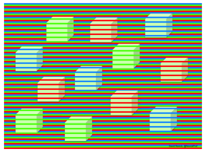
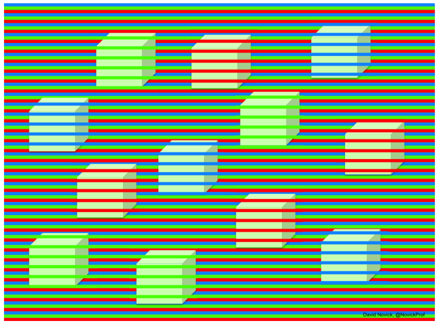
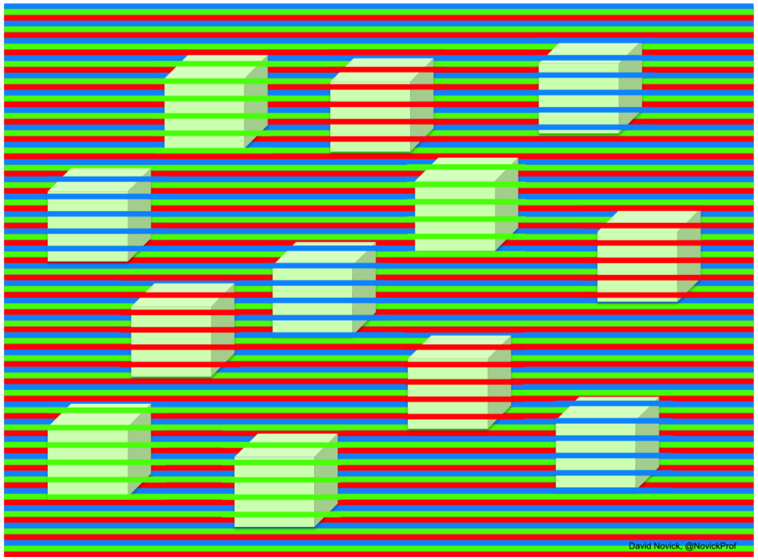
