Science Data Visualization & Formatting for Publication

John D'Ignazio
School of Information Studies
Syracuse University

IST 400/600 Spring 2008

Outline

- Data Formatting for Publication
- Data Visualization
- Visualization Techniques
- Visualization Tools
- Examples from your projects
- Attempts of our own

After analysis: go for the glory

- Formatting for publication
  - Requirements, Standards
  - Set by publication or linked repository
  - Based on community preservation practices and publication needs
- Tables for print/Electronic files
- Enabling tools

Example from the field

- *Science* Magazine / www.sciencemag.org
  - Published by the American Association for the Advancement of Science
  - AAAS serves some 262 affiliated societies and academies of science, serving 10 million individuals
  - the largest paid circulation of any peer-reviewed general science journal in the world, est. readership of 1 Million

Science submission policy

- large data sets, including:
  - microarray data
  - protein or DNA sequences
  - atomic coordinates or electron microscopy maps for macromolecular structures
- deposited
  - in an approved database
  - an accession number provided for inclusion
- ***"before publication"***

Science approved databases

- Molecular structure data:
  - Worldwide Protein Data Bank [through the Research Collaboratory for Structural Bioinformatics, Macromolecular Structure Database (MSD EMBL-EBI), or Protein Data Bank Japan]
  - BioMag Res Bank
  - Electron Microscopy Data Bank (MSD-EBI)
  - and for synthetic molecules, the Cambridge Crystallographic Data Center
  - Other databases supporting DNA or microarray data

Science – what if there is no established repository?

- must be housed as supporting online material
  - Supporting tables – data tables used to assess the paper's arguments
  - Supporting figures – figures that can't be printed but are integral to the paper
  - linked database presentations more complex than a flat text file or table (need an editorial consult):
    - hyperlinked to public sequence, array, or protein databases
    - collections of hypertext tables or Excel files linked to explanatory image files or tables

Science – allowed formats

- Essential article components
  - materials and methods, supporting text, supporting figures, supporting tables, supporting references and notes
  - All put in a Microsoft Word-compatible file
  - Excel spreadsheets/figures embedded directly in file
  - LaTex users submit one file too – in EPS or PDF
- The takeaway!
  - To help get that article published, put your data in an community repository, hosted on the Web, with a DOI
  - For a journal, the text is still most important: data tables and figures serve the text.

What is LaTex? www.tug.org

- A computer program created by Donald Knuth for typesetting documents
  - Suited for production of long articles and books
  - Built to handle tables and mathematical equations
  - Produces a DVI file
- An example of tricky things it does well

```
\[ \begin{array}{lcr}
\mbox{First number} & x & 8 \\
\mbox{Second number} & y & 15 \\
\mbox{Sum} & x + y & 23 \\
\mbox{Difference} & x - y & -7 \\
\mbox{Product} & xy & 120 \end{array}\]
```

First number	x	8
Second number	y	15
Sum	$x+y$	23
Difference	$x-y$	-7
Product	xy	120

Science Data Visualization: the what

The inputs

- Data can be floating-point data, integer data, image data, and text data
- Format are various.
- Data dimensions (1-D, 2-D, 3-D or more)

The outputs

- A sweet, sexy, powerful graphic image!

Data Visualization

- Technique often, but not necessarily, dependent on computer-based tools to help researchers understand and/or interpret data
- Science is a specific, innovative area in visualization – also borrows the same or similar techniques used in other domains
- Frames a problem, allows investigation with the data, goal is to afford interaction and communication of models/observed phenomenon.
- Closely tied to data analysis methods and techniques

Science Data Visualization: the why

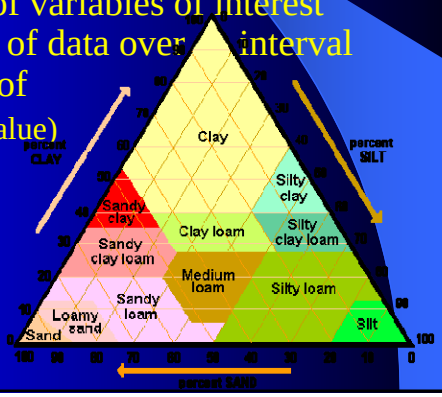
Because we're built that way

- Number of neocortical neurons males = 22.8 billion & females = 19.3 billion (Pakkenberg et al., 1997; 2003)
- Some estimates suggest that 40 percent of brain neurons are devoted to visual processing
- Gives EVERYONE the ability to "see" patterns, synthesize over multiple dimensions

