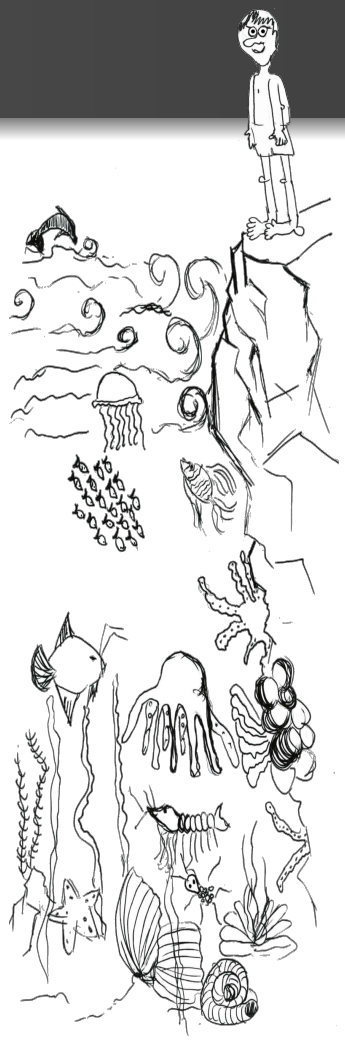
Życie w oceanie danych: wczoraj, dziś i jutro

8 Studencki Festiwal Informatyczny, 8-10 marca 2012

Przemyslaw.Biecek@gmail.com

IBM Polska / MIM UW / SmarterPoland



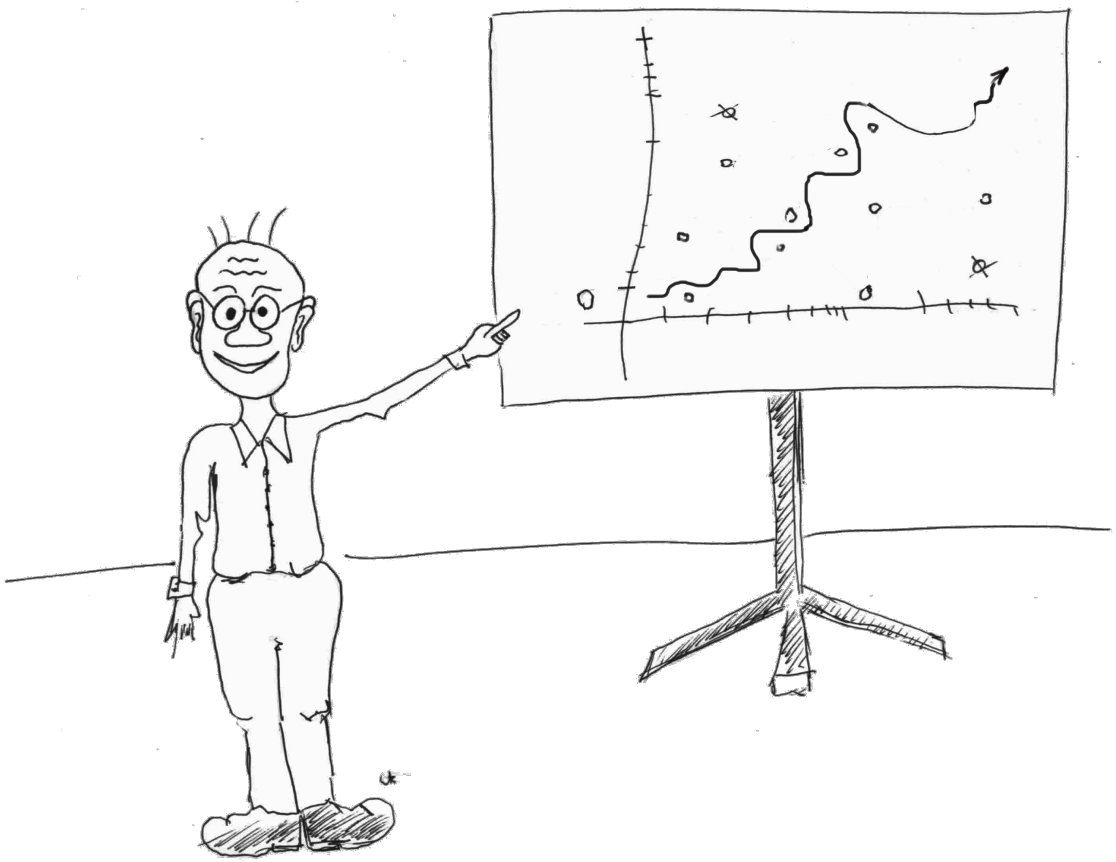
Czy rozmiar się liczy?

~ MB Dane nie kłamią ale wykresy mogą,
czyli dobre i złe wizualizacje.

~ GB USOS wie lepiej,
czyli rekomendacje
na podstawie 2 mln ocen.

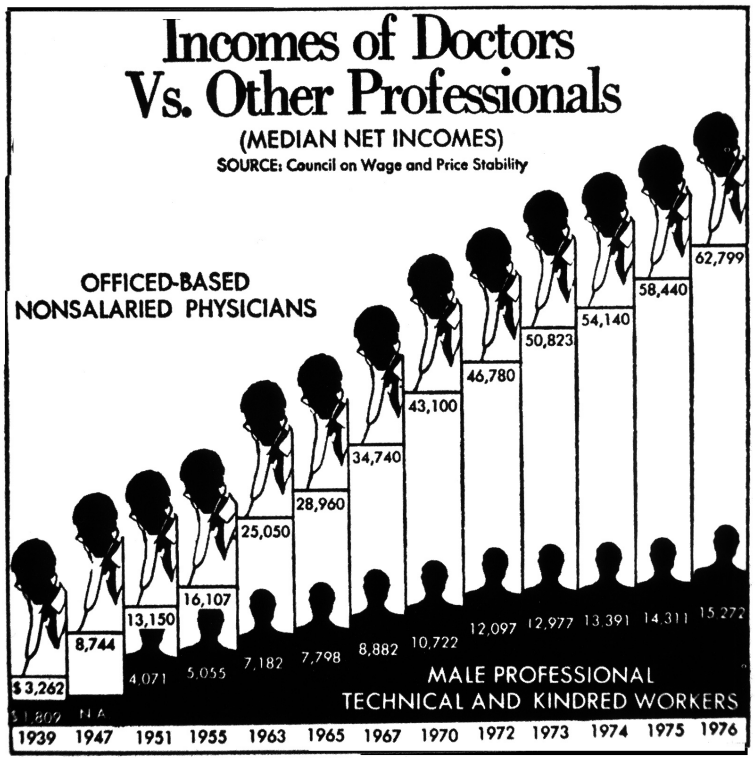
~ TB Co widzi bioinformatyk
w ciągu 20 000 000 000 000 literek.

Część I. Wizualizacja



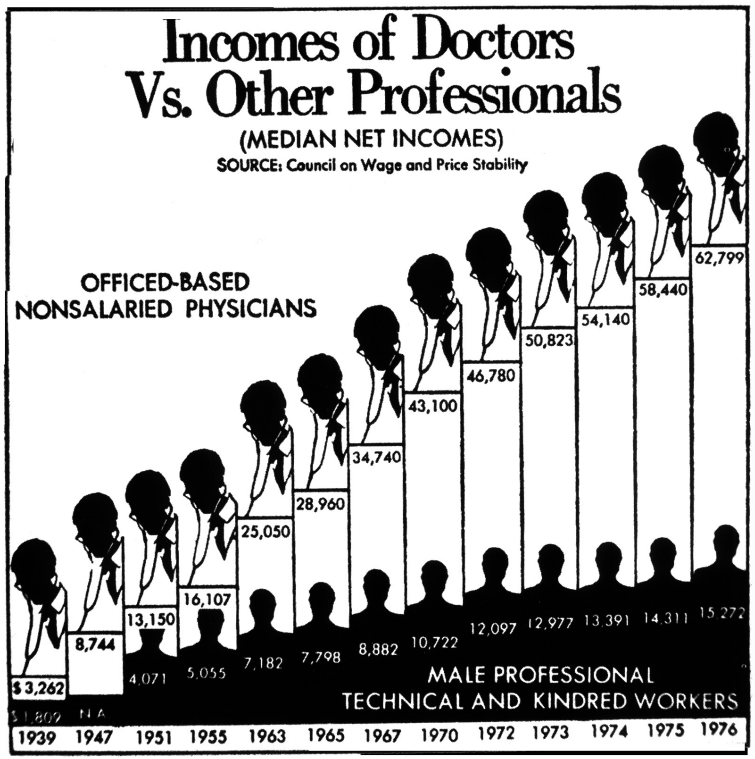
Śledzenie błędów i manipulacji

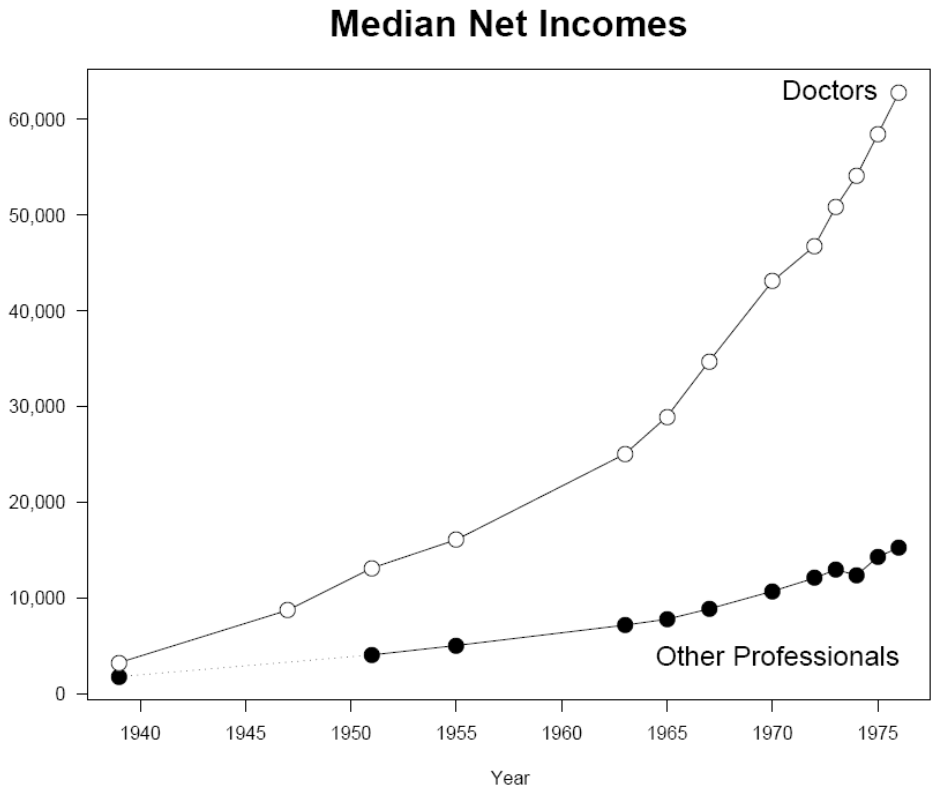
Wszystko zaczęło się od lektury książek Edwarda Tuftego pokazujących wiele wykresów o wysokiej wartości współczynnika „Lie Factor”.



