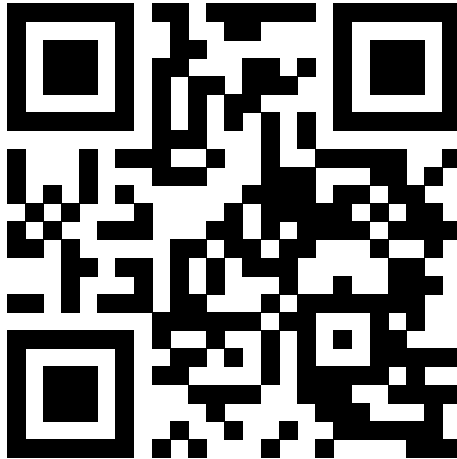# Representation

Jörg Cassens

Data and Process Visualization
SoSe 2017

medieninformatik

IMAI – Institut für
Mathematik und
Angewandte Informatik

☞ http://pingo.upb.de/650660

Literal vs. Abstract

- When you visualize data, you represent it with a combination of visual cues that are scaled, colored, and positioned according to values
- Dark-colored shapes mean something different from light-colored shapes, or dots in the top right mean something different than dots in the bottom left
- Visualization is what happens when you make the jump from raw data to bar graphs, line charts, and dot plots
- Process taking you from a grid of photos to a bar graph
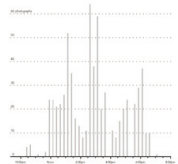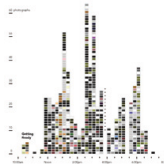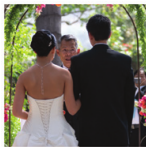
← More literal   More abstract →

Source: Yau (2013)

- In many ways, visualization is like cooking
- You are the chef, and data-sets, geometry, and color are your ingredients
- A skilled chef, who knows the process of how to prepare and combine ingredients and plate the cooked food, is likely to prepare a delicious meal
- A less skilled cook, who heads to the local freezer section to see what microwave dinners look good, might nuke a less savory meal
- Of course, some microwave dinners taste good, but there are a lot that taste bad.

- Whereas the person who is only familiar with entering the time and power level on a microwave must either endure poor-tasting meals or stick only to the handful of good ones, people who understand the ingredients and actually know how to cook have fewer limitations
- The skilled chef might even transform an average frozen dinner into a gourmet meal
- Likewise, with visualization, when you know how to interpret data and how graphical elements fit and work together, the results often come out better than software defaults

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Components

Visual Cues
Coordinate Systems
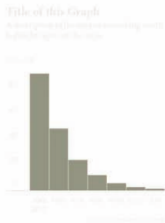Scales
Context

Example

Tutorial

References

Ingredients

- What are the ingredients of visualization?
- Breakdown into four components, with data as the driving force behind them:
    - visual cues,
    - coordinate system,
    - scale, and
    - context
- Each visualization, regardless of where it is on the spectrum, is built on data and these four components
- Sometimes, they are explicitly displayed, and other times they form an invisible framework
- The components work together, and your choice with one affects the others.

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Ingredients

Visual Cues

Visualization involves encoding data with shapes, colors, and sizes. Which cues you choose depends on your data and your goals.

Scale

Increments that make sense can increase readability, as well as shift focus.

Coordinate System

You map data differently with a scatterplot than you do with a pie chart. It's x- and y-coordinates in one and angles with the other; it's cartesian versus polar.

Context

If your audience is unfamiliar with the data, it's your job to clarify what values represent and explain how people should read your visualization.

Source: Yau (2013)

Visual Cues

Visual Cues

- In its most basic form, visualization is simply mapping data to geometry and color
- It works because we have seen that the brain is wired to find patterns, and you can switch back and forth between the visual and the numbers it represents
- You must make sure that the essence of the data isn't lost in that back and forth between visual and the value it represents because if you can't map back to the data, the visualization is just a bunch of shapes
- You must choose the right visual cue, which changes by purpose, and you must use it correctly, which depends on how you perceive the varied shapes, sizes, and shades

Components
Visual Cues
Coordinate Systems
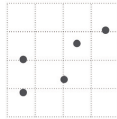Scales
Context

Example

Tutorial

References

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Visual cues

When you visualize data, you encode values to shapes, sizes, and colors.
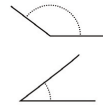
Position
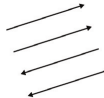Where in space the data is

Length
How long the shapes are

Angle
Rotation between vectors

Direction
Slope of a vector in space

Shapes
Symbols as categories

Area
How much 2-D space

Volume
How much 3-D space

Color saturation
Intensity of a color hue

Color hue
Usually referred to as color

Source: Yau (2013)

Position

Components
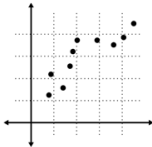Visual Cues
Coordinate Systems
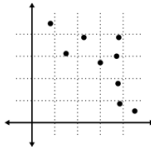Scales
Context

Example

Tutorial

References

- When you use position as a visual cue, you compare values based on where others are placed in a given space or coordinate system

- For example, when you look at a scatterplot, as shown in the next Figure, you judge a data point based on its x- and y-coordinate and where it is relative to others
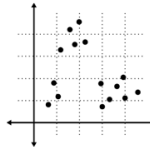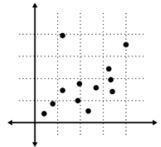
Upward trend   **Downward trend**   **Clustering**   **Outlier**

Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- One of the advantages of using just position is that it tends to take up less space than the other visual cues because you can draw all the data within the x- and y-plane, and you can represent each point with a dot

- Unlike other visual cues that use size to compare values, all points in a position-based plot are the same size

- In turn, you can spot trends, clusters, and outliers by plotting a lot of data at once

- However, the advantage of using position alone can also be a disadvantage

- If you look at a lot of points at once in a scatterplot, it can be a challenge to identify what each point represents

- Even in an interactive plot, you still must mouse over or select a point to find out more information, and overlap can cause more problems

Components
Visual Cues
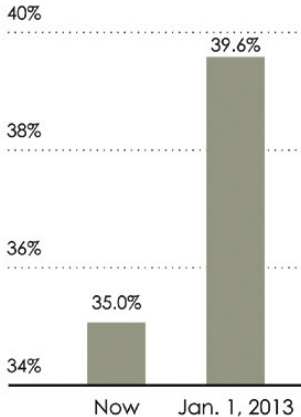Coordinate Systems
Scales
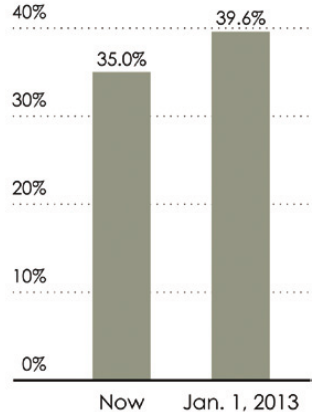Context

Example

Tutorial

References

- Length is most commonly used in the context of bar charts
- The longer a bar is, the greater the absolute value, and it can work in all directions: horizontal, vertical, or even at different angles on a circle
- How do you judge length visually?
- You figure out the distance from one end of a shape to the other end, so to compare values based on length, you must see both ends of the lines or bars
- Otherwise, you end up with a skewed view of maximums, minimums, and everything in between
- As a simple example, as shown next, a major news outlet displayed a bar graph on television that compared a tax rate before and after a date

Length Example

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Axis starting at 34 percent

Axis starting at 0 percent

Source: Yau (2013)

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Length Problems

- The difference between the two values looks like a huge increase
- The length of the right bar is about five times the length as the other
- But the value axis starts at 34 percent
- The chart on the right shows the change when the axis starts at zero, which looks less dramatic
- Of course, you can always look at the axis to verify what you see (and you always should), but that defeats the purpose of showing the values with length, and if the chart is shown quickly on television, most people won't notice the misstep

Angle

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- For each angle in between zero and 360 degrees, there is an implied opposite angle that completes the rotation, and together those two angles are considered conjugates
- This is why angles are commonly used to represent parts of a whole, using the fan favorite, but often maligned, pie chart
- The sum of the wedges makes a complete circle
- Although the donut chart is often considered the pie chart's close cousin, arc length is the former's visual cue because the center of the circle, which indicates angles, is removed

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Angle Example

Pies

The visual cue is the relative degrees in the circle.

