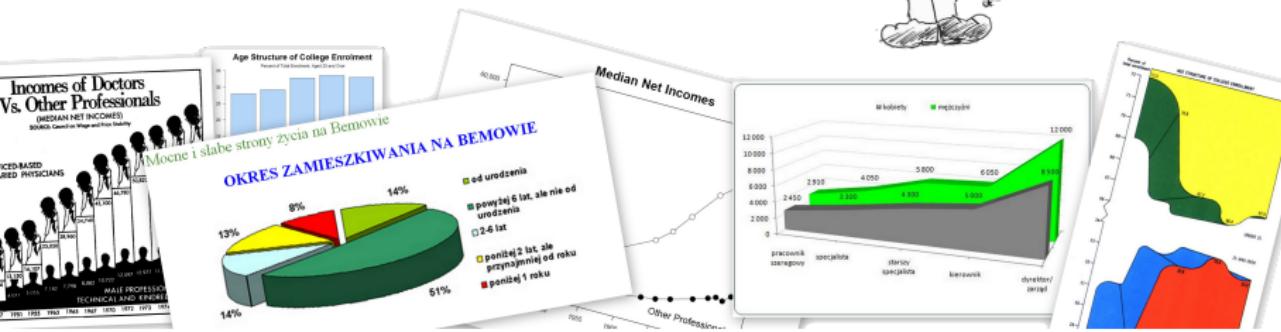
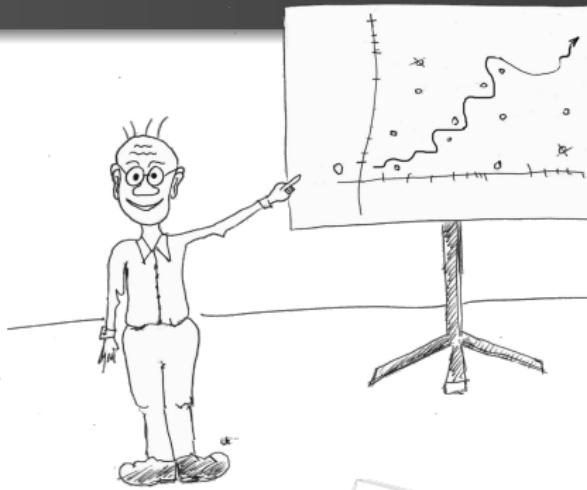


# Wizualizacja danych

Przemysław.Biecek@gmail.com, MIM UW / SmarterPoland

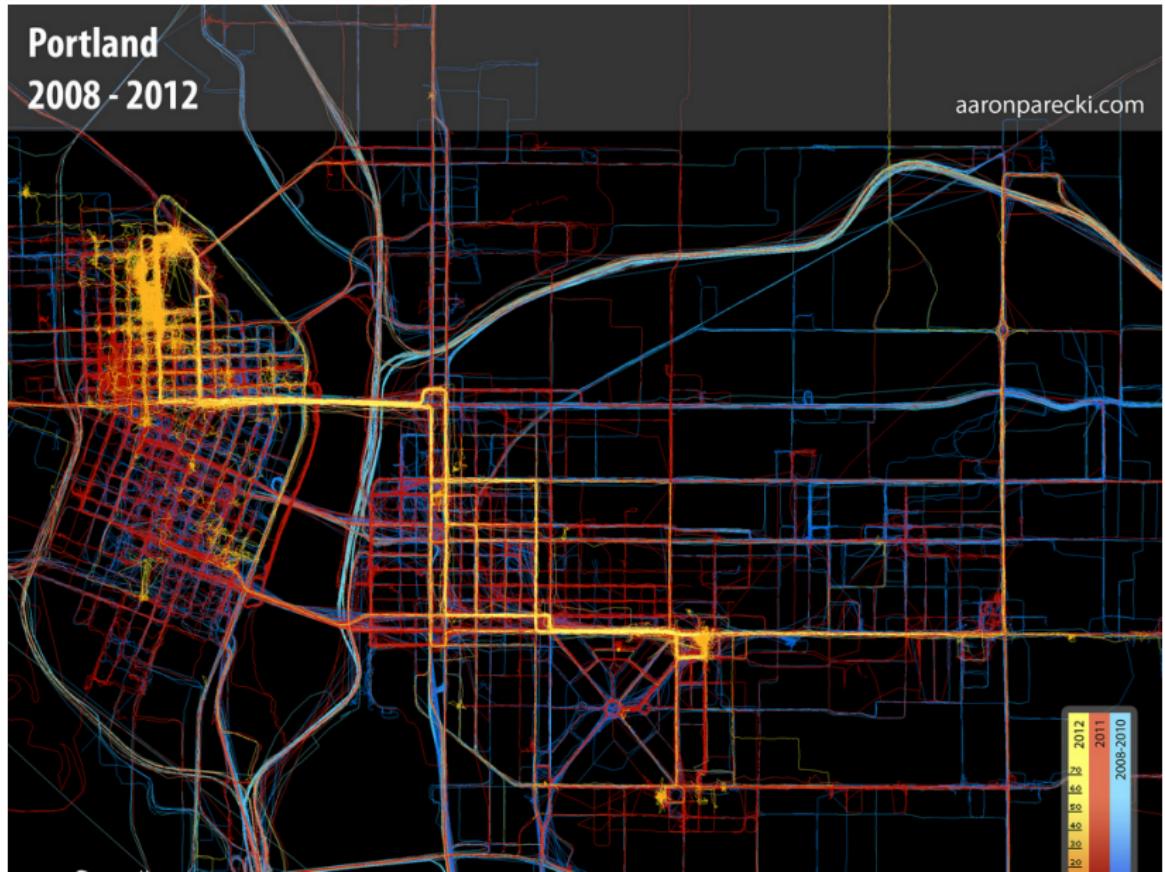
WUM, 15 kwietnia 2013

- 1 Przykładowe „bogate” wizualizacje.
- 2 Ładniej nie znaczy lepiej, czyli typowe oszustwa percepcji.
- 3 Po pierwsze „nie kłamać”, czyli co mierzy „Lie factor”.
- 4 Dobre rady wujka Edwarda Tufte'go.



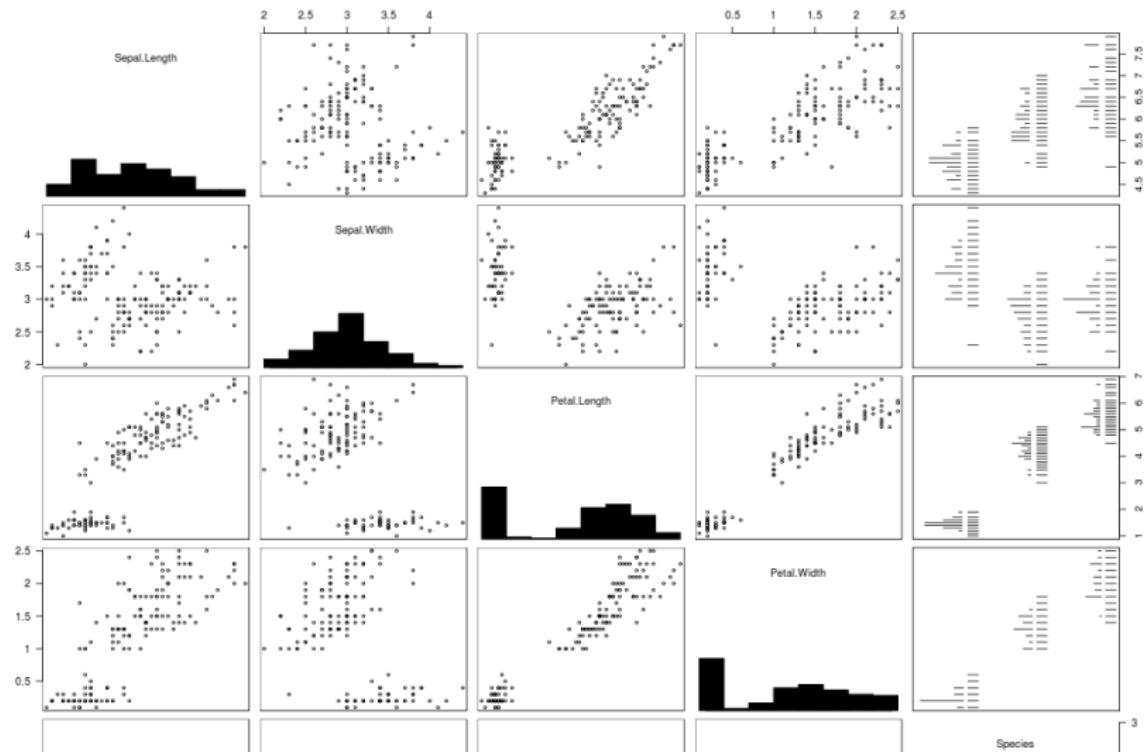
# Po co nam wizualizacja? - Artystyczne wizualizacje

