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.

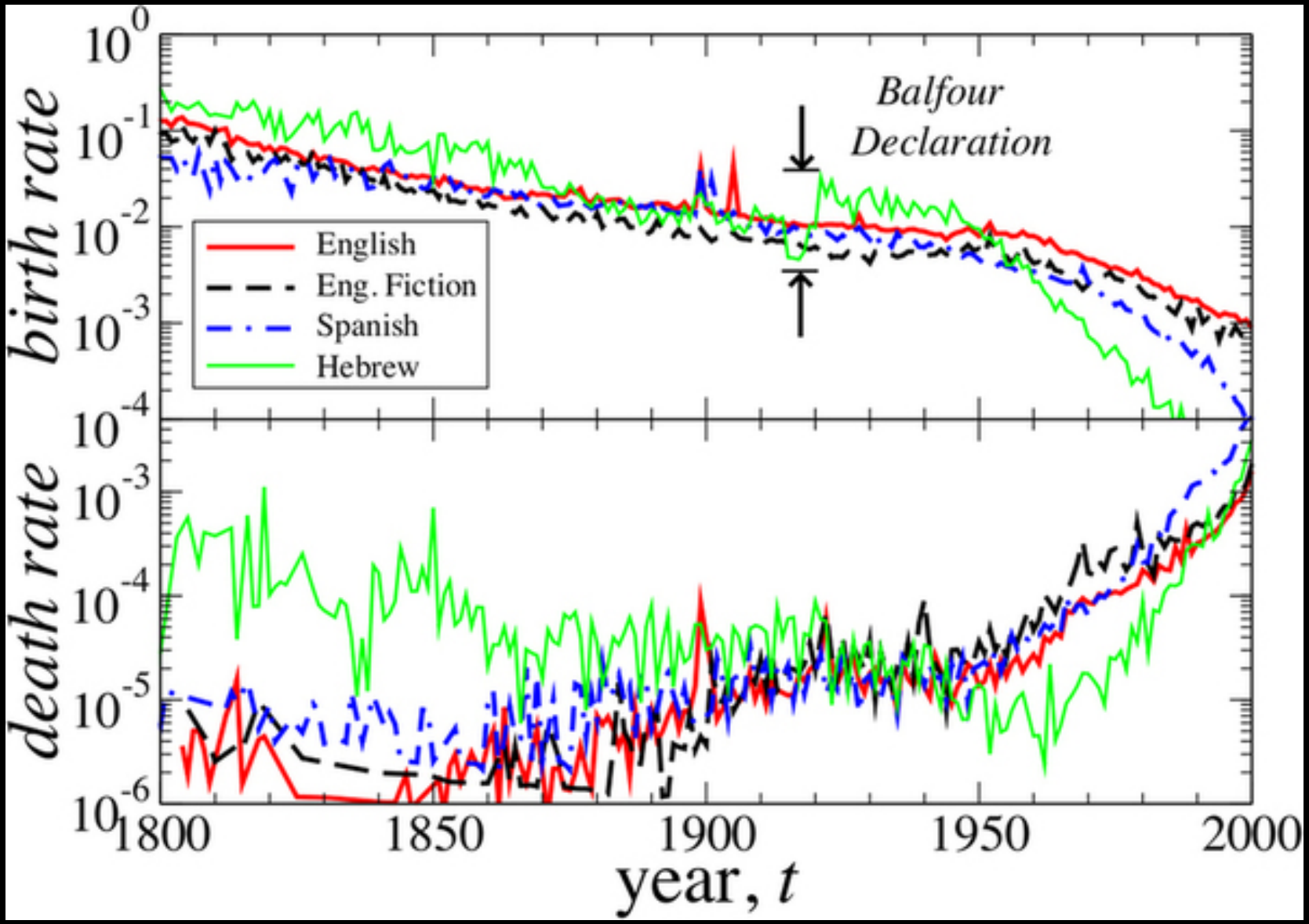


Figure: Word birth and death.
(http://www.nature.com/srep/2012/120315/srep00313/full/srep00313.html)

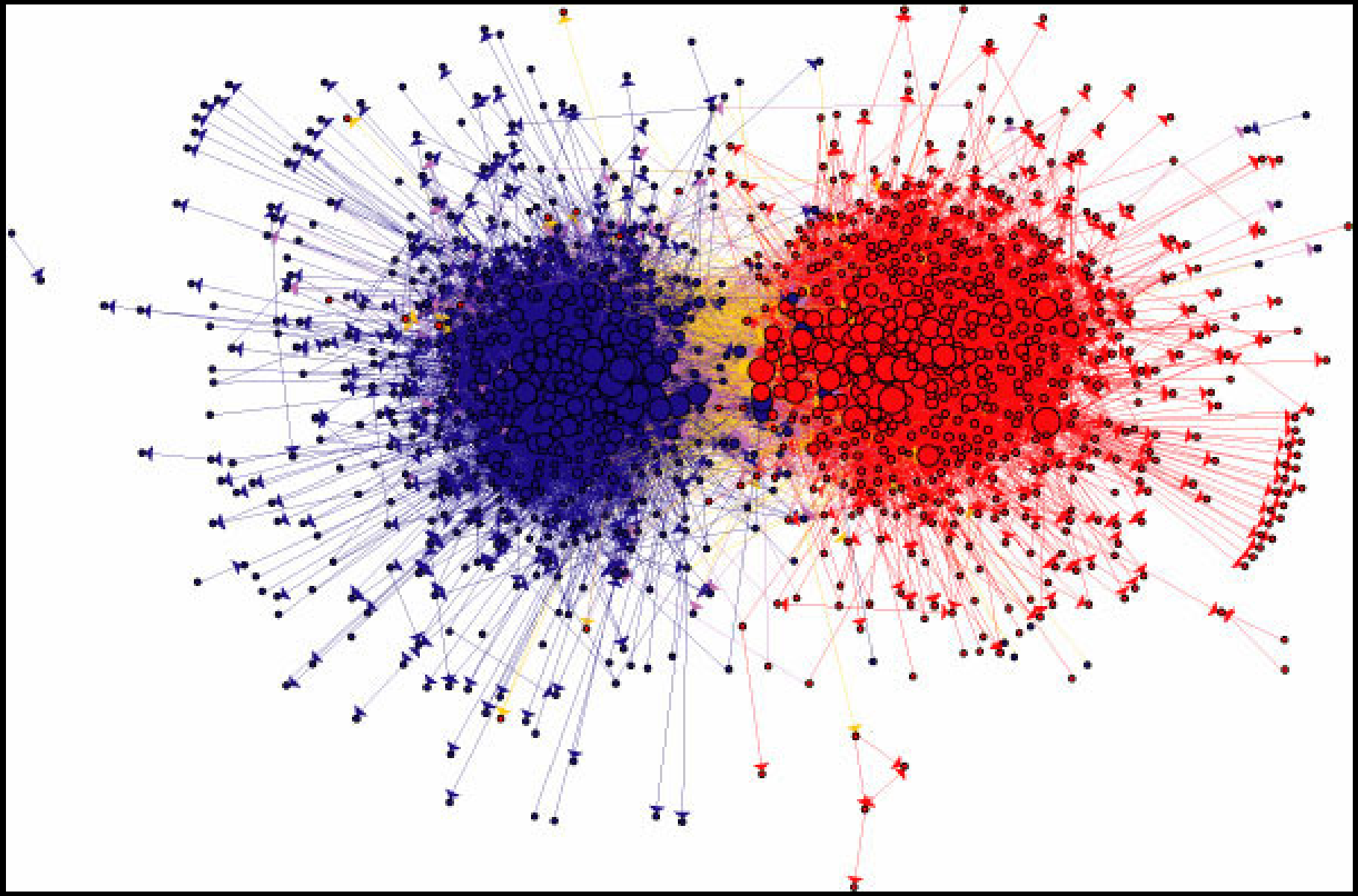


Figure: Linking patterns between US political blogs

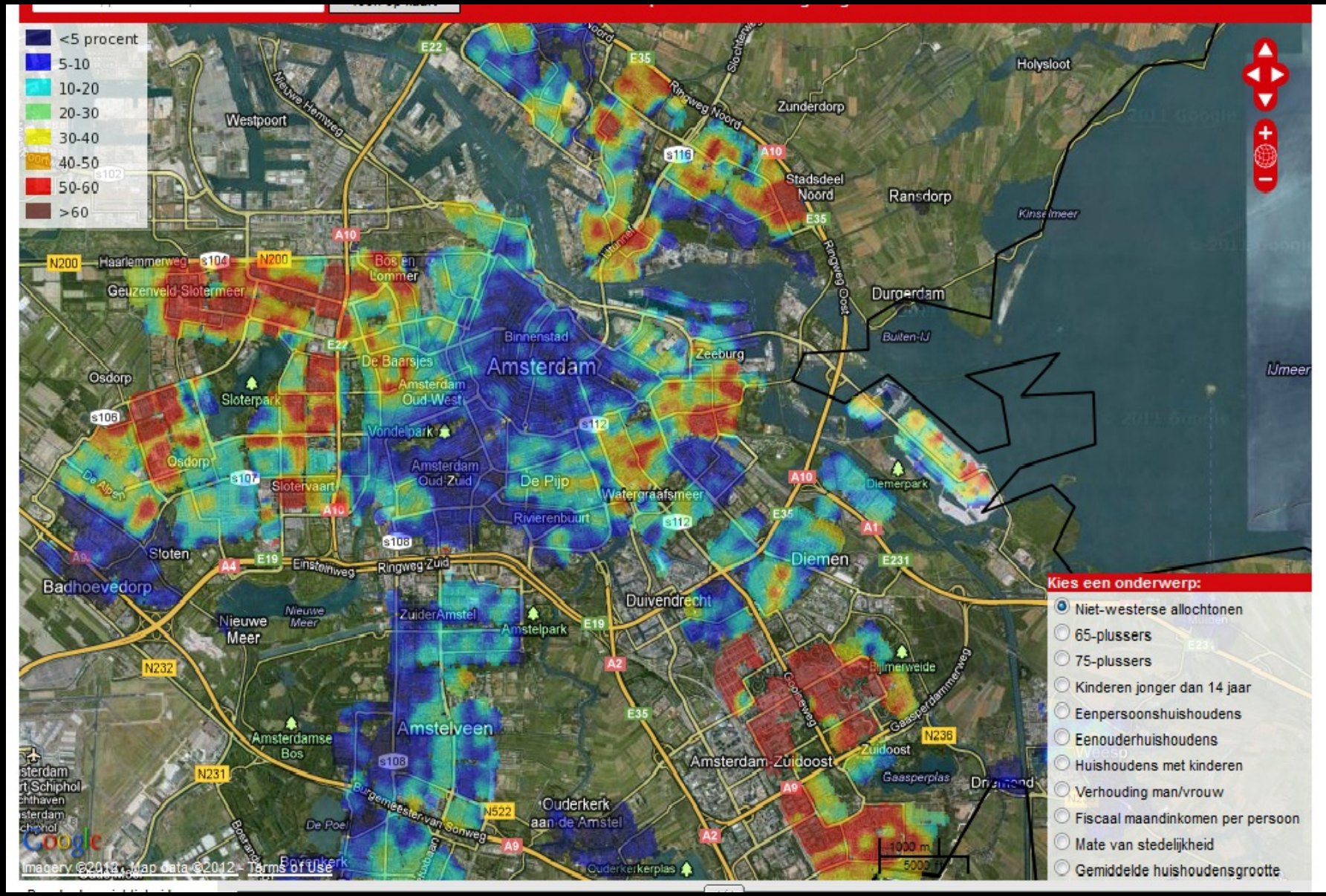


Figure: Immigrant clusters in Amsterdam

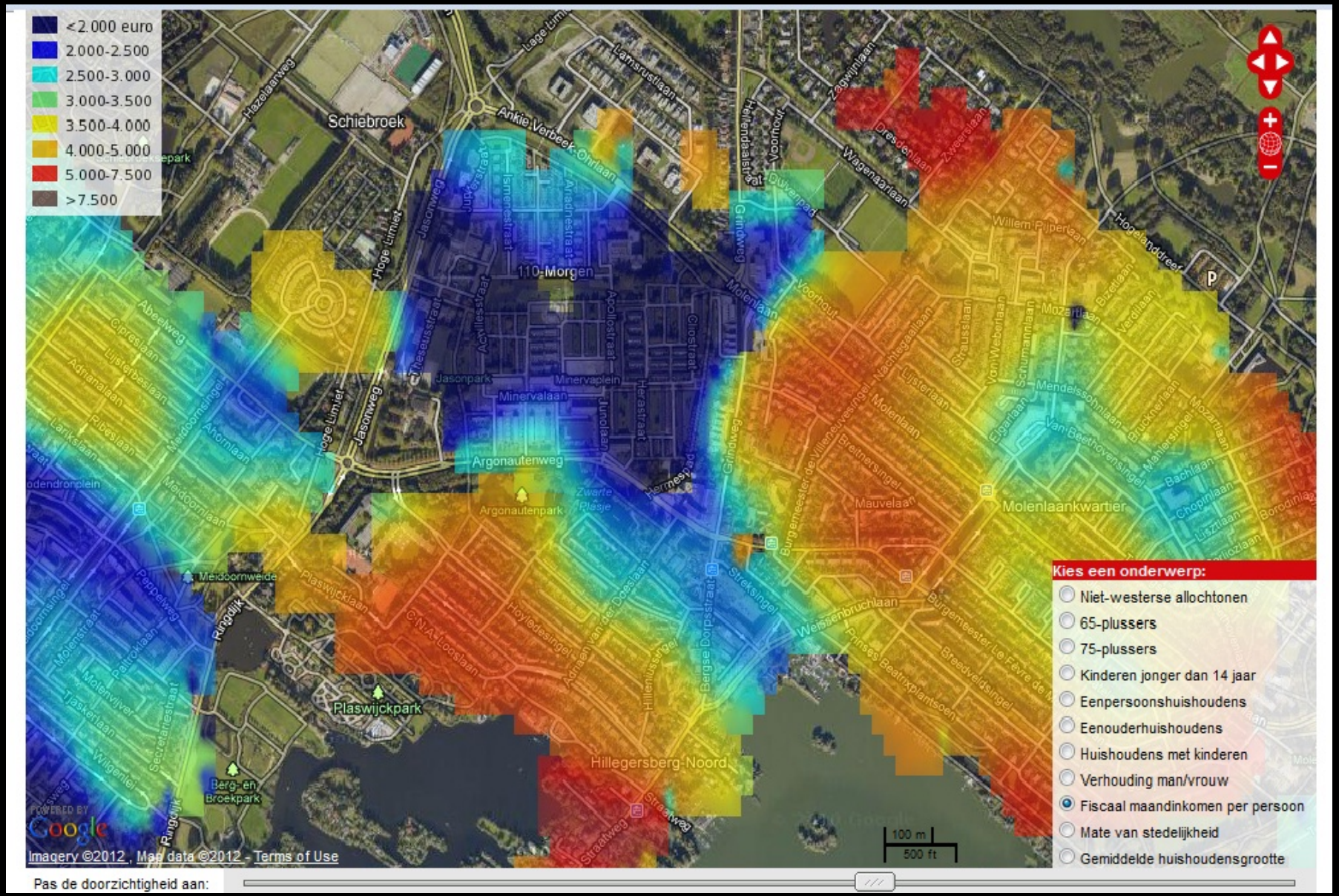


Figure: Income clusters in Rotterdam

Importance

We also observe a phenomenal level of growth in individual-level data: Internet, smartphones, automated sensors etc.

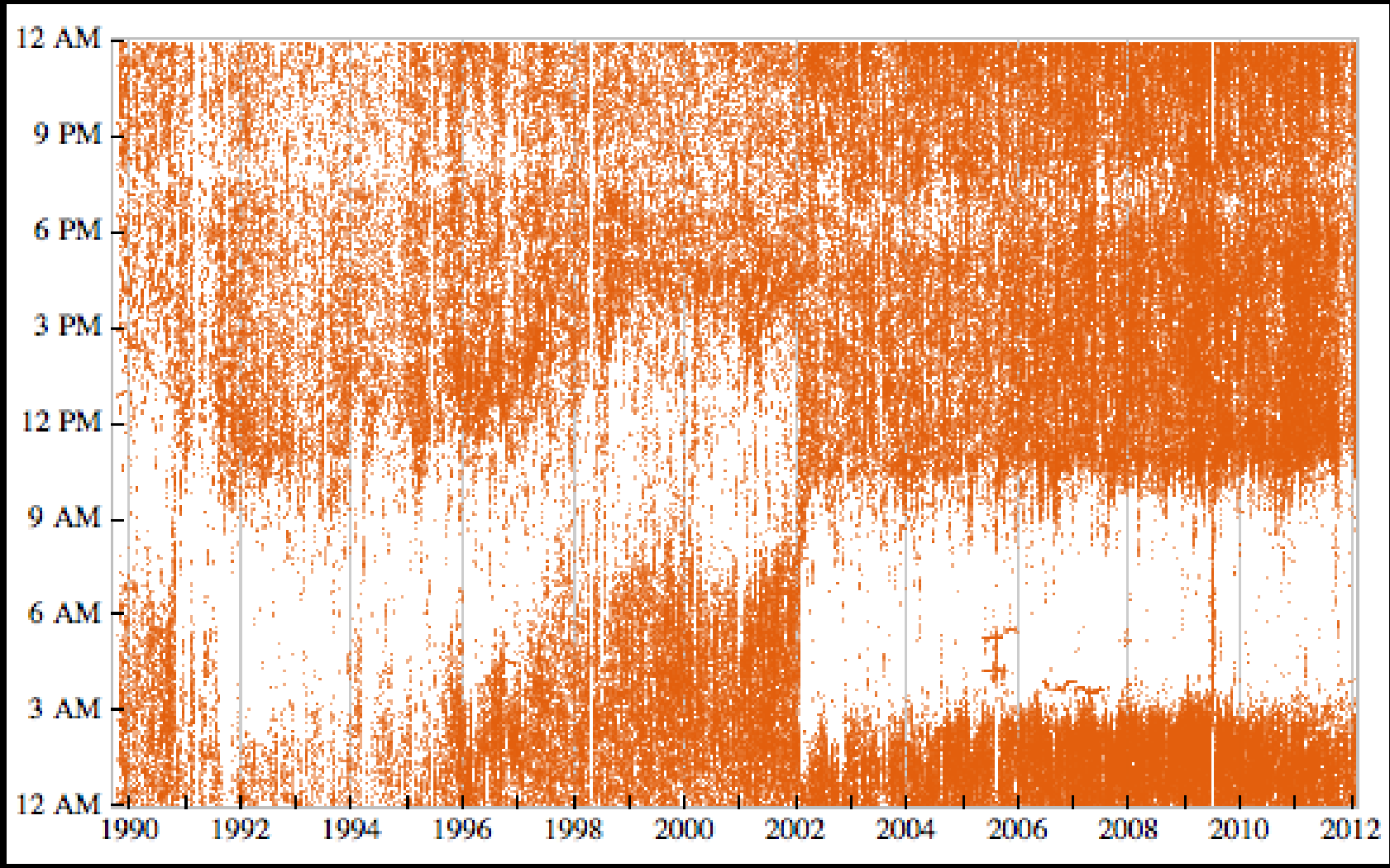


Figure: Stephen Wolfram's outgoing e-mail (approximately 300.000)

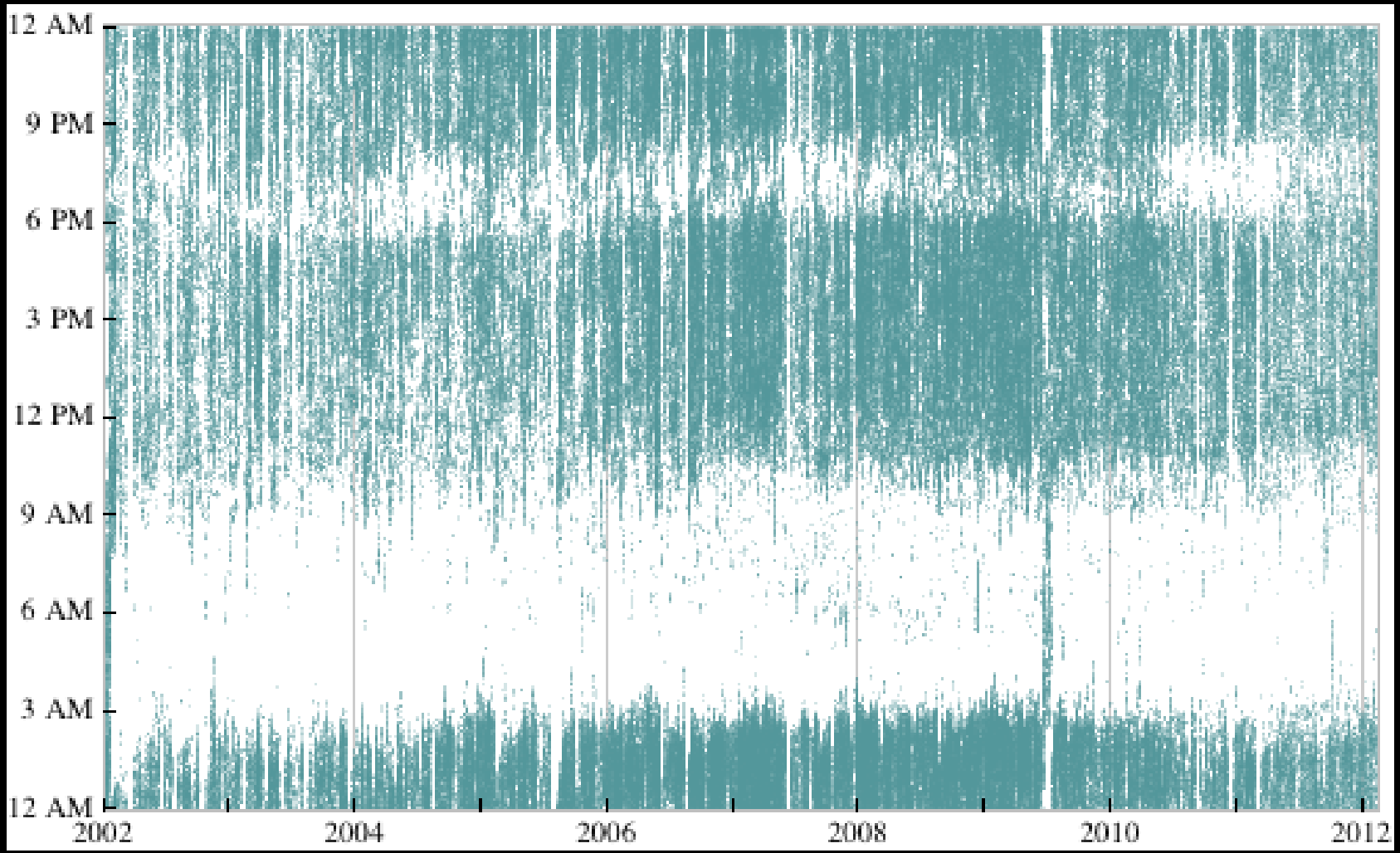


Figure: Stephen Wolfram's keystrokes (approximately 100 million)

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.

1

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 ($9 * 10^{11}$) to 2 trillion ($2 * 10^{12}$) 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.[2]

[2] Adapted from Tufte (2001)

ACCENT principles I

- 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

ACCENT principles II

- 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[3]

[3] *Source*: D. A. Burn (1993), "Designing Effective Statistical Graphs". In C. R. Rao, ed., *Handbook of Statistics*, vol. 9, Chapter 22.

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

Table: Estimates from a logistic regression model predicting likelihood of turnout (Sweden, EES 2009)

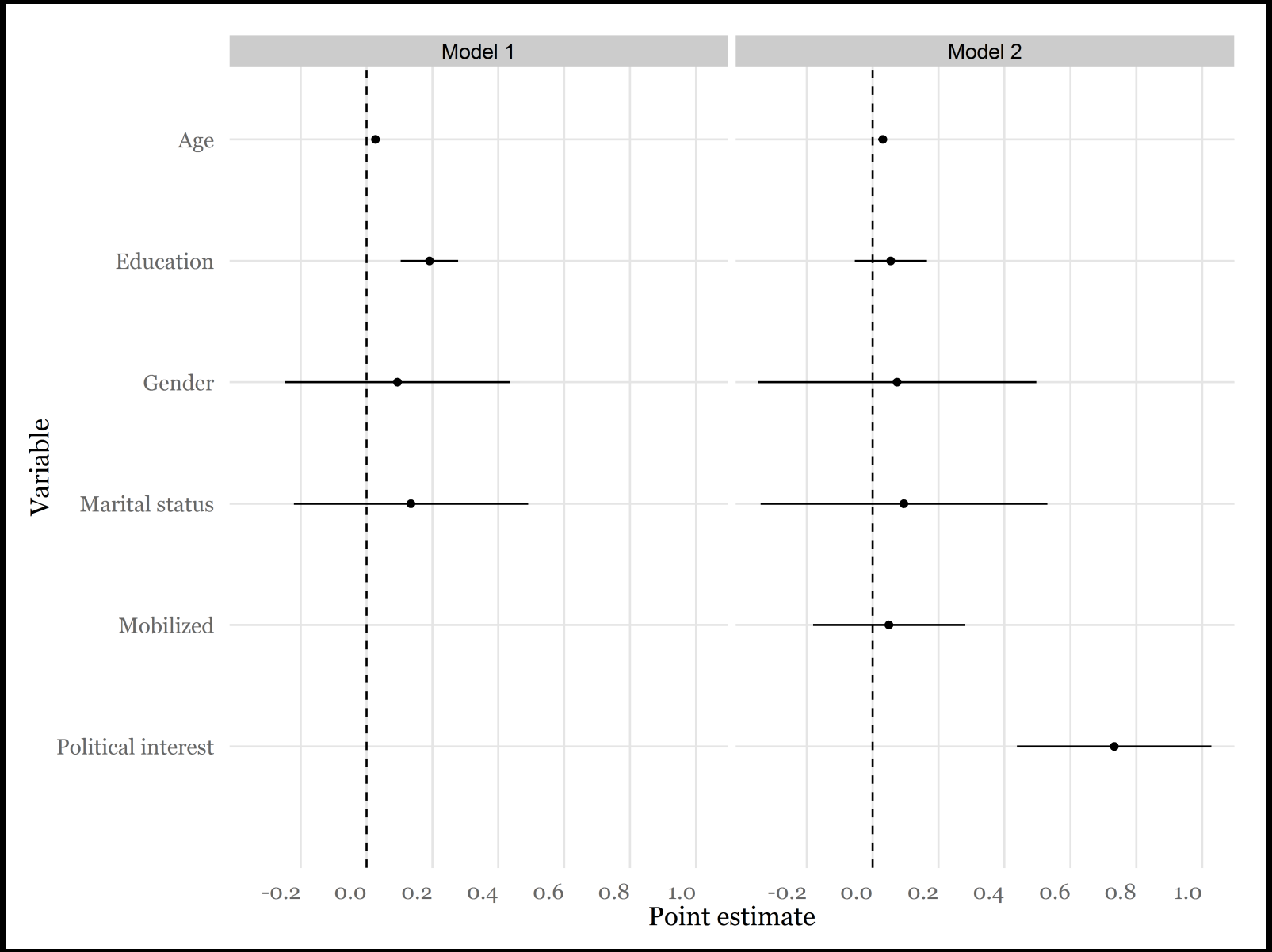


Figure: Estimates from the regression model in graphical form

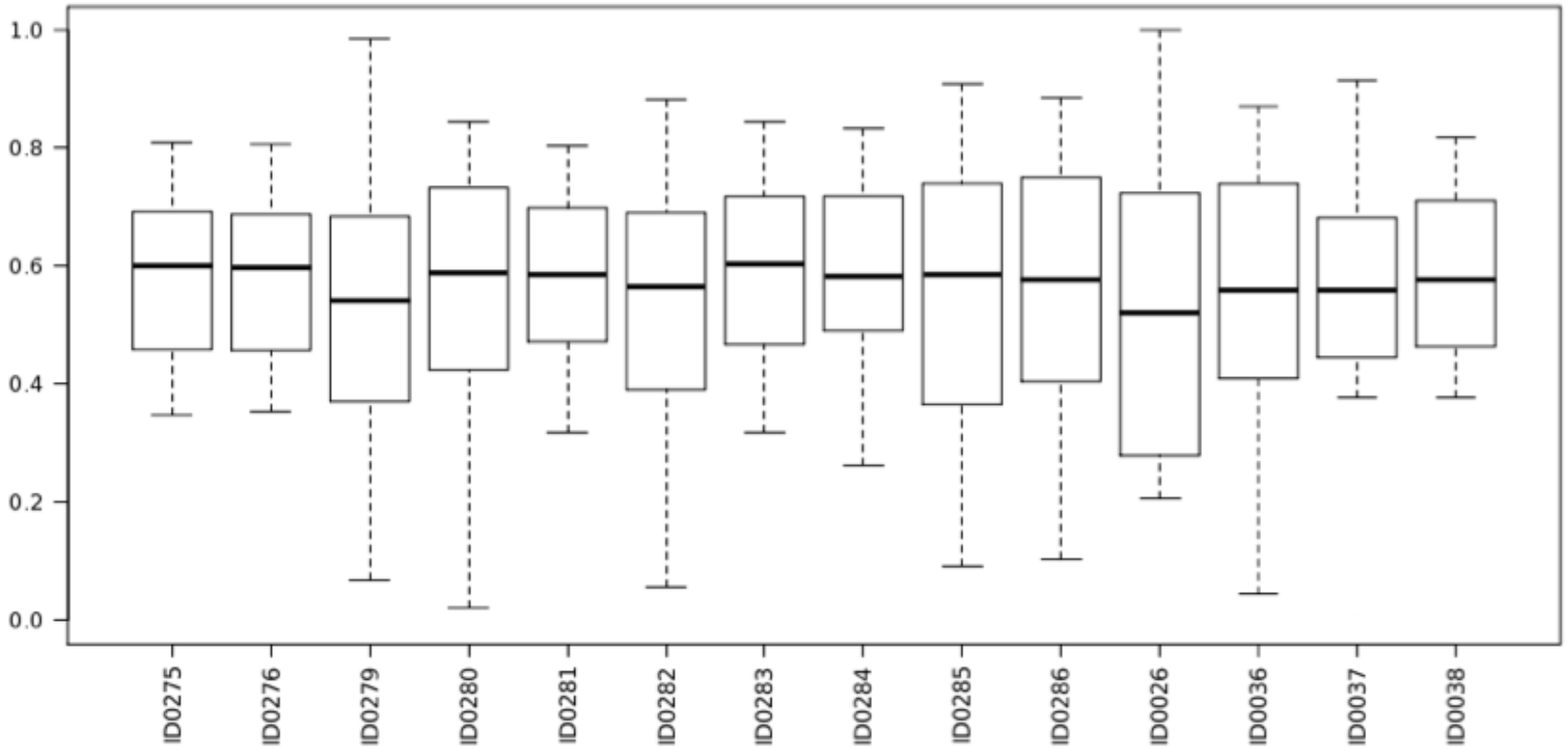


Figure: Traditional boxplot

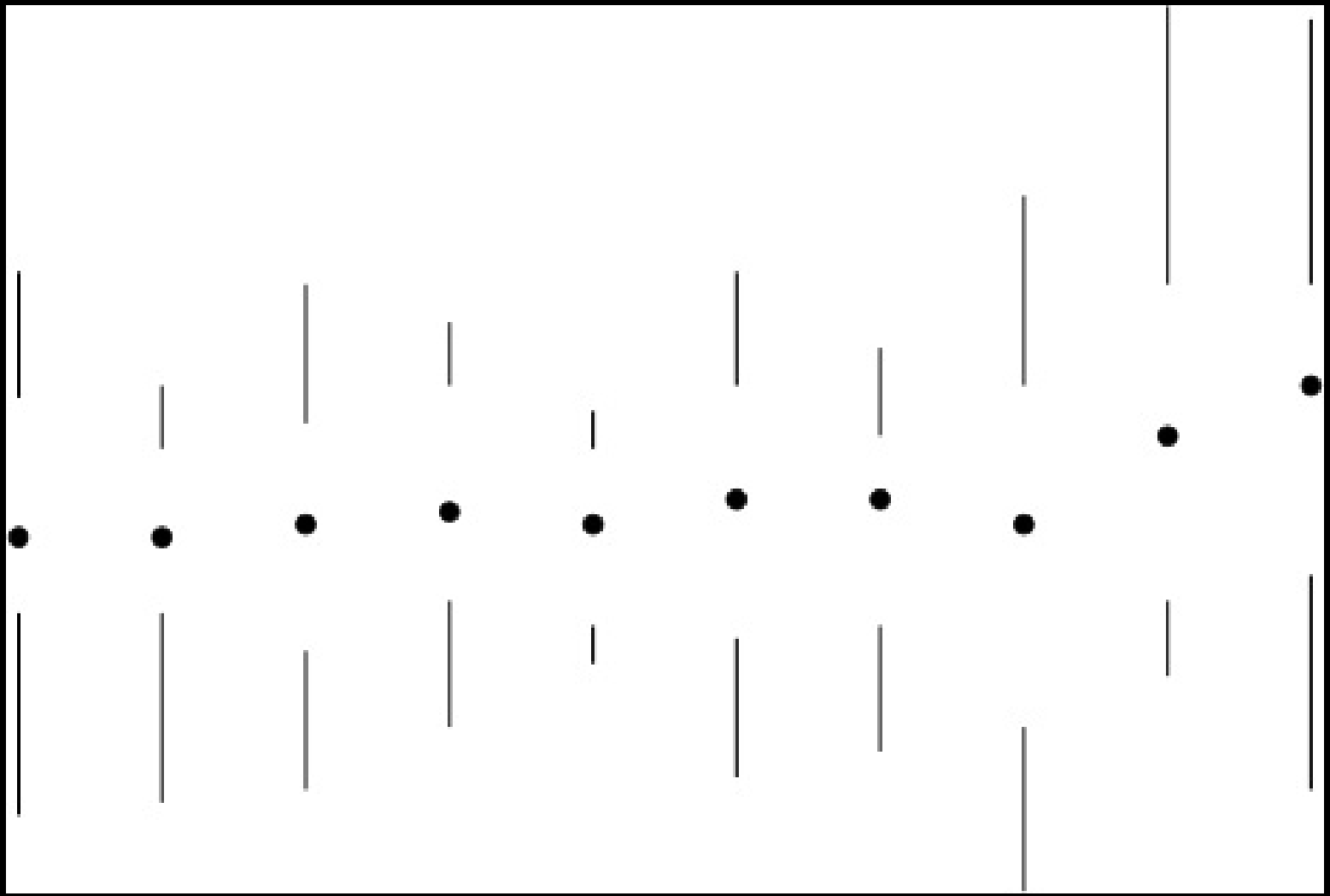
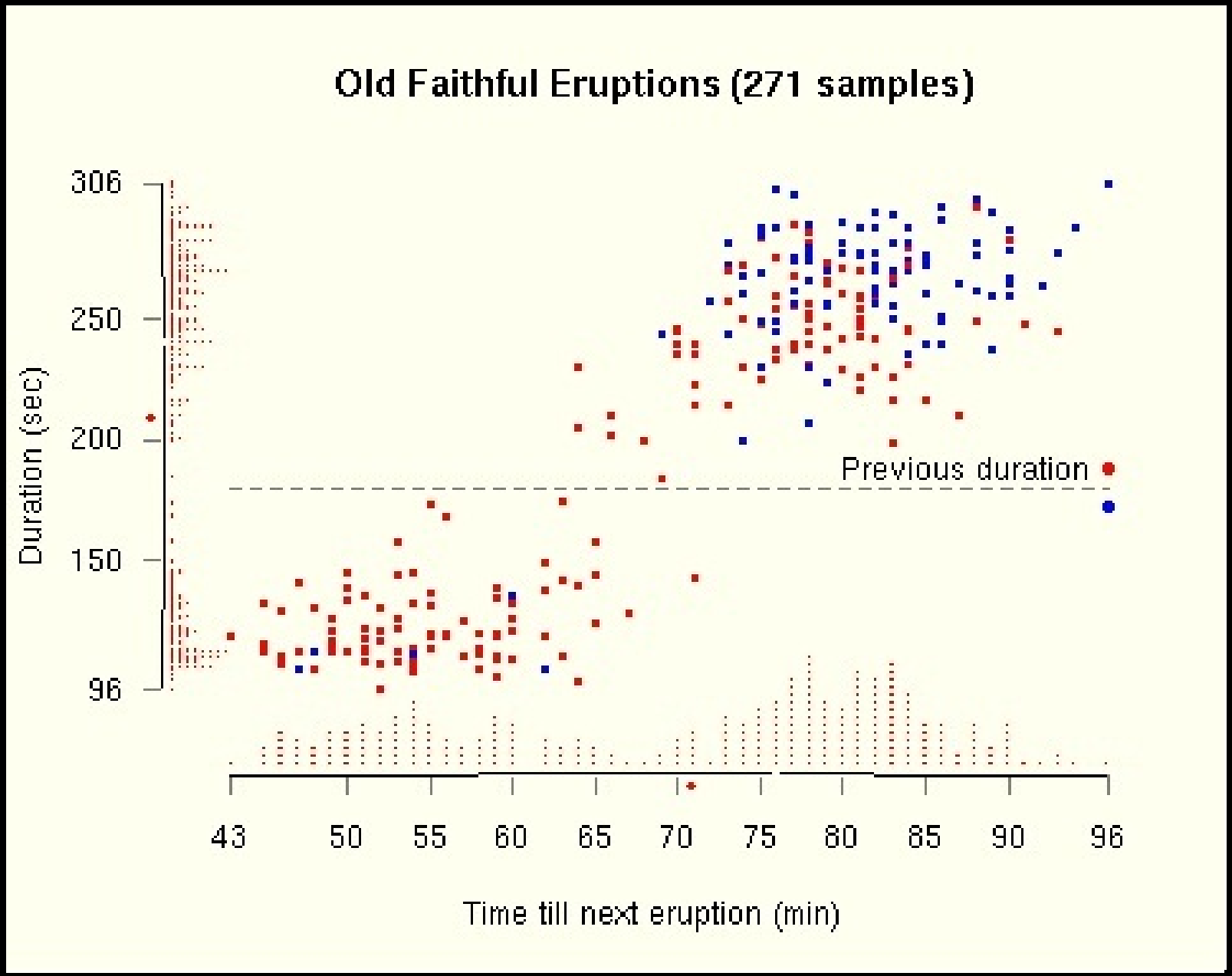


Figure: Quartile plot


Old Faithful Eruptions (271 samples)
306
250
200
150
96
Duration (sec)
Previous duration
43
50
55
60
65
70
75
80
85
90
96
Time till next eruption (min)

2

2.1

Napoleon's 1812-1813 Russian campaign - Charles Joseph Minard.

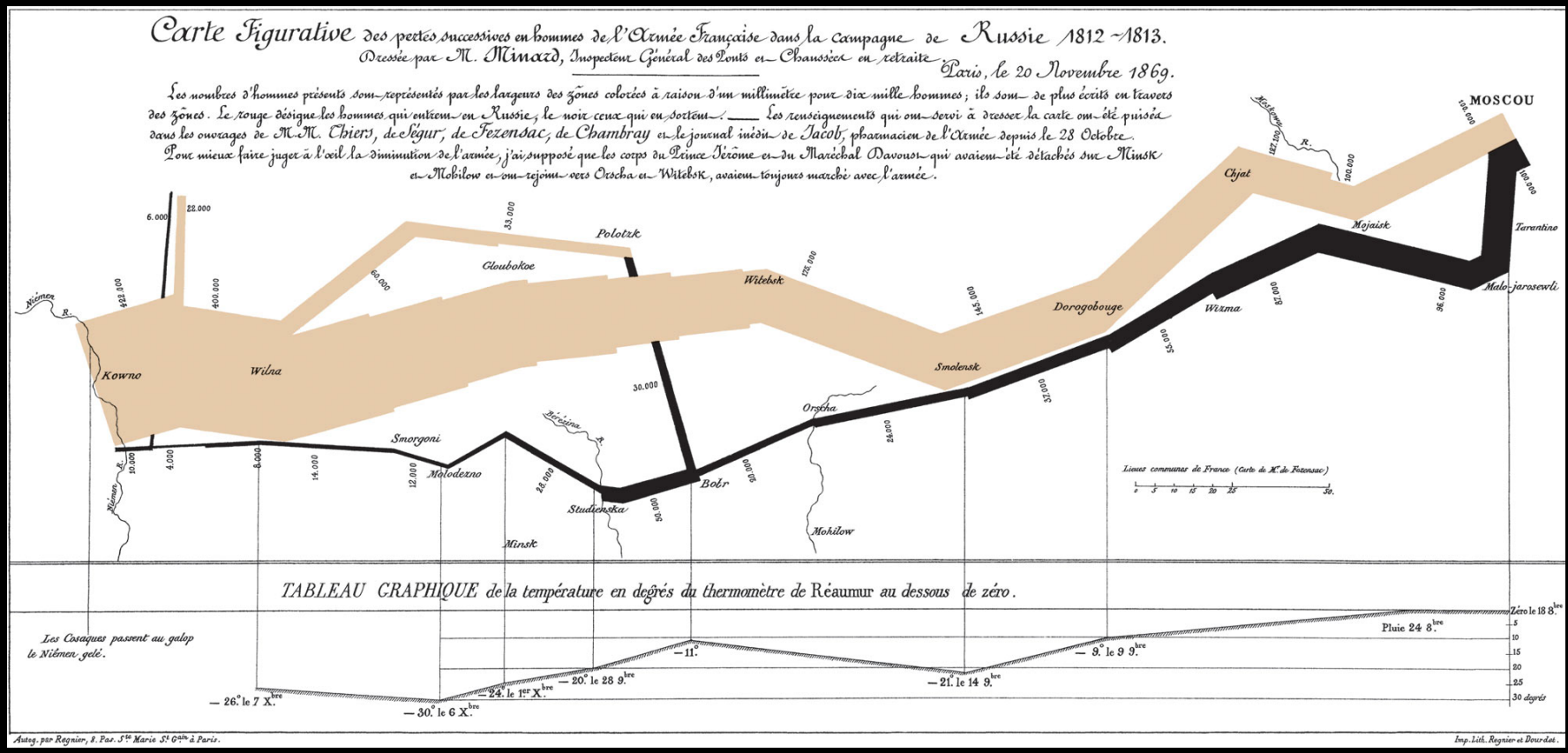


Figure: Campaign map

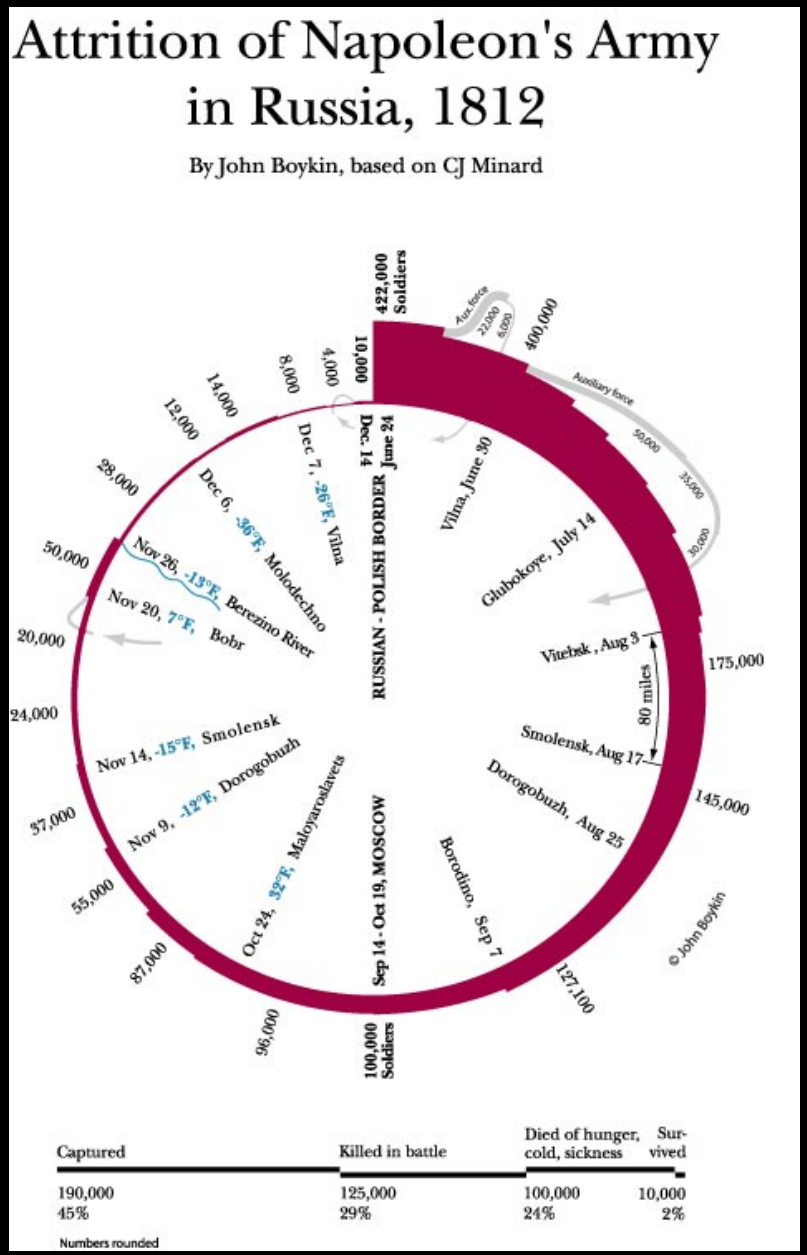


Figure: Alternative to the map

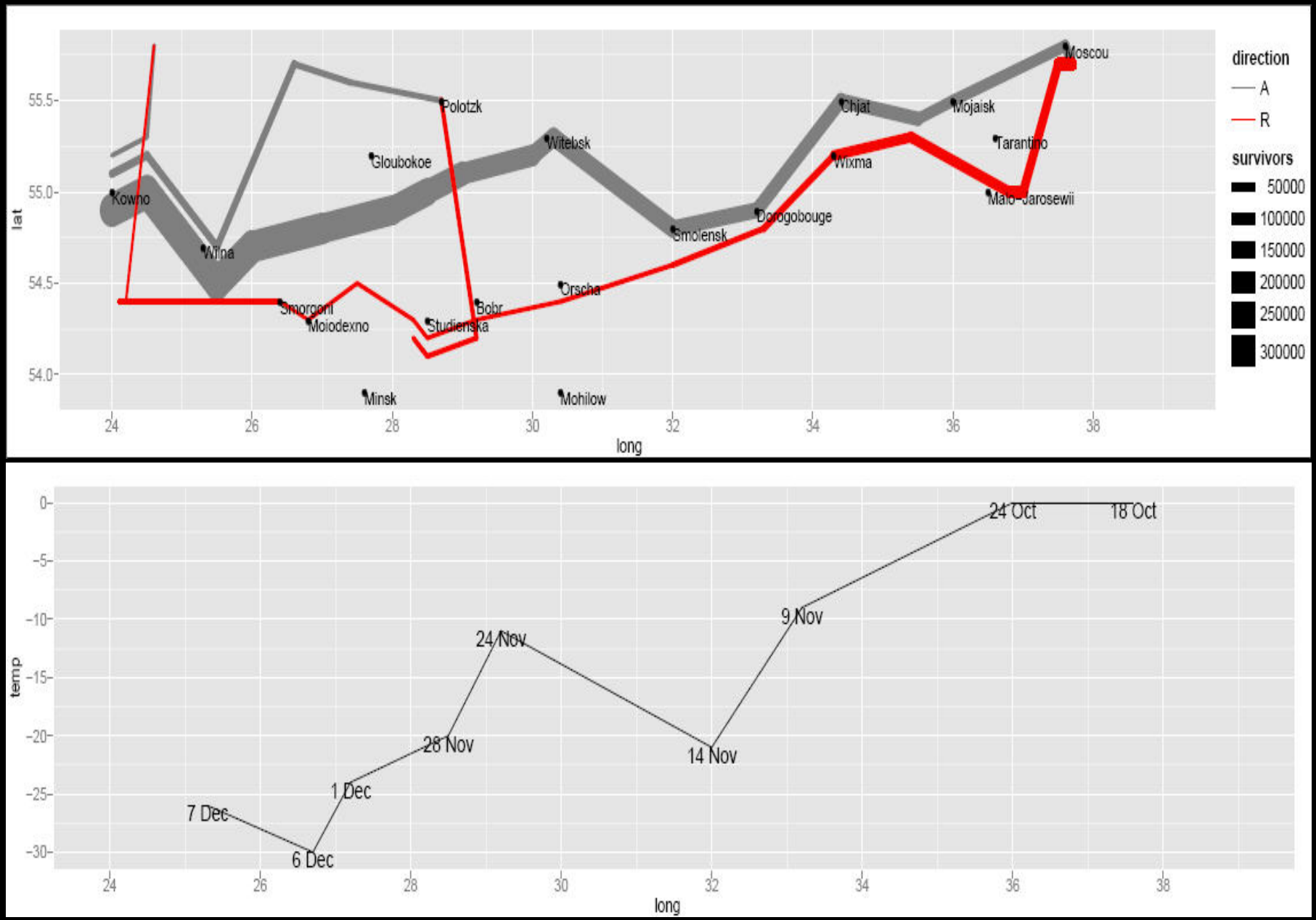
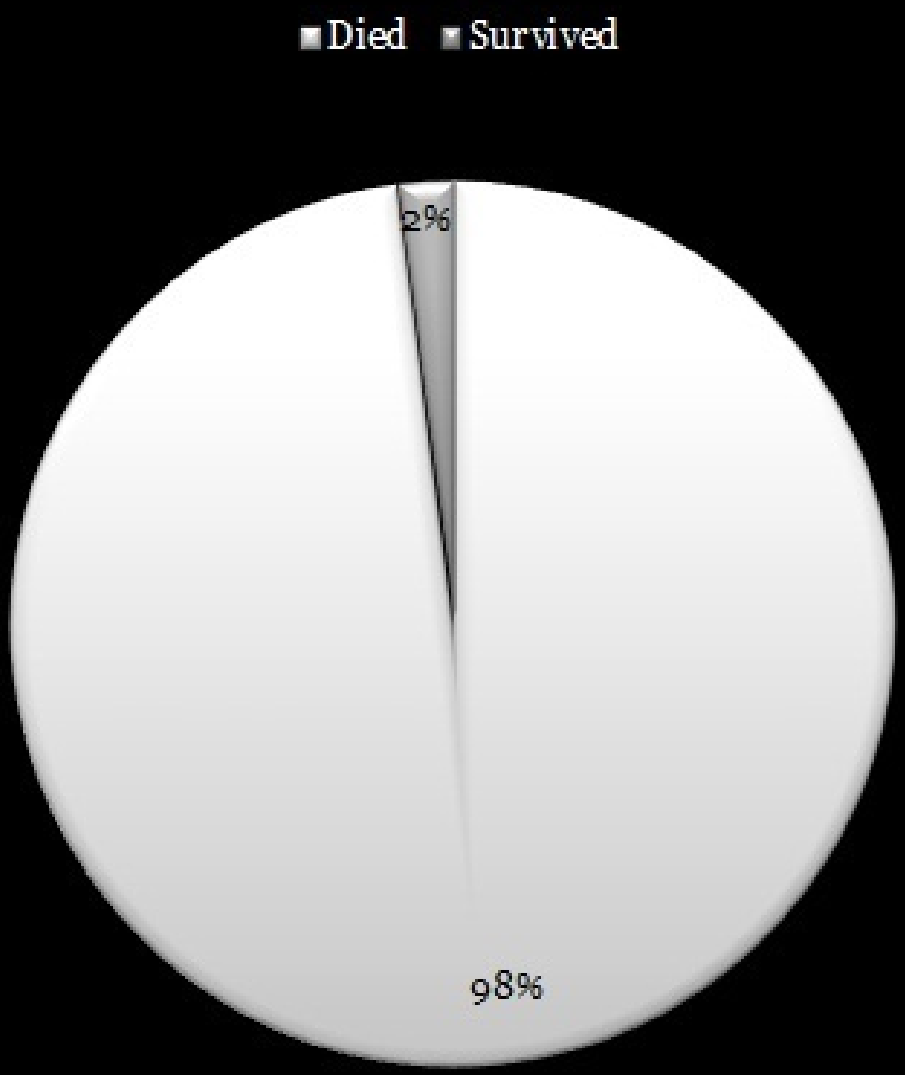


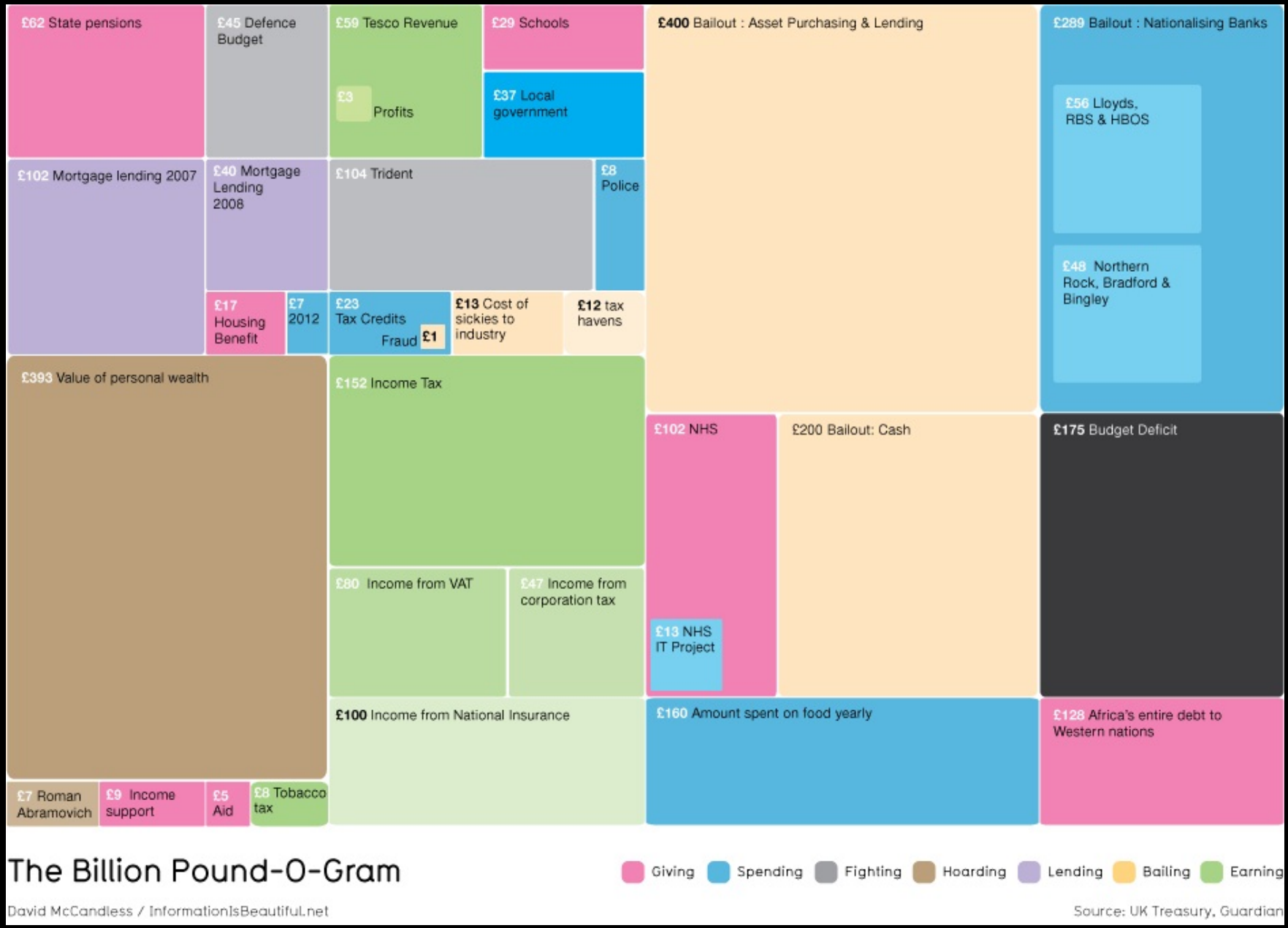
Figure: Alternative to the map

Napoleon's 1812-1813 Russian Campaign



2.2

The UK Budget - David McCandless.


£62 State pensions
£45 Defence Budget
£59 Tesco Revenue
£3 Profits
£29 Schools
£37 Local government
£400 Bailout : Asset Purchasing & Lending
£289 Bailout : Nationalising Banks
£56 Lloyds, RBS & HBOS
£48 Northern Rock, Bradford & Bingley
£102 Mortgage lending 2007
£40 Mortgage Lending 2008
£104 Trident
£8 Police
£17 Housing Benefit
£7 2012
£23 Tax Credits
Fraud £1
£13 Cost of sickies to industry
£12 tax havens
£393 Value of personal wealth
£152 Income Tax
£102 NHS
£13 NHS IT Project
£200 Bailout: Cash
£175 Budget Deficit
£80 Income from VAT
£47 Income from corporation tax
£100 Income from National Insurance
£160 Amount spent on food yearly
£128 Africa's entire debt to Western nations
£7 Roman Abramovich
£9 Income support
£5 Aid
£8 Tobacco tax
The Billion Pound-O-Gram
Giving
Spending
Fighting
Hoarding
Lending
Bailing
Earning
David McCandless / InformationIsBeautiful.net
Source: UK Treasury, Guardian

2.3

Commuters in the US - SENSEable City Laboratory, MIT.

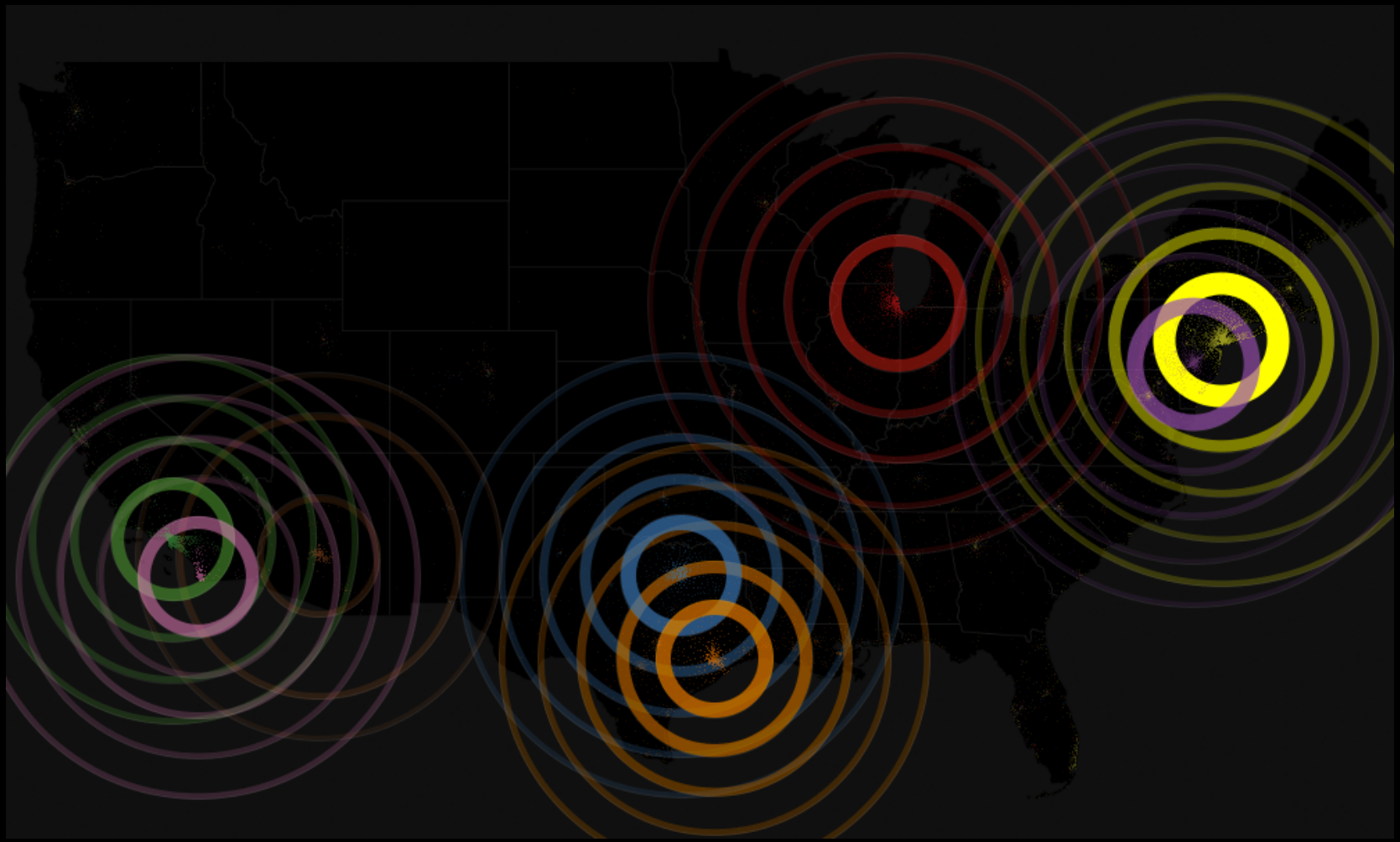


Figure: Commuters - July 2010, AT&T cell phone data

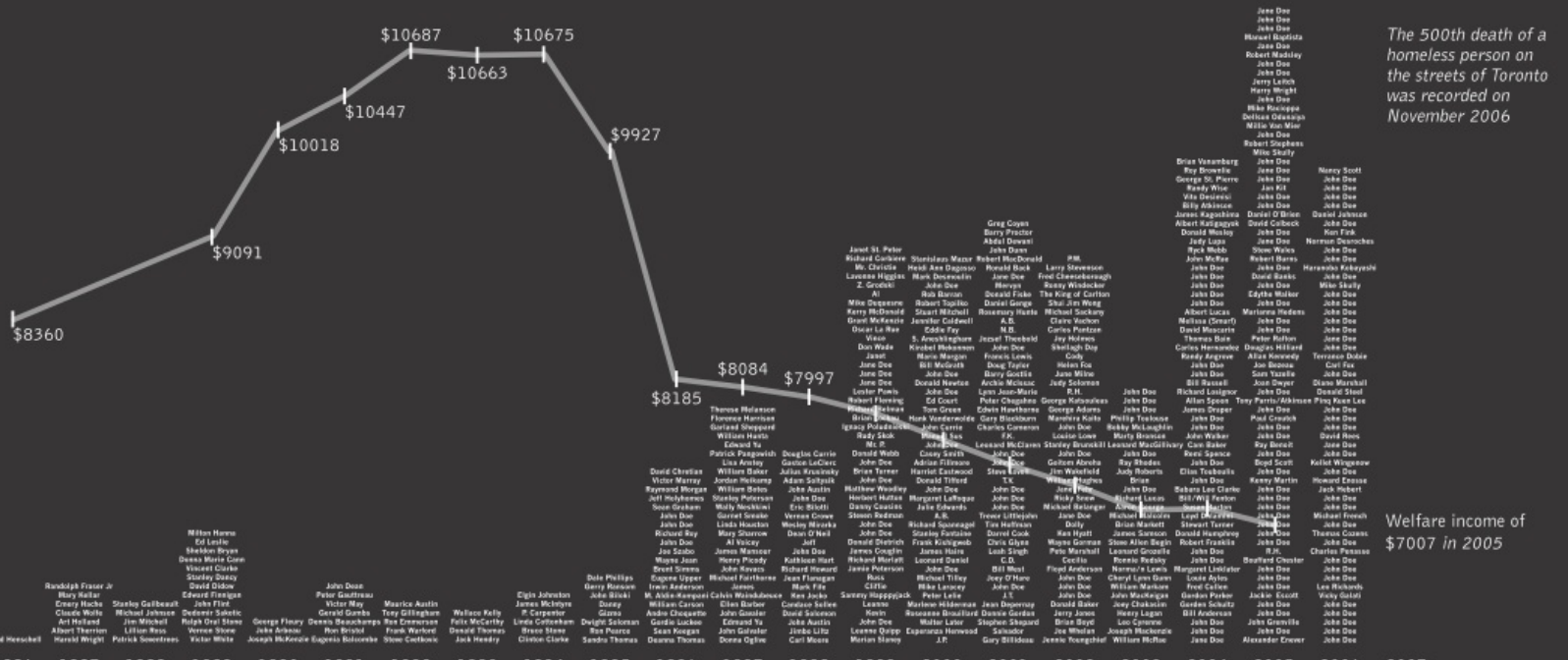
2.4

Welfare benefits in Ontario

Common Sense Revolution

Ontario Welfare Income for a Single Person in 2005 Constant Dollars & Homeless Persons Who Have Died on the Streets of Toronto 1985–2006

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



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

3.1

‘Chartjunk’

Figure: Prominent example

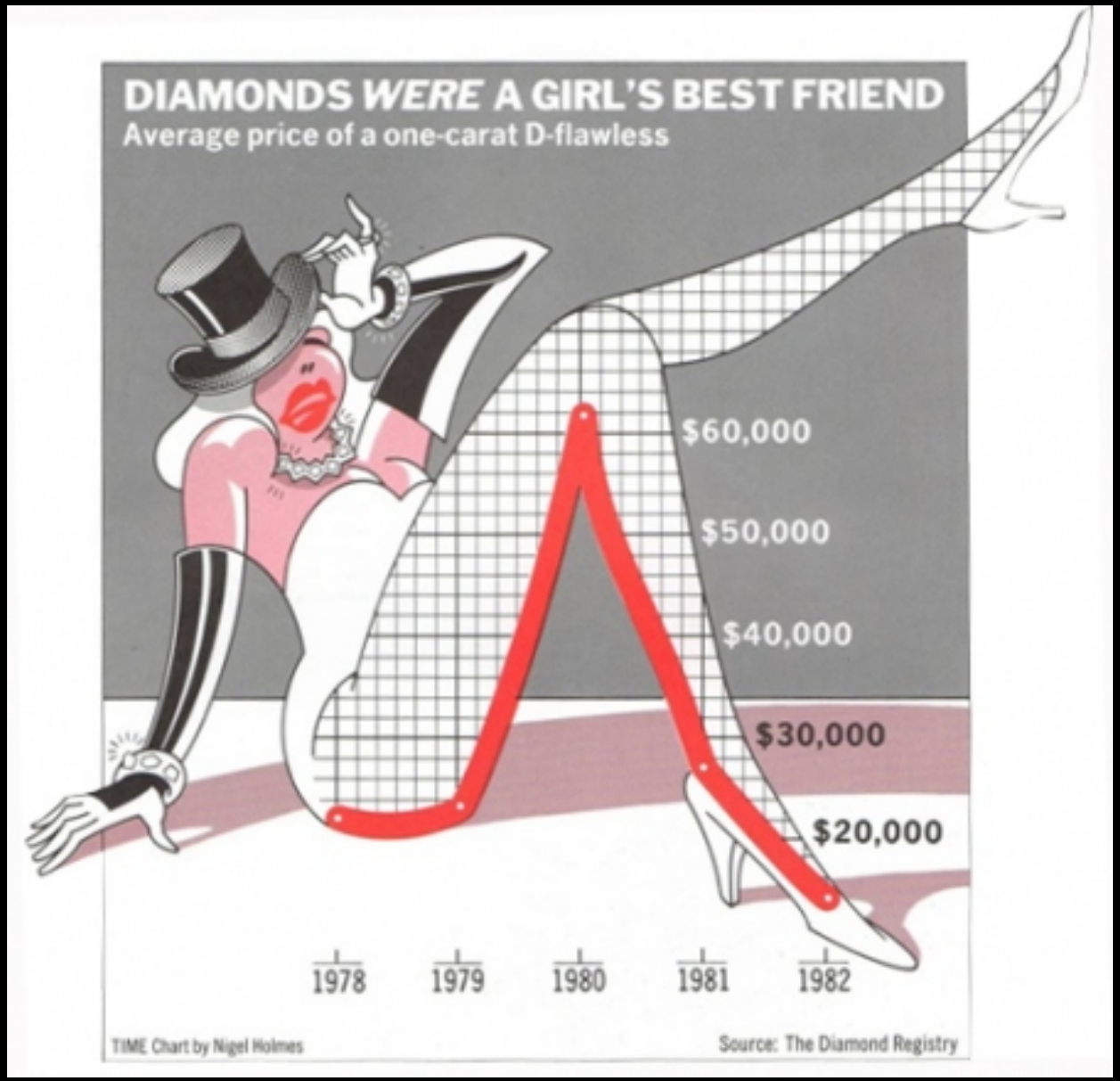


Figure: Prominent example

3.2

Misleading graphs

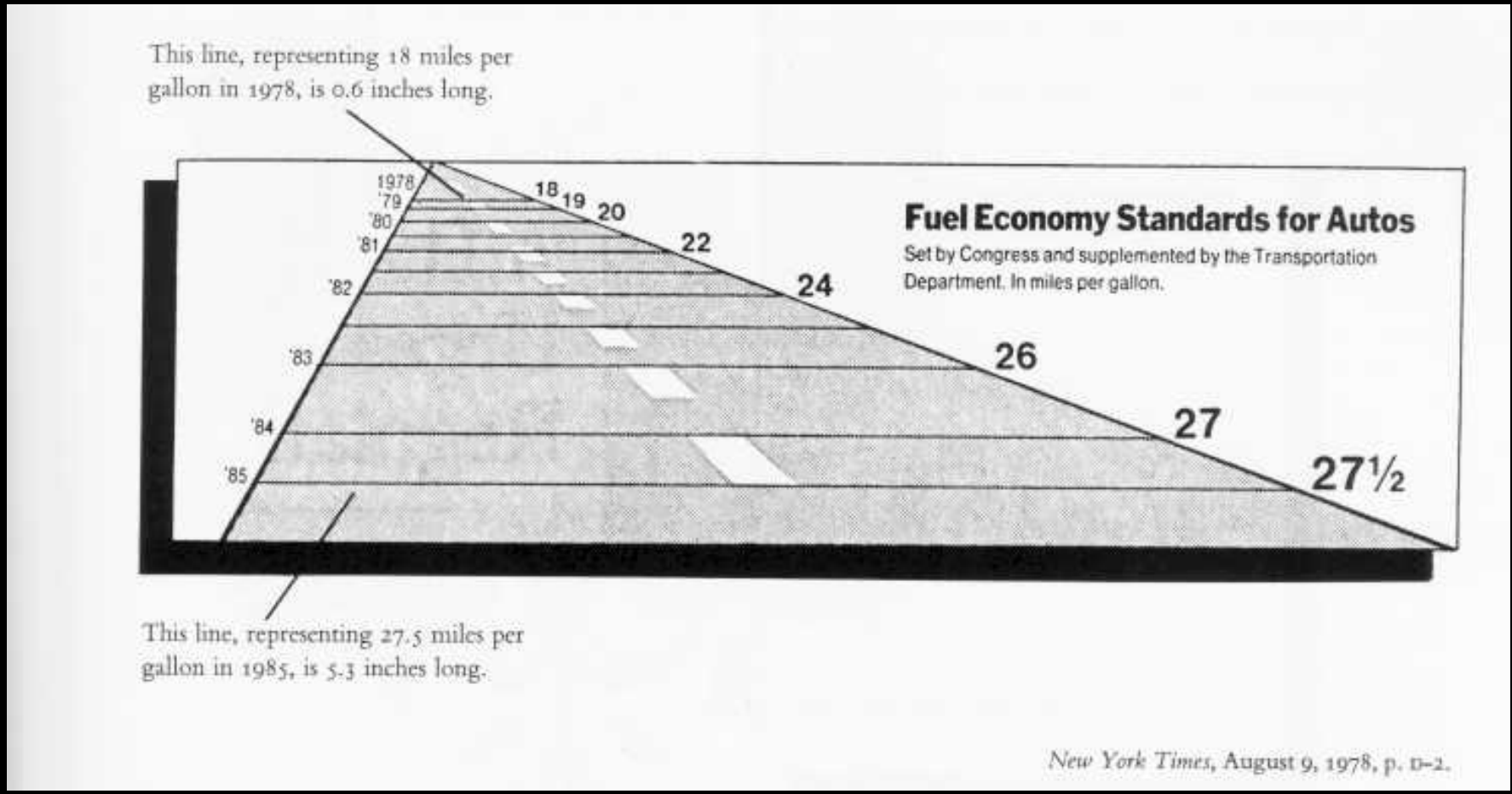
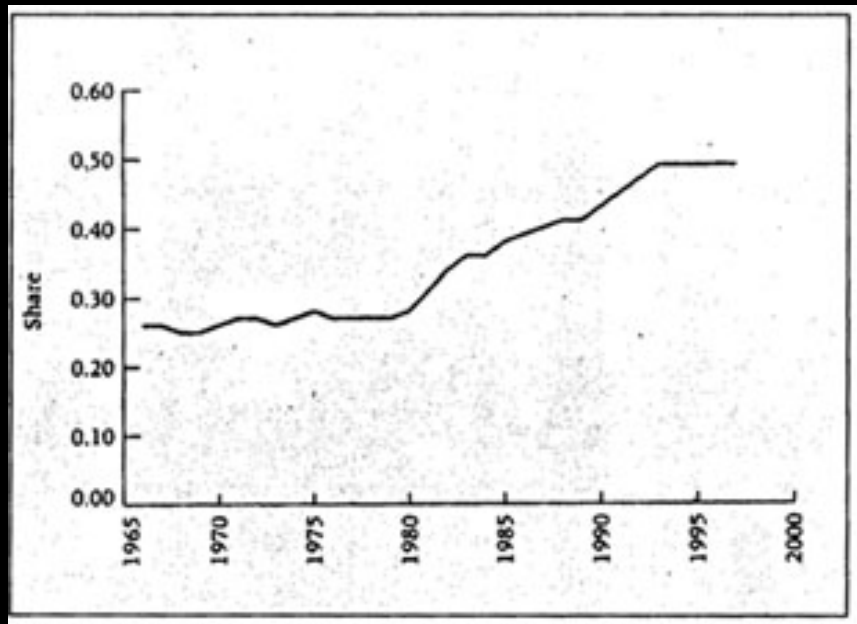
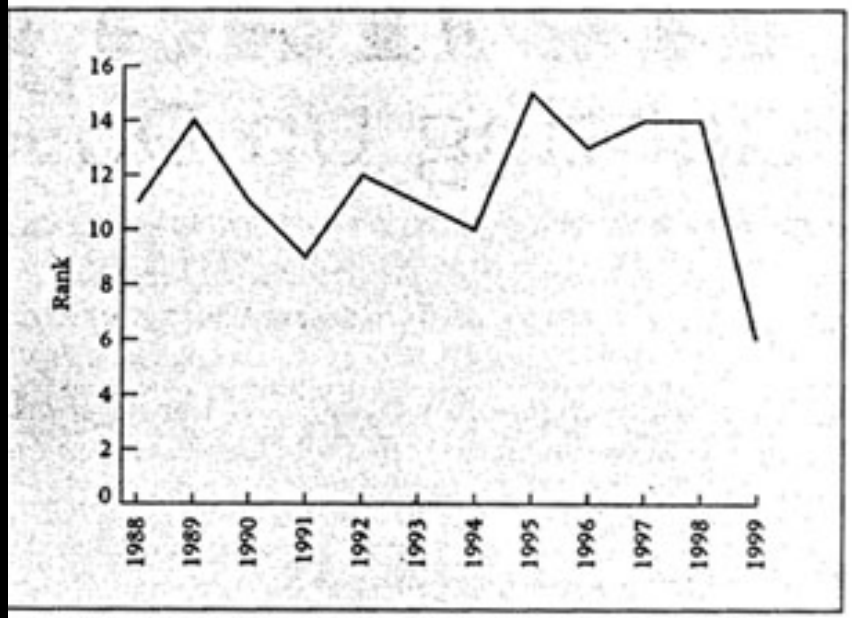


Figure: First example



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.

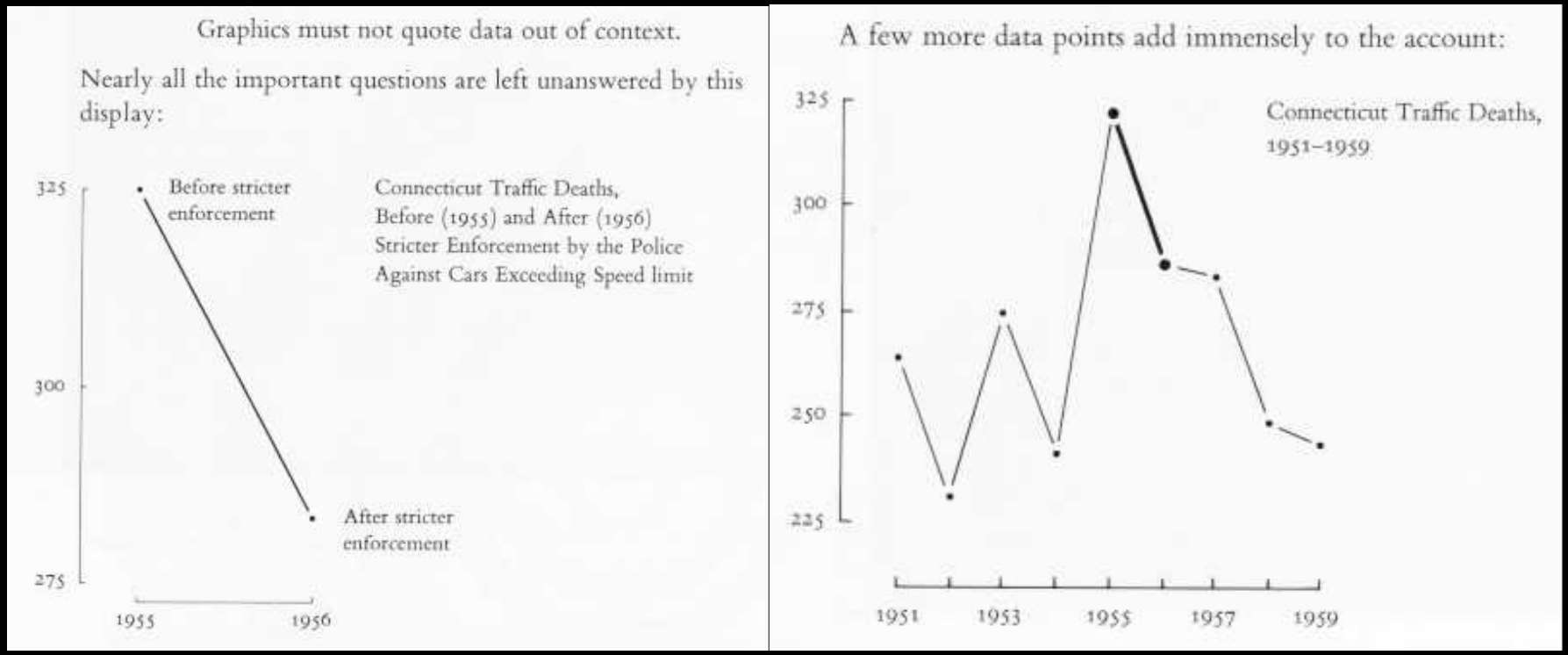


Figure: Third example

3.3

Poor understanding of statistics

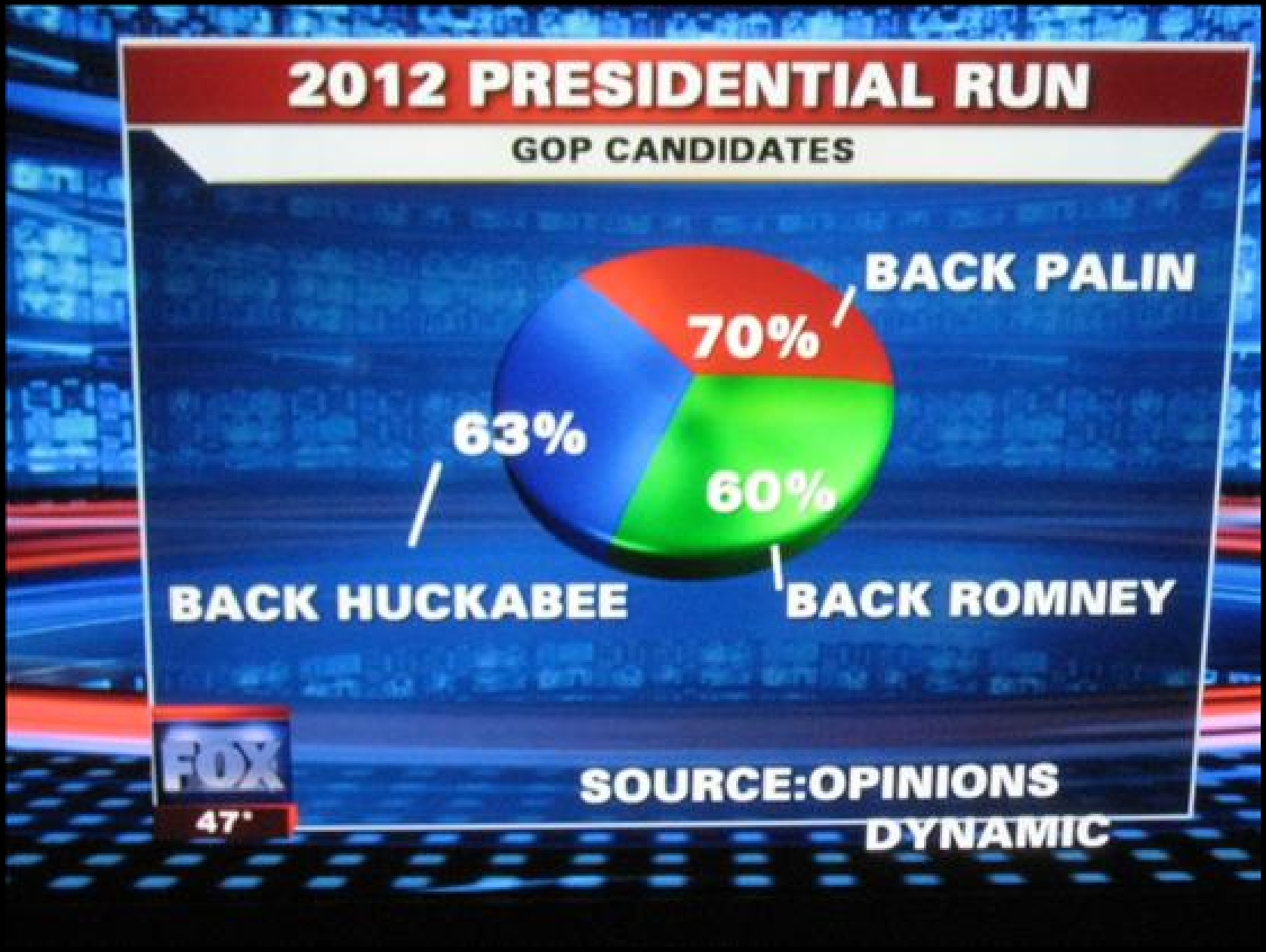


Figure: First example

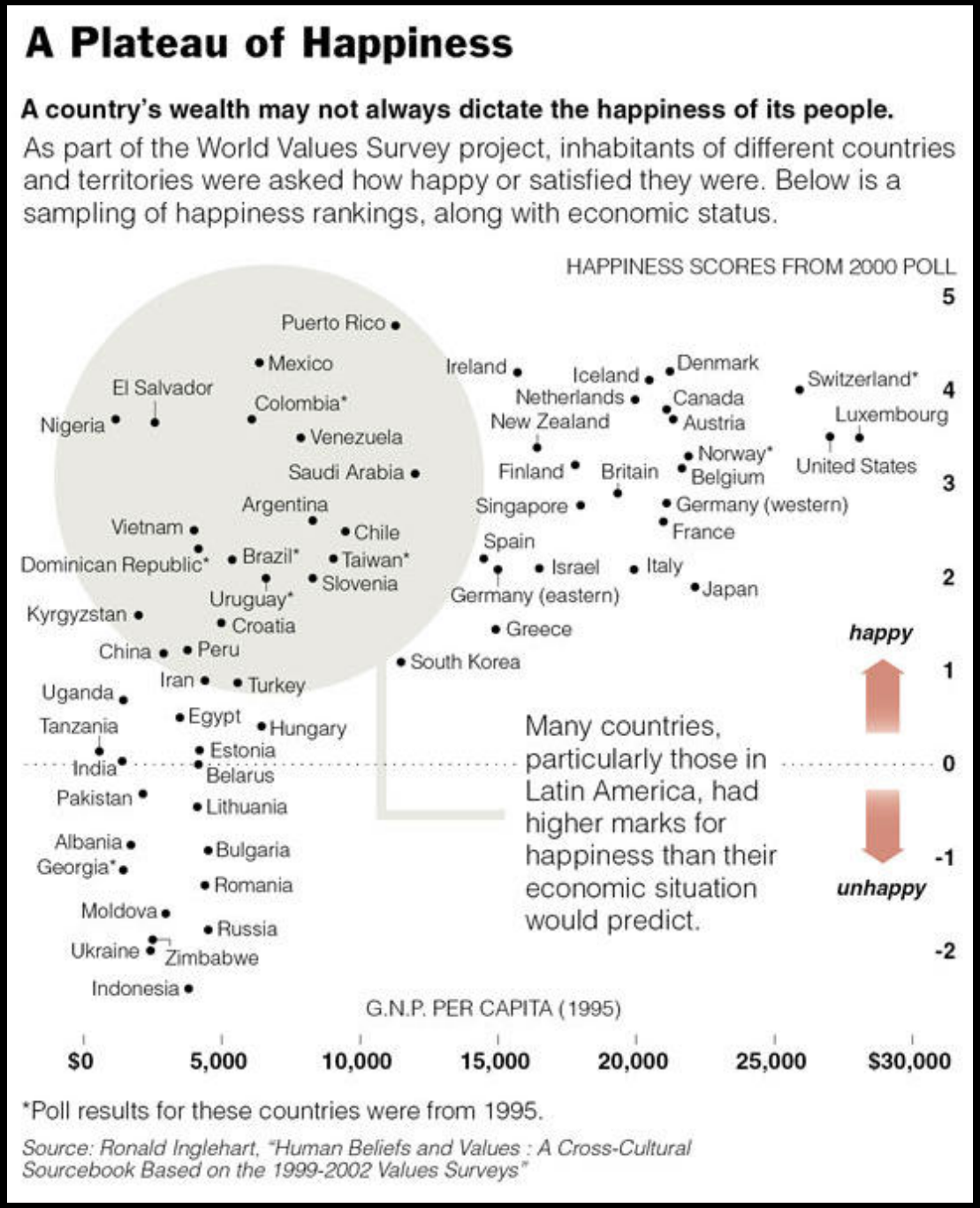


Figure: Second example

3.4

Poor choice of graphical display

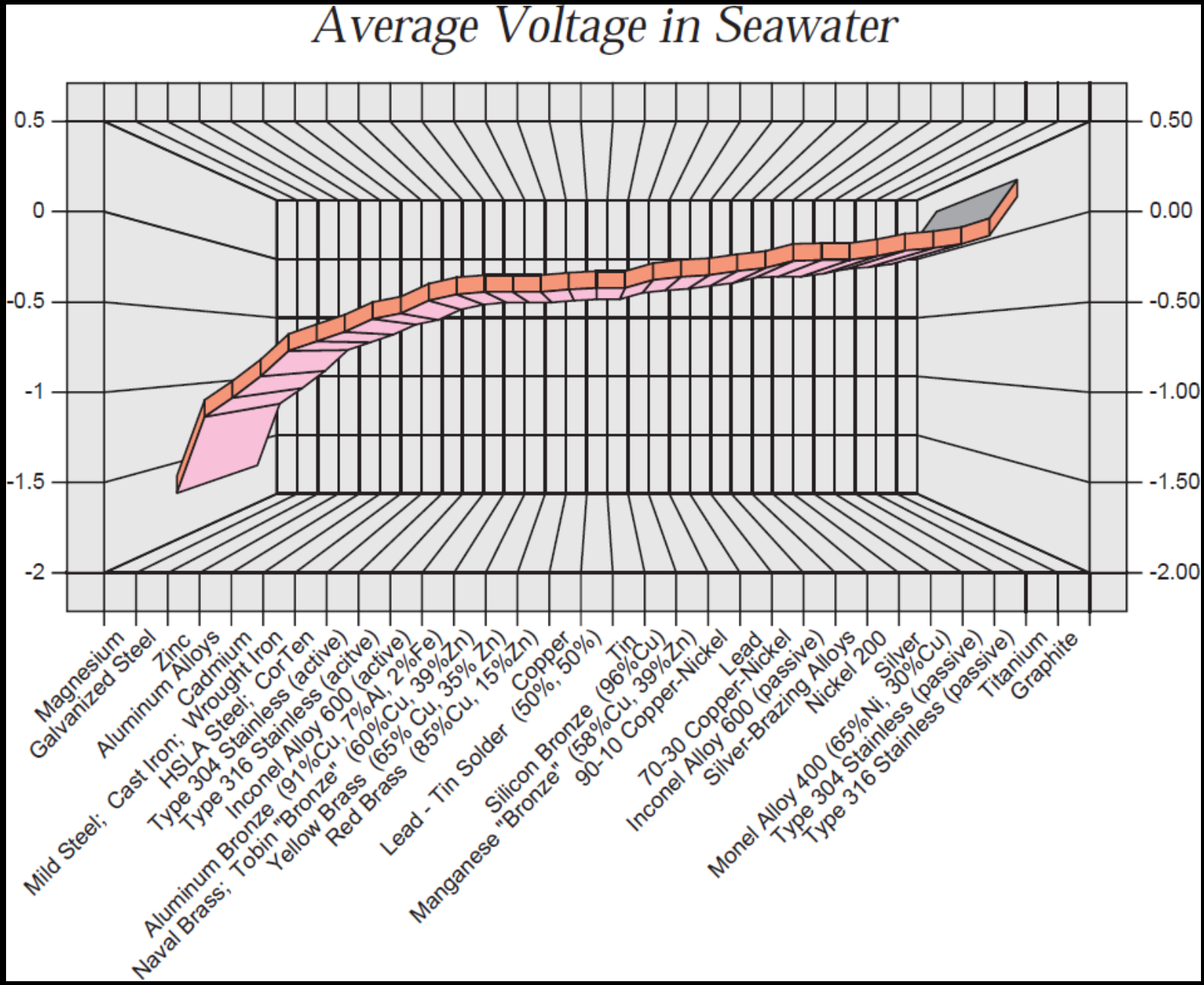


Figure: First example

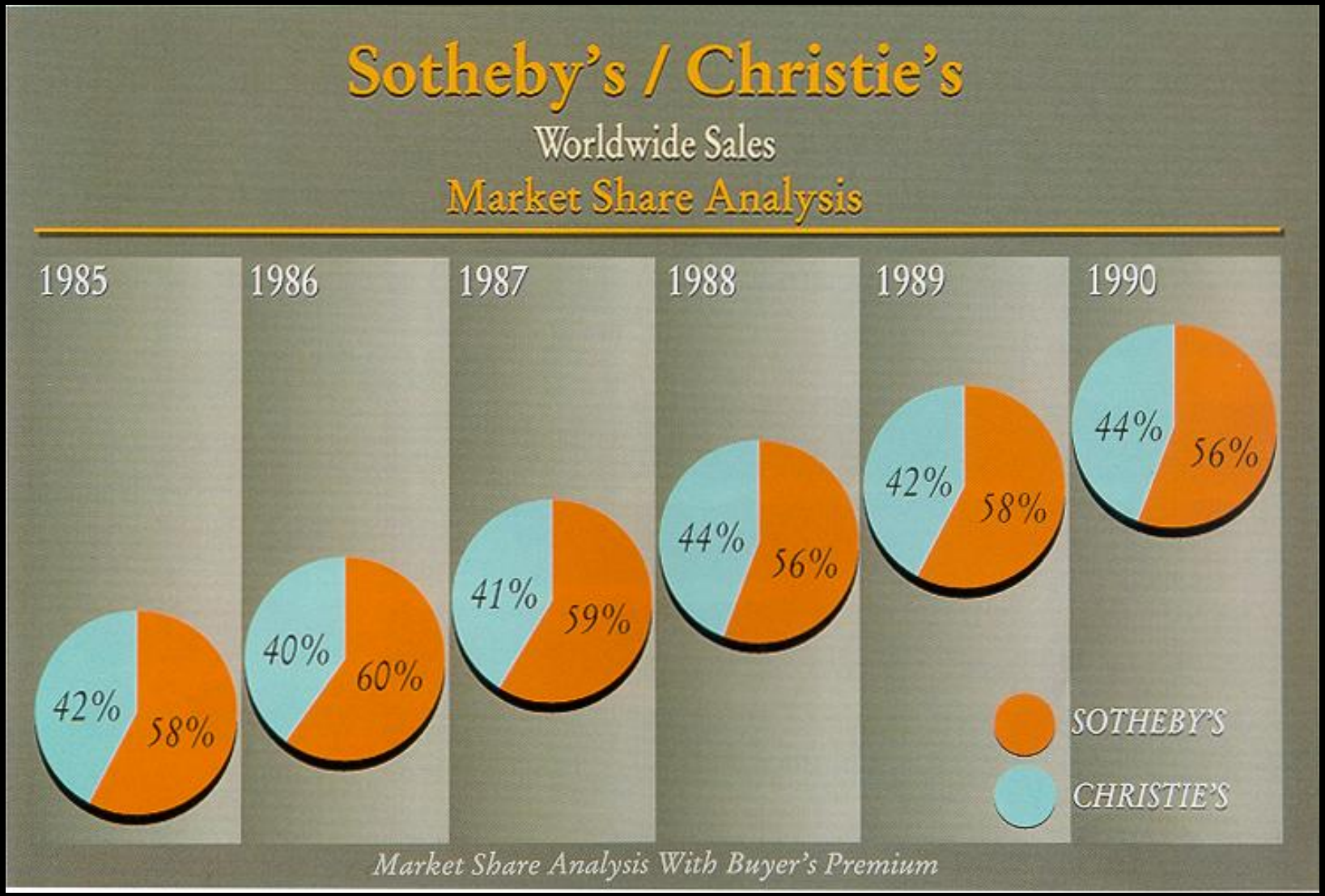


Figure: Second example

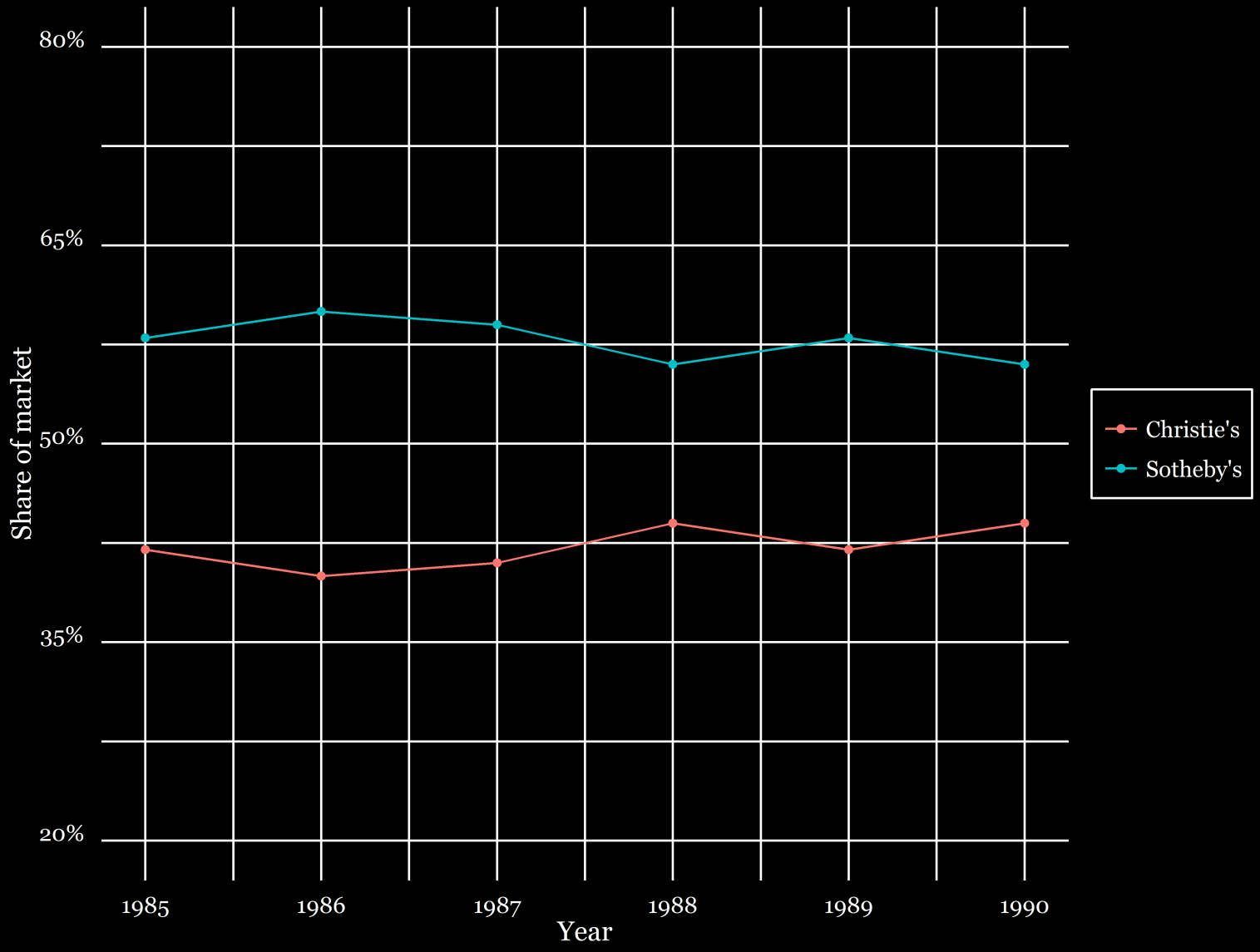


Figure: Alternative to second example

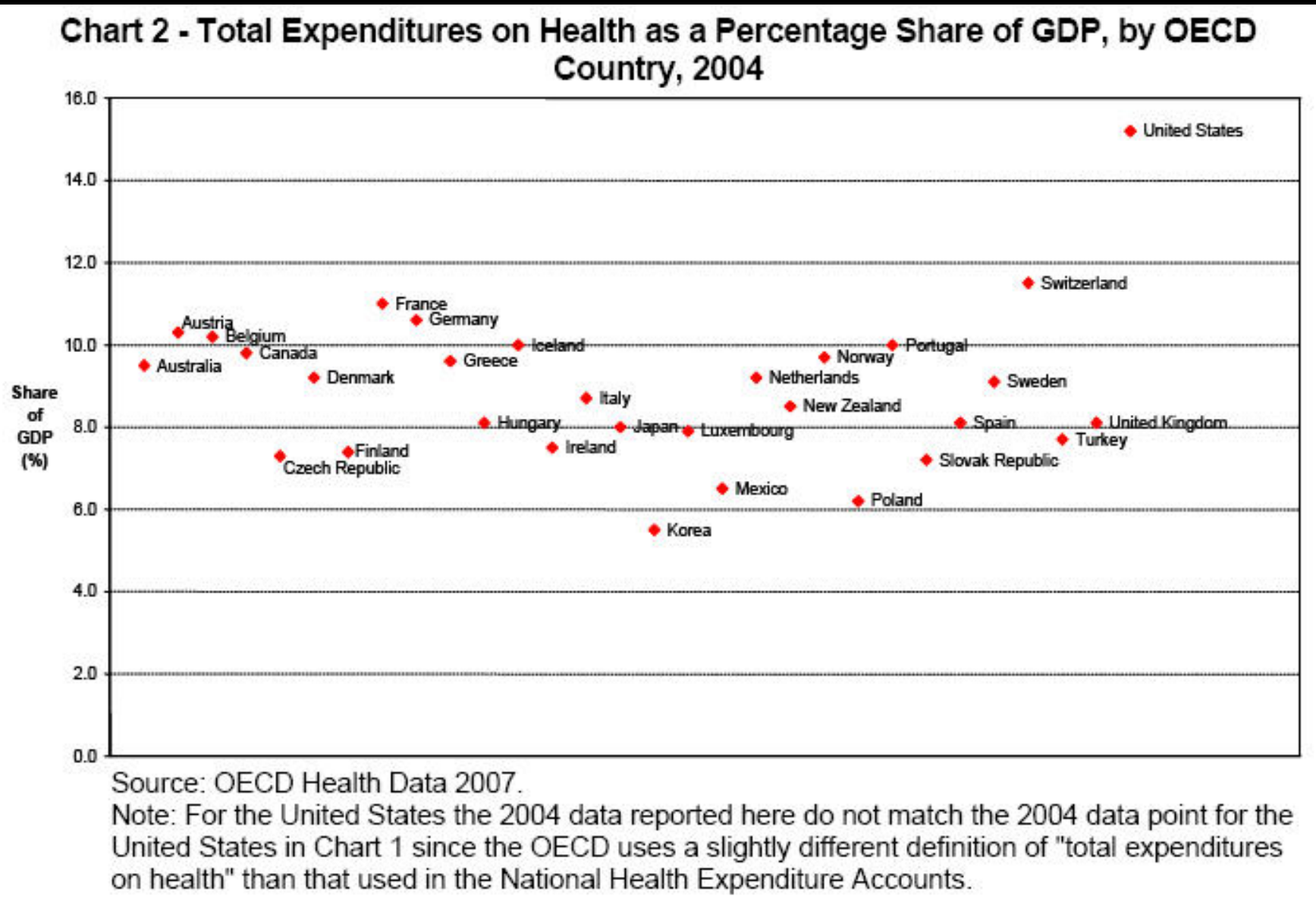


Figure: Third example

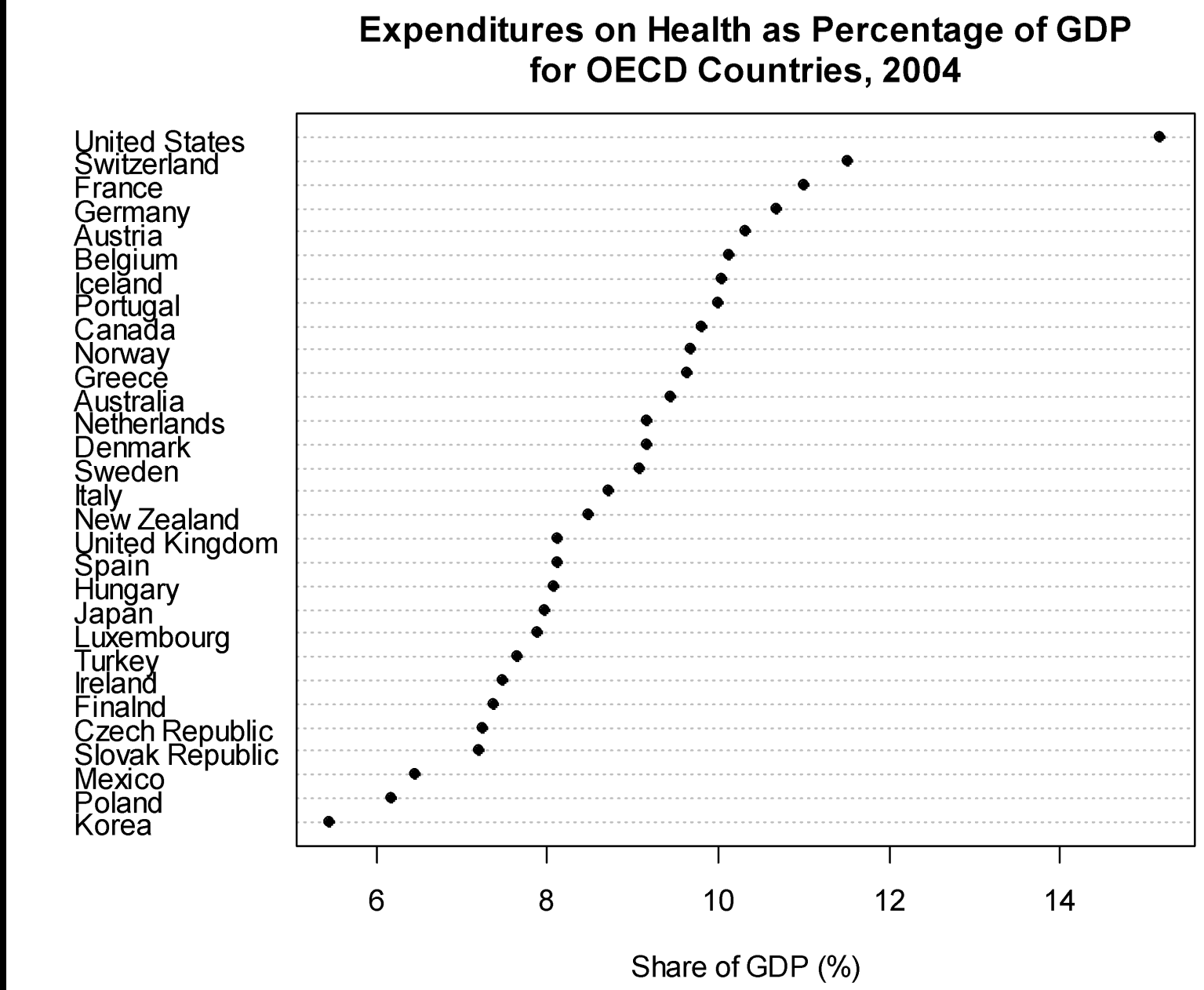


Figure: Reworked graph

4

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

Gapminder

- 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

ggplot2

- 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)

5

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!

References I

Books used for ideas or graphs:

- Tufte, Edward R. 1997. *Visual Explanations - Images and Quantities, Evidence and Narrative*. Cheshire, CT: Graphics Press.
- Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
- Tufte, Edward R. 2006. *Beautiful Evidence*. Cheshire, CT: Graphics Press.
- Wickham, Hadley. 2009. *ggplot2 - Elegant Graphics for Data Analysis*. New York: Springer.
- Wilkinson, Leland. 2005. *The Grammar of Graphics*. New York: Springer.

References II

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/

References III

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/)