EKSPLORACJA DANYCH

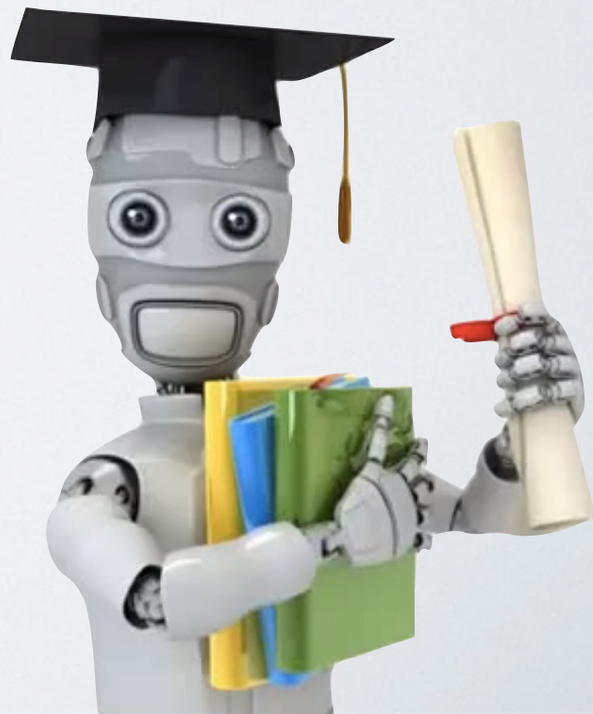
Statystyka biomedyczna

Artur Klepaczko

MAŁE ENTRÉ

- Eksploracja danych – dziedzina informatyki stosowanej obejmująca algorytmy przetwarzania informacji w celu rozpoznawania istniejących lub konstrukcji nowych reguł (potencjalnie trudnych do sformalizowania) pomiędzy cechami opisującymi obiekty rzeczywistości.

- Ze względu na różne rozłożenie akcentów, eksploracja danych może być także nazywana:
  - uczenie maszynowe (ang. *machine learning*),
  - odkrywanie wiedzy (ang. *knowledge discovery*),
  - rozpoznawanie wzorców (ang. *pattern recognition*).



CO TO SĄ DANE?

- **Dane tworzą opis obiektów świata rzeczywistego**, wyrażony za pomocą ich **cech** (atrybutów, parametrów).
- **Obiektami** mogą być ludzie, przedmioty lub zjawiska fizyczne oraz ich reprezentacje, np. obrazy.
- Cechy mogą być **jakościowe** lub **ilościowe** – tak, jak w przypadku zmiennych losowych.
- Zbiór cech obiektu stanowi odpowiadający mu **wektor danych**.
- Kolekcja wektorów danych zmierzonych dla serii obiektów stanowi **zbiór danych**.

Dlaczego mówimy
o eksploracji danych
w ramach przedmiotu
Statystyka biomedyczna?

EKSPLORACJA DLA INŻYNIERII BIOMEDYCZNEJ

- **Predykcja** rozwoju choroby w badaniach przesiewowych na podstawie szeregu charakterystyk.

- **Detekcja** obszarów tkanki zmienionych chorobowo w obrazach biomedycznych (np. detekcja owrzodzeń na ścianach jelita cienkiego w obrazach endoskopowych).

- **Wspomaganie diagnostyki** medycznej i procesu podejmowania decyzji (np. o tym, czy mamy do czynienia z nowotworem złośliwym, czy niezłośliwym).

- **Rozpoznawanie** zespołów genów o wspólnej funkcji w badaniach ekspresji i funkcji wielu (tzn. tysięcy) genów.

STATYSTYKA DLA EKSPLORACJI

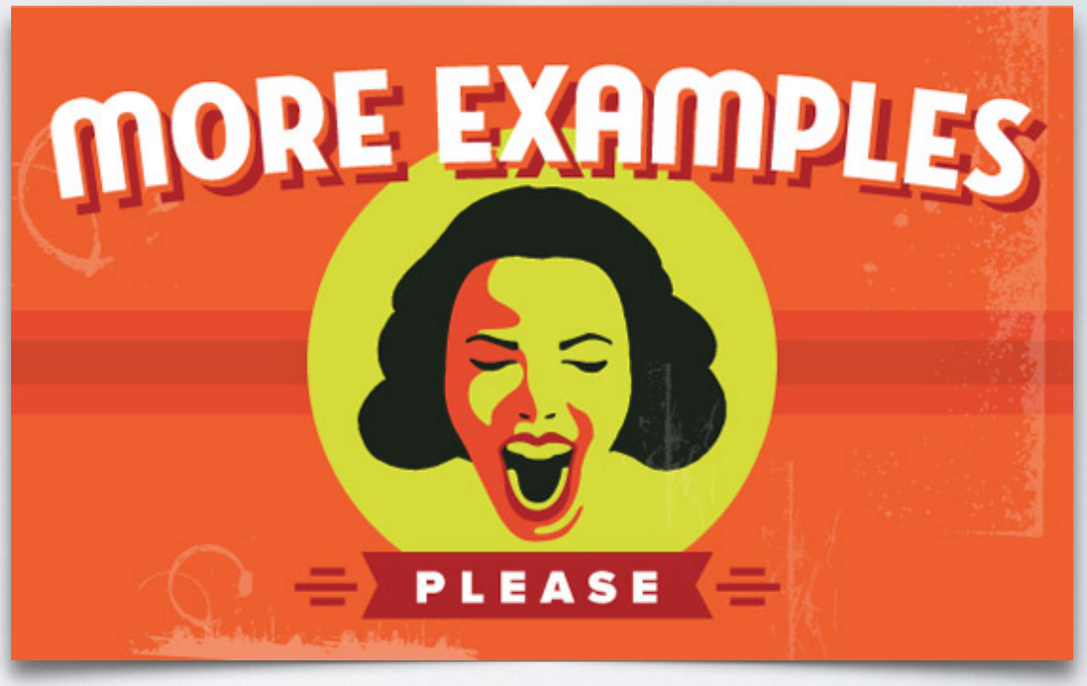
- Rozpoznawanie cech znaczących i relacji między cechami – analiza regresji, analiza korelacyjna, analiza wariancji, analiza dyskryminacyjna.

- Klasyfikacja – regresja logitowa, twierdzenie Bayesa o prawdopodobieństwie warunkowym.

- Ocena jakości klasyfikacji – rozkład Bernouliego, szacowanie przedziału ufności.