Z serwisu <http://www.flowingdata.com>



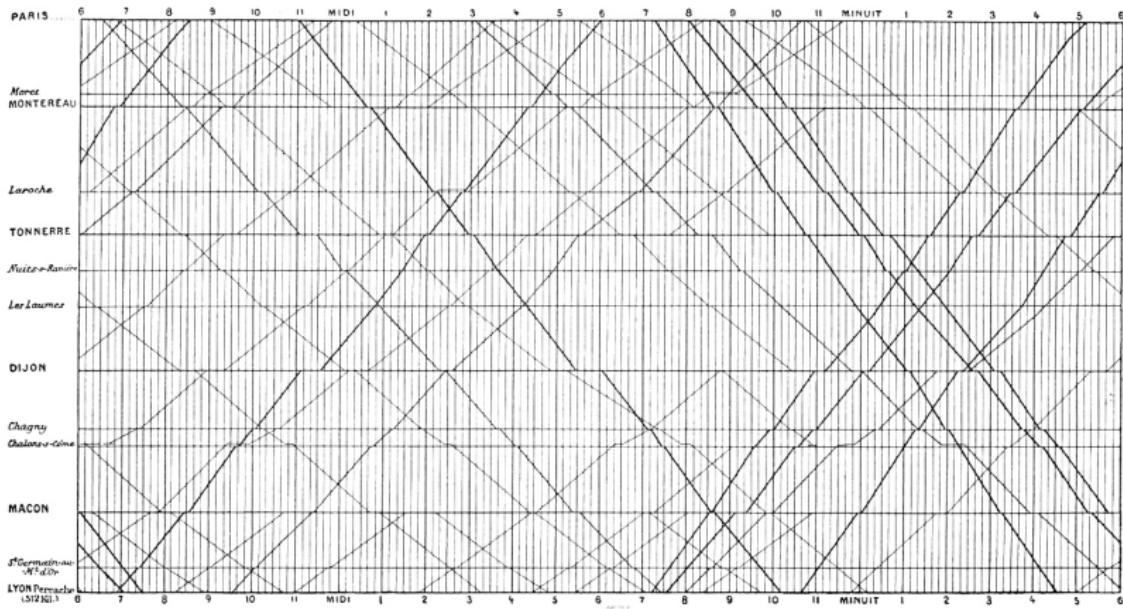
# Po co nam wizualizacja? - Eksploracja danych

Z dokumentacji dla funkcji gpairs()



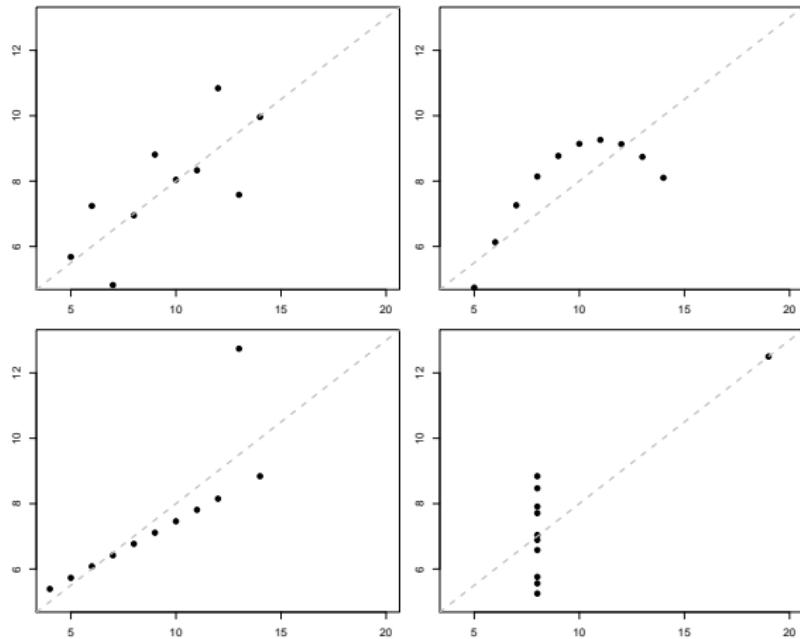
# Po co nam wizualizacja? - Prezentacja informacji

Train Schedule by E.J. Marey, <http://c82.net/posts.php?id=66>



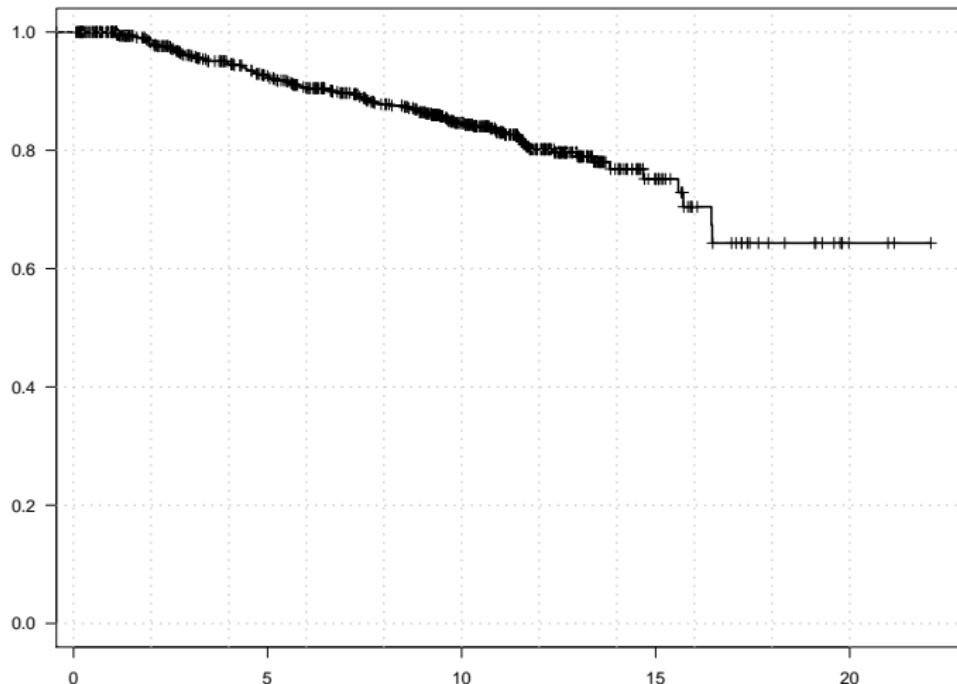
# Wizualizacja - jest potrzebna

Oglądając dane często łatwiej nam zrozumieć charakter zależności niż patrząc na surowe liczby.



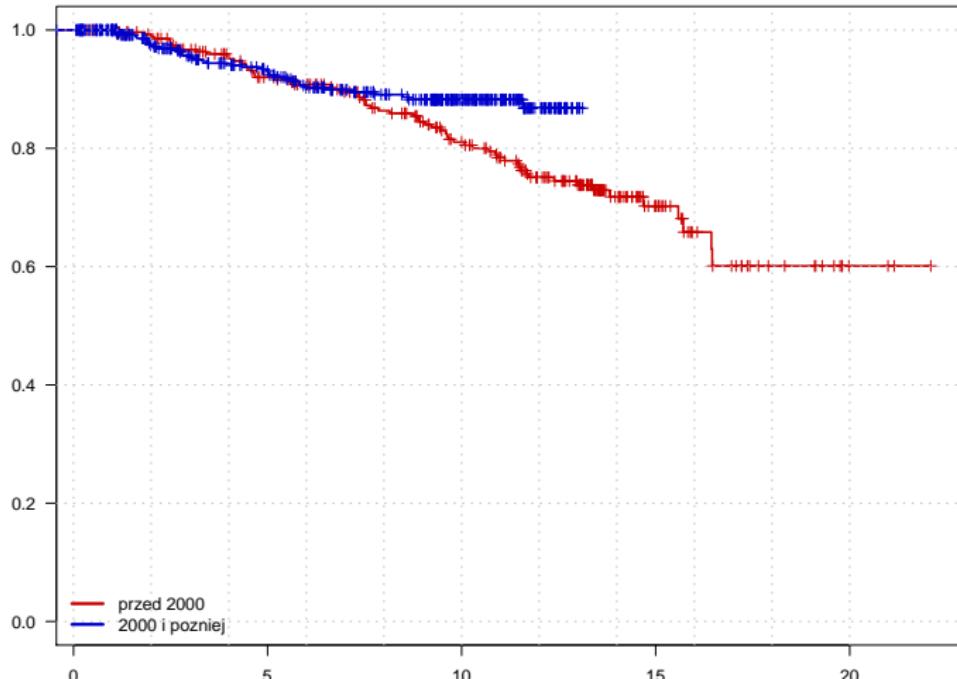
# Interesujące prezentacje złożonych informacji

Krzywa przeżycia przedstawia zarówno informację o dynamice zgonów jak i informację o czasach cenzorowania.



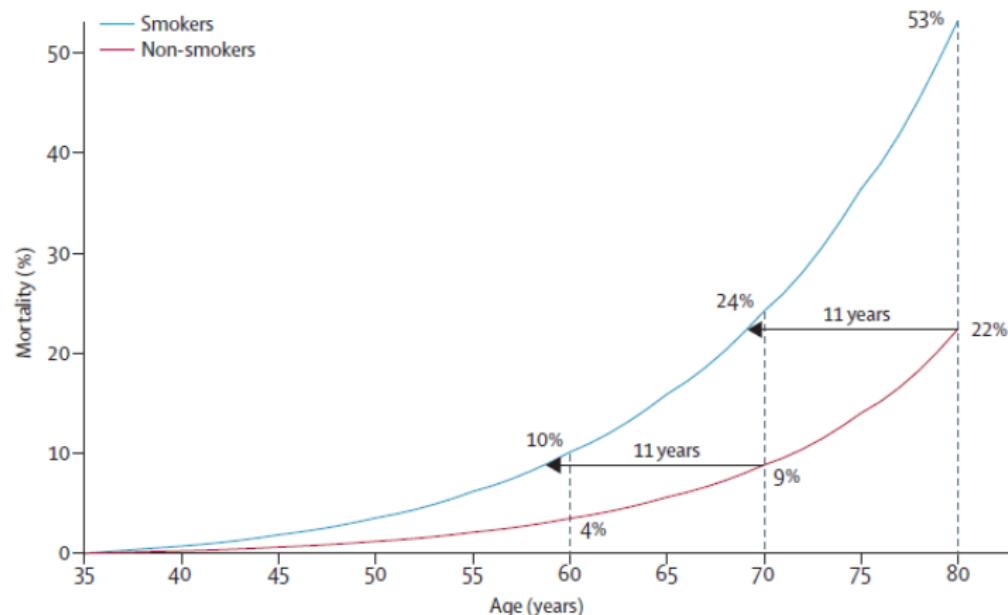
# Interesujące prezentacje złożonych informacji

Krzywą przeżycia można przedstawiać w podziale na grupy. Często z dobrego przedstawienia danych możemy dowiedzieć się o danym zjawisku więcej niż z samych liczb.



