New developments in Statistics

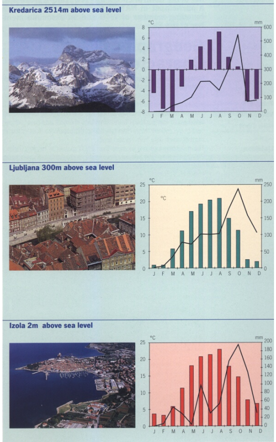
Data visualization

Vladimir Batagelj

IMFM Ljubljana and IAM UP Koper

PhD program on Statistics
University of Ljubljana, March 25, 2025

Outline



Vladimir Batagelj: vladimir.batagelj@fmf.uni-lj.si

Current version of slides (March 24, 2025 at 15 : 32): slides PDF

Measurement scales

A *measurement scale* is a function $f : A \to \mathbb{R}$ that is a homomorphism from the observed physical system into the corresponding relational system on $\mathbb{R}$. In data analysis, the measured values of a selected property on selected objects (*units*) form a *variable*.

Measurement scales measuring the same property are equivalent – for example, the price of an item expressed in USD or EUR. For usual scales, the equivalent scales can be obtained from each other using the *admissible transformations*. They determine different *scale types* (see next slide).

Scale types have an important role in data analysis – they determine what can we do with given data.

VB: Teorija merjenja. F.S. Roberts: Discrete Mathematical Models

Scale types

admissible transformations	**scale type**	**example**
$\varphi(x) = x$ (identity)	absolute	counting
$\varphi(x) = a.x, a > 0$ similarity	ratio	mass temperature (K)
$\varphi(x) = a.x + b, a > 0$	interval	temperature (C,F) time (calendar)
$x \geq y \Leftrightarrow \varphi(x) \geq \varphi(y)$ monotone increasing	ordinal	school grades, air quality, hardness
φ is one-to-one	nominal	hair color, nationality

Stevens (1946) *On the theory of scales of measurement*.

Data exploration and visualization

In the nineties with the development of multimedia computers the *data visualization* started to emerge as an important tool in the *data exploration* and the *presentation* of obtained results.

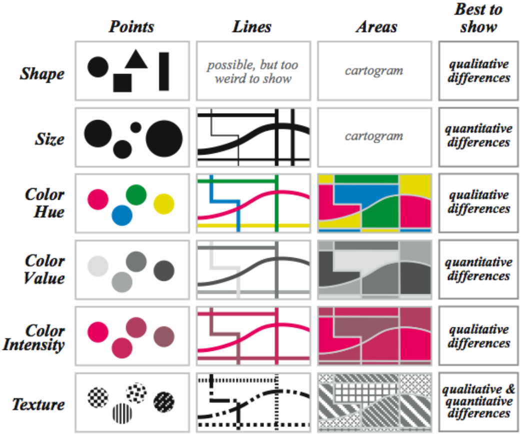
The main task of the *data analysis and visualization* is discovering of properties of (large) data sets using the capabilities of our (human) visual system on visualizations (nowadays produced by computer graphic systems) – *a picture is worth a thousand words*. The final goal is to get the *insight* into data and to understand the relations among their components.

Milestones

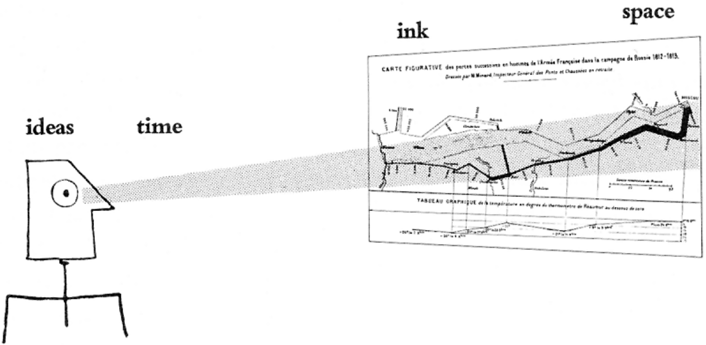
Clay tablets
Mesopotamia 3000 BC

Foundations of visualization

Theoretical foundations of data visualization were given by Jacques Bertin in his book *Sémiologie graphique* (1967).

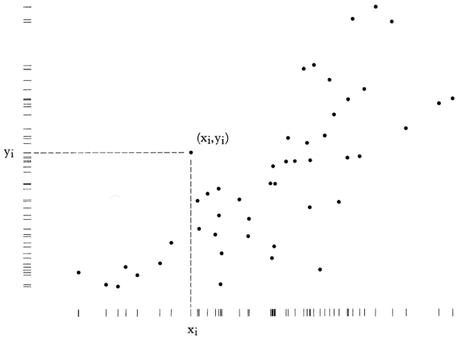


Tufte's visualization principles
(Edward R. Tufte)



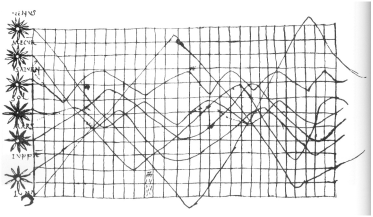
Graphical *excellence* consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*. It gives the viewer the greatest number of *ideas* in the shortest *time* with the least *ink* in the smallest *space*. It is telling the *truth* about the data.

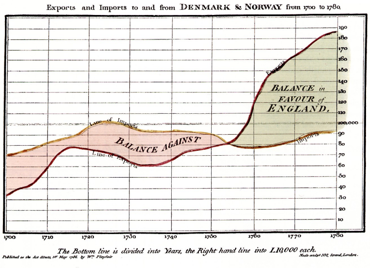
Tufte's principles



- Above all else show the data.
- Maximize the data-ink ratio.
- Erase non-data-ink.
- Erase redundant data-ink.
- Revise and edit.

Early examples





An old (10th century) attempt to display a continuous change – planetary movements shown as cyclic inclinations over time.

