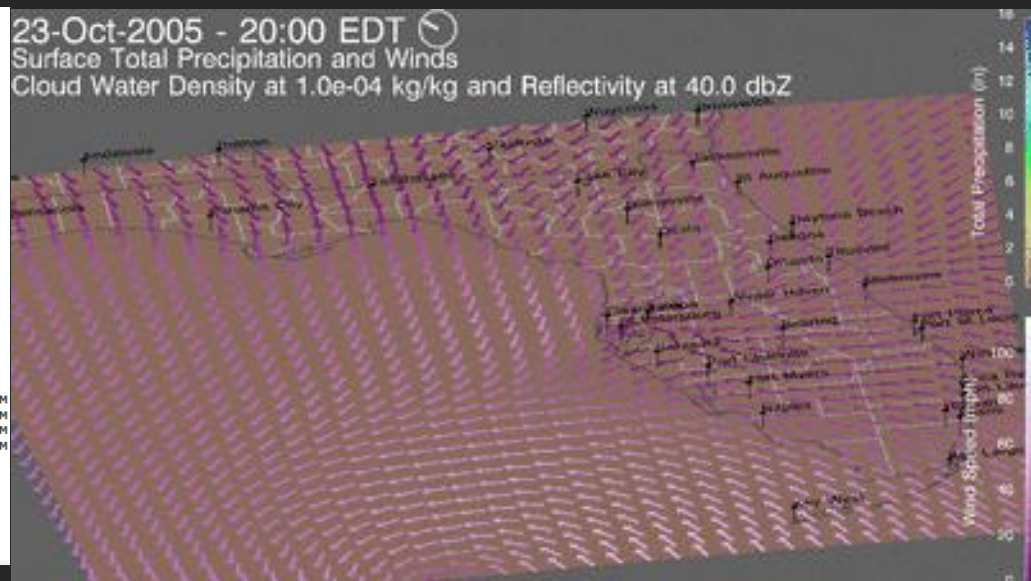


Introduction to Data Visualization

Alark Joshi

What is Data Visualization?

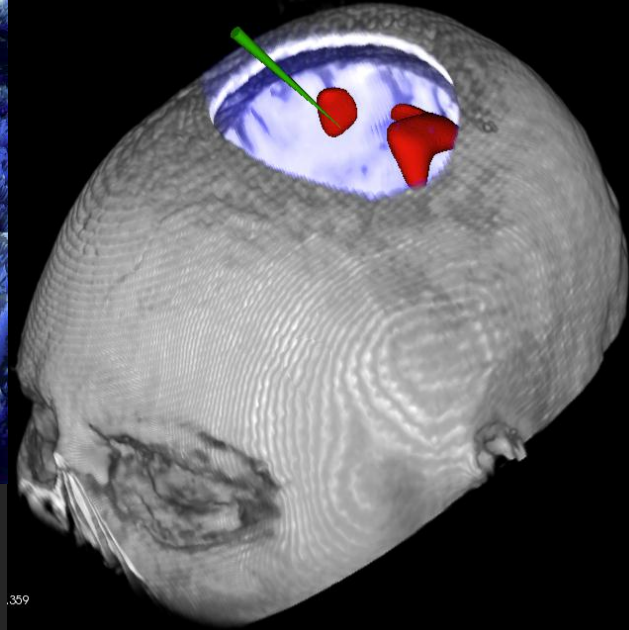
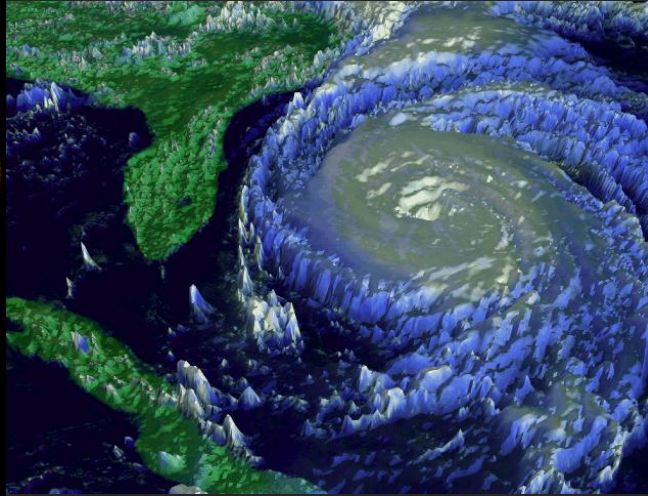
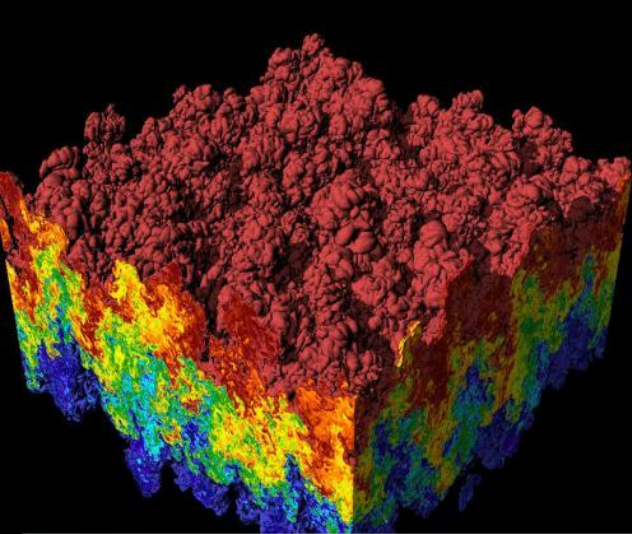
- Visual Representation of Data
- For exploration, discovery, insight, ..
- Interactive component provides more insight as compared to a static image



Types of Data Visualization

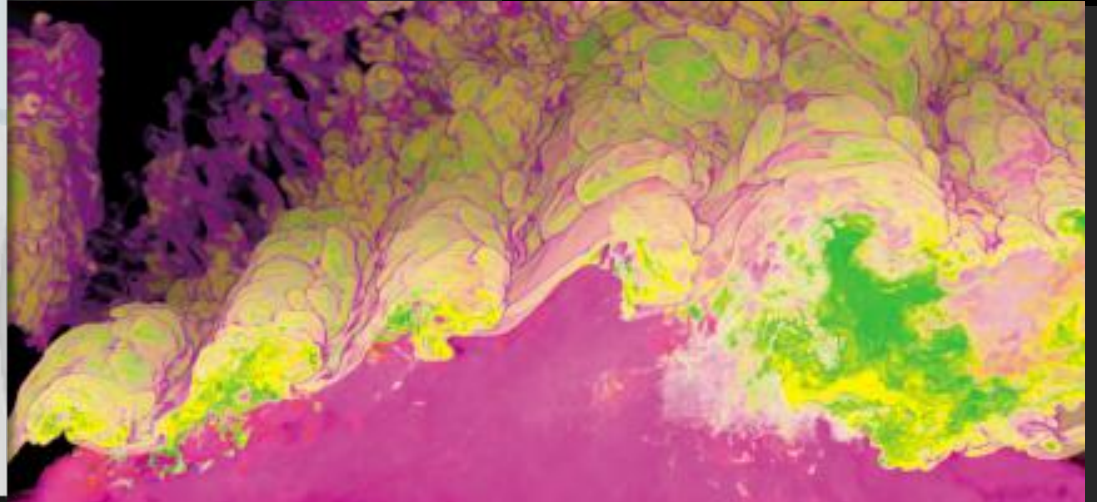
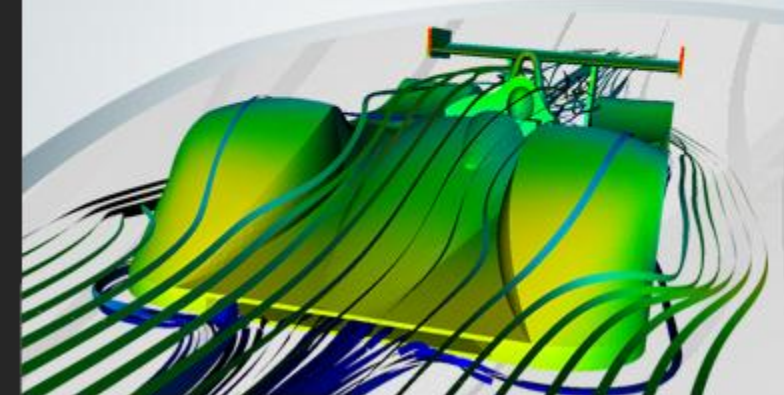
- Scientific Visualization –
 - Structural Data – Seismic, Medical, ..
- Information Visualization
 - No inherent structure – News, stock market, top grossing movies, facebook connections
- Visual Analytics
 - Use visualization to understand and synthesize large amounts of multimodal data – audio, video, text, images, networks of people ..

Scientific Visualization



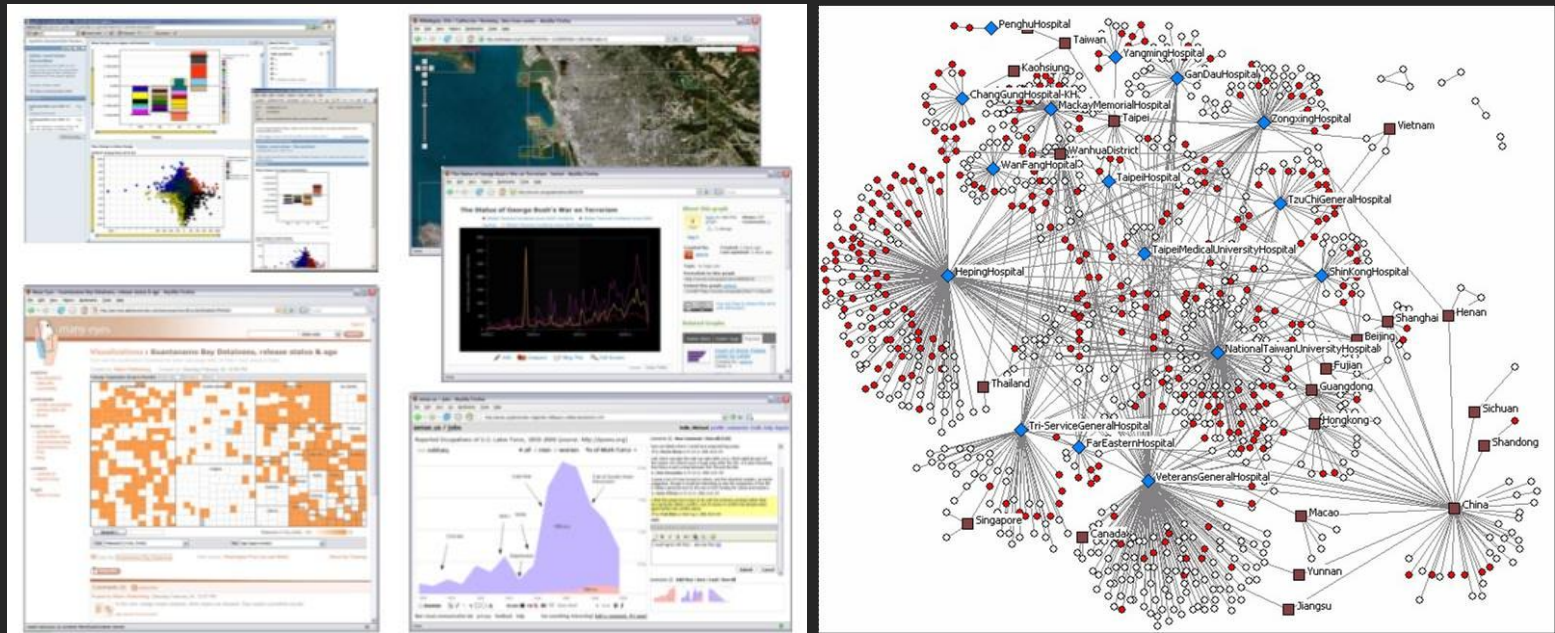
ParaView

Visualize data sets of size, from small to very large on desktop computers or high-performance clusters, using this open-source, multi-platform application.