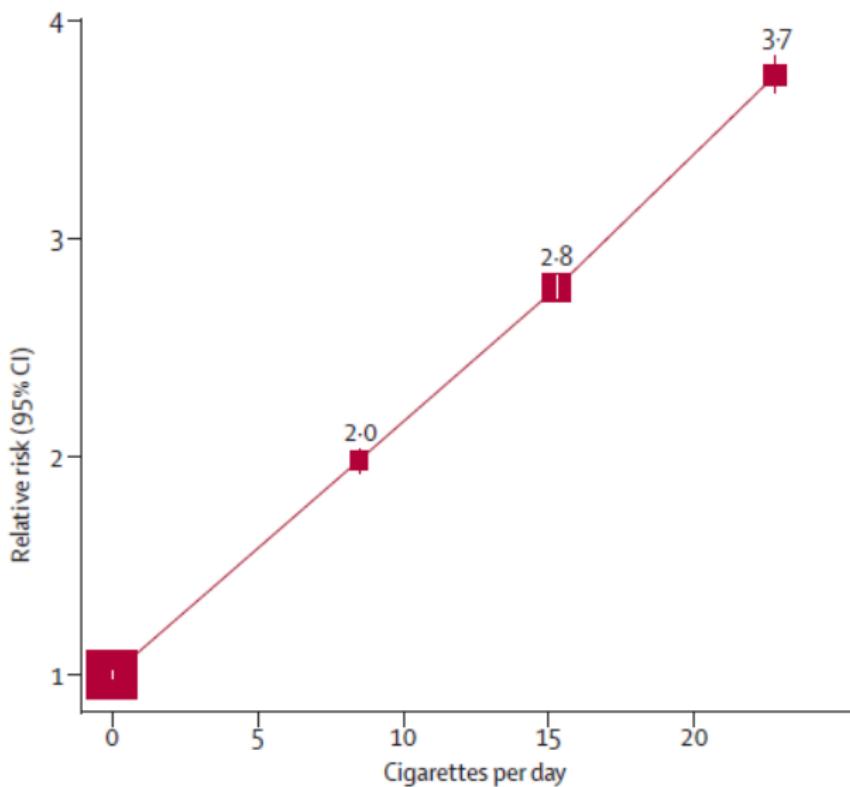
# Krzywa przeżycia inaczej.

The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK



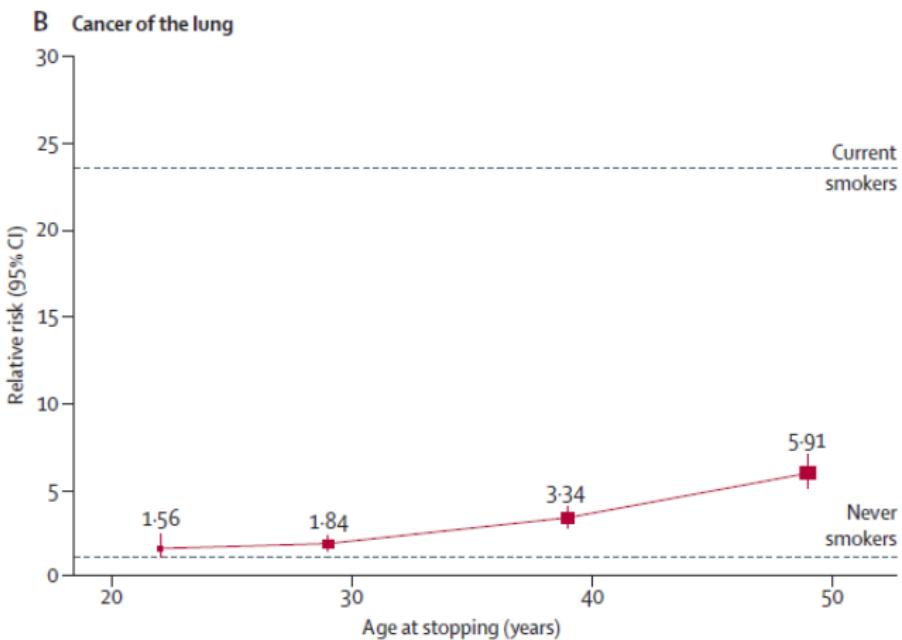
[źródło The 21st century hazards of smoking and benefits of stopping, K. Pirie et all, Lancet 27 X 2012  
[http://dx.doi.org/10.1016/S0140-6736\(12\)61720-6](http://dx.doi.org/10.1016/S0140-6736(12)61720-6)

# Względne ryzyko nowotworu a palenie



[źródło The 21st century hazards of smoking and benefits of stopping, K. Pirie et all, Lancet 27 X 2012  
[http://dx.doi.org/10.1016/S0140-6736\(12\)61720-6](http://dx.doi.org/10.1016/S0140-6736(12)61720-6)]

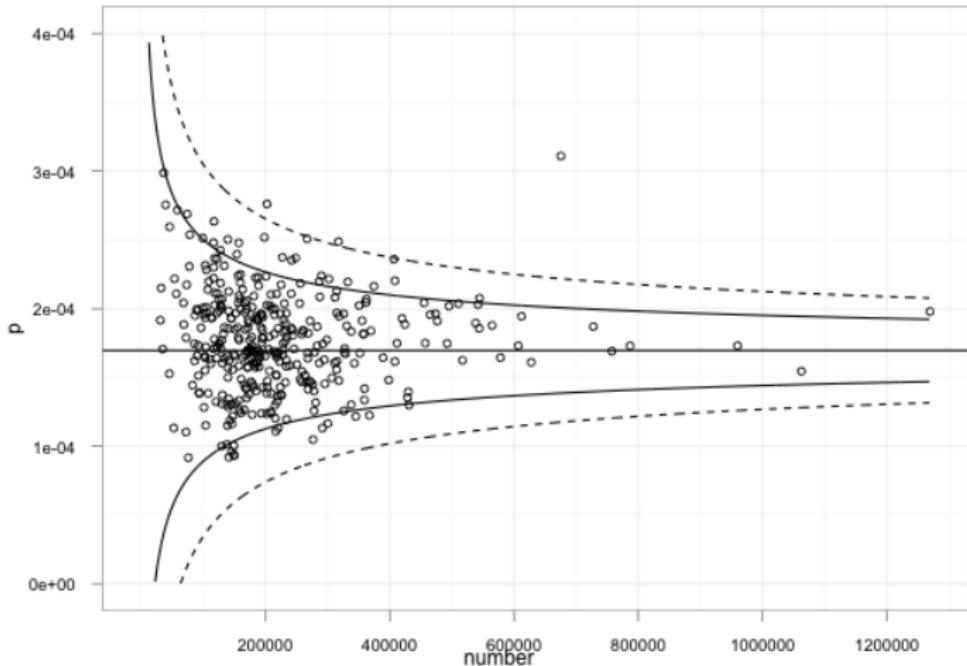
# Względne ryzyko nowotworu a palenie



Age at stopping (mean), years	<25 (22)	25-34 (29)	35-44 (39)	45-54 (49)
Number of deaths	24	86	159	243
RR (95% CI)	1.56 (1.03-2.37)	1.84 (1.45-2.34)	3.34 (2.76-4.03)	5.91 (5.01-6.97)

[źródło The 21st century hazards of smoking and benefits of stopping, K. Pirie et al, Lancet 27 X 2012  
[http://dx.doi.org/10.1016/S0140-6736\(12\)61720-6](http://dx.doi.org/10.1016/S0140-6736(12)61720-6)]

# Wykres tunelowy wykorzystywany w meta-analizie



# Po co nam wizualizacja?

Czy poniższy obrazek mówi coś więcej niż dwie liczby?



# Prezentacja informacji

Celem komunikacji z użyciem wizualizacji jest: przedstawienie wartości liczbowych lub zależności obecnych w danych, za pomocą graficznych wzorców.

Ważny jest **czas komunikacji** oraz **dokładność/precyzja** z jaką przedstawiamy dane.

Walory estetyczne mają drugorzędne znaczenie.

Hipokrates: „po pierwsze nie szkodzić” .

Uwaga na iluzje optyczne i przekłamania, powodowane przez kolory (czerwony - efekt „uwypuklenia”), skale szarości (duże obszary wyglądają na ciemniejsze), rozmieszczenie porównywanych obiektów (obiekty o różnej wielkości im są sobie bliżej tym są bardziej różne).

