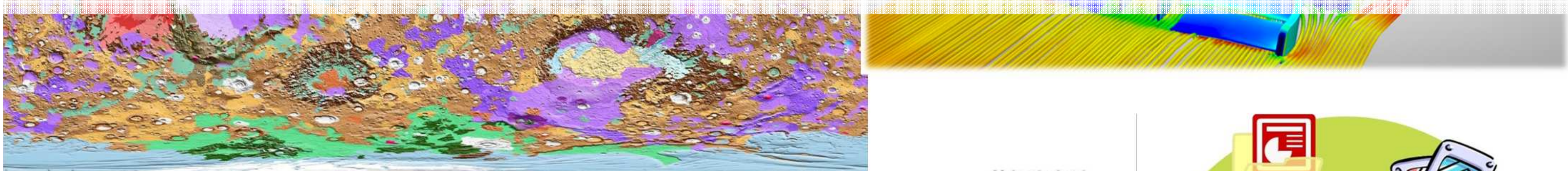
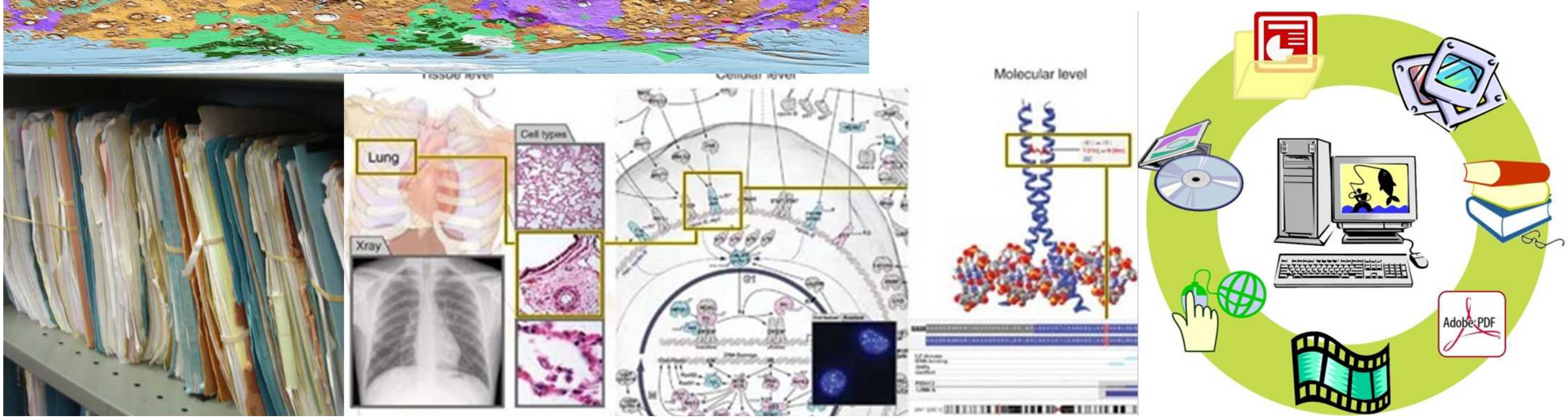
A Short Introduction on **Data Visualization**

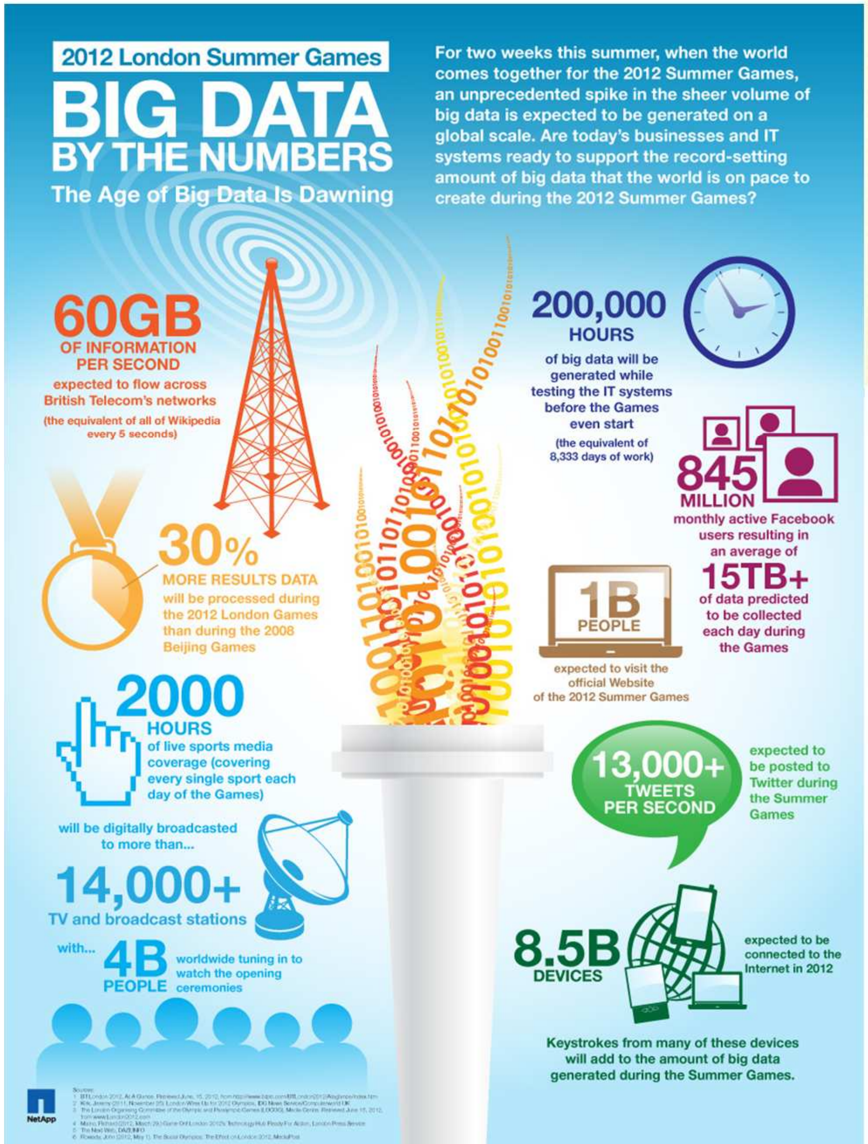
Guoning Chen

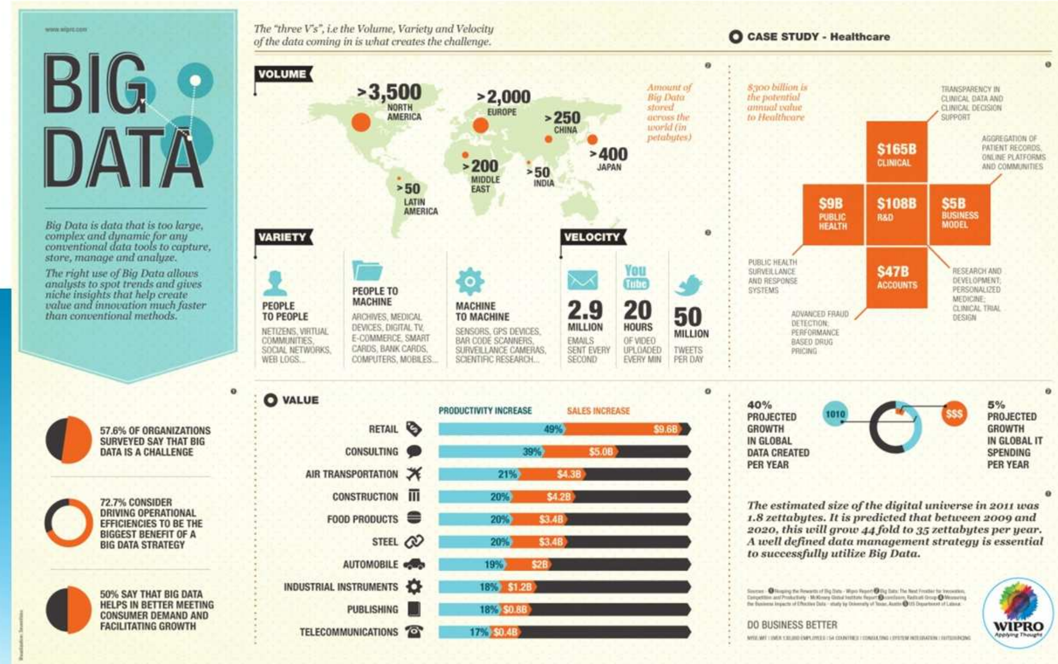
Data is generated everywhere and everyday





