Vizualizace dat

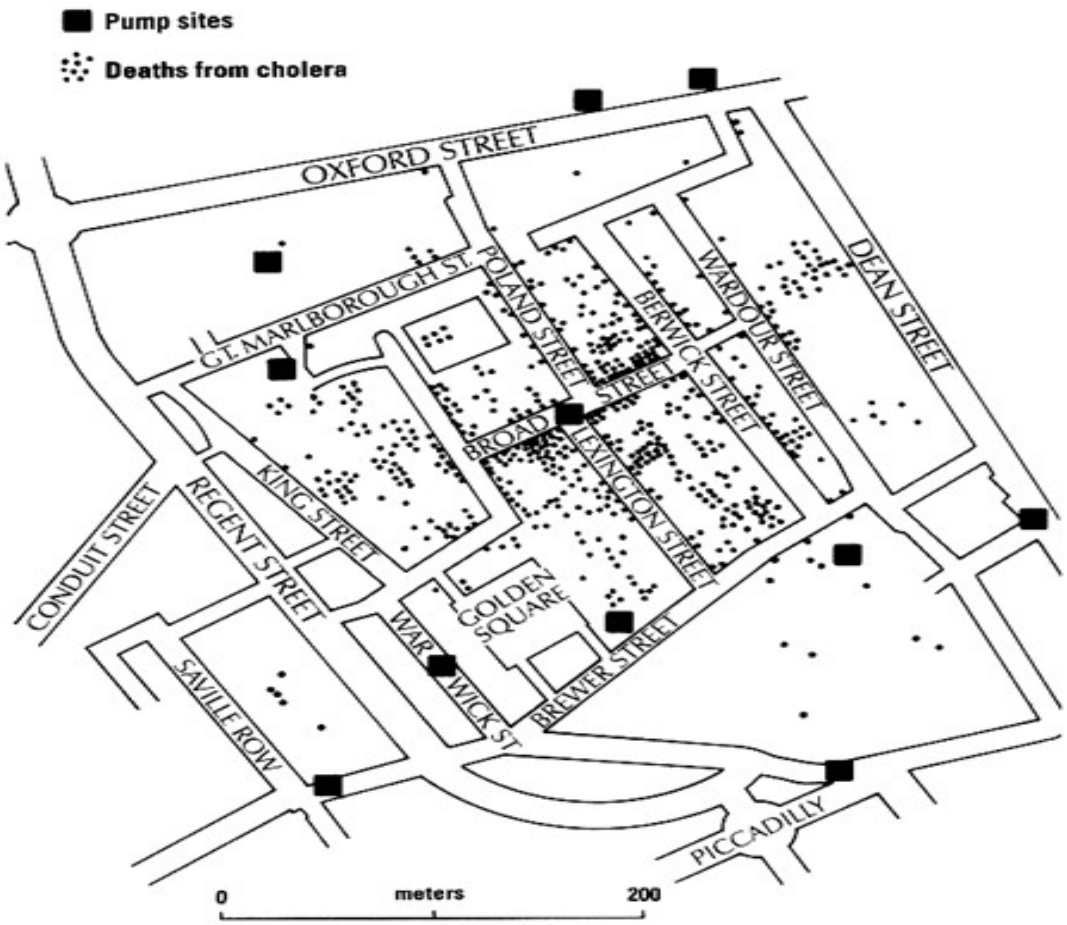
Lenka Vysloužilová

Olga Štěpánková

Nature Inspired
Technologies Group

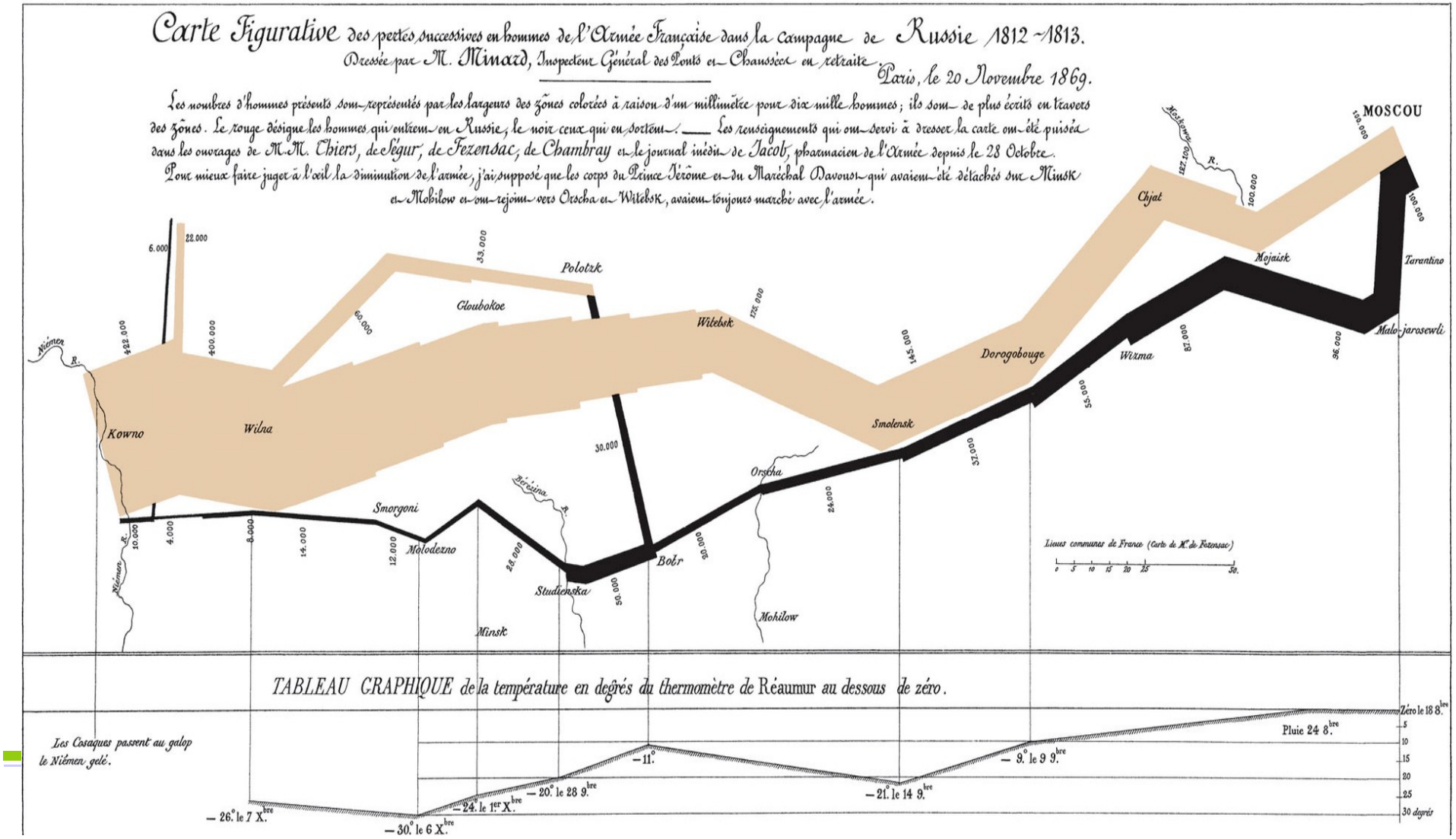
Historie

- 1855 – Dr. John Snow – dot map of cholera

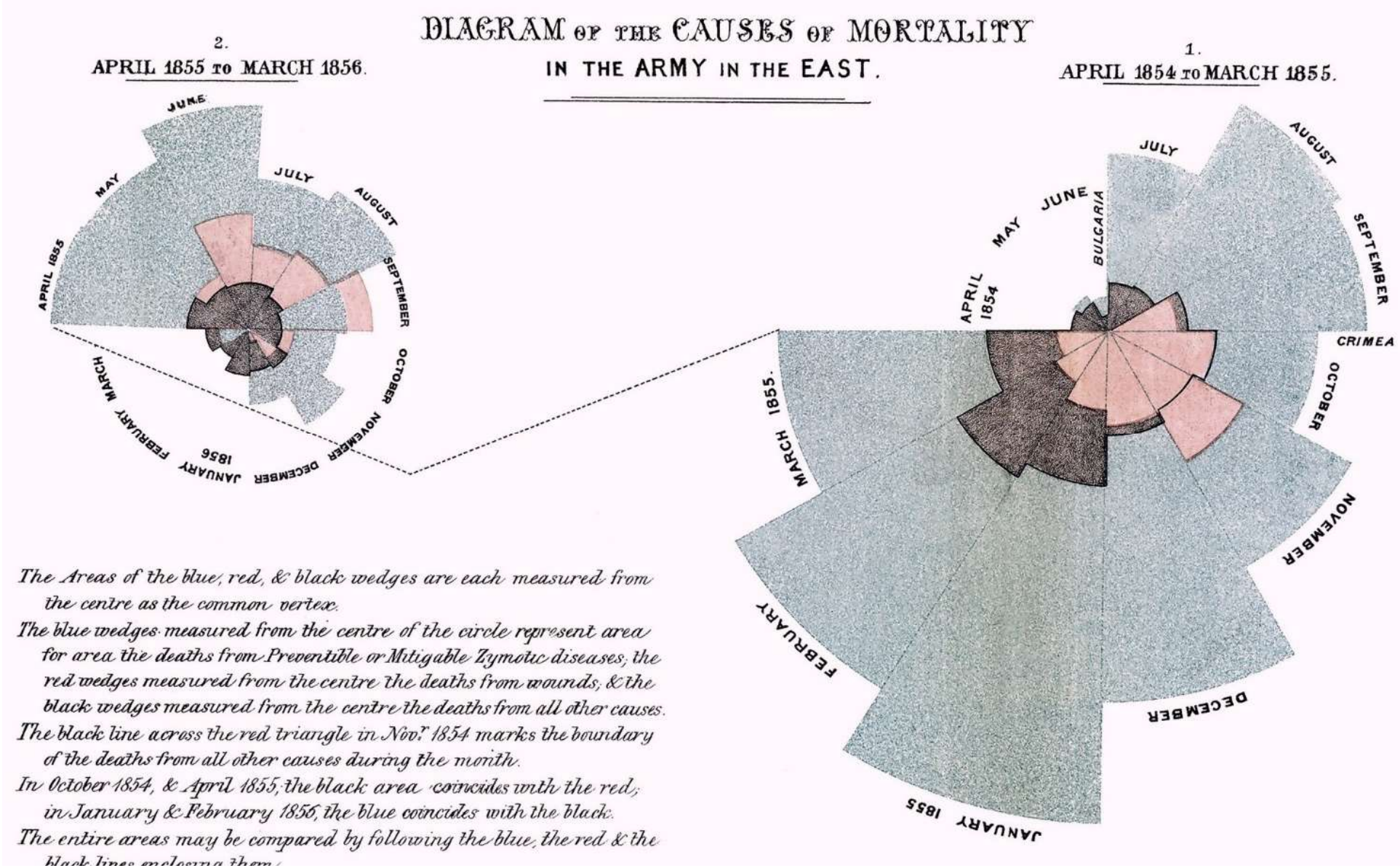


https://www.tableau.com/learn/articles/best-beautiful-data-visualization-examples

