Visualizing non-Spatial Data aka, Infovis

Mark Van Moer
Senior Visualization Programmer
Data Analysis and Visualization, NCSA
mvanmoer@illinois.edu
Overview

• Audience
  • Users/generators of non-spatial data
  • Beginning info-visualizers

• Topics
  • What is non-spatial data/infovis
  • Visualization options
  • Tools and training resources

• Goals
  • Know where to start looking for more detail
  • Common vocabulary
SciVis – Infovis Spectrum

SciVis

- Specific geometry and topology – meshes, coordinates, …
- Physical data – temp, pressure, charge densities, …
- Natural processes – weather, CFD, MD, …

Infovis

- Arbitrary spatial arrangement – network graphs, …
- Cultural data – text, census data, music, …
- Relationships – statistical, social, hierarchical, …
Technical Challenges

• No inherent spatial layout
  • Must be imposed
  • Lots of freedom (too much?)
• Reducing multi-dimensional data to 2D, 3D
• Mixed data types
  • Text, numbers, continuous, categorical
• Text issues
  • mix-and-match character sets, Unicode, OCR errors
Design Challenges

• Substance and style
  • Substance driven by a research question
    • Is it useful?
    • Does it clarify or just looks cool?
  • Aesthetics can matter
    • If unappealing, no second looks
    • Graphic design to simplify and clarify
Seven Stages of Visualizing Data (Fry 2008)

- Not strict divisions, interaction among stages
- Rough guide for modularizing for scripts, etc.
- (Acquisition almost always done by PI’s team)
Parsing/Preprocessing

• Best case: file maps directly to representation
  • but don’t be constrained by this
• Analysis output format may not make sense for vis, requiring additional preprocessing.
  • E.g., some networks can be pre-computed making less work for the vis tools.
  • Drawback – might lose ability to interactively query data
• Structured formats provide easy ingestion
  • Tables, XML, JSON, etc.
Some Parsing Tools

- If data is table-like:
  - Tableau, Excel importer
  - Python csv module
    - https://pymotw.com/2/csv/index.html#module-csv

- Unstructured data, semi-structured
  - R tidytext tools
    - https://cran.r-project.org/web/packages/tidytext/index.html
  - Unix command line tools
      - Advice: Capture one-liners in a shell script
  - Python XML, JSON modules
Parsing Case Study

- Comparative Text Mining (Zhai 2004)
- Outputs flat text file
  - Mix of document and term probabilities/weights
- Two hierarchies
  - Arbitrary branching in each
- Original vis app parsed
  - Slow startup!

Solution: preprocess into XML
Parsing Case Study (cont’d)

Read directly into tree data structure.
Filter (make the haystack smaller)

- Data for vis often much smaller
- Cull everything you can get away with
  - Files with meta data sometimes can select which vars to load
- Reduce memory requirements
  - Interactive vis should fit in RAM or VRAM
- Finite number of pixels
  - 8K res is ≈ 33.2 megapixels
Mine (characterize needle-ness)

- Explicitly identify attributes for emphasis
  - Bring out the needle!
- Hard to comprehend hundreds of objects let alone millions
- Stats help identify possible areas of interest
Represent

Decide on basic representation

- Also see (Yau 2010, 2013)

- Standard charts: scatter, bar, line, histogram, etc.
  - Great for showing stats, distributions, etc.

- Network graphs
  - Show relationships, but can clutter quickly

- Trees, Treemaps
  - Hierarchical structures
Scatter Plot

- “go to” basic plot for 2, 3 variables
- Can compare continuous or categorical

Anscombe’s Quartet:
- Same means of x’s
- Same means of y’s
- Same regression lines
- Same x,y correlations
- And more!

Anscombe’s Quartet (Anscombe 1973)
Bar Chart, Histogram
Line graph

Champaign County Elevation Profile, South-North
Aspect ratio!

- What’s appropriate?
  - Context
  - Bank to 45° (Cleveland 1994)
  - Sparklines (Tufte 2006)
Network Graph  (Mallet case study)
Adjacency Matrix/Heat Map (Mallet cont’d)
Wide-ranging network gallery: visualcomplexity.com
Treemap
Parallel Coordinates
Refine

• Graphic design and UI principles assist in engaging viewers
• Leverage perceptual hierarchy, pre-attentive cognition tasks (Mackinlay 1986, Ware 2004)
  position > length > angle/slope > area > density > saturation > hue > shape
• General rule of thumb, tone down colors
  • ColorBrewer
Refine (cont’d)

- Position
- Length
- Slope
- Area
- Density
- Saturation
- Hue
- Shape
Muted colors afford increased info density

Interact

• Exploration
• Real-time response required, may have to cull data even further
• Highly recommend system with built-in GUI support
  • from-scratch UI programming can be a tedious time sink (Processing, looking at you…)
• Mapping mouse coords into data structure
  • Might require search, run time cost depends on implementation
Infovis Tool Resources (non-comprehensive)

Starting from scratch, fairly clean, table-ISH data
  • Tableau training videos: https://www.tableau.com/learn/training

Starting from scratch, data needs cleaning
  • Text Mining with R (tidytext) http://tidytextmining.com/
  • R Graphics cookbook http://www.cookbook-r.com/Graphs/

Need web interactivity
  • D3 tutorials http://alignedleft.com/tutorials/d3
  • Britecharts, D3-based lib http://eventbrite.github.io/britecharts/index.html
  • VEGA https://vega.github.io/vega/examples
  • Google Charts https://developers.google.com/chart/
Infovis Tools (cont'd)

Already have a Python workflow

- Pyplot https://matplotlib.org/users/pyplot_tutorial.html
- Matplotlib https://www.labri.fr/perso/nrougier/teaching/matplotlib/

Want total freedom and to blue-sky vis metaphors

- Processing https://processing.org/tutorials/

Familiar with VTK vis pipeline model and C++

- VTK infovis video https://www.youtube.com/watch?v=cXPdKaUc0BQ

Networks

- Gephi
- Cytoscape
- GraphViz
ParaView, VisIt, IDL

• Support vis analytics on mesh data
  • less support for pure infovis workflows

• ParaView filters (GUI Help menu | Reader, Filter, and Writer Reference)
  • Compute Quartiles, Descriptive Statistics, Histogram, Multicorrelative Statistics, Plot Data, Plot Global Variables Over Time, Plot On Intersection Curves, Plot on Sorted Lines, Plot Over Line, Plot Selection Over Time, Principal Component Analysis, Scatter Plot, Table to Points, Table to Structured Grid, Temporal Statistics, Contingency Statistics, K Means, etc.

• VisIt plots and operators (GUI Help menu)
  • Plots: Curve, Histogram, MultiCurve, Parallel Coordinates, Scatter, Spreadsheet
  • Analysis Operators: Connected Components, DataBinning, DeferExpression, Flux, Lineout, ModelFit, StatisticalTrends

• IDL
  • http://www.idlcoyote.com/gallery/index.html
References


Any of Tufte’s books are worthwhile reading.

[FlowingData](http://www.flowingdata.com) has a wealth of information.