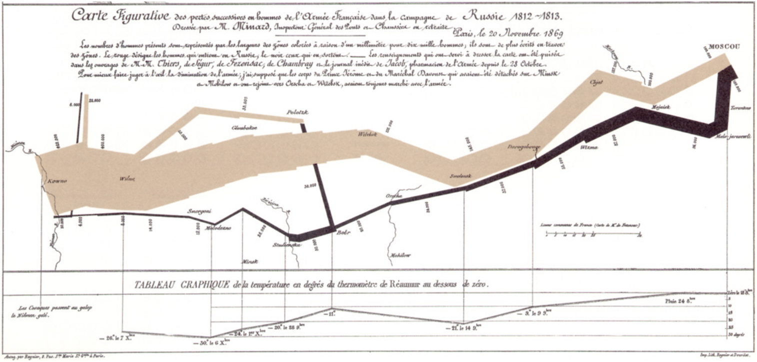
An early display of economic data (Playfair 1785) – *The Commercial and Political Atlas; Representing, by Means of Stained Copper-Plate Charts, the Exports, Imports, and General Trade of England, at a Single View.*

Example: Cholera

On 31 August 1854, after several other outbreaks had occurred elsewhere in the city, a major outbreak of cholera reached Soho. A physician John Snow eventually linked the outbreak to contaminated water.

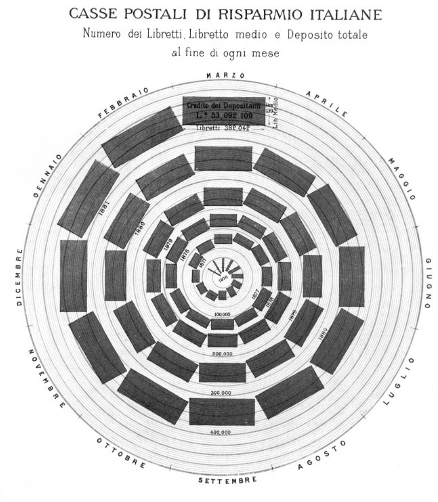
By talking to residents, he identified the source of the outbreak as the public water pump on Broad Street (now Broadwick Street). His studies of the pattern of the disease were convincing enough to persuade the local council to disable the well pump by removing its handle. This action has been commonly credited as ending the outbreak. Snow's study can be regarded as the founding event of the science of epidemiology.

Example: Napoleon's Russian campaign of 1812



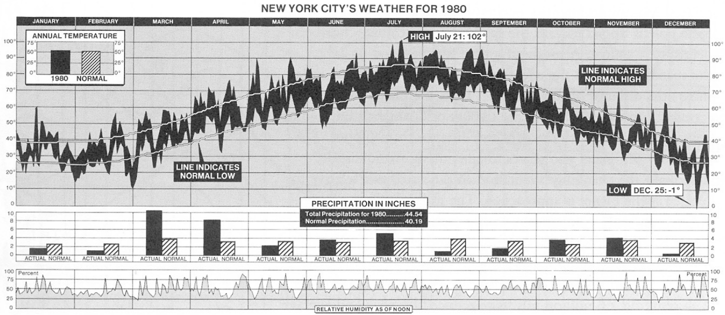
Charles Minard's (1861) map of Napoleon's disastrous Russian campaign of 1812-13. The graphic is notable for its representation in two dimensions of six types of data: the number of Napoleon's troops; distance; temperature; latitude and longitude; direction of travel; and location relative to specific dates.

Example: Growth of savings



The multivariate history of the Italian post office savings uses two dimensions, with the number of postal savings books issued and the average size of deposits multiplying up to total deposits at the end of each month from 1876 to 1881 (Gabaglio A.: Teoria Generale della Statistica.)

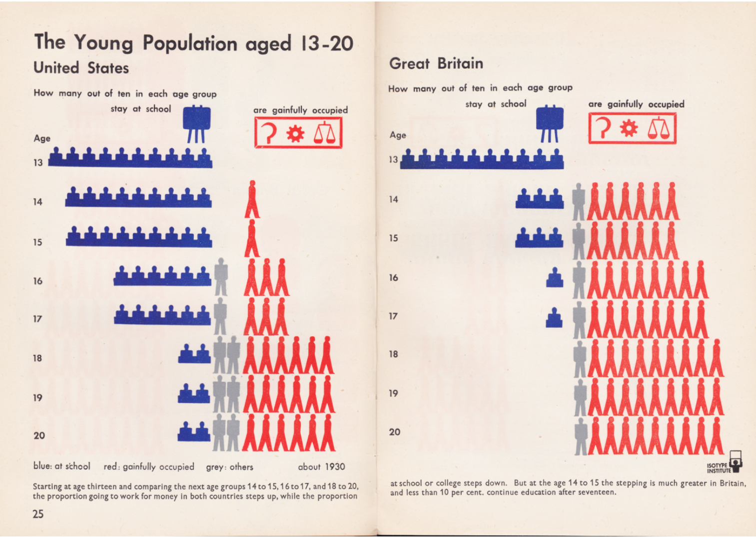
Example: The New York weather history 1980



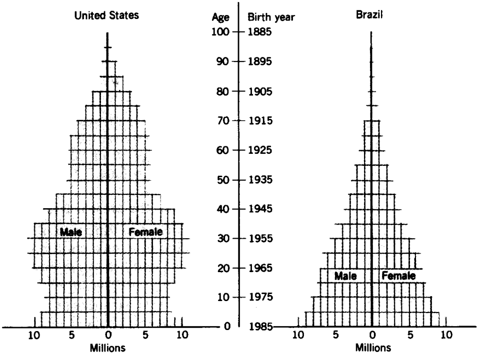
New York Times, January 11, 1981, p. 32.

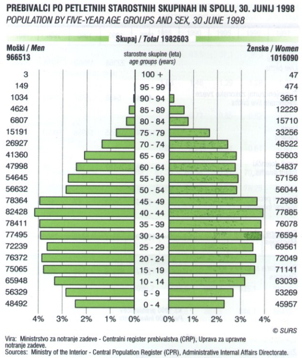
This graph, from the New York Times (Jan. 11, 1981) shows 2200 numbers that summarize the trends and patterns in weather in New York City in 1980. The three aligned charts show temperature, precipitation, and relative humidity. In the graph of temperature, the area is filled between the daily low and daily high.

