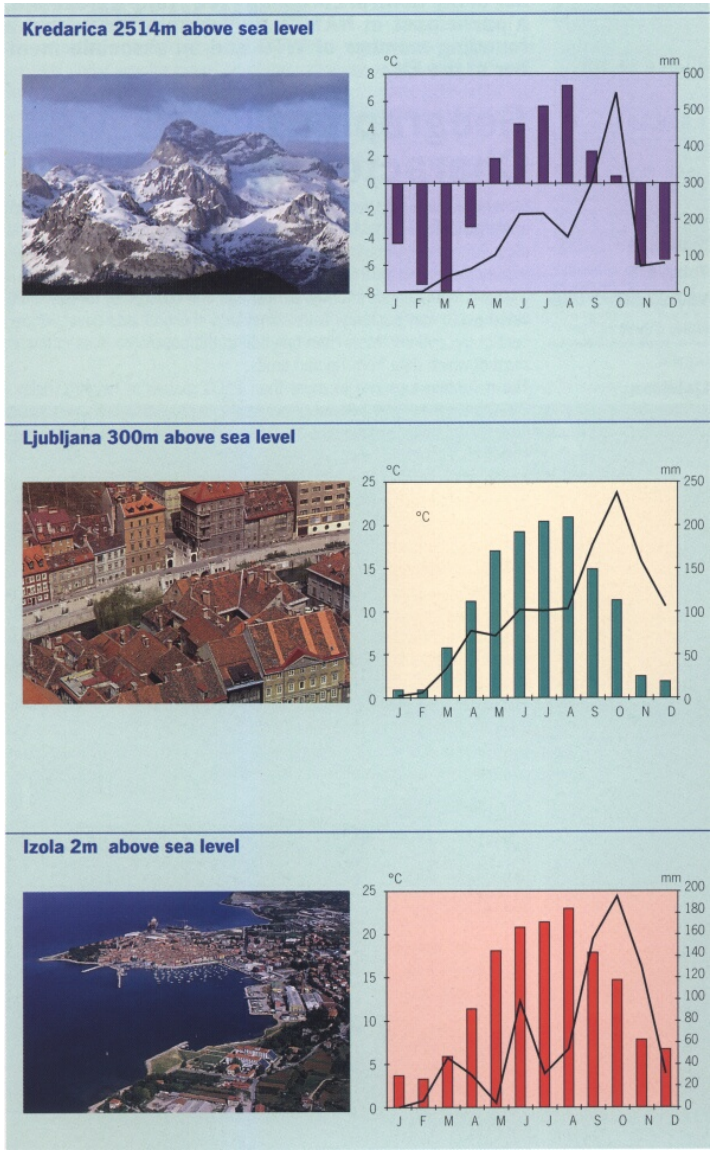


Univerza v Ljubljani
Doktorski študijski program
Statistika

Sodobni statistični pristopi

Prikazi podatkov

Vladimir Batagelj
IMFM, Ljubljana

19. februar 2015

Kazalo

http://zvonka.fmf.uni-lj.si/netbook/doku.php?id=pub:stat

Merske lestvice

Merska lestvica je predpis $f : A \to \mathbb{R}$, ki posameznemu objektu x iz množice A priredi neko realno število $f(x)$. V analizi podatkov izmerjene vrednosti dane lastnosti na izbranih objektih določajo spremenljivko.

Merske lestvice, ki merijo isto lastnost so enakovredne – npr. vrednost izražena v USD oziroma v EUR. Za običajne lestvice velja, da lahko enakovredni lestvici predelamo eno v drugo z *dopustnimi transformacijami*. Te določajo *vrsto* (tip) lestvice (glej naslednjo prosojnico).

Vrste lestvic imajo v analizi podatkov pomembno vlogo – določajo, kaj lahko z danimi podatki počnemo.

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

Vrste merskih lestvic

dopustne transformacije	**vrsta lestvice**	**primeri**
$\varphi(x) = x$ (identiteta)	absolutna	štetje
$\varphi(x) = a.x, a > 0$	razmernostna	masa
podobnost		temperatura (K)
$\varphi(x) = a.x + b, a > 0$	razmična	temperatura (C,F)
		čas (koledar)
$x \geq y \Leftrightarrow \varphi(x) \geq \varphi(y)$	urejenostna	šolske ocene, kakovost
strogo naraščajoča		zraka, trdost kamnin
φ je povratno enolična	imenska	barva las, narodnost

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

Preiskovanje in predstavitve podatkov

Z rastjo zmogljivosti namiznih računalnikov se je v 90. letih *prikazovanje* podatkov začelo uveljavljati kot pomembno orodje za *preiskovanje* podatkovij in za *predstavitve* dobljenih rezultatov.

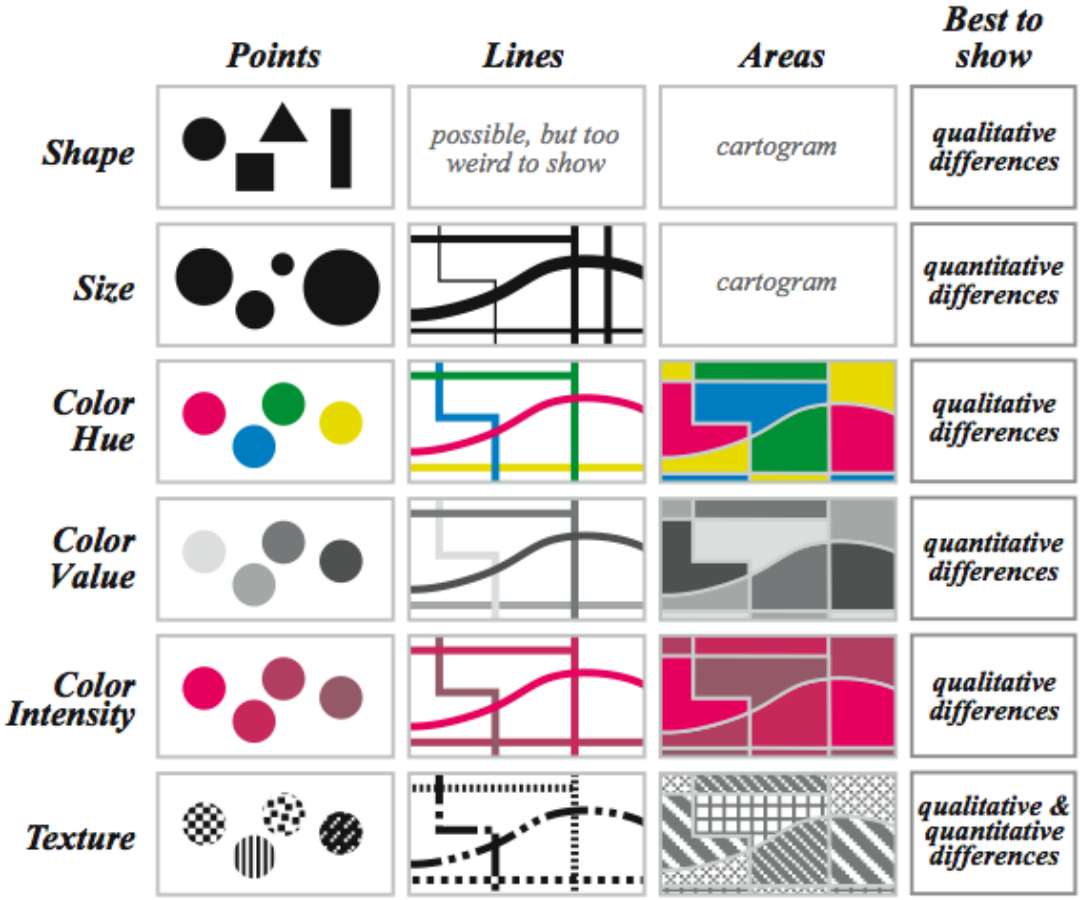
Osnovna naloga *analize in prikazov podatkov* je razkrivanje lastnosti (obsežnih) podatkovij z uporabo zmogljivosti človekovega vida na prikazih ustvarjenih s sodobno računalniško grafiko. Končni cilj je dobiti vpogled v podatke in razumeti odnose med posameznimi deli.

Cilj (slikovnih) predstavitev je kar se da razumljivo posredovanje rezultatov. Pri (znanstvenih) prikazih uporabljamo slike, da bi razkrili posamezne značilnosti podatkov. Pogosto se oba vidika prekrivata.

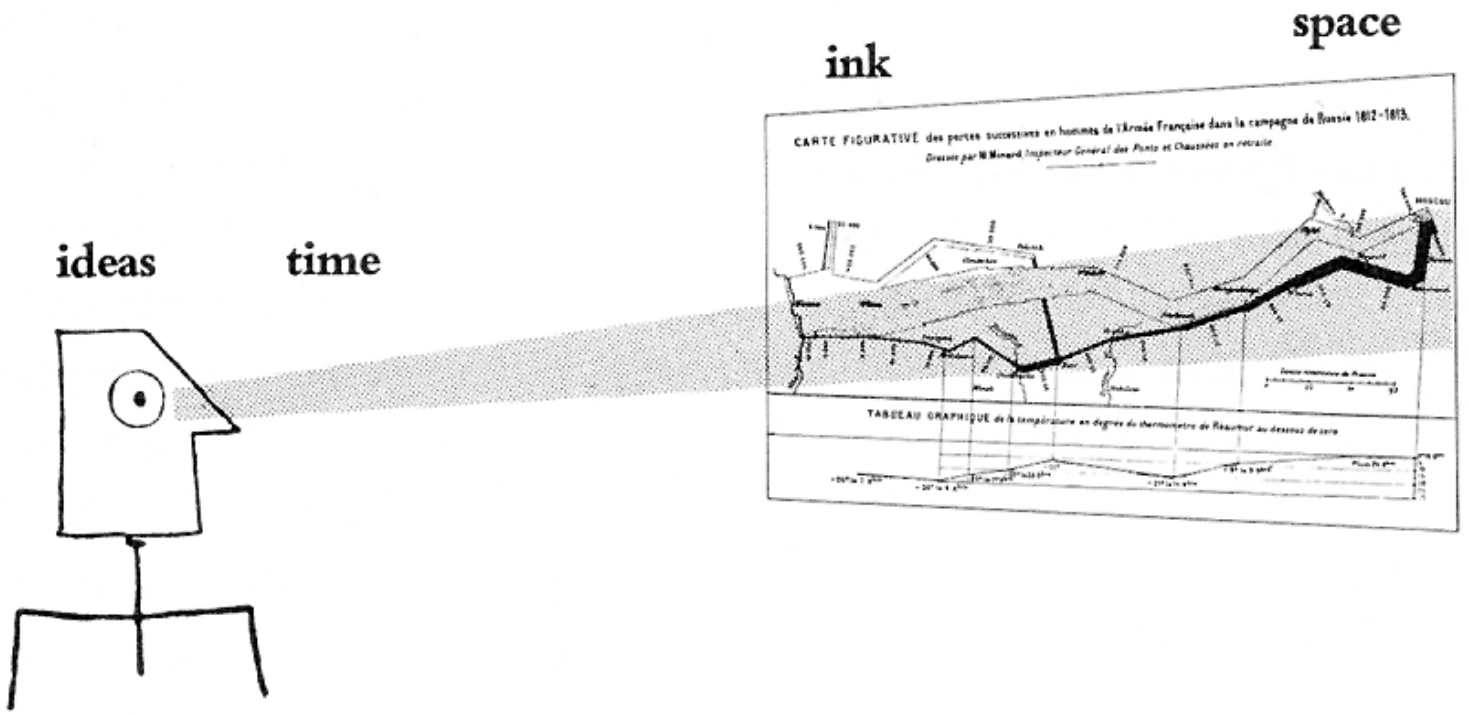
Milestones

Osnove prikazov podatkov

Teoretične osnove prikazov podatkov je postavil Jacques Bertin v svoji knjigi *Sémiologie graphique* (1967).

