Information Visualization

Introduction

Dr. David Koop
What is Data Visualization?
Data Visualization
Data Visualization

[Rock 'N' Roll is Here to Pay, R. Garofalo, 1977 (via Tufte)]
What is **Information Visualization** (InfoVis)?
Compared to Statistical Graphics?
Streamgraphs

- Visualize movie ticket sales by time
- Stack films that are in theaters on top of each other
- Area = the total sales
- "You can see Oscar contenders attract a smaller audience than the holiday and summer blockbusters and kind of slowly build an audience." — N. Yau, FlowingData

[Byron and Wattenberg, 2012]
Streamgraphs

- [Gelman & Unwin] Instead use two plots:
  1. Total sales over time
  2. Trajectories for individual movies

- "Discussion burst out across the Web . . . that I am convinced would not have come about if instead of a Streamgraph, they used say, a stacked bar chart." — N. Yau

[Byron and Wattenberg, 2012]
"That Puzzle-Solving Feeling" — [Gelman & Unwin]
Nightigale's Coxcomb Diagram

The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent areas for death from Preventible or Mitigable Zymotic diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red, & the black lines in each.

[F. Nightingale, 1858]
Mortality rates in the Crimean War from April 1854 to March 1856

Sanitary commission arrives

[Diagram showing mortality rates with categories: Zymotic diseases, Wounds, injuries, All other causes]

British Army Size in the Crimean War from April 1854 to March 1856

[A bar chart showing average army size with years 1854-1856]

Gelman and Unwin's Remake

D. Koop, CSCI 628, Fall 2021

[Reference: Gelman and Unwin, 2014]
Compared to Infographics?
America's Most Popular Charts

[The Onion, 2007]
Decision Tree: The Obama-Clinton Divide

In the nominating contests so far, Senator Barack Obama has won the vast majority of counties with large black or highly educated populations. Senator Hillary Rodham Clinton has a commanding lead in less-educated counties dominated by whites. Follow the arrows for a more detailed split.

- Is a county more than 20 percent black?
  - NO There are not many African-Americans in this county.
  - YES This county has a large African-American population.

- And is the high school graduation rate higher than 78 percent?
  - NO This is a county with less-educated voters.
  - YES This is a county with more educated voters.

- And where is the county?
  - Northeast or South
  - West or Midwest

- Clinton wins these counties 704 to 89.
- Obama wins these counties 383 to 70.

- In 2006, were many households poor?
  - YES At least 47% earned less than $50,000.
  - NO At least 53% earned more than $50,000.

- What’s the population density?
  - Very rural: <61.5 people per sq. mile
  - Clinton wins these counties 52 to 25.

- In 2004, did Bush beat Kerry badly? (by more than 16.5 percentage points)
  - YES
    - Very Republican
    - Clinton wins these counties 48 to 13.
  - NO
    - Obama wins these counties 56 to 35.

Note: Chart excludes Florida and Michigan. County-level results are not available in Alaska, Hawaii, Kansas, Nebraska, New Mexico, North Dakota or Maine. Texas counties are included twice; once for primary voters and once for caucus participants.

Sources: Election results via The Associated Press, Census Bureau, Dave Leip’s Atlas of U.S. Presidential Elections

[A. Cox, NYTimes, 2008]
Infographics Embellish Boring Plots?

African Countries by GDP

South Africa
Egypt
Nigeria
Algeria
Morocco
Angola
Libya
Tunisia
Kenya
Ethiopia
Ghana
Cameroon

GDP in billions of US dollars

[Gelman & Unwin]
Infographics Embellish Boring Plots?