Donuts

Arc length is the visual cue because the center is cut out.
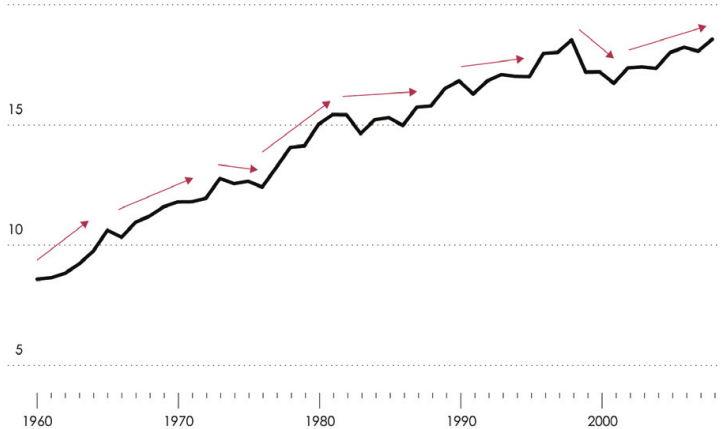
Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Direction

- Direction is similar to angle, but instead of relying on two vectors joined at a point, direction relies on a single vector's orientation in a coordinate system
- You can see which way is up, down, left, and right and everything in between
- This helps you determine slope
- You can see increases, decreases, and fluctuations
- A rule of thumb is to scale your visualization so that direction fluctuates mostly around 45 degrees, but this is hardly a concrete rule
- The best thing to do is to start with this suggestion and then adjust accordingly based on context
- If a small change is significant, then it might be appropriate to stretch the scale so that you can see the shift
- In contrast, if a small change is not significant, don't stretch out the scale just to make a shift look dramatic.

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Direction: Example

Direction in a time series

20 metric tons of CO2 per capita in Australia

Source: The World Bank

Source: Yau (2013)
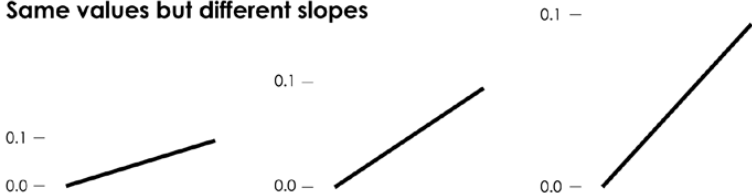
Scale

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- The amount of perceived change depends a lot on the scale
- For example, you can make a small change in percentage look like a lot by stretching out the scale
- Likewise, you can make a big change look like a little by compressing the scale

Same values but different slopes

Source: Yau (2013)

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Shapes

- Shapes and symbols are commonly used with maps to differentiate categories and objects
- Location on a map can be directly translated to the real world, so it makes sense to use icons to represent things in the real world
- You might represent forests with trees or residential areas with houses
- In a chart context, shapes to show variation are used less frequently than they used to be
- For example, triangles and squares could be used in a scatterplot, which is quicker to draw than to switch between colored pencils and pens or fill a single shape with a solid or cross-hatched pattern
- Nevertheless, varied shapes can provide context that points alone can't, and it's typically not more difficult to try with your favorite software

Shapes: Example

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Source: Yau (2013)
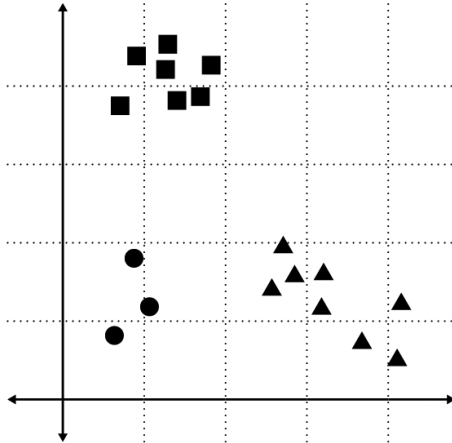
Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- Bigger objects represent greater values
- Like length, area and volume can be used to represent data with size, but with two and three dimensions, respectively
- For the former, circles and rectangles are commonly used, and with the latter, cubes and sometimes spheres
- You can also size more detailed icons and illustrations
- Be sure to mind how many dimensions you use
- The most common mistake is to size a two- or three-dimensional object by only one dimension, such as height, but to maintain the proportions of all dimensions
- This results in shapes that are too big and too small, which makes it impossible to fairly compare values

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Area and Volume: Example

Sizing by area

This is one unit.

Four units sized by area

4 times the area as unit square

Four units *incorrectly* sized by side length

16 times the area as unit square

Sizing by volume

This is one unit.

Four units sized by volume

4 times the volume as unit cube

Four units *incorrectly* sized by edge length
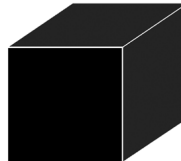
32 times the volume as unit square

Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- Color as a visual cue can be spilt into two categories: hue and saturation
- They can be used individually or in combination
- Color hue is what you usually just refer to as color
  - That's red, green, blue, and so on
- Differing colors used together usually indicates categorical data, where each color represents a group
- Saturation is the amount of hue in a color, so if your selected color is red, high saturation would be very red, and as you decrease saturation, it looks more faded
- Used together, you can have multiple hues that represent categories, but each category can have varying scales

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- Careful color selection can lend context to your data, and because there is no dependency on size or position, you can encode a lot of data at once
- However, keep color blindness in mind if you want to make sure that everyone can interpret your graphics
- When you encode your data only with red and green, people with a red-green deficiency might have trouble decoding your visualization, if they can at all

Ranking

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- In 1985, William Cleveland and Robert McGill published a paper on graphical perception and methods
- The focus of the study was to determine how accurately people read the visual cues above (excluding shapes), which resulted in a ranked list from most accurate to least accurate

| Position | Length | Angle | Direction | Area | Volume | Saturation | Hue |

◄ More accurate   Less accurate ►
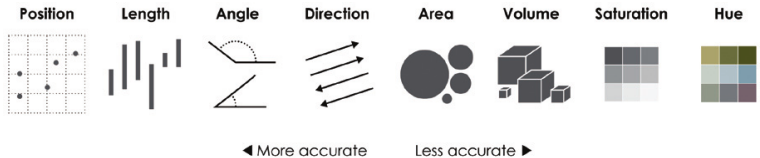
Source: Yau (2013)

Ranking: Problems

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- A lot of visualization suggestions (and current research) stem from this list
- It places bar charts above pie charts, heat maps at the bottom, and so on
- This is sound advice, but remember that this list doesn't mean that dot plots are always better than bubble plots or that pie charts are evil
- Following this list blindly is an oversimplification of what visualization is
- Efficiency and exactness are not always the goal
- That said, regardless of what you want to visualize data for, it's good to know how well people can read your visual cues and what information they can extract
- In other words, use these rankings as a guide rather than a rule book

Coordinate Systems

Universität
Hildesheim

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Placement of Objects

- When you encode data, you eventually must place the objects somewhere

- There's a structured space and rules that dictate where the shapes and colors go

- This is the coordinate system, which gives meaning to an x-y coordinate or a latitude and longitude pair

- There are several systems, but there are three that cover most of your bases: Cartesian, polar, and geographic

Coordinate systems

There are a variety of them, from cylindrical to spherical, but these three will cover most of your bases.