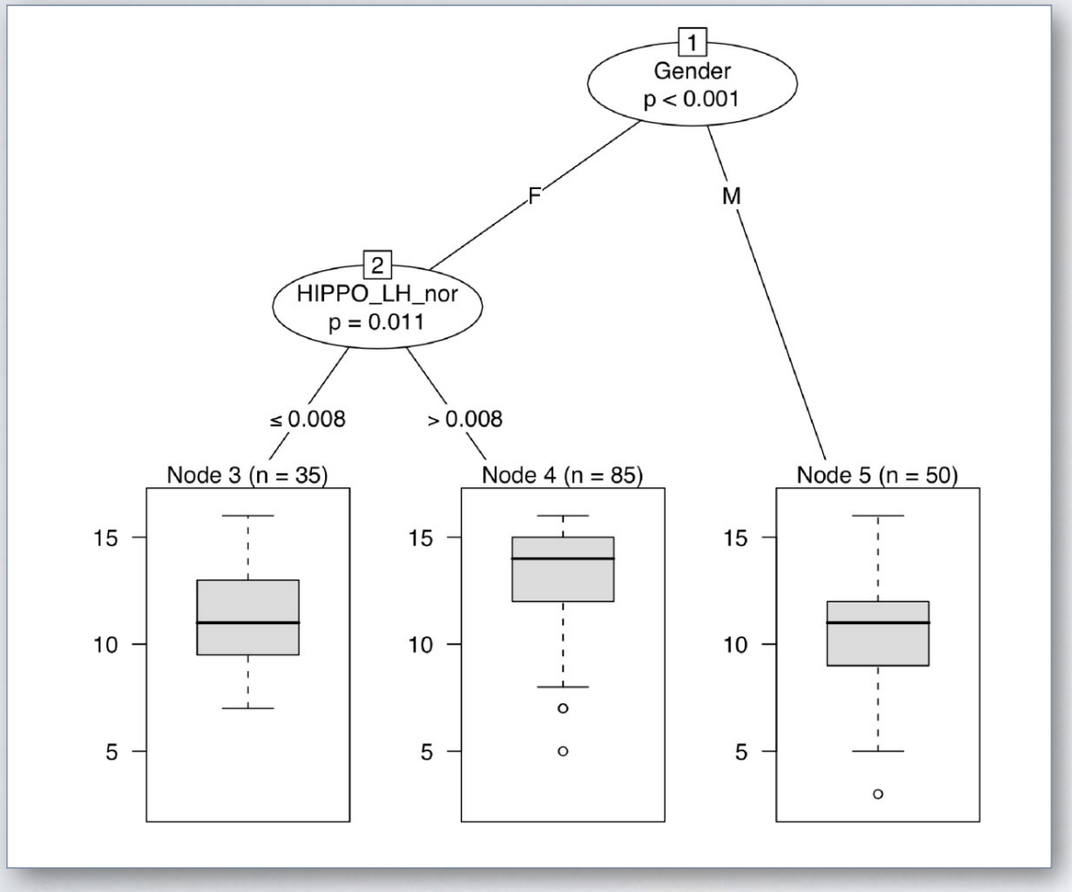
- Porównywanie klasyfikatorów (sparowany test Studenta).

Yyy… To może jakieś
przykłady?



ANALIZA PROCESÓW STARZENIA

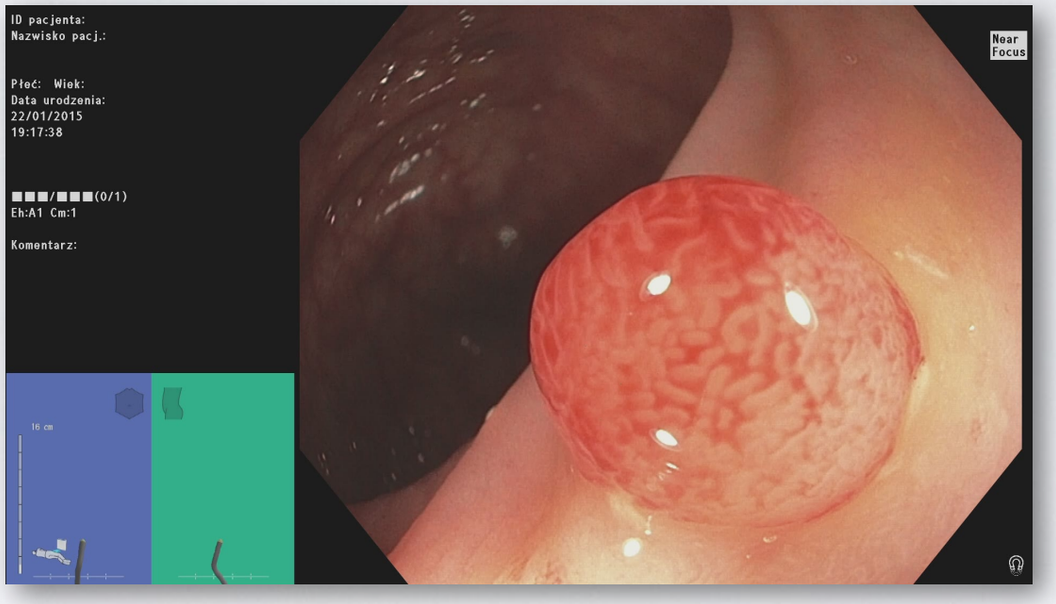
Drzewo wnioskowania warunkowego



M. Ystad et al., *Hippocampal volumes are important predictors for memory function in elederly women*, BMC Biomedical Imaging, 9(1), 2009.

- Odpowiedź: wynik testu neuropsychologicznego na zdolność zapamiętywania mowy.
- Zmienne towarzyszące: płeć, wiek, objętość hippokampa, objętość lewego i prawego hippokampa (oddzielnie), ekspresja genu ApoE**ε**4.
- Wykryto małą, ale znaczącą zależność pomiędzy wynikami testu a całkowitą objętością hippokampa.

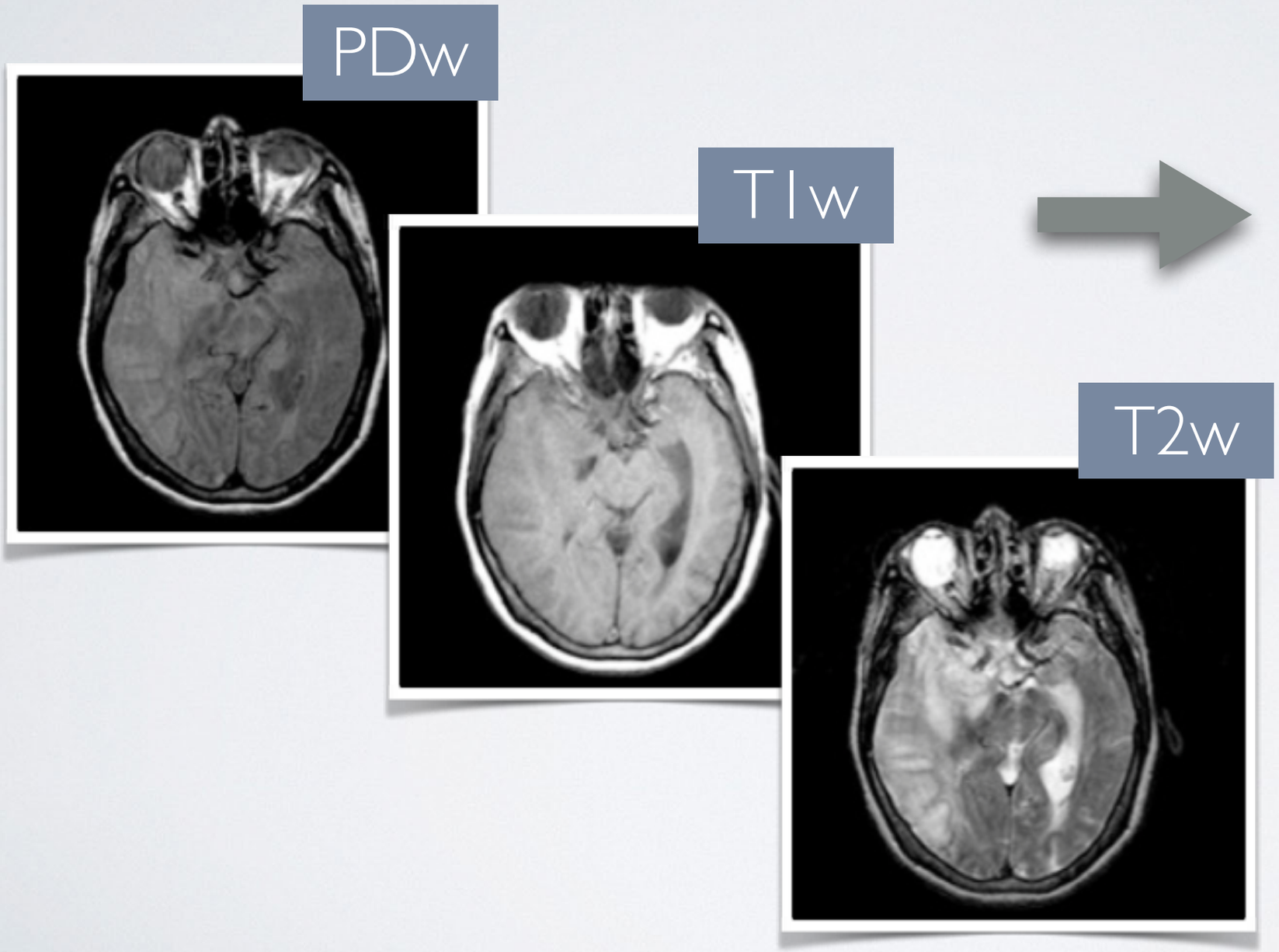
DETEKCJA POLIPÓW W OBRAZACH KOLONOSKOPOWYCH