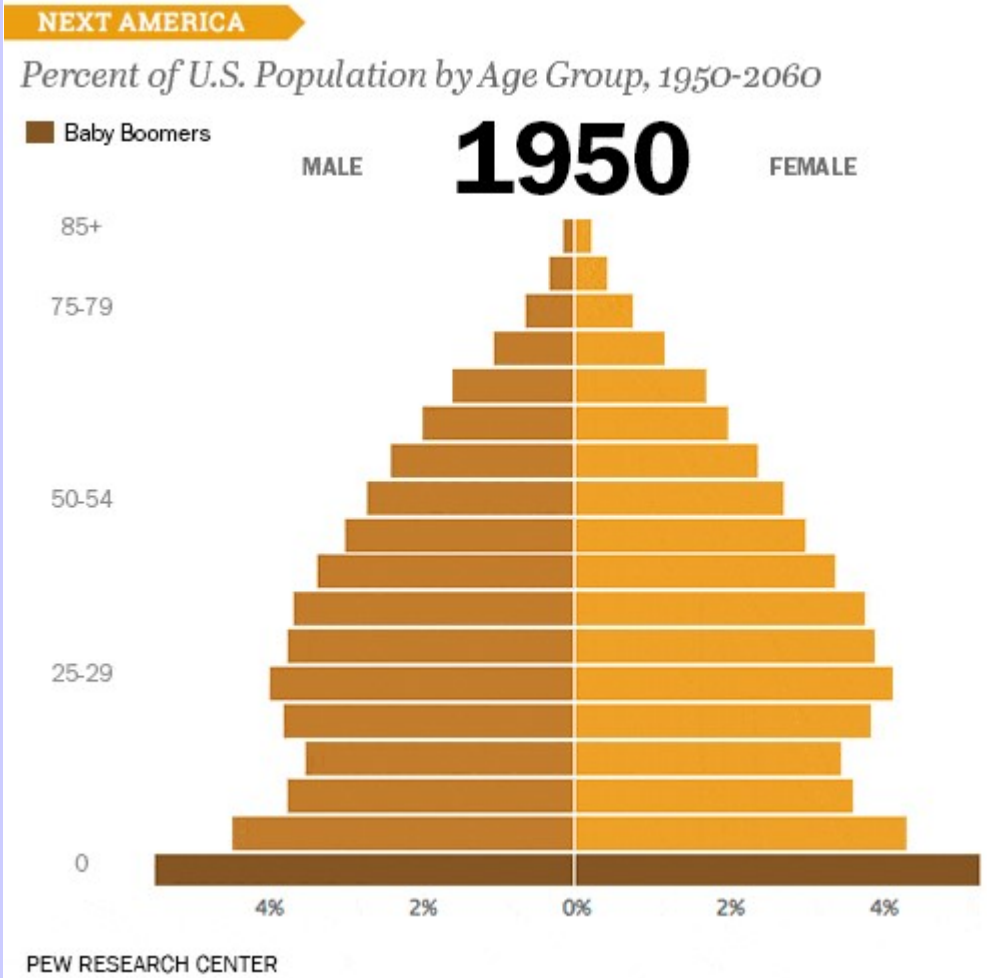
Napoleonův pochod na Moskvu



Florence Nightingale - **Krymská válka**



USA za 100 let (The Next US)

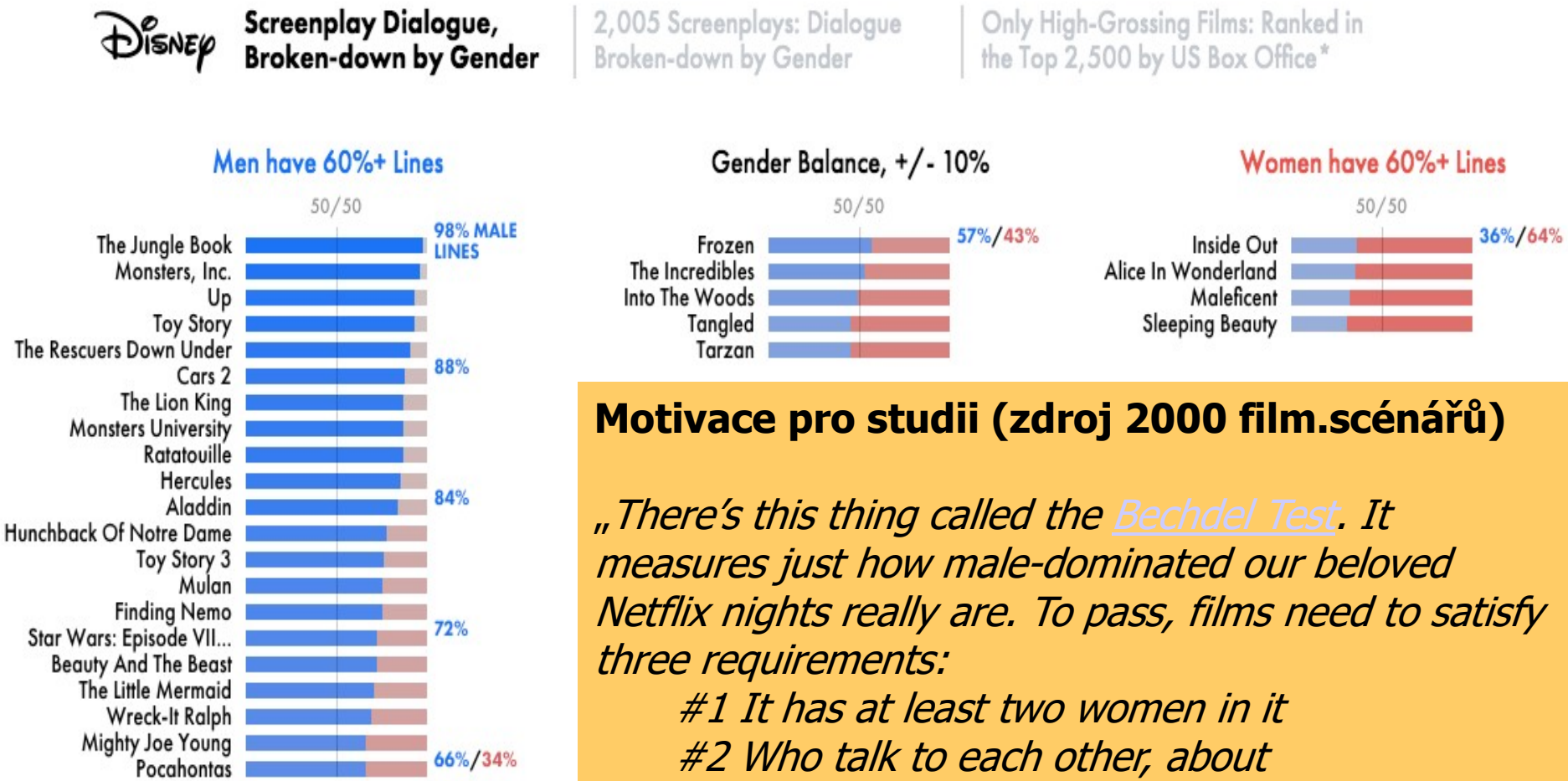


2060

?

Filmové dialogy a gender

https://pudding.cool/2017/03/film-dialogue/



Motivace pro studii (zdroj 2000 film.scénářů)

„There's this thing called the Bechdel Test. It measures just how male-dominated our beloved Netflix nights really are. To pass, films need to satisfy three requirements:

#1 It has at least two women in it
#2 Who talk to each other, about
#3 Something besides a man."

Osnova

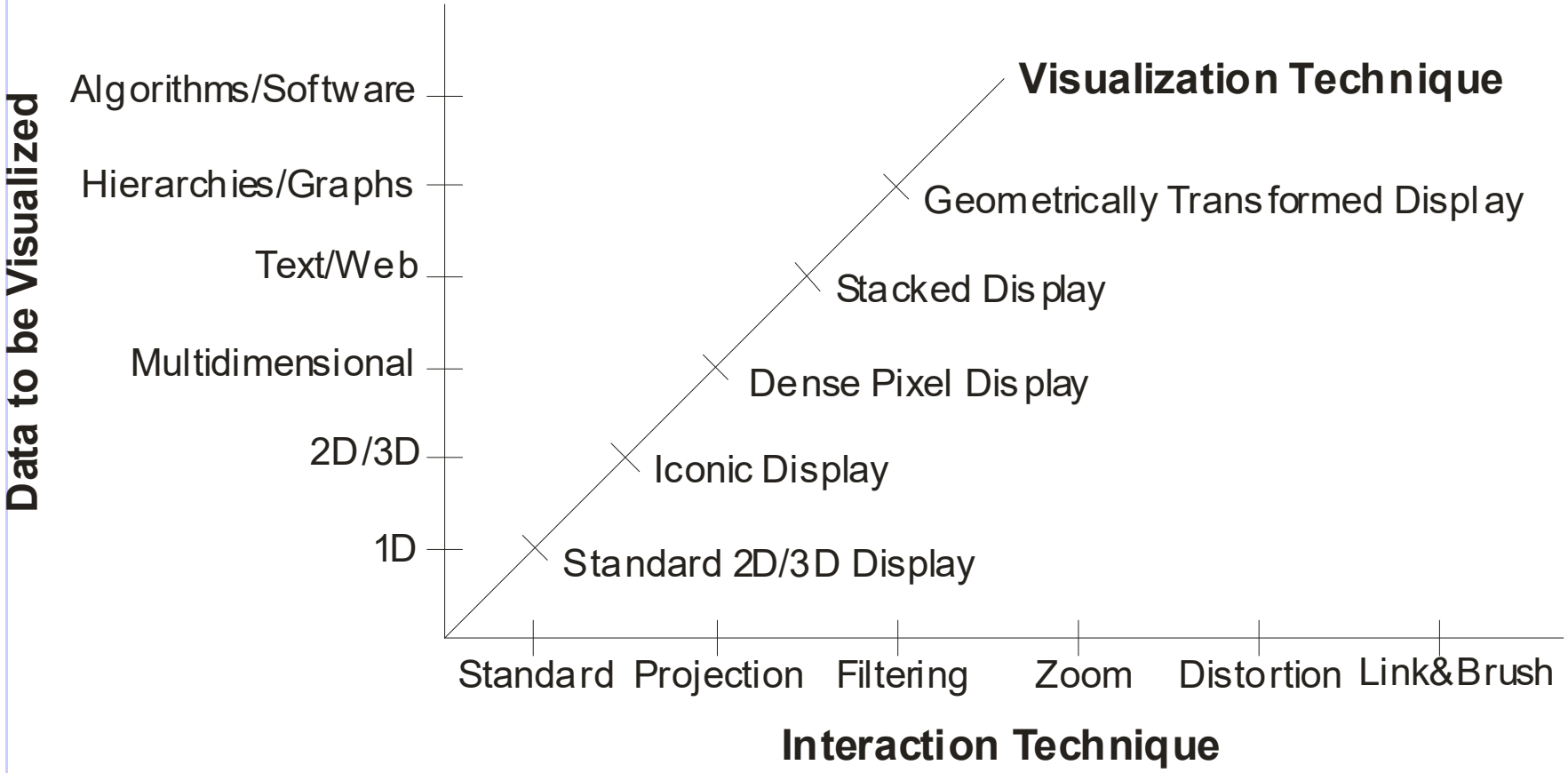
- Role vizualizace
- Klasické grafy, mapy
- Statistika – box graf, histogram
- Multidimensionální vizualizace dat

Role vizualizace

- Pomoc při prezentaci výsledků
- Pomoc při průzkumu velkého množství dat
- Možná iterakce s uživatelem

- Potřeba lidského oka a jeho zkušeností
- Pozor: vizualizace může být zavádějící

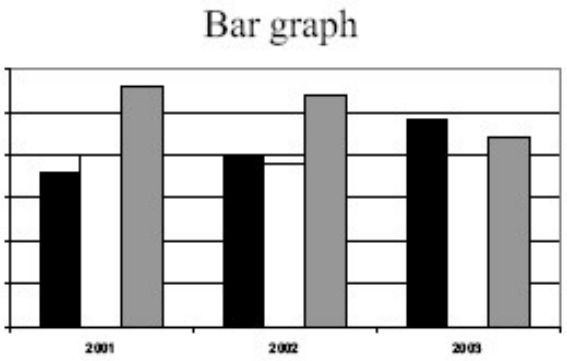
Information visualization

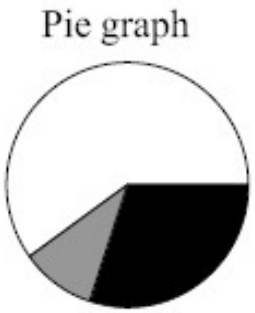


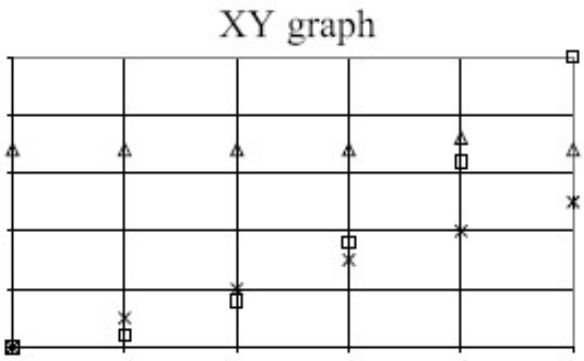
D.A. Keim. Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics, 8(1):1--8, 2002.

