

dr inż. Artur Klepaczko

Eksploracja danych Wprowadzenie

Zadanie nr 13 – Studia podyplomowe „Przetwarzanie i analiza obrazów biomedycznych”



KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY



Prezentacja multimedialna
współfinansowana przez Unię Europejską
w ramach Europejskiego Funduszu Społecznego
w projekcie

*„Innowacyjna dydaktyka bez ograniczeń
– zintegrowany rozwój Politechniki Łódzkiej –
zarządzanie Uczelnią,
nowoczesna oferta edukacyjna
i wzmacniania zdolności do zatrudniania
osób niepełnosprawnych”*



Politechnika Łódzka
Instytut Elektroniki

90-924 Łódź, ul. Żeromskiego 116,
tel. 042 631 28 83
www.kapitalludzki.p.lodz.pl

Eksploracja danych i terminy pokrewne



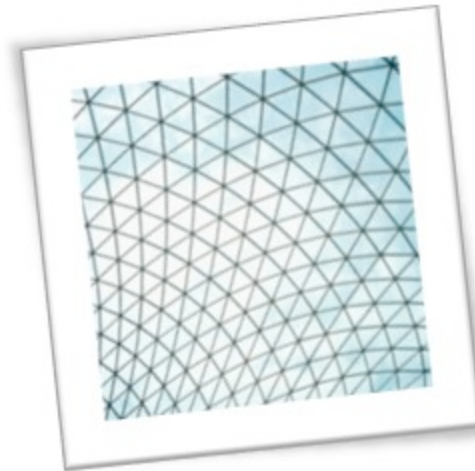
Data mining



Knowledge discovery



Machine learning



Pattern recognition

Cel nauki – zdobywanie wiedzy



„Wiedza to prawdziwe, uzasadnione przekonanie.”

Platon, *Teajtet*

„Ludzie rodzą się z niezapisanym umysłem, który zapelnia się myślami na skutek codziennych doświadczeń życiowych.”

Arystoteles

Źródło: pl.wikipedia.org.

Grafika dostępna na zasadach GFDL

Eksploracja danych – przykład 1.



Czym różnią się te obrazy?
Jaka regularność występuje w powyższym zestawie?

Klasyfikacja obrazów

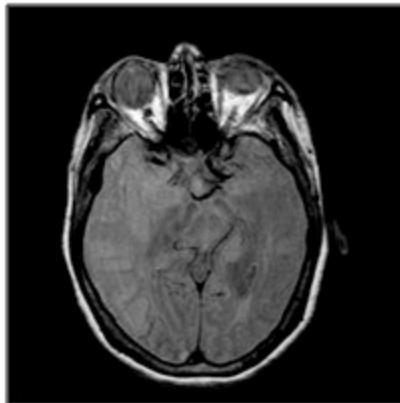
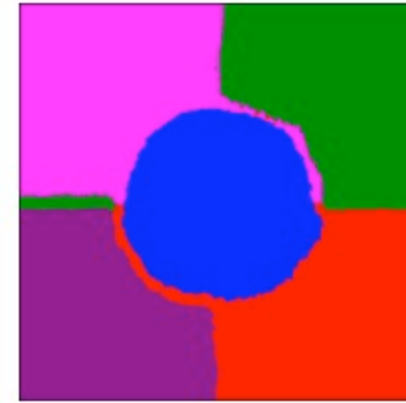
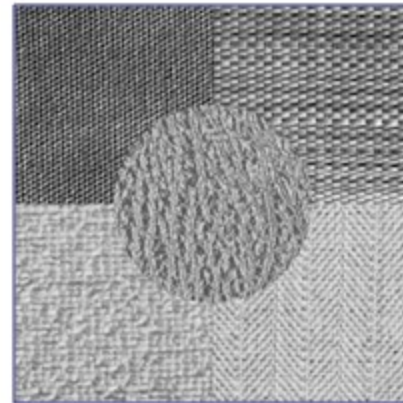
Ze względu na kształt

Ze względu na kolor

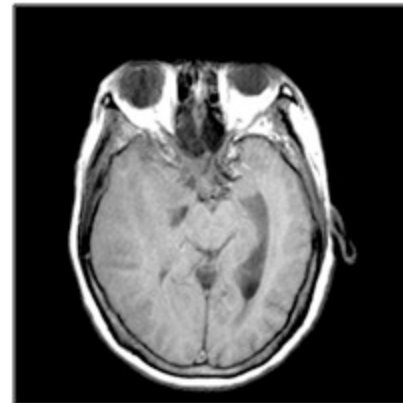


Eksploracja danych – przykład 2.

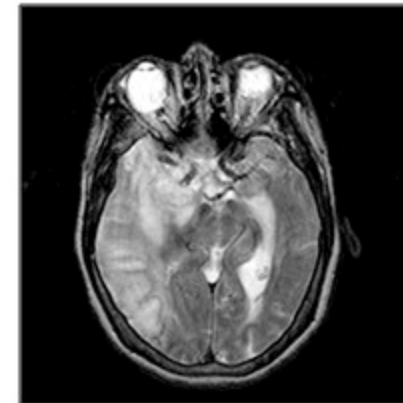
Segmentacja
obrazów



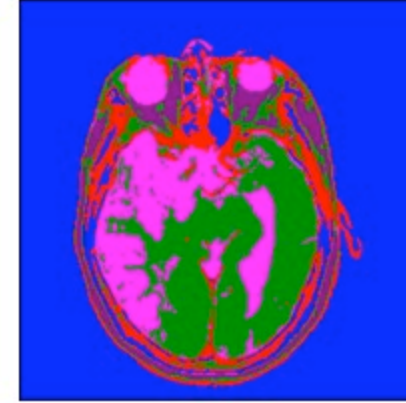
Proton Density



T1-zależna



T2-zależna





Co to są dane?