Otto Neurath (1882 -1945) ISOTYPE



Example: Population pyramids



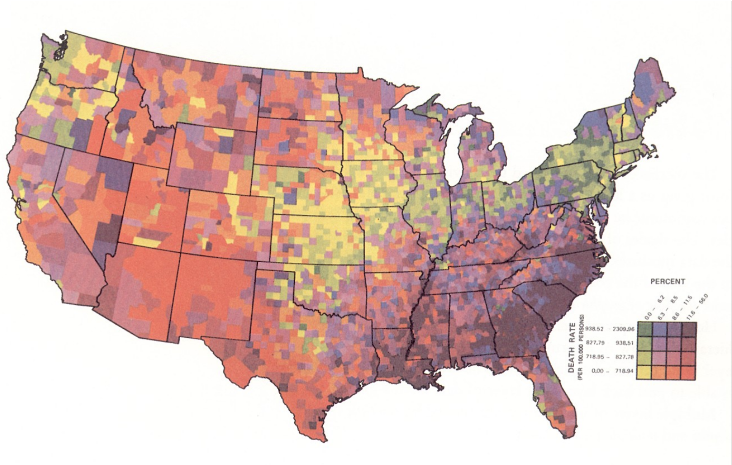


The shape of the pyramid contains information about the history and the level of development of a country.

London Tube

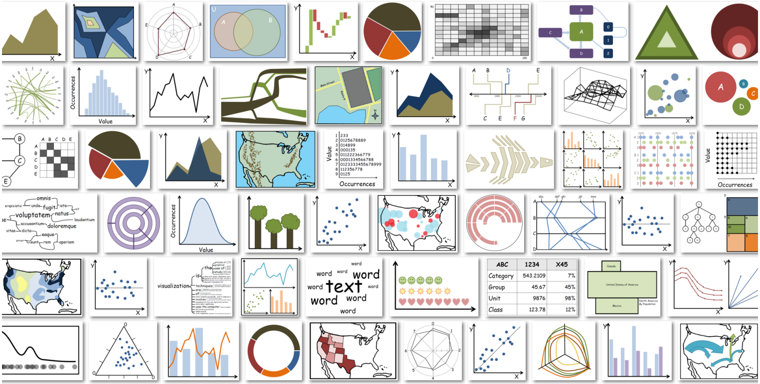
Tube map
Check before you travel
Key to lines
Bakerloo
Central
Circle
District
Hammersmith & City
Jubilee
Metropolitan
Northern
Piccadilly
Victoria
Waterloo & City
DLR
Emirates Air Line cable car (special fares apply)
London Overground
TfL Rail
London Trams
Key to symbols
Explanation of zones
MAYOR OF LONDON
tfl.gov.uk
24 hour travel information
0343 222 1234*
Sign up for email updates
tfl.gov.uk/emailupdates
@TfLTravelAlerts
UNDERGROUND
TRANSPORT FOR LONDON
EVERY JOURNEY MATTERS
Online maps are strictly for personal use only. To license the Tube map for commercial use please visit tfl.gov.uk/maplicensing

Example: Data on the map



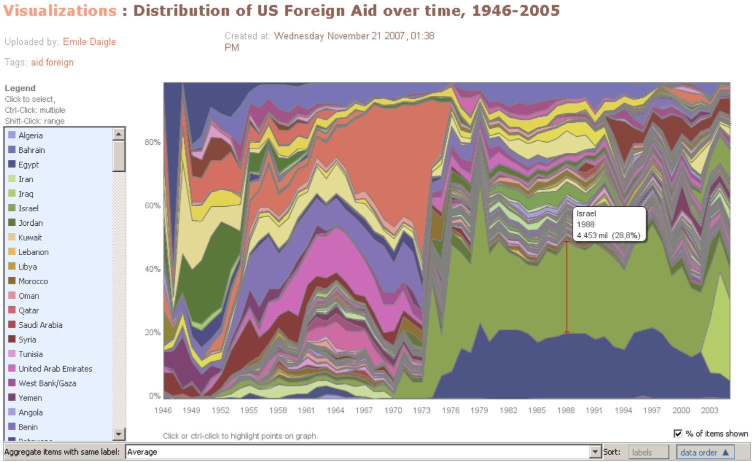
Global Administrative Areas.

Many charts

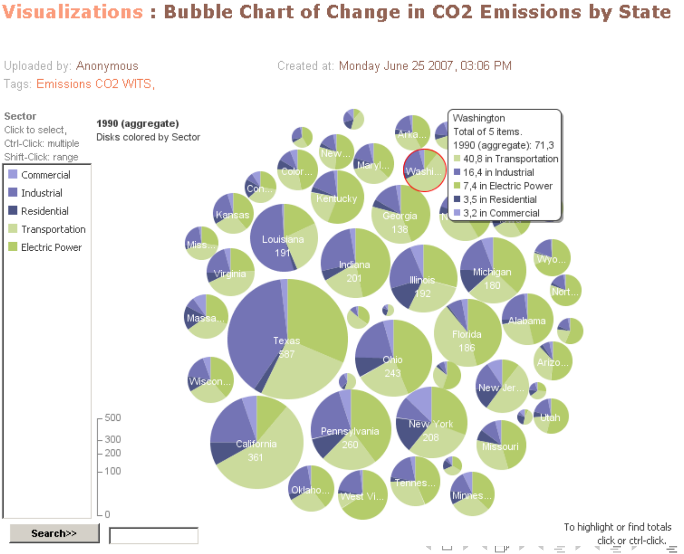


Example: ManyEyes

ManyEyes / IBM Visual Communication Lab: Martin Wattenberg, Fernanda Viégas, Frank van Ham, ... Open source alternative **RAW**Graphs



ManyEyes 2


Visualizations : Bubble Chart of Change in CO2 Emissions by State
Uploaded by: Anonymous
Created at: Monday June 25 2007, 03:06 PM
Tags: Emissions CO2 WITS,
Sector
Click to select,
Ctrl-Click: multiple
Shift-Click: range
Commercial
Industrial
Residential
Transportation
Electric Power
1990 (aggregate)
Disks colored by Sector
500
300
200
100
0
Search>>
Washington
Total of 5 items.
1990 (aggregate): 71,3
40,8 in Transportation
16,4 in Industrial
7,4 in Electric Power
3,5 in Residential
3,2 in Commercial
Texas 587
California 361
Pennsylvania 260
Ohio 243
New York 208
Indiana 201
Illinois 192
Louisiana 191
Florida 186
Michigan 180
Georgia 138
Kentucky
Alabama
Missouri
Virginia
Kansas
Utah
To highlight or find totals
click or ctrl-click.