Śledzenie błędów i manipulacji

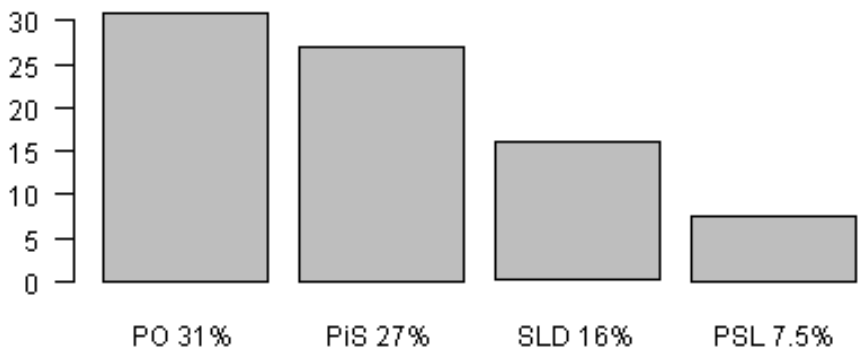
Wszystko zaczęło się od lektury książek Edwarda Tuftego pokazujących wiele wykresów o wysokiej wartości współczynnika „Lie Factor”.





Śledzenie błędów i manipulacji

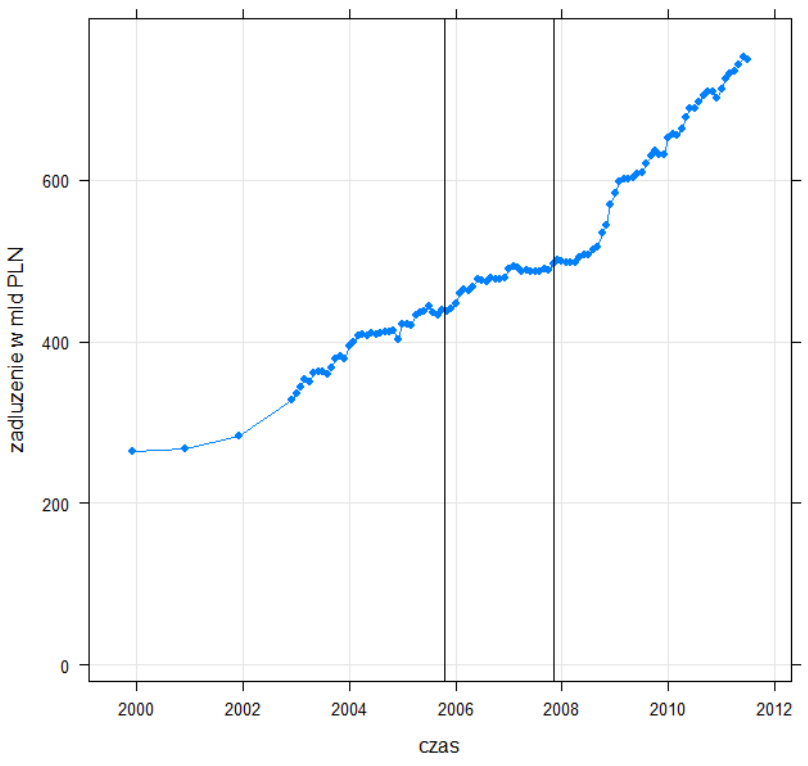
Jeżeli jakieś liczby przedstawione są w sposób graficzny i tekstowy to pierwsze wrażenie / sugestia dotyczące charakteru zależności oparte jest zazwyczaj o grafikę.
Nie ważne ile pieniędzy sie wydało na zebranie liczb opisujących rzeczywistość, źle opracowana grafika może zupełnie zmienić percepcję wyników.

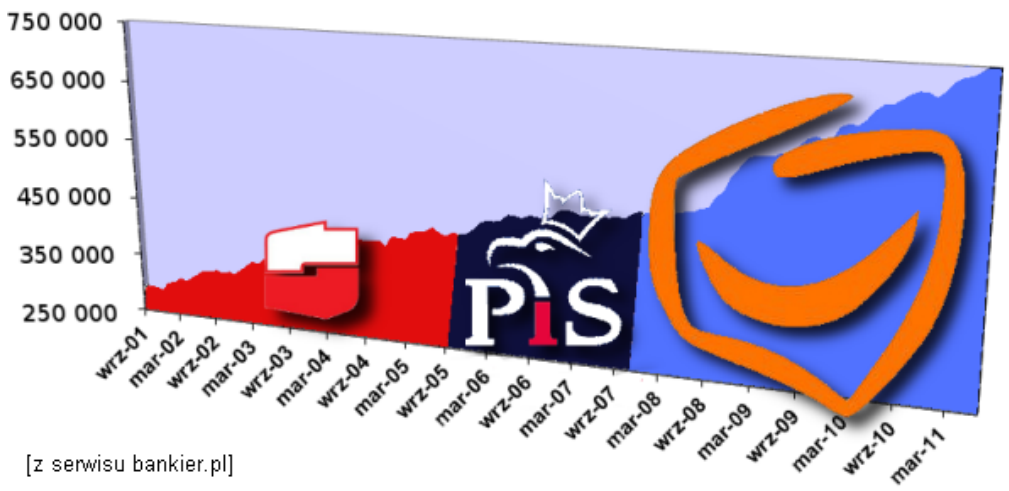


[Z serwisu szczecinek.pl]

Śledzenie błędów i manipulacji

Typowe błędy polegają na dodawaniu ozdobników w stylu perspektywy lub pseudo trzeciego wymiaru ...

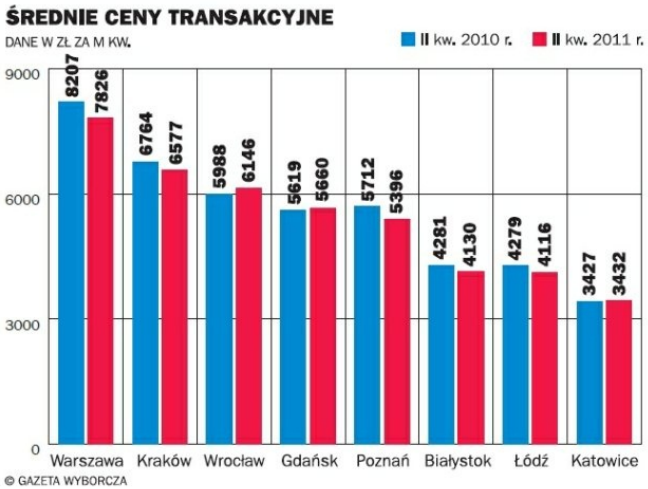


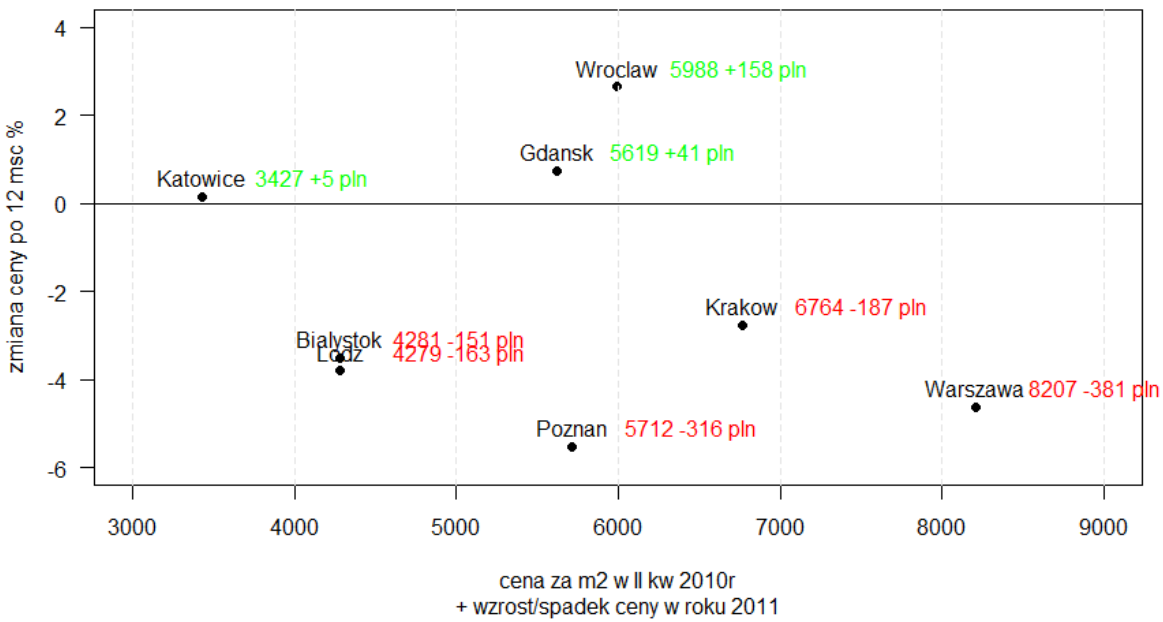


[z serwisu bankier.pl]

Śledzenie błędów i manipulacji

... ustawianie kolejności obiektów by wywołać dodatkowe wrażenia lub używanie osi ukrywającej potencjalne różnice ...

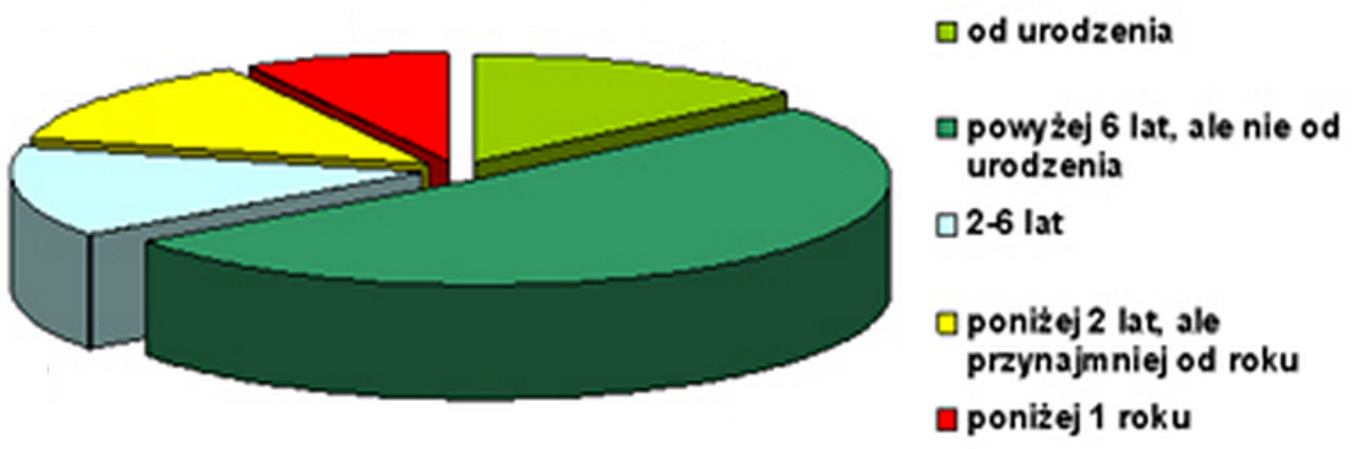




Śledzenie błędów i manipulacji

... czy wykresów kołowych z dodanym pseudo trzecim wymiarem ...

