Zaawansowane Metody Analizy Danych w Biologii Molekularnej

Semest Letni 2018

Prowadzący

Zespół Biologii Obliczeniowej IPI PAN

- dr Magdalena Mozolewska (5 zajęć)
- dr Michal J. Dąbrowski (5 zajęć)
- **dr inż. Michał Dramiński** (5 zajęć)

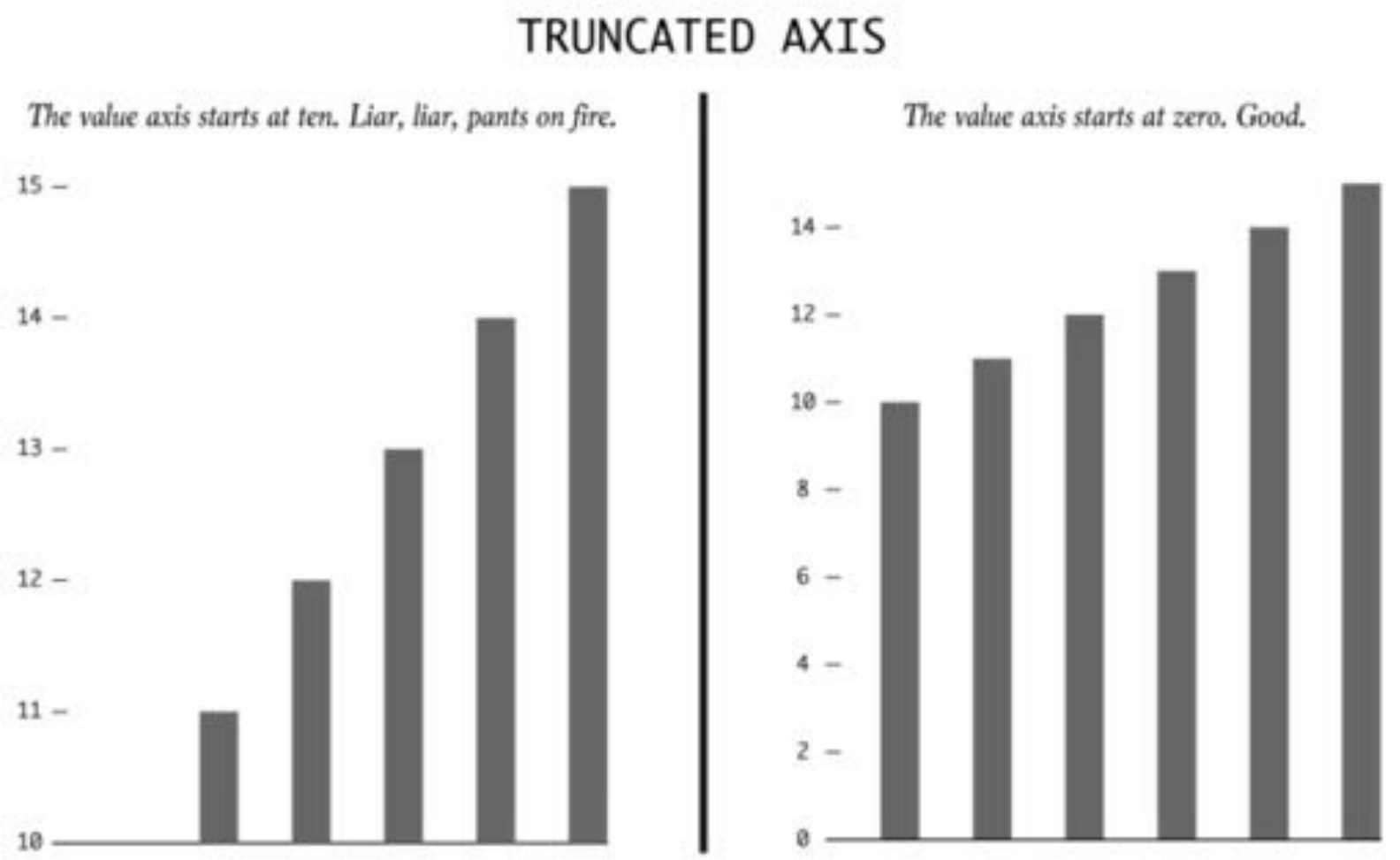
http://zmadbm.ipipan.waw.pl/

Agenda zajęć – R + Analiza danych

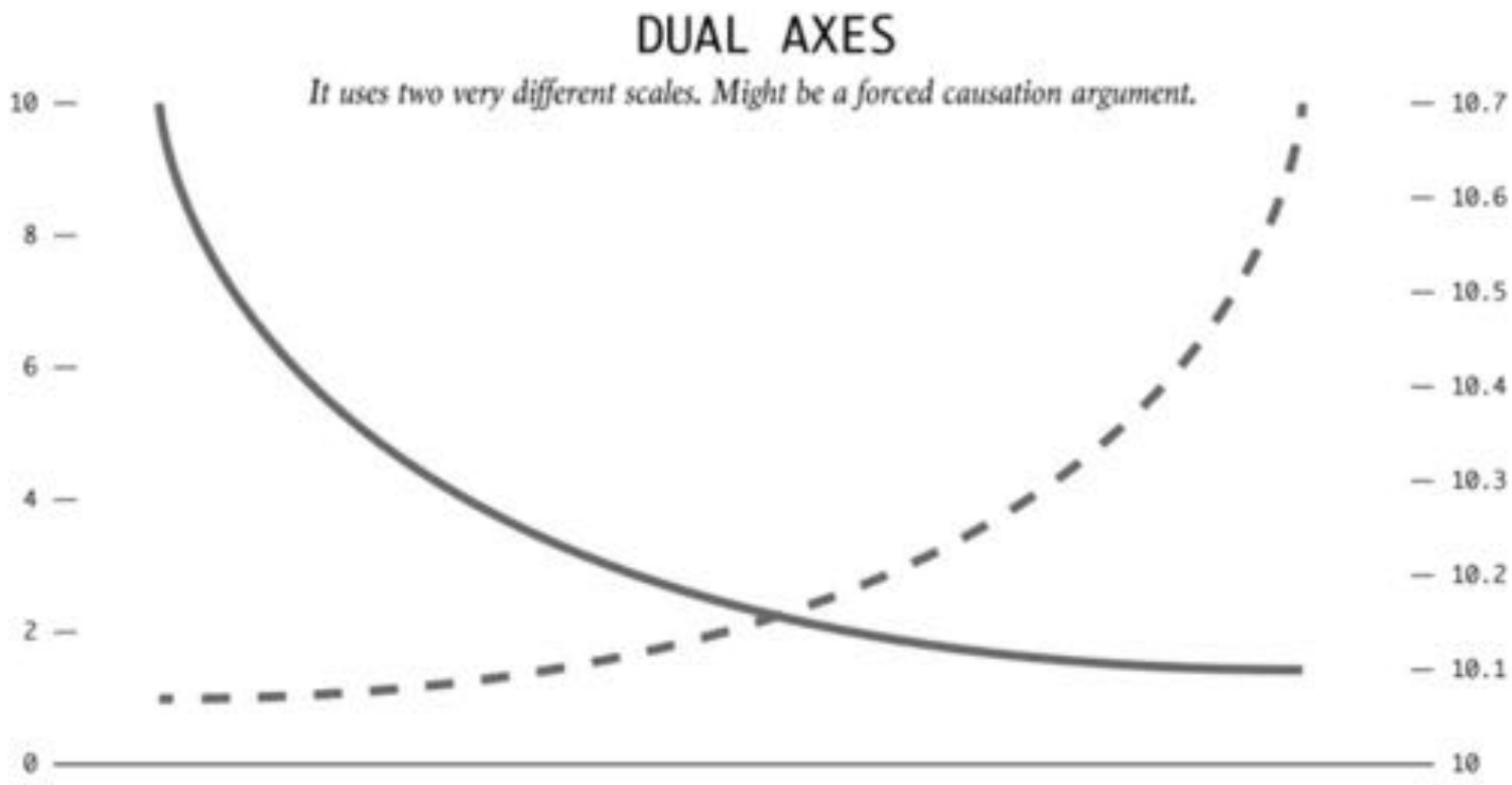
1. Wprowadzenie do przedmiotu. Wprowadzenie do R i RStudio.
2. Wstępna analiza danych. Wprowadzenie do statystyki.
3. Selekcja oraz budowanie nowych cech.
4. Modelowanie danych i analiza wyników selekcji cech.
5. **Pułapki w analizie danych o dużym rozmiarze.**

