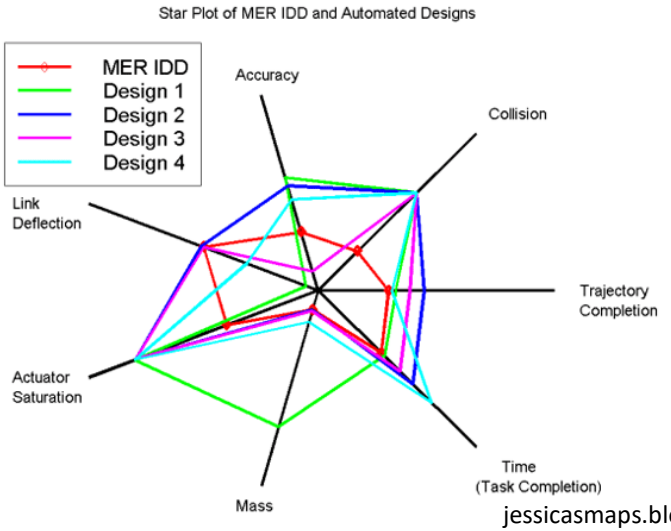
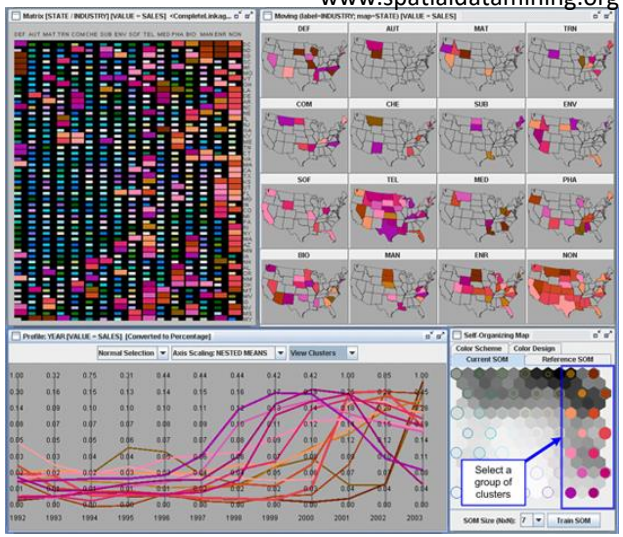
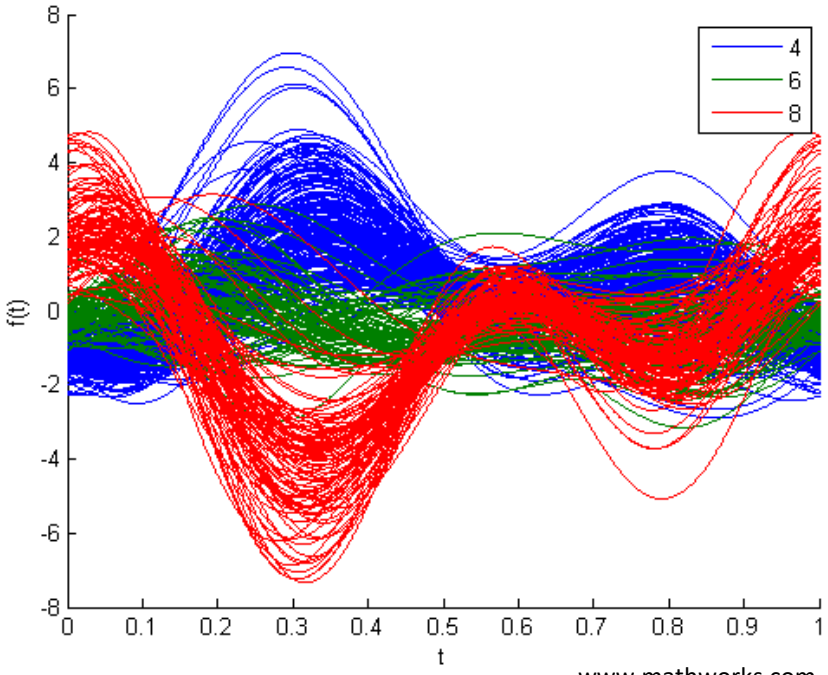


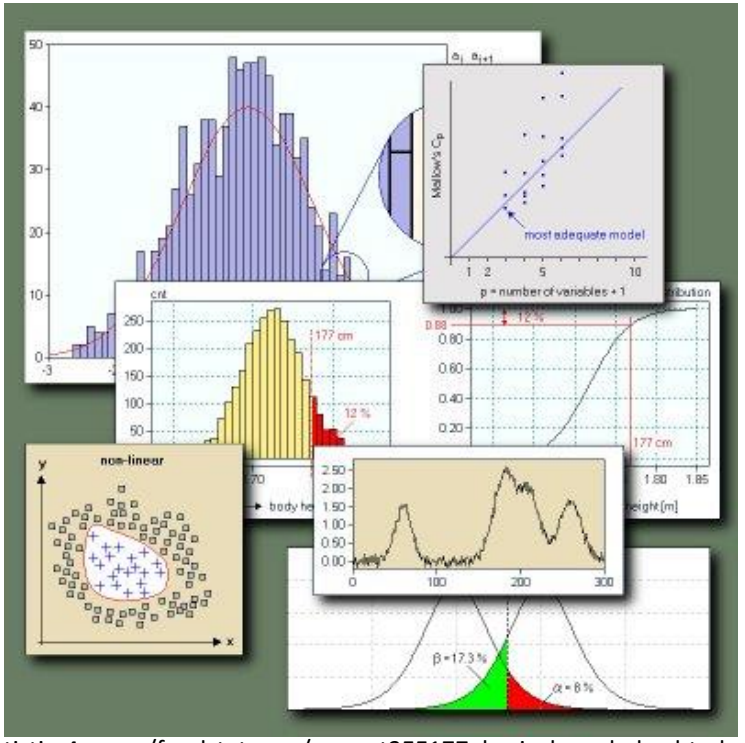
jessicasmaps.blogspot.com



6. Vizualizace multivariate dat



www.mathworks.com



http://www.statistics4u.com/fundstat_eng/wrapnt3EE177_basic_knowledge.html

Multivariate data

- Skládají se z různých typů atributů
  - Např. váha *w*, výška *h*, číslo bot *s* u náhodného vzorku osob
  - Pak trojice $(w_1, h_1, s_1), (w_2, h_2, s_2)$ je sada multivariate dat
- Techniky vizualizace seznamů a tabulek dat, které obecně nemají explicitní prostorové atributy

Techniky pro bodová data

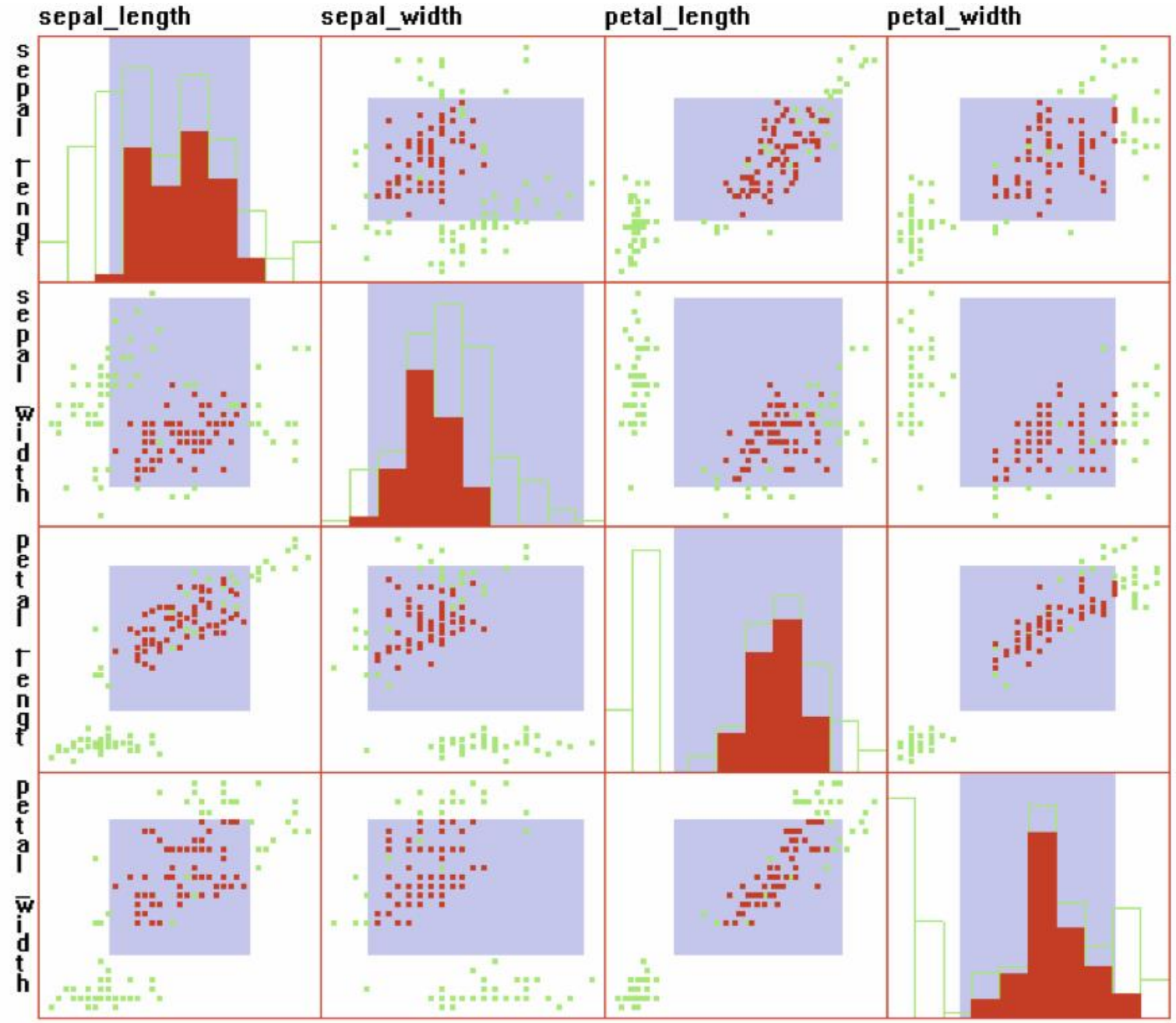
- Bodové grafy = promítání záznamů z n-dimenzionálního datového prostoru do libovolného k-dimenzionálního prostoru výstupního zařízení
- Datové záznamy jsou mapovány na k-dimenzionální body
- Každý záznam je asociován s určitou grafickou reprezentací

Bodové grafy (scatterplots)

- Jedny z prvních a nejrozšířenějších vizualizačních technik při analýze dat
- Analýza se skládá z:
  1. Hledání podmnožiny vstupních dimenzí
  2. Redukce dimenze (PCA, multidimensional scaling)
  3. Ukotvení dimenze (embedding) – mapování na další grafické atributy (barva, velikost, tvar)
  4. Násobné zobrazení – zobrazení několika grafů najednou (superimposition, juxtaposition)

Násobné zobrazení

- **Matice bodových grafů** (scatterplot matrix)
  - Mřížka obsahující bodové grafy
  - N^2 buněk, kde N je počet dimenzí
  - Každá dvojice dimenzí vykreslena dvakrát – pouze rotována o 90°
  - Obvykle symetrická podle hlavní diagonály
  - Hlavní diagonála zobrazuje
    - Popis odpovídajících dimenzí nebo
    - Histogram dané dimenze


sepal_length
sepal_width
petal_length
petal_width
sepal_length
sepal_width
petal_length
petal_width

Force-based metody

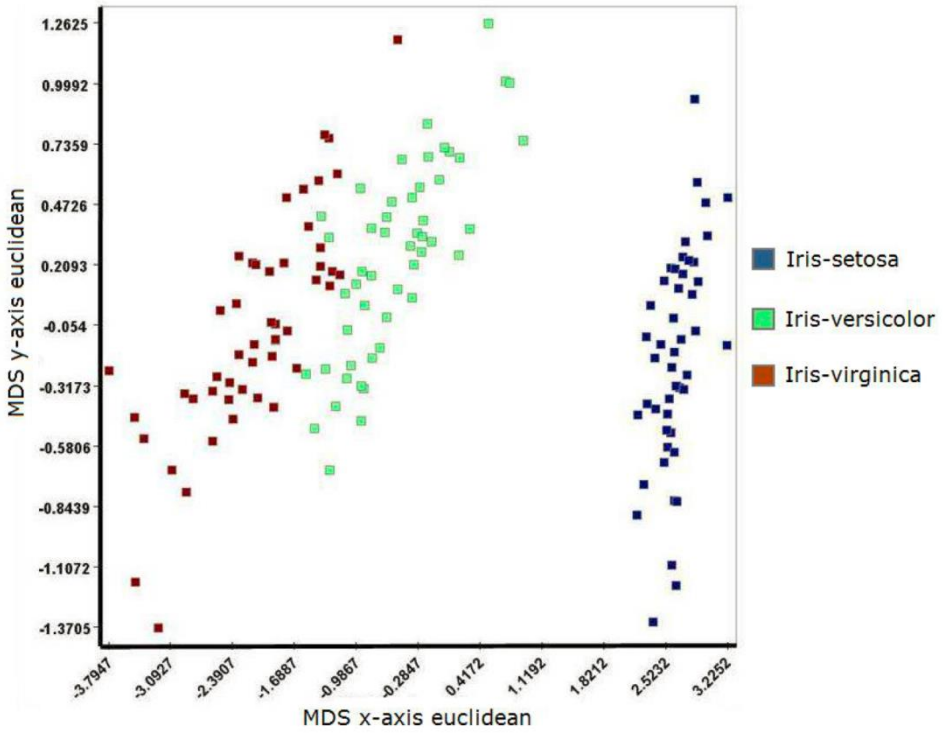
- Projekce bodů o velkých dimenzích do 2D nebo 3D prostoru
- Pokus o zachování vlastností N-dimenzionálních dat při jejich projekci do jiné dimenze
- Projekce může zavést nežádoucí artefakty projevující se ve výsledné vizualizaci

Multidimensional scaling

1. Datová množina o M záznamech a N dimenzích. Vytvoříme MxM matici D_s obsahující výsledky měření podobnosti mezi jednotlivými páry dat.
2. Předpokládejme, že vstupní data chceme promítnout do K dimenzí. Sestrojíme matici L o rozměrech MxK, která obsahuje umístění promítnutých bodů.
3. Spočteme matici L_s o rozměrech MxM obsahující podobnost mezi všemi páry bodů z L.
4. Spočteme hodnotu *stress* S – měřením rozdílů mezi D_s a L_s.
5. Pokud je S dostatečně malé, algoritmus končí.
6. Jinak posuneme pozice bodů L ve směru, ve kterém je hodnota *stress* redukována.
7. Návrat na krok 3.

Multidimensional scaling

- Řada variant tohoto algoritmu. Rozdíl spočívá v:
  - Odlišném způsobu výpočtu podobnosti a stress
  - Odlišné definici počátečních a koncových podmínek
  - Různé strategii updatování pozice bodů

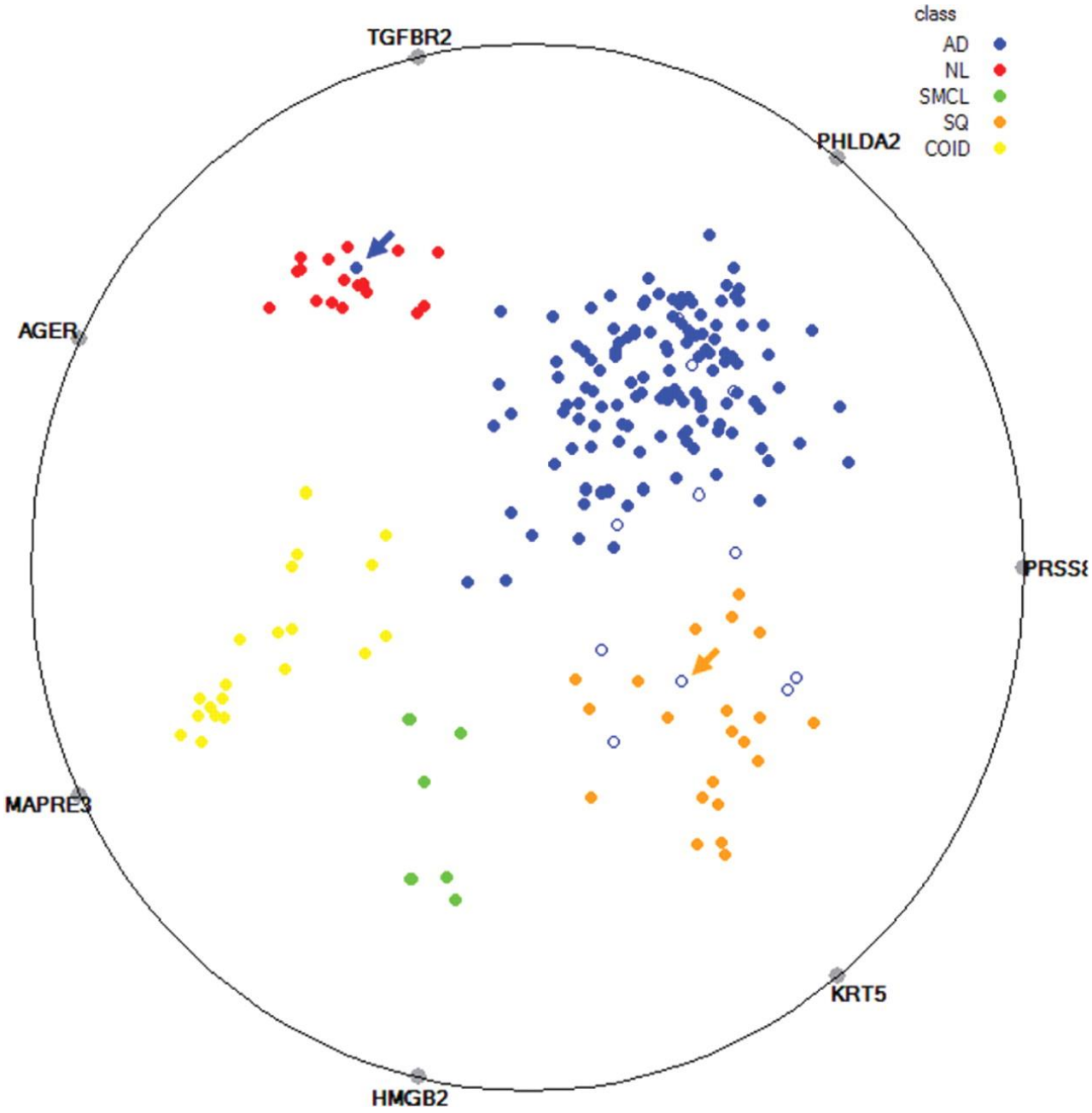


Problémy

- Výsledky nejsou unikátní – drobné změny v počátečních podmínkách mohou vést k odlišným výsledkům
- Souřadný systém po projekci nemusí být pro uživatele zcela „smysluplný“ – s ohledem na dimenze původních dat
  - Důležitá je relativní pozice jednotlivých bodů, nikoliv absolutní. Ta se může u různých algoritmů lišit.