2D

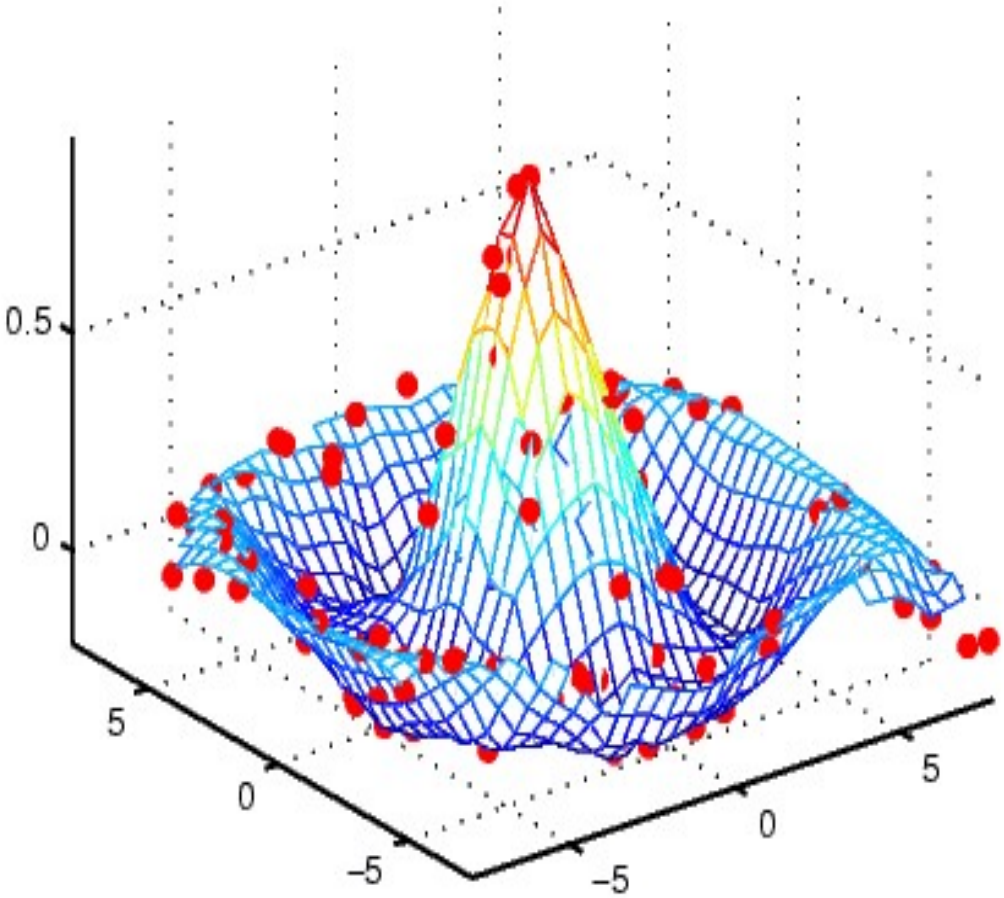
- sloupcové grafy
- koláčové grafy
- X-Y grafy



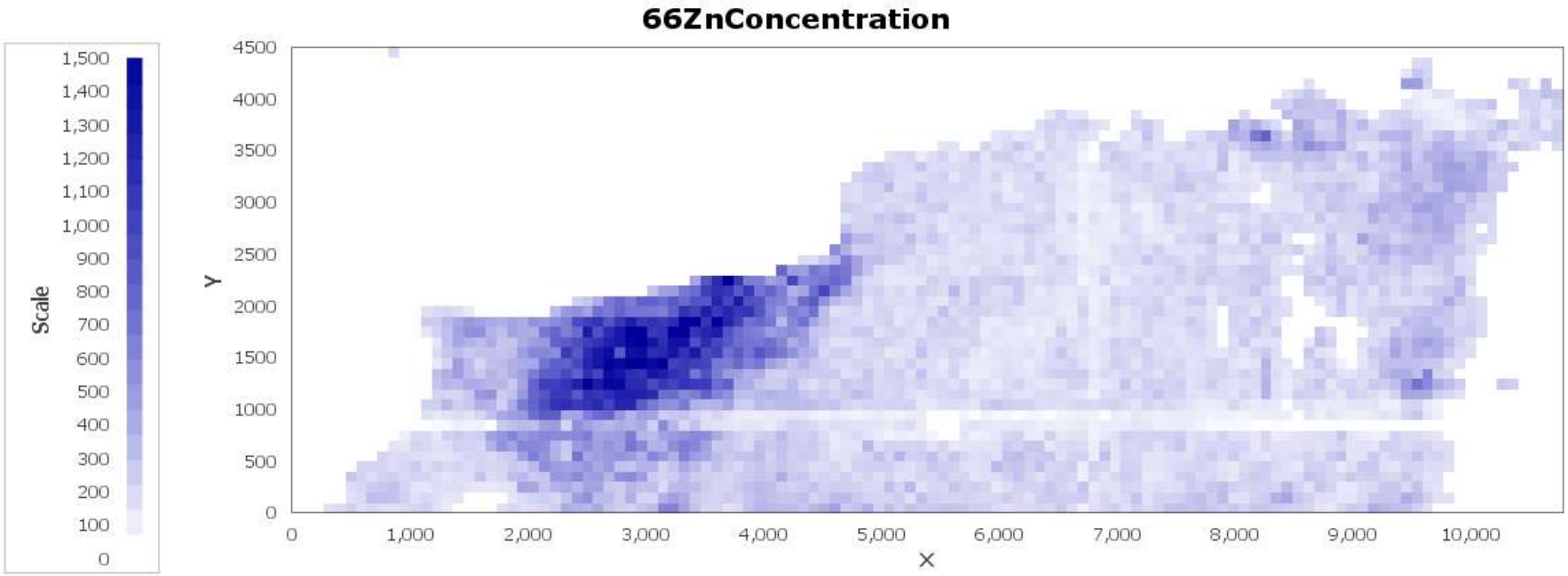




3D



Mapy


66ZnConcentration
Scale
1,500
1,400
1,300
1,200
1,100
1,000
900
800
700
600
500
400
300
200
100
0
Y
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
X

