Science of Science Research and Tools
Tutorial #02 of 12

Dr. Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
http://info.slis.indiana.edu/~katy

With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duohon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Hanning Guo, Mark A. Price, Angela M. Zoss, and Scott Weingart

Invited by Robin M. Wagner, Ph.D., M.S.
Chief Reporting Branch, Division of Information Services
Office of Research Information Systems, Office of Extramural Research
Office of the Director, National Institutes of Health

Suite 4090, 6705 Rockledge Drive, Bethesda, MD 20892
10a-noon, July 7, 2010

12 Tutorials in 12 Days at NIH—Overview

1. Science of Science Research
2. Information Visualization
3. CIShell Powered Tools: Network Workbench and Science of Science Tool
4. Temporal Analysis—Burst Detection
5. Geospatial Analysis and Mapping
6. Topical Analysis & Mapping
7. Tree Analysis and Visualization
8. Network Analysis
9. Large Network Analysis
10. Using the Scholarly Database at IU
11. VIVO National Researcher Networking
12. Future Developments
[02] Information Visualization

- Introduction
- Designing Effective Visualizations
- Visualization Layers
- Visual Languages
- Promising Research Directions

Recommended Reading

- Information Visualization class at Indiana University, http://ella.slis.indiana.edu/~katy/S637-S10
Information Visualization - Definition

“Information Visualization is a process of transforming data and information that are not inherently spatial, into a visual form allowing the user to observe and understand the information.”
(Source: Gershon and Eick, First Symposium on Information Visualization)

Scientific Visualization

Information Visualization

Information Visualization – Potential

- Rooted in geography, scientific visualization.
- Not even 20 years old.
- Growing fast.
- Interdisciplinary nature: computer graphics, electronic engineering, information systems, geography, information science, …

Well designed visualizations …
- Provide an ability to comprehend huge amounts of data.
- Reduce search time and reveal relations otherwise not being noticed (perception of emergent properties).
- Often reveal things not only about the data but how the data was collected - errors and artifacts jump out.
- Facilitate hypothesis formulation.
- Are effective sources of communication.
Information Visualization – Why Now?

- Information explosion.
- Work is becoming more ‘knowledge-oriented’.
- Increasing computing power (doubles every 18 months - Moore’s Law).
- Decreasing cost of storage.
- Fast graphics processors.
- Larger hard disk sizes -> more information available quickly.
- High resolution color monitors.
- Alternative user interfaces Idesk, CAVE (2 hands, audio, 3D).
- Connectivity between systems is expanding rapidly.
- Increasing visual intelligence.
- There is a bad mismatch between computer displays and the human perceptual system and between computer controls and human motor functions.

Information Visualization – Conferences and Journal

- IEEE Symposium on Information Visualization
- International Conference on Information Visualization
- Conference on Visual Data Exploration and Analysis
- SIGGRAPH
- Conference on Human Factors in Computing Systems
- International Conference on Human-Computer Interaction
- Intelligent User Interfaces
- Network Science Conference
- Publications of the ACM include IEEE symposium and conference on IV, SIGGRAPH, SIGIR, SIGCHI
Information Visualization – Major Books

Readings in Information Visualization: Using Vision to Think
by Stuart K. Card, Jock D. MacKinlay, Ben Shneiderman, 1999
Information Visualization: Perception for Design
by Colin Ware, 1999
Information Visualisation and Virtual Environments
by Chaomei Chen, Nov 1999
Information Visualization
http://www.ee.ic.ac.uk/research/information/www/Bobs.html

Mapping Cyberspace
by Martin Dodge and Rob Kitchin, 2000
http://www.mappingcyberspace.com/

The Craft of Information Visualization: Readings and Reflections
by Benjamin B. Bederson, Ben Shneiderman, 2003

More are listed on http://ella.slis.indiana.edu/~katy/S637-S09

Information Visualization – Recent Books

Beautiful Data: The Stories Behind Elegant Data Displays
by Nathan Yau
$40.09

Now You See It: Simple Visualization Techniques to Demonstrate Complex Ideas
by Stephen Few
$29.70