Cartesian

If you've ever made a graph, the x- and y-coordinate system will look familiar to you.
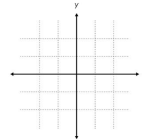
Polar

Pie charts use this system. Coordinates are placed based on radius r and angle θ.
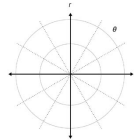
Geographic

Latitude and longitude are used to identify locations in the world. Because the planet is round, there are multiple projections to display geographic data in two dimensions. This one is the Winkel tripel.

Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Cartesian

- The Cartesian coordinate system is the most commonly used one with charts
- You typically think of coordinates in the system as an x and y pair that is denoted as (x, y)
- Two lines that are perpendicular to each other, and range from negative to positive, form the axes
- The place the lines intersect is the origin, and the coordinate values indicate the distance from that origin
- You can also extend the Cartesian space to more than two dimensions
- The takeaway is that you can describe geometric shapes using Cartesian coordinates, which makes it easier to draw in the space
- From an implementation standpoint, the coordinate system enables you to encode values to paper or a computer screen

Polar

- The polar coordinate system consists of a circular grid, where the rightmost point is zero degrees
- The greater the angle is, the more you rotate counter-clockwise
- The farther away from the circle you are, the greater the radius is
- Place yourself on the outer-most circle, and increase the angle
- This rotates you counterclockwise toward the vertical line (or the y-axis if this were Cartesian coordinates), which is 90 degrees (that is, a right angle)
- Rotate one-quarter more, and you get to 180 degrees
- Rotate back to where you started, and that's a 360-degree rotation

Universität
Hildesheim

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Geographic

- Location data has the added benefit of a connection to the physical world, which in turn lends instant context and a relationship to that point, relative to where you are
- A geographic coordinate system can map these points
- Location data comes in many forms, but it's most commonly described as latitude and longitude, which are angles relative to the Equator and Prime Meridian, respectively
  - Sometimes elevation is also included
- Latitude lines run east and west, which indicates north and south position on a globe
- Longitude lines run north and south and indicate the east and west position
- Elevation can be thought of as a third dimension
- Compared with Cartesian coordinates, latitude is like the horizontal axis, and longitude is like the vertical axis
  - That is, if you use a flat projection

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Geographic: Projections

- The tricky part about mapping the surface of Earth is that it's wrapped around a spherical mass, but you usually need to display it on a two-dimensional surface
- The variety of ways to do this are called projections, each has its advantages and disadvantages
- When you project something that is three-dimensional onto a two-dimensional plane, some information is lost, whereas other information is preserved
- The Mercator projection, for example, preserves angles in local regions
- It was created in the 16th century by cartographer Geradus Mercator primarily for navigation on the seas and is still the most-used projection for online direction lookup
- On the other hand, the Albers projection preserves area but distorts shape
- So the projection depends on what you want to focus on

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Map projections

Equirectangular
Typically used for thematic mapping, but doesn't preserve area or angle

Albers
Scale and shape not preserved; angle distortion is minimal

Mercator
Preserves angles and shapes in small areas, making it good for directions

Lambert conformal conic
Better for showing smaller areas and often used for aeronautical maps.

Sinusoidal
Preserves area; useful for areas near the prime meridian

Polyconic
Was often used to show US in the mid-1900s; little distortions in small areas near meridian

Winkel Tripel
Minimized area, angle, and distance distortion; good choice for world map

Robinson
A compromise between preserving areas and angles; good to show world map

Orthographic
Representing a 3-D object in 2-D; need to rotate to area of interest

Source: Natural Earth

Source: Yau (2013)

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Scales

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- Whereas coordinate systems dictate the dimensions of a visualization, scale dictates where in those dimensions your data maps to
- There's a variety of them, and you can even define your own scales based on mathematical functions, but most likely you'll rarely stray from the ones in the Following Figure 3-15
- These can be grouped into three categories: quantitative/numerical, categorical, and time
  - Compare slide set "Data"

Scales

Along with coordinate systems, they dictate where the shapes are placed and how objects are shaded.

Linear
Values are evenly spaced

0   1   2   3   4
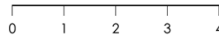
Logarithmic
Focus on percent change

1   10   100   1,000   10,000
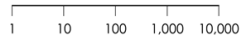
Categorical
Discrete placement in bins

A   B   C   D   E
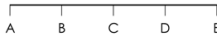
Ordinal
Categories where order matters

Horrible   Bad   Okay   Good   Great

Percent
Representing parts of a whole

0%   25%   50%   75%   100%
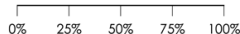
Time
Units of months, days, or hours
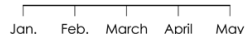
Jan.   Feb.   March   April   May

Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Quantitative

- The visual spacing on a linear scale is the same regardless of where you are on the axis
- So if you were to measure the distance between two points on the lower end of the scale, it'd be the same if they were at the high end of the scale
- On the other hand, a logarithmic scale condenses as you increase values
- This scale is used less than the linear scale and is not as well understood or straightforward for those who don't regularly work with data, but it's useful if you're interested in percent differences more than you are raw counts or your data has a wide range
- For example, when you compare state populations in the United States, you deal with numbers from the hundreds of thousands up to the tens of millions

Example: Logarithmic Scale

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- California has a population of approximately 38 million people, whereas Wyoming has a population of approximately 600,000

- With a linear scale, states with smaller populations are clustered on the bottom, and then a few states rest on top

- Easier to see points on the bottom with a logarithmic scale
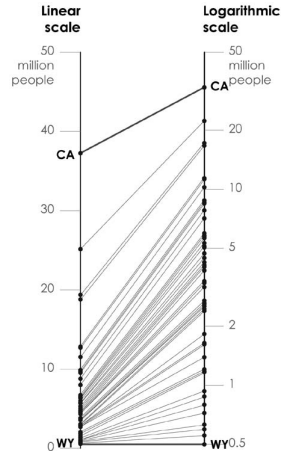
Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Scale: Percentage

- A percent scale is usually linear, but when it's used to represent parts of a whole, its maximum is 100 percent
- As shown in the next Figure, the sum of all the parts is 100 percent
- This seems obvious – that the sum of percentages in a pie chart, represented with wedges, should not exceed 100 percent – but the mistake seems to come up occasionally
- Sometimes it's due to mislabeling, but some people just aren't familiar with the concept

Percentage: Example

Components

Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial
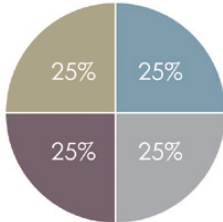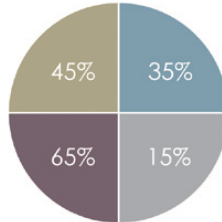
References

Source: Yau (2013)

Categorical

- Data doesn't always need to be numeric
- It can be categorical, such as people's cities of residence or the political parties of government officials
- A categorical scale provides visual separation for these different groups and often works with a numeric scale
- A bar plot for example, can use a categorical scale on the horizontal axis and a numeric scale on the vertical to show counts or measurements for different groups
- Spacing between each category is arbitrary because it does not depend on a numeric value, but it is typically adjusted to increase clarity
- Ordering should be used in the context of the data
- Although this can also be arbitrary, for an ordinal scale that uses categories, order of course matters
- If your data is categorical ordinal it makes sense to keep that order visually, which makes it easier to compare