Dane tworzą opis obiektów rzeczywistości wyrażony za pomocą ich cech (atrybutów, parametrów).

Obiektami mogą być ludzie, wszelkie przedmioty lub zjawiska fizyczne oraz ich reprezentacje, np. obrazy.

Cechy jakościowe to np. kolor, smak, stan pogody; pochodzą najczęściej z obserwacji.

Cechy ilościowe to np. waga, rozmiar, wartość funkcji; pochodzą najczęściej z pomiaru oraz obliczeń.

Zbiór cech danego obiektu stanowi odpowiadający mu **wektor danych**.

Kolekcja wektorów danych wyznaczonych dla pewnej serii obiektów stanowi **zbiór danych**.



Notacja matematyczna

Wektor danych

$$\mathbf{x} = (\xi^1, \xi^2, \dots, \xi^N)$$

Zbiór danych

$$\mathbf{X} = \{\mathbf{x}_i, i = 1..Q\}$$

Cecha j obiektu i

$$\xi_i^j$$

Egzemplarz a z klasy obiektów A

$$a \in A$$

Cechy ilościowe (ciągłe)

$$\xi_{num} : (a \in A) \mapsto \mathbb{R}$$

Cechy jakościowe (nominalne)

$$\xi_{nom} : (a \in A) \mapsto \{v_1, \dots, v_r\}$$



Czym jest eksploracja danych?

Eksploracja danych obejmuje zespół metod pozwalających na wydobywanie informacji użytecznej nt. ludzi, przedmiotów bądź zjawisk lub ich reprezentacji (np. obrazów) na podstawie zgromadzonych cech tych obiektów i z zastosowaniem metod uczenia maszynowego. Uzyskana informacja ma umożliwiać m.in. klasyfikację danych, aproksymację funkcji, predykcję zachowań. Wynik tych działań ma **wzbogacać** wiedzę o przedmiocie zainteresowania.





Interpretacja klasycznej definicji wiedzy wg Gettier'a

Podmiot S wie, że P wtedy i tylko wtedy, gdy:

- P jest zdaniem prawdziwym
- S jest przekonany, że P
- S posiada uzasadnienie swojego przekonania, że P

Edmund Gettier, „Is Justified True Belief Knowledge”, *Analysis*, 23, 121-123, 1963





Pojęcie wiedzy

Encyklopedia Powszechna PWN

- ogół wiarygodnych informacji o rzeczywistości wraz z umiejętnością ich wykorzystywania.

Wiedza a priori

- jest niezależna od zmysłów i dotyczy prawd "absolutnych" lub uniwersalnych jakimi są prawa logiki, prawa matematyki.

Wiedza a posteriori

- jest wiedzą nabytą poprzez zmysły i jej prawdziwość może być obalona poprzez następne obserwacje.

W psychologii w szerokim rozumieniu wiedza to np.

- ogół treści utrwalonych w umyśle ludzkim w wyniku kumulowania doświadczenia oraz uczenia się. (...) W węższym znaczeniu wiedza stanowi osobisty stan poznania człowieka w wyniku oddziaływania na niego obiektywnej rzeczywistości. Często zamiast pojęcia „wiedza” używa się terminu „pamięć”.





Pojęcie wiedzy w uczeniu maszynowym

W badaniach nad sztuczną inteligencją wiedzę traktuje się jako materiał wejściowy albo efekt działania algorytmów sztucznej inteligencji.

Część metod sztucznej inteligencji zajmuje się próbami formalizacji ludzkiej wiedzy celem automatycznego wnioskowania (np. systemy ekspertowe).

Znane jest też odwrotne podejście. Dzięki zastosowaniu tych metod w dużych zbiorach danych można odnaleźć wiedzę o ich naturze (np. sieci neuronowe, algorytmy genetyczne, ekstrakcja cech).



Czym jest uczenie się maszyn?

Uczenie się systemu

- Każda **autonomiczna zmiana** w systemie zachodząca na podstawie **doświadczeń**, która prowadzi do **poprawy** jakości jego działania.

zmiana — każdy uczący się system zmienia się w czasie;

poprawa – zwiększenie skuteczności, sprawności systemu biorąc pod uwagę preferencje konstruktora lub użytkownika lub obu na raz.

autonomiczność — system, który się uczy, *sam* zmienia się na lepsze (a nie jest zmieniany przez kogoś).

doświadczenia —korzystna zmiana następuje pod wpływem czynników zewnętrznych, które można traktować jako doświadczenia zdobywane przez system.

Motywacja do uczenia się maszyn

Złożoność środowiska

- Złożone zadania nie posiadają prostych rozwiązań
- Praca w warunkach niedookreślonych
- Trudność w sformułowaniu modelu rzeczywistości

Potrzeba autonomii

- Programy uczące się posiadają zdolność adaptacji do zmiennych warunków środowiska pracy.

Potrzeba automatyzacji

- Bazy danych zawierają ogromne zasoby informacji.
- Wydobycie z nich informacji użytecznej poprzez zwykłe przeszukiwanie stanowi bardzo czasochłonne zadanie.