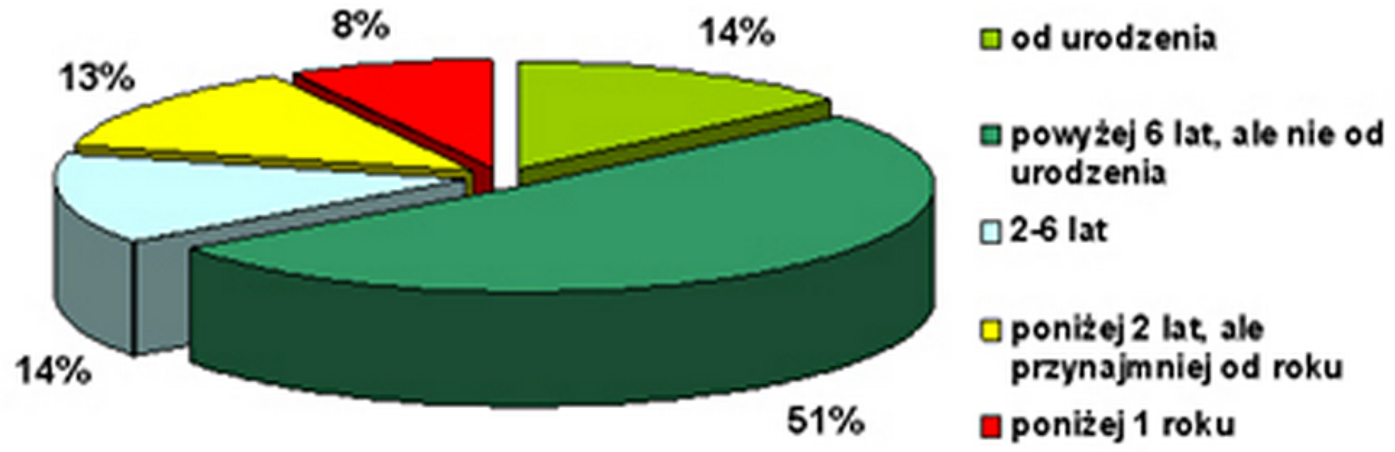
Mocne i słabe strony życia na Bemowie



Śledzenie błędów i manipulacji

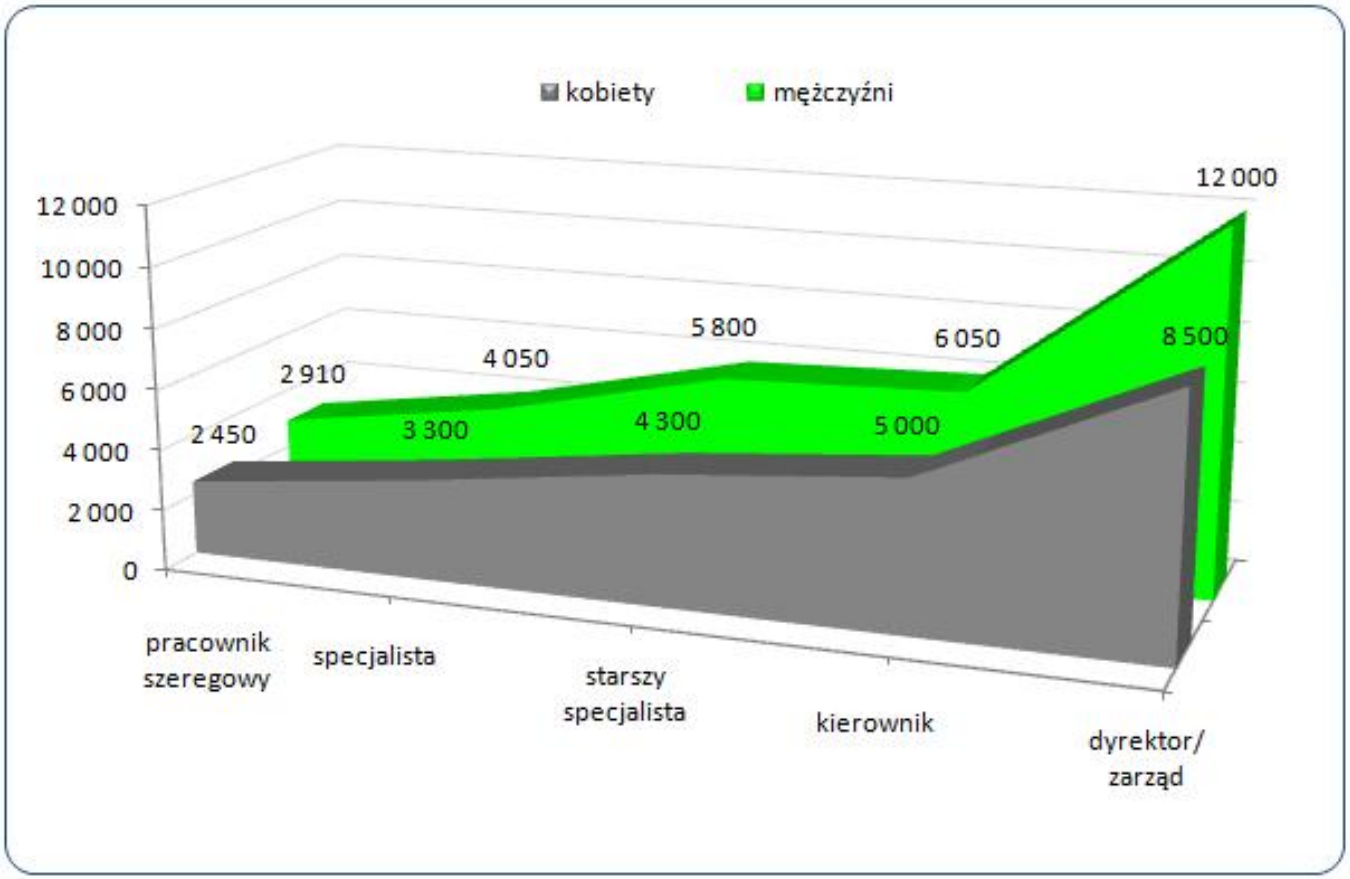
... czy wykresów kołowych z dodanym pseudo trzecim wymiarem ...

Mocne i słabe strony życia na Bemowie



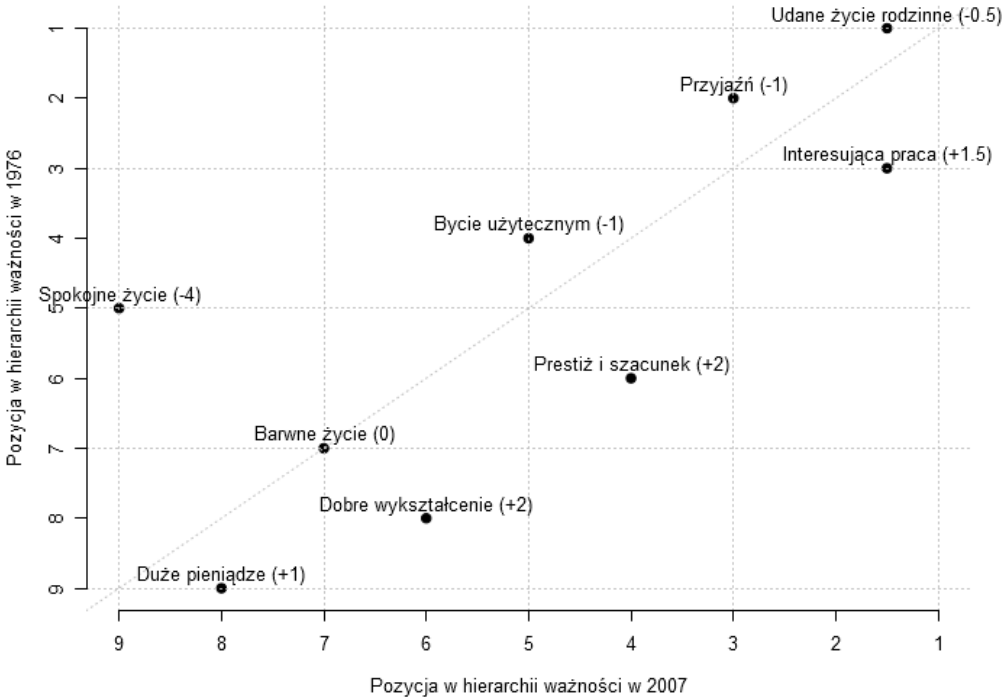
Śledzenie błędów i manipulacji

... pomysłowość autorów jest tak wielka, że zwrot „podkręcenie wyników” nabiera nowego znaczenia.

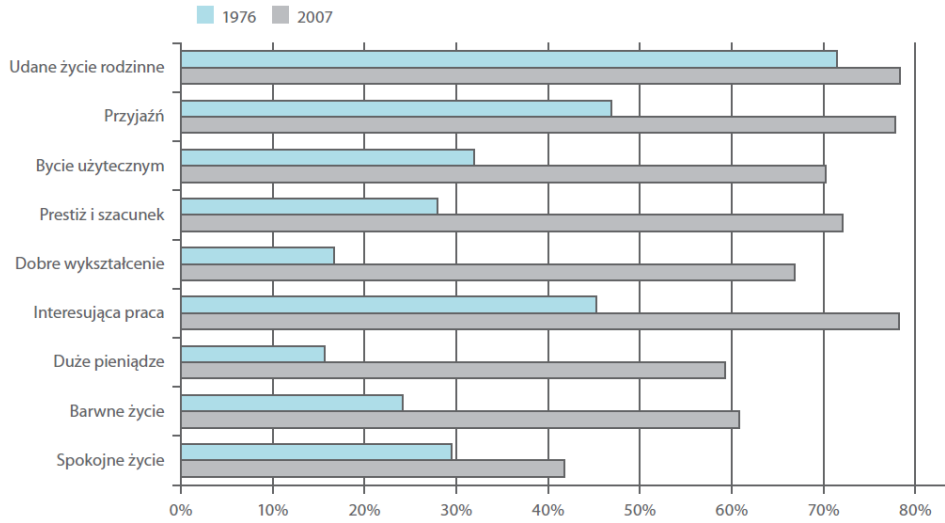


Wyszukiwanie błędów

Czasem błąd polega na zwykłej nieumiejętności czytania danych.
Co niestety może mieć spore konsekwencje.



Rys.2.1. *Co jest w życiu ważne?* Odpowiedzi 19-letniej młodzieży w 1976 i 200



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

[Raport Mlodzi 2011, ministerstwo MAC strona 39]

Wyszukiwanie błędów

Czasem błąd polega na przekonaniu badacza, że jest inaczej niż to pokazują dane.
Co niestety może mieć spore konsekwencje.