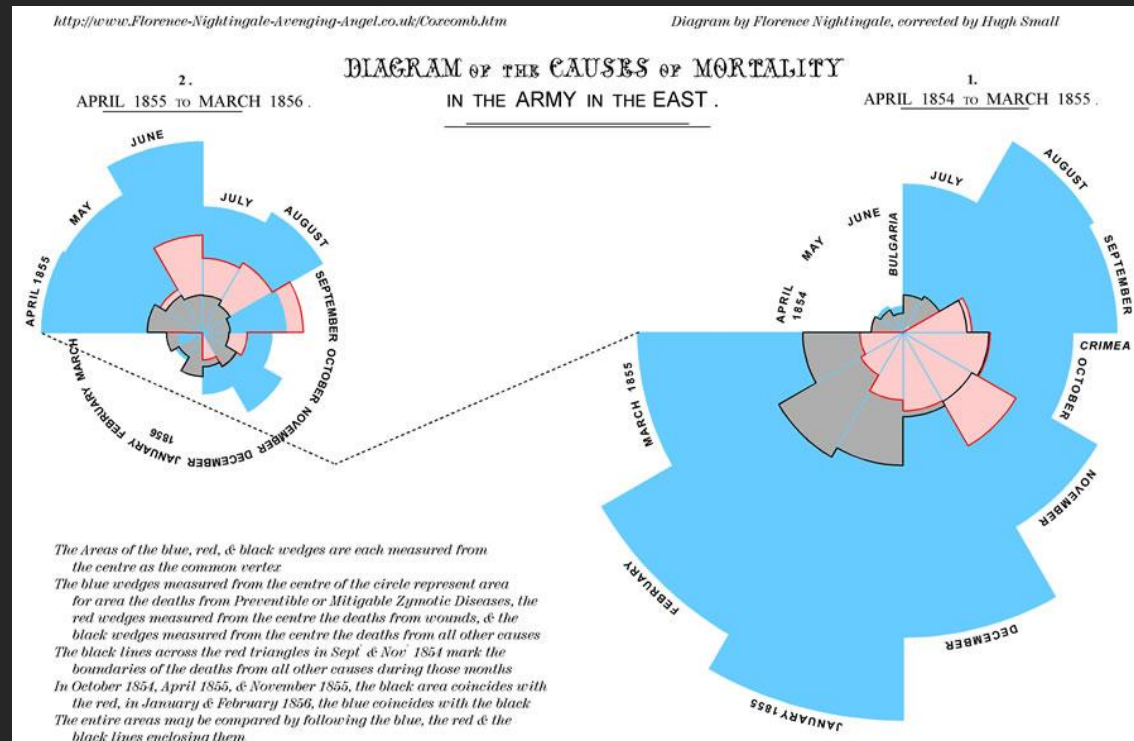
Visual Analytics

- Integration of interactive visualization with analysis techniques to answer a growing range of questions in science, business, and analysis.
- Making sense of multimodal data -audio clips, video, photographs, transcripts, ...



Impact of Visualization

- Huge impact on policy, planning and disaster avoidance.
- Florence Nightingale's visualization of casualties during the Crimean War



Impact of Visualization

- Hurricane Visualization for the common man

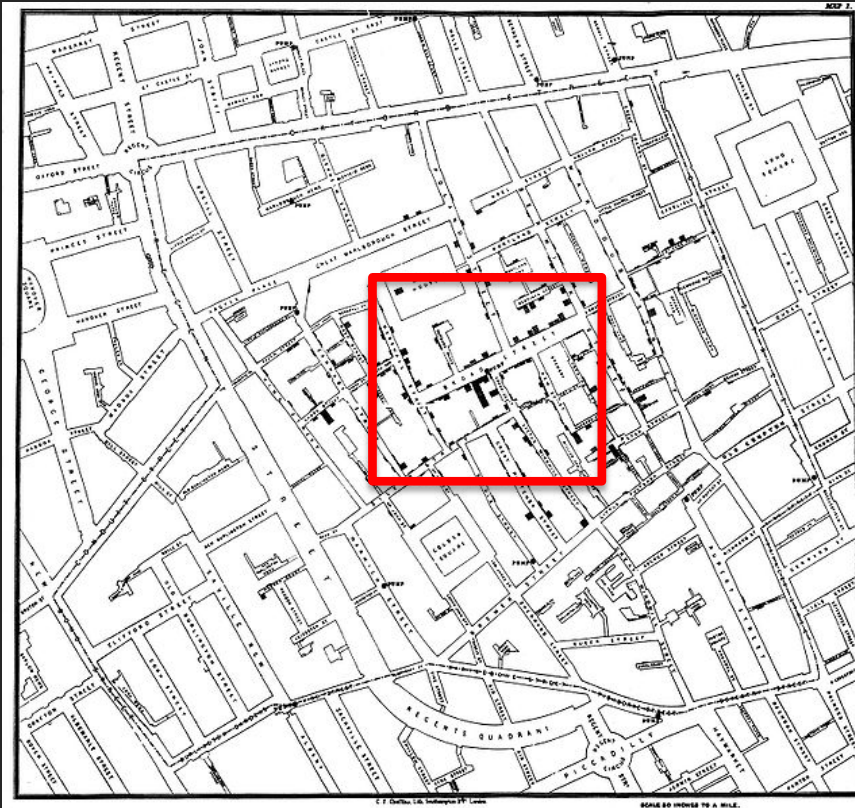


Demo:

<http://www.msnbc.msn.com/id/26295161?preferredName=Gustav>

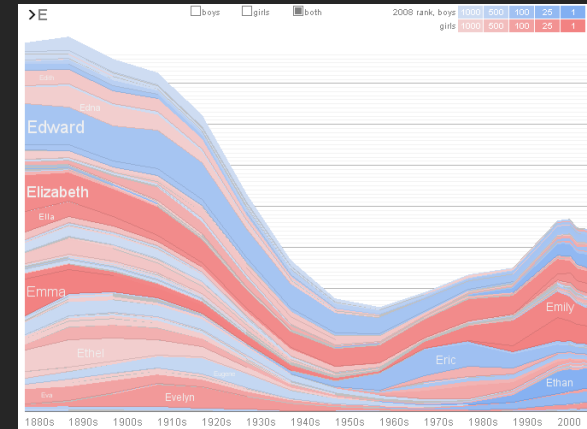
Impact of Visualization

- John Snow's Cholera Map
- Snow used a spot map to illustrate how cases of cholera clustered around the pump



Why are we here?

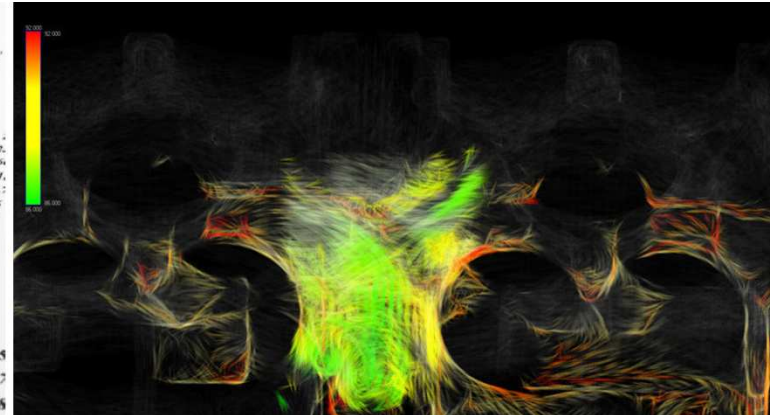
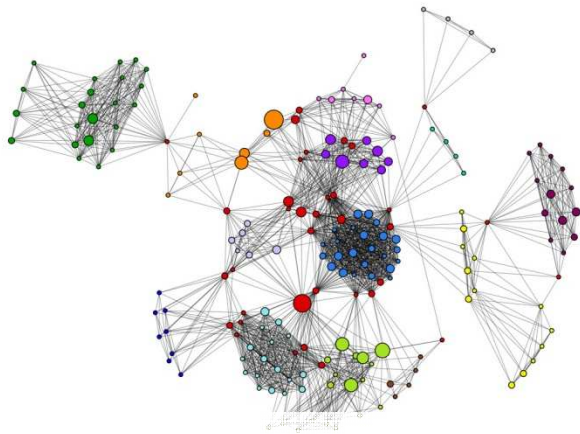
- Baby Name Wizard
 - <http://www.babynamewizard.com/voyager>
- Origin of Species – Edits
 - <http://benfry.com/traces/>
- Netflix Queues
 - <http://www.nytimes.com/interactive/2010/01/10/nyregion/20100110-netflix-map.html?ref=nyregion>
- Unemployment Visualization (NYTimes)
 - <http://www.nytimes.com/interactive/2009/11/06/business/economy/unemployment-lines.html>



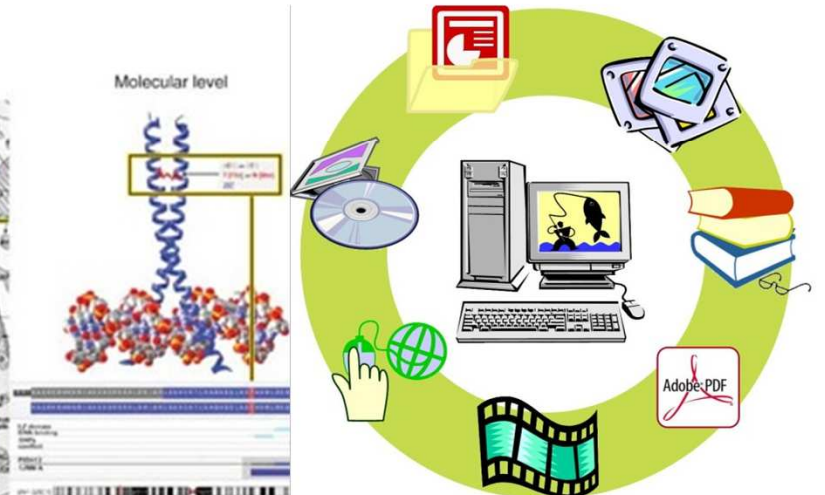
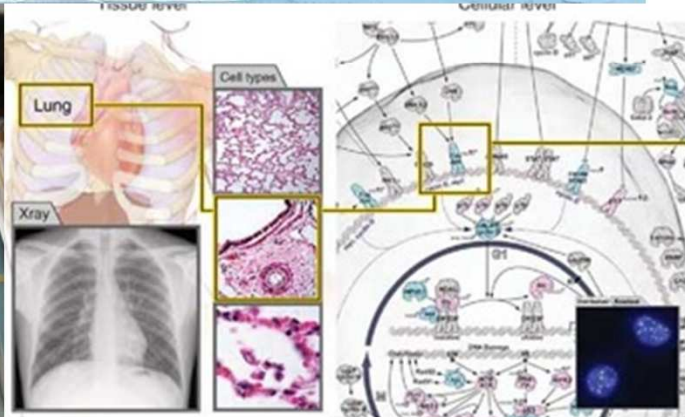
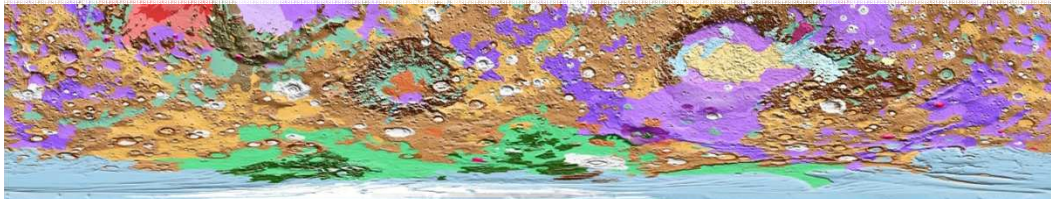
A Short Introduction on **Data** **Visualization**