Iris Data

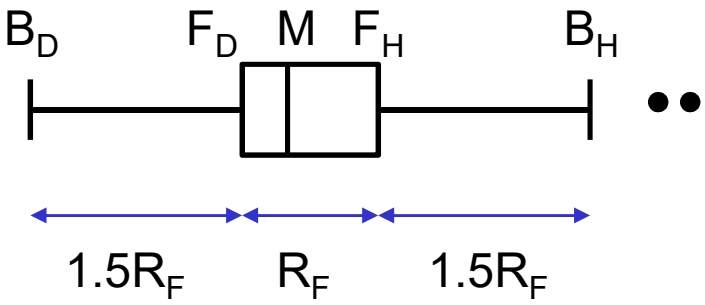
Iris setosa

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
…	…	…	…
5.9	3	5.1	1.8

Iris versicolor

Iris virginica

Box graf – Quartile Color



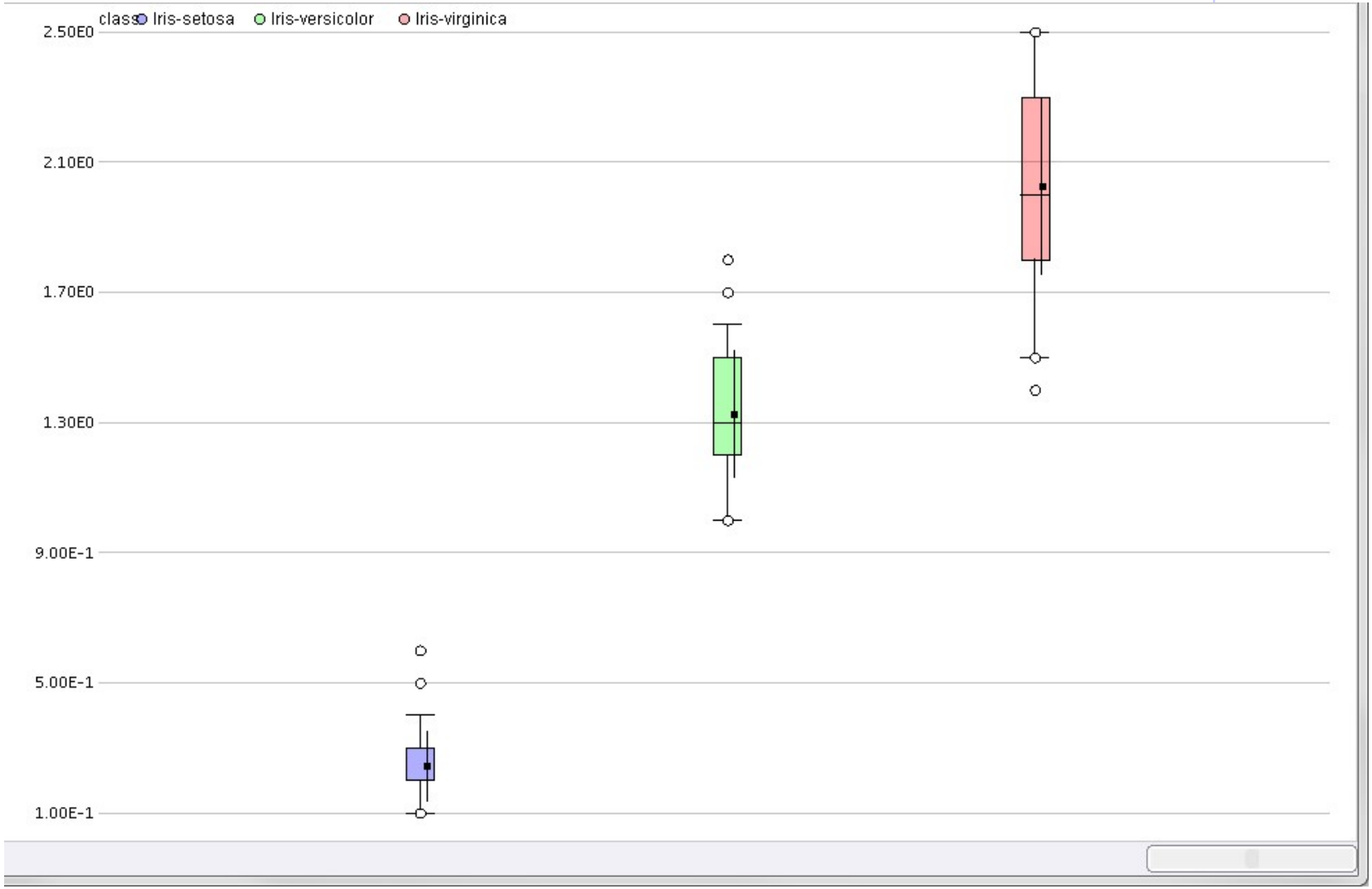
M – medián

F – horní a dolní kvartil

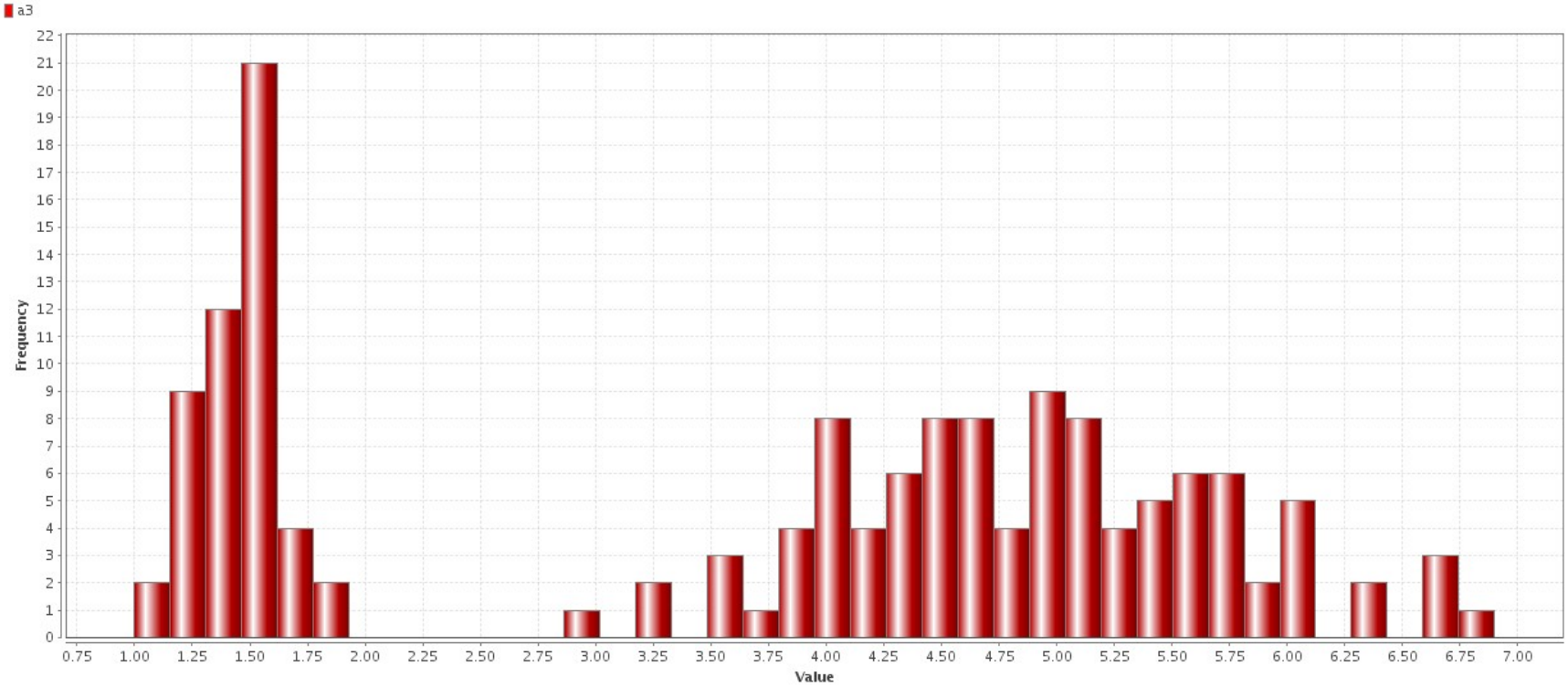
$$R_F = F_H - F_D$$

$$B_D = F_D - 1.5 R_F$$

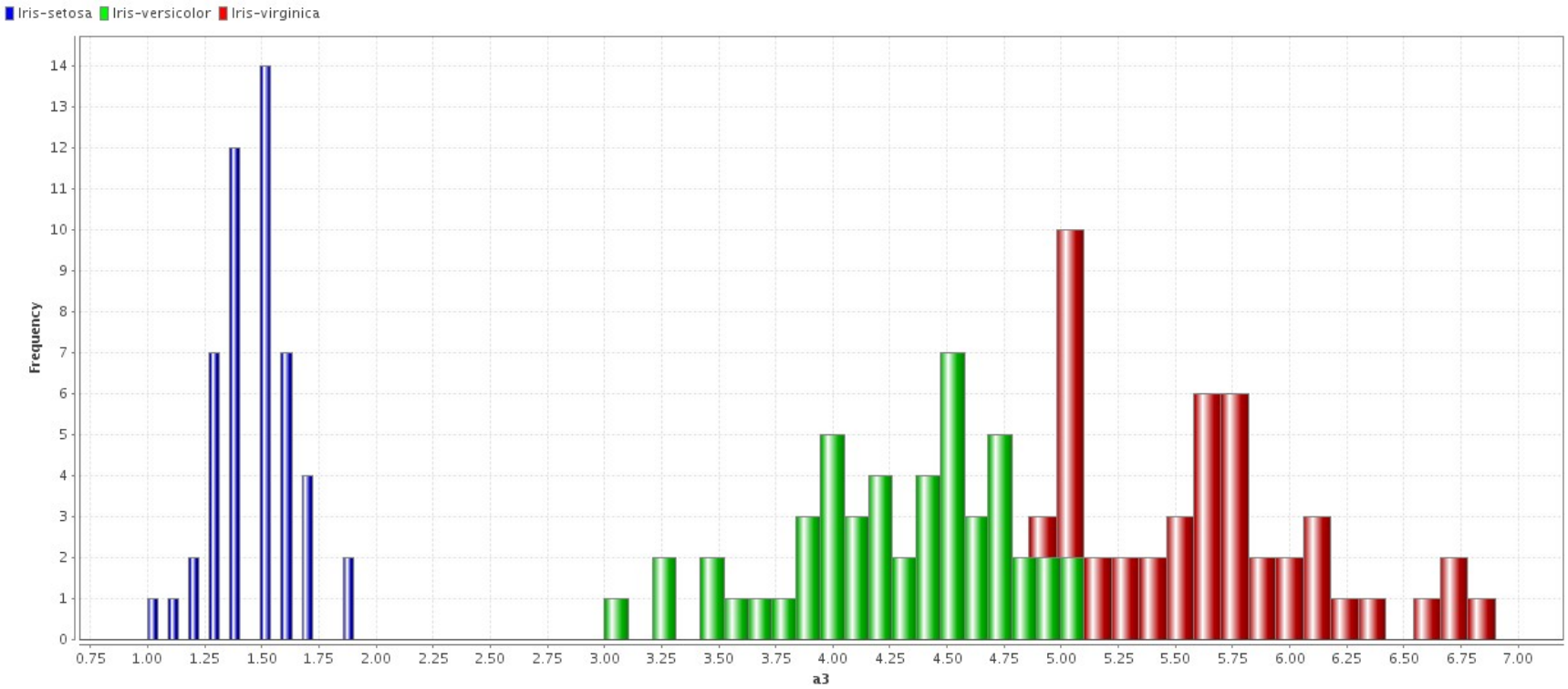
Box graf – Quartile Color



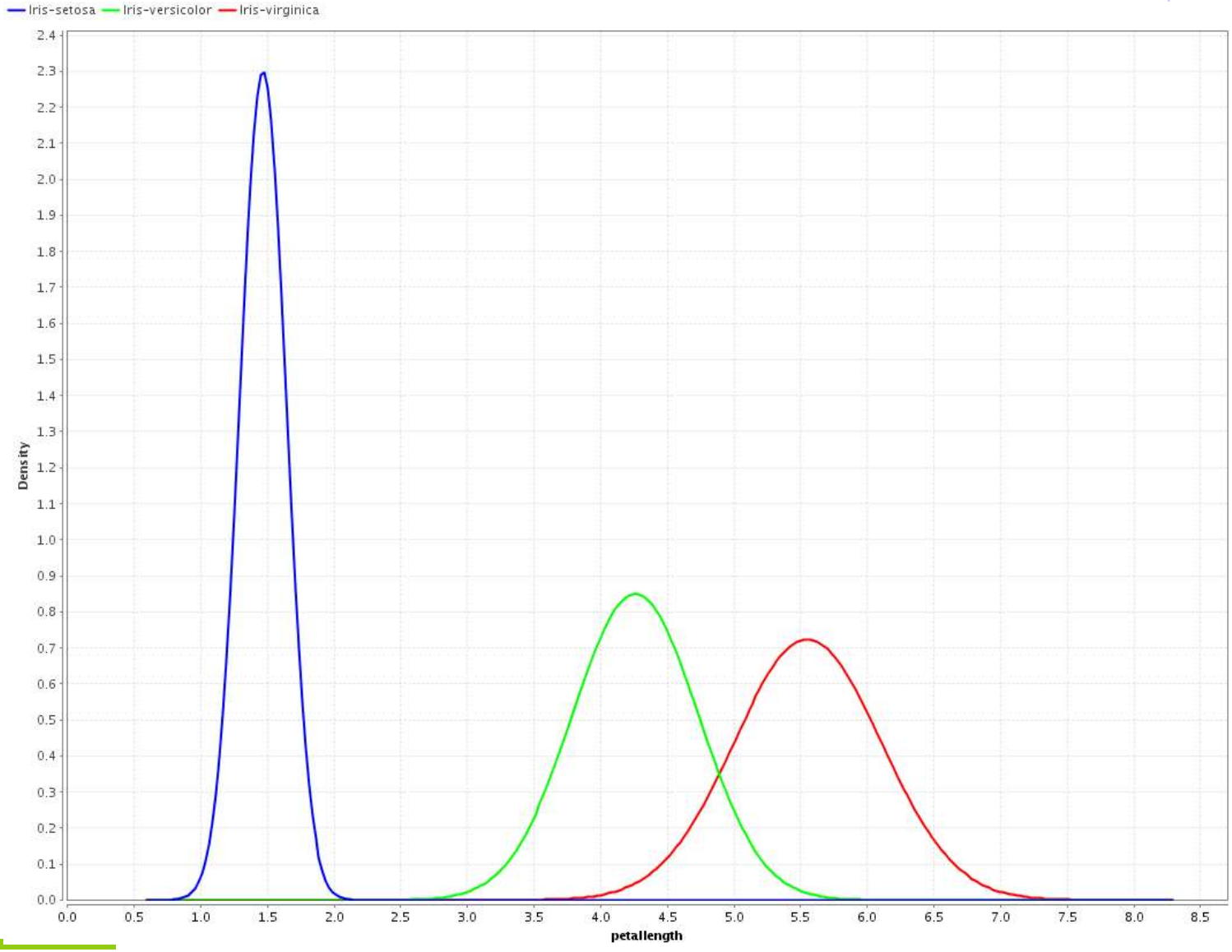
Histogram


a3
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
Value

Histogram

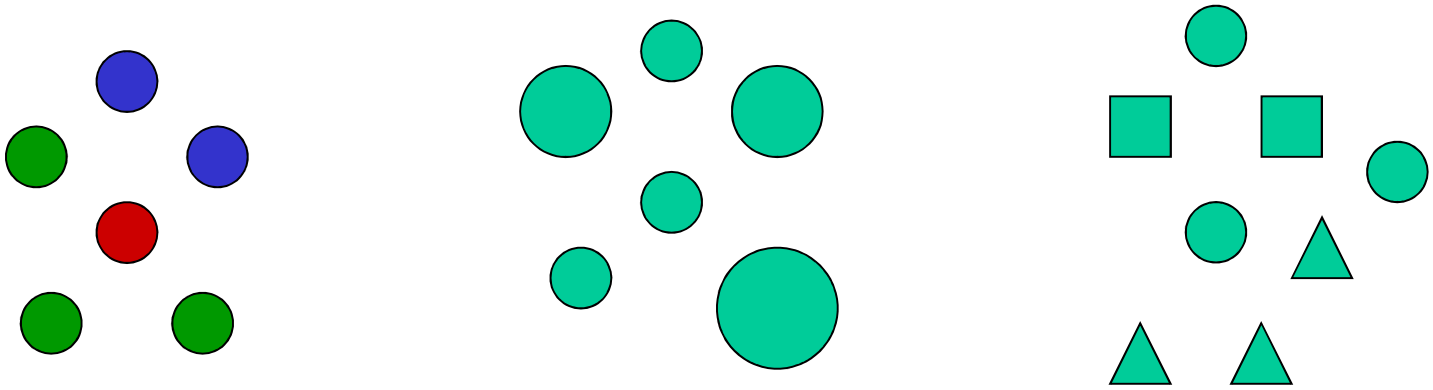

Iris-setosa
Iris-versicolor
Iris-virginica
Frequency
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
3.25
3.50
3.75
4.00
4.25
4.50
4.75
5.00
5.25
5.50
5.75
6.00
6.25
6.50
6.75
7.00
a3

Distribution

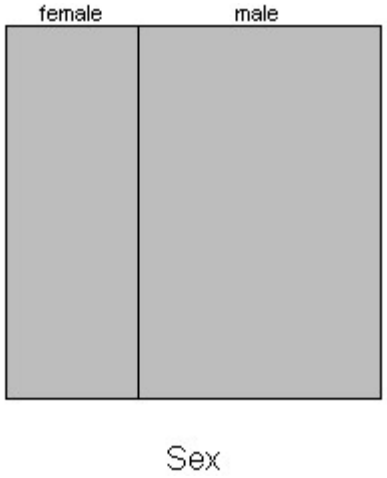


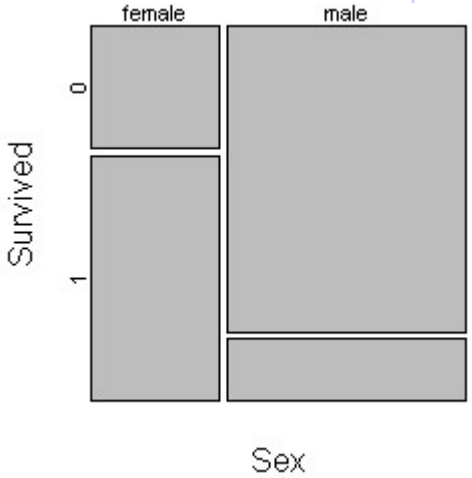
Multidimensionální vizualizace dat

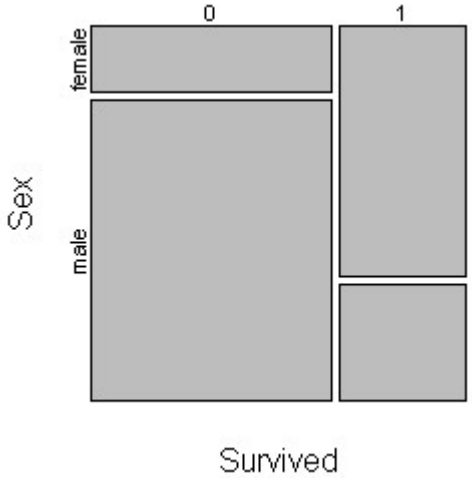
- člověk je omezen světem kolem sebe na 3D vnímání
- nD -> 2D (3D)
- další dimenze nahrazujeme
  - přidáním barev, velikostí objektů a tvarem objektů
  - transformací dat
  - projekcí dat