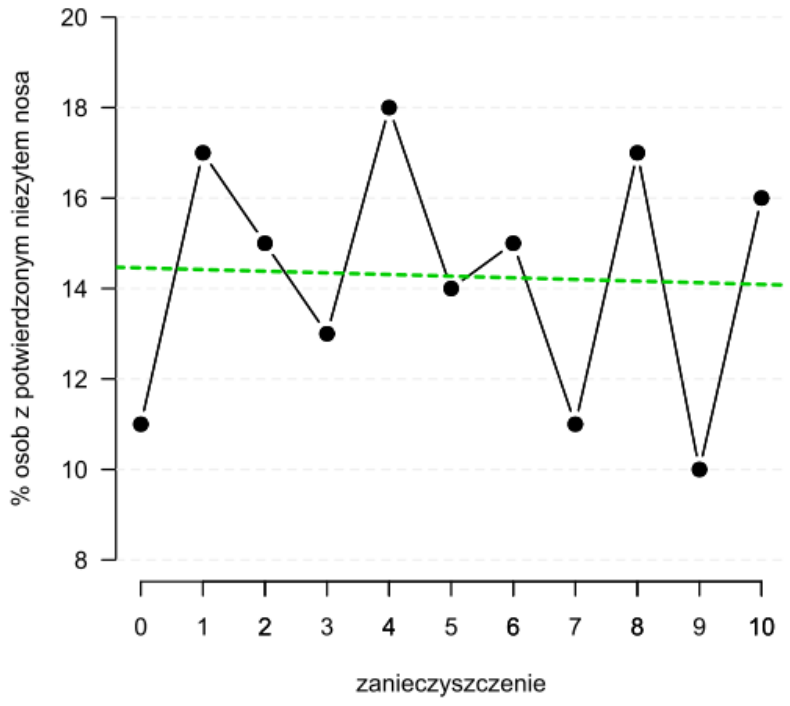


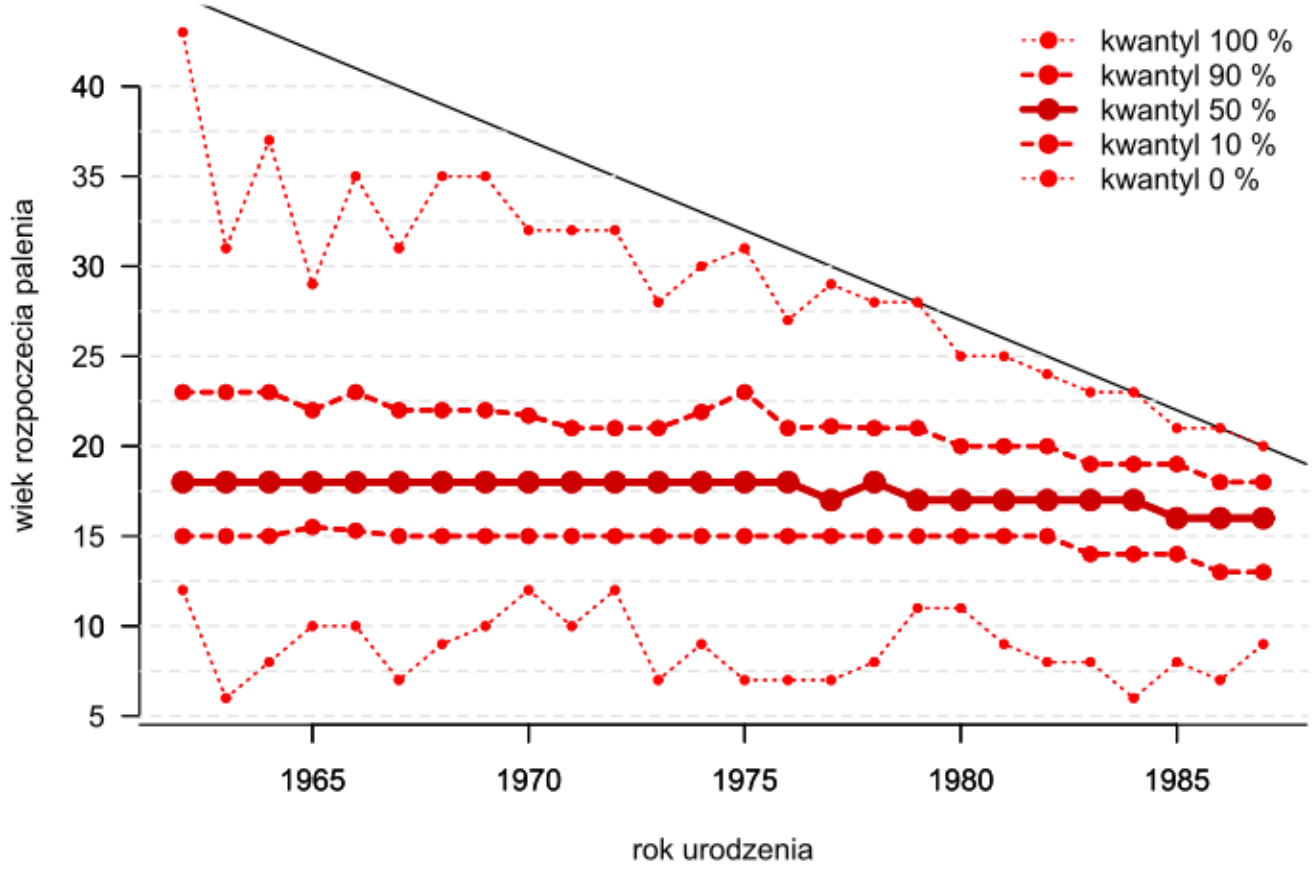
Tabela 2. Iloraz szans (OR) dla potwierdzonego klinicznie okresowego alergicznego nieżytu nosa ze względu na oddziaływanie zanieczyszczonego powietrza z zewnątrz budynku.

Wpływ zanieczyszczonego powietrza	Potwierdzony klinicznie okresowy alergiczny nieżyt nosa		Iloraz szans (OR)	95% przedział ufności		p
	Nie N-liczba	Tak N-liczba				
brak	933	123				
słabo	1272	221	**1,32**	**1,04**	**1,67**	**0,02**
umiarkowanie	1194	187	1,19	0,93	1,51	0,16
silnie	715	129	**1,37**	**1,05**	**1,78**	**0,02**

[KSZTAŁCENIE PODYPLOMOWE WUM Rok IV, nr 1/2011]
[https://ckp.wum.edu.pl/sites/ckp.wum.edu.pl/files/periodyk_nr_1-2011_0.pdf]

Wyjaśnianie rzeczywistości

Oczywiście to nie jest tak, że wizualizacja jest zawsze zła. Wizualizacja może być bardzo użyteczna.



Ocena wpływu wdrożenia technologii teleinformatycznych na poszczególne dziedziny - Ogółem

H1. Jakie zmiany w rezultacie zastosowania technologii teleinformatycznych w Państwa Urzędzie zaobserwowano w poszczególnych dziedzinach?

N=1601, odpowiadały wszystkie urzędy.

Zdaniem około 40% respondentów dzięki wdrożeniu technologii teleinformatycznych wzrosła sprawność załatwiania spraw, liczba innowacyjnych rozwiązań związanych z organizacją pracy, efektywność wykorzystania zasobów urzędów oraz innowacyjność pracowników. Jednocześnie podobny odsetek badanych był zdania, że wdrożenie nowych technologii przyniosło wzrost obciążenia pracą. Ponad połowa respondentów uznało, że proces ten nie miał wpływu na uciążliwość procedur obsługi klienta ani liczbę dokumentów w postaci papierowej przetwarzanych przez urząd.

„Badanie wpływu informatyzacji na działanie urzędów administracji publicznej w Polsce w 2011 roku"
Raport z badania ilościowego dla MSWiA

[Raport MAC „Badanie wpływu informatyzacji na działanie urzędów administracji publicznej w Polsce w 2011 roku"
http://www.mswia.gov.pl/download.php?s=1&id=13759]

Konkursik nr 1

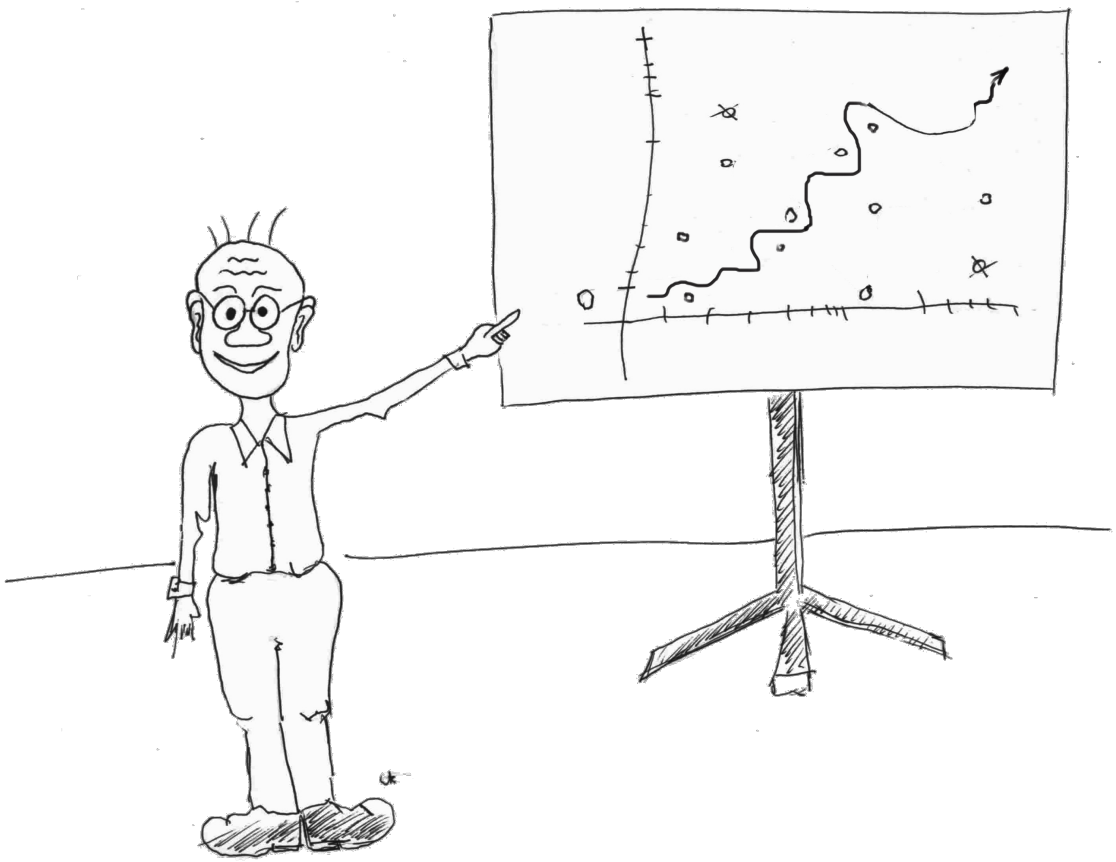
Znajdź artykuł w gazecie, wizualizację lub grafikę na stronie www, który jest zupełnie niecztelny, źle przygotowany, zmanipulowany lub przeciwnie, który Ci się bardzo podoba, jest Twoim zdaniem świetną, interesującą wizualizacją.

Podejdź do mnie z tym obrazkiem lub wyślij informację o nim emailem na adres `Przemyslaw.Biecek@gmail.com`.