Guoning Chen

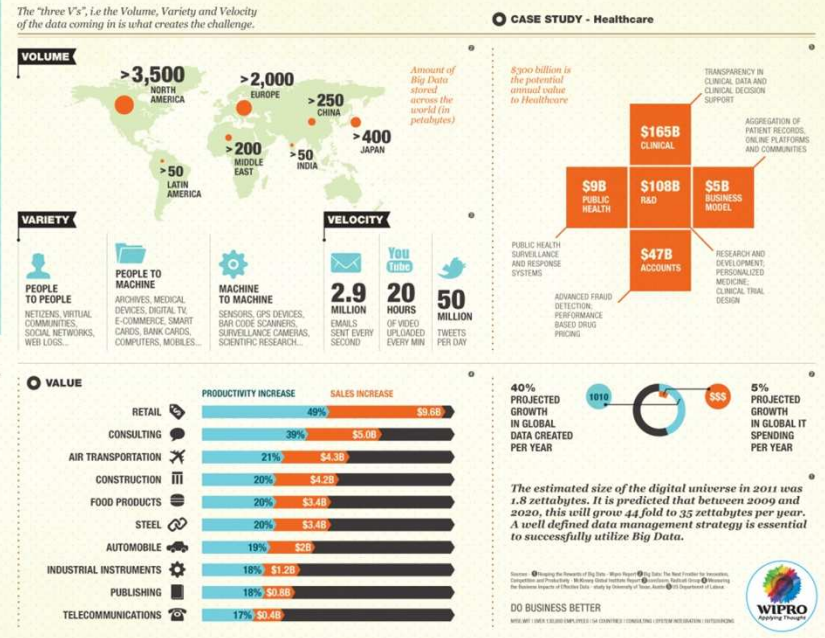
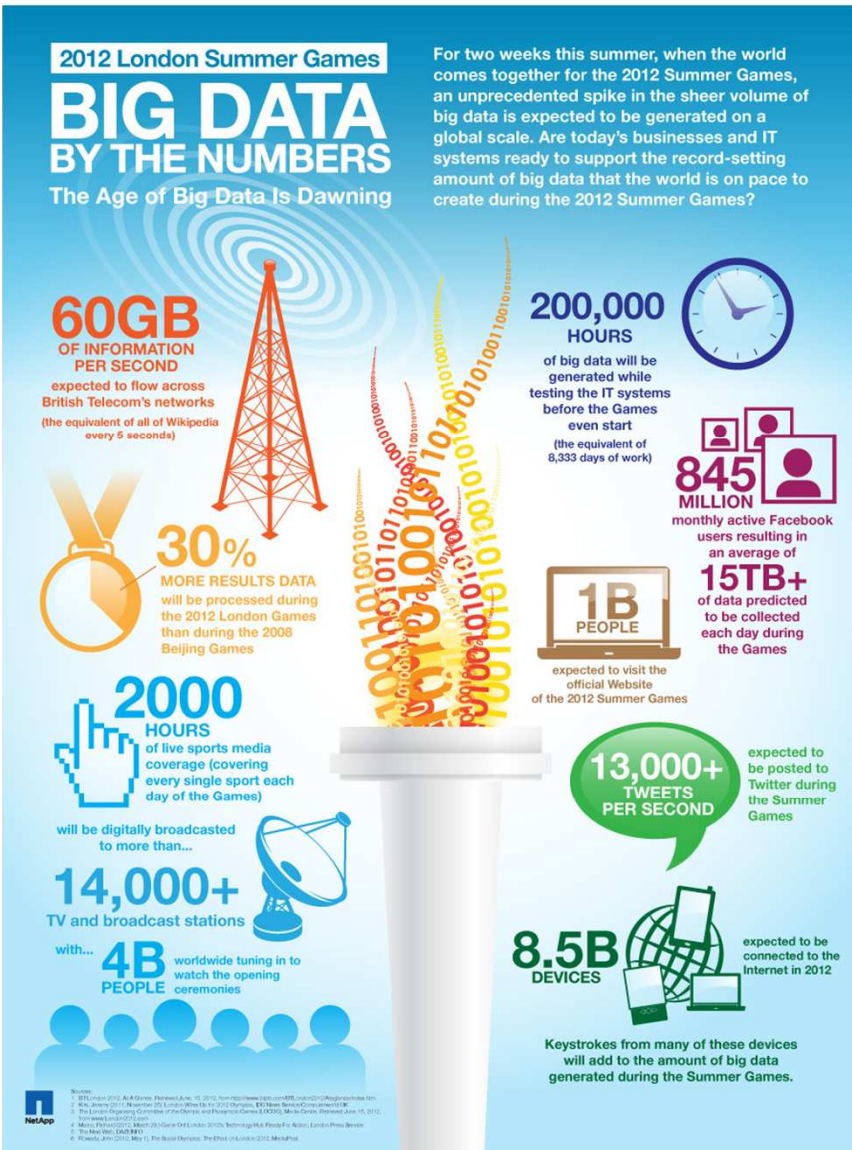




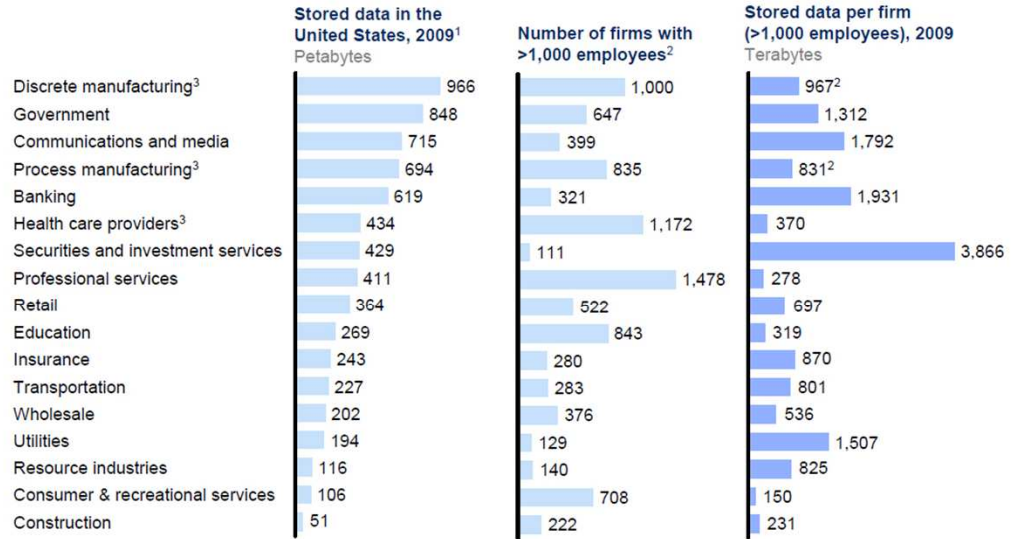
Data is generated everywhere and everyday



Age of Big Data



Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



1 Storage data by sector derived from IDC.
 2 Firm data split into sectors, when needed, using employment
 3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

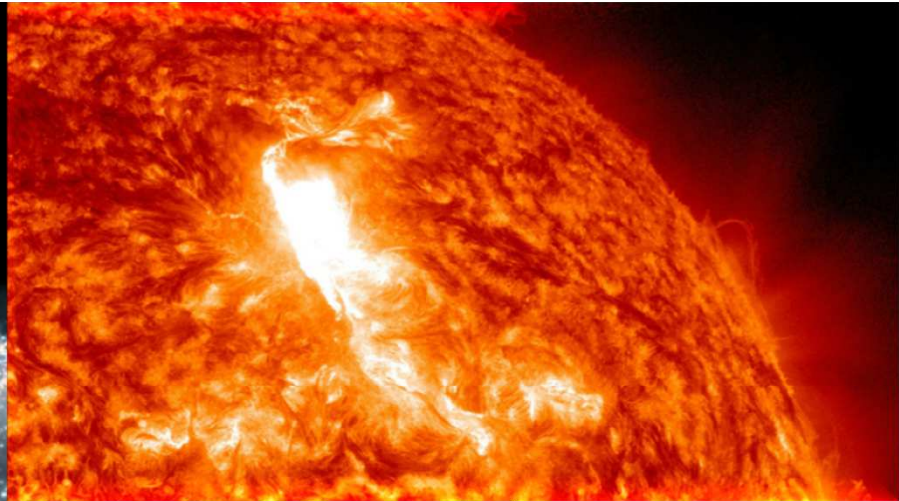
SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

What is Visualization?

- In 1987
 - the National Science Foundation (of the U.S.) started “Visualization in scientific computing” as a new discipline, and a panel of the ACM coined the term “scientific visualization”
 - Scientific visualization, briefly defined: The use of computer graphics for the analysis and presentation of computed or measured scientific data.
- Oxford Engl. Dict., 1989
 - to form a mental vision, image, or picture of (something not visible or present to the sight, or of an abstraction); to make visible to the mind or imagination
- Visualization transforms data into images that effectively and accurately represent information about the data.
 - Schroeder et al. The Visualization Toolkit, 2nd ed. 1998

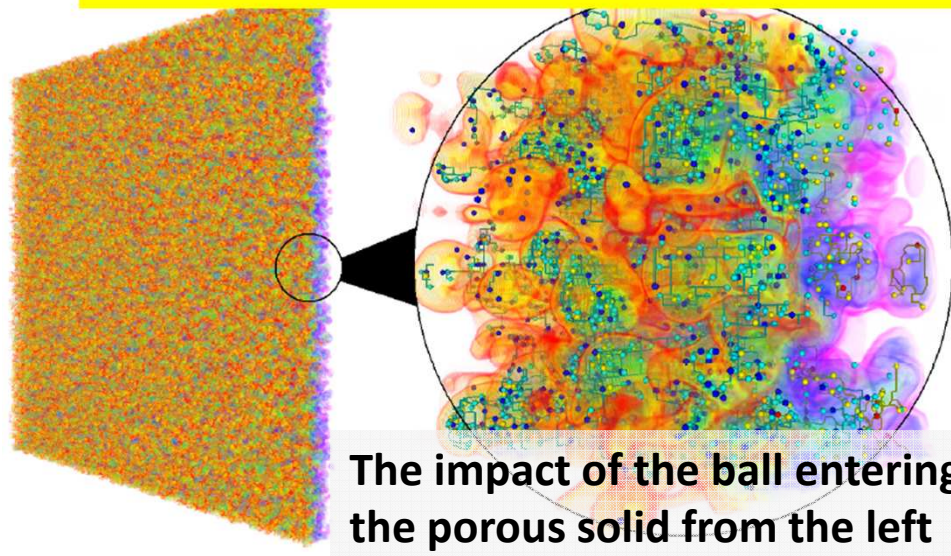
Tool to enable a User *insight* into Data

Large scale systems and events



Source: NASA

Turning invisible into visible that people can understand intuitively

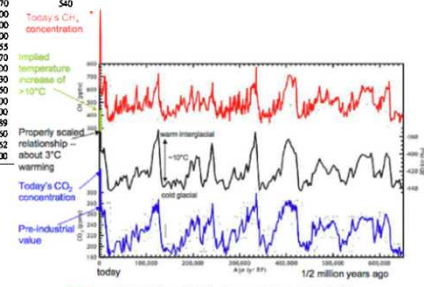


The impact of the ball entering the porous solid from the left

Table 7-8 Direct global warming potentials of several well-mixed trace gases relative to CO₂. The GWPs of the various non-CO₂ species are calculated for each of five time horizons (20, 50, 100, 200 and 500 years) using, as in IPCC, the carbon cycle model of Siegenthaler (1983). (Note that IPCC contained a typographical error which led to incorrect values for the direct GWP of methane.)

Gas	Time Horizons				
	20 years	50 years	100 years	200 years	500 years
CO ₂	1	1	1	1	1
CH ₄	10.5	35	19	11	7
N ₂ O	132	260	270	240	170
HFC-11	35	4500	4100	3400	2400
HFC-12	116	7100	7400	7100	4100
HFC-22	15.8	4200	2400	1600	970
HFC-113	110	4600	4700	4500	3900
HFC-114	230	6100	6700	7000	7000
HFC-115	590	5500	6300	7000	7800
HFC-123	1.71	330	150	90	55
HFC-124	6.9	1500	750	440	270
HFC-125	40.5	5300	4500	3400	2200
HFC-134	15.6	3100	1900	1200	730
HFC-141b	10.8	1800	950	580	350
HFC-142b	22.4	4000	2800	1800	1100
HFC-143a	64.2	4700	4500	3800	2800
HFC-152a	1.8	330	250	150	89
CCL ₄	47	1800	1600	1300	860
CH ₃ Cl	6.1	340	170	100	100
CF ₃ Br	77	5400	5300	5200	3800

SAOD Table 7.2 (p. 6)



Methane, temperature (from hydrogen isotope ratios (δD)) and carbon dioxide from the Dome C ice core. (EPICA Project members, 2006).

What Does Visualization Do?

