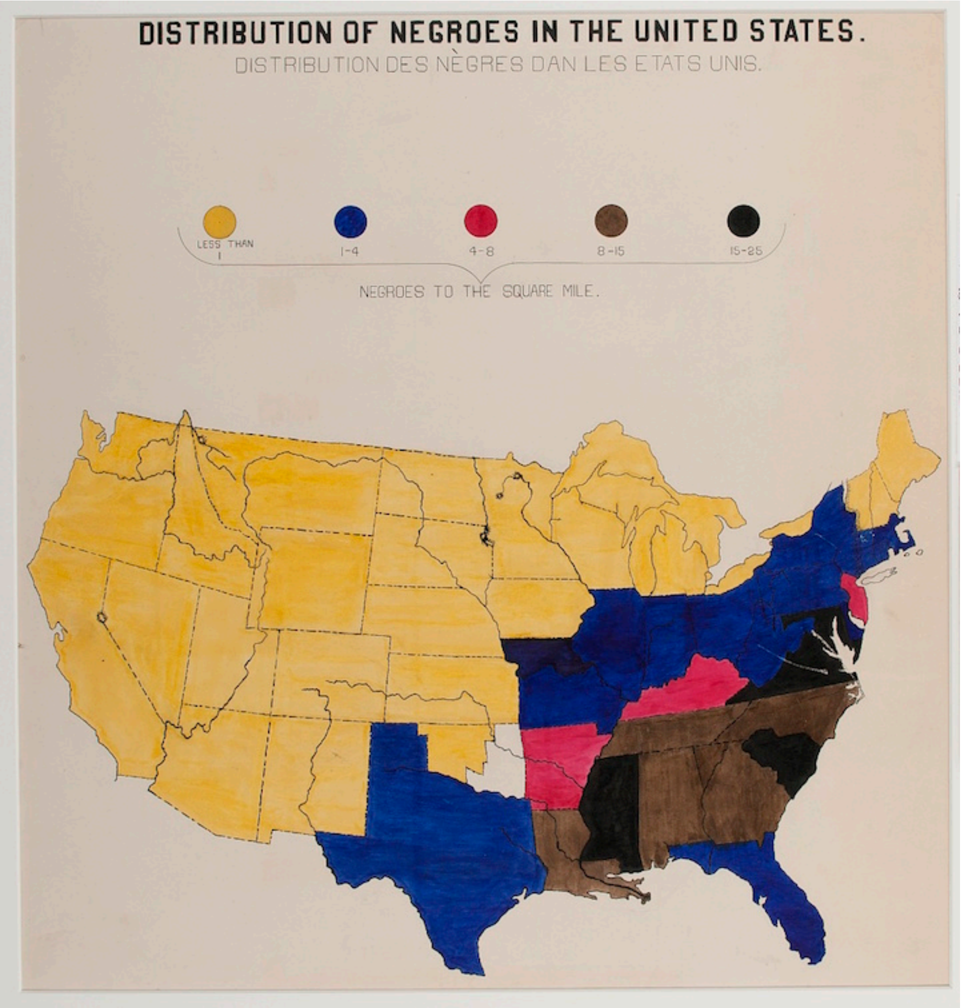
Wilke, Chapter 3

Nice graphic, naïve about color

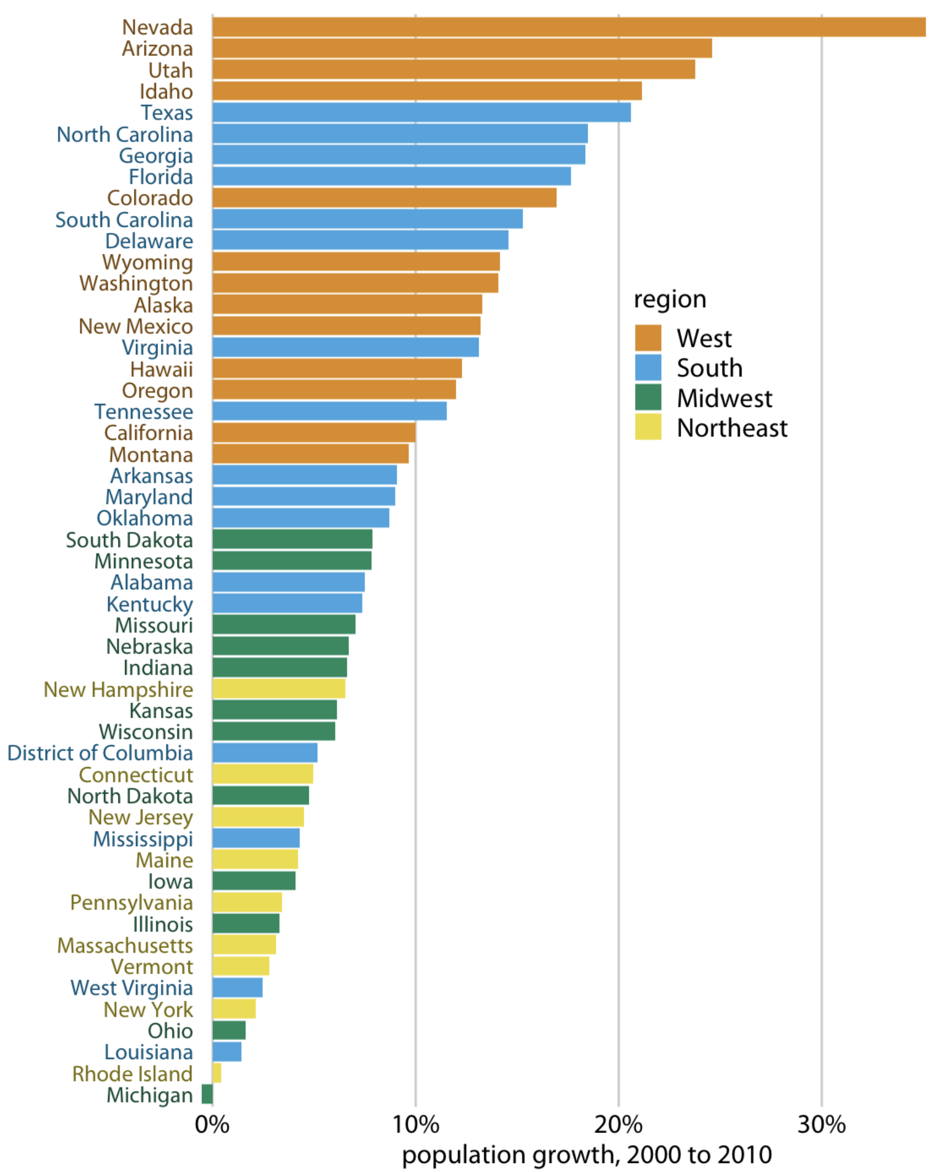
W.E.B. Du Bois presented this as part of an exhibition on The American Negro at the 1900 Paris Exposition.

It is a landmark graphic, but shows no understanding of the use of color for a quantitative variable.

Q: Are there more Negroes per sq. mile in Texas or Louisiana?



Color as a tool to distinguish: Categories



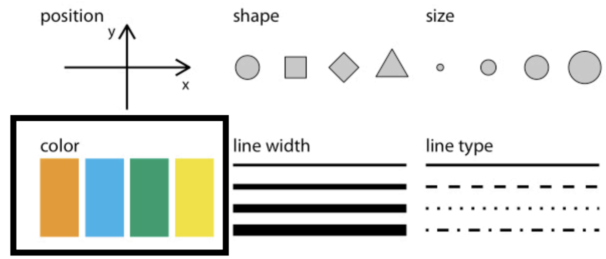
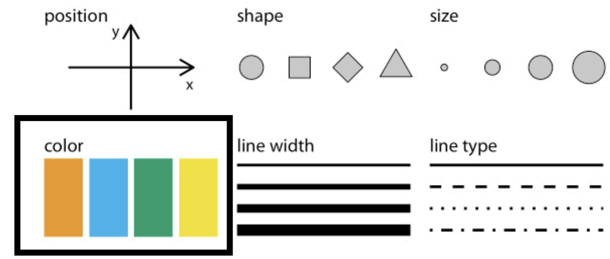


Figure 4.1: Example qualitative color scales. The Okabe Ito scale is the default scale used throughout this book (Okabe and Ito 2008). The ColorBrewer Dark2 scale is provided by the ColorBrewer project (Brewer 2017). The ggplot2 hue scale is the default qualitative scale in the widely used plotting software ggplot2.