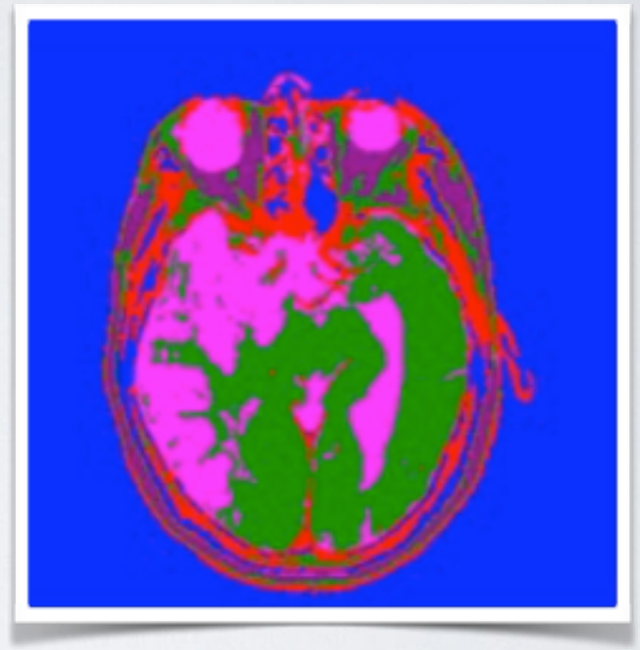


Ekstrakcja cech – analiza tekstury w obszarach zainteresowania

Klasyfikacja obszarów zainteresowania

SEGMENTACJA TKANKI ZMIENIONEJ CHOROBOWO





Analiza skupień – automatyczny pomiar objętości mózgu w obszarze niedokrwienia

Skoro mamy mówić o algorytmach wykrywania wiedzy, to jakie są metody jej reprezentacji?

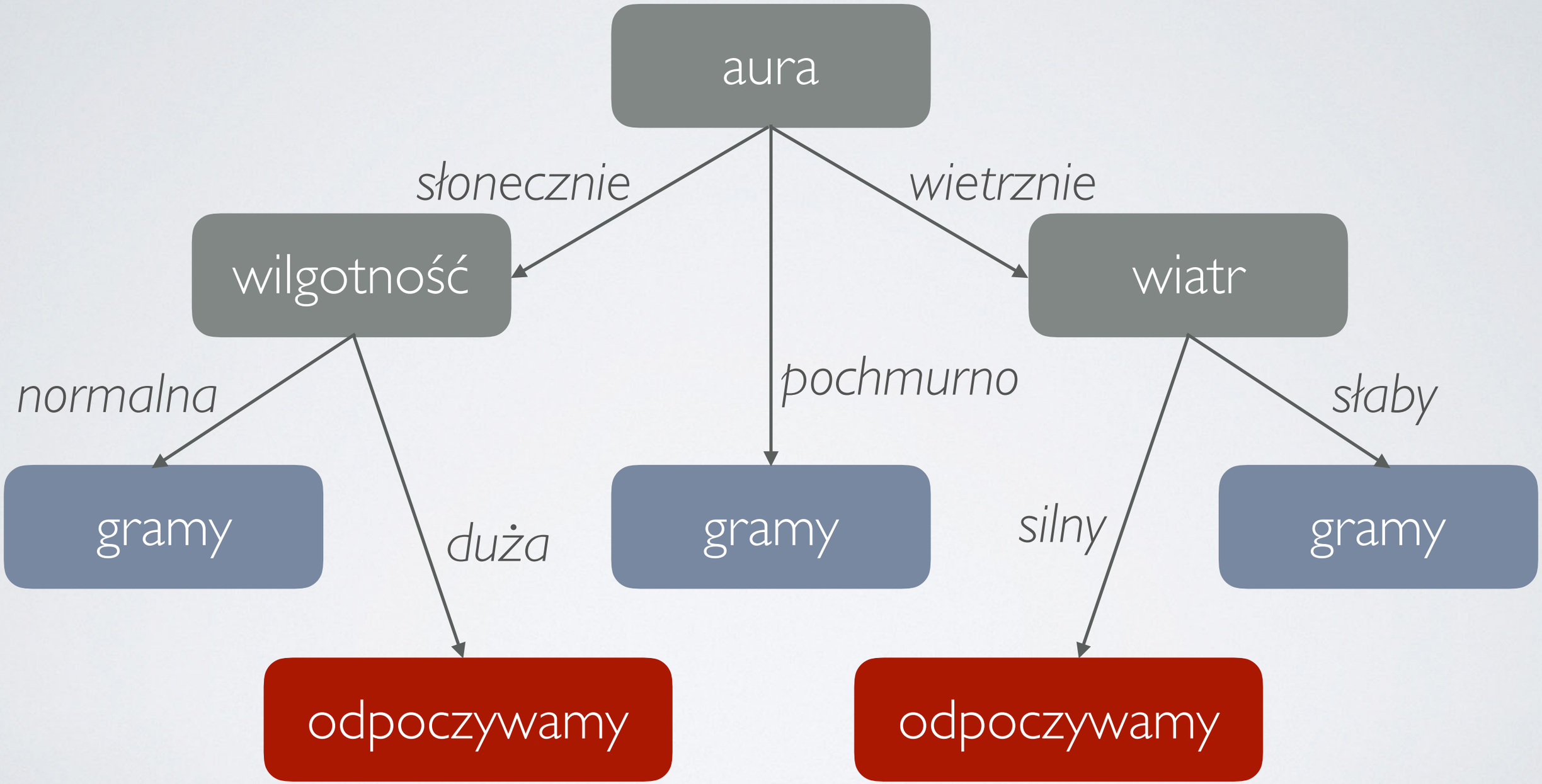
Arystoteles
384–322 p.n.e

Platon
427–347 p.n.e

TABELE DECYZYJNE

x	aura	wilgotność	wiatr	decyzja
1	słoneczna	duża	słaby	odpoczywamy
2	słoneczna	duża	silny	odpoczywamy
3	pochmurna	duża	słaby	gramy
4	deszczowa	duża	słaby	gramy
5	deszczowa	normalna	słaby	gramy
6	deszczowa	normalna	silny	odpoczywamy
7	pochmurna	normalna	silny	gramy
8	słoneczna	normalna	silny	gramy
9	słoneczna	normalna	słaby	gramy
10	pochmurna	duża	silny	gramy

