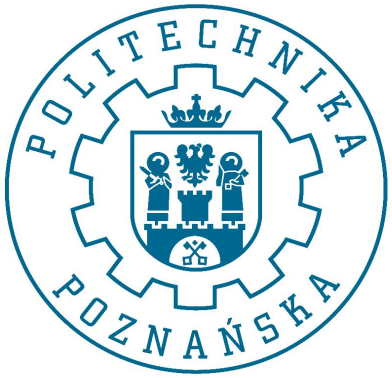
Wizualizacja danych

Prezentowanie wielowymiarowych danych

Część 1



Jerzy Stefanowski
Instytut Informatyki
Politechnika Poznańska

Wykład TWO -- listopad 2015

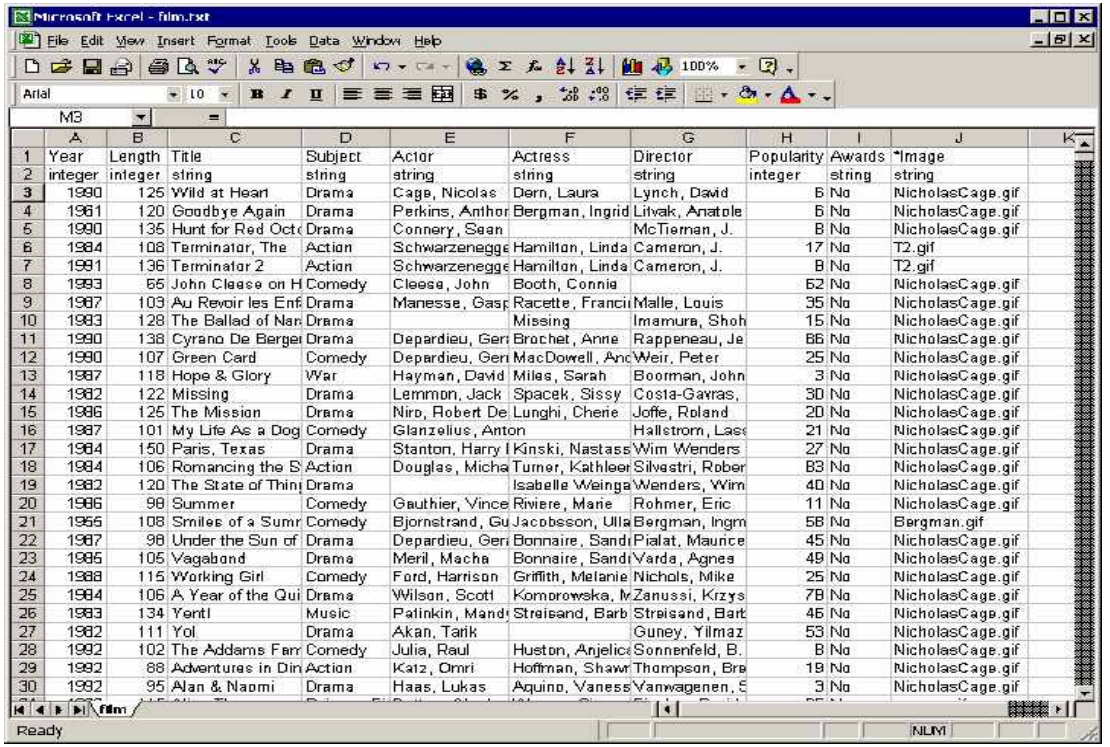
Plan wykładu

Dzisiaj

- Dane wielowymiarowe
- Prezentacje danych liczbowych
  1. Wykresy rozrzutu – od 2D do 3D
  2. Połączenie wielu elementów
  3. Macierze wykresów (Scatterplot Matrices)
  4. Mapy poziomicowe
  5. Parallel coordinates
  6. Wykresy profilowe (analiza skupisk)
  7. Inne rozwiązania (glyphs)
- Dane mieszane – kolejny wykład
  1. Trellis display
  2. Mapy ciepła
- Dane jakościowe – kolejny wykład

Dane wielowymiarowe

- Na ogół więcej niż 3 zmienne (często min. kilkanaście)
  - Zastosowanie tekstowe, obrazowe – znacznie więcej
- Podstawowy model – tabela danych (*X*×*A*)
  - Obserwacje (obiekty) opisane zbiorem zmiennych (atrybutów)



Microsoft Excel - film.txt

File Edit View Insert Format Tools Data Window Help

Arial 10 100%

M3 =

	A	B	C	D	E	F	G	H	I	J
1	Year	Length	Title	Subject	Actor	Actress	Director	Popularity	Awards	*Image
2	integer	integer	string	string	string	string	string	integer	string	string
3	1990	125	Wild at Heart	Drama	Cage, Nicolas	Dern, Laura	Lynch, David	6	No	NicholasCage.gif
4	1961	120	Goodbye Again	Drama	Perkins, Anthor	Bergman, Ingrid	Litvak, Anatole	6	No	NicholasCage.gif
5	1990	135	Hunt for Red Oct	Drama	Connery, Sean		McTiernan, J.	8	No	NicholasCage.gif
6	1984	108	Terminator, The	Action	Schwarzenegge	Hamilton, Linda	Cameron, J.	17	No	T2.gif
7	1991	136	Terminator 2	Action	Schwarzenegge	Hamilton, Linda	Cameron, J.	8	No	T2.gif
8	1993	65	John Cleese on H	Comedy	Cleese, John	Booth, Connie		62	No	NicholasCage.gif
9	1987	103	Au Revoir les Enf	Drama	Manesse, Gasp	Racette, Francii	Malle, Louis	35	No	NicholasCage.gif
10	1983	128	The Ballad of Nar	Drama		Missing	Imamura, Shoh	15	No	NicholasCage.gif
11	1990	138	Cyrano De Berger	Drama	Depardieu, Ger	Brochet, Anne	Rappeneau, Je	86	No	NicholasCage.gif
12	1990	107	Green Card	Comedy	Depardieu, Ger	MacDowell, And	Weir, Peter	25	No	NicholasCage.gif
13	1987	118	Hope & Glory	War	Hayman, David	Miles, Sarah	Boorman, John	3	No	NicholasCage.gif
14	1982	122	Missing	Drama	Lemmon, Jack	Spacek, Sissy	Costa-Gavras,	30	No	NicholasCage.gif
15	1986	125	The Mission	Drama	Niro, Robert De	Lunghi, Cherie	Joffe, Roland	20	No	NicholasCage.gif
16	1987	101	My Life As a Dog	Comedy	Glanzelius, Anton		Hallstrom, Las	21	No	NicholasCage.gif
17	1984	150	Paris, Texas	Drama	Stanton, Harry I	Kinski, Nastass	Wim Wenders	27	No	NicholasCage.gif
18	1984	106	Romancing the S	Action	Douglas, Micha	Turner, Kathleer	Silvestri, Rober	83	No	NicholasCage.gif
19	1982	120	The State of Thin	Drama		Isabelle Weinga	Wenders, Wim	40	No	NicholasCage.gif
20	1986	98	Summer	Comedy	Gauthier, Vince	Riviere, Marie	Rohmer, Eric	11	No	NicholasCage.gif
21	1955	108	Smiles of a Sumr	Comedy	Bjornstrand, Gu	Jacobsson, Ulla	Bergman, Ingm	58	No	Bergman.gif
22	1987	98	Under the Sun of	Drama	Depardieu, Ger	Bonnaire, Sandi	Pialat, Maurice	45	No	NicholasCage.gif
23	1985	105	Vagabond	Drama	Meril, Macha	Bonnaire, Sandi	Varda, Agnes	49	No	NicholasCage.gif
24	1988	115	Working Girl	Comedy	Ford, Harrison	Griffith, Melanie	Nichols, Mike	25	No	NicholasCage.gif
25	1984	106	A Year of the Qui	Drama	Wilson, Scott	Komorowska, M	Zanussi, Krzys	78	No	NicholasCage.gif
26	1983	134	Yentl	Music	Patinkin, Mand	Streisand, Barb	Streisand, Barb	46	No	NicholasCage.gif
27	1982	111	Yol	Drama	Akan, Tarik		Guney, Yilmaz	53	No	NicholasCage.gif
28	1992	102	The Addams Fam	Comedy	Julia, Raul	Huston, Anjelic	Sonnenfeld, B.	8	No	NicholasCage.gif
29	1992	88	Adventures in Din	Action	Katz, Omri	Hoffman, Shawr	Thompson, Bra	19	No	NicholasCage.gif
30	1992	95	Alan & Naomi	Drama	Haas, Lukas	Aquino, Vaness	Vanwagenen, S	3	No	NicholasCage.gif

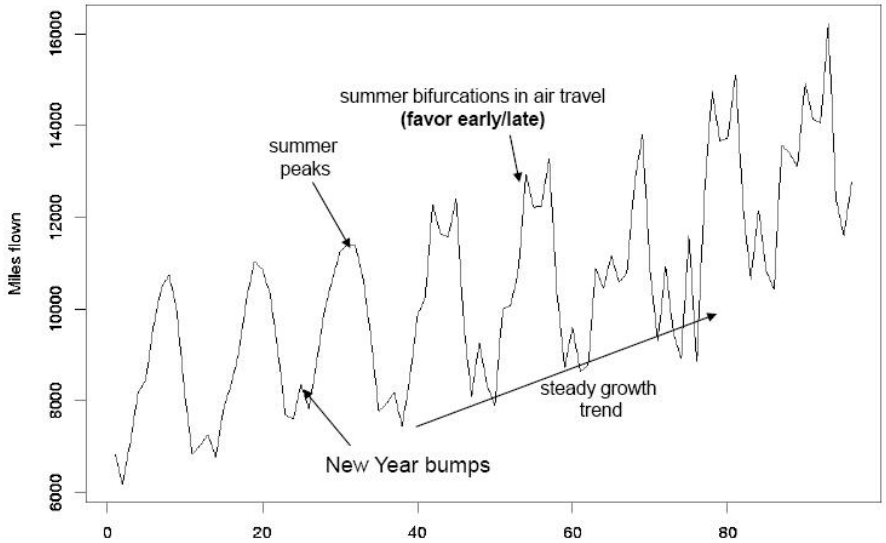
film

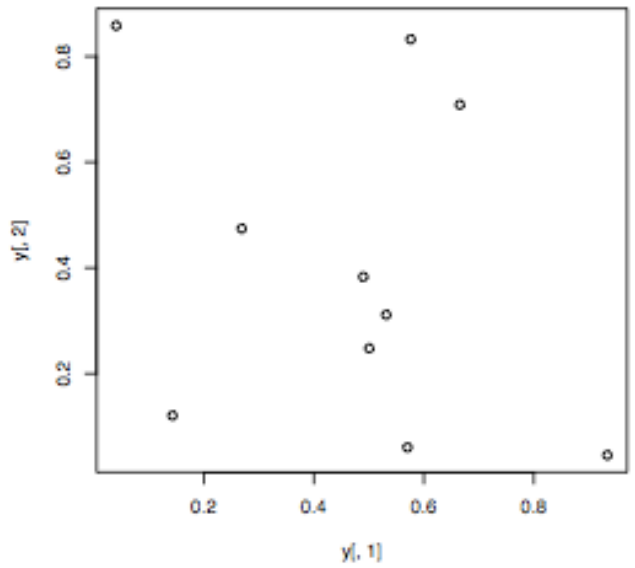
Ready NUM

Eksploracja zmiennych liczbowych

2D - podstawowe

- Wykresy liniowe (trend lines)
- Wykres rozrzutu (scatterplots)
- Standardowo wspierane przez wszystkie narzędzia

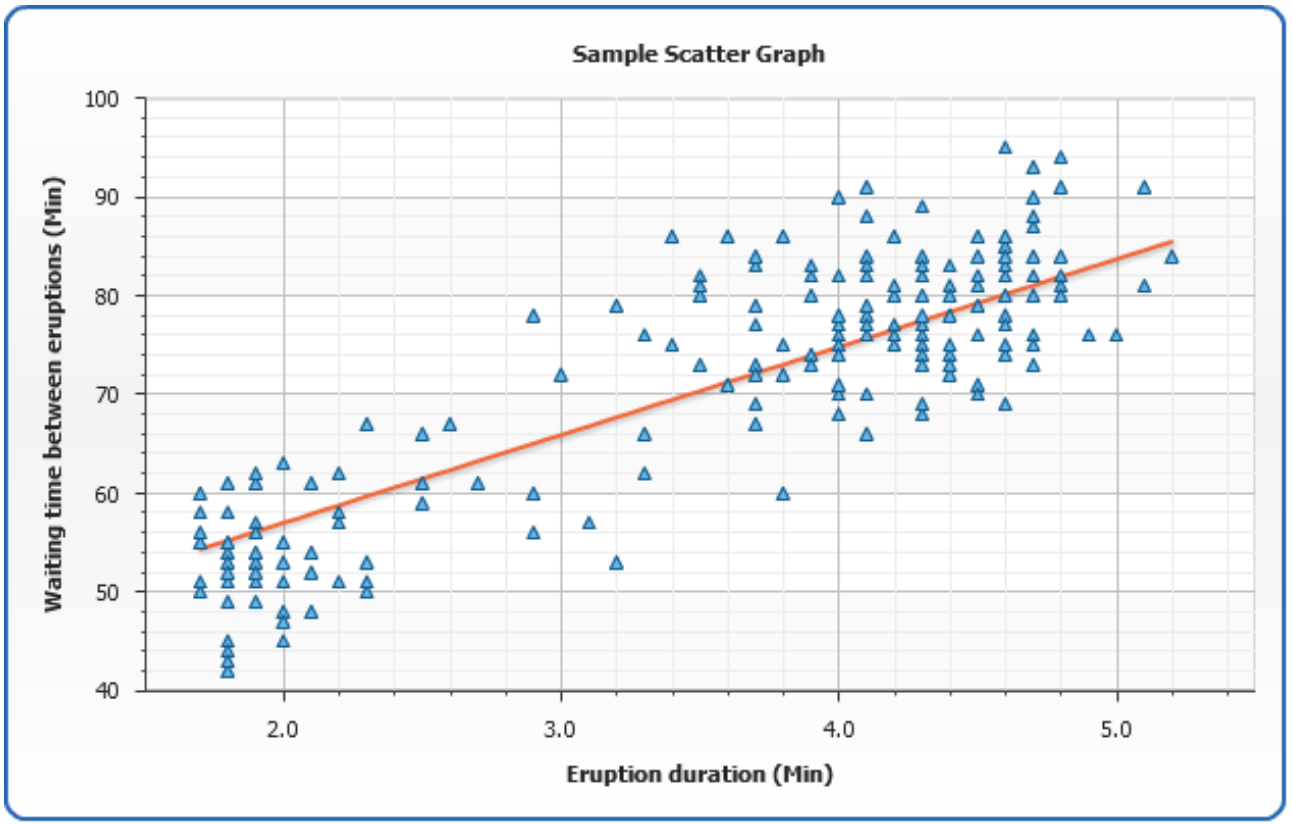




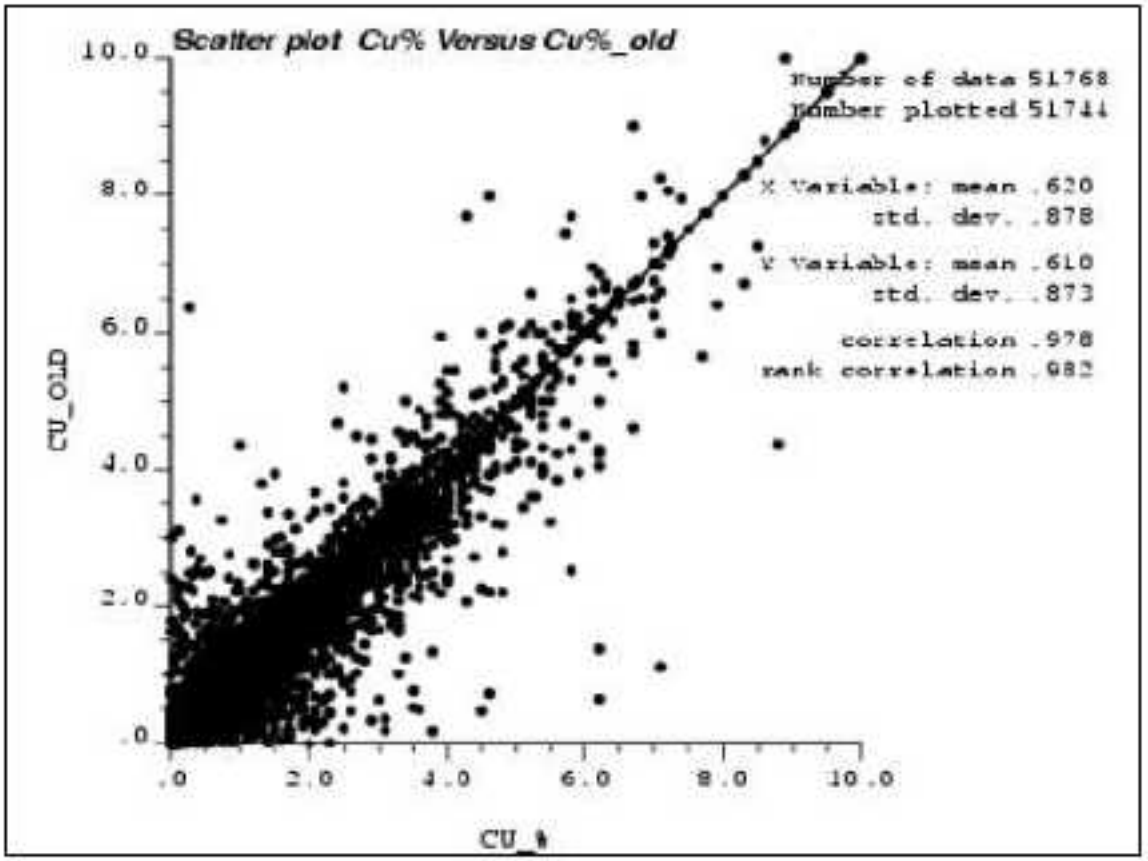

```
> set.seed(1410)
> y <- matrix(runif(30), ncol=3, dimnames=list(letters[1:10], LETTERS[1:3]))
> plot(y[,1], y[,2])
```

