Statistical Computing and Data Visualization in R

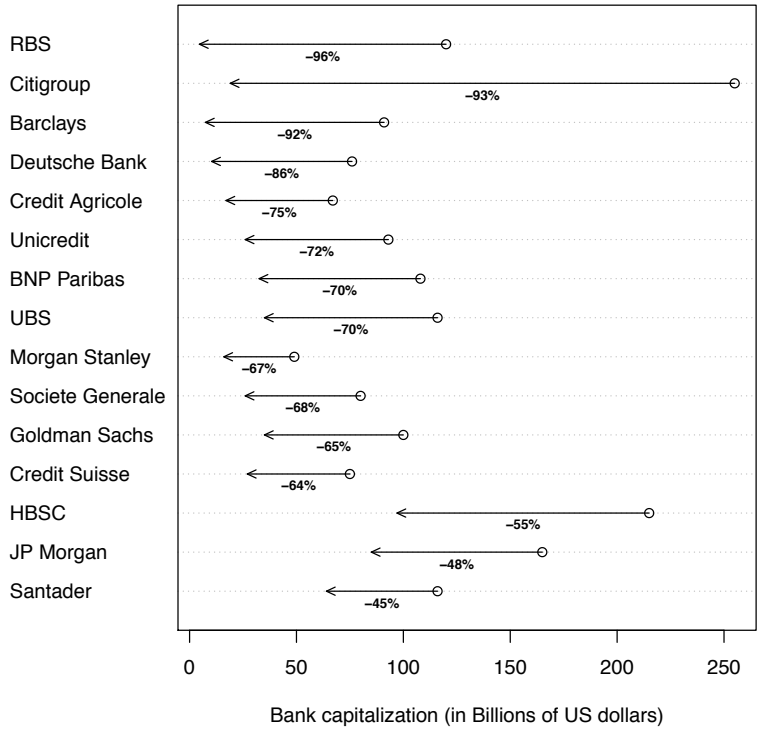
Lecture 5

Principles of Data Visualization

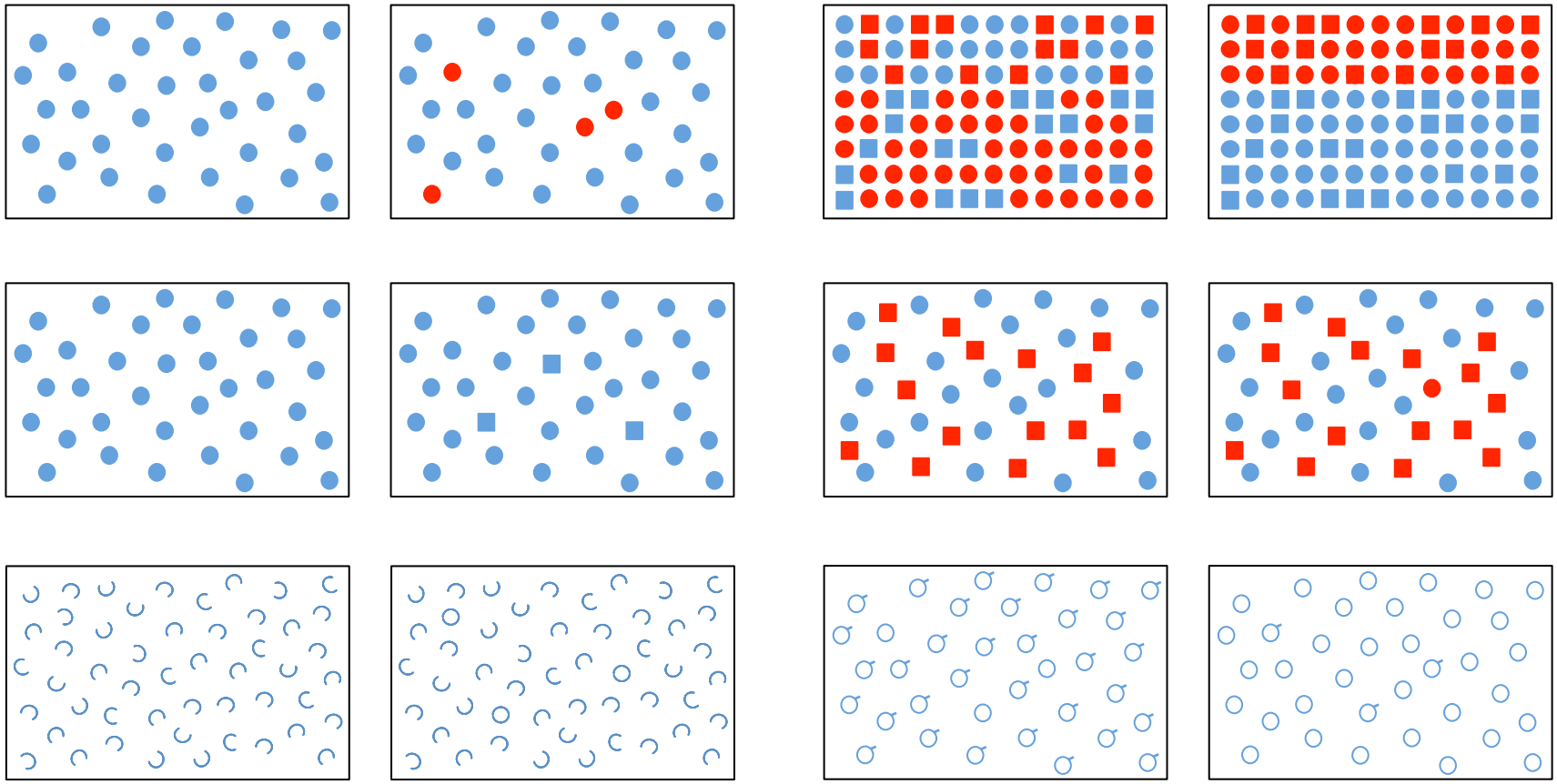
A picture is worth a 1,000 words …

Evolution has made our brains good at grasping visual cues but not so good at looking for patterns in raw numbers!

Bank	*January 2007*	*January 2009*
RBS	120.0	4.6
Citigroup	255.0	19.0
Barclays	91.0	7.4
Deutsche Bank	76.0	10.3
Credit Agricole	67.0	17.0
Unicredit	93.0	26.0
BNP Paribas	108.0	32.5
UBS	116.0	35.0
Morgan Stanley	49.0	16.0
Societe Generale	80.0	26.0
Goldman Sachs	100.0	35.0
Credit Suisse	75.0	27.0
HBSC	215.0	97.0
JP Morgan	165.0	85.0
Santader	116.0	64.0



But not all visual cues are born equal



Pre-attentive processing

- Pre-attentive processing is the ability of the low-level human visual system to rapidly identify certain basic visual properties.
- Typically, tasks that can be performed on large multi-element displays in less than 200 to 250 milliseconds (msec) are considered pre-attentive (processed in parallel by the low-level visual system).
- The term pre-attentive processing is a misnomer.

What is data visualization?

- "Information visualization is the use of computer-supported interactive visual representations of abstract data to amplify cognition." – Card, Mackinlay and Schneiderman (1999) Readings in Information Visualization: Using Vision to Think.
  - I would slightly modify it to say "… of abstract data to amplify cognition **and/or communicate information** ."

Exploration vs. Presentation

- This definition of data visualization highlights its two main roles:
  - Amplify cognition (Exploration): how do we find hidden patterns in data, facts worth highlighting that are not obvious by simply looking at data. EDA has a long history in statistics (Tukey)!
  - Communicate information (Presentation): once we have found interesting patterns, how do we make other people aware of them?
- The two tasks are interrelated, but the sometimes require slightly different solutions.

Insights from multiple disciplines …

- Graphic design: Emphasizes aesthetics.
- Computer science: Emphasizes algorithms.
- Cognitive psychology: Provides insights into the most effective tools.
- Journalism: Emphasizes storytelling.
- Statistics: Emphasizes quantification of information.