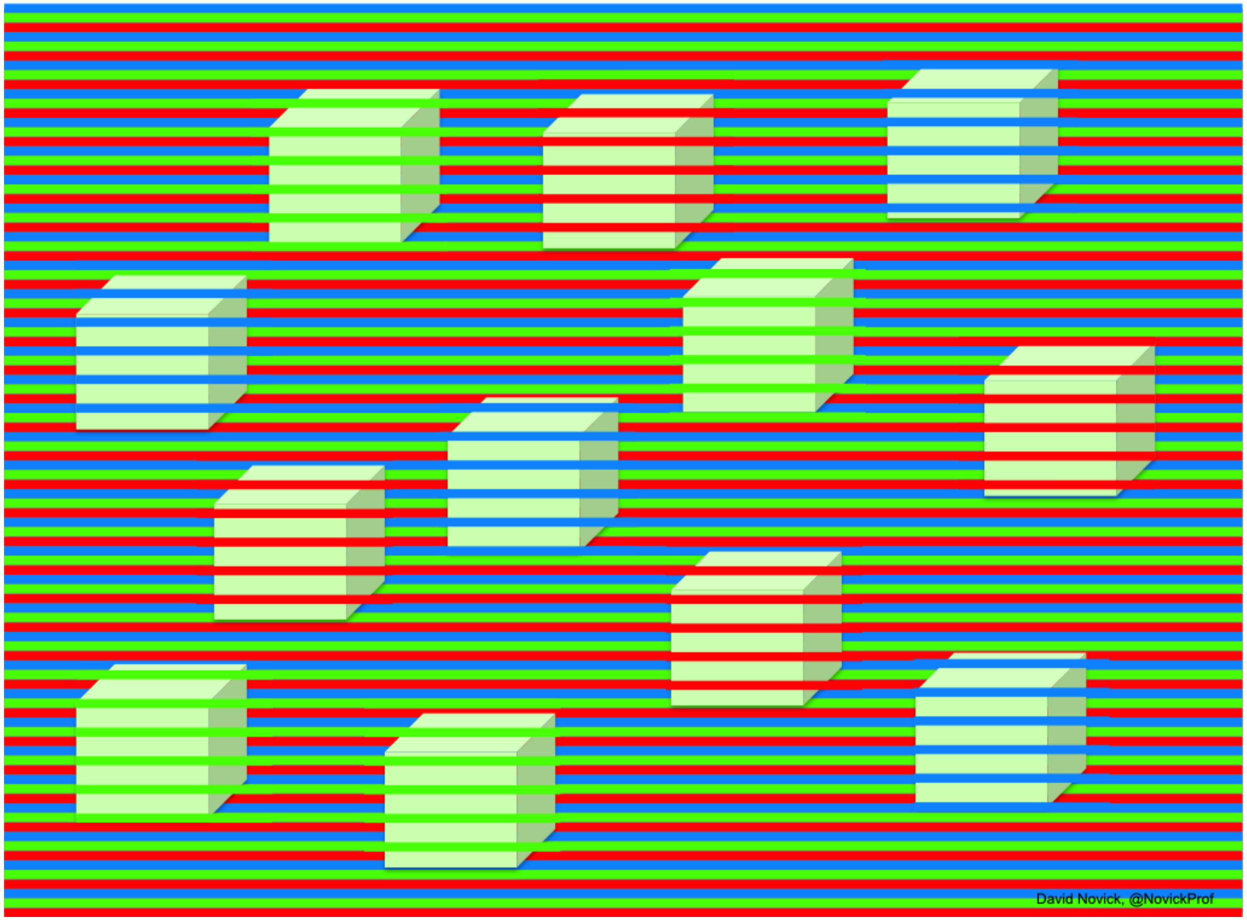
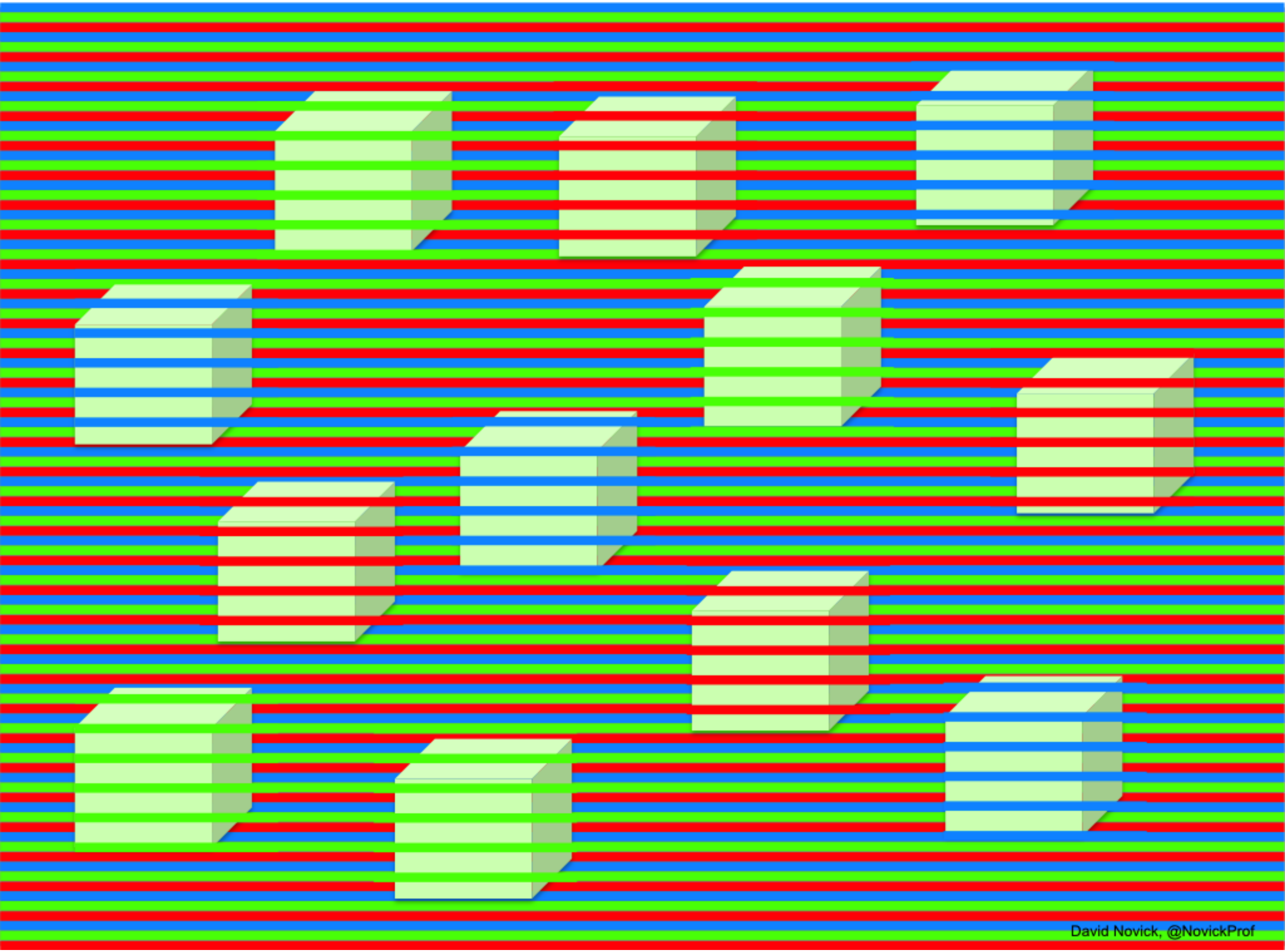
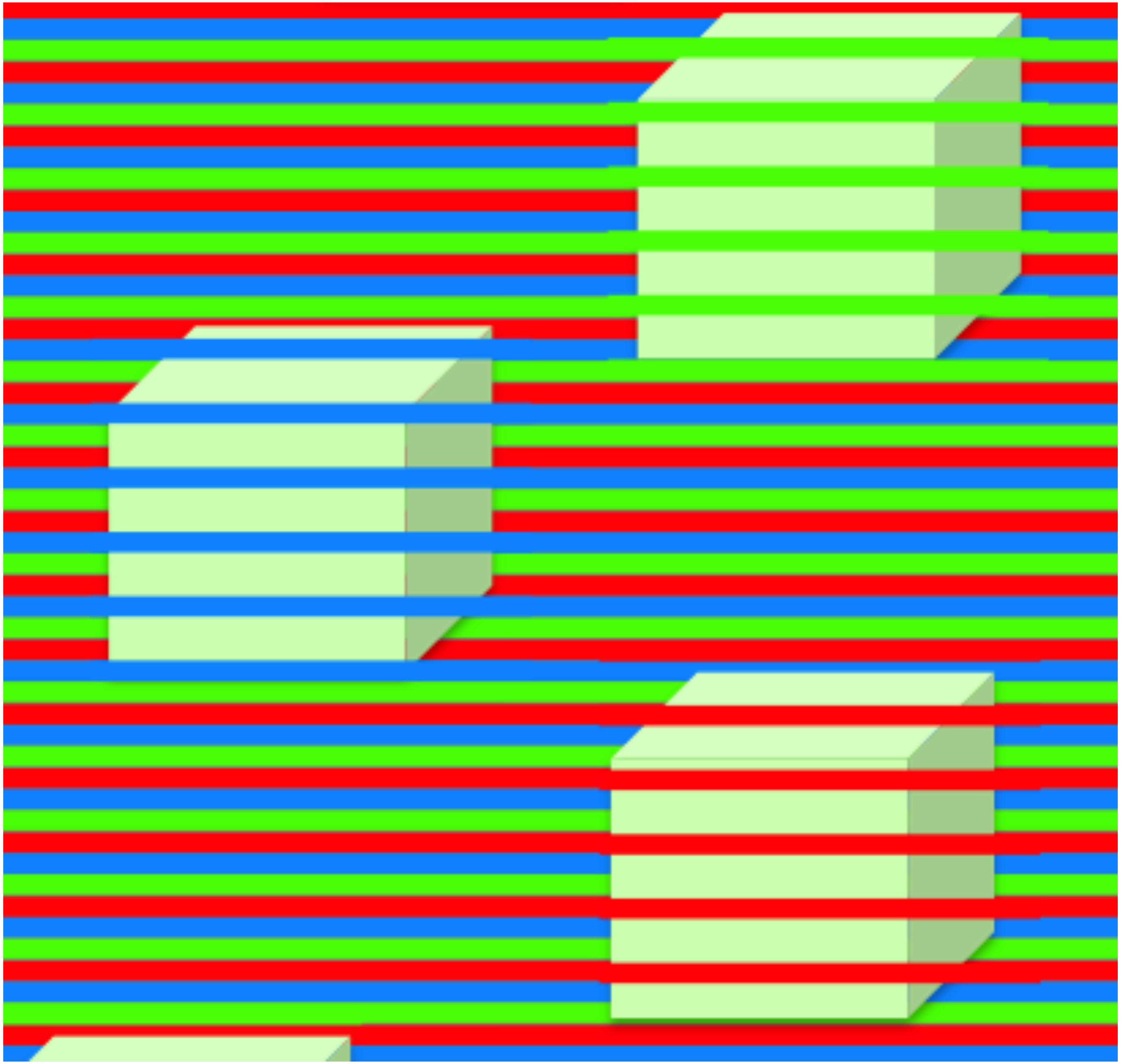