Na podstawie: Paweł Cichosz, *Systemy uczące się*. WNT, Warszawa 2000.



Zadania eksploracji danych

Klasyfikacja

Przypisywanie wektorów danych do kategorii

Regresja

Aproksymacja funkcji rzeczywistej

Predykcja

Przewidywanie klasy, zmiany stanu systemu lub trendu

Modelowanie

Matematyczny model zjawisk i obiektów

Wizualizacja

Prezentacja wektorów danych wielowymiarowych

Optymalizacja

Inteligentne rozwiązywanie skomplikowanych zagadnień matematycznych (polioptymalizacyjnych, zwłaszcza w przestrzeni wielowymiarowej)

Przeszukiwanie

Wydobywanie informacji

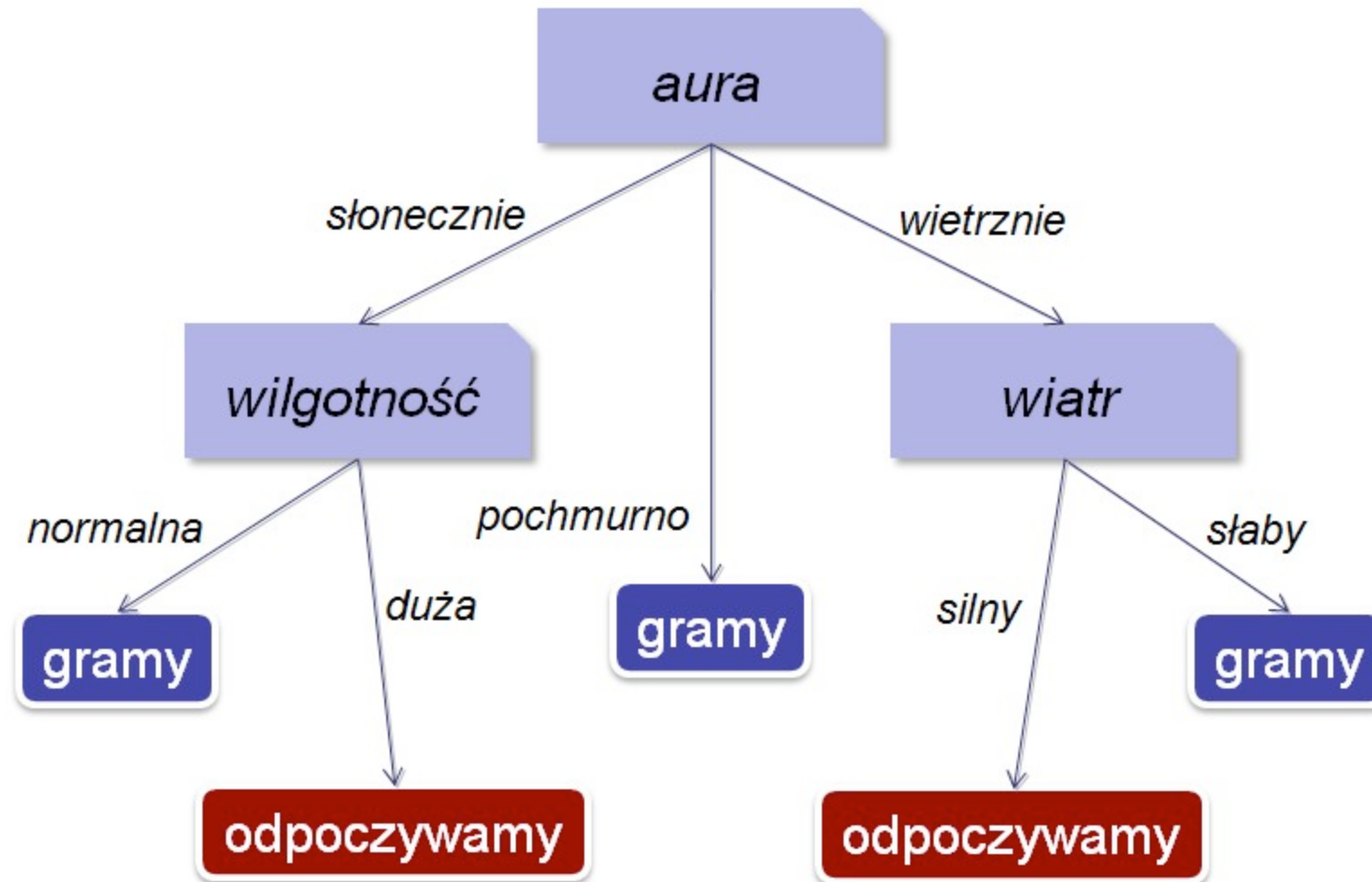
Odnajdywanie właściwych dokumentów lub obrazów



Metody reprezentacji wiedzy — tabele decyzyjne

<i>x</i>	<i>aura</i>	<i>wilgotność</i>	<i>wiatr</i>	<i>decyzja</i>
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

Metody reprezentacji wiedzy — drzewa decyzyjne



Metody reprezentacji wiedzy — reguły

- Reguły klasyfikacyjne

- część warunkowa
- część decyzyjna
- predykcja klasy:

JEŚLI *aura=słonecznie* **I** *wilgotność=duża*

TO *decyzja=odpoczywamy*

- Reguły asocjacyjne

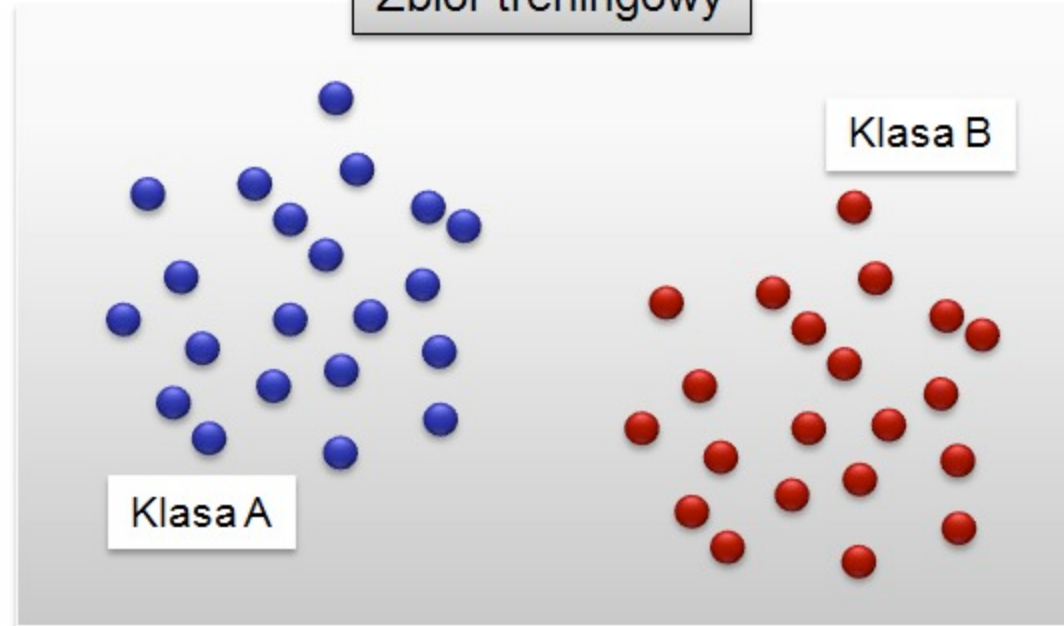
- predykcja dowolnego atrybutu:

JEŚLI *wiatr=słaby* **I** *decyzja=odpoczywamy*

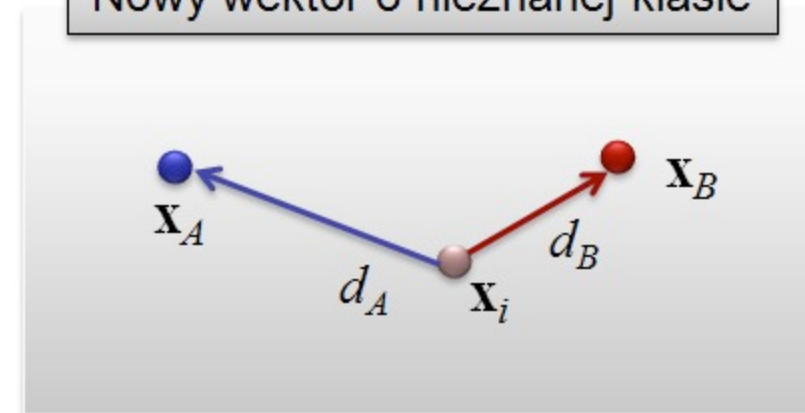
TO *aura=słonecznie* **I** *wilgotność=duża*