A (Very) Short History of Visualization

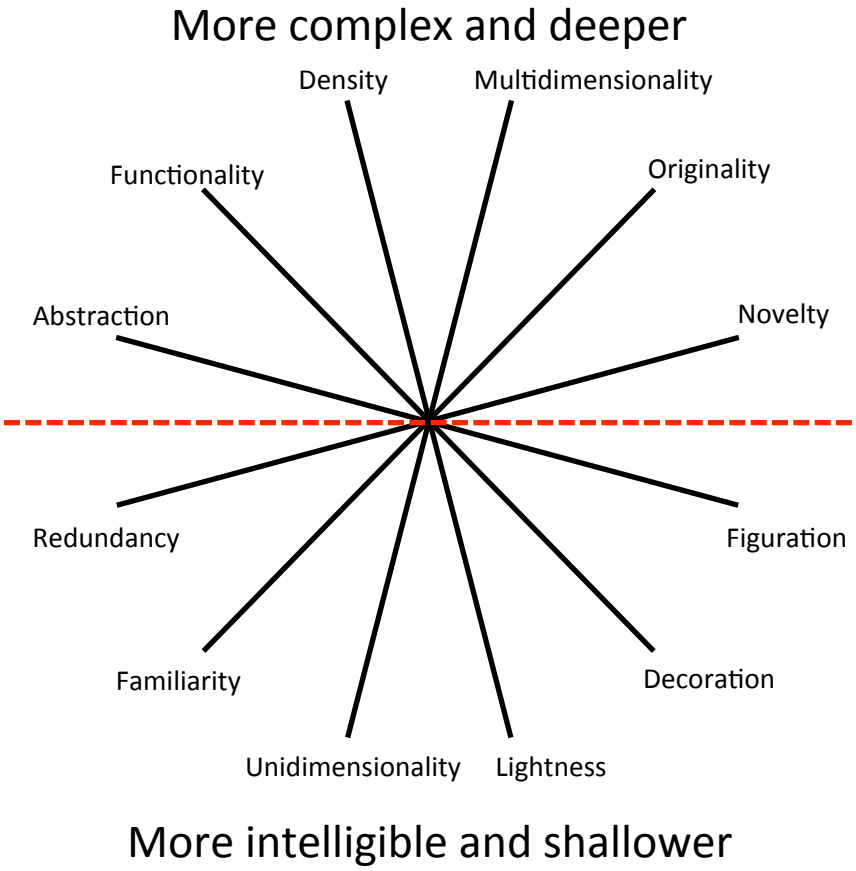
- 1637 – Descartes first uses 2D grids to visually encode numbers.
- 1786 – William Playfair's "The Commercial and Political Atlas".
- 1855 – John Snow uses maps to link the 1854 London cholera epidemic to contaminated drinking water.
- 1857 – Florence Nightingale uses stacked bar and pie charts to persuade Queen Victoria to improve conditions on British military hospitals.
- 1954 – Darrel Huff's "How to Lie with Statistics".
- 1977 – John Tukey introduces boxplots.
- 1983 – Edward Tufte's "Visual Displays of Quantitative Information".
- 1994 – William Cleveland's "The Elements of Graphing Data".
- 2004 – Stephen Few "Show me the Numbers".
- Nowadays dominated by computer scientists (on the technical side) and business analytics (on the more applied side).

Are there situations where a table is better than a graph?

- Yes, but these are relative exceptions.
  - To convey a handful of numbers.
  - To report precise values for lookup.
  - To present many different types quantities (dimensions) for a small number of cases.
- Tables are usually a bad idea if comparison is important.
- I will not discuss principles of table design, but they are similar to the ones for visualization!

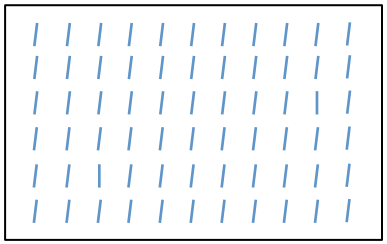
Complexity vs. comprehension

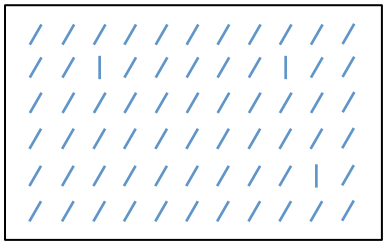
- There are clear trade-offs in creating graphs.
- More complex and deeper graphs are usually also harder to understand.
- The complexity of a graph can be evaluated using a six-component scale (due to Alberto Cairo).

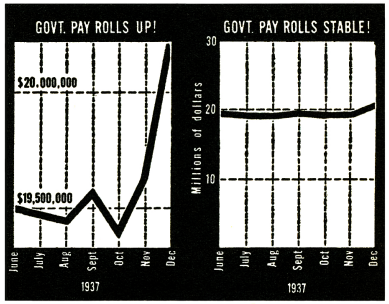


A few lessons from cognitive psychology …

- Attention is drawn to large perceptible differences: humans think in terms of differences.
- People expect changes in properties to carry information.
- Form and meaning need to be compatible.
- People can only hold in mind up to four groups of information at once.
- People automatically group elements into units.
- Try to maximize data/ink ratio.
- When possible, interactivity is your friend.







Blue Red

[_] _][

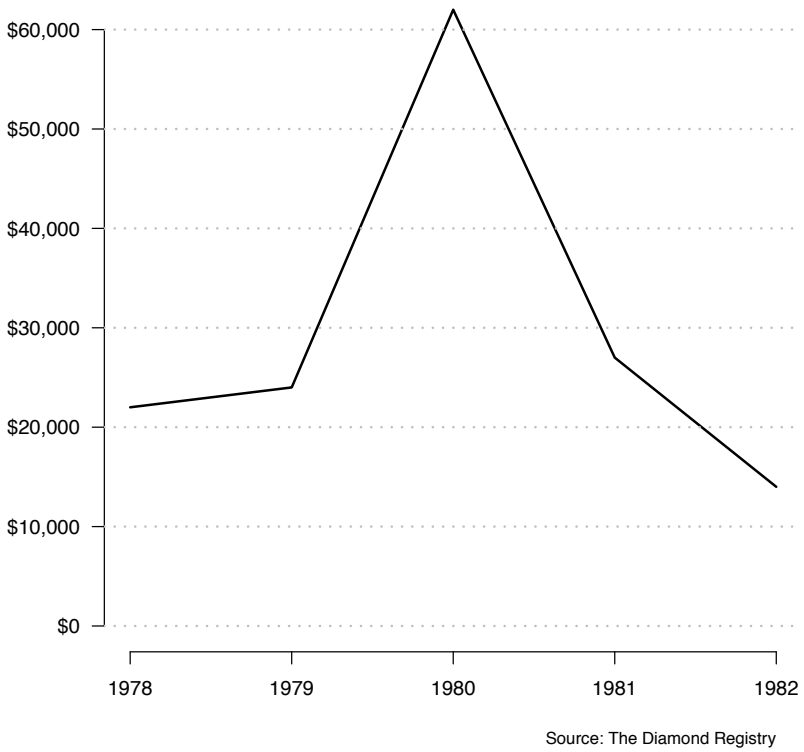
XXOO XO XK

http://www.nytimes.com/interactive/2014/04/24/upshot/facebook-baseball-map.html

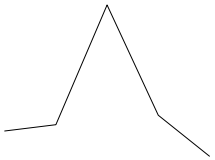
The structure of visualizations

- Visualizations can be divided into four components:
  - The **visual cues** that are used to represent each variable in the data.
  - The **coordinate system** that is used for each variable.
  - The **scale** that is used to represent each variable.
  - The **contextual information** that is used to help readers understand the visualization.

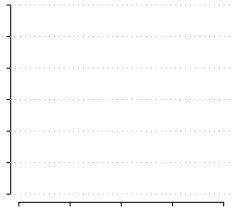
The structure of visualizations



Visual cues



Coordinate system



Scale

Context

Visual cues

- The key to making graphs that are visually appealing and easier to understand is to use visual cues that engage **pre-attentive processing** (position, color, shape, etc.) to transmit the most important information.
- However, not all visual cues are equally effective, and their effectiveness depends on whether you represent quantitative or qualitative variables.

