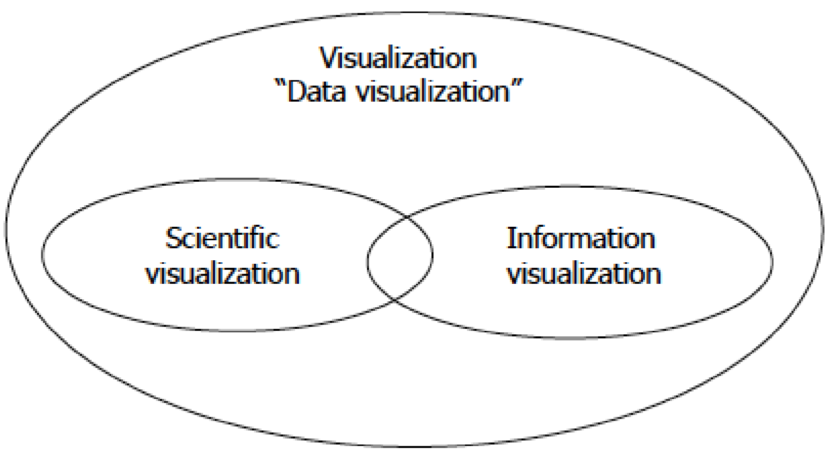
Week 4: Information Visualization

"A picture is worth a thousand words"

Complicated and huge data sets can be reduced to a few numbers using statistics. Many technical statistical terms can understand only by the trained/skills professionals. On the other hand, the graphical presentations are accessible to a wider audience [Few, 2009].

The terms data visualization or visualization refers to the visual representations that facilitate the exploration, examination and communications of data. Two subcategories of data visualization are information visualization and scientific visualization. Card, Mckinlay and Shneiderman defined information visualization (InfoVis) as "*the use of computer-supported, interactive, visual representations of data to amplify cognition*". Typically, cognition is related to the knowledge usage. Scientific visualization mainly deals with volumes, surfaces, or sources of physical or geometric related, for example, gravity waves, fluid flow, climate visualization, etc. The emphasis of this chapter is about information visualization. The history of cartography, statistical graphics, and data visualization can be read from the paper authored by Michael Friendly at http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf



Source: http://www.cc.gatech.edu/~stasko/7450/courses.html

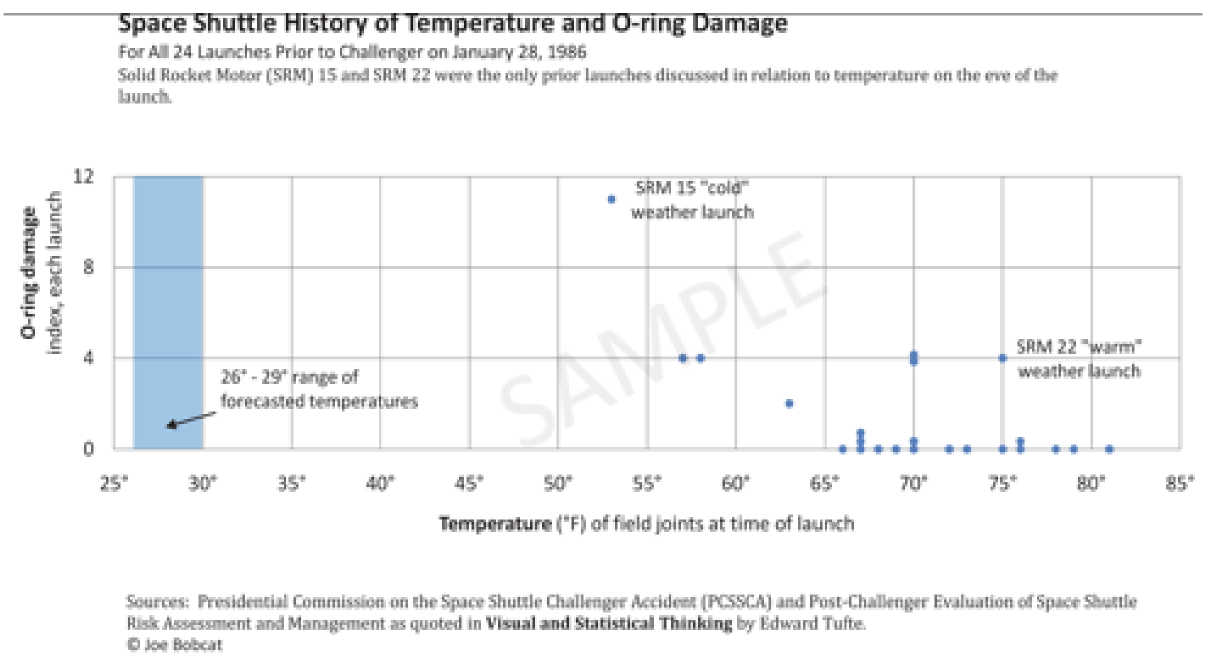
Due to Stasko, " *The purpose of visualization is insight not pictures*". Visualization is a process of making a graphic or an image. It is a cognitive process of forming a mental image to internalize an understanding. Visualization helps us think and communicate information. In addition, graphics should complement what humans do well.