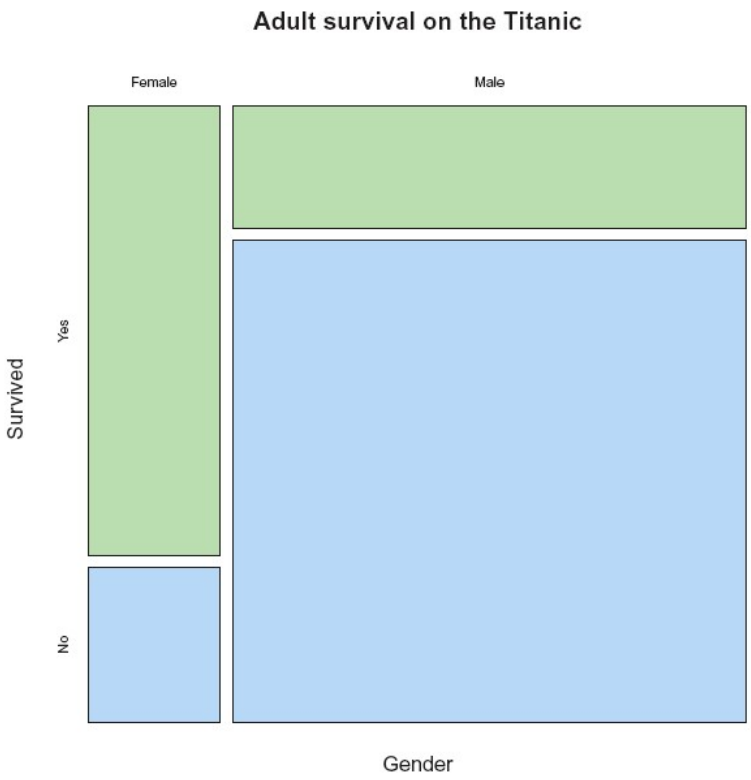
Mosaic Plots

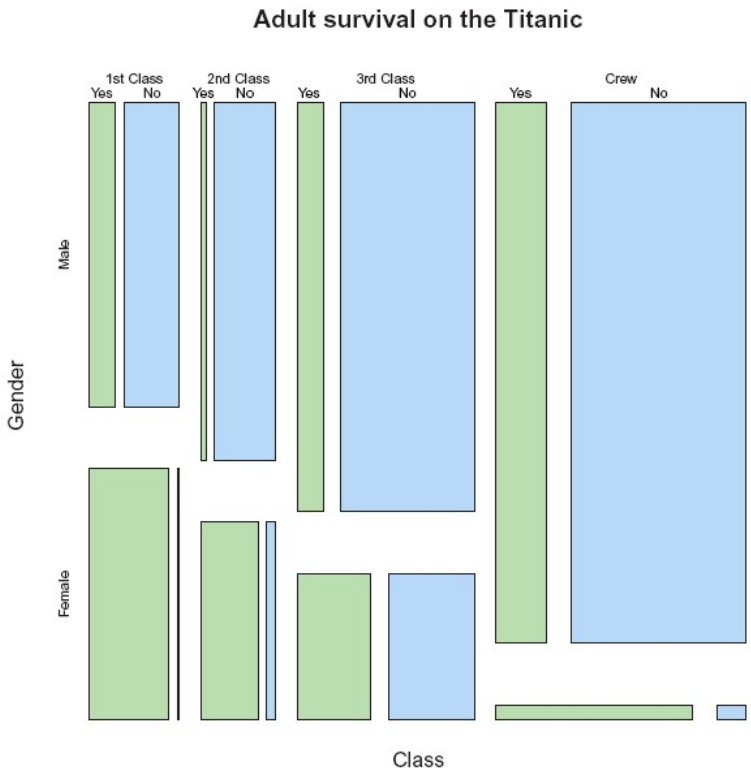




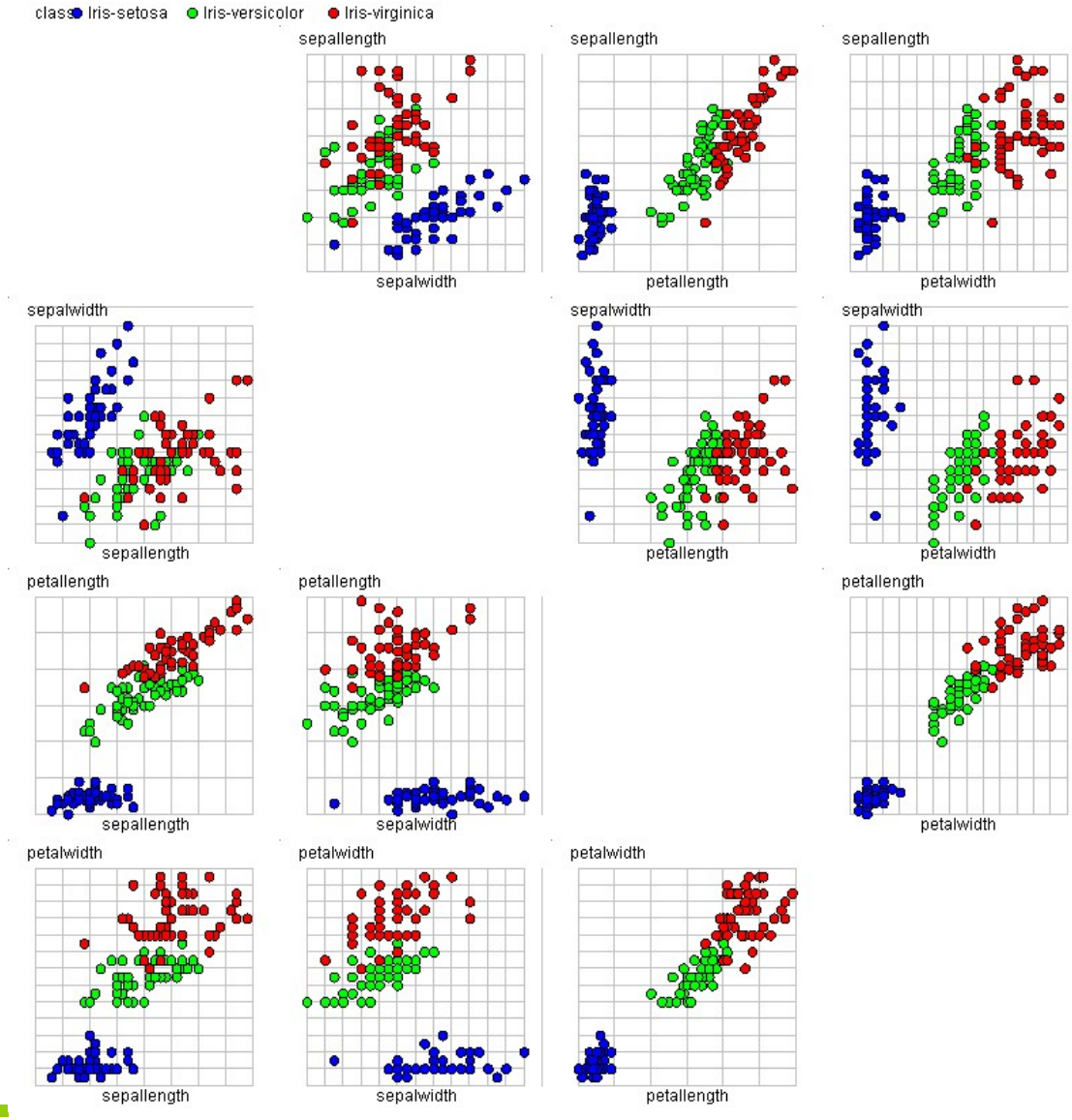


Mosaic Plots

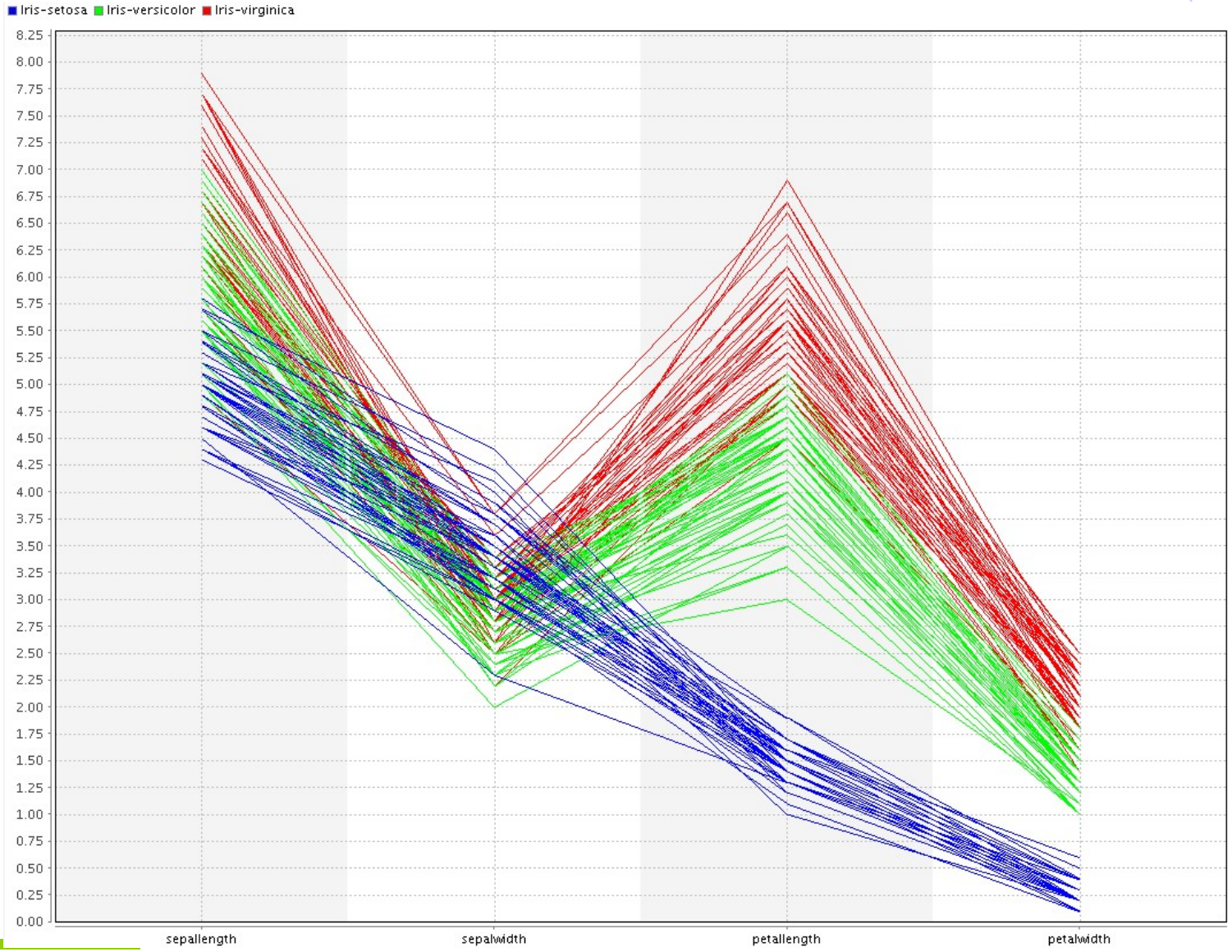




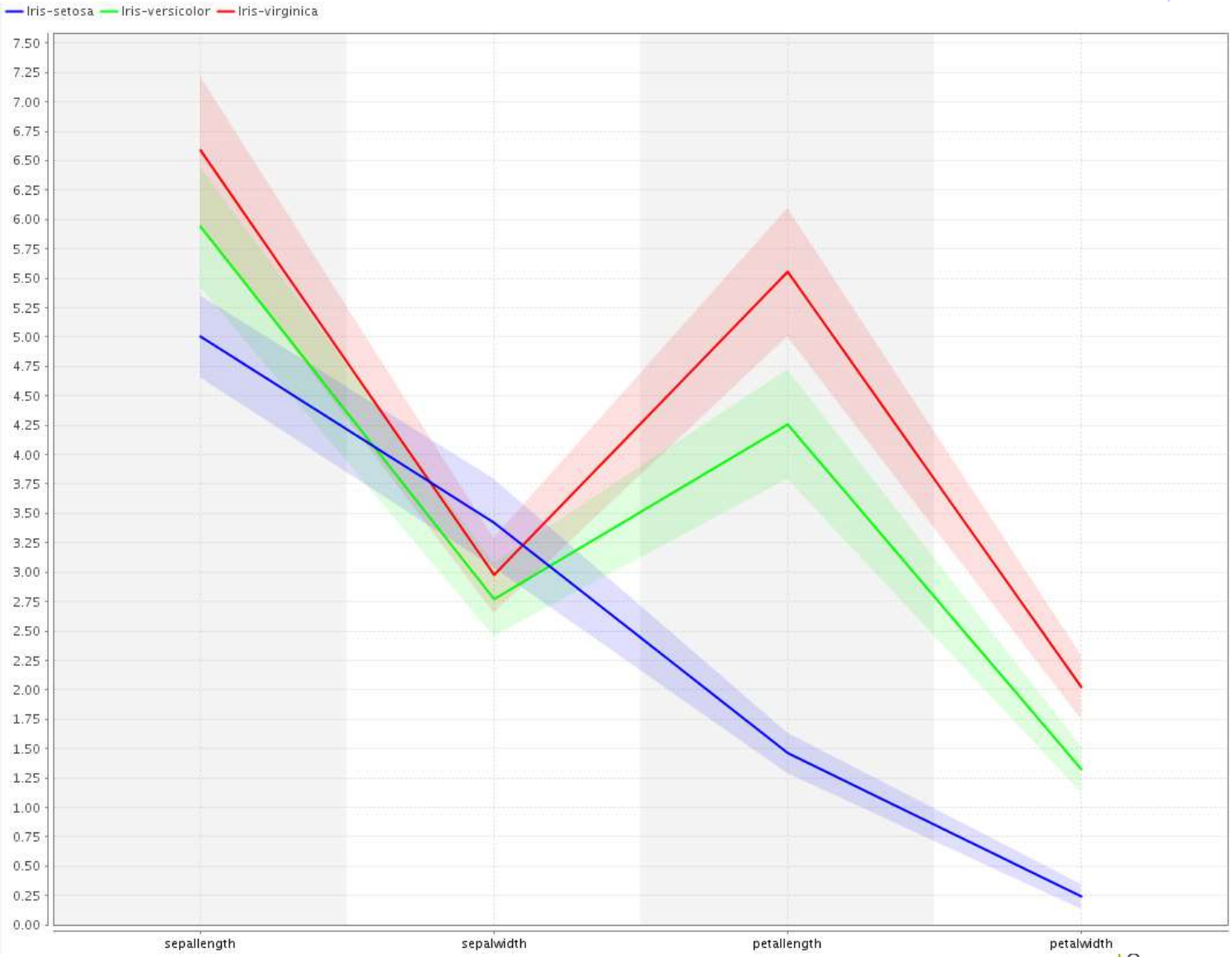
Scatter Plot matrix



Parallel coordinates



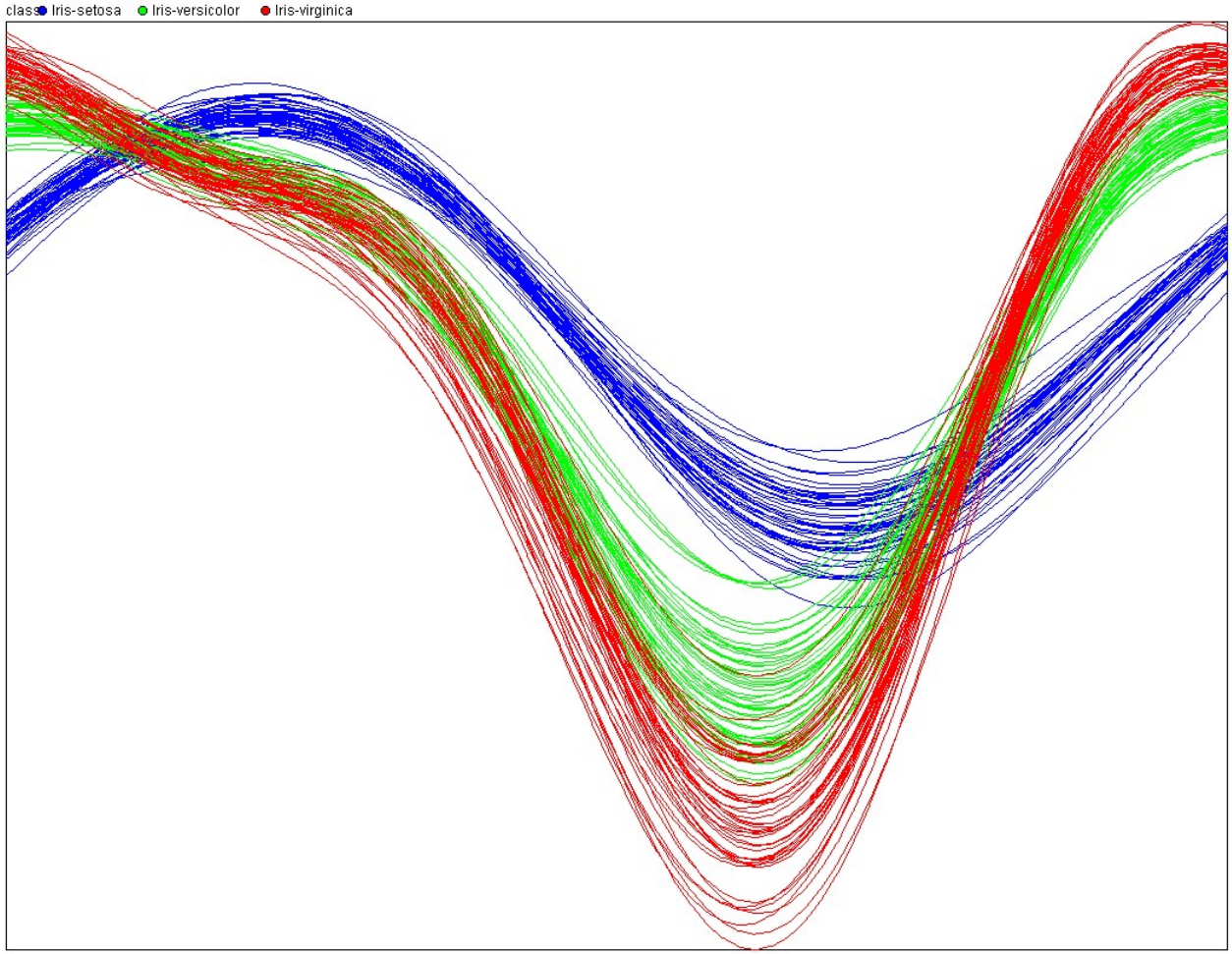
Deviation



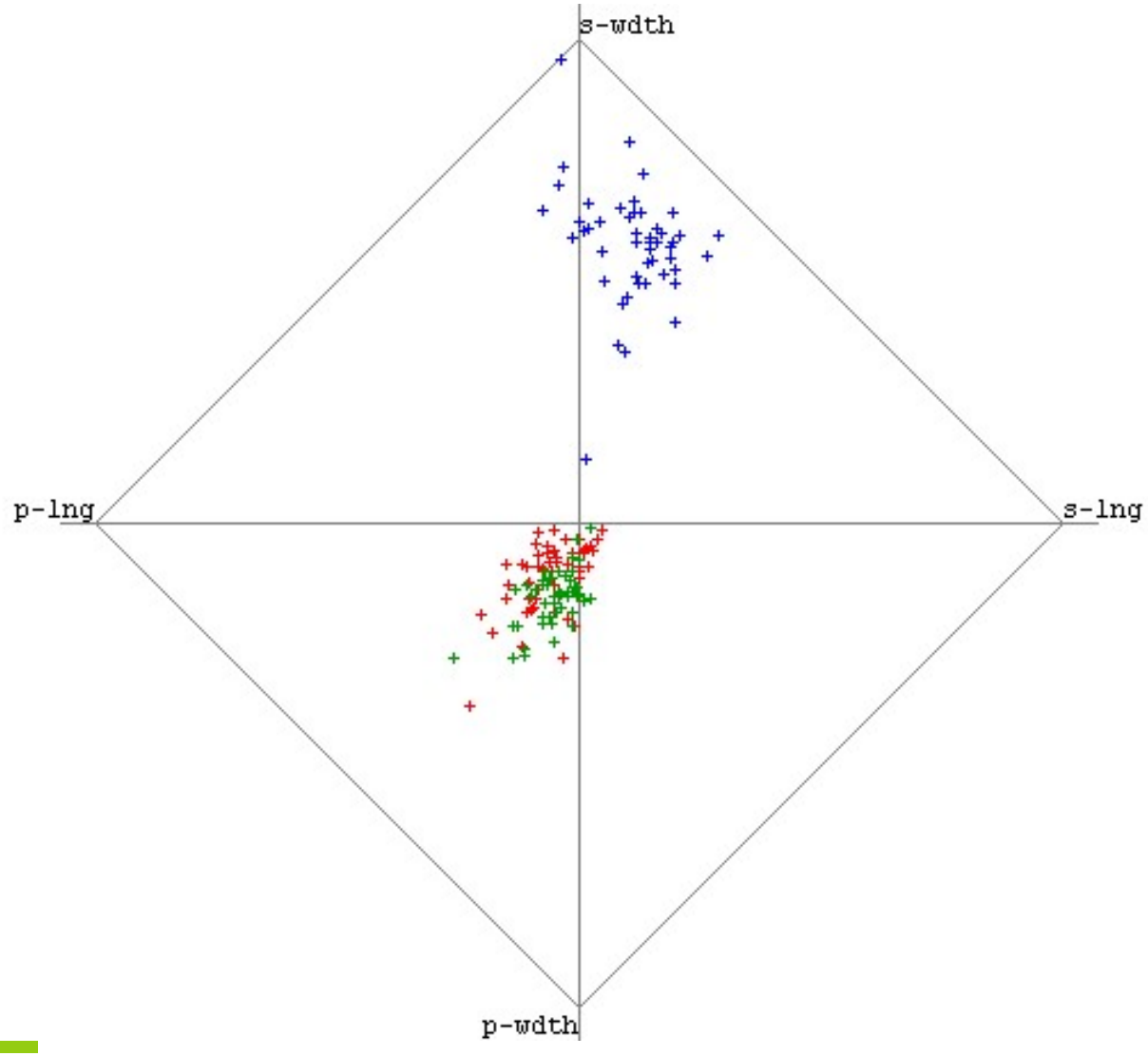
Nature Inspired
Technologies Group