Dobry styl wykresów

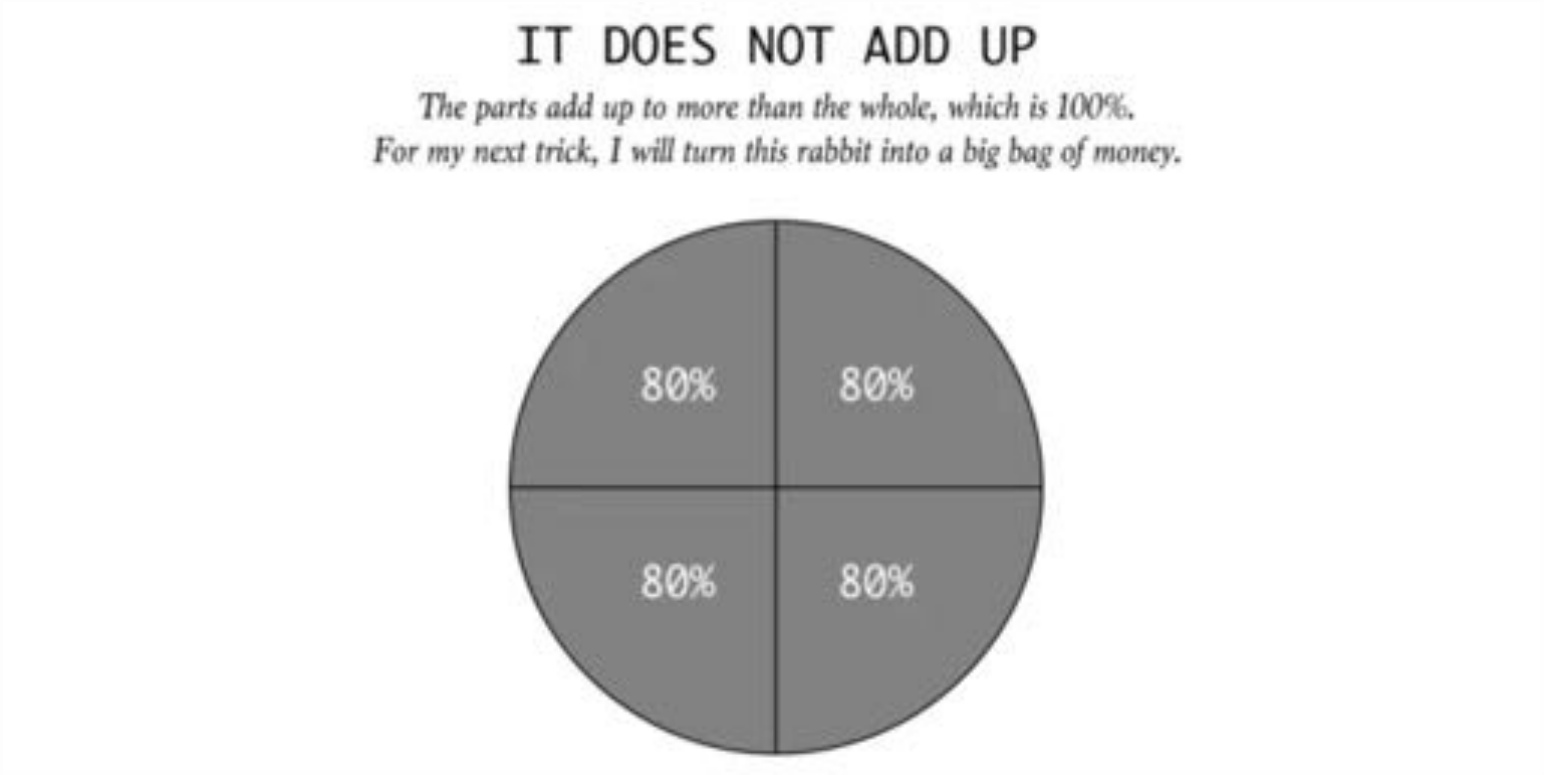
Dobry styl wykresów



Dobry styl wykresów



Dobry styl wykresów



Oraz taka sama dokładność prezentowanych wartości: np.
Tak: 80.3%; 12,4%; 1.1%
Nie: 70%; 12.56%; 2.1%

DEATHS LINKED TO OUTDOOR AND HOUSEHOLD AIR POLLUTION

7 million people die prematurely every year from air pollution – both household and outdoor.

Among these deaths:

21% are due to pneumonia

20% from stroke

34% from ischaemic heart disease

19% from chronic obstructive pulmonary disease (COPD)

7% from lung cancer

CLEAN AIR FOR HEALTH #AirPollution World Health Organization

Dobry styl wykresów

AIR POLLUTION – THE SILENT KILLER

Every year, around **7 MILLION DEATHS** are due to exposure from both outdoor and household air pollution.

Air pollution is a major environmental risk to health. By reducing air pollution levels, countries can reduce:

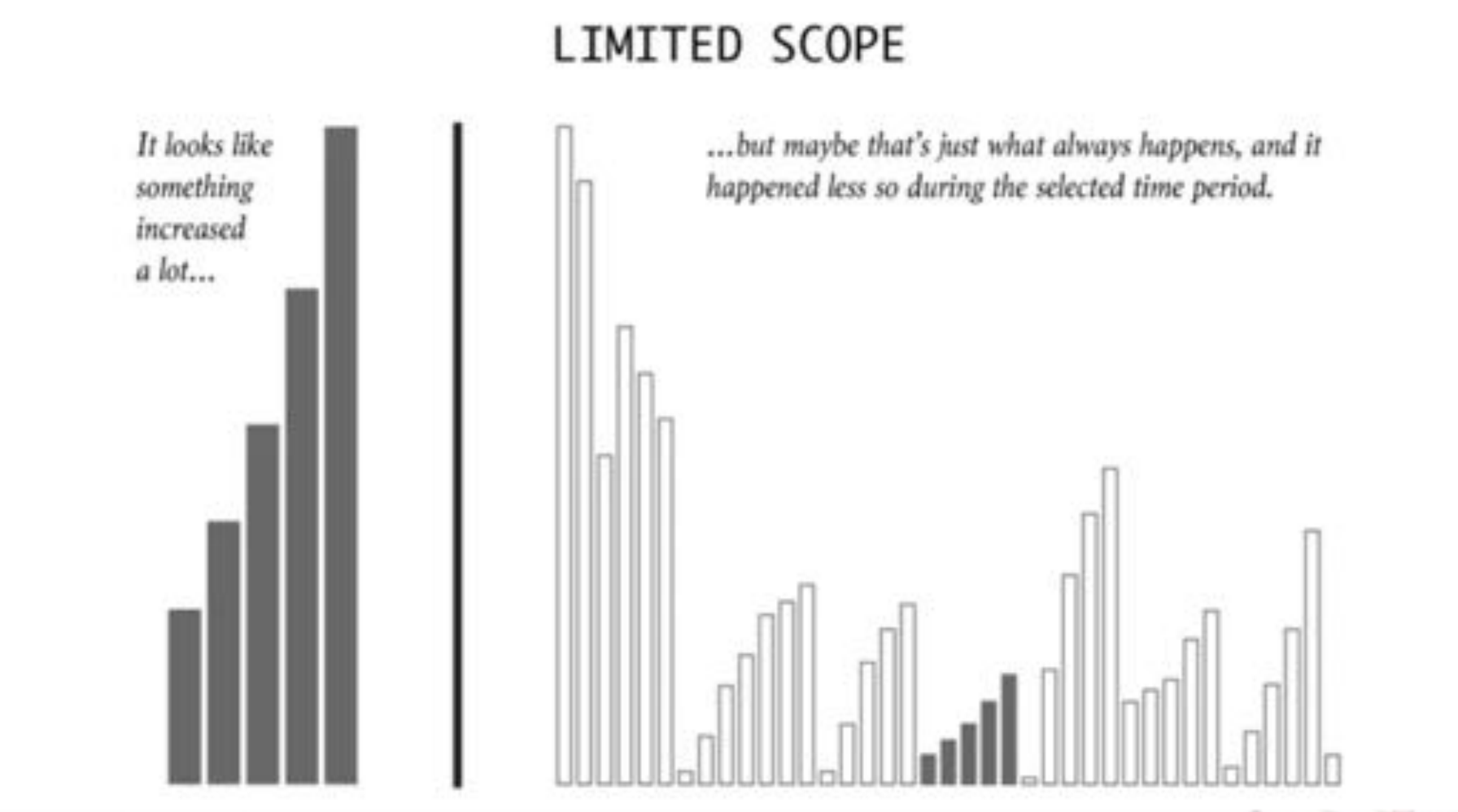
- **Stroke**
- **Heart disease**
- **Lung cancer, and both chronic and acute respiratory diseases, including asthma**