ManyEyes 3

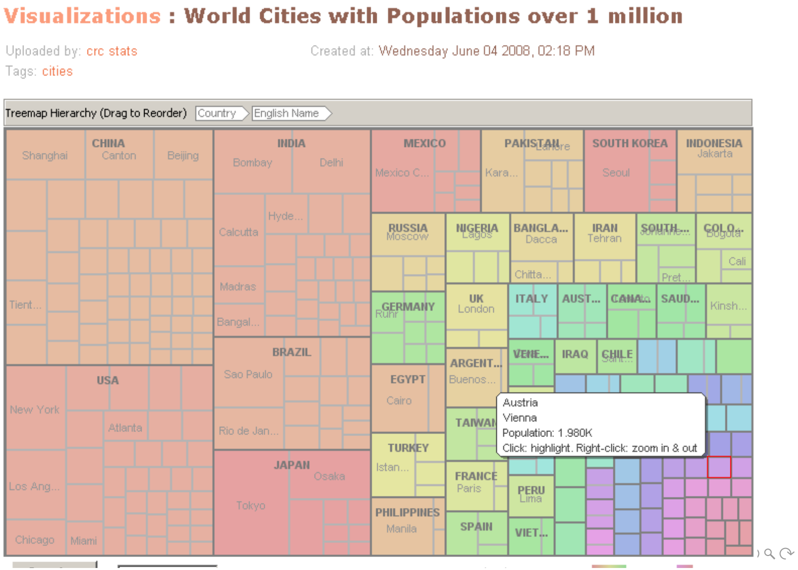
Visualizations : The Future Begins

Uploaded by: Susanne

Created at: Tuesday January 20 2009, 01:26 PM

Description:

ManyEyes 4

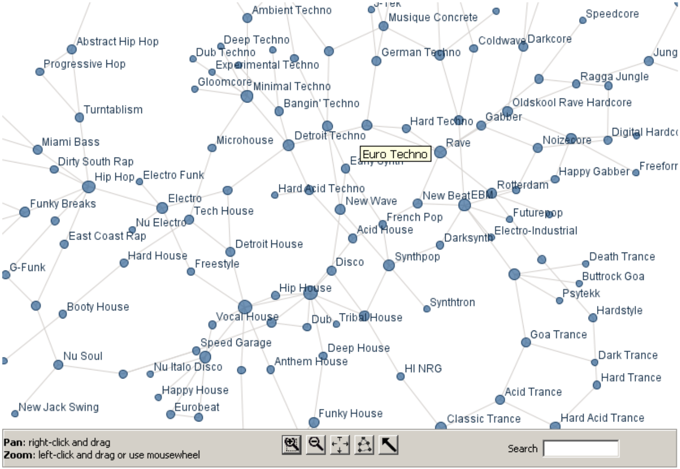

Visualizations : World Cities with Populations over 1 million
Uploaded by: crc stats
Created at: Wednesday June 04 2008, 02:18 PM
Tags: cities
Treemap Hierarchy (Drag to Reorder)
Country
English Name
CHINA
Shanghai
Canton
Beijing
Tient...
INDIA
Bombay
Delhi
Calcutta
Hyde...
Madras
Bangal...
MEXICO
Mexico C...
PAKISTAN
Lahore
Kara...
SOUTH KOREA
Seoul
INDONESIA
Jakarta
RUSSIA
Moscow
NIGERIA
Lagos
BANGLA...
Dacca
Chitta...
IRAN
Tehran
SOUTH...
Pret...
COLO...
Bogota
Cali
GERMANY
Ruhr
UK
London
ITALY
AUST...
CANA...
SAUD...
Kinsh...
USA
New York
Atlanta
Los Ang...
Chicago
Miami
BRAZIL
Sao Paulo
Rio de Jan...
JAPAN
Osaka
Tokyo
EGYPT
Cairo
ARGENT...
Buenos...
VENE...
IRAQ
CHILE
TAIWAN
TURKEY
Istan...
FRANCE
Paris
PERU
Lima
PHILIPPINES
Manila
SPAIN
VIET...
Austria
Vienna
Population: 1.980K
Click: highlight. Right-click: zoom in & out

ManyEyes 5

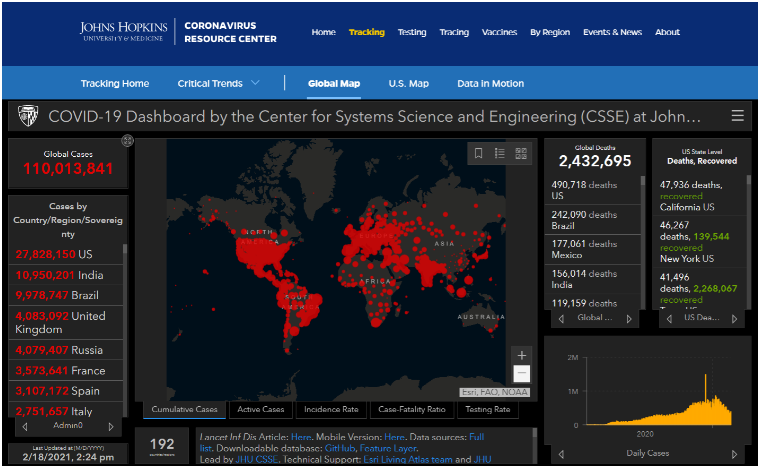
Visualizations : The Family Tree of Electronic Music

