The Graphs They Are a-Changin’
Principles, Examples, Software for Data Visualization

Constantin Manuel Bosancianu and Joost van Beek

Doctoral School and Center for Media and Communication Studies, Central European University

April 26, 2012
Plan

Things to speak about:

1. Basics of good data visualization;
2. “The good, the bad, and the ugly” when it comes to data visualization - examples;
3. Software (open-source, web-based...);
4. Discussion time.
Importance

There is more data than ever waiting to be analyzed, mined for patterns, summarized, or linked to other data.
Figure: Word birth and death.
(http://www.nature.com/srep/2012/120315/srep00313/full/srep00313.html)
**Figure:** Linking patterns between US political blogs
Figure: Immigrant clusters in Amsterdam
**Figure:** Income clusters in Rotterdam
Importance

We also observe a phenomenal level of growth in individual-level data: Internet, smartphones, automated sensors etc.
**Figure:** Stephen Wolfram’s outgoing e-mail (approximately 300,000)
Figure: Stephen Wolfram’s keystrokes (approximately 100 million)
Importance

Presenting this information in an accurate and intuitive way for the purpose of highlighting causal connections will be crucial for our ability to make adequate choices in a democracy.
Data visualization (DV)

- At the confluence between statistics and design, dealing with the search for the most effective and graphically intuitive way of making an argument on the basis of data.

- In 2000, an estimated 900 billion \((9 \times 10^{11})\) to 2 trillion \((2 \times 10^{12})\) graphs were generated every year (Tufte 2001).
Goals of DV

Multiple:

- Making an argument;
- Minimizing any distractions from the central argument;
- Ensuring the integrity of the argument;\(^1\)
- Summarizing a lot of information in a reduced space;
- Encouraging comparison.

\(^1\)“Making a presentation is a moral act as well as an intellectual activity.” (Tufte 2006, 141)
Principles of DV

- The overarching purpose is to show the data;
- Minimize the data-ink ratio, as much as possible;
- Erase non-data-ink, as much as possible;
- Minimize redundant data-ink, as much as possible;
- Revise and edit;
- Mobilize every graphical element needed.  

Adapted from Tufte (2001)
• **Apprehension:** Ability to correctly perceive relations among variables

• **Clarity:** Ability to visually distinguish all the elements of a graph

• **Consistency:** Ability to interpret a graph based on similarity to previous graphs
• **Efficiency**: Ability to portray a possibly complex relation in as simple a way as possible

• **Necessity**: The need for the graph, and the graphical elements

• **Truthfulness**: Ability to determined the true value represented by any graphical element by its magnitude relative to the implicit or explicit scale

---

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>.027*** (.005)</td>
<td>.031*** (.006)</td>
</tr>
<tr>
<td>Gender</td>
<td>.094 (.174)</td>
<td>.074 (.215)</td>
</tr>
<tr>
<td>Education</td>
<td>.191*** (.044)</td>
<td>.055 (.056)</td>
</tr>
<tr>
<td>Marital status</td>
<td>.135 (.181)</td>
<td>.095 (.222)</td>
</tr>
<tr>
<td>Mobilized</td>
<td>- (.117)</td>
<td>.049 (.150)</td>
</tr>
<tr>
<td>Political interest</td>
<td>- (.733*** (.150)</td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Estimates from a logistic regression model predicting likelihood of turnout (Sweden, EES 2009)
Figure: Estimates from the regression model in graphical form
Figure: Traditional boxplot
Figure: Quartile plot
2.1
Napoleon’s 1812-1813 Russian campaign - Charles Joseph Minard.
Figure: Campaign map
Figure: Alternative to the map
Figure: Alternative to the map
Napoleon's 1812-1813 Russian Campaign

- Died
- Survived

98%
2.2

The UK Budget - David McCandless.
The Billion Pound-O-Gram

David McCandless / InformationIsBeautiful.net

Source: UK Treasury, Guardian
2.3

Commuters in the US - SENSEable City Laboratory, MIT.
Figure: Commuters - July 2010, AT&T cell phone data
2.4

Welfare benefits in Ontario
Common Sense Revolution


(National Council of Welfare & the Toronto Disaster Relief Committee)

© Scott Spiili

The 500th death of a homeless person on the streets of Toronto was recorded in November 2006

Welfare income of $7007 in 2005

2.5
Web-based and interactive
The new frontier

- New York Times’ *Mapping America*
- Washington Post’s *Top Secret America*
- Wall Street Journal’s *What They Know*
- Harvard’s Berkman Center for Internet & Society
  *Mapping the Persian Blogosphere*
3.1

‘Chartjunk’
Figure: Prominent example
Figure: Prominent example
3.2

Misleading graphs
This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

**Figure:** First example

By the Numbers: Over 35 years, Cornell’s tuition has taken an increasingly larger share of its median student family income.

Pecking Order: Over 12 years, Cornell’s ranking in US News & World Report has risen and fallen erratically.
Figure: Third example
3.3

Poor understanding of statistics
Figure: First example
A Plateau of Happiness

A country’s wealth may not always dictate the happiness of its people. As part of the World Values Survey project, inhabitants of different countries and territories were asked how happy or satisfied they were. Below is a sampling of happiness rankings, along with economic status.

HAPPINESS SCORES FROM 2000 POLL

Many countries, particularly those in Latin America, had higher marks for happiness than their economic situation would predict.


*Poll results for these countries were from 1995.

3.4
Poor choice of graphical display
Figure: First example
Figure: Second example
Figure: Alternative to second example
Chart 2 - Total Expenditures on Health as a Percentage Share of GDP, by OECD Country, 2004

Note: For the United States the 2004 data reported here do not match the 2004 data point for the United States in Chart 1 since the OECD uses a slightly different definition of "total expenditures on health" than that used in the National Health Expenditure Accounts.

Figure: Third example
Expenditures on Health as Percentage of GDP for OECD Countries, 2004

Figure: Reworked graph
Tools

To cover in the remaining minutes:

- Gapminder;
- IBM’s Many Eyes;
- Web interface for ggplot2;
4.1

IBM’s Many Eyes
http://www-958.ibm.com/software/data/cognos/manyeyes/

A “shared visualization and discovery” service, still in experimental phase
4.2

Hans Rosling’s Gapminder project
**Figure:** Hans Rosling, Professor of International Health, Karolinska Institute, Stockholm, Sweden
The problem he identifies: there is an abundance of yearly indicators for phenomena, scattered in the public domain

Creates Gapminder Foundation and develops the Trendalyzer software (later sold to Google)

Recently: Gapminder Desktop
Gapminder

Google develops, on the basis of Trendalyzer, Google Public Data Explorer
(http://www.google.com/publicdata/directory)
4.3

Jeroen Ooms’ ggplot2 interface
• R package developed by Hadley Wickham, on the basis of Leland Wilkinson’s ideas regarding visualization (*The Grammar of Graphics*)

• Heavily code-based

• Jeroen Ooms adds a simple web-based interface to the package (other packages: IRT, lme4)
Honorable mentions

Still worthy to explore for a bit:

- Drillet (basic, but free)
- StatSilk (maps with indicators)
- GNU Octave (high-level interpreted language for numerical computations)
- IBM’s Many Bills (specialized) (http://manybills.researchlabs.ibm.com/)
- Wordle (word clouds)
Conclusion

Good data visualization involves thinking about the argument to be made, making choices among alternatives, and taking into consideration issues such as audience, parsimony, integrity. It will rarely result from canned routines and default options found in statistical packages.
Thank you!
Books used for ideas or graphs:

Internet sources where some of the graphs can be found:

- http://www.informationisbeautiful.net/ (David McCandless, UK)
- http://www.datavis.ca/gallery/index.php (Michael Friendly, York University)
- http://flowingdata.com/
- http://www.infosthetics.com/
- http://senseable.mit.edu/ (SENSEable City Laboratory, MIT)
- http://chartporn.org/2012/03/02/improving-on-minard/
- http://igraphicsexplained.blogspot.com/
Web-based software:

- Gapminder Desktop (http://www.gapminder.org/downloads/)
- IBM’s Many Eyes (http://www-958.ibm.com/software/data/cognos/manyeyes/)
- Jeroen Ooms’ ggplot2 interface (http://rweb.stat.ucla.edu/ggplot2/)
- StatSilk (http://www.statsilk.com/)
- Wordle (http://www.wordle.net/)