REGIONAL ESTIMATES ACCORDING TO WHO REGIONAL GROUPINGS:

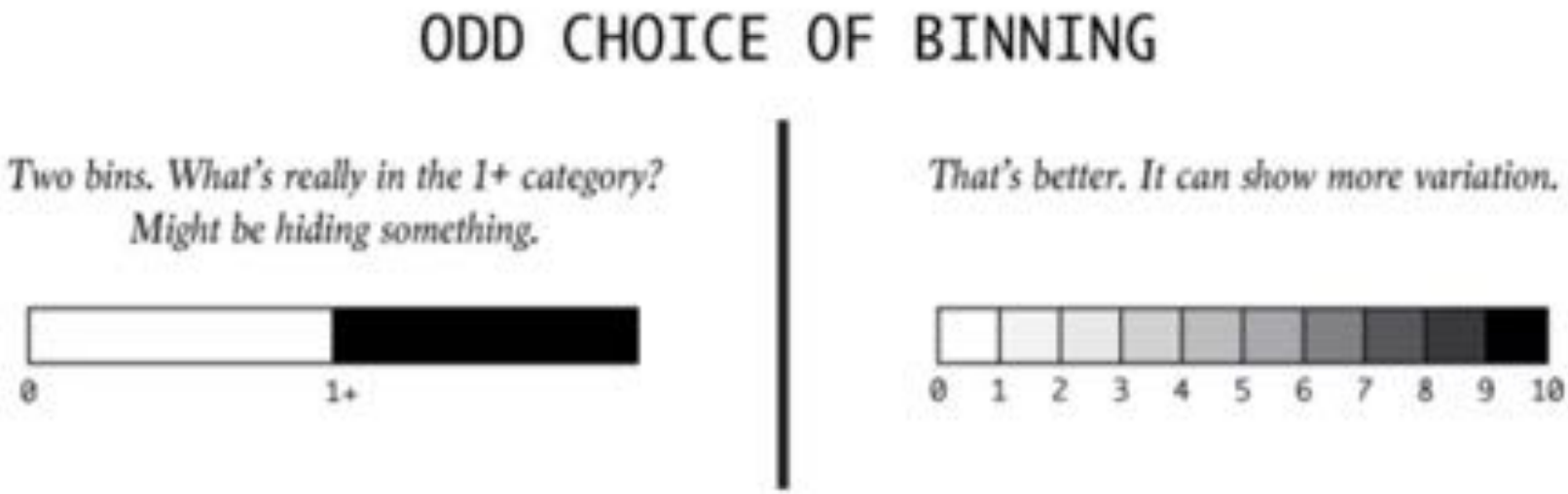
- **Over 2 million** in South-East Asia Region
- **Over 2 million** in Western Pacific Region
- **Nearly 1 million** in Africa Region
- **About 500 000** deaths in Eastern Mediterranean Region
- **About 500 000** deaths in European Region
- **More than 300 000** in the Region of the Americas

