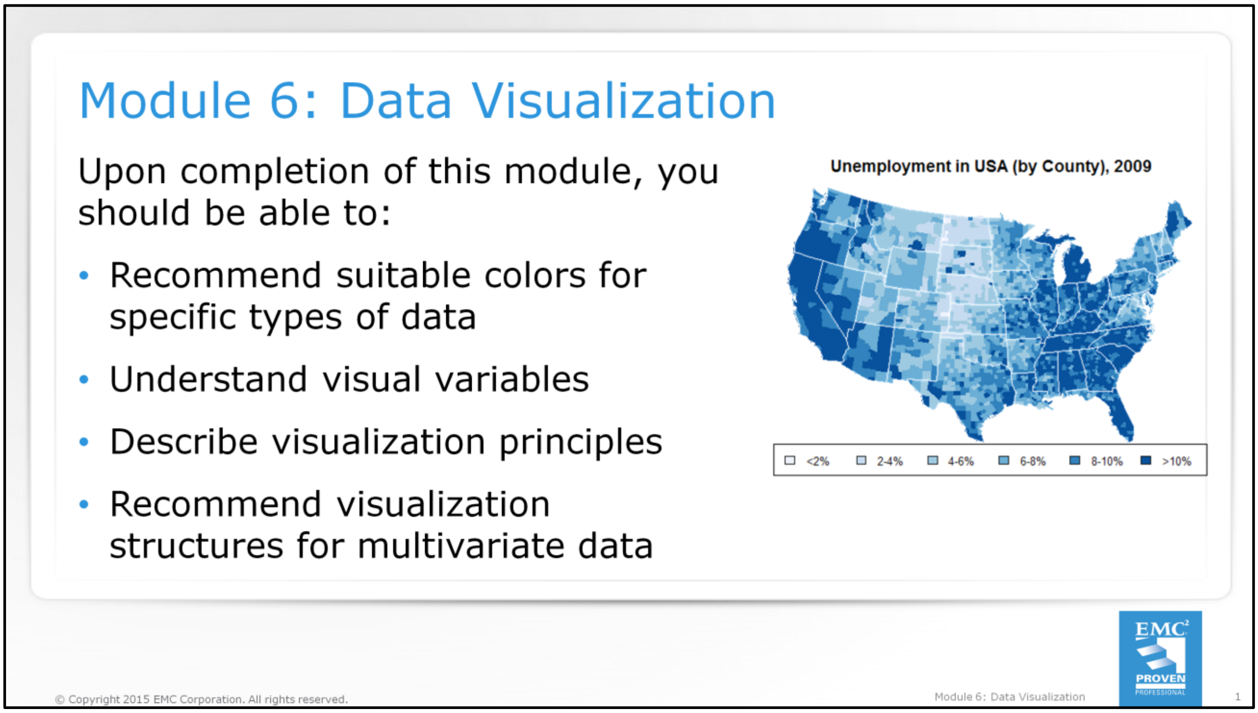


This module focuses on Data Visualization.

Lesson 1: Perception and Visualization

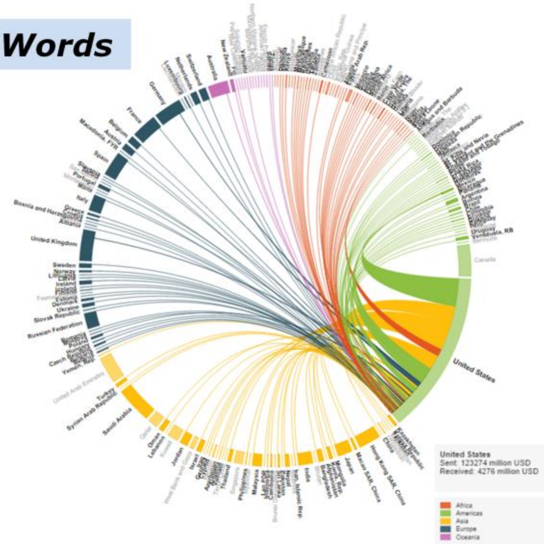
This lesson covers the following topics:

- Perception and colors
- Eight visual variables
- Visualization principles

Why is Data Visualization Important?

One Look is Worth a Thousand Words

- We are visual beings
  - Sight is a key sense for information understanding
  - We have been using visuals for many centuries
- Data is easier to read in visual form
- Helps discover new knowledge
- Applies to any domain
- Assist in analysis and communication



Global remittance flow: http://www.torre.nl/remittances/

Before we get into the module, we need to understand why data visualization is important.

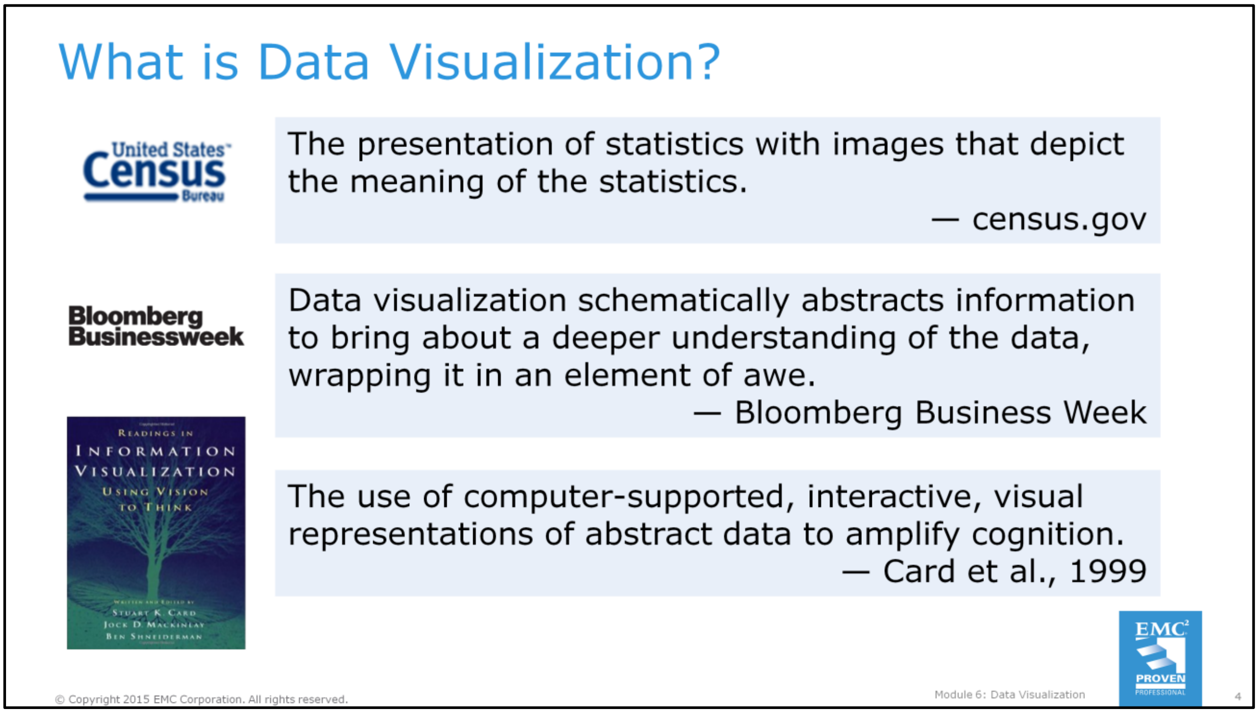
"One look is worth a thousand words," as the Chinese proverb puts it. We are all visual beings. A baby forms sight as early as in the womb, long before the child learns to speak or read. Sight is one of the key senses for understanding information. Humans have been using visuals for many, many centuries. In prehistoric times, ancient humans already knew how to use cave drawing to communicate. The oldest known cave art comes from the Cave of El Castillo in Cantabria, Northern Spain, dated back to at least 40,800 years.

Data is often much, much easier to read in a visual form and it can also help discover new knowledge. For example, a histogram of household income data makes it much easier for us to identify the distribution of the income compared to reading a gigantic Excel spreadsheet.

Data visualization can be applied to any domain. For example, the same visualization tool (such as a scatterplot) can visualize two variables from any domain, whether we are plotting car fuel efficiency versus the weight of the car, or the country GDP versus the unemployment rate.

Data visualization assists in analysis and communication. The slide shows the global movements of money due to remittances (money sent by migrants to their home countries), with United States migrants highlighted. You can view the visualization at: http://www.torre.nl/remittances/. The graph shows that, the majority of money from the US went to Asia (yellow), Americas (green), and Europe (dark blue). This visualization is based on the World Bank data, which can be downloaded from http://go.worldbank.org/JITC7NYTT0.

Last but not least, data visualization may distort the truth if it's not used correctly. There are many examples of how visualizations from distorted data have been used to sway opinions and lie to the audience. These "visualization lies" can be found everywhere from prestigious journals and newspapers to company portfolios. We will take a look at some of these examples later in the module. This Gizmodo article also shows a few examples: http://gizmodo.com/how-to-lie-with-data-visualization-1563576606.



The slide shows three different definitions of Data visualization.

The US Census Bureau defines data visualization as a term that means "the presentation of statistics with images that depict the meaning of the statistics."[1]

Bloomberg Business Week says "Data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe."[2]

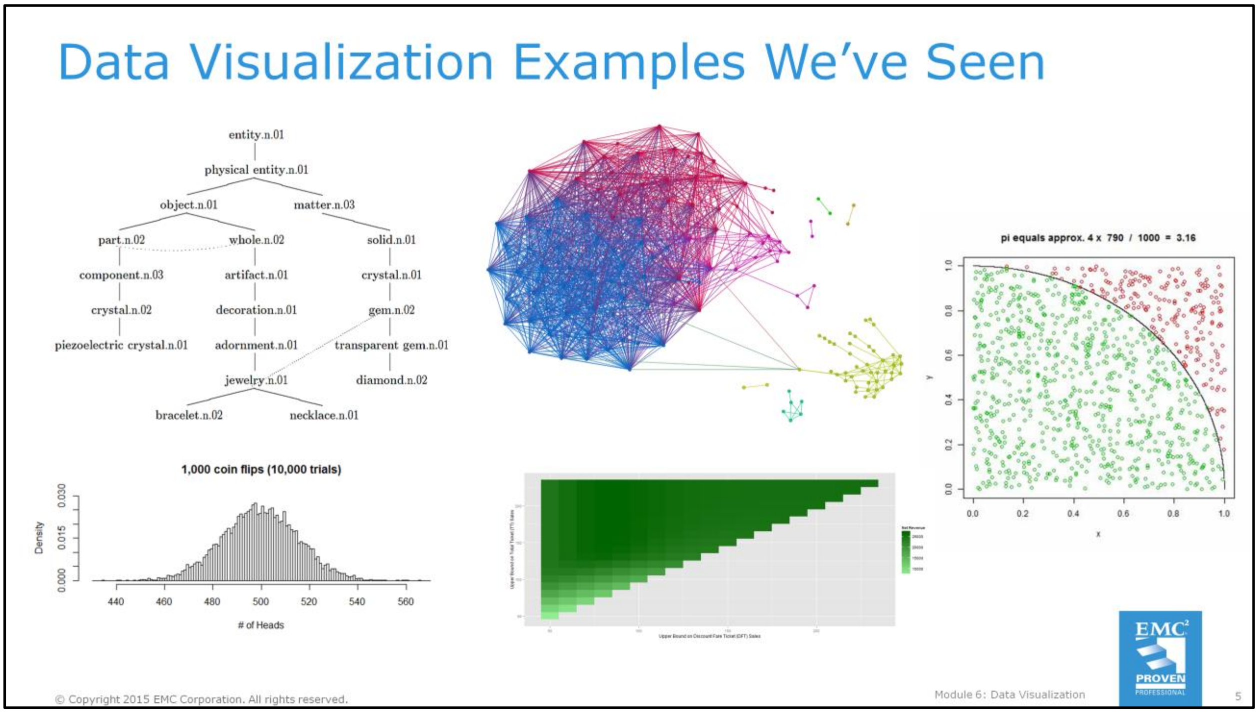
A better definition is perhaps from Card et al.'s book in 1999. They define data visualization as "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition."[3]

References:

[1]Yau, N. Visualizing Census Data. *The United States Census Bureau*. Last modified September 6, 2013. http://www.census.gov/multimedia/www/videos/data_visualization.php

[2]Popova, M. (2012). Data Visualization: Stories for the Information Age. *Bloomburg BusinessWeek Innovation and Design*. http://www.businessweek.com/innovate/content/aug2009/id20090811_137179.htm

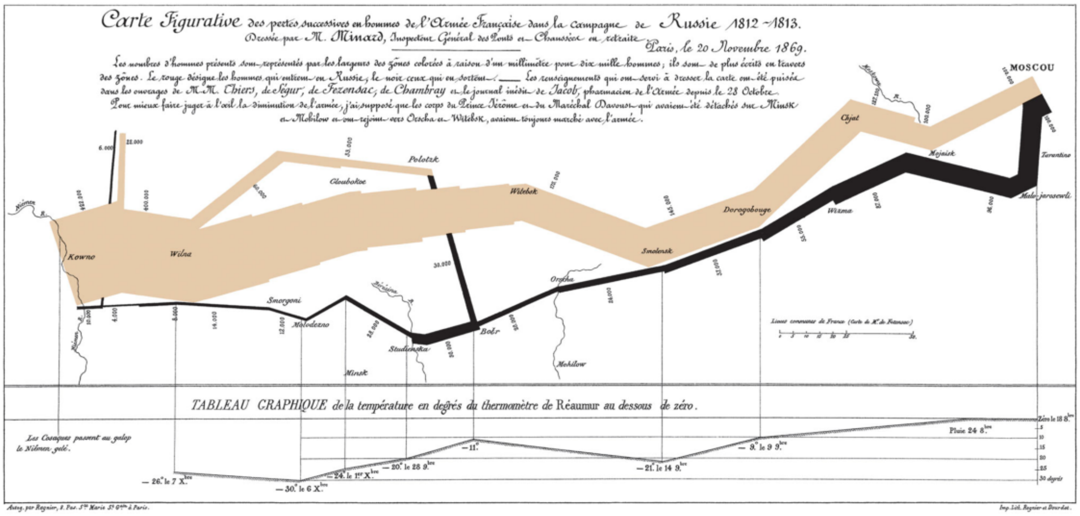
[3]Card, S. K., Mackinlay, J. D., & Shneiderman, B. (Eds.). (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.



Here are some of the visualization examples we have seen in the earlier modules.