- Three types of goals for visualization

- ... to **explore**

- Nothing is known,
- Vis. used for data exploration

- ... to **analyze**

- There are hypotheses,
- Vis. used for Verification or Falsification

- ... to **present**

- “everything” known about the data,
- Vis. used for Communication of Results

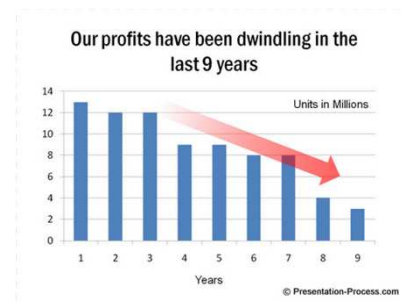
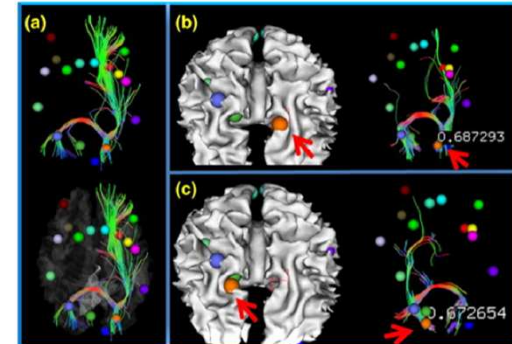
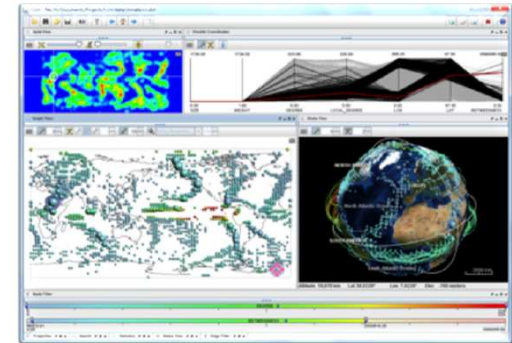
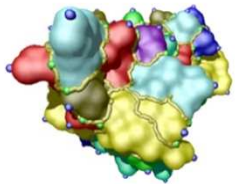
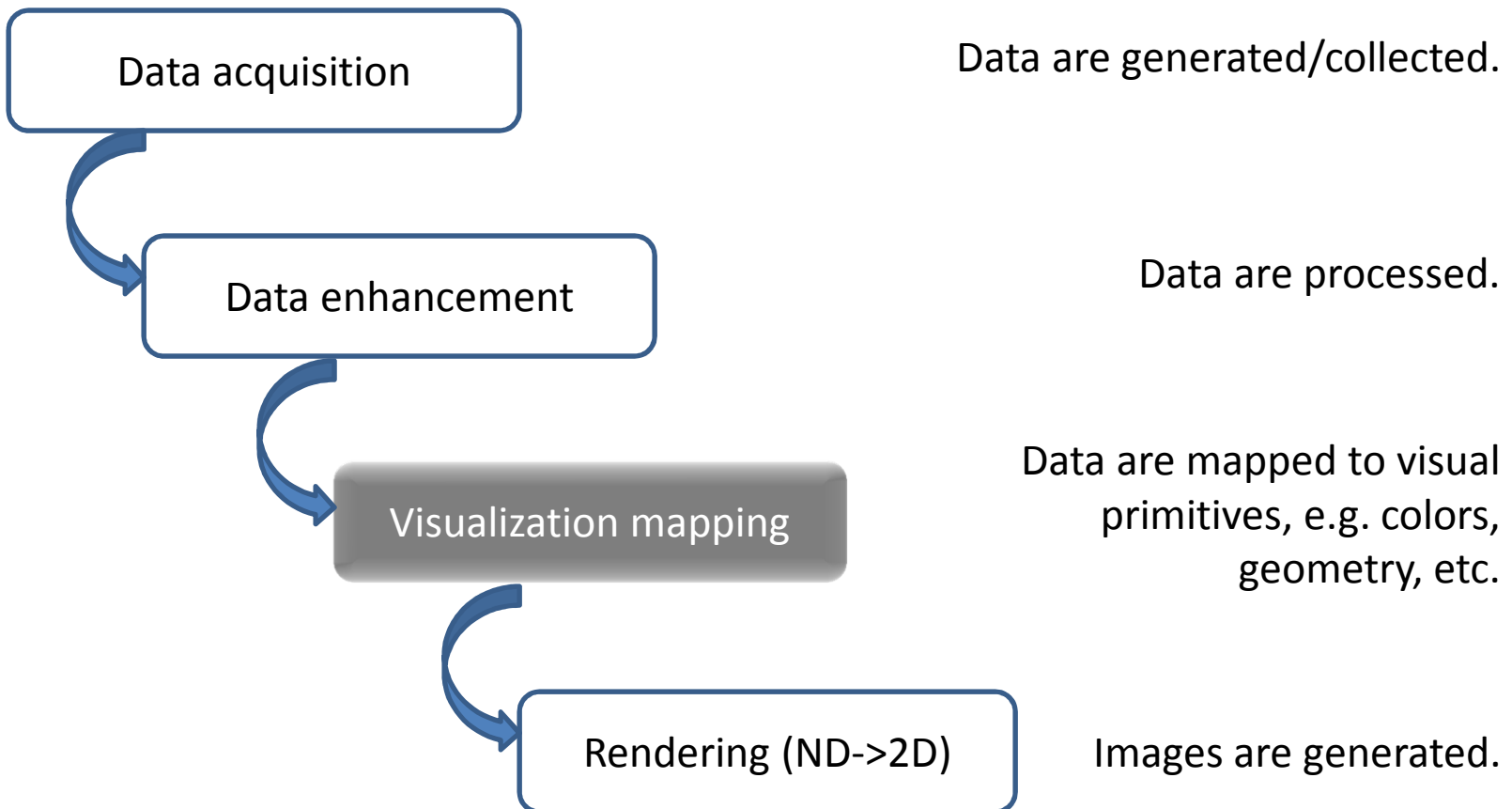


Image source: Google images

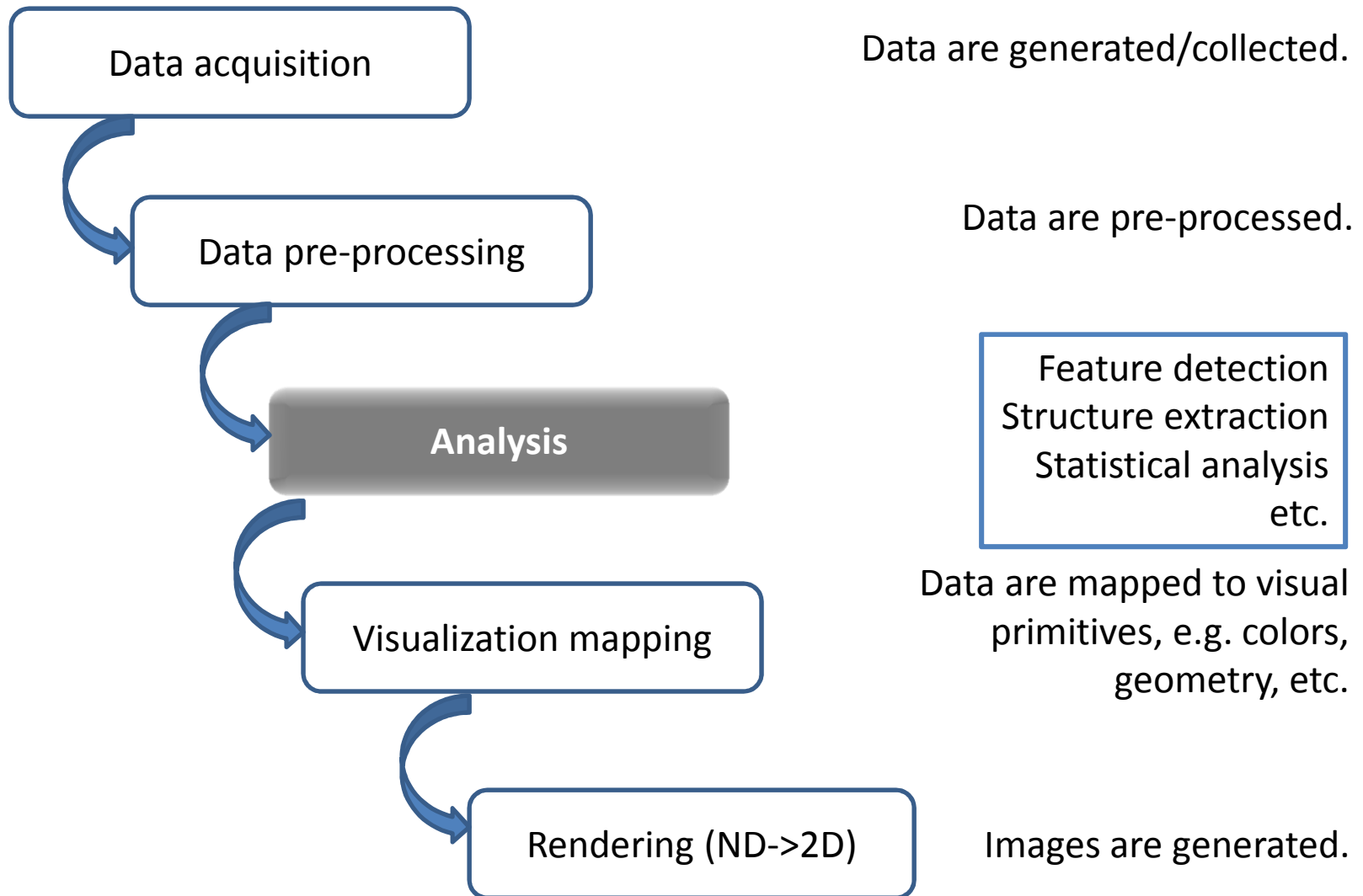
This is a well rich and inter-disciplinary area that combines knowledge from various disciplines

A Visualization Pipeline



This pipeline represents only the lecturer's opinion and need not reflect the opinions of NSF or UH!

Data Visual Analytic Pipeline

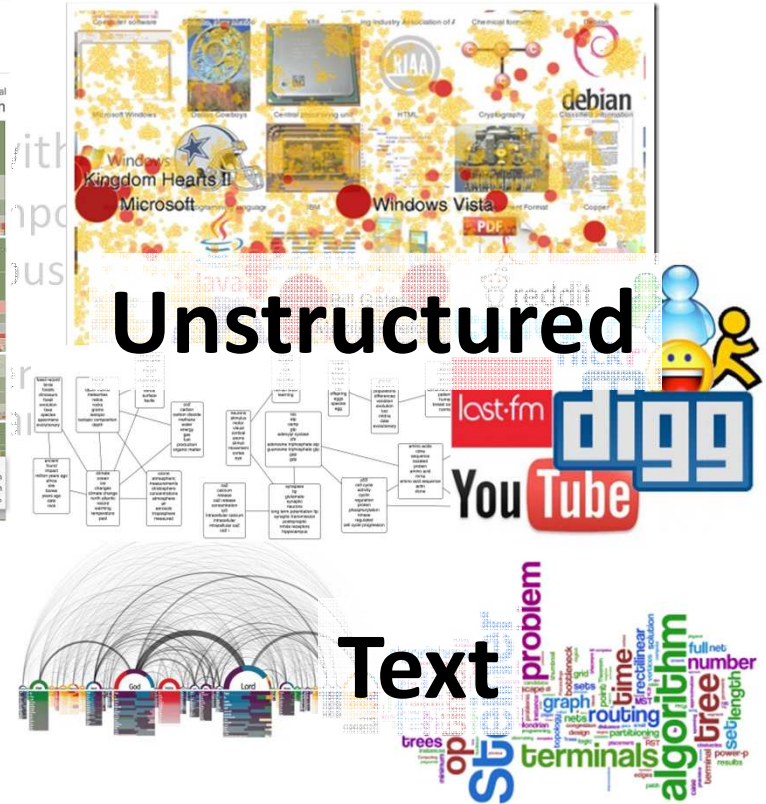
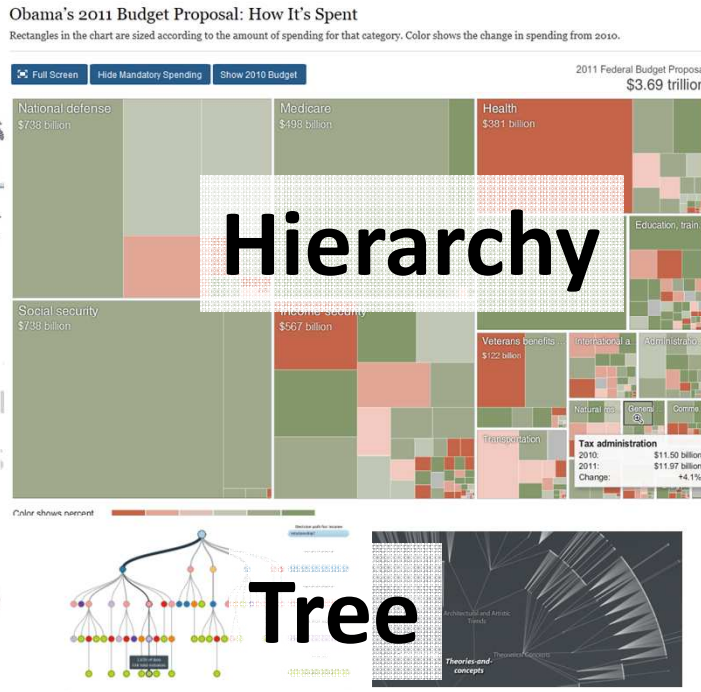
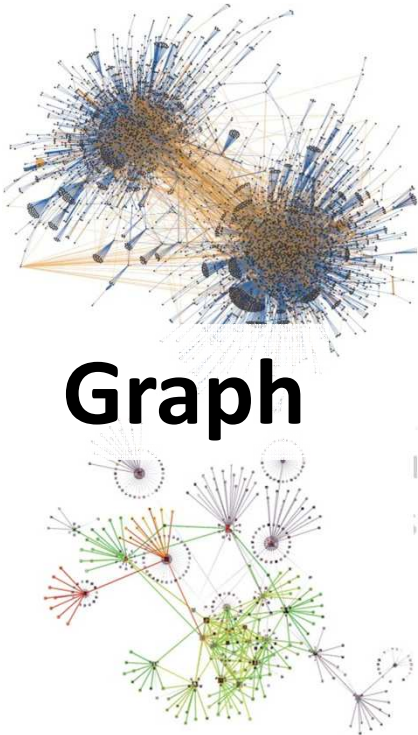


This pipeline represents only the lecturer's opinion and need not reflect the opinions of NSF or UH!