Information visualization provides presentation tools to helps people understand and gain insight. The visual information is useful for analysis and decision-making. The challenge is how to transfer the information overload to human vision. Stasko characterizes human vision as the highest bandwidth sense: fast and parallel, pattern recognizing, pre-attentive, and memory-extending. People think visually.

Visualization concepts are the essential foundations for communicating data. The visualized data can generate many profound effects such as to make data easy to understand. Understanding is critical since the decision-making or action based upon that understanding. Analyzing the data is not worthwhile if it is misinterpreted or not well understood by users.

Visualization is particularly useful in exploratory data analysis when we do not know what to look for, when there are no a priori questions, and search for the right questions to ask.

Tufte [1997] illustrated that part of the failure that happened with the Space shuttle challenger incident, was due to the poor visualization. The breached O-ring in the rocket boosters is the cause of the explosion. The O-ring failed because it was not durable enough to maintain the seal when the temperatures are below freezing at the shuttle launch. The decision makers may not have needed information regarding to O-ring. Therefore, that the shuttle was launched in cold temperatures is the key mistake of this disaster. Exploiting all capabilities of the visual systems to perceive and think is very critical in the decision making process. Tufte redesigned the visualization with the same data, simplifying the display dramatically. From the graph, it appeared that a shuttle had never been launched in cold temperatures before.



Tufte suggests that all data should be presented when making a decision. Particularly, data presentation should be in a meaningful way. He also acknowledged that this incident could have been prevented if this graph had been considered. This is one of many examples of good visualization that can lead to a better decision.

The book: “The Visual Display of Quantitative Information” by Edward R. Tufte is a classical visualization book. People who are interested in visualization and analytical design are encouraged to read this book. His design principles of graphical excellence greatly impact the visualization field to today.

Tufte’s Design Principles (Graphical excellence is …):

- The well designed presentation of interesting data.
- Comprised of complex ideas that are communicated with clarity, precision and efficiency.
- The greatest number of ideas generated in the shortest time (using the least ink and smallest space).
- Telling the truth about data (aka graph integrity) such as baseline, scale and context.
- Generally, data is multivariate (not necessarily using 3D projection).
- Visual attribute data are directly proportion to data attribute value. It is known as lie factor = size of effect shown in graphic/size of effect in data.
- The design should maximize the ink ratio

  Data ink ratio = data ink/ total ink used in graphic.
- Show the data.
- Avoid chart-junk (i.e. extraneous elements that distract the message).
- Content is the key (e.g. quality, relevance and integrity).
- Using color effectively (e.g. color is used for label, measure, imitate reality, or decorate).

In summary, graphical displays should serve these purposes:

- Show the data.
- Encourage viewers to think about substance rather than graphic design, methodology, etc.
- Prevent distortion of the data.
- Effective space usage (many numbers can be presented within a small space).
- Make large data sets clear.
- Aiding the eyes to compare different pieces of data.
- Disclose data from several levels (e.g. overview to fine structure).
- Assist with purposes of description, exploration, tabulation or decoration.
- Easily incorporate with statistical and verbal descriptions of a data set.