Early Visualization: Minard's Flow Map (1869)

- Charles Minard's flow map of Napoleon's March
  - Displays army size, geo coordinate, date and weather all in a 2D image



6

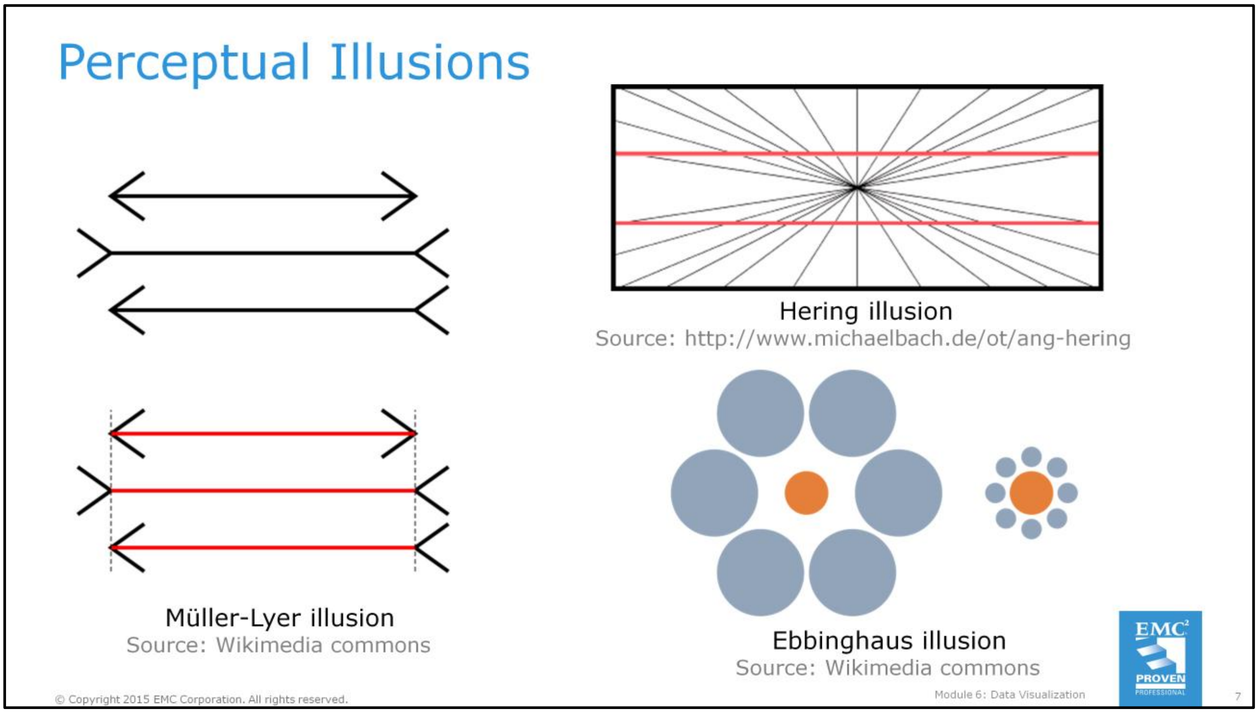
This chart tells the story of a war: Napoleon's Russian campaign of 1812. It was drawn half a century afterwards by Charles Joseph Minard, a French civil engineer who worked on dams, canals and bridges. Charles Minard is considered the inventor of information graphics. The breakthrough of this chart lies in its abstract representation for the axes.

The graph displays several variables in a single two-dimensional image:

- the size of the army - providing a strong visual representation of human suffering, e.g. the sudden decrease of the French army's size, from over 400,000 to 10,000, at the crossing of the Berezina river on the retreat
- the geographical co-ordinates, latitude and longitude, of the army as it moved
- the direction that the army was traveling, both in advance and in retreat, showing where units split off and rejoined
- the location of the army with respect to certain dates
- the weather temperature along the path of the retreat

Reference:

Charts: Worth A Thousand Words, economist.com, December 19, 2007, http://www.economist.com/node/10278643.



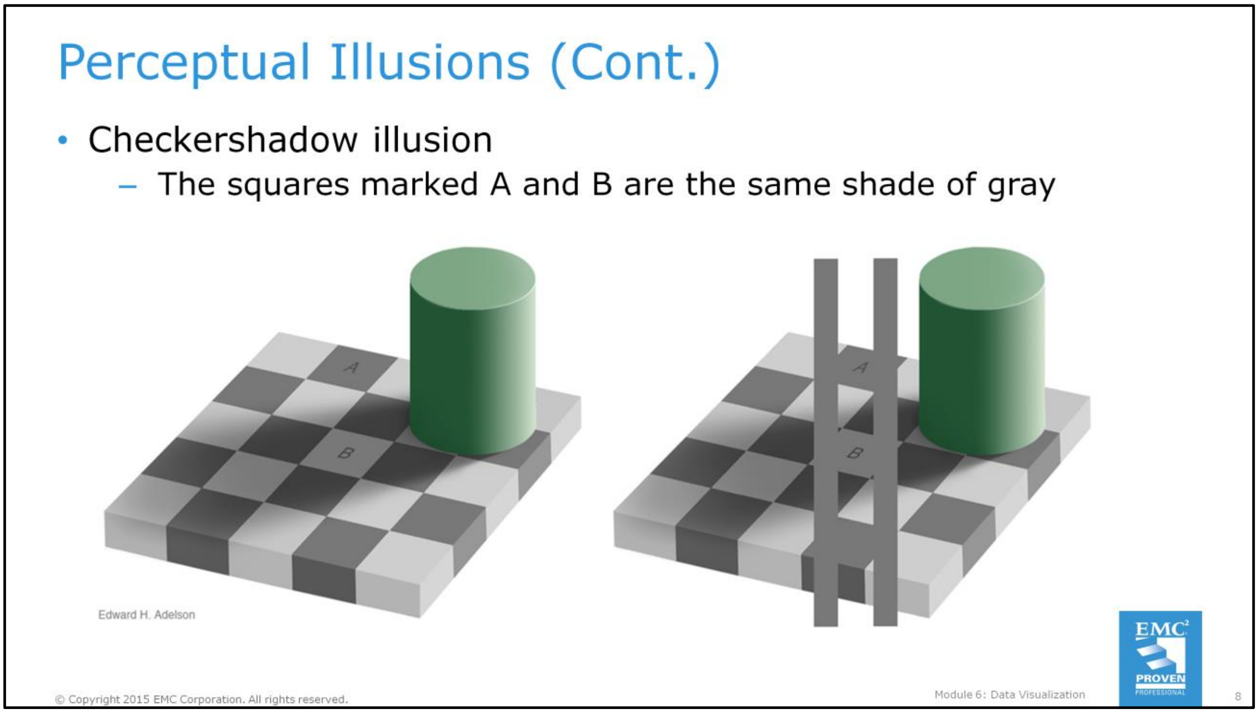
Sometimes we cannot perceive what does exist. Sometimes we perceive things that do not exist. Sometimes we perceive what cannot be there. Illusions show that much happens automatically in our visual system. There are many famous perceptual illusions out there. This slide shows three of them.

The Müller-Lyer illusion, on the left, was devised by Franz Carl Müller-Lyer, a German sociologist, in 1889. It is an optical illusion that consists of stylized arrows. Straight line segments of equal length comprise the shafts of the arrows, while shorter line segments (called the fins) protrude from the ends of the shaft. The picture shows two sets of arrows that exhibit the Müller-Lyer optical illusion. For most people, the line with the fins of the arrow protruding outward appears to be the longest while the line with the arrow fins pointing inwards appears to be the shortest. But the set on the bottom shows that all the shafts of the arrows are of the same length.

The Hering illusion, top right, is a geometrical-optical illusion and was discovered by the German physiologist Ewald Hering in 1861. The two red lines are straight and in parallel. But when they are presented in front of a radial background, like the spokes of a bicycle, the lines appear as if they were bowed outwards. Even the black rectangle frame appears to be bowed outwards as well.

The Ebbinghaus illusion, bottom right, is an optical illusion of relative size perception. It was named after the German psychologist Hermann Ebbinghaus (1850–1909). Two circles of identical size are placed near each other, and one is surrounded by large circles while the other is surrounded by small circles. As a result of the juxtaposition of circles, the central circle surrounded by large circles appears smaller than the central circle surrounded by small circles.

Studying perceptual illusion can help us understand human perception and highlight what to avoid when designing a visualization system. For example, the Hering illusion may occur when overlaying polygons in a geographical information system (GIS) and the Ebbinghaus illusion may occur in a scatter plot or bubble chart.



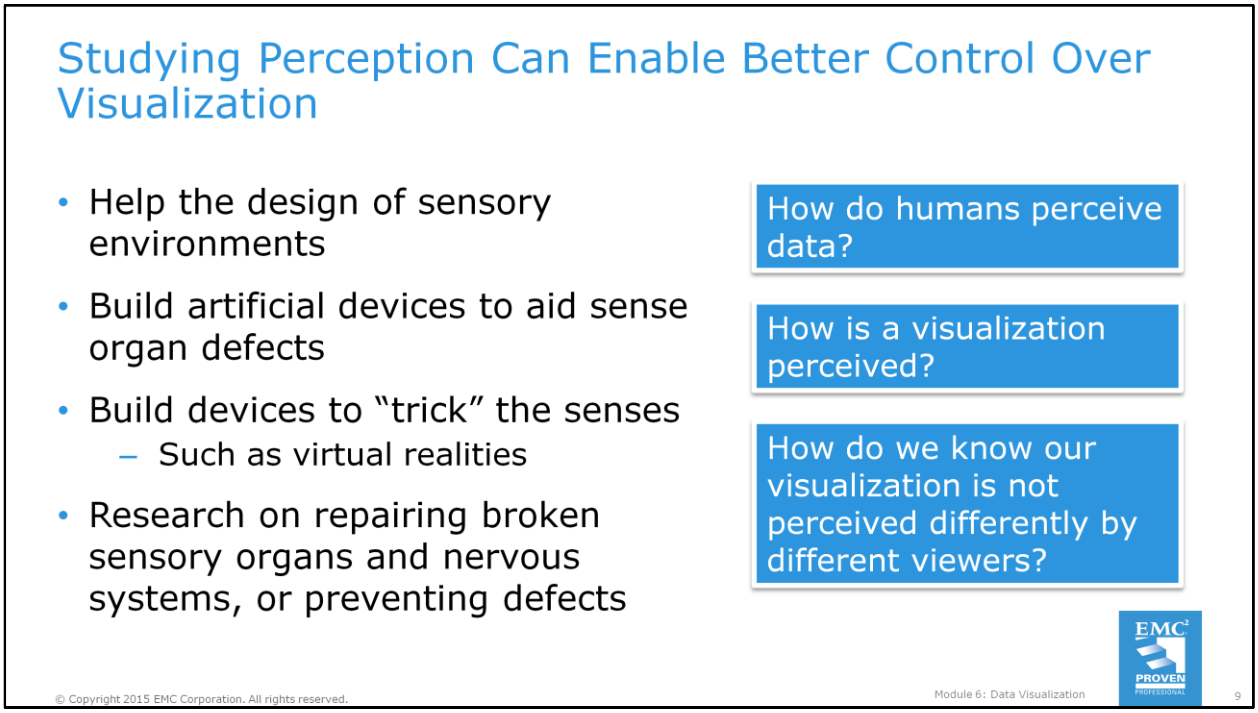
The checker shadow illusion was published by Edward Adelson, a professor of vision science at MIT in 1995.

The original image on the left shows a checkerboard with light and dark squares. The square labeled A appears to be darker than the square labeled as B on the 2D plane of the rendered 3D projection, but they actually share the same color on the 2D plane of the image. The image on the right adds two stripes to the original image. By joining the squares marked A and B with two vertical stripes of the same shade of gray, it becomes apparent that both squares are the same color.

The human visual system needs to determine the color of objects in the world. In this case the problem is to determine the shade of the squares on the checkerboard. Just measuring the light coming from a surface (i.e., the luminance) is not enough: a shadow dims a surface, so that a white surface in a shadow may be reflecting less light than a black surface in full light.

Reference:

Adelson, E. (1995) Checker Shadow Illusion. http://web.mit.edu/persci/people/adelson/checkershadow_illusion.html.



Next, we will discuss perception. Why do we study perception? We know that humans perceive data. But how is the data perceived? More importantly, how is a visualization perceived? How do we know the visualization we have created is not perceived differently by different users?

Studying perception will help us control the data presentation better and even harness the power of human perception. Studying perception can:

- Help the design of sensory environments, such as the design of lighting and acoustics of a cockpit
- Help build artificial devices to aid sense organ defects, such as hearing aids and glasses
- Help build devices to "trick" the senses, such as virtual realities
- Aid medicine study, such as repairing broken sense organs and nerve systems, or preventing defects

What is a Color Model?

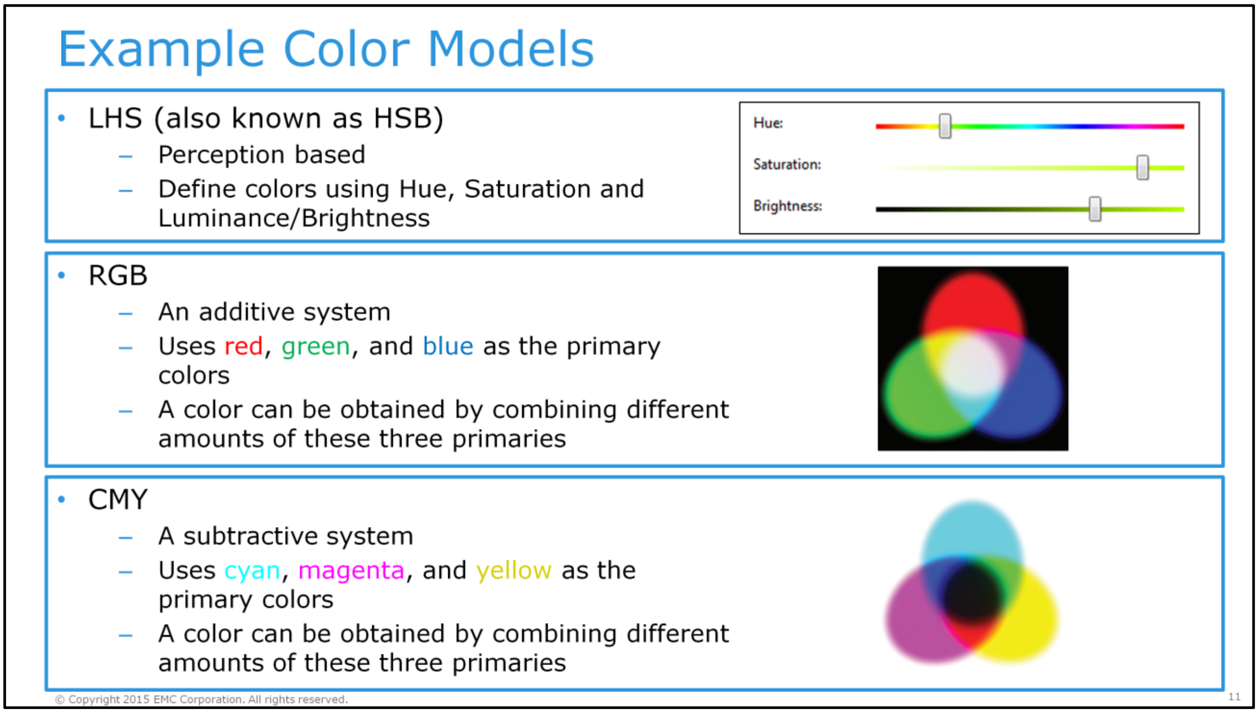
- The way the human eye perceives colors suggests that color is a 3-dimensional entity
- Color model
  - Is an abstract 3-D mathematical system for representing color
  - Defines three primary colors along three dimensions
  - Is typically limited in the range of colors they can represent and often can't represent all colors in the visible spectrum
- Example color models
  - Perception based (such as LHS)
  - Additive systems (such as RGB)
  - Subtractive systems (such as CMY)

Rods and cones are two types of photoreceptors in the human eyes. The rod system is extremely sensitive to light and allows us to see objects in low-light environments. Conversely, the cone system is relatively insensitive to light and allows us to see color.

The retina's three types of cones allow us to see the three primary colors, blue, green, and red, and their combined colors. This suggests that color is a three-dimensional entity.

Color model is an abstract mathematical system for representing colors in three-dimensional abstractions. The three primary colors are defined along the three dimensions. Note that color models are typically limited in the range of colors they can represent and hence often can't represent all colors in the visible spectrum.

Some example color models include perception-based (such as LHS), additive systems (such as RGB) and subtractive systems (such as CMY).