RadViz

- Založena na Hookově zákonu rovnováhy, nalezení rovnovážné polohy bodu.
- Pro N-dimenzionální datovou množinu umístíme na obvod kružnice (pro zjednodušení jednotková, umístěná v počátku souřadné soustavy) tzv. „kotevní“ body – reprezentují fixní konce N strun přiřazených každému datovému bodu.


class
AD
NL
SMCL
SQ
COID
TGFBR2
PHLDA2
AGER
PRSS8
MAPRE3
KRT5
HMGB2

RadViz

- Pro daný normalizovaný vektor dat $D_i = (d_{i,0}, d_{i,1}, \ldots, d_{i,N-1})$ a sadu vektorů *A*, kde A_j představuje j-tý kotevní bod, dostáváme pro rovnováhu:

$$\sum_{j=0}^{N-1} (A_j - p) d_j = 0$$

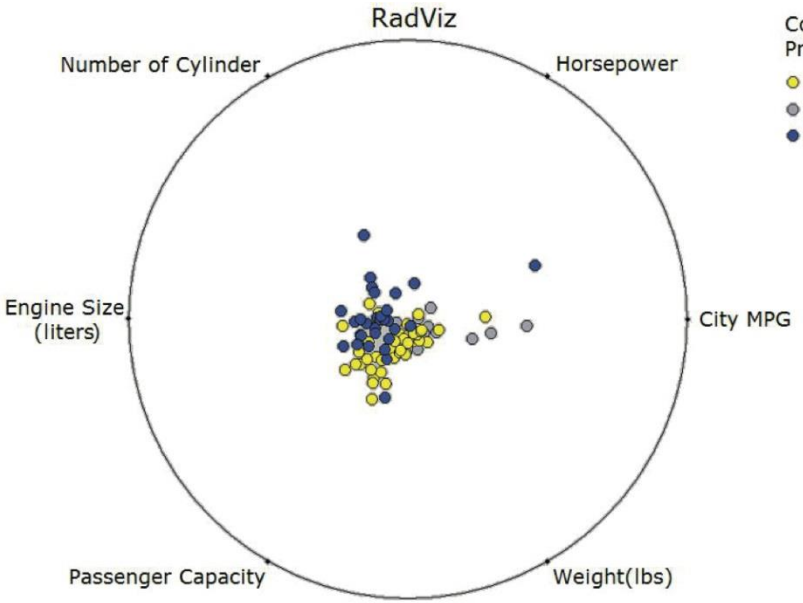
kde p je vektor pro bod rovnovážné poloze a dostaneme jej jako

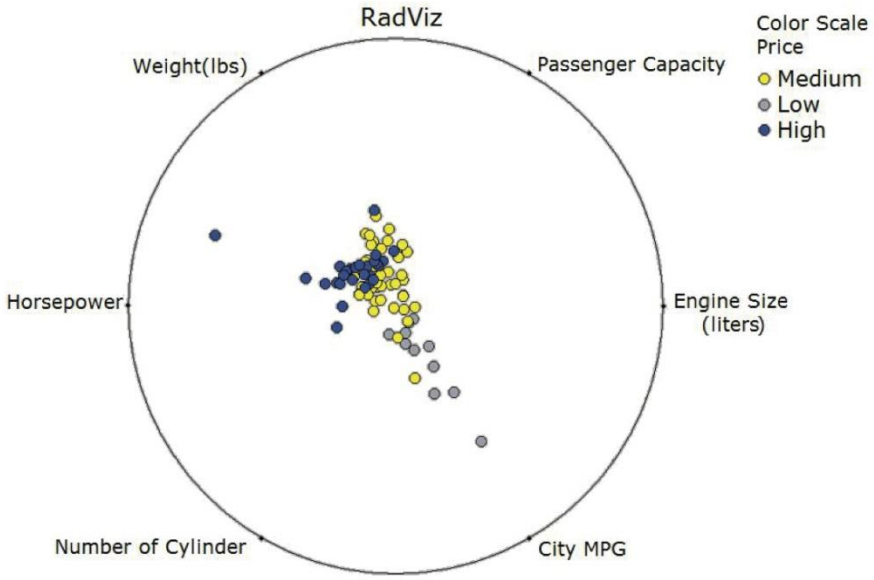
$$p = \frac{\sum_{j=0}^{N-1} (A_j d_j)}{\sum_{j=0}^{N-1} d_j}$$

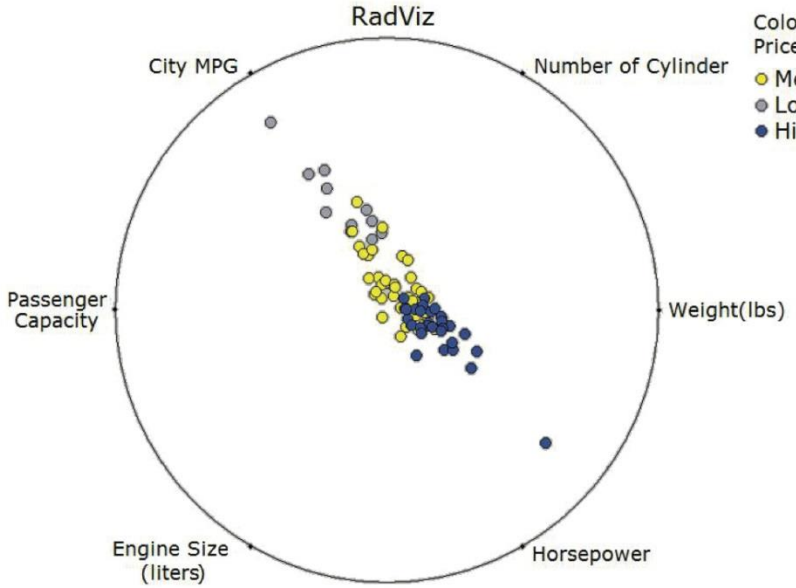
RadViz

- Různé rozmístění a uspořádání kotev vede k odlišným výsledkům
- Body, které jsou v N dimenzích odlišné, mohou být mapovány na stejné místo ve 2D
- Toto jsou problémy všech technik pro projekce a redukci dimenze
- RadViz má jednoduché řešení – zpřístupnění interakce (umožnit manipulaci s kotvami)

RadViz

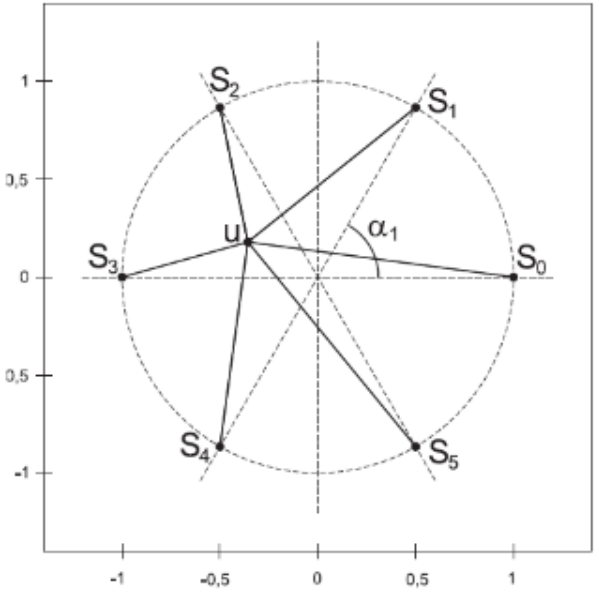

RadViz
Number of Cylinder
Horsepower
Engine Size (liters)
City MPG
Passenger Capacity
Weight(lbs)
Color Scale
Price
Medium
Low
High


RadViz
Weight(lbs)
Passenger Capacity
Horsepower
Engine Size (liters)
Number of Cylinder
City MPG
Color Scale
Price
Medium
Low
High


RadViz
City MPG
Number of Cylinder
Passenger Capacity
Weight(lbs)
Engine Size (liters)
Horsepower
Color Scale
Price
Medium
Low
High

RadViz – definice trochu jinak

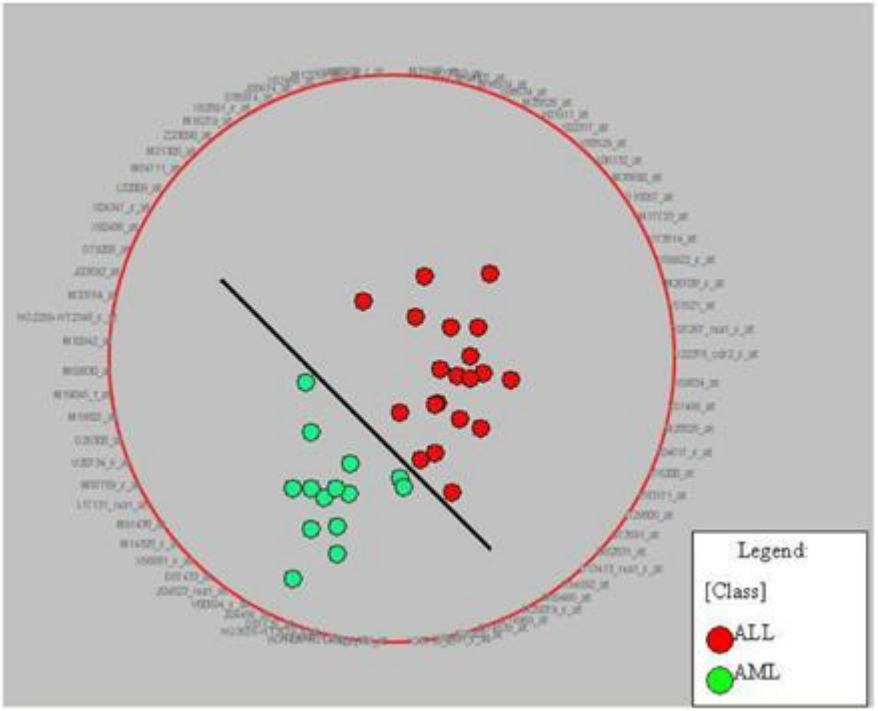
- Bod v n-dimenzionálním prostoru $[y_1, y_2, \ldots, y_n]$
- Ke každé kotvě S_j je připevněna virtuální pružina o tuhosti y_j – mění se podle hodnoty daného parametru
- Všechny pružiny jsou spojeny v jednom bodě *u*
- Požadován je vyvážený systém pružin



https://cyber.felk.cvut.cz/research/theses/papers/216.pdf

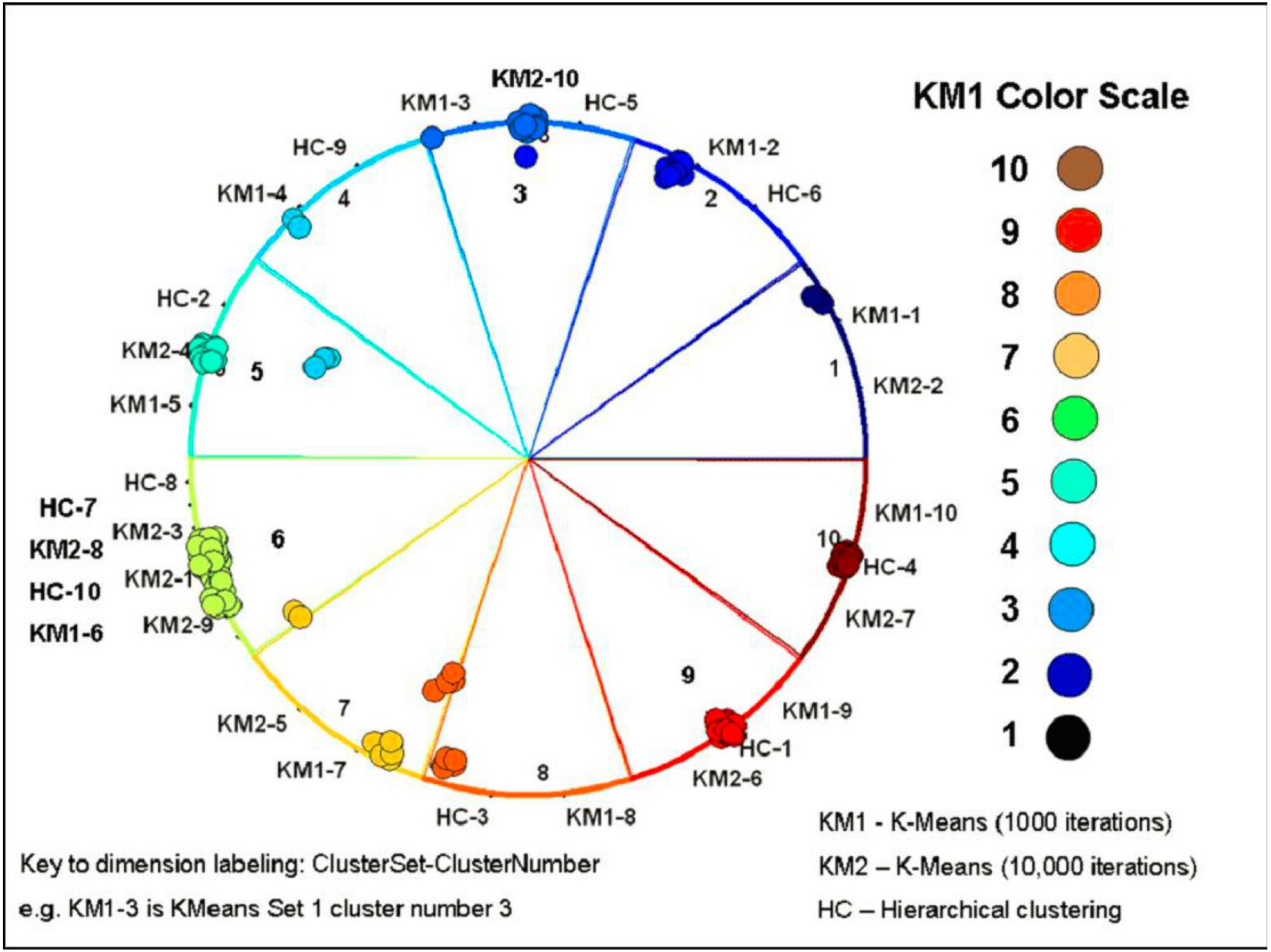
RadViz

- Vyhledávací algoritmus hledající rozložení dimenzí po kružnici, které vede k maximálnímu rozptýlení dat



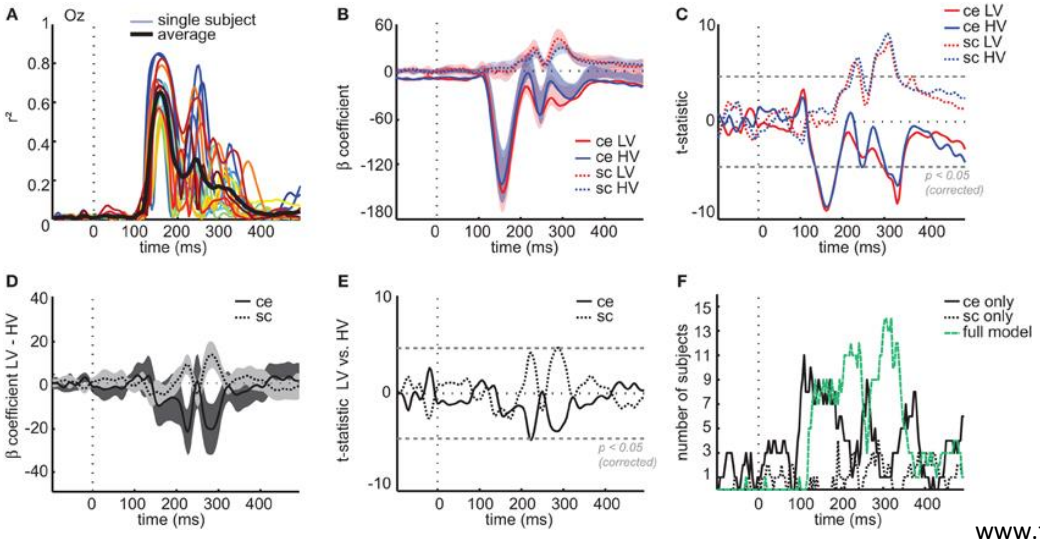
Vectorized RadViz (VRV)

- Konstrukce násobných dimenzí pro jednotlivé vstupní dimenze
- Podobné metodě třídění dat do „košů“
- Každá původní dimenze je reprezentována vektorem nových dimenzí – každá nová souřadnice ve vektoru nabývá hodnotu 0 nebo 1 podle toho, zda daný záznam obsahuje hodnotu odpovídající této dimenzi nebo ne
  - Pro každý nový vektor obsahuje právě jednu dimenzi s hodnotou 1 a všechny ostatní mají hodnotu 0


KM2-10
KM1-3
HC-5
HC-9
KM1-2
KM1-4
4
3
2
HC-6
HC-2
KM1-1
KM2-4
5
1
KM2-2
KM1-5
HC-8
HC-7
KM2-8
HC-10
KM1-6
KM2-3
6
KM1-10
10
HC-4
KM2-1
KM2-9
KM2-7
9
KM1-9
KM2-5
7
HC-1
KM1-7
8
KM2-6
HC-3
KM1-8
KM1 Color Scale
10
9
8
7
6
5
4
3
2
1
Key to dimension labeling: ClusterSet-ClusterNumber
e.g. KM1-3 is KMeans Set 1 cluster number 3
KM1 - K-Means (1000 iterations)
KM2 – K-Means (10,000 iterations)
HC – Hierarchical clustering