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Time

- Time is a continuous variable, which lets you plot temporal data on a linear scale, but you can divide it into categories such as months or days of the week, which lets you visualize it as a discrete variable
- Also, it cycles
- There's always another noon time, Saturday, and January
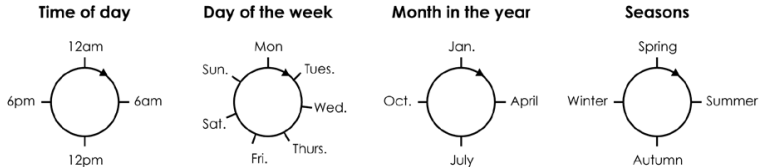
Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- You saw this when we showed fatal crashes over time, by year, by month, by day, and by hour
- Data was plotted continuously in these cases
- However, aggregates by time of day, day of the week, and month (over multiple years) showed a different picture
- When communicating data to an audience, the time scale, like geographic maps, gives you an advantage of lending a reader connection because time is a part of everyday life
- You feel and experience time internally and through your clocks and calendars, and as the sun rises and sets

Context

Components
Visual Cues
Coordinate Systems
Scales
Context
Example
Tutorial
References

- Context can make the data clearer for readers and point them in the right direction
  - Context here: information that lends to better understanding the who, what, when, where, and why of your data
- At the least, it can remind you what a graph is about when you come back to it a few months later
- Sometimes context is explicitly drawn, and other times it's implied through the medium

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Sample

- Matt Robinson and Tom Wrigglesworth drew "sample" on a wall, with ballpoint pens and different typefaces
- Because ink usage varies by typeface, each pen had a different amount of ink left, which made for an interesting bar graph

Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

iPad

- George Kokkinidis approached iPad usage in a similar way
- he looked at fingerprint traces while he used different apps
- In Mail, he typed messages most of the time, so the keyboard pattern is most evident
- Most interaction is in the bottom-left corner for the game Angry Birds.
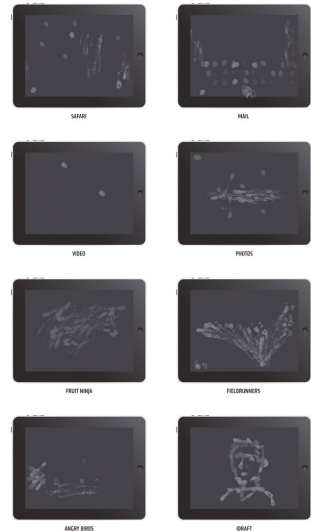
Source: Yau (2013)

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- Of course, you can't always draw on familiar physical objects for context, so you must provide familiarity and a sense of scale in other ways
- The easiest and most straightforward way is to label your axes and specify units of measure, or provide a description that tells others what each visual cue represents
- Otherwise, when the data is abstracted, there's no way to decode the shapes, sizes, and colors, and you might as well show an amorphous blob

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

- A descriptive title is a small but easy thing you can create to set up readers for what they're about to look at
- Imagine you produce a time series plot for gas prices that shows an upward trend
- You could just title it "Gas Prices" and that would be a fair title
- That's what it is, but you could also title it "Rising Gas Prices," which says what data is used and what is shown
- You could also include lead-in text underneath the title that describes fluctuations or by how much gas prices rose

Components
Visual Cues
Coordinate Systems
Scales
Context

Example

Tutorial

References

Context: Implicit

- Your choice of visual cues, a coordinate system, and scale can implicitly provide context
- Bright, cheery, and contrasting colors says something different than dark, neutral, and blending colors
- Similarly, a geographic coordinate system places you within the context of physical space, whereas an x-y plot using Cartesian coordinates keeps you within a virtual space
- A logarithmic scale could suggest a focus on percentage changes and reduce focus on absolute values
- This is why it's important to pay attention to software defaults

Context: Other

- Programs are designed to be flexible and fast and they work outside the context of the data
- This is great to draw a visualization base and explore your data, but it's up to you to make the right decisions along the way and to make the computer output something for humans
- This comes partly from knowing how you perceive geometry and colors, but mostly it comes from practice and the experience gained from seeing a lot of data and evaluating how others, who aren't familiar with your data, interpret your work
- Common sense also goes a long way

Source: Yau (2013)

Visual cues

	Position	Length	Angle	Direction	Shapes	Area or Volume	Color

Coordinate systems

Cartesian

Polar

Geographic

Outline

Components

Example

Tutorial

References

- You know what ingredients are available
- Now it's time to cook the meal
- Viewed separately, the visualization components aren't that useful because they are just bits of geometry floating in an empty space without context
- However, when you put the components together, you get a complete visualization worth looking at

Cooking the Meal (contd.)

- What do you get when you use length as a visual cue, a Cartesian coordinate system, and a categorical scale on the horizontal axis and a linear scale on the vertical?
  - You get a bar chart
- Use position with a geographic coordinate system?
  - You get points on a map
- What do you get when you use a polar coordinate system with the area as the visual cue, a percentage scale on the radius, and a time scale on the rotation?
  - That's a polar area diagram

Nightingale: Polar Area Diagram
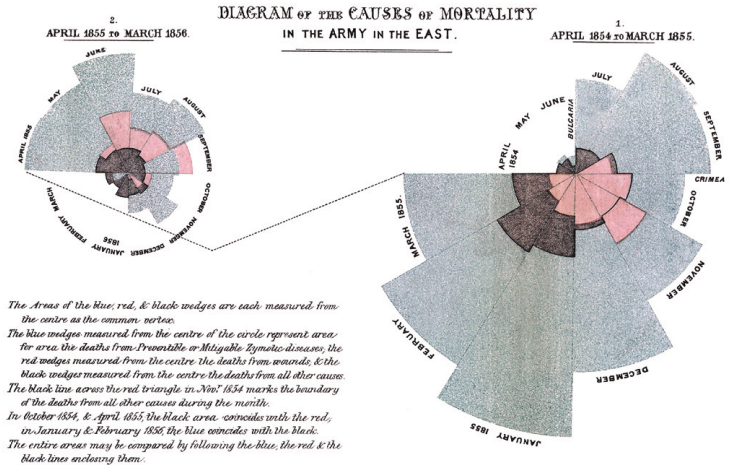
Source: Yau (2013)

Cooking the Meal (contd.)

- On the Origin of Species: The Preservation of Favoured Traces, designer and developer Ben Fry uses color and length, Cartesian coordinates, and a linear scale
    - ☞ fathom.info/traces
- The interactive and animated visualization shows how Charles Darwin's theory of evolution changed through six editions
- The gray blocks represent the original text, and each subsequent color represents a revision in an edition, so you can see what changed and by how much

Fry: Origin of Species

Source: Yau (2013)

Reset Pause Slow Fast
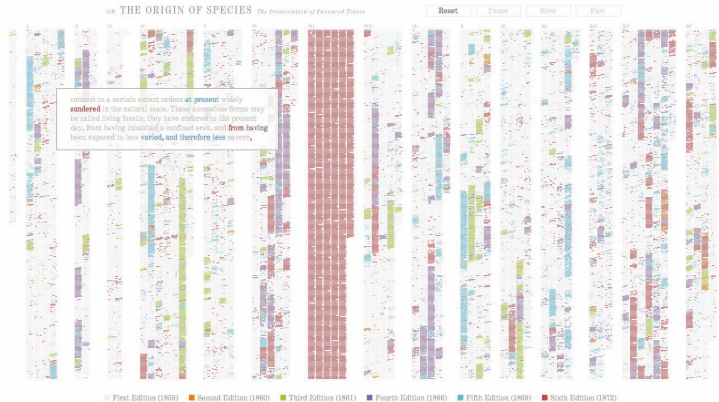
connect to a certain extent, orders **at present** widely **sundered** in the natural scale. These anomalous forms may be called living fossils; they have endured to the present day, from having inhabited a confined area, and **from having** been exposed to less **varied, and therefore less** severe,