Andrews' Curves

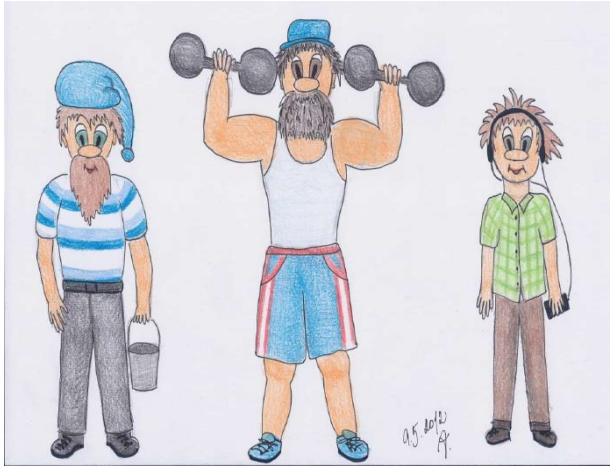
$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin(t) + x_3 \cos(t) + x_4 \sin(2t) + x_5 \cos(2t) + \ldots,$$



RadViz – Iris dataset

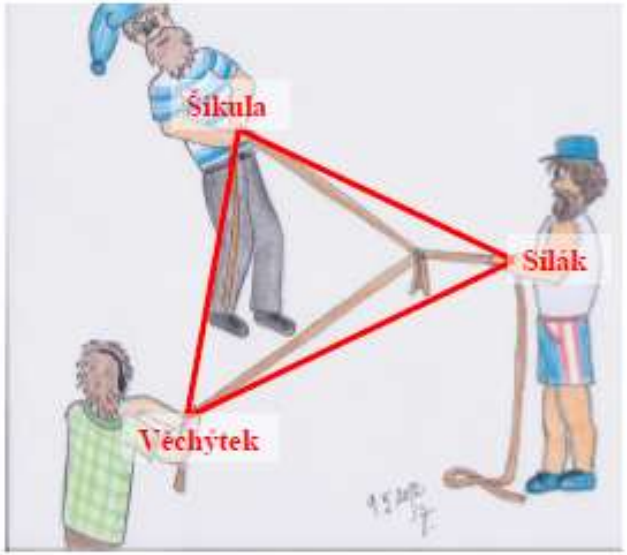


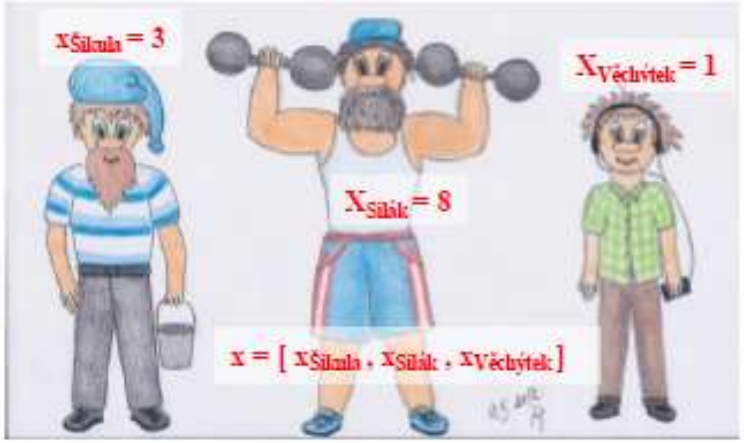
RadViz



Valentová K.: Vizualizace multidimenzionálních dat metodou RadViz, diplomová práce, 2012, vedoucí práce L.Vysloužilová