Techniky pro čárová data

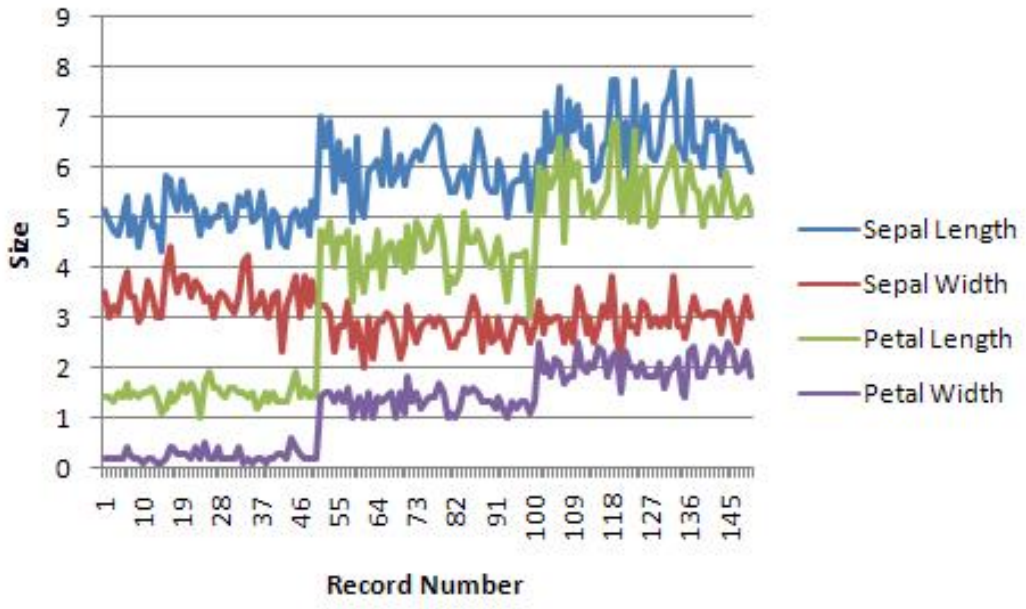
- Záznamy jsou zobrazeny tak, že odpovídající body jsou spojeny přímou či zakřivenou čarou
- Tyto čáry mohou navíc pomocí dalších vlastností, jako je zakřivení, křížení apod. zobrazit další vztahy mezi daty



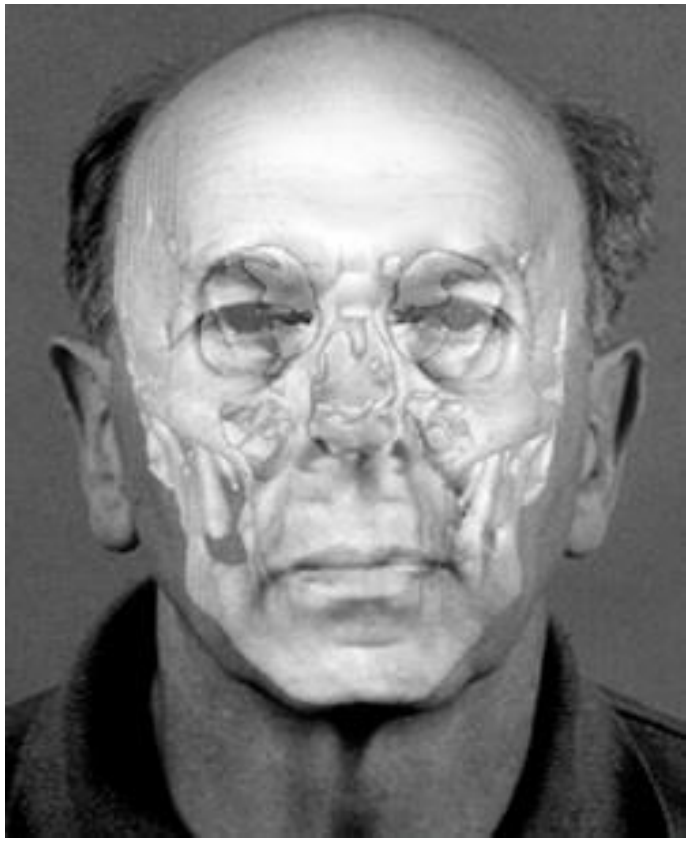
www.frontiersin.org

Čárové grafy

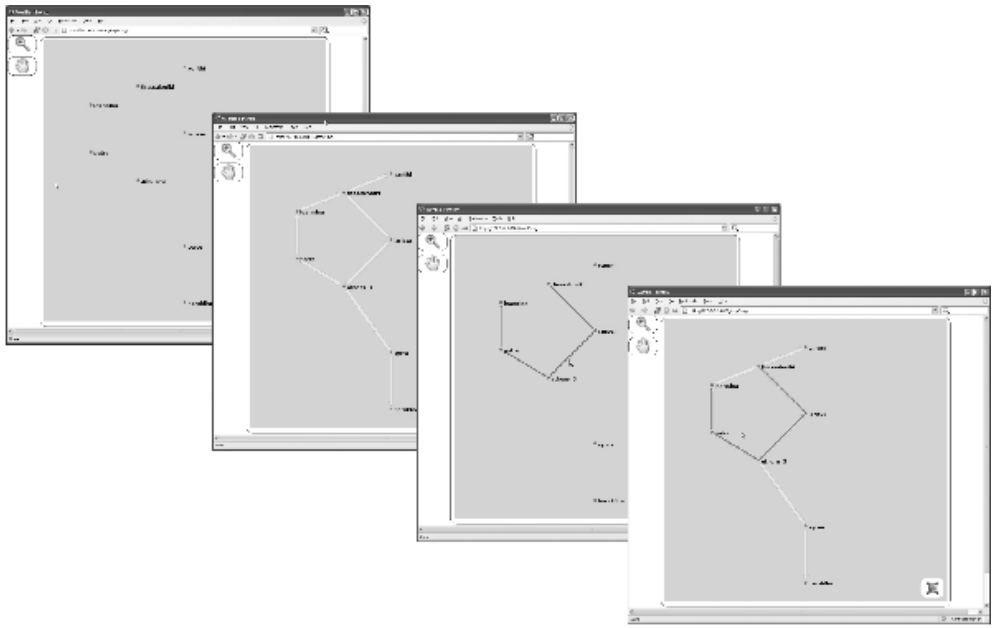
- Vizualizační technika o jedné proměnné, kdy vertikální osa reprezentuje možný rozsah hodnot proměnných a horizontální osa reprezentuje jisté uspořádání záznamů v dané datové množině
- Rozšíření na více proměnných – superimposition, juxtaposition



Superimposition vs. juxtaposition



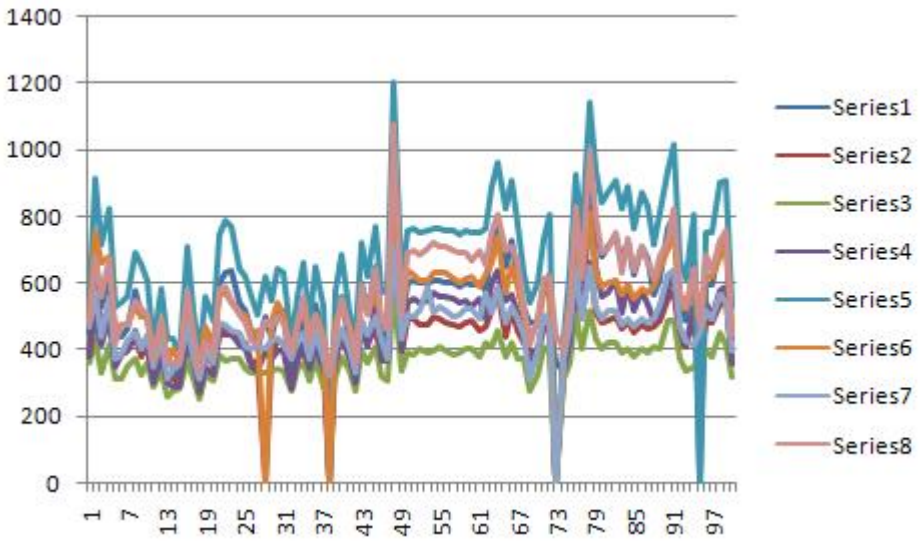
www.craniofacial-id.com

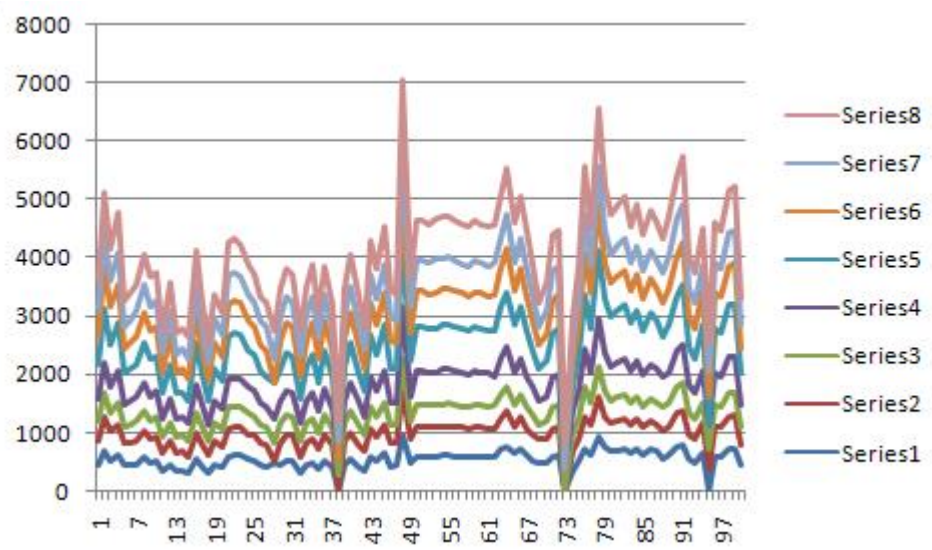


www.usenix.org

Čárové grafy

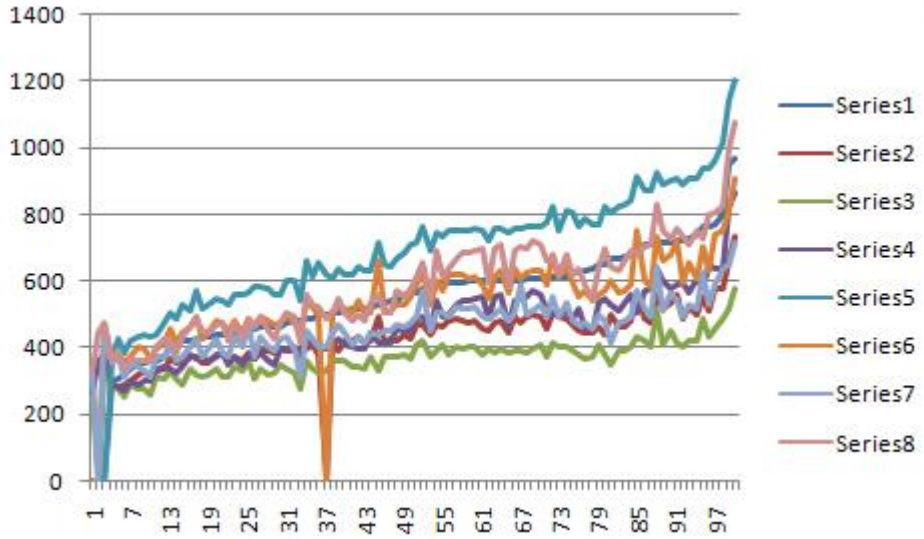
- Klasický čárový graf pro 8-mi dimenzionální datovou množinu vs. vrstvený čárový graf (pro každou další dimenzi je použit jako základ graf předchozí dimenze)

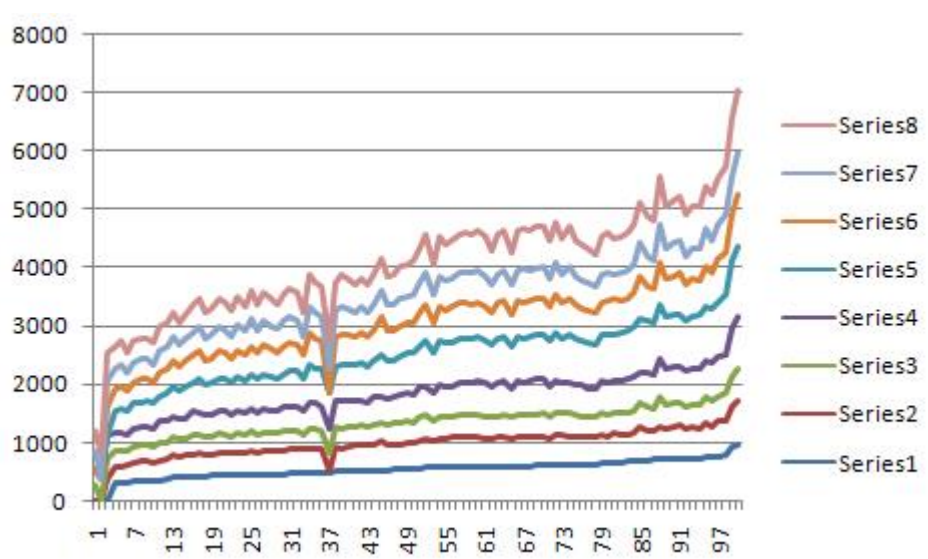




Čárové grafy

- Třídění záznamů podle jedné dimenze





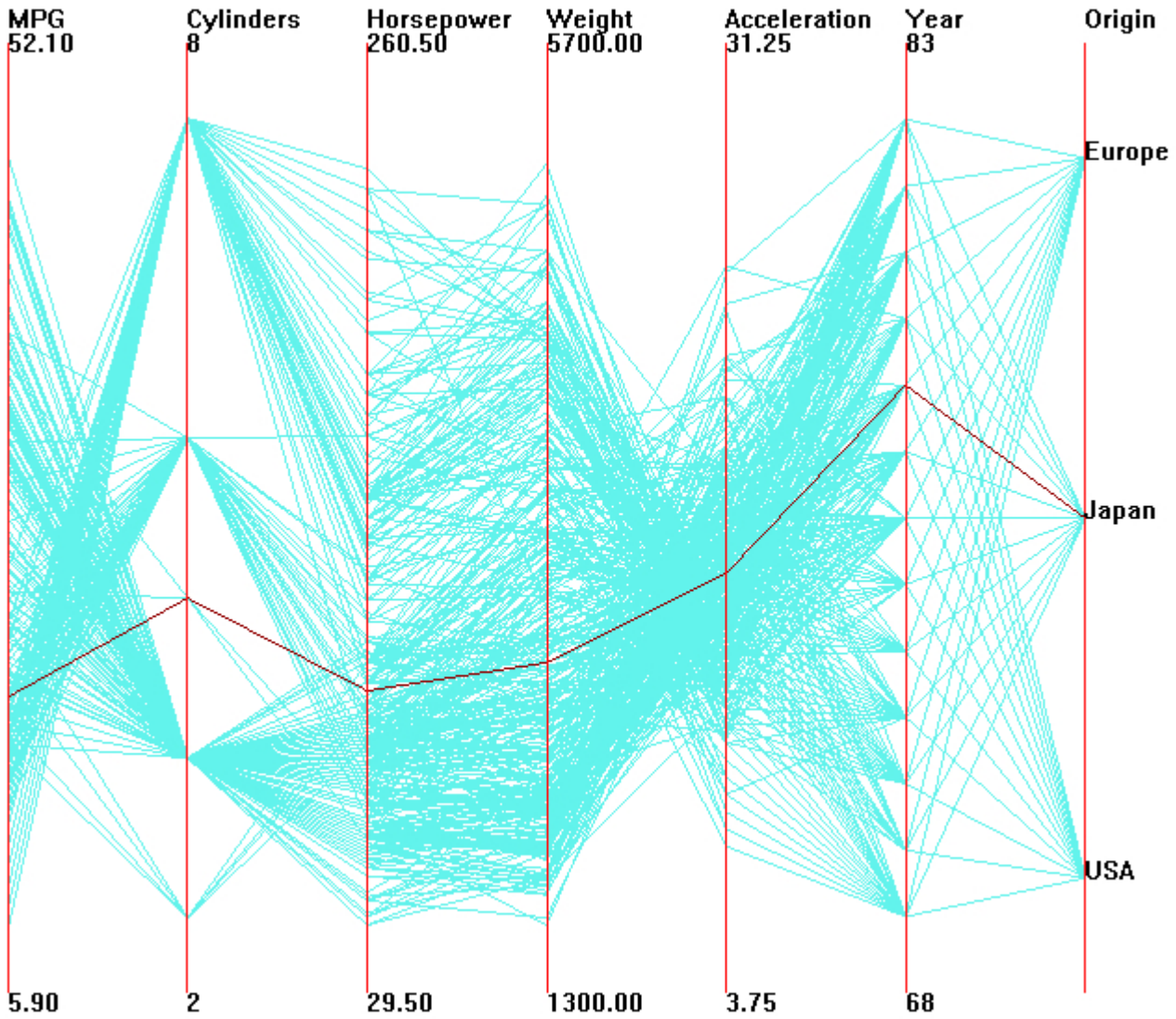
Čárové grafy

- Pokud mají dimenze společné jednotky pro osy, pak je možné využít výše uvedených technik
- Pokud však mají jednotlivé proměnné různé jednotky, je nutné použít jiné metody, např:
  - Využití násobných vertikálních os
  - Vertikální skládání grafů pro jednotlivé dimenze

Paralelní souřadnice

- Zavedeny v roce 1985 (Inselberg) jako mechanismus pro studium geometrie o vyšších dimenzích
- Rozšiřující metody pro analýzu multivariate dat
- Osy jsou místo ortogonálního umístění rozmístěny paralelně za sebou
- Datový bod je vykreslen jako polyčára, která protíná každou osu na pozici úměrné své hodnotě v odpovídající dimenzi