# Prezentacja informacji

Percepcja wzorców obecnych na wykresie odbywa się w trzech krokach (zobacz też: Statistical presentation graphics, Frank Harrell)

- identyfikacja - jakie geometrie kodują prezentowane wartości (kąty, powierzchnie, długości),
- grupowanie - zestawienie listy obiektów prezentujących dane wartości,
- ocena / szacowanie różnic, porządku lub względnych proporcji pomiędzy przedstawianymi geometriami.

Jeżeli wizualizacja nie jest przemyślana to w każdym z tych kroków coś może pójść źle.

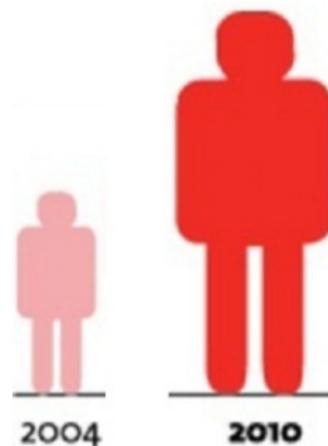
# Identyfikacja geometrii

Przykład z <http://biznes.interia.pl/raport/emerytury/news/reforma-reformy-emerytalnej-bo-dane-sie-zdezaktualizowaly>'

Poniższy rysunek porównuje dwie liczby: liczby polaków na emigracji w 2004 i 2010 roku.

Czy można z niego odczytać o ile % liczba polaków na emigracji wyrosła?

Która właściwość „ludzika” przedstawia liczbę osób na emigracji?



# Identyfikacja geometrii

Przykład z <http://biznes.interia.pl/raport/emerytury/news/reforma-reformy-emerytalnej-bo-dane-sie-zdezaktualizowaly>'

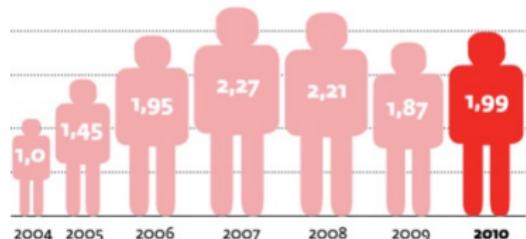
## Pole czy wysokość?

Nasza percepcja lepiej ocenia różnice w wysokości. Intuicyjnie jednak te obrazki różnią się polami.



Liczba emigrantów przebywających poza granicami Polski

Dane w miln



źródło: GUS

[Z serwisu interia.pl, usunięto liczby i linie poziome siatki  
<http://biznes.interia.pl/news/reforma-reformy-emerytalnej-bo-dane-sie-zdezaktualizowaly,1771747,4265>]

# Grupowanie obiektów

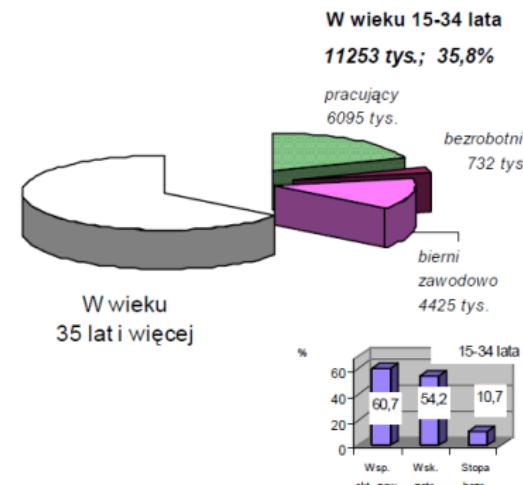
Przykład z raportu „Wejście ludzi młodych na rynek pracy”, GUS 2010

Na dobrym wykresie powinno być oczywiste, który element odpowiada jakiej grupie.

## Aktywność ekonomiczna<sup>1</sup> ludności w wieku 15-34 lata

Zbiorowość osób młodych, zdefiniowanych tu jako ludność w wieku 15-34 lata, liczyła w II kwartale 2009 r. 11253 tys., co stanowiło nieco ponad 1/3 globalnych zasobów w pracy (w wieku 15 lat i więcej). W badanym okresie nieco częściej niż **co druga młoda osoba pracowała** – 6095 tys. (wskaźnik zatrudnienia – 54,2%), a kryteria **bezrobotnego** wg MOP spełniała częściej niż co piętnasta - **735 tys.**. Łącznie, trzy z pięciu młodych osób były aktywne zawodowo, tj. pracowały lub aktywnie poszukiwały pracy (współczynnik aktywności zawodowej - 60,7%).

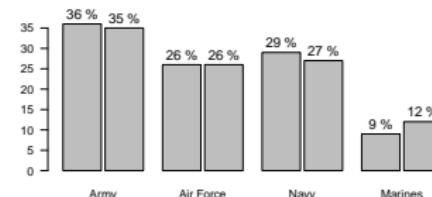
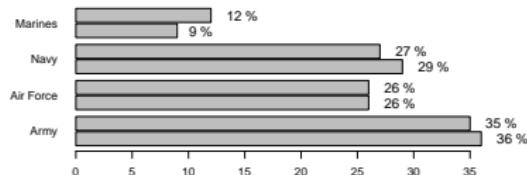
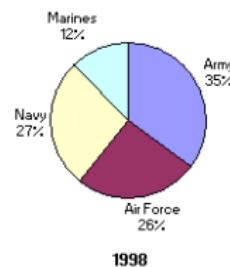
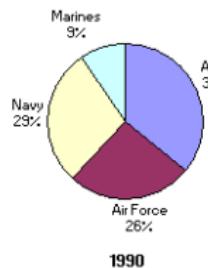
Ludność w wieku 15 lat i więcej,  
w tym 15-34 lata według aktywności ekonomicznej  
- II kw. 2009 r.



# Ocena proporcji

Przykład z <http://lilt.ilstu.edu/gmklass/pos138/datadisplay/badchart.htm>

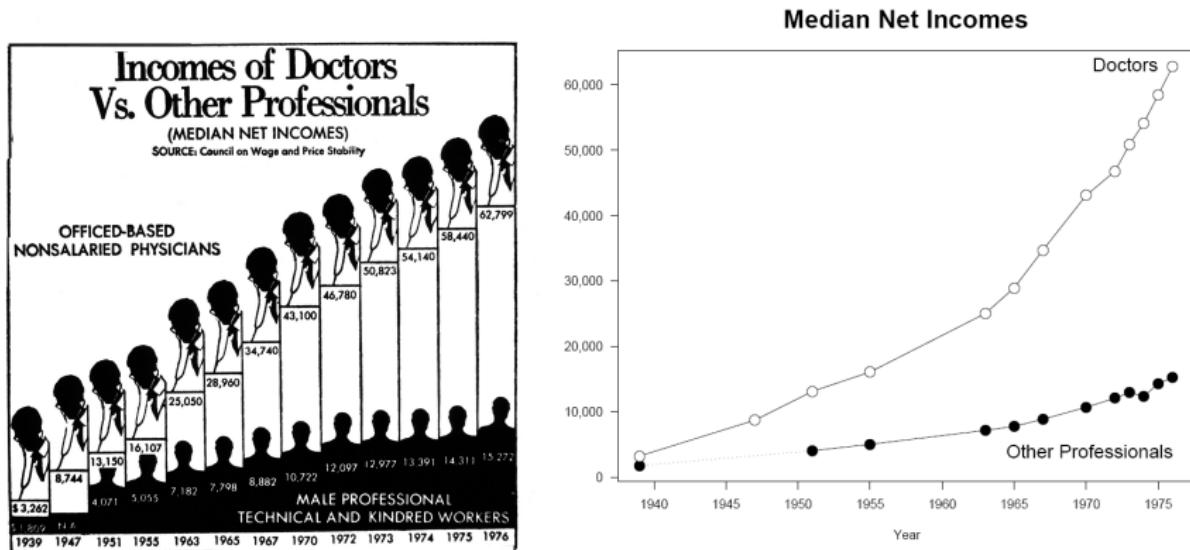
Na dobrym wykresie różnice i podobieństwa powinny być łatwe do oceny.



# Zniekształcanie wykresu: gumowe osie OX

Przykład z „The Visual Display of Quantitative Information”. E. Tufte

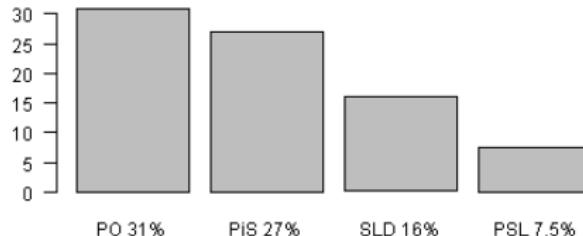
Zniekształcanie może być wynikiem zarówno nieumyślnego błędu jak i celowej manipulacji.



# Zniekształcanie wykresu: gumowe osie OY

Przykład z serwisu <http://www.szczecinek.pl>

Nawet jeżeli wizualizacji towarzyszą liczby, jeżeli te same wartości przedstawione są w sposób i graficzny i liczbowy to pierwsze wrażenie dotyczące charakteru zależności oparte jest zazwyczaj o grafikę.