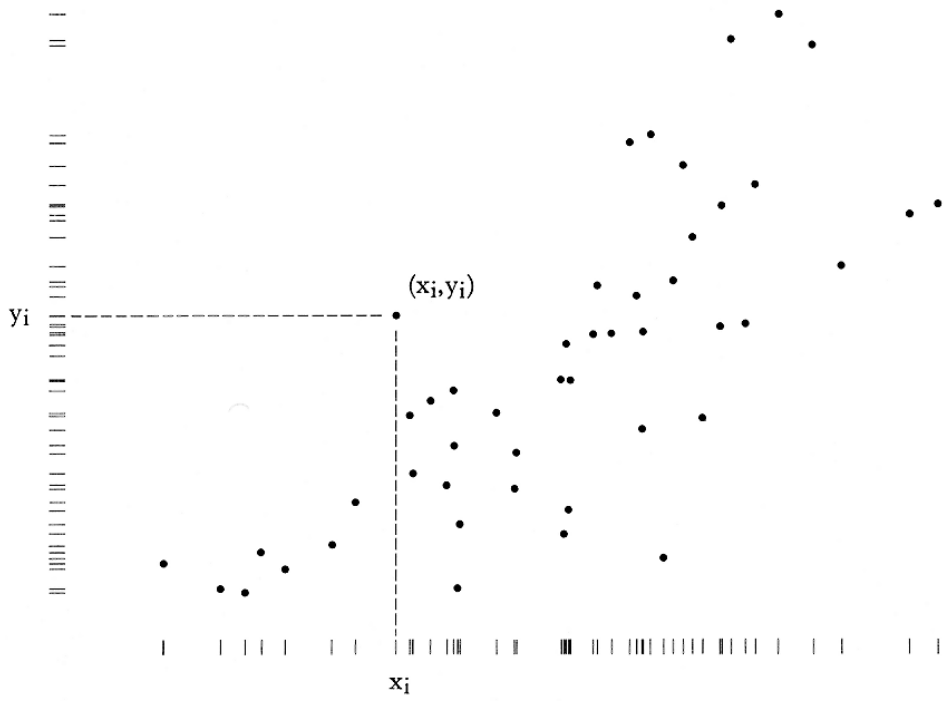


Osnove prikazov podatkov (Edward R. Tufte)



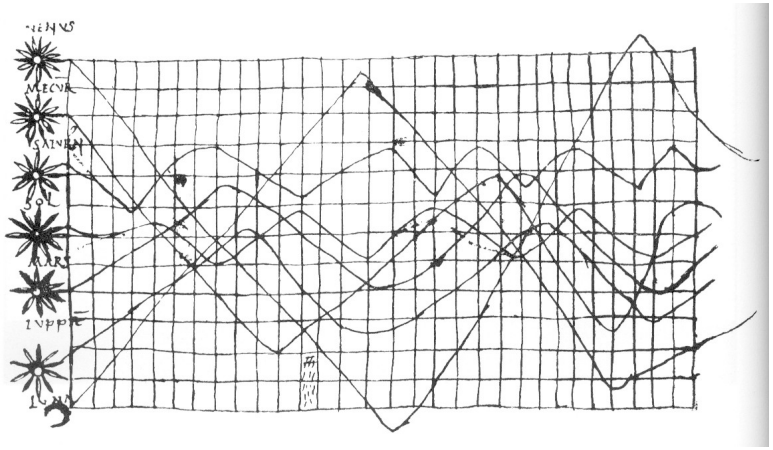
Graphical *excellence* consists of complex ideas communicated with *clarity*, *precision*, and *efficiency*. It gives to the viewer the greatest number of *ideas* in the shortest *time* with the least *ink* in the smallest *space*. It is telling the *truth* about the data.

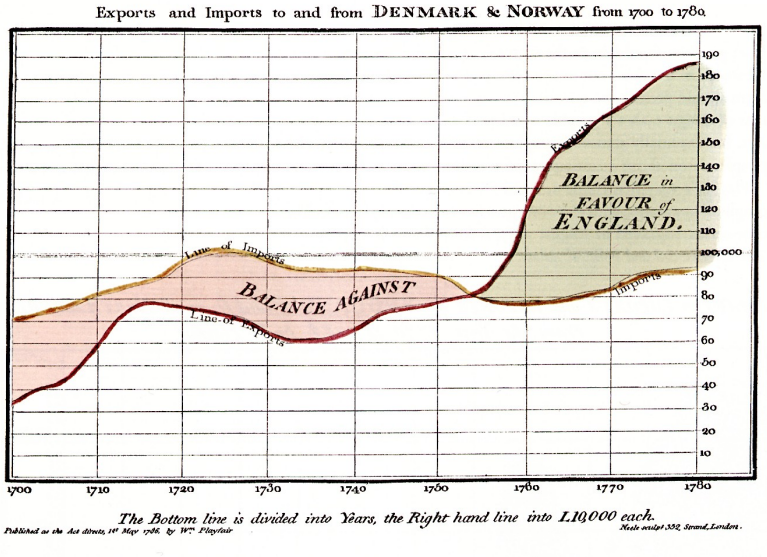
Tufteova načela



- Predvsem prikaži podatke.
- Povečuj razmerje med podatki in črnilom.
- Odstranjuj nepodatkovno črnilo.
- Odstranjuj odvečno podatkovno črnilo.
- Preglej in preuredi.

Zgledi





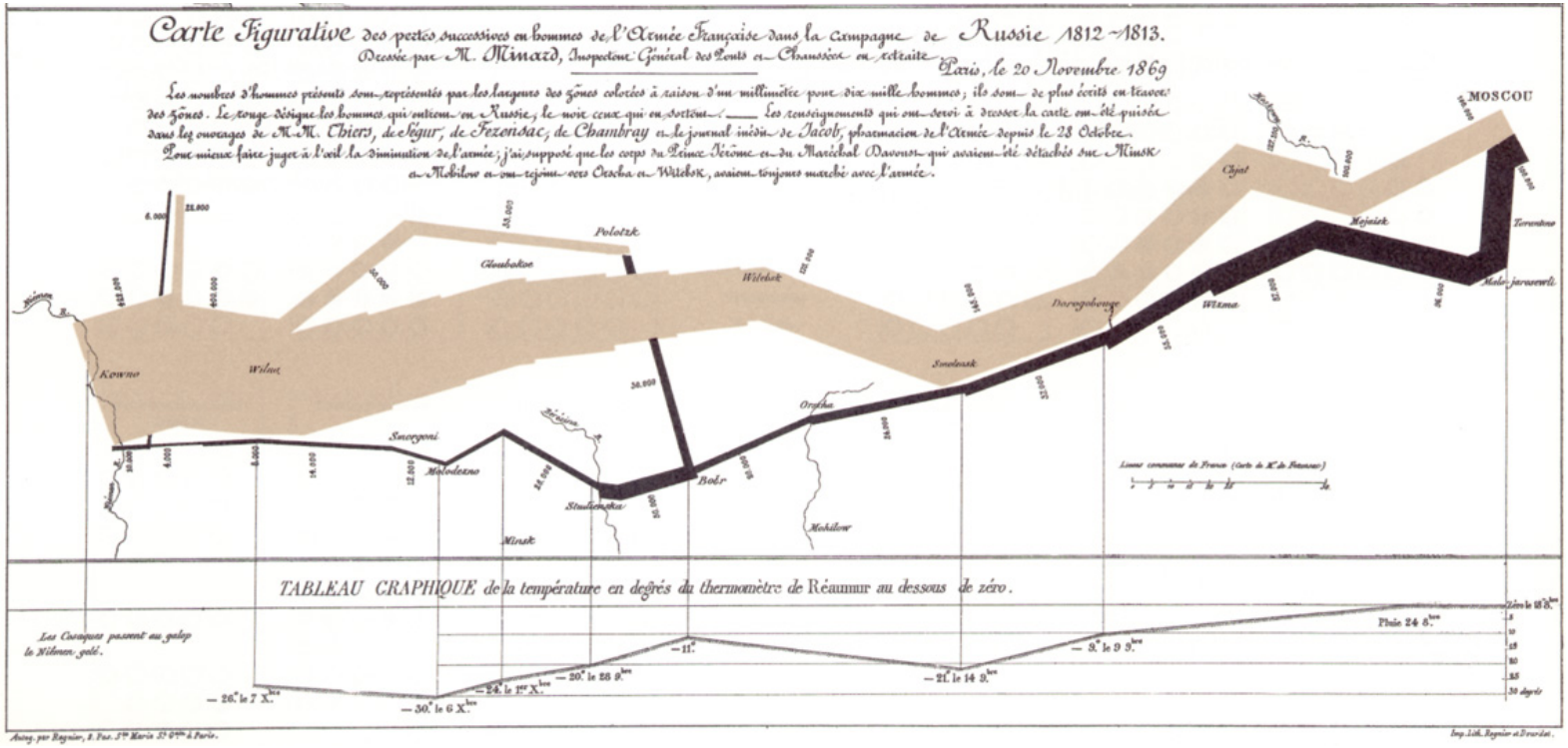
Najstarejši (10. stoletje) znani poskus prikaza zveznega spreminjanja neke količine je prikaz spreminjanja naklona planetarnih tirnic.

Eni izmed zgodnejših (1785) prikazov ekonomskih podatkov so Playfair-ovi prikazi izvoza in uvoza med državami.

Zgled: Kolera

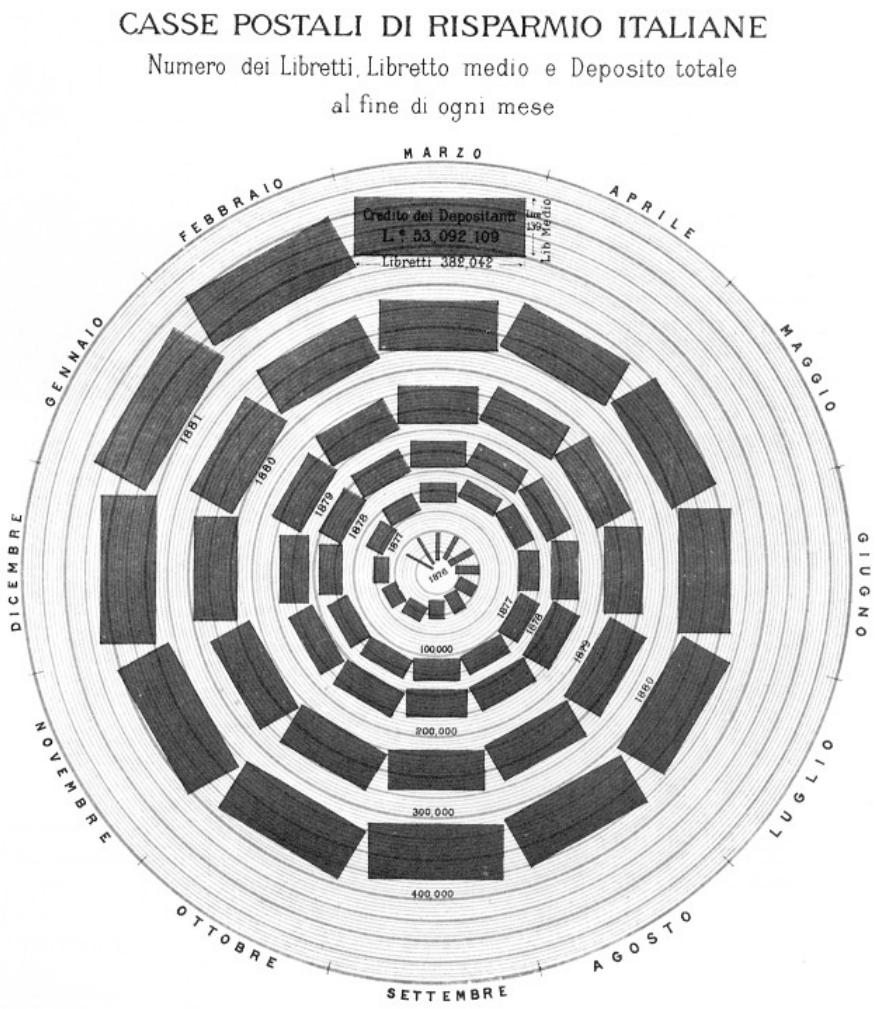
Leta 1854 je dr. John Snow na zemljevid mestnega predela, v katerem je izbruhnila kolera, vrisal s pikami mesta, kjer so posamezniki umrli. V središču pik se je nahajal vodnjak, iz katerega so se okoliški prebivalci oskrbovali z vodo. Dal ga je zapreti – bolezen je izginila.