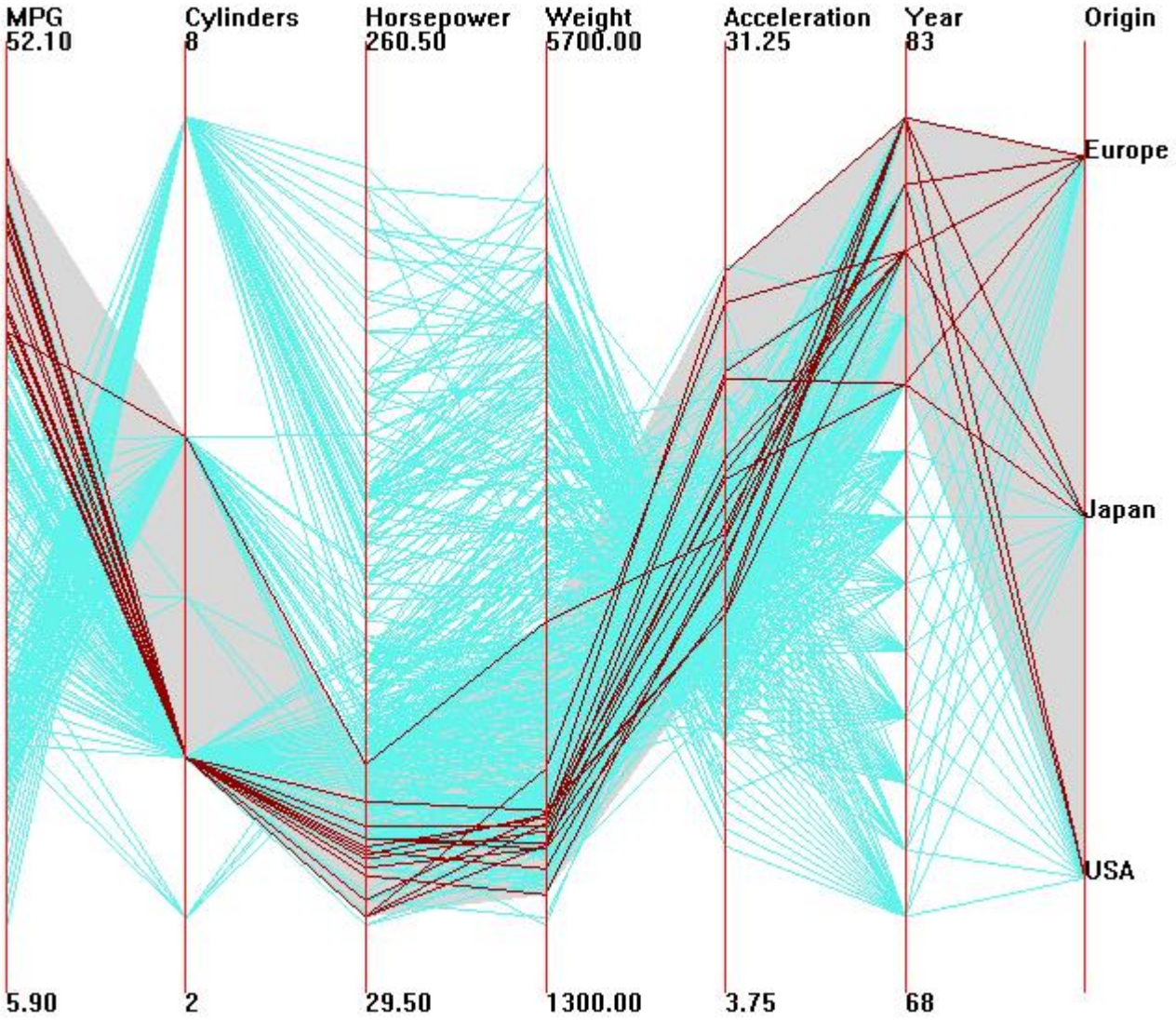
Paralelní souřadnice



Paralelní souřadnice

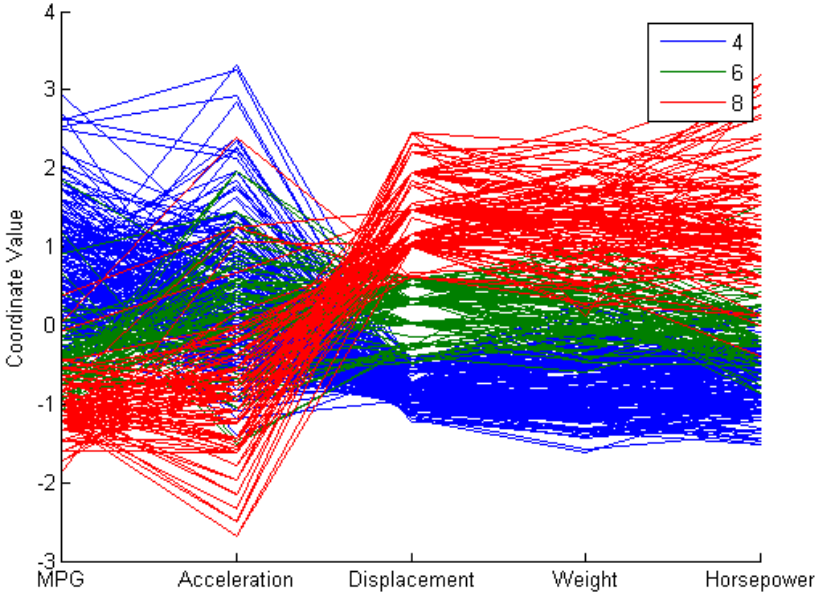
- Interpretace grafu – díváme se na:
  - Podobné čáry
  - Podobné průsečíky a čáry, které jsou buď izolované nebo mají výrazně odlišný sklon od svých sousedů
- Problém: paralelní souřadnice dokáží zobrazit pouze vztahy mezi dvojicemi dimenzí
- Pomocí interaktivního výběru a zvýrazňování záznamů může uživatel pozorovat vztahy, které pokrývají všechny dimenze

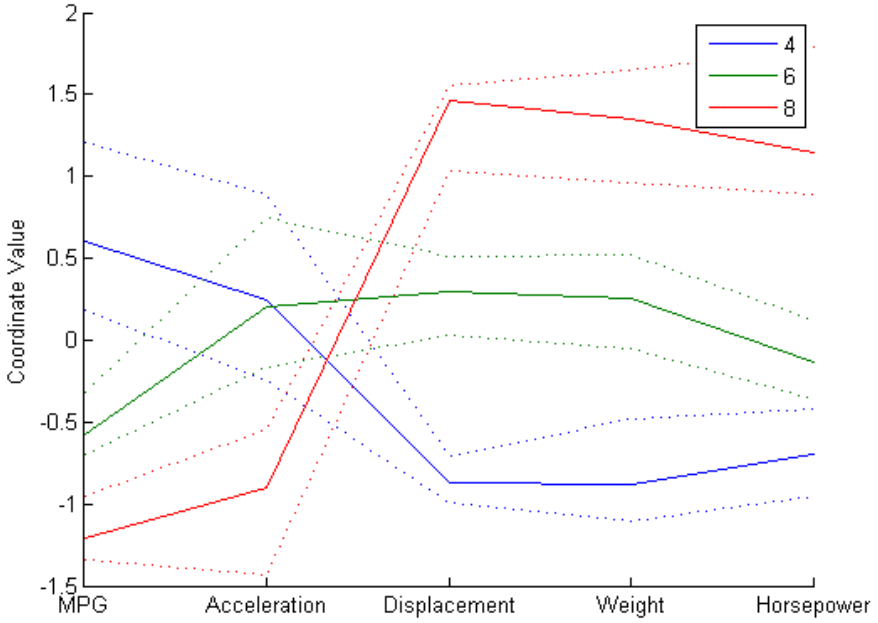
Paralelní souřadnice – interaktivní výběr



Paralelní souřadnice - medián

- Při velkém množství dat nepřehledné





Paralelní souřadnice - vylepšení

- Hierarchické paralelní souřadnice
- Použití poloprůhledných čar
- Klastrování, přeskupování
- Shlukování dat do pásu klastrů
- Zahrnutí histogramů
- Napasování křivek na průsečíky
- …

Andrewsovy křivky

- Vyvinuty v roce 1972 Davidem F. Andrewsem
- Každý multivariate datový bod $D = (d_1, d_2, ..., d_N)$ je využit pro vytvoření křivky ve tvaru

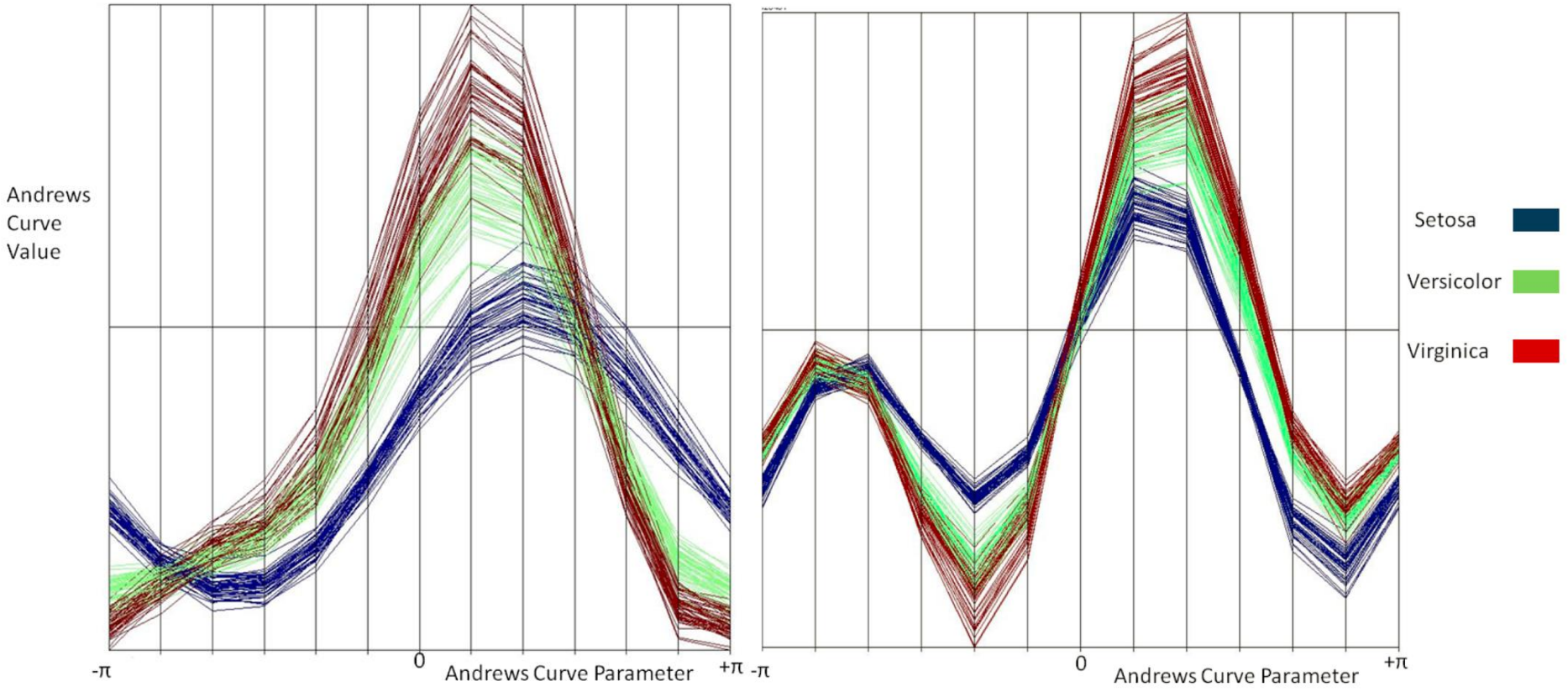
$$f(t) = \frac{d_1}{\sqrt{2}} + d_2 \sin(t) + d_3 \cos(t) + d_4 \sin(2t) + d_5 \cos(2t) + \ldots$$

pokud je počet dimenzí lichý, pak poslední člen je ve tvaru: $\cos\left(\frac{N-1}{2}t\right)$

pokud je sudý: $\cos\left(\frac{N}{2}t\right)$

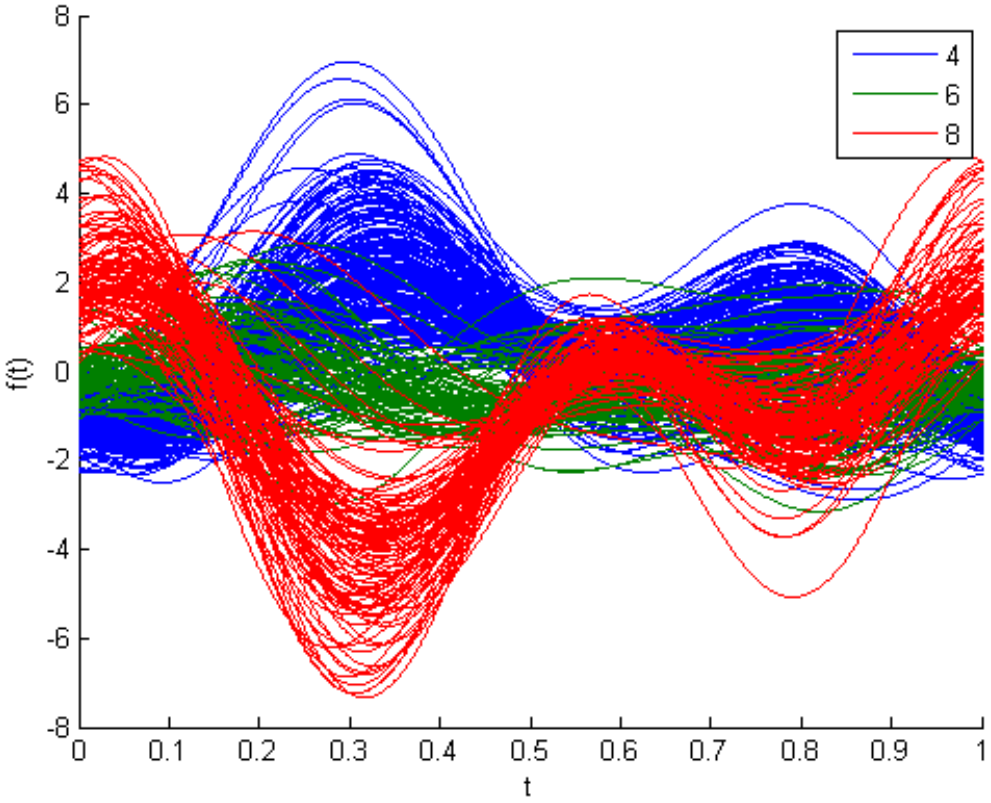
Andrewsovy křivky

- Pořadí dimenzí má vliv na výsledný tvar křivky



Andrewsovy křivky

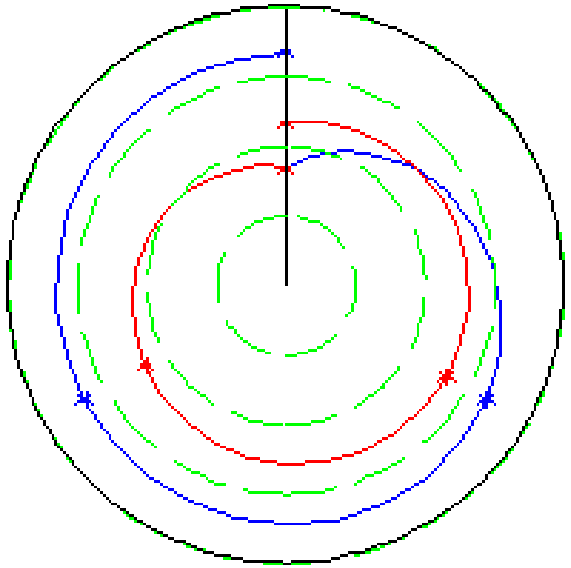
- Vyhlazení



http://www.mathworks.com/products/statistics/examples.html?file=/products/demos/shipping/stats/mvplotdemo.html#7

Techniky radiální osy

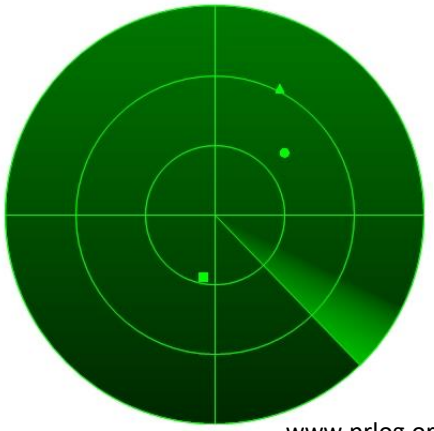
- Pro každou techniku s horizontální a/nebo vertikální orientací souřadného systému existuje ekvivalentní technika využívající radiální orientaci
- Kruhový čárový graf



publib.boulder.ibm.com

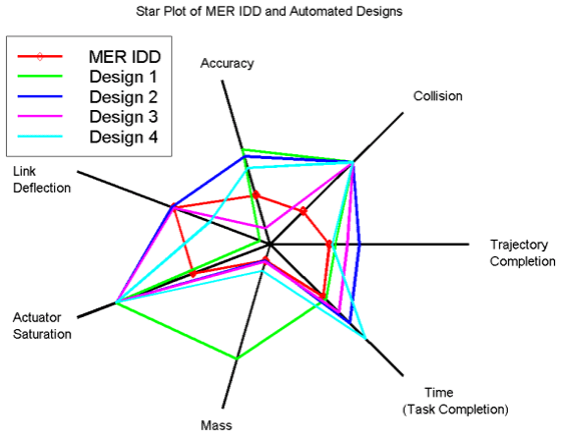
Další techniky

- Radar



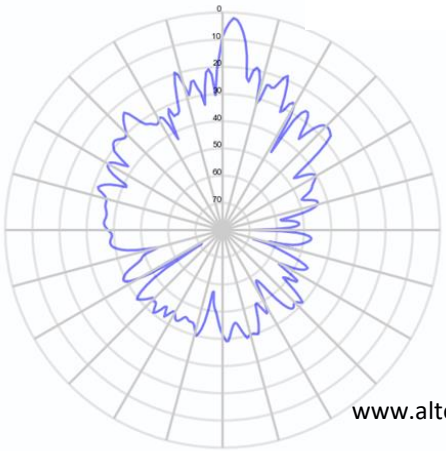
www.prlog.org

- Hvězdicové grafy



commons.wikimedia.org

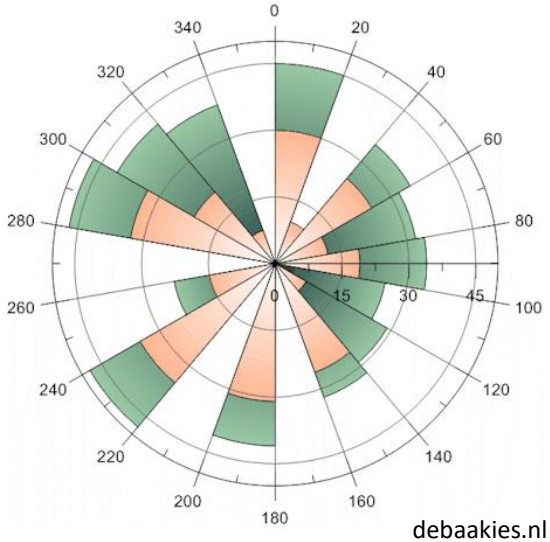
- Polární grafy
  - Zobrazení polárních souřadnic