Metody reprezentacji wiedzy — pamięć przykładów

Zbiór treningowy



Nowy wektor o nieznanej klasie



Podobieństwo wektorów —
odległość Euklidesowa

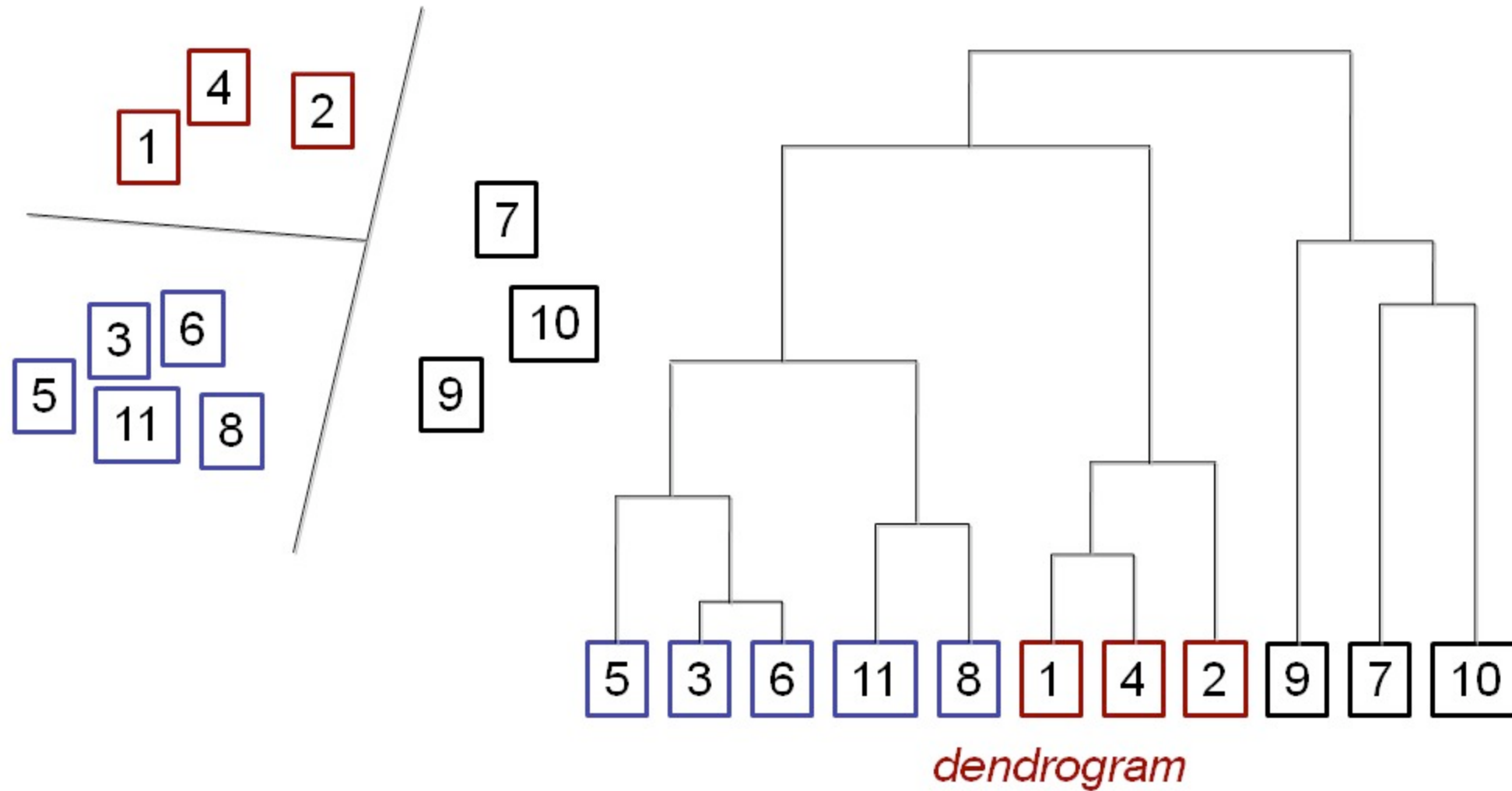
Metoda k najbliższych sąsiadów (k -NN)

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

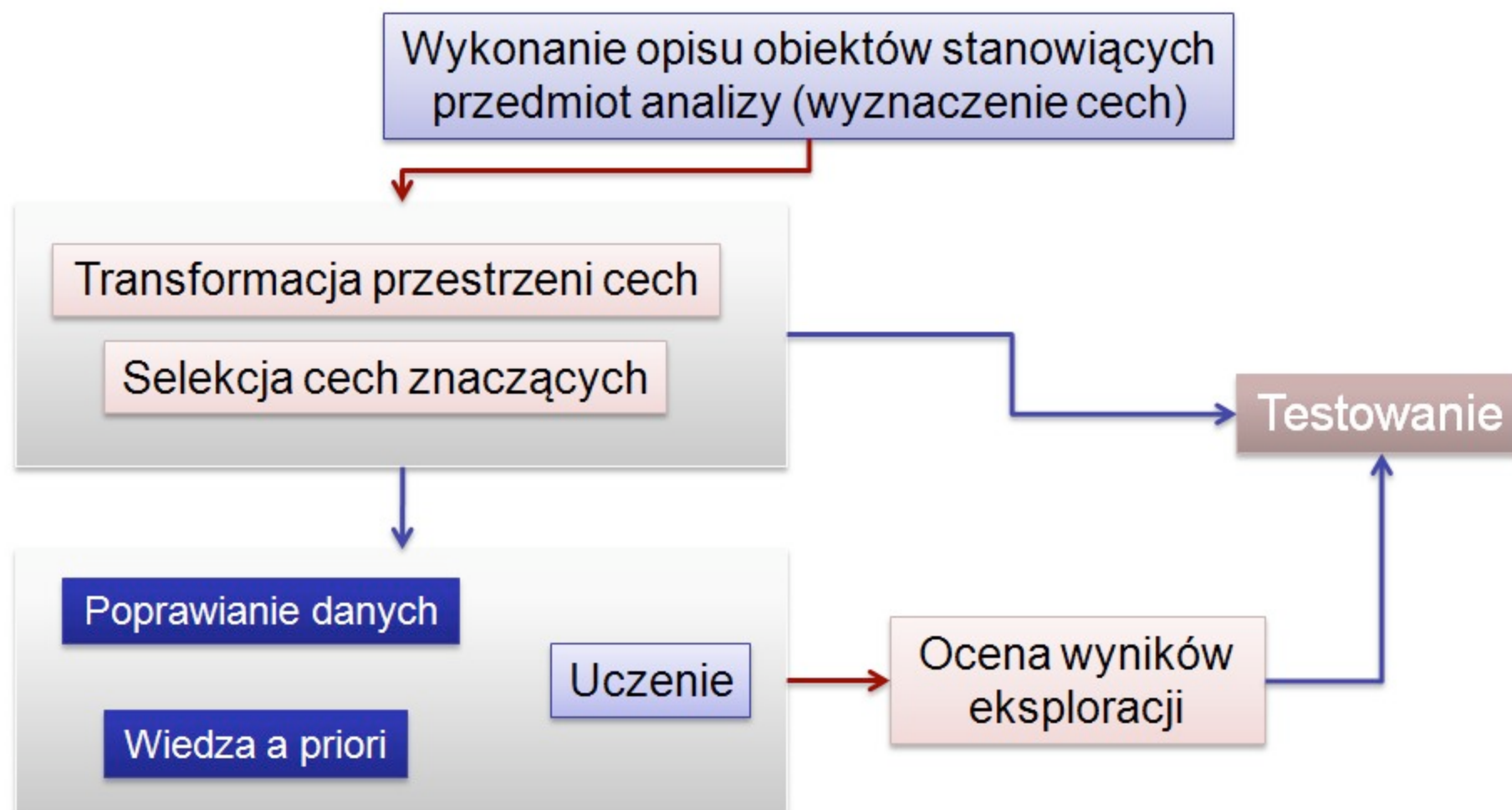
$$d_A = \sqrt{(x_i^1 - x_A^1)^2 + (x_i^2 - x_A^2)^2}$$