African Countries by GDP

- South Africa
- Egypt
- Nigeria
- Algeria
- Morocco
- Angola
- Libya
- Tunisia
- Kenya
- Ethiopia
- Ghana
- Cameroon

GDP in billions of US dollars

- South Africa: $285.4 b
- Egypt: $188.4 b
- Nigeria: $173 b
- Algeria: $140.6 b
- Morocco: $91.4 b
- Angola: $75.5 b
- Libya: $62.3 b
- Tunisia: $39.6 b
- Kenya: $29.4 b
- Ethiopia: $28.5 b
- Ghana: $26.2 b
- Cameroon: $22.2 b

Top Countries by GDP in U.S. $ Billions

GDP Calculation: Private consumption + gross investment + government spending + exports - imports

[Gelman & Unwin]

D. Koop, CSCI 628, Fall 2021
Compared to Scientific Visualization?
SciVis

[Google Image Search for "scientific visualization", 2017]
InfoVis

[Google Image Search for "information visualization", 2017]
SciVis → Fields

Scalar Fields
(Order-0 Tensor Fields)

Vector Fields
(Order-1 Tensor Fields)

Tensor Fields
(Order-2+)

Each point in space has an associated...

\[ s_0 \]
Scalar

\[
\begin{bmatrix}
  v_0 \\
  v_1 \\
  v_2
\end{bmatrix}
\]
Vector

\[
\begin{bmatrix}
  \sigma_{00} & \sigma_{01} & \sigma_{02} \\
  \sigma_{10} & \sigma_{11} & \sigma_{12} \\
  \sigma_{20} & \sigma_{21} & \sigma_{22}
\end{bmatrix}
\]
Tensor
## Visualization Taxonomy Structure

### Display Attributes

<table>
<thead>
<tr>
<th>Given</th>
<th>Constrained</th>
<th>Chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Continuous</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images (e.g., medical)</td>
<td>Distortions of given / continuous ideas (e.g., flattened medical structures, 2D geographic maps, fish-eye lens views)</td>
<td>Continuous (high-dimensional) mathematical functions</td>
</tr>
<tr>
<td>Fluid / gas flow, pressure distributions</td>
<td>Arrangement of numeric variable values</td>
<td>Continuous time-varying data, when time is mapped to a spatial dimension</td>
</tr>
<tr>
<td>Molecular structures (distributions of mass, charge, etc.)</td>
<td></td>
<td>Regression analyses</td>
</tr>
<tr>
<td>Globe – distribution data (e.g., elevation levels)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Discrete** | | |
| Classified data / images (e.g., segmented medical images) | Distortions of given / discrete ideas (e.g., 2D geographic maps, fish-eye lens views) | Discrete time-varying data, when time is mapped to a spatial dimension |
| Air traffic positions | Arrangement of ordinal or numeric variable values | Arbitrary entity-relationship data (e.g., file structures) |
| Molecular structures (exact positions of components) | | Arbitrary multi-dimensional data (e.g., employment statistics) |
| Globe – discrete entity data (e.g., city locations) | | |

*Table 1: High-level visualization taxonomy, illustrated by examples. Design models are classified based on whether they are discrete or continuous.*

- **Discrete** design models assume data cannot be interpolated, whereas continuous models assume data can be interpolated.
- Users develop user models as they obtain more information, which can be used to support or contradict hypotheses.
- Users may have preconceived ideas about the object of study, which can influence their choice of visualization techniques.
- Design models are developed and refined through interaction with visualization tools, allowing users to verify their ideas.
- Discrete and continuous models are classified based on the type of data used, with discrete models using entities that cannot be interpolated and continuous models using data that can be interpolated.

### 4.1 High Level Taxonomy Structure

- **4.1.1 Discrete / Continuous Classification**
  - **Discrete** models assume data cannot be interpolated, whereas continuous models assume data can be interpolated.
  - Design models are classified according to whether they are discrete or continuous.
  - Interval and ratio data can be interpolated, whereas discrete models assume data cannot be interpolated.
  - Continuous models assume data can be interpolated, whereas discrete models assume data cannot be interpolated.
  - Nominal and ordinal data can only be visualized by discrete models, whereas continuous models can be used for visualization.