www.alteryx.com

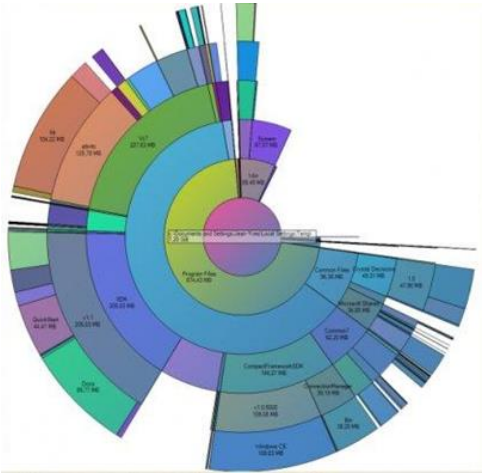
Další techniky

- Kruhové sloupcové diagramy



debaakies.nl

- Kruhové sloupcové grafy

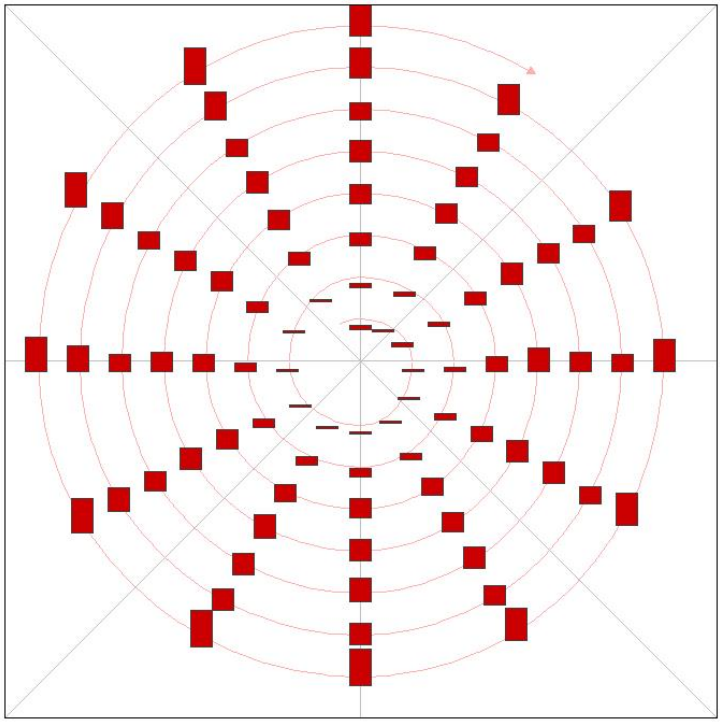


pcprogramasgratis.com

- Kruhové plošné grafy

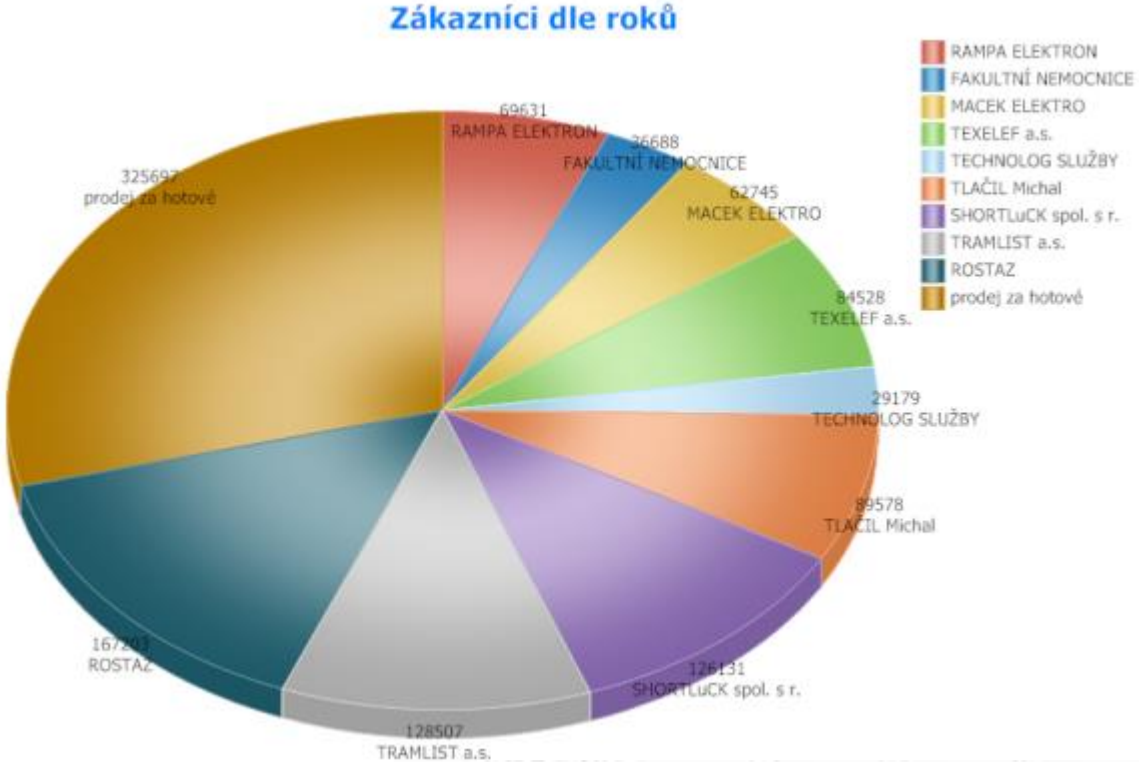
Typy technik pro radiální osy

- Soustředné kružnice
- Spojitá spirála – nevykazuje nespojitosti na konci každého cyklu
- Oproti tradičnímu sloupcovému vyjádření dovoluje sledovat vzory mezi prvky na stejné pozici v různých cyklech

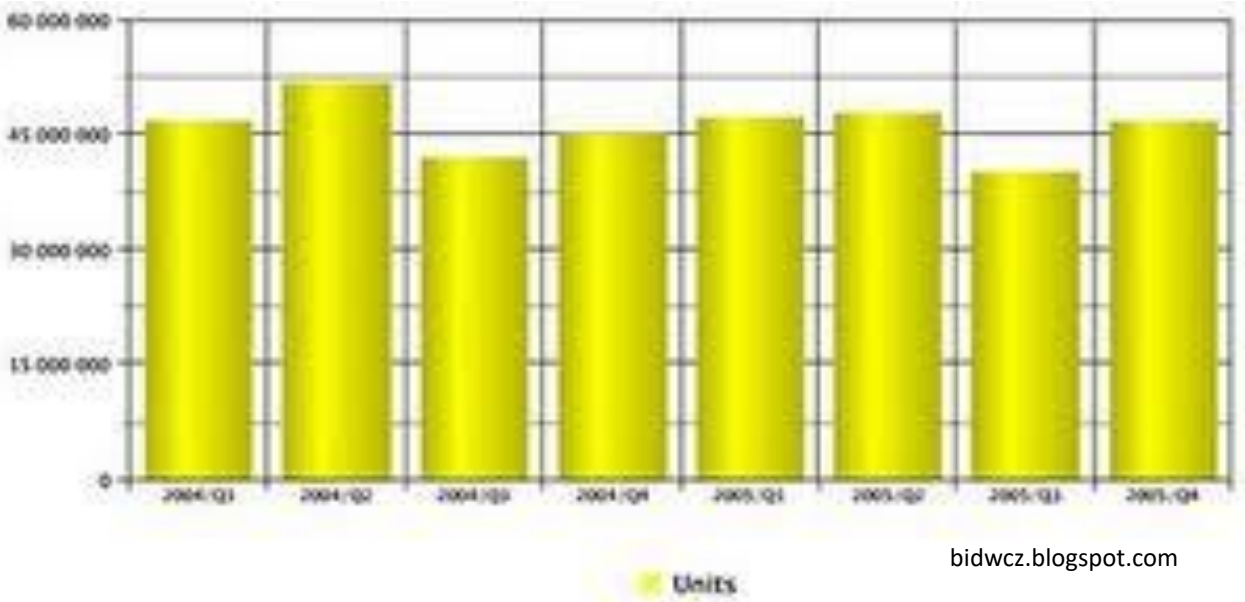


Techniky pro plošná data

- Využití vyplněných polygonů o dané velikosti, tvaru, barvě, …
- Cílem těchto technik není ukazovat jednotlivá data, ale jejich shluky a rozložení
- Původně navrženo pro univariate data (jedna proměnná) – koláčové a sloupcové grafy. Následně rozšířeno do více dimenzí



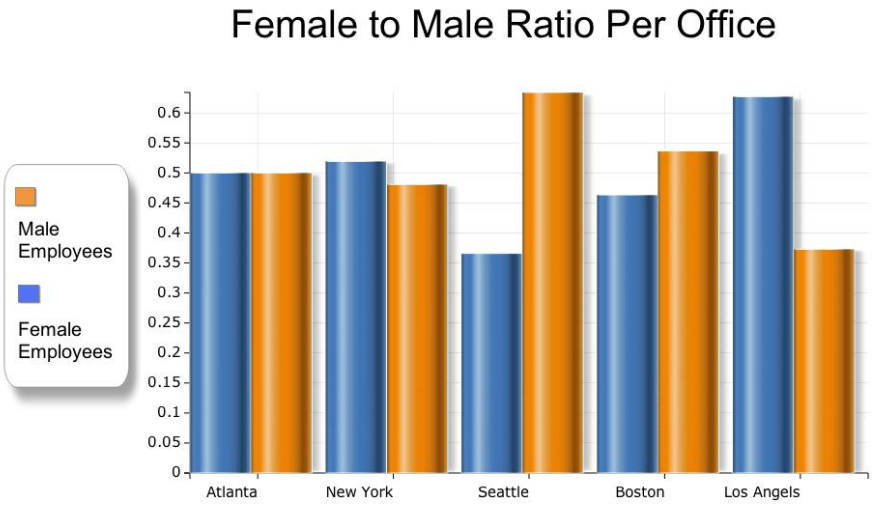
red.helios.eu



bidwcz.blogspot.com

Sloupcové diagramy/histogramy

- Pro zobrazení numerických hodnot využity obdélníkové sloupce
- Efektivní díky schopnostem lidského vnímání dobře rozpoznat délku a obecné lineární vlastnosti
- Sloupcům jsou běžně přiřazovány textové popisky

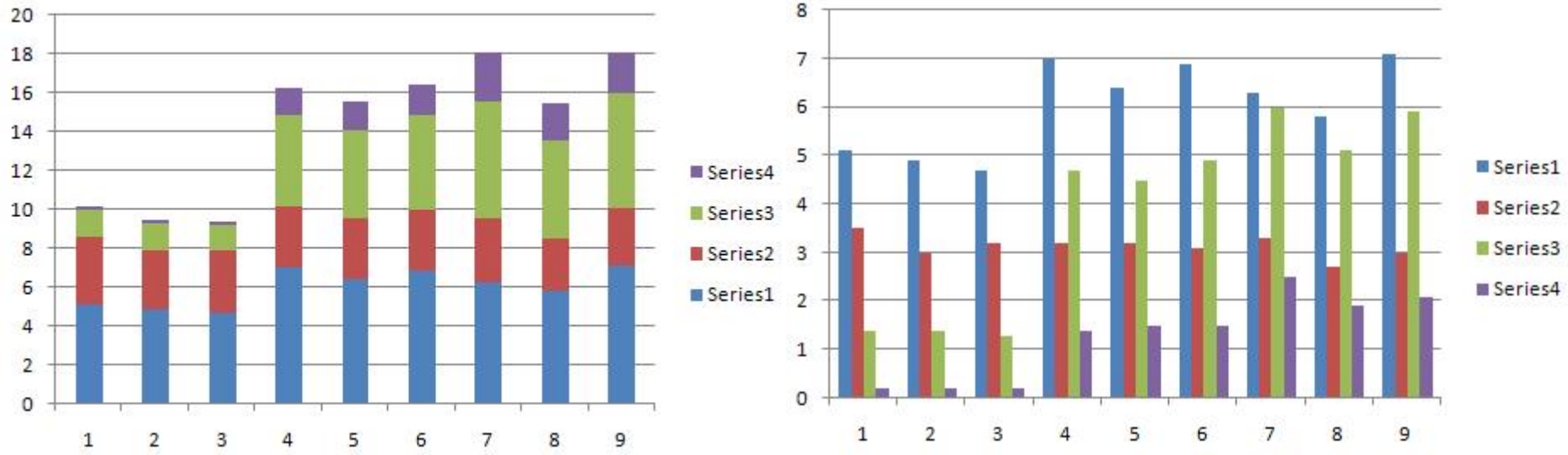


Sloupcové diagramy/histogramy

- Zásadní je určení, kolik sloupců je zapotřebí pro co nejlepší reprezentaci dat
- Mějme N proměnných. Pokud N není velké, můžeme využít mapování 1:1
- Chceme-li zobrazit souhrn nebo rozložení datové množiny, využijeme **histogram**
- Nominální hodnoty – tolik sloupců, kolik je různých hodnot
- Ordinální hodnoty – vytvoření intervalů hodnot, každý interval odpovídá jednomu sloupci

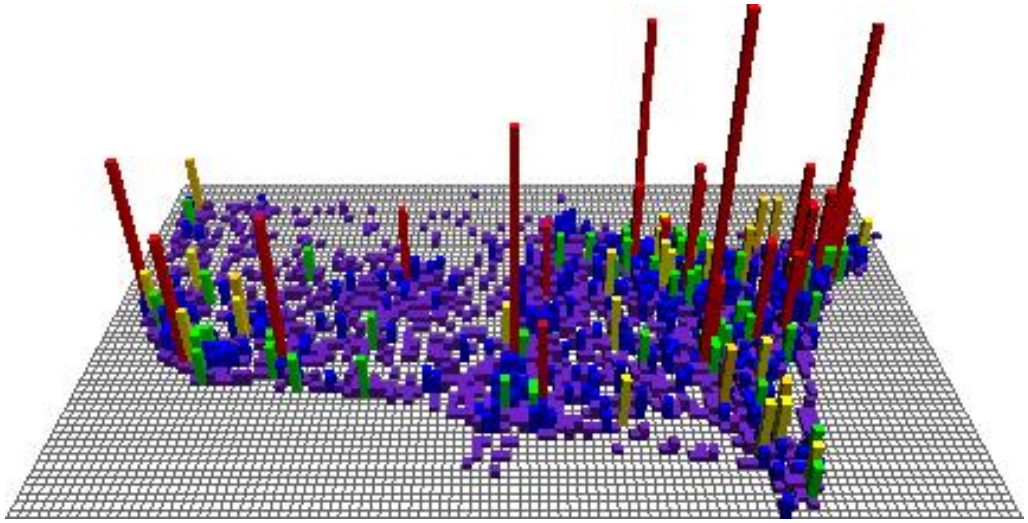
Sloupcové diagramy/histogramy

- Multivariate data – vrstvený sloupcový graf



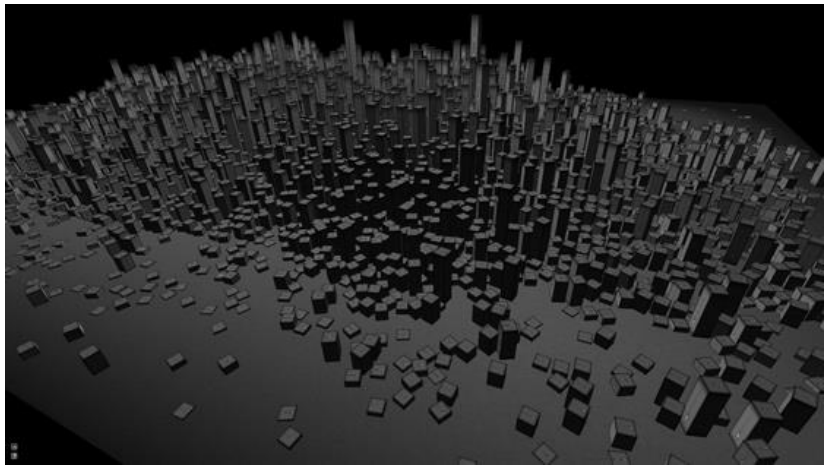
Cityscapes

- Využití 3D kvádrů namísto 2D obdélníků
- Sloupce rozloženy v mřížce, 2 dimenze určují pozici, další dimenze barvu, výšku
- Název odvozen od vzhledu – připomíná budovy ve městě
- Obsazeny všechny buňky mřížky = **3D histogram**



Problémy 3D sloupcových grafů

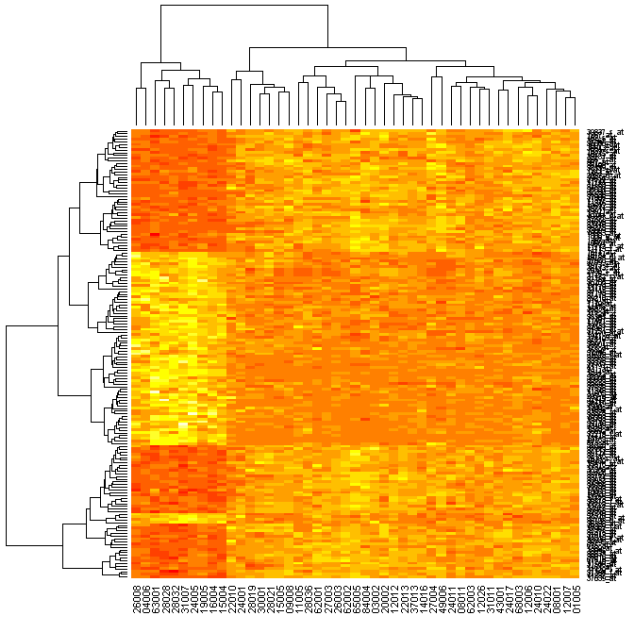
- Částečné překrytí
- Možná řešení:
  - Umožnit uživateli rotovat se scénou
  - Zmenšení tloušťky sloupců
  - Změna průhlednosti jednotlivých sloupců