DRZEWA DECYZYJNE

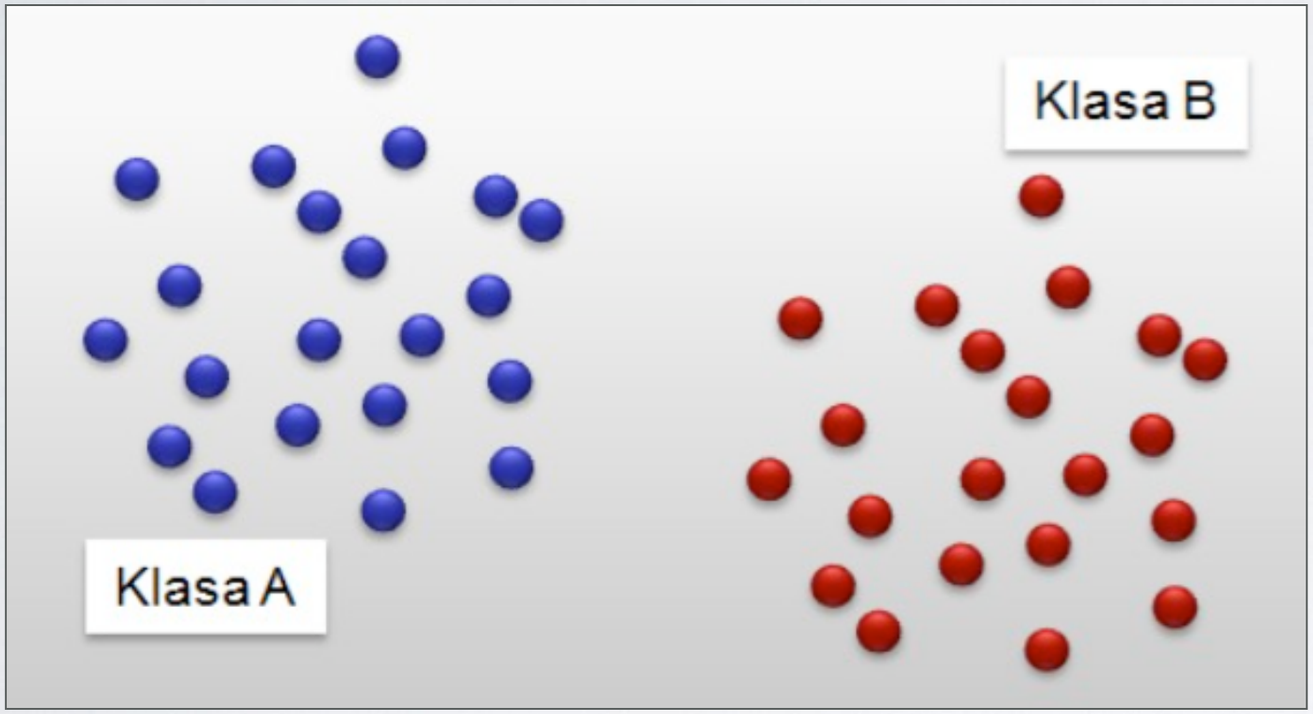


REGUŁY DECYZYJNE

- Reguły klasyfikacyjne
  - część warunkowa
  - część decyzyjna
  - predykcja klasy:
    - **JEŚLI** *aura=słonecznie* **I** *wilgotność=duża*
      TO *decyzja=odpoczywamy*
- Reguły asocjacyjne
  - predykcja dowolnego atrybutu:
    - **JEŚLI** *wiatr=słaby* **I** *decyzja=odpoczywamy*
      TO *aura=słonecznie* **I** *wilgotność=duża*

PAMIĘĆ PRZYKŁADÓW

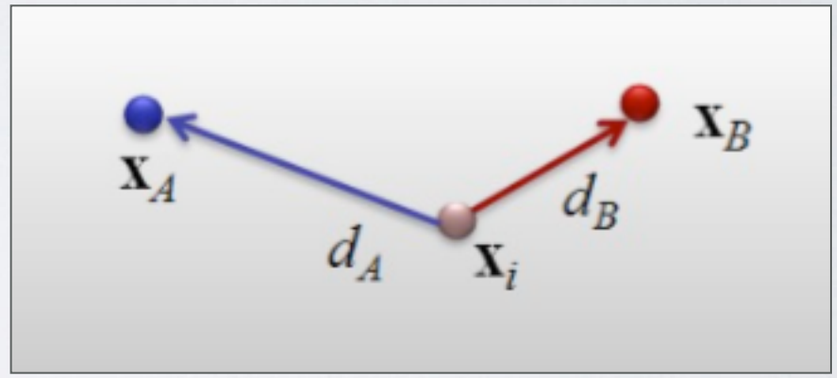
Zbiór treningowy



Metoda *k* najbliższych sąsiadów

Nowy wektor przypisywany jest do klasy, do której należy większość z jego *k* sąsiadów.