Evolution of Visualization Research

- From direct visualization to derived information visualization.
- From simple data to more complex ones.
- From represent the data with fidelity to reveal new findings.
- From scientific visualization to information visualization, bio-visualization, geographical data visualization, and beyond.

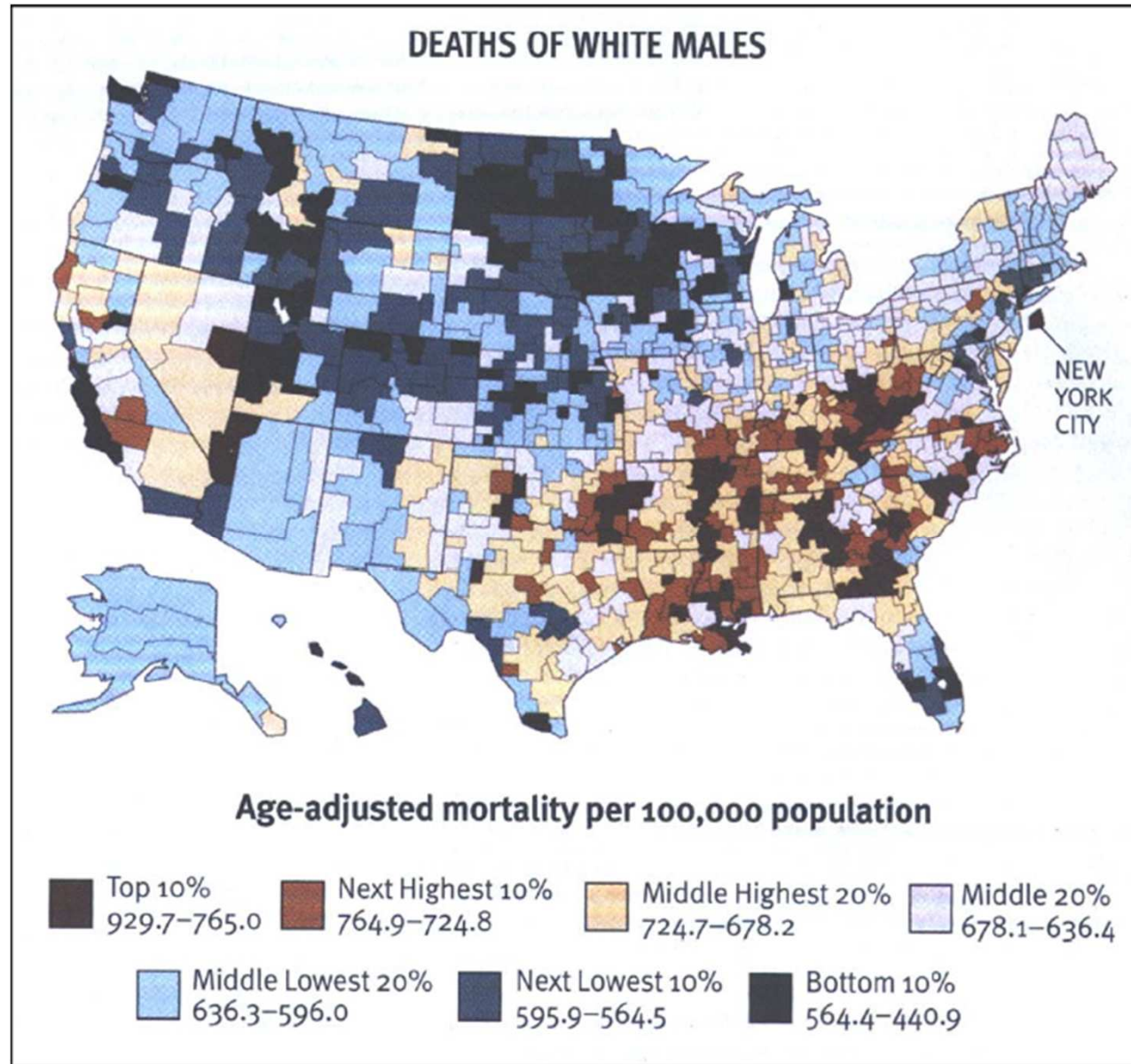
SciVis vs. InfoVis



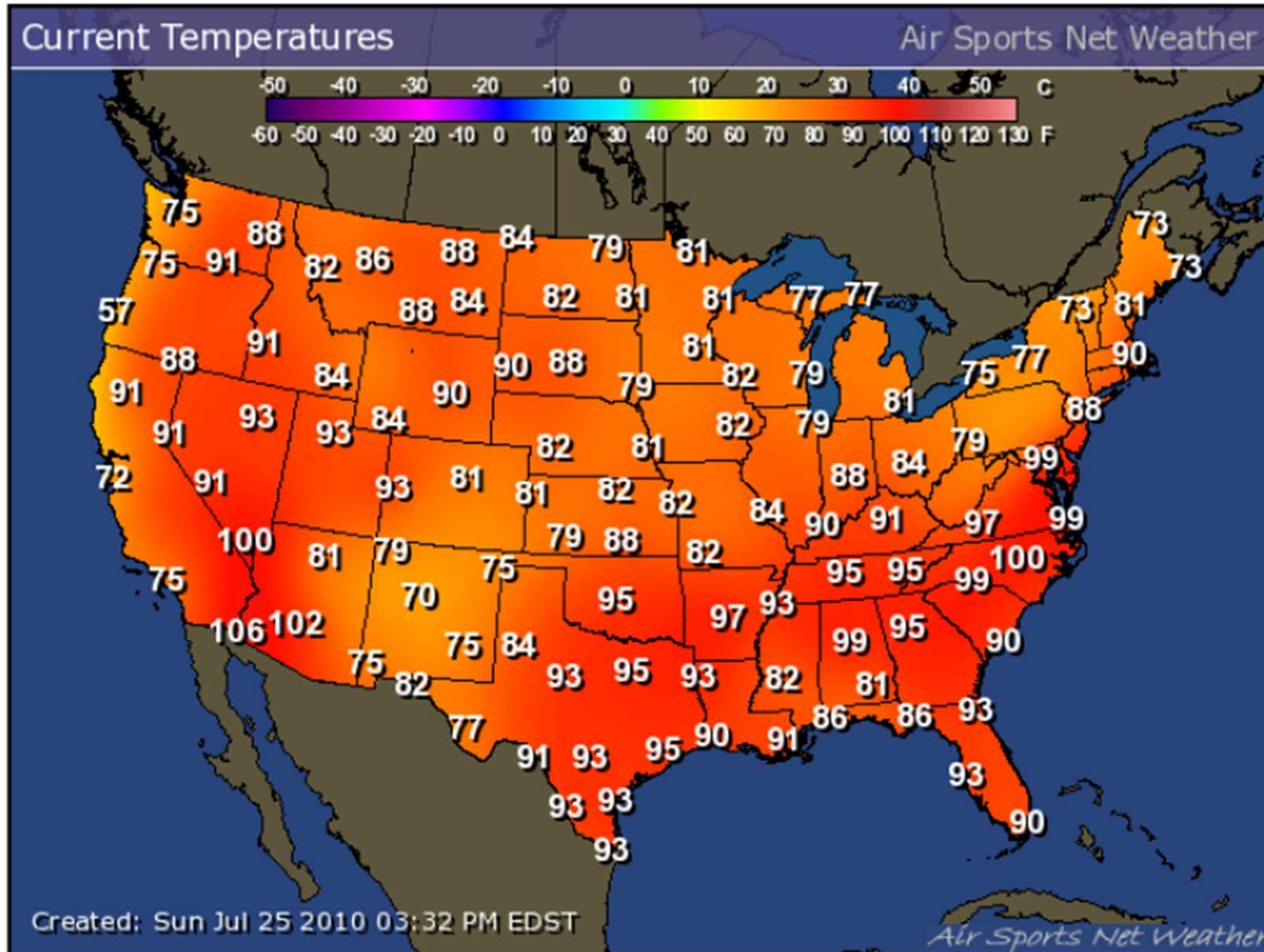
- **Information visualization** focuses on:
 - high-dimensional ($\gg 4$), abstract data (i.e. tree, graphs, hierarchy, ...)
 - Data is discrete in the nature
 - Examples include financial, marketing, HR, statistical, social media, political,
 - Feature are not well-defined, the typical analysis tasks including finding patterns, clusters, voids, outliers

Use Colors Wisely

What is Wrong with this Color Scale

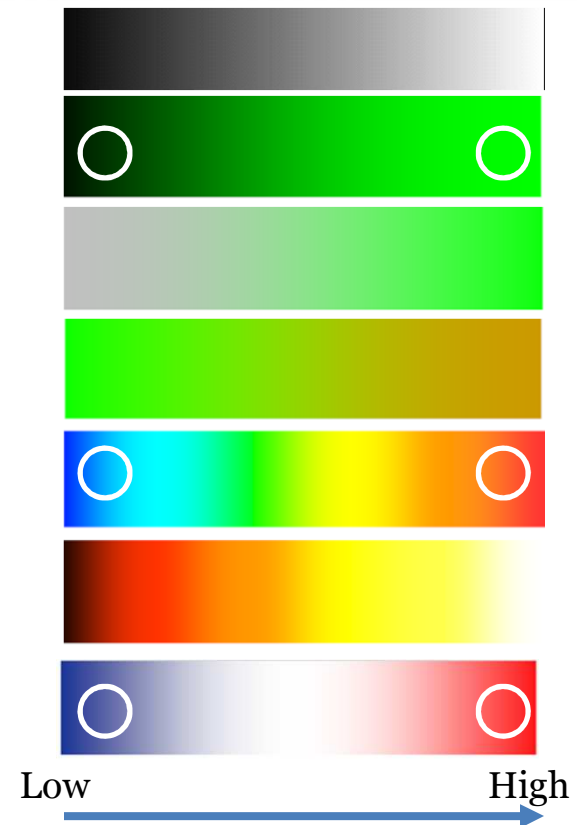


**Not a bad choice of color scale,
but the Dynamic Range needs some work**



Use the Right Transfer Function Color Scale to Represent a Range of Scalar Values

- Gray scale
- Intensity Interpolation
- Saturation interpolation
- Two-color interpolation
- Rainbow scale
- Heated object interpolation
- Blue-White-Red



Given any 2 colors, make it *intuitively obvious* which represents “higher” and which represents “lower”