Wykresy rozrzutu (Scatterplots)

- Wykres rozrzutu – dobre narzędzie oglądu dwu-wymiarowych danych liczbowych
- Pozwala sprawdzić współzależności zmiennych (korelacja) oraz oceni model (regresja)



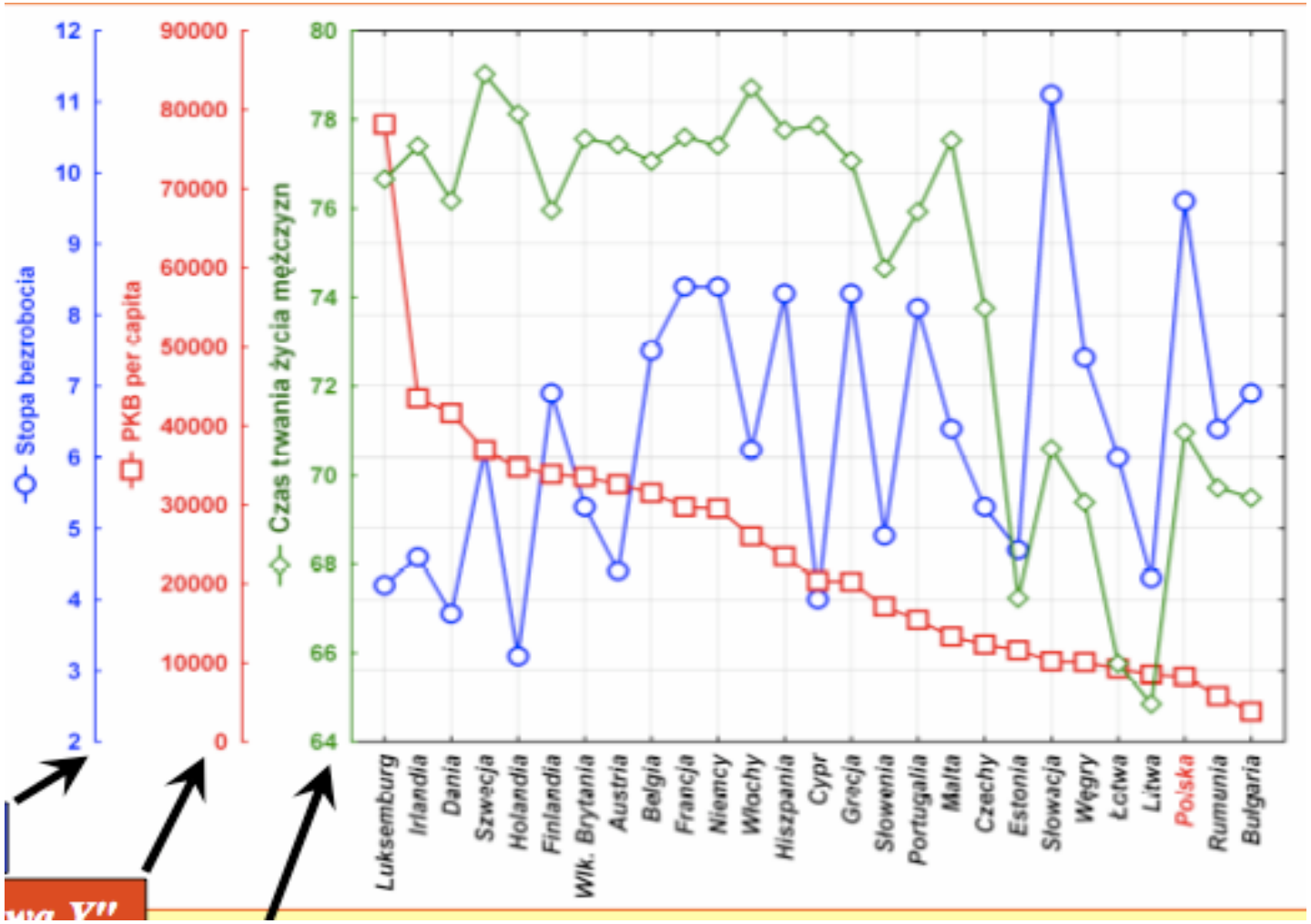
Wykresy – lecz czy czytelne dla masywnych danych



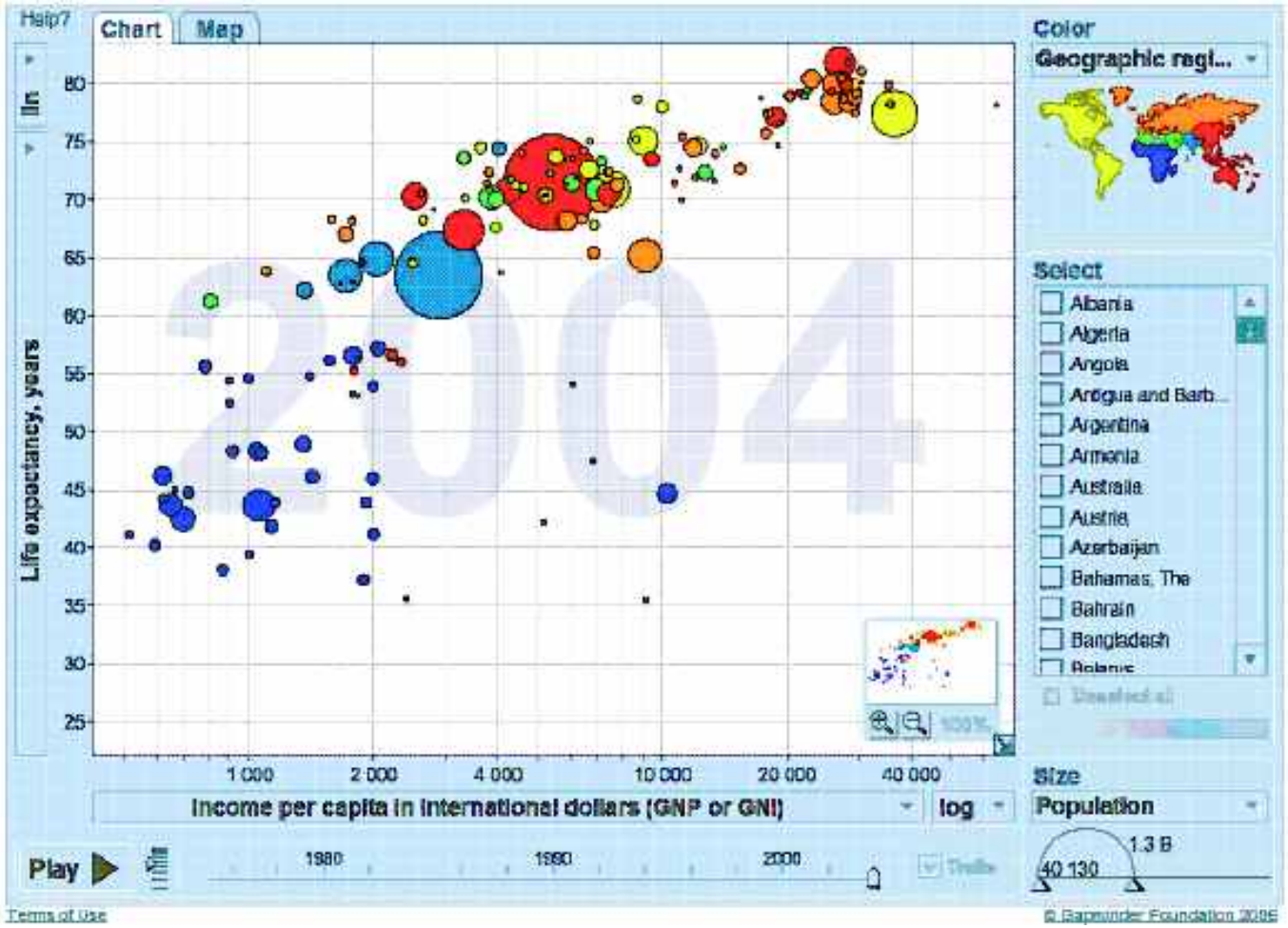
Golder, 2010.