Age of Big Data



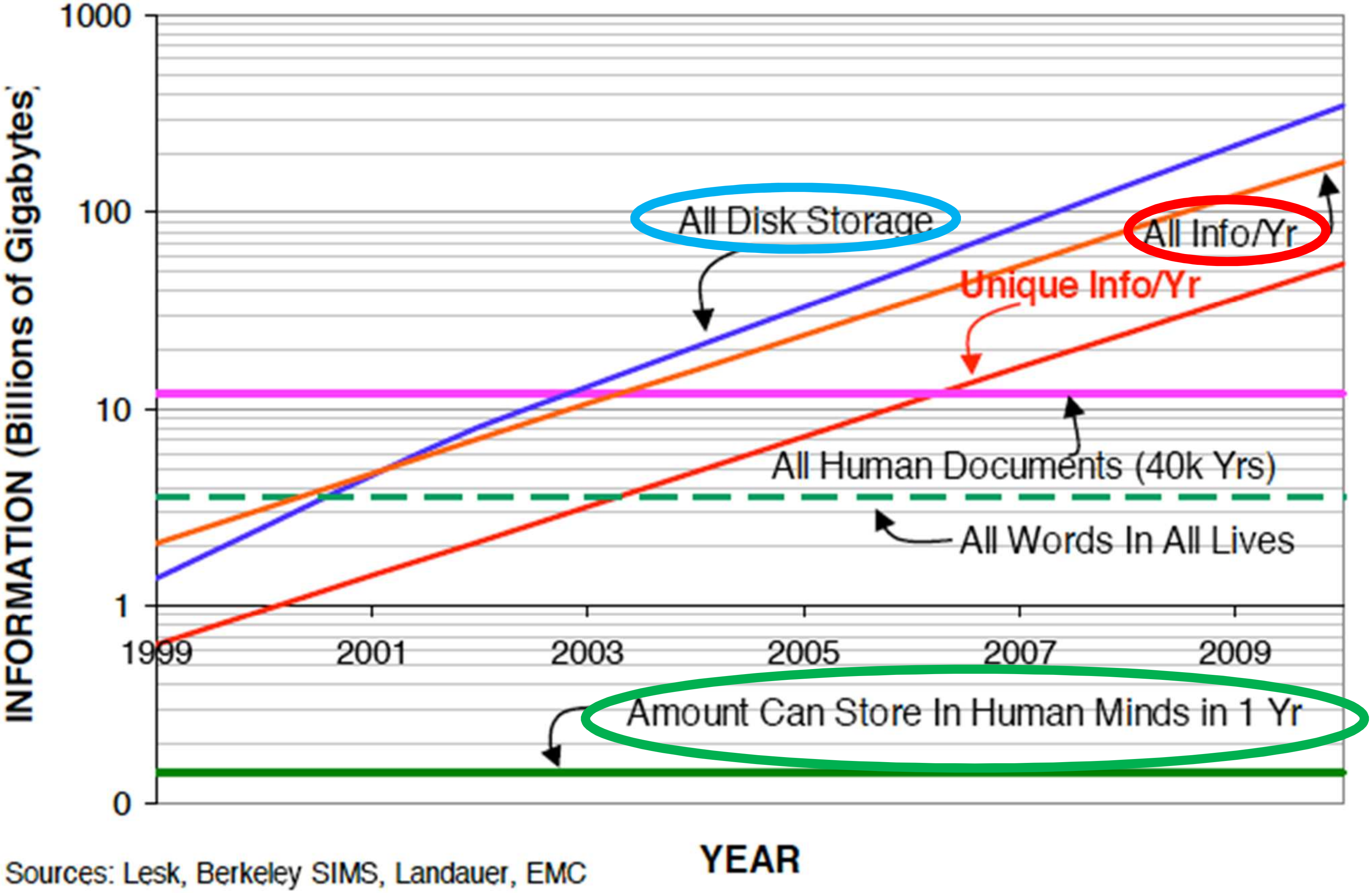


Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte

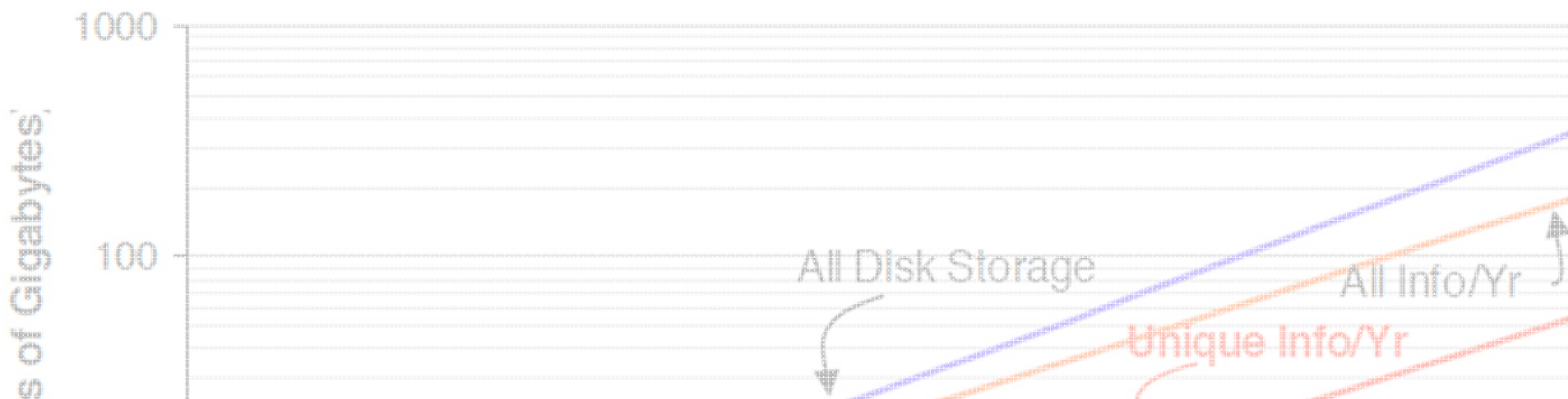
	Stored data in the United States, 2009[1] Petabytes	Number of firms with >1,000 employees[2]	Stored data per firm (>1,000 employees), 2009 Terabytes
Discrete manufacturing[3]	966	1,000	967[2]
Government	848	647	1,312
Communications and media	715	399	1,792
Process manufacturing[3]	694	835	831[2]
Banking	619	321	1,931
Health care providers[3]	434	1,172	370
Securities and investment services	429	111	3,866
Professional services	411	1,478	278
Retail	364	522	697
Education	269	843	319
Insurance	243	280	870
Transportation	227	283	801
Wholesale	202	376	536
Utilities	194	129	1,507
Resource industries	116	140	825
Consumer & recreational services	106	708	150
Construction	51	222	231

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.

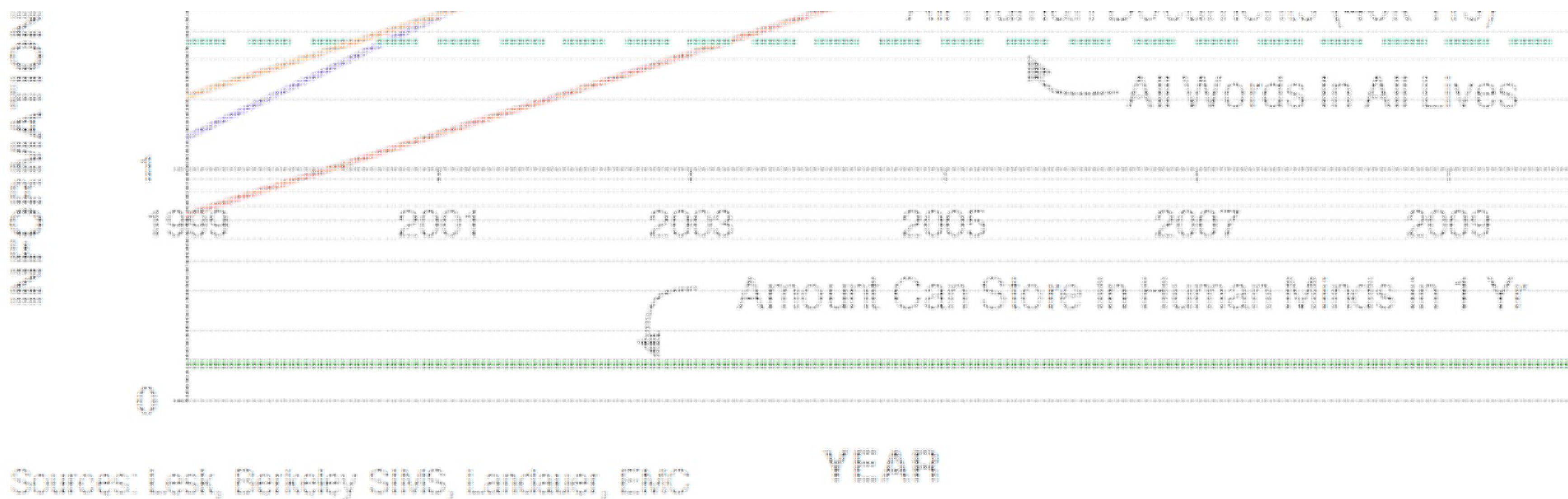
SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis



INFORMATION (Billions of Gigabytes)
1000
100
10
1
0
1999
2001
2003
2005
2007
2009
YEAR
All Disk Storage
All Info/Yr
Unique Info/Yr
All Human Documents (40k Yrs)
All Words In All Lives
Amount Can Store In Human Minds in 1 Yr
Sources: Lesk, Berkeley SIMS, Landauer, EMC

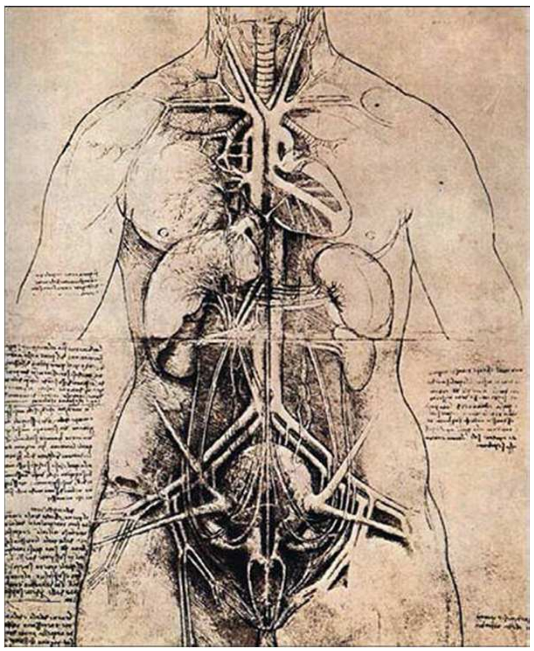


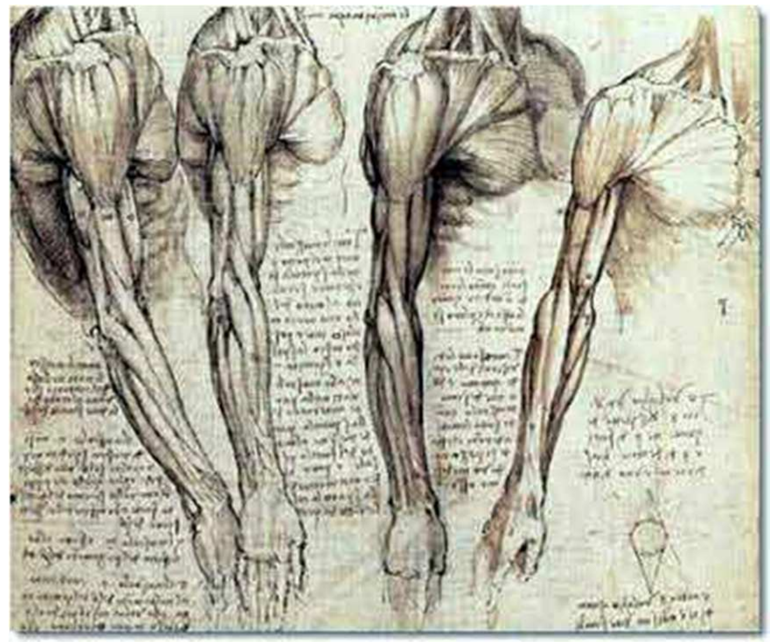
Data in ever increasing sizes ⇒ need an effective way to understand them



History of Visualization

- Visualization = rather old L. da Vinci (1452-1519)





- Often an intuitive step: graphical illustration

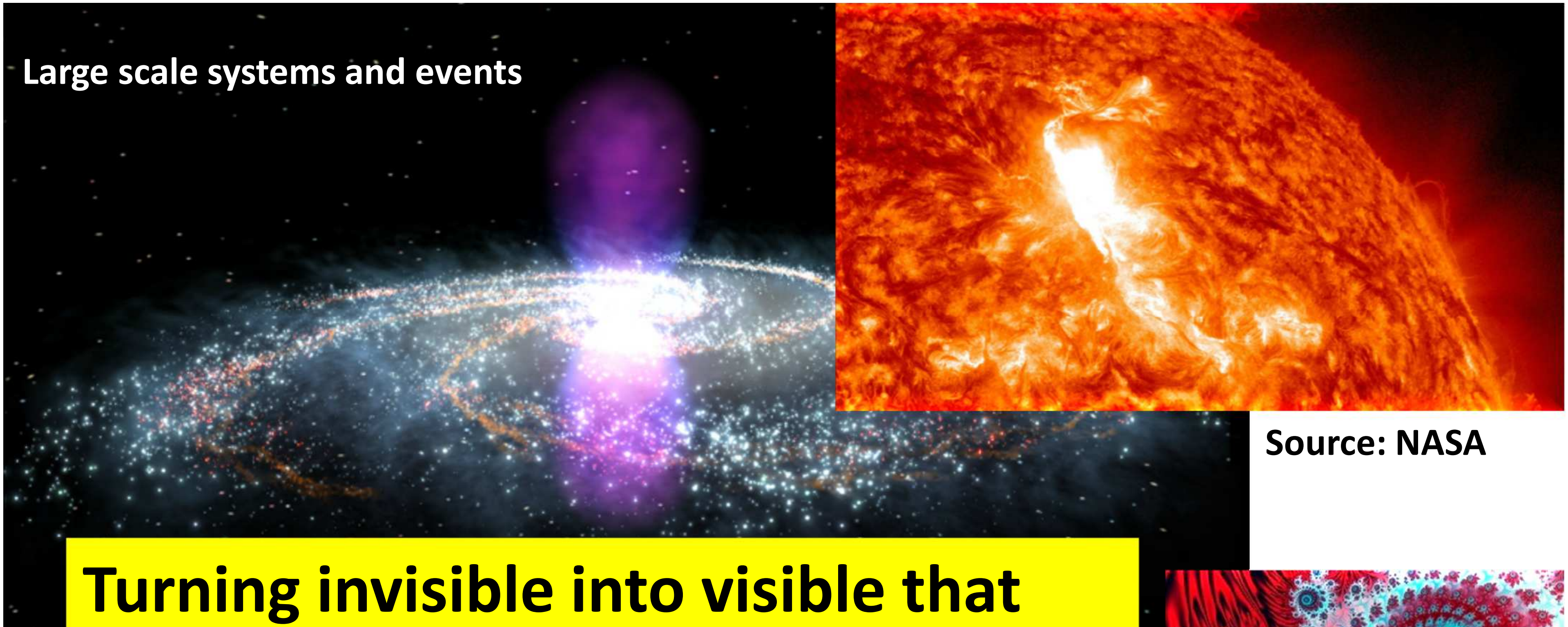
Image source: http://www.leonardo-da-vinci-biography.com/leonardo-da-vinci-anatomy.html