**Do Not Attempt to Fight Pre-Established
Color Meanings**

COLOR MEANINGS

Examples of Pre-Established Color Meanings

Red

Stop
Off
Dangerous
Hot
High stress
Oxygen
Shallow
Money loss

Green

On
Plants
Carbon
Moving
Money

Blue

Cool
Safe
Deep
Nitrogen

Use good contrast as human eye is good
at **difference**

at **difference**

Color Alone Doesn't Cut It

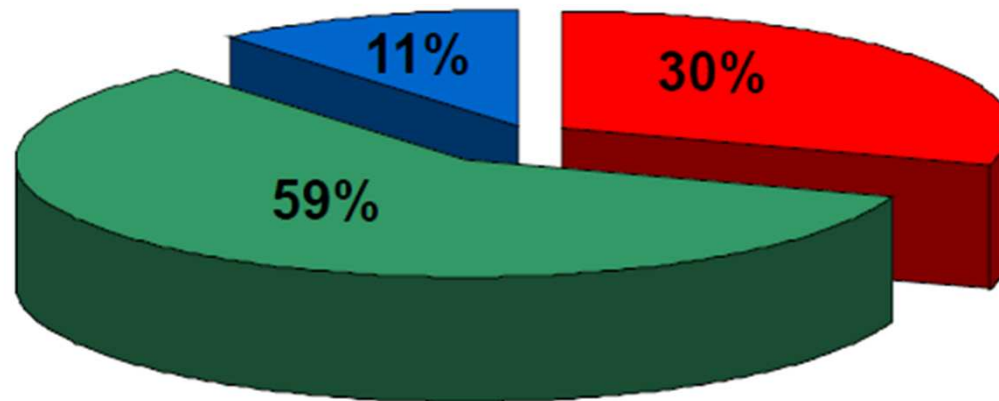
**I sure hope that my
life does not depend
on being able to read
this quickly and
accurately!**

Luminance Contrast is Crucial

**I would prefer that
my life depend on
being able to read *this*
quickly and
accurately!**

The Luminance Equation

$$Y = 0.3 \times \textit{Red} + 0.59 \times \textit{Green} + 0.11 \times \textit{Blue}$$



≈ Contrast Table

	Black	White	Red	Green	Blue	Cyan	Magenta	Orange	Yellow
Black	0.00	1.00	0.30	0.59	0.11	0.70	0.41	0.60	0.89
White	1.00	0.00	0.70	0.41	0.89	0.30	0.59	0.41	0.11
Red	0.30	0.70	0.00	0.29	0.19	0.40	0.11	0.30	0.59
Green	0.59	0.41	0.29	0.00	0.48	0.11	0.18	0.01	0.30
Blue	0.11	0.89	0.19	0.48	0.00	0.59	0.30	0.49	0.78
Cyan	0.70	0.30	0.40	0.11	0.59	0.00	0.29	0.11	0.19
Magenta	0.41	0.59	0.11	0.18	0.30	0.29	0.00	0.19	0.48
Orange	0.60	0.41	0.30	0.01	0.49	0.11	0.19	0.00	0.30
Yellow	0.89	0.11	0.59	0.30	0.78	0.19	0.48	0.30	0.00

ΔL^* of about 0.40 are highlighted and recommended

Use good contrast

	Black	Black	Black	Black	Black	Black	Black	Black
White		White	White	White	White	White	White	White
Red	Red		Red	Red	Red	Red	Red	Red
Yellow	Yellow	Yellow		Yellow	Yellow	Yellow	Yellow	Yellow
Green	Green	Green	Green		Green	Green	Green	Green
Blue	Blue	Blue	Blue	Blue		Blue	Blue	Blue

ΔL^* of about 0.40 makes good contrast

Other Rules...

- Limit the total number of colors if viewers are to discern information quickly.
- Be aware that our perception of color changes with: 1) surrounding color; 2) how close two objects are; 3) how long you have been staring at the color; 4) sudden changes in the color intensity.
- Beware of Mach Banding.
- Be Aware of Color Vision Deficiencies (CVD)

It is not possible to list all the useful rules. They come with a lot of experience!

Beware of Color Pollution

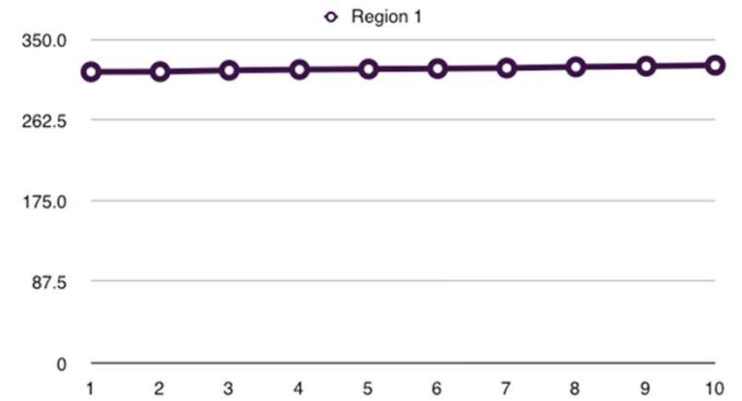
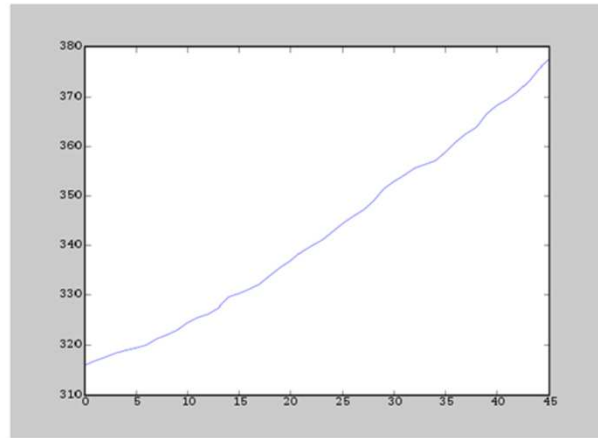
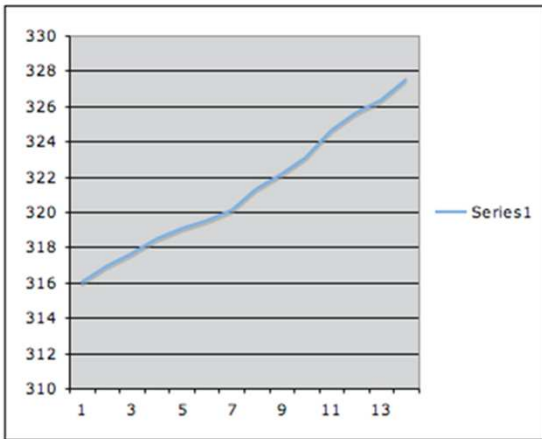
Just because you have millions of colors to choose from

doesn't mean you must use them all ...

Some Principles for Plots

Visualizing Data [Cleveland 93] and *Elements of Graphing Data*
[Cleveland 94] by William S. Cleveland

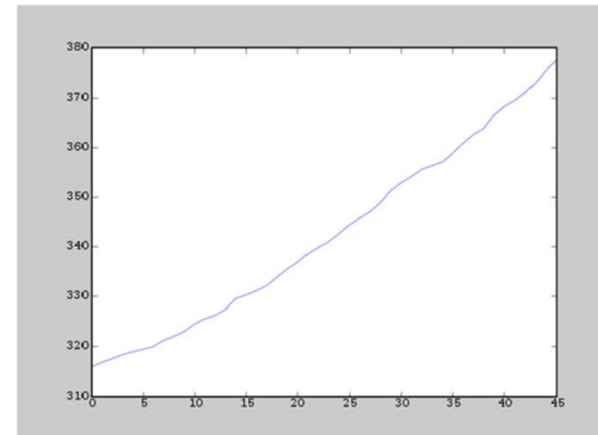
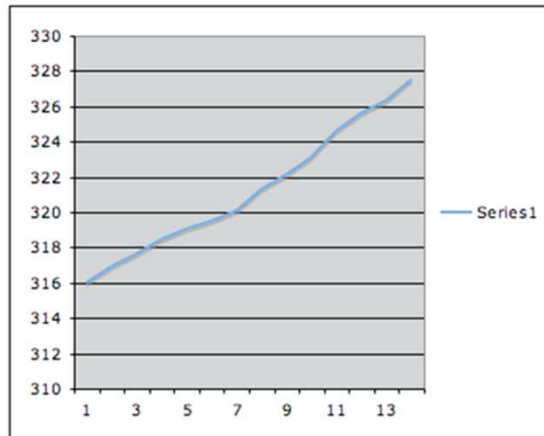
The information provided here should be considered as guidelines



- Why are they all different?
- What is good/bad about each?

Improving the Vision

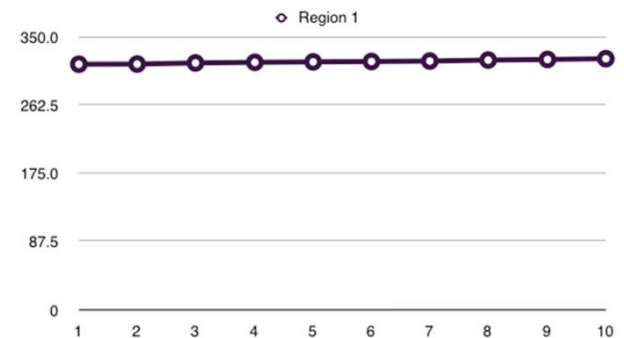
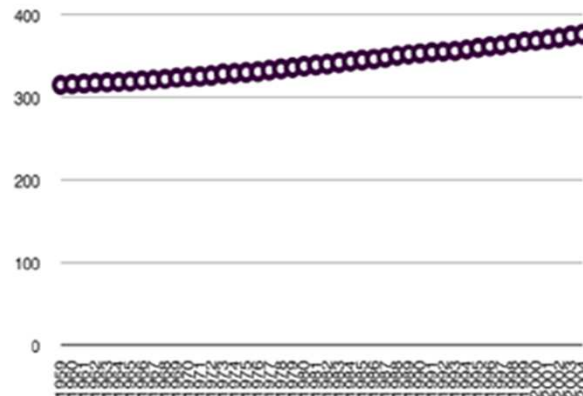
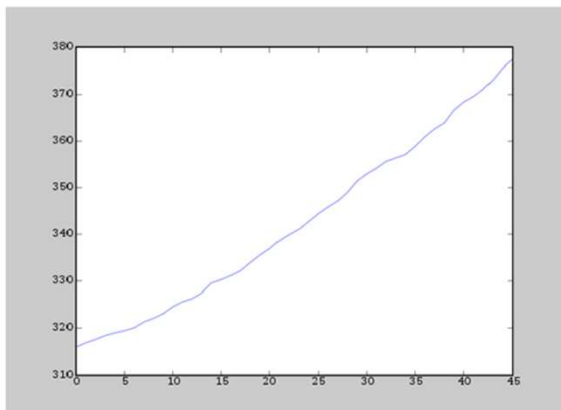
- Principle 1: Reduced clutter, Make data stand out
 - The main focus of a plot should be on the data itself, any superfluous elements of the plot that might obscure or distract the observer from the data needs to be removed.



Which one is better?