David Novick, @NovickProf

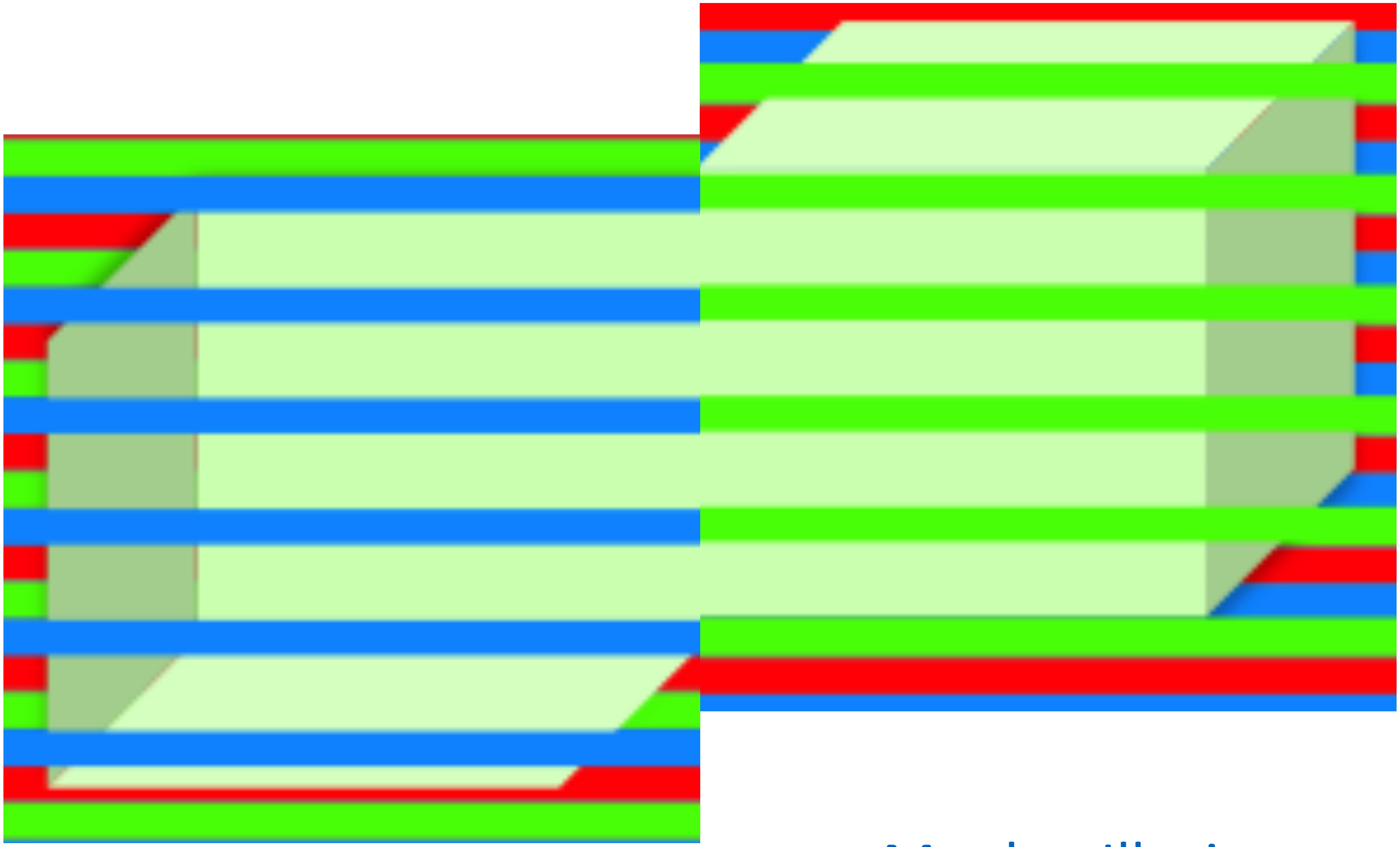

David Novick, @NovickProf


David Novick, @NovickProf

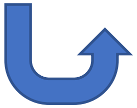

David Novick, @NovickProf


David Novick, @NovickProf

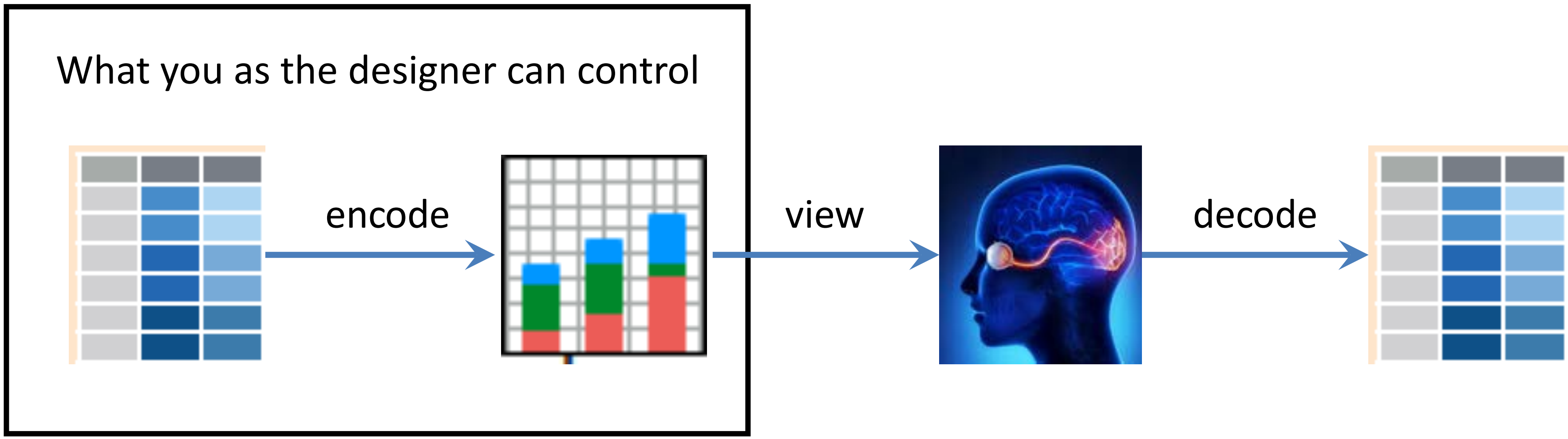




[Munker Illusion]



Error or efficient?


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*).

