01 Introduction

36-721 Statistical Graphics and Visualization

Jerzy Wieczorek

9/1/15

Examples and context

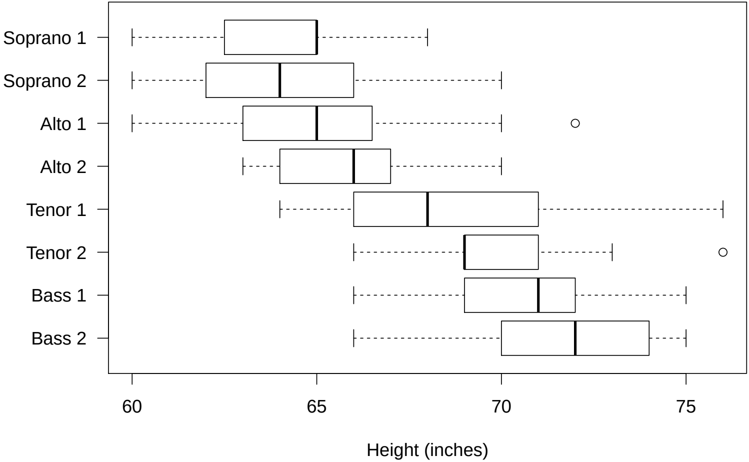
What good is data visualization?

What can we aspire to?

How do statistical graphics fit in with other flavors of visualization and information design?

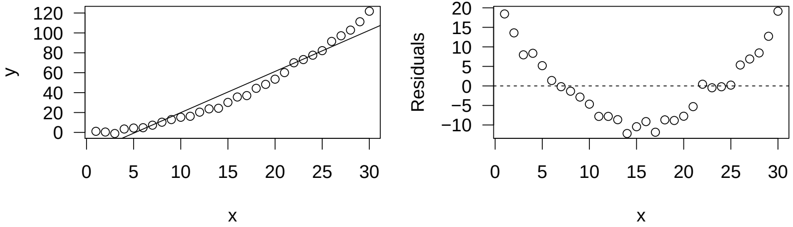
Statistical graphics

EDA



Statistical graphics

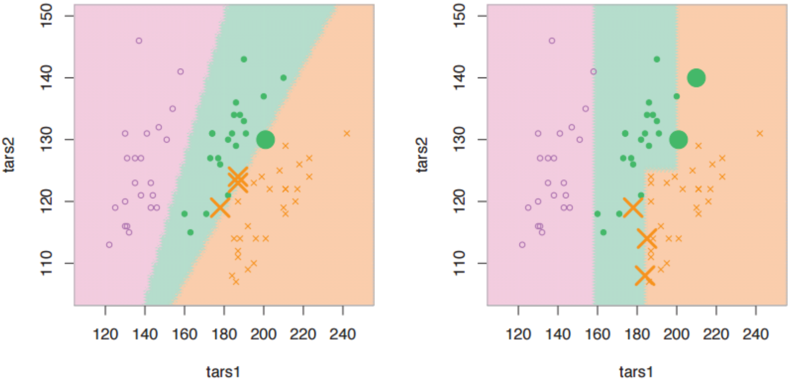
Regression diagnostics



Statistical graphics

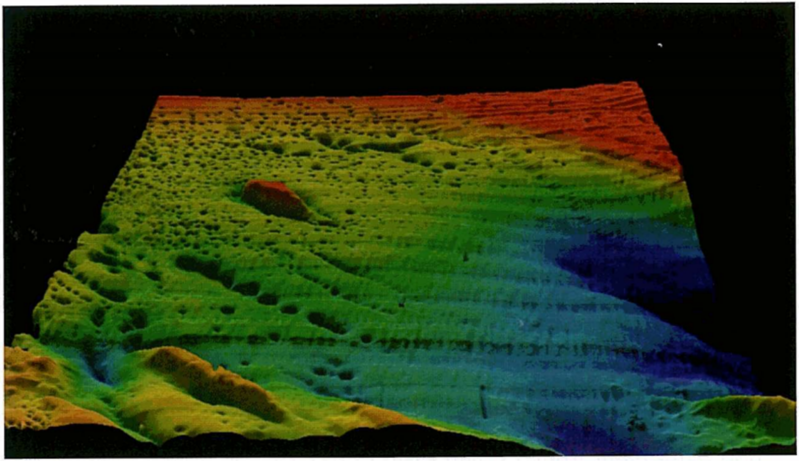
Classifier diagnostics

Cook & Swayne, *Interactive and Dynamic Graphics for Data Analysis, With R and Ggobi*



Scientific visualization

Ware, *Information Visualization*



Passamoquoddy Bay visualization. *Data courtesy of the Canadian Hydrographic Service.*

Statistical graphics and scientific visualization

- See a huge dataset all at once
- Find interesting features at different scales
- Notice data anomalies
- Propose scientific hypotheses

Infographics & data journalism

Static infographic by **Alberto Cairo**

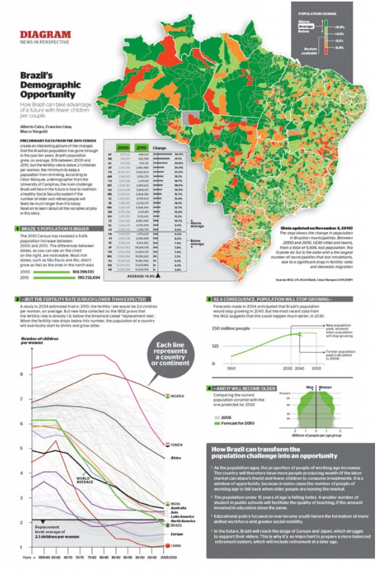


DIAGRAM
NEWS IN PERSPECTIVE

Brazil's Demographic Opportunity

How Brazil can take advantage of a future with fewer children per couple.

Alberto Cairo, Francine Lima, Marco Vergotti

PRELIMINARY DATA FROM THE 2010 CENSUS create an interesting picture of the changes that the Brazilian population has gone through in the past ten years. Brazil's population grew, on average, 10% between 2000 and 2010, but the fertility rate is below 2.1 children per woman, the minimum to keep a population from shrinking. According to César Marques, a demographer from the University of Campinas, the main challenge Brazil will face in the future is how to maintain a healthy Social Security system if the number of older and retired people will likely be much larger than it is today. Read on to learn about all the variables at play in this story.