Nowy wektor
o nieznanej klasie

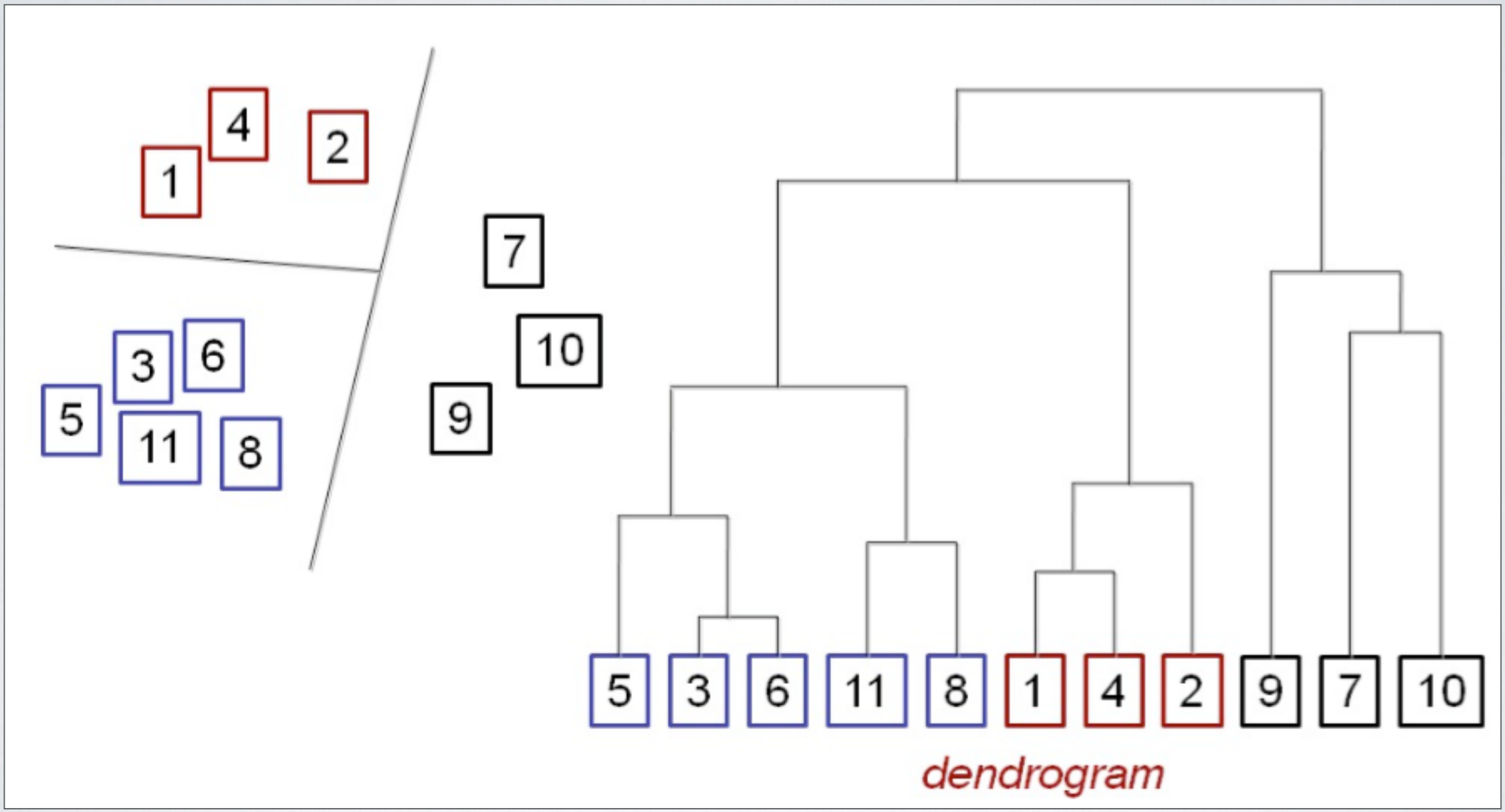


Podobieństwo wektorów
– miara Euklidesowa

$$d_A = \sqrt{(x_i^1 - x_A^1)^2 + (x_i^2 - x_A^2)^2}$$

$$d_B = \sqrt{(x_i^1 - x_B^1)^2 + (x_i^2 - x_B^2)^2}$$

KLASTERY



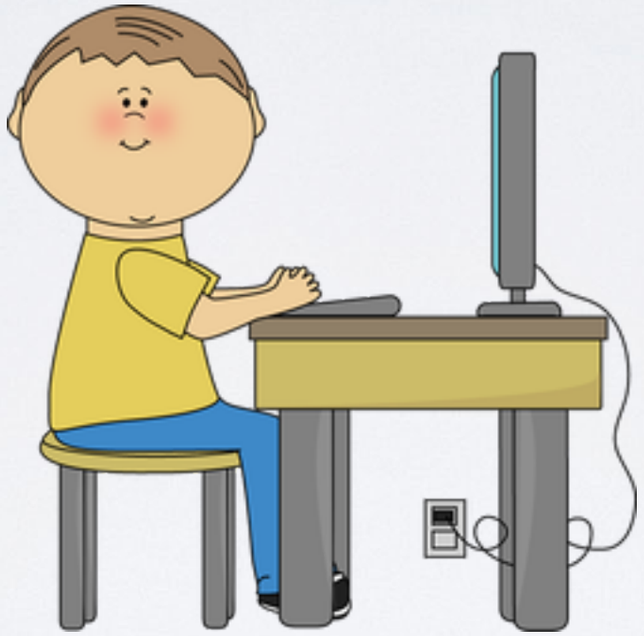
ETAPY EKSPLORACJI DANYCH

FORMY UCZENIA

Wykład

Nauczyciel przekazuje wiedzę uczniowi

Laboratorium

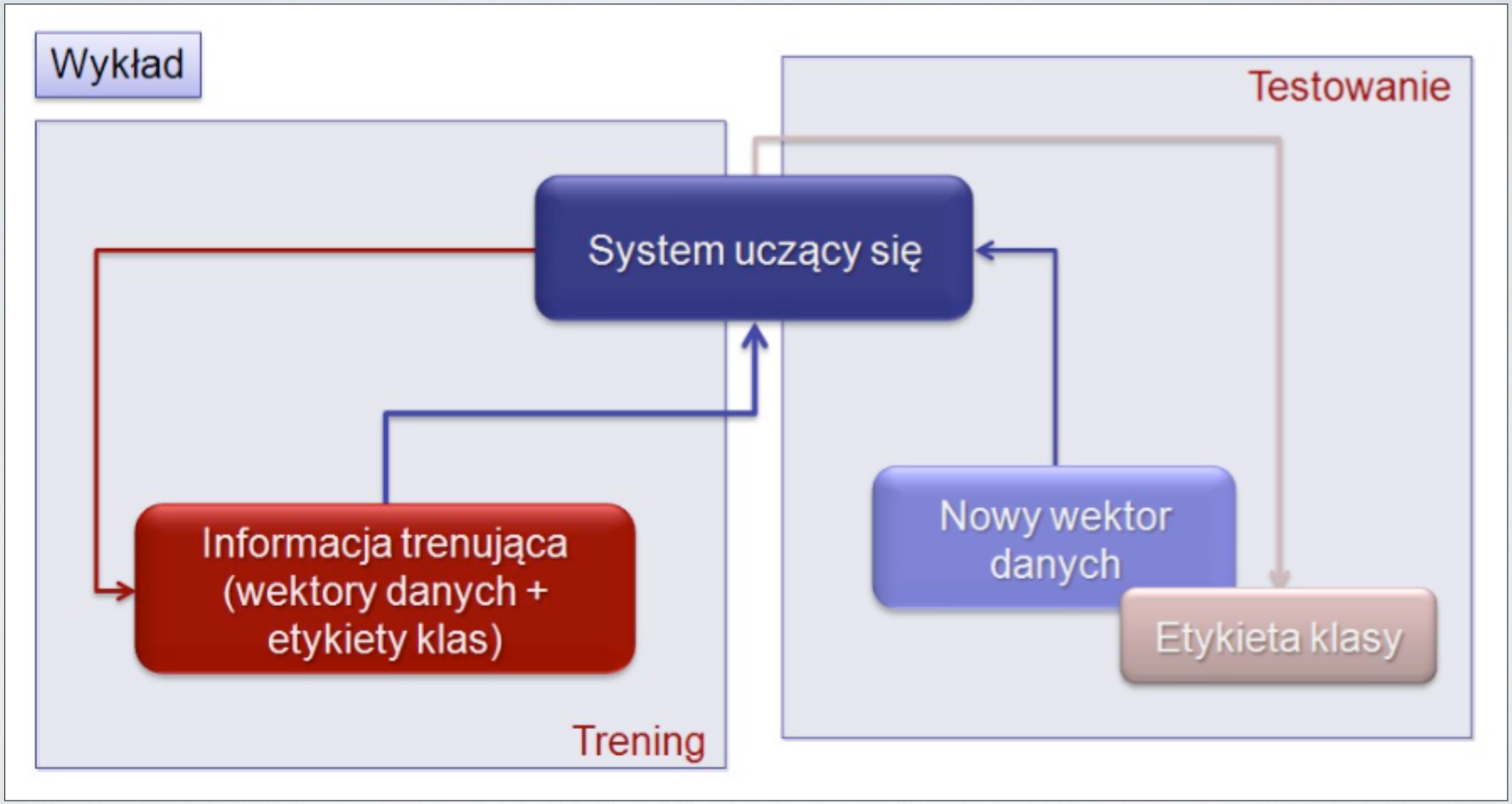