Cognitive scale of visual cues for quantitative variables

Position

Length

Angle

Direction

Shape

Area

Volume

Color saturation

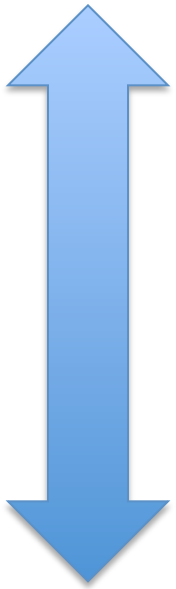
Color Hue

Dot charts
Bar plots
Scatterplots

Piecharts

Heatmaps

More accurate comparisons



Less accurate comparisons

Cognitive scale of visual cues for qualitative variables

Color Hue

Orientation

Shape

Color Intensity

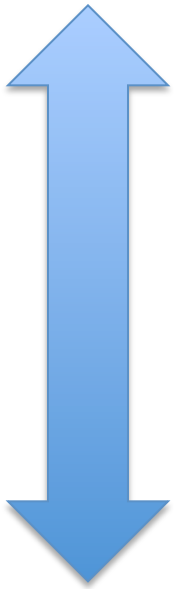
Size

Curvature

Added marks

Closure

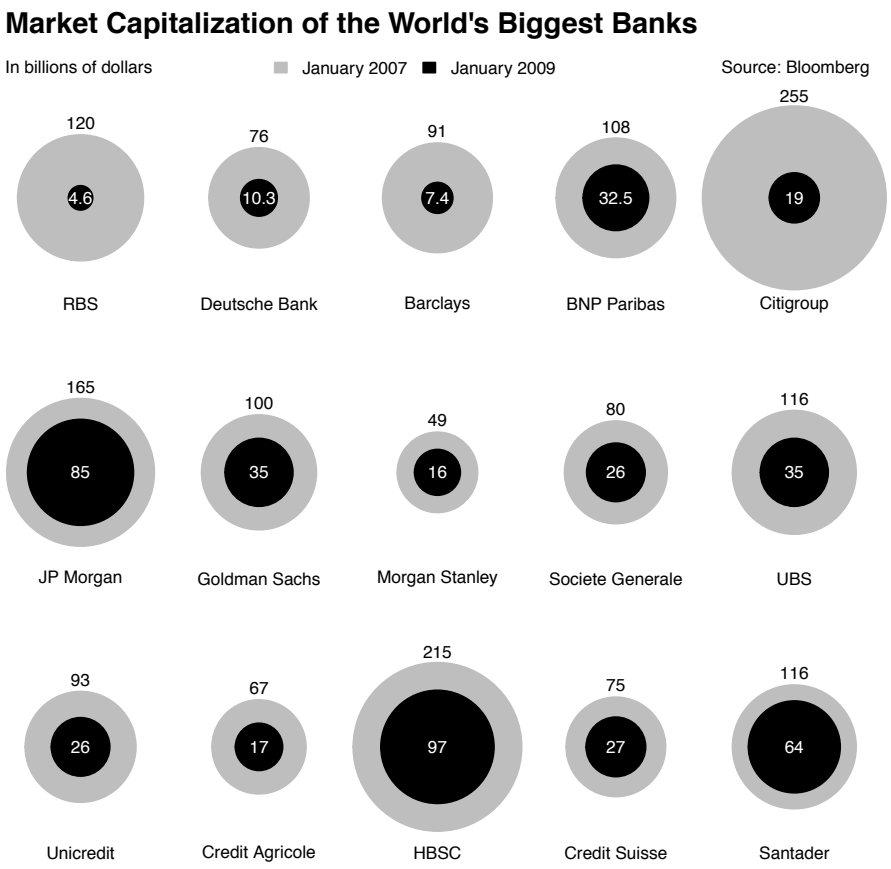
More accurate comparisons



Less accurate comparisons

The accuracy of visual clues

- This graph shows the market capitalization of the worlds biggest banks in January 2007 and January 2009.
- The original version was published by J.P. Morgan. This is a reinterpretation of the original graph (which we will see later).

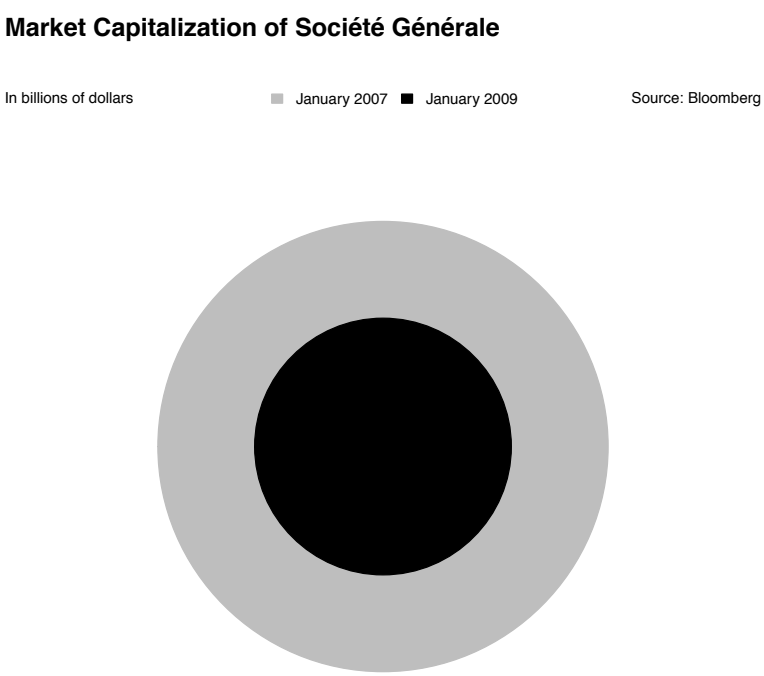


The accuracy of visual clues

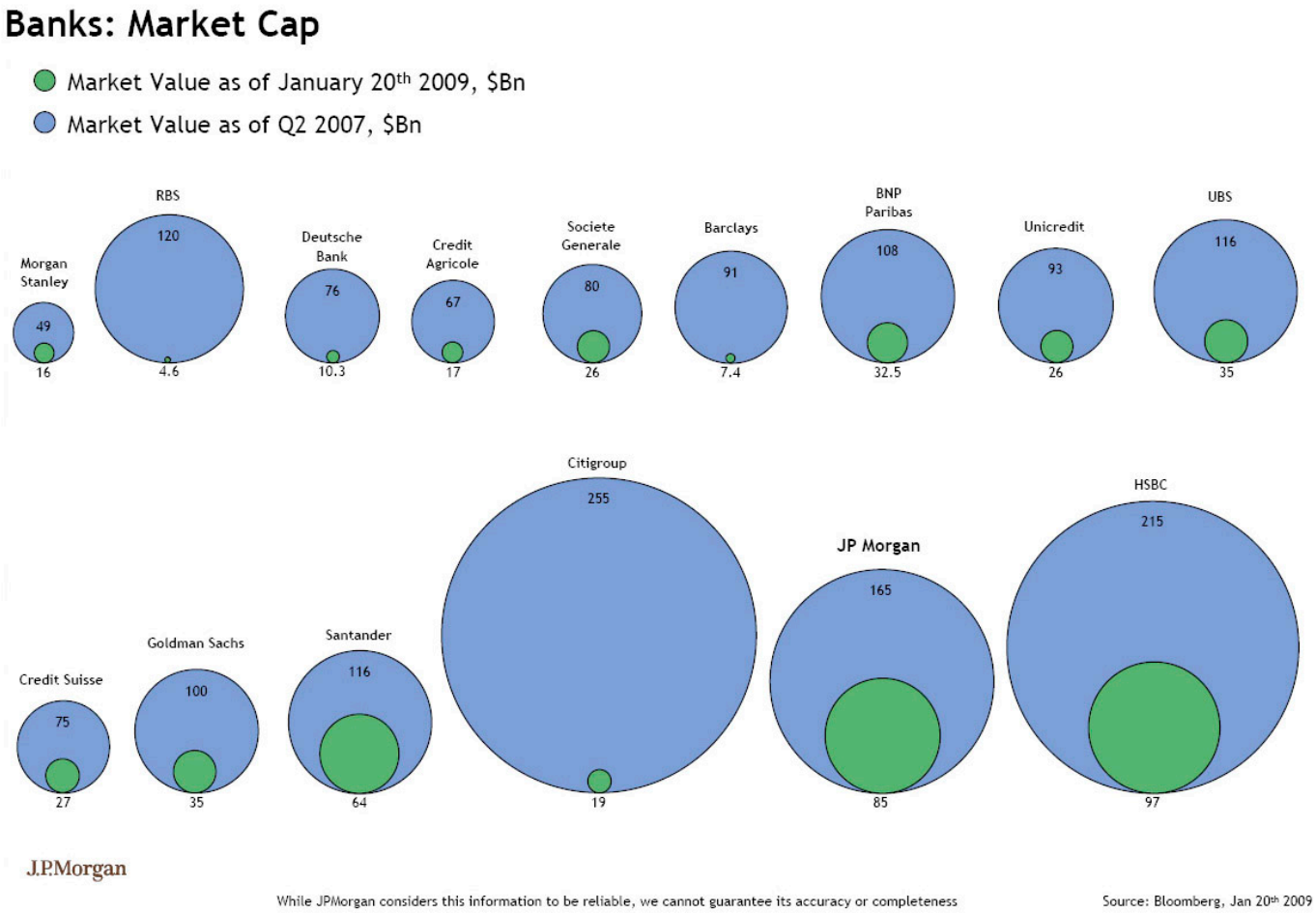
- What is the main point of the visualization?
  - Did banks increase or lost market capitalization?
  - Which banks lost the most?
  - Which banks lost the least?
  - How much market values was lost by the biggest losers?
  - How much market values was lost by the banks that fared better?
- The previous graph is good main message (banks lost market capitalization, the bigger losers were RBS and Citigroup), but not appropriate for making detailed comparisons!

The accuracy of visual clues

- Let's focus on one specific bank.
- If the largest bubble represents $80 billion, how much money does the second bubble represents?
  - Slightly less than than $40 billion?
  - Slightly more than $25 billion?
  - Slightly more than $50 billions?