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

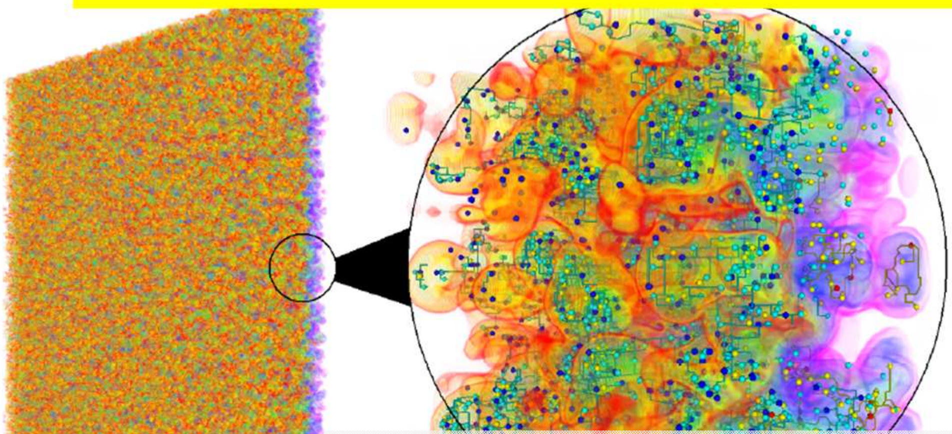
- Visualization transforms data into images that effectively and accurately represent information about the data.
  - Schroeder et al. The Visualization Toolkit, 2nd ed. 1998

Tool to enable a User insight into Data



Source: NASA

Turning invisible into visible that people can understand intuitively

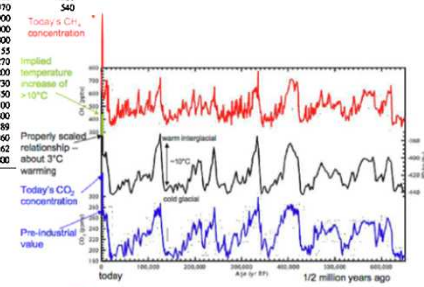


The impact of the ball entering the porous solid from the left

Table 7-2 Direct global warming potentials of several well-mixed trace gases relative to CO_2. The GWPs of the various non-CO_2 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	Lifetime (years)	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	4
N_2O	132	260	270	270	240	170
CFC-11	55	4500	4100	3400	2400	1400
CFC-12	116	7100	7400	7100	6200	4100
HCFC-22	15.8	4200	2600	1600	970	540
CFC-113	110	4600	4700	4500	3900	
CFC-114	220	6100	6700	7000	7000	
CFC-115	550	5500	6200	7000	7800	
HCFC-123	1.71	330	150	90	55	
HCFC-124	6.9	1500	760	440	270	
HFC-125	40.5	5200	4500	3400	2200	
HFC-134a	15.6	3100	1900	1200	730	
HCFC-141b	10.8	1800	980	580	350	
HCFC-142b	22.4	4000	2800	1800	1100	
HFC-143a	64.2	4700	4500	3800	2800	
HFC-152a	1.8	530	250	150	89	
CCl_4	47	1800	1600	1300	860	
CH_3CCl_3	6.1	360	170	100	62	
CF_3Br	77	5600	5500	[illegible]	3800	

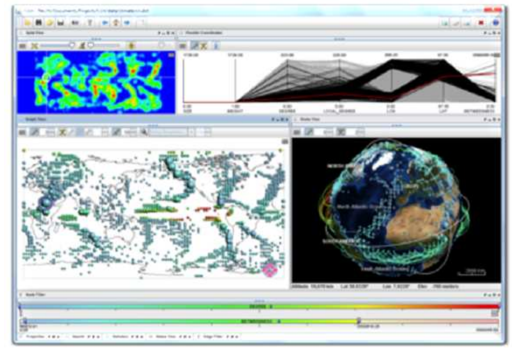
SAOD Table 7.2 (p[illegible].6)

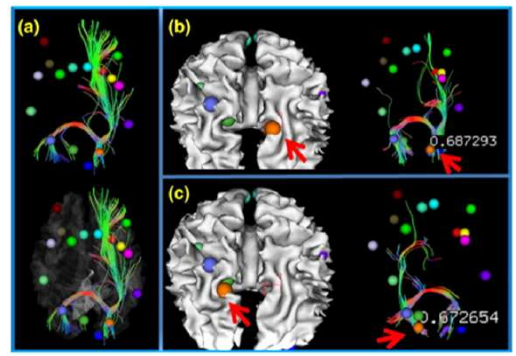


Methane, temperature (from hydrogen isotope ratios ("δD") and carbon dioxide from the Dome C Ice core. (EPICA Project members, 2006).

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





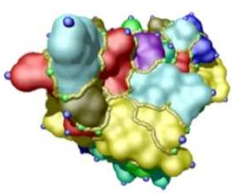
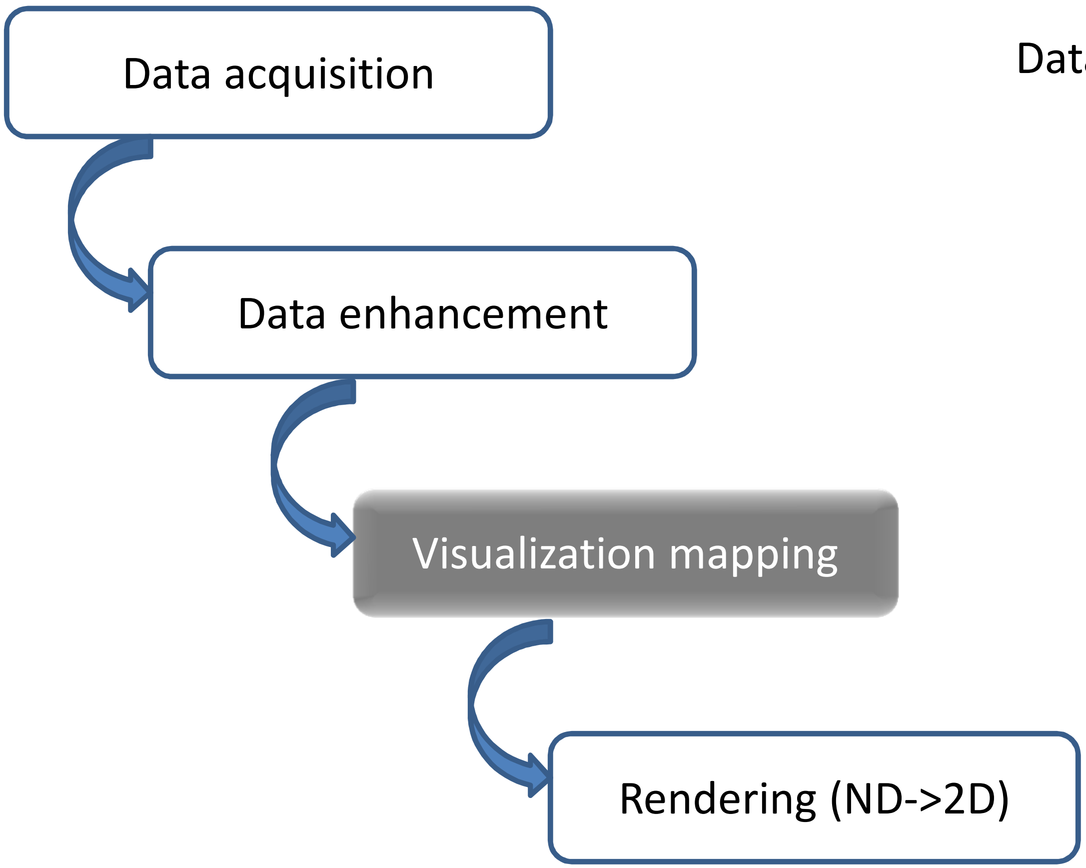


Image source: Google images

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

A Visualization Pipeline



Data are generated/collected.

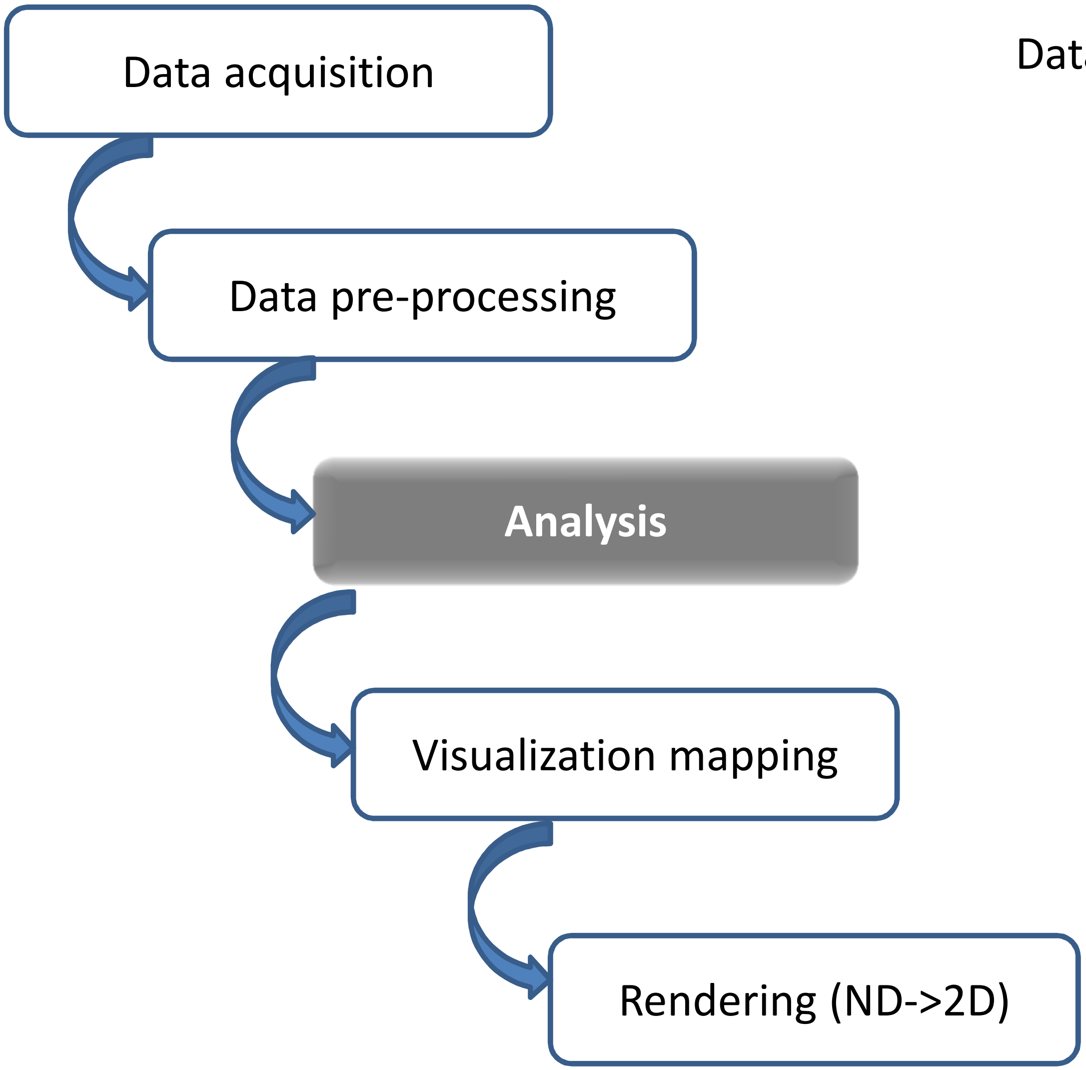
Data are processed.