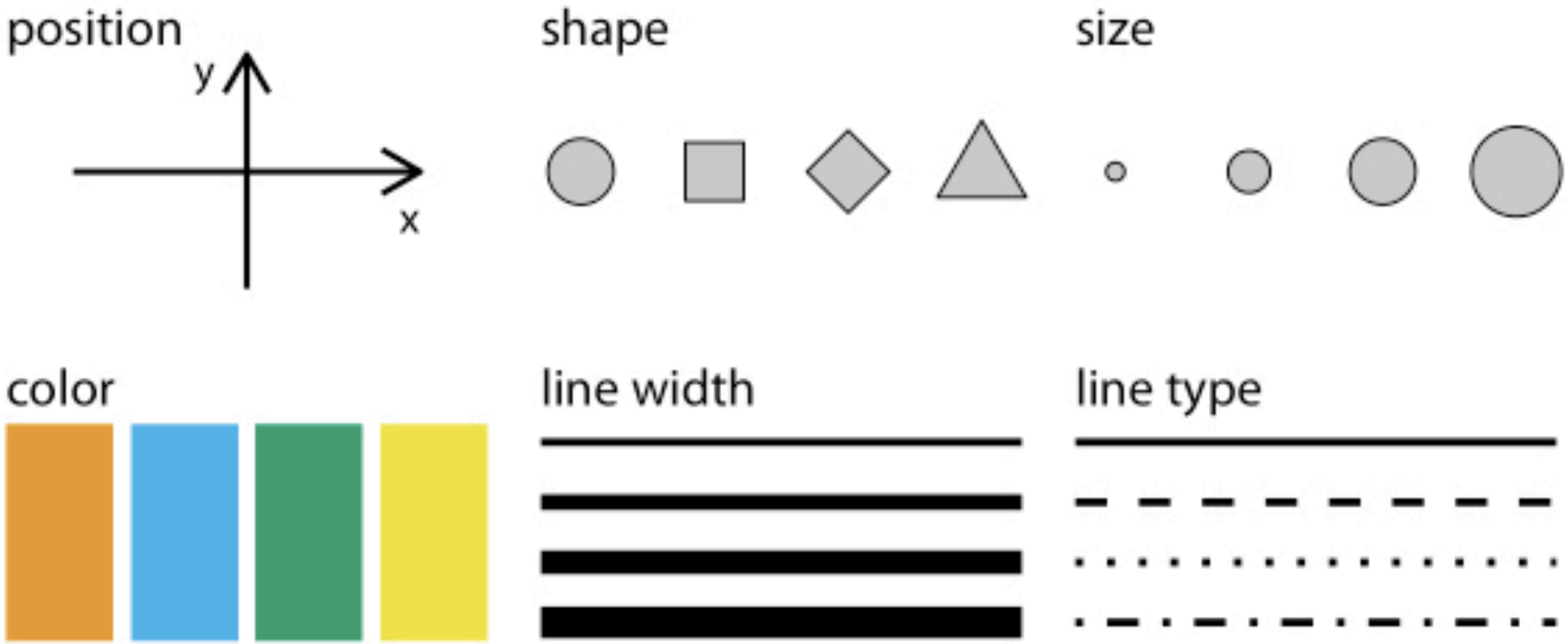


Figure 2.1: Commonly used aesthetics in data visualization: position, shape, size, color, line width, line type. Some of these aesthetics can represent both continuous and discrete data (position, size, line width, color) while others can usually only represent discrete data (shape, line type).

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, 1.5×10^{-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

?	?	?	?	?

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

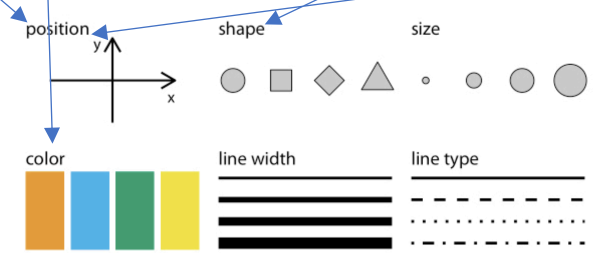
What is **encoding**? **Mapping** data variables onto aesthetics

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.

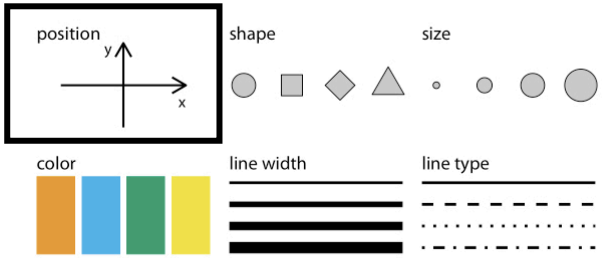


This is your job as a Data Vis **designer**.

Wilke, Chapter 2

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

Position is encoded in **coordinate systems**.



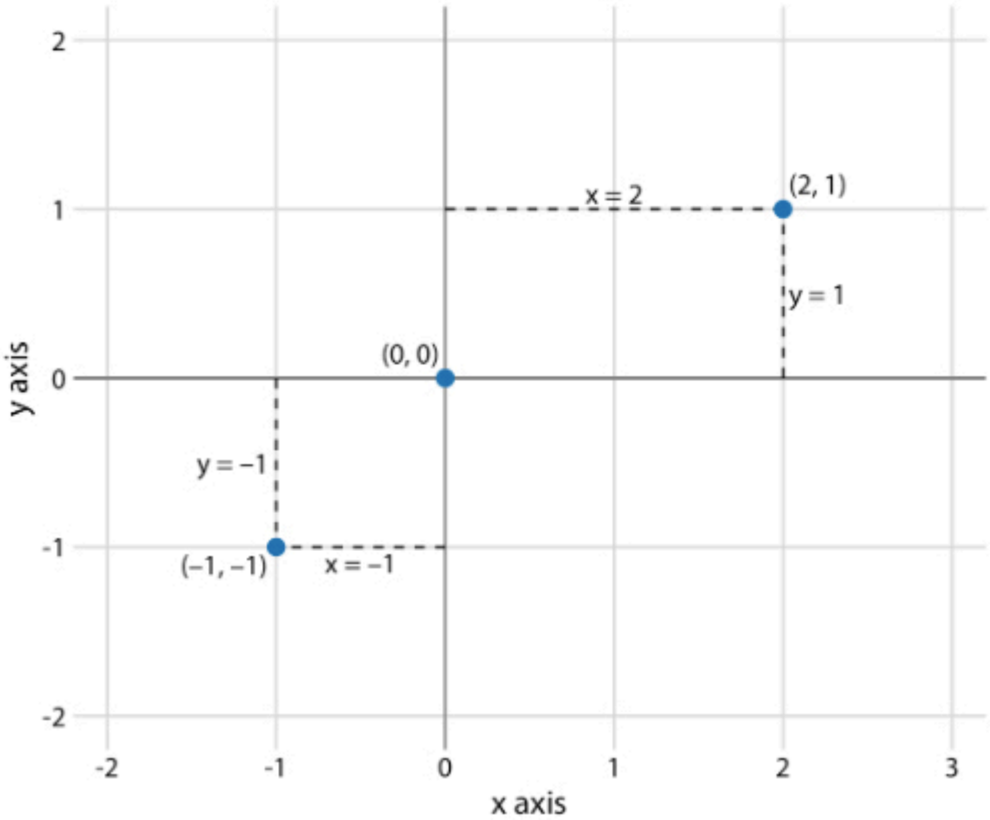
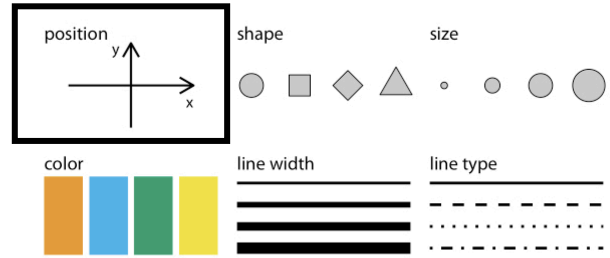