Uploaded by: Belarius Created at: Monday August 04 2008, 06:44 AM

Tags: music music,techno,electronic



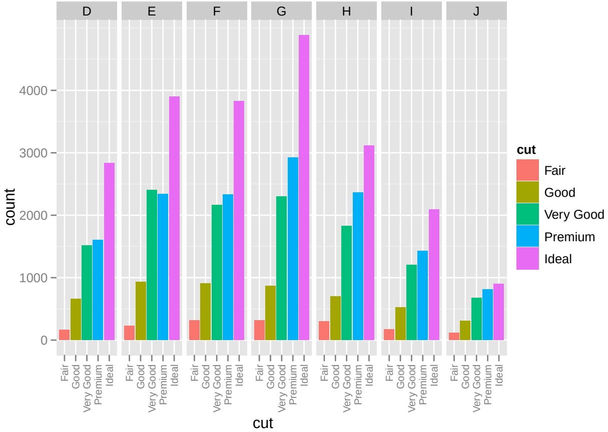
Dashboards



jhu

Example: ggplot2

```
> library(ggplot2)
> qplot(cut, data = diamonds, geom = "bar", fill = cut) + facet_grid(. ~ color) +
+  opts(axis.text.x = theme_text(angle = 90, hjust = 1, size = 8, colour = "grey50"))
> ggsave("cut.pdf", width = 7, height = 5)
```



Software support – libraries

In the programming system R some visualization libraries are available. Besides the basic library `graphics` we can use also a variant of the S-library `lattice`. A (future basic) library `grid` is under development. 3D interactive visualization are supported by libraries `rgl` and `rggobi` (an interface to program GGobi).

Special visualizations of results (clustering dendrograms, maps, etc.) are included also in other libraries.

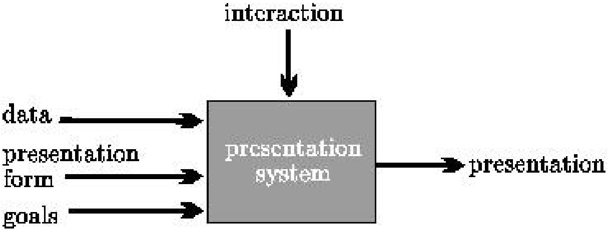
Leland Wilkinson wrote a fundamental book on data visualization The grammar of graphics and with his collaborators he developed a corresponding program nViZn (for SPSS).

Wilkinson's approach is based also on the R-library `ggplot2`, `ggraph`. For a visualization of data on the web a javascript library d3 was developed. The data for d3 scripts are usually supplied in JSON format. Tableau, BI, Plotly (R, python, javascript), Vega, seaborn, infographics (Isotype / Neurath), dashboards, ...

For visualization of data streams a sedimentation approach was proposed.

Visualization system

With the growth of the computing power of desktop computers, data visualization is gaining popularity among researchers as a tool for data exploration and the presentation of results. Using a data visualization system a researcher usually adopts different goals to: identify, locate, distinguish, categorize, cluster, rank, compare, associate, or correlate some data.



Output devices

Older visualization literature is dominated by the 'sheet of paper' paradigm – the solutions and techniques are mainly based on the assumption that the final result is a static picture on a sheet of paper. The computer screen is a different medium, which offers many new possibilities:

- parallel views (global and local);
- interactivity: brushing and linking, zooming and panning, views, 3D effects (movement);
- temporary elements (additional information about the selected elements, labels, legends, markers, etc.), highlighted selections, sounds; and others;
- program support (computation, searching, animation, . . .).

These features can and should be maximally leveraged to support data analytic tasks; or repeating Shneiderman's mantra: overview first, zoom and filter, then details on-demand (extended with relate, history, and extract). The future is interactive, dynamic, 3D, multiview displays on the computer screen(s).

3D printing. GGobi

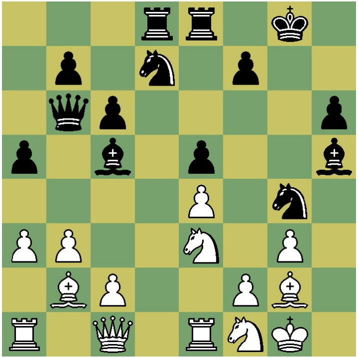
Graphic elements

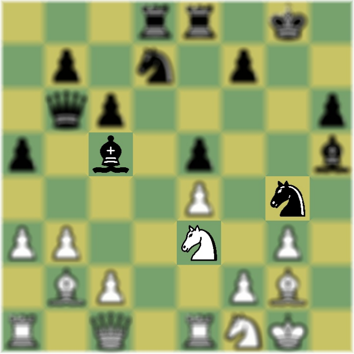
Graphic elements are used to represent data attributes (variables) values

- **fills**: color (hue), saturation, brightness, sharpness, pattern/texture;
- **spatial elements**:
  - general: length, size, shape, location, direction, symmetry, position (front/ back), rotation, distance (between elements)
  - special 3D: occlusion, perspective, transparency, depth, movement
- **dimension**: dot, curve, shape, solid (3D shape)

Graphic elements are used to construct different data representations. catalogue.
colors in R

Graphic elements – depth





Semantic Depth of Field

Weber-Fechner law

Psychophysics studies the relations between our perceptions and the corresponding physical stimuli. Ernst Weber (1795-1878) noticed that the *just-noticeable difference* ψ between two stimuli is proportional to the magnitude of the *stimuli* φ, (and the subject's sensitivity)

$$d\psi = k\frac{d\varphi}{\varphi}$$

The resolution of perception diminishes for stimuli of greater magnitude.

From this Gustav Theodor Fechner (1801-1887) derived the relation

$$\psi = k \ln \frac{\varphi}{\varphi_0}$$

where φ_0 is that threshold of stimulus below which it is not perceived at all: vision: $k = \frac{1}{60}$; pain: $k = \frac{1}{30}$; smell: $k = \frac{1}{4}$; taste (salt): $k = \frac{1}{3}$.

Example: the perception of a difference of two lengths depends on their ratio (proportion); and not on their difference. It can be substantially increased by adding a grid. Wikipedia

Stevens' power law

Experiments ($\sim$ 1930) showed that in most cases Fechner's law doesn't hold. Instead the *power law* proposed by Stanley Smith Stevens (1906–1973) holds:

$$\psi = k(\varphi - \varphi_0)^c$$

where k, c and φ_0 are constants.

For perceptions/stimuli with $c \neq 1$ our perceptions are biased.