Wilke, Chapter 4

Color to represent data values: Numerical



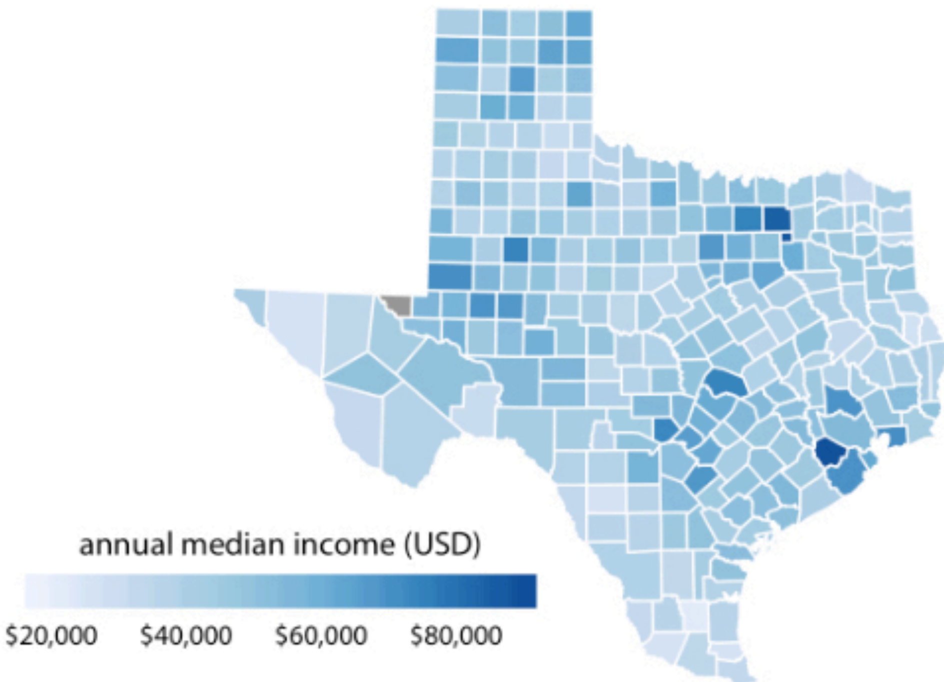
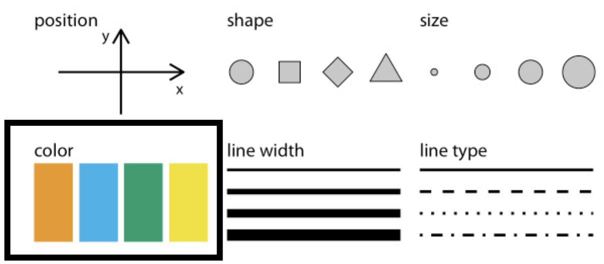


Figure 4.4: Median annual income in Texas counties. The highest median incomes are seen in major Texas metropolitan areas, in particular near Houston and Dallas. No median income estimate is available for Loving County in West Texas and therefore that county is shown in gray. Data source: 2015 Five-Year American Community Survey

Figure 4.3: Example sequential color scales. The ColorBrewer Blues scale is a monochromatic scale that varies from dark to light blue. The Heat and Viridis scales are multi-hue scales that vary from dark red to light yellow and from dark blue via green to light yellow, respectively.

Wilke, Chapter 4

Color to represent data values: Numerical



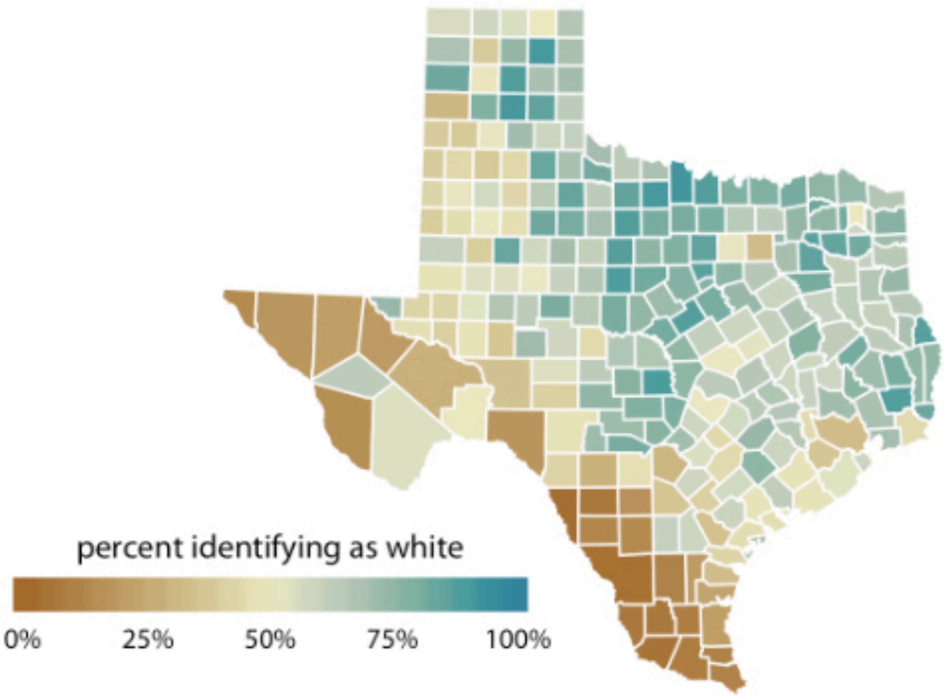


Figure 4.6: Percentage of people identifying as white in Texas counties. Whites are in the majority in North and East Texas but not in South or West Texas. Data source: 2010 Decennial U.S. Census

Divergent scales

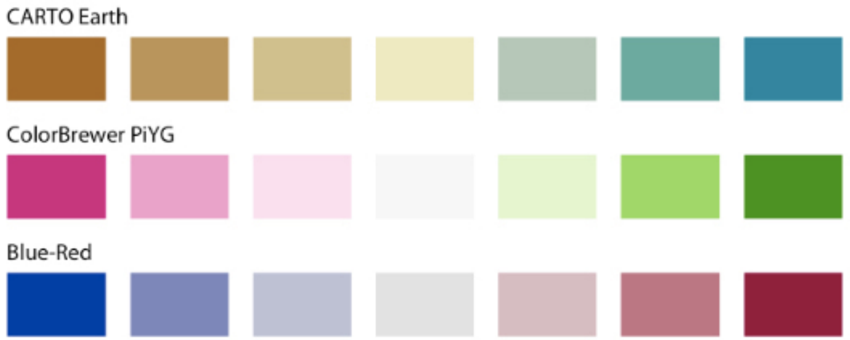
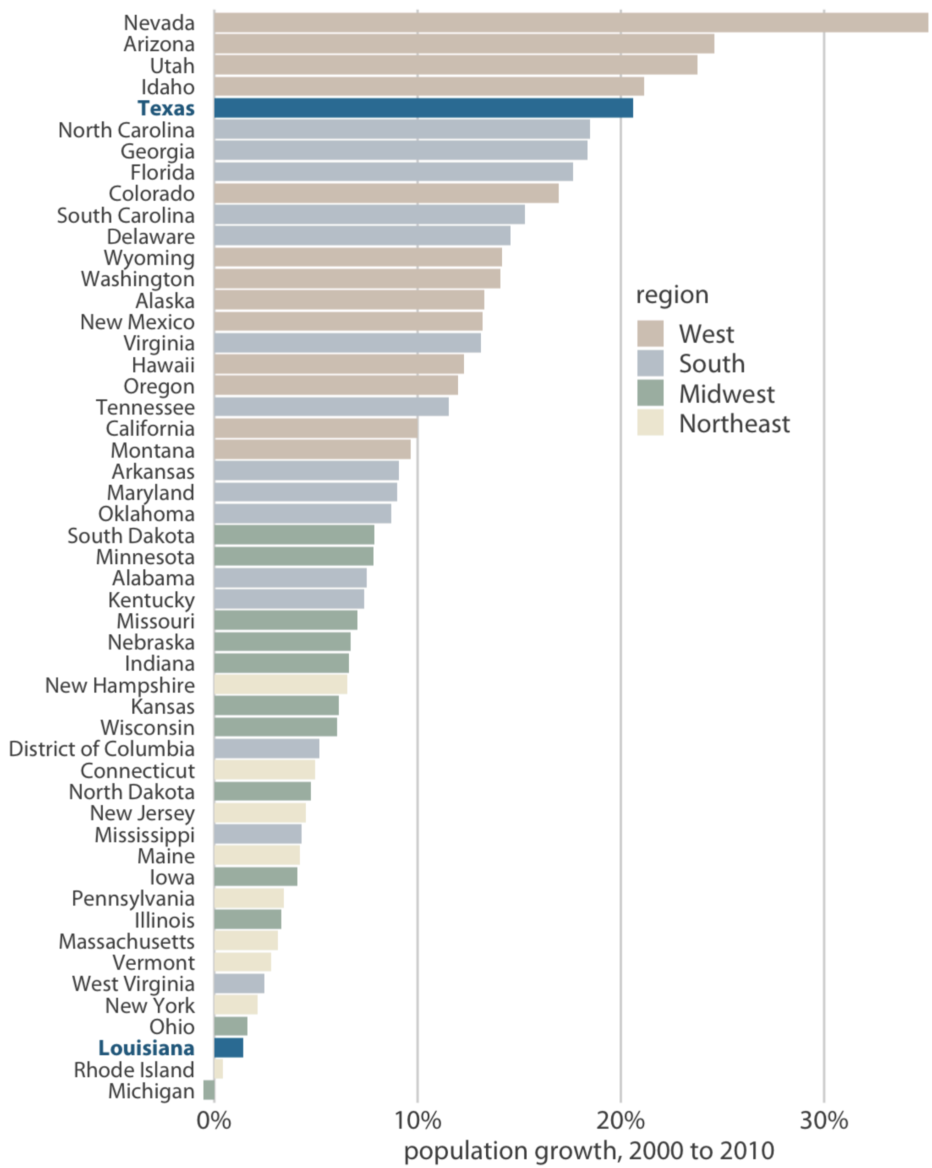


Figure 4.5: Example diverging color scales. Diverging scales can be thought of as two sequential scales stiched together at a common midpoint color. Common color choices for diverging scales include brown to greenish blue, pink to yellow-green, and blue to red.