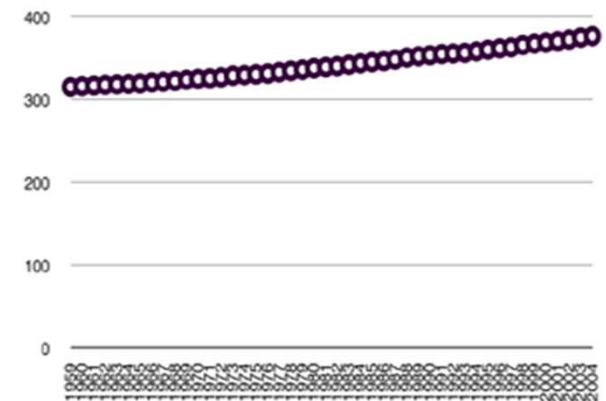
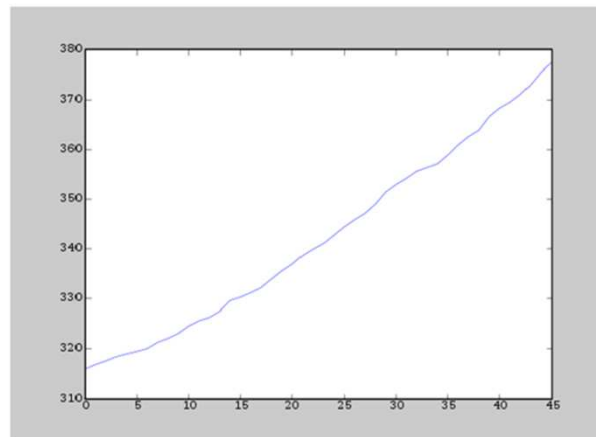
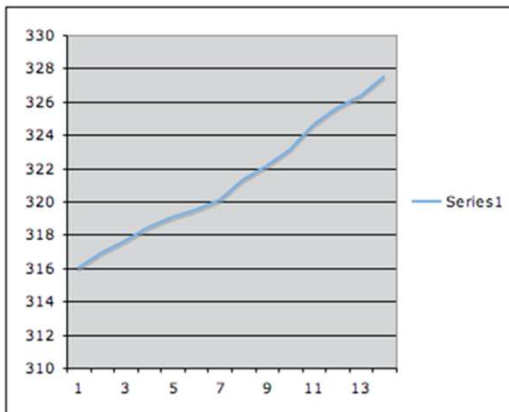
Improving the Vision

- Principle 2: Use visually prominent graphical elements to show the data.
 - Connecting lines should never obscure points and points should not obscure each other.
 - If multiple samples overlap, a representation should be chosen for the elements that emphasizes the overlap.
 - If multiple data sets are represented in the same plot (superposed data), they must be visually separable.
 - If this is not possible due to the data itself, the data can be separated into adjacent plots that share an axis



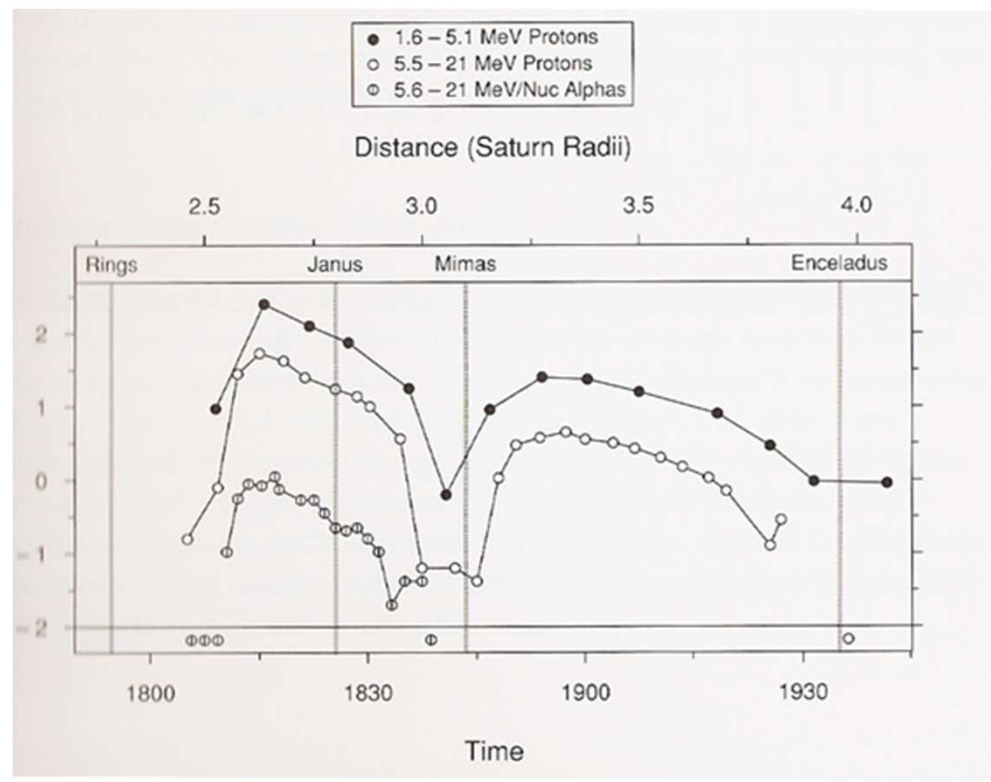
Improving the Vision

- Principle 3: Use proper scale lines and a data rectangle.
 - Two scale lines should be used on each axis (left and right, top and bottom) to frame the data rectangle completely.
 - Add margins for data
 - Tick-marks out and 3-10 for each axis



Improving the Vision

- Principle 4: Reference lines, labels, notes, and keys.
 - Only use them when necessary and don't let them obscure data.



Improving the Vision

- Principle 4: Reference lines, labels, notes, and keys.
 - Only use them when necessary and don't let them obscure data.

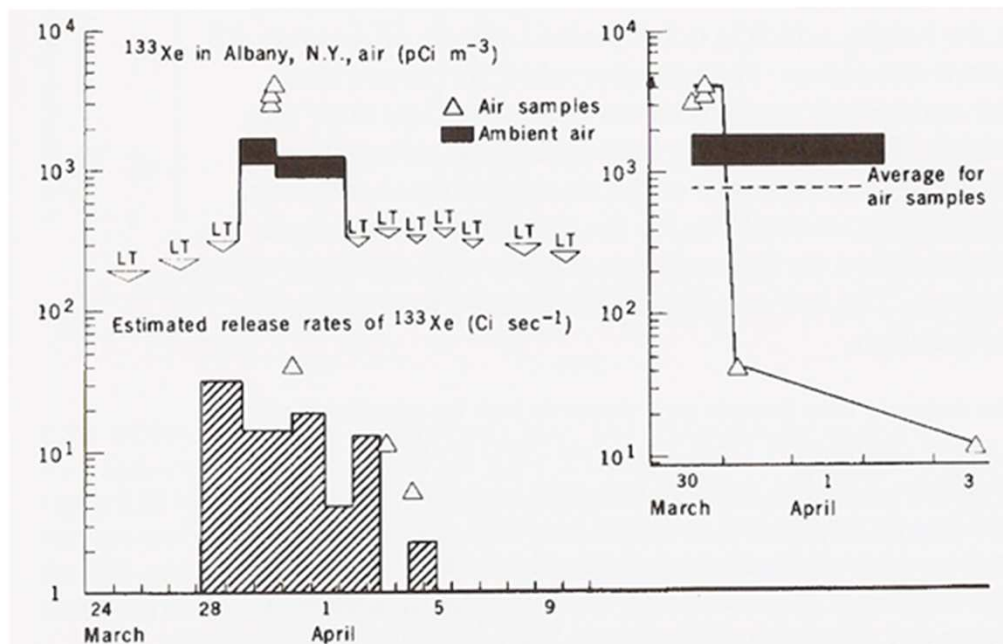
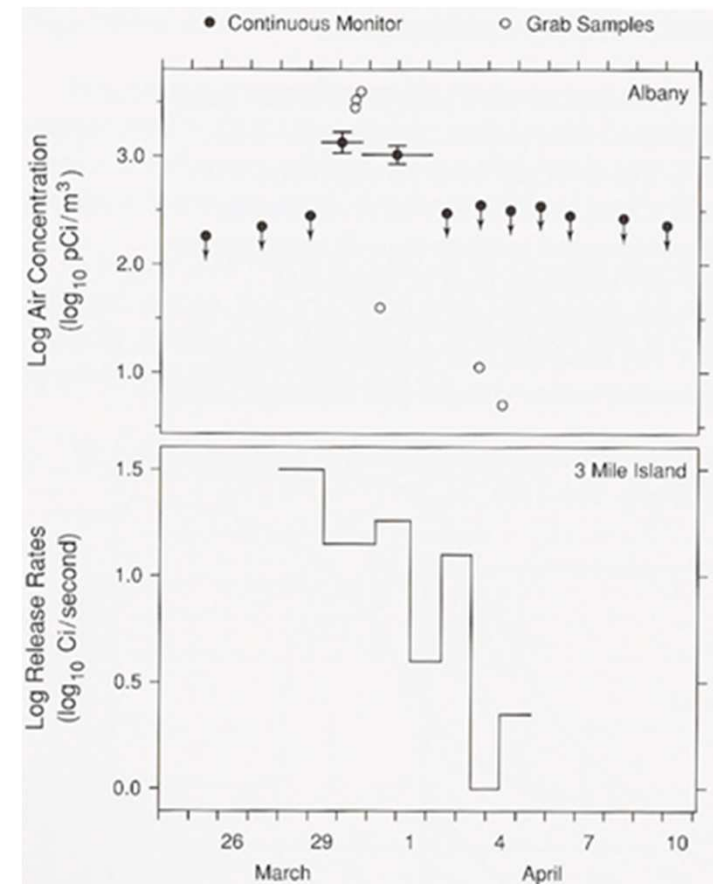
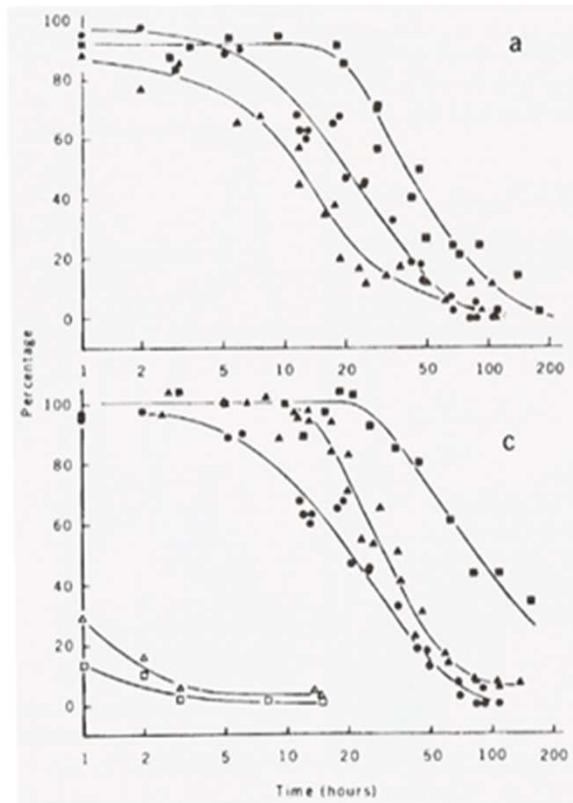


Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (2). The inset shows detailed values for air samples (gas counting) and concurrent average values for ambient air (Ge diode). Abbreviation: LT, less than.



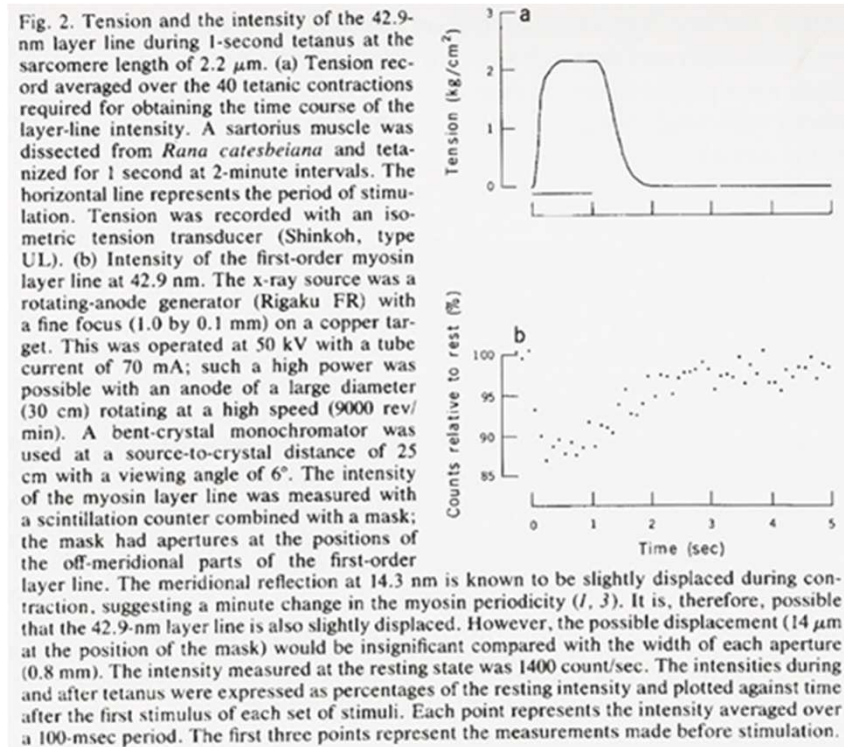
Improving the Vision

- Principle 5: Superposed data set
 - Symbols should be separable and data sets should be easily visually assembled.