Other issues with areas

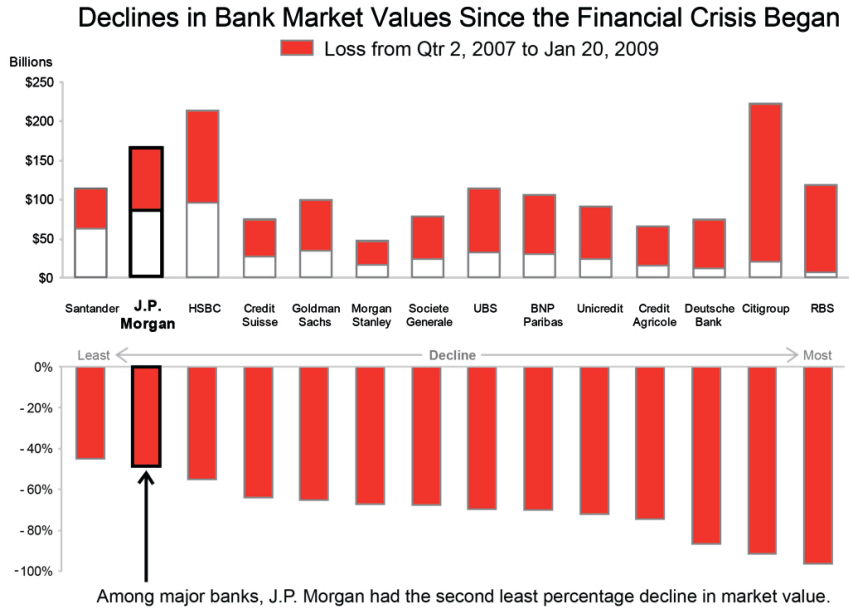


Original visualization created by J.P. Morgan Bank. Taken from http://www.perceptualedge.com/example18.php

Other issues with areas

- Why does the graphs appear to give the wrong message?
  - Values are encoded through the **diameters**, but the eye is might be trying compare the areas! This is a common problem with bubble plots.
  - This "trick" helps in making the differences across banks appear bigger than they really are.
  - What makes it worse is that there is no hint in the graph that suggests that you should be looking at the diameters rather than the areas.

Stephen Few's visualization for the Bank's market capitalization

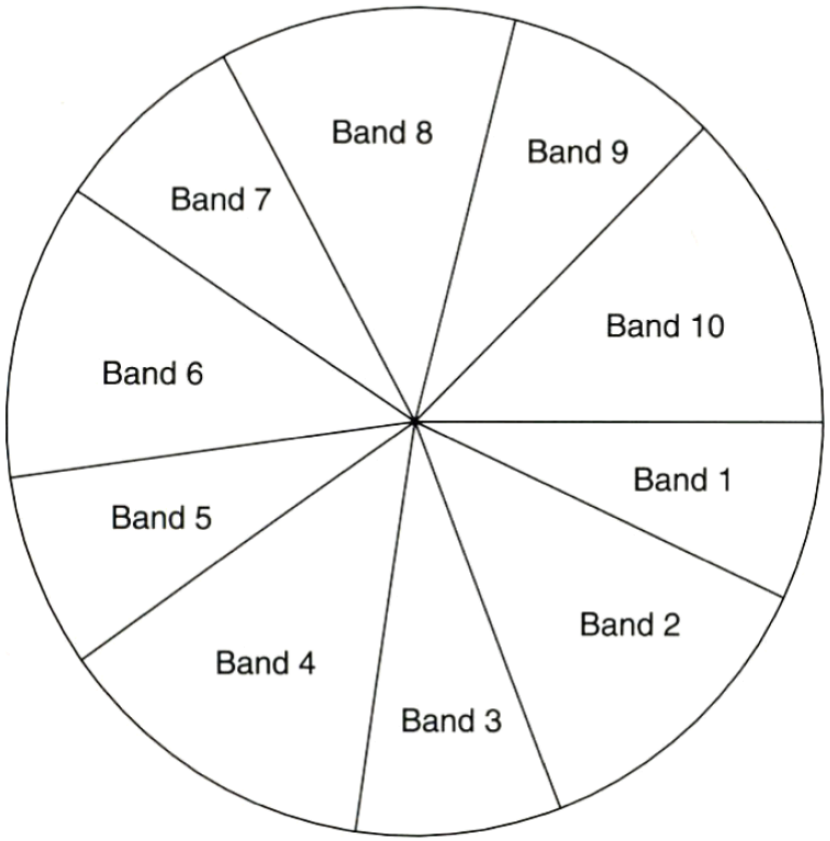


- Stephen Few suggests using two sets of bar plots, one showing absolute values, and a second one showing relative declines.
- Much easier to make comparisons because cue is higher in the cognitive scale.
- Additional text highlights J.P. Morgan!

Taken from http://www.perceptualedge.com/example18.php

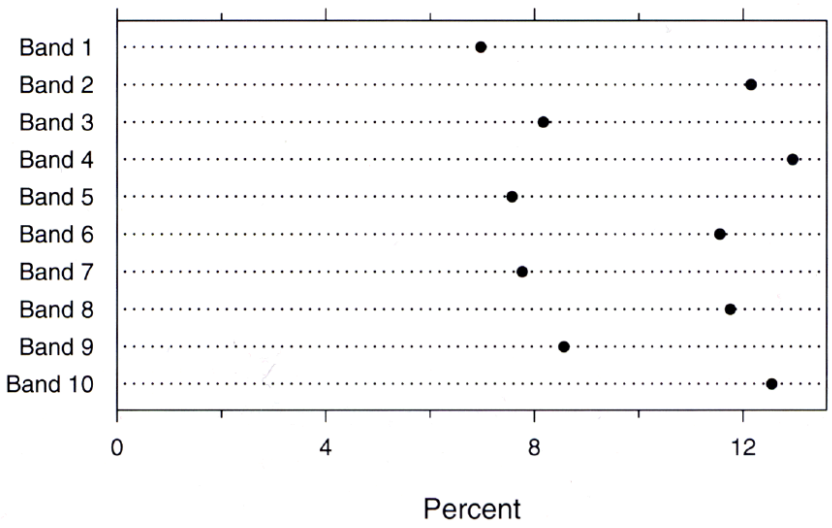
Areas and pie charts

- This is an example of a pie chart, which are extremely popular.
- Can you describe these data? What are the relative sizes of the slices?



From “The Elements of Graphing Data”, by Cleveland.

The accuracy of visual clues



From "The Elements of Graphing Data", by Cleveland.

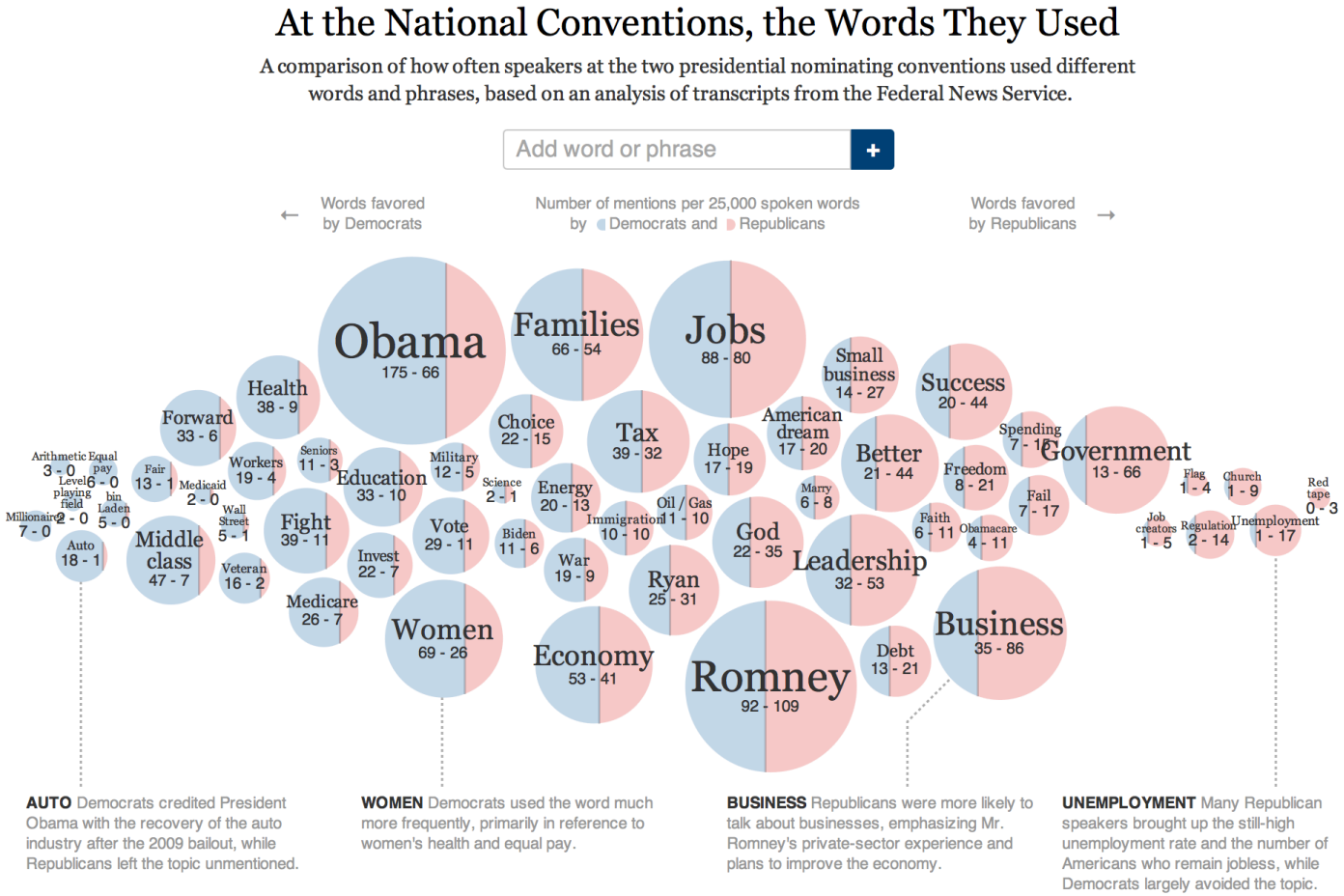
- An alternative solution is to use a dot chart(which uses visual cues that are higher in our cognitive scale).
- Did you realize that some slices are 50% bigger than others?