Science Data Visualization: the why

Because of the power of design

- Project rhetorical arguments of science mission
- Clear visual statements of variables of interest where changes in values of data over [illegible] interval are reflected in changes of
  - Color (hue, brightness, value)
  - Shape
  - Contrast
  - Motion



Techniques & Technologies

- Plotting (used in data analysis)
- Mapping (used in graphics)
- Color image interpreting (used in image processing)
- Volume rendering (used in volume visualization)
- Graphics (Glut, OpenGL, …)
- Animation
- Virtual reality (CaveLib, openGL, …)
- Internet & database management software

Science Data Visualization: the how

Because of the power of design

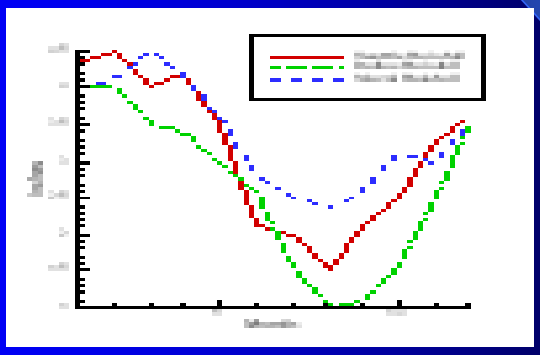
- data visualization solutions can be developed
  - based on specific needs
  - according to community practices
- Understanding relationship between data and tools is an important step in deploying specific data into and out of visualization environment
  - Formats
  - File sizes, iterations, versions
  - Speed of production

Data Visualization Tools

- Many to choose from, but disciplines/communities/researchers have their preferences
- Some of them are built inside data analysis packages (toolboxes)
  - mathematica (www.wolfram.com/),
  - Matlab, (www.mathworks.com/)
- commercial or individual visualization packages
  - Tecplot 9.0 from AMTEC (www.amtec.com)
- Social Networks & Internet resources
  - flowingdata.com / many-eyes.com