CLEAN AIR FOR HEALTH #AirPollution World Health Organization

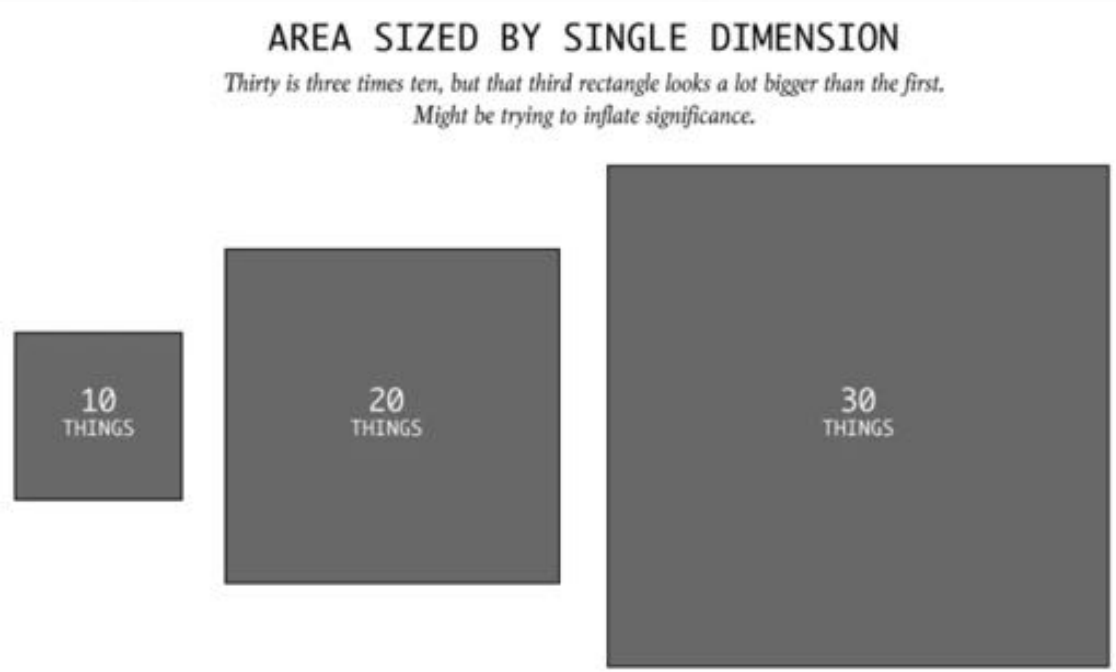
Dobry styl wykresów

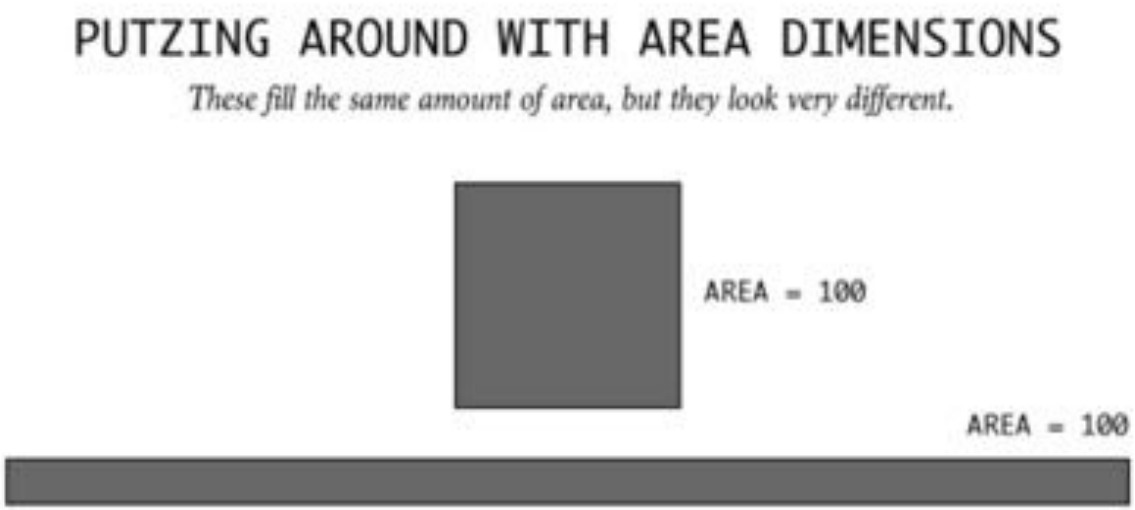


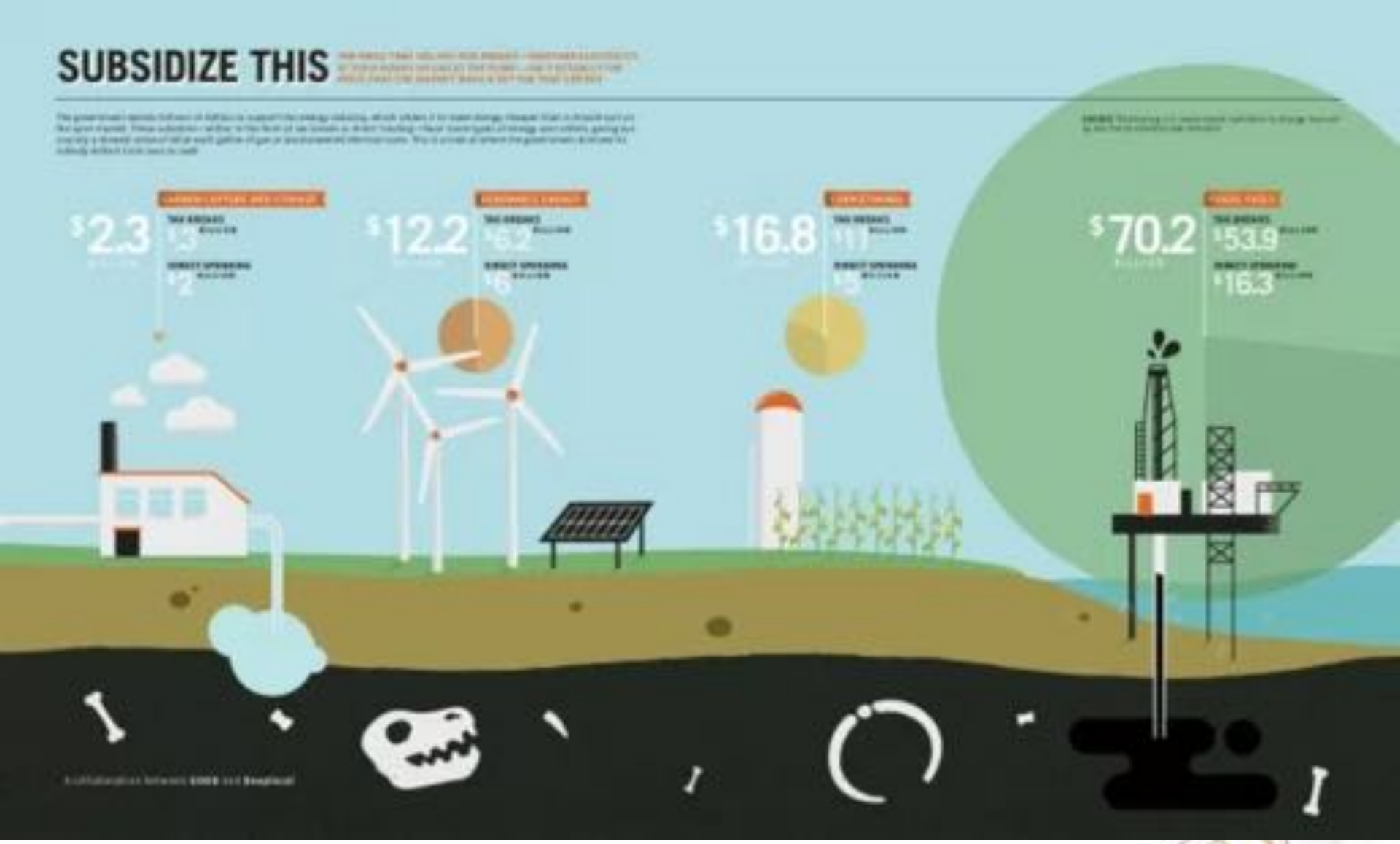
Dobry styl wykresów

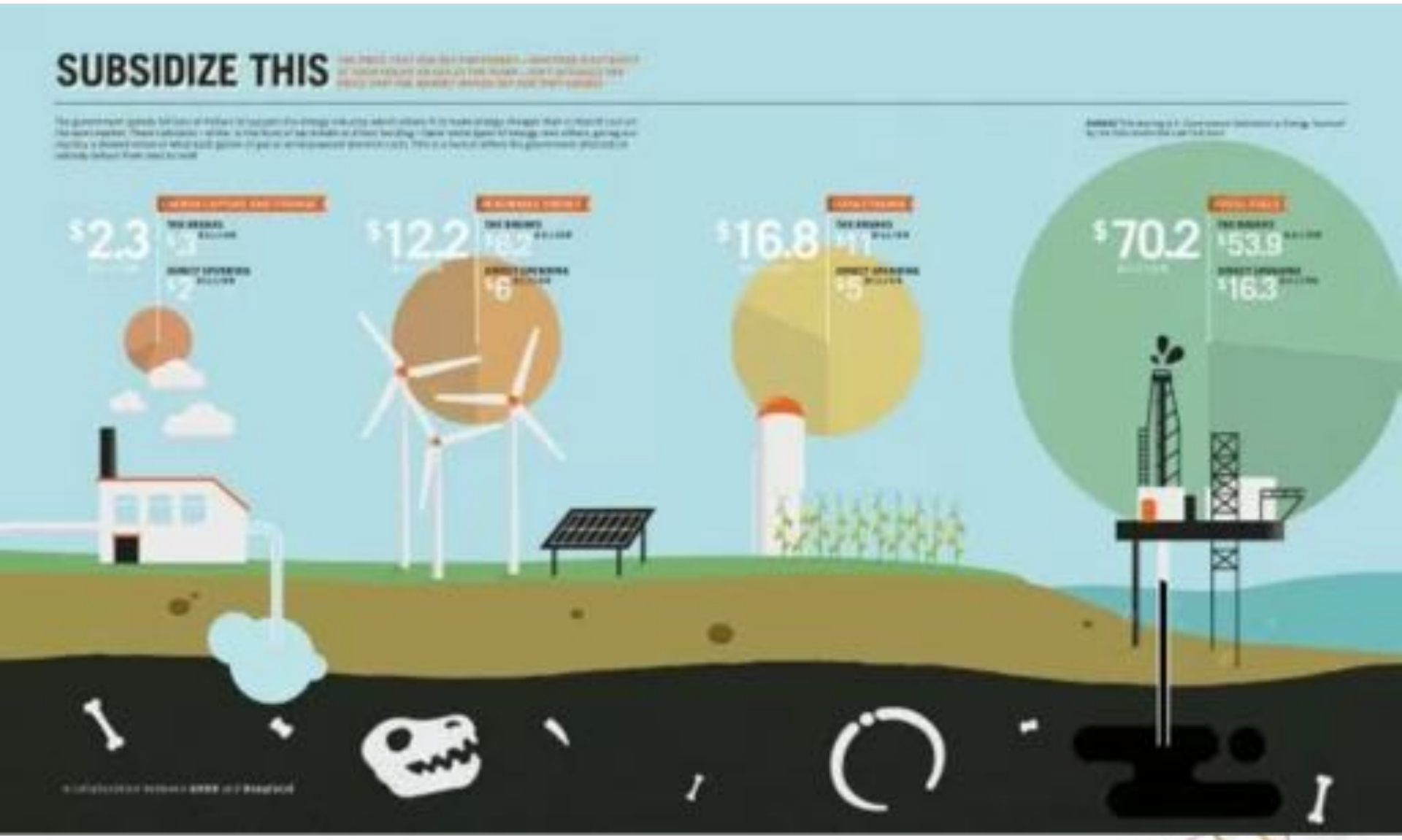


Dobry styl wykresów

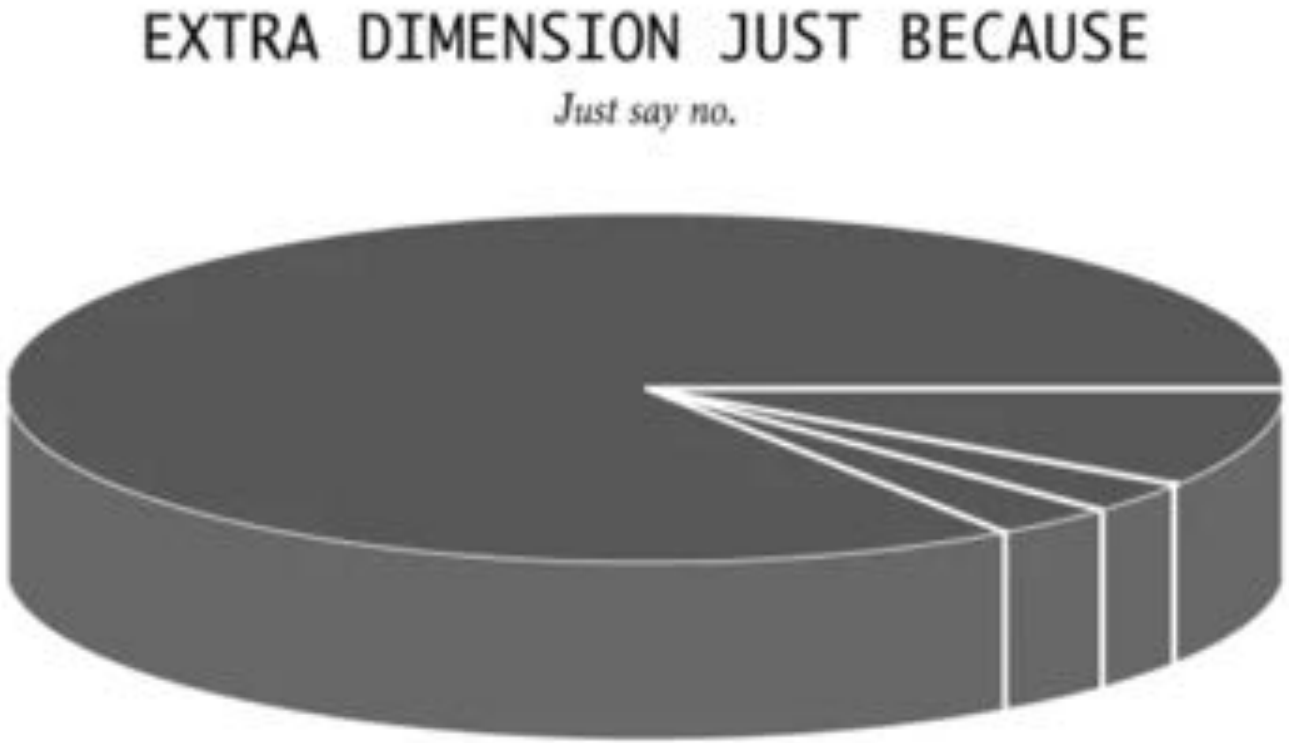





SUBSIDIZE THIS
$2.3
$12.2
$16.8
$70.2
$53.9
$16.3


SUBSIDIZE THIS
$2.3
$12.2
$6
$16.8
$5
$70.2
$53.9
$16.3

Dobry styl wykresów

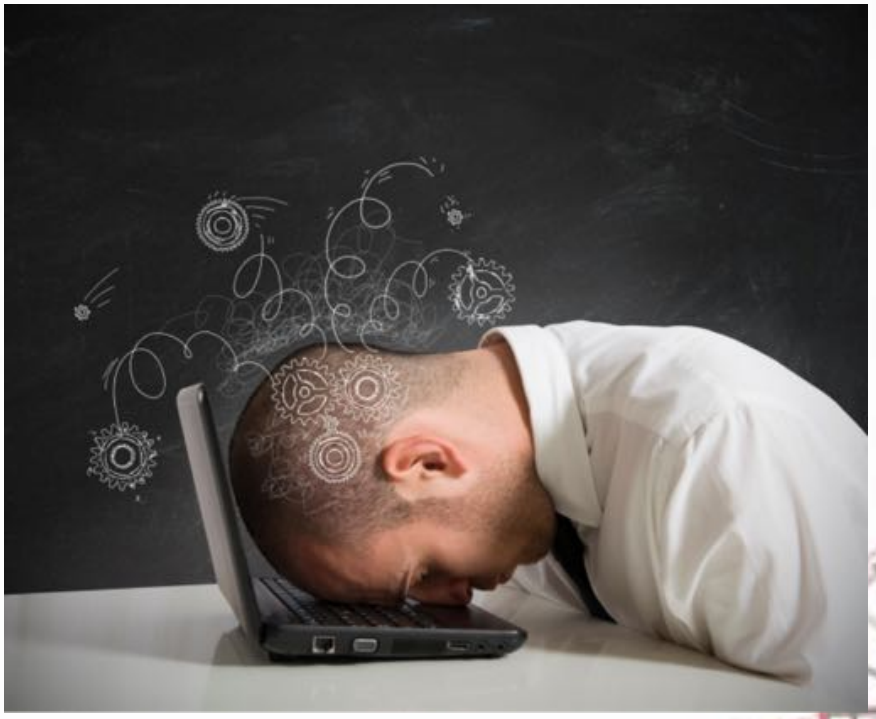


https://flowingdata.com

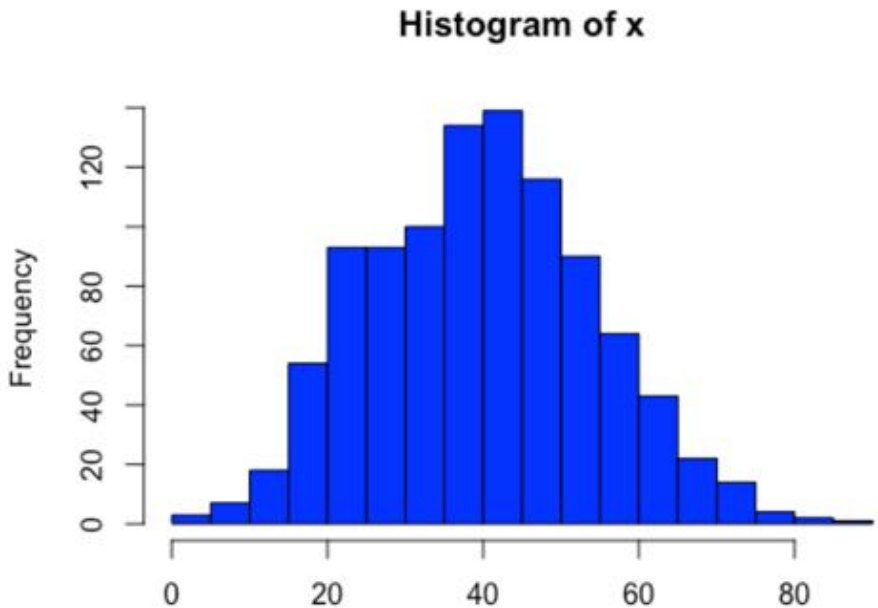
Problemy z danymi

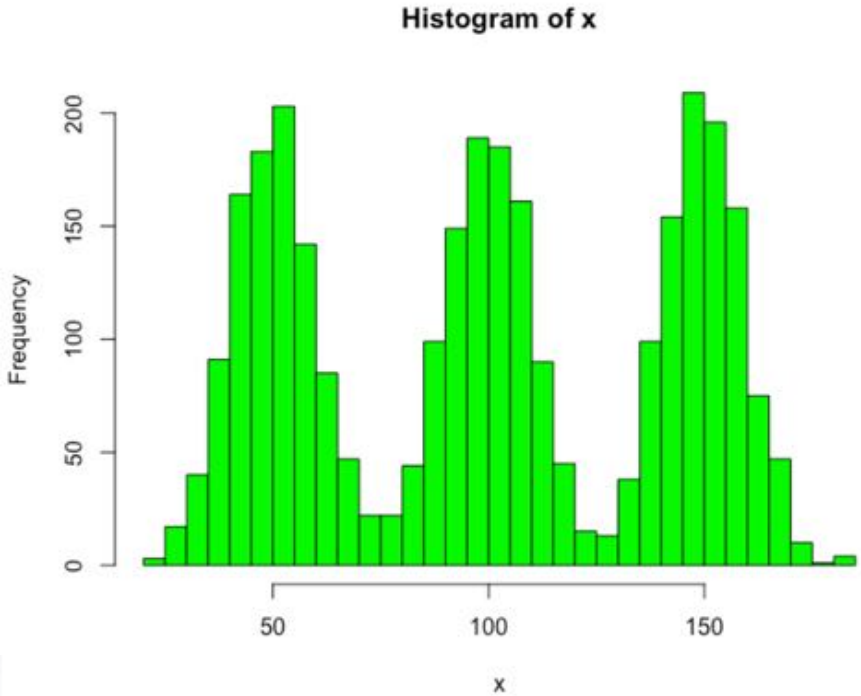
Problemy z danymi

- Brak idealnych rozkładów w świecie realnym
- Niezbalansowane liczności klas
- Braki w danych
- Szum
- Obciążenie



Brak idealnych rozkładów w świecie realnym





Console Markers

~/Dropbox/_PRACA/IPIPAN/ZMADBM/

```
W = 0.99752, p-value = 0.1345

> hist(x,20,col='blue')
> x <- x[x>0]
> hist(x,20,col='blue')
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.99583, p-value = 0.008418

>