1 BRAZIL'S POPULATION IS BIGGER

The 2010 Census has revealed a 9.4% population increase between 2000 and 2010. The differences between states, as you can see on the chart on the right, are noticeable. Most rich states, such as São Paulo and Rio, didn't grow as fast as the ones in the north east.

2000	169.799.170
2010	190.732,694

	2000	2010	Change
AP	477,032	648,553	36.0%
RR	324,397	425,398	31.1%
AC	557,526	707,125	26.8%
DF	2,051,146	2,469,489	20.4%
PA	6,192,307	7,443,904	20.2%
AM	2,812,557	3,350,773	19.1%
TO	1,157,098	1,373,551	18.7%
MT	2,504,353	2,954,625	18.0%
GO	5,003,228	5,849,105	16.9%
MS	2,078,001	2,404,256	15.7%
SC	5,356,360	6,175,603	15.4%
SE	1,784,475	2,036,277	14.1%
MA	5,651,475	6,424,340	13.7%
RN	2,776,782	3,121,451	12.4%
RO	1,379,787	1,535,625	11.3%
CE	7,430,661	8,180,087	10.1%
AL	2,822,621	3,093,994	9.6%
ES	3,097,232	3,392,775	9.5%
PB	3,443,825	3,753,633	9.0%
PI	2,843,278	3,086,448	8.6%
PE	7,918,344	8,541,250	7.9%
SP	37,032,403	39,924,091	7.8%
PR	9,563,458	10,266,737	7.4%
MG	17,891,494	19,159,260	7.1%
RJ	14,391,282	15,180,636	5.5%
BA	13,070,250	13,633,969	4.3%
RS	10,187,798	10,576,758	3.8%

(Data updated on November 4, 2010) *The map shows the change in population in Brazilian municipalities. Between 2000 and 2010, 1,630 cities and towns, from a total of 5,506, lost population. Rio Grande do Sul is the state with a the largest number of municipalities that lost inhabitants, due to a significant drop in fertility rates and domestic migration*

Sources: IBGE, UN, World Bank, César Marques (UNICAMP)

2 —BUT THE FERTILITY RATE IS MUCH LOWER THAN EXPECTED

A study in 2004 estimated that in 2010, the fertility rate would be 2.4 children per woman, on average. But new data collected by the IBGE prove that the fertility rate is already 1.9, below the threshold called "replacement rate". When the fertility rate drops below this number, the population of a country will eventually start to shrink and grow older.

3 AS A CONSEQUENCE, POPULATION WILL STOP GROWING—

Forecasts made in 2004 anticipated that Brazil's population would stop growing in 2040. But the most recent data from the IBGE suggests that this could happen much earlier, in 2030.

250 million people
125
0
1950
2030
2040
2050
New population peak: moment when population will stop growing
Former population peak (calculated in 2004)

4 —AND IT WILL BECOME OLDER

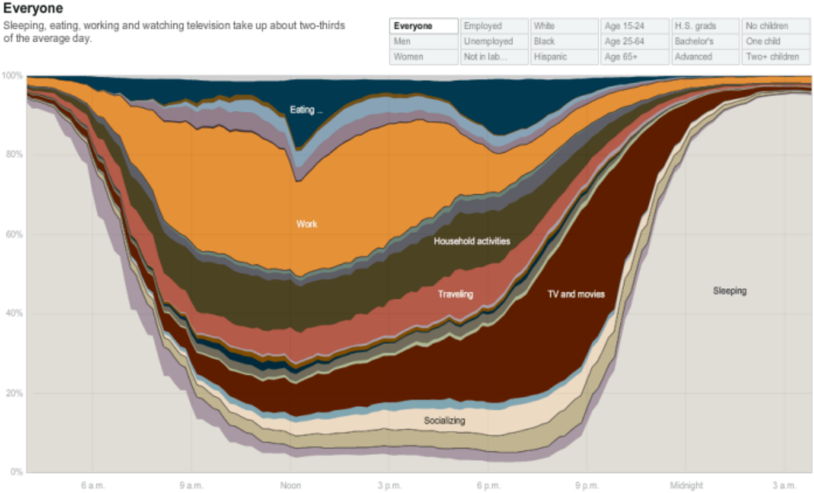
Comparing the current population pyramid with the one predicted for 2050

How Brazil can transform the population challenge into an opportunity

- As the population ages, the proportion of people of working age increases. The country will therefore have more people producing wealth (if the labor market can absorb them) and fewer children to consume investments. It is a window of opportunity, because in some cases the number of people of working age to fall back when older people are leaving the market.
- The population under 15 years of age is falling today. A smaller number of student in public schools will facilitate the quality of teaching, if the amount invested in education stays the same.
- Educational policy focused on low-income youth favors the formation of more skilled workforce and greater social mobility.
- In the future, Brazil will reach the stage of Europe and Japan, which struggle to support their elders. This is why it's so important to prepare a more balanced retirement system, which will include retirement at a later age.