Duża liczba obserwacji

Więcej informacji w wykresach liniowych

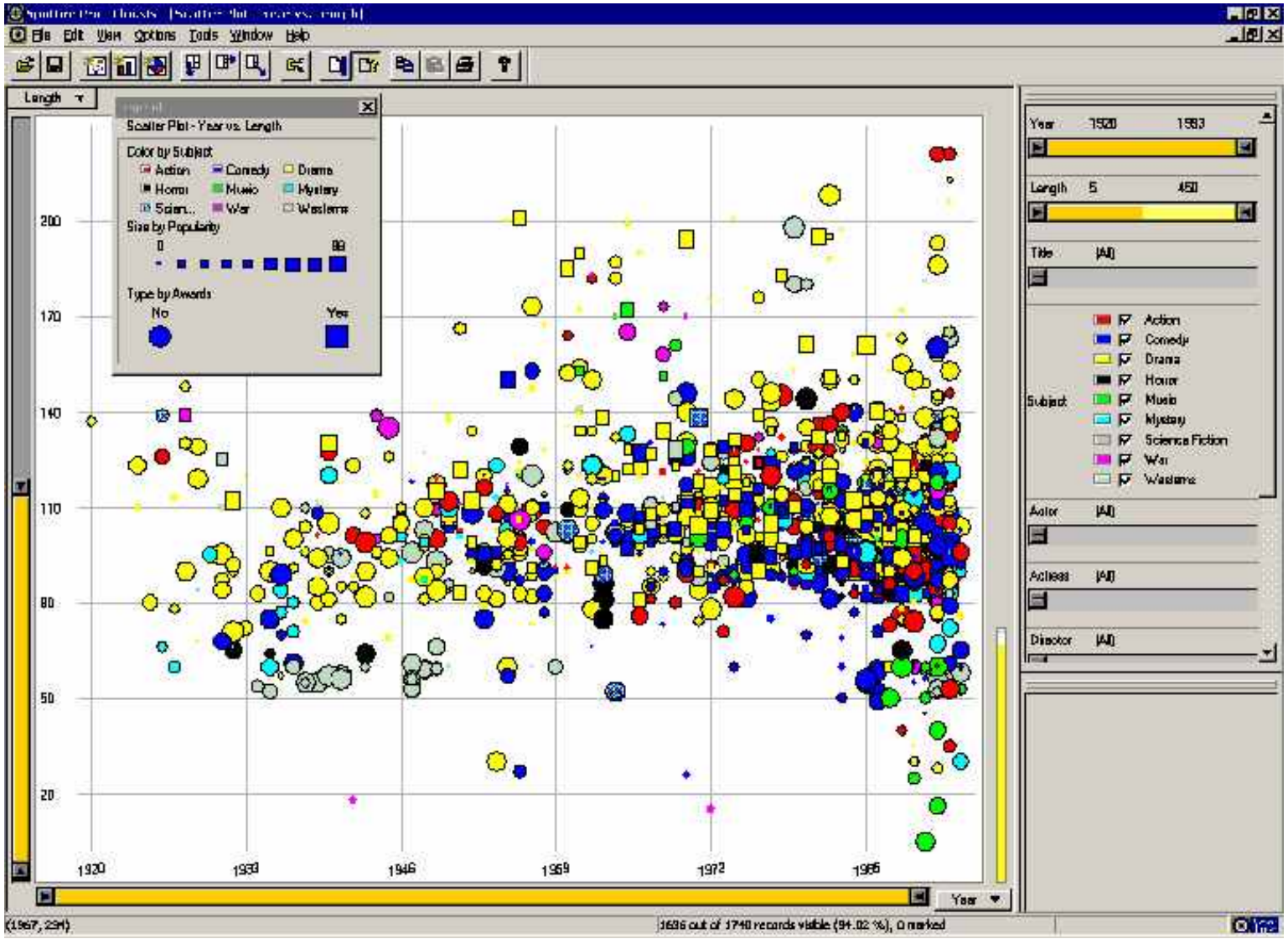


Wykresy bąbelkowe



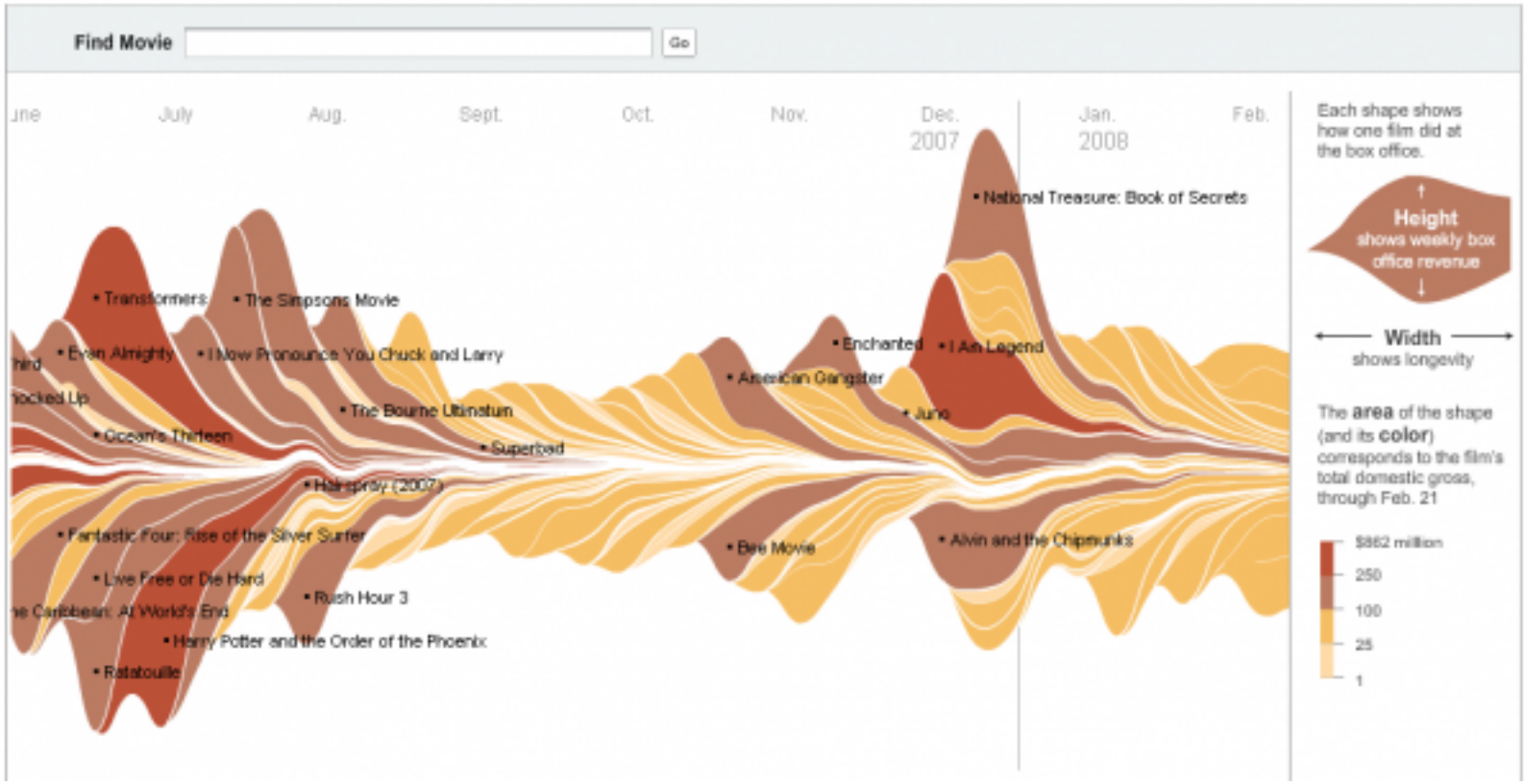
http://www.gapminder.org/

Inne rozwiązanie w oprogramowaniu Spotfire

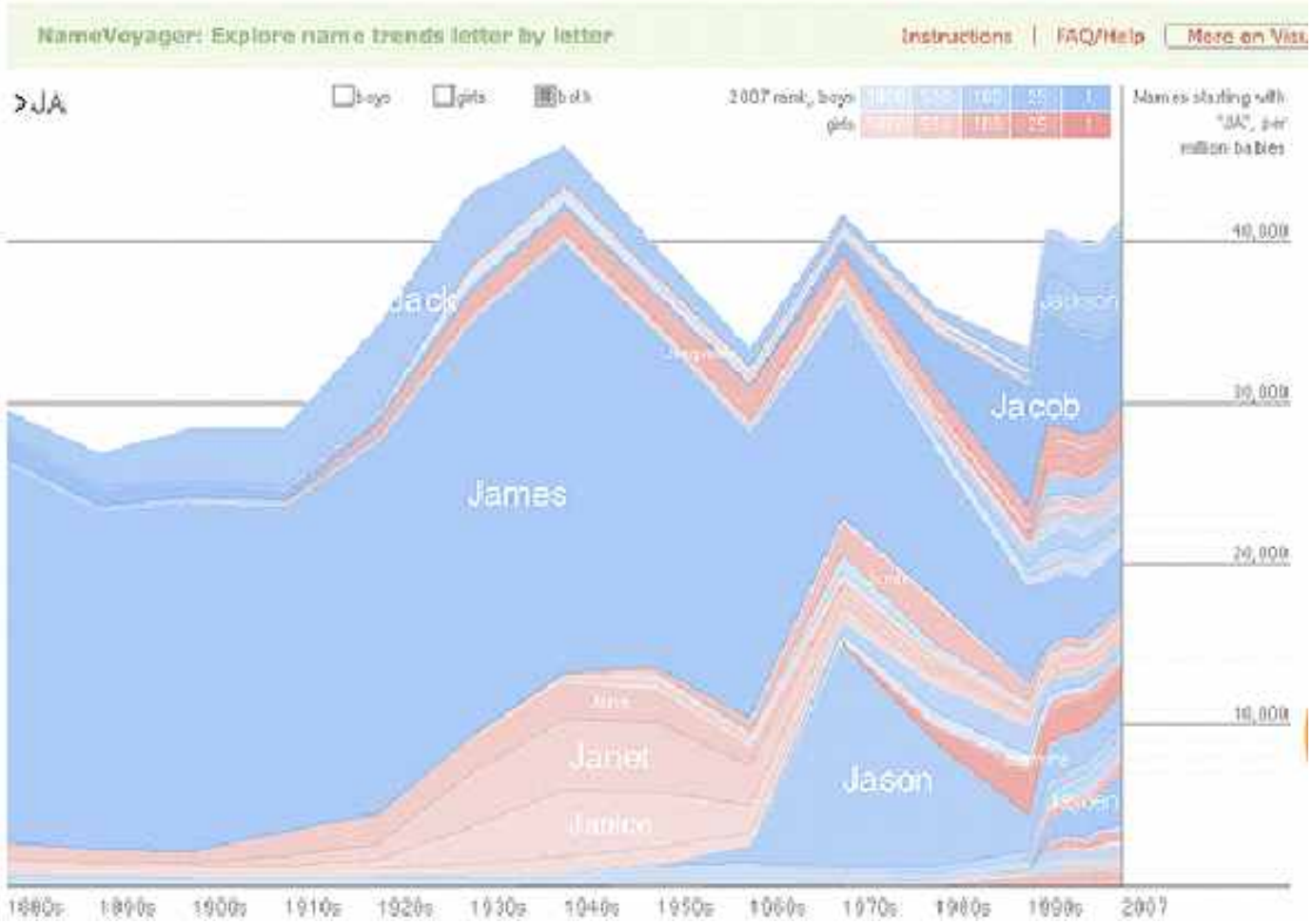


Wykresy strumieniowe

- Stream graphs – zastosowanie warstw (dodatkowa kategoryzacja) oraz aspekt zmian w zależności od czasu
- Spopularyzowane w czasopismach amerykańskich (NY)



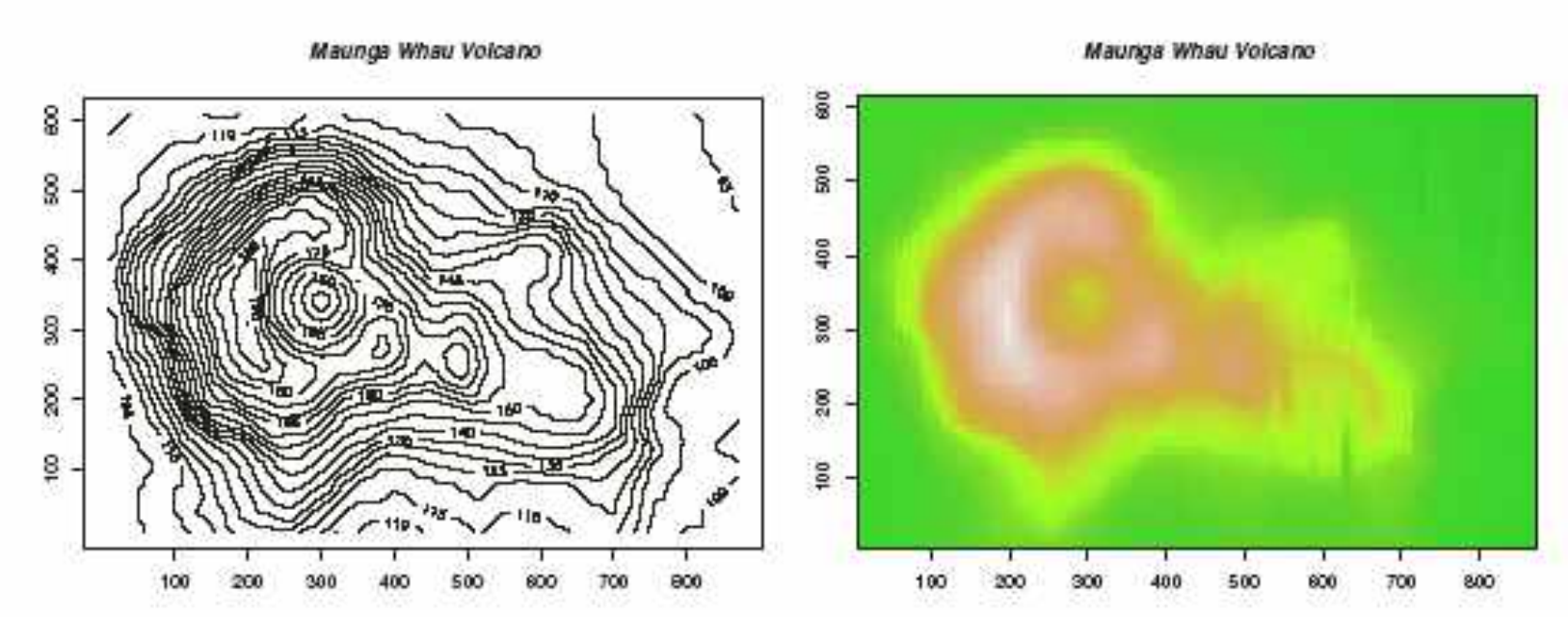
Warstwowy strumien czasowy Demo – Baby Names Voyager



http://www.babynamewizard.com/voyager

Wykresy poziomicowe

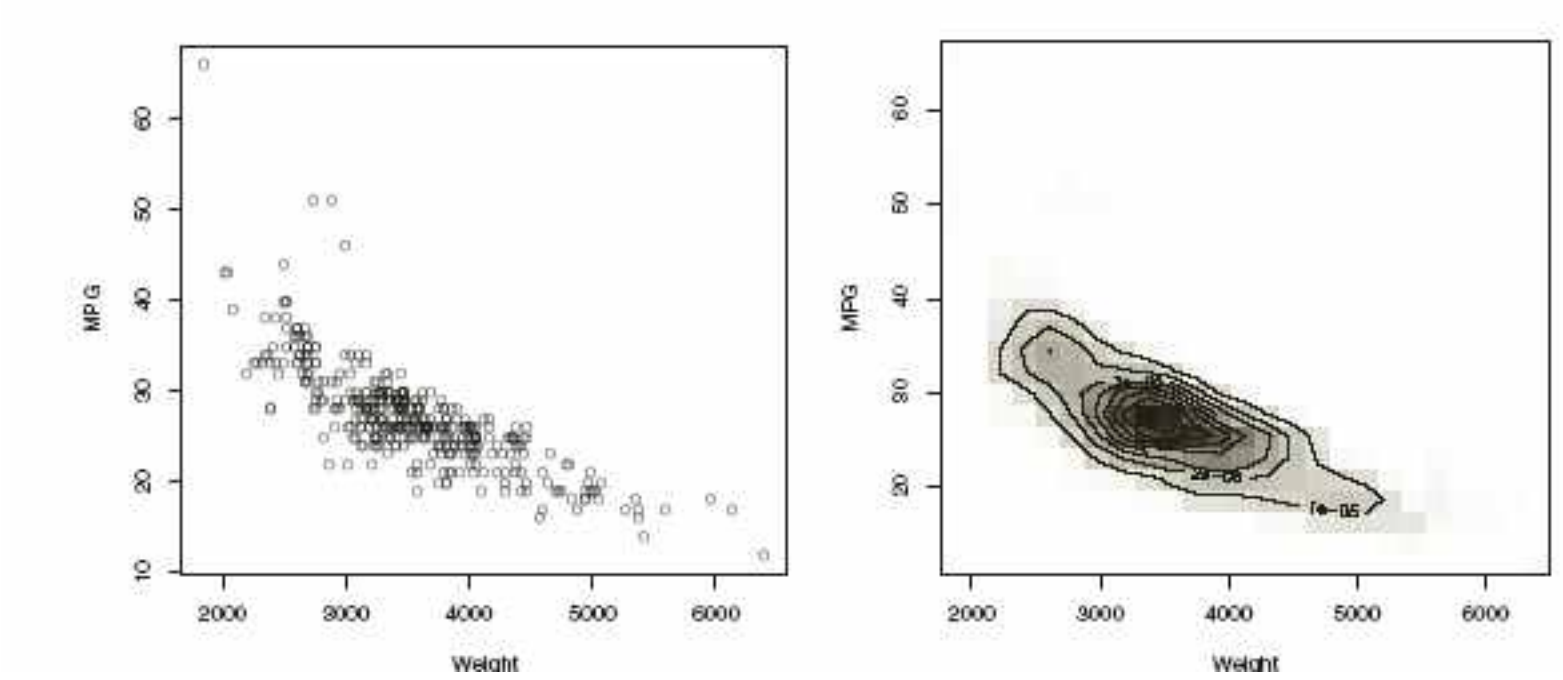
- Contour plots – analogia do tworzenie map
- Poziomice (opisane, lub ich kolorystyka) wyrażają 3 wymiar



Za – Unwin, Theus, Hofmann :
Graphics of Large Datasets. Visualizing a Million

Wykresy poziomicowe

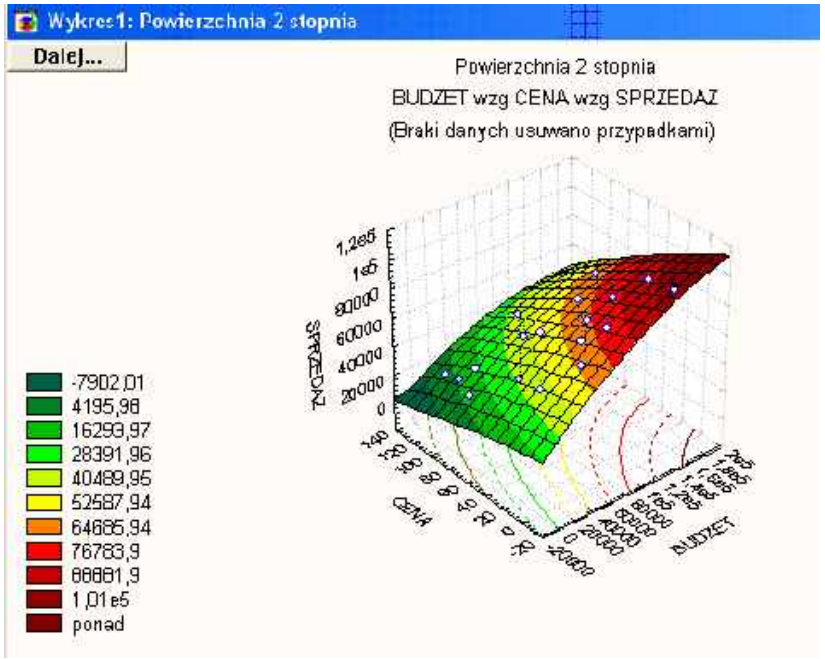
- Contour plots – Cars2004 Data sets

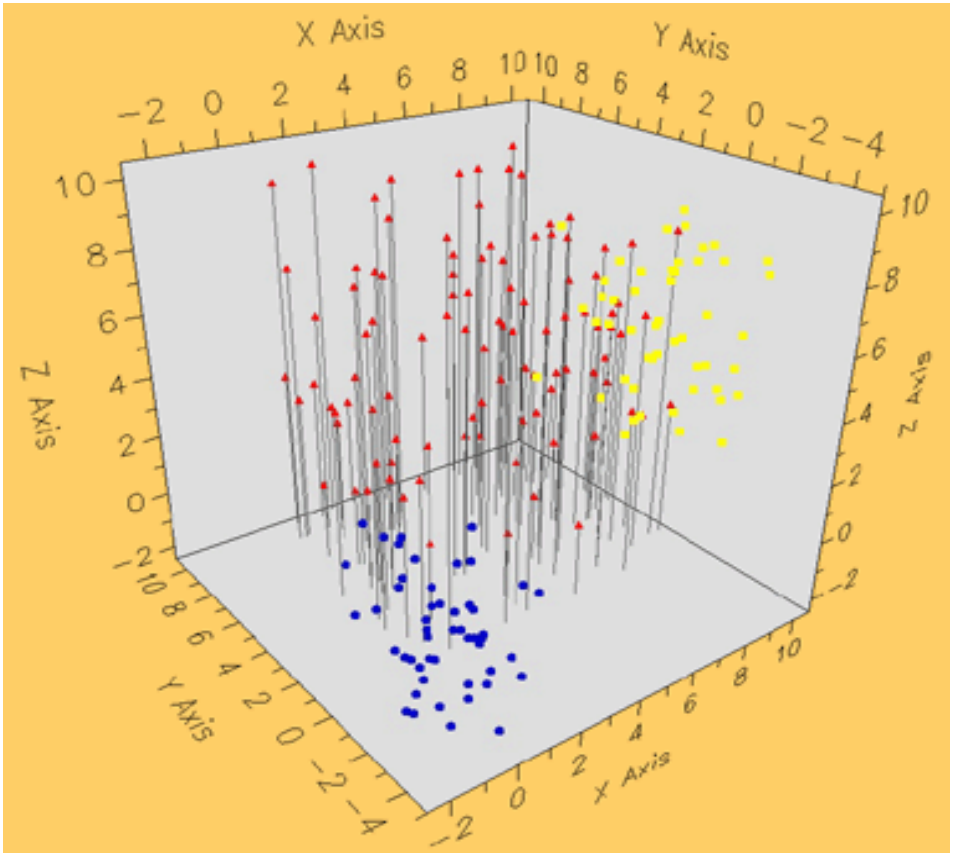


3D Scatter Plots

- Rozszerzenie wykresów rozrzutu do 3 wymiarów (XYZ)
- Użycie koloru i innych prostych metafor graficznych

- … czy dostatecznie czytelne?