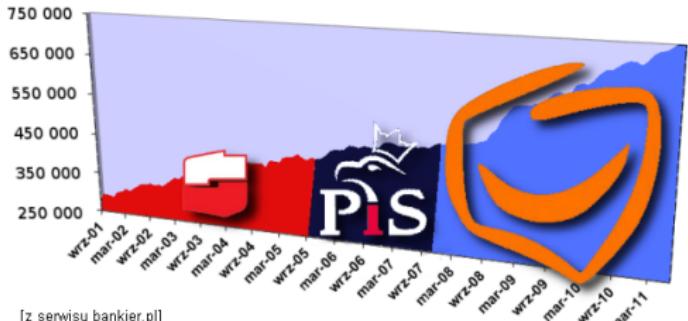
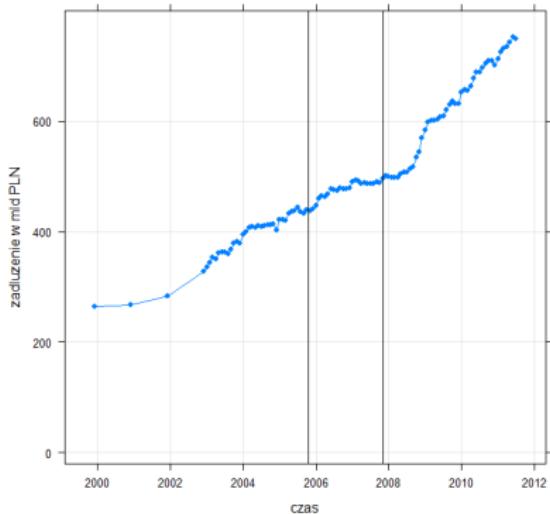


[Z serwisu [szczecinek.pl](http://www.szczecinek.pl)]

# Zniekształcanie wykresu: obroty

Przykład z serwisu <http://www.bankier.pl>

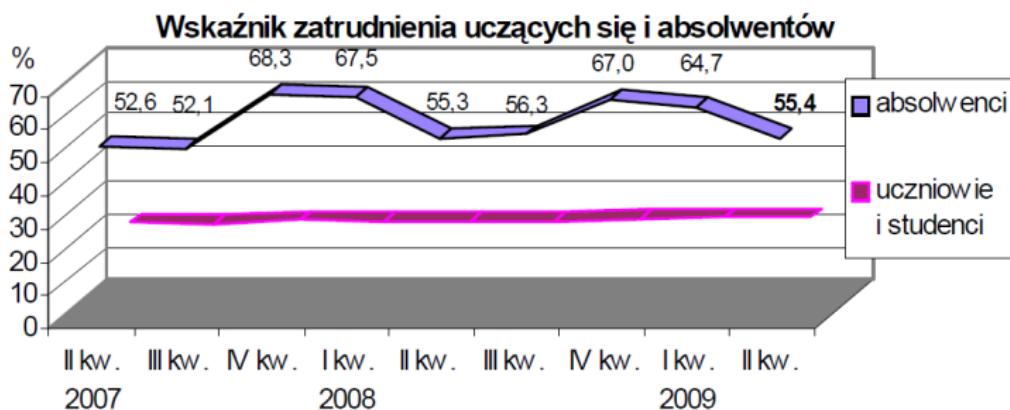
Obroty wykresu to szczególnie zły pomysł, gdy chcemy porównywać nachylenia krzywych lub wysokości punktów.



# Zniekształcanie wykresu: pseudo perspektywa

Przykład z raportu „Wejście ludzi młodych na rynek pracy”, GUS 2010

Dodawanie perspektywy bardzo rzadko jest dobrym pomysłem.  
Czasem perspektywa uniemożliwia odczytanie informacji z wykresu.

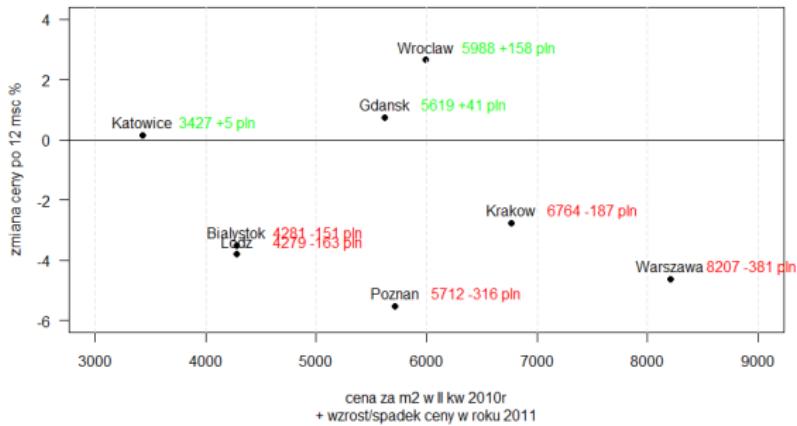
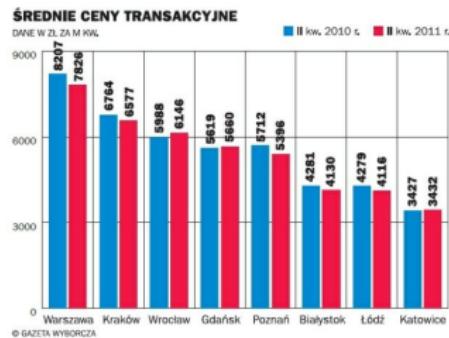


# Zniekształcanie wykresu: kolejność obiektów

Przykład z serwisu <http://www.gazeta.pl>

Porządkowanie obiektów to dobry pomysł, gdy chcemy przedstawić ranking wartości.

Ale zły pomysł gdy chcemy przedstawić różnice pomiędzy wartościami.



# Zniekształcanie wykresu: perspektywa zmienia kąty

Przykład z serwisu dzielnicy warszawa Bemowo

Wykres kołowy przedstawia liczby za pomocą kątów.

Rzuty „pseudo 3D” zmieniają kąty. Więc połączenie perspektywy i wykresu kołowego to bardzo zły pomysł.

Mocne i słabe strony życia na Bemowie



# Zniekształcanie wykresu: perspektywa zmienia kąty

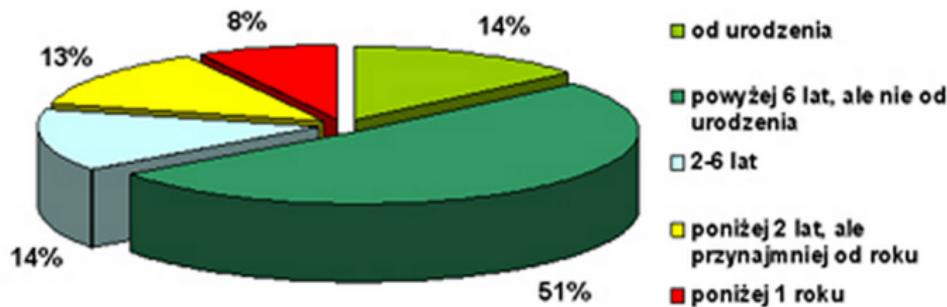
Przykład z serwisu dzielnicy Warszawa Bemowo

Wykres kołowy przedstawia liczby za pomocą kątów.

Rzuty „pseudo 3D” zmieniają kąty. Więc połączenie perspektywy i wykresu kołowego to bardzo zły pomysł.

Mocne i słabe strony życia na Bemowie

## OKRES ZAMIESZKIWANIA NA BEMOWIE

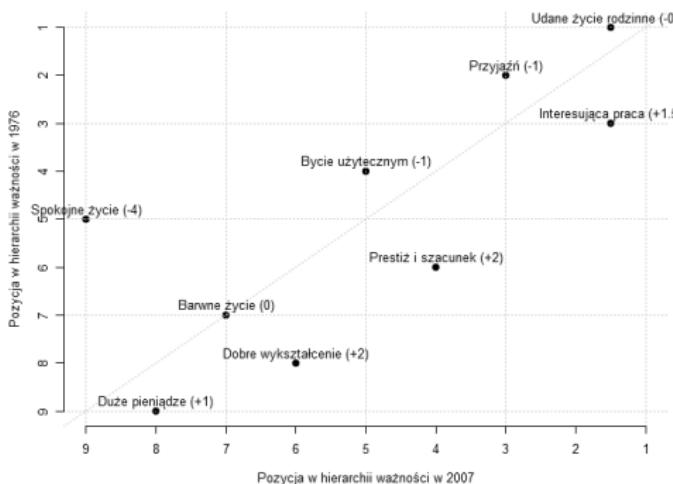


# Zła identyfikacja charakterystyki do oceny

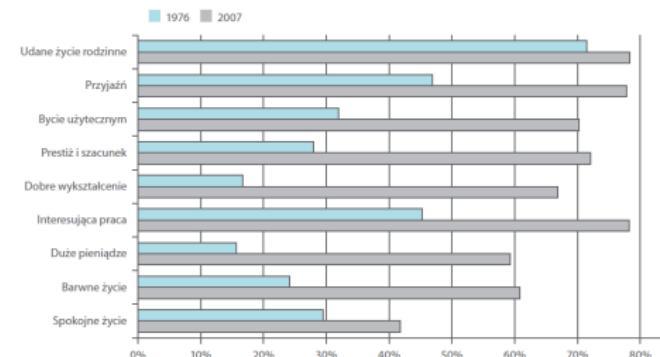
Przykład z raportu MAC „Młodzi 2011”

Nawet jeżeli wykres przedstawia dane poprawnie, wciąż pozostaje kwestia odczytania wykresu.

Z poniższych danych można mieć inne wnioski patrząc na ranking wartości a coś innego patrząc na wartości bezwzględne.



Rys.2.1. Co jest w życiu ważne? Odpowiedzi 19-letniej młodzieży w 1976 i 2007



Źródło: Badania warszawsko-kieleckie S. Nowaka (lata 70. XX w.), badania własne: „Porządkowanie – ścieżki edukacyjne i wchodzenie w dorosłość” (N = 1096).