Odbierz kubek, 2GB pendrive lub termos*.

*Oferta ważna do wyczerpania zasobów.

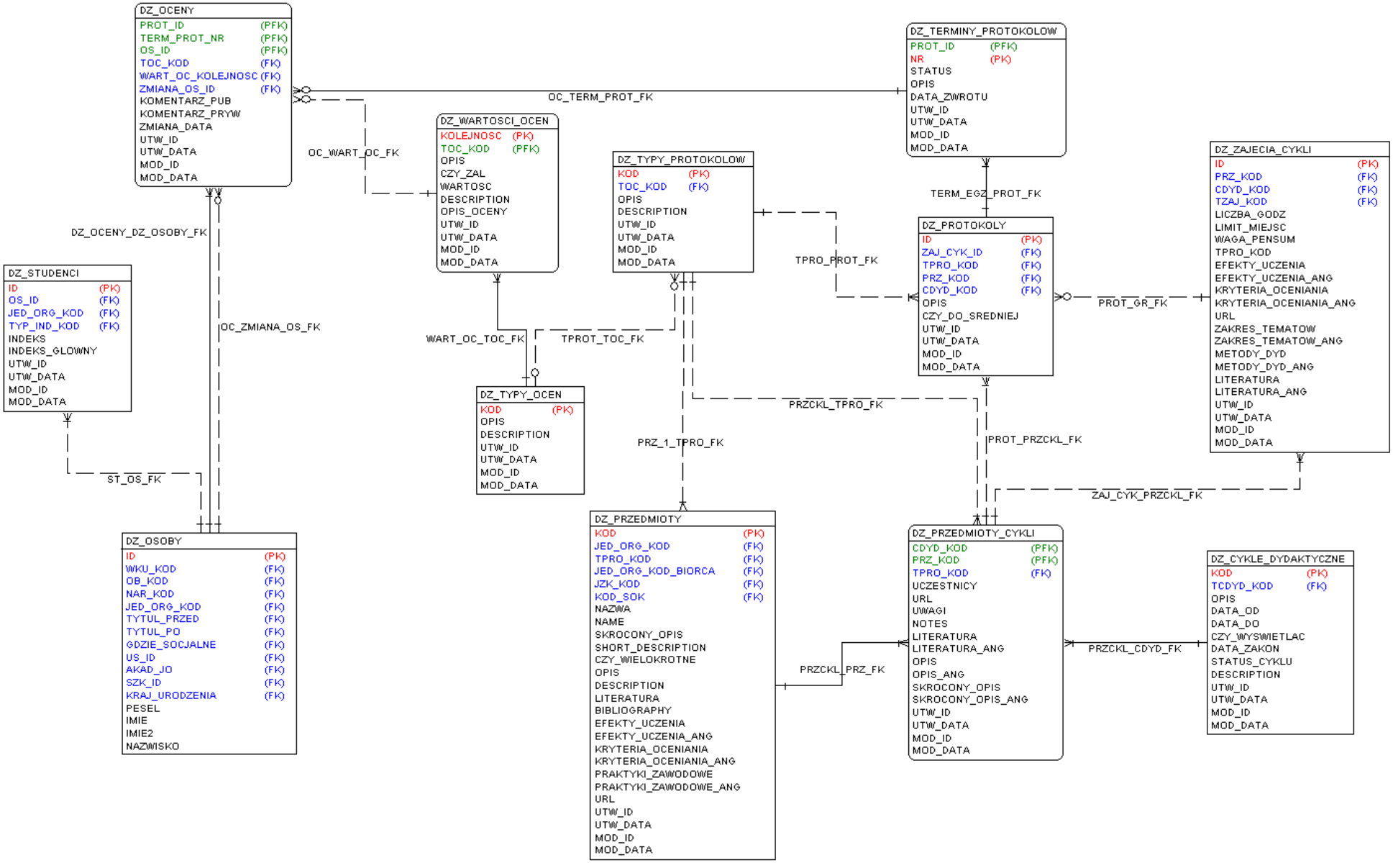
Część II. Drążenie w USOSie



O USOSie faktów kilka

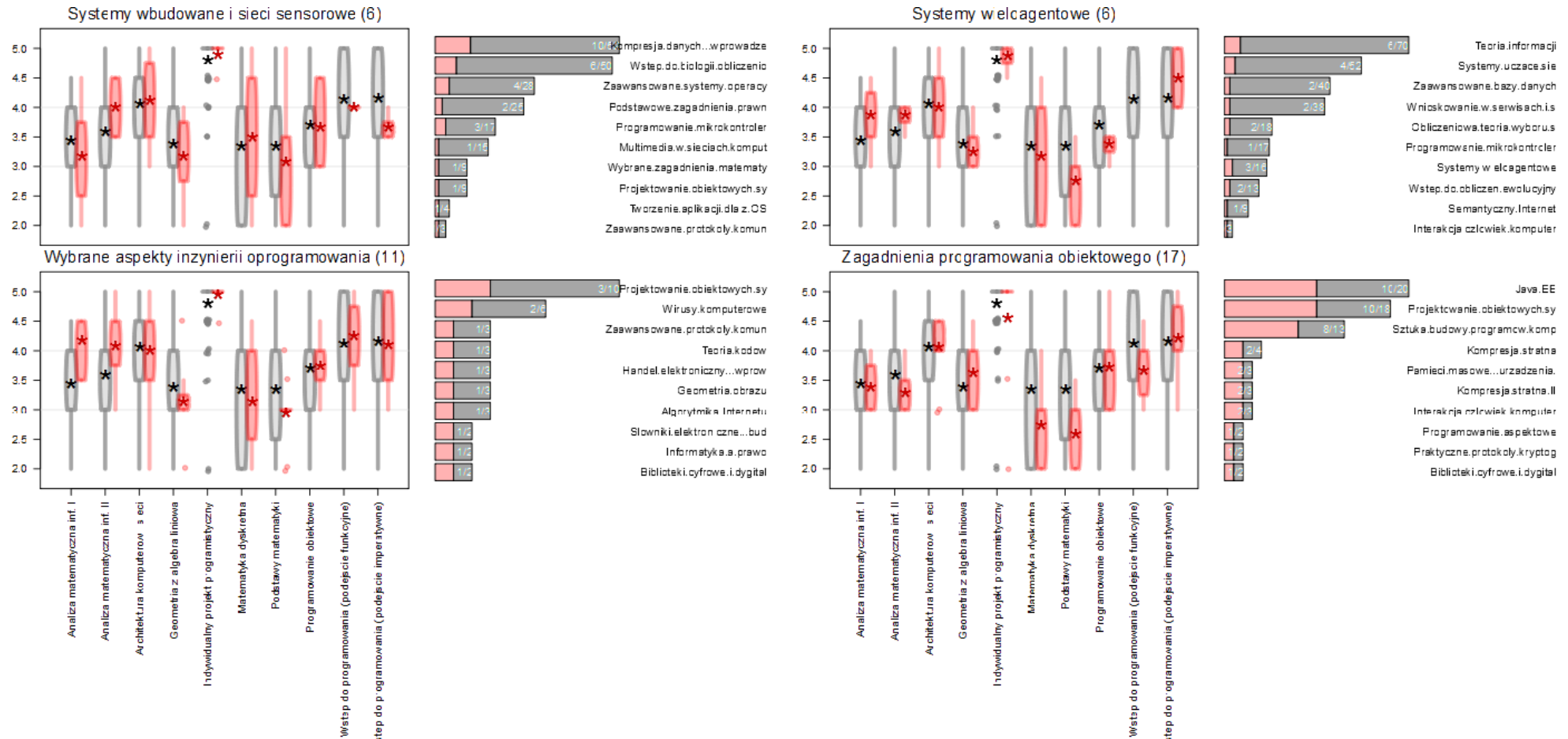
- USOS jest rozwijany od 1999/2000 roku. Początkowo na wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego, obecnie też przez UJ, UMK i PR.
- Z systemu USOS korzystają obecnie 23 szkoły wyższe.
- Liczba studentów w tych szkołach to około 470 tys. co stanowi 1/4 wszystkich studentów w Polsce,
- Część uczelni przechowuje w USOSie dane począwszy od 1999 roku.
- Główna baza danych ma ponad 360 tabel.
- Stowarzyszone projektu to między innymi Krajowy Rejestr Matur, Elektroniczny Katalog Studiów, Internetowa Rejestracja Kandydatów.