Infographics & data journalism

Interactive graphic by **New York Times**



Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

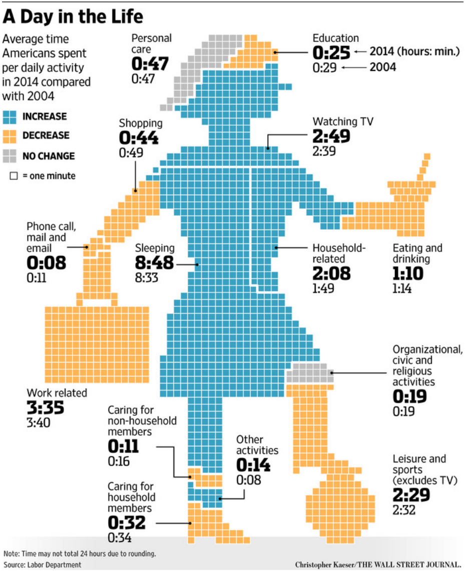
Animation and narration

Hans Rosling's TED talks

Watch **2:15-5:01 here**

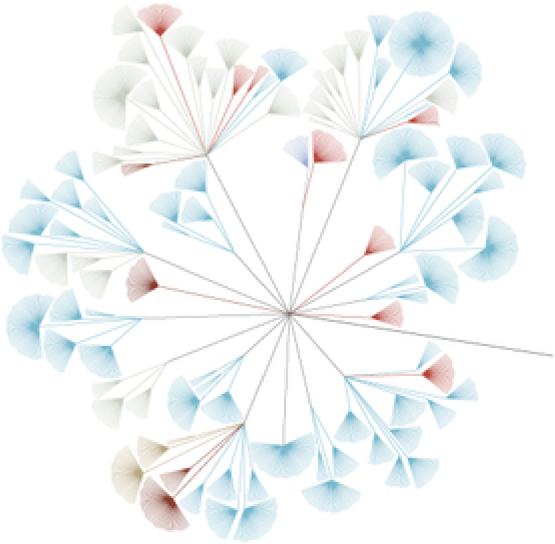
Data art

Wall Street Journal



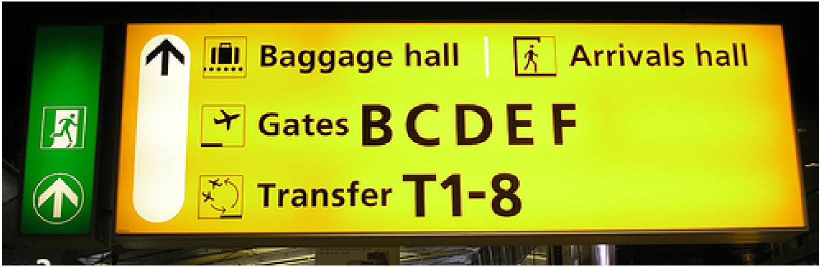
Data art

Stephanie Posavec



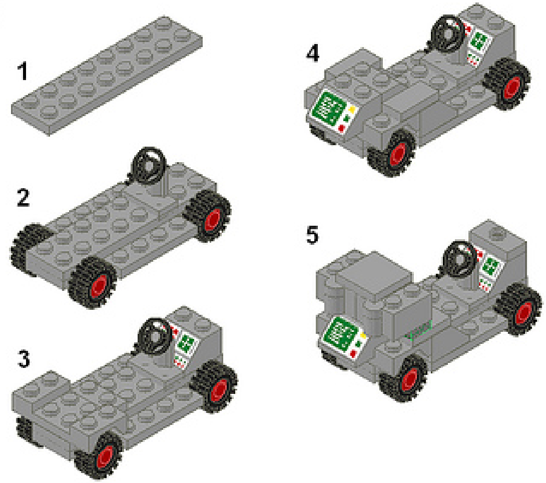
Information design, visual explanation

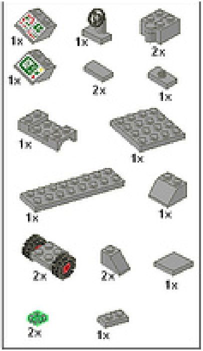
Wayfinding



Information design, visual explanation

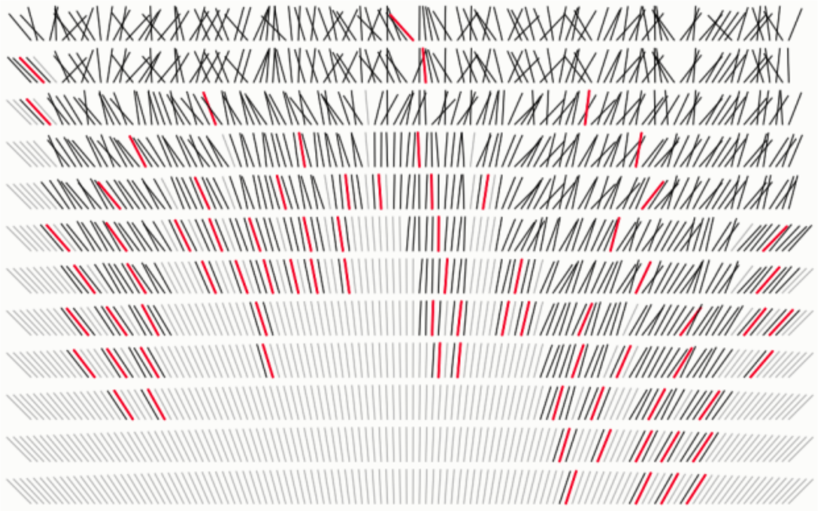
Instruction





Information design, visual explanation

Visualizing algorithms



Course info

How is the class organized? How is it graded?

How would you organize this course?

What topics to cover? How to group them?

How I've organized this course

Course objectives (see Syllabus) are based on frameworks and principles

In class we will

- Look at graphics
- Discuss whether they work
- Propose reasons why they (don't) work
- Compare to frameworks/principles in the literature
- Critique graphs and make new ones using our principles
- Learn software tools as needed for implementation

Assessments

Each assessment (homework, critique, project) targets one of the learning objectives. No points—just a rubric for each assignment.

Final grade depends on which assignments you completed and how well (see Syllabus).

You can revise and resubmit. But first submissions must be on time and show sincere effort! *(to keep grading manageable for us)*

Syllabus

- Office hours and due dates
- Objectives
- Texts and software
- Assessments
- Administrivia
- Schedule

Be sure to note:

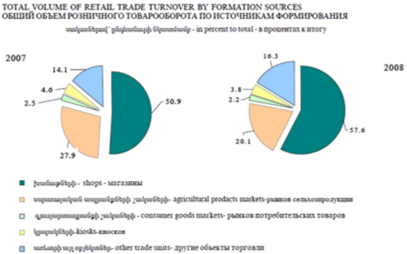
- Passing grade for CMU graduate students is B- or above.
- R software is required for Statistics MSPs.

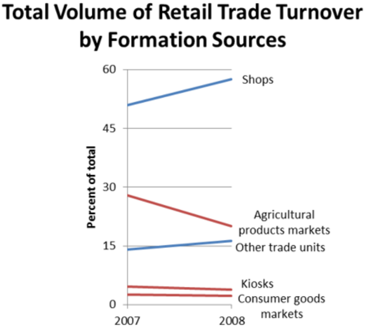
Schedule of topics

What will we study?