Zgled: Napoleonov pohod v Rusijo



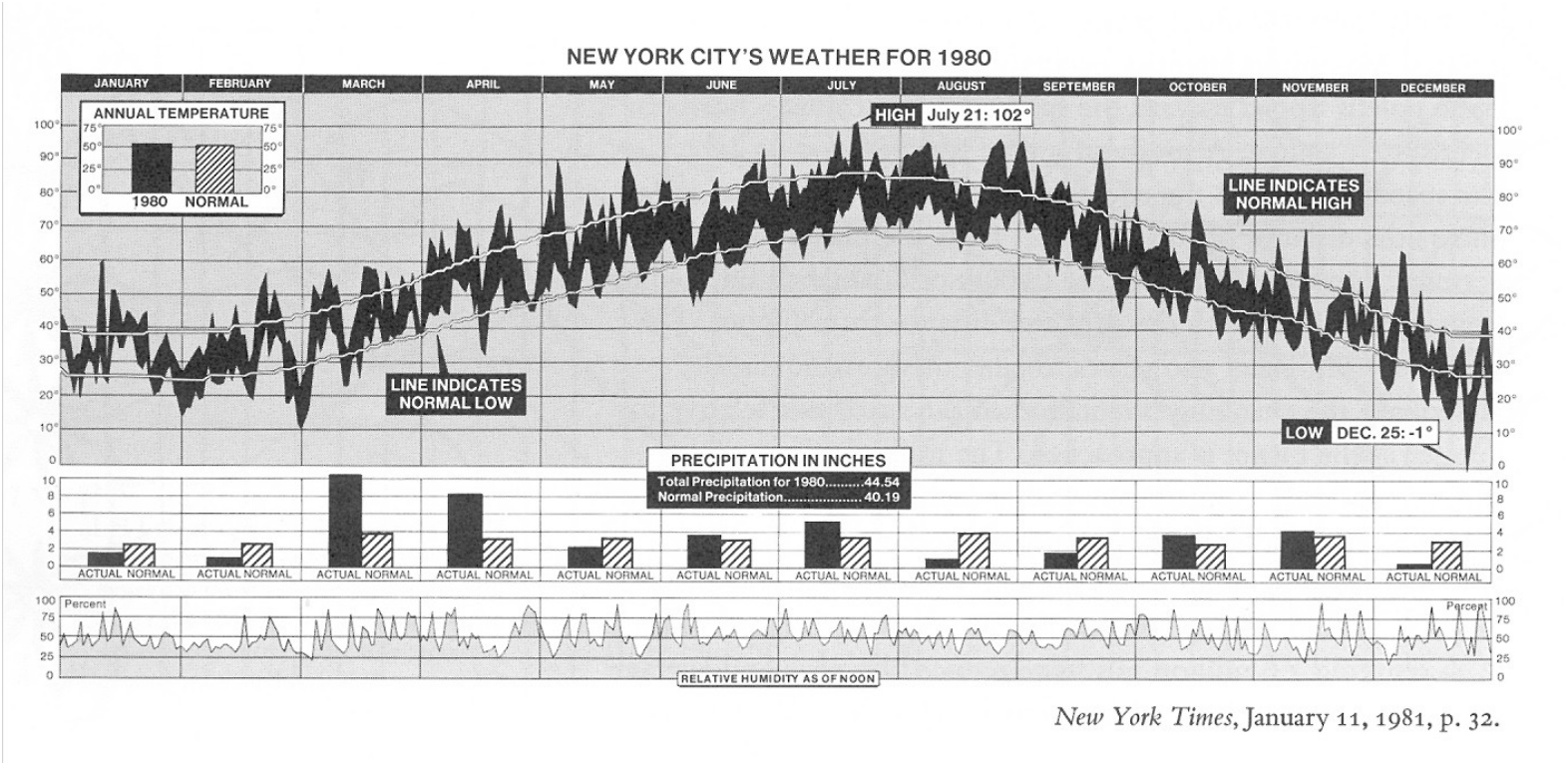
Za pravo mojstrovino velja Minardov (1861) prikaz Napoleonovega pohoda v Rusijo 1812-1813.

Zgled: Rast prihrankov



Zanimiv je prikaz mesečne rasti prihrankov v italijanski poštni hranilnici v letih 1876 do 1881 objavljen v knjigi Gabaglio A.: Teoria Generale della Statistica.

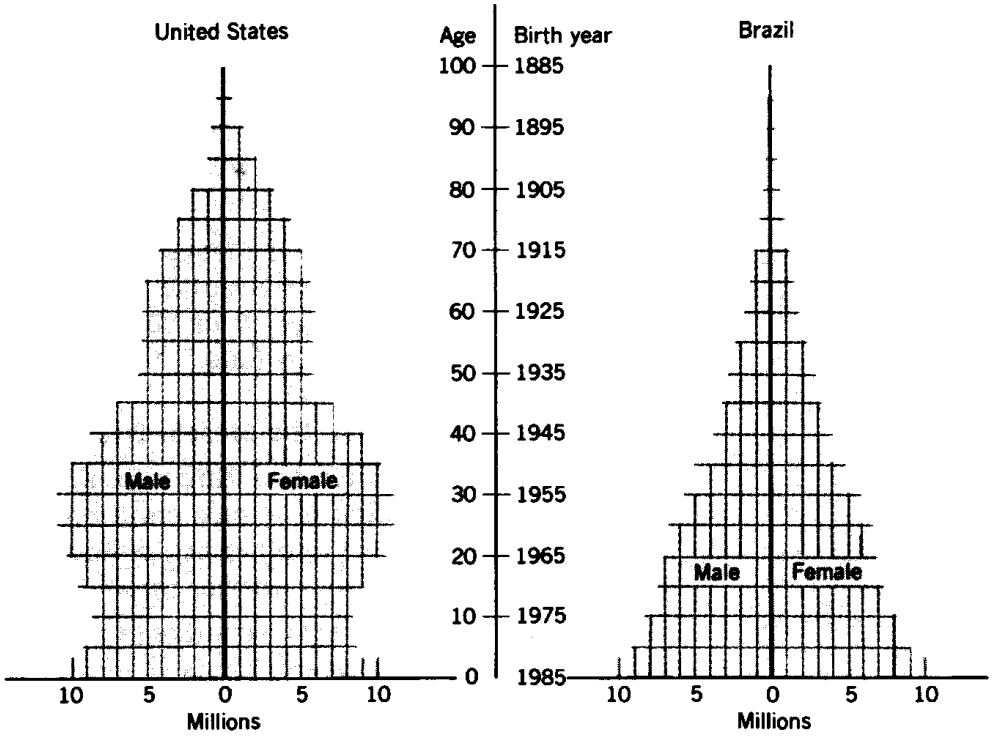
Zgled: Letni pregled vremena

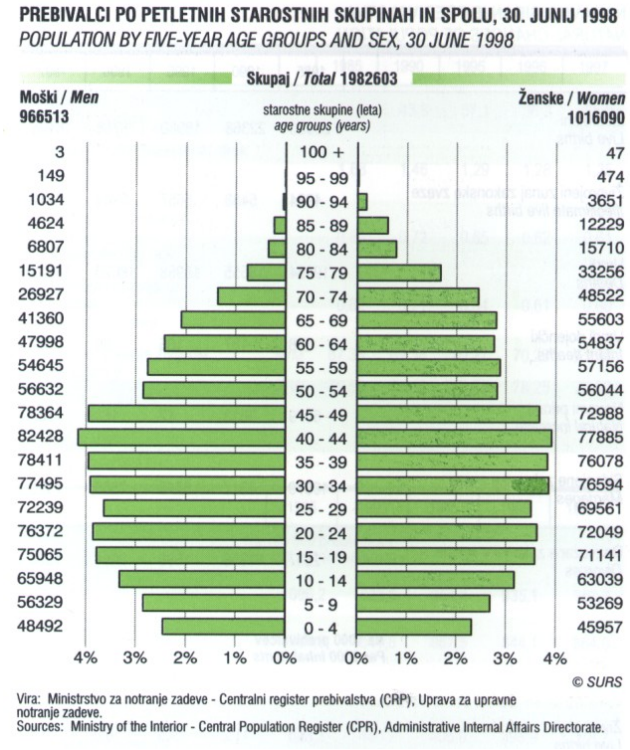


New York Times, January 11, 1981, p. 32.

Zelo veliko podatkov je vgrajenih tudi v prikaz vremena v New Yorku v letu 1980, objavljen v New York Timesu, 11. januarja 1981.

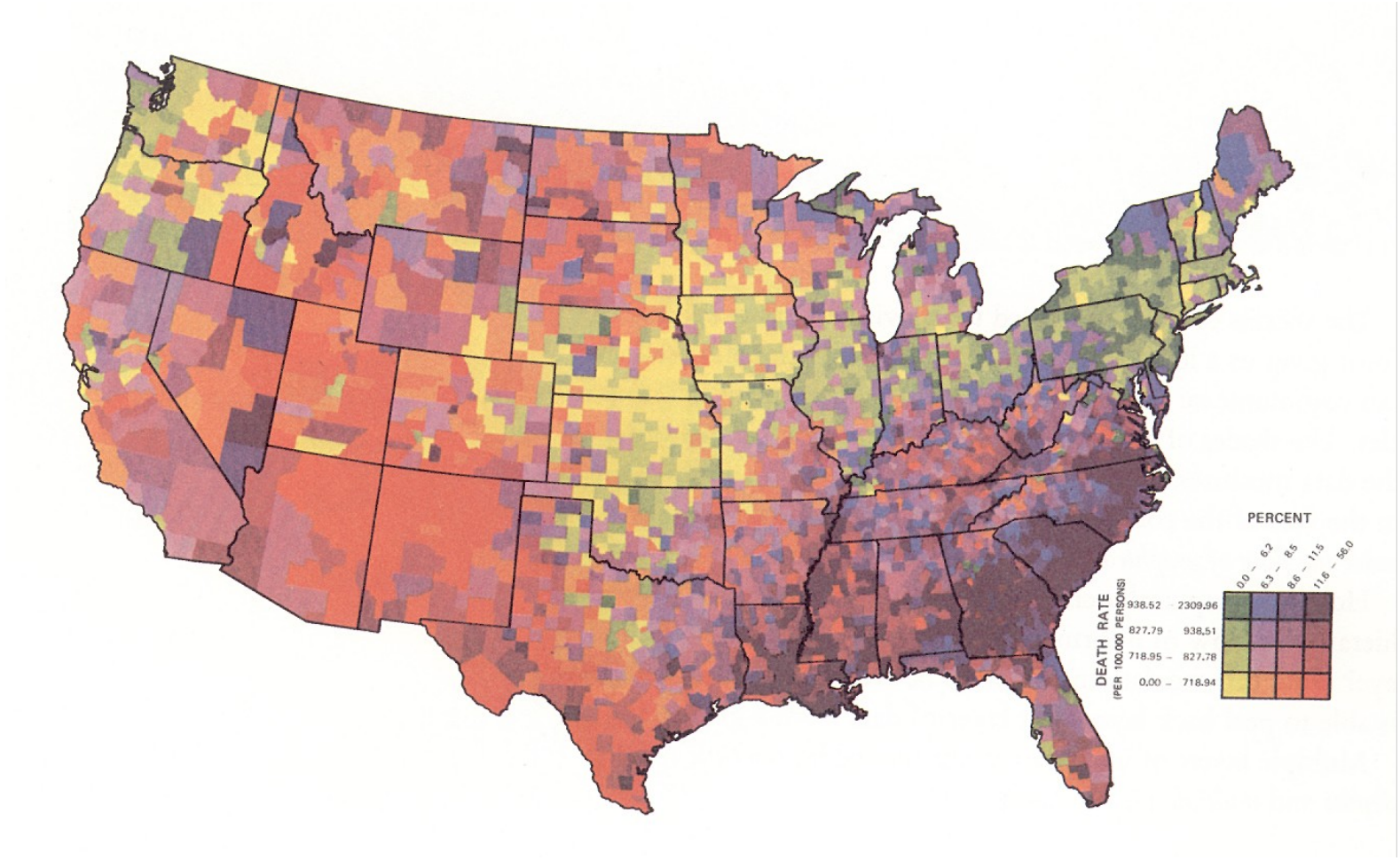
Zgled: Starostna porazdelitev prebivalstva





Iz oblike porazdelitev lahko marsikaj izvemo o prebivalstvu in zgodovini neke dežele.