Experimentally the following values of c were determined: length 0.9-1.1, area 0.6-0.9, volume 0.5-0.8.
Wikipedia

Luce's theorem

Theorem (Luce 1959) Let ψ be a continuous function and the scales p and f satisfy the conditions from the table, then the function ψ has to satisfy the functional equation and has the solution given in the table

f	p	equation	solution ψ
$f : A \to \mathbb{R}^+$ ratio	$p : A \to \mathbb{R}^+$ ratio	$\psi(kx) = K(k)\psi(x)$ $k > 0,\ K(k) > 0$	$\psi(x) = \alpha x^\beta$
$f : A \to \mathbb{R}^+$ ratio	$p : A \to \mathbb{R}$ interval	$\psi(kx) = K(k)\psi(x) + C(k)$ $k > 0,\ K(k) > 0$	$\psi(x) = \alpha \ln x + \beta$ or $\psi(x) = \alpha x^\beta$
$f : A \to \mathbb{R}$ interval	$p : A \to \mathbb{R}^+$ ratio	$\psi(kx + c) = K(k, c)\psi(x)$ $k > 0,\ K(k, c) > 0$	$\psi(x) =$const
$f : A \to \mathbb{R}$ interval	$p : A \to \mathbb{R}$ interval	$\psi(kx + c) =$ $K(k, c)\psi(x) + C(k, c)$ $k > 0,\ K(k, c) > 0$	$\psi(x) = \alpha x + \beta$

All scales p are considered in the wide sense.

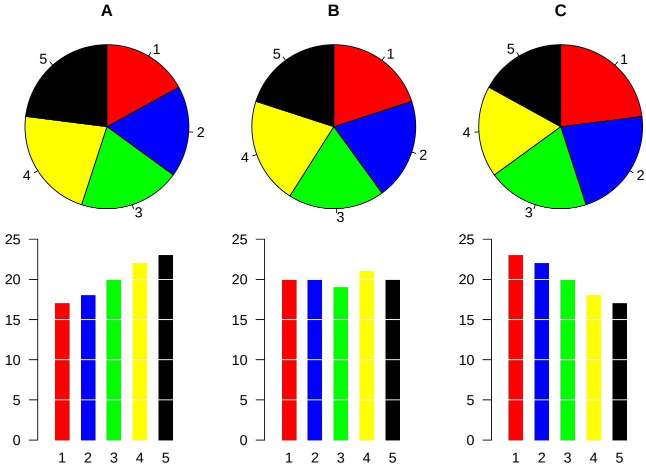
Stevens' power law – implications

Stevens' law implies the following ordering of graphical elements concerning the quality of perception:

- location to a common axis (best)
- location on two not-aligned axes
- length
- angle – inclination
- area
- volume
- color (hue) – saturation – brightness (worst)

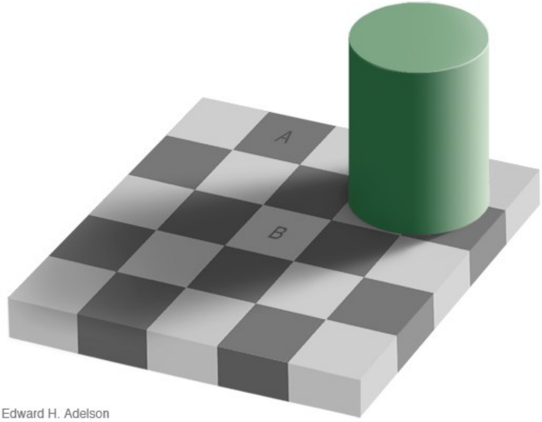
People are different: color blindness, habits, training, education, ...

Stevens' power law – implications



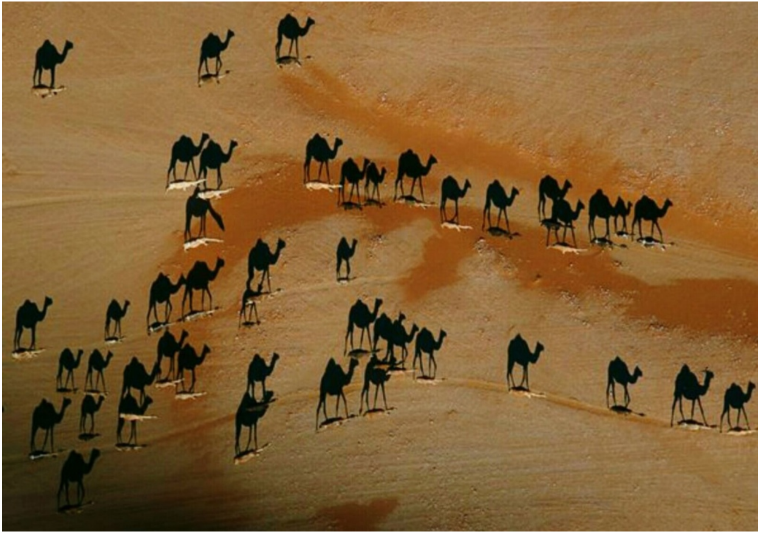
Wikipedia: Pie chart

Illusions, culture

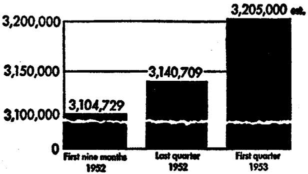


Blog, 8 best, Colors

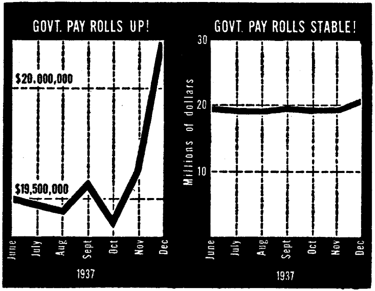
George Steinmetz: Camels



Carefully



From an April 24, 1953, newspaper advertisement for Collier's



In journals only interesting parts of a data chart are often presented. We have to read/observe it carefully – don't trust the first impression.