LHS decomposes color according to perception rather than according to how it is physically sensed. A color in LHS is defined by:

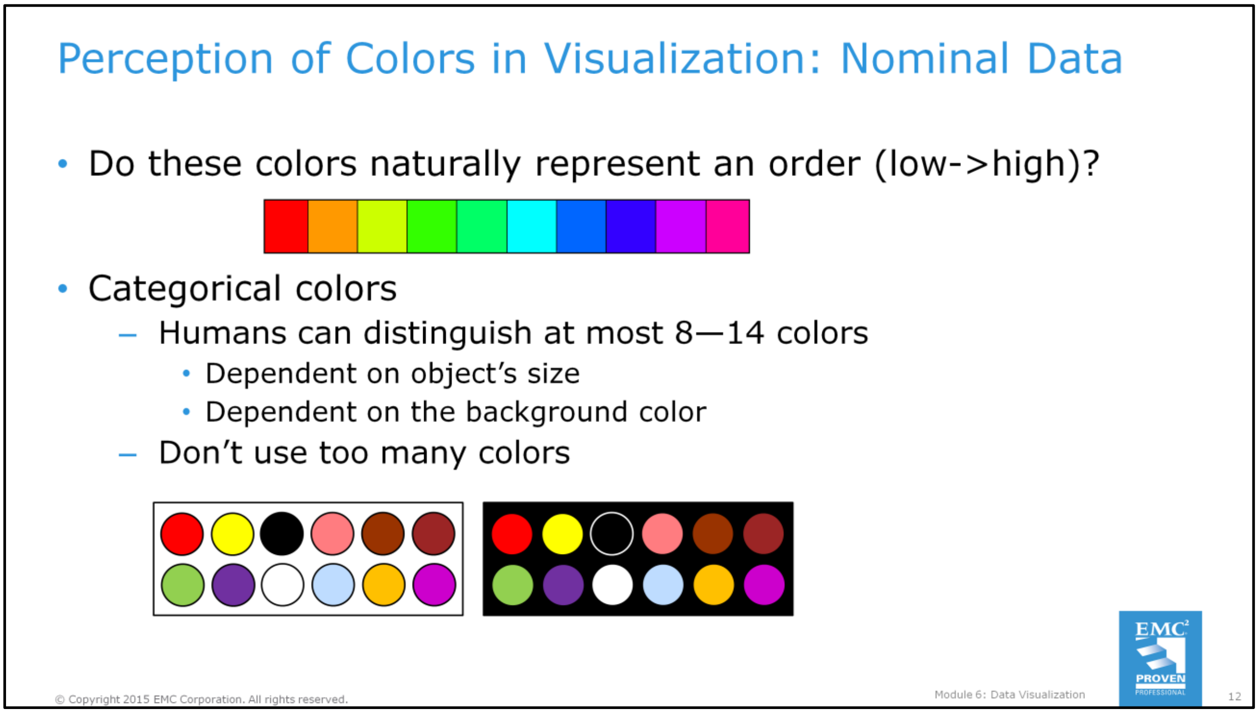
- Hue (the chromaticity or pure color; the vertical spike in the spectral density curve)

- Saturation (the vividness or dullness of the color; the proportion of the color lying within the spike)

- Luminance/Brightness (the intensity of the color; the area under the spectral density curve)

RGB is an additive color model system. It assumes that light is used to generate colors. It uses red, green and blue as the primary color. Human perception is additive since black is the absence of light and white the presence of all wavelengths of light. Example devices that use RGB include computer monitor, LCD screens and projectors.

CMY is a subtractive color model system. It assumes that pigment is used to generate colors. It uses cyan, magenta, and yellow as the primary colors. White light is reflected off of an object which “absorbs” or “subtracts” certain wavelengths. The viewer then sees the light that is not absorbed. The absence of pigment is “white” (blank piece of paper) while the presence of all types of pigment is “black.” Example devices that use CMY include paint, pencils, pen, and printers.

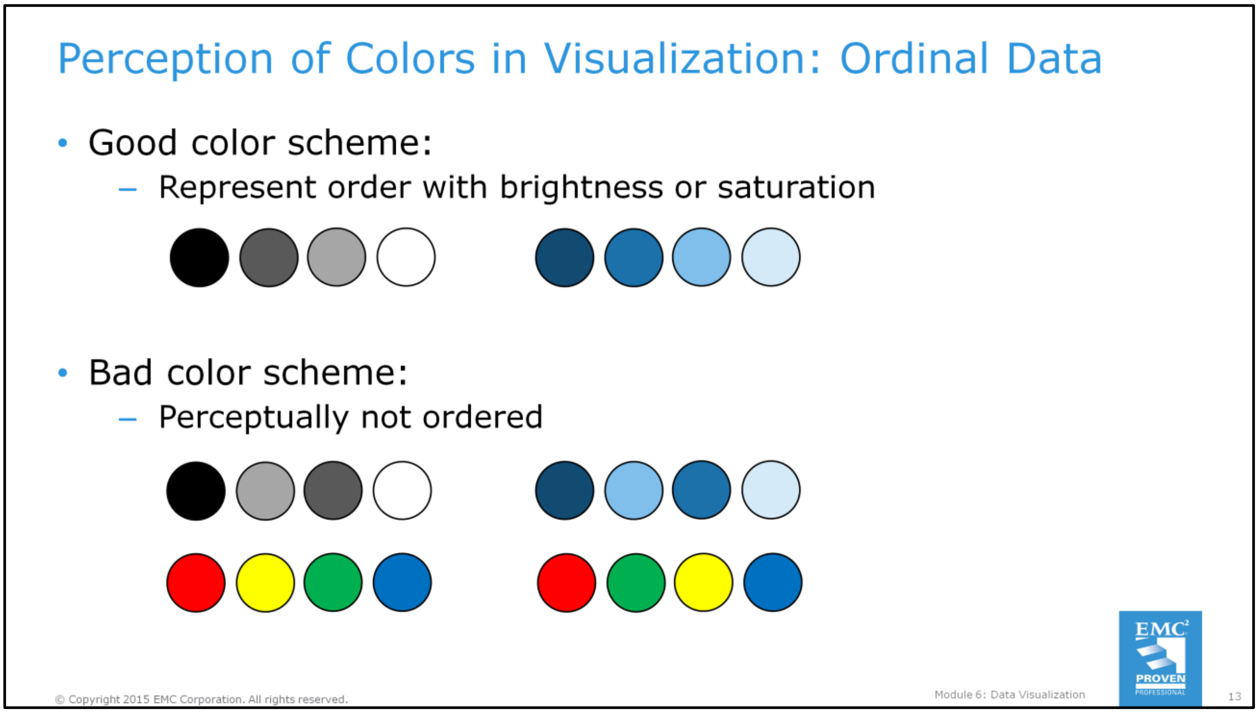


The rainbow color map is not a good choice for ordinal data, because perceptually we cannot associate rainbow colors with orders. In fact, research has shown that the rainbow color map is rarely the optimal choice when displaying ordinal data. The rainbow color map confuses viewers through its lack of perceptual ordering, obscures data through its uncontrolled luminance variation, and actively misleads interpretation through the introduction of non-data-dependent gradients.

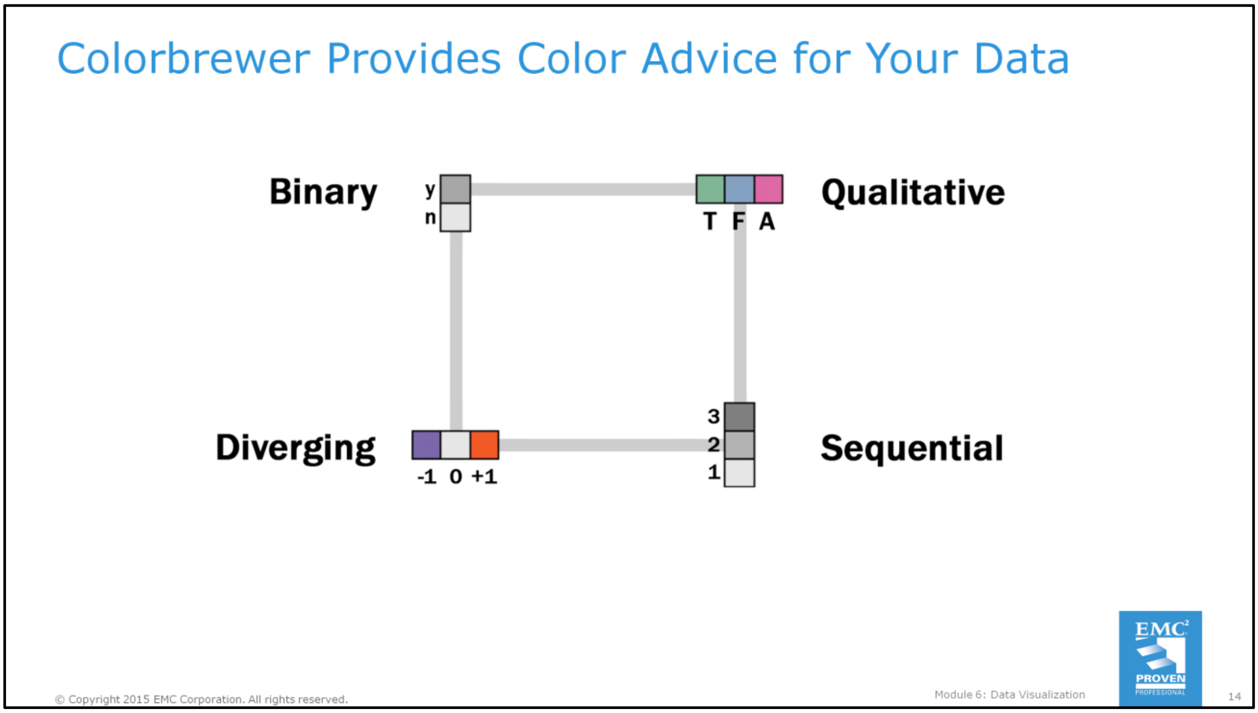
The rainbow color map is OK for nominal (or categorical) data. However, you need to be cautious not to go overboard with colors. As the number of colors being chosen increases, it’s likely more of these colors are similar to each other. Therefore, it would bring more confusion to viewers when more colors are used. Humans can generally only distinguish between 8-14 colors; however, this range may change depending on the object size and the background color.

Reference:

Borland, D., & Taylor II, R. M. (2007). Rainbow color map (still) considered harmful. *IEEE computer graphics and applications*, *27*(2), 14-17. http://people.renci.org/~borland/pdfs/RainbowColorMap_VisViewpoints.pdf.



A good color model for visualizing ordinal data avoids the use of rainbow-like color maps, and represents the order of the data through brightness or saturation. In addition, a good color scheme uses perceptually ordinal colors with linear, perceptual increments, even though the entire color spectrum of the scheme may contain many just-noticeable differences (JNDs).



The colorbrewer provides advice for the preferable color scheme of a specific situation. It's widely used for coloring geospatial maps although the idea can be used in other places such as scatter plots. The colorbrewer contains four basic colors schemes: binary, qualitative, sequential, and diverging.

Binary color schemes show nominal differences that are divided into only two categories, such as yes or no, true or false, etc.

Qualitative color schemes use differences in hue to represent nominal differences. For example, the class labels could be:

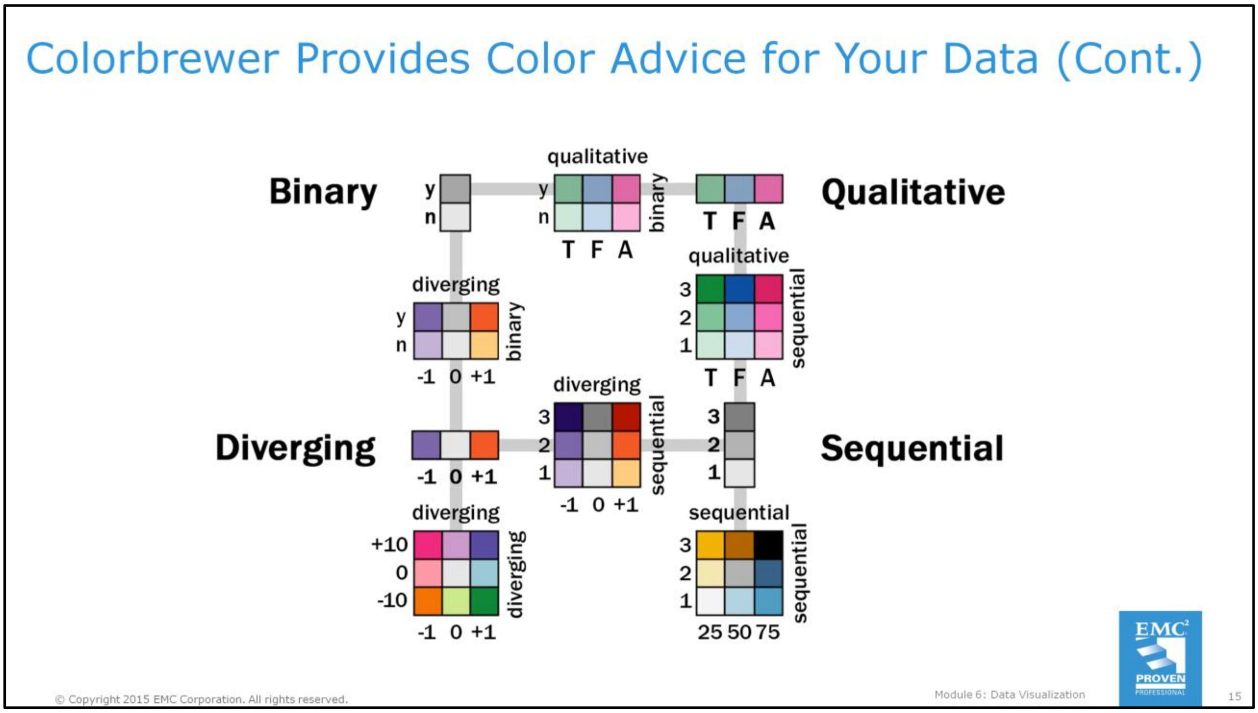
(1) true (T), false, (F) and ambiguous (A), or

(2) Races such as Hispanic, White, Black and Asian.

When using a qualitative color scheme for the map, assign the darkest, and most saturated hues in the scheme to categories that warrant emphasis.

In **sequential color schemes**, data classes are logically arranged from high to low. This stepped sequence of categories is represented by sequential lightness steps. Low data values are usually represented by light colors and high values are represented by dark colors. Transitions between hues may be used in a sequential scheme, but the light-to-dark progression should dominate the scheme.

Diverging color schemes allow the emphasis of a quantitative data display to show progression outward from a critical midpoint of the data range. A typical diverging scheme pairs sequential schemes based on two different hues so that they diverge from one common midpoint, a light color, toward dark colors of different hues at each extreme. For example, a green to yellow diverging color scheme can be used on the map to visualize deviations above and below the median death rate from a disease.



By mixing and matching the four basic colors schemes, we have these additional color schemes:

In **qualitative-binary color schemes**, light and dark versions of each hue of the qualitative variable correspond to the binary variable categories. A voter demographic map with hues indicating ethnic groups, in which darker hues are used for registered voters and lighter hues are used for non-voters, is well suited for a qualitative-binary color scheme.

In **qualitative-sequential color schemes**, the qualitative variable is represented with hues and the quantitative variable is represented with sequences of lightness steps within each hue. Population percentages (sequential) of varied dominant ethnic groups or religions (qualitative), for example, are well represented by a qualitative-sequential color scheme.