Readability and best practices

Armenian National Statistical Service and my own remake





Human visual perception

Cleveland and McGill (1984)

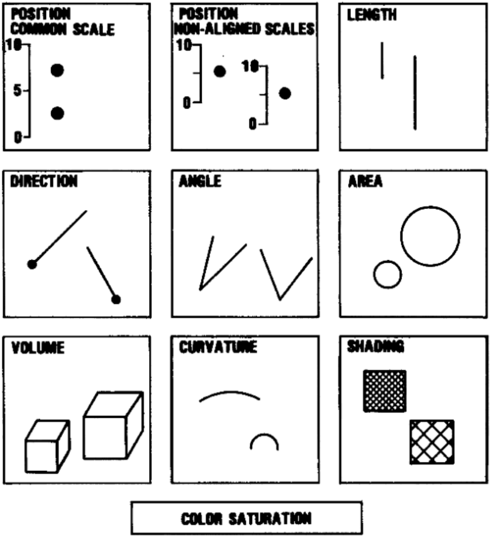
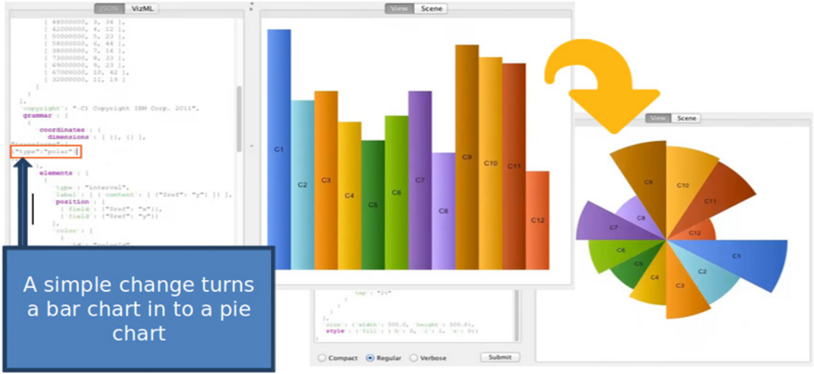


Figure 1. Elementary perceptual tasks.

The Grammar of Graphics

IBM's VizJSON, R's ggplot2, SPSS's GPL and Visualization Designer, Tableau. . .



```
        [ 48000000, 3, 34 ],
        [ 42000000, 4, 12 ],
        [ 50000000, 5, 23 ],
        [ 58000000, 6, 44 ],
        [ 38000000, 7, 14 ],
        [ 73000000, 8, 33 ],
        [ 69000000, 9, 23 ],
        [ 67000000, 10, 42 ],
        [ 32000000, 11, 19 ]
      ]
    }
  ],
  "copyright" : "(C) Copyright IBM Corp. 2011",
  "grammar" : [
    {
      "coordinates" : {
        "dimensions" : [ {}, {} ],
"transforms":[
{"type":"polar"}
]
      },
      "elements" : [
        {
          "type" : "interval",
          "label" : [ { "content" : [ {"$ref": "y"} ]} ],
          "position" : [
            { "field" : {"$ref": "x"}},
            { "field" : {"$ref": "y"}}
          ],
          "color" : [
            {
              "id" : "colorId",
```

```
              "top" : "2%"
            }
          }
        }
  ],
  "size" : {"width" : 500.0, "height" : 500.0},
  "style" : {"fill" : {"h" : 0, "l" : 1, "s" : 0}}
}
```

○ Compact ● Regular ○ Verbose Submit

A simple change turns a bar chart in to a pie chart

Graphic design

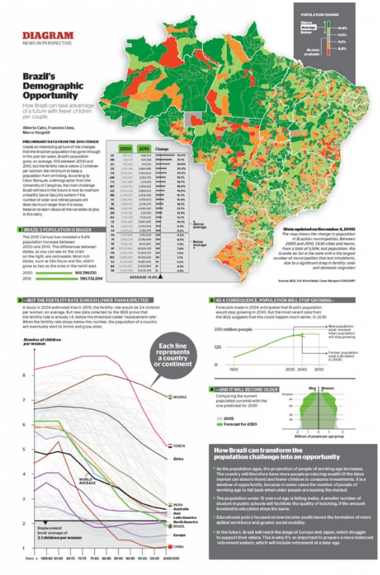


DIAGRAM
NEWS IN PERSPECTIVE

Brazil's Demographic Opportunity

How Brazil can take advantage of a future with fewer children per couple.

Alberto Cairo, Francine Lima, Marco Vergotti

PRELIMINARY DATA FROM THE 2010 CENSUS create an interesting picture of the changes that the Brazilian population has gone through in the past ten years. Brazil's population grew, on average, 10% between 2000 and 2010, but the fertility rate is below 2.1 children per woman, the minimum to keep a population from shrinking. According to César Marques, a demographer from the University of Campinas, the main challenge Brazil will face in the future is how to maintain a healthy Social Security system if the number of older and retired people will likely be much larger than it is today. Read on to learn about all the variables at play in this story.

1 BRAZIL'S POPULATION IS BIGGER

The 2010 Census has revealed a 9.4% population increase between 2000 and 2010. The differences between states, as you can see on the chart on the right, are noticeable. Most rich states, such as São Paulo and Rio, didn't grow as fast as the ones in the north east.