Data visualization

V. Batagelj

Scale types

Visualization

Examples

Tools

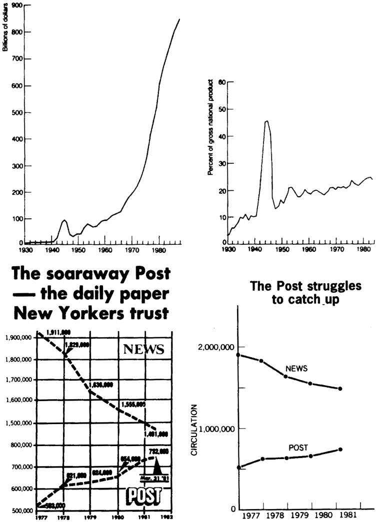
Foundations

Carefully

Dimensions

Hans Rosling

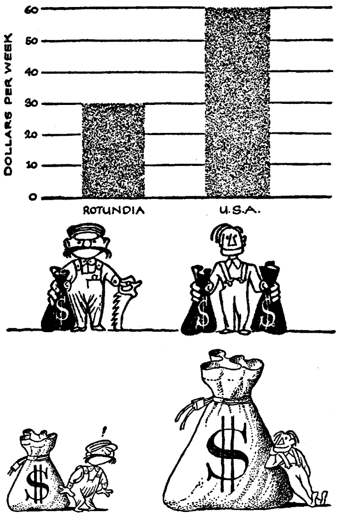
Resources



When the meaning of the data changes through time we have to put them on the “common denominator”.

Continuous axes are mis-leading.

. . . carefully: journal

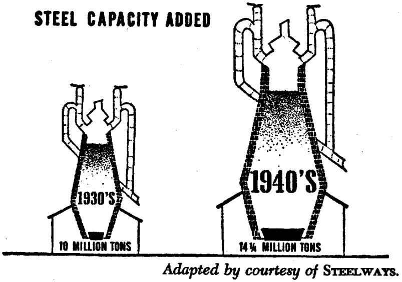


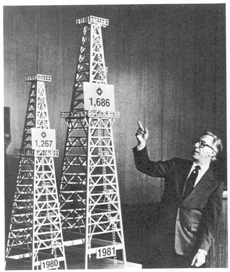
To increase the attractiveness of a representation it is (in journals or presentations for the general public) often enhanced by using pictures.

But, be careful: a shape with double height is perceived as four times larger.

Graphical illusions

Our perception strongly depends on the properties of our (human) visual system. They should be considered in the design of data representations.. Optical illusions 1, 2, 3.
Example: Steel production (from Darrrell H.: *How to Lie with Statistics*).

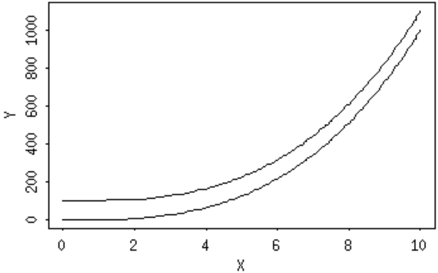




The right icon is 1.4 times higher than the left icon. A viewer is not comparing heights but areas – he perceives that the steel production doubled ($= 1.4^2$).

. . . illusions: two curves

Comparing two functions we have to look at their **vertical distance** – difference of their values. Our visual system is considering the **shortest/smallest distance** between the curves.



What was your *first impression* about presented functions? Are they “converging”?

Data sets

Properties of the data set that crucially influence forms of its representation are:

- *size*: small, large, unbounded;
- *density*: sparse, dense, cluttered; and
- *activity*: static, dynamic (deterministic, random).

Small data sets can be presented in totality and in detail in a single view. In an overall view of large data sets details are lost; and a detailed view can encompass only a part of the data set.

Exocentric and egocentric view

Standard data visualizations supported by general-purpose programs (Excel, PowerPoint, ...) are mostly *exocentric* – the viewer is positioned outside the presentation. Often the third dimension is used only to make the presentation fancier, and not to get better insight about the data.

The basic feature of VR (Virtual Reality) is the support of *egocentric* view – the user is immersed in the presentation as its active part; he can travel inside the data scene, and the view is determined by his position in the scene.

VR and visualization

Control

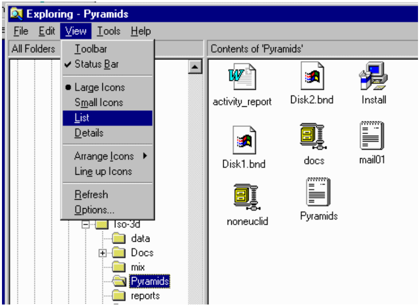
In an exploration of large data sets a serious problem appears: How to avoid being “lost within the forest”? Several solutions help the user’s orientation:

- *restart option*: returns the user to the starting position;
- introduction of additional *orientation elements*: coordinates display, grids, shadows, landmarks (static / user set). These elements can be switched on/off.
- *multiview*: consists of at least two views (windows/screens):
  - *map view*: overall view (usually exocentric) which contains the current position and allows ’long’ moves (jumps). For very large data sets it can be combined with zooming or fisheye.
  - *local view*: which displays the selected portion of the data set.

Additional support can be achieved by implementing **`trace`** / **`backtrack`** / **`replay`** mechanism and *guided tours*.

Glasses and lenses

Closely related to the multiview idea are the concepts of *glasses*, *lenses* and *zooming* (**Pad++**, **inXight**).



Selecting different glasses we obtain different views on the same data. Glasses affect the entire window, and lenses only on the selected region.

The display can be augmented with *supporting elements*: (temporary) labels, grids, infos, . . .

Visualization of multivariate data

In visualizing multivariate data we usually deal with small or large, sparse, and static data sets. Let $E = \{X_i\}$ be a *set of units*. A *unit* X is usually described by a list of values of selected attributes (properties) – *variables*

$$X_i = (V_1 = x_{i1}, V_2 = x_{i2}, \ldots, V_m = x_{im})$$

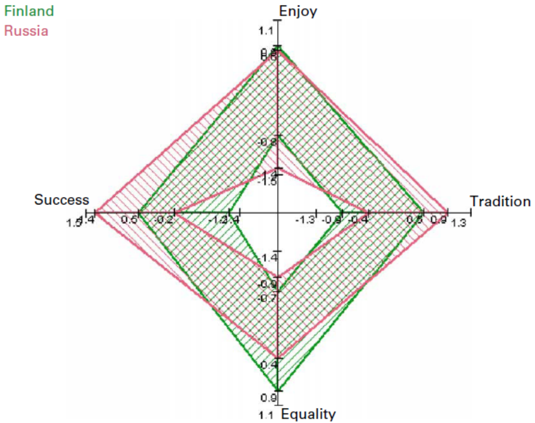
It is usually represented by a *glyph* which integrates, as its components, elements representing the unit's variables.