Zgled: Podatki na zemljevidu

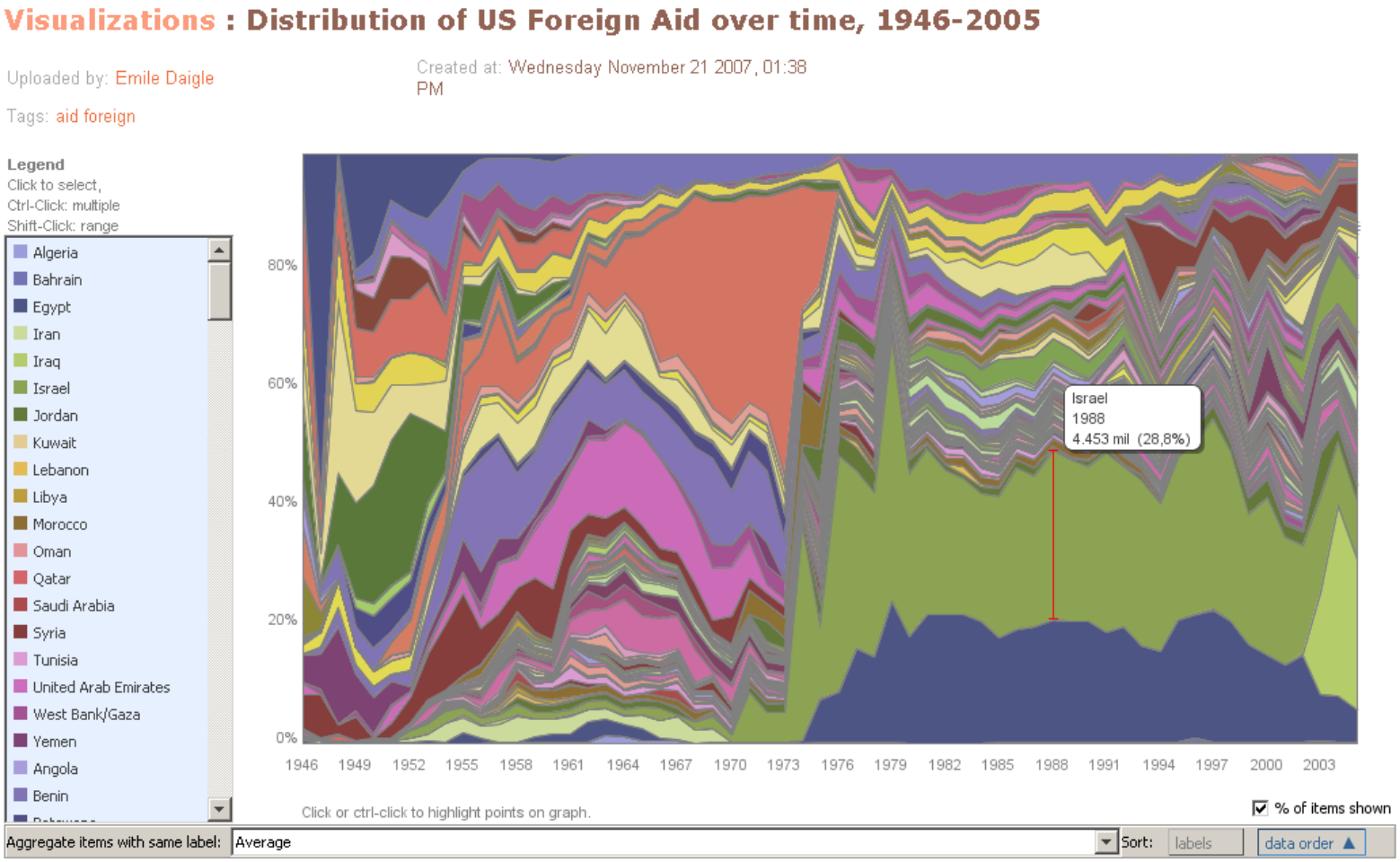


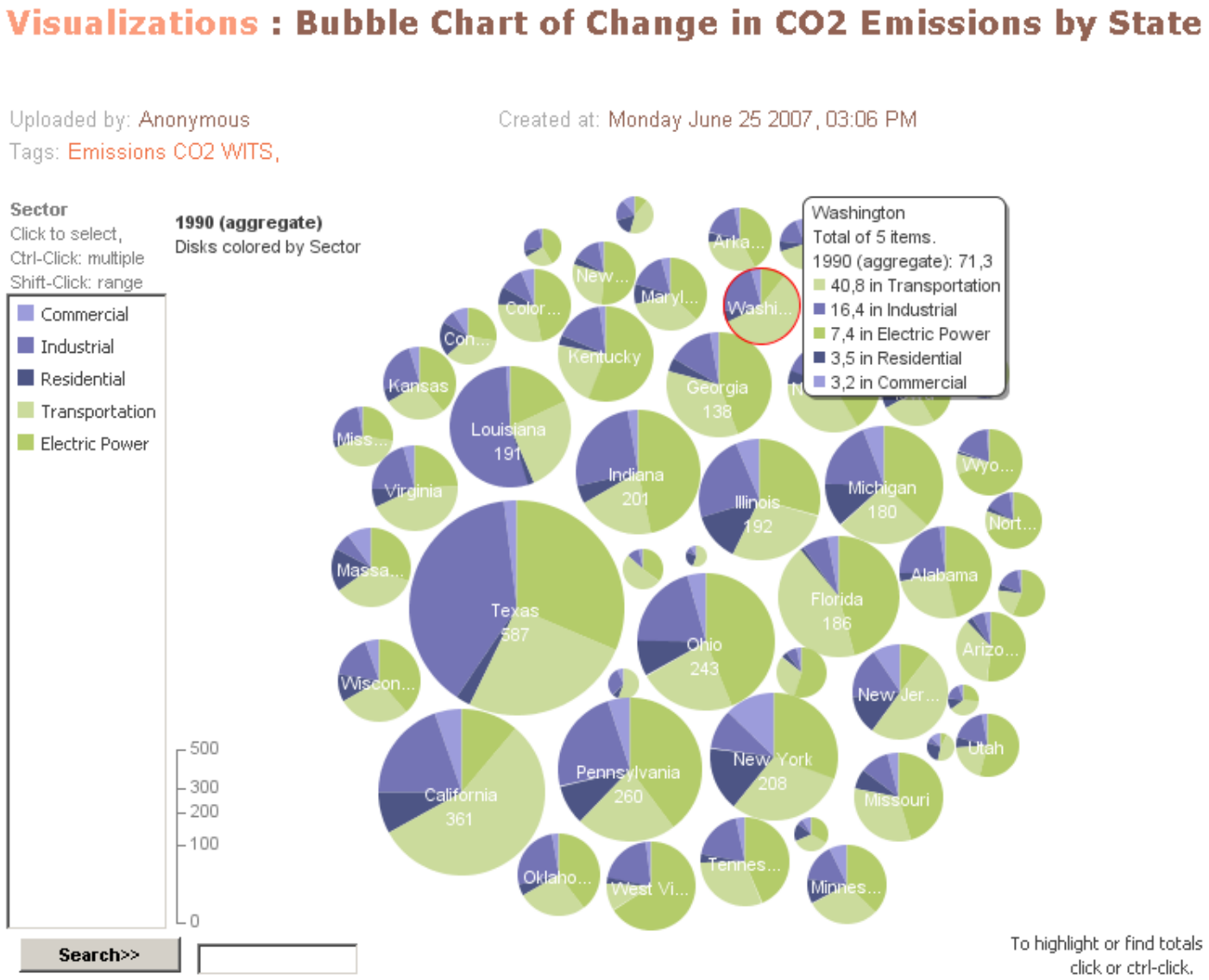
Zelo učinkovito je tudi prepletanje zemljepisnih in statističnih podatkov.

Global Administrative Areas.

Zgledi: ManyEyes

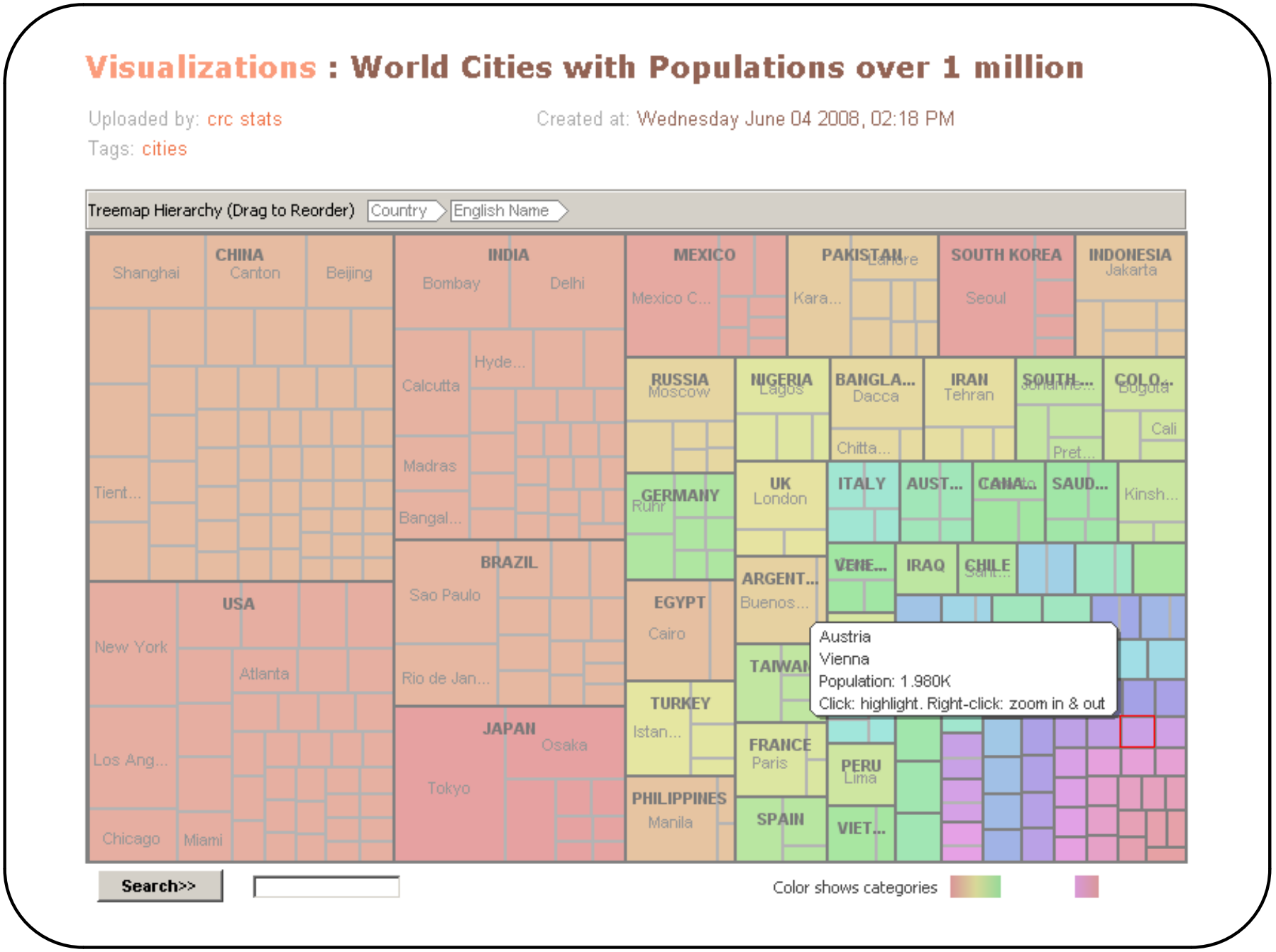
ManyEyes; IBM Visual Communication Lab: Martin Wattenberg, Fernanda Viégas, Frank van Ham, . . .