2000	169.799.170
2010	190.732,694

	2000	2010	Change
AP	477,032	648,553	36.0%
RR	324,397	425,398	31.1%
AC	557,526	707,125	26.8%
DF	2,051,146	2,469,489	20.4%
PA	6,192,307	7,443,904	20.2%
AM	2,812,557	3,350,773	19.1%
TO	1,157,098	1,373,551	18.7%
MT	2,504,353	2,954,625	18.0%
GO	5,003,228	5,849,105	16.9%
MS	2,078,001	2,404,256	15.7%
SC	5,356,360	6,175,603	15.4%
SE	1,784,475	2,036,277	14.1%
MA	5,651,475	6,424,340	13.7%
RN	2,776,782	3,121,451	12.4%
RO	1,379,787	1,535,625	11.3%
CE	7,430,661	8,180,087	10.1%
AL	2,822,621	3,093,994	9.6%
ES	3,097,232	3,392,775	9.5%
PB	3,443,825	3,753,633	9.0%
PI	2,843,278	3,086,448	8.6%
PE	7,918,344	8,541,250	7.9%
SP	37,032,403	39,924,091	7.8%
PR	9,563,458	10,266,737	7.4%
MG	17,891,494	19,159,260	7.1%
RJ	14,391,282	15,180,636	5.5%
BA	13,070,250	13,633,969	4.3%
RS	10,187,798	10,576,758	3.8%

(Data updated on November 4, 2010)
The map shows the change in population in Brazilian municipalities. Between 2000 and 2010, 1,630 cities and towns, from a total of 5,506, lost population. Rio Grande do Sul is the state with a the largest number of municipalities that lost inhabitants, due to a significant drop in fertility rates and domestic migration

Sources: IBGE, UN, World Bank, César Marques (UNICAMP)

2 —BUT THE FERTILITY RATE IS MUCH LOWER THAN EXPECTED

A study in 2004 estimated that in 2010, the fertility rate would be 2.4 children per woman, on average. But new data collected by the IBGE prove that the fertility rate is already 1.9, below the threshold called "replacement rate". When the fertility rate drops below this number, the population of a country will eventually start to shrink and grow older.

Each line represents a country or continent

Number of children per woman
8
7
6
5
4
3
2
1
Years ▸ 1955-60 60-65 65-70 70-75 75-80 80-85 85-90 90-95 95-00 00-05 2005-2010
WORLD AVERAGE
Replacement level: average of 2.1 children per woman
NIGERIA
YEMEN
Africa
INDIA
Australia
Asia
Latin America
North America
BRAZIL
Europe
CHINA

3 AS A CONSEQUENCE, POPULATION WILL STOP GROWING—

Forecasts made in 2004 anticipated that Brazil's population would stop growing in 2040. But the most recent data from the IBGE suggests that this could happen much earlier, in 2030.

250 million people
125
0
1950
2030
2040
2050
New population peak: moment when population will stop growing
Former population peak (calculated in 2004)

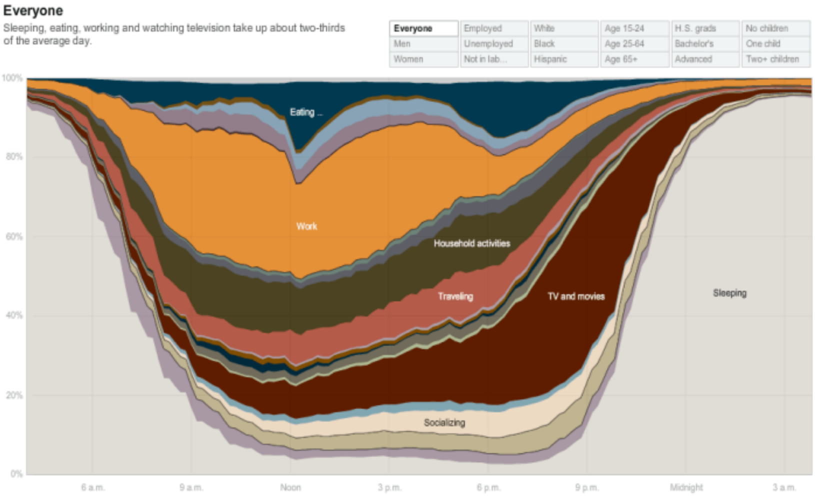
4 —AND IT WILL BECOME OLDER

Comparing the current population pyramid with the one predicted for 2050

How Brazil can transform the population challenge into an opportunity

- As the population ages, the proportion of people of working age increases. The country will therefore have more people producing wealth (if the labor market can absorb them) and fewer children to consume investments. It is a window of opportunity, because in some cases the number of people of working age to fall back when older people are leaving the market.
- The population under 15 years of age is falling today. A smaller number of student in public schools will facilitate the quality of teaching, if the amount invested in education stays the same.
- Educational policy focused on low-income youth favors the formation of more skilled workforce and greater social mobility.
- In the future, Brazil will reach the stage of Europe and Japan, which struggle to support their elders. This is why it's so important to prepare a more balanced retirement system, which will include retirement at a later age.

Interaction design

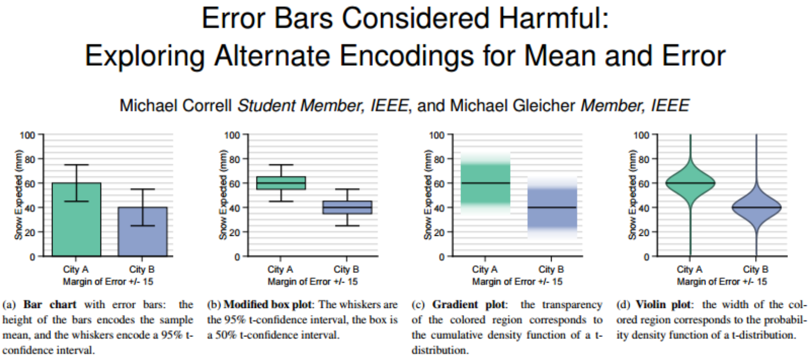


Everyone

Sleeping, eating, working and watching television take up about two-thirds of the average day.

The dataviz research literature

Correll and Gleicher (2014)



Error Bars Considered Harmful:
Exploring Alternate Encodings for Mean and Error

Michael Correll *Student Member, IEEE*, and Michael Gleicher *Member, IEEE*

100
80
60
40
20
0
Snow Expected (mm)
City A
City B
Margin of Error +/- 15

(a) **Bar chart** with error bars: the height of the bars encodes the sample mean, and the whiskers encode a 95% t-confidence interval.