In some cases, we need to visualize the deviation of data values in one of two directions relative to a **neutral midpoint**.

Wilke, Chapter 4

Color as a tool to highlight



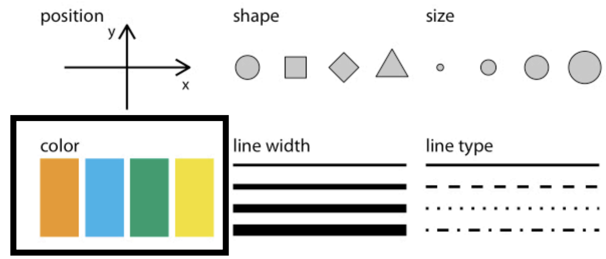


Figure 4.7: Example accent color scales, each with four base colors and three accent colors. Accent color scales can be derived in several different ways: (top) we can take an existing color scale (e.g., the Okabe Ito scale, Fig 4.1) and lighten and/or partially desaturate some colors while darkening others; (middle) we can take gray values and pair them with colors; (bottom) we can use an existing accent color scale, e.g. the one from the ColorBrewer project.

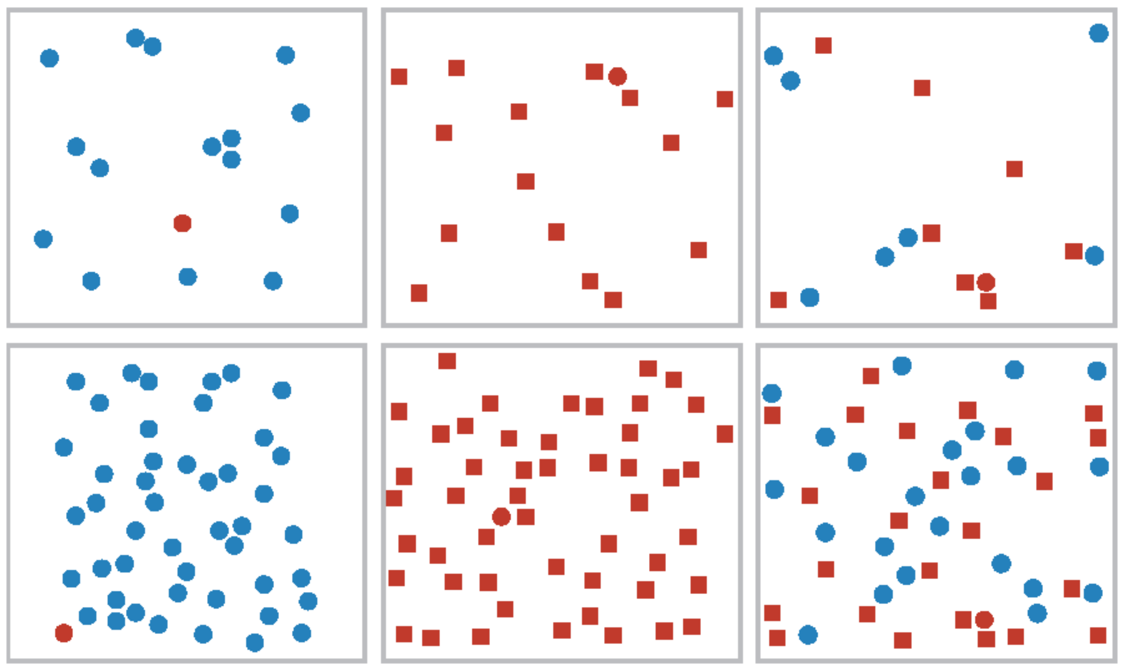
Wilke, Chapter 4

Anomaly detection

Find the red dot in each of the following displays

- This task is easiest when all the rest are blue dots
- Next easiest when only shape distinguishes the red dot
- Hardest when both color and shape vary

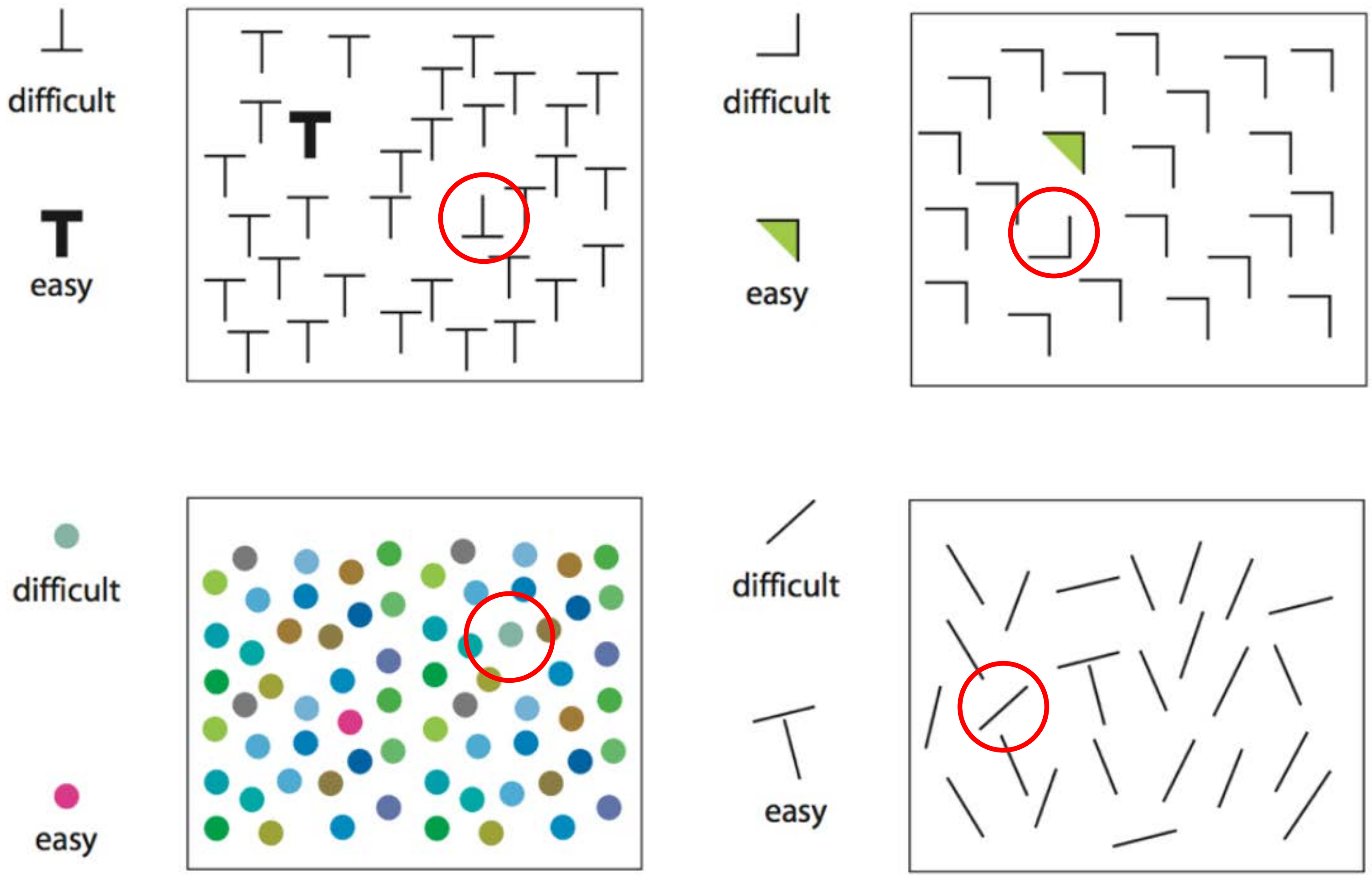
Sometimes called
“popout” effect.
Not a good term.



Anomaly detection

For each display, find the anomaly shown at the left

Color and shape: What is easy or hard depends on the background



Encodings: Types & ranks

Based on this, Munzner (2015) proposes a ranking of visual attributes for ordered & categorical variables in data displays

These hold when the task is to estimate a magnitude.