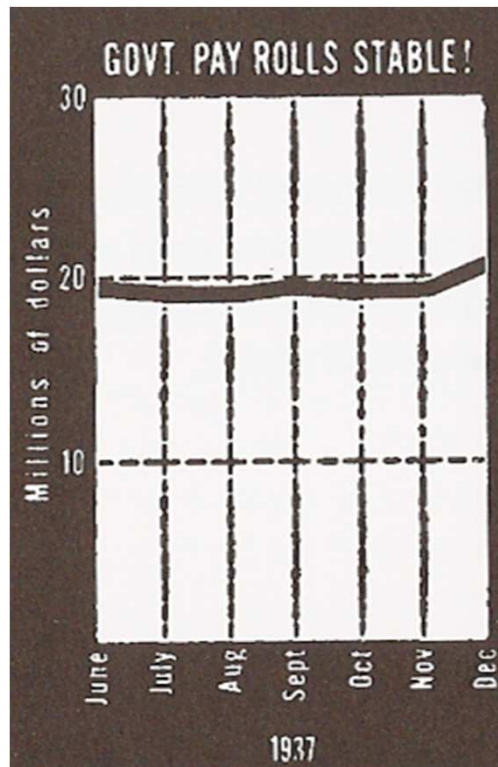
Improving the Understanding

- Principle 1: Provide explanations and draw conclusions
 - A graphical representation is often the means in which a hypothesis is confirmed or results are communicated.
 - Describe everything, draw attention to major features, describe conclusions



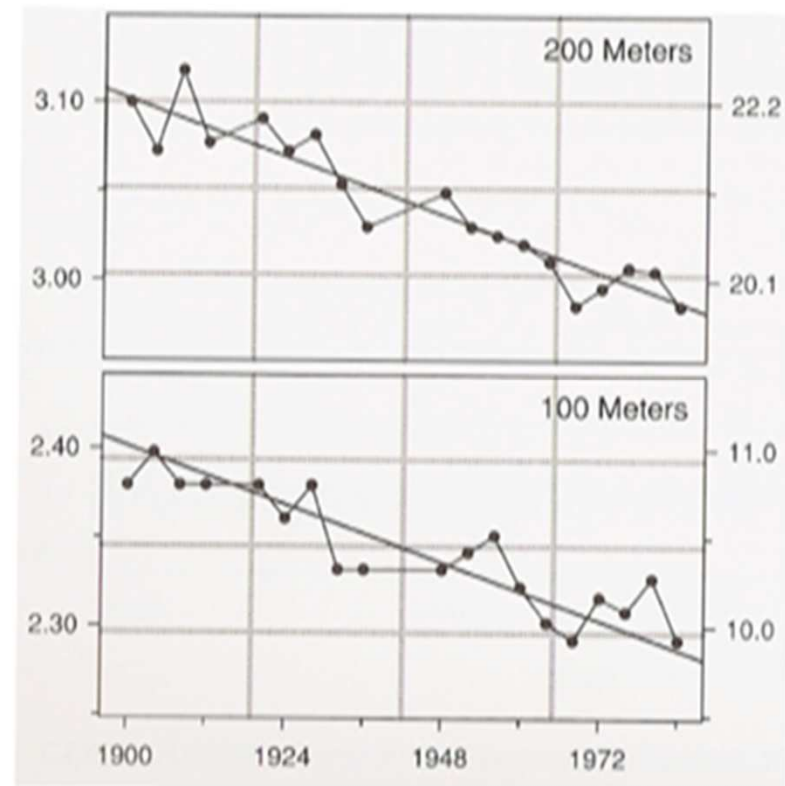
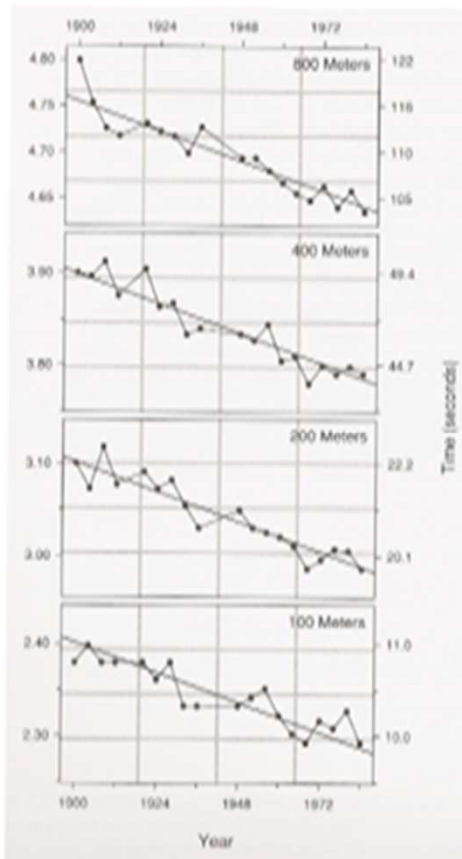
Improving the Understanding

- Principle 2: Use all available space.
 - Fill the data rectangle, only use zero if you need it



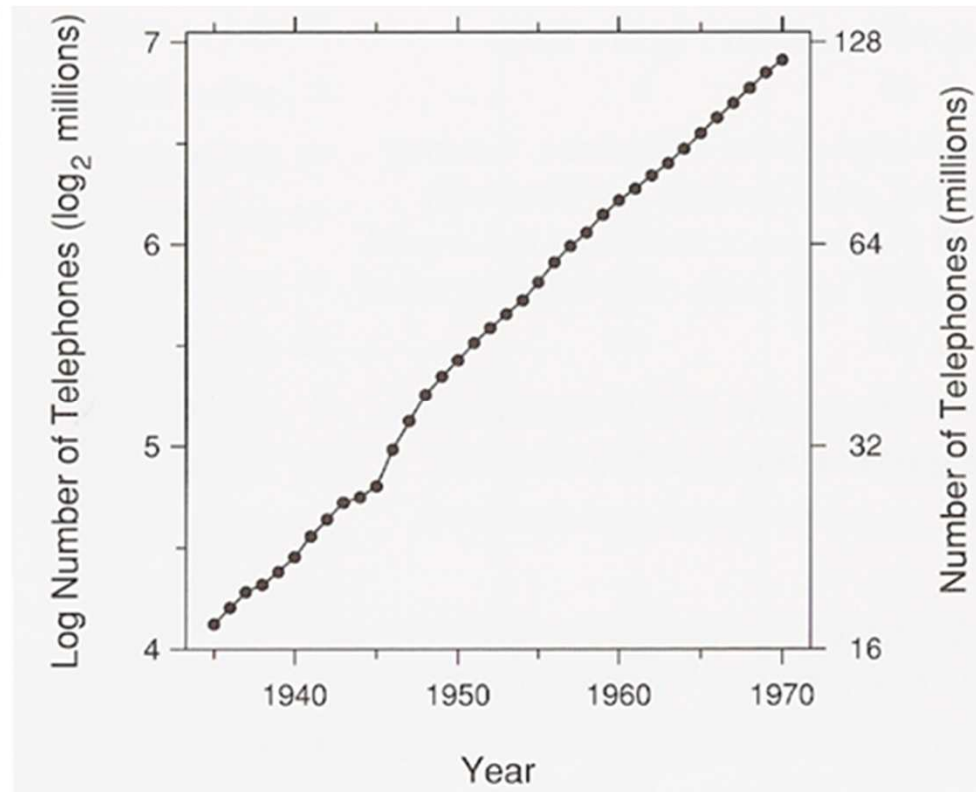
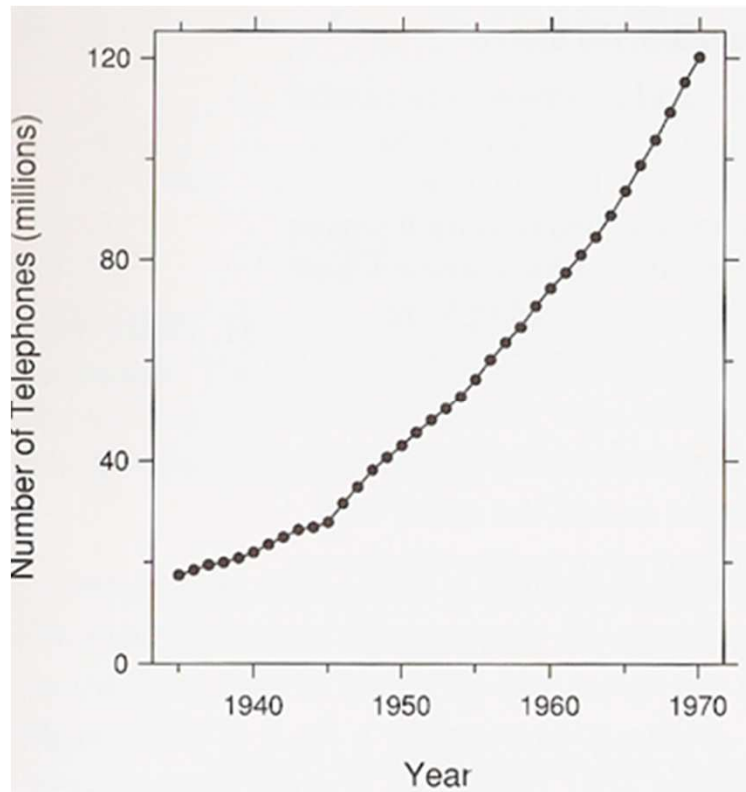
Improving the Understanding

- Principle 3: Align juxtaposed plots
 - Make sure scales match and graphs are aligned



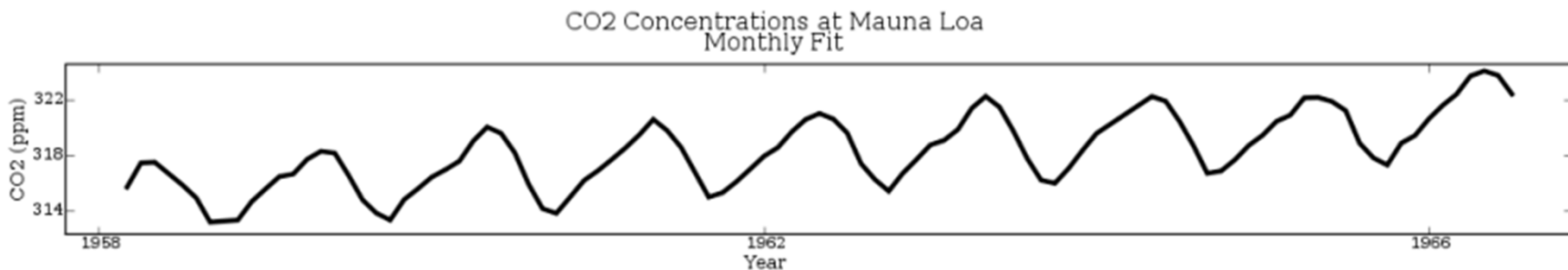
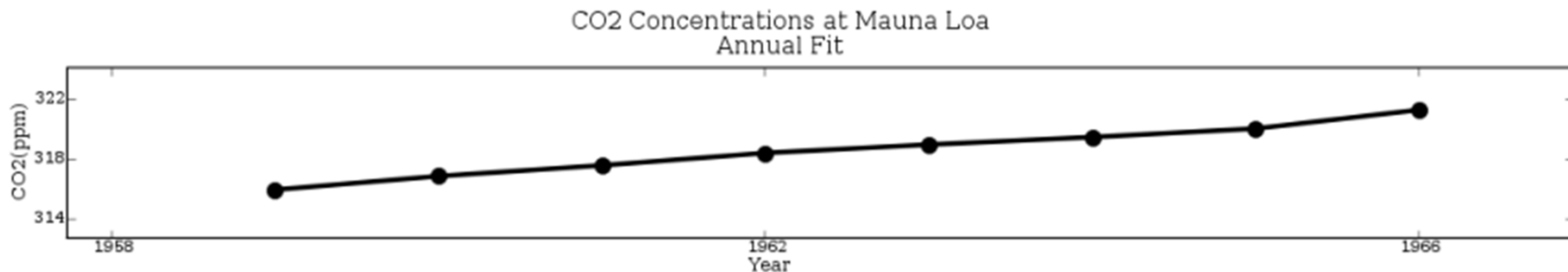
Improving the Understanding

- Principle 4: Use log scales when appropriate
 - Used to show percentage change, multiplicative factors and skewness



Improving the Understanding

- Principle 5: Bank to 45°
 - Optimize the aspect ratio of the plot



Summary of Principles

- Improve vision
 1. Reduced clutter, Make data stand out
 2. Use visually prominent graphical elements
 3. Use proper scale lines and a data rectangle
 4. Reference lines, labels, notes, and keys
 5. Superposed data set
- Improve understanding
 1. Provide explanations and draw conclusions
 2. Use all available space
 3. Align juxtaposed plots
 4. Use log scales when appropriate
 5. Bank to 45°