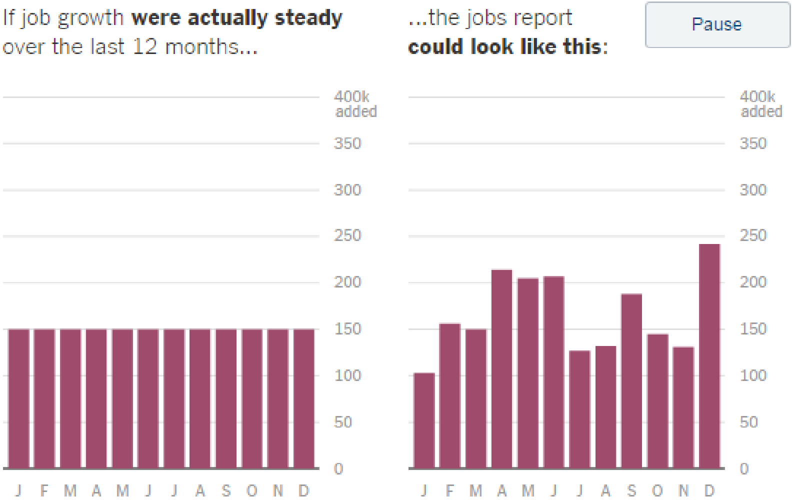
(b) **Modified box plot**: The whiskers are the 95% t-confidence interval, the box is a 50% t-confidence interval.

(c) **Gradient plot**: the transparency of the colored region corresponds to the cumulative density function of a t-distribution.

(d) **Violin plot**: the width of the colored region corresponds to the probability density function of a t-distribution.

Communicating statistical ideas

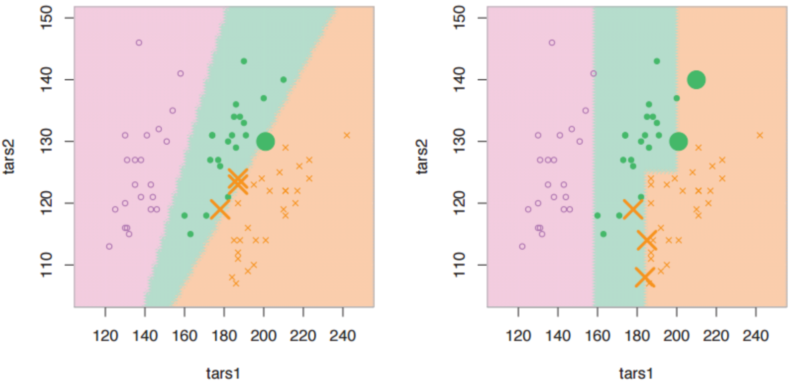
Sampling variation, via **New York Times**



If job growth **were actually steady** over the last 12 months...

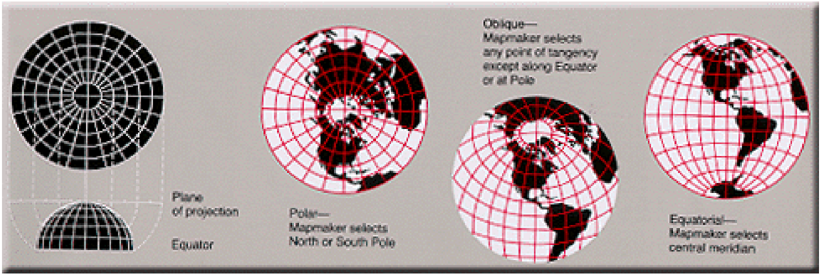
...the jobs report **could look like this:**

Graphics for statistical analysis



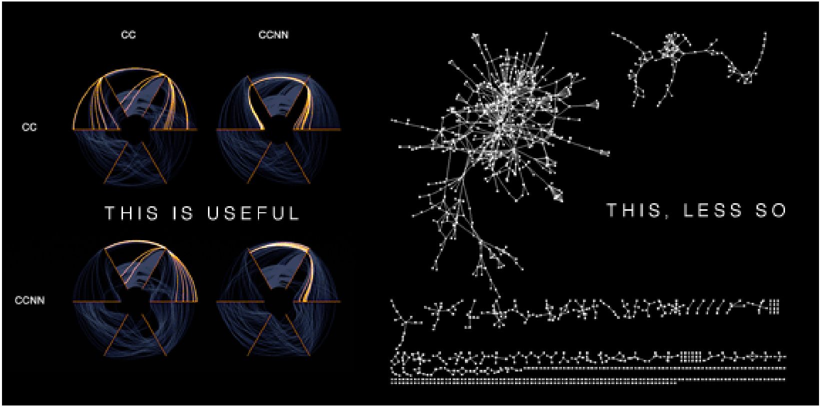
Maps and cartography

Map projections, choropleths, cartograms



Trees and networks

Hiveplot network diagrams



TBD—ideas

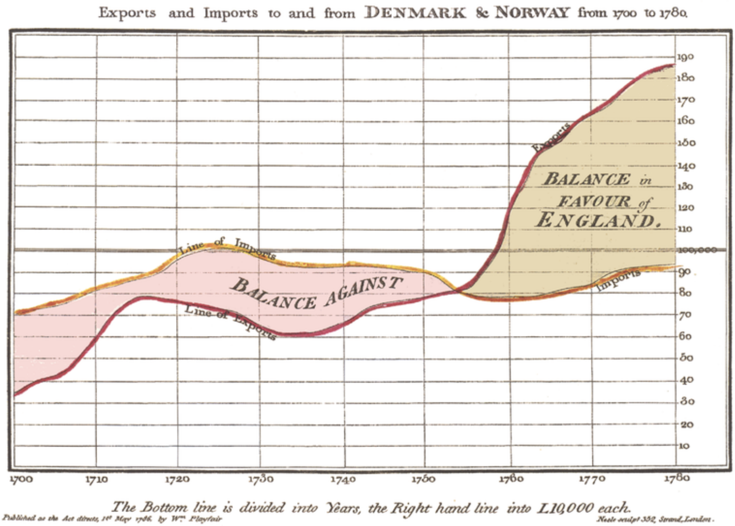
- Vector fields
- Data sonification
- D3.js practice
- Chart zoo
- Table design
- History of visualization
- *Your topic here?*

Historical classics

So you can nod and say, “Oh yeah, I know that one”