Beautiful Mapping: Leading Thinkers Demonstrate... by Turner
Andrew
$36.57

by Dona M. Wong
$19.77

Data Flow 2: Visualizing Information in Graphic Design
by R. Kahan
$49.14

Diagrams: Innovative Solutions for Graphic Design
by Jessica Gleen
$26.40

Data Visualization by Alexander C. Teets
$55.06

Surveys: The Hidden Pattern Behind Every... by Albert Lusin Barabasi
$17.70

Show Me the Numbers: Designing Tables and Graphs
by Stephen Few
$29.70

Effective UI: The Art of Building Great User Experiences... by Eik
$20.69

Visual Thinking: for Design (Morgan Kaufmann Series... by Colin Ware
$33.33

Cognitive Surplus: Creativity and Generosity in a... by Clay Shirky
$17.33

Information Dashboard Design: The Effective Visual... by Stephen Few
$23.09

Getting Started with Processing by Casey Reas
$13.99

Data Flow: Visualizing Information in Graphic Design... by Eik
$49.14
[02] Information Visualization

- Introduction
- Designing Effective Visualizations
- Visualization Layers
- Visual Languages
- Promising Research Directions
Designing Effective Visualizations

"The success of a visualization is based on deep knowledge and care about the substance, and the quality, relevance and integrity of the content." 
(Tufte, 1983)

Principle of Graphical Excellence

- Well-designed presentation of interesting data: substance, statistics, design.
- Complex ideas communicated with clarity, precision, and efficiency.
- Conveying the most knowledge in the shortest time with the least ink in the smallest space.
- It requires telling the truth about the data.
- It is nearly always multivariate.

(Tufte, 1983)

Five Principles in the Theory of Graphic Display

- Above all else show the data.
- Maximize the data-ink ratio, within reason.
- Erase non-data ink, within reason.
- Erase redundant data-ink.
- Revise and edit.

Visualizations should strive towards the following goals

- Focus on content of data not the visualization technique.
- Strive for integrity.
- Utilize classic designs and concepts proven by time.
- Comparative rather than descriptive visualizations.
- High resolution.

(Tufte, 1983)
Aesthetics

- Properly choose format and design
- Use words, numbers, drawings in close proximity
- Use lines of different weights as an attractive and compact way to display data.
- Reflect a balance, a proportion, a sense of relevant scale.
- Display an accessible complexity of detail.
- Let the graphics tell a story about the data.
- Avoid content-free decoration.
- Make use of symmetry to add beauty (although someone once said that "all true beauty requires some degree of asymmetry").
- Draw graphics an a professional manner, with the technical details of production done with care.

(Tufte, 1983)

Labeling

- Words spelled out.
- Words run left to right.
- Little messages explain data.
- Labels on the graphic; no legend needed.
- Graphic provokes curiosity.
- Blue contrasted with others.
- Clear, precise, modest type.
- Type is mixed case, with serifs

(Tufte, 1983)
User Needs Driven Approach: General Tasks

Visualization can help to identify

- Trends in the data.
- Outliers.
- Jumps in the data (gaps).
- Maxima and minima like largest, smallest, most recent, oldest, etc.
- Boundaries (not the same as maxima or jumps).
- Clusters in the data.
- Structure in heterogeneous information.
- A particular item of interest within the context of an enormous amount of contextual data.

Each of these tasks requires a different visualization design!

Visual Encoding of Data (e.g., in a network)

- What data entities should be represented as nodes?
- What nodes are important?
- What relationships are important and should be represented as edges?
- What node/edge attributes are important and need to be encoded?
- What subset of nodes, edges, subgraphs need to be labeled and how?
- Are there aggregate attributes, e.g., clusters, that need to be communicated?
- Is there a temporal, geospatial, or semantic substrate that should augment the layout of nodes?
- Are there any existing metaphors that can guide the visual encoding of nodes, edges, and their attributes?