A different ranking may occur for other graph-based tasks.

angle (pie charts) – good for % of total judgments

color (mosaic plots) – good for pattern perception

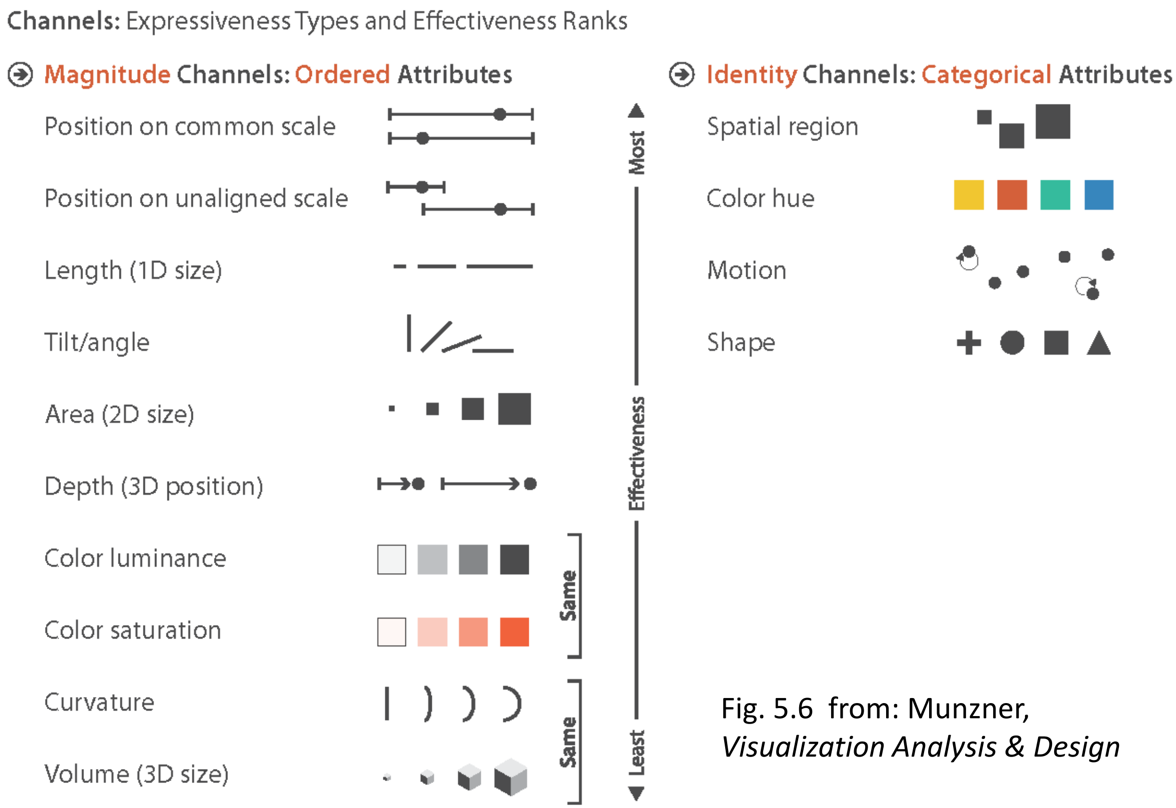
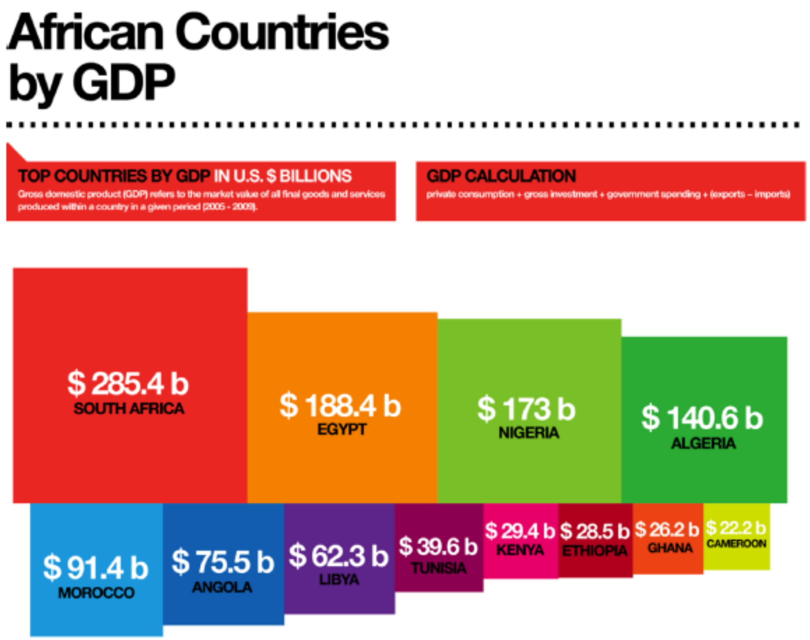
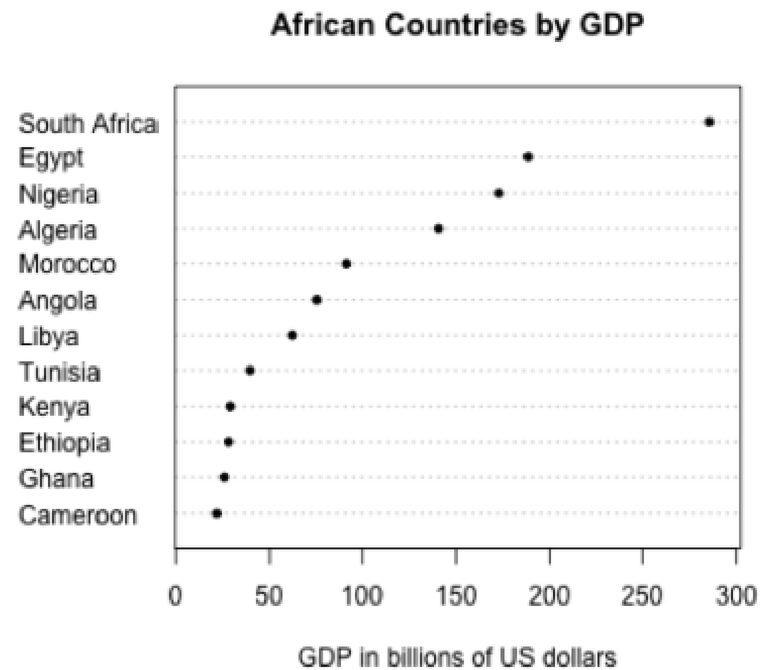


Fig. 5.6 from: Munzner, *Visualization Analysis & Design*

Area vs. length judgments

How much larger is South Africa than Egypt?





Judgments here based on area

Judgment here based on position along a scale

Magnitude estimation

How large are transport accidents?
How much bigger than non-transport accidents?

transport

non-transport

Estimation of length or ratios of length are more accurate than the same judgments of area.

Accuracy: Experimental evidence

Cleveland & McGill (1984) and later Heer & Bostock (2010) carried out experiments to assess the relative accuracy of magnitude judgments for different visual encodings

The task here is to estimate the %age of the smaller highlighted portion.

The details of these studies are interesting & important – more next week

The graph of these results is a great model for data display

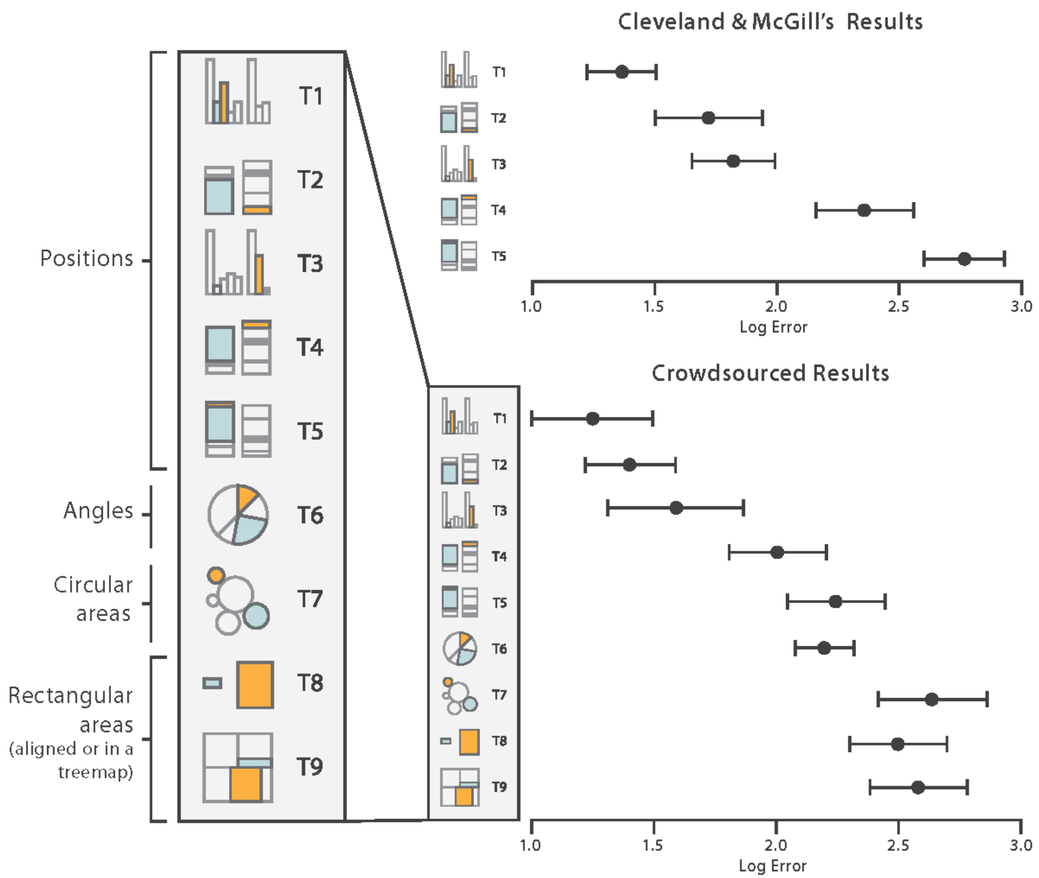
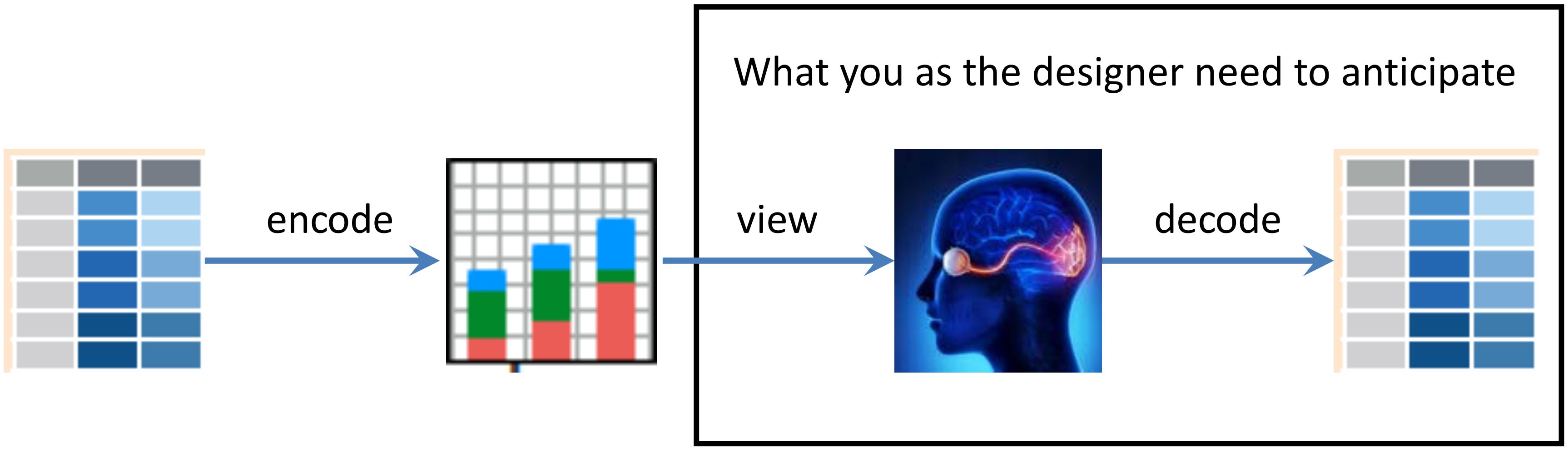