Figure 3.1: Standard cartesian coordinate system. The horizontal axis is conventionally called *x* and the vertical axis *y*. The two axes form a grid with equidistant spacing. Here, both the *x* and the *y* grid lines are separated by units of one. The point (2, 1) is located two *x* units to the right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



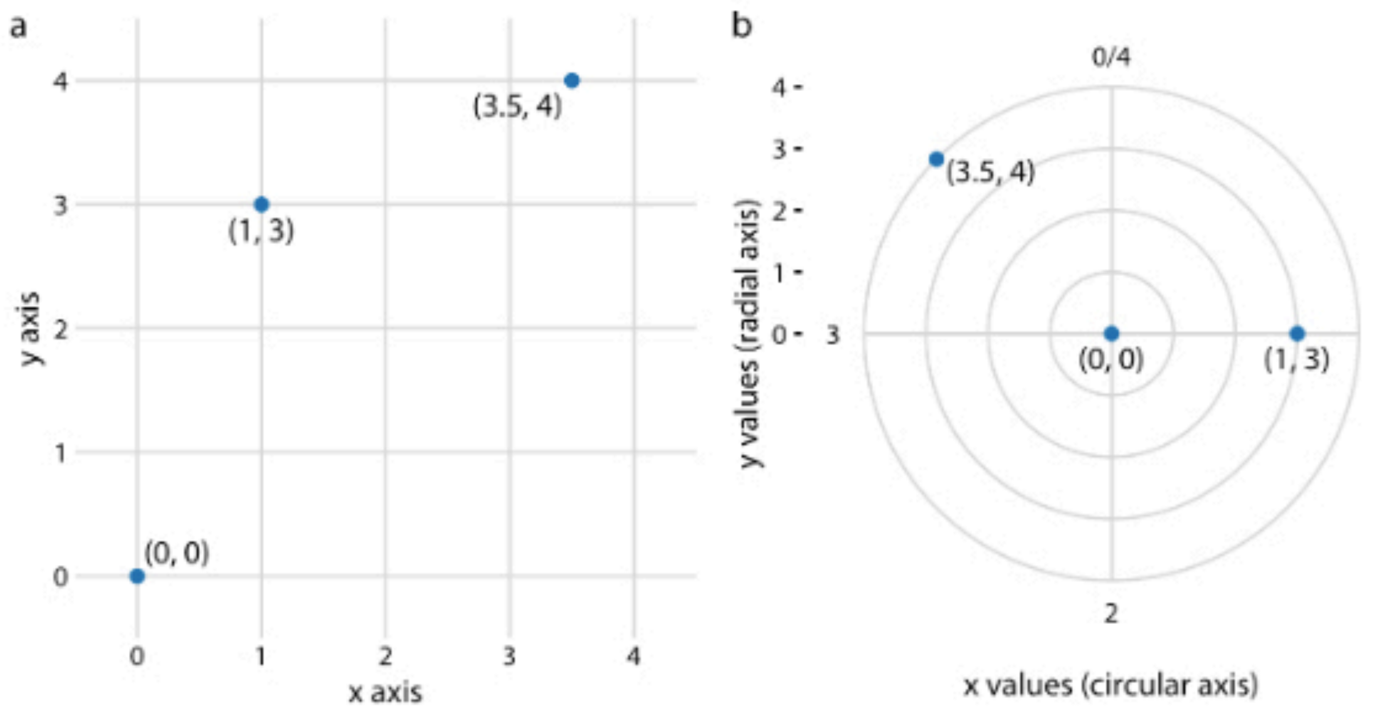
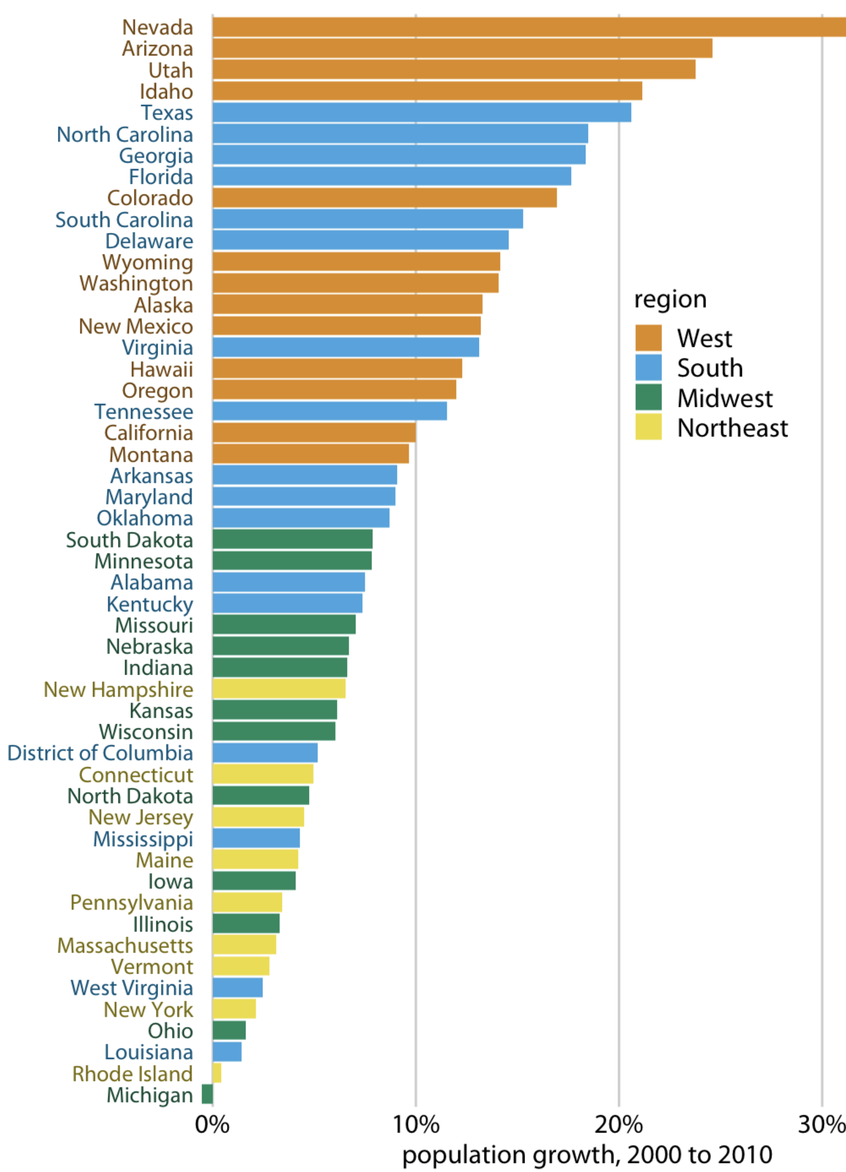