Data are mapped to visual primitives, e.g. colors, geometry, etc.

Images are generated.

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

Data Visual Analytic Pipeline



Data are generated/collected.

Data are pre-processed.

Feature detection
Structure extraction
Statistical analysis
etc.

Data are mapped to visual primitives, e.g. colors, geometry, etc.

Images are generated.

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.

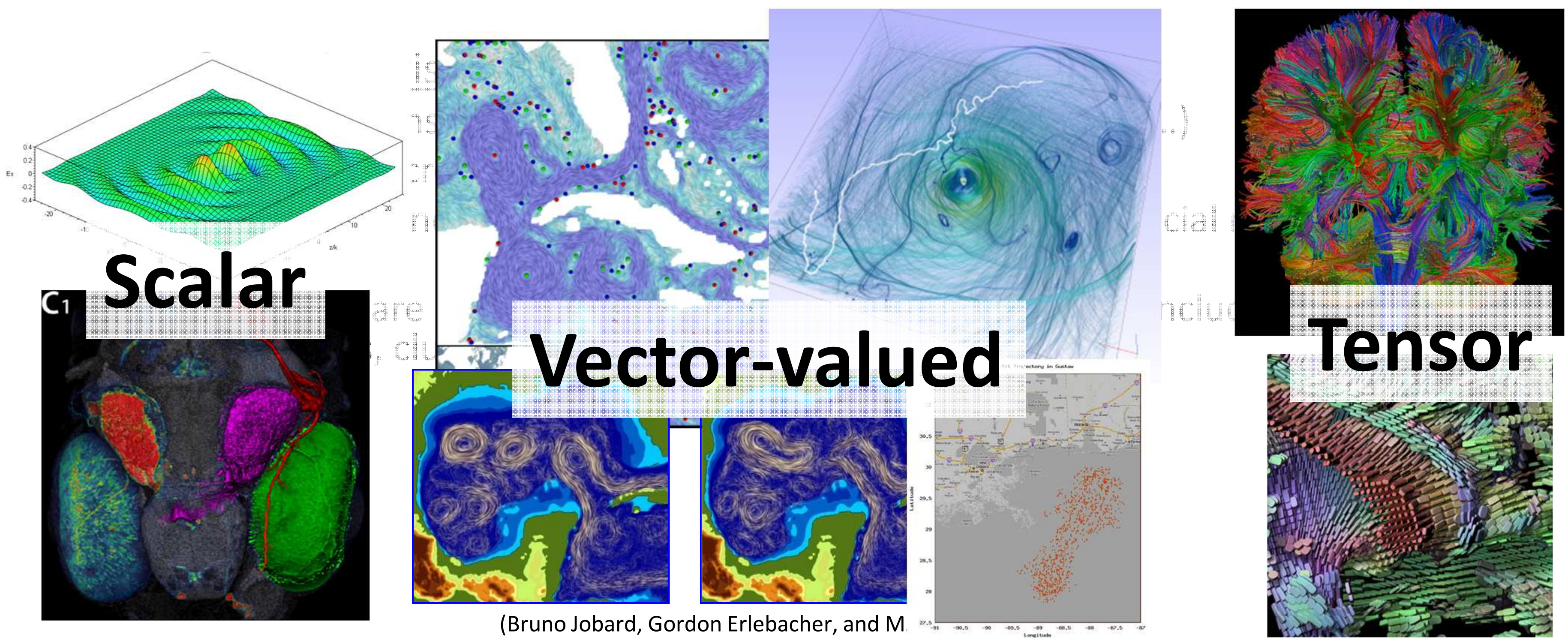
- From simple data to more complex ones.

- From represent the data with fidelity to reveal new findings.

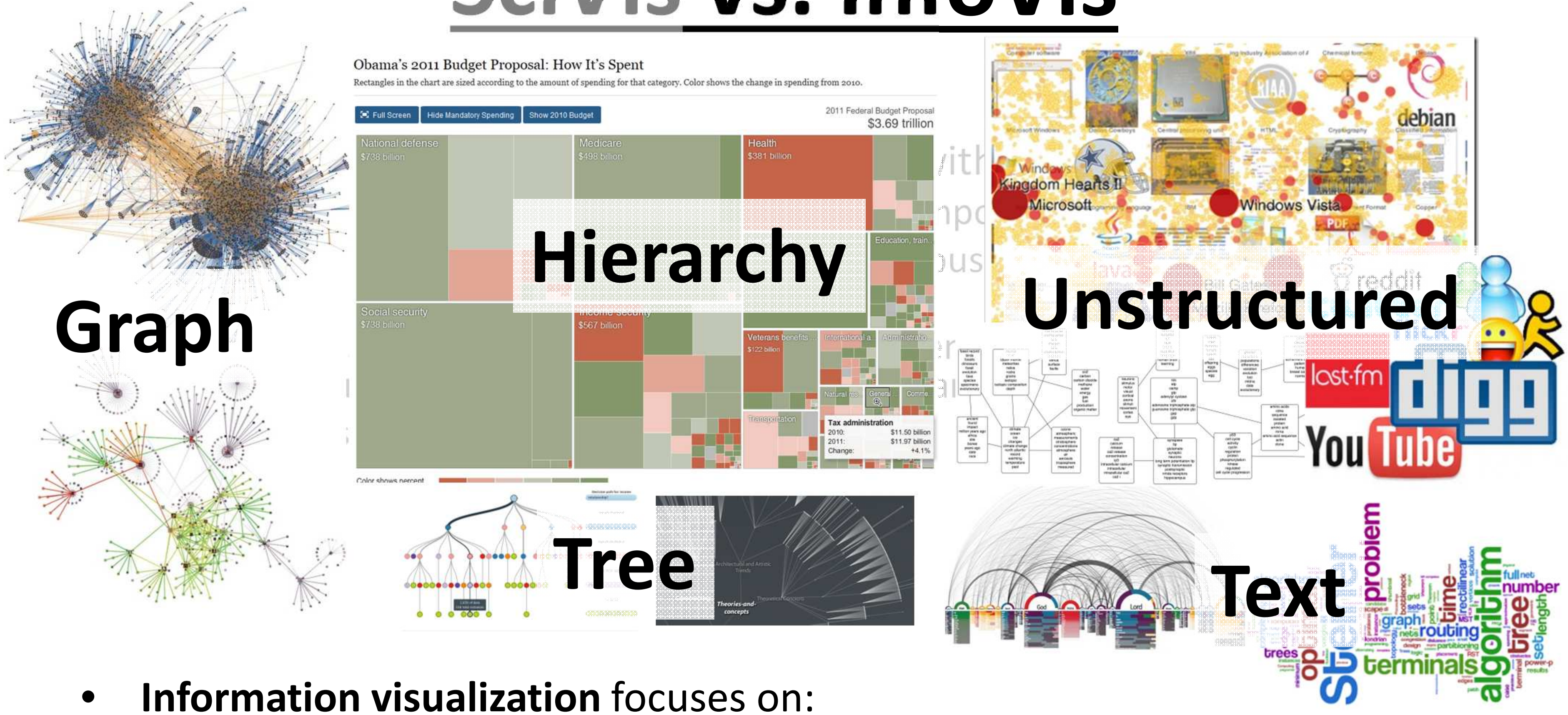
- From scientific visualization to information visualization, bio-visualization, geographical data visualization, and beyond.

SciVis vs. InfoVis

- **Scientific visualization** is mostly concerned with:
  - Data defined in physical space, i.e. spatio-temporal data (2~4 dimensions)
  - Data describes continuous events in continuous space, however, the representation is discrete (i.e. sampled data)
  - Examples include simulation and measurement data from physics, chemistry, geo-science, medical-biological, climate, oceanography, energy, ……
  - Features are well-defined