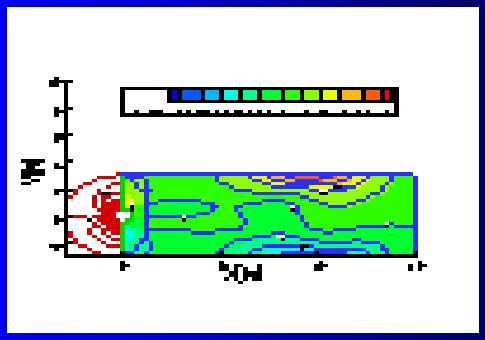
Basic TecPlot guide

- **Creating an X-Y Plot**



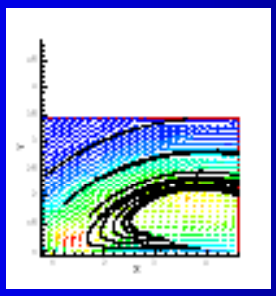
Basic TecPlot guide

- **Creating a Contour Plot with Structured Data**



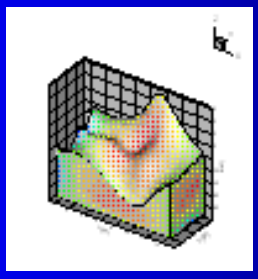
Basic TecPlot guide

- **Creating a Vector Plot**



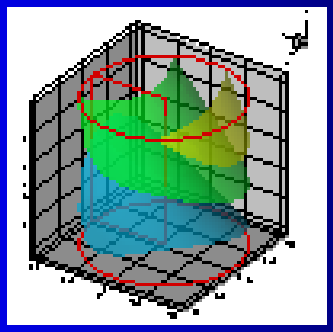
Basic TecPlot guide

- **Creating a Shaded Contour Plot**



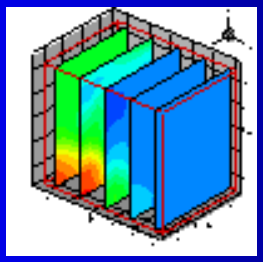
Basic TecPlot guide

- **Creating an Iso-Surface Plot in Tecplot 9.0**



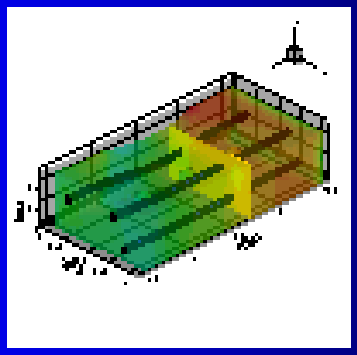
Basic TecPlot guide

- **Creating a Slice Plot in Tecplot 9.0**



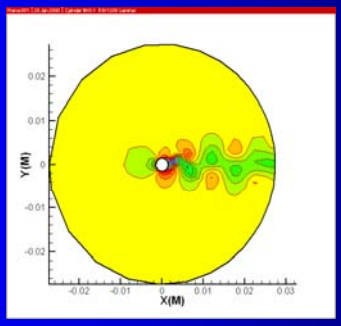
Basic TecPlot guide

- **Creating a Streamtrace Plot in Tecplot 9.0**



Basic TecPlot guide

- **Creating an Animation**



Graphical Integrity

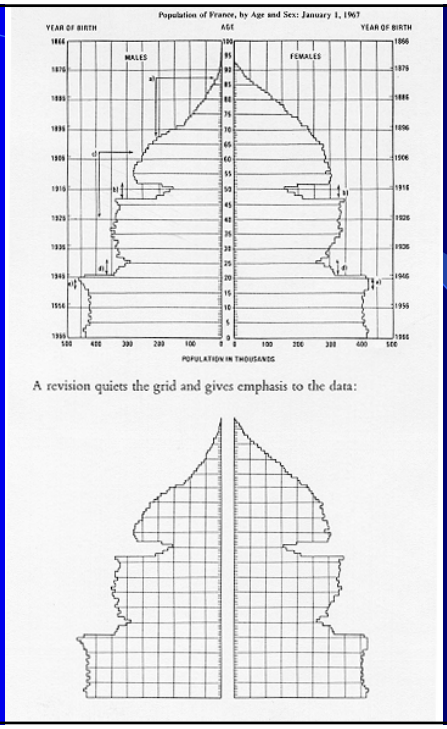
- Numbers should be proportional to numeric quantities
- Clear, detailed, and thorough labeling
- Show data variation, not design variation
- Show money in deflated units -- in time series
- Quote data in context

Data - Ink

- Above all else, show the data
- Data - Ink -- nonerasable core of a graphic, nonredundant ink
- Data-ink ratio = data-ink/total ink used to print the graphic

Chart-Junk

- All the extra stuff
  - moire vibration
  - grids
  - duck (the whole structure is decoration)

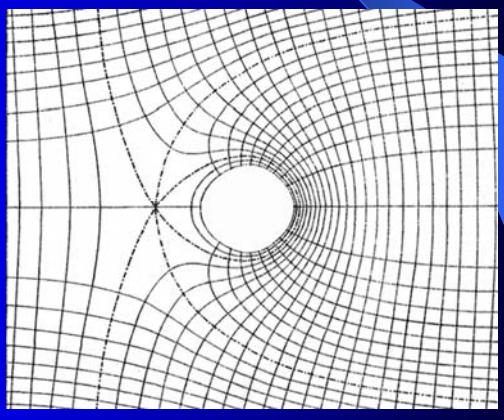


Complex Information

- Showing complexity is hard work
- Need to show proper relationships among information levels
- Use second color
- Subtract weight
  - Finer lines
  - Gray it down
  - Delete altogether

Complex Information

- Depend/respect on community's use of forms:



Physics Project - Prof. Suter

- Put Data in Context: