Visual Communications

The art of showing complex data and messages effectively

Roberto Foa

foa@fas.harvard.edu

today

1. Principles of data presentation

2. A practical guide to GIS mapping

today

1. Principles of data presentation

2. A practical guide to GIS mapping

factblink

Clients and Projects



Kofi Annan, Africa Progress Panel



Al Gore, Generation IM



Why data visualisation matters

When was the last time you saw a good presentation?

(with statistics that made sense)





What these presentations have in common...

In general, we are **visual** learners

Concentrate!



Now draw what you just saw...

How about now...



We tend to recognise patterns and find relationships.

This helps us **remember**

we didn't evolve to process data



[timeline of human evolution]

So there is a strong case to go visual

So there is a strong case to go visual

What are the benefits of visual communication?

















Source: Cognac Communications - http://www.cognac.co.uk

benefits of visual display

- 1. Getting a sense of the data
- 2. Getting understood by others
- 3. Getting noticed
- 4. Getting published

Publication in top science journals (Science, Nature) often requires an **advanced visual infographic** and the same is increasingly true in the social sciences.



Source: Michalopoulos and Papaioannou (2013) "Precolonial Ethnic Institutions and Contemporary African Development", *Econometrica*

Data visualisation also has value during the analysis stage

 Let's suppose we have a table of data (homicide rates in provinces of Russia)

RU-AD	6.132182603
RU-AL	21.29719264
RU-ALT	10.92263136
RU-AMU	18.36192317
RU-ARK	13.13963927
RU-AST	10.6888361
RU-BA	9.897099634
RU-BEL	5.481239804
RU-BRY	10.1944793
RU-BU	26.12631146
RU-CE	7.998745295
RU-CHE	13.57062764
RU-CHU	17.85714286
RU-CU	8.156737305
RU-DA	9.928221641
RU-IN	14.69879518
RU-IRK	22.32381894
RU-IVA	10.46084252

Russia, homicide rate



Similarly for India...



visualisation is essential for inductive reasoning







The Pump Today in London – a Memorial to John Snow

benefits of visual display

- 1. Getting a sense of the data
- 2. Getting understood by others
- 3. Getting noticed
- 4. Getting published

How not to display data

TABLES STORE DATA, NOT PRESENT IT

Motivation

Tables are natural containers for data. Whenever information is presented, chances are excellent that it is communicated by means of a table. In many cases, however, when this information is complex (and the table, therefore, is large) a tabular presentation is difficult to parse visually and patterns in the tabulated data remain opaque.

In other words - a useful container isn't automatically a useful presenter. The table presents individual data points very well and patterns that they compose very poorly.

TYPES OF CONTAINERS Table 1. Number of homopeptide repeats and RCPs in GENPEPT, Eukaryotes, and Prokaryotes Other (viruses/environmental GENPEPT Eukaryote Prokaryote sequences) Repeats Proteins Repeats Proteins Repeats Proteins Repeats Proteins 6132 5045 117 5465 4425 83 351 256 370 Alanin 416 149 Valine 46 122 20 463 106 20 414 Leucine 1638 1602 1446 1426 56 3931 22 186 koleurine 57 34 4157 Proline 4837 3333 217 184 Methionin 27 18 175 13 Pherofalaning 20 Tryptophan 0 571 463 372 14 105 184 9 222 329 127 214 5981 5020 4168 310 5000 281 Clycine 669 581 442 26 133 260 13 247 5463 2415 52 3731 5699 4742 1984 38 6383 2997 258 59 5424 2492 378 63 Serine Threonine 64 38 Cystine 6962 8022 39 1554 7126 3597 5464 Asparagine 31 52 4 34 67 Clutamine 8334 51 38 1451 3912 1774 56 1835 51 Tyrosine apartic Acid 4334 1920 378 136 229 59 Chinamic Acid 4779 4302 1926 714 2081 25 Lysine 751 Arginine 462 443 971 1061 1049 32 32 58 3339 Histidine Total \$4,566 37,355 45 691 32,628 4268 Faux, N. G. et al. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Res 15, 537-51 (2005). for eggs for data

Figure If your data are eggs, then the table is the egg crate, which keeps data ordered, separated and easily accessible. But, eggs aren't served out of egg crates ... perhaps data shouldn't be served out of a table either.