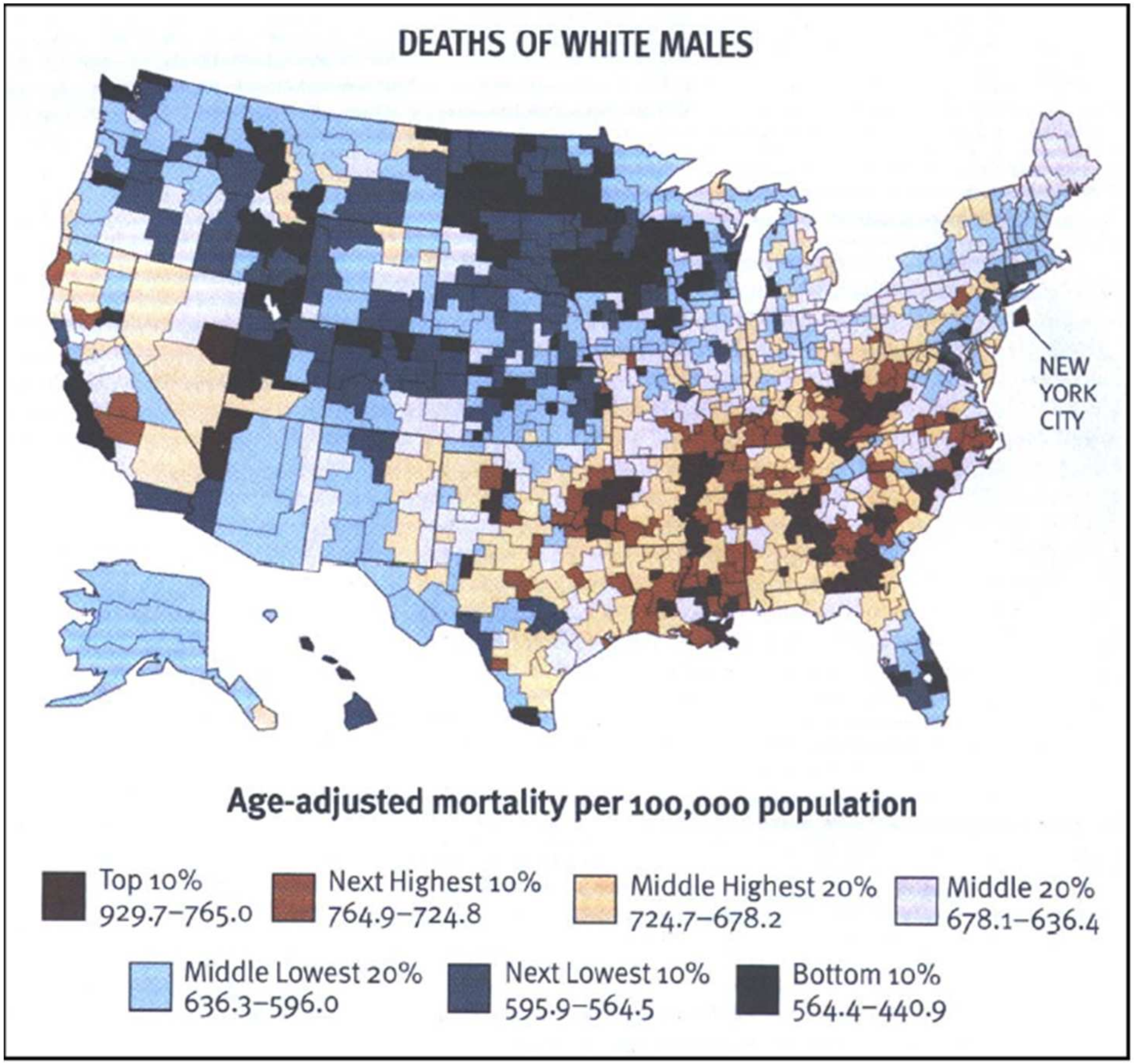
SciVis vs. InfoVis



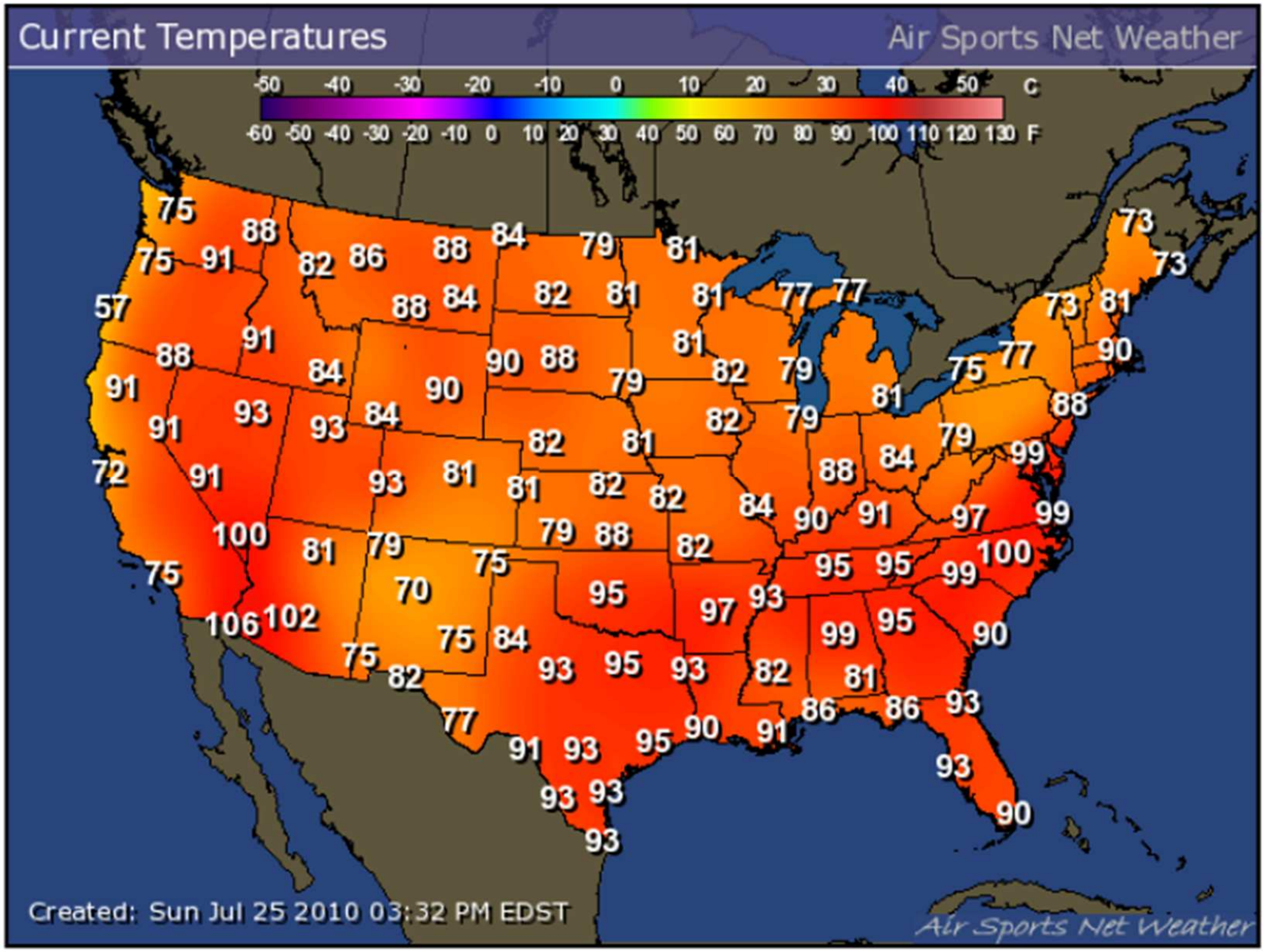
- **<u>Information visualization</u>** focuses on:
  - high-dimensional (>>4), abstract data (i.e. tree, graphs, hierarchy, …)
  - Data is discrete in the nature
  - Examples include financial, marketing, HR, statistical, social media, political, ……
  - Feature are not well-defined, the typical analysis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

Use Colors Wisely

What is Wrong with this Color Scale

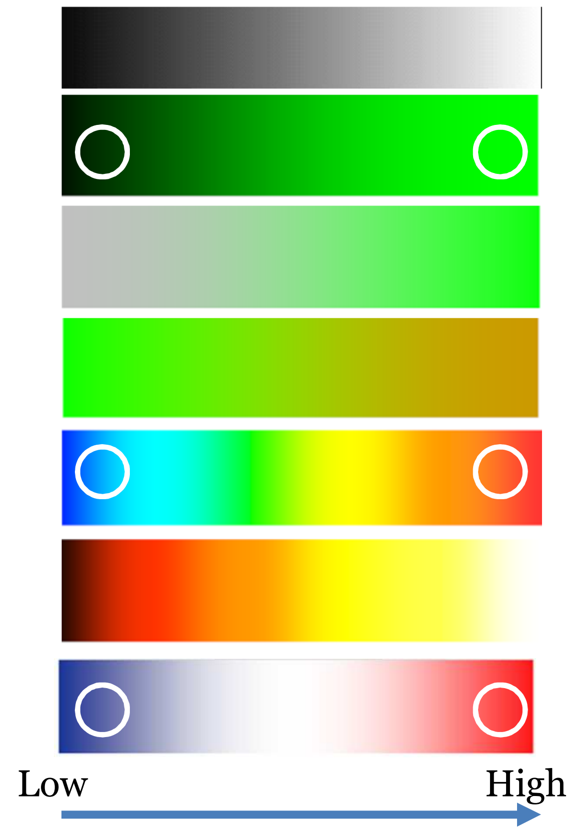


Not a bad choice of color scale, but the Dynamic Range needs some work



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

Examples of Pre-Established Color Meanings

Red	Green	Blue
Stop	On	Cool
Off	Plants	Safe
Dangerous	Carbon	Deep
Hot	Moving	Nitrogen
High stress	Money	
Oxygen		
Shallow		
Money loss		

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

Color Alone Doesn't Cut It

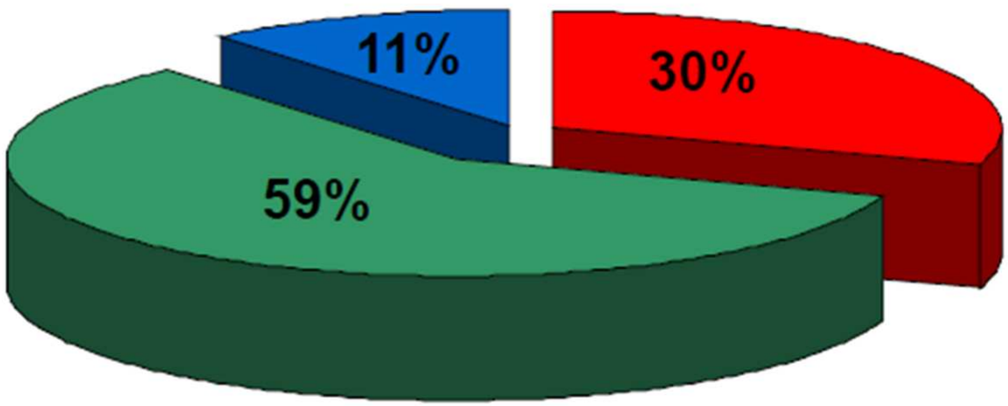
I sure hope that my life does not depend on being able to read this quickly and accurately!

Luminance Contrast is Crucial

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

The Luminance Equation

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



Material from Dr. Mike Bailey, Oregon State Univ.

≈ 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

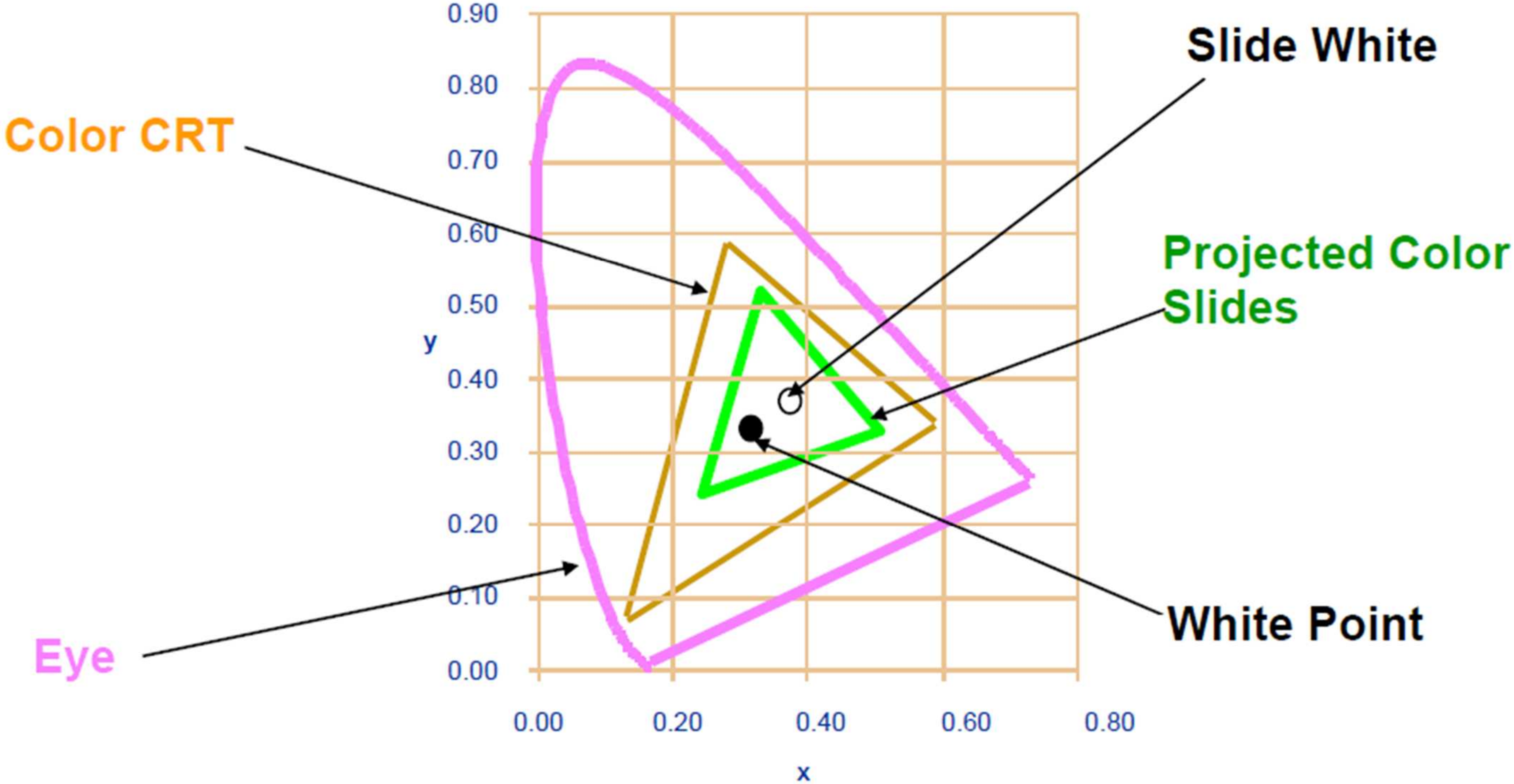
Use good contrast

ΔL* of about 0.40 makes good contrast

Material from Dr. Mike Bailey, Oregon State Univ.