```

Braki w danych

- Braki w danych uzupełniamy:
  - Wartością średnią
  - Wartością mediany
  - Najczęstszą wartością
  - Budujemy model do predykcji wartości
  - Uzupełniamy ww metodami w ramach klasy decyzyjnej
  - Uzupełniamy ww metodami w ramach grupy wynikającej z uczenia bez nadzoru
  - Uzupełniamy ww metodami w ramach grupy przykładów najbardziej podobnych (jak kNN)

Szum w danych

- Szum w danych mogą wynikać z:
  - Błędów w pomiarze
  - Błędy ludzkie w przepisywaniu danych np. z formularzy
  - Błędów w konwersji
  - Ograniczenia technologii
- Objawia się on także jako:
  - Outliery
  - Błędne etykiety
  - Redundancje
  - Przypadki graniczne

Obciążenie

- Obciążenie próby to taka cecha próby badawczej, że nie jest ona reprezentatywna dla badanej populacji, pomimo, że została dobrana w sposób losowy. Przyczyną obciążenia próby mogą być rozmaite błędy poczynione w trakcie doboru losowego.
- Przykładowo dane z różnych szpitali. Gdzie każdy szpital może wprowadzać swój własny unikalny rodzaj szumu.
- Dane uzyskane różnymi metodami.