The purpose of a table is...

- To show the audience that you know how to run a regression?
 - OR
- To showcase your findings?

Problems with Regression Tables in Particular

- Too much irrelevant information
- Difficulty of reading the effect sizes ('is a significant result really significant?')
- Easier to interpret standard errors when displayed visually
- Unable to identify individual cases (are there leverage points?)
- Masks heteroskedasticity in the data
- Ultimately can leave more questions than answers

The purpose of a table

• If the purpose is to show a relationship between two variables, then ... the best thing is to show the relationship between those variables.

• For example, a post-regression residual plot.

• If you have a treatment and control group, then just show the difference in means and the standard error around the estimates. Such a (simple) graphic is almost mandatory in many disciplines (hard sciences)
Don't distract the viewer



[BusinessWeek, Oct 4-10]

I am pulling for the new management to do a better job, but I'm not very hopeful. Want to make this chart better? Simply remove the third dimension and make it a simple, plain old line graph. If I had to guess, I'd say this might just be a case where the functionality lets me create it in the software, so it must be good, right? Wrong. Maybe Stephen Few is right in that great data visualizations start with well designed software that do not let the creator choose an option like this.

What do you think? Blame it on poorly designed software or uneducated designers? As a substitute for substance, one can try lots of color, 3D effects, or disguised redundancy. This graph uses all three techniques, to display just five numbers. Note the clever use of mirror-imaging -- the top series is just (100 - the bottom series) and the interesting use curved lines, front and back to avoid the appearance that there's a lot less here than meets the eye.

Tufte (1983, p.118) says, "This may well be the worst graphic ever to find its way into print."



Alternative version:

AGE STRUCTURE OF COLLEGE ENROLLMENT

Percent of Total Enrollment 25 and Over

1972	•	•	•	'	'	'	'	'	•	•	'	•	•	'	•	•	•		28	3.0			
1973	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		2	9.2		
1974	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	32.8	
1975	·	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	• •	3 3.6	3
1978																					•	33.0	



https://wikis.uit.tufts.edu/confluence/display/comp15004visualization/Margarita+Parasi

When Graphic Design Goes Badly

The most general standards of charting data are thus the following:

- Present meaningful data.
- Define the data unambiguously.
- Do not distort the data.
- Present the data efficiently.

To see what happens when these rules are violated, consider figure 3, taken from Robert Putnam's Bowling Alone (where it is labeled figure 47), a work that contains many good and bad examples of graphical data display (and unfortunately, no tables at all). In just one chart, Putnam violates the three fundamental rules of data presentation: the chart does not depict meaningful data; the data it does depict are ambiguous, and the chart design is seriously inefficient. One can't accuse Putnam of distorting the data only because his main conclusions are not derived from the data presented in the chart.





Figure 3: Very bad graphical display

Of these, let's consider the inefficiency first: the first thing you notice about the chart is that the graphical elements are represented in three dimensions. On both efficiency and truthfulness this is unfortunate; the 3-D effect is entirely unnecessary and in this case serves to distort the visual representation of the data. Had not the data labels been shown on the top of each bar, it would not be readily apparent that column A is in fact bigger than column F, or that C is the same size as B. In addition the chart suffers from what might be called "numbering inefficiency": Putnam uses 13 numbers to represent 6 data points. Eliminating the 3-D, as shown in figure 4, offers a more exact representation of the data with a lot less ink.

http://www.synergy.com/klass.htm



Note the extraneous features of this in this graphic:

•A completely irrelevant map of the world.

•Two entirely different kinds of 3-D charts displayed at two different perspectives.

•Country names are repeated three times.

•To display 24 numeric data points, 28 numbers are used to define the scales.

The countries are sorted in no apparent order (not even alphabetically).
Note the use of the letter " I " to separate the countries on the bottom chart.



http://lilt.ilstu.edu/gmklass/pos138/datadisplay/images/phillips1.jpg

This image, from the graphic design book, *Diagraphics II*, attempts to show the relative market shares of Sotheby's vs. Christies over time.