Be Aware of the Different Color Ranges on Different Devices

Color Gamut for a Monitor and Color Slides



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

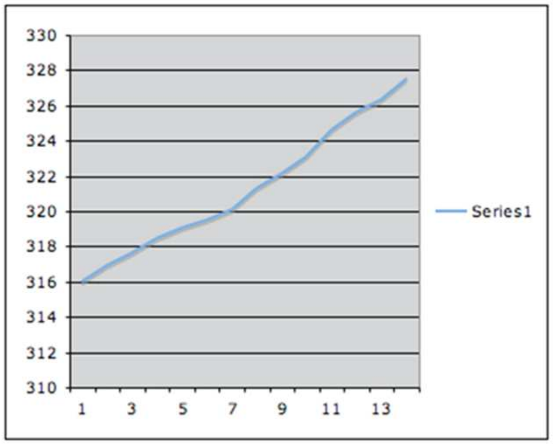
Just because you have millions of colors to choose from

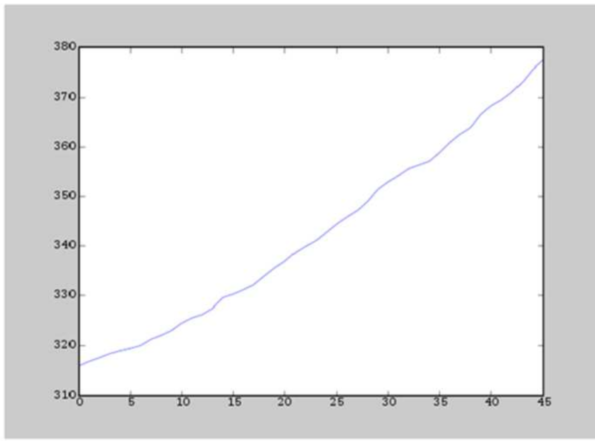
doesn't mean you must use them all •••

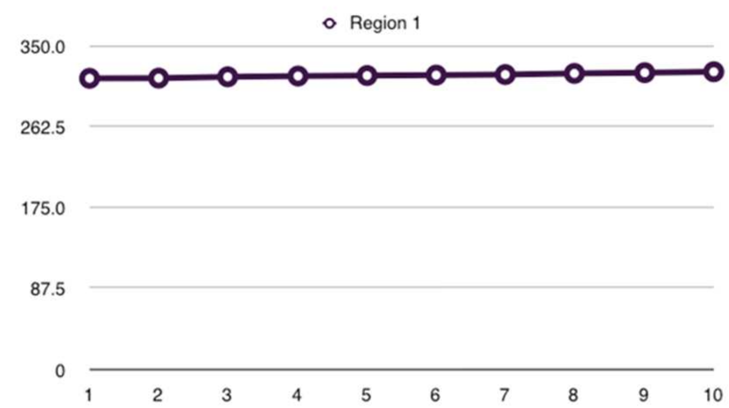
Some Principles for Plots

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

The information provided here should be considered as guidelines



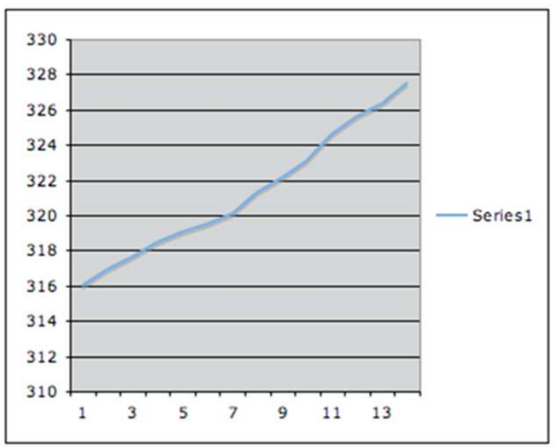


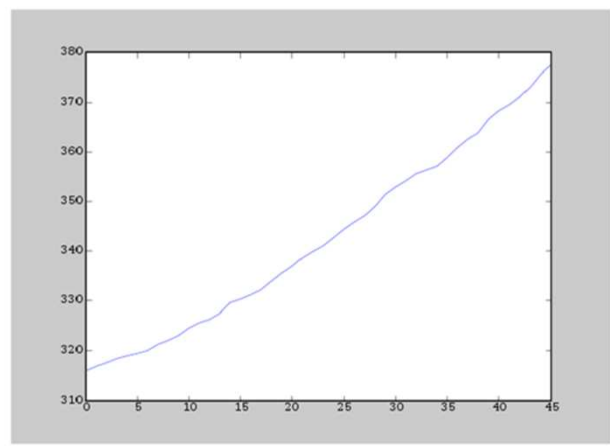


- Why are they all different?
- What is good/bad about each?

Improving the Vision

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

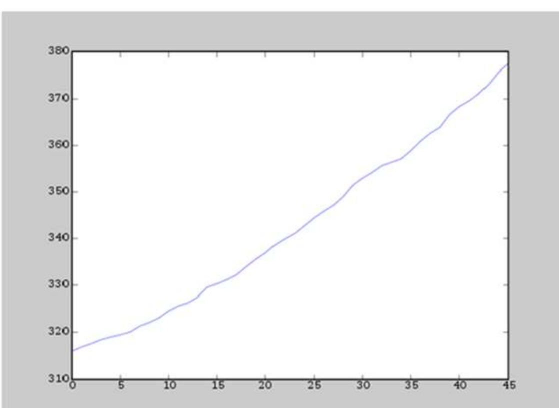


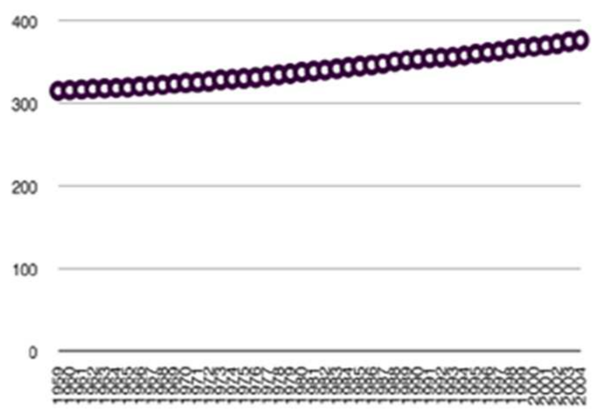


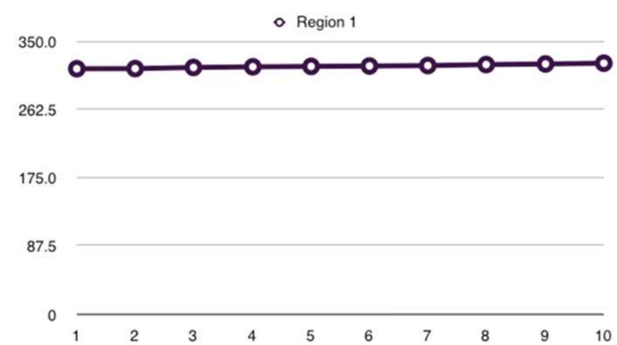
Which one is better?

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

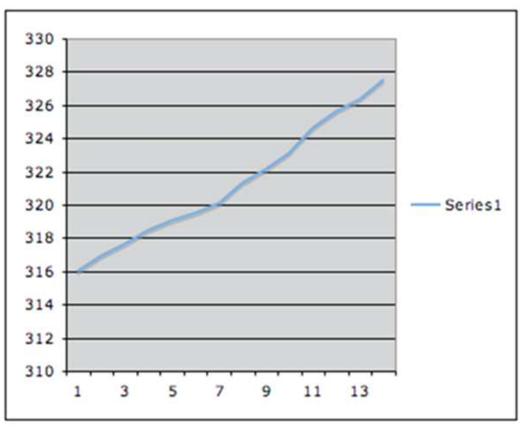


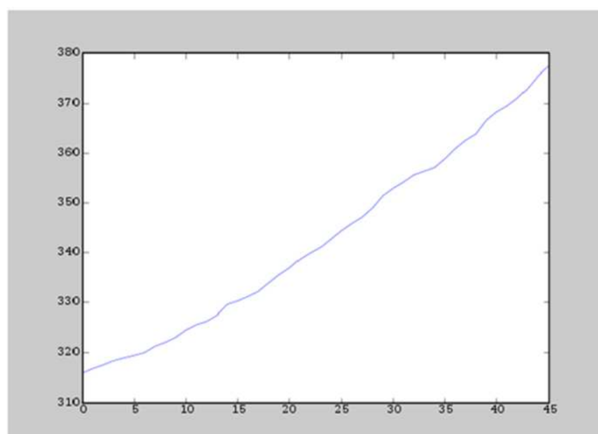


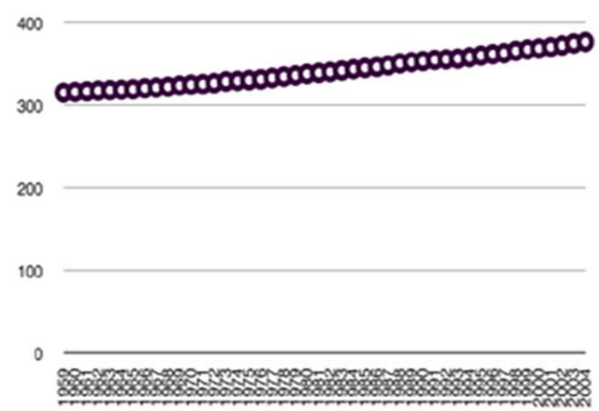


Improving the Vision

- <u>Principle 3</u>: 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 to data rectangle completely.
  - Add margins for data
  - Tick-marks outs and 3-10 for each axis

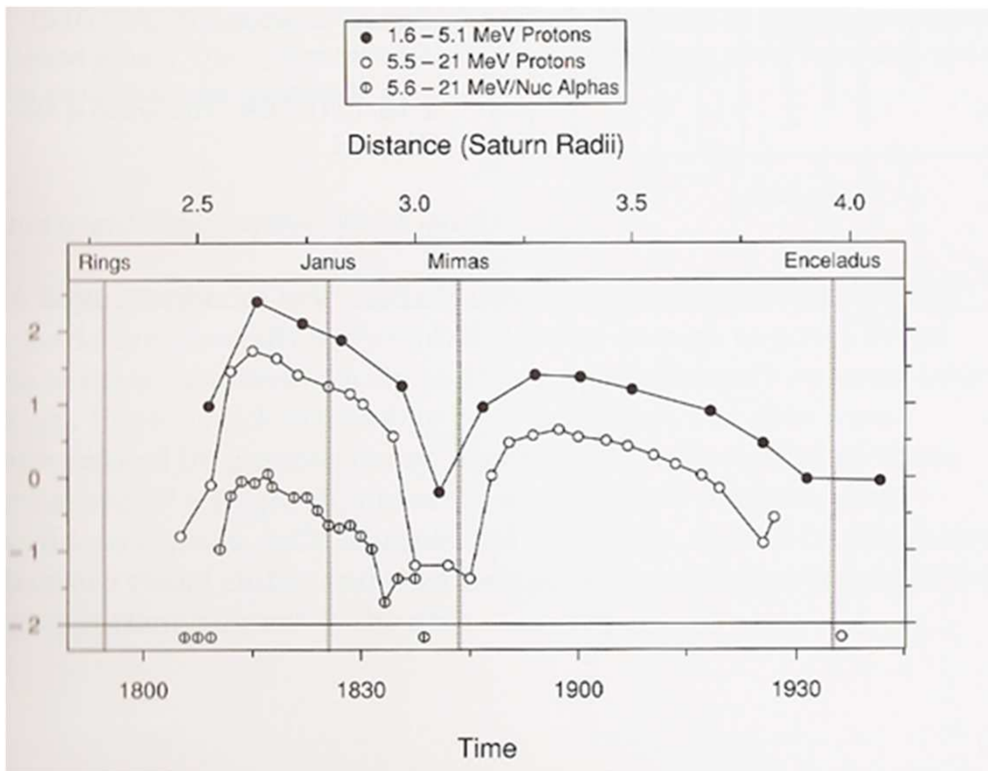






Improving the Vision

- 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.