First Edition (1859)   Second Edition (1860)   Third Edition (1861)   Fourth Edition (1866)   Fifth Edition (1869)   Sixth Edition (1872)

Death Charts

- In the deaths chart shown next, from the Statistical Atlas of the United States published in 1874, length is used to show the distribution of deaths for each state, by age and gender
- The horizontal axis on each plot represents the number of deaths on a linear scale, and the vertical axis represents numeric categories that represent age groups

Death Chart (Example)
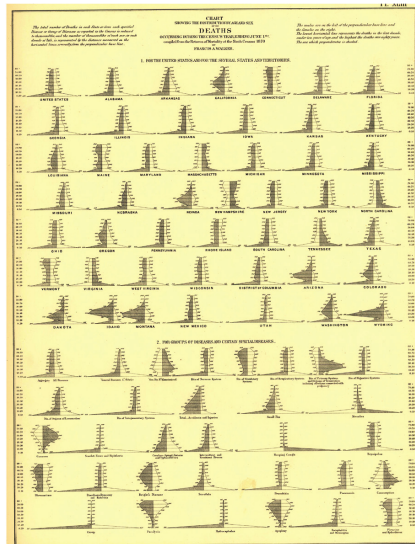
Source: Yau (2013)

Death Chart (Example, Detail)

The total number of Deaths in each State or from each specified Disease or Group of Diseases as reported in the Census is reduced to thousandths, and the number of thousandths of each sex, in each decade of life, is represented by the distance measured on the horizontal lines, severally, from the perpendicular base line.
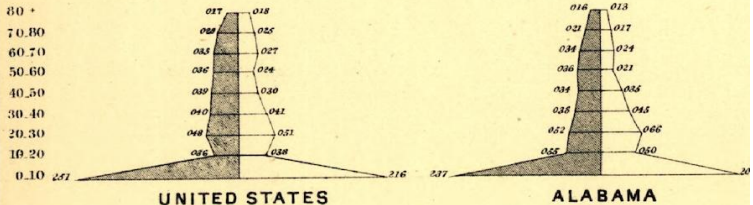
Source: Statistical Atlas of the United States, 1874

Death Chart (Example, Detail)

Components

Example
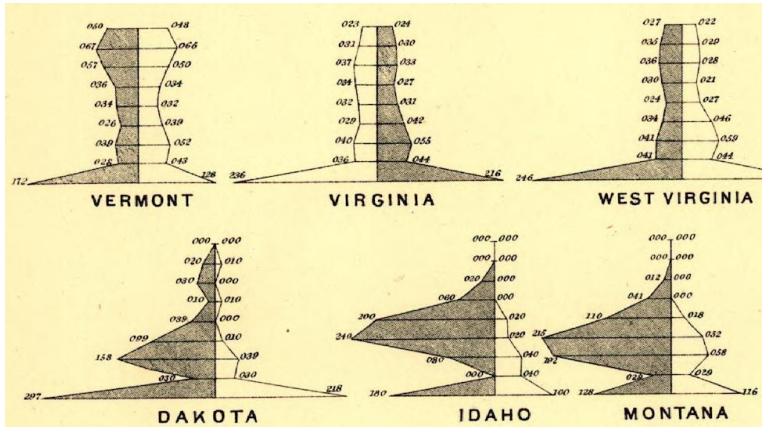
Tutorial

References

The males are on the left of the perpendicular base line and
the females on the right.
The lowest horizontal line represents the deaths in the first decade,
under ten years of age, and the highest the deaths over eighty years.
The sex which preponderates is shaded.

Source: Statistical Atlas of the United States, 1874

Death Chart (Example, Detail)

Source: Statistical Atlas of the United States, 1874

Death Chart (Example, Detail)
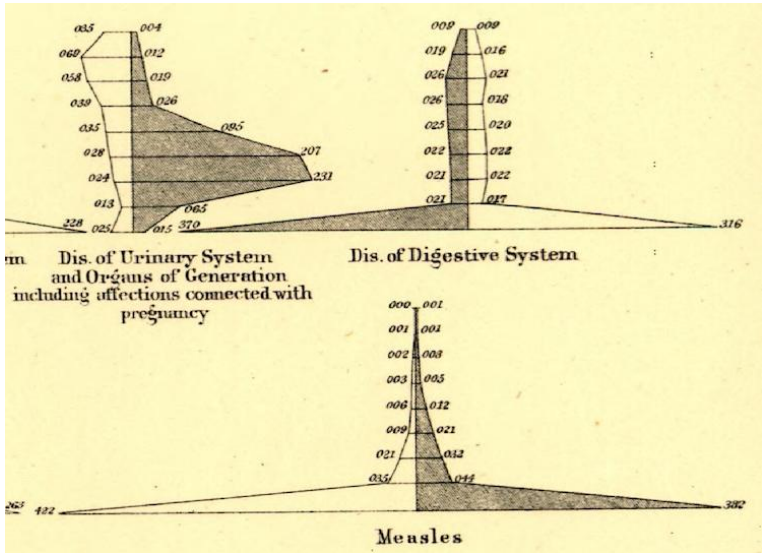
Source: Statistical Atlas of the United States, 1874

- We have seen how others have used specific combinations of components identified
  - Chart types
  - Context
- Now try to fit components together, starting with the data and then building on that foundation
- Starting with a data table from the United States Census Bureau that shows educational attainment (high school graduate or more, bachelor's degree or more, and advanced degree or more) by state, in 1990, 2000, and 2009
- Values are percentages for people 25 years old and over

Educational Attainment

Table 233. Educational Attainment by State: 1990 to 2009

[In percent. 1990 and 2000 as of April. 2009 represents annual averages for calendar year. For persons 25 years old and over. Based on the 1990 and 2000 Census of Population and the 2009 American Community Survey, which includes the household population and the population living in institutions, college dormitories, and other group quarters. See text, Section 1 and Appendix III. For margin of error data, see source]