Fig. 5.8 from: Munzner, *Visualization Analysis & Design*

Encodings: Lessons

- Best to show quantitative variables with position or length
- Bar charts:
  - Best encoding via length → start at 0
  - Avoid stacked bars (not aligned), where possible
- Dot charts:
  - Best encoding via position along a scale → start at 0
- Frequency data:
  - area/color encoding to show patterns
  - sqrt or log scale often useful to show magnitude
- Color: choose sensibly ordered hues or saturation
- Arrangement
  - make comparisons easier by placing things to be compared nearby

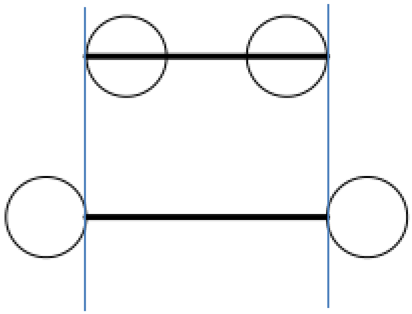


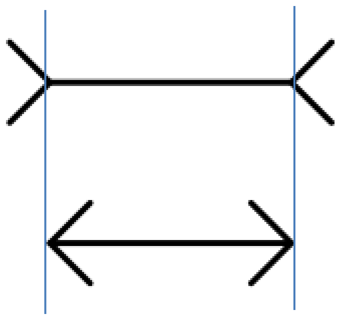
And where the human mind can be fooled and misinterpret your data visualization…

Illusions: Length

Surrounding context matters in judging the size of objects.

Which line is longer? Or are they the same?



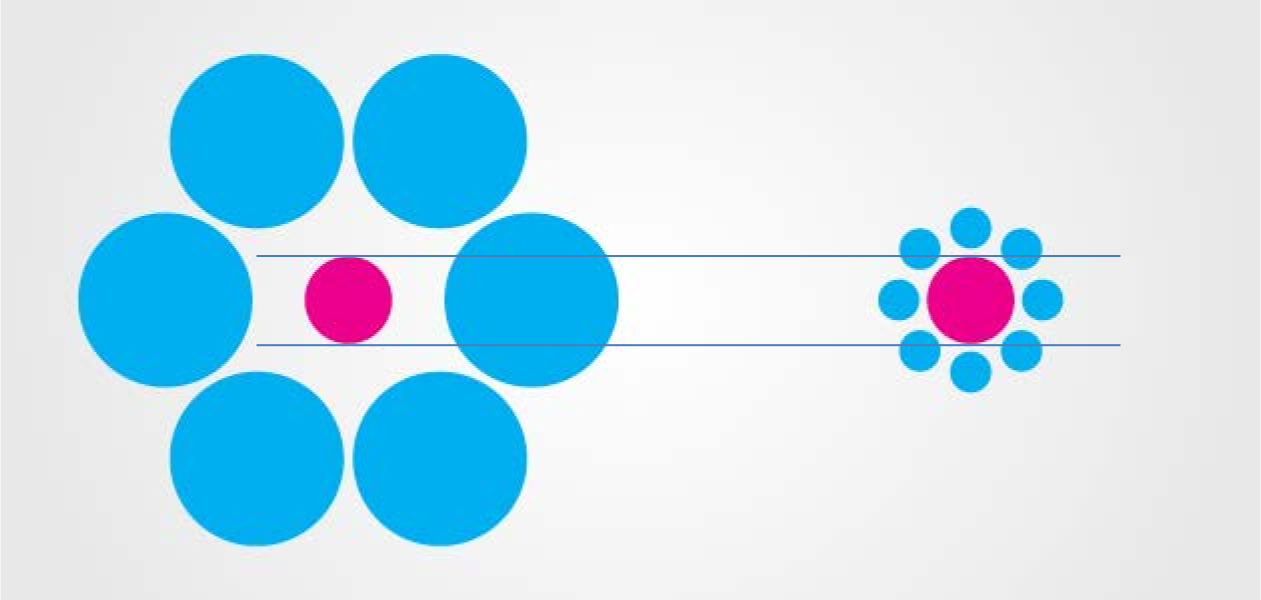


Surrounding context pulls perception of length in its direction
This is the famous Müller-Lyre illusion

Illusions: Area

Surrounding context matters in judging the size of objects.

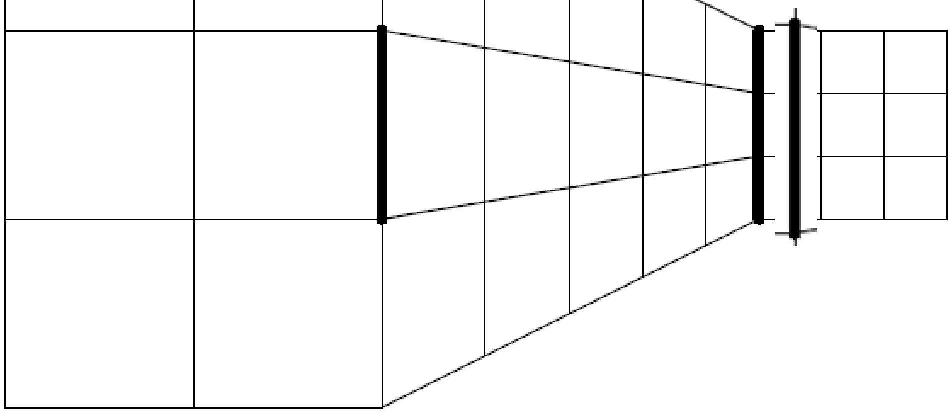
Which red circle is larger? Or are they the same?



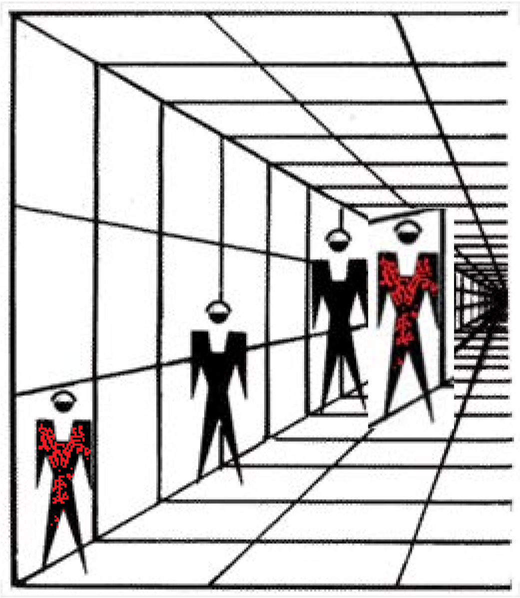
Surrounding context pulls perception of area against the background
This is often called the Ebbinghaus illusion or the Tichener illusion

Illusions: Perspective

Which **thick** line is longer? Or, both the same?