RadViz

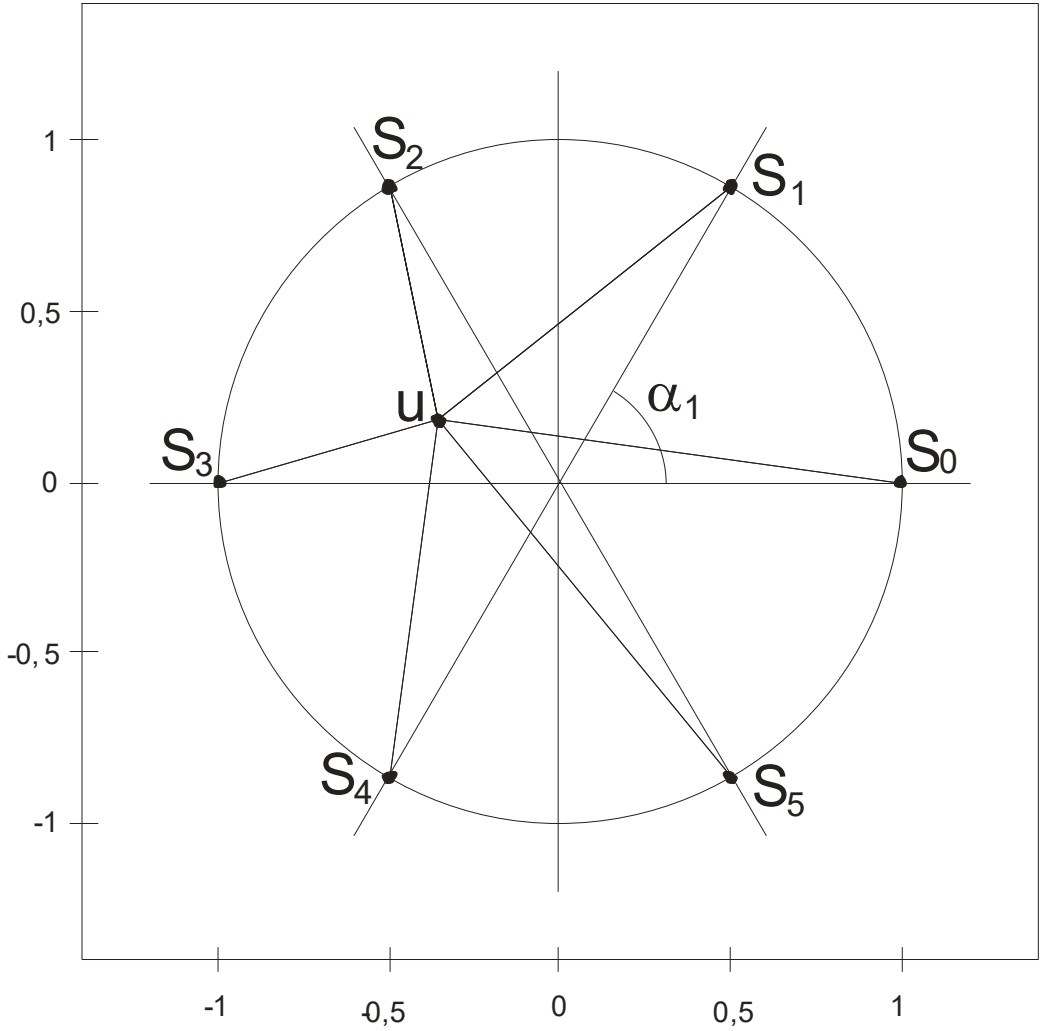




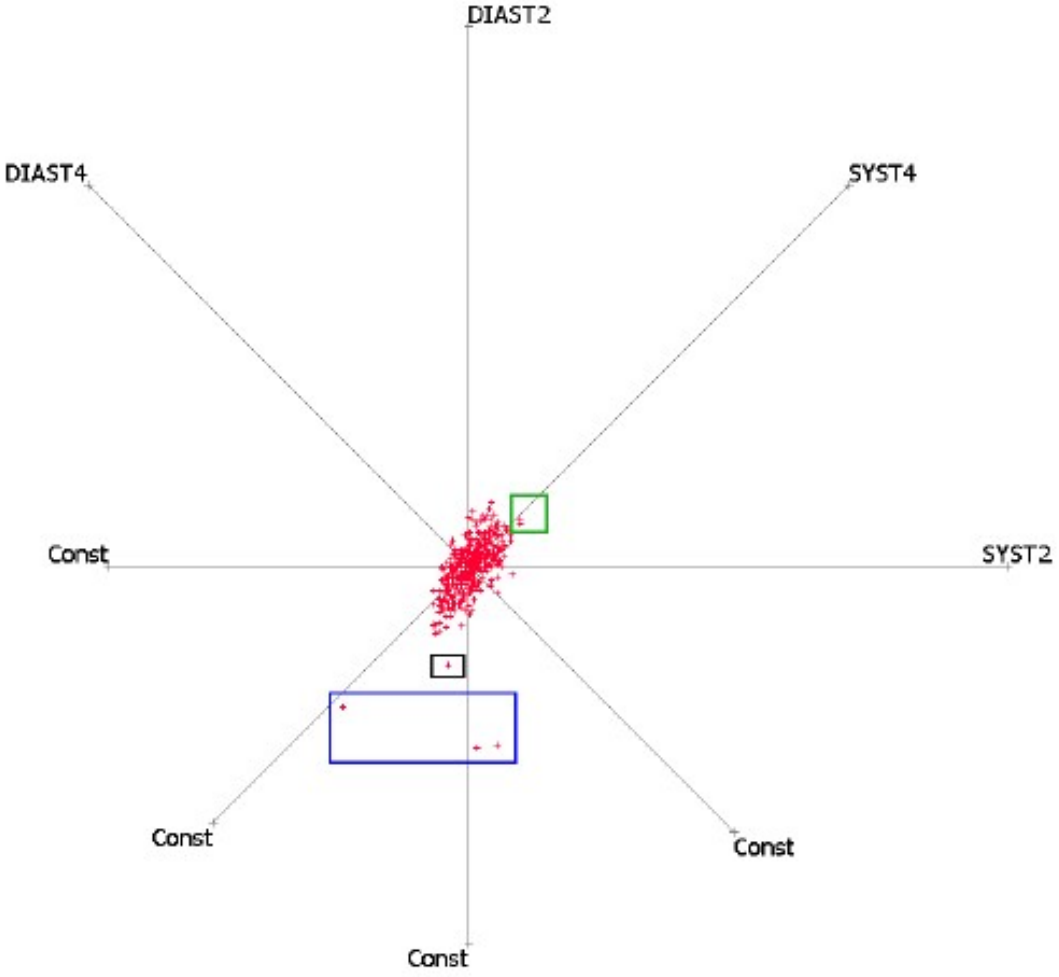
Iris data = 4 atributy,
tj. můžeme si představit čtyři přetahující se trpaslíky,
hodnoty jejich síly, budou z následující tabulky.

sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
...	...	...	...
5.9	3	5.1	1.8

RadViz



Identifikace odlehlých hodnot

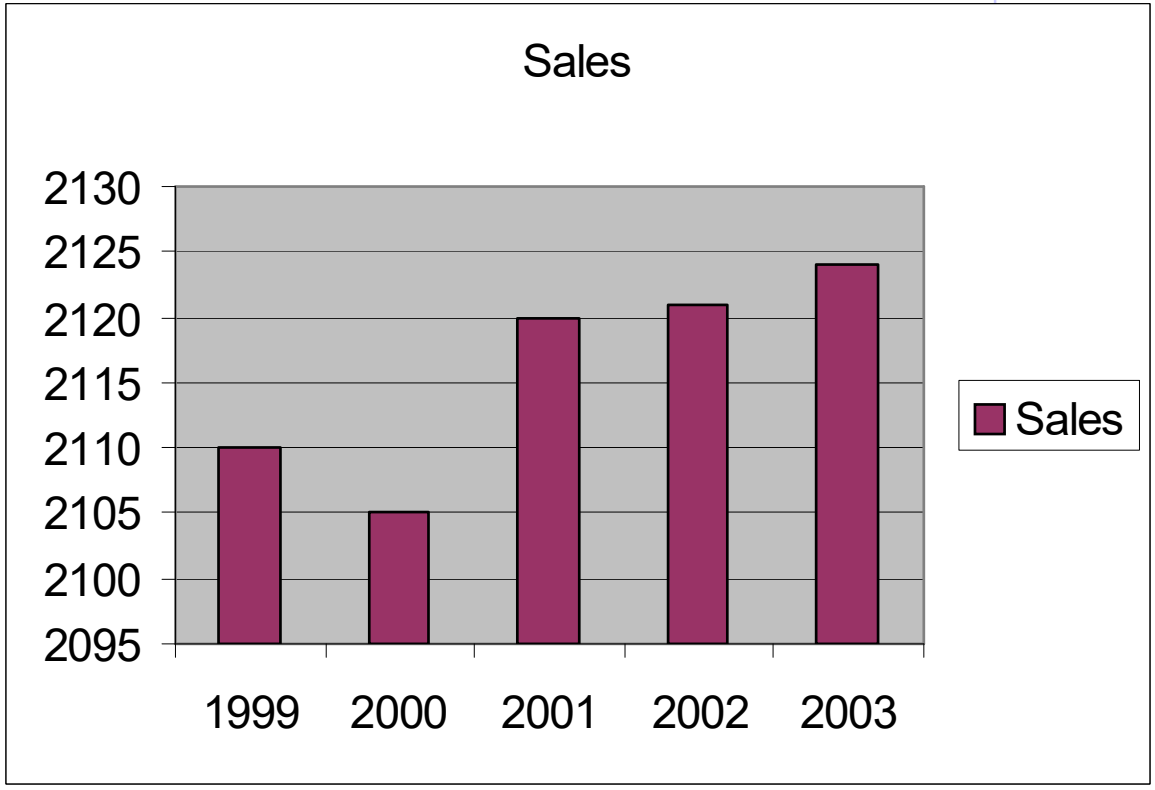


Problémy

- Vliv měřítka
- Vliv setřídění a pořadí zobrazených dat
- Skryté shluky

Vliv měřítka

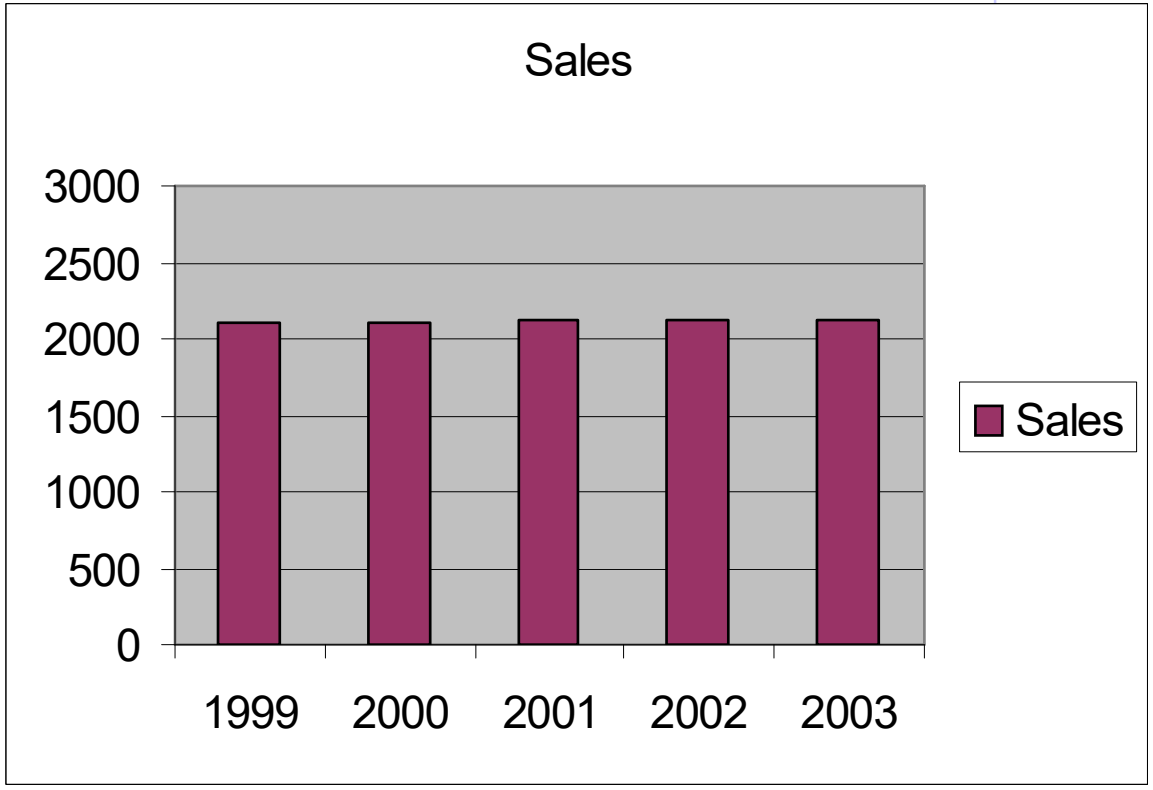
Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124