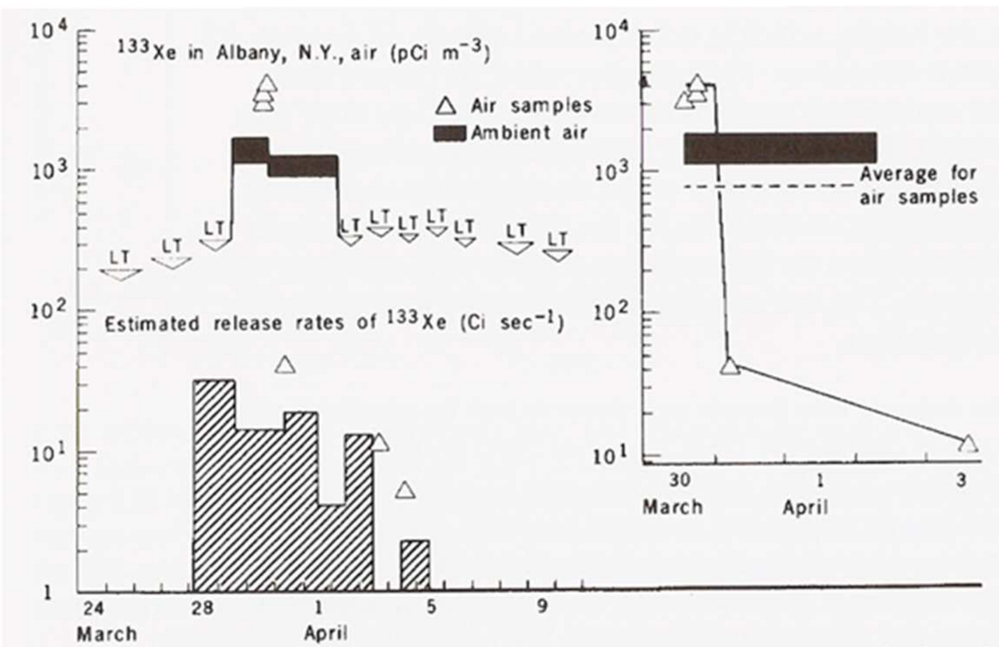
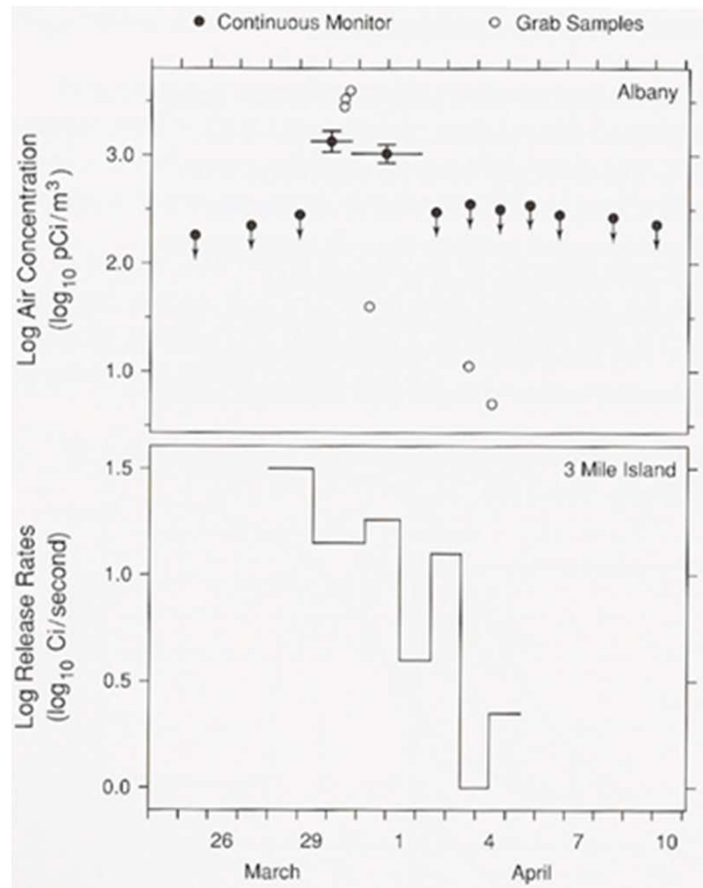
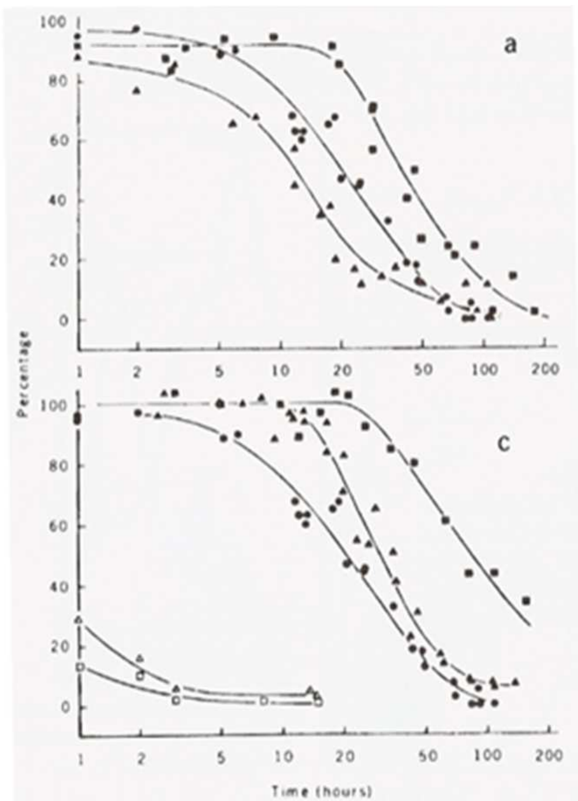


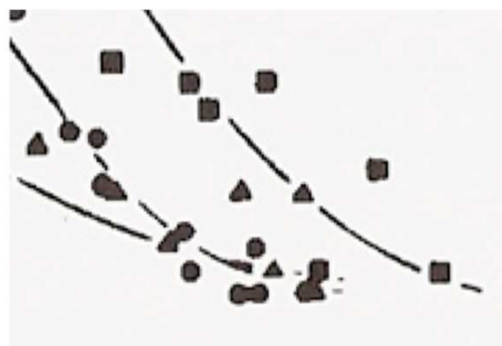
Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (2). The inset shows detailed values for air samples (gas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



Improving the Vision

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





Improving the Understanding

- <u>Principle 1</u>: 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

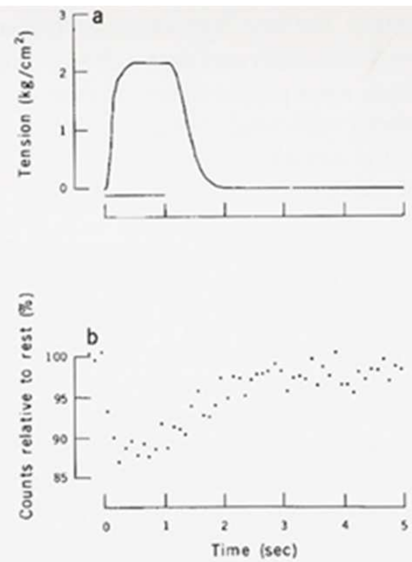
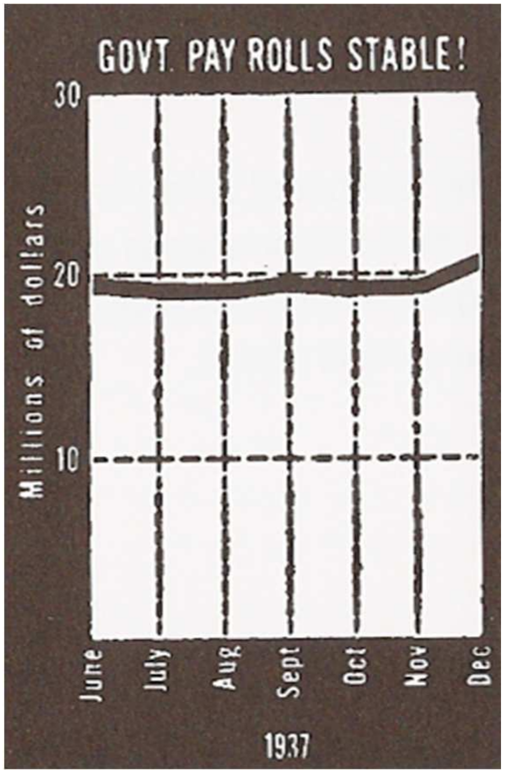


Fig. 2. Tension and the intensity of the 42.9-nm layer line during 1-second tetanus at the sarcomere length of 2.2 μm. (a) Tension record averaged over the 40 tetanic contractions required for obtaining the time course of the layer-line intensity. A sartorius muscle was dissected from *Rana catesbeiana* and tetanized for 1 second at 2-minute intervals. The horizontal line represents the period of stimulation. Tension was recorded with an isometric tension transducer (Shinkoh, type UL). (b) Intensity of the first-order myosin layer line at 42.9 nm. The x-ray source was a rotating-anode generator (Rigaku FR) with a fine focus (1.0 by 0.1 mm) on a copper target. This was operated at 50 kV with a tube current of 70 mA; such a high power was possible with an anode of a large diameter (30 cm) rotating at a high speed (9000 rev/min). A bent-crystal monochromator was used at a source-to-crystal distance of 25 cm with a viewing angle of 6°. The intensity of the myosin layer line was measured with a scintillation counter combined with a mask; the mask had apertures at the positions of the off-meridional parts of the first-order layer line. The meridional reflection at 14.3 nm is known to be slightly displaced during contraction, suggesting a minute change in the myosin periodicity (*1, 3*). It is, therefore, possible that the 42.9-nm layer line is also slightly displaced. However, the possible displacement (14 μm at the position of the mask) would be insignificant compared with the width of each aperture (0.8 mm). The intensity measured at the resting state was 1400 count/sec. The intensities during and after tetanus were expressed as percentages of the resting intensity and plotted against time after the first stimulus of each set of stimuli. Each point represents the intensity averaged over a 100-msec period. The first three points represent the measurements made before stimulation.