Balans klas

real	prediction
T	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F
F	F

real

prediction

	T	F
T	0	1
F	0	19

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \qquad \text{(a)}$$

$$\text{Overall Accuracy} = \frac{(TP + TN)}{(P + N)} \qquad \text{(b)}$$

Accuracy = 19/20 = 95%

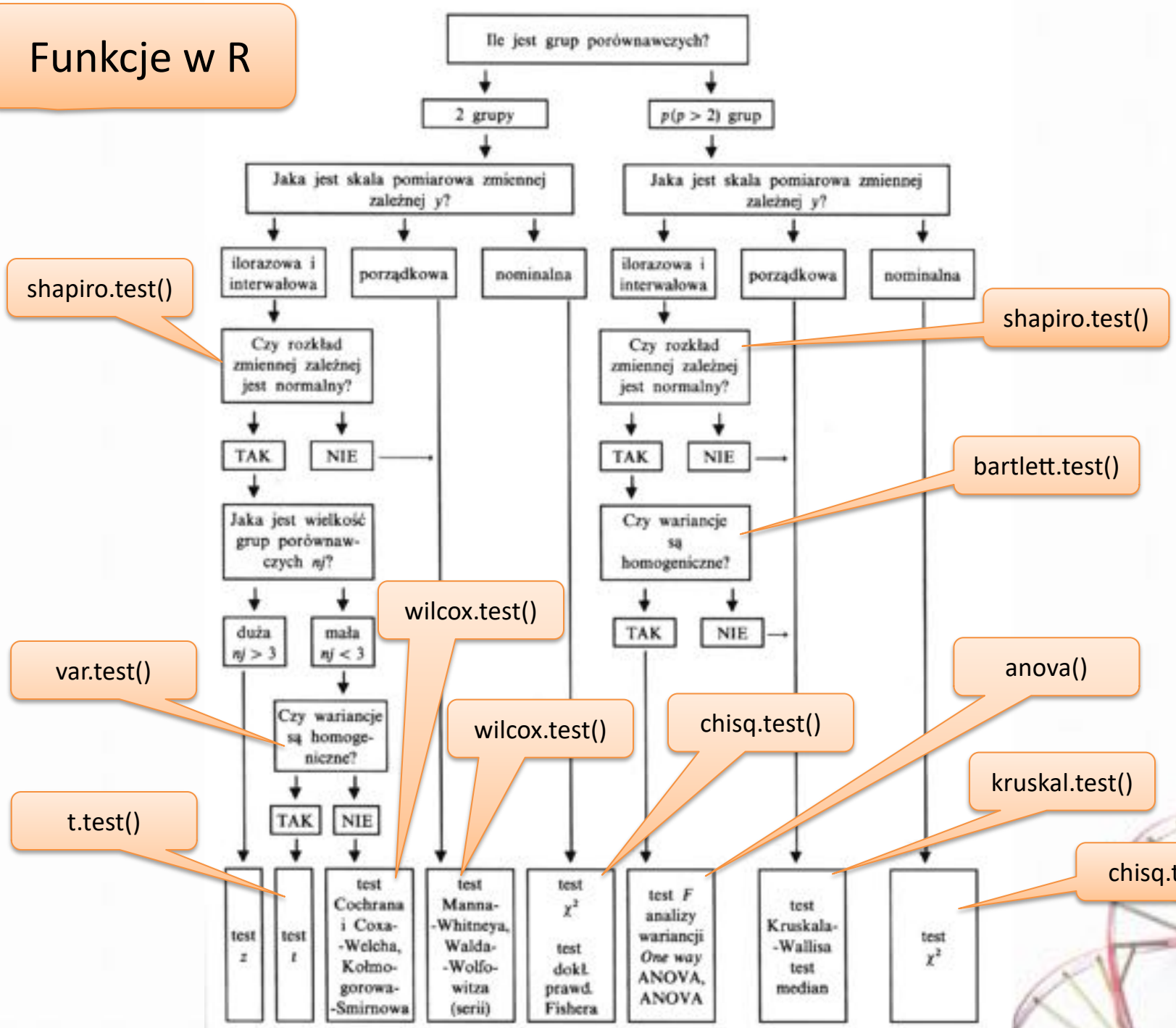
Balanced Accuracy = ½ (0/1 + 19/19) = 50.0%

Problemy w analizie

Problemy w analizie

- Aplikacja odpowiednich testów statystycznych do rozkładów
- Z korelacji nie wynika przyczynowość skutkowość
- Ograniczenia modeli
- Problem nadmiernego dopasowania
- Standaryzacja
- Wielokrotne testowanie
- Problemy dla wysokowymiarowych danych

Funkcje w R

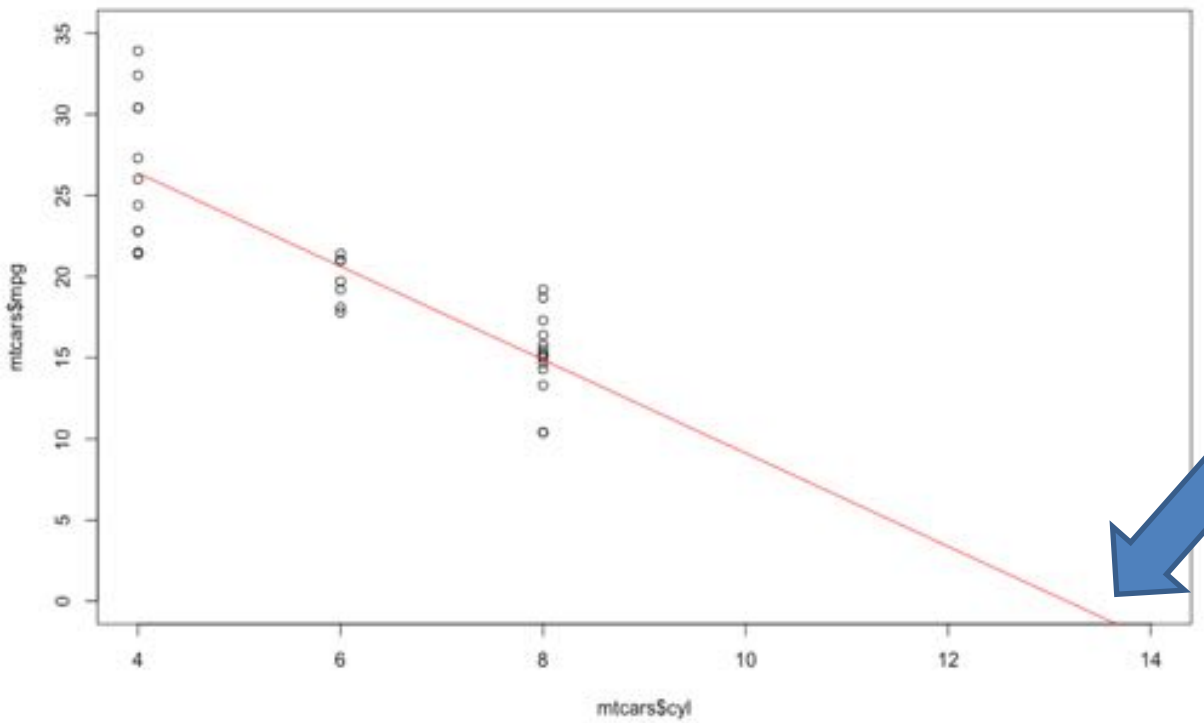

Ile jest grup porównawczych?
2 grupy
p(p > 2) grup
Jaka jest skala pomiarowa zmiennej zależnej y?
ilorazowa i interwałowa
porządkowa
nominalna
Czy rozkład zmiennej zależnej jest normalny?
TAK
NIE
Jaka jest wielkość grup porównawczych nj?
duża nj > 3
mała nj < 3
Czy wariancje są homogeniczne?
test z
test t
test Cochrana i Coxa-Welcha, Kołmogorowa-Smirnowa
test Manna-Whitneya, Walda-Wolfowitza (serii)
test χ² test dokł. prawd. Fishera
test F analizy wariancji One way ANOVA, ANOVA
test Kruskala-Wallisa test median
test χ²
shapiro.test()
shapiro.test()
bartlett.test()
wilcox.test()
var.test()
wilcox.test()
chisq.test()
anova()
kruskal.test()
t.test()
chisq.test()