*Source: Tory & Möller*
### Visualization Taxonomy Structure

#### SciVis

<table>
<thead>
<tr>
<th>Given</th>
<th>Display Attributes</th>
<th>Chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Images (e.g., medical)</td>
<td>Distortions of given / continuous ideas</td>
<td>Continuous (high-dimensional) mathematical functions</td>
</tr>
<tr>
<td>Fluid / gas flow, pressure</td>
<td>Distortions of given / continuous</td>
<td>Continuous time-varying data, when time is mapped to a spatial dimension</td>
</tr>
<tr>
<td>distributions</td>
<td>ideas (e.g., flattened medical structures, 2D geographic maps, fish-eye lens views)</td>
<td></td>
</tr>
<tr>
<td>Molecular structures (distributions of mass, charge, etc.)</td>
<td>Arrangement of numeric variable values</td>
<td>Regression analyses</td>
</tr>
<tr>
<td>Globe – distribution data (e.g., elevation levels)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Discrete

<table>
<thead>
<tr>
<th>Classified data / images (e.g., segmented medical images)</th>
<th>Distortions of given / discrete ideas (e.g., 2D geographic maps, fish-eye lens views)</th>
<th>Discrete time-varying data, when time is mapped to a spatial dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air traffic positions</td>
<td>Arrangement of ordinal or numeric variable values</td>
<td>Arbitrary entity-relationship data (e.g., file structures)</td>
</tr>
<tr>
<td>Molecular structures (exact positions of components)</td>
<td></td>
<td>Arbitrary multi-dimensional data (e.g., employment statistics)</td>
</tr>
<tr>
<td>Globe – discrete entity data (e.g., city locations)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

[Tory & Möller]
Visualization Taxonomy Structure

### Table 1: High-level visualization taxonomy, illustrated by examples.

<table>
<thead>
<tr>
<th><strong>Design Models</strong></th>
<th><strong>Given</strong></th>
<th><strong>Constrained</strong></th>
<th><strong>Chosen</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Images (e.g., medical)</td>
<td>Distortions of given / continuous ideas (e.g., flattened medical structures, 2D geographic maps, fish-eye lens views)</td>
<td>Continuous (high-dimensional) mathematical functions</td>
</tr>
<tr>
<td></td>
<td>Fluid / gas flow, pressure distributions</td>
<td>Arrangement of numeric variable values</td>
<td>Continuous time-varying data, when time is mapped to a spatial dimension</td>
</tr>
<tr>
<td></td>
<td>Molecular structures (distributions of mass, charge, etc.)</td>
<td></td>
<td>Regression analyses</td>
</tr>
<tr>
<td></td>
<td>Globe – distribution data (e.g., elevation levels)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>Classified data / images (e.g., segmented medical images)</td>
<td>Distortions of given / discrete ideas (e.g., 2D geographic maps, fish-eye lens views)</td>
<td>Discrete time-varying data, when time is mapped to a spatial dimension</td>
</tr>
<tr>
<td></td>
<td>Air traffic positions</td>
<td>Arrangement of ordinal or numeric variable values</td>
<td>Arbitrary entity-relationship data (e.g., file structures)</td>
</tr>
<tr>
<td></td>
<td>Molecular structures (exact positions of components)</td>
<td></td>
<td>Arbitrary multi-dimensional data (e.g., employment statistics)</td>
</tr>
<tr>
<td></td>
<td>Globe – discrete entity data (e.g., city locations)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**SciVis**

- **Display Attributes**
  - **Given**
    - Images (e.g., medical)
    - Fluid / gas flow, pressure distributions
    - Molecular structures (distributions of mass, charge, etc.)
    - Globe – distribution data (e.g., elevation levels)
  - **Constrained**
    - Distortions of given / continuous ideas (e.g., flattened medical structures, 2D geographic maps, fish-eye lens views)
    - Arrangement of numeric variable values
  - **Chosen**
    - Continuous (high-dimensional) mathematical functions
    - Continuous time-varying data, when time is mapped to a spatial dimension
    - Regression analyses