Uczeń wykonuje zadania i jest „prowadzony” przez nauczyciela

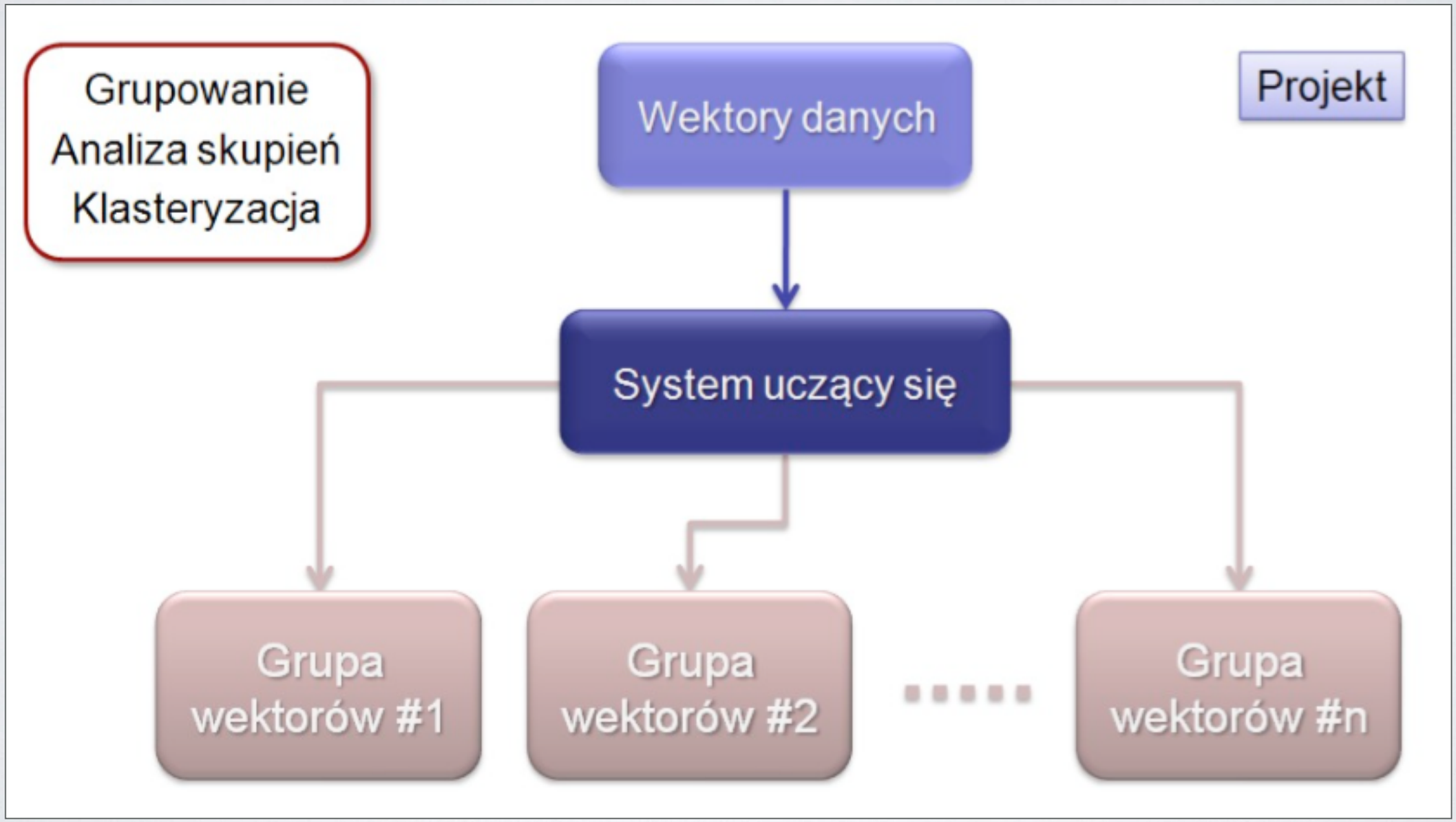
Projekt

Uczeń samodzielnie zdobywa wiedzę

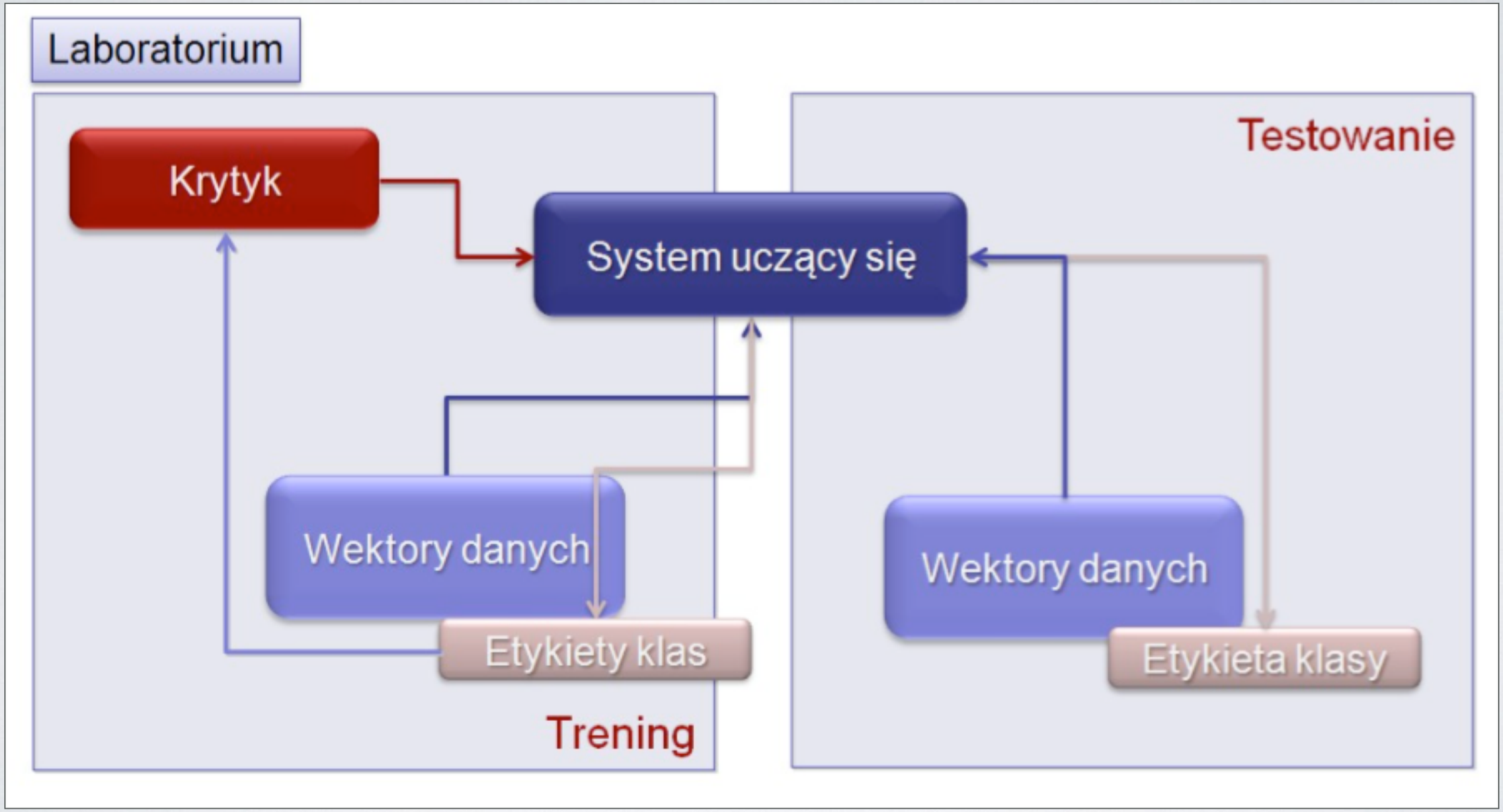
UCZENIE NADZOROWANE



UCZENIE NIENADZOROWANE



UCZENIE ZE WZMOCNIENIEM



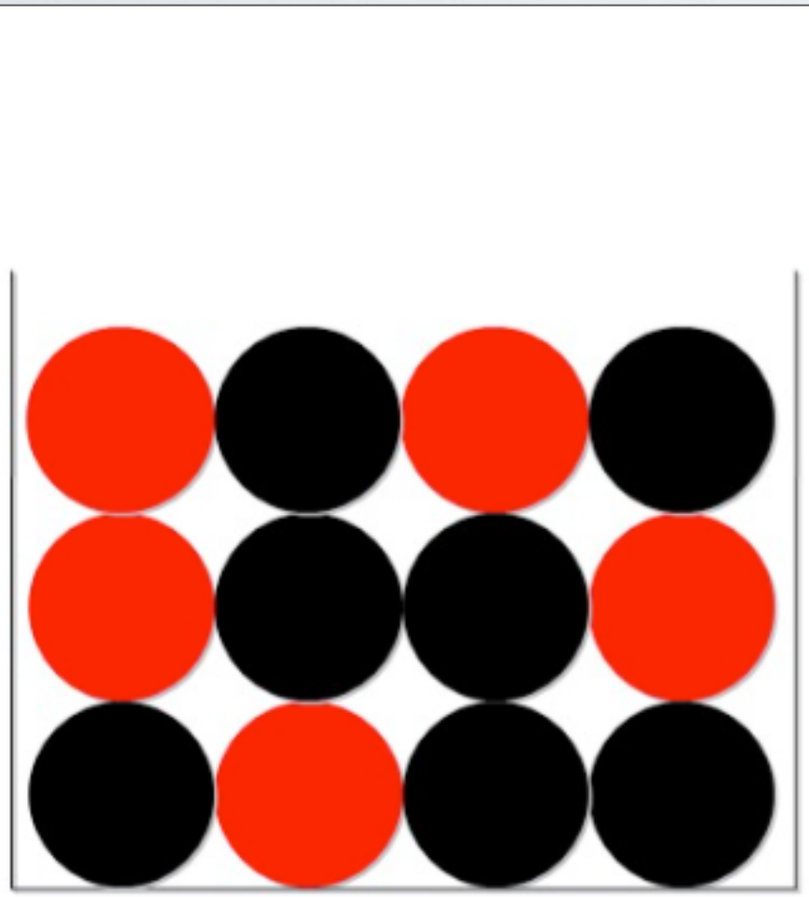
To było ogólne
wprowadzenie do tematu.

Należy zapamiętać:

- **dane to zbiór cech** jakichś obiektów,
- uczenie może być **nadzorowane** lub **nienadzorowane**,
- podstawowe zadania eksploracji danych to **klasyfikacja** i **regresja**,
- w przypadku uczenia nadzorowanego zbiór danych dzieli się na część **treningową** i **testową**,
- klasyfikatory, niestety, popełniają **błędy klasyfikacji**.

Pora na jakiś algorytm…

MINI-RESUMÉ Z PROBABILISTYKI



Zmienna losowa

- Kula, X

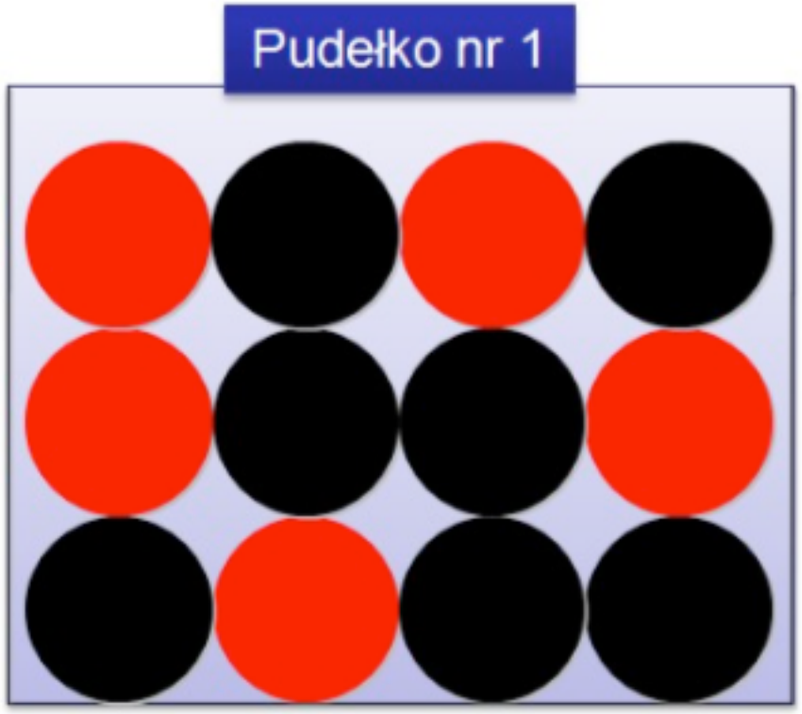
Realizacja zmiennej losowej

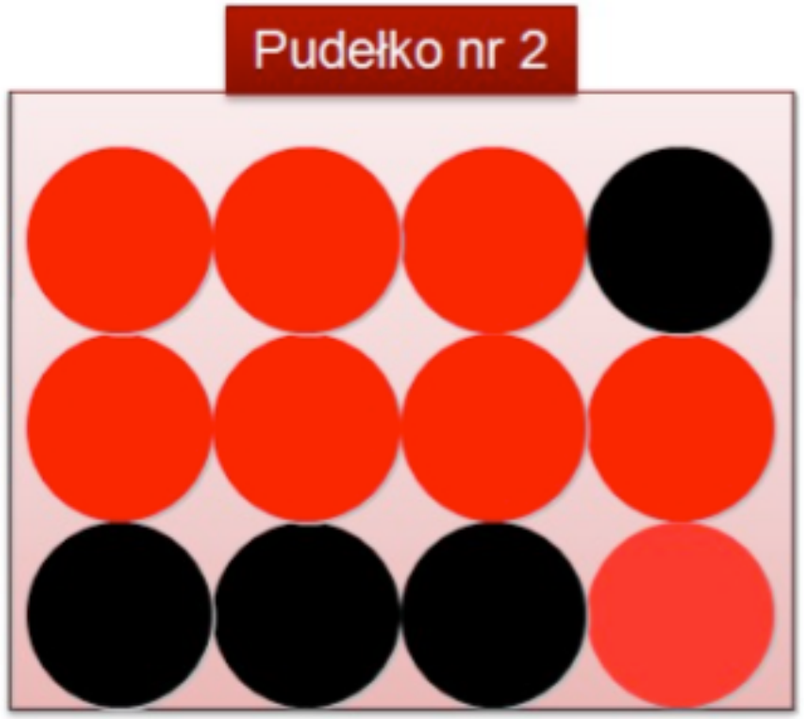
- Czarna kula, $X = b$
- Czerwona kula, $X = r$