From standard data analysis, we know several types of 2D-glyphs: point in the plane, pie charts, bar charts, columns, stars, Chernoff faces, Andrews curves, . . . Most of the 2D-glyphs can be extended to 3D-glyphs, and some additional should be invented.

Graphical examination of data VisualComplexity

Zoom stars

For visualization of symbolic data **Zoom Stars** were developed (Noirhomme-Fraiture M., 1997).



Levels in the data

The representation elements support associative, selective, ordering and/or quantifying tasks. In the visualization task, there are several levels of detail represented by the hierarchy

$$\textit{attribute} \subset \textit{unit} \subset \textit{group} \subset \textit{groups} \subset \textit{data set}$$

Most data analysis procedures can be seen as transformations on or relations between these levels.

Representations and measurement scales

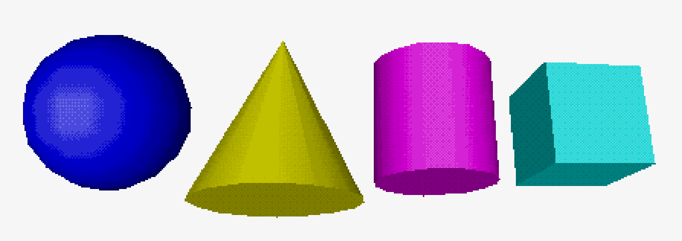
Different scale types are represented by different graphical elements:

scale	**representation**
nominal	color, shape
ordinal	grade, lightness, texture, arrangement (position)
numeric	size, location, direction, angle

Since *numeric* is included in *ordinal*, and ordinal is included in *nominal*, the representations compatible with higher scales could be used also for lower scales – e.g., direction to represent nationality. A general rule is that this should be avoided because they can suggest unsubstantial associations.

3D visualizations

For several years for describing 3D visualizations the web format VRML (Virtual Reality Modeling Language) was used. Later it was replaced by the new formats X3D, Kinemage, WebGL, . . .



. . . 3D visualizations

In a presentation of multivariate data different graphical elements can be used:

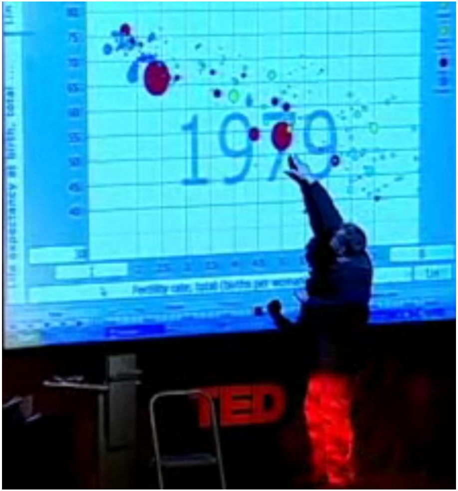
- location in space (x, y, z);
- shape (sphere, cube, cone, cylinder, plane, triangle, . . .);
- color;
- size, angle, slope, area, volume;
- pattern (texture);
- direction (orientation);
- text;
- lights (different light sources, shadowing, transparency, reflections, . . .);
- rotation of objects;
- different views and ways of moving in the obtained scene; camera properties (orthographic, perspective, stereoscopic; field of view);
- interactivity.

. . . 3D visualizations

Interactive visualization in the analysis of multiway networks. As an example, let's look at the EU airports and air companies (Cardillo, A. et al. (2013) Emergence of network features from multiplexity, Scientific Reports 3, 1344.) Clustering of airports.

A cube in the 3D layout represents a link/triple (ap_1, ap_2, co) - the airport ap_1 is linked to the airport ap_2 with a line provided by the company *co*. Lines provided by the same company are of the same color: All airports; 13-diversity core; Reordered core.

Hans Rosling (1948-2017) – Gapminder



- Gapminder
- Youtube: Hans Rosling
- The best stats you have ever seen
- New insights on poverty
- Let my dataset change your mindset
- Asia's rise – how and when
- The good news of the decade?

Data storytelling [12].

Resources I

- Jacques Bertin: Graphics and Graphic Information-Processing
- Edward R. Tufte: The Visual Display of Quantitative Information
- Leland Wilkinson: The Grammar of Graphics
- Colin Ware: Information Visualization: Perception for Design
- Matthew O. Ward, Georges Grinstein, Daniel Keim: Interactive Data Visualization: Foundations, Techniques, and Applications
- Paul Murrell: R Graphics
- Hadley Wickham: ggplot2: Elegant Graphics for Data Analysis
- Riccardo Mazza: Introduction to Information Visualization
- Noah Iliinsky, Julie Steele: Designing Data Visualizations

Resources II

- Jock Mackinlay: Information Visualization
- Richard Brath, David Jonker: Graph Analysis and Visualization
- Cole Nussbaumer Knaflic: Storytelling with Data: A Data Visualization Guide for Business Professionals
- Nancy Duarte: Slidedocs
- Steve Wexler, Jeffrey Shaffer, Andy Cotgreave: The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios
- Marti Hearst: Information Visualization and Presentation
- John Stasko: Information Visualization
- Wojciech Basalaj: Proximity Visualization of Abstract Data
- Chris North: Information Visualization

Resources III

- Gitta Domik: Computer-generated Visualization
- Huff D.: How to lie with statistics, Norton, New York, 1993/1954.
- Cleveland W.S.: The Elements of Graphing Data, AT&T, 1994
- Friendly M., Denis D.J.: Milestones in . . . Data Visualization
- Graham Wills: Visualizing Time
- Elijah Meeks: D3.js in Action
- Nick Qi Zhu: Data Visualization with D3.js Cookbook
- Chun-houh Chen, Wolfgang Härdle, Antony Unwin Handbook of Data Visualization
- Toward a Perceptual Science of Multidimensional Data Visualization

Resources IV

- A Periodic Table of Visualization Methods
- Datavisualization tools
- Types of charts