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

Metody reprezentacji wiedzy — klastery



Etapy eksploracji danych



Formy uczenia się

Wykład



Nauczyciel przekazuje
wiedzę uczniowi

Laboratorium



Uczeń wykorzystuje wiedzę
z wykładu do rozwiązania zadań
i sam uczy się nowych zjawisk

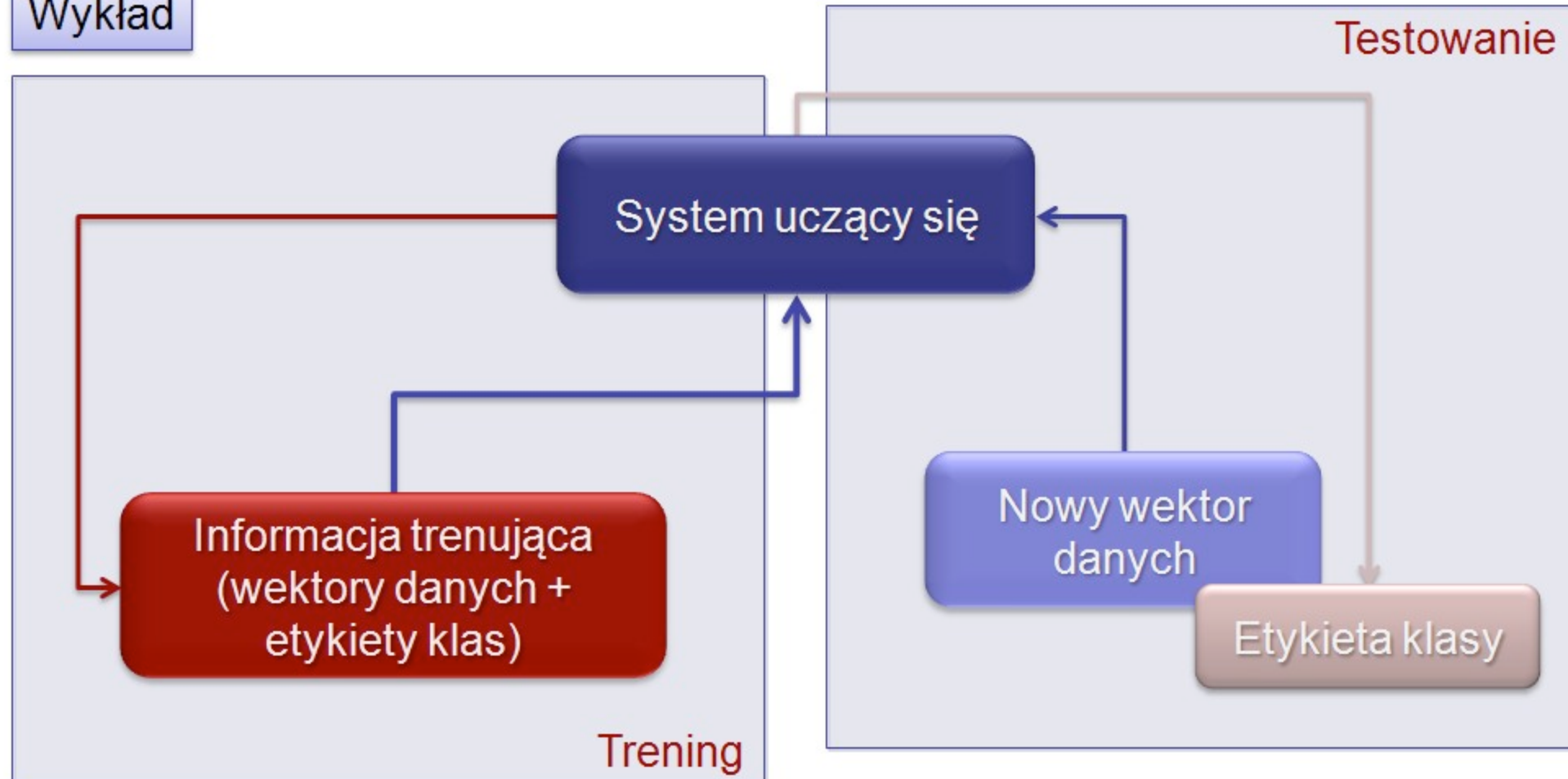
Projekt



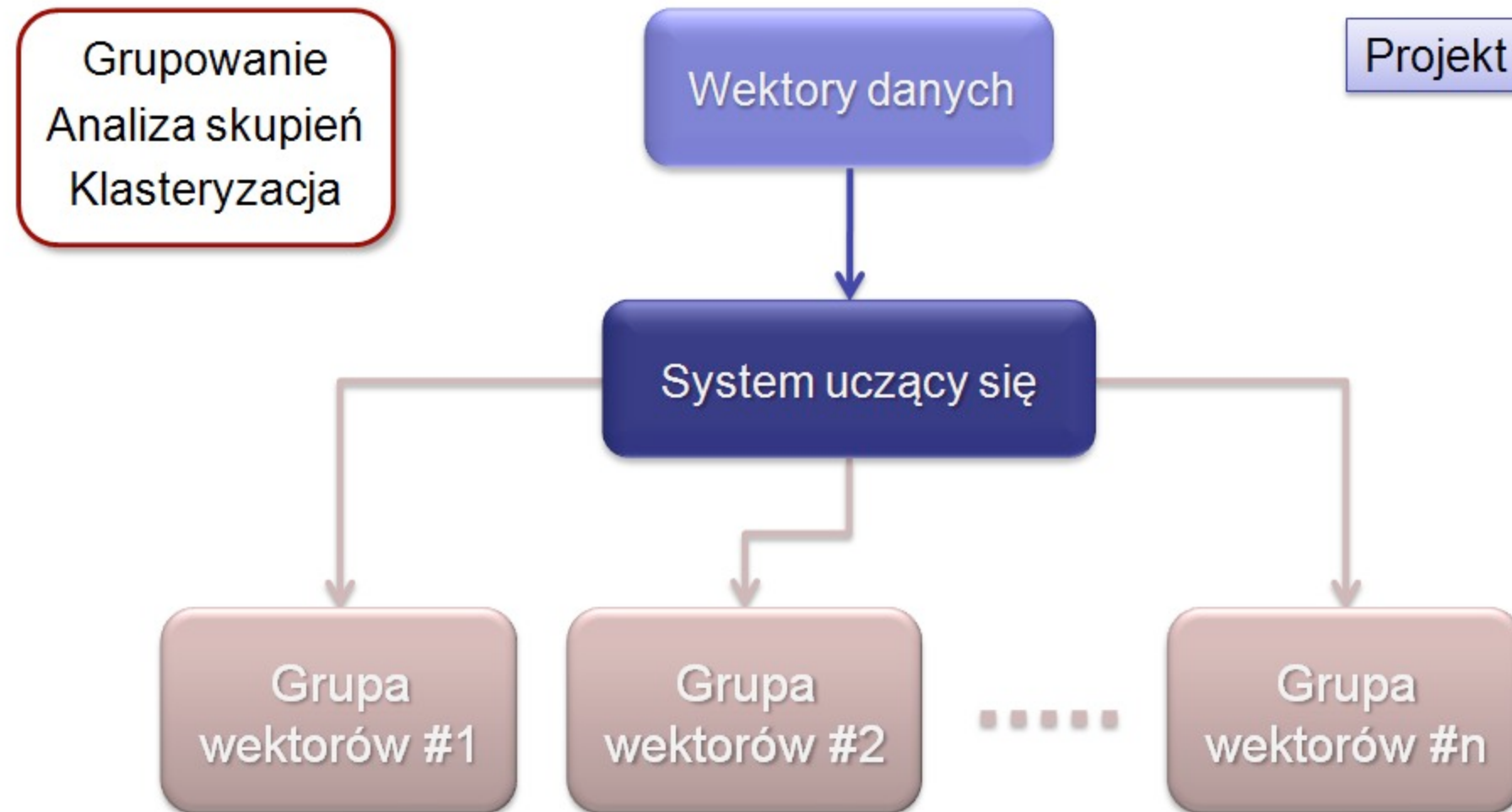
Uczeń samodzielnie
zdobywa wiedzę