**InfoVis**

- **Display Attributes**
  - **Given**
    - Classified data / images (e.g., segmented medical images)
    - Air traffic positions
    - Molecular structures (exact positions of components)
    - Globe – discrete entity data (e.g., city locations)
  - **Constrained**
    - Distortions of given / discrete ideas (e.g., 2D geographic maps, fish-eye lens views)
    - Arrangement of ordinal or numeric variable values
  - **Chosen**
    - Discrete time-varying data, when time is mapped to a spatial dimension
    - Arbitrary entity-relationship data (e.g., file structures)
    - Arbitrary multi-dimensional data (e.g., employment statistics)
Kosara's Definition of Information Visualization

- **It is based on (non-visual) data.** "The data to be visualized must come from outside the program, and the program must be able (at least in principle) to work on different data sets. Also, visualization is not image processing or photography; if the source data is an image and is used as an image in the result, it is not being visualized."

- **It produces an image.** "Clearly, each visualization has the goal of producing one or more images from the data, and the visual must be the primary means of communicating the data. Other media can be part of a visualization, but the visualization must be able to stand on its own."
Kosara's Definition of Information Visualization

• **The result is readable and recognizable.** "There are many ways to transform data into images, most of which do not allow the viewer to understand the underlying data. A visualization must produce images that are readable by a viewer, even if that requires training and practice. Visualization images must also be recognizable as such, and not appear to be something else. The use of additional elements (or even “eye candy”) is certainly possible, but must not take precedence over the communication goals of the visualization."
4.1 The Sublime

One aesthetic criterion of particular interest is the sublime. The sublime can be understood as that which inspires awe, grandeur, and evokes a deep emotional and/or intellectual response. Works of art generally possess a sublime quality, making them enigmatic and captivating at the same time. Sack [16] equates its opposite, the anti-sublime, with user friendliness, which is a central concept in computer science. In fact, visualization is generally understood to be a part of human-centered computing [11], and techniques that are published at the main conferences and in journals usually need to be evaluated in user studies [13]. They are thus designed to remove any sublimity, and instead foster immediate understanding.

While the sublime is just one criterion in aesthetics, it is an incredibly useful one for this discussion. The data-based visualization examples discussed above and shown in Figure 1 can be easily classified using a measure of their sublimity: while the classical technical information visualization is entirely anti-sublime, artistic visualizations are primarily sublime.

4.2 Pragmatic Visualization

Pragmatic visualization is what we term the technical application of visualization techniques to analyze data. The goal of pragmatic visualization is to explore, analyze, or present information in a way that allows the user to thoroughly understand the data. Card et al. describe this process as knowledge crystallization [3], and the recent initiatives in visual analytics [19] have used the slogan Detecting the Expected, Discovering the UnexpectedTM.

Visual efficiency is of course a key criterion for work in visualization. The goal is to produce images that convey the data as quickly and effortlessly as possible. User studies are conducted to measure the speed and accuracy of users, and to compare different methods and tasks [13].

Pragmatic <-> Artistic Visualization

Parallel Sets, etc. Informative Art Music Visualization, etc.

Recognizable Readable

Not Recognizable Readable

Not Recognizable Not Readable

Pragmatic Visualization Utilitarian Artistic Visualization Sublime

[R. Kosara]
Rules

• We saw many rules in CSCI 627 (Data Visualization)
• How do we use those to think about visualization?
Visualization Mistakes

Original

**Left-click**
Average number of likes per Facebook post 2016

- **Jeremy Corbyn**: 5,210
- **Labour Party**: 750
- **Momentum**: 250
- **Owen Smith**: 750
- **Andy Burnham**: 500
- **Saving Labour**: 0

Source: Facebook
Visualization Mistakes

**Original**

**Left-click**
Average number of likes per Facebook post 2016

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>250</th>
<th>500</th>
<th>750</th>
<th>1,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeremy Corbyn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>5,210</strong></td>
</tr>
<tr>
<td>Labour Party</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Momentum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owen Smith</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andy Burnham</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saving Labour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Facebook