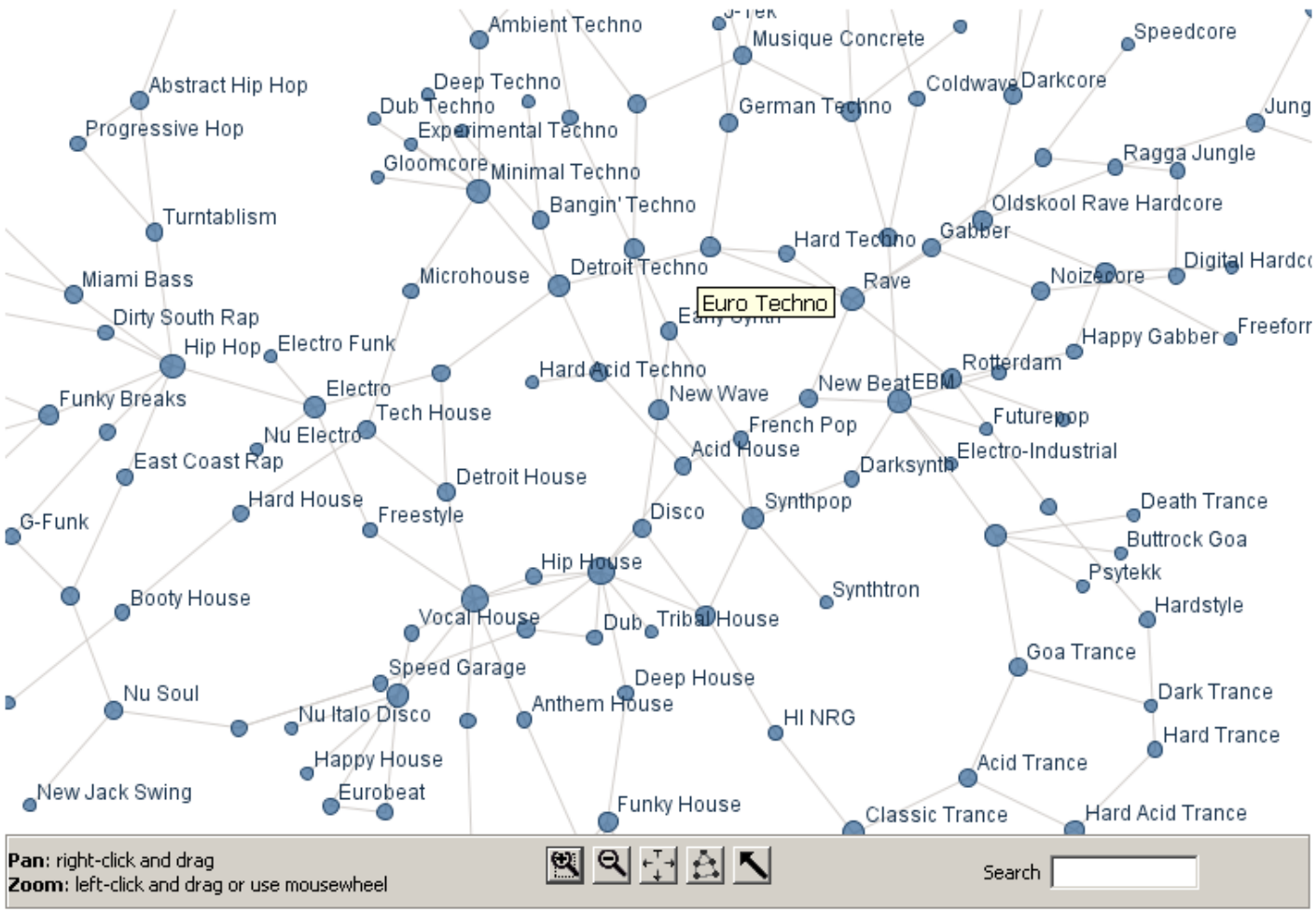



Visualizations : Bubble Chart of Change in CO2 Emissions by State
Uploaded by: Anonymous
Tags: Emissions CO2 WITS,
Created at: Monday June 25 2007, 03:06 PM
Sector
Click to select,
Ctrl-Click: multiple
Shift-Click: range
Commercial
Industrial
Residential
Transportation
Electric Power
1990 (aggregate)
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 387
California 361
Pennsylvania 260
New York 208
Ohio 243
Florida 186
Illinois 192
Indiana 201
Michigan 180
Georgia 138
Louisiana 191
Kentucky
Alabama
Missouri
Tennes...
Minnes...
West Vi...
Oklaho...
Wiscon...
Massa...
Virginia
Kansas
Miss...
Con...
Color...
New...
Maryl...
Washi...
Arka...
Wyo...
Nort...
Arizo...
New Jer...
Utah
To highlight or find totals
click or ctrl-click.

Visualizations : The Future Begins

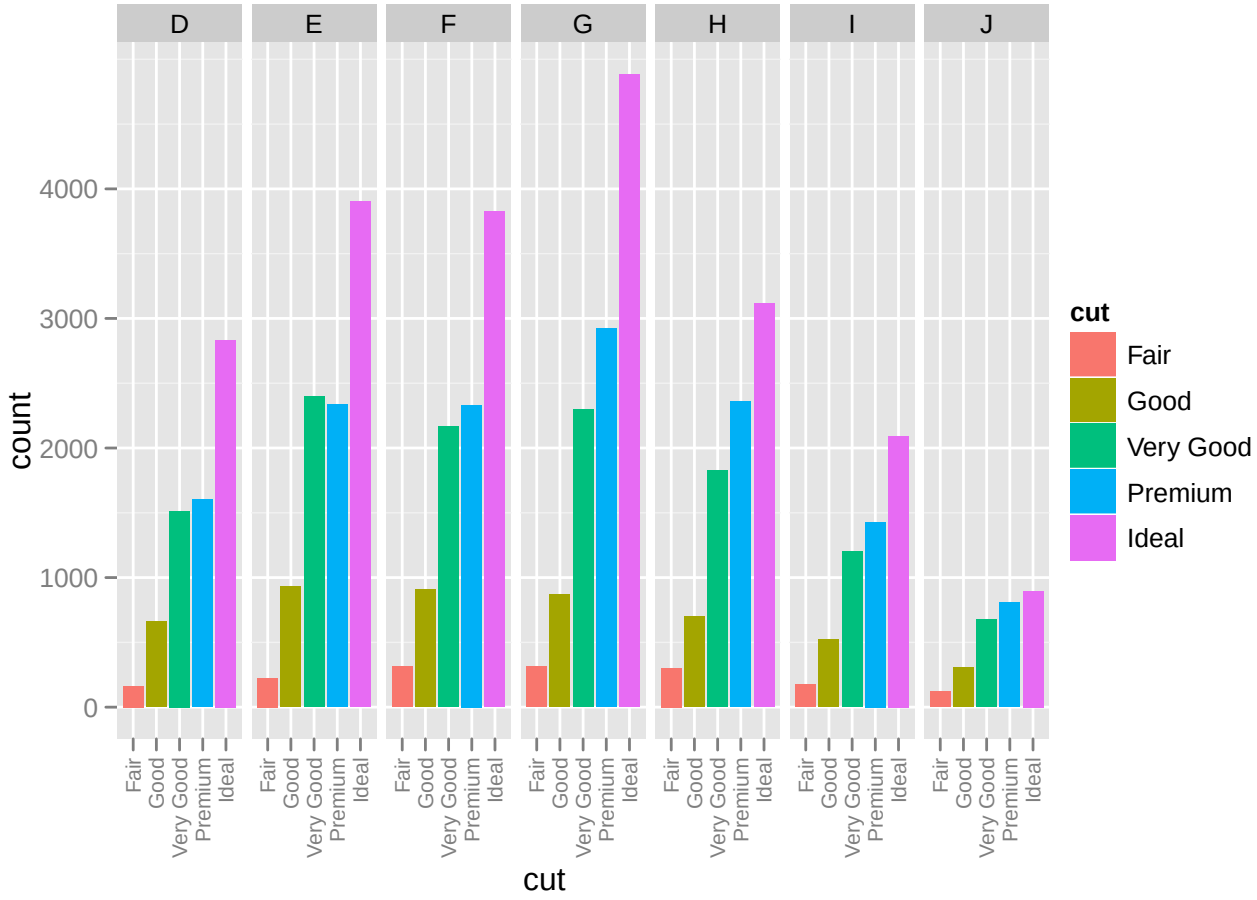
Uploaded by: Susanne Created at: Tuesday January 20 2009, 01:26 PM
Description:


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
Kara...
Lahore
SOUTH KOREA
Seoul
INDONESIA
Jakarta
RUSSIA
Moscow
NIGERIA
Lagos
BANGLA...
Dacca
Chitta...
IRAN
Tehran
SOUTH...
COLO...
Bogota
Cali
Pret...
GERMANY
Ruhr
UK
London
ITALY
AUST...
SAUD...
Kinsh...
BRAZIL
Sao Paulo
Rio de Jan...
USA
New York
Atlanta
Los Ang...
Chicago
Miami
EGYPT
Cairo
ARGENT...
Buenos...
VENE...
IRAQ
CHILE
TAIWAN
TURKEY
Istan...
JAPAN
Osaka
Tokyo
FRANCE
Paris
PERU
Lima
PHILIPPINES
Manila
SPAIN
VIET...
Austria
Vienna
Population: 1.980K
Click: highlight. Right-click: zoom in & out
Search>>
Color shows categories


Visualizations : The Family Tree of Electronic Music
Uploaded by: Belarius
Created at: Monday August 04 2008, 06:44 AM
Tags: music music,techno,electronic
Ambient Techno
Musique Concrete
Speedcore
Abstract Hip Hop
Deep Techno
Coldwave
Darkcore
Dub Techno
German Techno
Jung
Progressive Hop
Experimental Techno
Ragga Jungle
Gloomcore
Minimal Techno
Turntablism
Bangin' Techno
Oldskool Rave Hardcore
Hard Techno
Gabber
Miami Bass
Microhouse
Detroit Techno
Rave
Noisecore
Digital Hardc
Euro Techno
Dirty South Rap
Happy Gabber
Freeforr
Hip Hop
Electro Funk
Rotterdam
Hard Acid Techno
Funky Breaks
Electro
New Wave
New Beat
EBM
Tech House
Futurepop
French Pop
Nu Electro
Acid House
Electro-Industrial
East Coast Rap
Darksynth
Detroit House
Death Trance
Hard House
Synthpop
G-Funk
Freestyle
Disco
Buttrock Goa
Hip House
Psytekk
Synthtron
Booty House
Hardstyle
Vocal House
Dub
Tribal House
Goa Trance
Speed Garage
Deep House
Nu Soul
Dark Trance
Anthem House
Nu Italo Disco
HI NRG
Hard Trance
Happy House
Acid Trance
New Jack Swing
Eurobeat
Funky House
Classic Trance
Hard Acid Trance
Pan: right-click and drag
Zoom: left-click and drag or use mousewheel
Search