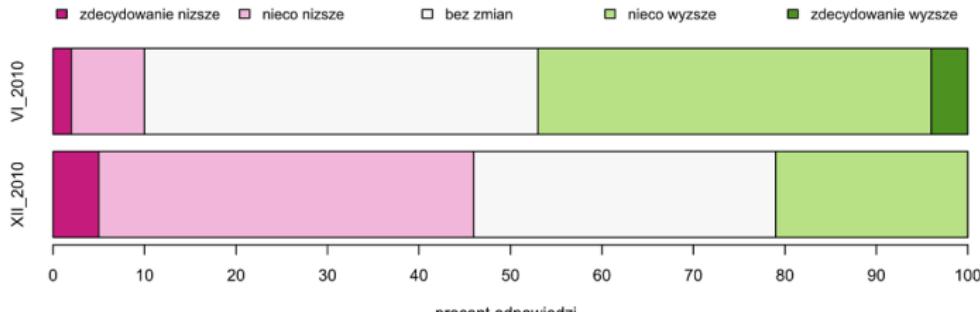
[Raport Młodzi 2011, ministerstwo MAC strona 39]

# Nie wszystkie charakterystyki są sobie równe

Na przykładzie <http://www.rp.pl/galeria/8,2,641431.html>

Rodzaj wykresu powinien uwzględniać rodzaj prezentowanych zmiennych.

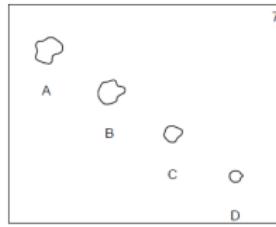
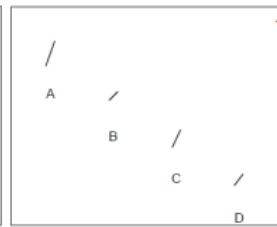
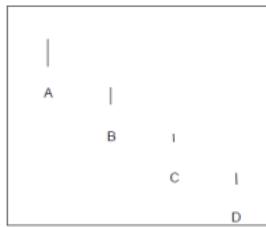
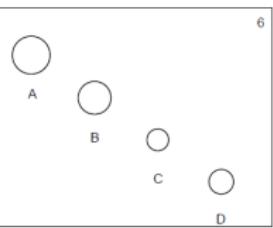
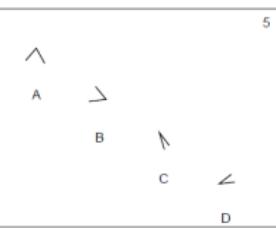
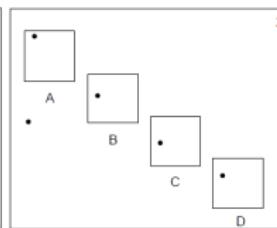
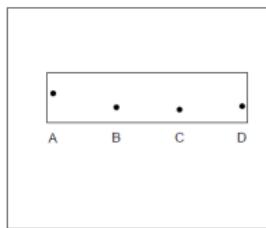
Jak zmienią się ceny mieszkań według deweloperów w ciągu 12 miesięcy



# Nie wszystkie charakterystyki są sobie równe

Eksperyment Cleveland and McGill, na podstawie „Information Visualisation”, Ross Ihaka

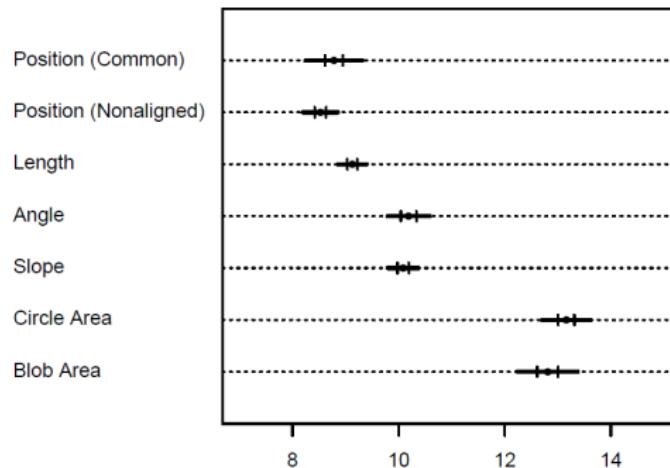
W łatwy sposób można zbadać jak dokładnie ludzie są w stanie oceniać różnice pomiędzy długościami, polami, kolorami, kątami itp.



# Nie wszystkie charakterystyki są sobie równe

Eksperyment Cleveland and McGill, na podstawie „Information Visualisation”, Ross Ihaka

W łatwy sposób można zbadać jak dokładnie ludzie są w stanie oceniać różnice pomiędzy długościami, polami, kolorami, kątami itp.



# Prezentacja informacji

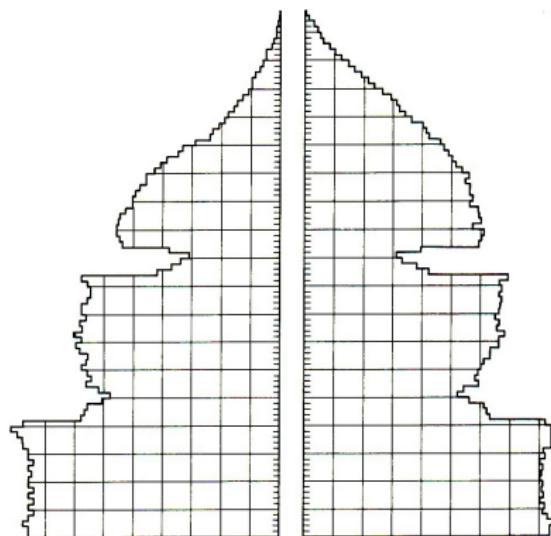
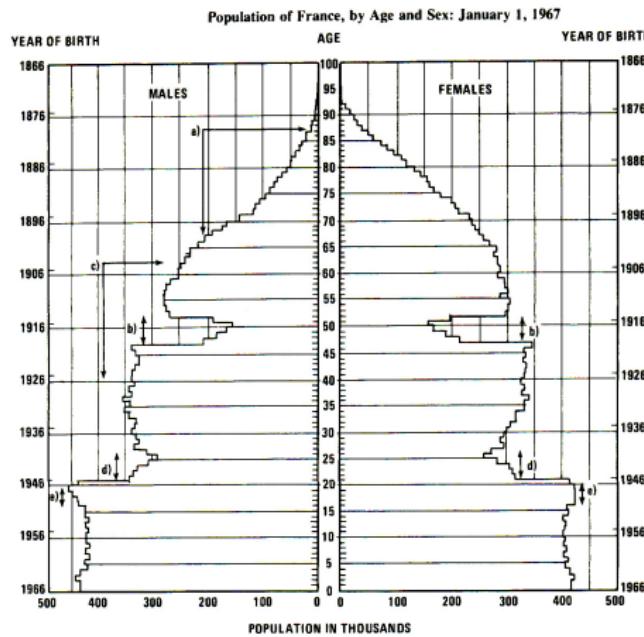
## Dobre rady wujków Edwarda Tufte'go i Rosa Ihaka

- Maksymalizuj współczynnik dane / atrament (ang. data - ink ratio).
- Minimalizuj współczynnik przekłamania (ang. lie - factor).
- Proste dane przedstawiaj prosto (ang. If the „story” is simple, keep it simple).
- Złożone dane przedstawiaj w czytelny i łatwy do interpretacji sposób (ang. If the „story” is complex, make it look simple).

# Współczynnik dane / atrament

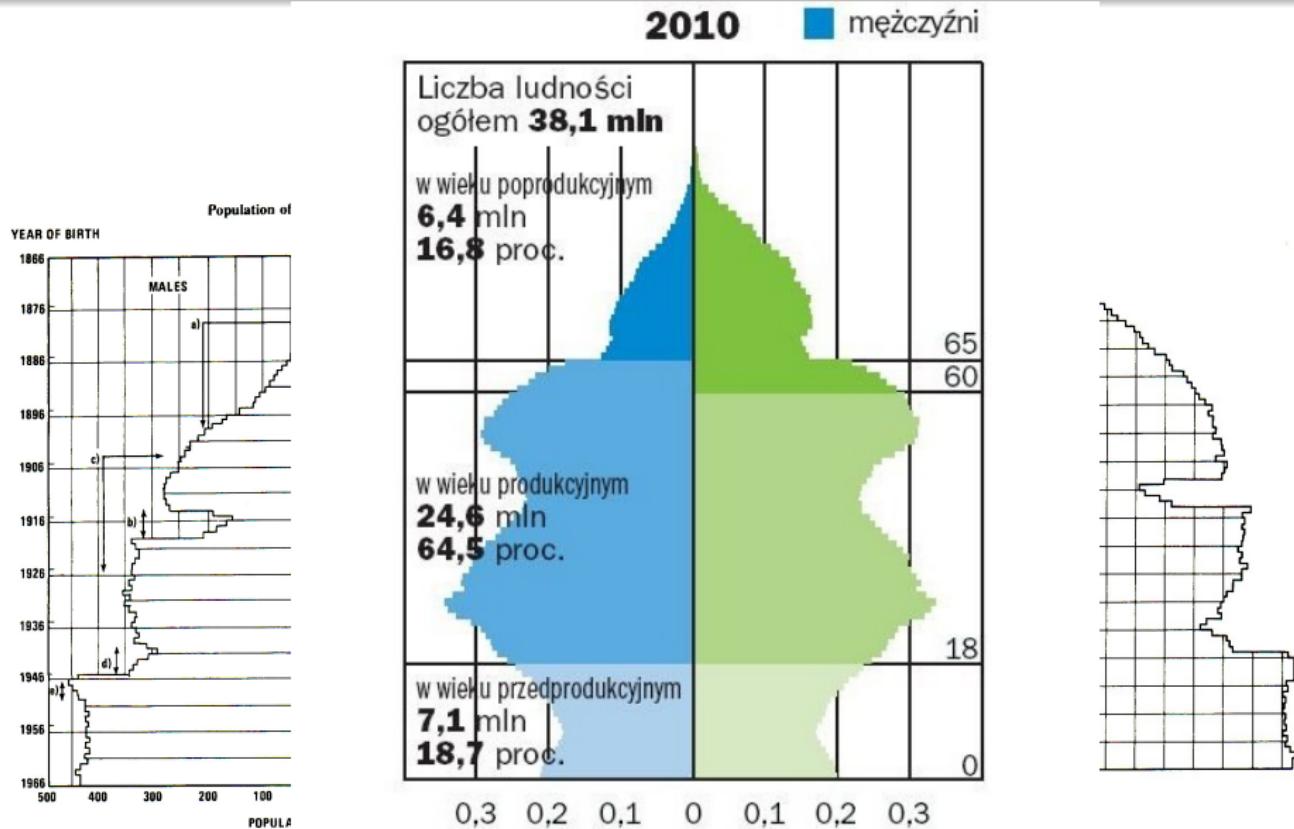
Przykład z „The Visual Display of Quantitative Information”. E. Tufte