- How large is the network? What data can be omitted to provide users with a meaningful overview of the dataset?
Images and Words

- Words (mathematical symbols, natural language, music) are better for representing procedural information, logical conditions, abstract verbal concepts (freedom).
- Images (graphics, abstract & figurative imagery) are better for spatial structures, location, detail.
- Animation brings graphics closer to words in expressive capacity (causality, disassembly).

Images and words can be linked via
- Proximity
- Continuity/connectedness
- Common region
- Combinations thereof

Rules of thumb to integrate words and images:
- In written text - give text first then link to image.
- Highlight relevant part of info just before the start of relevant speech segment.
- Move viewpoint in visualization to draw attention to different features. Cinematography: Static scenes 'go dead' visually after a few glances.

#02 Information Visualization

- Introduction
- Designing Effective Visualizations
- Visualization Layers
- Visual Languages
- Promising Research Directions
**Needs-Driven Workflow Design** using a modular data acquisition/analysis/modeling/visualization pipeline as well as modular visualization layers.

- **Deployment** of results is enabled through paper printouts, online animations, or interactive, three-dimensional, audiovisual environments.
- The **Legend Design** delivers guidance on the purpose, generation, and visual encoding of the data. Mapmakers should proudly sign their visualizations, adding credibility as well as contact information.
- In many cases, it is desirable to **Interact** with the data, that is, to zoom, pan, filter, search, and request details on demand. Selecting a data entity in one view might highlight this entity in other views.
- Sometimes it is beneficial to show multiple simultaneous views of the data, here referred to as **Combination**.
- Frequently, **Aggregation/Clustering** techniques are applied to identify data entities with common attribute values or dense connectivity patterns.
- **Graphic Design** refers to the visual encoding of data attributes using qualities such as size, color, and shape coding of nodes, linkages, or surface areas.
- Placing the **Raw Data** in a reference system reveals spatial patterns.
- **Projections/Distortions** of the reference system help emphasize certain areas or provide focus and context.
- **Reference Systems** organize the space.
**Reference System:** organizing the display space.

- Use known reference systems as much as possible.
- Provide overview map if space is large.
- Indicate user location and direction of view in map.
- Provide imagery of key landmarks and discrete but separately identifiable objects—there must be enough landmarks/objects that several are always visible at any instant.
- Strong visual cues indicating paths and regions help users understand structure of a space. Borders, boundaries and gridlines significantly improve navigation performance.

---

**Projections/Distortions:** Emphasize certain areas or provide focus and context.

Many (cartographic) projections exist. Projections are chosen such that distortions are minimized in accordance with map purpose.

**Distortion** techniques such as equal-area cartograms (see below) are widely used for distorting the surface areas of countries according to given variables (for example, number of papers published). Given our familiarity with the world or U.S. map, these maps can be easily interpreted despite their distortion. Polar coordinates and hyperbolic spaces are sometimes used to provide focus and context.

**Projections/Distortions**: Emphasize certain areas or provide focus and context.

Polar coordinates (left) and hyperbolic spaces (right) are used to provide focus and context.

---

Raw Data: Reveal spatial patterns.

Density patterns and outliers may become visible, but data records having identical coordinates will appear as one data point.

http://www.mzande.net/~zande/statistiek/stat-online

VIVO User Activity, see Tutorial 11
**Graphic Design:** Visually encoding data attributes using qualities such as size, color, and shape coding of nodes, linkages, or surface areas.

Most data records have multiple attributes, which can be represented by size-, color-, and shape-coding. Size-coding is made with the same coordinates; however, different attribute values make multiple records visible. Textual labels for major graphical elements help interpret a map. Landmarks ease navigation and exploration.