Prawdopodobieństwo

- $\Pr(X=b) = 7 / 12$
- $\Pr(X=r) = 5 / 12$

PRAWDOPODOBIEŃSTWO WARUNKOWE





Zmienna losowa: pudełko, Y

Realizacja zmiennej losowej: $Y=1$ lub $Y=2$

Prawdopodobieństwo wylosowania kuli czerwonej pod warunkiem wylosowania pudełka nr 2:

$\Pr(X=r \mid Y=2) = 8/12$

A PRIORI / A POSTERIORI

Przypuśćmy, że…

1. Eksperyment polegający na wielokrotnym losowaniu pudełka wykazał, że pudełko nr 1 wybrano w 40% przypadków.
2. W jednym z kolejnych losowań wybrano kulę czarną.

Prawdopodobieństwo a priori — przed wykonaniem kroku 2.

$\Pr(Y=1)=0,4$ $\Pr(Y=2)=0,6$

Całkowite prawdopodobieństwo wylosowania kuli czarnej:

$$\Pr(X=b) = \Pr(X=b \mid Y=1)\times\Pr(Y=1) + \Pr(X=b \mid Y=2)\times\Pr(Y=2)$$
$$\Pr(X=b) = 7/12\times0,4 + 4/12\times0,6 = 0,43$$

Jakie jest prawdopodobieństwo a posteriori tego, że wylosowano pudełko nr 1po wykonaniu kroku 2.

$\Pr(Y=1 \mid X=b) = ?$

TWIERDZENIE BAYESA

Thomas Bayes, 1702–1761

Źródło: en.wikipedia.org.
Grafika dostępna na zasadach GFDL

$$\Pr(Y \mid X) = \frac{\Pr(X \mid Y)\Pr(Y)}{\Pr(X)}$$

Prawdopodobieństwa a posteriori

Prawdopodobieństwo a priori

Prawdopodobieństwo całkowite

$$\Pr(Y = 1 \mid X = b) = \frac{\Pr(X = b \mid Y = 1)\Pr(Y = 1)}{\Pr(X = b)}$$

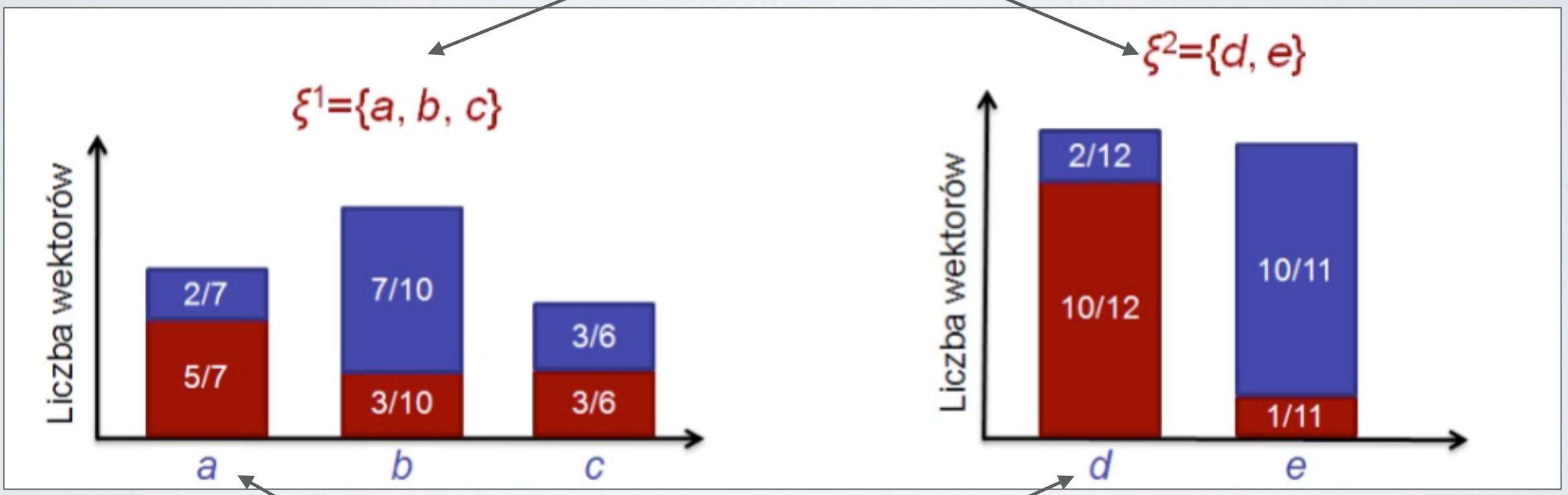
$$= \frac{7/12 \cdot 0{,}4}{0{,}43} \approx 0{,}54$$

PRZYKŁAD OBLICZENIOWY

Klasa A – 11 wektorów

Klasa B – 12 wektorów

Cechy



Wartości cech

DECYZJE NA PODSTAWIE PRAWDOPODOBIEŃSTWA

ξ^1			ξ^2			Klasa	
	A	B		A	B	A	B
a	5/11	2/12	*d*	10/11	2/12	11/23	12/23
b	3/11	7/12	*e*	1/11	10/12		
c	3/11	3/12					

Prawdopodobieństwo tego, że dla wektora danych należącego do klasy „A" atrybut ξ^1=a.

Prawdopodobieństwo tego, że wektor danych należy do klasy „B"

Jaka będzie klasa nowego wektora $x_{test} = \{b, d\}$?

NAIWNY KLASYFIKATOR BAYESA

Wektor testowy: $x_{test}=\{b, d\}$

Prawdopodobieństwo *a priori* klasy „A”
Pr(A) = 11/23

Prawdopodobieństwo *a priori* klasy „B”
Pr(B) = 12/23

Prawdopodobieństwo *a posteriori* klasy „A”
Pr(Klasa „A” | $\mathbf{x}_{test}$)

Prawdopodobieństwo *a posteriori* klasy „B”
Pr(Klasa „B” | $\mathbf{x}_{test}$)

Iloczyny prawdopodobieństw atrybutów

3/11 × 10/11 × 11/23 ≈ 0,12

7/12 × 2/12 × 12/23 ≈ 0,05

Obliczenia są poprawne przy (naiwnym) założeniu niezależności atrybutów.

CDN.