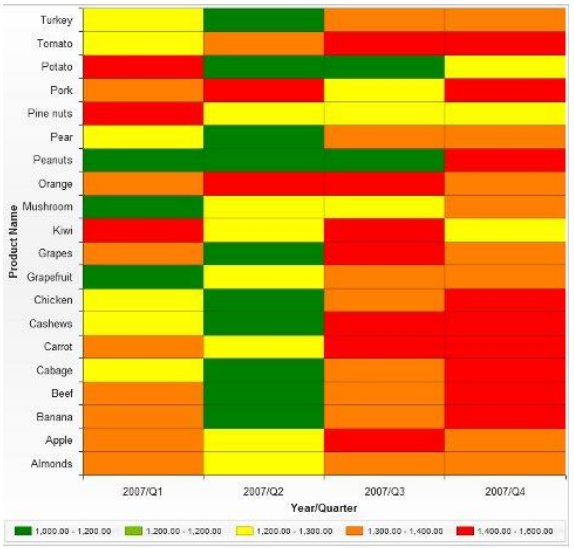
todaycreate.com

Tabulková zobrazení

- Multivariate data často v tabulkách
- **Heatmapy**
  - zobrazení záznamů pomocí barvy místo textu
  - každá hodnota renderována jako barevný čtverec

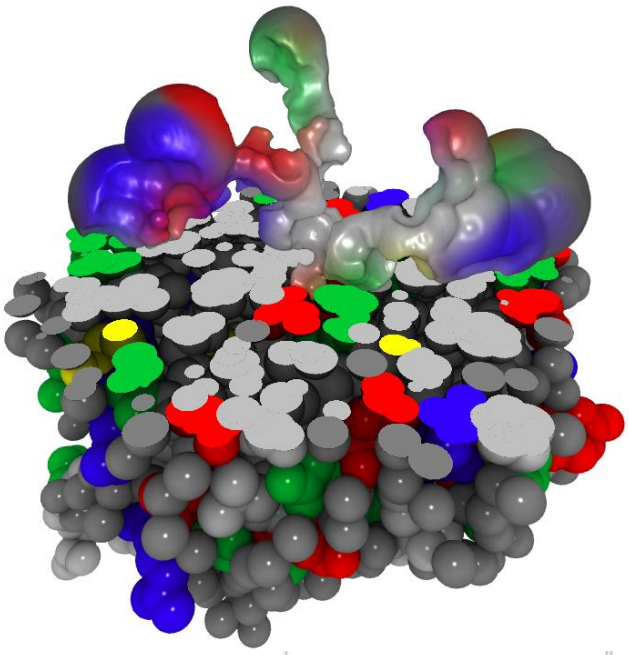


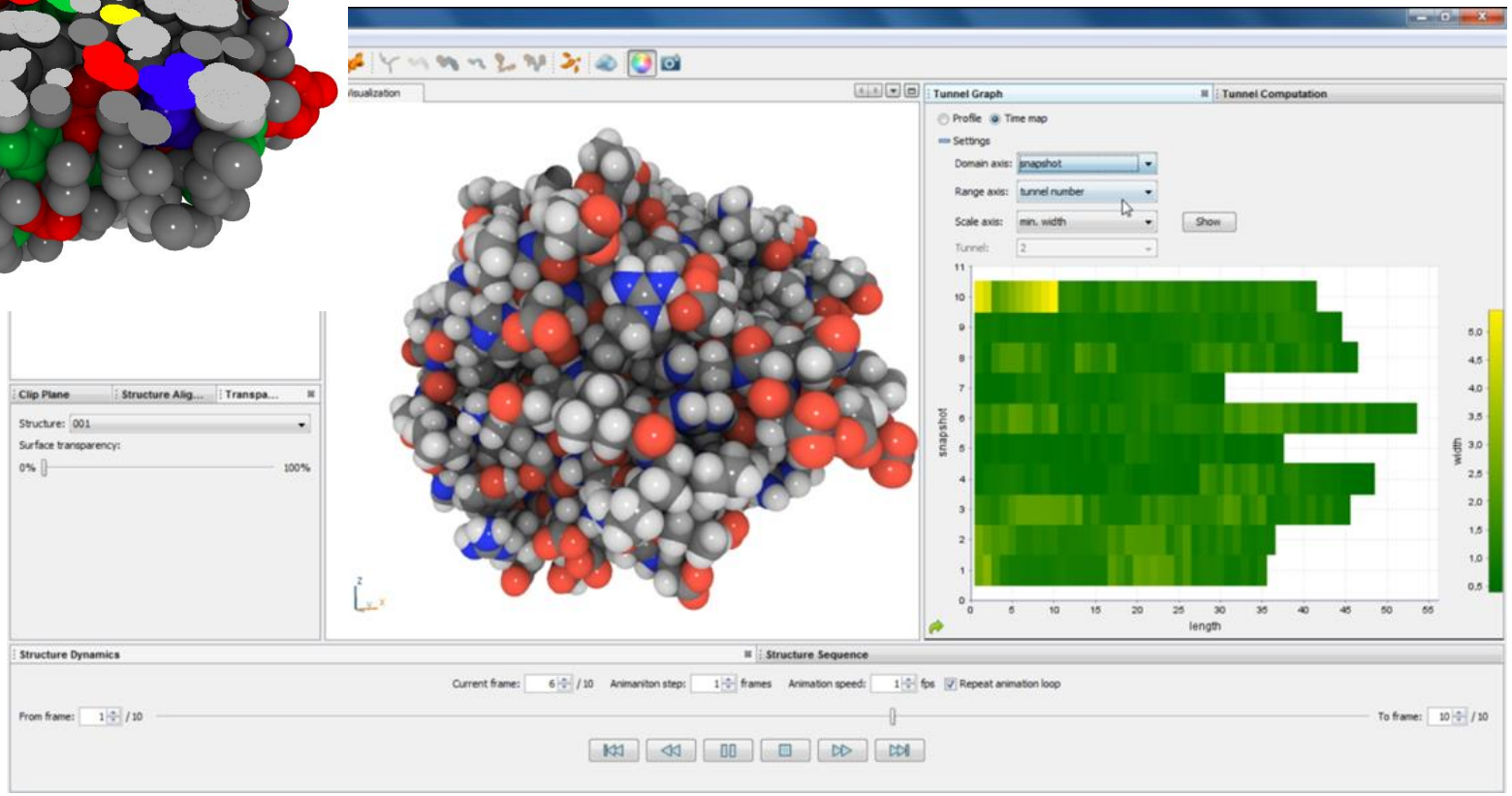
www2.warwick.ac.uk



akweebeta.com

Příklad použití



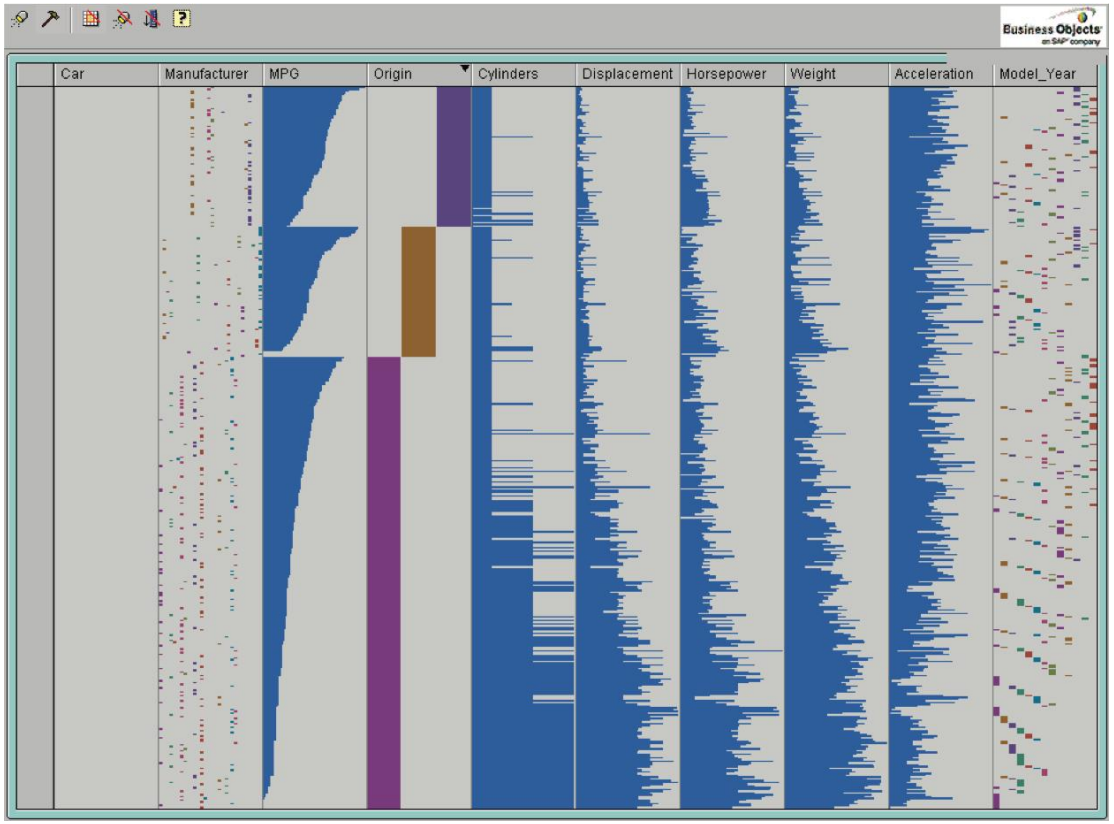


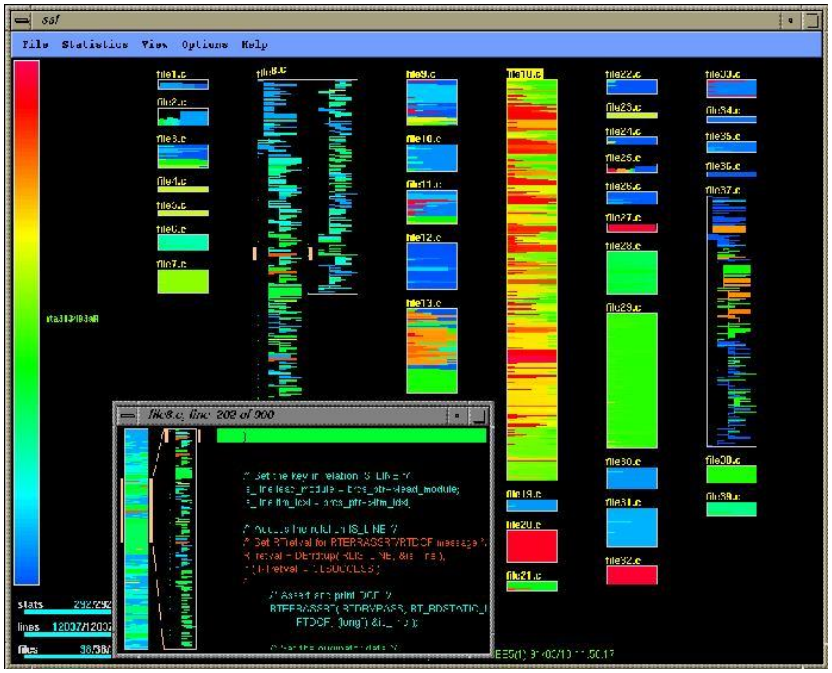
Tabulková zobrazení

- **Survey plot**
  - Namísto barvy buněk pracujeme s jejich velikostí
  - Středy buněk zarovnány na jednotlivé atributy
  - Měření plochy je ale více náchylné k chybě než měření délky

Tabulková zobrazení

- Kombinace dosavadních přístupů do level-of-detail techniky





http://ds.cc.yamaguchi-u.ac.jp/~ichikay/pfp7/iv/pics/SeeSoft-line.jpg

http://ds.cc.yamaguchi-u.ac.jp/~ichikay/pfp7/iv/pics/SeeSoft-line.jpg

Skládání (stacking) dimenzí

- Mapování dat z diskrétního N-dimenzionálního prostoru do 2D obrázku takovým způsobem, že je minimalizováno překrytí dat za současného zachování většiny prostorové informace

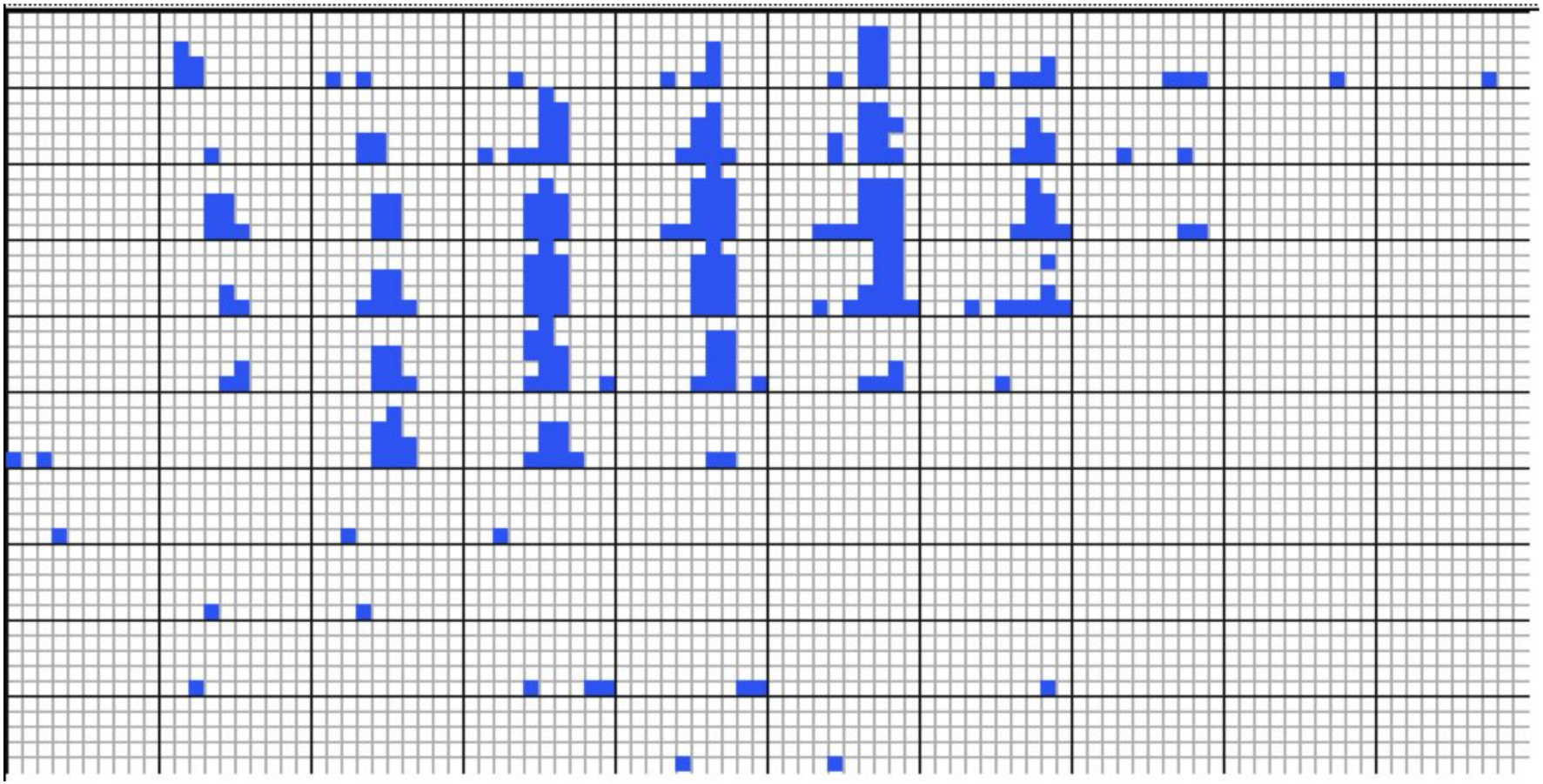
Skládání (stacking) dimenzí

- Data o dimenzi 2N+1
- Vybereme konečnou kardinalitu pro každou dimenzi
- Jednu z dimenzí zvolíme jako závislou proměnnou, ostatní dimenze jsou nezávislé
- Vytvoříme uspořádané dvojice nezávislých proměnných (N párů) a každému páru přiřadíme jeho jedinečnou hodnotu (rychlost) – od 1 do N
- Dvojice odpovídající rychlosti 1 vytvoří virtuální obraz o velikosti odpovídající kardinalitě jejích dimenzí
- V každé pozici tohoto virtuálního obrazu je vytvořen další virtuální obraz, který odpovídá dimenzím o rychlosti 2
- Proces je opakován, dokud nejsou zahrnuty všechny dimenze

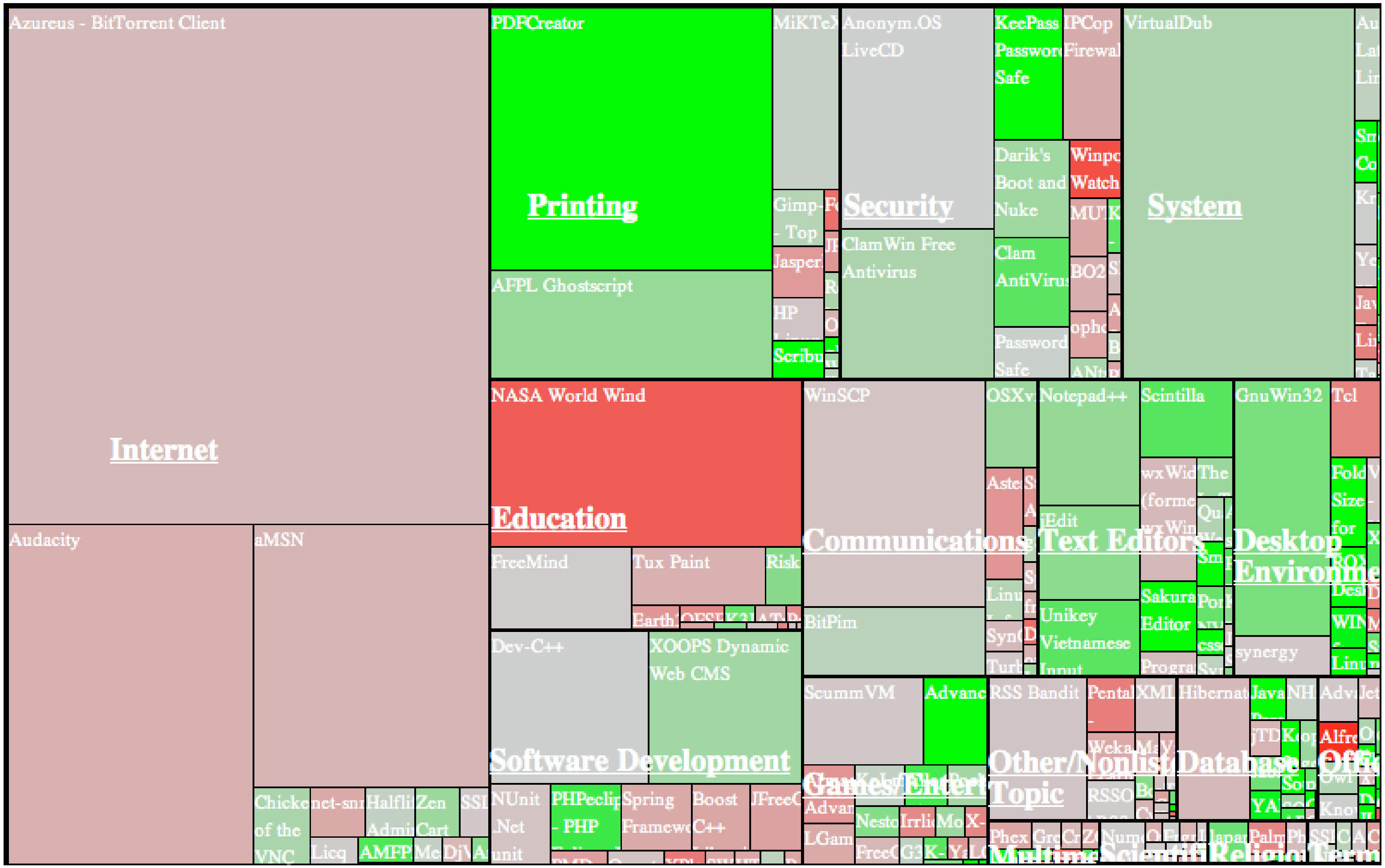
Skládání (stacking) dimenzí

- Začíná se diskretizací rozsahů v každé dimenzi. Každé dimenzi je pak přiřazena orientace a uspořádání. Dimenze se dvěma nejnižšími uspořádáními se použijí pro rozdělení virtuální obrazovky na sekce, přičemž kardinalita dimenzí určuje, kolik sekcí v horizontální a vertikální ose je generováno. Další takto vytvořená sekce je použita pro rekurzivní definici virtuální obrazovky v dalších dvou dimenzích stejným způsobem. Tento proces se opakuje, dokud nejsou zpracovány všechny dimenze a data nejsou umístěna na jejich pozici v obrazovce.