Take home message …

- Try to avoid bubbles/areas if accurate comparisons are important.
- There is almost always a better alternative than pie charts!
- However, areas and bubbles can still be a helpful tool if accurate comparisons are not key and either:
  - You want to build in redundancy in your visualization.
  - Other visual cues have been used for other variables.

Redundancy in visualization

The relative usage of the word by democrats is encoded by both the blue area in each circle and by the position of the circle in the x axis.



http://www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html?_r=0

Using color

- When used to represent continuous variables, color is at the bottom of the cognitive hierarchy, as it is does not allow for very effective comparisons. However, it can still be helpful in relaying the “big picture”.
- When dealing with categorical variables, color can be a very useful visual clue that allows us to quickly differentiate among groups, particularly when their number is moderate.

How the eye perceives color

- Human eyes contain two kinds of photo-receptive elements:
  - Rods: sensitive to brightness. Single photon receptors, little use in sunlight.
  - Cones: Come in three varieties.
    - Red: most sensitive.
    - Green: moderately sensitive.
    - Blue: weak.

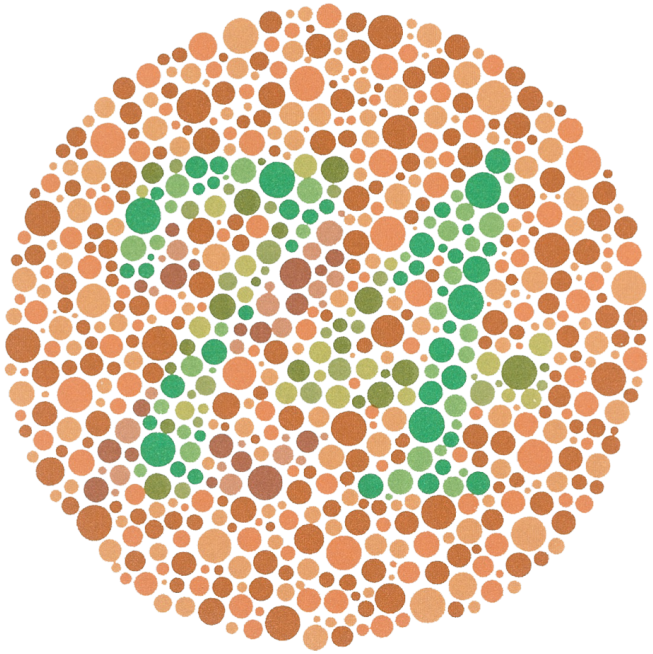
    Humans are best at seeing red, worst at seeing blue!

    Mixing the three (human) primary colors yield any color humans can see.

How the eye perceives color

- The number of cones determines how many primary colors a species perceives:
  - Dogs have only two cones and are red-green colorblind.
  - Chickens have 12 different types of cones!
- Colorblindness (sometimes called daltonism) is a genetic condition that manifest in inability to perceive certain colors.
  - Different types, but red-green is by far the most common, and affects between 8% and 10% of males. Very few women suffer from color blindness.

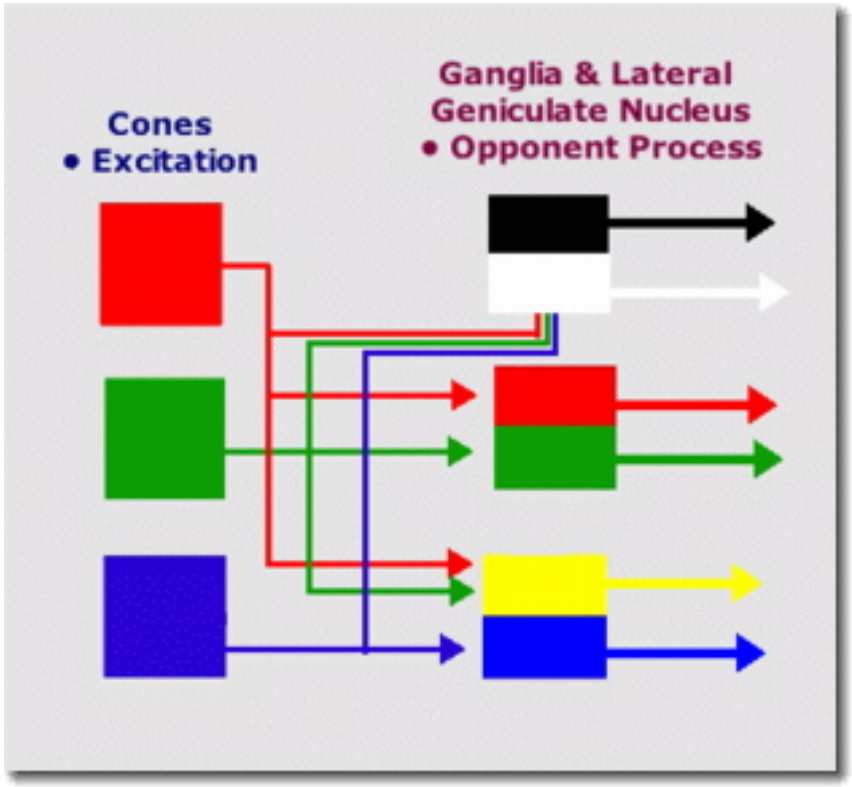
How the eye perceives color



- Ishihara color test: can you see the number 74?
- Consider color blind individuals when creating your visualization: avoid using both green and red in your visualization, particularly if they appear adjacent to each other.

How the eye perceives color

- Opponent color (or antagonistic color) theory : the human visual system interprets information about color by processing signals from cones and rods in an antagonistic manner.
  - Red is contrasted to green.
  - Blue is contrasted to yellow.
  - Black is contrasted to white.
- Mutually exclusive (nobody ever says “greenish red” or “yellowish-blue”).