Zgledi: 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)
```



Knjižnice

Za prikaze podatkov imamo v R-ju na voljo nekaj knjižnic. Osnovna knjižnica je `graphics`. Pozna tudi S-ovo knjižnico `lattice`. V razvoju je nova knjižnica `grid`. Trirazsežne interaktivne prikaze omogočata knjižnici `rgl` in `rggobi` (vmesnik za program GGobi).

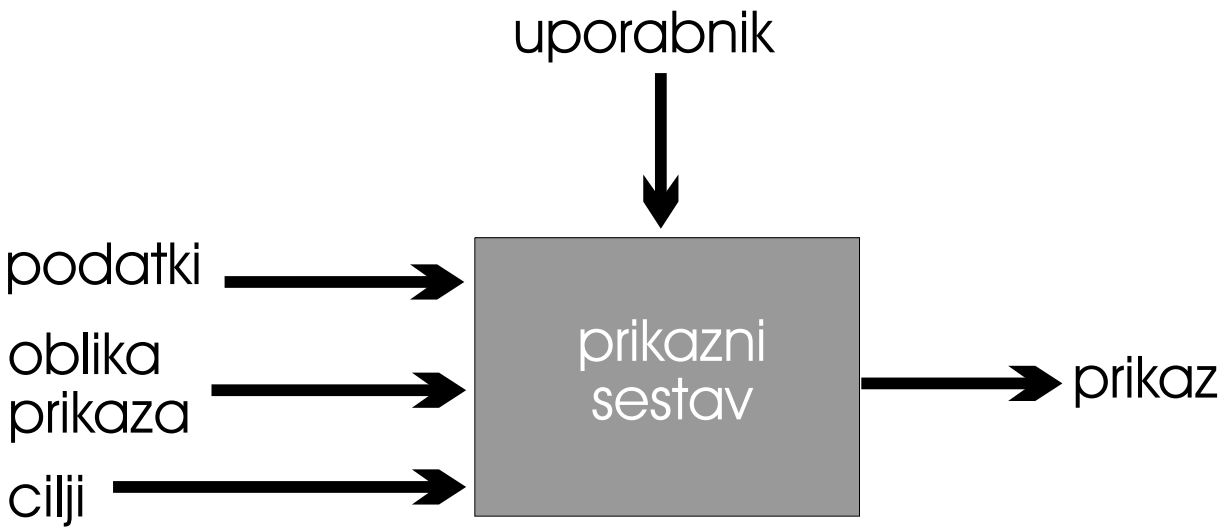
Različne posebne slikovne prikaze najdemo tudi v drugih knjižnicah. Na primer: drevesa razvrstitev in prikaze na zemljevidih.

Leland Wilkinson je napisal temeljno knjigo o prikazih podatkov The grammar of graphics in s sodelavci razvil okolje nViZn (za SPSS).

Na Wilkinsonovi knjigi temelji tudi R-jeva knjižnica `ggplot2`. Za prikaze podatkov na spletu so pred kratkim razvili v javascriptu knjižnico Protovis in njeno naslednico d3.

Prikazni sestav

Pri prikazovanju podatkov uporabnik sledi različnim ciljem: najde/določi, razlikuje, uvrsti, razvrsti, uredi, primerja, poveže, ...



Izhodne naprave

Večina starejših virov o prikazovanju podatkov je usmerjena na 'list papirja'.

Računalniški zaslon ponuja veliko novih možnosti

- vzporedni prikazi
- začasne sestavine (oznake, pojasnila, zastavice, . . .)
- poudarjene izbire, povezave med prikazi
- sodejnost (interaktivnost)
- občutek trirazsežnosti (gibanje)
- programski nadzor (gibljive slike - animacije)

Te možnosti je potrebno kar se le da izkoristiti za podporo reševanja nalog analize podatkov. Prihodnost je v sodejnih, dinamičnih, trirazsežnih in večsličnih (lokalni/celotni pogled, izbire, primerjave, . . .) prikazih.

GGobi

Slikovne sestavine

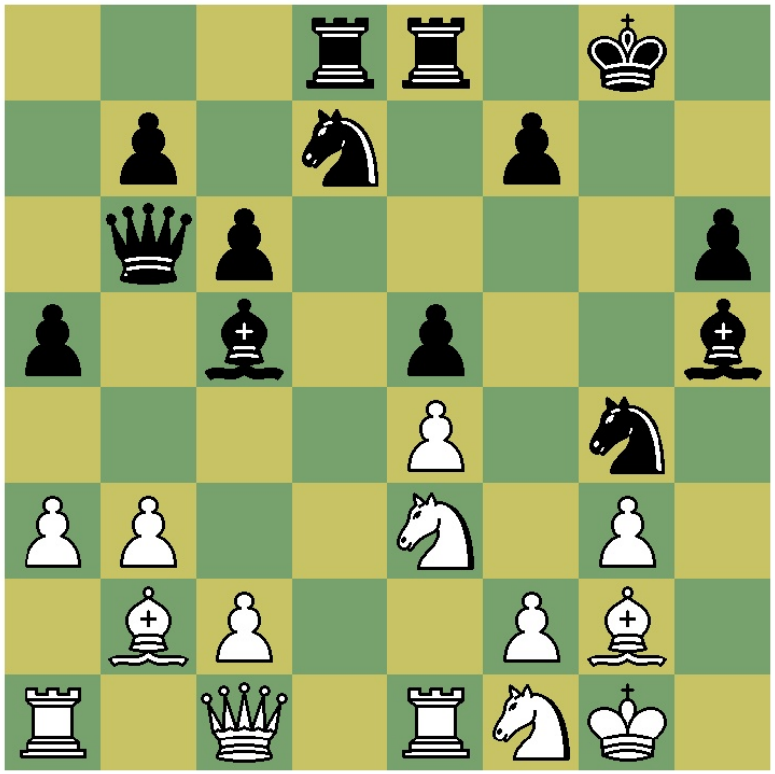
Slikovne sestavine so lastnosti s katerimi lahko predstavimo izbrano lastnost podatkov.

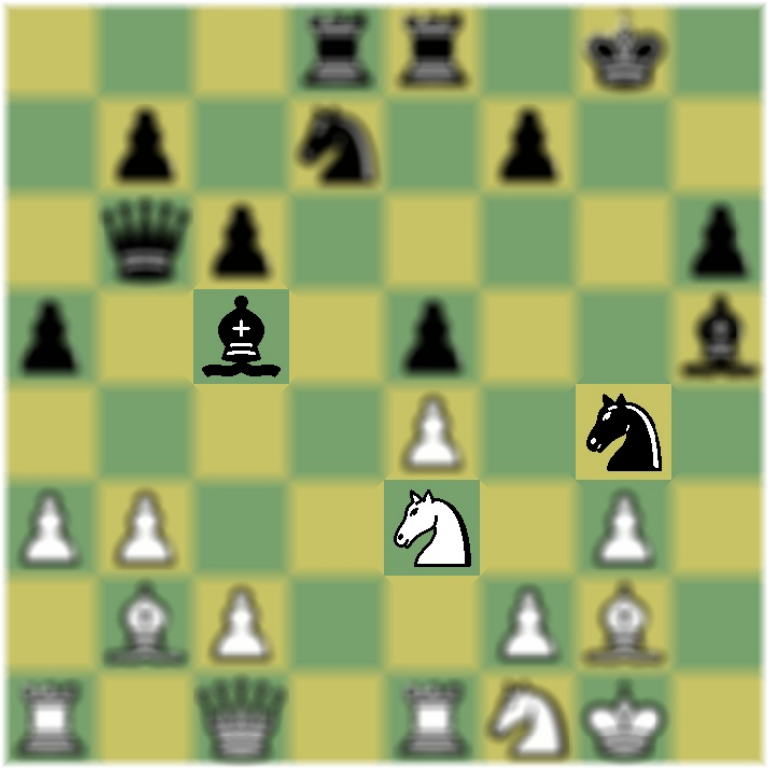
- polnila: barva (hue), nasičenost, svetlost, ostrina, vzorec
- prostorske sestavine:
  - splošne: dolžina, velikost, lik, mesto, smer, simetrija, položaj (spredaj / zadaj), zasuk, razdalja (med gradniki)
  - posebne 3D : zakrivanje, perspektiva, prozornost, globina, gibanje
- razsežnosti: pika, daljica, lik, telo

Slikovne sestavine so osnova različnih slikovnih prikazov.

Barve v R-ju

Slikovne sestavine – globinska ostrina





Semantic Depth of Field

Weber-Fechner-jev zakon

Ernst Weber (1795-1878) je opazil, da naj bi bila *najmanjša še zaznavna sprememba* občutka ψ sorazmerna razmerju spremembe dražljaja φ in njegove velikosti

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

Zaznavamo torej relativne spremembe dražljajev.

Od tu je Gustav Theodor Fechner (1801-1887) izpeljal zvezo

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

kjer je φ_0 mejna velikost (ne)zaznavnosti dražljaja.

vid: $k = \frac{1}{60}$; bolečina: $k = \frac{1}{30}$; vonj: $k = \frac{1}{4}$; slanost: $k = \frac{1}{3}$.

Na primer: opaznost različnosti dveh dolžin je odvisna od razmerja teh dveh dolžin, ne pa od njune razlike. Opaznost razlik pri primerjavi dolžin lahko bistveno povečamo z vpeljavo oporne mreže.

Wikipedia

Stevens-ov zakon

Poskusi v 30. letih 20. stoletja so pokazali, da za večino dražljajev Fechnerjev zakon ne velja. Izkazalo se je, da velja potenčna zveza, ki jo je predlagal Stanley Smith Stevens (1906–1973):

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

kjer so k, c in φ_0 konstante.

Natančnost naših zaznav je za lastnosti s c različnim od 1 odvisna od velikosti vrednosti – zaznave so pristranske.

Poskusi so pokazali, da je c za dolžino med .9 in 1.1; za ploščino med .6 in .9; in za prostornino med .5 in .8 .

Wikipedia

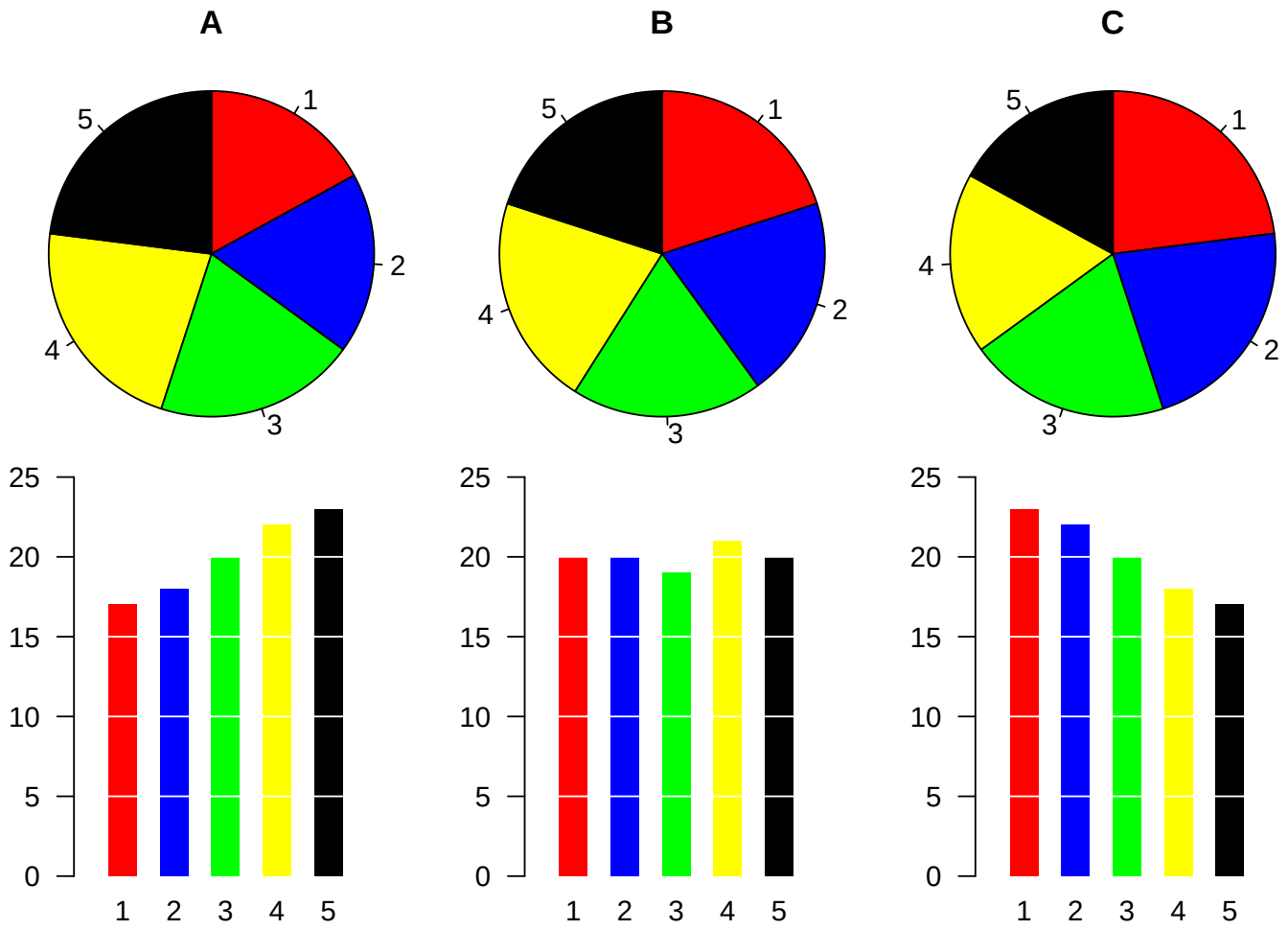
Posledice

Poskusi na osnovi Stevensovega zakona so dali naslednjo urejenost slikovnih sestavin glede na natančnost zaznavanja:

- mesto glede na skupno os (najbolj natančno)
- lega na dveh enakih, a neporavnanih oseh
- dolžina
- kot – naklon
- ploščina
- prostornina
- barva (hue) – barvna nasičenost – barvna svetlost (najmanj natančno)

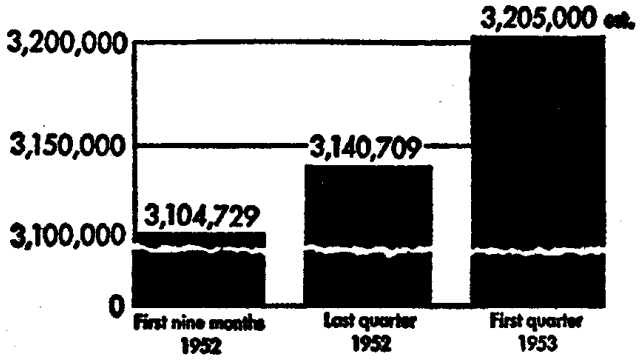
Ljudje smo različni: barvna slepota, navajenost, izurjenost, izobraženost, ...

Posledice

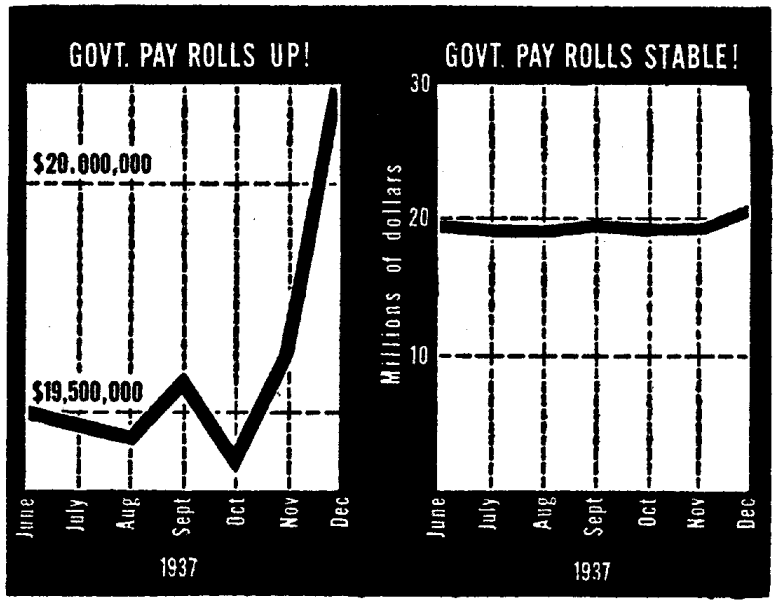


Wikipedia: Pie chart

Previdno



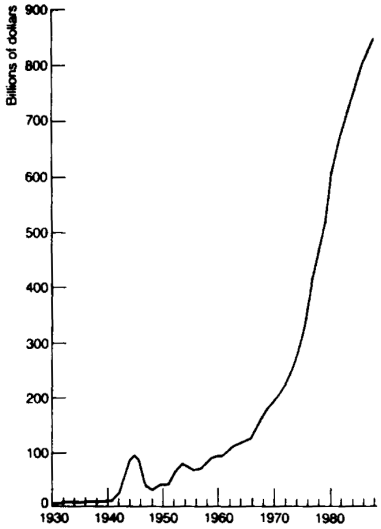
From an April 24, 1953, newspaper advertisement for COLLIER'S

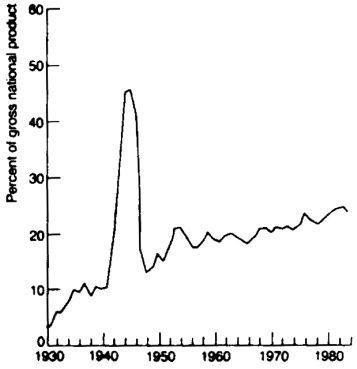


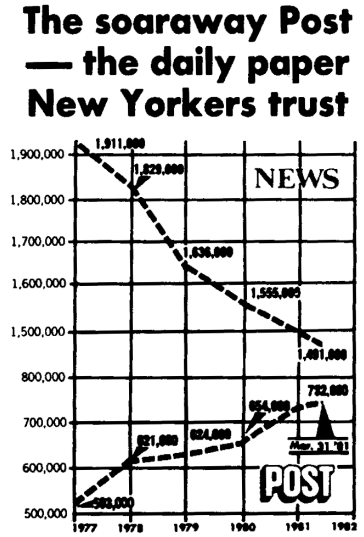
Pri tolmačenju prikazov moramo biti previdni in natančno pregledati prikaz, ne pa se kar prepustiti prvemu vtisu.

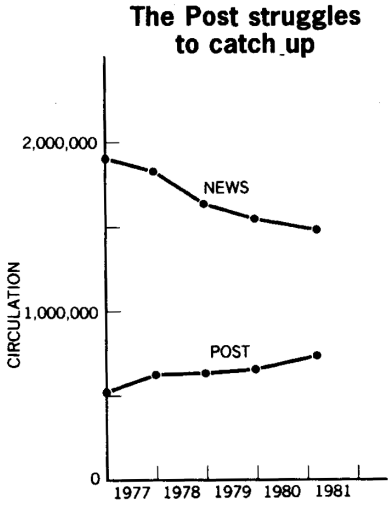
Pogosto je v prikazih podan samo zanimiv izrez iz podatkov.

. . . Previdno







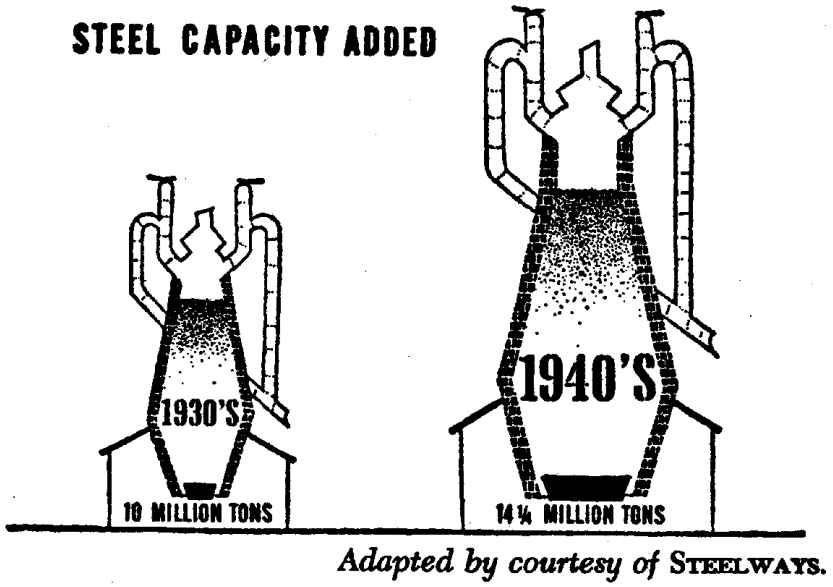


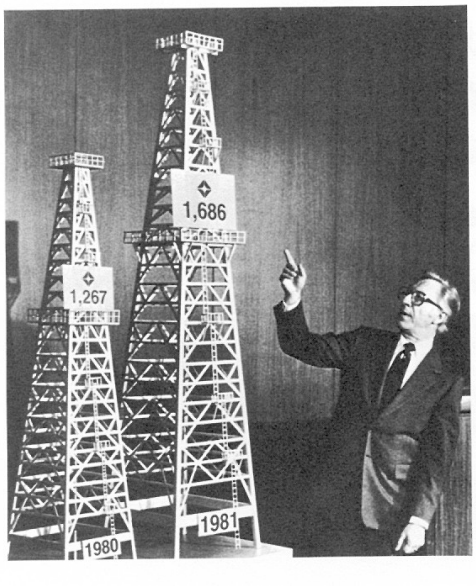
Kadar se pomen surovih podatkov spreminja skozi čas, jih moramo postaviti na 'skupni imenovalec'.

Neprekinjenost osi zavaja.

Slikovne prevare

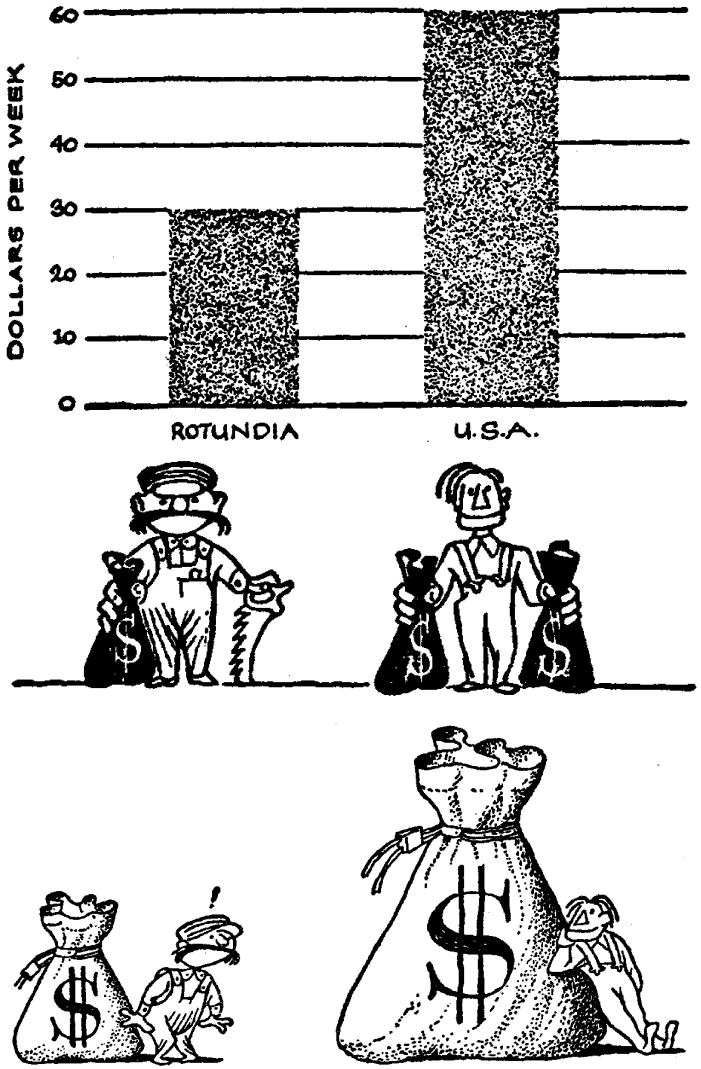
Naše opažanje je odvisno tudi od lastnosti našega (človeškega) sistema videnja. Te lastnosti je potrebno upoštevati pri načrtovanju prikazov. 1, 2, 3, 4; iluzije. Poglejmo za primer prikaz količine proizvodnje jekla (iz Darrell H.: *How to Lie with Statistics*).





Desna sličica je 1.4 krat višja od leve, ker se je proizvodnja povečala za 1.4 krat. Toda, ker pri primerjanju likov ljudje primerjamo njihove ploščine, zaznamo razliko kot dvakratno ($= 1.4^2$).

. . . Previdno: časopisi

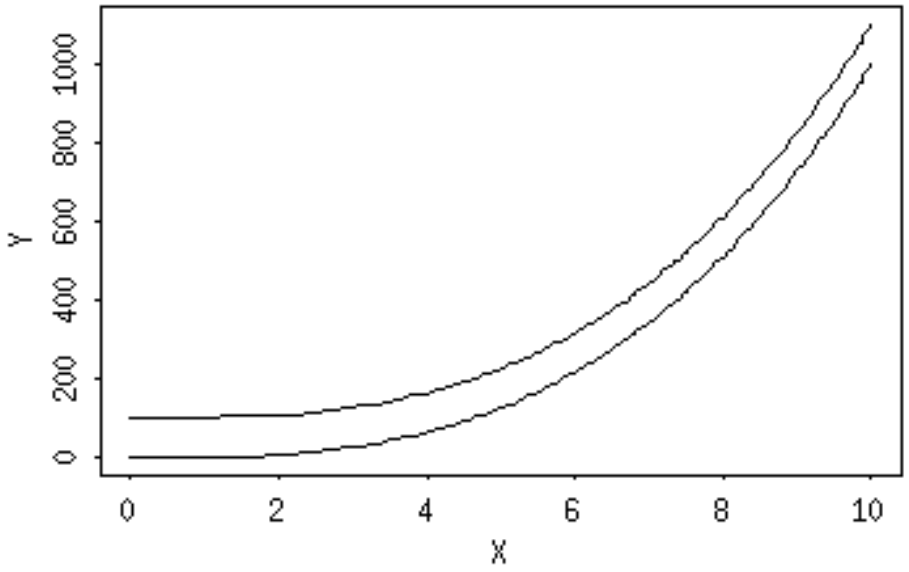


Pogosto, naprimer v časopisih in predstavitvah, je prikaz 'okinčan', da bi bil privlačnejši.

Toda pri tem moramo biti previdni – lik s podvojeno višino zaznamo kot štirikrat večjega.

. . . Slikovne prevare: krivulji

Pri primerjavi dveh krivulj nas običajno zanima **navpična razdalja** – razlika funkcijskih vrednosti; tisto, kar običajno opazimo pa je **najmanjša razdalja** med krivuljama.



Kaj je vaš *prvi vtis* o obeh funkcijah? Težita skupaj?

Podatkovja

Na izbiro oblik prikaza močno vplivajo:

- *velikost*: majhno, veliko, ogromno, 'neskončno';
- *gostota*: razpršeno, gosto, več skupin; in
- *dejavnost*: ustaljeno, spreminjajoče (deterministično, naključno).

Majhna podatkovja lahko prikažemo v celoti in v vseh podrobnostih na eni sliki. V celovitem prikazu obsežnega podatkovja se podrobnosti zgubijo; podrobni prikaz pa lahko zaobjame le del podatkovja.

Zunanji in notranji pogled

Običajni prikazi podatkov, kakršne najdemo v splošno namenskih programih (Excel, PowerPoint, ...) omogočajo praviloma pogled *od zunaj* – uporabnik je postavljen izven prikaza. V teh primerih je tretja razsežnost uporabljena predvsem za 'okraske', in ne za ustreznejši prikaz.

Ena od osnovnih značilnosti navidezne resničnosti (Virtual Reality) je podpora *notranjega* pogleda – uporabnik je vključen v prikaz kot njegov dejavni del in lahko potuje po prikazu. Slika je določena z njegovim mestom v prikazu.

Krmilje

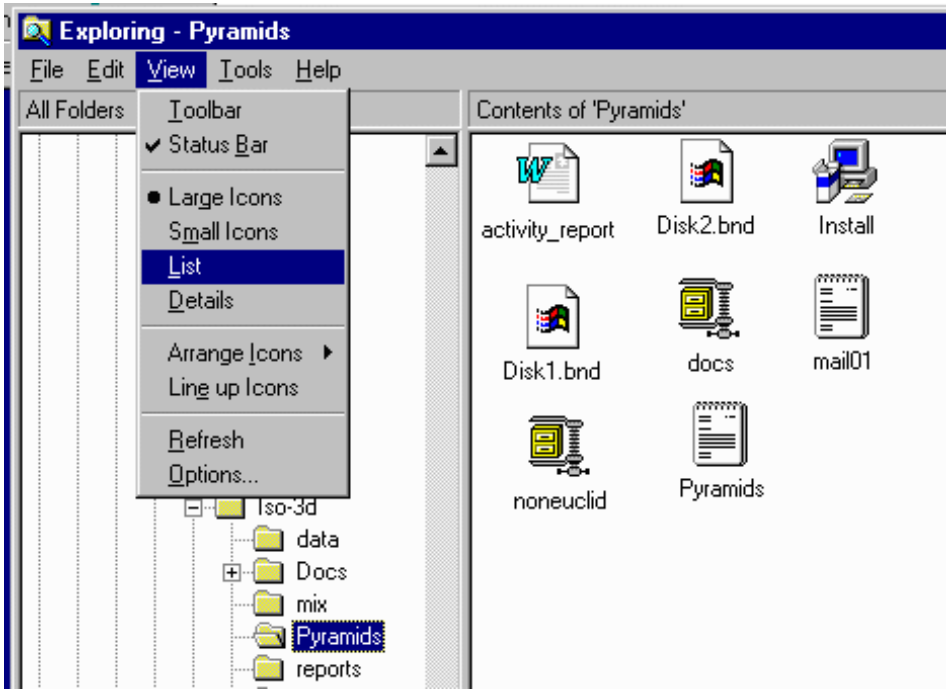
Pri pregledovanju obsežnih podatkovij se lahko uporabnik "*izgubi v gozdu*". Pri usmerjanju uporabnika so na voljo različni prijemi:

- *ponoven začetek*: vrne uporabnika v izhodišče;
- vpeljava dodatnih *orientacijskih sestavin*: prikaz osi, mreže, sence, znamenja (stalna / uporabniška). Te sestavine lahko uporabnik vklaplja in izklaplja.
- *več slik*: prikaz na vsaj dveh slikah (oknih):
  - *zemljevid*: celovit (praviloma zunanji) pogled, ki prikazuje tudi trenutno mesto uporabnika in omogoča 'dolge' korake (skoke). Za zelo obsežna podatkovja lahko vključuje možnost povečave oziroma 'ribjega očesa'.