Volba špatného měřítka u osy Y
budí dojem velkých rozdílů mezi prodeji

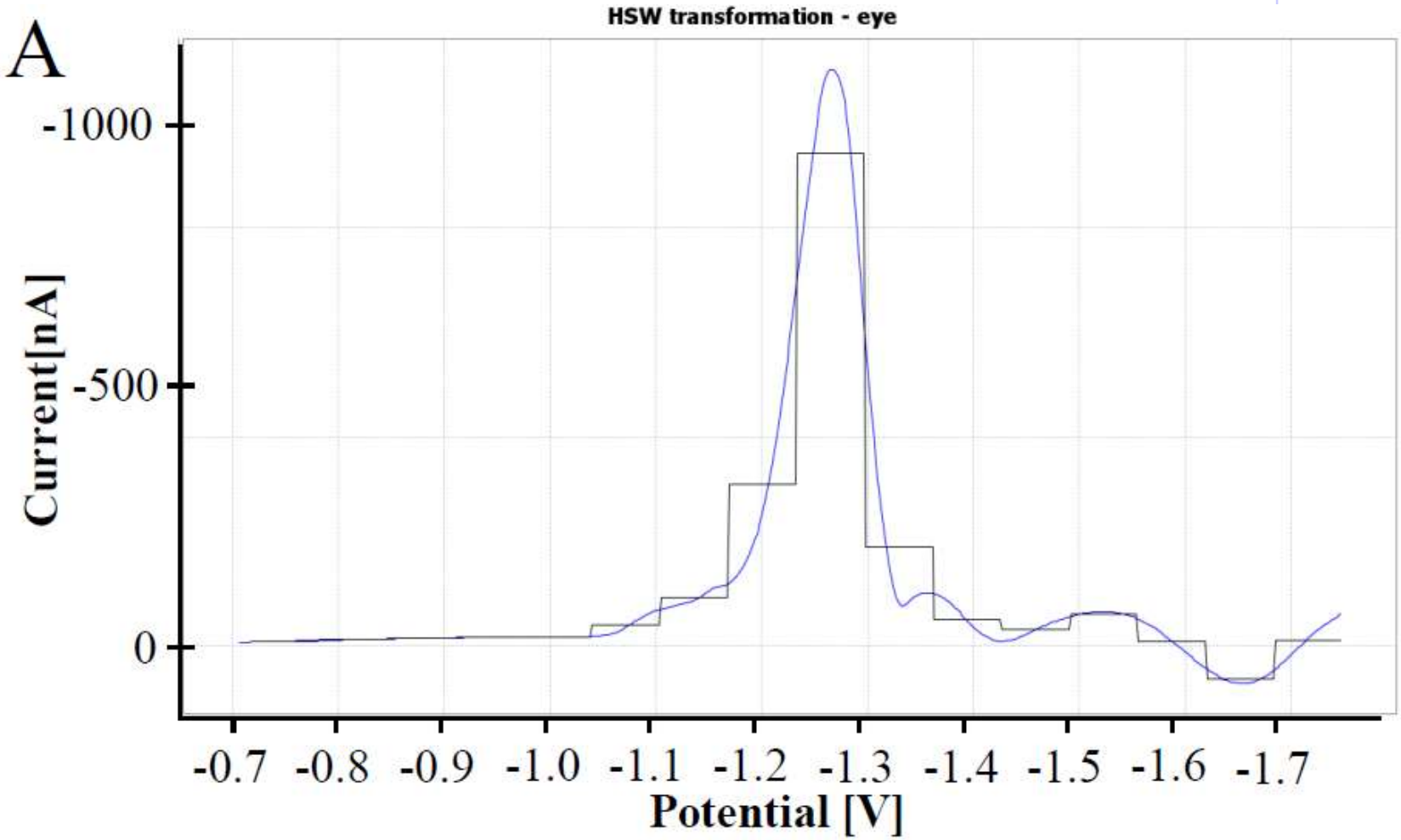
Vliv měřítka

Year	Sales
1999	2110
2000	2105
2001	2120
2002	2121
2003	2124



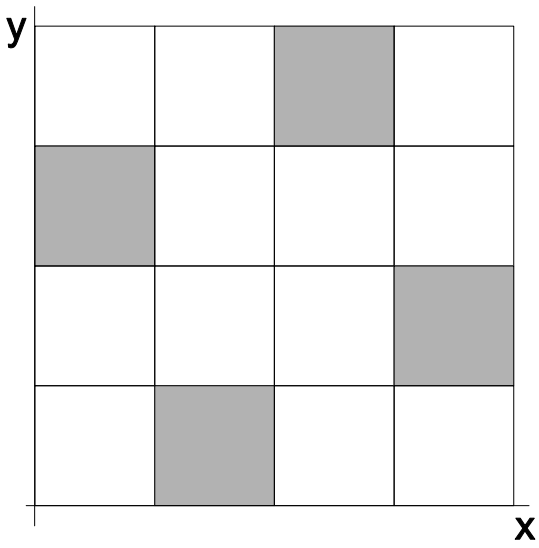
Volba měřítka od 0 do 2000 u osy Y
nyní ukazuje malé rozdíly mezi prodeji

Vliv setřídění a pořadí

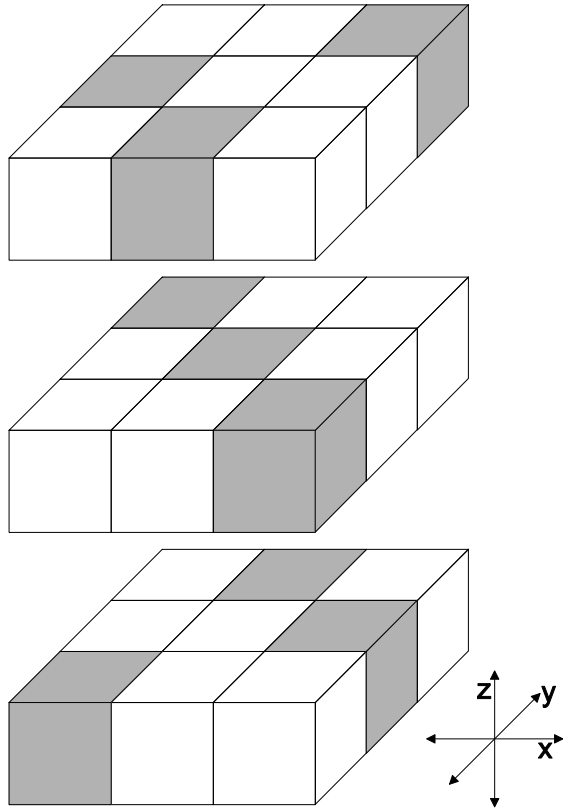


I uspořádání řady na osách může být doménově závislé!

Zobrazení shluků



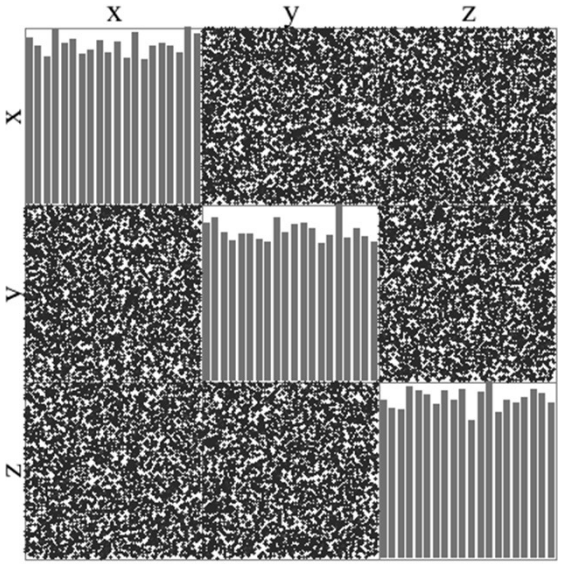
2D

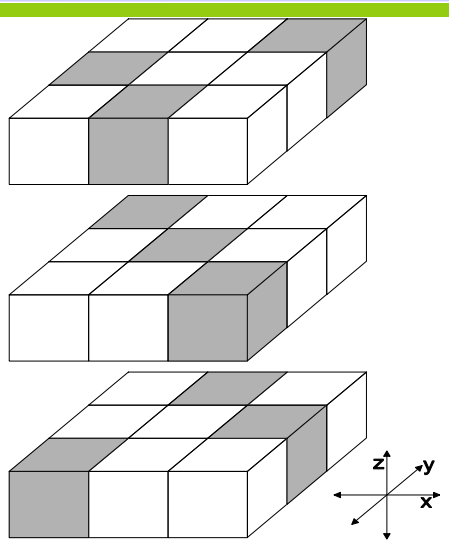


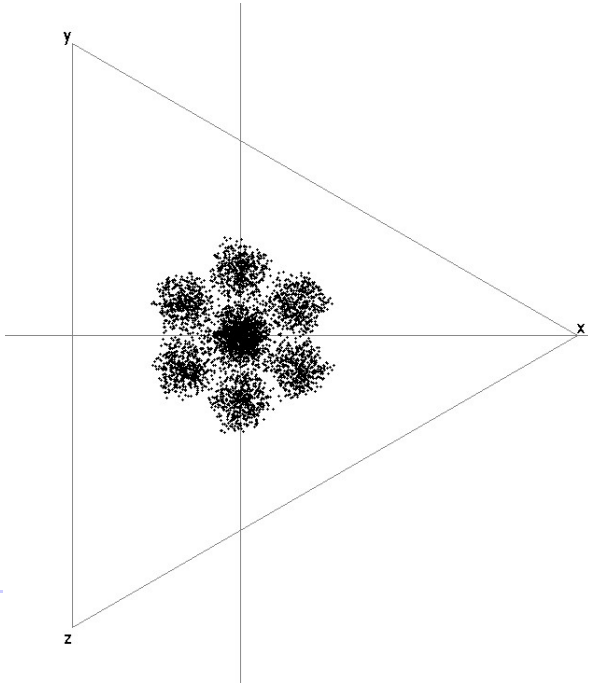
3D

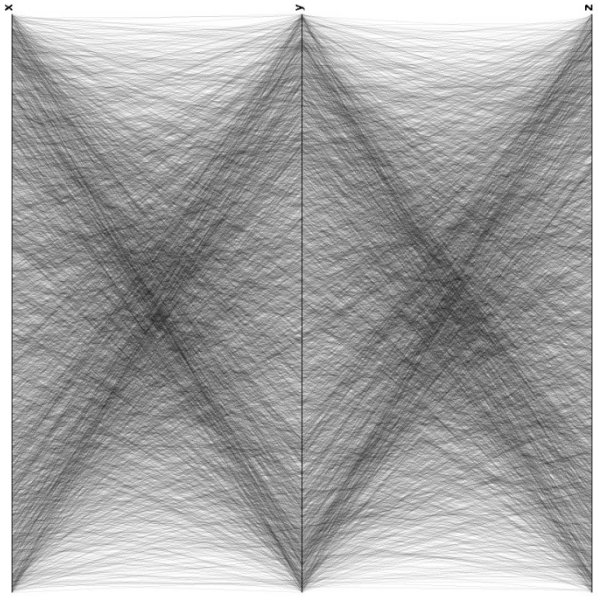
4D ...

Zobrazení shluků









Závěrečné poznámky

- vizualizace je účinný nástroj pro průzkum dat, ale není všemocný
- nutnost zachovat zvyky obvyklé v dané komunitě uživatelů
- vždy je důležitá správná interpretace