So, this example shows the use of several tricks to show that *time* is increasing over time.

http://www.math.yorku.ca/SCS/Gallery/

Principles of Infographic Design



'a species understands just enough of the constant and calculable to base a scheme of behaviour upon it' - Nietzsche

Map projections







Clive Billson,"A History of the London Tube Maps" http://homepage.ntlworld.com/clivebillson/tube/tube.html reality is already quite complex

a graphic is by definition

a simplification of reality, so:

show what you want to show

and nothing else

How Long Do Your Relationships Usually Last?

people who use twitter every day vs. everyone else



source: 833,987 OkCupid users

Overlapping priorities

Percentage of Chinese who say they are a person for whom it is important to...



Born before one-child policy

Born under one-child policy

Data source:

World Values Surveys, 2010-2013: www.worldvaluessurvey.org. Percentage of respondents who said that a person for whom it was important "to have a good time" and "to be rich was 'very much' or 'somewhat like me'

"Oligarchy Index"





Principles of data graphics design [Tufte]

Above all else show the data

Maximize the data-ink ratio

Erase non-data-ink

Erase redundant data-ink

Revise and edit

The Lie Factor

size of graphic size of data

- Can be used to exaggerate
 differences or similarities
 → deceptive graphics
- Use it when evaluating your own designs

Minimize it

THE SHRINKING FAMILY DOCTOR In California

Percentage of Doctors Devoted Solely to Family Practice



"The shrinking family doctor in California", Los Angeles Times, p. 3, August 5, 1979

Chartjunk

Decorative elements that provide no information provide distractions



Stephen Few, "Common Mistakes in Data Presentation", August 7, 2004 http://intelligent-enterprise.informationweek.com

Chart Junk Considered Useful After All

Robert Kosara; April 22, 2010 Criticism Design Representation Visual Communication



There is almost universal agreement that any extraneous elements in a chart or visualization, elements that do not represent numbers, are detrimental to understanding the data. A paper that was presented at CHI recently described a study to figure out just how bad all this chart junk really was. As it turns out, it's actually rather helpful.

The Study

Quick show of hands: which chart is better, this one:



Or this one:





If you subscribe to the Tuffe school of thinking, the first chart probably made you cringe. But imagine somebody who just looks at this chart on a newspaper page for a few seconds before reading another article. Which chart is more likely to get its message across?

9. Children 12-17 Years Old Who Have Repeated Grades by Residential Characteristics

(Percent of children repeating grades)



Survey of Income and Program Participation (SIPP), US Census Bureau, April 1998

http://lilt.ilstu.edu/gmklass/pos138/datadisplay/chart_clutter_examples.htm

striking visual communication ('getting noticed')



This map drawn by Charles Joseph Minard portrays the losses suffered by Napoleon's army in the Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen, the thick band shows the size of the army (422,000 men) as it invaded Russia. The width of the band indicates the size of the army at each position. In September, the army reached Moscow with 100,000 men. The path of Napoleon's retreat from Moscow in the bitterly cold winter is depicted by the dark lower band, which is tied to temperature and time scales. The remains of the Grande Armée struggled out of Russia with 10,000 men. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and temperature on various dates during the retreat from Moscow. It may well be the best statistical graphic ever drawn. Napoleon's March poster \$14 postpaid; English/French version \$18 postpaid.

Frenchmen



totally screwedonly somewhat screwed

General guidelines for graphing data

Maximize the data density

Minimize the data-ink ratio

Minimize the lie factor

Be mindful of the visual metaphor

Show data in context

Label well



M. Ericson, New York Times via Hanspeter Pfister, Harvard



TOP 100 TOTAL

132.632.524

89,34

No Rights Reserved This work is placed in the Public Domain



Source: http://www.mint.com/blog/finance-core/visualizing-uschina-trade-relations/

When Sea Levels Attack!

Which cities will flood when?

years	sea level		
8000	80m	8000 years	Antarctic ice sheet (South Pole) 61m
1000	20m	the trê dat	Greenland Ice sheet 7m West Arctic Ice sheet 6m
	8m 7m	New York London Taiwan	Heating ocean expanding 1m per century
400	6m Sm Sm	***	
300	South London 4m		
200	3m		l l
	2m Houlywood Hamburg St Petersburg Manhattan	80 years	
100	Venice	Sec.	Already happened 20-40cm

)avid McCandless // v1.0 // Jan 2010 Illustrations: Laura Sullivan & Joe Swainson

InformationIsBeautiful.net

note: Heights above sea level vary across cities. Lowest points used. source: IPCC, NASA,

NewScientist.com,

Sea Level Explorer

Potsdam Institute,

1

Find Ways to 'Layer' Data

What if there are too many variables/dimensions?