Figure 3.9: Relationship between Cartesian and polar coordinates. (a) Three data points shown in a Cartesian coordinate system. (b) The same three data points shown in a polar coordinate system. We have taken the *x* coordinates from part (a) and used them as angular coordinates and the *y* coordinates 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

Color as a tool to distinguish: Categories



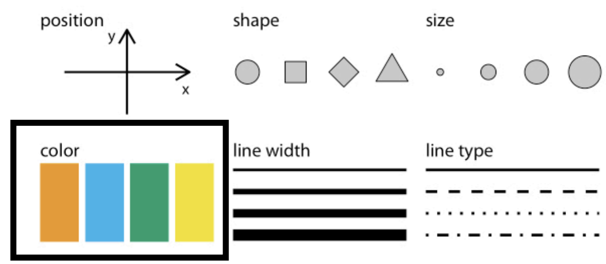
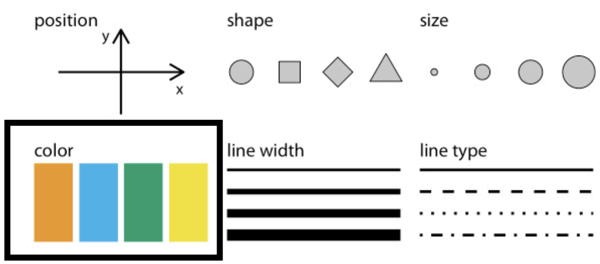


Figure 4.1: Example qualitative color scales. The Okabe Ito scale is the default scale used throughout this book (Okabe and Ito 2008). The ColorBrewer Dark2 scale is provided by the ColorBrewer project (Brewer 2017). The ggplot2 hue scale is the default qualitative scale in the widely used plotting software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

Wilke, Chapter 4

Color to represent data values: Numerical



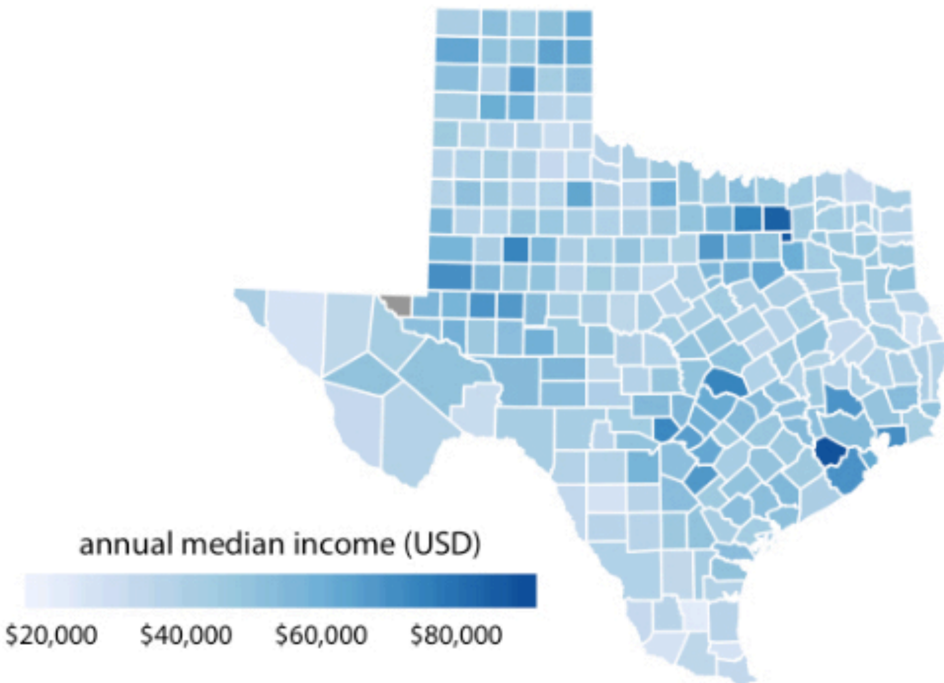
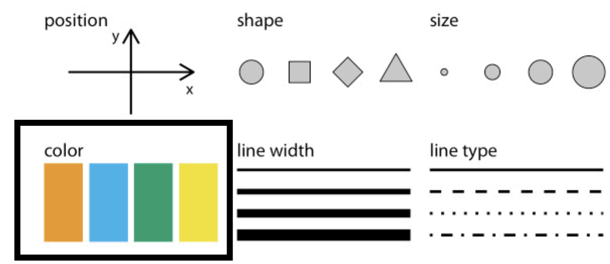


Figure 4.4: Median annual income in Texas counties. The highest median incomes are seen in major Texas metropolitan areas, in particular near Houston and Dallas. No median income estimate is available for Loving County in West Texas and therefore that county is shown in gray. Data source: 2015 Five-Year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