Stephen Few is another expert in visualization. One of his well-known books is “Now you see it: simple visualization techniques for quantitative analysis” His design principles are: (http://www.perceptualedge.com/articles/Whitepapers/Visual_Communication.pdf) :

- “Display neither more or less than what is relevant to your message”
- “Do not include visual differences in a graph that do not correspond to actual differences in the data”

- “Use the lengths or 2-D locations of objects to encode quantitative values in graphs unless they have already been used for other variables”
- “Differences in the visual properties that represent values (that is, differences in their lengths or 2-D locations) should accurately correspond to the actual differences in the values they represent”
- “Do not visually connect values that are discrete, thereby suggesting a relationship that does not exist in the data”
- “Make the information that is most important to your message more visually salient in a graph than information that is less important”
- “Augment people’s short term memory by combining multiple facts into a single visual pattern that can be stored as a chunk of memory and by presenting all the information they need to compare within eye span”

According to Few, eight types of quantitative messages that users attempt to understand and communicate from a set of data are: (http://en.wikipedia.org/wiki/Data_analysis)

1. Time Series: focus on a particular variable over a period of time, for instance, the price of stock A over a six-month time period. The trend can be illustrated using a line chart.
2. Ranking: data are ranked in order (ascending or descending). For example, ranking of the top merchandise sales. A bar chart is used to compare the sales of different merchandises.
3. Part-to-Whole: data is measured as a ratio to the whole. Commonly, pie chart and bar chart are used for the ratio comparisons.
4. Deviation: data is compared with its reference such as comparison of the actual number of voters with expected number of voters. The bar chart can be used for this illustration.
5. Frequency distribution: number of observations of a specified variable is displayed for a given interval such as number of years for mutual fund performance at 0-5%, 6-10%, 11-15%, etc. Histogram is recommended for this type of display.
6. Correlation: two variables (e.g. X, Y) are determined whether they move in the same or opposite directions. Scatter plot is the typical illustration.
7. Nominal comparison: compare categorical data without any order. For instance, the interest rate of the saving account by banks. The comparison can be shown using bar chart.
8. Geographic or geospatial: a variable is compared across a layout or map. For example, high school graduated rate by US states. A cartogram (value by area or isodemographic map) is a common graphical type.

Few expressed that quantitative information in visual form representations can enhance human ability to think. Patterns, trends, and outliers in numbers become visible and understandable by visual representations. Through brain research, the connection between vision and cognition starts to be revealed and better apprehended. The rules for applying visual perception effectively are critical for success in data exploration, analysis and insight. Good decisions usually derive from good understanding. Therefore the goal is to the design of visual presentations to accomplish the tasks. Few also points out two questions that leads us to data understanding are descriptive and predictive questions. Descriptive questions strive to understand cause and effect:

- "*What is happening?*"
- "*What is causing this to happen?*"

Predictive questions lead us to identify the future with the questions:

- "*What do we want to happen?*"
- "*What actions would likely to this desired outcome?*"

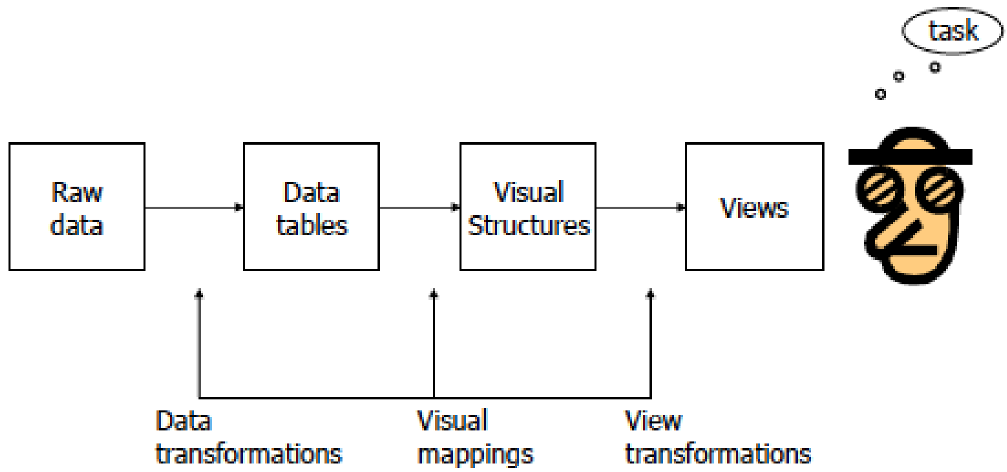
Source: Few, S. (2009)