Zawiłości korelacji

- Z korelacji nie wynika przyczynowość skutkowość:
  - Obserwujemy wysoką korelację pomiędzy liczbą książek w domu a ilorazem inteligencji dzieci w nim wzrastających.
  - Czy to oznacza że wystarczy kupić bardzo dużo książek do domu i mamy zapewnione bardzo inteligentne potomstwo?
  - NIE! ☺

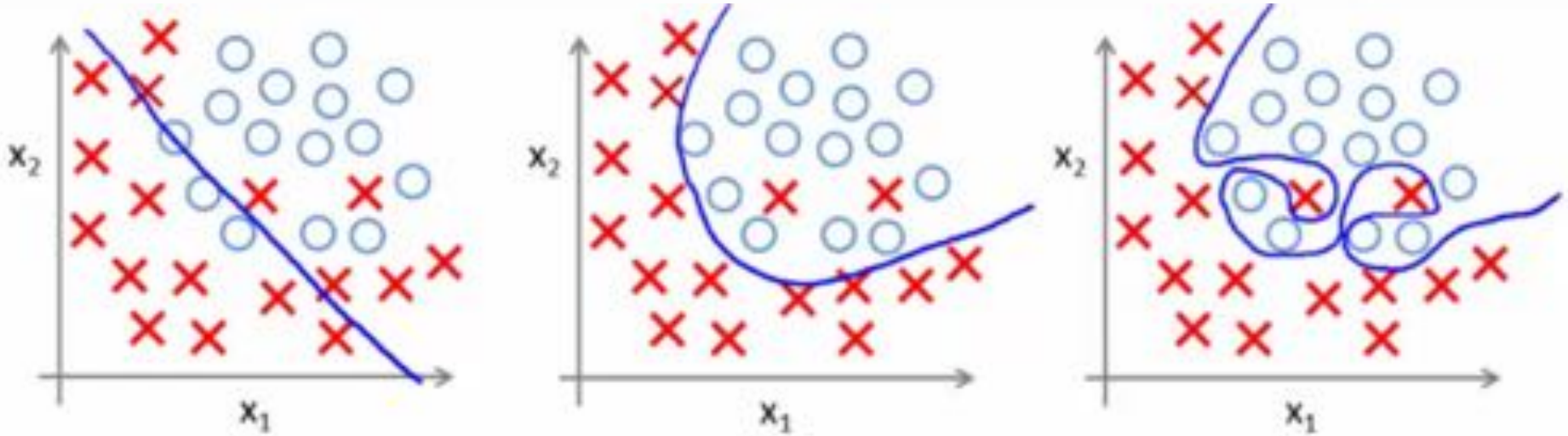
Ograniczenia modeli

- Regresja liniowa mpg (miles per galon) vs liczba cylidrów. Model liniowy: y = 37.885 - 2.876 * x
- Z modelu wynika że: ~13.16 cylindrów nie pozwoli na przejechanie więcej niż zero mil ☺



Nadmierne dopasowanie

- ang. overfitting



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g = sigmoid function)

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \ldots$

UNDERFITTING
(high bias)

OVERFITTING
(high variance)

Normalizacja i standaryzacja

- Grupujemy ludzi w.g.
  - Wzrostu
  - Wagi
  - Płci (1=M, 2 = F)
  - Kolor włosów (jasne = 1, ciemne = 2, brak = 0)
- Każda ze zmiennych ma inny zakres wartości oraz rozkład!
- W przypadku używiania k-means + miary Euklidesowej **MUSIMY** dane odpowiednio przygotować!
- W przeciwnym razie wartości duże o dużej dynamice zdominują miarę odległości.

Standaryzacja (centrowanie)

- **Normalizacja** – procedura wstępnej obróbki danych w celu umożliwienia ich wzajemnego porównywania i dalszej analizy [wikipedia].
- **Standaryzacja** – rodzaj normalizacji zmiennej losowej, w wyniku której zmienna uzyskuje średnią wartość oczekiwaną zero i odchylenie standardowe jeden. Od zmiennej odejmujemy wartość średnią i następnie dzielimy przez odchylenie standardowe.

$$z = \frac{x - \mu}{\sigma}$$

- W R standaryzację realizuje funkcja: `scale()`

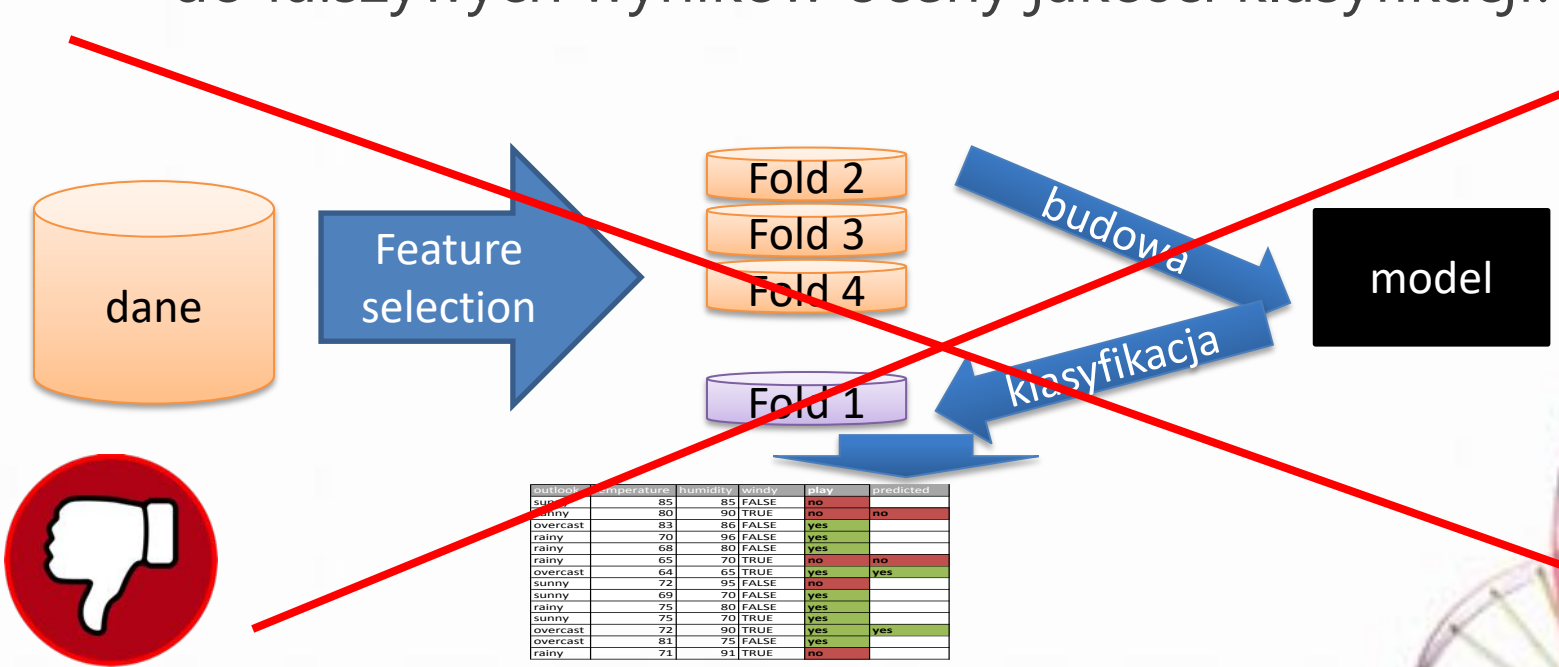
Wielokrotne testowanie

- Poprawka Bonferroniego (klasyka) inne np. Holma–Bonferroniego lub procedura Benjaminiego–Hochberga.
- Statystyczna procedura przeciwdziałania problemowi porównań wielokrotnych, polegające na zmniejszeniu poziomu istotności każdego z przeprowadzanych testów wprost proporcjonalnie do ich ogólnej liczby.
- Na przykład, przy wykonywaniu 5 powiązanych testów z poprawką Bonferroniego, należy użyć w każdym przypadku krytycznej wartości *alfa* aby całe badanie zachowało poziom istotności 0,05.