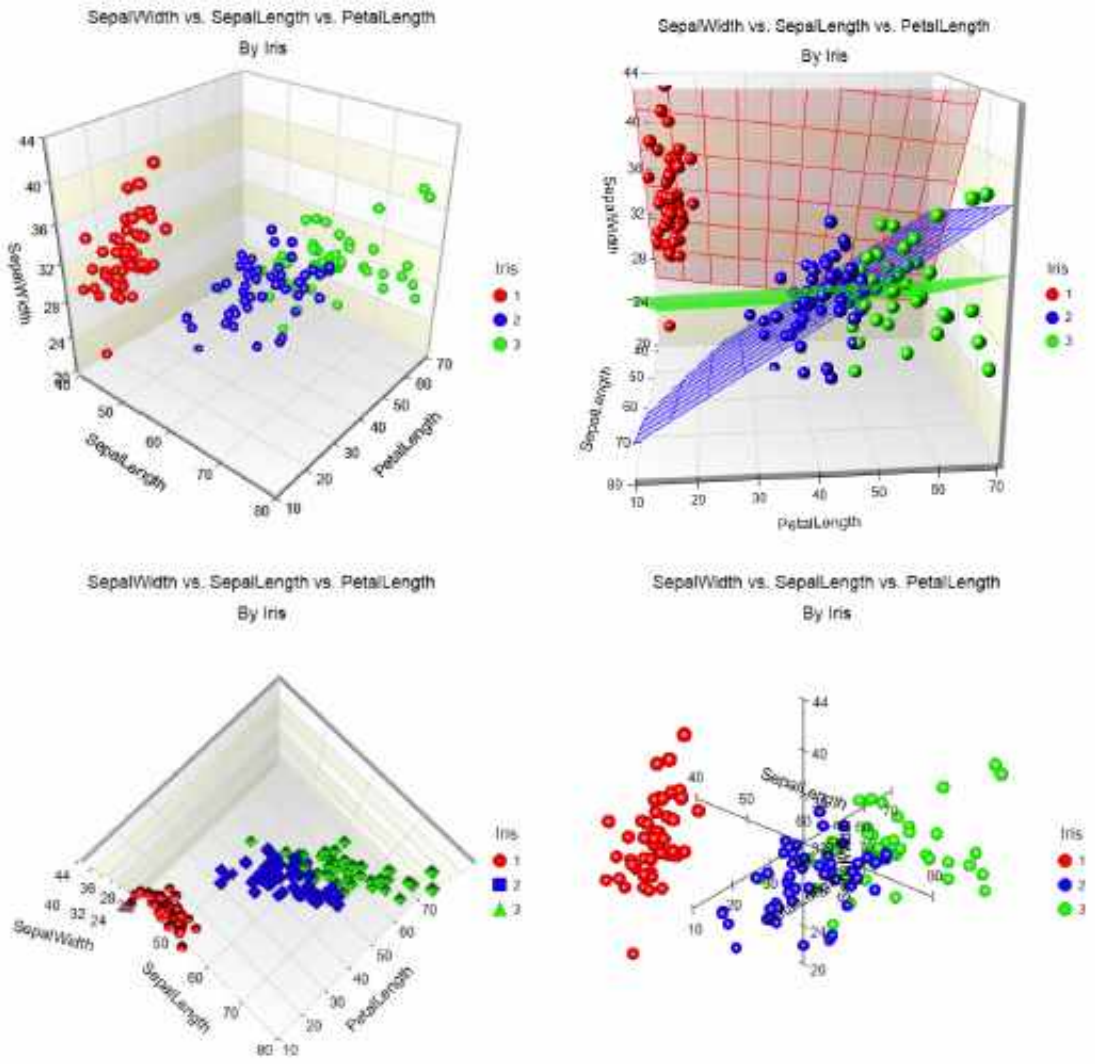


http://www.ist.co.uk/XRT/xrt3d.html

3D Rotating Plots

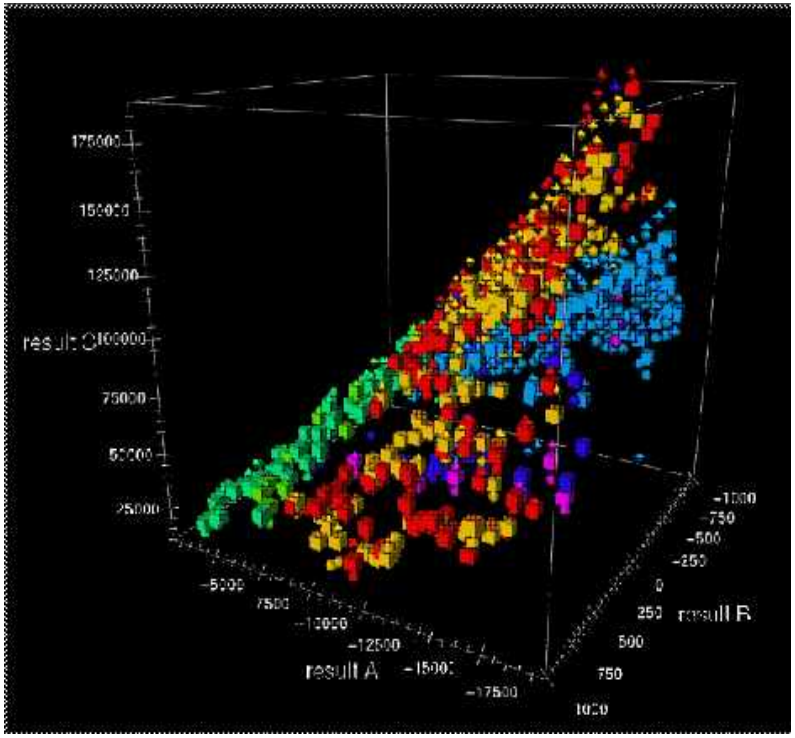
Wykres rozrzutu - 3D scatterplots

- Elementy obrotu i ruchu obserwatora
- Możliwości oglądu przekrojów
- Dostępny w niektórym oprogramowaniu



Jeszcze więcej zmiennych

- Więcej pomysłów z wykorzystaniem kodowania kolorem i obiektami geometrycznymi
- IRIS Explorer (a scientific visualization system!)
  - Five variables displayed using spatial arrangement for three, colour and object type for others
  - Rotations
  - Poszukiwanie skupisk i obserwacji nietypowych …
- Lecz ograniczenia takich rozszerzeń

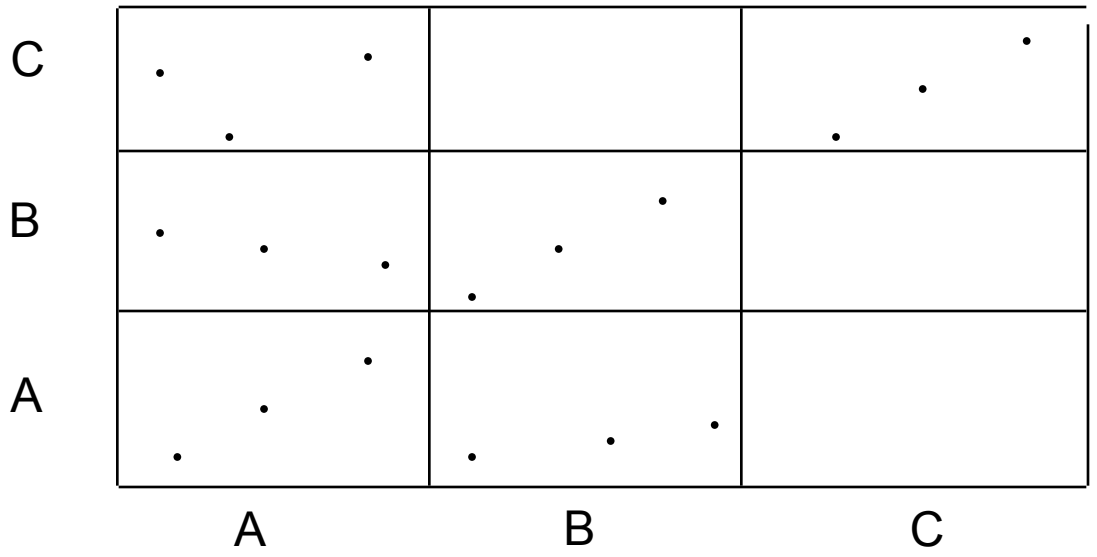


Kraus & Ertl, U Stuttgart
http://wscg.zcu.cz/wscg2001/Papers_2001/R54.pdf

Scatter Plot Matrices

Alternatywne podejście – wyłącznie „karta” dwuwymiarowa

- Tabela wielowymiarowa, jednoczesna prezentacja wielu 2D scatter plots pomiędzy wszystkimi (lub wybranymi) parami zmiennych
- Dostępne w większości oprogramowania

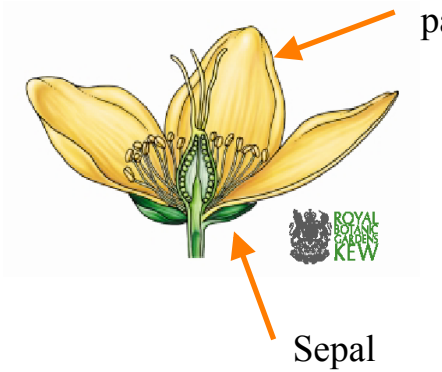


Można zauważyć korelacje zmiennych, nietypowe obserwacje, a czasami skupiska

Przykład wielowymiarowej wizualizacji – dane IRIS

R. Fisher iris data set

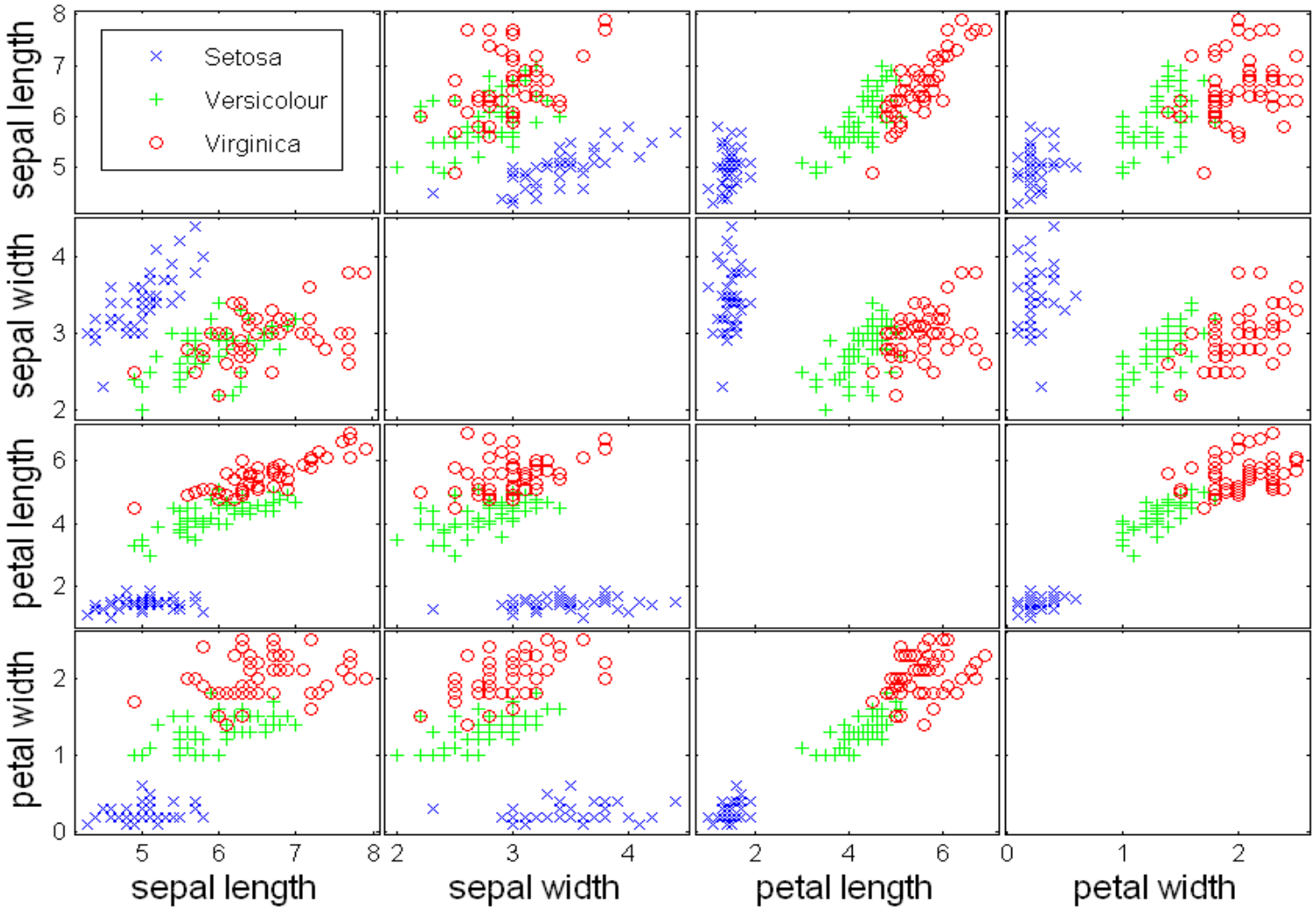
- 150 observations of 4 variables (length, width of petal and sepal)



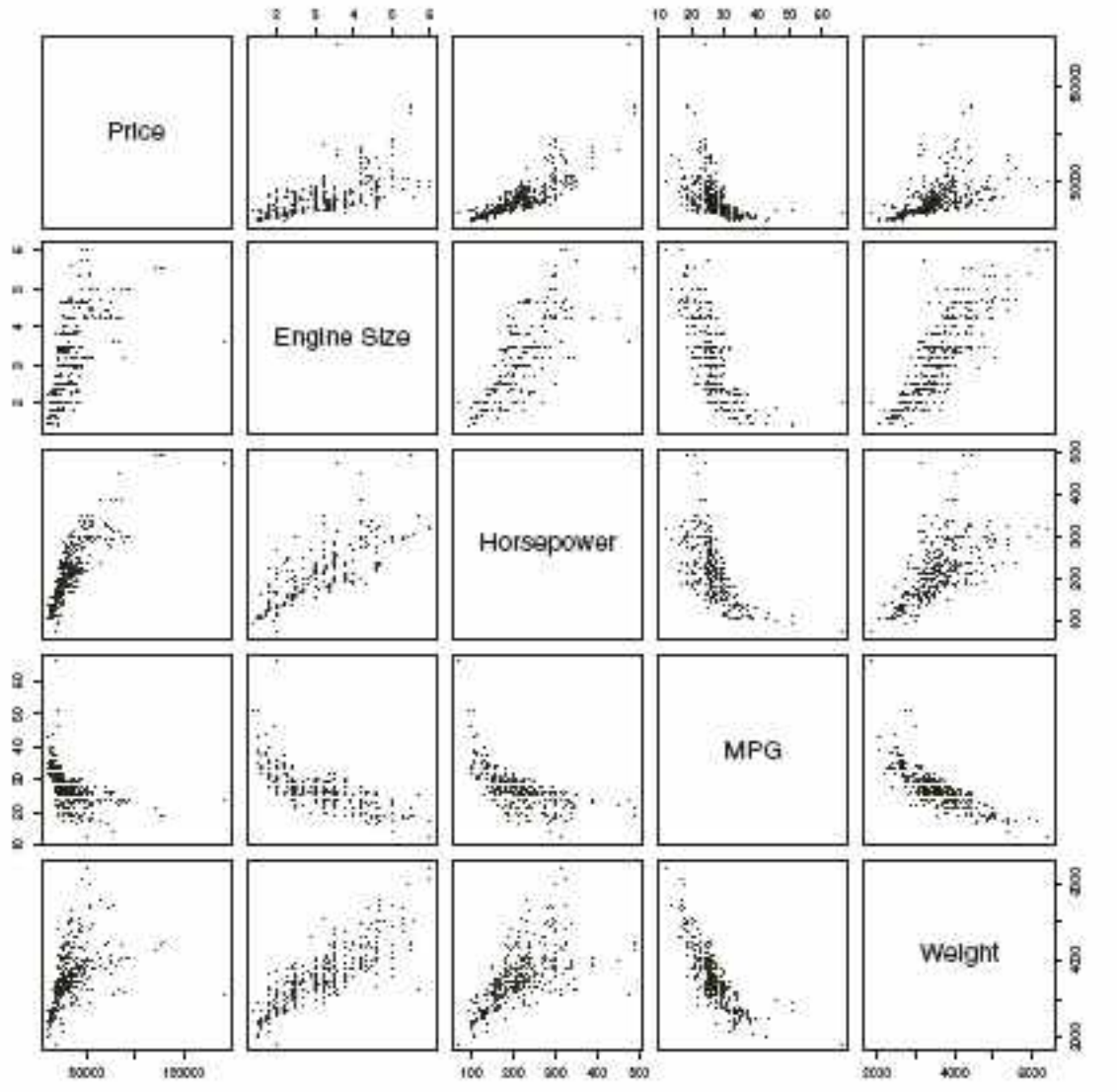
```
4 150
sepal_length
sepal_width
petal_length
petal_width
4.3 7.9 5
2.0 4.4 5
1.0 6.9 5
0.1 2.5 5
5.1   3.5   1.4   0.2
4.9   3     1.4   0.2
4.7   3.2   1.3   0.2
4.6   3.1   1.5   0.2
5     3.6   1.4   0.2
5.4   3.9   1.7   0.4
4.6   3.4   1.4   0.3
5     3.4   1.5   0.2
4.4   2.9   1.4   0.2
4.9   3.1   1.5   0.1
5.4   3.7   1.5   0.2
4.8   3.4   1.6   0.2
4.8   3     1.4   0.1
4.3   3     1.1   0.1
5.8   4     1.2   0.2
5.7   4.4   1.5   0.4
```

Challenge in visualization is to design the visualization to match the analytical task

Scatter Plot Array of Iris Attributes

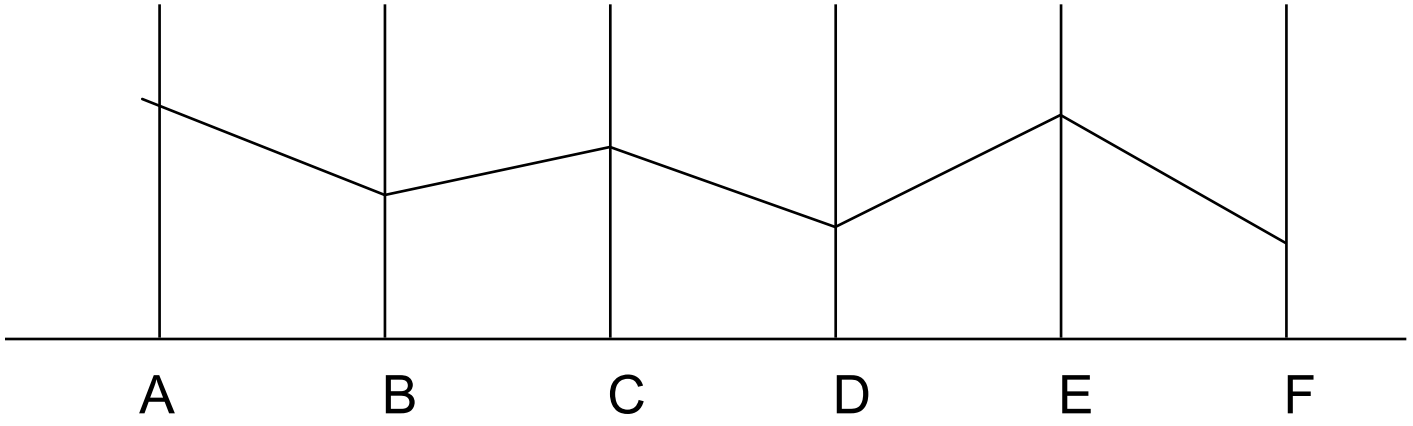


Analiza danych Cars2004



Za – Unwin, Theus, Hofmann : Graphics of Large Datasets. Visualizing a Million

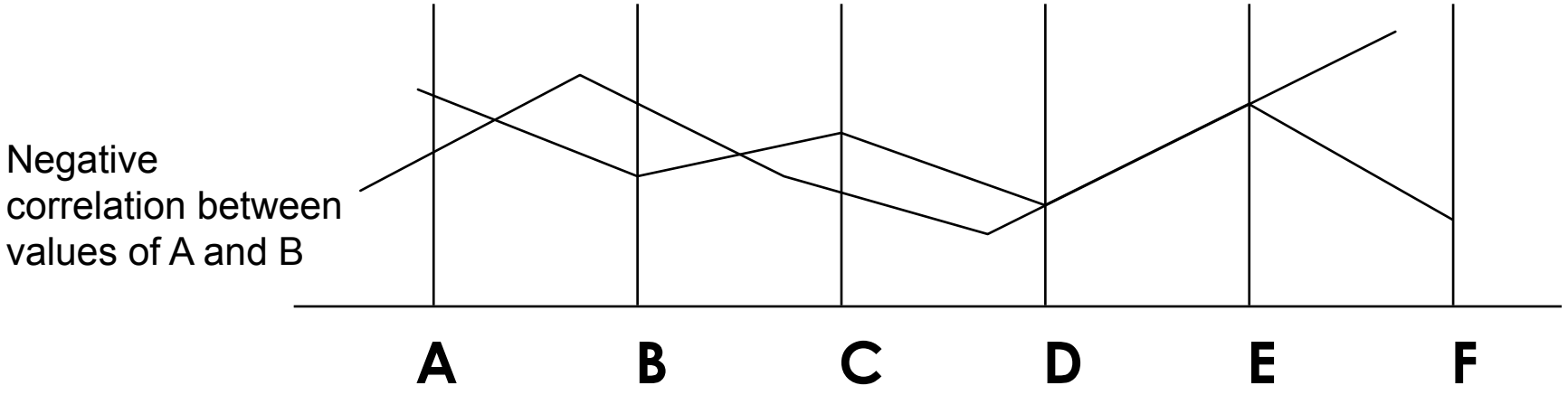
Parallel Coordinates – równoległe współrzędne



- “Coordinates” – układ współrzędnych (lecz inny niż tradycyjne)
- Prezentacja wielu wymiarów za pomocą równoległych linii współrzędnych
- Każda linia / współrzędna odnosi się do pojedynczej zmiennej
  - Równo-odległe od siebie
  - Posiadają wspólną skalę!