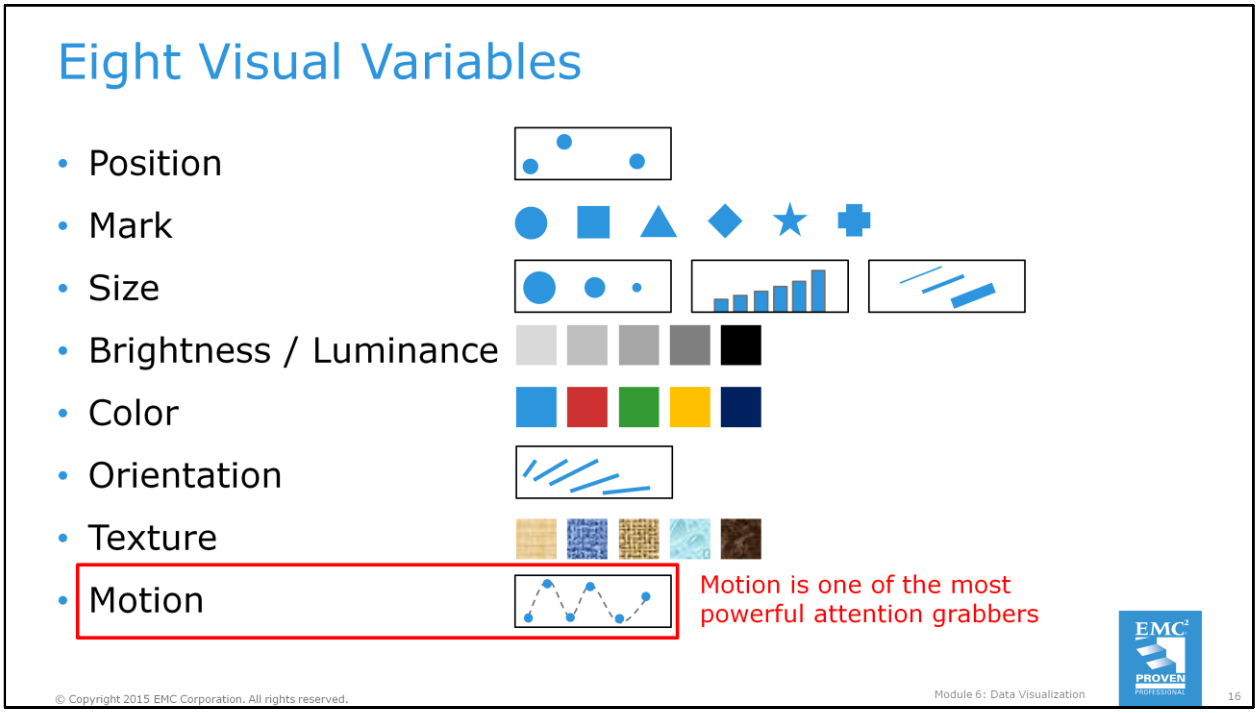
Sequential-sequential color schemes are the logical mix of all combinations of the colors in two sequential schemes. For example, data on educational attainment crossed with crime rate categories is well represented by a sequential-sequential color scheme.

Diverging-diverging color schemes are the only two-variable schemes that depart from the idea of a direct overlay of the component one-variable schemes. Different moderately-dark hues at each of the four corners of the legend represent categories that are extremes for both variables. A very light or white color at the center of the legend creates an appropriately light color for the class that contains the critical value or midpoint of both variables. The remaining colors are lighter than the corners, because they contain the midpoint of one of the two variables, and they are transitional hues that lie between their adjacent hues. Areas above and below the poverty line in 1970 and 1990, for example, are well represented by a diverging-diverging color scheme.

Diverging-binary and **diverging-sequential color schemes** have the same perceptual characteristics. The success of the schemes hinge on the large contrast range available in the lightness dimension. Large lightness steps are used for the binary or sequential variable. Smaller lightness steps, that are bolstered by a change in hue, represent the diverging component of the scheme within each large lightness step of the comparison variable. For example, data on cancer rates above and below a mean rate (diverging) and air pollution levels (sequential) are well represented by a diverging-sequential color scheme.

References:
Brewer, C. Color Scheme Types and combinations: Overview. *Color Use Guidelines for Mapping and Visualization*. http://www.personal.psu.edu/faculty/c/a/cab38/ColorSch/Schemes.html

Brewer, C, & Harrower, M. (2013). Color Advice for Cartography. *ColorBrewer2.0*. http://colorbrewer2.org/



There are eight basic visual variables: position, mark (or shape), size, brightness, color, orientation, texture, and motion. A visualization in practice usually combines several visual variables. Depending on the type of visualizations, some of the combinations are more effective than the others.

Of the eight visual variables, motion is one of the most powerful attention grabbers. However, you need to be careful not to use too much motion in order to avoid confusion and ensure the correct message is delivered.

Features to Categorize the 8 Visual Variables

- Selective
  - Is a change enough to allow us to select it from a group?
- Associative
  - Is a change enough to allow us to perceive them as a group?
- Ordinal
  - Are changes in this variable perceived as ordered?
- Quantitative
  - Is there a numerical reading obtainable from changes in this variable?
- Separating
  - Across how many changes in this variable are distinctions perceptible?

Module 6: Data Visualization 17

Here we list five features that we can use to categorize the eight visual variables.

The **selective** visual variables allows us to spontaneously divide different data values into distinguished groups. They are used to code nominal data.

For **associative** visual variables, all factors have same visibility. They are used to code nominal data.

The **ordinal** visual variables allow us to spontaneously place different data values in an order. They are used to code ordinal data.

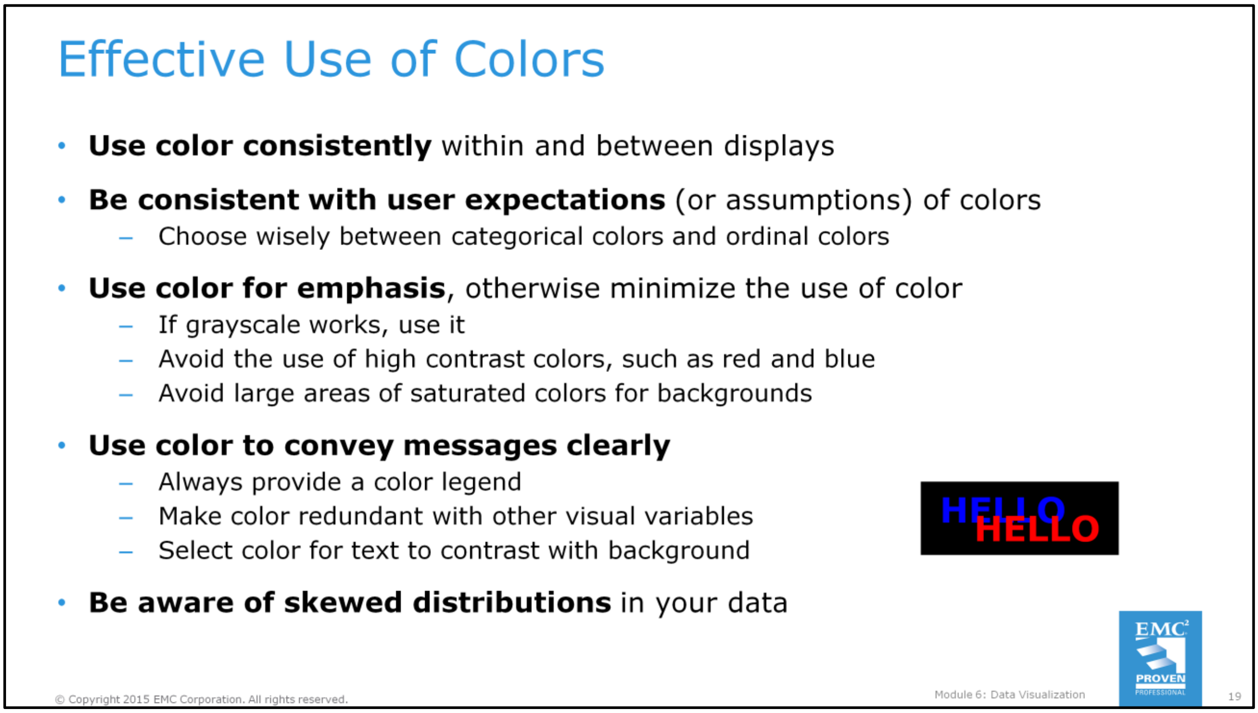
The **quantitative** visual variables obtain a direct association between numerical values and relative sizes. The are used to code quantitative or proportional data.

The **separating** feature defines the number of distinct changes in a visual variable that are human perceptible.

Feature Selection Guide for the 8 Visual Variables

	Position	Mark	Size	Bright-ness	Color	Orientation	Texture	Motion
Selective	YES	YES	YES	YES	YES	YES	YES	YES
Associative	YES	YES	YES	YES	YES	YES	YES	YES
Ordinal	YES	NO	YES	YES	NO	YES	NO	?
Quantitative	YES	NO	CAUTION	CAUTION	NO	CAUTION	NO (density YES)	YES
Separating	?	?	≤ 5	≤ 10	8—14 max	≤ 5	?	?

This table shows how the eight visual variables are categorized using the five features. The table provides a guide for selecting appropriate visualizations depending on the data, while the highlighted cells portray caution areas. Chart junk is likely to emerge if the caution areas (in the table) are not handled properly. Consider size as a quantitative visual variable in a scatter plot, as an example. Let's say we associate a quantitative variable, such as income, with sizes of the circles in the scatterplot. Instead of using a different circle size for each distinct income value, we should place income into 5 bins and use only 5 different sizes to represent the data. The distinctions are more perceptible in this way.



Color is a powerful tool. It provides an additional dimension to visualize the data. But this doesn't mean that you can go overboard with colors as doing so will only introduce confusion to the viewers.

If your visualization is going to be displayed or printed on different devices, make sure these devices match the outputs. For example, the color scale, the color spectrum, the primary colors and the combination of colors should all match across the devices. In some cases, it's good to provide device calibration to guarantee the cross-device compatibility.

When designing your visualization, keep in mind the user's expectations of colors. For example, if your user is from military, they may associate red color with alert. If your user is from the finance industry, they may assume red means loss. It would be counter-intuitive to design a visualization on stock prices that use red for prices going up and green for prices going down.

You need to be cautious about when to use colors. As long as grayscale works, use grayscale instead of colors – this will warrant the correct display of color scales whether your visualization will be displayed on screen, scanned, photocopied, or printed on black-and-white printers. Also minimize the use of colors. As discussed earlier, humans can only distinguish at most 8-14 different colors.

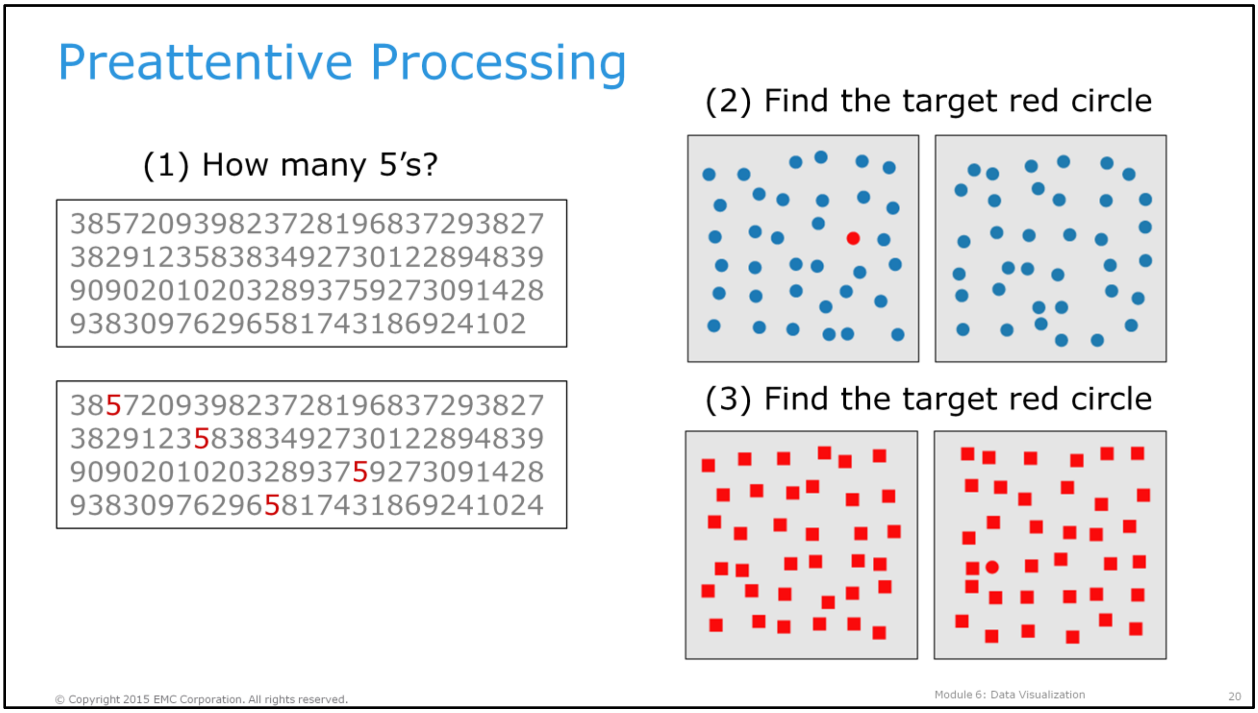
When color is used in your visualization, make sure to provide a color legend for the user.

Be conscious of the type of data you are visualizing. If the data is nominal, don't use colors that imply an order. On the other hand, if the data is ordinal, don't used categorical colors (such as rainbow colors).

It's often a good practice to make color redundant with other visual variables, such as position, mark, size, or texture. For example, you can use both color and shape to visualize nominal data, such as race. In case color cannot be perceptually comprehended (due to color-blindness of the user or the absent of colors on a display, for example), the other visual variable such as shape can provide the backup to reinforce the delivery of the message.

You should avoid the large areas of saturated colors for backgrounds. For example, a bright-red background is generally less aesthetically-pleasing than a light gray background. Texts should use colors that are in contrast with the background. Also avoid the use of high contrast colors. For example, it's not a good practice to use red and blue colors on top of the black background.

Be aware of skewed distributions in your data. The distribution of colors should always match the distribution of your data. If you used an evenly distributed color scheme to represent your skewed data, it would be difficult to tell the real distribution of the data. As a result, it would create distortions of the data and you may send the wrong message to your viewers.



Preattentive processing is a kind of uncontrolled perception. Certain basic visual properties are detected immediately by a low-level visual system. Preattentive perception happens fast and automatically. Such a task is performed in parallel, often less than 200 to 250 milliseconds on a complex display. In comparison, eye movements take at least 200 milliseconds to initiate. Some effects pop out and are the result of preconscious visual processes.

The slide shows three examples of preattentive tasks. The first two examples demonstrate that color (or hue) is preattentive.

The first example involves a task of telling how many 5's are in the text. By coloring the 5's red, it allows us to quickly find where they are.

The second example of a preattentive task is the detection of a red circle in a group of blue circles. The target object has a red visual property that the blue distractor objects do not. A viewer can tell at a glance whether the target is present or absent.

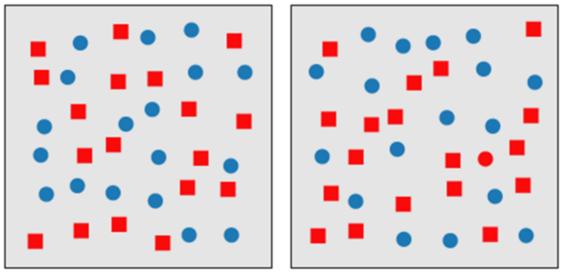
The third example demonstrates that form, such as curvature, is also preattentive. The task is to detect a target red circle among the distractors of red squares. This is a little less obvious than the previous example, but still we can tell at a glance whether the target is absent or present.

Reference:

Healey, C. G., & Enns, J. T. (2012). Attention and visual memory in visualization and computer graphics. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(7), 1170-1188. Available: http://www.csc.ncsu.edu/faculty/healey/PP/index.html.