$$\alpha = \frac{0,05}{5} = 0,01,$$

Wysoki wymiar danych

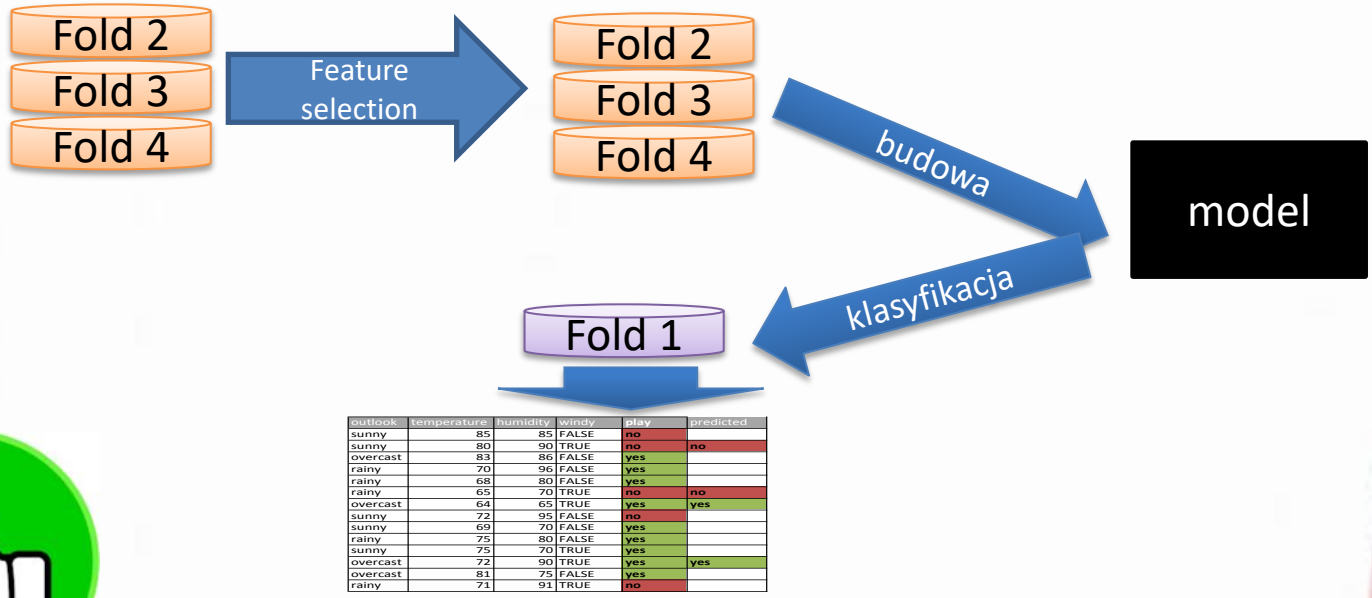
- Wszelkie manipulacje danych (np. selekcja cech, dyskretyzacja, uzupełnianie etc.) należy wykonywać w pętli walidacji krzyżowej na zbiorze trenującym.
- Wszelkie nadzorowane operacje na danych są elementem uczenia!
- Dokonywanie tych operacji na pełnym zbiorze prowadzi do fałszywych wyników oceny jakości klasyfikacji!



outlook	temperature	humidity	windy	play	predicted
sunny	85	85	FALSE	no	
sunny	80	90	TRUE	no	no
overcast	83	86	FALSE	yes	
rainy	70	96	FALSE	yes	
rainy	68	80	FALSE	yes	
rainy	65	70	TRUE	no	no
overcast	64	65	TRUE	yes	yes
sunny	72	95	FALSE	no	
sunny	69	70	FALSE	yes	
rainy	75	80	FALSE	yes	
sunny	75	70	TRUE	yes	
overcast	72	90	TRUE	yes	yes
overcast	81	75	FALSE	yes	
rainy	71	91	TRUE	no	

Wysoki wymiar danych

- **Wszelkie manipulacje danych (np. selekcja cech, dyskretyzacja, normalizacja etc.) należy wykonywać w pętli walidacji krzyżowej na zbiorze trenującym.**



outlook	temperature	humidity	windy	play	predicted
sunny	85	85	FALSE	no	
sunny	80	90	TRUE	no	no
overcast	83	86	FALSE	yes	
rainy	70	96	FALSE	yes	
rainy	68	80	FALSE	yes	
rainy	65	70	TRUE	no	no
overcast	64	65	TRUE	yes	yes
sunny	72	95	FALSE	no	
sunny	69	70	FALSE	yes	
rainy	75	80	FALSE	yes	
sunny	75	70	TRUE	yes	
overcast	72	90	TRUE	yes	yes
overcast	81	75	FALSE	yes	
rainy	71	91	TRUE	no	

R dla bardziej zaawansowanych

dplyr

```
>library(dplyr)
#filter do selekcji wierszy przy użyciu warunków
>filter(mtcars, cyl == 4, gear == 5)
 mpg cyl  disp  hp drat    wt qsec vs am gear carb1
26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    22
30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
>filter(mtcars, cyl == 4 | gear == 5)
#slice do selekcji wierszy przy użyciu indeksu
>slice(mtcars, 2:5)
mpg cyl disp  hp drat    wt  qsec vs am gear carb1
21.0   6  160 110 3.90 2.875 17.02  0  1    4    42
22.8   4  108  93 3.85 2.320 18.61  1  1    4    13
21.4   6  258 110 3.08 3.215 19.44  1  0    3    14
18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
#ostatni wiersz
>slice(mtcars, n())
```

dplyr

```
#do sortowania funkcja arrange
>arrange(mtcars, cyl, desc(gear))
#do wybierania kolumn select
>select(mtcars, cyl, gear)
#znajdz unikalne kombinacje wartości
>distinct(mtcars, cyl, gear)
#dodaj nową kolumnę
>mutate(mtcars, l_km = 282.48/mpg)
#wybierz losowo wiersze
>sample_n(mtcars, 5)
#dwie tabele można łatwo łączyć funkcjami join
>inner_join(x, y, by = 'nazwa_kolumny')
>left_join(x, y, by = 'nazwa_kolumny')
>right_join(x, y, by = 'nazwa_kolumny')
>full_join(x, y, by = 'nazwa_kolumny')
```

dplyr

```
#siła dplyra polega na potokowym przetwarzaniu
>new_mtcars <- mtcars %>%
      mutate(l_km = 282.48/mpg) %>%
      group_by(cyl, gear) %>%
      summarise(total.count=n(),
            avg_mpg = mean(mpg, na.rm = TRUE),
            avg_l_km = mean(l_km, na.rm = TRUE)) %>%
            filter(avg_l_km > 15) %>%
      data.frame
>new_mtcars
 cyl gear total.count avg_mpg avg_l_km
1   8    3          12   15.05 19.44137
2   8    5           2   15.40 18.35524
```

Ćwiczenia 13

Ćwiczenia 13

1. Wygeneruj zbiór 1000 numerycznych losowych kolumn (runif(n,1,10)) i jednej nominalnej (10A+10B). Całość jako data.frame o 1001 kolumnach i 20 wierszach **(1pkt)**.
2. Na całym zbiorze przykładów dokonaj selekcji cech i dla topowych 10 zbuduj dowolny klasyfikator. Sprawdź jego jakość klasyfikacji (predict) na zbiorze danych **(2pkt)**.
3. Podziel losowo dane na dwa równe podzbiory 10+10 przykładów. Na pierwszym trenującym dokonaj selekcji cech dowolnie wybraną metodą i dla topowych 10 zmiennych zbuduj dowolny klasyfikator. Sprawdź jego jakość na drugim zbiorze (testującym) nie użytym do trenowania klasyfikatora i selekcji cech **(3pkt)**.
4. Napisz wnioski z ćwiczenia**(4pkt)**.