Każda obserwacja – profil / linia łącząca punkty wartości zmiennych

Parallel Coordinates



- Prezentują dużo informacji o obserwacjach i zmiennych je opisujące
- Zalecane do eksploracji graficznej danych
  - Ideal tool to get a first overview of a data set
- Możliwość analizy indywidualnych profili vs. inne obserwacje
  - Odpowiednie podświetlenie
  - Wykrywanie obserwacji odstających i skupisk

Ocena korelacji między zmiennymi na podstawie własności geometrycznych

Parallel Coordinates – własności geometryczne

- Oryginalne dane – układ kartezjański / linie przecinają się w punkcie
- Zmienne o korelacji negatywnej – przecinają się; punkt przecięcia między współrzędnymi
- Zmienne o korelacji pozytywnej – punkt potencjalnego przecięcia poza współrzędnymi

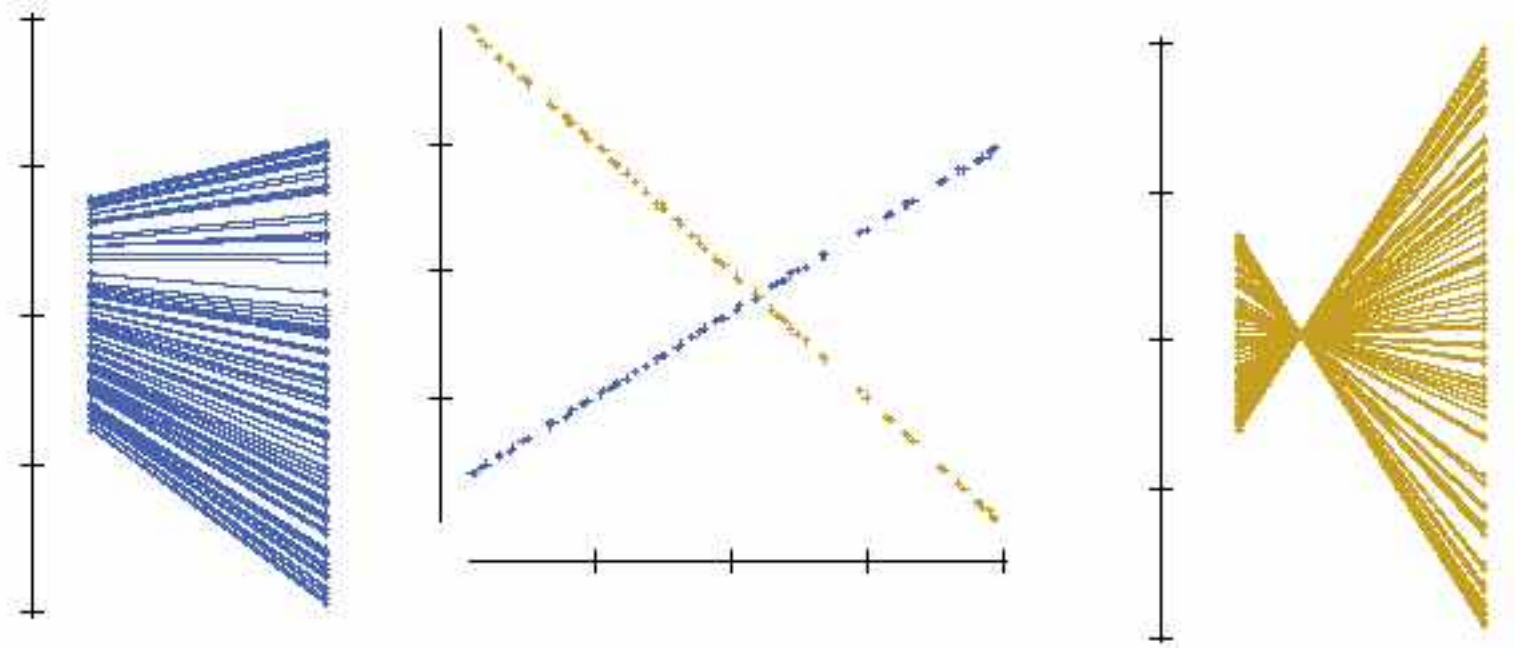
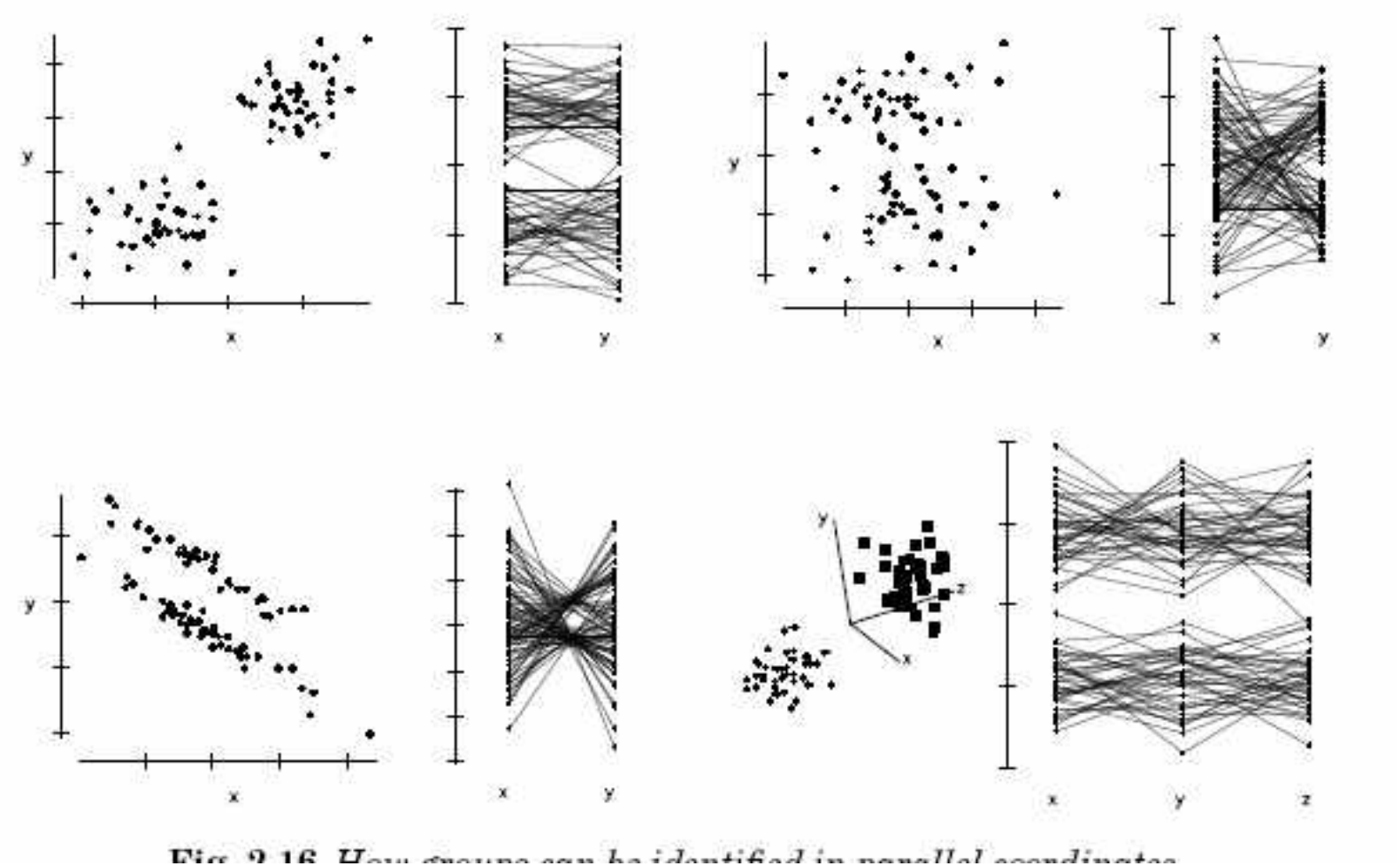


Fig. 2.15. *How lines in two dimensions translate into parallel coordinates.*

Za – Unwin, Theus, Hofmann : Graphics of Large Datasets. Visualizing a Million

Parallel Coordinates – własności geometryczne



Różne rozkłady punktów w układzie kartezjańskich współrzędnych

Za – Unwin, Theus, Hofmann : Graphics of Large Datasets. Visualizing a Million

Parallel Coordinates – przykład ilustracyjny

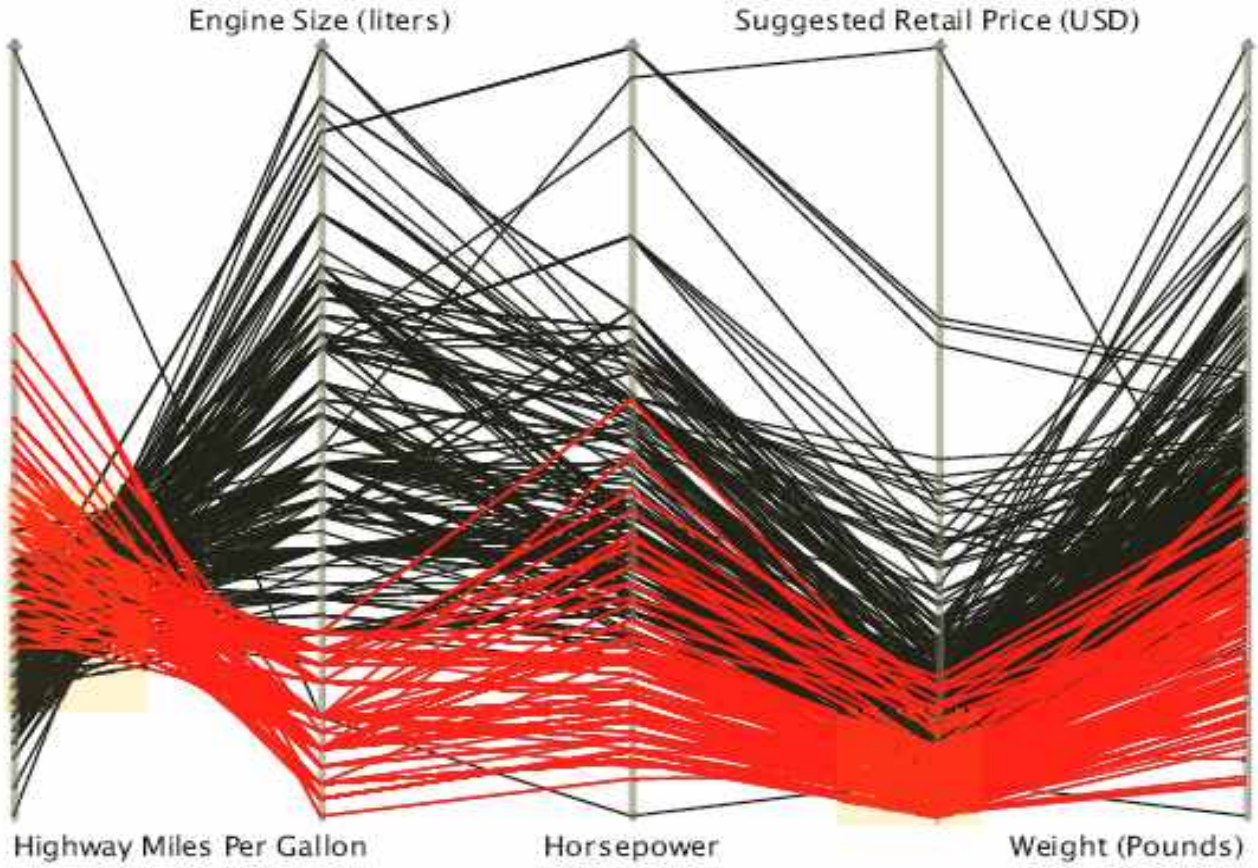


Fig. 2.17. *A parallel coordinate plot for the five variables of the Cars2004 dataset in Figure 2.12. All 4-cylinder cars are highlighted.*

Highlighting -
Skupienie uwagi
na grupie samochodów

Oryginalne dane – Cars2004 pięć wybranych (skorelowanych) zmiennych, np.

- Zmienne o korelacji negatywnej – Mpg i Enginge Size
- Zmienne o korelacji pozytywnej – Retail price i Weight

Za – Unwin, Theus, Hofmann : Graphics of Large Datasets. Visualizing a Million

Parallel Coordinates – Historia

Pierwsze pomysły

- P.Maurice d'Ocagne po raz pierwszy użył terminu „Coordonnees paralleles" w swojej książce 1885 roku
- Henry Gannettes „General Summary, Showing the Rank of States by Ratios" 1880

Systematyczne podejście i popularyzacja

Alfred Inselberg (od lat 60tych poprzedniego wieku)