Należy unikać rysowania elementów nie niosących dodatkowej informacji o danych. Istotne elementy nie powinny być „zalane” pomocniczymi ozdobnikami.



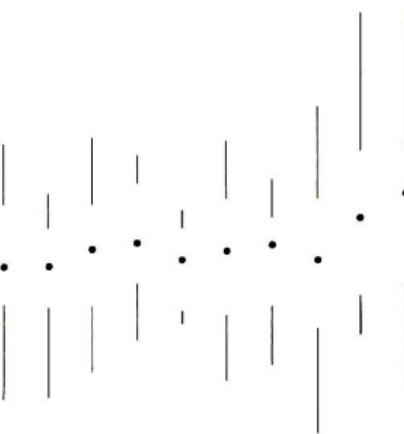
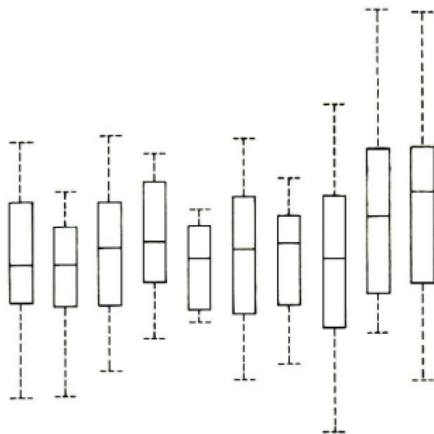
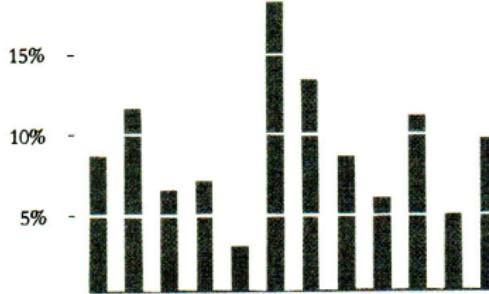
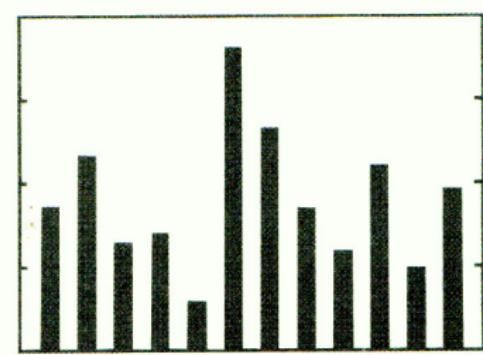
# Współczynnik dane / atrament

Przykład z „The Visual Display of Quantitative Information”. E. Tufte



# Współczynnik dane / atrament

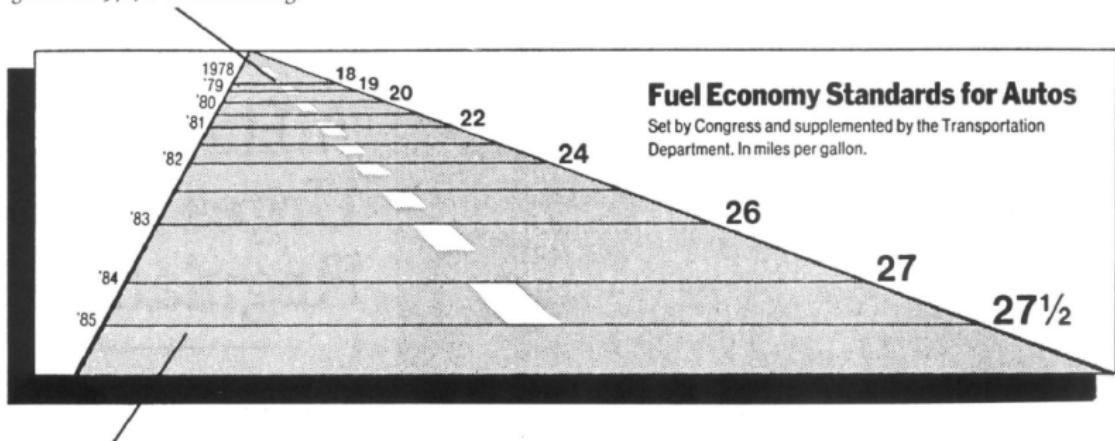
Przykład z „The Visual Display of Quantitative Information”. E. Tufte



# Współczynnik przekłamania (lie - factor)

Przykład z „The Visual Display of Quantitative Information”. E. Tufte

This line, representing 18 miles per gallon in 1978, is 0.6 inches long.



This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

$$\text{Data Effect} = \frac{27.5 - 18}{18} = 0.53,$$

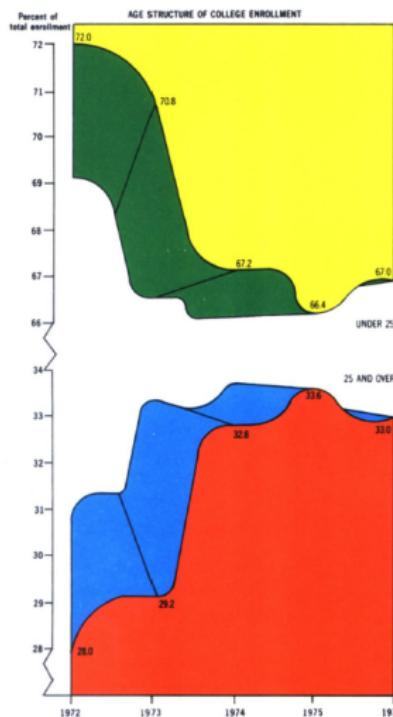
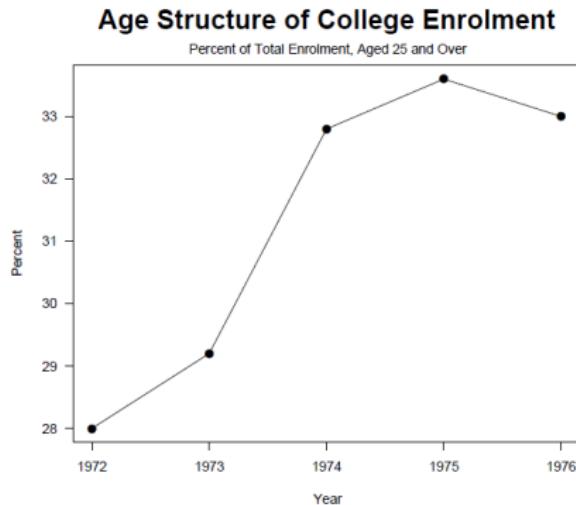
$$\text{Graph Effect} = \frac{5.3 - .6}{.6} = 7.83,$$

$$\text{Lie Factor} = 14.8$$

# Proste dane przedstawiaj prosto

Przykład z „Information Visualisation”, Ross Ihaka

Ile kolorów i elementów potrzeba by przedstawić 5 liczb?

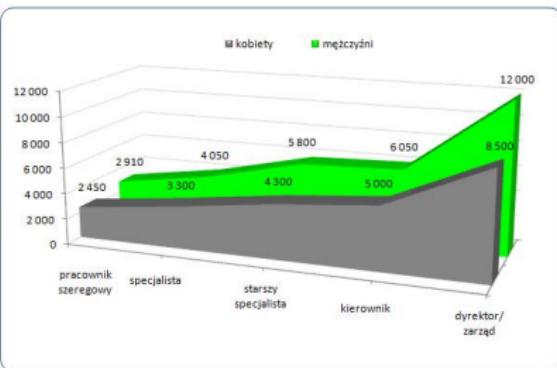


# Przykłady złych wykresów można mnożyć.

Kolekcjonuję je na blogu SmarterPoland.pl, jeżeli znajdziecie inne ciekawe podeślijcie!