**Aggregation/Clustering:** Identify data entities with common attribute values or dense connectivity patterns.

- High-density areas and limited screen/paper space often mandate the grouping of records into higher-level structures.
- For example, authors can be grouped by their geographic location or institution.
- Semantic spaces are often split into topic areas or network communities.
- Cluster boundaries can help to visually separate them.
- Network layouts often benefit from the identification of communities using betweenness centrality clustering, and the highlighting of backbone structures is calculated using pathfinder network scaling.
Combination: Show multiple simultaneous views of the data.

It is often beneficial to examine a data set from different perspectives—using multiple, coupled windows. For example, to look at the growth of a nation it might be beneficial to examine a geographic map of exported goods and a science map of federal funding with resulting patents.

Small multiples are graphical depictions of different attributes of a data set using the identical reference system—for example, a scatterplot. They can be examined within a user’s eye span to support comparisons.

Interact: Zoom, pan, filter, search, request details on demand. Selecting a data entity in one view might highlight this entity in other views.

Often, data is too vast to be understood at once. Interaction via zooming and panning, exploration via brushing and linking, and access to details via search and selection are important. Ben Shneiderman’s visual information seeking mantra—“Overview first, zoom and filter, then details-on-demand”—summarizes the major visual design guidelines.

Principles of interaction design

- Mapping between data and their visual representation should be fluid and dynamic. -> Principle of transparency - 'the tools itself disappear’ (Rutkowsky, 1982).
- User obtains illusion of direct control.
- Provide visual feedback within 1/10 seconds (Shneiderman, 1987).
- Object constancy - use animation between displays instead of jumps.
Legend Design: Communicate purpose, generation, and visual encoding of the data.

No visualization is complete without information on what data is shown and how it was processed, by whom, and when. As more advanced data preprocessing and analysis algorithms are developed, it becomes necessary to educate viewers on the effect of parameters and visualization layer instantiation decisions, which add credibility and support interpretation. Mapmakers should proudly sign their visualizations, adding credibility as well as contact information.

Each visualization should have a
- Title
- Name of map maker
- Date of creation
- Explanation of all visual encodings, i.e., what do nodes, edges, colors, etc. represent?
- Information on dataset, dataset preparation, analysis.
- Short explanation of unique features and insights (if space permits).
- Web link(s) and/or reference(s) to additional information.

Deployment of results is enabled through paper printouts, online animations, or interactive, three-dimensional, audiovisual environments.

Static printouts
- High resolution of print
- No computer is in the way

Animations
- Show change over time

Interactive displays
- Zoom, pan, filter, details on demand
- Different simultaneous (coupled) views

Hands-on physical display
- Exploit spatial memory, touch sense

Hybrids
- Combine the best of different worlds – Illuminated Diagram
Illuminated Diagram Display

Questions:
➢ Who is doing research on what topic and where?
➢ What is the ‘footprint’ of interdisciplinary research fields?
➢ What impact have scientists?

Contributions:
➢ Interactive, high resolution interface to access and make sense of data about scholarly activity.

Large-scale, high resolution prints illuminated via projector or screen.

Interactive touch panel.

A single person’s spreading influence is shown as a series of four snapshots. First, we highlight topics and places relating to that person’s papers — papers that are well-cited today. The second light everything that cites that original work. Note that the first-generation impact extends to far more topics than did the original work. The third snapshot highlights scholars who cite the second, and the fourth, scholars that cites she cited.
纳米技术

探索科学学科的相互关联性

<table>
<thead>
<tr>
<th>科学学科</th>
<th>纳米技术</th>
</tr>
</thead>
<tbody>
<tr>
<td>物理学</td>
<td>无机化学科</td>
</tr>
<tr>
<td>材料科学</td>
<td>有机化学科</td>
</tr>
<tr>
<td>生物学</td>
<td>高分子材料科</td>
</tr>
<tr>
<td>化学</td>
<td>纳米光学科</td>
</tr>
</tbody>
</table>