Data represents some phenomena from the world. Raw data comes in various different forms. It is necessary to transform the data into a more workable form. Data are characterized using these 3 components: 1) objects or items of interest (e.g. patients, cars, web pages, courses) 2) Attributes or properties of data (e.g. name, SSN, address, age, symptoms) 3) relations with other objects (e.g. patients treated by doctors, doctors affiliate with insurance-networks). Each object contains many cases (patient 1, patient 2…patient N). Each case has its associated attributes (also known dimensions). Also, there are relations between cases.

Three main variables types are: N-nominal (equal or not equal to other values), O-ordinal (it is ordered set), and Q-quantitative (calculation is meaningful).

Common representations for multivariate data are tables and graphs. Tables are suitable for looking up individual values, comparing individual values, requiring precise values, and communicating more than one unit of measure. Graphs are utilized when the shape of the values contains the message and express relationships among values.

Shneiderman a professor at University of Maryland, who has done many works in infoviz, had suggested the principles of interactive visualization design as "Overview, zoom&filter, details on demand". Overview is the big picture for users to see all the data and important relationship at a glance. Zoom and filter are used to reduce the data subset that they want to study in detail. Details on demand brings up more details when needed such as hovering over a point to display a title or clicking on the point for more details.



Source: Card, Mackinley, Shneiderman 1999

According to Stasko, visualization systems can be broadly classified as:

- Infrastructure & toolkits for building systems such as InfoVis Toolkit (http://philogb.github.io/jit/), prefuse/flare, Protovis, Piccolo, D3, InfiView, Google Chart Tools, and Processing. These systems provide tools/environments for building data visualization systems.

- Systems /Tools provide a view or multiview such as Many Eyes, Improvise, MSPivot/PivotViewer, Polaris, ILOG Discovery, etc.

- Commercial systems such as:

- Table Lens/Eureka (http://www.sai.msu.su/sal/D/1/TABLE_LENS.html)
- Spotfire (http://spotfire.tibco.com/)
- InfoZoom (http://www.infozoom.com/nc/en/home-page.html)
- InfoScope (http://www.macrofocus.com/public/products/infoscope)
- QlikView (http://www.qlik.com/)
- Tableau (http://www.TableauSoftware.com)

etc.

For more visualization software: http://www.kdnuggets.com/software/visualization.html

Reference:

Card, S. Mackinlay, J. and Shneiderman, B (1999). *Readings in Information Visualization*
-Using Vision to Think. Morgan Kaufman.

Few, S. (2009) "*Now you see it: Simple Visualization Techniques for Quantitative Analysis*" Analytics Press. Oakland, CA.

J. Heer, M. Bostock, and V. Ogievetsky, "A Tour through the Visualization Zoo", Communications of the ACM, Vol. 53, No. 6, June 2010, pp. 59-67.

Friendly, M. (Aug 24, 2009). *Milestones in the history of thematic cartography, statistical, graphics, and data visualization*. Retrieved from: http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf

Miller, R. (2011) *Information Visualization – Lecture 19* (User Interface Design and Implementation) MIT OpenCourseWare. Retrieved from: http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-831-user-interface-design-and-implementation-spring-2011/lecture-notes/MIT6_831S11_lec19.pdf

Shneiderman, B (1996). *The Eyes Have It: A task by data type taxonomy for information visualizations*. Retrieved from: http://www.cs.ubc.ca/~tmm/courses/cpsc533c-05-fall/readings/shneiderman96eyes.pdf

Stasko, J. (2013) CS 7450 - Information Visualization. Georgia Technological University Retrieved from: http://www.cc.gatech.edu/~stasko/7450/syllabus.html

Tufts, E. (2001) *The Visual Display of Quantitative Information*. Graphic Press. Connecticut 06410

Vul, Ed, and Mike Frank (2009). RES.9-0002 *Statistics and Visualization for Data Analysis and Inference, January IAP 2009.* (MIT OpenCourseWare: Massachusetts Institute of Technology), Accessed: Nov 21, 2014. Retrieved from: http://ocw.mit.edu/resources/res-9-0002-statistics-and-visualization-