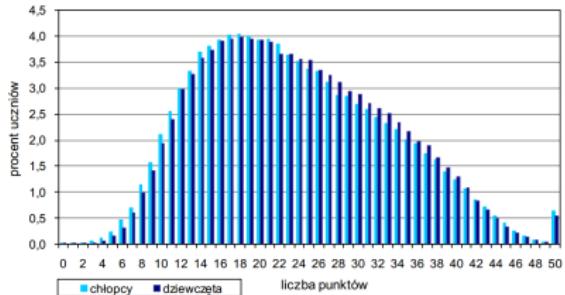


Źródło: Ogólnopolskie Badanie Wynagrodzeń przeprowadzone przez Sedlak & Sedlak w 2010 roku



Źródło: Home Broker, GUS

[Z portalu biznes.interia.pl <http://biznes.interia.pl/nieruchomosci/news/drozejaca-energia-podnosi-koszty-utrzymania-mieszkani,1722094,4205>]

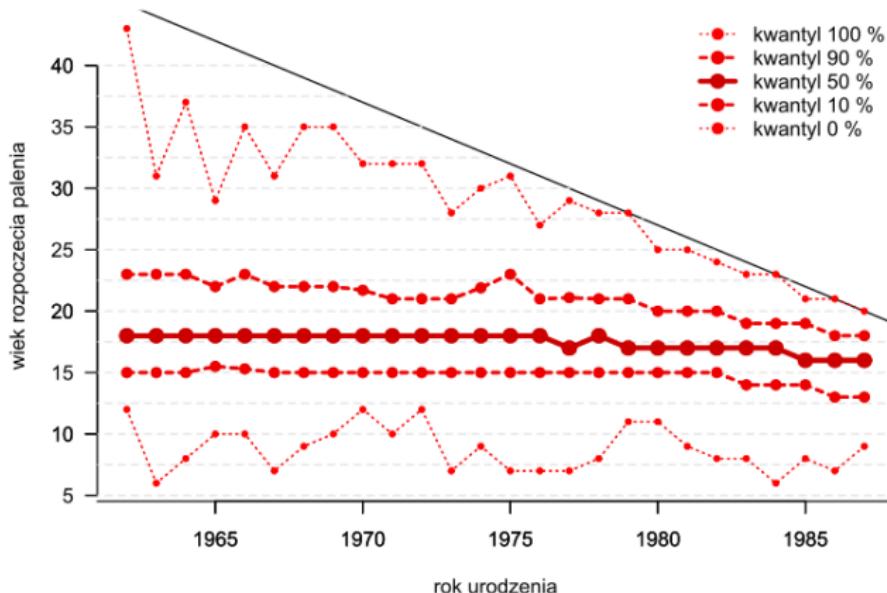


Wykres 8 z [http://www.sse.edu.pl/images/Muzeum0001/Grafika\\_2011gr.pdf](http://www.sse.edu.pl/images/Muzeum0001/Grafika_2011gr.pdf) [Dostępnie uzupełnieni konkretnymi dane]

# Wizualizacja pozwala na wyjaśnienie rzeczywistości

Czy to zdanie wyjaśnia coś nt. wieku palenia:

Średni wiek rozpoczęcia palenia to  $17.7 \pm 2.3$  lat.

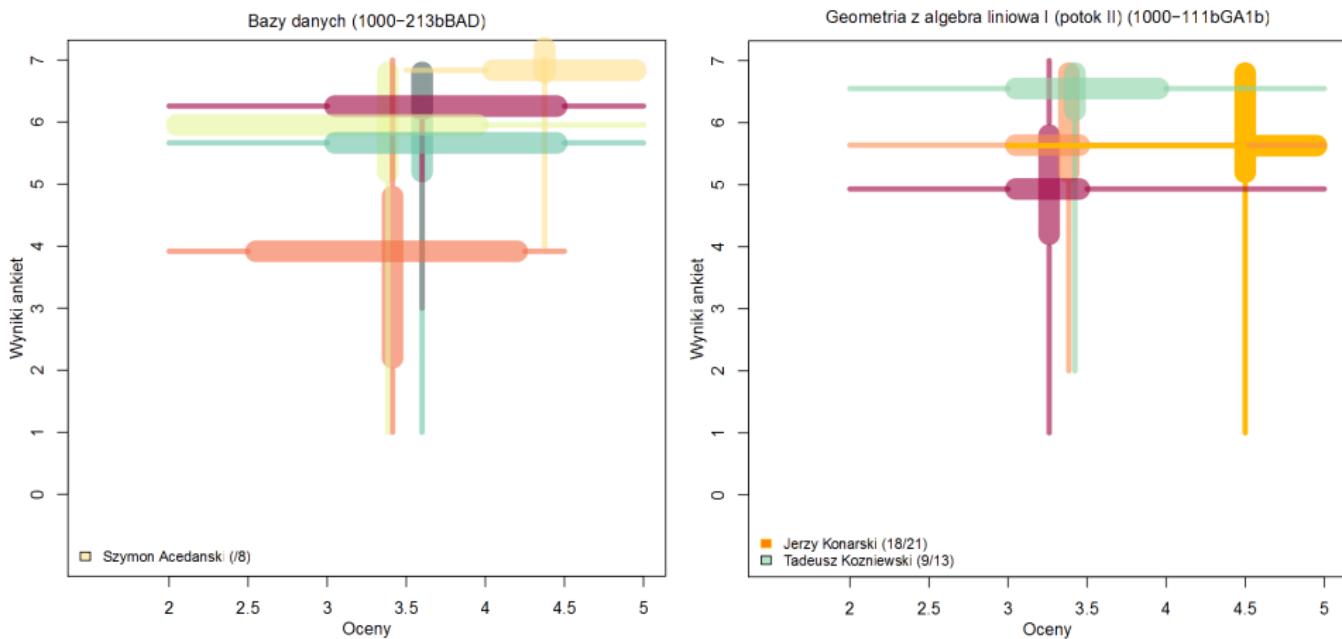


# Wizualizacja bardziej przyciąga uwagę niż tabela liczb

Wizualizacja danych z systemu uniwersyteckiego USOS

Czy lepiej wybrać miłego prowadzącego czy efektywnego?

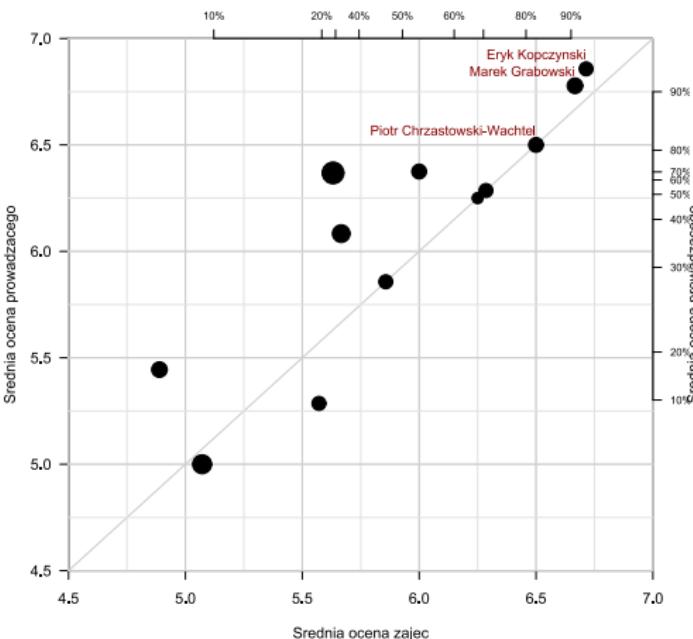
Czy mili prowadzący są bardziej efektywni od tych niemiłych?



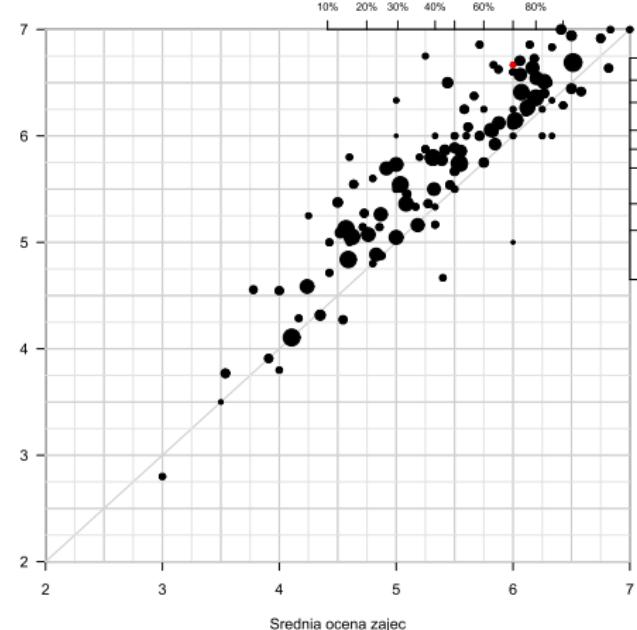
# Czy wybieramy przedmiot czy prowadzącego?

Którego ćwiczeniowca wybrać do Wstęp do programowania?

Wstęp do programowania, ćwiczenia i laboratoria, semestr 2010Z



Wykłady, semestr 2010Z



# Profile seminariów magisterkich

Wybór seminarium magisterskiego nie zawsze jest prosty. Opinie kolegów są często obciążone subiektywnymi doświadczeniami.

Pewne informacje, które mogą pomóc w tym wyborze znajdują się w USOSie, trzeba je tylko wyciągnąć i przedstawić.

Wybrane aspekty inżynierii oprogramowania (11)