Rosnące zainteresowania od końca lat 80tych

Developed by
Alfred Inselberg
while at IBM

Przykład wizualizacji Iris Data

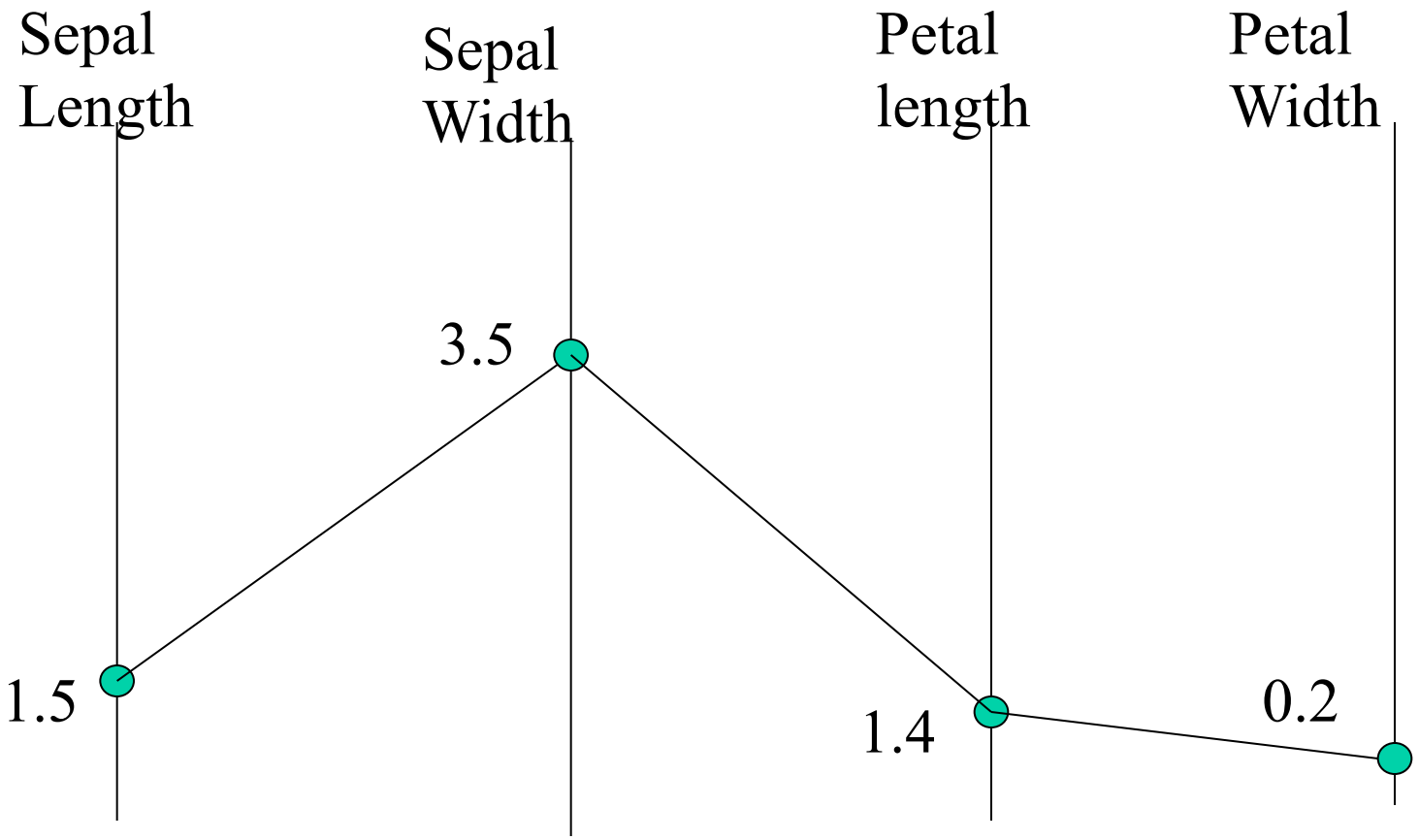
Iris setosa

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

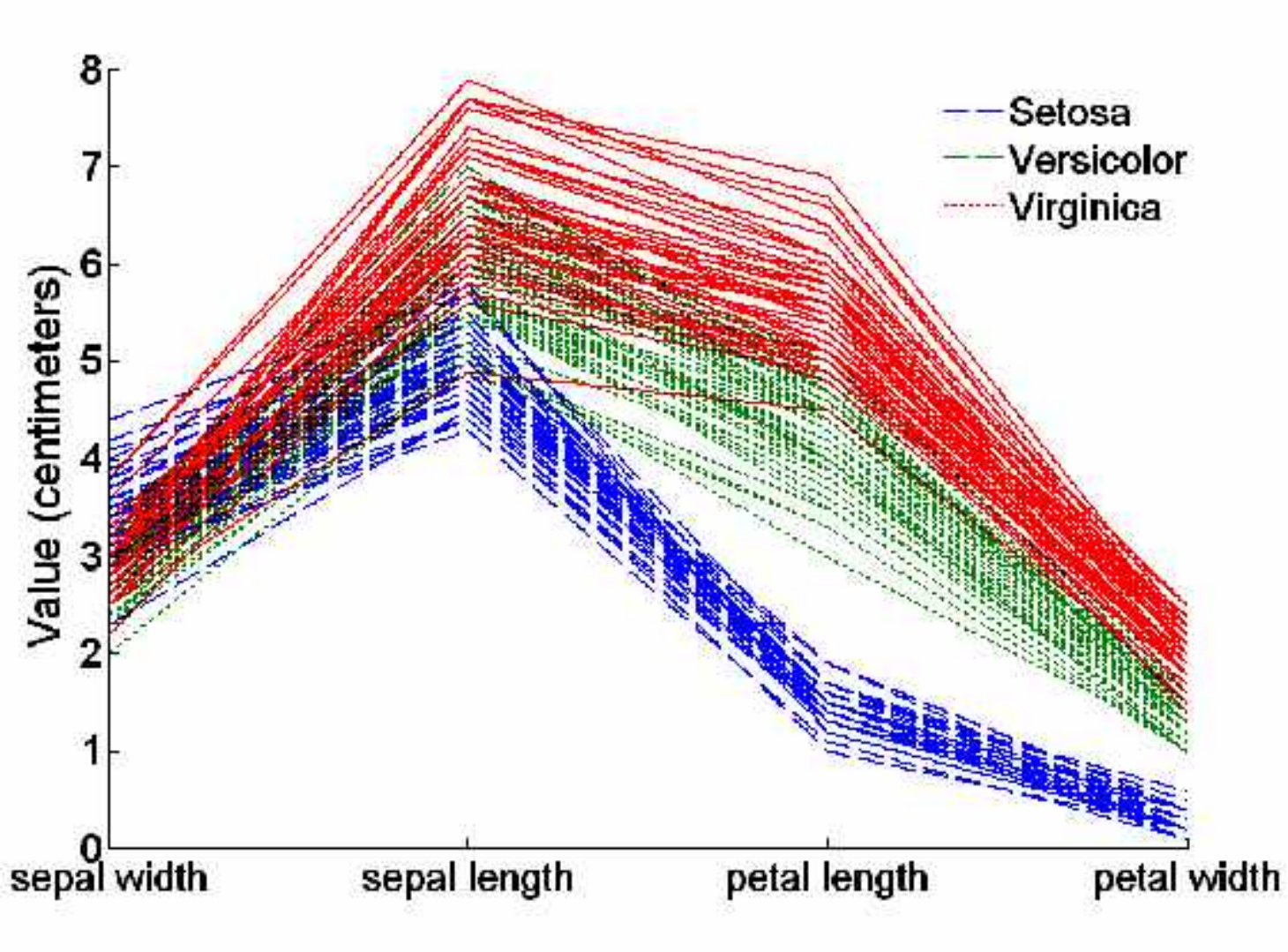
Parallel Coordinates: 4 D (Iris)



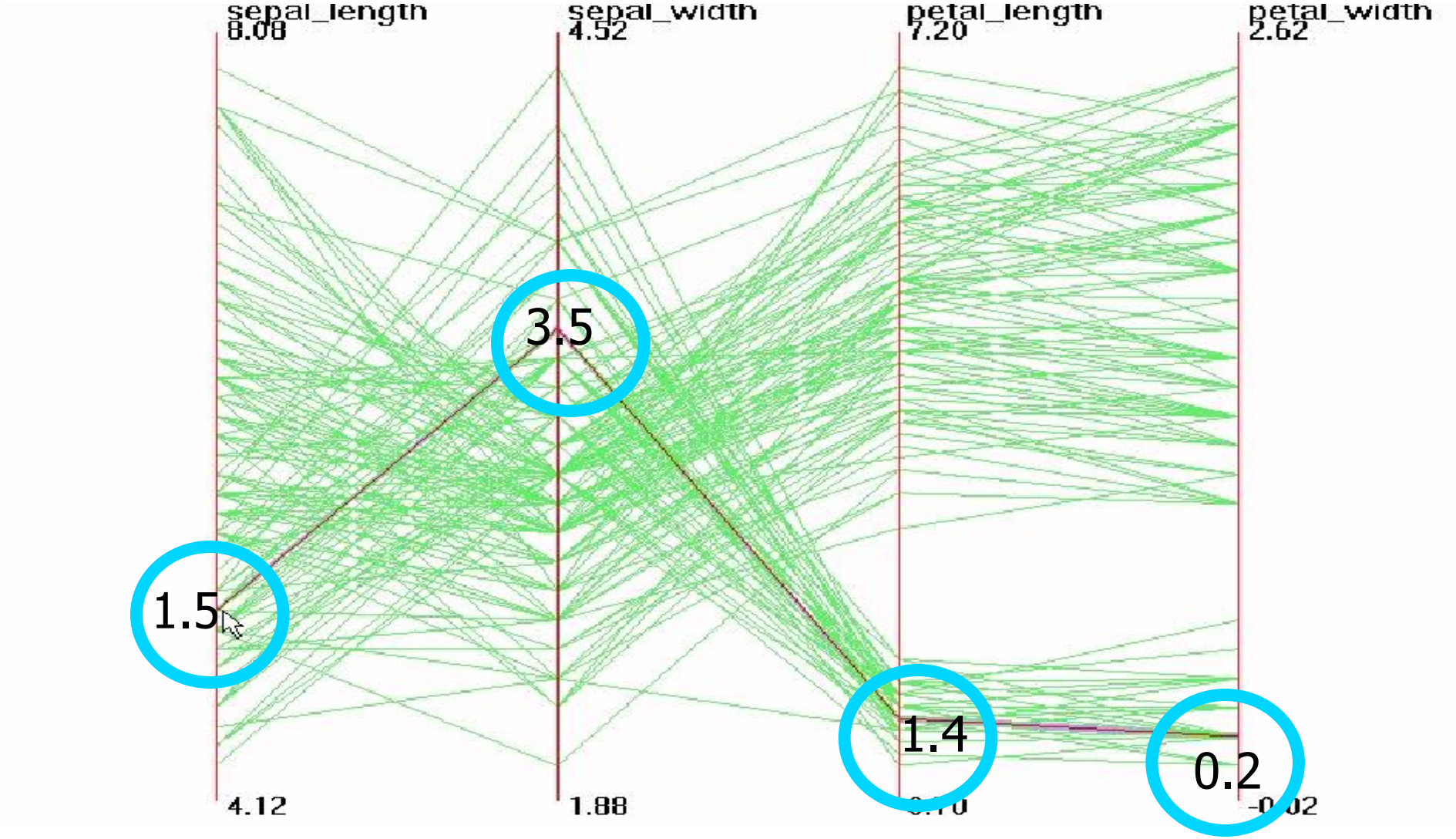
Wspólna skala w cm

sepal length	sepal width	petal length	petal width
1.5	3.5	1.4	0.2

Parallel Coordinates Plots for Iris Data

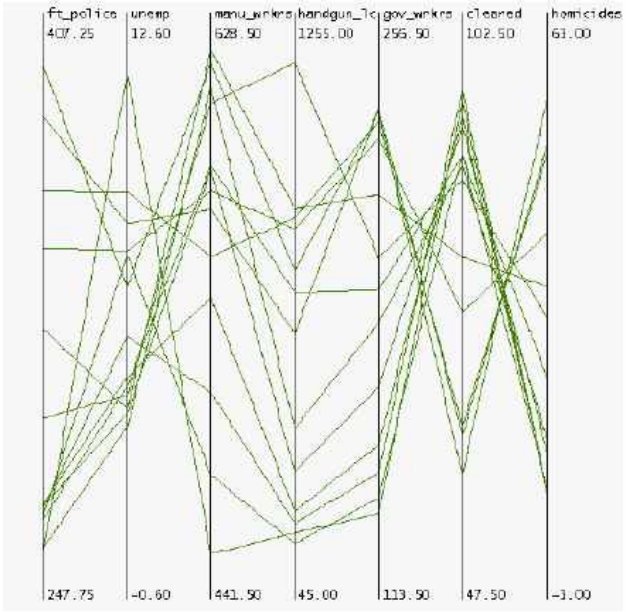


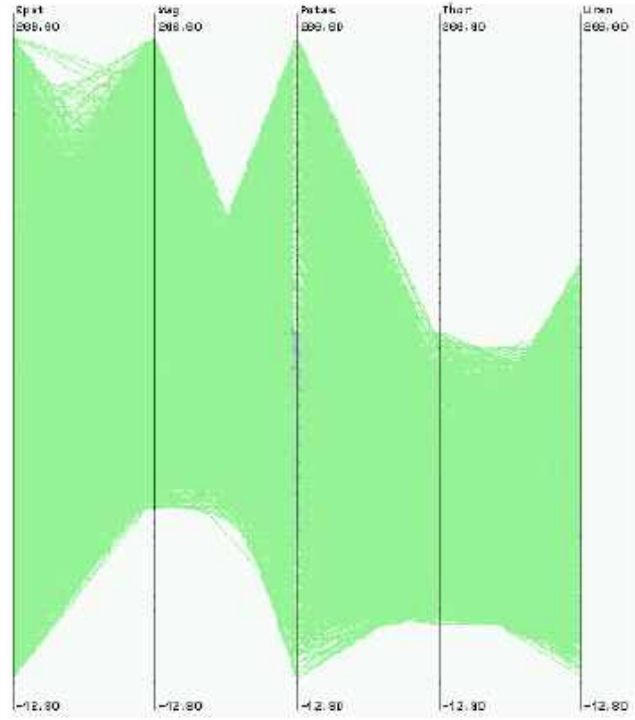
Pojedynczy profil przykładu Iris



Trudności z wizualizacją zbyt dużej liczby przykładów

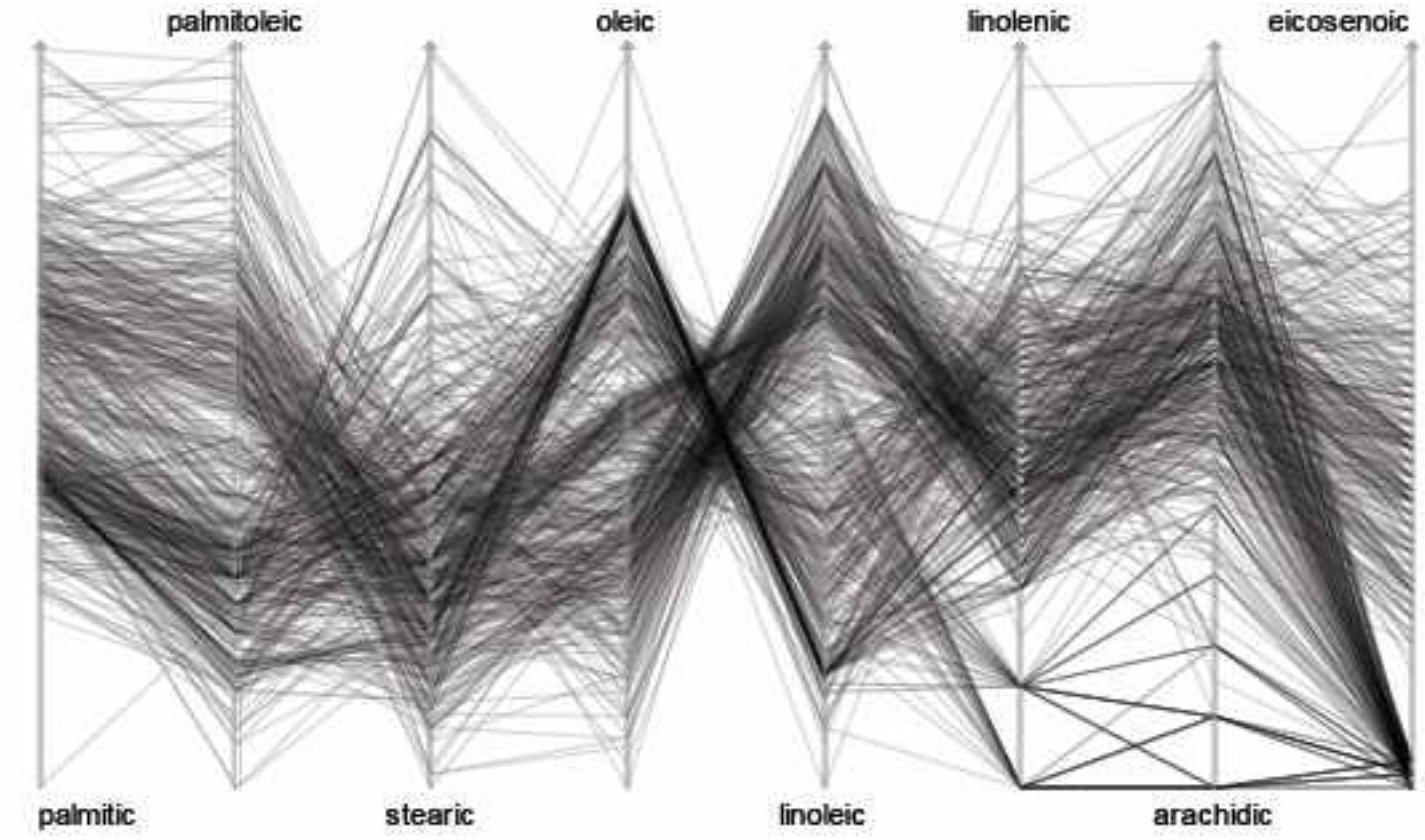
- Nieczytelność zbyt licznych danych
- Parallel co-ordinates (when run out of space)
  - Usable for up to 150 observations
  - Unworkable greater than few hundreds ob.