Skládání (stacking) dimenzí



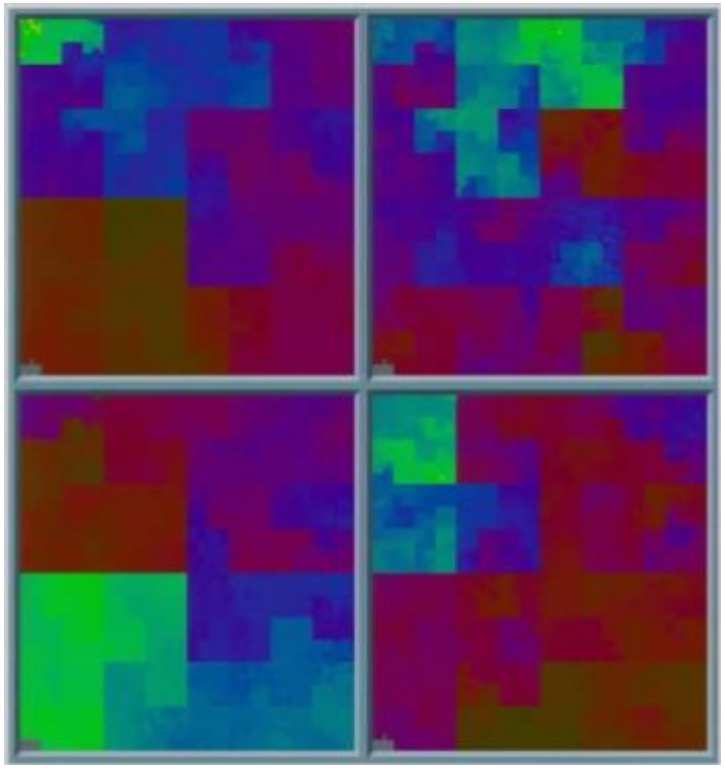
Treemap



Azureus - BitTorrent Client
Internet
Audacity
aMSN
PDFCreator
Printing
AFPL Ghostscript
Anonym.OS LiveCD
Security
ClamWin Free Antivirus
KeePass Password Safe
IPCop Firewall
Darik's Boot and Nuke
Clam AntiVirus
VirtualDub
System
NASA World Wind
Education
FreeMind
Tux Paint
WinSCP
Communications
BitPim
Notepad++
Scintilla
Text Editors
Unikey Vietnamese Input
Sakura Editor
GnuWin32
Tcl
Desktop Environment
synergy
Dev-C++
XOOPS Dynamic Web CMS
Software Development
ScummVM
Games/Entertainment
RSS Bandit
Other/Nonlisted Topic
Database

Kombinace technik

- Hybridní techniky založené na kombinaci předchozích technik pro body, čáry a plochy
- Nejznámější:
  - Glyfy (piktogramy)
  - Dense pixel displays

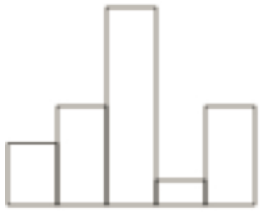


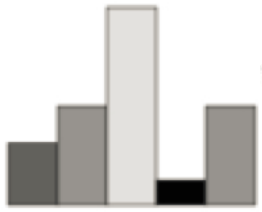
Glyfy a ikony

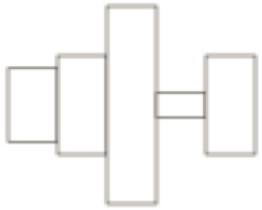
- Vizuální reprezentace části dat nebo informace, kde je grafická entita a její atributy řízena jedním nebo více atributy vstupních dat
- Grafické atributy, na které mohou být datové hodnoty mapovány:
  - pozice, velikost, tvar, orientace, materiál, styl čáry, dynamika

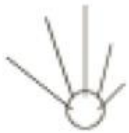
Glyfy a ikony

- Typy mapování:
  - 1:1 – každý datový atribut je mapován na jednoznačný grafický atribut
  - 1:N – sada redundantních mapování (např. na velikost i barvu zároveň)
  - M:N – několik nebo všechny datové atributy mapovány na společný typ grafického atributu





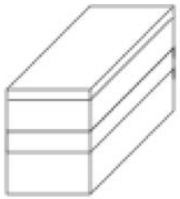


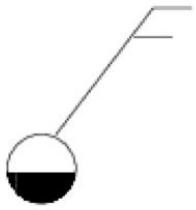


PROFILE GLYPHS

STARS AND METROGLYPHS

STICKS AND TREES





AUTOGLYPH/BOX GLYPH

FACE GLYPHS

ARROWS/WEATHERVANES

Glyfy a ikony

- Musíme si být vědomi řady nepřesností a omezení:
  - Nepřesnosti ve vnímání – závisí na typu použitých grafických atributů
  - Vzdálenost mezi grafickými atributy ovlivňuje přesnost jejich porovnání – bližší porovnávány přesněji
  - Počet dimenzí dat a záznamů, které je možné efektivně zobrazit pomocí glyfů, je omezen

Glyfy a ikony

- Po vybrání typu glyfu existuje N! různých uspořádání dimenzí, které mohou být při mapování použity
- Existuje několik strategií pro volbu vhodného uspořádání:
  - Třídění dimenzí na základě jejich korelace
  - Zvýšení vlivu glyfů se symetrickým tvarem
  - Třídění podle hodnot dimenzí v jednom záznamu
  - Ruční třídění na základě znalostí domény

Rozmístění glyfů

- Tři základní typy strategií pro rozmístění glyfů na obrazovce:

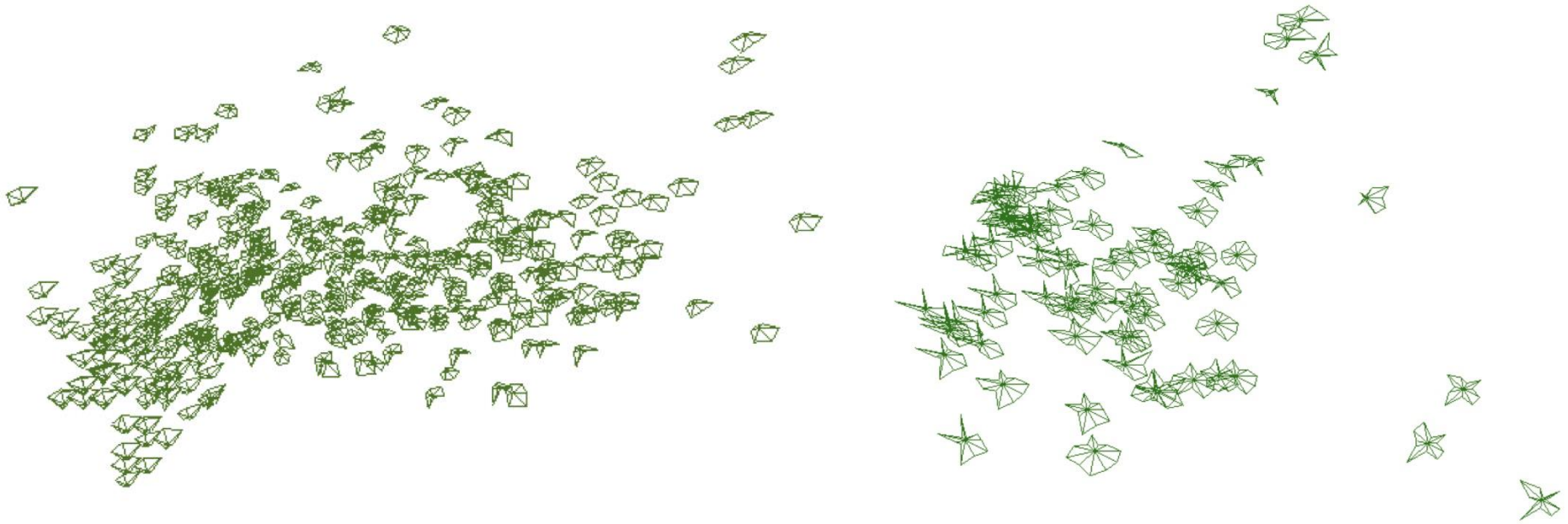
1. Uniformní
2. Řízené daty
3. Řízené strukturou

Uniformní rozmístění

- Rovnoměrné umístění na obrazovce
- Eliminace překryvů, efektivní využití obrazovky

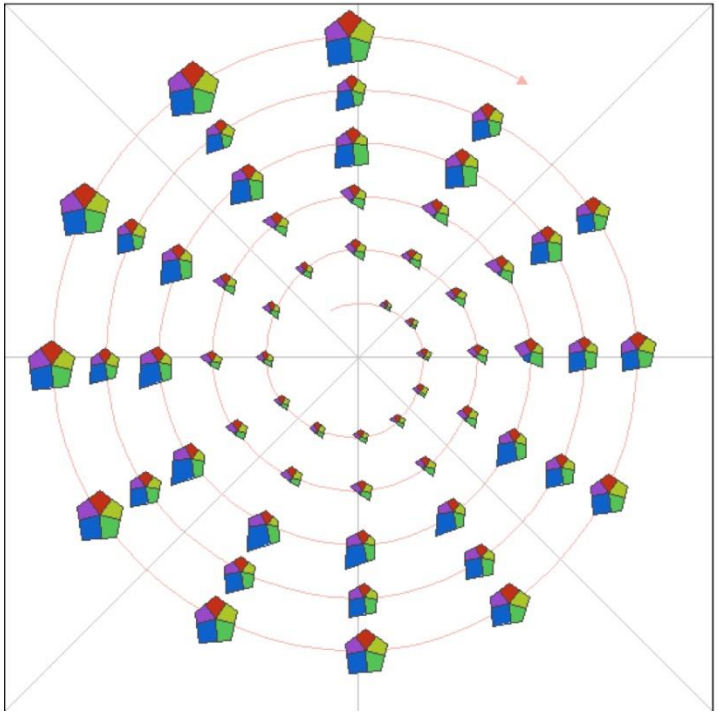
Rozmístění řízené daty

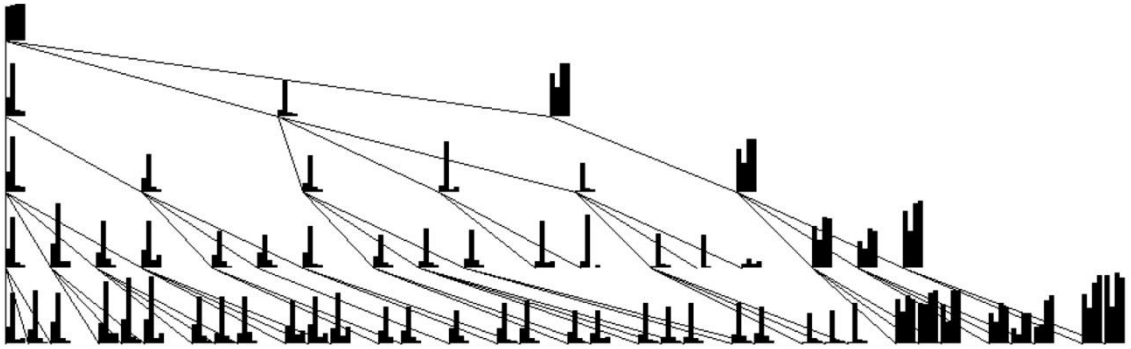
- Dva přístupy:
  - Vybrány dvě dimenze pro řízení rozmístění (vlevo)
  - Pozice odvozeny pomocí PCA, MDS (vpravo)



Rozmístění řízené strukturou

- Využití struktury dat – cyklická, hierarchická





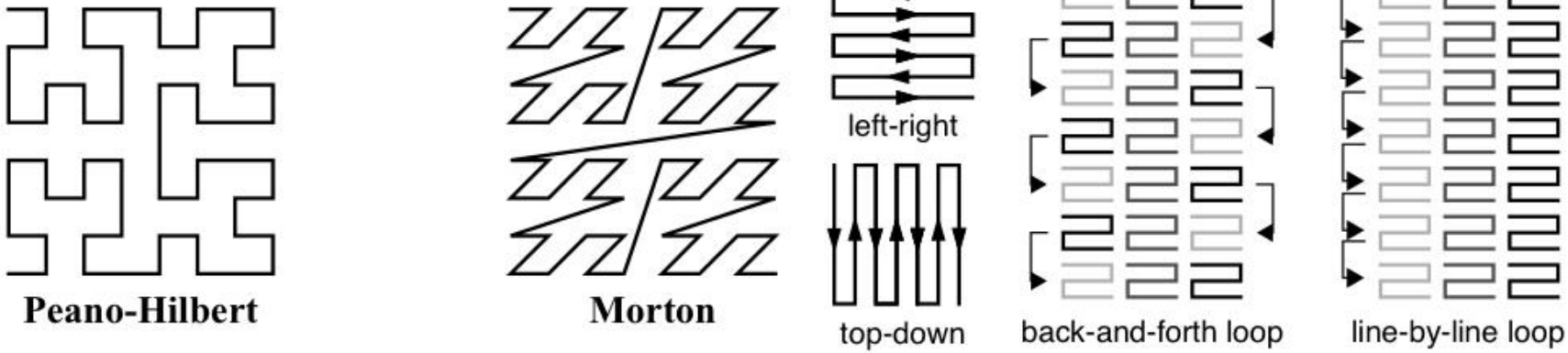
Dense Pixel Displays

- Hybridní metoda na pomezí bodových a regionálních (plošných) metod
- Mapuje každou hodnotu na jednotlivé pixely a pro každou dimenzi vytváří vyplněný polygon
- Zobrazení milionů hodnot na jediné obrazovce
- Počet bodů určuje počet jednotlivých položek v obrázku
- Technika spoléhá na využití barvy

Dense Pixel Displays

- Nejjednodušší forma:
  - Každá dimenze datové množiny generuje samostatný oddělený „podobrázek“ na obrazovce
  - Každou dimenzi můžeme považovat za nezávislou sadu čísel, každá sada řídí barvu odpovídajících pixelů
  - Rozmístění prvků v sadách (zdůraznění vztahů mezi blízkými body): střídáme průchod zleva doprava a zprava doleva; pokud dosáhneme okraje podobrázku, posuneme se o řadu níže