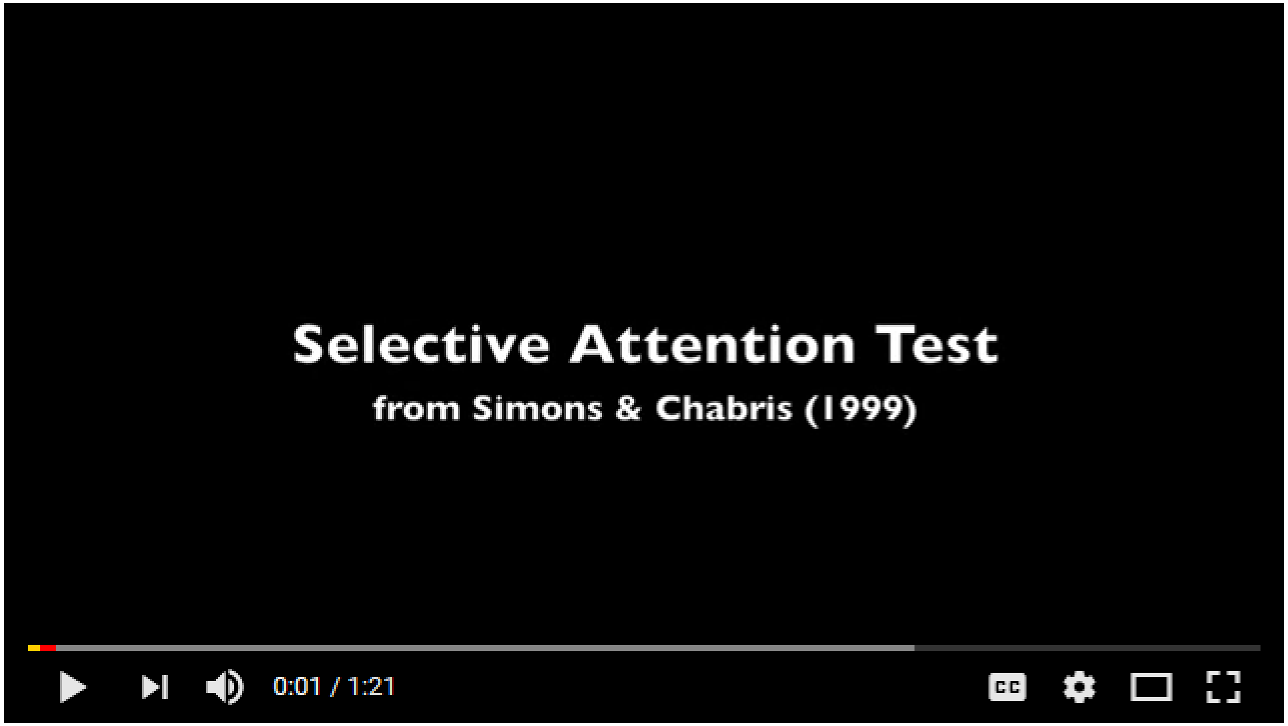


Which figure is tallest?
Or, all the same?



This is often called the Ponzo illusion: We judge the size of real-world objects relative to their background.

Selective attention



https://www.youtube.com/watch?v=vJG698U2Mvo

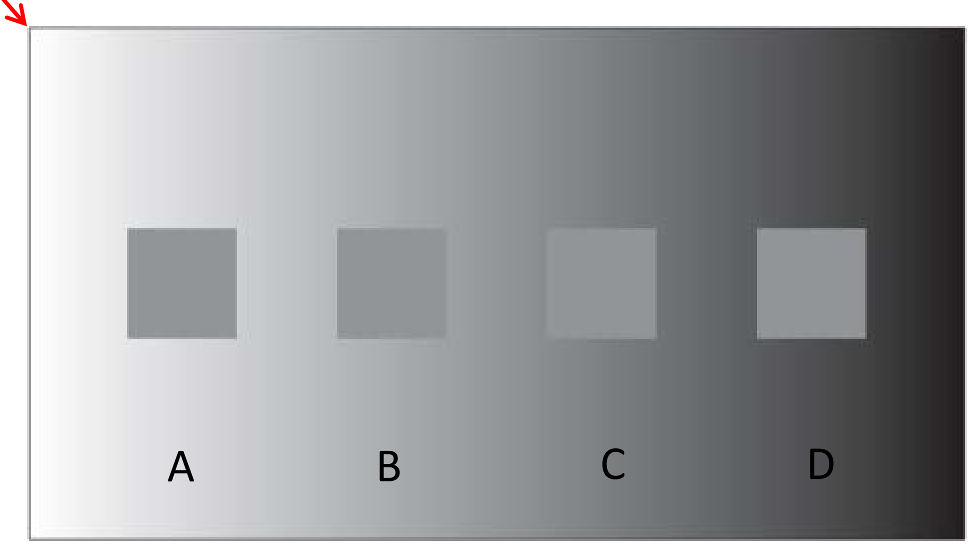
Perception: Contrast

Color perception, even of gray, is influenced by contrast against a background

Q: Which gray square at right is most similar to that at the left?

A: it is the same gray square against a changing background

gray square



Most people say **A**, because it is shown on a white background

Luminance contrast

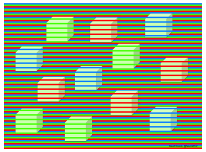
Showing blue text on a black background doesn't work very well. There is insufficient luminance contrast.

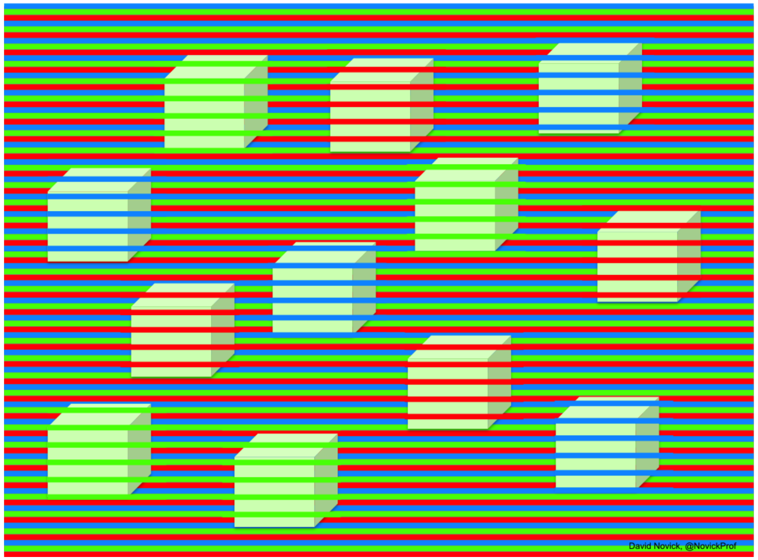
Showing blue text on a white background works better. There is sufficient luminance contrast.

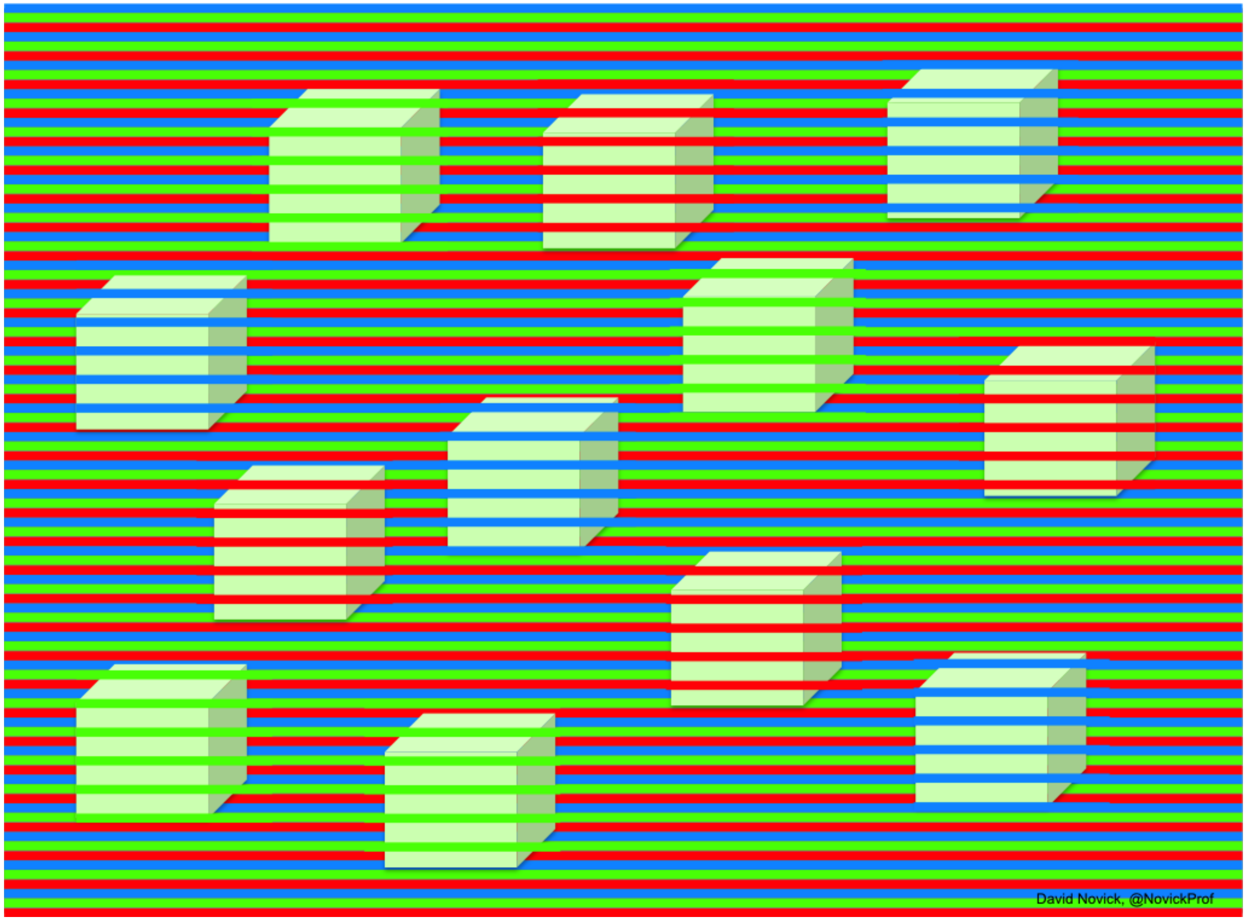
Showing yellow text on a white background doesn't work very well. There is insufficient luminance contrast.