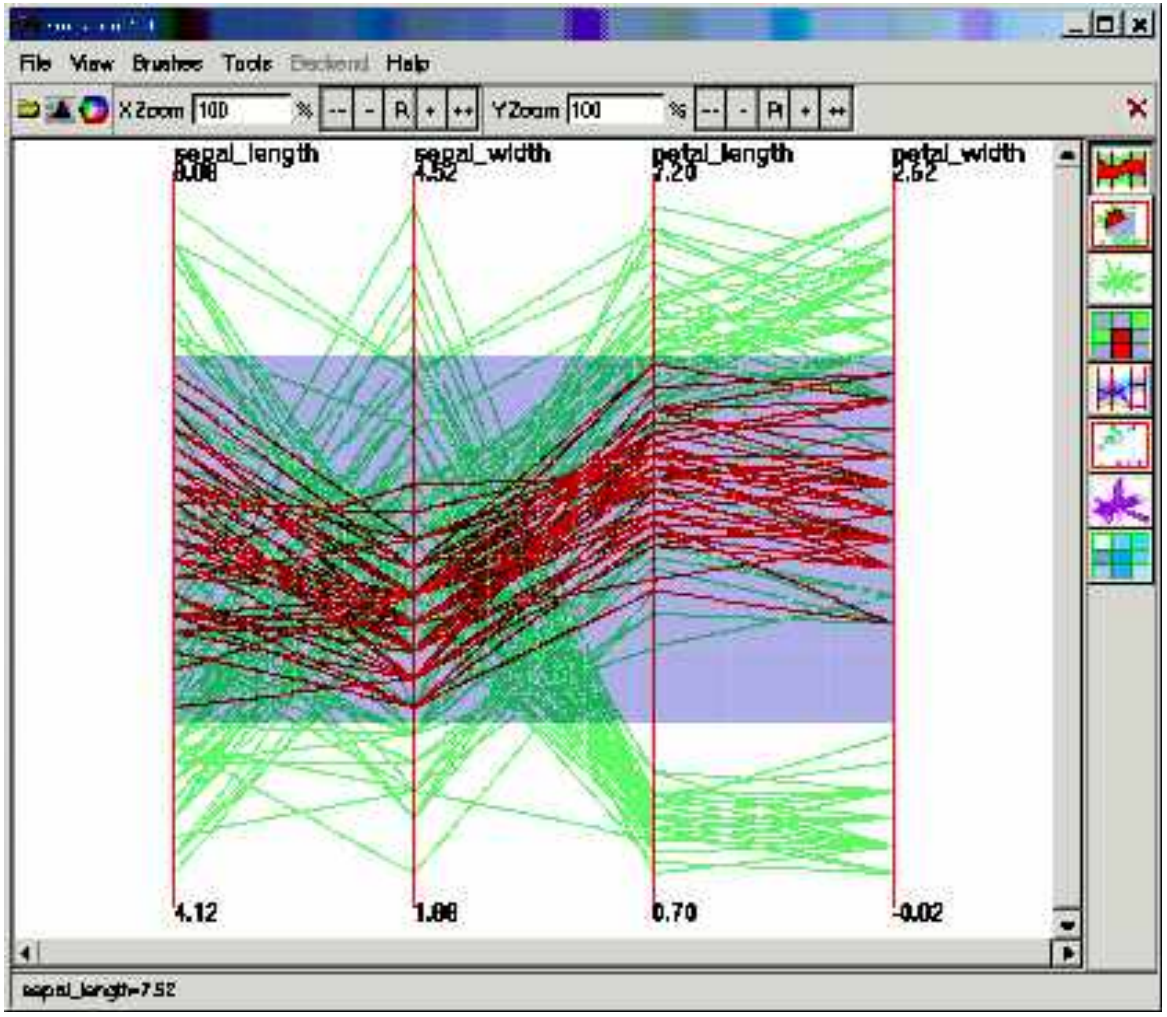
Remote sensing: 5 variables, 16,384 observations)

- Alpha blending

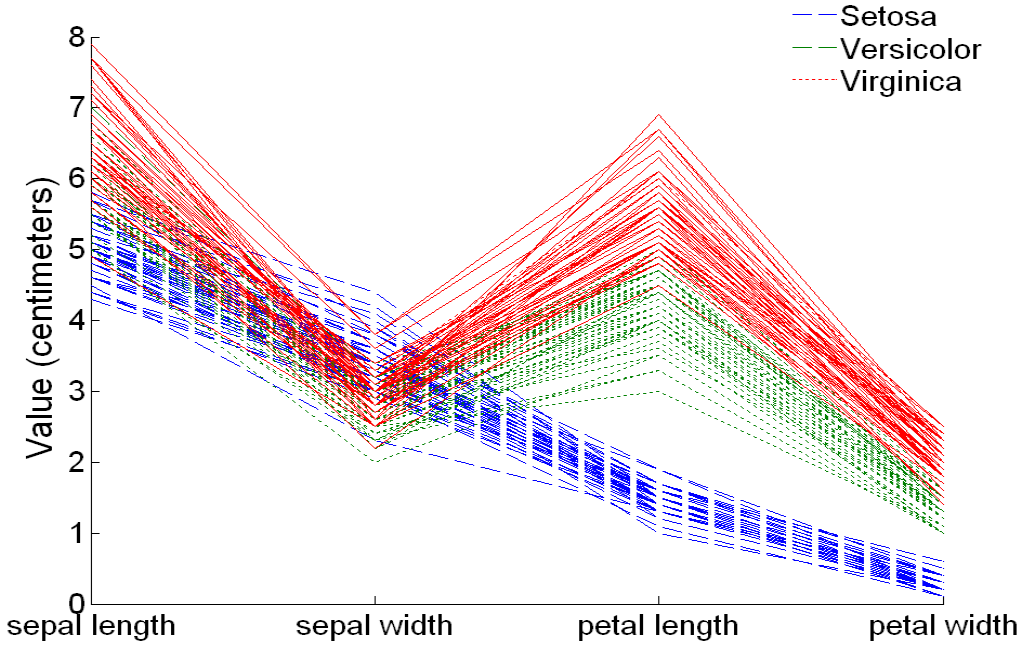


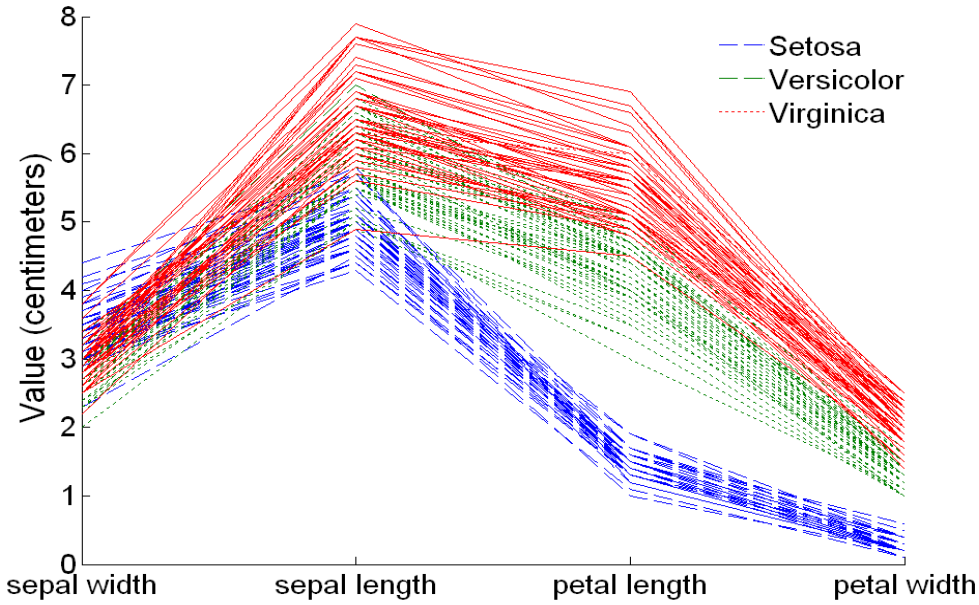
Brushing Tools

- Brushing selects a restricted range of one or more variables
- Selection then highlighted



Sortowanie współrzędnych





Dobór skali współrzędnych

- Stosuje się różne podejścia
- Wpływa na wizualizacje

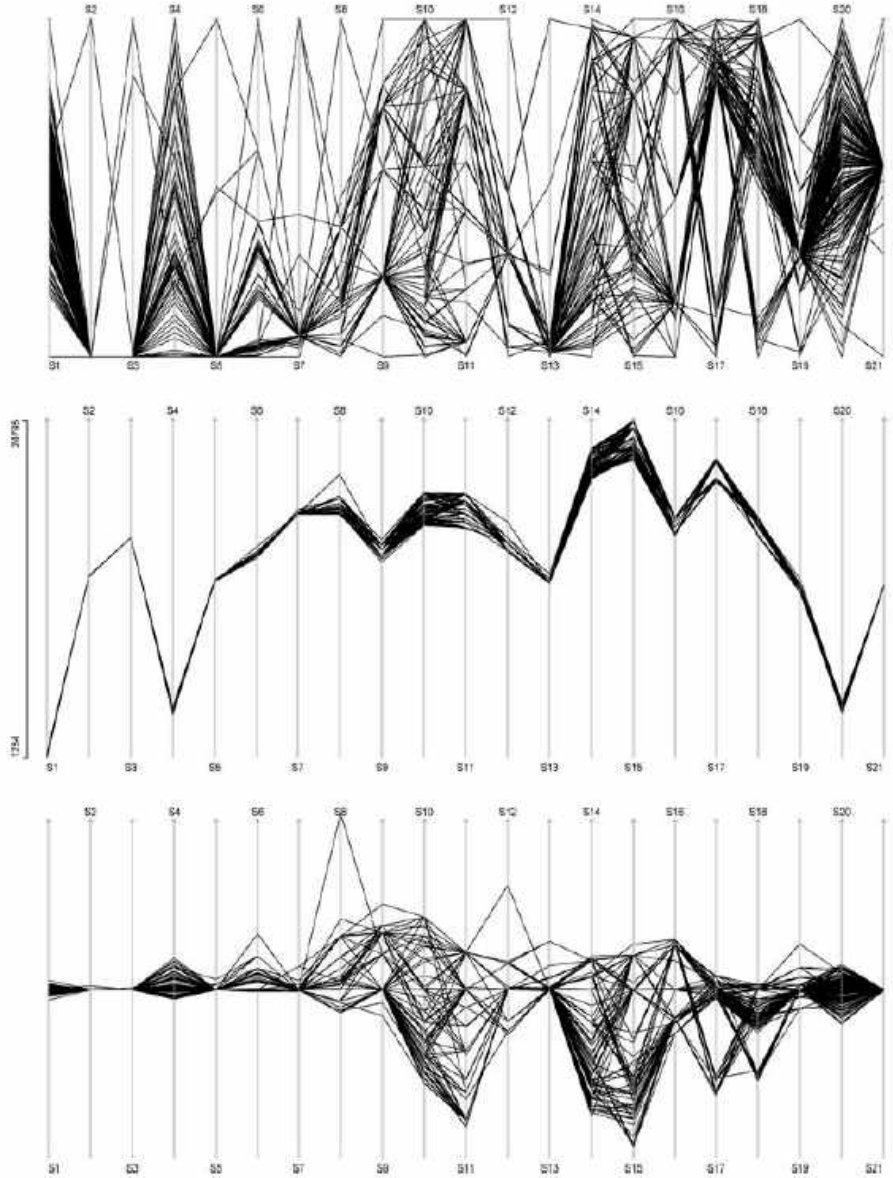
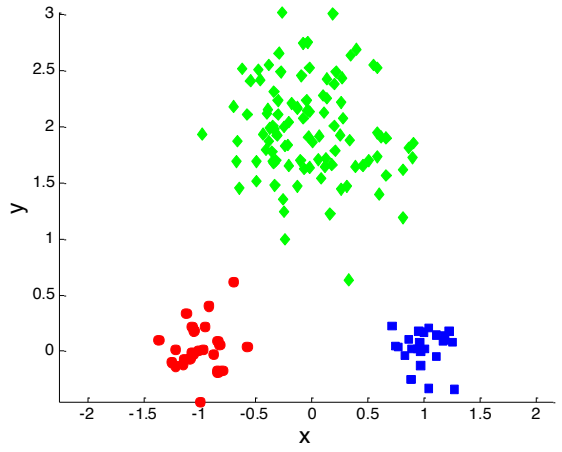
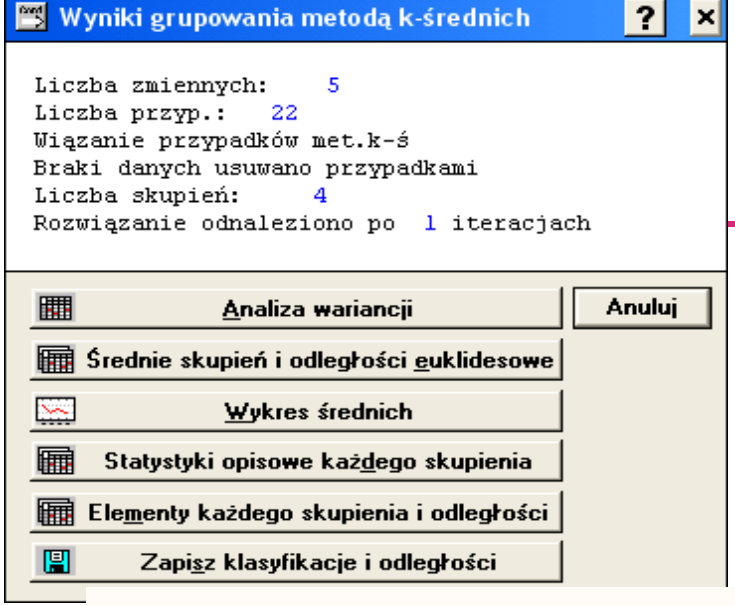


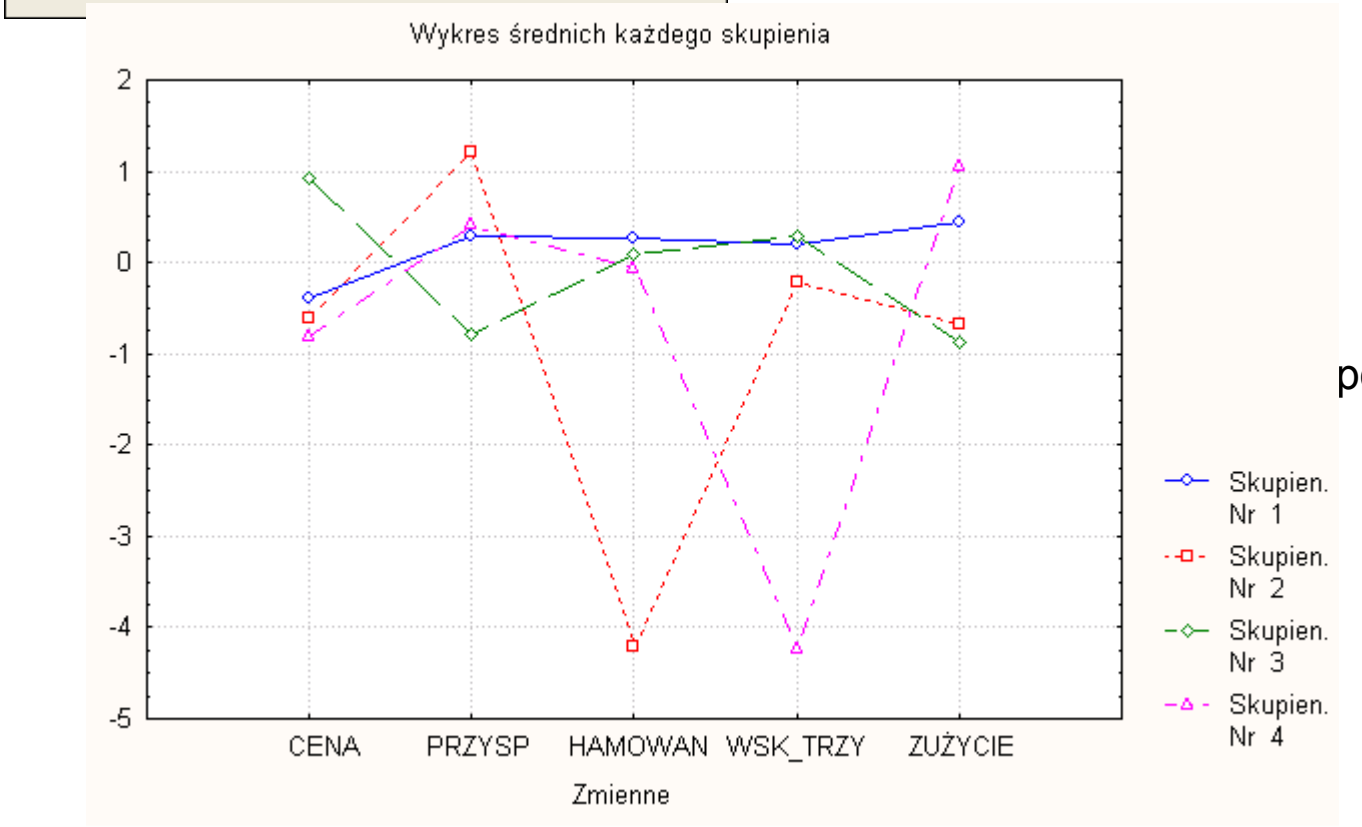
Figure 6.16. Three scaling options for the stage times in the Tour de France 2005: *Top*: all stages are scaled individually between minimum and maximum value of the stage (usual default for parallel coordinate plots) *Middle*: a common scale is used, i.e., the minimum/maximum time of all stages is used as the global minimum/maximum for all axes (this is the only scaling option where a global and common axis can be plotted) *Below*: common scale for all stages, but each stage is aligned at the median value of that stage, i.e., differences are comparable, locations not

PC do reprezentacji skupisk obserwacji

- Skupienie – zbiór obserwacji podobnych do siebie, i równocześnie mniej podobnych do obiektów w innych skupieniach
- Analiza skupień – często stosowane w nienadzorowanej eksploracji danych
  - Wiele algorytmów
  - Przydział obserwacji do skupisk
- Lecz jak opisać utworzone skupiska