In order to add extra dimensions to data, we can find ways to layer using colours, size, and time

This helps remove text



Western Europe
 Eastern Europe
 Middle East / North Africa
 Sub-Saharan Africa
 South Asia
 East Asia
 Australia / Oceania
 North America
 Central America
 South America
 South America

source: 1,701,312 OkCupid users. GDP per capita data from the IMF. R² = .47 for the best-fit line.

Largest Cities by Population, 1500-2012



factblink

Data source:

Chandler, T. (1987) Four Thousand Years of Urban Growth: An Historical Census, Lewiston, NY: Edwin Mellen Press Geohive Statistics (2012) http://www.geohive.com

curvy women vs. skinny women




Find shortcuts to visualisation

e.g. automated mapping software







40 LEGIBILITY AND SIMPLIFICATION

42 LEGIBILITY AND SIMPLIFICATION



3. Novoselok village before the Stolypin Reform

5. Novoselok village after the Stolypin Reform

Source: James Scott, Seeing Like a State pp. 40-42

Nature and Space 41



4. Khotynitsa village before the Stolypin Reform



6. Khotynitsa village after the Stolypin Reform

Source: James Scott, Seeing Like a State pp. 41-43

The Rules

- 1. Focus on the **message**
- 2. Show the data, not the design
- 3. Eliminate excess **text**
- 4. Create multiple layers where needed
- 5. Minimise the **lie factor**
- 6. Find shortcuts to visualise data quickly

Refined Petroleum Exports









- · Data source: United Nations Conference on Trade and Development, 2002.
- *There were no net refined petroleum exports recorded for 138 territories
- · Refined petroleum includes gasolines, kerosene, distillates, lubricating oils and diesel fuels.
- See website for further information

MOST AND LEAST US\$ OF NET REFINED PETROLEUM EXPORTS

lank	Territory	Value	Rank	Territory	Value
1	Bahrain	3160	53	Bosnia Herzegovina	4.07
2	Kuwalt	1995	54	Kenya	3.66
3	United Arab Emirates	1285	55	Angola	3.44
4	Qatar	869	56	Australia	3.06
5	Trinidad & Tobago	853	57	Morocco	2.08
6	Netherlands	447	58	Peru	1.25
7	Singapore	380	59	India	0.48
В	Lithuania	278	60	Cameroon	0.20
9	Belgium	230	62	Democratic Republic of Congo	0.04
10	Libvan Arab Jamahiriya	225	61	Equatorial Guinea	0.04

US\$ worth of annual net refined petroleum exports per person living in that territory*



Petroleum refinement includes various steps: fractionation separates the hydrocarbon compounds within the crude petroleum; conversion changes the hydrocarbon structures; treatment to remove impurities and further separation processes; blending and formulating produces the finished products. The products of refinement include fuel oils, kerosene, gasoline and lubricating oils. Refined petroleum is 2.4% of worldwide earnings from exports.

The Middle East has the highest net refined petroleum exports (US\$), this region is where most extraction occurs and has the largest known oil reserves.

Territory size shows the proportion of worldwide net exports of refined petroleum (in US\$) that come from there. Net exports are exports minus imports. When imports are larger than exports the territory is not shown.

REGIONAL NET REFINED PETROLEUM EXPORTS



"The rapid economic growth of the 1960s was driven by expansion of energy-consuming industries based on lavish use of cheap Middle East oil." Yoko Kitazawa, 1990 Map 065

www.worldmapper.org @ Copyright 2006 SASI Group (University of Sheffield) and Mark Newman (University of Michigan)

This report was presented at the training methodological workshop "Economic and Social Changes: values effects across Eurasia".

March 31 - April 6, 2015 – Turkey.

http://lcsr.hse.ru/en/seminar m2015

Настоящий доклад был представлен на методологическом учебном семинаре «Экономические и социальные изменения: оценка эффектов по всей Евразии».

31 марта – 6 апреля 2015 года – Турция.

http://lcsr.hse.ru/seminar m2015