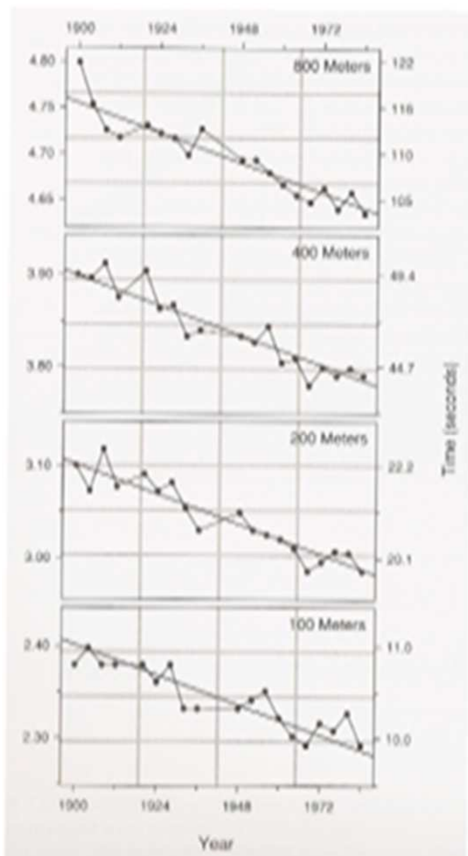
Improving the Understanding

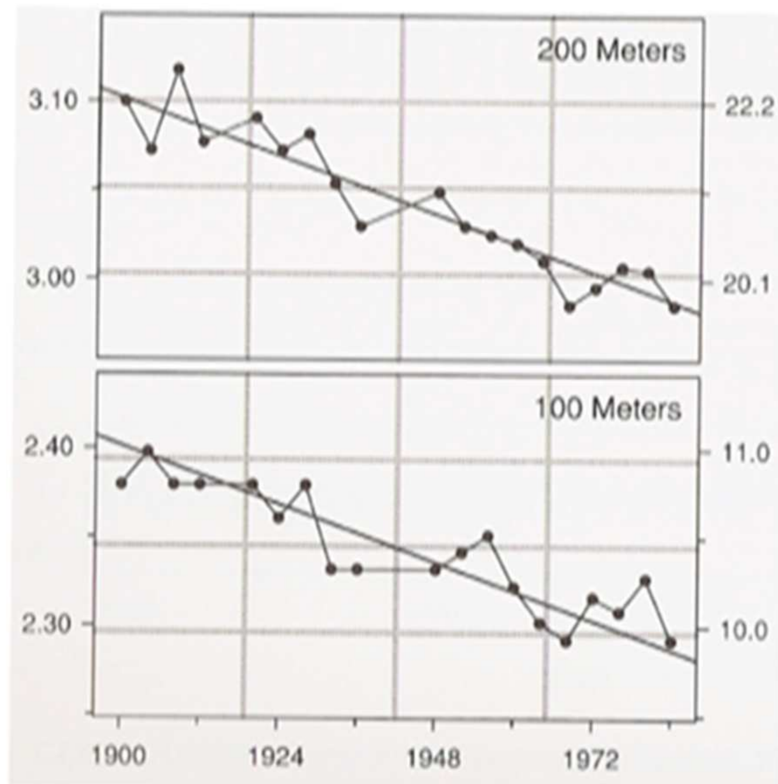
- Principle 2: Use all available space.
  - Fill the data rectangle, only use zero if you need it



Improving the Understanding

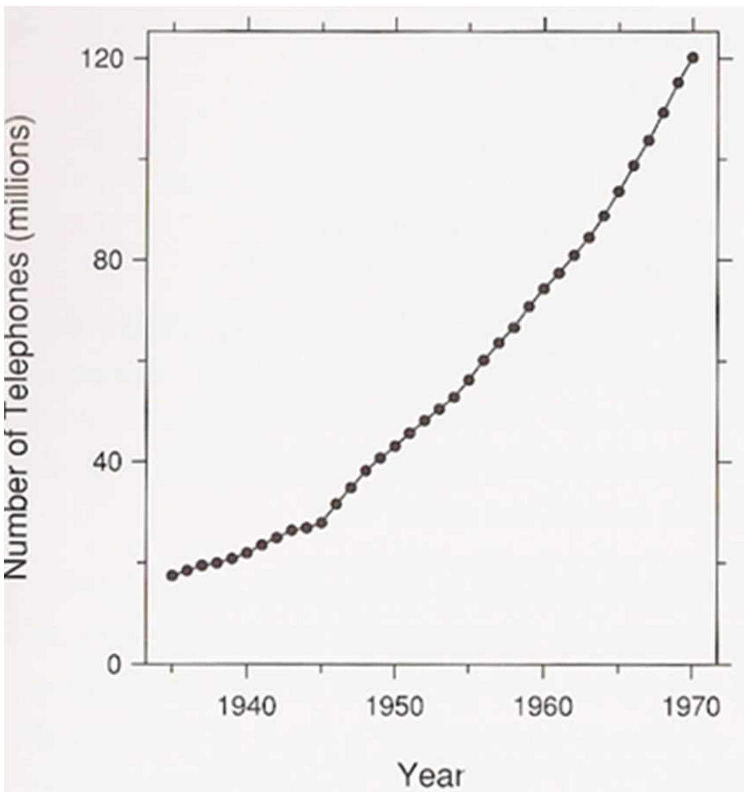
- Principle 3: Align juxtaposed plots
  - Make sure scales match and graphs are aligned

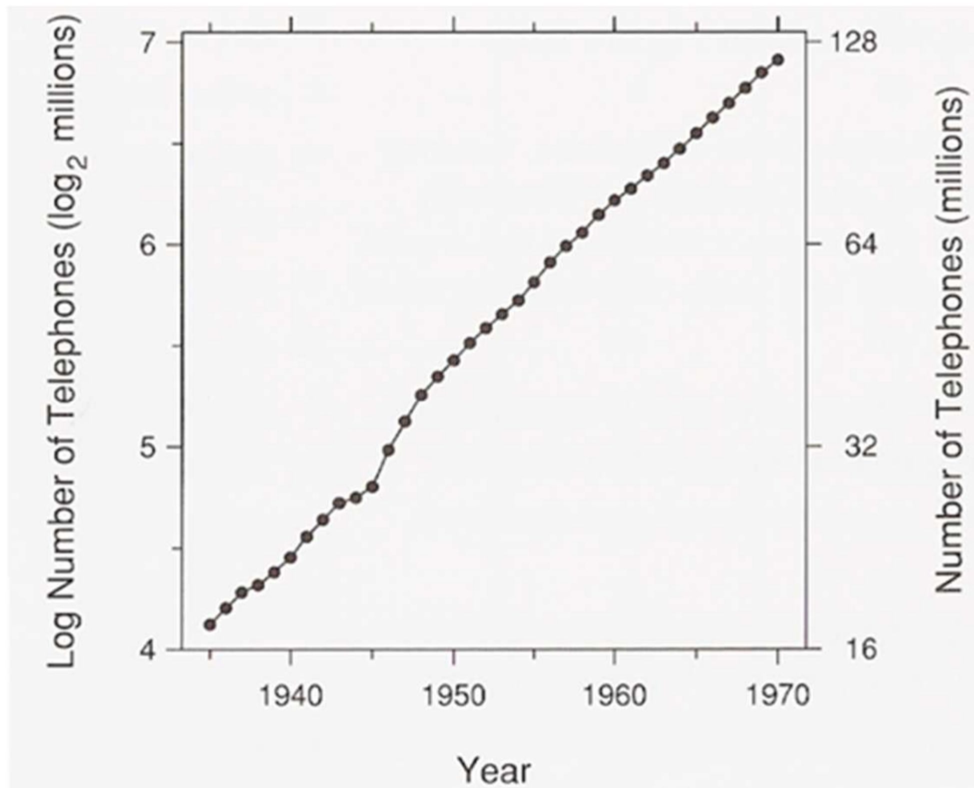




Improving the Understanding

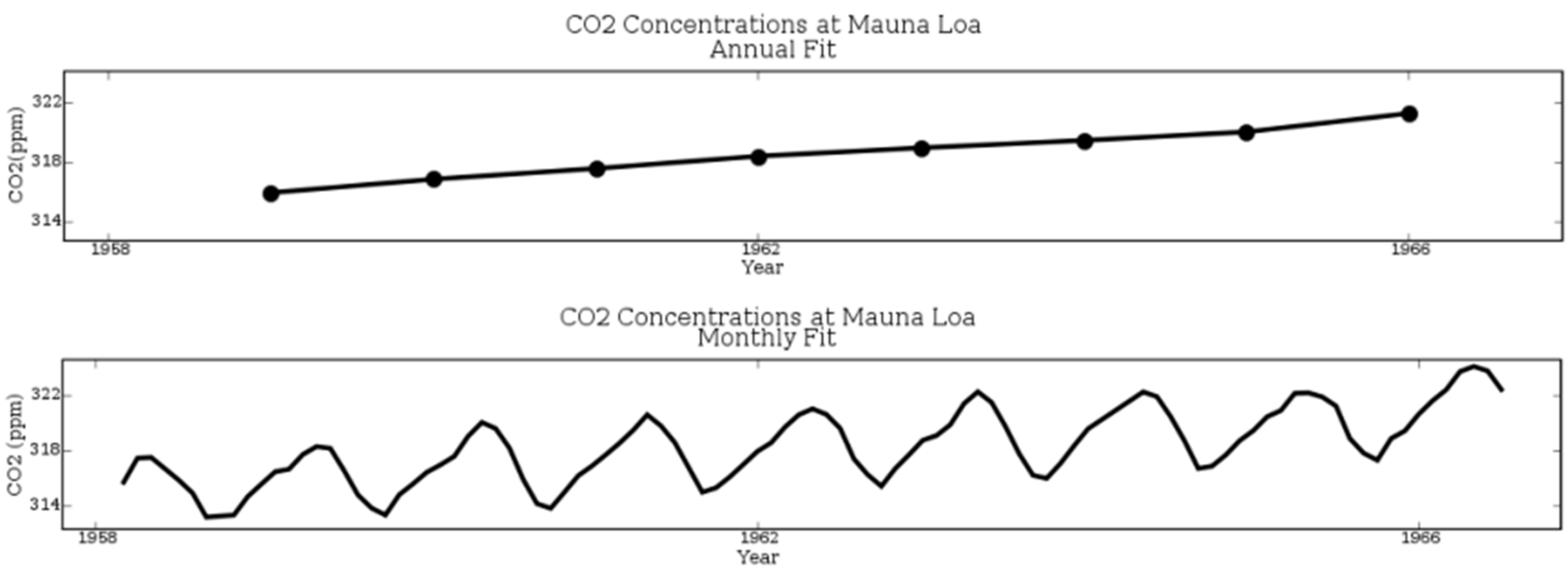
- Principle 4: Use log scales when appropriate
  - Used to show percentage change, multiplicative factors and skewness





Improving the Understanding

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

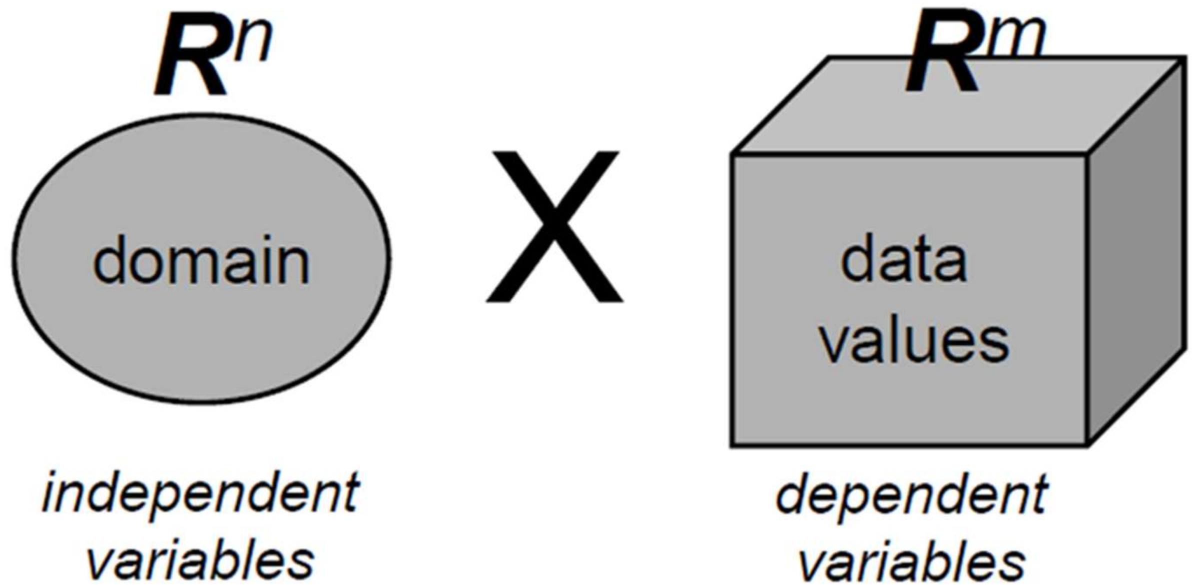


Summary of Principles

- 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^o

Data we are discussing



Source: VIS, University of Stuttgart

Scientific data	3**D**+time (*n*<4)	Scalar/vector/tensor
Information data	*n***D** (*n*>3)	Heterogeneous