State	1990 High school graduate or more	1990 Bachelor's degree or more	1990 Advanced degree or more	2000 High school graduate or more	2000 Bachelor's degree or more	2000 Advanced degree or more	2009 High school graduate or more	2009 Bachelor's degree or more	2009 Advanced degree or more
United States	75.2	20.3	7.2	80.4	24.4	8.9	85.3	27.9	10.3
Alabama	66.9	15.7	5.5	75.3	19.0	6.9	82.1	22.0	7.7
Alaska	86.6	23.0	8.0	88.3	24.7	8.6	91.4	26.6	9.0
Arizona	78.7	20.3	7.0	81.0	23.5	8.4	84.2	25.6	9.3
Arkansas	66.3	13.3	4.5	75.3	16.7	5.7	82.4	18.9	6.1
California	76.2	23.4	8.1	76.8	26.6	9.5	80.6	29.9	10.7
Colorado	84.4	27.0	9.0	86.9	32.7	11.1	89.3	35.9	12.7
Connecticut	79.2	27.2	11.0	84.0	31.4	13.3	88.6	35.6	15.5
Delaware	77.5	21.4	7.7	82.6	25.0	9.4	87.4	28.7	11.4
District of Columbia	73.1	33.3	17.2	77.8	39.1	21.0	87.1	48.5	28.0
Florida	74.4	18.3	6.3	79.9	22.3	8.1	85.3	25.3	9.0
Georgia	70.9	19.3	6.4	78.6	24.3	8.3	83.9	27.5	9.9
Hawaii	80.1	22.9	7.1	84.6	26.2	8.4	90.4	29.6	9.9
Idaho	79.7	17.7	5.3	84.7	21.7	6.8	88.4	23.9	7.5
Illinois	76.2	21.0	7.5	81.4	26.1	9.5	86.4	30.6	11.7
Indiana	75.6	15.6	6.4	82.1	19.4	7.2	86.6	22.5	8.1
Iowa	80.1	16.9	5.2	88.1	21.2	6.5	90.5	25.1	7.4
Kansas	81.3	21.1	7.0	86.0	25.8	8.7	89.7	29.5	10.2
Kentucky	64.6	13.6	5.5	74.1	17.1	6.9	81.7	21.0	8.5
Louisiana	68.3	16.1	5.6	74.8	18.7	6.5	82.2	21.4	6.9
Maine	78.8	18.8	6.1	85.4	22.9	7.9	90.2	26.9	9.6
Maryland	78.4	26.5	10.9	83.8	31.4	13.4	88.2	35.7	16.0
Massachusetts	80.0	27.2	10.6	84.8	33.2	13.7	89.0	38.2	16.4
Michigan	76.8	17.4	6.4	83.4	21.8	8.1	87.9	24.6	9.4
Minnesota	82.4	21.8	6.3	87.9	27.4	8.3	91.5	31.5	10.3
Mississippi	64.3	14.7	5.1	72.9	16.9	5.8	80.4	19.6	7.1
Missouri	73.9	17.8	6.1	81.3	21.6	7.6	86.8	25.2	9.5
Montana	81.0	19.8	5.7	87.2	24.4	7.2	90.8	27.4	8.3
Nebraska	81.8	18.9	5.9	86.6	23.7	7.3	89.8	27.4	8.8
Nevada	78.8	15.3	5.2	80.7	18.2	6.1	83.9	21.8	7.6
New Hampshire	82.2	24.4	7.9	87.4	28.7	10.0	91.3	32.0	11.2
New Jersey	76.7	24.9	8.8	82.1	29.8	11.0	87.4	34.5	12.9
New Mexico	75.1	20.4	8.3	78.9	23.5	9.8	82.8	25.3	10.4
New York	74.8	23.1	9.9	79.1	27.4	11.8	84.7	32.4	14.0
North Carolina	70.0	17.4	5.4	78.1	22.5	7.2	84.3	26.5	8.8
North Dakota	76.7	18.1	4.5	83.9	22.0	5.5	90.1	25.8	6.7
Ohio	75.7	17.0	5.9	83.0	21.1	7.4	87.6	24.1	8.8
Oklahoma	74.6	17.8	6.0	80.6	20.3	6.8	85.6	22.7	7.7
Oregon	81.5	20.6	7.0	85.1	25.1	8.7	89.1	29.2	10.4
Pennsylvania	74.7	17.9	6.6	81.9	22.4	8.4	87.9	26.4	10.2
Rhode Island	72.0	21.3	7.8	78.0	25.6	9.7	84.7	30.5	11.7
South Carolina	68.3	16.6	5.4	76.3	20.4	6.9	83.6	24.3	8.4
South Dakota	77.1	17.2	4.9	84.6	21.5	6.0	89.9	25.1	7.3
Tennessee	67.1	16.0	5.4	75.9	19.6	6.8	83.1	23.0	7.9
Texas	72.1	20.3	6.5	75.7	23.2	7.6	79.9	25.5	8.5
Utah	85.1	22.3	6.8	87.7	26.1	8.3	90.4	28.5	9.1
Vermont	80.8	24.3	8.9	86.4	29.4	11.1	91.0	33.1	13.3
Virginia	75.2	24.5	9.1	81.5	29.5	11.6	86.6	34.0	14.1
Washington	83.8	22.9	7.0	87.1	27.7	9.3	89.7	31.0	11.1
West Virginia	66.0	12.3	4.8	75.2	14.8	5.9	82.8	17.3	6.7
Wisconsin	78.6	17.7	5.6	85.1	22.4	7.2	89.8	25.7	8.4
Wyoming	83.0	18.6	5.7	87.9	21.9	7.0	91.8	23.8	7.9

Source: U.S. Census Bureau, 1990 Census of Population, CPH-L-96; 2000 Census of Population, P37, "Sex by Educational Attainment for the Population 25 Years and Over"; 2009 American Community Survey, R1501, "Percent of Persons 25 Years and Over Who Have Completed High School (Includes Equivalency)"; R1502, "Percent of Persons 25 Years and Over Who Have Completed a Bachelor's Degree"; and R1503, "Percent of Persons 25 Years and Over Who Have Completed an Advanced Degree"; <http://factfinder.census.gov/>, accessed February 2011.

Source: Mullin and O'Brien (2012)

Educational Attainment (Detail)

State	1990		
	High school graduate or more	Bachelor's degree or more	Advanced degree or more
United States	**75.2**	**20.3**	**7.2**
Alabama	66.9	15.7	5.5
Alaska	86.6	23.0	8.0
Arizona	78.7	20.3	7.0
Arkansas	66.3	13.3	4.5
California	76.2	23.4	8.1

Source: Mullin and O'Brien (2012)

Data Properties

- The "or more" for each column means you can't just add the values from each column because there's overlap between them
- If you want to make a pie chart that shows the values of each column, you must do some math
- For example, the United States estimate for people with a high school degree (or equivalent) or more is 75.2 percent
- Subtract those with a bachelor's degree or more, 20.3 percent, to get rid of the "or more" part of the high school value, which gives you 54.9 percent of people with only a high school degree

- It's also useful to know the sample population. If it were everyone in America, the percentages would be lower
- If for some odd reason the sample was those under 18, the percentages for an advanced degree or more would represent a tiny group of people who skipped or advanced quickly through elementary and high school

Components

Example

Tutorial

References

- So you have the most important part of any visualization: the data
- There are nine columns, spread out over 3 years and three subcategories, plus one more column for state names, so you can visualize the data on multiple dimensions
- You might want to focus on educational attainment in 2009, in which case, a few bar charts, as shown in Figure next, could work

Attainment: Bar Chart

Source: Yau (2013)

Components

Example

Tutorial

References
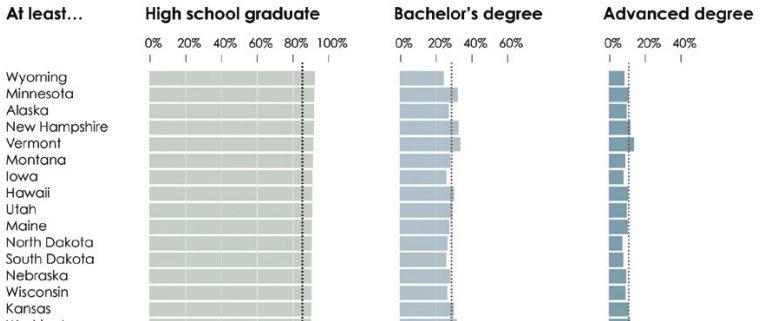
Source: Yau (2013)

- This is practically a direct translation of the last three columns in the table
- Each row represents the values for a state, and each column is a level of attainment
- Each bar chart has its own linear scale, but the increments are spaced equally and start at zero percent
- States are sorted by estimated percent of people with a high school diploma or equivalent, in descending order, rather than alphabetically, like in the table
- Instead of giving the national average its own row, it's presented as a vertical dotted line to provide a sense of low and high
- Color hue – gray, light blue, and blue – is used to indicate three separate estimates

Break it Down

- We use length (bars), color (each bar chart), and position (lines for national averages) as visual cues
- We have a Cartesian coordinate system
- We use linear scales for each of the bar charts, and a categorical scale for the sorted states
- The title and subtitles provide context for what the data is about