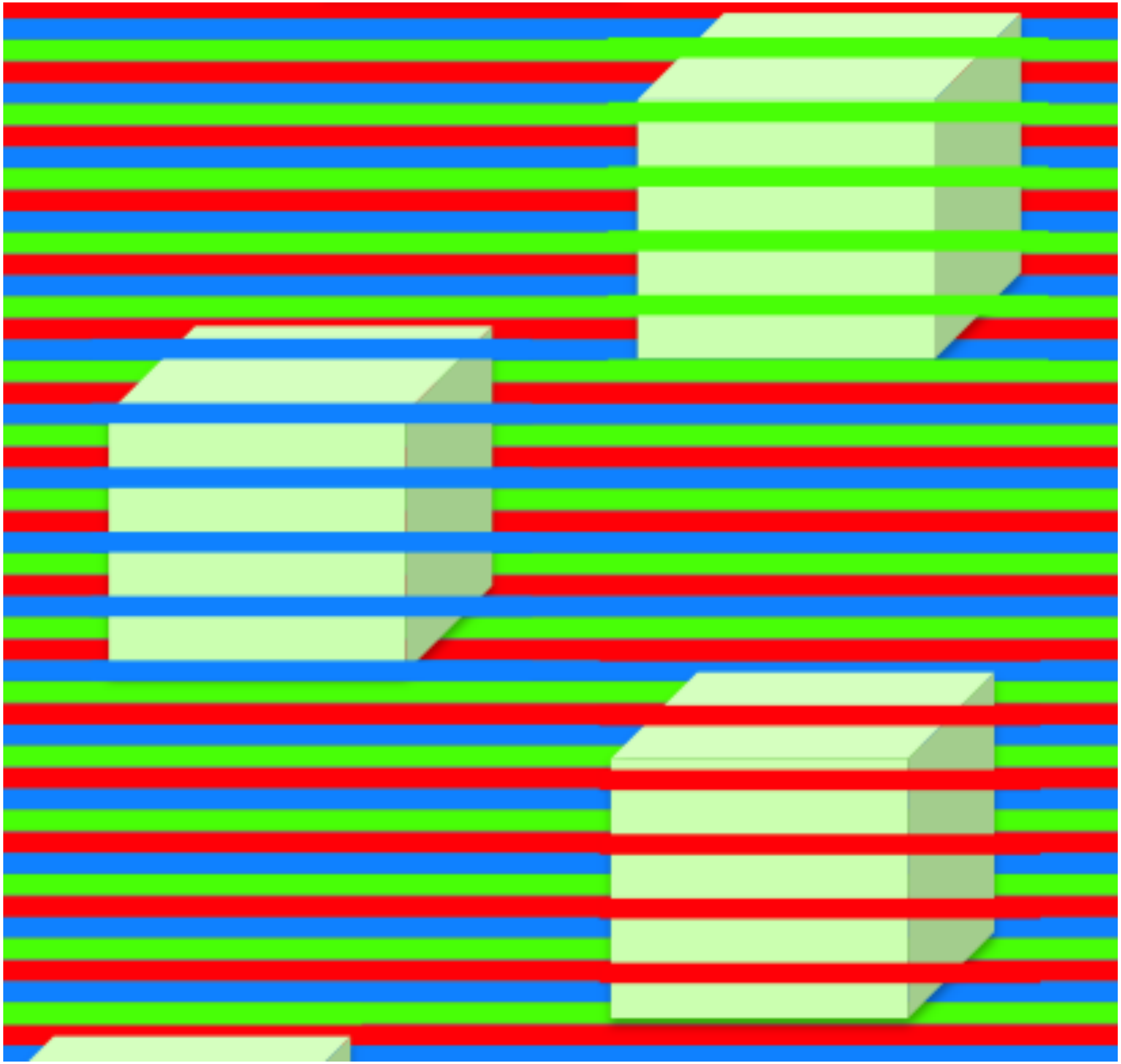
Showing yellow text on a black background works better. There is sufficient luminance contrast.

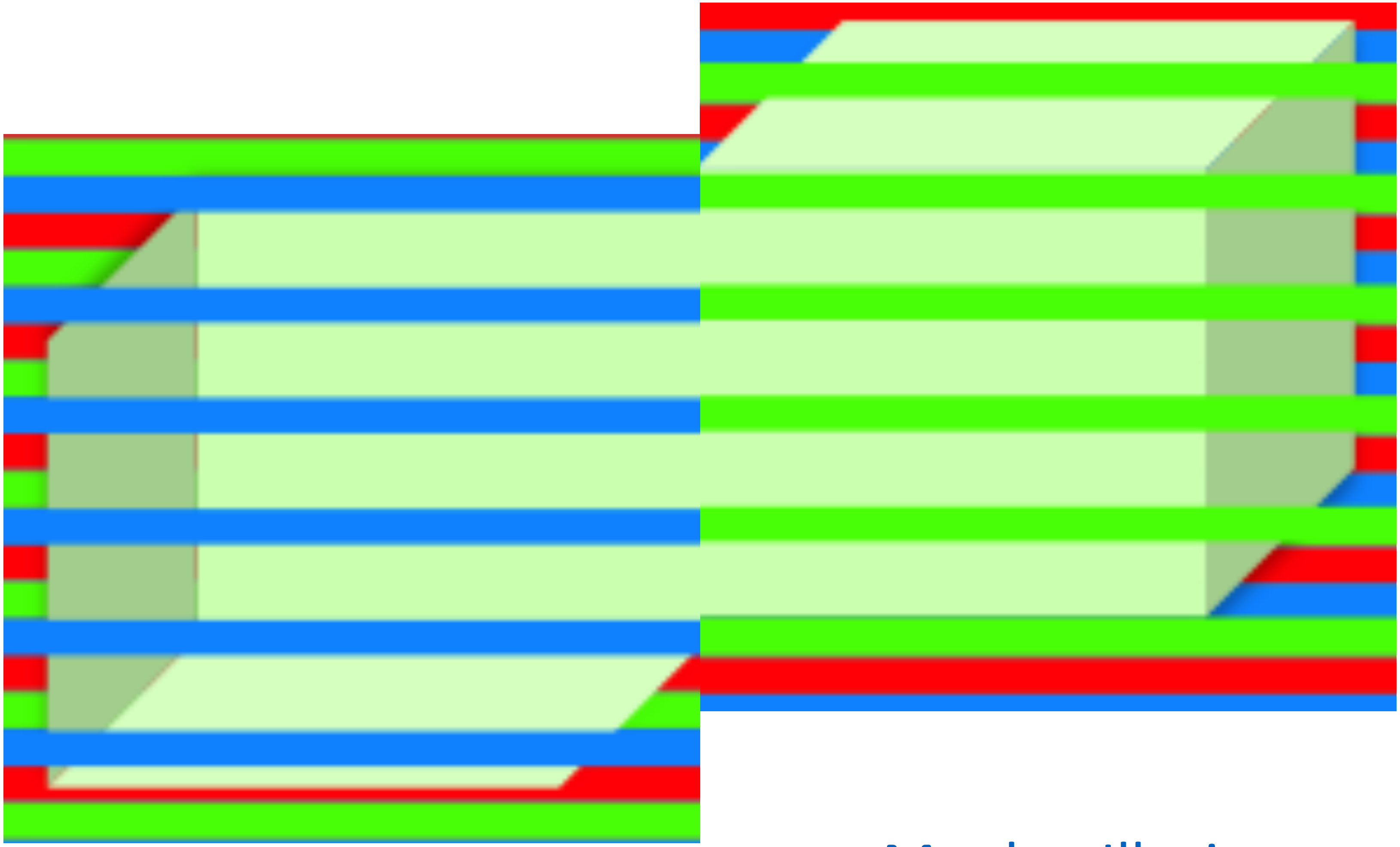
TIP: For presentations, light text on a dark background is often preferred.
I don't do this, because I'm also concerned with printing slides.
(With LaTeX Beamer, it is easy to have separate setups for presentation & print)