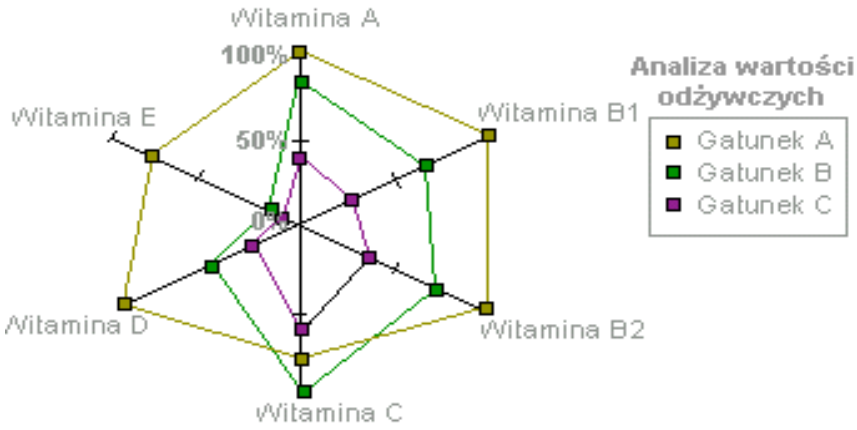




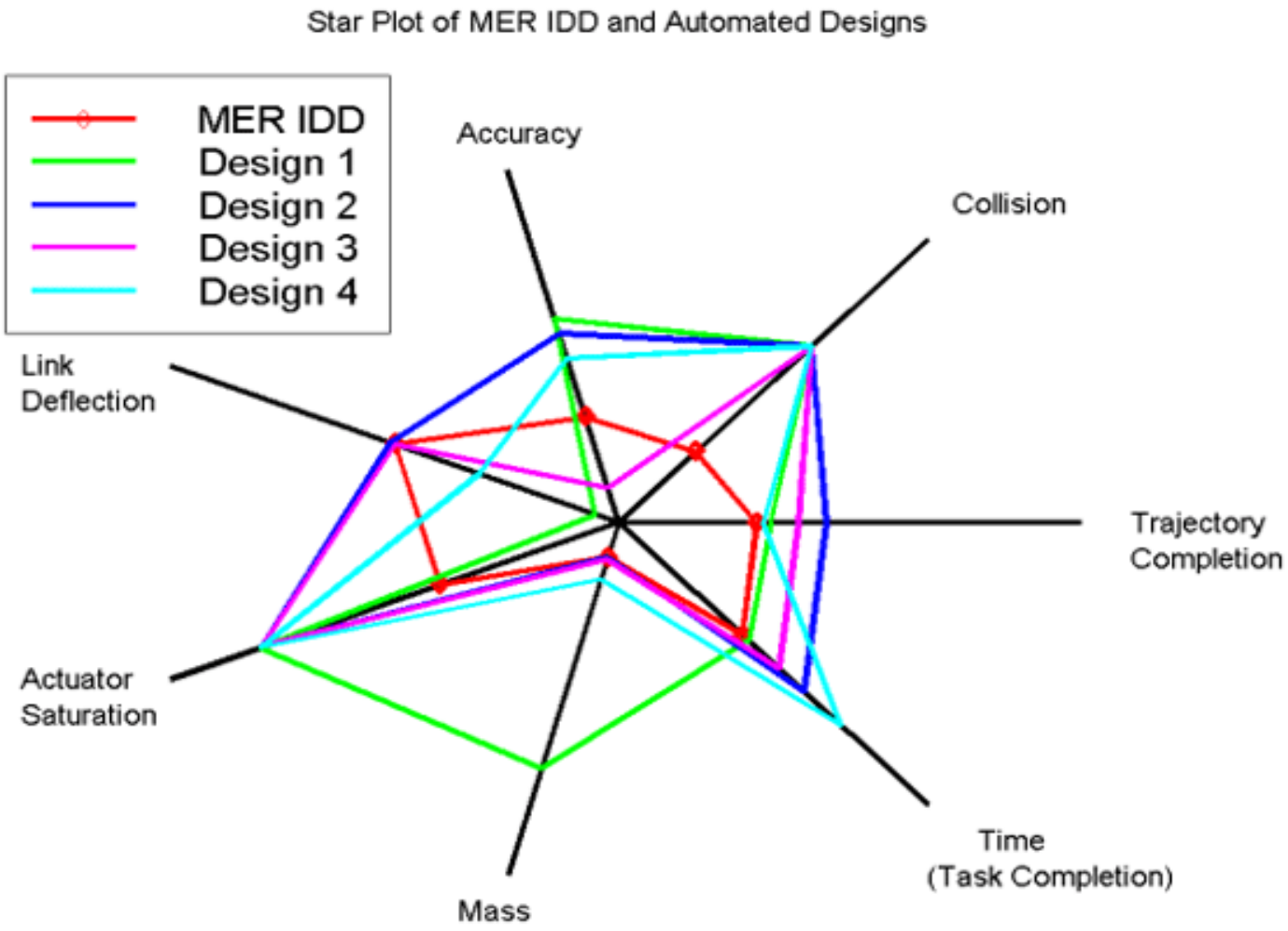
Analiza Skupień
– algorytm k-średnich
Centroid skupiska – jako reprezentant / średni obiekt na podstawie obiektów w skupieniu

Wykresy radarowe

- Porównuje się zagregowane wartości kilku serii danych (obserwacji wielowymiarowych)
- Każda seria danych na wykresie ma unikatowy kolor lub wzór i jest reprezentowana w legendzie wykresu
- Na wykresie można wykreślić jedną lub kilka serii danych



Rader plot from NASA, with some of the most desirable design results



Glyph Techniques

- Odwzoruj wartości danych to prostych elementów geometrycznych, tzw. glif
- Sprawdź definicje “a glyph”
- Zaproponowano wiele podstawowych glifów
  - Star glyphs
  - Faces
  - Arrows
  - Sticks
  - Shape coding

Angielski

glyph | glɪf |

noun

1 a hieroglyphic character or symbol. *flanges painted with esoteric glyphs.*

- a sculptured symbol (e.g. as forming the ancient Mayan writing system). *these glyphs refer to an ancient Olmec ruler.*
- Computing a small graphic symbol.

2 Architecture an ornamental carved groove or channel, as on a Greek frieze.

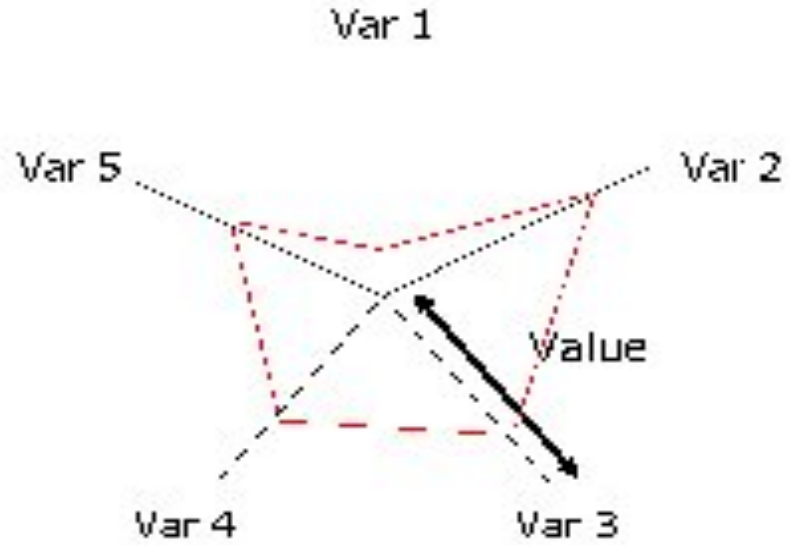
Glyph Layouts

- How do we place the glyphs on a chart?
- Sometimes there will be a natural location – for example?

- If not… two of the varieties can be allocated to spatial position, and the remainder to the attributes of the glyph

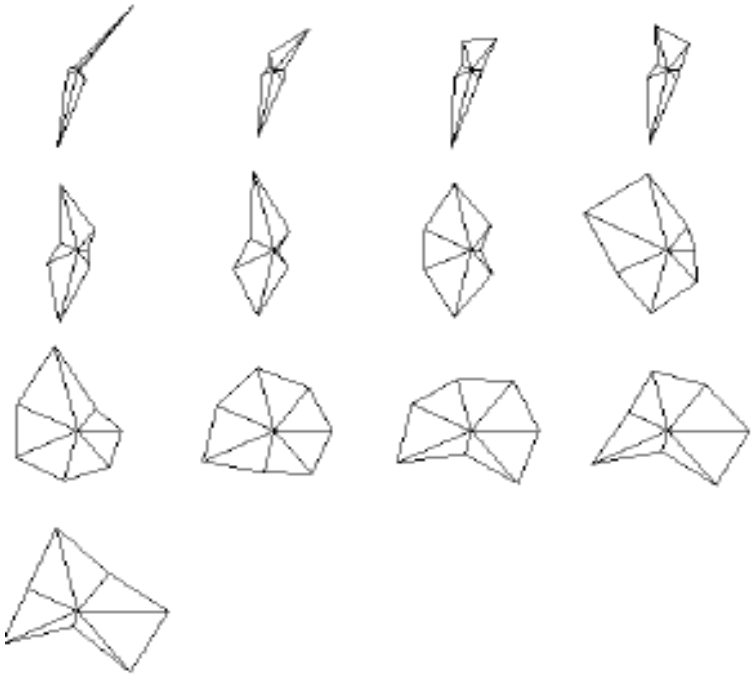
Glyph Techniques – Star Plots

- Each observation represented as a ‘star’
- Each spike represents a variable
- Length of spike indicates the value

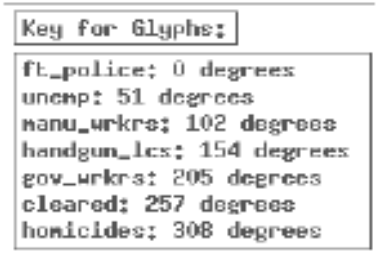


Glyph Techniques – Star Plots

- Each observation represented as a ‘star’
- Each spike represents a variable
- Length of spike indicates the value



Crime in Detroit



Comparing several cars

The variable list for the sample star plot is:

Price

Mileage (MPG)

1978 Repair Record (1 = Worst, 5 = Best)

1977 Repair Record (1 = Worst, 5 = Best)

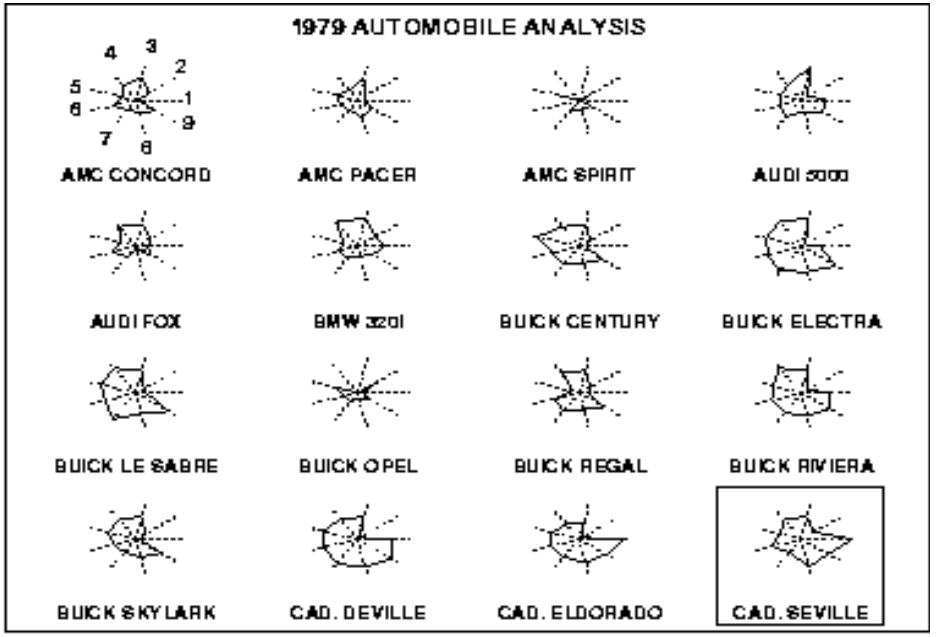
Headroom

Rear Seat Room

Trunk Space

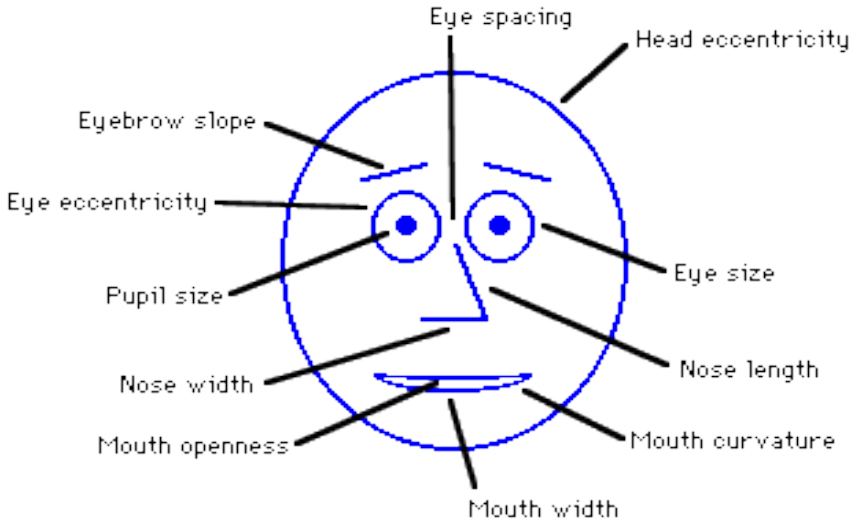
Weight

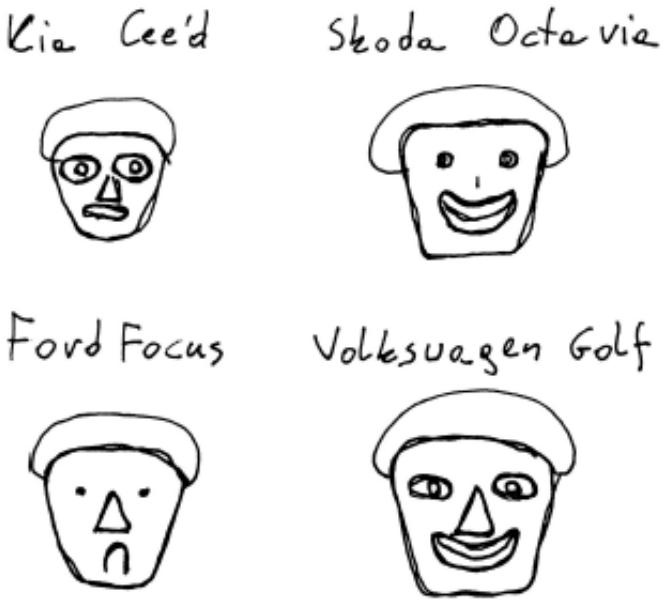
Length



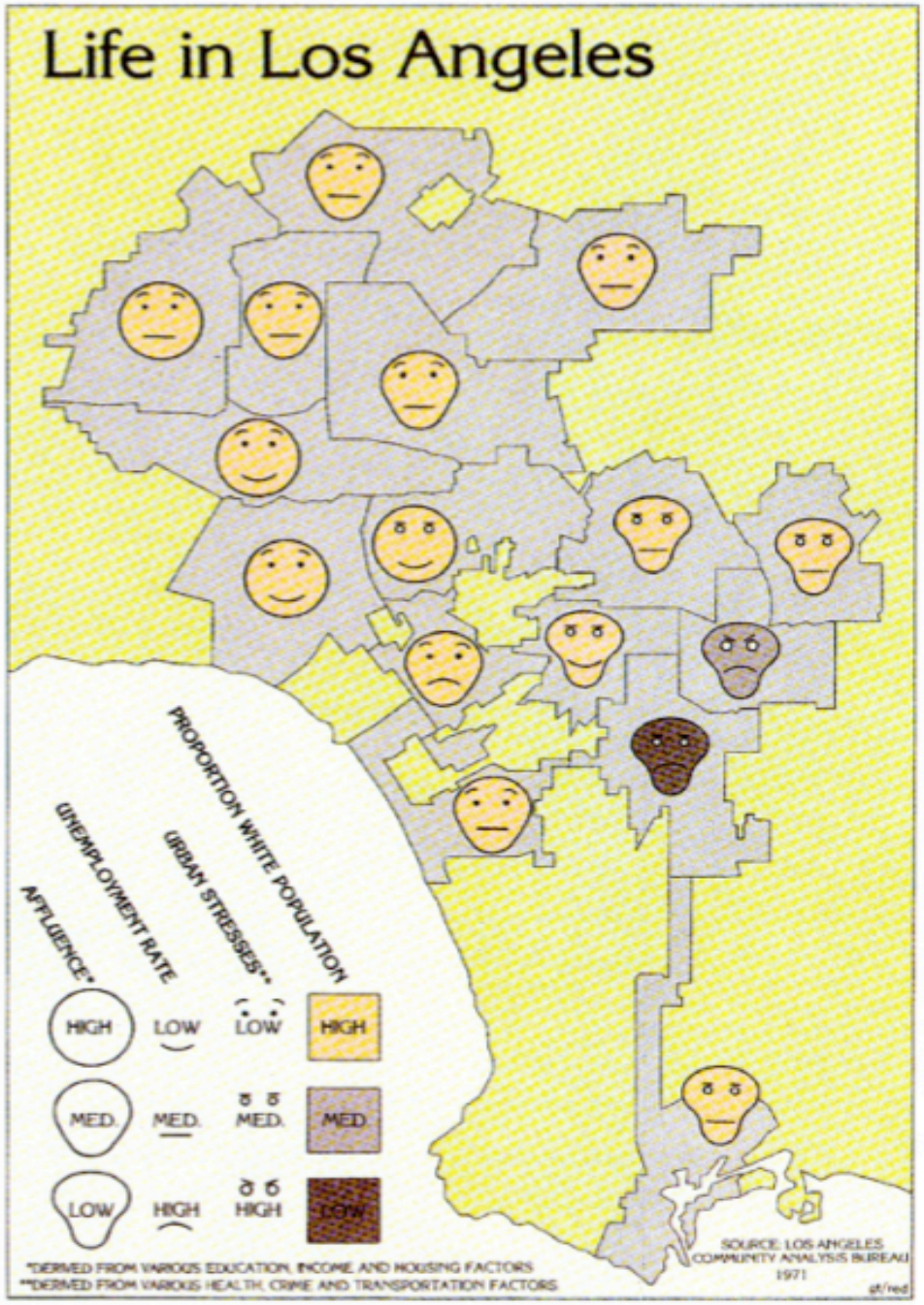
Chernoff Faces

Zakoduj zmienne jako charakterystyczne elementy ludzkiej twarzy





Cute applets: `http://www.cs.uchicago.edu/~wiseman/chernoff/`
`http://hesketh.com/schampeo/projects/Faces/chernoff.html`



http://kspark.kaist.ac.kr/Human%20Engineering.files/Chernoff/Chernoff%20Faces.htm

Chernoff's Face

- .. And here is Chernoff's face ☺

http://www.fas.harvard.edu/~stats/People/Faculty/Herman_Chernoff/Herman_Chernoff_Index.html

Bardzo dziękuję za uwagę i obecność na wykładzie!