探索某位学者的科学著作的影响力

<table>
<thead>
<tr>
<th>学者</th>
<th>作品</th>
</tr>
</thead>
<tbody>
<tr>
<td>弗朗西斯·克里克</td>
<td>DNA双螺旋模型的发现</td>
</tr>
<tr>
<td>阿尔伯特·爱因斯坦</td>
<td>相对论的创立</td>
</tr>
<tr>
<td>达尔文</td>
<td>《物种起源》</td>
</tr>
<tr>
<td>费朗克</td>
<td>量子力学的发展</td>
</tr>
</tbody>
</table>

纳米技术

纳米技术是与科学学科相互关联的学科，它与材料科学、物理学、化学、生物学等学科紧密相关。纳米技术的研究不仅限于实验室，其应用领域广泛，包括医学、电子学、能源、环境等领域。

Re-implementation of Illuminated Diagram Software
by Advanced Visualization Lab, Indiana University

Drives unlimited number of ID screens.

Touch screen for direct interaction.

Keyword and name search.

Selection of canned queries for:
- interdisciplinary research areas
- famous people
- activity patterns, e.g., bursts, trends, etc.
**Visual Languages**

Different sciences and arts use different visual encodings to communicate abstract data and concepts.

\[
\sum_{j=0}^{I_{\text{max}}} I_j = 0 \quad \sum_{v=0}^{V_{\text{max}}} V_v = 0 \quad V = 1 R \quad P = 1 V = 1 R = \frac{V^2}{R}
\]

\[
F = q \vee B_\perp = q \vee B = q \vee B \sin(\theta)
\]

\[
F = I L B_\perp = I L B = I L B \sin(\theta)
\]

\[
\sum_{\text{curv}} B_\parallel \Delta l = \mu_0 I_\perp
\]
# Exemplary Visual Encoding of Network Nodes and Edges

<table>
<thead>
<tr>
<th>Social (People, Institutions)</th>
<th>Three node symbols have same area size for same weight.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive (Terms, Papers, Patents, Journals)</td>
<td>Combinations of weighted+directed+dotted are possible.</td>
</tr>
<tr>
<td>Regulations (Funding, Laws)</td>
<td></td>
</tr>
</tbody>
</table>

- Undirected
- Directed
- Unweighted
- Weighted

- Direct link (citation)
- Co-occurrence (co-author, co-word)
- Co-citation (author CC, paper CC)

Time, geo, topic are attributes.
Use node/edge color coding for qualitative variables, e.g., type, gender, and area size coding for quantitative values, e.g., counts.

---

## [02] Information Visualization

- Introduction
- Designing Effective Visualizations
- Visualization Layers
- Visual Languages
- Promising Research Directions
Top Ten List of Challenges
Adapted from Chen 2002

1. Domain Specific vs. Domain Independent
2. Quality vs. Timeliness
3. Interdisciplinary Nature
4. Validation
5. Design Metaphor
6. Coverage
7. Scale-up
8. Automatic Labeling
9. Individual Differences
10. Ethical Constraints

http://www.llnl.gov/icc/sdd/img/images/Cr_Tiled_Small.mpg
Currently, diverse general tools (Excel, SPS, Pajek, etc.) and proprietary tools are used to study science and to gain science policy insight. The latter are patented, closed source, and rather expensive. Hence, most studies cannot be replicated due to price tags or legal issues.

A true science of science will benefit from tools that are:

- **Open source**—anybody can check and improve code.
- **Support many different data structures**—relevant static and streaming data comes in text or other format files, databases, RSS feeds.
- **Extensible**—new algorithms become available every day and it should be possible to integrate and use them.
- **Customizable**—different user groups have very different needs. It should be possible to quickly compile custom tools.
- **Scalable**—science is global and must be studied globally. Large scale datasets need to be processed using sufficient memory and processing power.
- **Workflow support**—different science studies require the application of many different algorithms and their parameter values in a specific sequence. It must be possible to log and share (ideally re-run) these workflows.

See Tutorial #3.

Please complete “Questionnaire #2” and “General Questionnaire.”