David Novick, @NovickProf


David Novick, @NovickProf

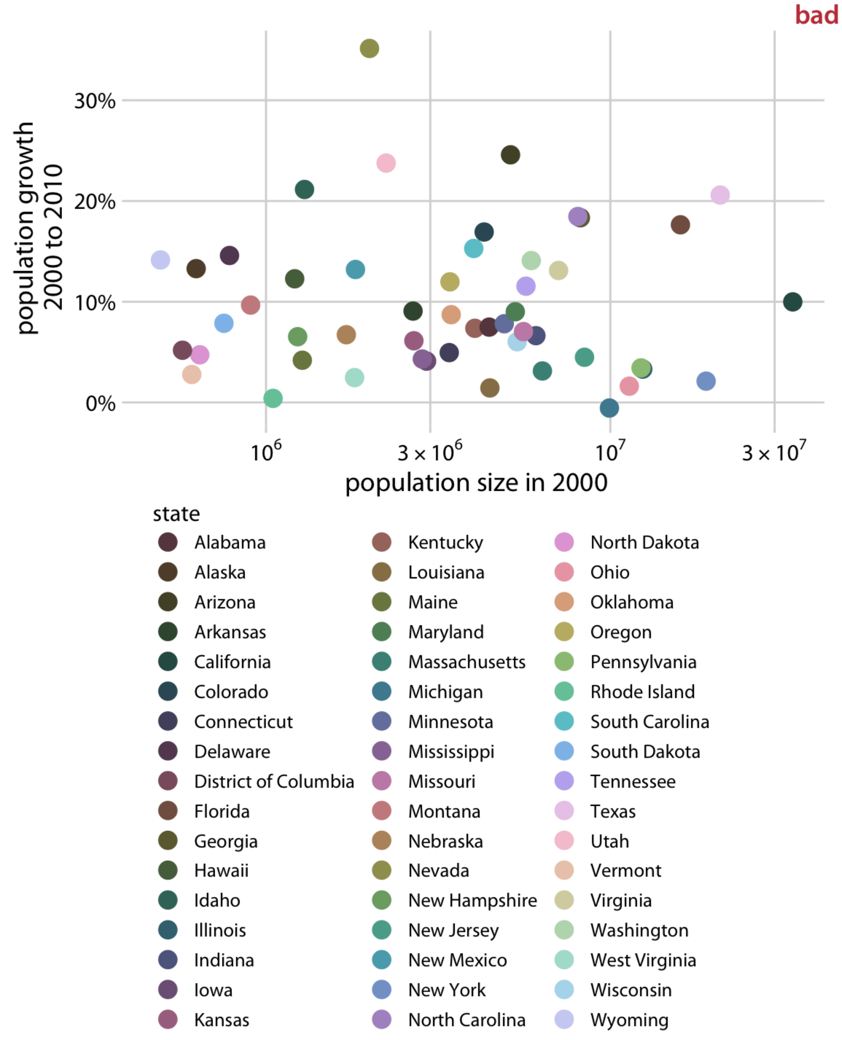

David Novick, @NovickProf





[Munker Illusion]

Error or efficient?

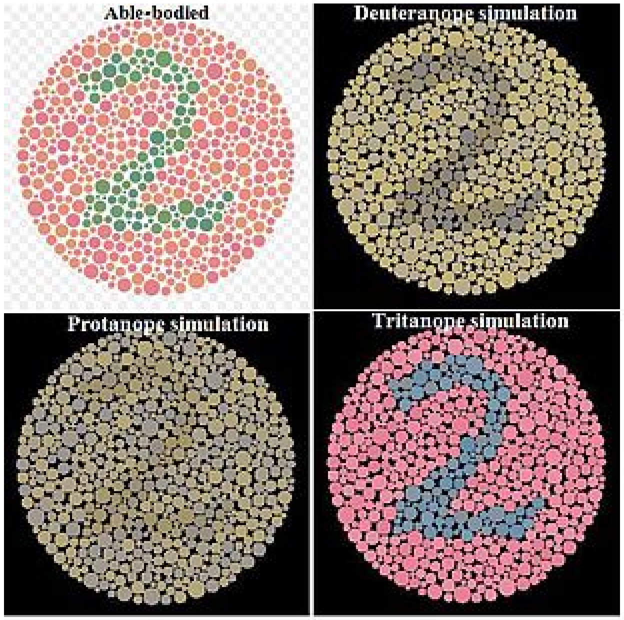


Encoding too much or irrelevant information

Use direct labeling instead of colors when you need to distinguish between more than about eight categorical items.

Wilke, Chapter 19

Colorblindness



Most common forms are genetic, and involve a deficiency in one of the cone type sensitivities

- Protanopia (red deficient: L cone absent)
- Deuteranopia (green deficient: M cone absent)
- Tritanopia (blue deficient: S cone absent)

Some form of red-green insensitivity is most common

- about 6-8% of population
- more common in males

TIP: Avoid color scales with main variation between red & green

Not designing for color-vision deficiency (cvd, aka colorblind).

Figure 19.7: A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly).

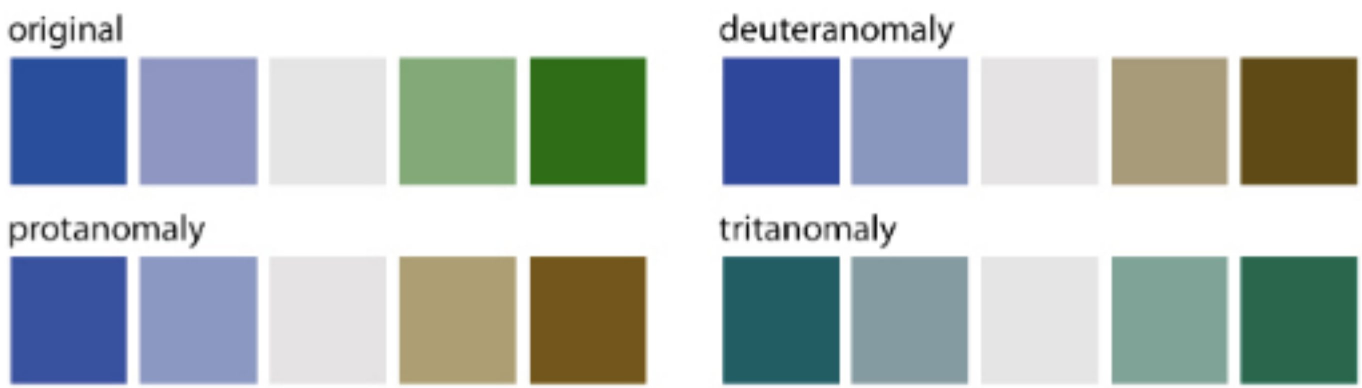
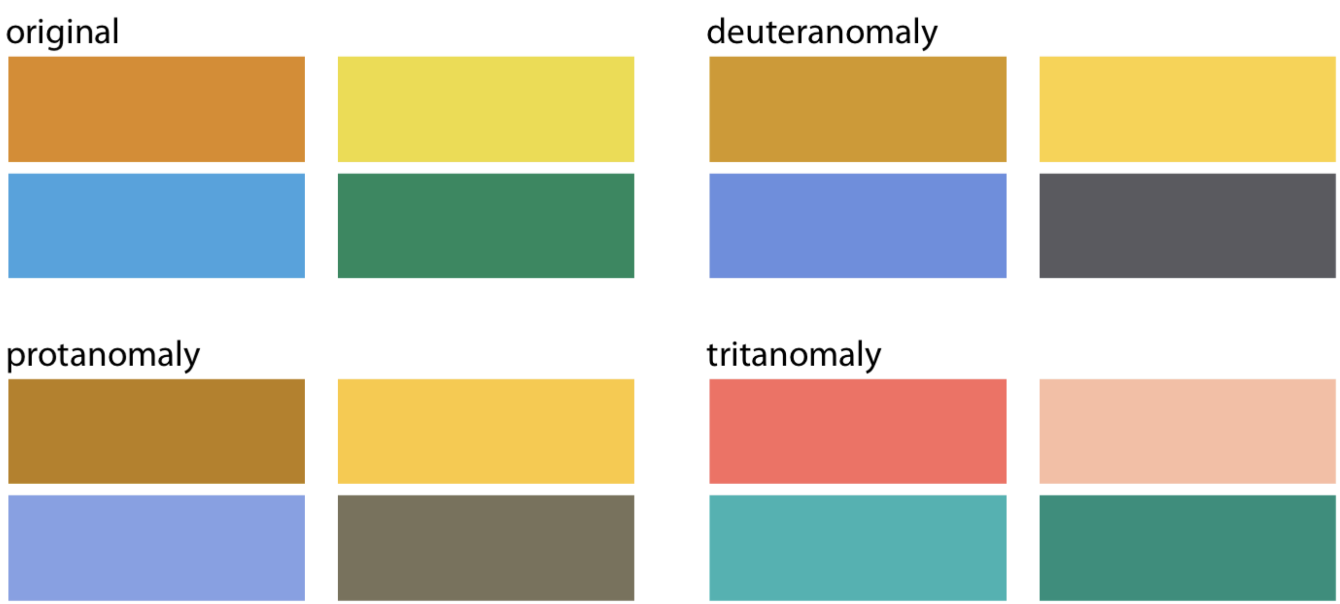


Figure 19.8: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

Wilke, Chapter 19

Not designing for color-vision deficiency (cvd, aka colorblind).

Figure 19.10: Qualitative color palette for all color-vision deficiencies (Okabe and Ito 2008). The alphanumeric codes represent the colors in RGB space, encoded as hexadecimals. In many plot libraries and image-manipulation programs, you can just enter these codes directly. If your software does not take hexadecimals directly, you can also use the values in Table 19.1.



Color: Lessons

- Use colors to represent differences in meaning
  - Avoid gratuitous use of multiple colors
  - Use consistent color scheme across multiple graphs of the same data
- Consider presentation goal:
  - Highlight one subset against the rest
  - Group a categorical variable
  - Encode a quantitative variable
- Consider differences in color perception, B/W printing

Summary

- In designing data graphics, consider the viewer
  - Info $\rightarrow$ encoding $\rightarrow$ image $\rightarrow$ decoding $\rightarrow$ understanding
- Perception: much is known, with ~ links to graphics
  - Bottom up: perceptual features, what grabs attention
  - Top down: expectations provide a context
  - Encoding attributes must consider what is to be seen
- Color: What is the presentation goal?
  - Color palettes for different purposes
  - Transparency increases the effective use of color