Dense Pixel Displays

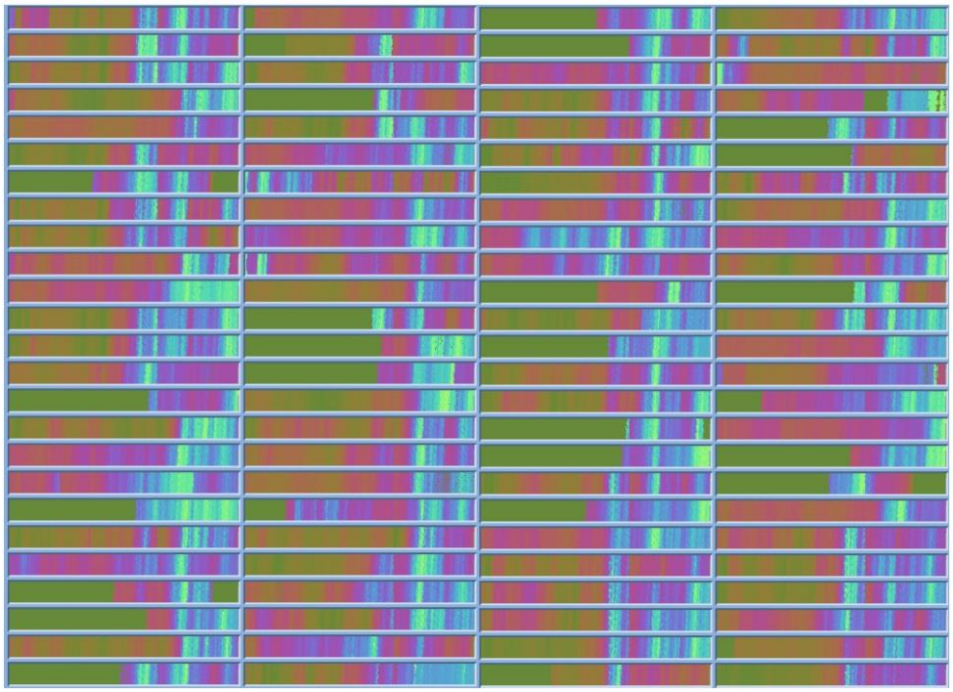


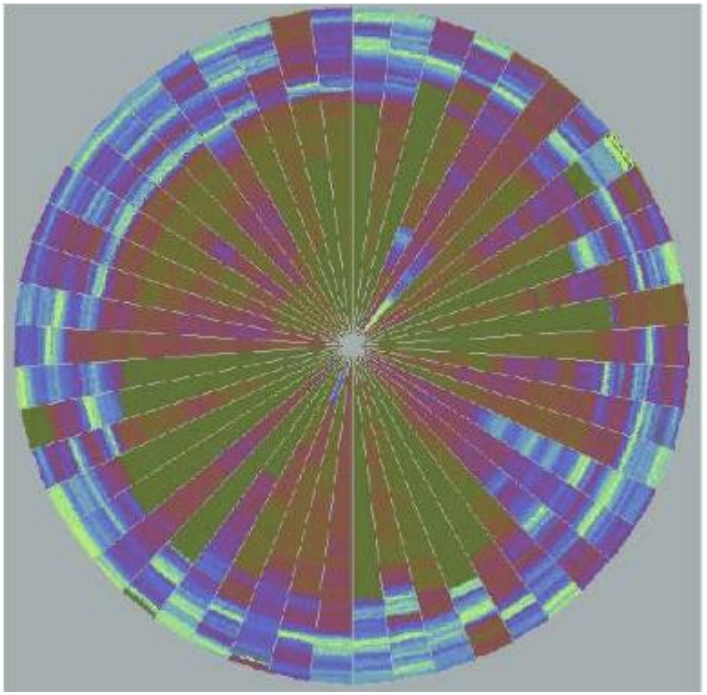
vyplnění obrazovky

rekurzivní vzory

Rekurzivní vzory, kruhové segmenty

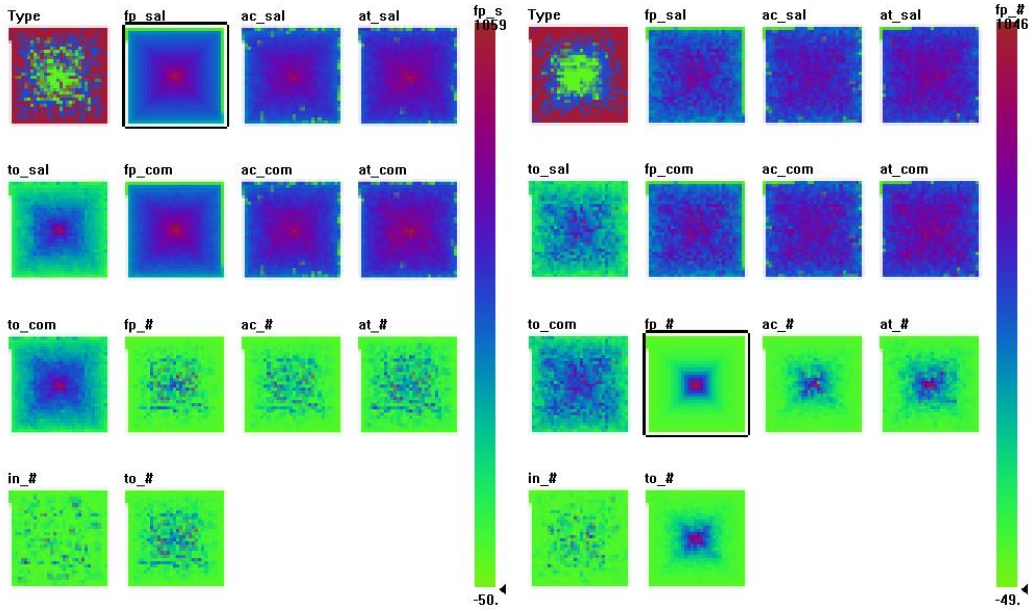
- Umístění podobrázků různými způsoby:





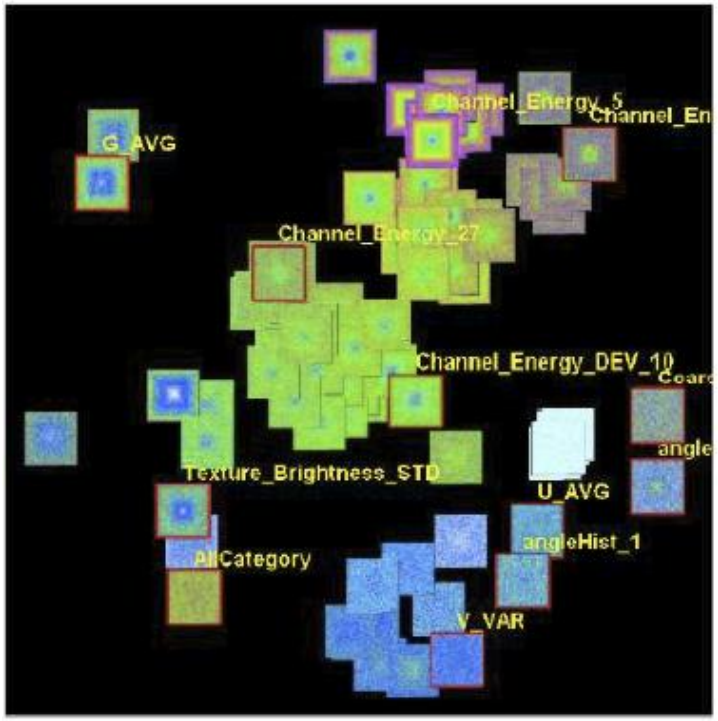
Dense Pixel Displays

- Poslední důležitou oblastí je uspořádání dat
- U časových sekvencí je uspořádání pevně dáno
- U ostatních může změna uspořádání záznamů odhalit zajímavé vlastnosti



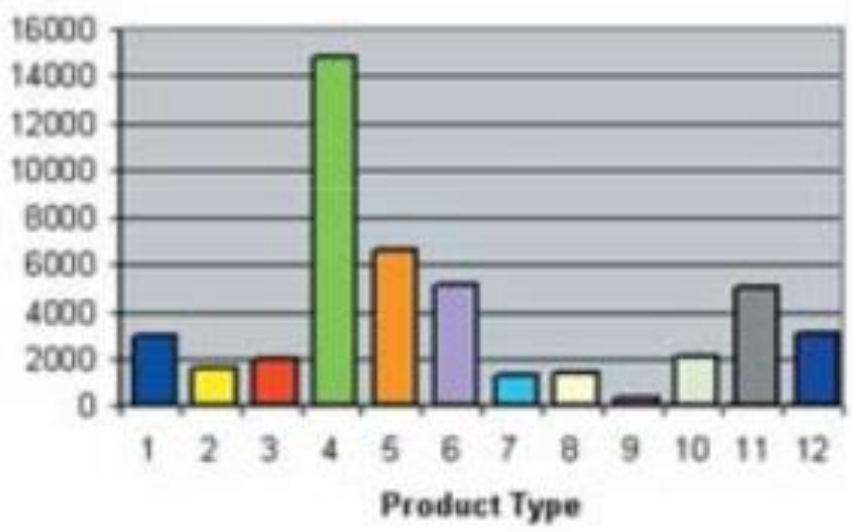
Další přístupy

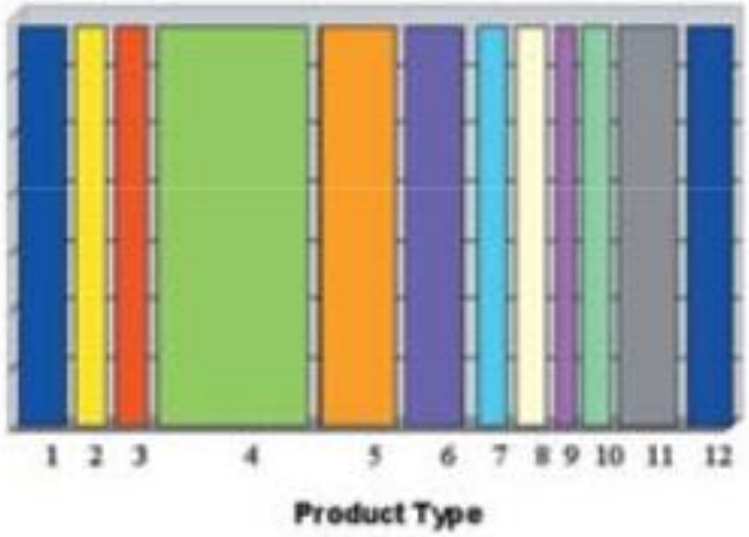
- Umožňují překrytí podobrázků:
  - „Value and Relation“ technika využívající multidimensional scaling



Pixelové sloupcové diagramy

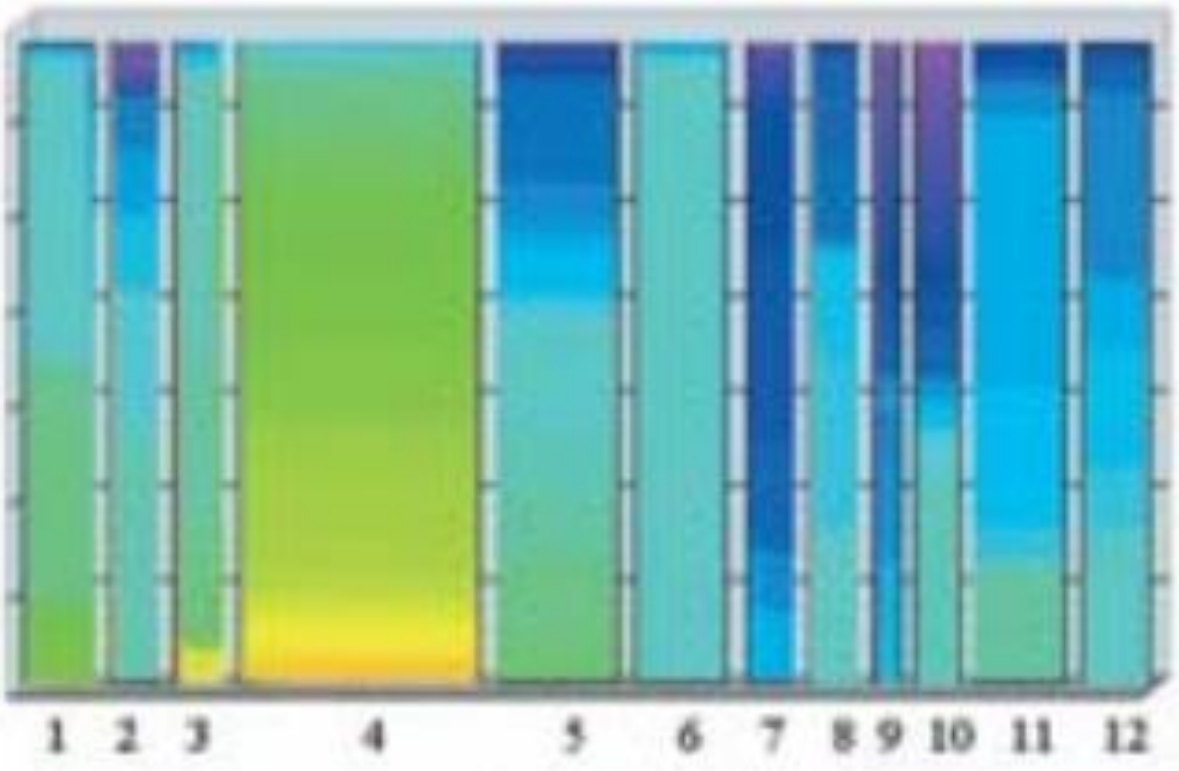
- Přetížení klasického sloupcového diagramu – zahrnutí více informací o jednotlivých prvcích





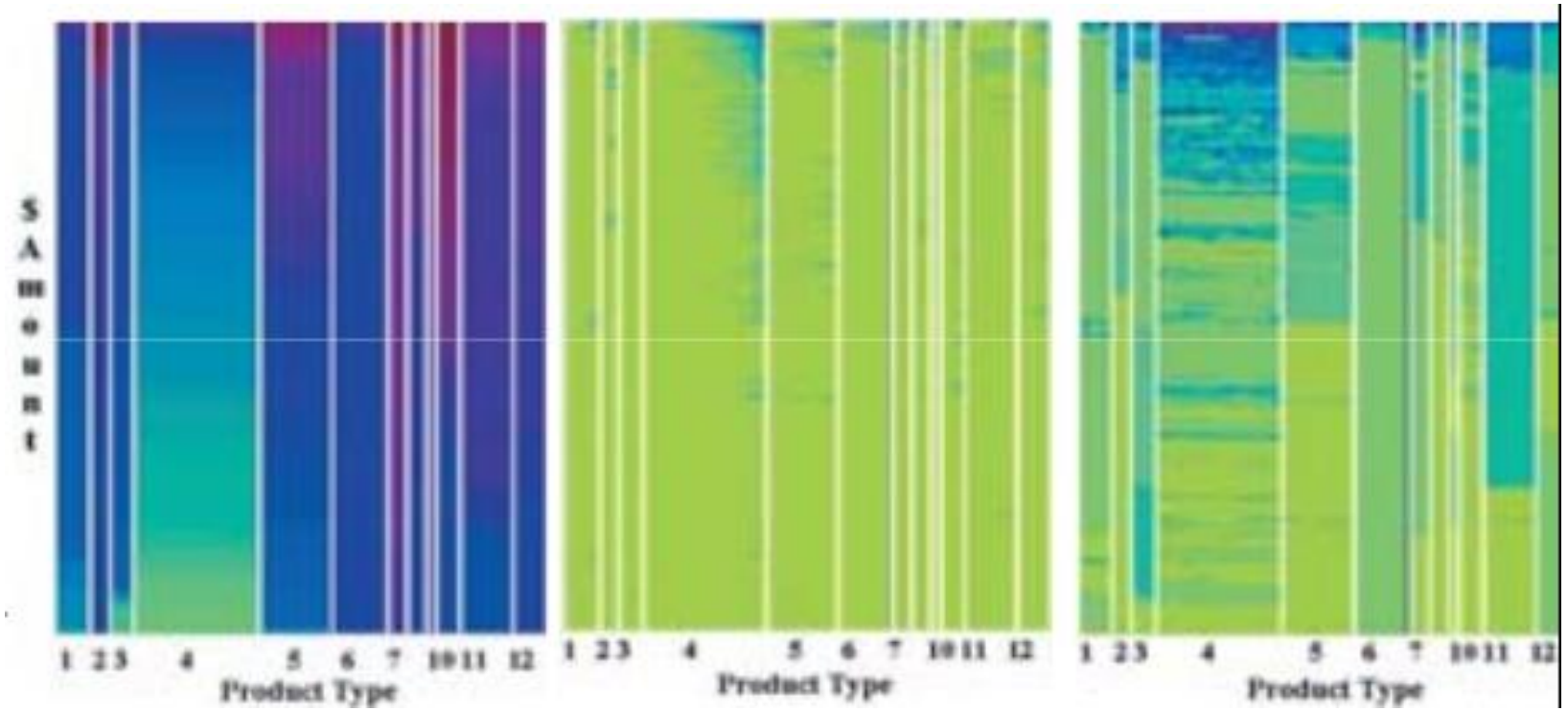
Pixelové sloupcové diagramy

- Každý pixel sloupce odpovídá datovému bodu patřícího do skupiny reprezentované tímto sloupcem



Pixelové sloupcové diagramy

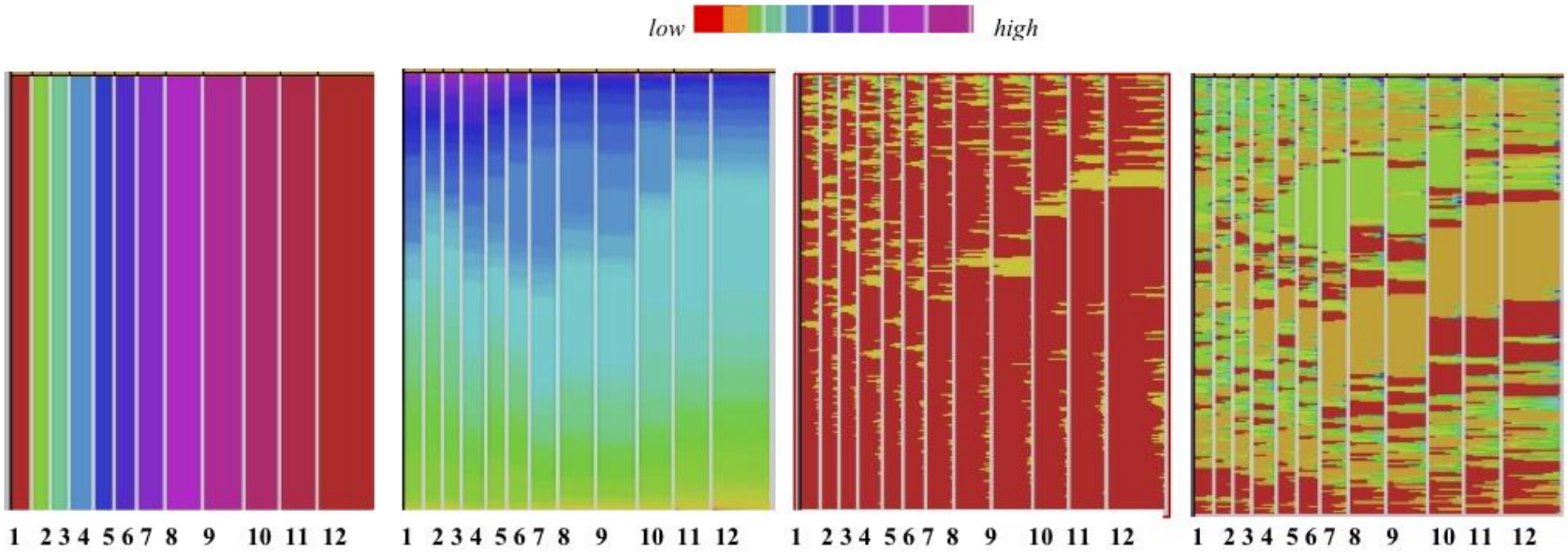
- Internetový nákup - vztah typu produktu vůči ceně. Barva je mapována na:



utracená částka počet návštěv velikost prodejů

Pixelové sloupcové diagramy

- Umístění dense pixels do sloupcového diagramu



a) Color: Month b) Color: Purchase Amount c) Color: # of Visits d) Color: Quantity

Pixelové sloupcové diagramy

- Dá se odvodit např.:
  - V prosinci byl největší počet zákazníků, zatímco v únoru, březnu a květnu jich bylo nejméně
  - Od února do května byl největší počet nákupů
  - Počet nákupů v prosinci je průměrný
  - Od března do června se zákazníci vraceli častěji než v jiných měsících. Prosincoví zákazníci byli většinou jednorázoví.
  - Zákazníci kupující nejvíce se vracejí častěji a kupují více věcí