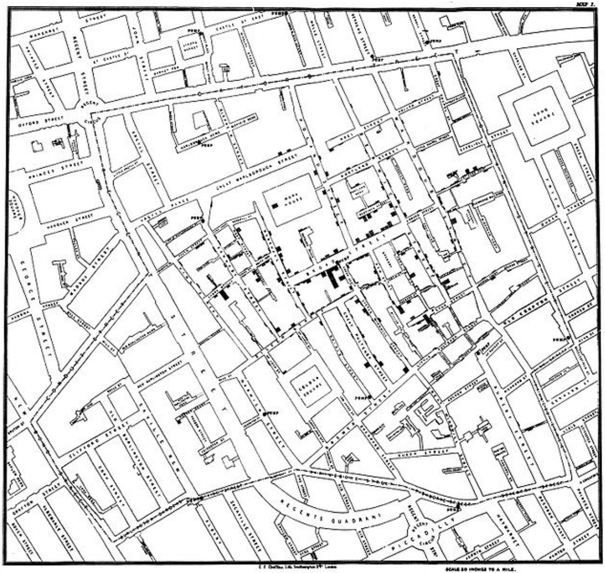
Playfair

Time series, 1786



Snow

Cholera map, 1854



Nightingale

Polar area diagram, 1858

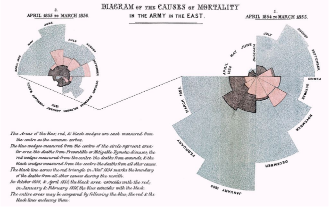
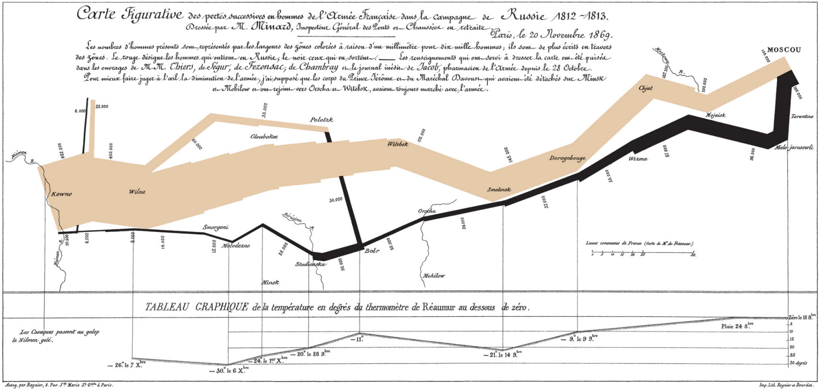


DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

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 area for area the deaths 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 Novr 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 enclosing them.

Minard

Napoleon's march, 1869



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.

Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

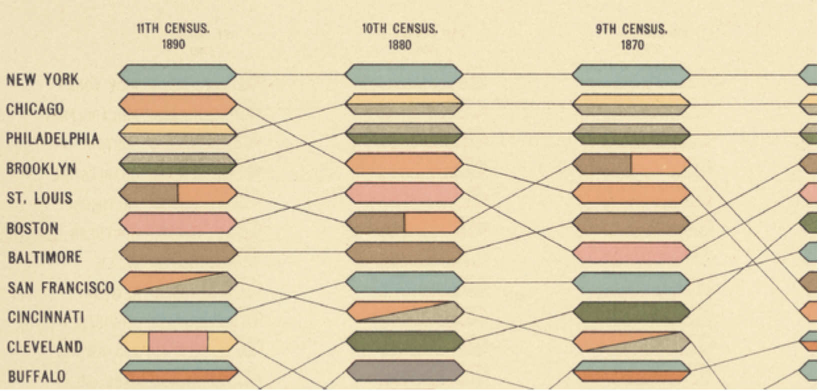
Les Cosaques passent au galop le Niémen gelé.

Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.

Imp. Lith. Regnier et Dourdet.

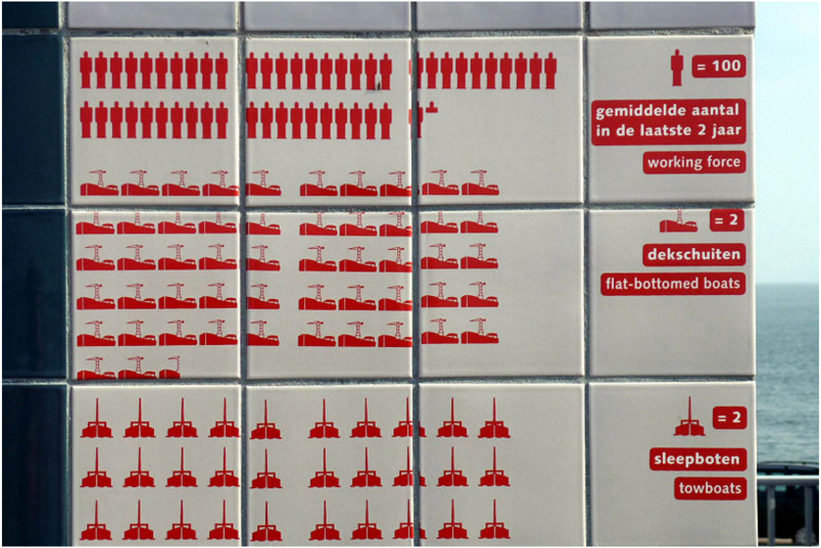
Statistical Atlas of the United States

Statistical Atlases by the U.S. Census Bureau, 1870-1890, including **ranks of city populations 1790-1890**



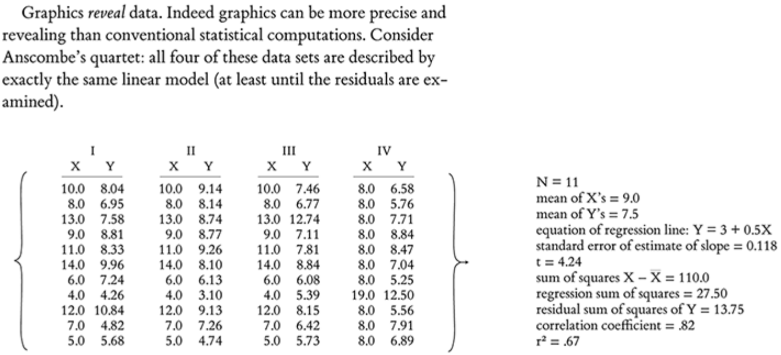
Neurath

Isotype, 1920s



Anscombe

Anscombe's quartet, 1973 (via **Tufte**, *The Visual Display of Quantitative Information*)