- If you are more interested in the changes between 2000 and 2009 than you are just the 2009 percentages, the next figure shows a few options that shift focus
- Length and position are still used, as well as a linear scale on the horizontal axis and a categorical scale on the vertical
- However, the context and layout are different than the bar charts
- Some other visual cues are also incorporated

Attainment: Changes

Source: Yau (2013)

Alphabetical
Values look scattered

2000 2009

Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

70%   80%   90%
High school or more

Greatest to least, endpoint
Focus on most recent numbers

2000 2009

Wyoming
Minnesota
Alaska
New Hampshire
Vermont
Montana
Iowa
Utah
Hawaii
Maine
North Dakota
South Dakota
Wisconsin
Nebraska
Washington
Kansas
Colorado
Oregon
Massachusetts
Connecticut
Idaho
Maryland
Pennsylvania
Ohio
New Jersey
Delaware
District of Columbia
Missouri
Virginia
Indiana
Illinois
Oklahoma
Florida
New York
North Carolina
Arizona
Nevada
Georgia
South Carolina
Tennessee
West Virginia
New Mexico
Arkansas
Louisiana
Alabama
Kentucky
California
Mississippi
Texas

70%   80%   90%
High school or more

Greatest to least, startpoint
Focus on past data

2000 2009

Alaska
Wyoming
Minnesota
Utah
New Hampshire
Montana
Washington
Colorado
Nebraska
Vermont
Iowa
Kansas
Maine
Oregon
Wisconsin
Massachusetts
South Dakota
Idaho
North Dakota
Maryland
Michigan
Connecticut
Delaware
New Jersey
Indiana
Pennsylvania
Virginia
Illinois
Missouri
Nevada
Oklahoma
Florida
New Mexico
New York
Arizona
Georgia
North Carolina
Rhode Island
District of Columbia
California
South Carolina
Tennessee
Texas
Arkansas
Alabama
West Virginia
Kentucky
Mississippi

70%   80%   90%
High school or more

Greatest to least, endpoint + Color
Reference to percent increase

2000 2009

Wyoming
Minnesota
Alaska
New Hampshire
Vermont
Montana
Iowa
Utah
Hawaii
Maine
North Dakota
South Dakota
Wisconsin
Nebraska
Washington
Kansas
Colorado
Oregon
Massachusetts
Connecticut
Idaho
Maryland
Pennsylvania
Ohio
New Jersey
Delaware
District of Columbia
Missouri
Virginia
Indiana
Illinois
Oklahoma
Florida
New York
North Carolina
Arizona
Nevada
Georgia
South Carolina
Tennessee
West Virginia
New Mexico
Arkansas
Louisiana
Alabama
Kentucky
California
Mississippi
Texas

Increase from
2000 to 2009
0%   6%   12%

70%   80%   90%
High school or more

Attainment: Changes (Detail)

Alphabetical

Values look scattered

Greatest to least, endpoint

Focus on most recent numbers

Source: Yau (2013)

Source: Yau (2013)

- An open circle represents the high school attainment in 2000 for each state, and the solid circles represent the same for 2009
- The dots are placed in the same position vertically, and a line is used to connect the two dots
- The longer the line is, the greater the change, by percentage points, was from 2000 to 2009

Visual Cues

- The shift from open circle to closed circle provides a sense of direction
- In this example, high school attainment in all states improved, so your eyes always shift from left to right
- If attainment decreased in one of the states, you could use the same visual cue
- For example, if there were a decrease from 80 percent to 70 percent, the solid dot would be on the left of the open one
- You can also use arrows if you want to highlight direction more prominently
- Given the data, a focus on the magnitude of the changes and the values of the endpoints was more appropriate

Sorting Alphabetically

- You can see how a change in sorting can shift focus
- States are sorted alphabetically in the first chart, and the lack of visual order makes it more challenging to make comparisons
- You can see the increases and it's easy to find a state of interest, but as an overall picture, you don't get much

Attainment: Changes (Detail)

Alphabetical

Values look scattered
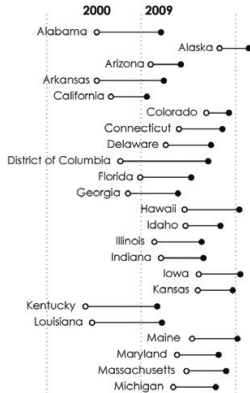
Greatest to least, endpoint

Focus on most recent numbers
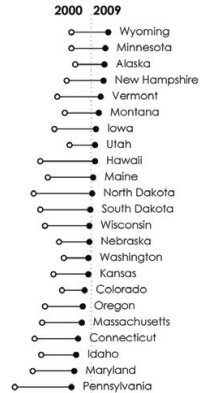
Source: Yau (2013)

Sorting by Attainment

- In contrast, the second chart shows the same data ordered instead by the highest percentage of attainment in 2009
- It starts with Wyoming and goes down to Texas
- This focuses on the more recent estimates, whereas still making it easy to pick out the values for 2000 because generally speaking, states with higher percentages in 2009 were higher in the rankings in 2000, too
- That said, you can also sort by the 2000 estimates and move the labels to the left to shift focus in this direction.

Attainment: Changes (Detail)

Greatest to least, startpoint

Focus on past data

Greatest to least, endpoint + Color
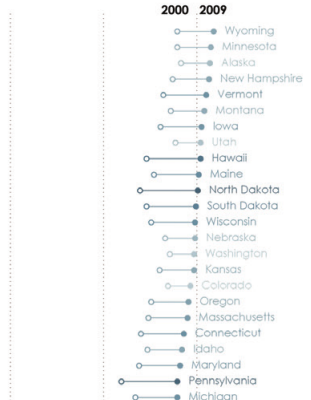
Reference to percent increase

Source: Yau (2013)

Visual Cue: Color

- The chart on the far right introduces color as a visual cue
- This is the same as the second chart that sorts by 2009 estimates, but color is used to highlight states that increase the most by percentage
- The District of Columbia had the greatest percentage increase, so it is shown in black
- The lower the increase, the lighter the states are shown
- States in between are shown with varying shades of green
- So if you look at the individual components of this chart, you get length, position, direction, and color used as visual cues; it uses a Cartesian coordinate system; and a linear numeric scale is used on the horizontal, with a categorical scale on the vertical

Different Representation

- What else can we do?
- As shown next, position and direction can be used differently to show the increases from 2000 to 2009
- Unlike the previous charts, states are plotted on a linear scale that represents high school attainment instead of on a categorical scale
- Values are categorized by year on the horizontal
  - This is essentially a couple of ticks on a time series plot
- If you were to show years in between, there would be more than two categories on the horizontal axis
- In any case, like in a time series plot, a greater slope from point to point means a greater rate of change
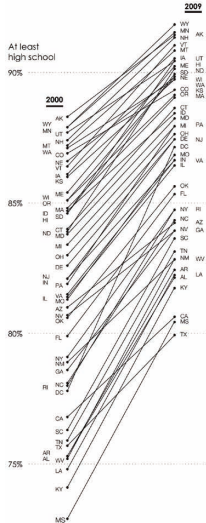
Attainment: Position (& Color)
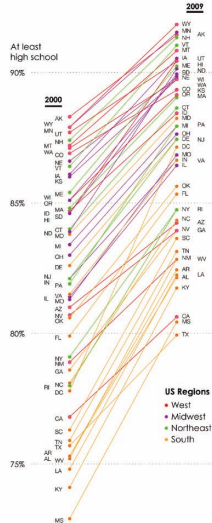
Source: Yau (2013)

- The chart on the right uses the same geometry as the one on the left, and uses color to represent regions in the United States
- So although you see improvement with all states, you also see a lot of the states in the South toward the bottom of the scale and Midwest and West states more toward the top
- Although, as is usually the case with real data, there are exceptions, such as California in the West that is toward the bottom and Maryland that is in the South is higher up