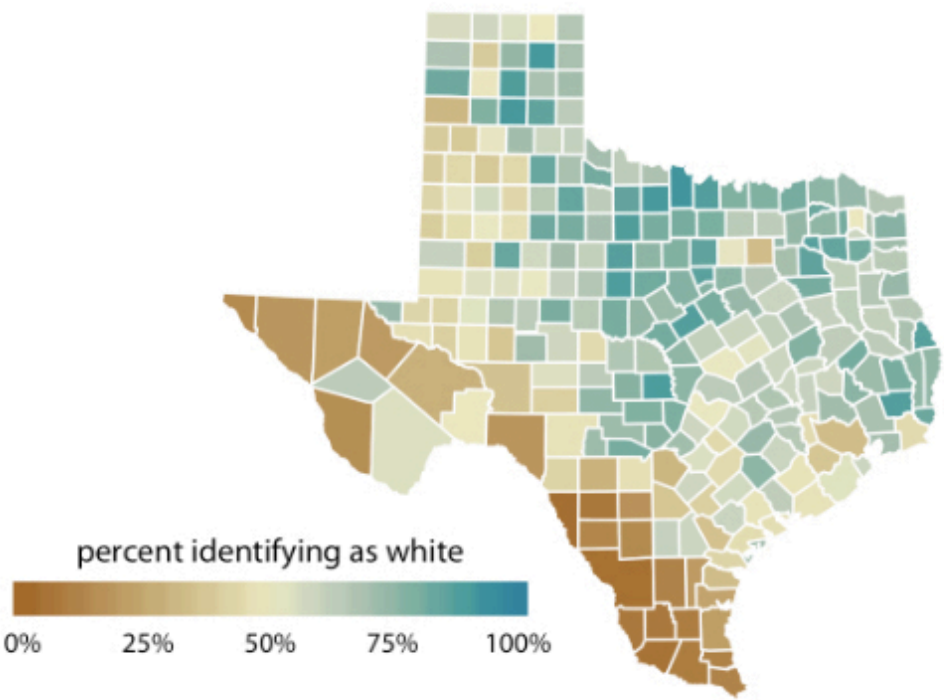


Figure 4.6: Percentage of people identifying as white in Texas counties. Whites are in the majority in North and East Texas but not in South or West Texas. Data source: 2010 Decennial U.S. Census

Divergent scales

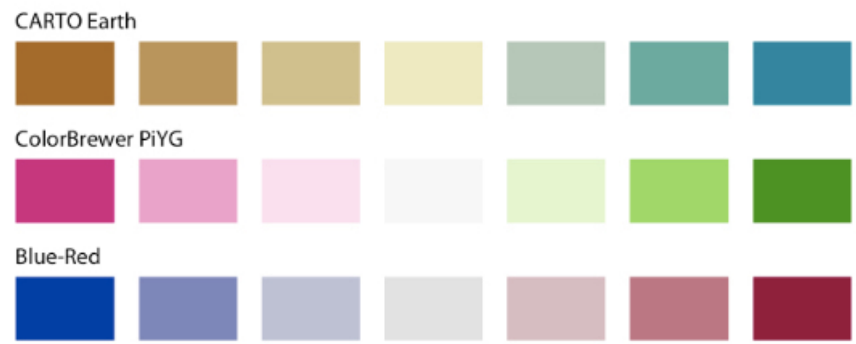
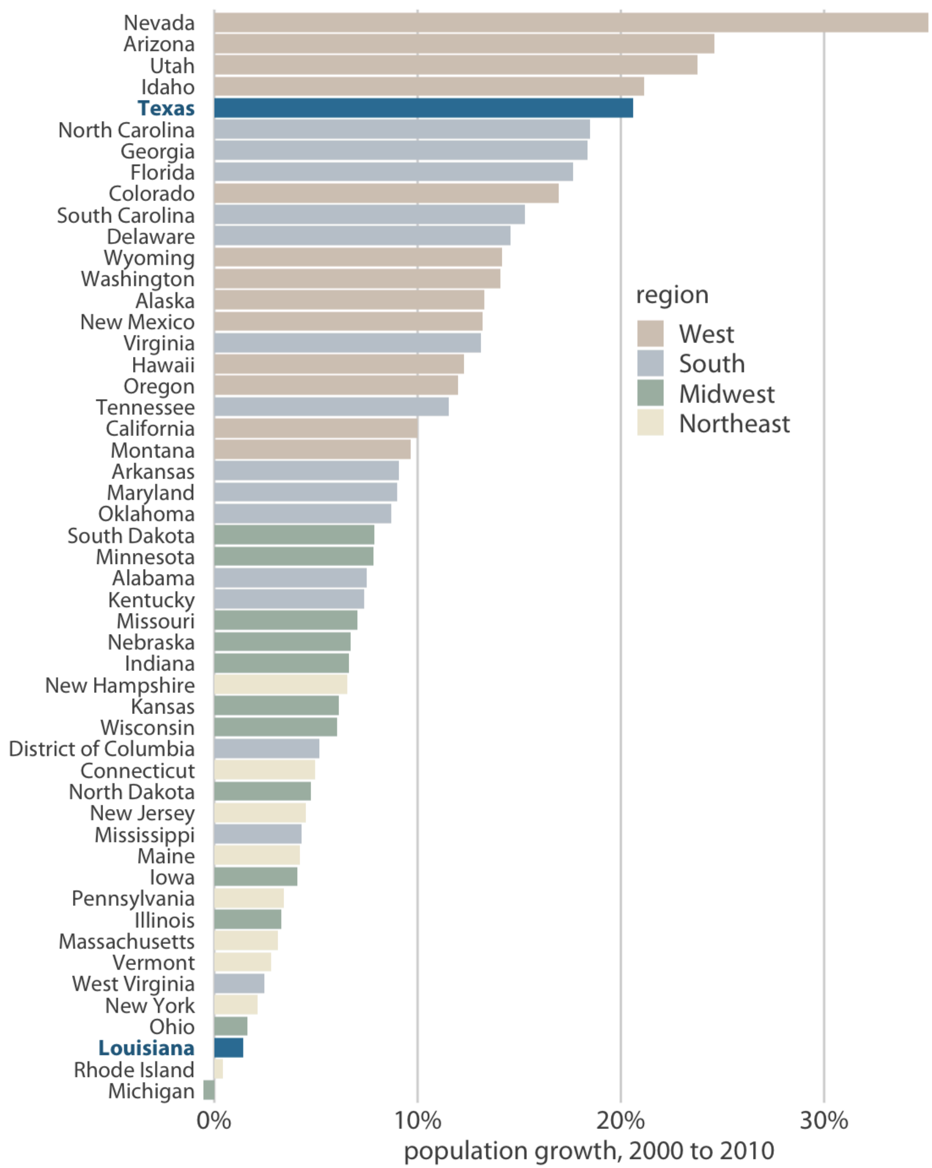


Figure 4.5: Example diverging color scales. Diverging scales can be thought of as two sequential scales stiched together at a common midpoint color. Common color choices for diverging scales include brown to greenish blue, pink to yellow-green, and blue to red.