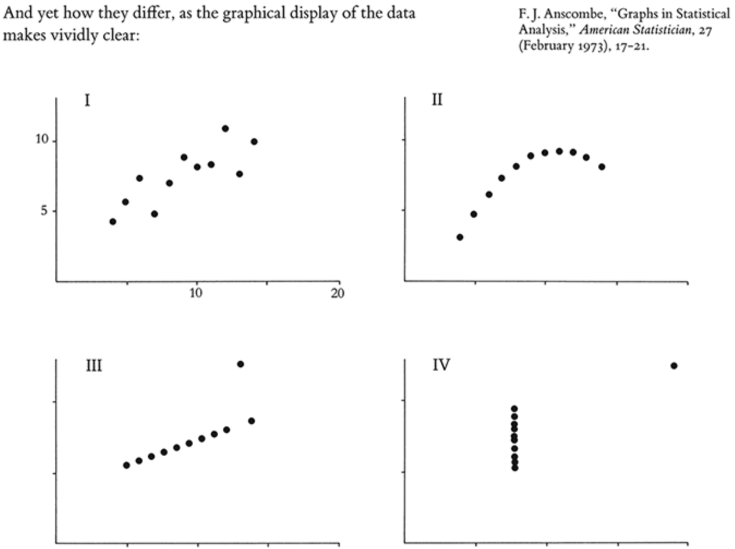


Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations. Consider Anscombe's quartet: all four of these data sets are described by exactly the same linear model (at least until the residuals are examined).

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

N = 11
mean of X's = 9.0
mean of Y's = 7.5
equation of regression line: Y = 3 + 0.5X
standard error of estimate of slope = 0.118
t = 4.24
sum of squares $X - \bar{X}$ = 110.0
regression sum of squares = 27.50
residual sum of squares of Y = 13.75
correlation coefficient = .82
r^2 = .67

Anscombe

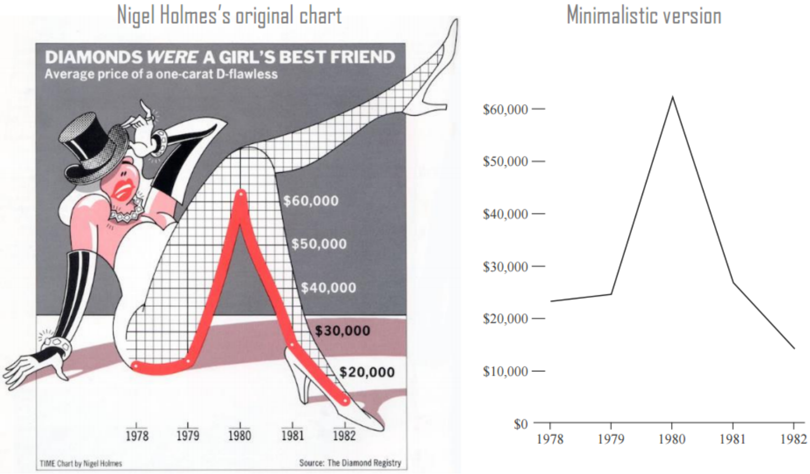


And yet how they differ, as the graphical display of the data makes vividly clear:

F. J. Anscombe, “Graphs in Statistical Analysis,” *American Statistician*, 27 (February 1973), 17–21.

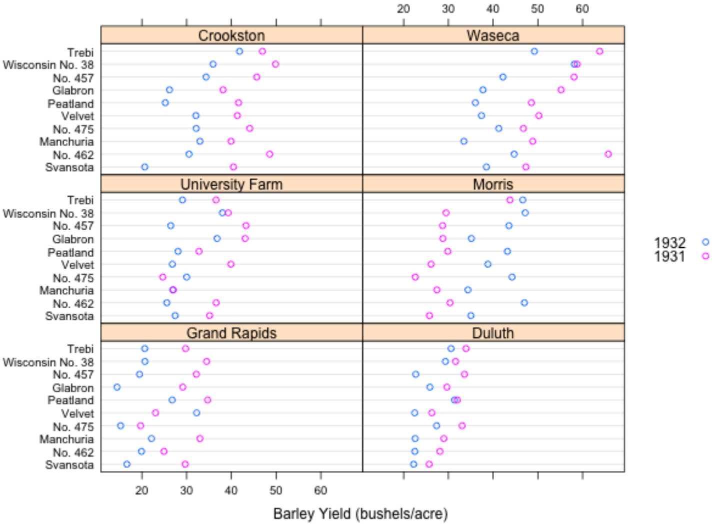
Holmes vs. Tufte

Holmes, *TIME*, 1982(?); Tufte, *Envisioning Information*, p. 34; Cairo, *The Functional Art*, p. 61-70



Cleveland

Trellis dotplot of **barley yield data** (Cleveland, 1993), though **see rejoinder**



Classic books

- Bertin, *Semiology of Graphics*
- Cleveland, *The Elements of Graphing Data* and *Visualizing Data*
- Wilkinson, *The Grammar of Graphics*
- Tufte, *The Visual Display of Quantitative Information* and *Envisioning Information*
- Wainer, *Visual Revelations* and others

Next time

What to prepare? What'll we cover?

Prepare for next time

- Install and test-drive your statistical graphics software
- Look at HW1 (data + rubric), let me know if unclear
- Readings: Cairo Ch 1-4; Donahue p. 1-23
- Blogs to follow
  - **Nathan Yau**
  - **Alberto Cairo**
  - **Robert Kosara**
  - **Kaiser Fung**
  - **Di Cook**

Next time we'll cover

- Best practices for most common 1D/2D charts and tables
- Image formats, resolution, saving plots
- A few handy tricks (logs, loess, jitter)
- R users: bring laptops to follow along

Software installation

Let's get everything installed and debugged to prepare for future classes.

Software installation

- Who'll use R? What else will be used?
- R users: install and test-drive
  - R and RStudio
  - `ggplot2`, `knitr`, `shiny` packages
- Tableau users: install student license
  - **www.tableau.com/academic/students**
- Also consider
  - D3.js
  - Inkscape