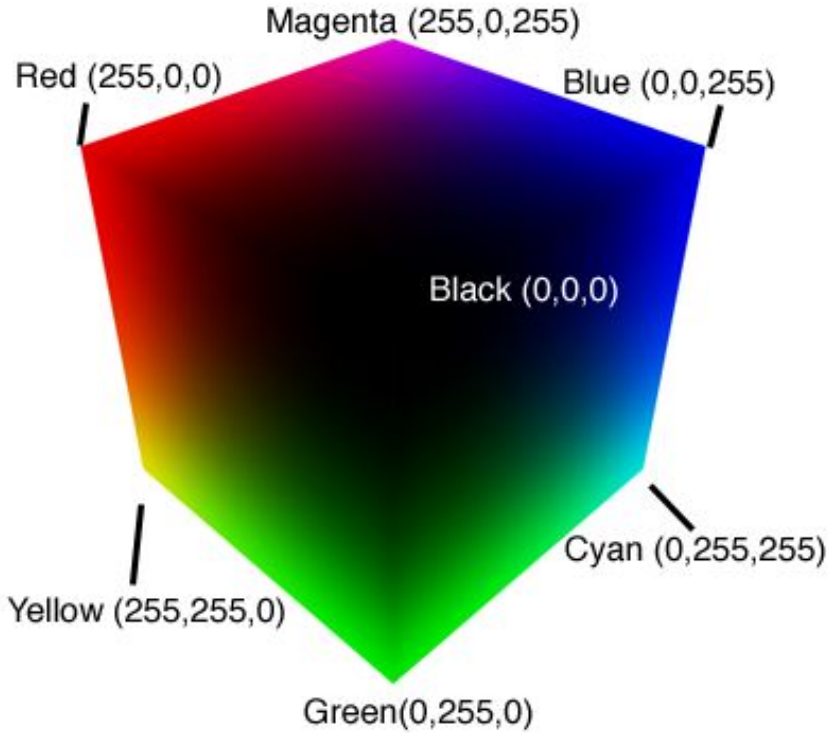


Representing colors

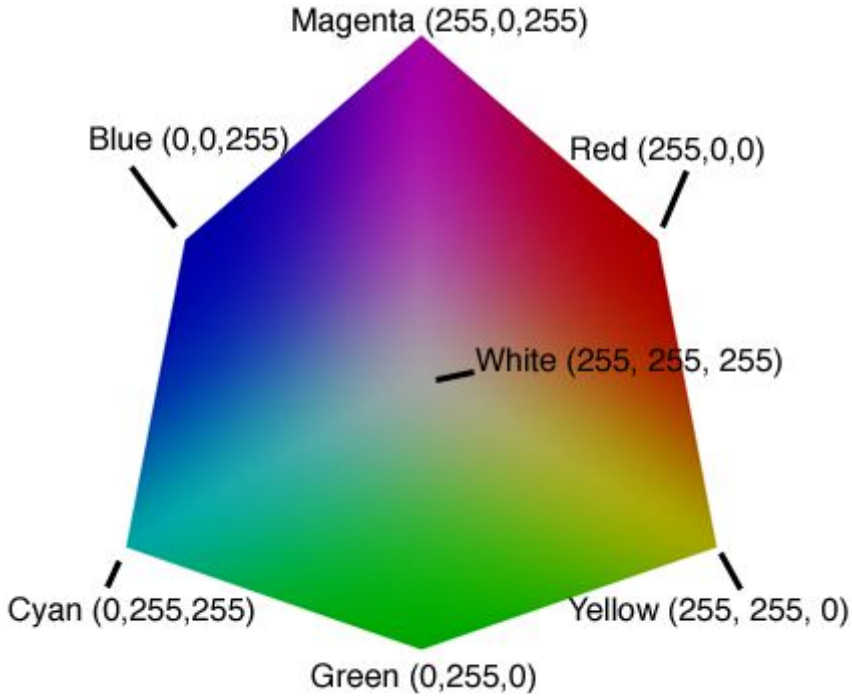
- Two main types of color scales:
  - Computer scales: mainly CMYK (for Cyan- Magenta-Yellow-Black) and **RGB** scale (for Red-Green-Blue).
  - Perceptual scales: PANTONE® Color System, **HSV** (for Hue-Saturation-Value), **CIElab** (where l stands for luminance – white vs. black –, a stands for red vs. green, and b stands for blue vs. green)
- Hue is what we usually call color (blue, green, etc), Saturation is the richness of the color (solids vs. pastels) and value is how much black or white is mixed in (stop-sign red vs. red wine).

The RGB scale

The black corner of the RGB cube

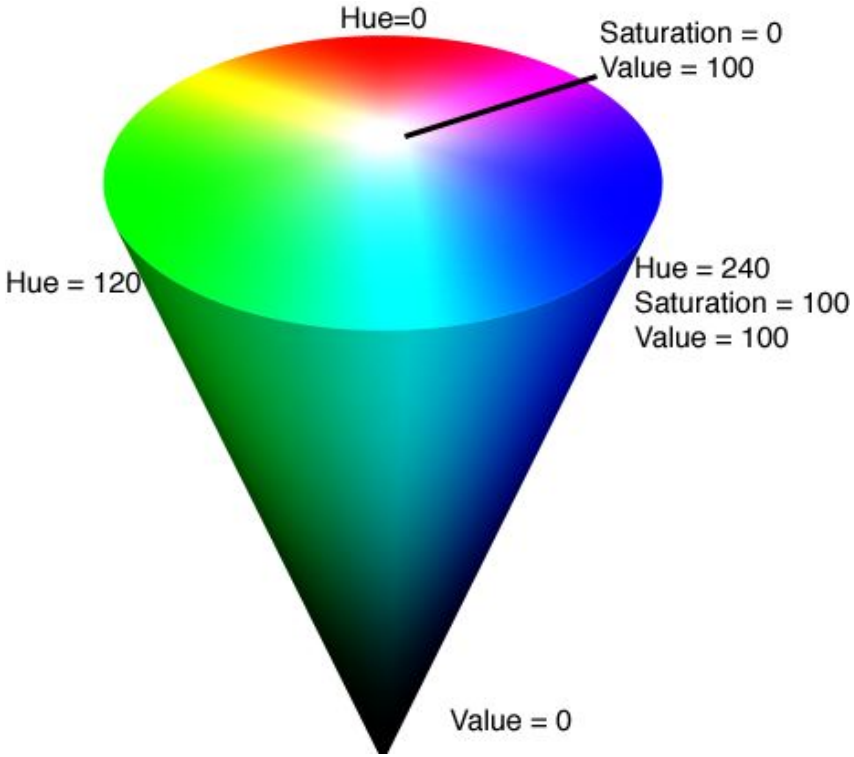


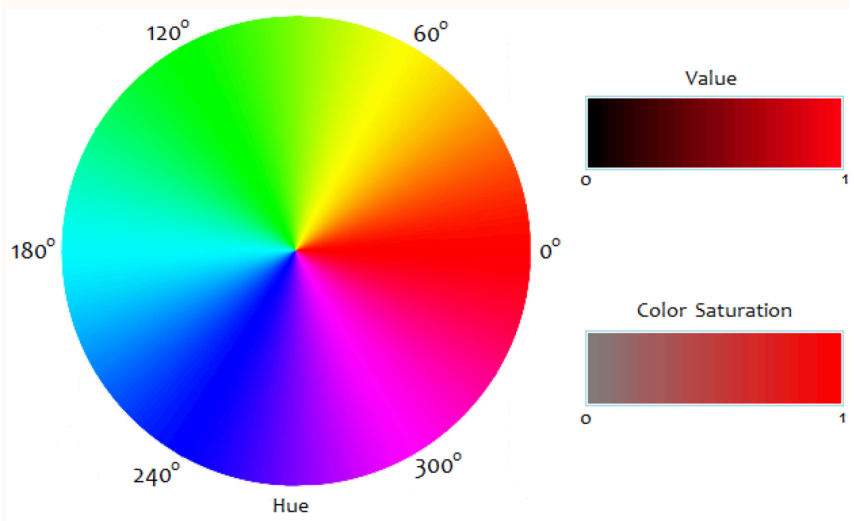
The white corner of the RGB cube



The RGB scale is mathematically beautiful, but not very intuitive for the regular individual.