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



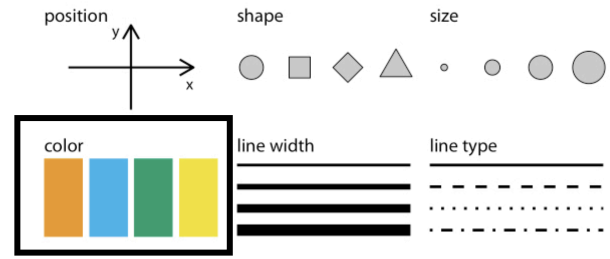
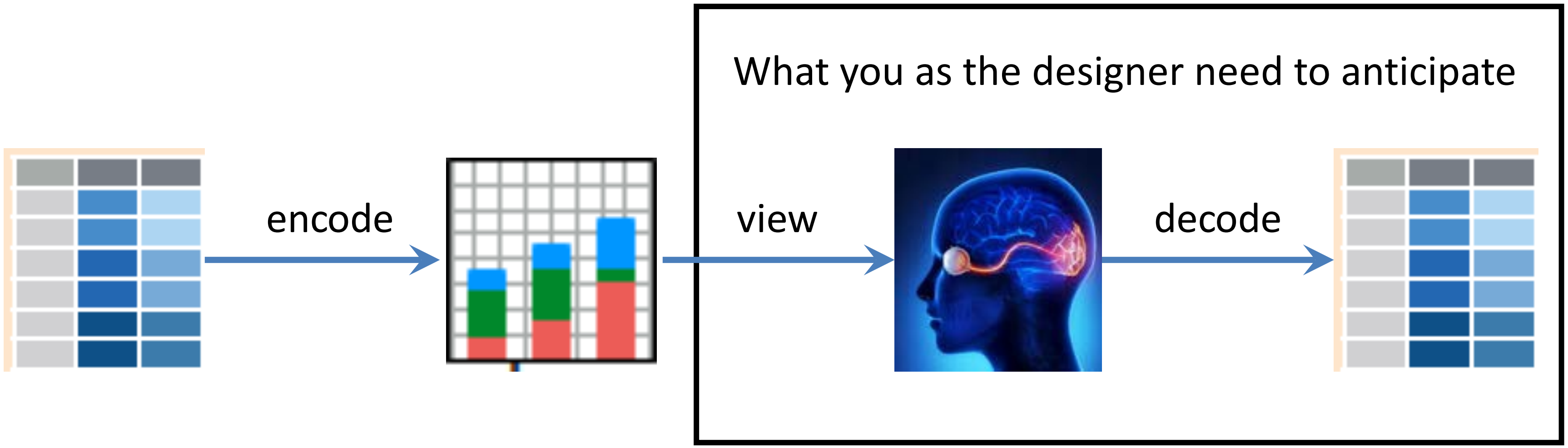
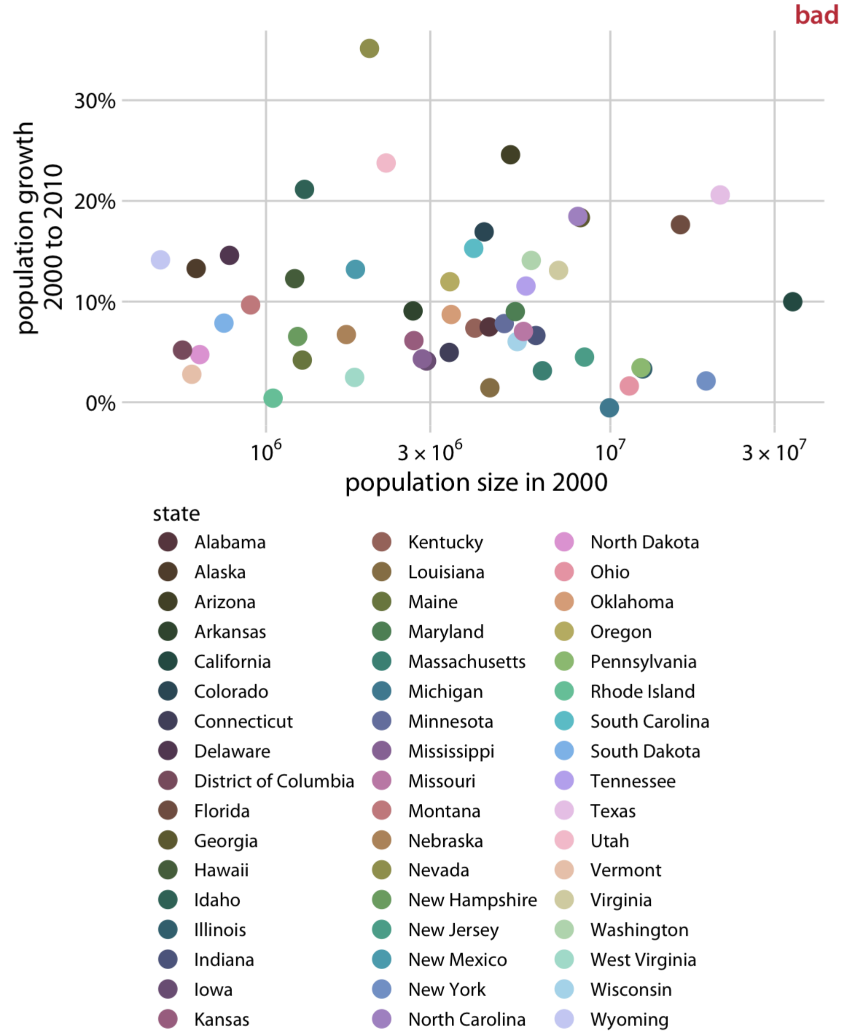


Figure 4.7: Example accent color scales, each with four base colors and three accent colors. Accent color scales can be derived in several different ways: (top) we can take an existing color scale (e.g., the Okabe Ito scale, Fig 4.1) and lighten and/or partially desaturate some colors while darkening 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



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



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

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).

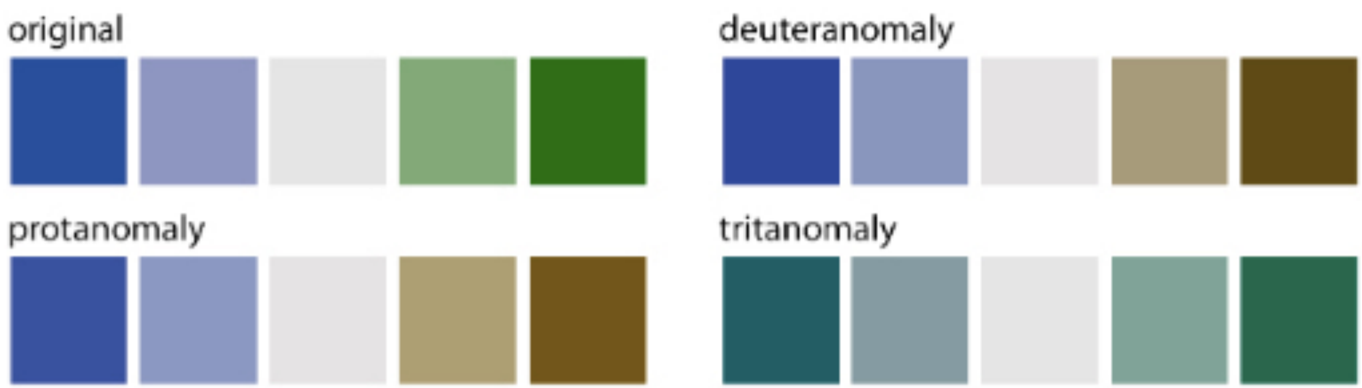


Figure 19.8: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

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.