Attainment: Position & Color (Detail)
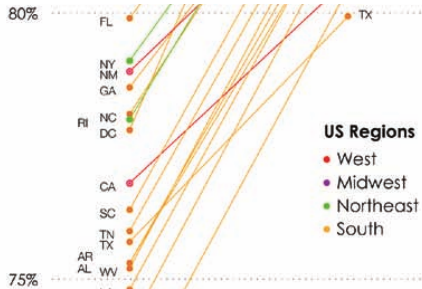
Source: Yau (2013)

Tendency

- Generally speaking though, the higher the attainment in 2000, the higher the attainment was in 2009
- This is obvious in the next figure, which uses position as a visual cue and linear scales on both axes
- High school attainment in 2000 is plotted on the horizontal axis, and attainment in 2009 is on the vertical
- There is an obvious upward trend, and you can spot Washington, DC sticking out somewhat, indicating the higher rate of improvement (and probably difference in demographics)
- You can also see Texas and California lagging around the bottom-left corner
- As shown in previous charts, you can incorporate other visual cues such as color, symbols, or both to provide additional dimensions of information

Attainment: Scatterplot
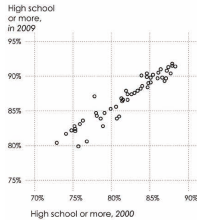
Source: Yau (2013)
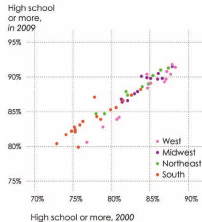
Components

Example

Tutorial
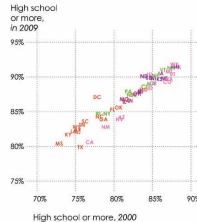
References

Source: Yau (2013)

Attainment: Scatterplot
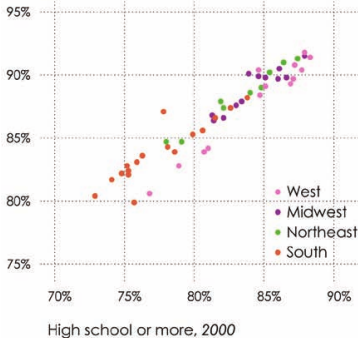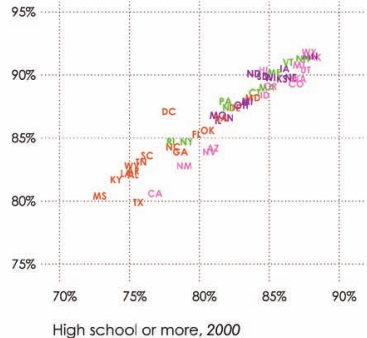
Source: Yau (2013)

- Remember this is geographic data, so you must map it, right?
- Actually, just because location is attached to your data, which seems like almost always these days, a map is not always the most useful view
- The next Figure shows a handful of maps with states colored using varying scales and metrics, which are called choropleth maps

Attainment: Map

Source: Yau (2013)

- Note that although each map uses the same method, the choice of scale can change the map's focus and message
- For example, the map on the top left uses a quartile scale, which means the states were split into four even groups based on a metric
- In this case, the metric is the percentage of people with a bachelor's degree in 2009
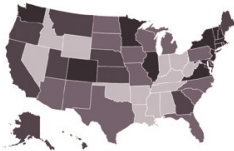- This makes a map with colors that are evenly distributed

Attainment: Map (Detail)

Source: Yau (2013)

Variation
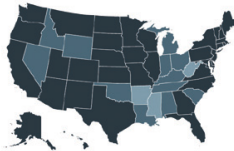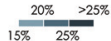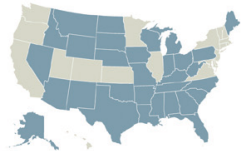
- However, the map that shows the same data on a linear scale, with just three shades of green, shows darker shades in the Midwest and Northeast regions
- Compare this with the quartile map, and you still get the lighter areas in the South, but the rest of the map tells a different story
- Likewise, you can further abstract the data by coloring states by whether they are below or above the aver- age (top right) or whether percentages increased or decreased (bottom right)
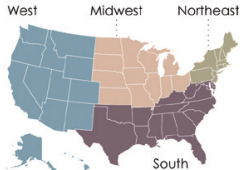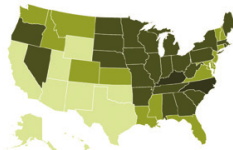
Categorical
Groups based on metadata, such as region

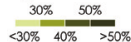**Difference**
A linear scale, but based on percent change between years

Categorical difference
Simple split based on increase or decrease (Good news: all increase in this example)

West   Midwest   Northeast

South

% change from 1990 to 2009

30%   50%

<30%   40%   >50%

Change from 1990 to 2009

■ Decrease   ■ Increase

Source: Yau (2013)

Multi-Focus

- As shown next, you can also show several maps at once to see how something has changed geographically over time
- Since you've looked the data from several perspectives already, you know that a high value in 2000 generally means a higher value in 2009, because the states improved at similar rates
- You see about the same thing when you compare 1990 to 2000
- In 1990, you see a more lightly colored map, where several states showed 15 percent or less of people 25 years or older with a bachelor's degree
- Only Wyoming, which had the highest percentage in 2009, shows a percentage higher than 25 percent
- As you move left to right, the map gets darker, like you'd expect
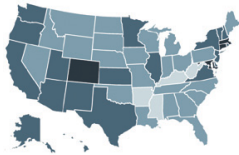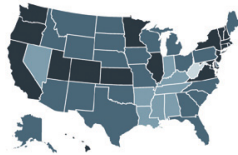
Attainment: Several Map

Geographic coordinates + Time scale + Color
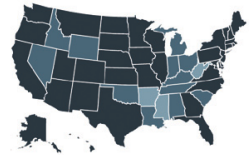
In 1990, about 20 percent of 25+ year olds had at least a bachelor's degree.

In 2000, the percentage was up to 24.

In 2009, the US average was up to 29 percent.

Bachelor's degree or more

15% or less   20%   More than 25%

Source: Yau (2013)

Outline

1 **Components**

2 **Example**

3 **Tutorial**

Assignment 7.1: Classification & Alternatives
Group Work

- Classify the Visualizations of the Educational Attainment according to the framework we have introduced
- What are differences in classification, if any, for the different chart types we have used?
- What other Visualizations would be interesting?
    - Data Preparation
    - Story
    - Chart types

Framework

Level

○ Micro level

◑ Meso level

● Macro level

Type

🛈 Profiling

🕓 Temporal

🌍 Geospatial

≡ Topical

▽ Network

Audience

⚥ Gender

⑤ Age

✍ Education

♿ Disability

☐ Context, e.g.

   ☕ Leisure

   🏭 Business

   ⚗ Scientific

   † Religious

   ☐ Other

Medium

✎ Printed

⌨ Digital

☣ Time-based

⊙ Spatial

Ⓐ With Text

🔔 With Sound

☞ Interactive

☐ Other

Assignment 7.2: Exam Questions

Group Work/Individual Assignment

- English or German?
- What kind of question do you expect in the exam?
- Design sample questions

- A discussion thread for exam questions has been opened in the LearnWeb

Representation

Jörg Cassens

Data and Process Visualization
SoSe 2017

medieninformatik
IMAI – Institut für
Mathematik und
Angewandte Informatik

Components

Example

Tutorial

References

Mullin, J. F. and O'Brien, I. R., editors (2012). *Statistical Abstract of the United States: 2012 (131st Edition)*. United States Census Bureau.

Yau, N. (2013). *Data Points – Visualization that means something*. Wiley.