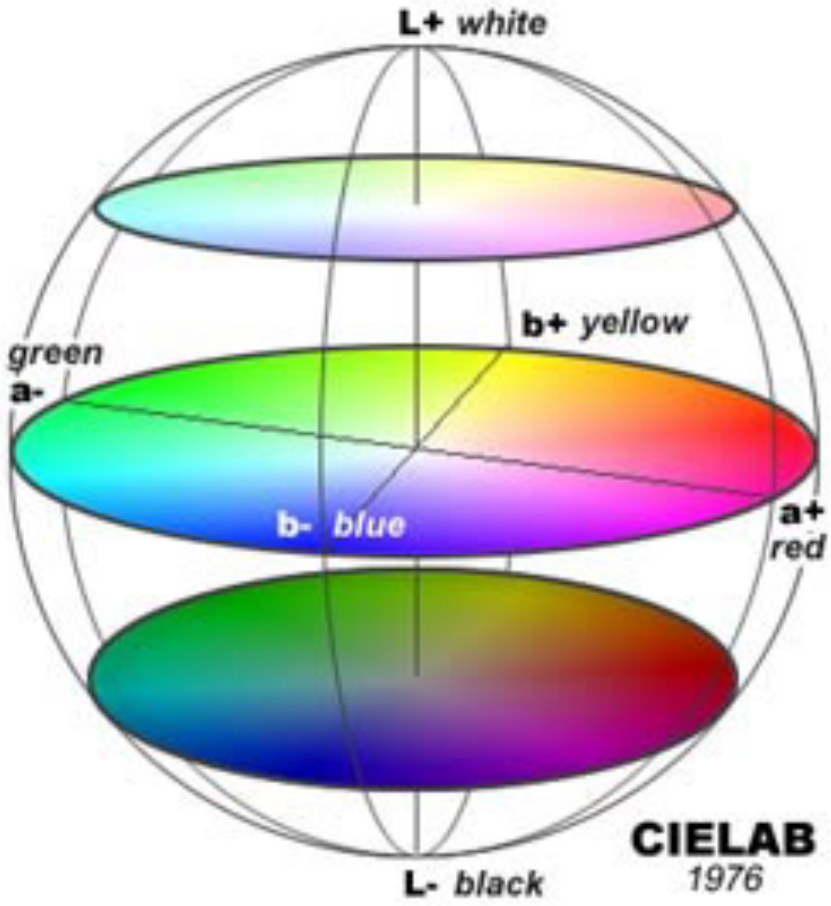
The HSV scale

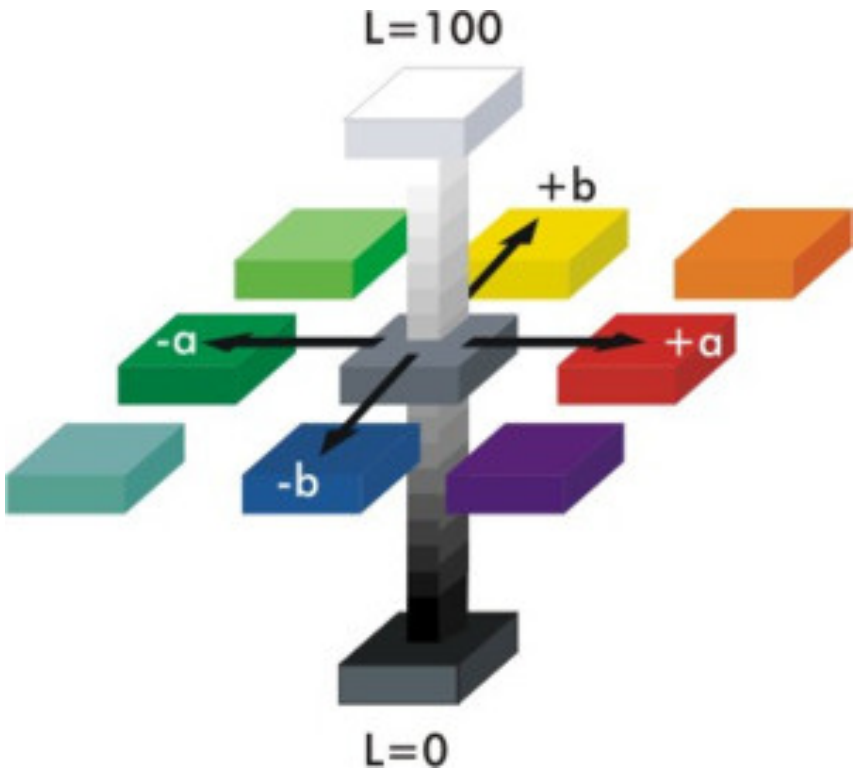




HSV is easier for humans to interpret (e.g., facilitates color selection and color matching) , but does not necessarily good differentiation.

The CIElab scale

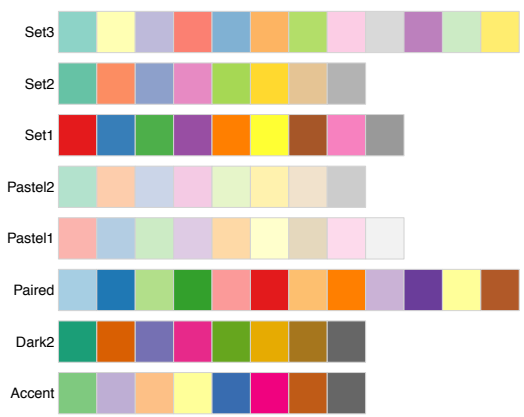




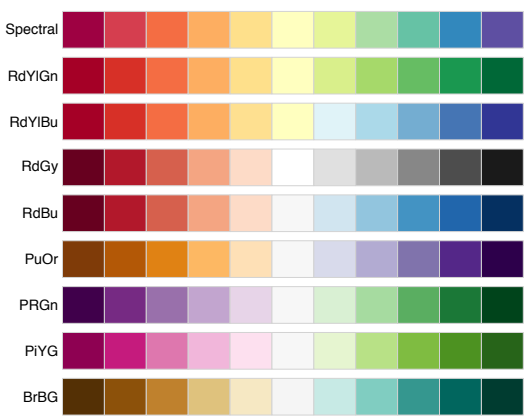
CIElab scale is the brain's color scale, and the best for picking colors for scientific visualization.

Examples of CIElab palettes

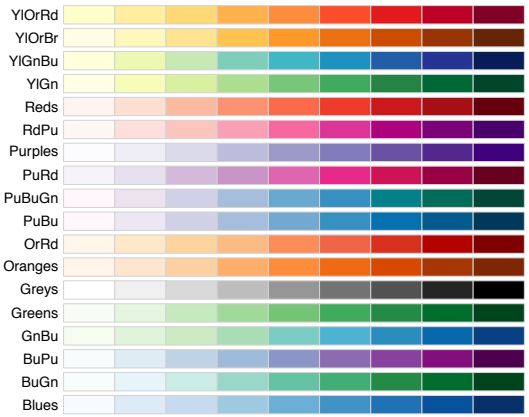
Qualitative



Quantitative (divergent)



Quantitative (sequential)



- These are Color Brewer palettes,

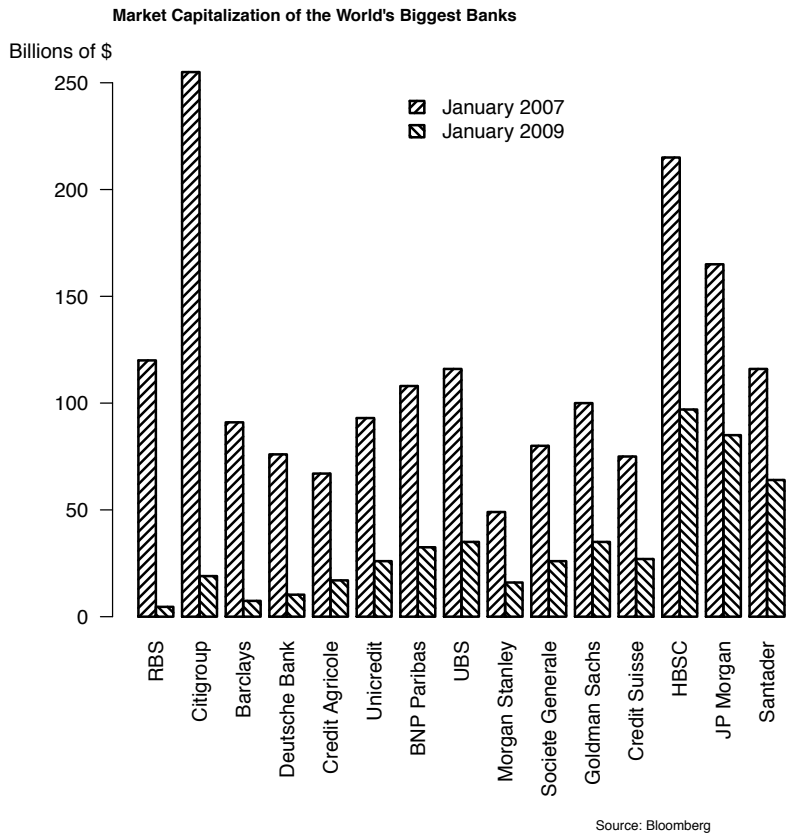
http://www.personal.psu.edu/cab38/ColorBrewer/ColorBrewer_intro.html

General rules

- When representing categorical variables, either use conventions (e.g., Republicans in red and Democrats in blue). If no convention exists use colors that achieve equal pairwise distinction.
- When representing continuous quantities, use smooth gradients.
- Avoid overlapping colors with similar brightness.
- Use pastels for large area colors and saturated colors for small points (e.g., see the “paired” palette in the previous slide).

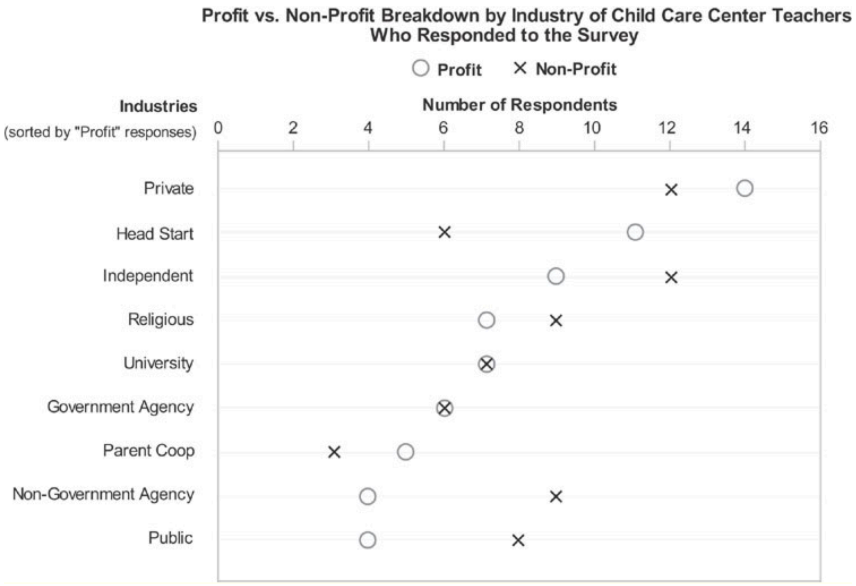
Shading lines as alternatives to color

- Shading lines are a common alternative to color:
  - Avoid shading lines, particularly dissonant combinations! The *moiré vibration* effect.
  - If you can't use color, use shades of grey for a more pleasant look.