Conjunction Target

- A target made up of a combination of non-unique features
- **Generally cannot be detected preattentively**
- Below is an example of a conjunction search for a target red circle (in a sea of red squares and blue circles as distractors)



Target red circle is absent   Target red circle is present

Conjunction target is a target made up of combinations of nonunique features. Studies have shown that most conjunction targets cannot be detected preattentively. Viewers have to perform a time-consuming serial search through the display to confirm the presence or absence of the target.

The slide shows an example of a conjunction search for a target red circle among distractors of many red squares and blue circles.

Reference:

Healey, C. G., & Enns, J. T. (2012). Attention and visual memory in visualization and computer graphics. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(7), 1170-1188. Available: http://www.csc.ncsu.edu/faculty/healey/PP/index.html.

Data Visualization Principles

- Show the data
- Provoke thought
- Avoid distorting what the data says
- Present many numbers in a small space
- Make large data sets coherent
- Encourage eyes to compare data
- Reveal data at several levels of detail
- Serve a reasonably clear purpose
- Be closely integrated with statistical and verbal descriptions of the dataset

Edward Tufte is an American statistician and professor at Yale University. He is noted for his writings on information design and as a pioneer in the field of data visualization. In the first chapter of his book *Visual Display of Qualitative Information*, he shows a set of principles good data visualization should follow.

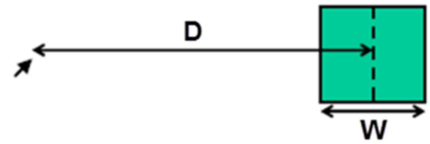
- Show the data. Let your visual *reveal* the data. There have certainly been successful "infographics" that don't actually display any real data, but that's not what data visualization is about.
- Induce the viewer to think about the substance rather than about methodology, graphic design, the tech of graphic production, or something else. In the modern times, good data visualization should in addition provoke thought and make the audience think.
- Avoid distorting what the data has to say. Avoid the use of “chart junk” such as 3D pie charts. We will discuss this later in the lesson.
- Present many numbers in a small space. A good example would be the Minard’s Flow Map we have seen earlier.
- Make large data sets coherent. Well-designed visualizations exist as a symbiosis between smart quantification, and beautiful and elegant design.
- Encourage eyes to compare data. Add controls to your visualization to make it interactive. Interactivity makes comparing data in a visualization fun and engaging.
- Reveal data at several levels of detail, from a broad overview to the fine structure. Allow your audience to understand the big picture quickly, but have the opportunity to pick out some details.
- Serve a reasonably clear purpose: description, exploration, tabulation or decoration. Use your visualization to tell a story to the audience that cares.
- Closely integrate the statistical and verbal descriptions of a data set. While your data visualization should be able to speak for itself, it’s best to include a link to the raw data, and some explanation of the interesting findings.

Reference:

- Tufle, E. R. (1983). The visual display of quantitative information. *CT Graphics, Cheshire*.
- Data Visualization Principles: Lessons from Tufte -- http://moz.com/blog/data-visualization-principles-lessons-from-tufte

Fitts's Law

- The time to select an item using a pointing device is proportional to the item's size and the distance to the item
- $MT = a + b \log_2 \frac{2D}{W}$
  - MT: movement time
  - a, b: constants
  - D: distance of movement from start to target center
  - W: width of the target



- The design of an interactive visualization system should account for Fitts's Law

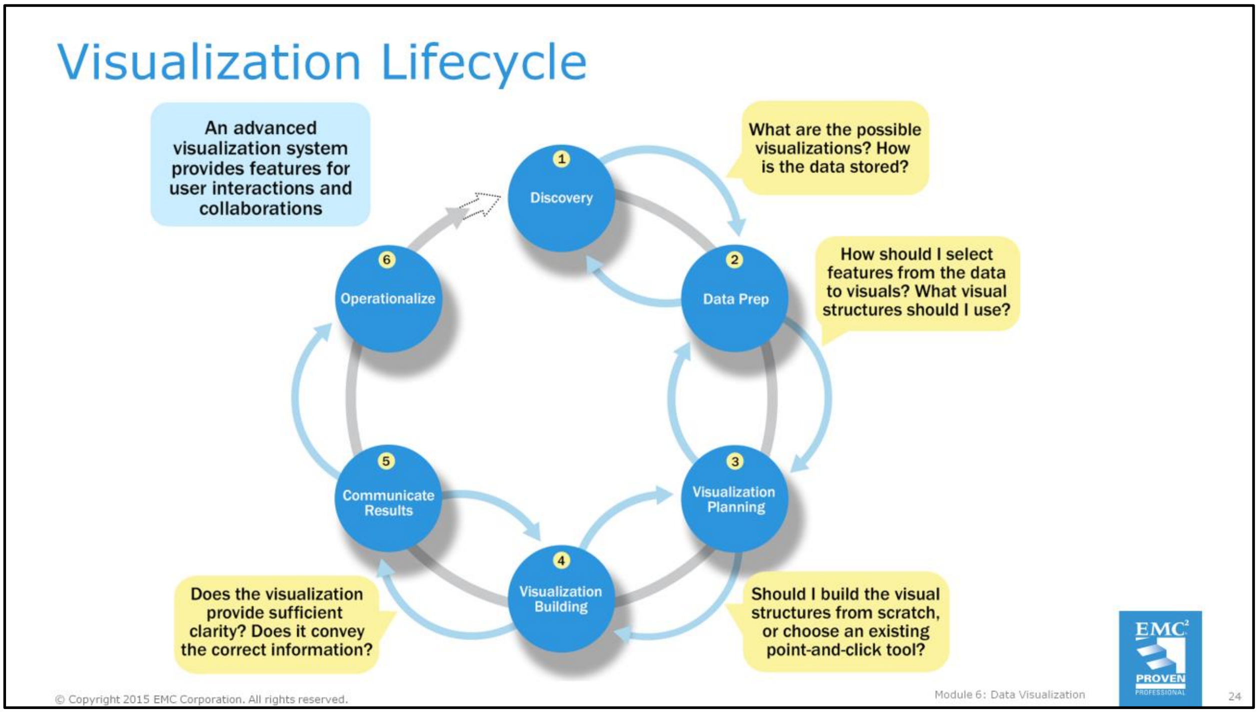
Fitts's law was developed in 1954 by Paul Fitts. It is considered a fundamental law of the human sensory-motor system. Fitts's law models human movement in human-computer interaction and ergonomics, and predicts that the time required to rapidly move to a target area is a function of the distance to and the size of the target. Fitts's law is used to model the act of pointing, either by physically touching an object with a hand or finger, or virtually, by pointing to an object on a computer display using a pointing device.

Keep Fitts's law in mind when you design an interactive visualization system. A few implications of Fitts's law include:

- Buttons and menus should be in reasonable sizes
- GUI of similar features should be grouped together or placed close to each other
- Edges and corners of the display should be easy to reach
- Popup menus can usually be opened faster than pull-down menus
- Pie menu, or dial, items are typically selected faster than linear menu items

Reference:

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47, 381-391.



On this slide is the visualization lifecycle. The callouts show some questions that need to be considered when designing a visualization system.

When moving from Phase 1 Discovery to Phase 2 Data Preparation, we need to decide what data to collect, and how the raw data should be transformed into something usable by the visualization system. In this step, we may need to deal with data issues such as missing values, error in input or data that is too large to process. We also need to decide the most suitable way to store the data. Datastores can be as simple as a CSV file, a traditional RDBMS, or NoSQL such as MongoDB and Hadoop HDFS.

From Phase 2 Data Preparation to Phase 3 Visualization Planning, we need to decide how we want to map the data to visuals. Should the output visual be a static image, or an interactive system? If the purpose is to design an advanced visualization system, then it’s better to provide features for user interactions and collaborations. What specific visual structures should be used to represent the data, e.g., histogram, heatmap, scatter plot, or geometry map? Without the careful planning in this step, it’s easy to develop a visualization that doesn’t make sense, or one that conveys the wrong information.

From Phase 3 Visualization Planning to Phase 4 Visualization Building, we need to decide how to build the visual structures. We may want to use an existing point-and-click tool (such as Tableau public or Weave) to provide a quick prototype of the visualization. Alternatively, we may want to build the visual structures from scratch using tools such as D3, Processing, ggplot, or matplotlib.

From Phase 4 to Phase 5, we have constructed the visuals and it’s time to verify the output. A good visualization should provide sufficient expressiveness and effectiveness. An expressive visualization presents all the information that’s needed. Expressing additional information may be dangerous in that it may introduce noise which would interfere with presenting the essential information. An effective visualization can be interpreted quickly and accurately and it can be rendered in a cost-effective manner. If the output visualization is not satisfactory, we have to rollback the workflow to previous phases.

The following video (Getting started with Weave Desktop) shows how to use an open source point-and-click software to import data sets and quickly turn into visualization for insights: http://youtu.be/dPXI3SJwOSM.

Check Your Knowledge

- What is data visualization and why is it important?
- What are the eight visualization variables? Which ones are not recommended for representing ordinal data?
- What is preattentive processing? Can conjunction target be detected preattentively?
- What are the data visualization principles?

Write your answers here.

Lesson 1: Summary

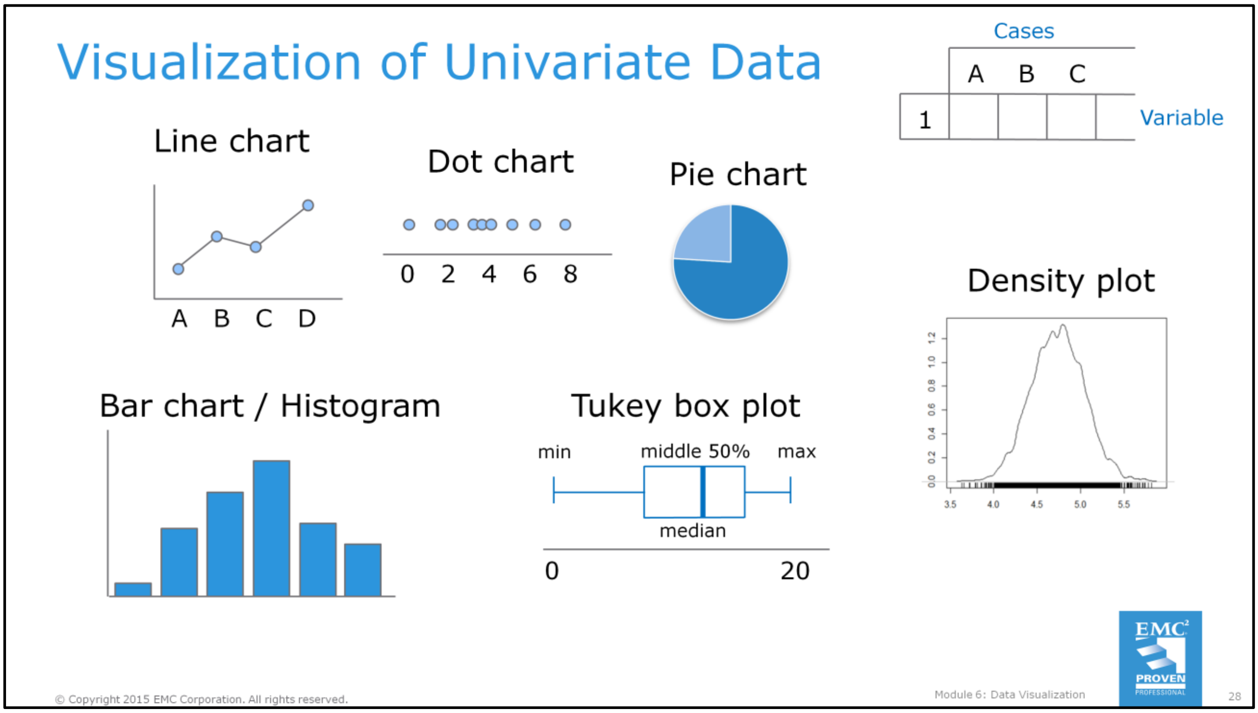
During this lesson the following topics were covered:

- Perception and colors
- Eight visual variables
- Visualization principles

Lesson 2: Visualization of Multivariate Data

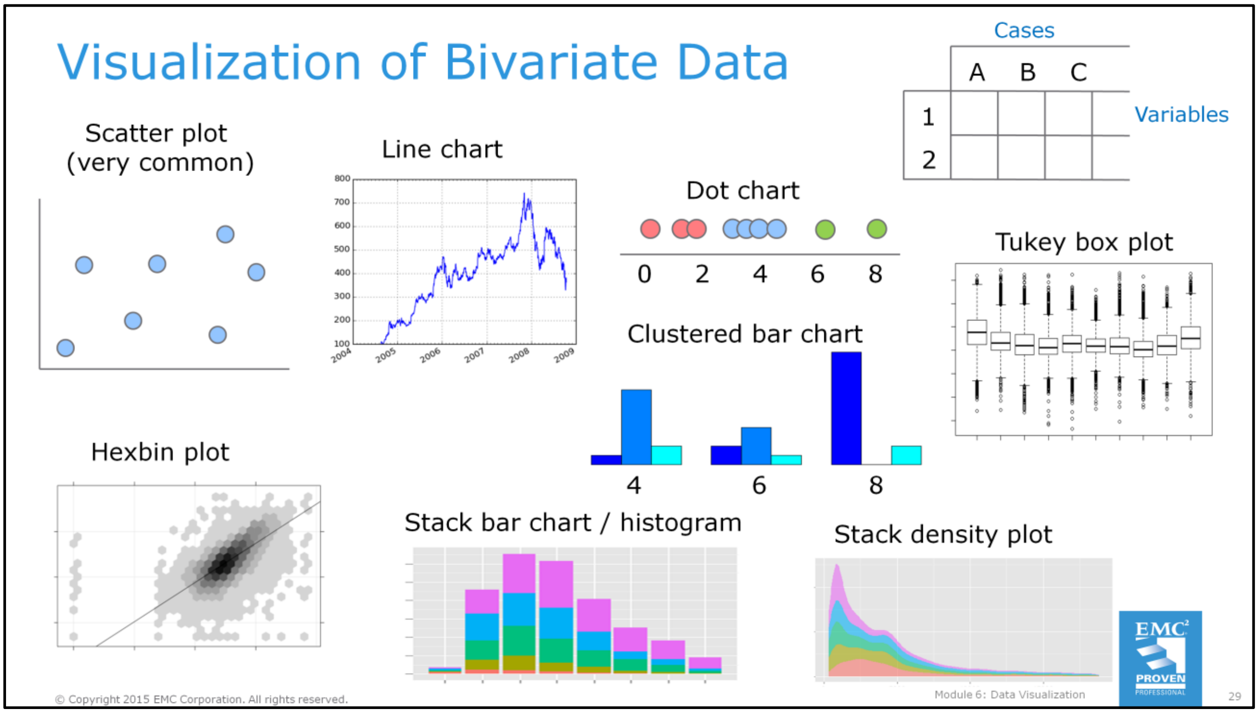
This lesson covers the following topics:

- Visualization of multivariate data
- Redesign of charts
- Lie factor



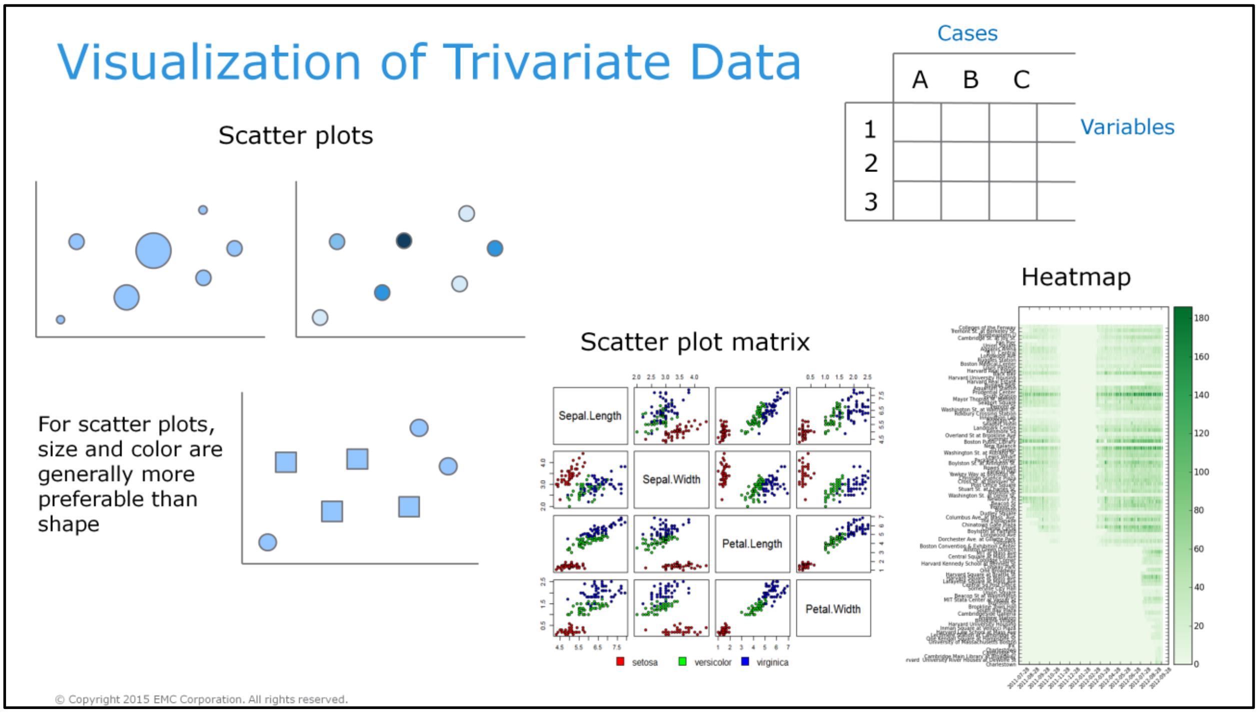
Here are some example visualization tools for univariate (1-D) data:

- Line chart
- Dot chart
- Pie chart
- Bar chart (or histogram)
- Tukey box plot
- Density plot



Here are some example visualization tools for bivariate (2-D) data:

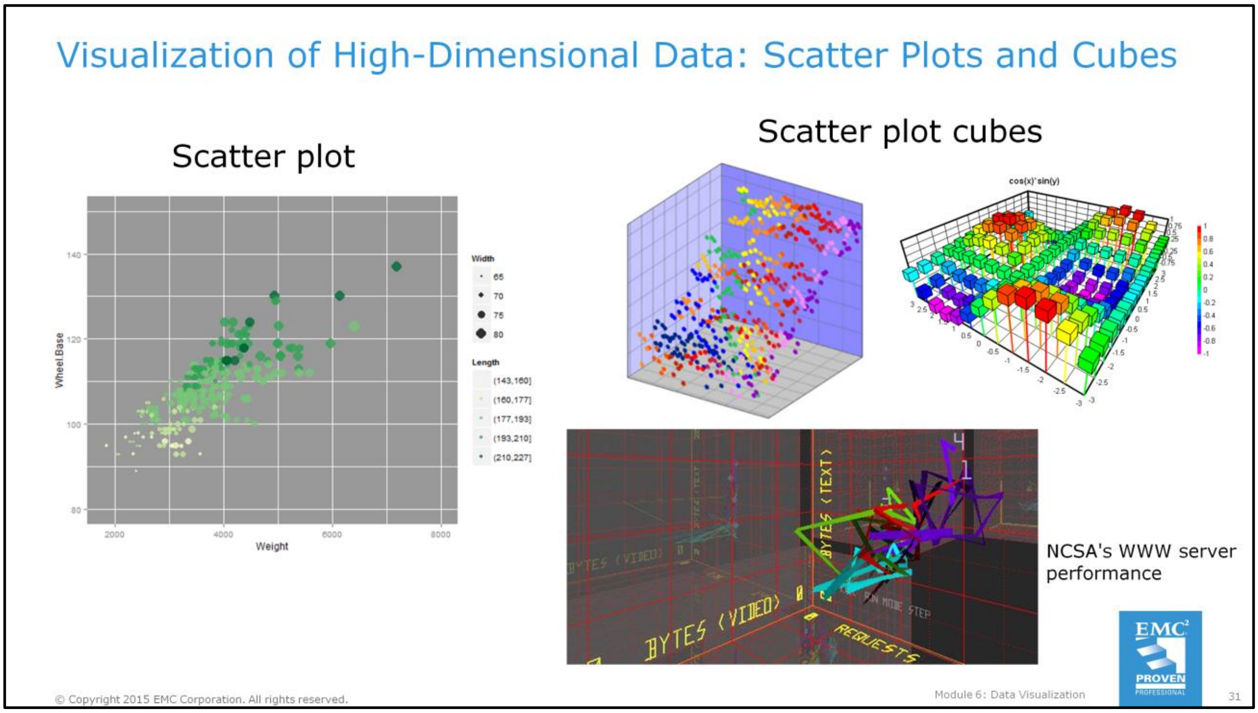
- Scatter plots are very commonly used
- Line chart
- Dot chart (with colors)
- Clustered bar chart
- Stacked bar chart
- Tukey box plot
- Hexbin plot
- Stack density plot



Here are some example visualization tools for trivariate (3-D) data:

- Scatter plots that use size, color, or shape as the third dimension
- Scatter plot matrix
- Heatmap

Scatter plot matrices are useful in visualizing the relationship between two attributes; however, they cannot visualize multiple attributes that are dependent. The next few slides show some tools that can visualize high-dimensional data.

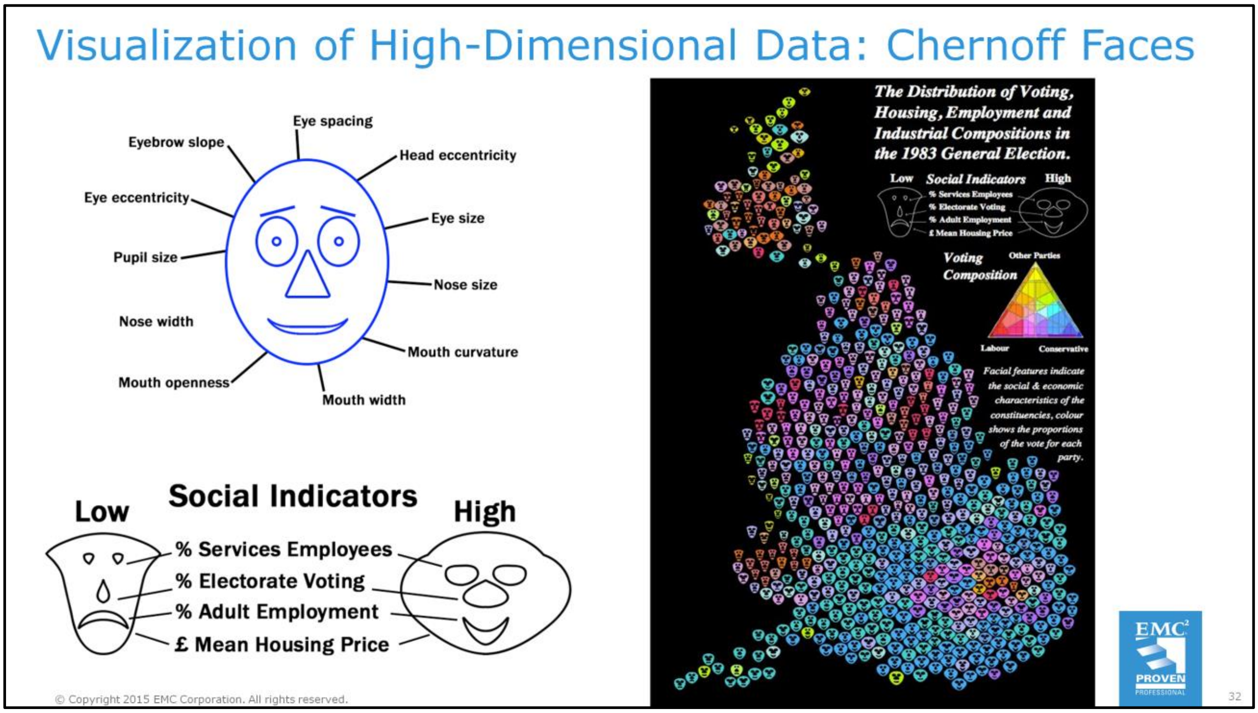


Here are some examples of using scatter plots and scatter plot cubes to visualize high-dimensional data.

Note that when using scatter plot cubes, it's better to make the visualization interactive so that the cubes can be rotated to better show the position of the data points. Also keep in mind that the rainbow colors are only suitable for showing categorical data, as we discussed earlier in the module.

Reference:

Lamm, S. E., Reed, D. A., & Scullin, W. H. (1996). Real-time geographic visualization of World Wide Web traffic. *Computer Networks and ISDN Systems*, 28(7), 1457-1468. http://www.ra.ethz.ch/CDStore/www5/www365/overview.htm.



Chernoff faces is another visualization tool for high-dimensional data.[1] Chernoff faces display multivariate data in the shape of a human face. The individual parts, such as eyes, ears, mouth and nose represent values of the variables by their shape, size, placement and orientation. The idea behind using faces is that humans easily recognize faces and notice small changes without difficulty. Chernoff faces handle each variable differently. Because the features of the faces vary in perceived importance, the way in which variables are mapped to each feature should be carefully chosen. For example, eye size and eyebrow slant have been found to be more perceptible than other facial features.

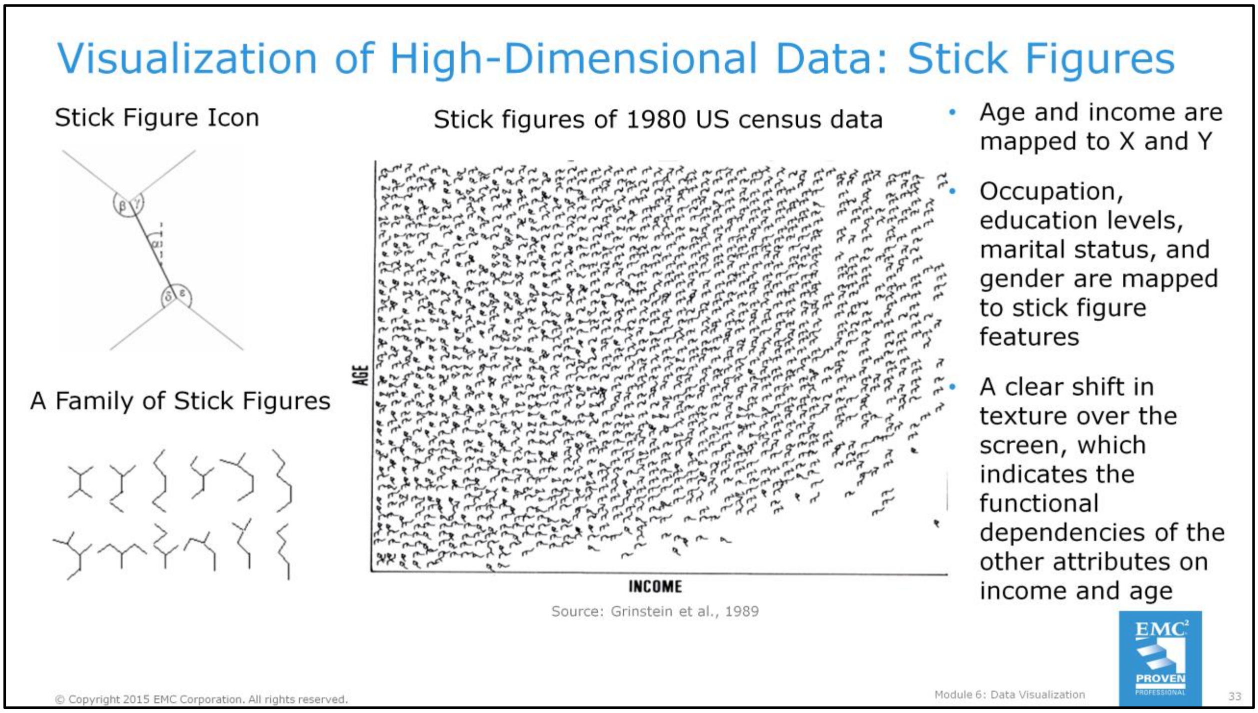
The right graph shows a visualization of using Chernoff faces to represent the results of data analysis about elections in Britain.[2]

One problem in using Chernoff faces is the danger of conveying unintentional emotional messages. When you are using faces, you are using emotions, so it’s important to be prepared to make emotional statements about your data.

References:

[1]Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, *68*(342), 361-368.

[2]Dorling, D. (2012). *The visualisation of spatial social structure*. John Wiley & Sons.



Stick figures are another visualization tool for high-dimensional data. Two variables are mapped to X, Y axes and other variables are mapped to limb lengths and angles. If the data is relatively dense with respect to the display, the resulting visualization of stick figures presents texture patterns that vary according to the characteristics of the data and are therefore, are detectable by preattentive perception.

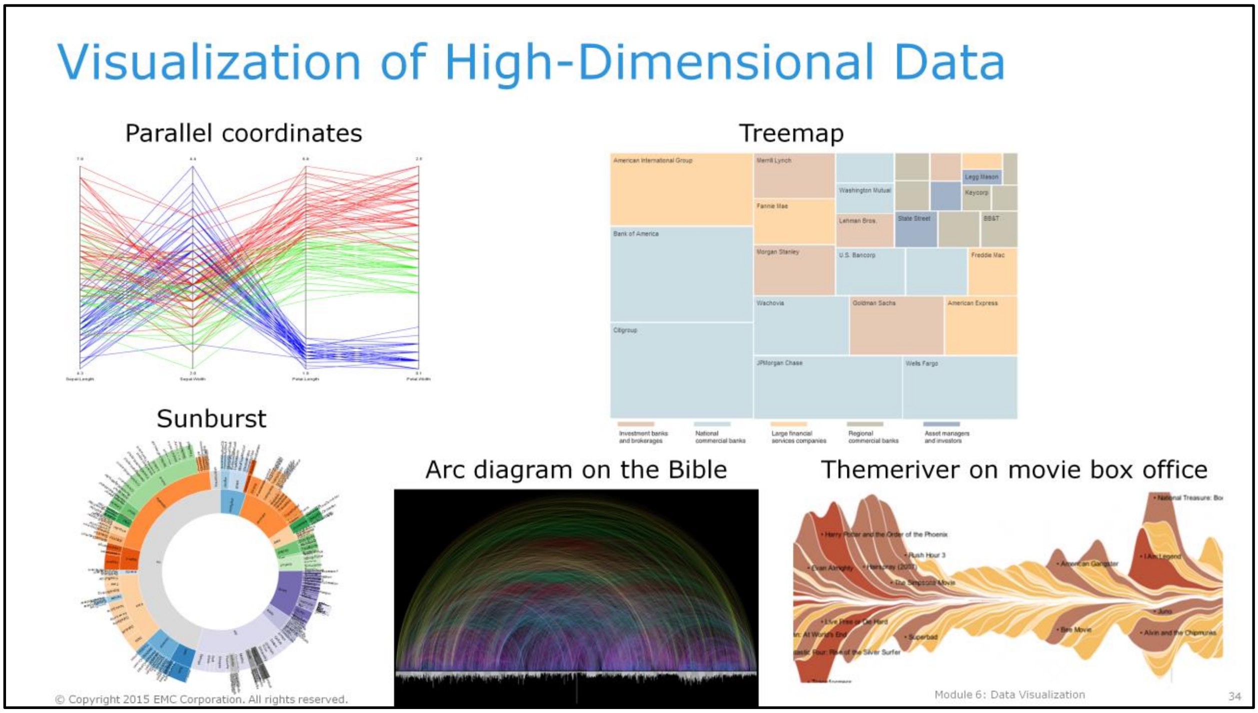
The slide shows an example of using stick figures to visualize the 1980 US census data. Age and income are mapped to X and Y axes. Occupation, education levels, marital status, and gender are mapped to stick figure features. The graph shows a clear shift in texture across the screen, which indicates the functional dependencies of the other attributes on income and age. The bottom right of the graph shows an outlier – a young and highly educated woman.