Schemat tabel opisujących oceny



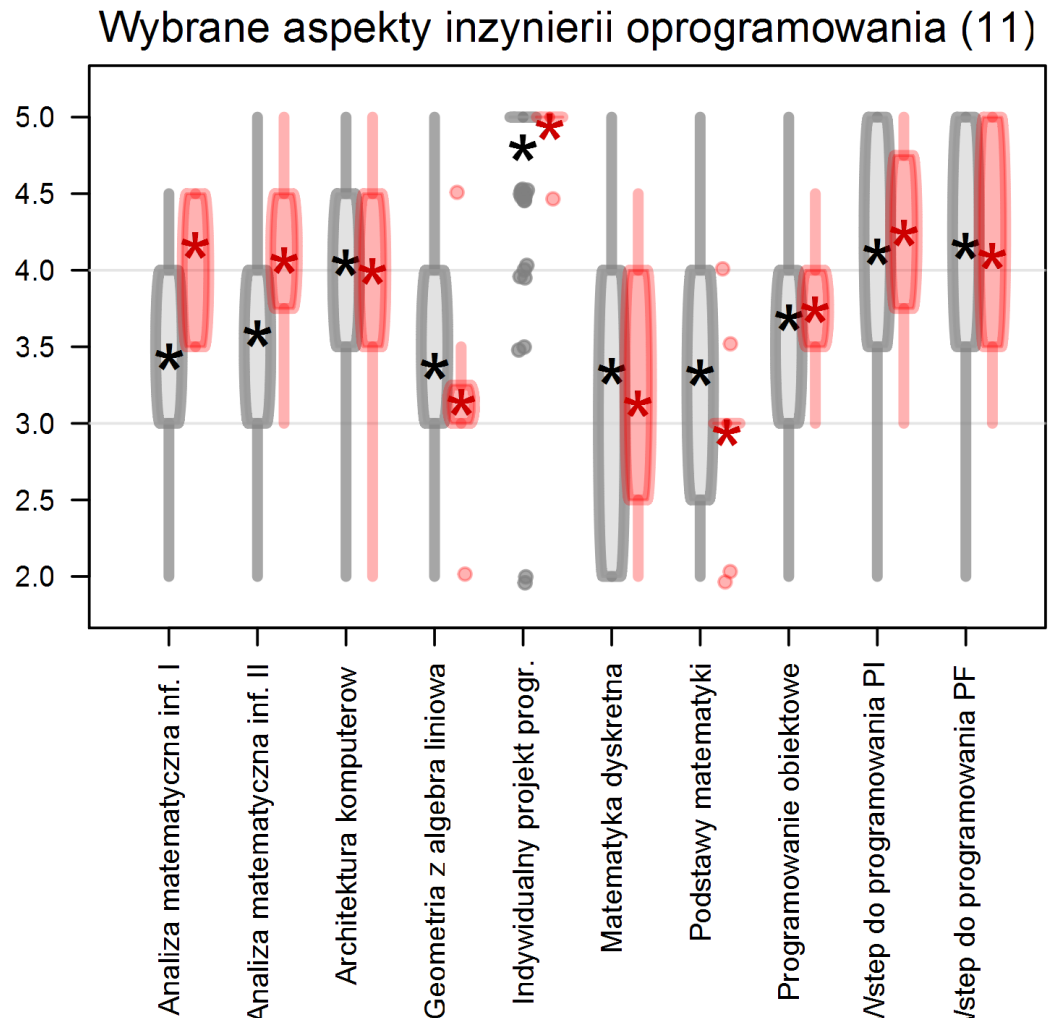
Profile seminariów magisterkich

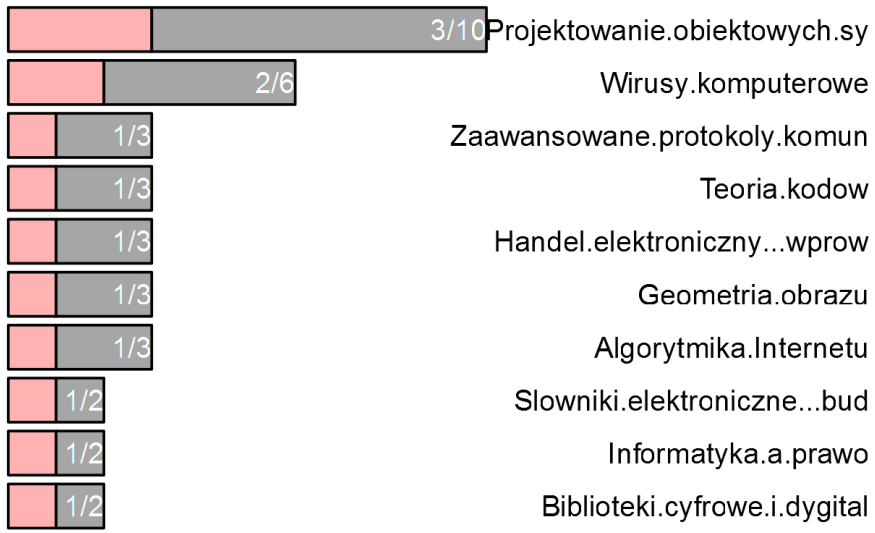
Wybór seminarium magisterskiego nie zawsze jest prosty. Opinie kolegów są często obciążone subiektywnymi doświadczeniami. Pewne informacje, ktre mogą pomóc w tym wyborze znajdują się w USOSie, trzeba je tylko wyciągnąć.



Profile seminariów magisterkich

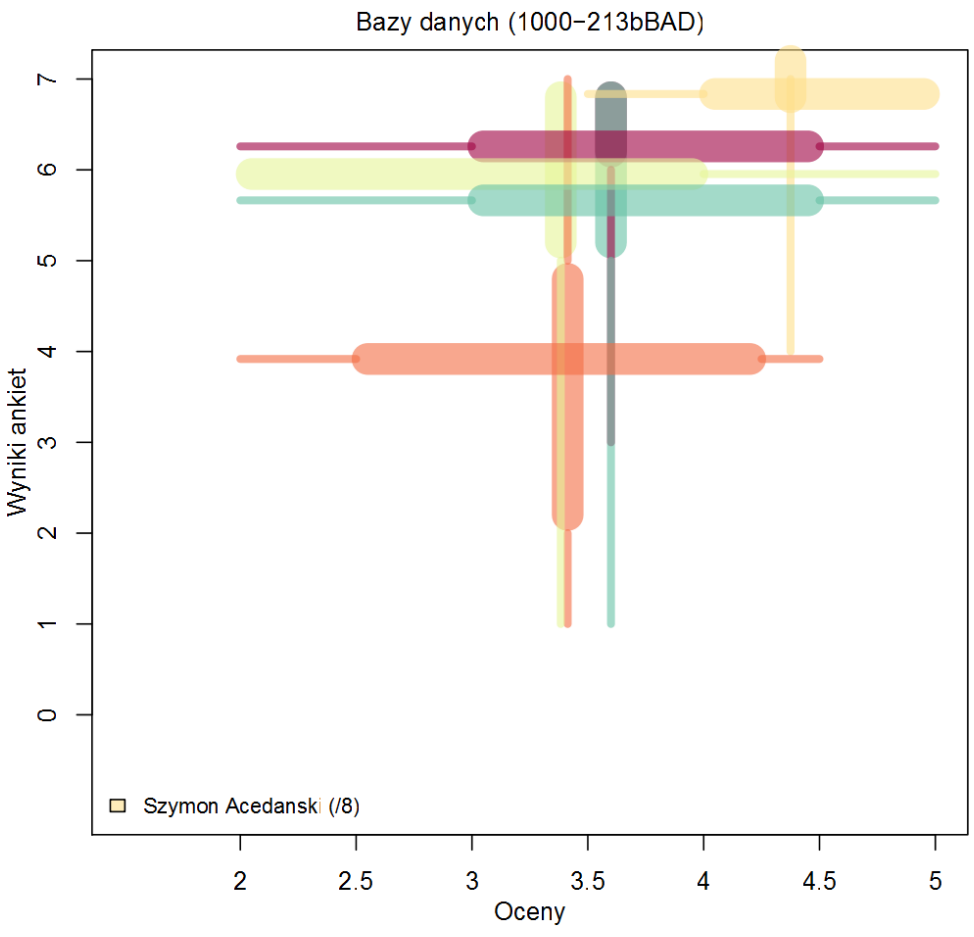
Po lewej rozkłady ocen z kursów obowiązkowych, po lewej stronie znajduje się lista kursów wybieranych częściej niż przez studentów z innych seminariów.

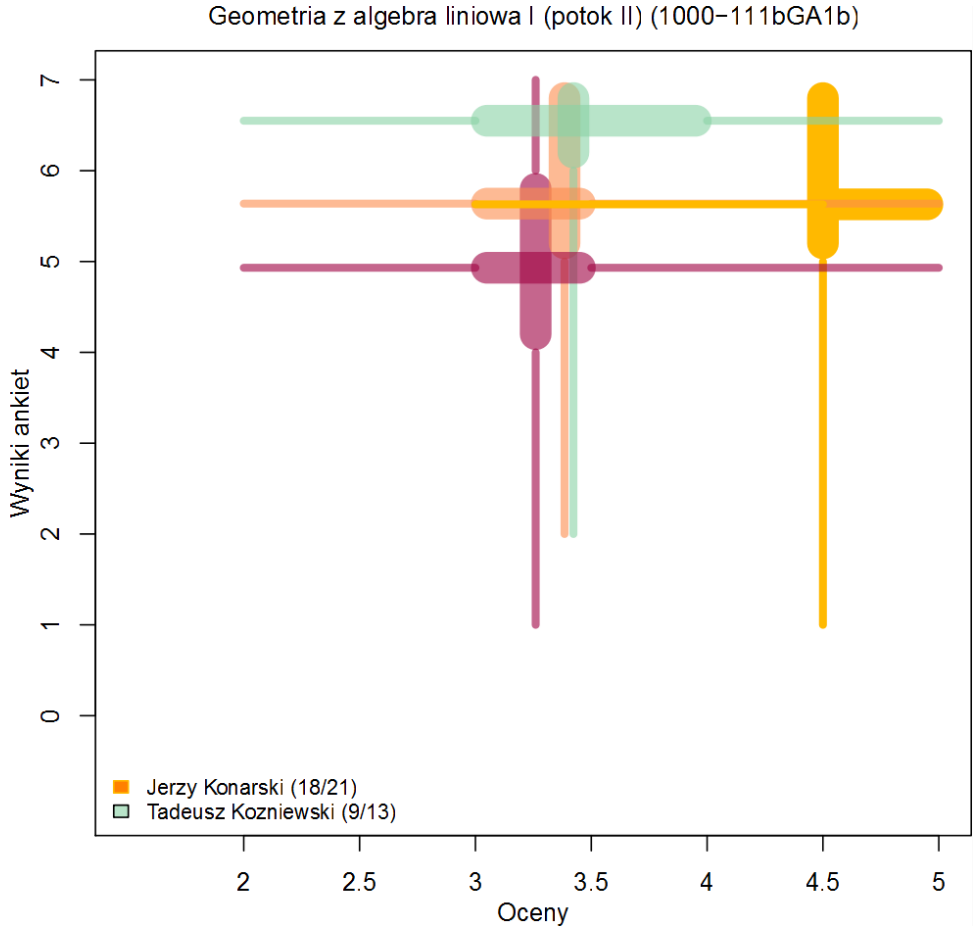




Lubiany czy uczący?

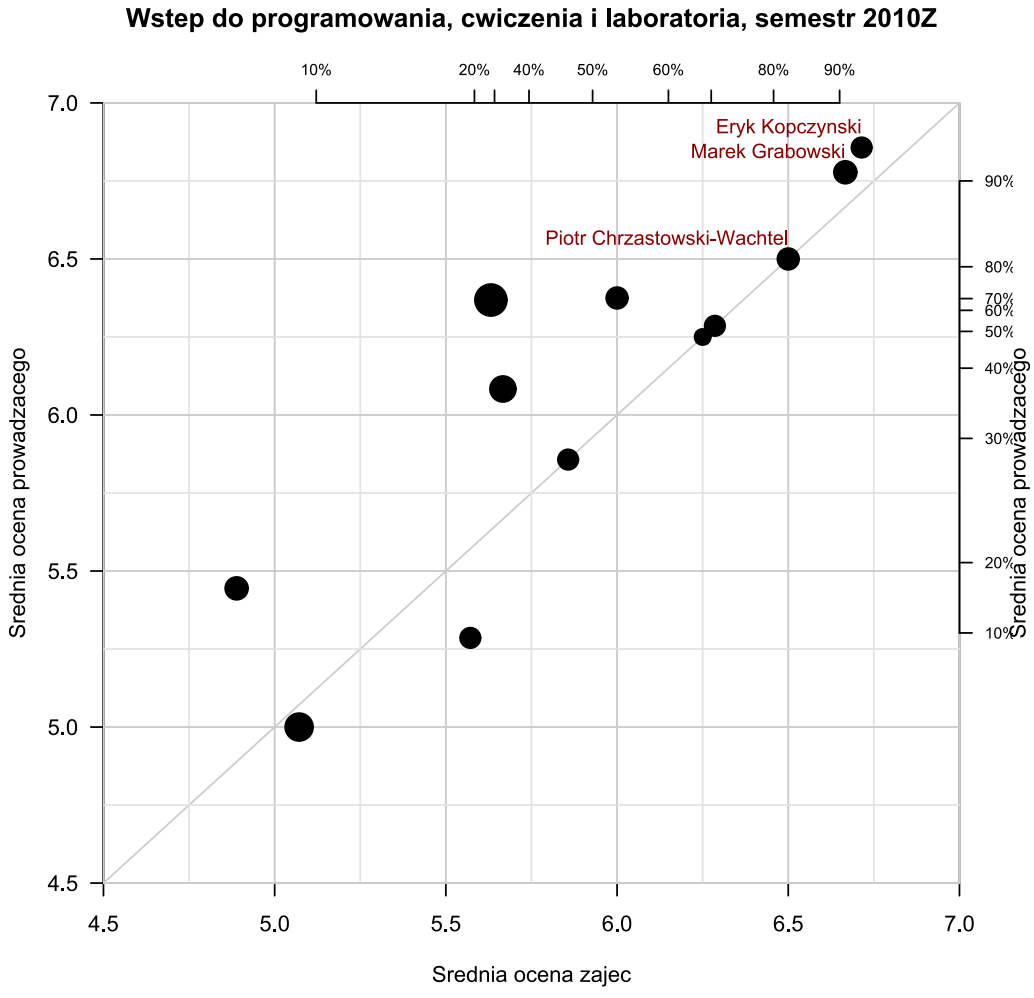
Czy lepiej wybrać miłego prowadzącego czy efektywnego?
Czy mili prowadzący są bardzie efektywni od tych niemiłych?