Encoding through shape

- As color, shape is most useful to encode categorical variables.
- When multiple points are plotted on each line and overlaps exist, shape can be more effective than color.
- However, redundancy would not hurt in this case.
- What would be good options if you were to plot a third variable?



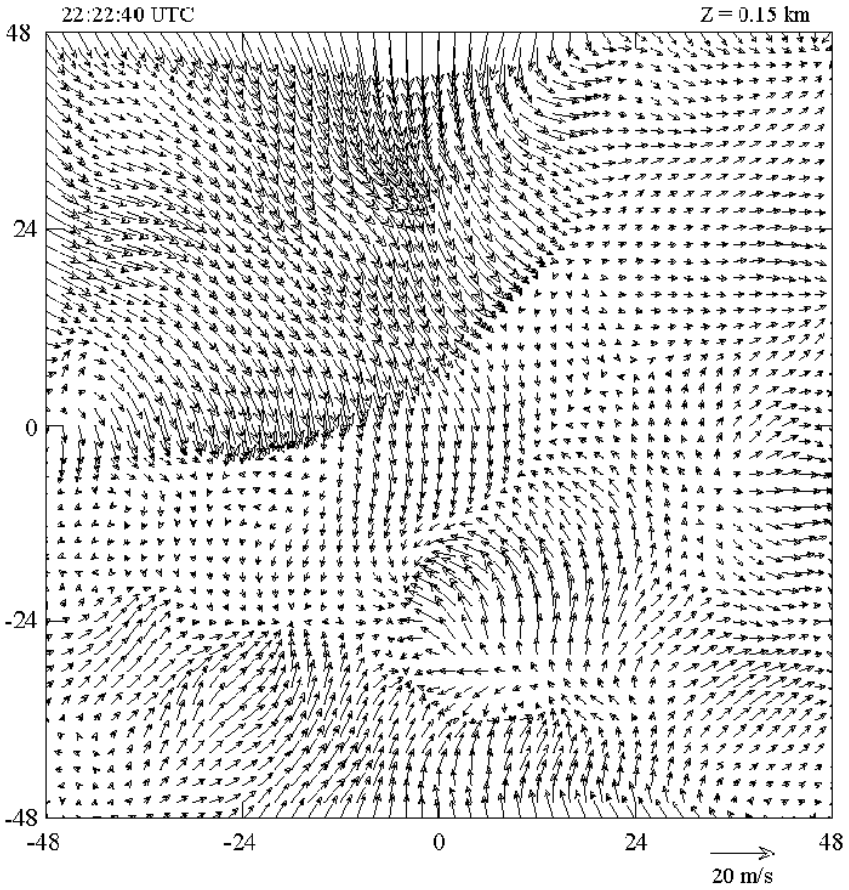
Encoding through shape

- However …

  "Pre-attentive symbols become less distinct as the variety of distracters increases. It is easy to spot a single hawk in a sky full of pigeons, but if the sky contains a greater vaiety of birds, the hawks will be more difficult to see. A number of studies have shown that the immediacy of any pre-attentive cue declines as the variety of alternative patterns increases, even if all the distracting patterns are individually distinct from the target."

  Colin Ware (2000) "Information Visualization: Perception and Design".

Encoding through shape

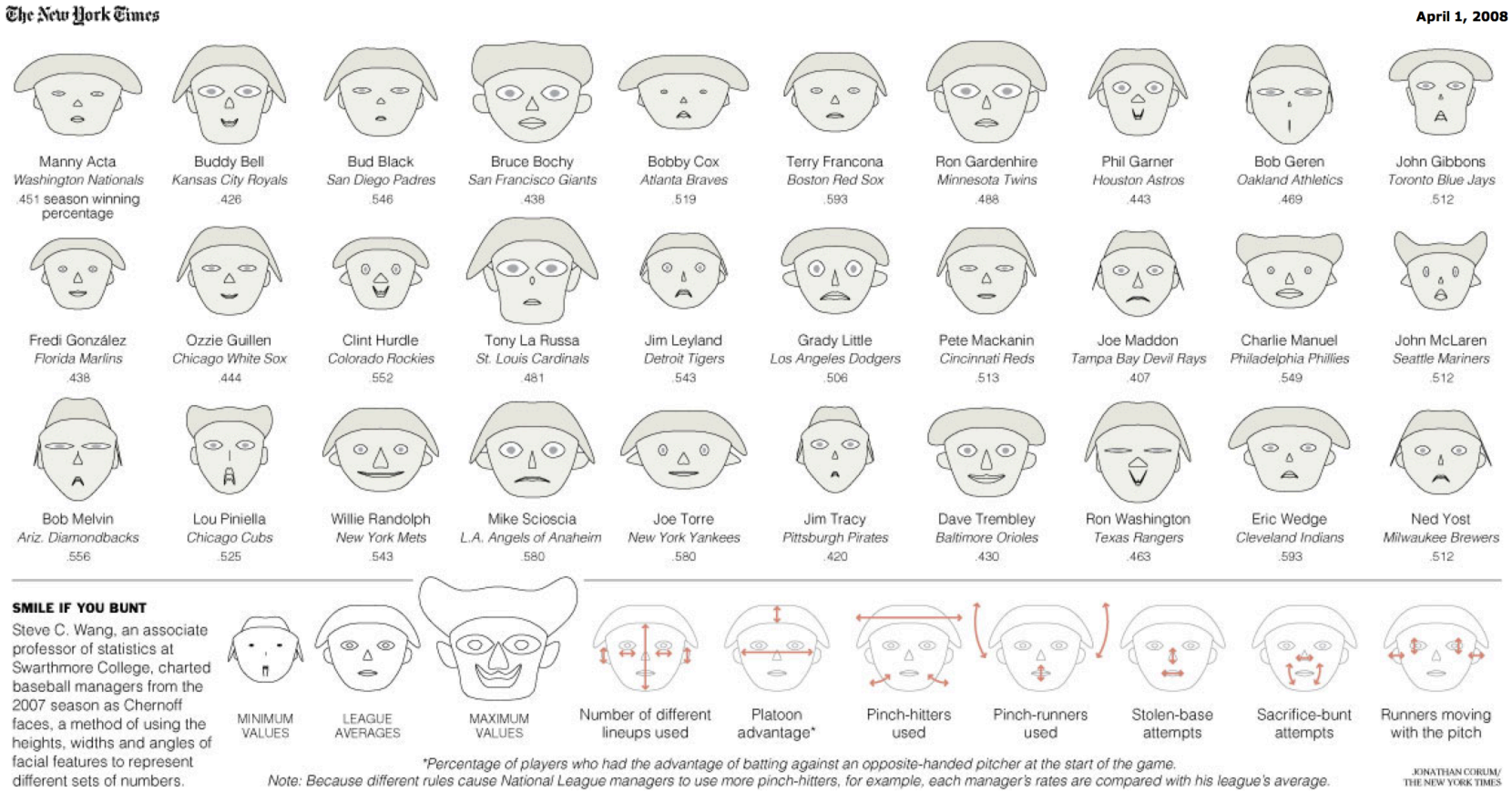


- Symbols can also be used to encode continuous data.
- Symbols have multiple dimensions that can be exploited to represent multiple variables simultaneously.
- Wind speed graph
  - Direction of the arrow tells you wind direction.
  - Length tells you wind speed.
- Using too many dimensions to encode data can lead to graphs that are hard to read.

Encoding through shape: Chernoff's faces

- Used to represent many variables simultaneously (up to 18).
- Each feature (height and width of face, size of eyes and ears, etc.) encodes a different variable.
- They are based on the idea that humans can pre-attentively differentiate faces.
- Note that we are not equally good at perceiving differences in all features.
  - Order of the variables might matter!
  - Redundancies can be important!

Chernoff's faces



http://www.nytimes.com/2008/04/01/science/01prof.html?ex=1364702400&en=c8f70d006f87dbd9&ei=5088&partner=rssnyt&emc=rss

Encoding through shape

- Remember that, when using lines to represent your data, additional information can be encoded through:
  - Categorical information: Line style (solid, dashed, dotted, dashed-dotted)
  - Either categorical or quantitative information: Line width.
- Line styles don't work as well as carefully chosen colors to encode categorical variables.
- Line widths can be effective at showing the "big picture" but do not allow for accurate comparisons (differences are not large enough).