Chernoff-Faces and stick figures are two examples of **icon-based visualization techniques**. These techniques take advantage of the preattentive perception and they have the following characteristics:

- Can handle small to medium datasets with a few thousand data records, as icons tend to use a screen space of several pixels
- Can be applied to datasets of high dimensionality, but interpretation is not straightforward and requires training
- Variables are treated differently, as some visual features of the icons may attract more attention than others
- The way data variables are mapped to icon features greatly determines the expressiveness of the resulting visualization and what can be perceived
- Defining a suitable mapping may be difficult and poses a bottleneck, particularly for higher dimensional data
- Data record overlapping can occur if some variables are mapped to the display positions

Reference:

Grinstein, G., Pickett, R., & Williams, M. G. (1989, June). EXVIS: an exploratory visualization environment. In *Proc. Graphics Interface,* Vol. 89, pp. 254-261.



Here are five visualization tools for high-dimensional data: parallel coordinates, treemap, sunburst, arc view, and theme river.

Parallel coordinates is a very powerful tool for understanding multi-dimensional numerical datasets. To show data points with n attributes, a backdrop is drawn consisting of n parallel lines, typically vertical and equally spaced. A data point is then represented as a polyline with vertices on the parallel axes. The position of the vertex on the i-th axis corresponds to the i-th axis. The example on the slide visualizes the iris dataset of four attributes (sepal length, sepal width, petal length, and petal width) and the red, blue and green colors highlight the three different iris species.

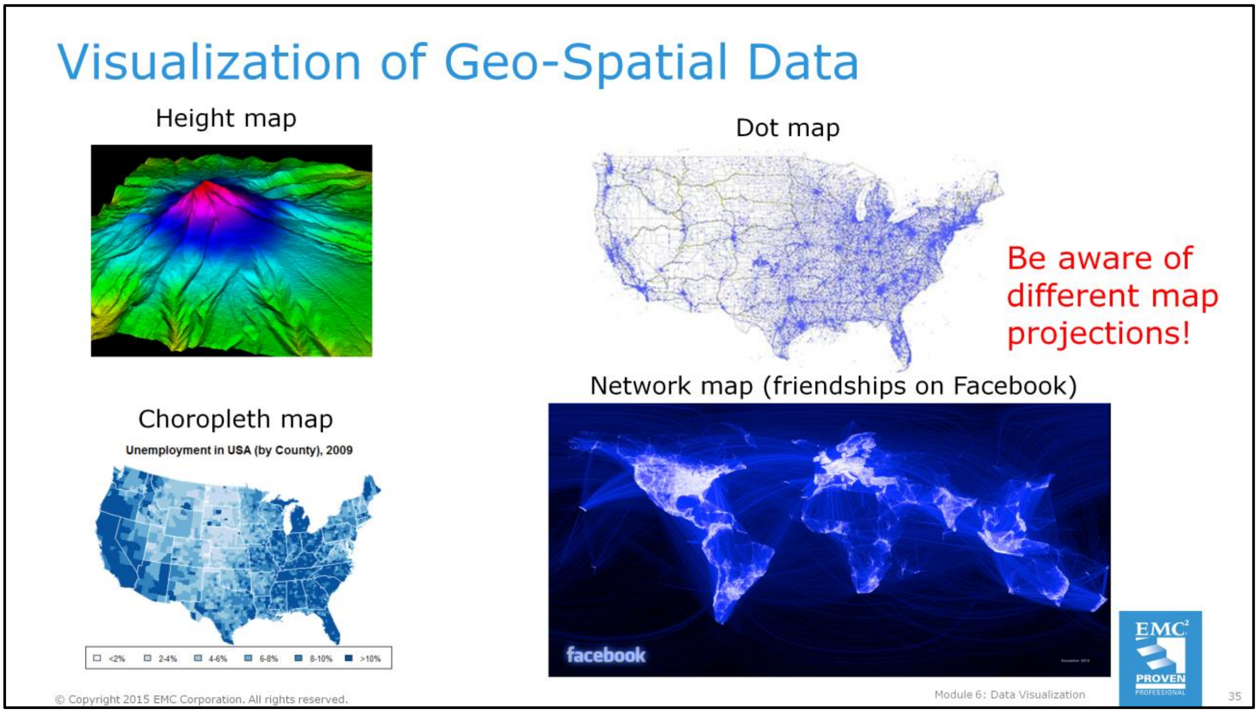
Both sunburst and treemap visualize hierarchical data. A treemap recursively subdivides an area into rectangles. The area of any node in the tree corresponds to its value. The example on the slide is taken from a New York Times article on a year of heavy losses in the financial industry.[1]

A sunburst is similar to the treemap, except it uses a radial layout. The root node of the tree is at the center, with leaves on the circumference. The area of each arc corresponds to its value.[2]

An arc diagram uses a one-dimensional layout of nodes, with circular arcs to represent links.[3] The example arc diagram on the slide visualizes the Bible.[4] Each of the 63,779 cross references found in the Bible is depicted by a single arc - the color corresponds to the distance between the two chapters, creating a rainbow-like effect.

Themeriver visualizes thematic variations over time across a collection of documents.[5] The horizontal axis represents the time and themes, or topics, are represented as colored currents within the river. The example on the slide shows the movie box office receipts from 1986 to 2008.[6]

References:
[1]http://www.nytimes.com/interactive/2008/09/15/business/20080916-treemap-graphic.html
[2]http://bl.ocks.org/mbostock/raw/910126/
[3]Wattenberg, M. (2002). Arc diagrams: Visualizing structure in strings. In*Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on* (pp. 110-116). IEEE.
[4]http://www.chrisharrison.net/index.php/Visualizations/BibleViz
[5]Havre, S., Hetzler, B., & Nowell, L. (2002). ThemeRiverTM: In search of trends, patterns, and relationships. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 9-20.
[6]http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html

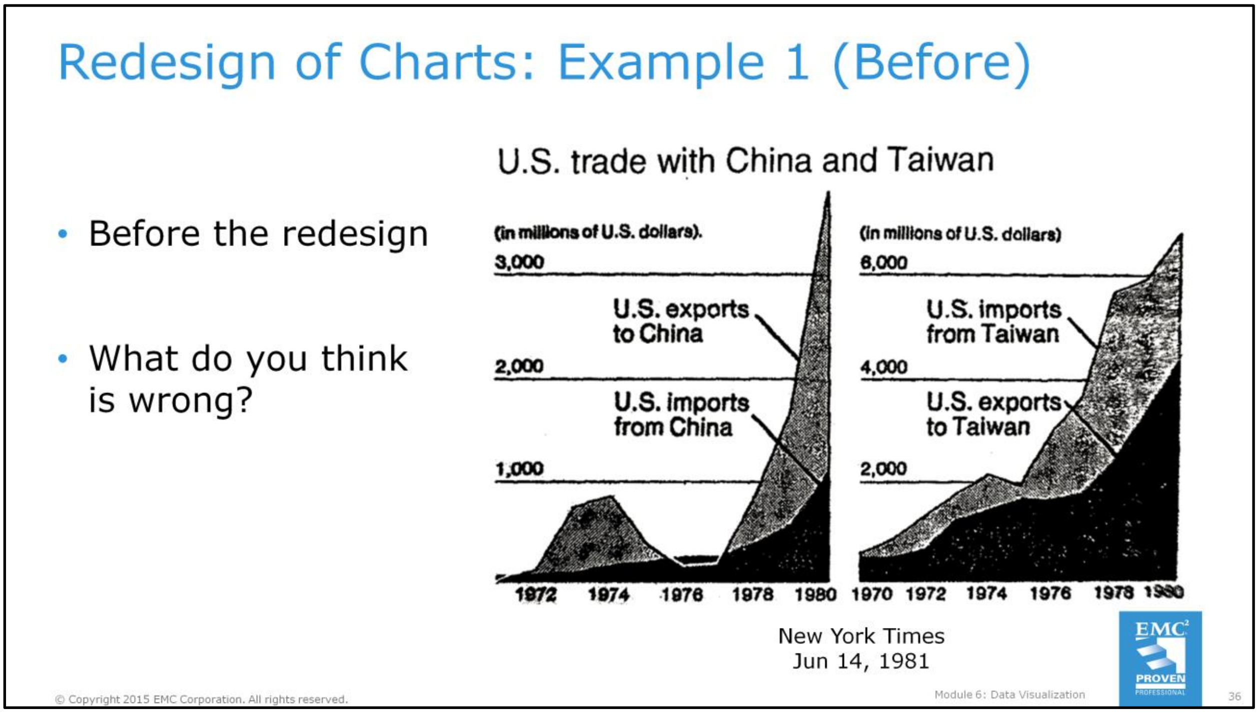


Here are some examples of visualizing geo-spatial data. When visualizing geo-spatial data, it's common to deal with coordinate systems. There are two common types of coordinate systems used in geographic information systems (GIS):

- A global or spherical coordinate system such as latitude–longitude. These are often referred to as geographic coordinate systems.
- A projected coordinate system based on a map projection such as transverse Mercator, Albers equal area, or Robinson, all of which (along with numerous other map projection models) provide various mechanisms to project maps of the earth's spherical surface onto a two-dimensional Cartesian coordinate plane. Projected coordinate systems are sometimes referred to as map projections.

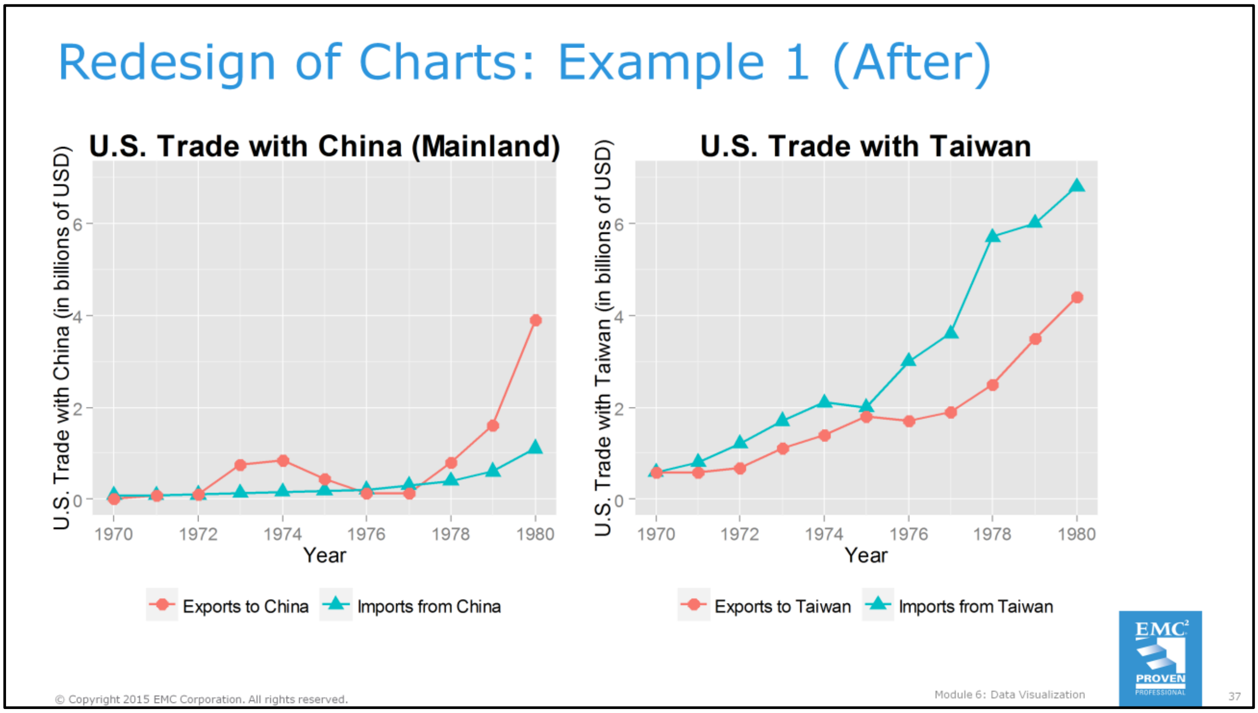
This site has a list of different map projections: http://egsc.usgs.gov/isb//pubs/MapProjections/projections.html.

Note that when visualizing data on a geo-spatial map, you need to be cautious of the different map projections. The data attributes and the shape file properties have to match the correct map projections in order to accurately visualize the data on the map.

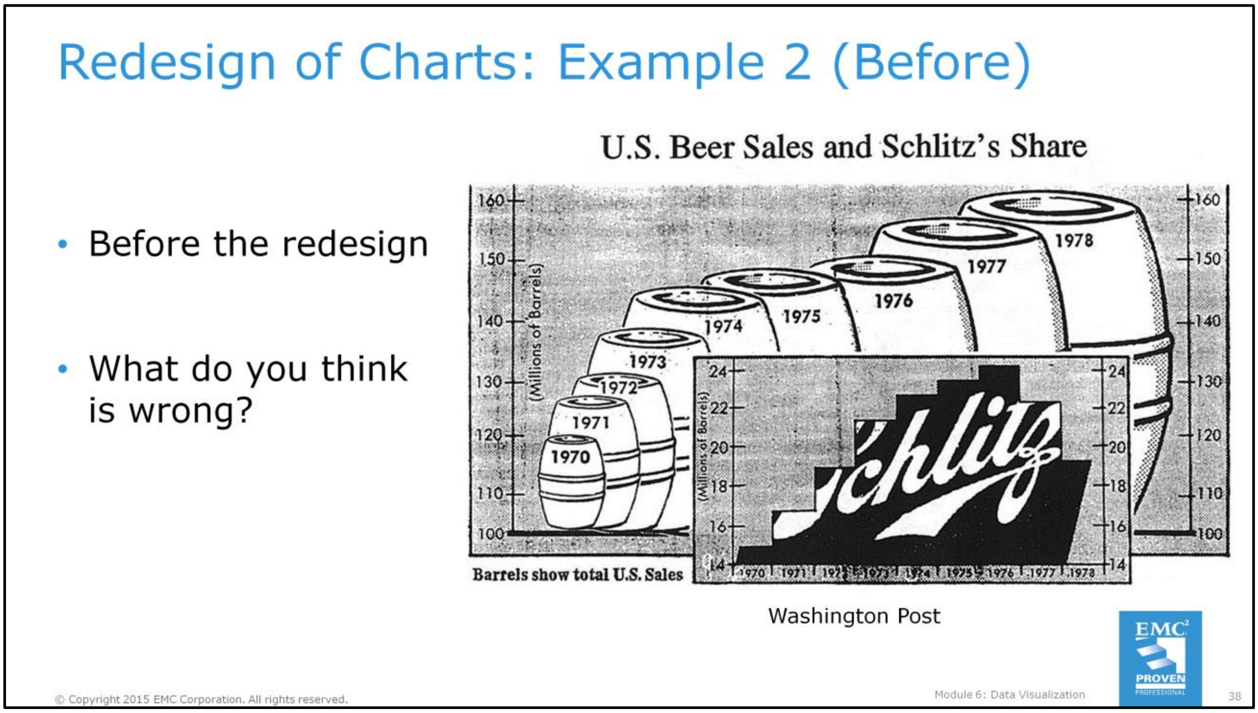


This graph is taken from a New York Times article, *The Shadow Of Taiwan Follows Haig To*, published on June 14, 1981. Some problems of the graph include:

- Different scales from the two charts
- Although the data has the same context, the labels and color scheme are reversed in the two charts



Here's an improved version of the graph. The y-axes of the two charts are in the same scale. The y-axis represents billions of USD instead of millions of USD. The color scheme is consistent across the charts. Data points are marked in different shapes on the two lines so the graph can represent data correctly even it's printed in black-and-white.