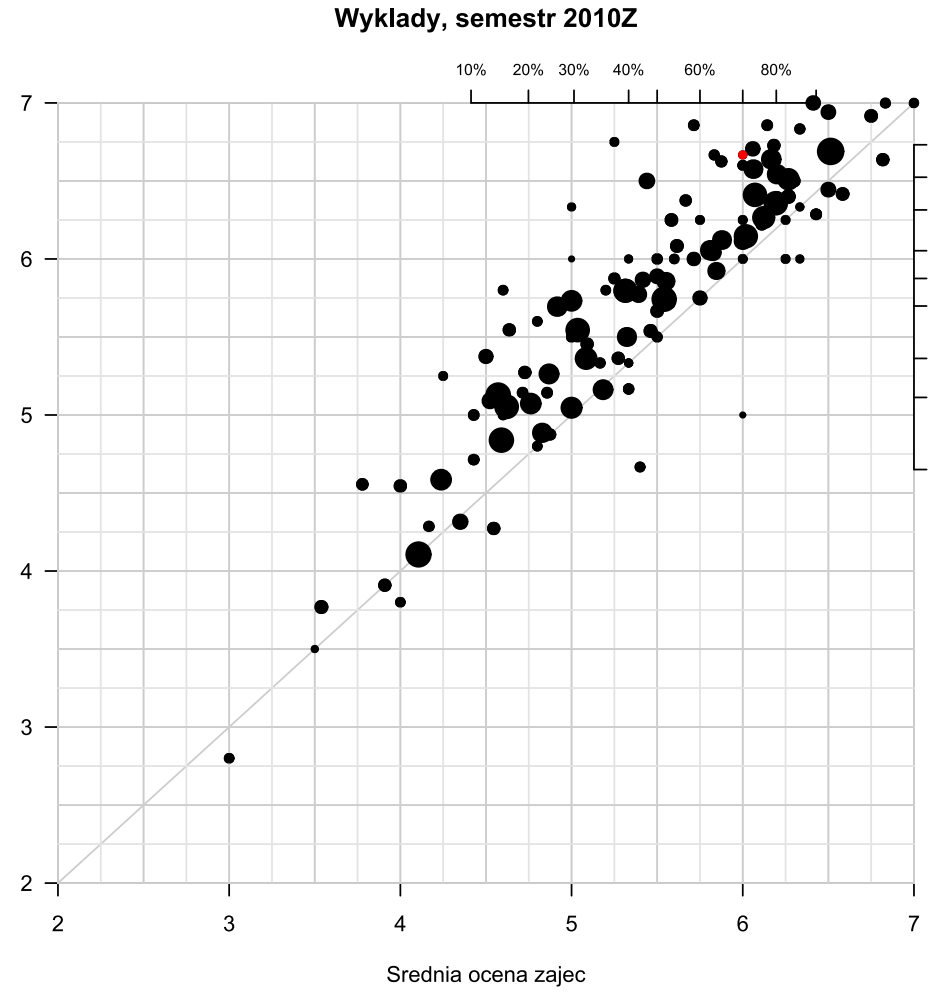




Którego ćwiczeniowca może mi Pani zaproponować?

Czy lepiej oceniamy prowadzącego, czy zajęcia które prowadzi?





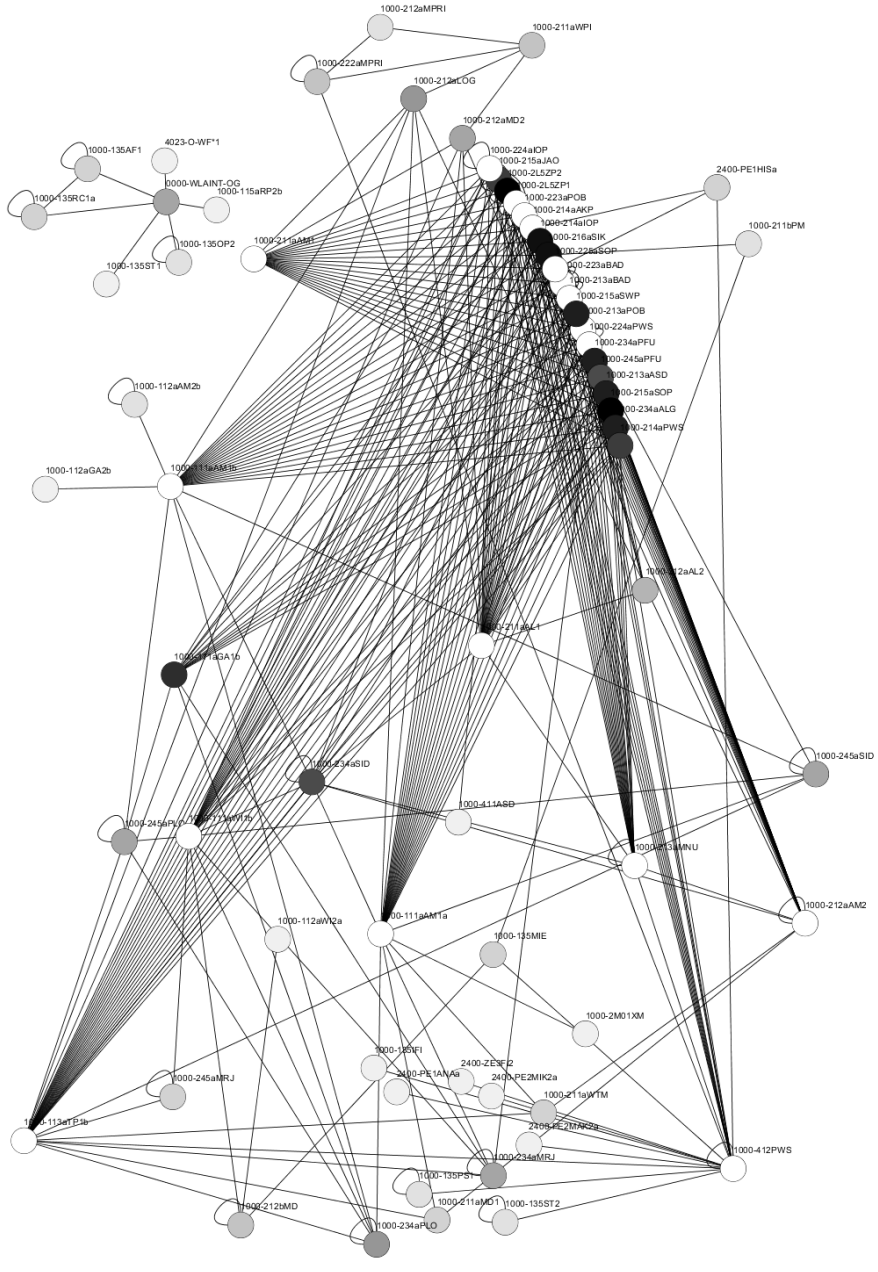
Systemy rekomendacyjne a warunkowe prawdopodobieństwo zaliczenia

W sylabusie można znaleźć informację o kursach wymaganych do zaliczenia danego przedmiotu.

Często jednak spotyka się sytuację, gdy jakiś kurs nie jest formalnie wymagany, ale znacząco ułatwia zaliczenie przedmiotu.

Lub też umiejętności z tego kursu procentują na wiele innych kursów.

Dobrze wiedzieć, które kursy są bardziej ważne.



Konkursik nr 2

Gdybyś ...
... miał dostęp do informacji o popularności wybierania kursów,
... wszystich ocenach studetnów,
... wynikach ankiet,
... danych o rekrutacji (IRKA),
... informacji o pracach dyplomowych (APD).

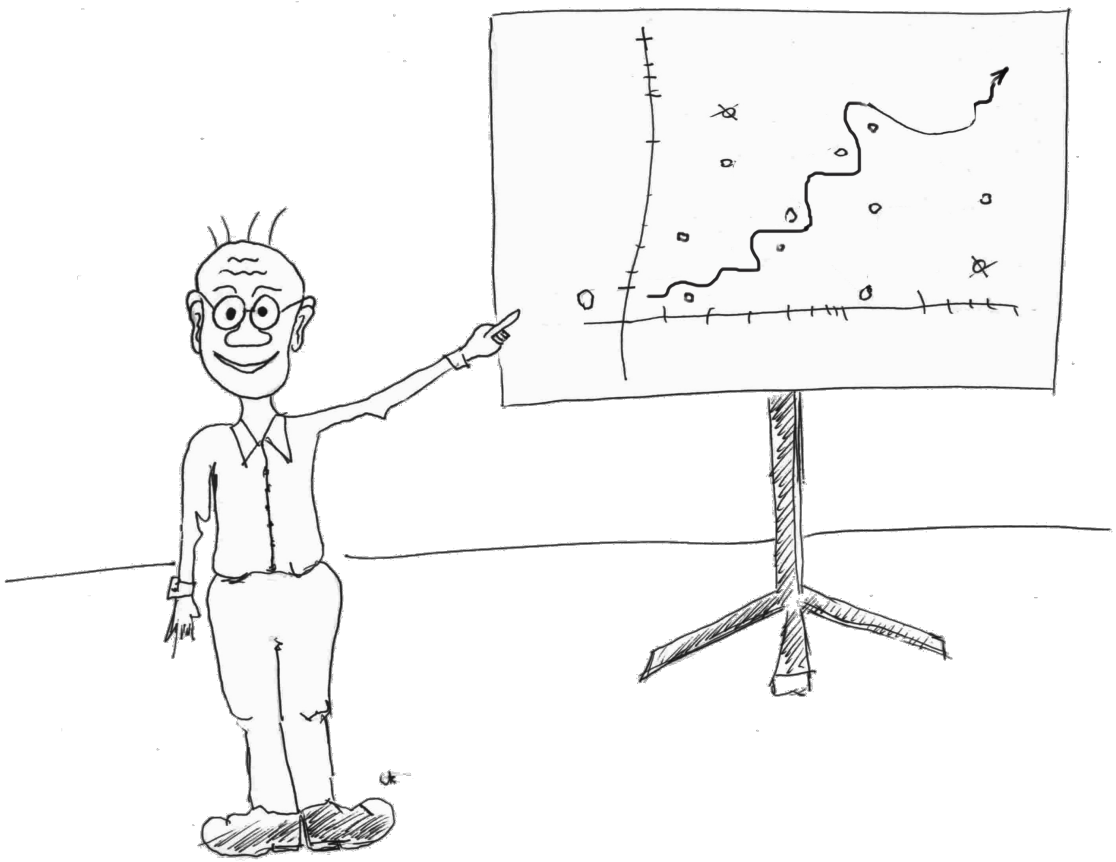
Jaką użyteczną informację potrafiłbyś z tych danych wyciągnąć?