  - *trenutna okolica*: prikaz izbranega dela podatkovja.

Koristna je uvedba **`sledi`** / **`nazaj`** / **`ponovi`** in *vodenih sprehodov*.

Očala in lupe

Z večsličnimi prikazi so tesno povezani pojmi *očal*, *lup* in *povečevanja* (**Pad++**, **inXight**).



Z izbiro različnih očal dobimo različne poglede na iste podatke – ki podpirajo različne vidike/cilje prikaza. Očala delujejo na celo sliko, lupe pa samo na izbrano območje slike. Pristop poznamo iz slikovnih vmesnikov.

Prikaz lahko dopolnimo s *podpornimi sestavinami*: oznake, mreže, pojasnila, ...

Prikazi večrazsežnih podatkovij

Naj bo $E = \{X_i\}$ *množica enot*. *Enota* X je običajno opisana z naborom vrednosti izbranih lastnosti – *spremenljivk*

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

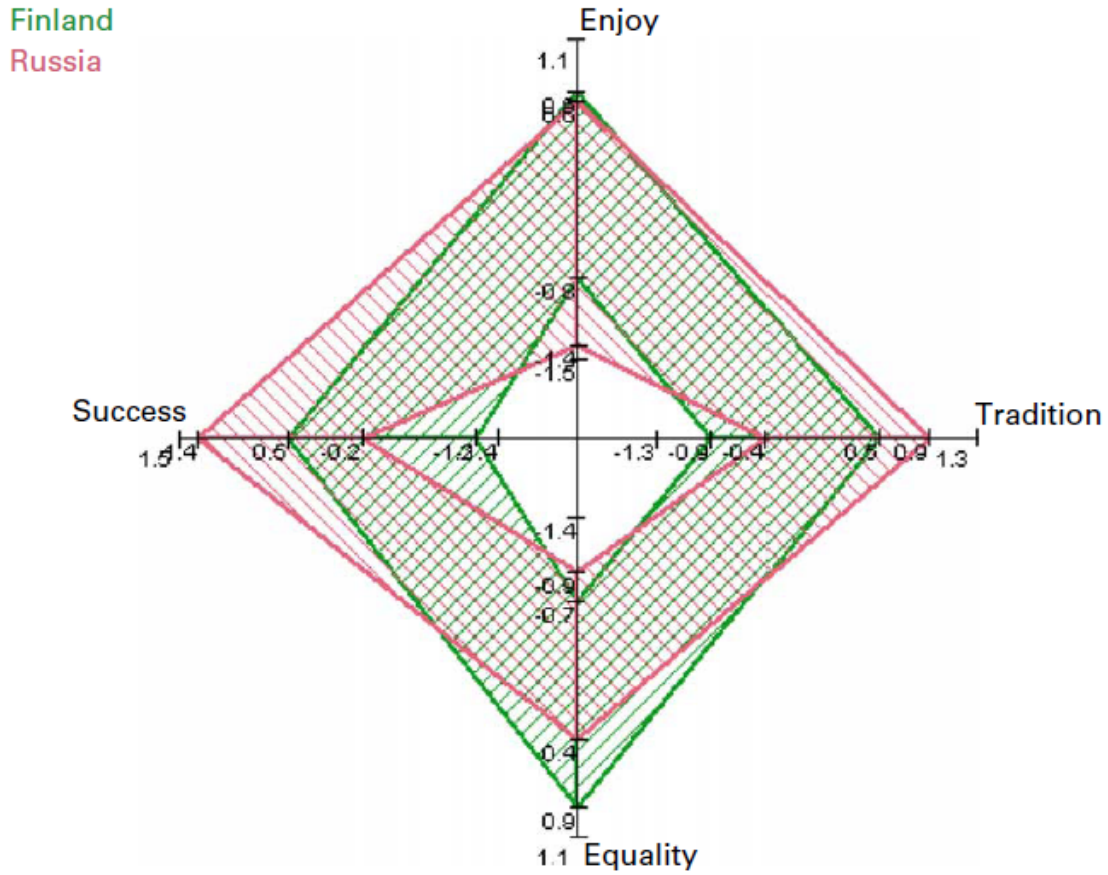
Posamezno enoto lahko prikažemo z uporabo *glif*a, ki v sebi združuje sestavine, ki predstavljajo vrednosti posameznih spremenljivk.

Znanih je več ravninskih (2D) glifov: točka v ravnini, krožni prikaz, stolpčni prikaz, zvezda, obraz Chernoffa, krivulja Andrewsa, tirnica, ... Večino teh rešitev je mogoče razširiti za prostorske prikaze. *

VisualComplexity

Zoom stars

Za prikaze simbolnih večrazsežnih podatkov so bile razvite **Zoom Stars** (Noirhomme-Fraiture M., 1997).



Ravni v podatkovju

Posamezne sestavine predstavitve enote podpirajo različna opravila (iskanje povezav, izbiranje, urejanje, vrednotenje, ...) nad enotami. Pri tem se pogosto prepletajo deli z različnih ravni

spremenljivka, enota, skupina, skupine, podatkovje

Večino postopkov analize podatkov je mogoče obravnavati kot transformacije ali odnose med temi ravnmi.

Slikovne sestavine in vrste lestvic

Posamezne vrste spremenljivk predstavljamo z ustreznimi slikovnimi sestavinami

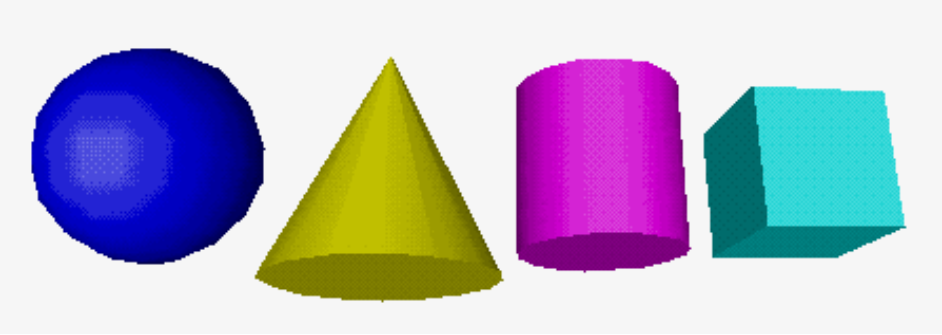
lestvica	**slikovne sestavine**
imenska	barva, oblika(lik)
urejenostna	stopnja, svetlost, gostota vzorca, razmestitev (mesto v)
številska	velikost, mesto, smer, kot

Ker so *številske* lestvice tudi *urejenostne*, in te tudi *imenske*, lahko ustrezne sestavine uporabimo tudi za predstavitve vsebovanih vrst lestvic – vendar pri tem zgubimo na natančnosti.

Uporaba neustreznih sestavin pa lahko privede do pojavitve neosnovanih povezav.

Trirazsežne predstavitve

Za razvoj trirazsežnih predstavitev se je uporabljal spletni jezik VRML (Virtual Reality Modeling Language), ki pa so ga v zadnjih letih nadomestili X3D, Kinemage, WebGL, ...

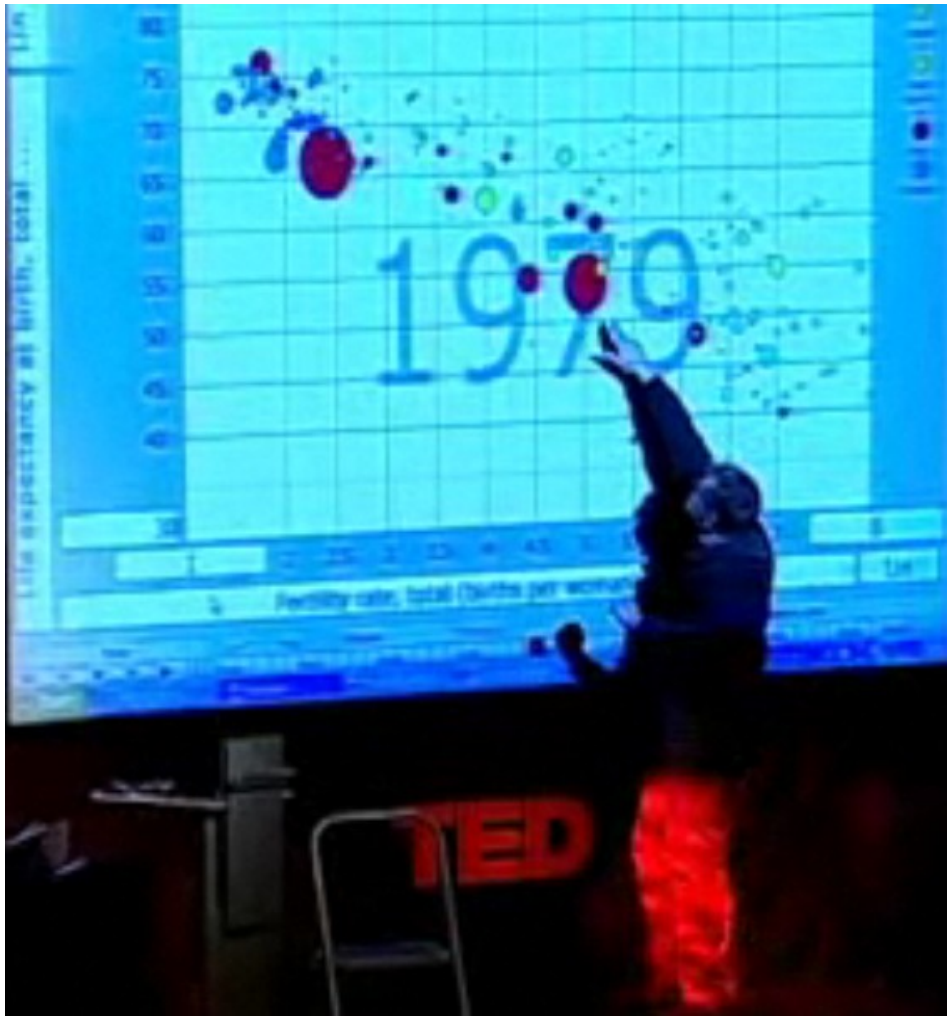


Pri tem lahko pri predstavitvah uporabimo naslednje sestavine:

- mesto v prostoru (x, y, z);
- obliko (krogla, kocka, stožec, valj, ravnina, ...;
- barva;
- velikost, kot, naklon, ploščina, prostornina;

- vzorec na ploskvah;
- smer (usmerjenost);
- besedilo;
- svetila (različni viri, sence, prozornost, odsevi, . . .);
- zasuki;
- različni pogledi in načini sprehajanja po prostoru; lastnosti kamere (pravokotna, perspektiva, stereoskopska; zorni kot).
- sodejnost

Hans Rosling – 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?

Viri

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

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

Jock Mackinlay: Information Visualization

Marti Hearst: Information Visualization and Presentation

John Stasko: Information Visualization

Wojciech Basalaj: Proximity Visualization of Abstract Data

Chris North: Information Visualization

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

Toward a Perceptual Science of Multidimensional Data Visualization

A Periodic Table of Visualization Methods

Datavisualization: izbor orodij