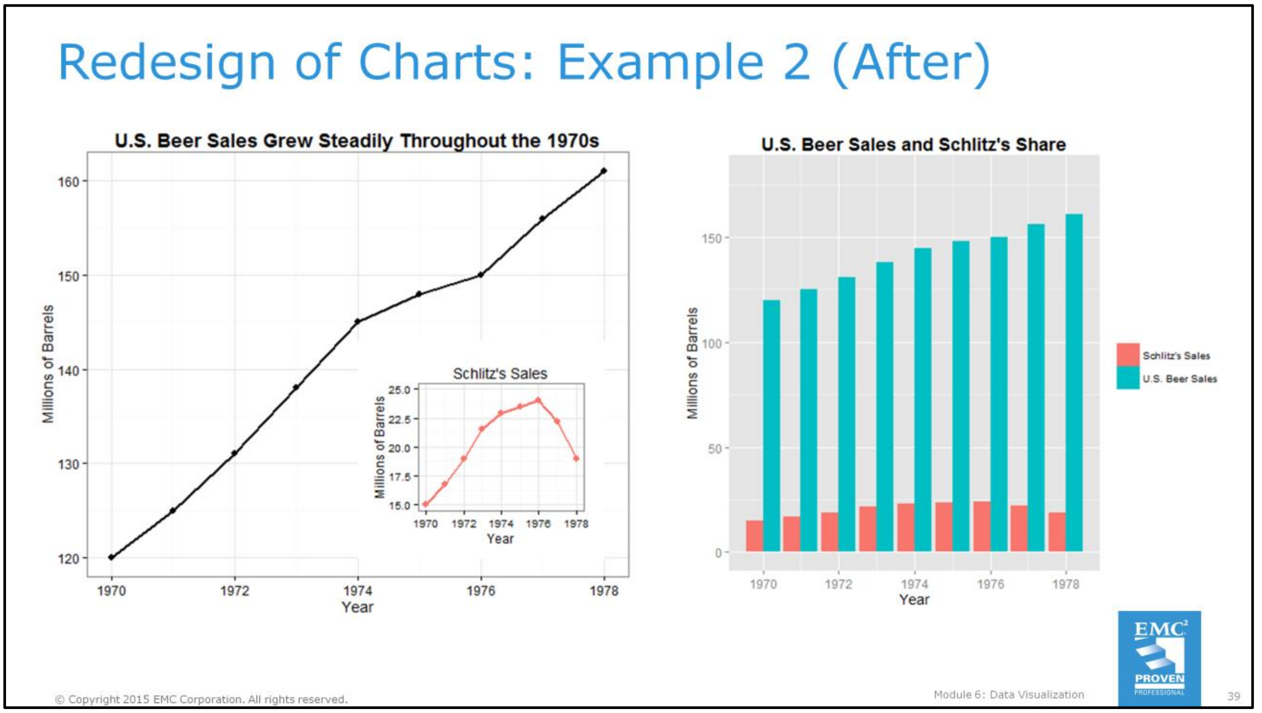


The graph of "U.S. Beer Sales and Schlitz's Share" consists of two separate plots. One table describes the U.S. Beer Sales and remains in the back while the other one describes the Schlitz's Share and remains in the front. This makes the overall graphic difficult to understand.

The method of using barrels as a design produces a misleading representation. 160 millions of barrels sold in the year 1978 and almost 120 millions barrels sold in the year 1970, which is a 25% difference. However, if you look at the sizes of the two barrels, the 1978 barrel appears much larger than the 1970 barrel.

The two tables also use different styles. U.S. Beer Sales is described by the use of barrels but Schlitz's Share is indicated by bars even though both plots refer to the same unit of measure, i.e., millions of barrels.

The contrast between the two plots is poor. The two plots represent the data in different ways (one with barrels, one with a bar chart with a logo overlay), making it difficult to compare them accurately.

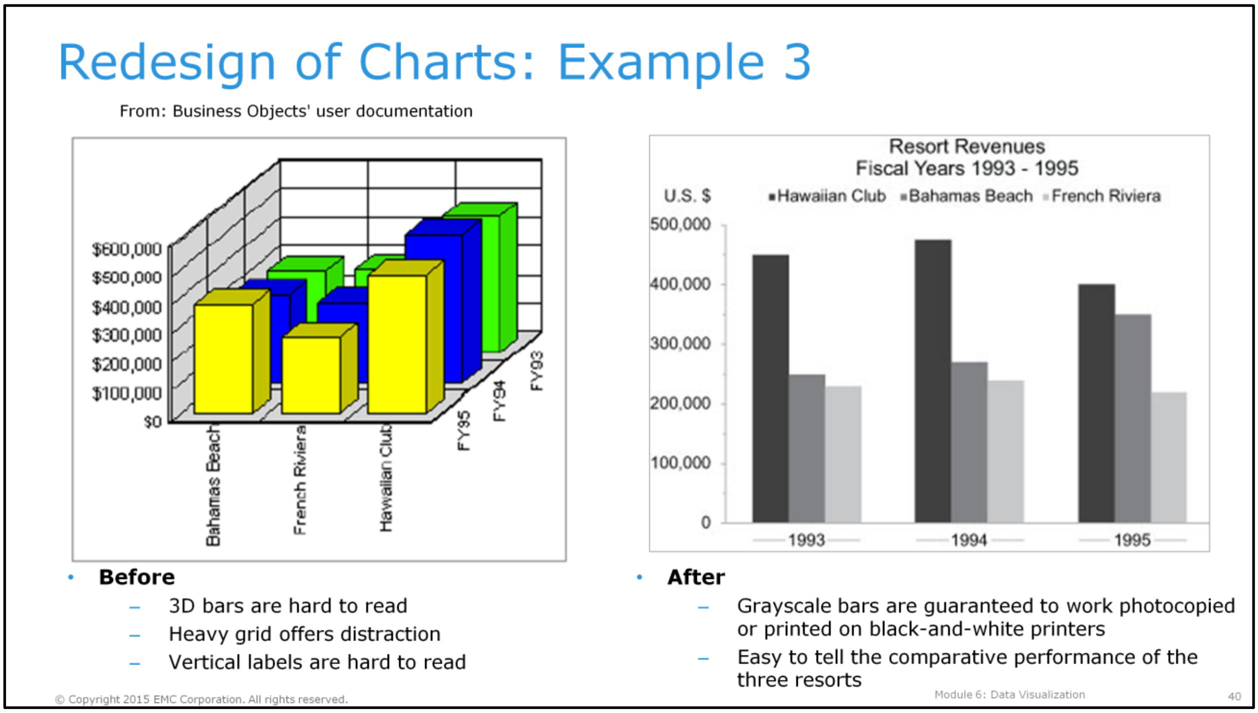


Here we show two versions of the improved graph. The graph on the left shows the U.S. beer sales in the main plot and Schlitz's sales in a subplot. Both are visualized using line charts. The subplot is differentiated by using the red color scheme. The title of the main plot summarizes the graph and makes it easier to comprehend.

The right graph combines both the U.S. sales and Schlitz's share into one graph to clearly show the viewer that the two original charts are semantically linked together. They now both share the same scale, making it easy to comprehend the proportions of Schlitz's share in the U.S. beer sales.

The scale of the y-axis starts with 0 so the sizes of the bars are now in the correct relation to each other, in contrast to the original graph where 160 million barrels was about five times bigger than 120 million barrels, which could convey a misleading understanding of the underlying data.

So which one should we prefer, the left graph or the right graph? Well it depends on what kind of message you want to deliver. The left graph shows the growth trend of Schlitz's share more clearly, while the right graph shows the proportions of Schlitz's share more easily.



In this example, the 3D bar chart shows the revenue of three resorts in 1993, 1994 and 1995. This graph, taken from Business Objects' user documentation, suffers the following problems:

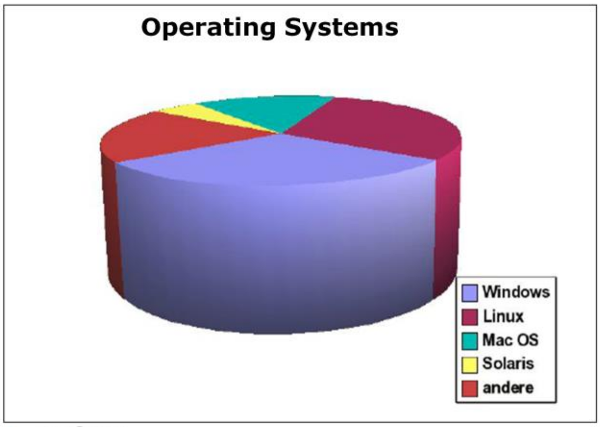
- The 3D bars are hard to read. Some bars are partially hidden and it’s difficult to tell the relative differences of the bar heights. For example, you may mistakenly conclude from the graph that the revenues of French Riviera are the same in 1993, 1994 and 1995, and the revenue of Bahamas Beach in 1995 is only slightly higher than the revenues from 1993 and 1994.
- The heavy grid offers distraction. With the partially hidden bars and grids, the grids do not help tell us the bar heights and only offers distraction.
- Vertical labels (such as the resort names and the fiscal years) are hard to read
- Years run from back to front, which is counter-intuitive

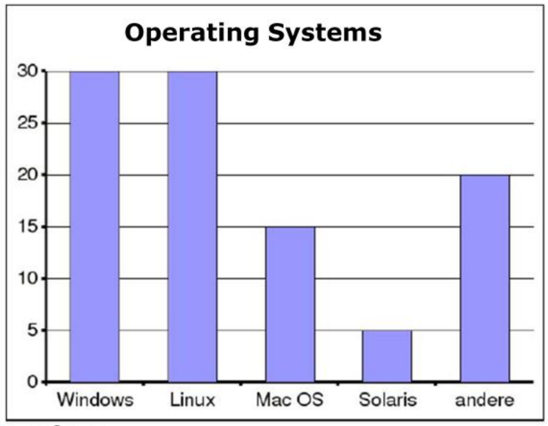
The 2D bar chart on the right address the issues in the original graph. The bar chart solution is ideal if people primarily want to focus on the magnitudes of a particular bar or if they want to compare the revenue of all three resorts in a given year. Notice the following improvements:

- The 2D bar chart substitutes the yellow-blue-green color scheme with grayscale. Grayscales are guaranteed to work photocopied or printed on black-and-white printers. If grayscale works, generally there’s no need to use colors.
- The 2D bar chart makes it easy to tell the comparative performance of the three resorts.
- The legend is placed on top, making it easy to read. The legend is also arranged in the same sequence as the bars, which eases the process of matching them up.
- Years are arranged from left to right, which is more intuitive

Reference: http://www.perceptualedge.com/example3.php

Redesign of Charts: Example 4



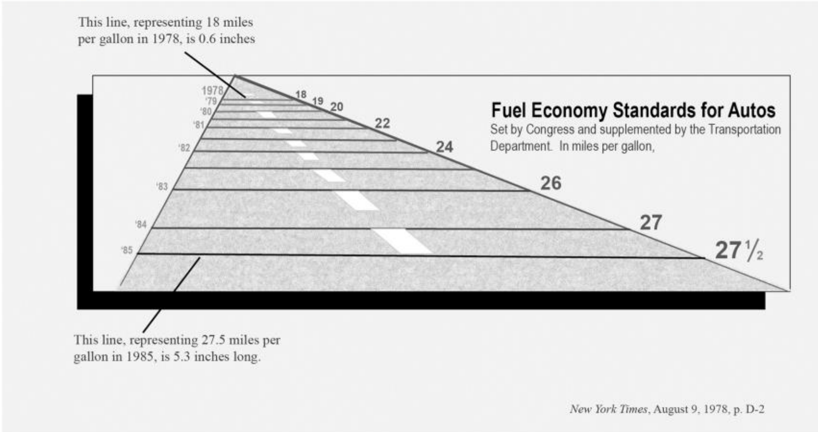


- **Before**
  - 3D pie charts are hard to read
  - Too many colors offers confusion
- **After**
  - Bar chart clearly shows the relative differences of the operating systems

The pie chart on the left shows the proportions of different operating systems. Although it looks fancy at the first glance, this pseudo 3D representation makes it difficult to tell the proportion of each slice and it's very difficult to compare the differences. For example, you may mistakenly conclude that Windows has a bigger market than Linux, and the market shares of Linux and andere are about the same. Also, the pie chart consists of too many colors. As the number of slices in a pie chart increases, it becomes more and more difficult to tell the colors apart.

The 2D bar chart on the right offers a much simpler representation of the same data. You can clearly tell that the market shares of Windows and Linux are the same, and the share of Linux is bigger than andere. The bar chart also removed the color scheme in the pie chart, making the representation cleaner, and more visually appealing.

Assessing the Lie Factor of a Graph



- $Lie\ factor = \frac{size\ of\ effect\ shown\ in\ graphic}{size\ of\ effect\ in\ data} = \frac{(5.3-0.6)/0.6}{(27.5-18)/18} = 14.8$
- Tufte requirement: 0.95<Lie Factor<1.05

This graph, from the NY Times, purports to show the mandated fuel economy standards set by the US Department of Transportation. The standard required an increase in mileage from 18 to 27.5, an increase of 53%. The magnitude of increase shown in the graph is 783%, for a lie factor = (783/53) = 14.8.

References:

The Lie Factor, http://www.datavis.ca/gallery/lie-factor.php.

Tufte, E. R., & Graves-Morris, P. R. (1983). The visual display of quantitative information (Vol. 2). Cheshire, CT: Graphics press.

Helberg, C. (1996). Pitfalls of Data Analysis. ERIC/AE Digest. Available: http://my.execpc.com/~helberg/pitfalls/.

Check Your Knowledge

- List three visualization tools for univariate data
- What is the lie factor?
- Why should we avoid the use of 3D pie charts?

Further reading:

Matthew Ward, Georges Grinstein, and Daniel Keim. *Interactive data visualization: foundations, techniques, and applications*. AK Peters, Ltd., 2010.

Nathan Yau. *Visualize this: the Flowing Data guide to design, visualization, and statistics*. John Wiley & Sons, 2011.

Open Source or Free Visualization Software

- Point-and-click
  - Weave (http://www.iweave.com/)
  - Tableau Public (http://www.tableausoftware.com/public/)
- Code-it-yourself
  - C3.js (http://c3js.org/)
  - D3.js (http://d3js.org/)
  - ggplot2 (http://ggplot2.org/)
  - Matplotlib (http://matplotlib.org/)
  - Processing (http://www.processing.org/)
  - OpenLayers (http://openlayers.org/)
  - Modest Maps (http://modestmaps.com/)

Module 6: Data Visualization 44

Reference:

StackExchange: Does anyone know any good open source software for visualizing data from database? http://stats.stackexchange.com/questions/44204/.

Note: Tableau Public is free, but not open source.

Lesson 2: Summary

During this lesson the following topics were covered:

- Visualization of multivariate data
- Redesign of charts
- Lie factor

Module 6: Summary

Key points covered in this module:

- How to choose the suitable colors for specific type of data
- The eight visual variables
- Data visualization principles
- Choices of visualization structures for multivariate data