Podejdź do mnie z pomysłem lub wyślij go emailem na adres `Przemyslaw.Biecek@gmail.com`.

Odbierz kubek, 2GB pendrive lub termos*.

*Oferta ważna do wyczerpania zasobów.

Część III. Bioinformatyka a zalew danych



Rozproszona baza danych i tłusty klaster do przetwarzania

Mając:

- Równoległą bazę danych Netezza TwinFin 1000,
- Tłusty klaster obliczeniowy z 12 GPU Tesla M2090 wspieranych kilkunastoma silnymi CPU,
- 20 TB surowych / treningowych danych o odczytach z sekwencerów,

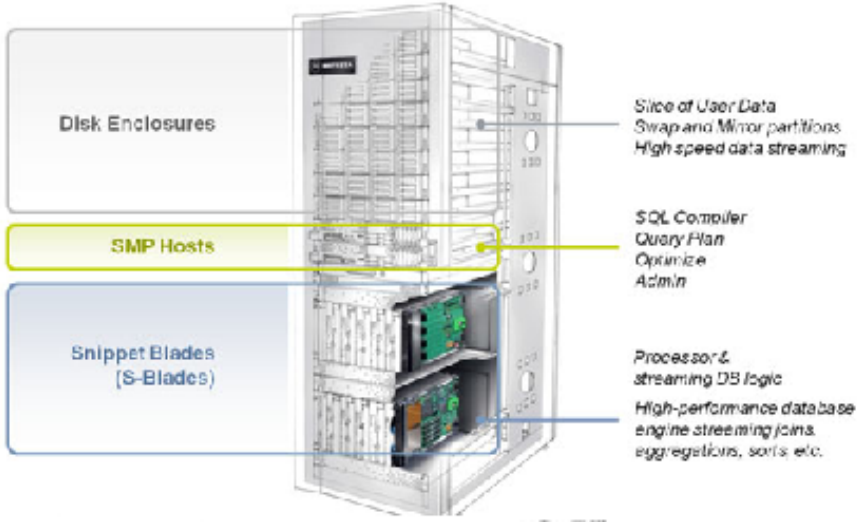
opracuj rozwiązanie pozwalające na znalezienie najlepszego kandydata na biorcę przeszczepu organów, które chirurg właśnie wycina z dawcy.

Sprzętowa architektura opracowanego rozwiązania

operator with the R interface

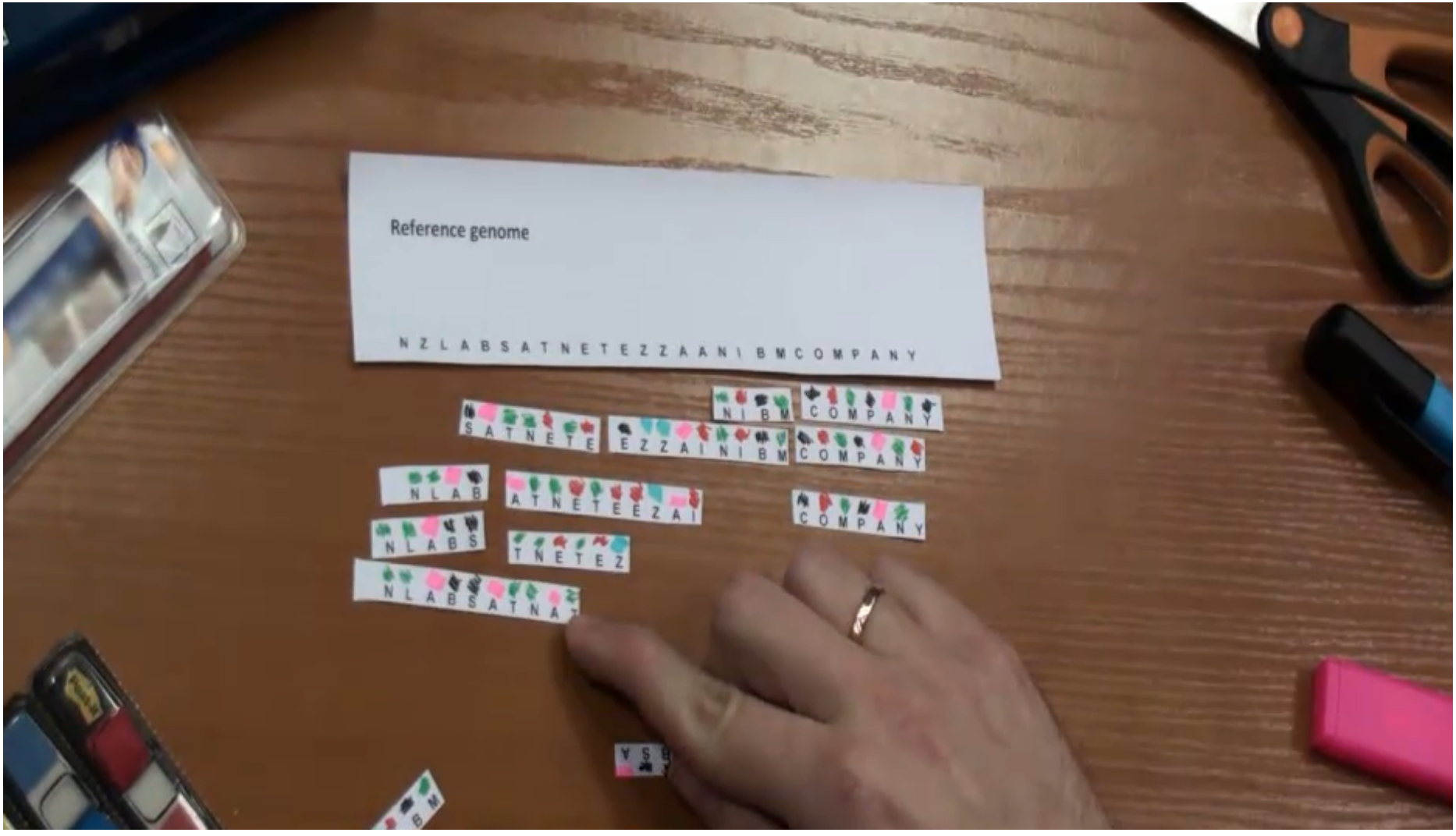
HPC server
(second line analitics)
(GPUs CPUs with lots of RAM)

Netezza Performance Server
(data source and first line analitics)
(share nothing parallel environment)

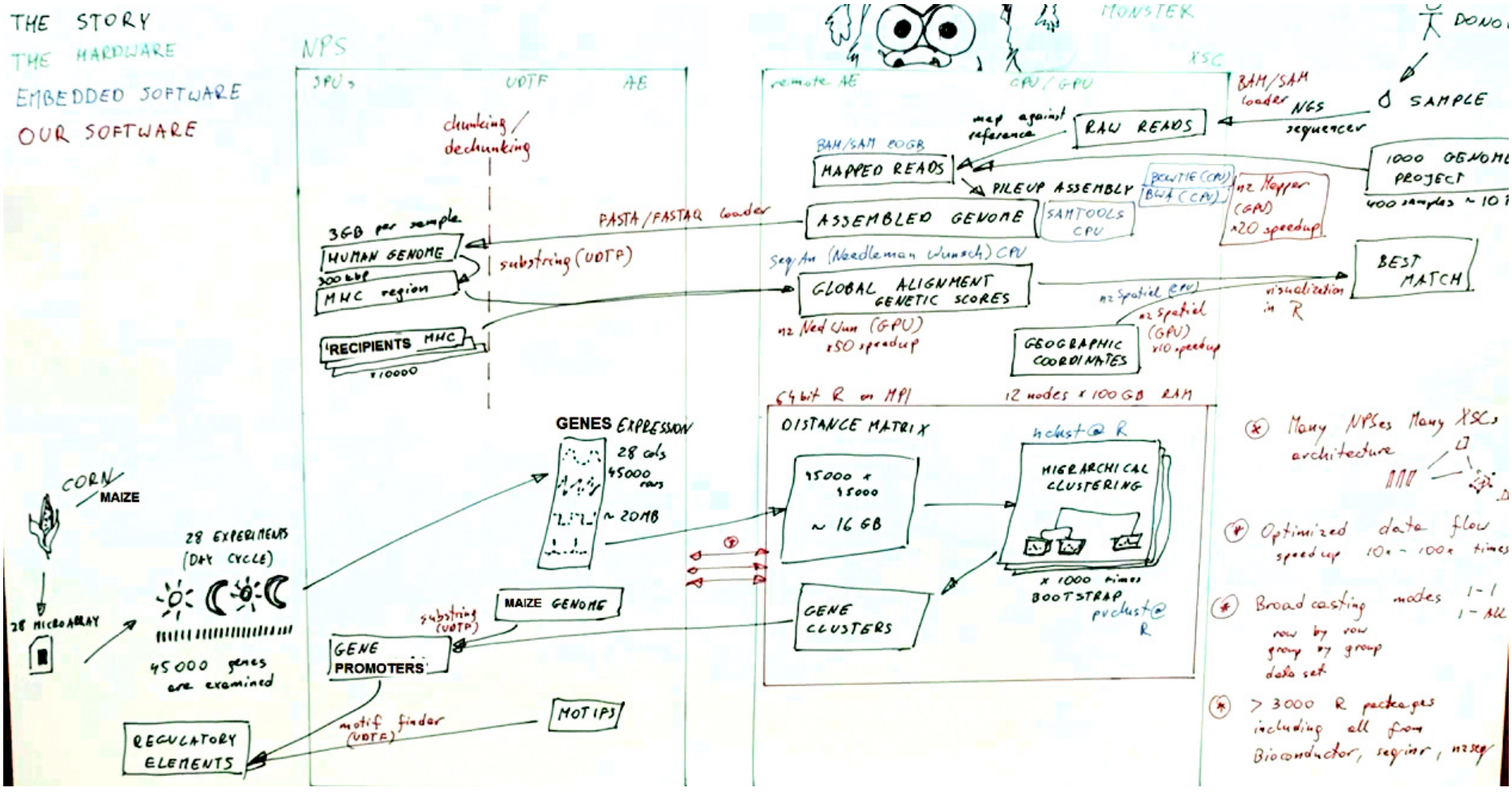


Filmik 1

Krótka historia o szukaniu dawcy i genotypowaniu.



Architektura sprzętowa i przepływ zadań w opracowanym rozwiązaniu.



Życie w oceanie danych: wczoraj, dziś i jutro

8 Studencki Festiwal Informatyczny, 8-10 marca 2012

Przemyslaw.Biecek@gmail.com

IBM Polska / MIM UW / SmarterPoland

Dziękuję organizatorom za zaproszenie.

Dziękuję Wam wszystkim za uwagę.

Czy macie pytania?

Nie zapomnijcie o konkursach!