**Better**

**Left-click**
Average number of likes per Facebook post 2016, '000

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeremy Corbyn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>5,210</strong></td>
<td></td>
</tr>
<tr>
<td>Labour Party</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Momentum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Owen Smith</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Andy Burnham</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saving Labour</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Facebook

[S. Leo]
Can we break the rules?
### Benefitting InfoVis with Visual Difficulties

<table>
<thead>
<tr>
<th>Cognitive efficiency</th>
<th>Visual difficulties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimize the cognitive steps required to process visualization</td>
<td>Induce constructive, self-directed cognitive activity on the part of the user</td>
</tr>
<tr>
<td>Maximize the ratio of data to ink</td>
<td>Design representations that are most likely to engage a user to actively process the information</td>
</tr>
</tbody>
</table>

---

**Table 1: Efficiency versus Difficulties Recommendations in InfoVis.**

- **Cognitive Operations**
  - Minimize the cognitive steps required to process visualization

- **Data-ink Ratio**
  - Maximize the ratio of data to ink

---

**Figure 3: Guessing game version**

---

**[J. Hullman et al.]**
# Benefitting InfoVis with Visual Difficulties

<table>
<thead>
<tr>
<th>Organization</th>
<th>Choose the format which makes important information most visually salient</th>
<th>Choose the format that best stimulates deep cognitive reflection on the important data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animation</td>
<td>Use animation to quickly and intuitively visualize important processes</td>
<td>Use static representations to induce interval visualization processing around causal mechanisms; consider animation in cases where mental animation lies beyond users’ capacities</td>
</tr>
<tr>
<td>Labeling</td>
<td>Use labels rather than legends to optimize immediate clarity</td>
<td>Use legends to stimulate deeper reflection on data</td>
</tr>
</tbody>
</table>

---

*J. Hullman et al.*
Administrivia
About Me

• Research Interests
  - Visualization
  - Computational Provenance
  - Geospatial Analysis

• Research Projects
  - Dataflow Notebooks
  - Geospatial Trajectory Data
  - Provenance for Web Applications

• See my web page for more information
  - http://faculty.cs.niu.edu/~dakoop/
About You

• Research Papers?
• Data Visualization
• Tools? JavaScript, D3, Tableau, Others?
• Research Experience?
• What topics do you want to see covered?
About this course

• Course web page is authoritative:
  - http://faculty.cs.niu.edu/~dakoop/cs628-2021fa/
  - Schedule, Readings, Assignments will be posted online
  - Check the web site before emailing me

• Lectures: TuTh 9:30-10:45am in PM 252

• This is an Advanced (Tier 2) Graduate Course
  - Present and discuss cutting-edge topics
  - Work on research problems

• Requires **participation**: readings and discussions
About this Course

• "Focus on advanced theory and methods for manipulating and visualizing the data of non-physical systems… Emphasis on the advanced study of the latest information visualization techniques developed by the research community. A computer programming background is required. Extensive laboratory work."
About this course

• Course Registration:
  - Make sure you have registered for the course
  - Email me if you are not registered but are interested in taking the course

• Review of course policies:
  - Plagiarism and academic honesty
  - If you have any concerns or questions, please email me as soon as possible

• If you are not sure if this course is a good fit, please email me or talk to me
Office Hours & Email

• Office hours will be held in person
  - Tu: 1:45-3pm, Th: 10:45am-12pm, or by appointment
• Please adhere to university regulations (Protecting the Pack)
• You do not need an appointment to stop by during scheduled office hours
• If you wish to meet virtually, please schedule an appointment
• If you need an appointment, please email me with details about what you wish to discuss and times that would work for you
• Many questions can be answered via email. Please consider writing an email before scheduling a meeting.
Expectations

• Be engaged:
  - Active participation
  - Constructive participation
• Work independently: self-directed and sustained
• Work collaboratively: learn from each other
• Put effort into the course:
  - Must put significant work in each week
  - Do not try to do everything before a deadline
  - Best effort (success or failure)
Next Class

- Review of Core Visualization Topics
- Topics Survey
- Bring Your Ideas