Uczenie nadzorowane

Wykład

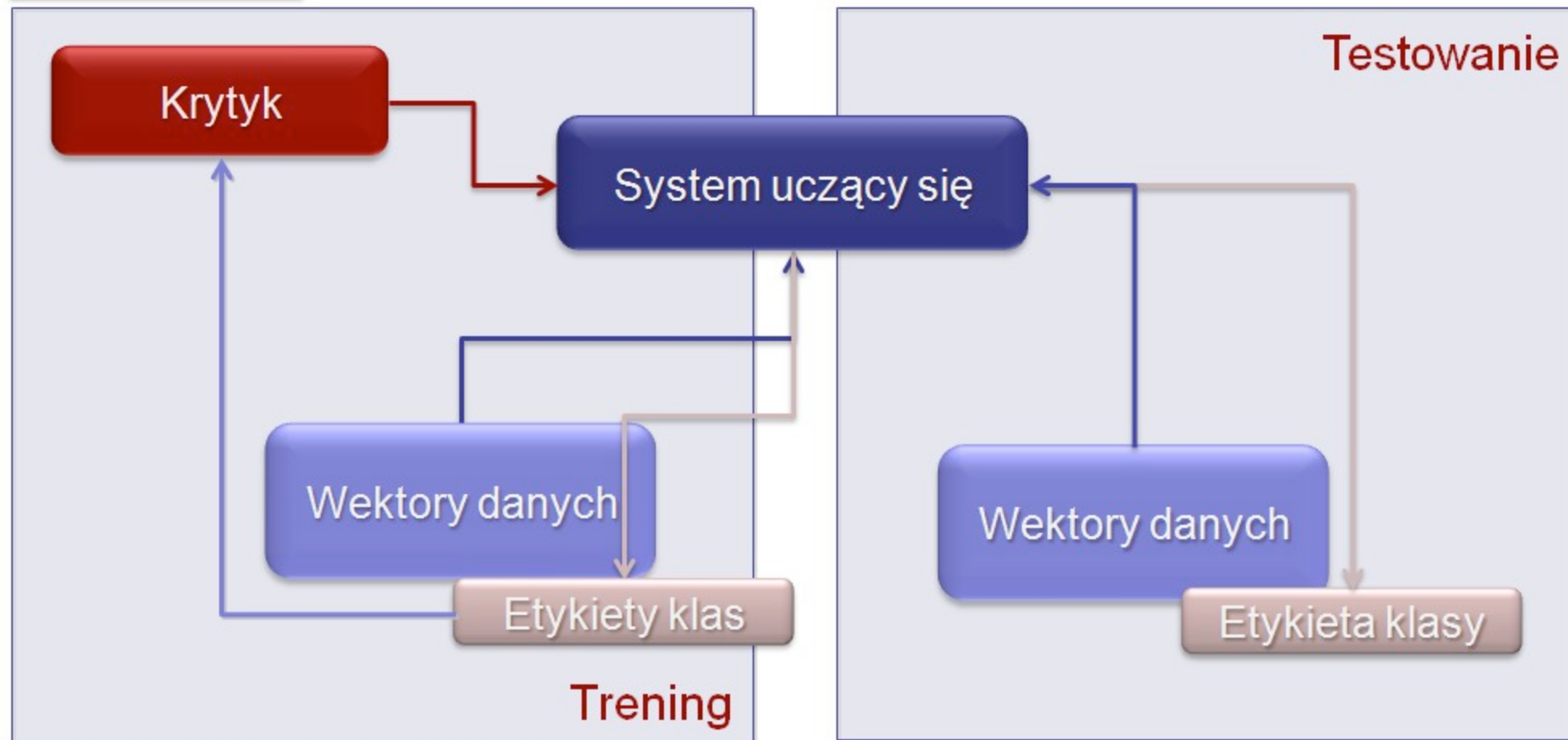


Uczenie nienadzorowane

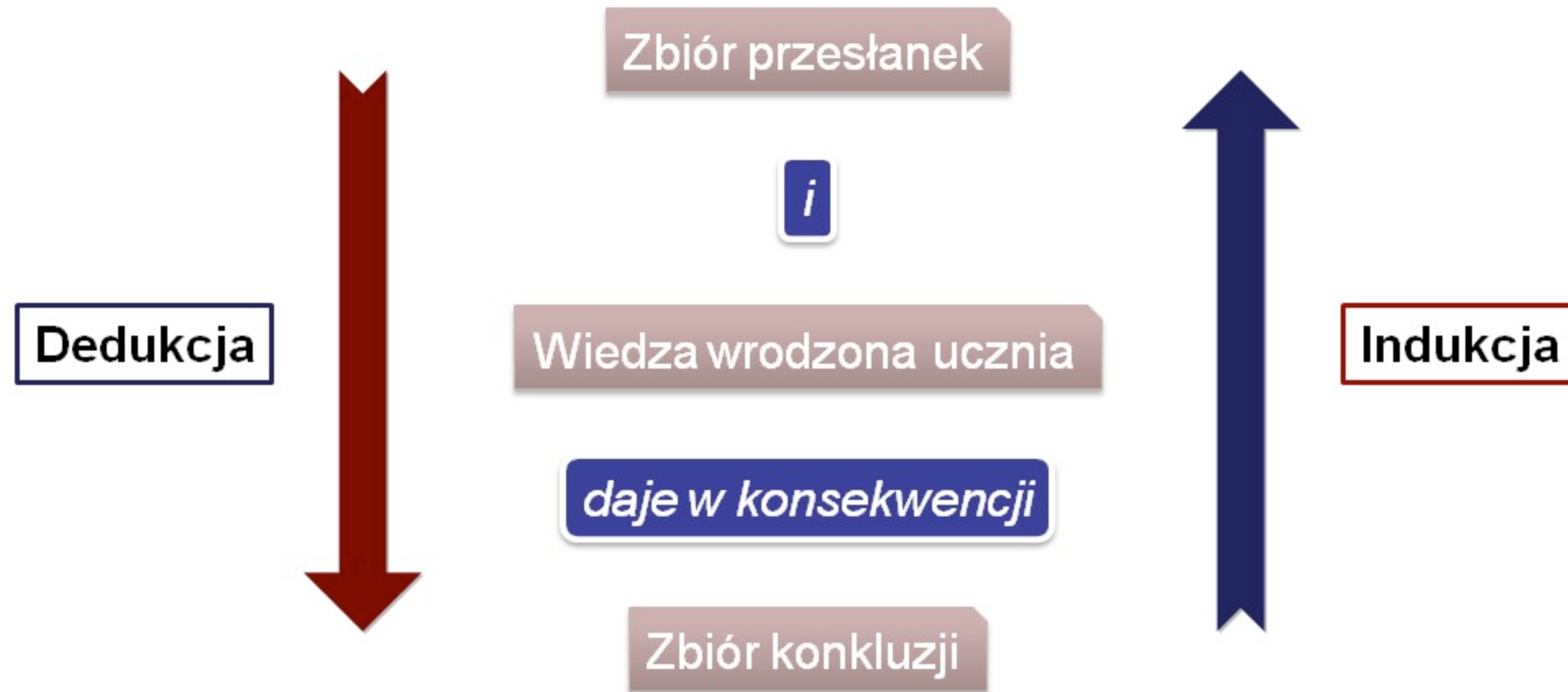


Uczenie ze wzmocnieniem

Laboratorium



Sposób nabywania wiedzy



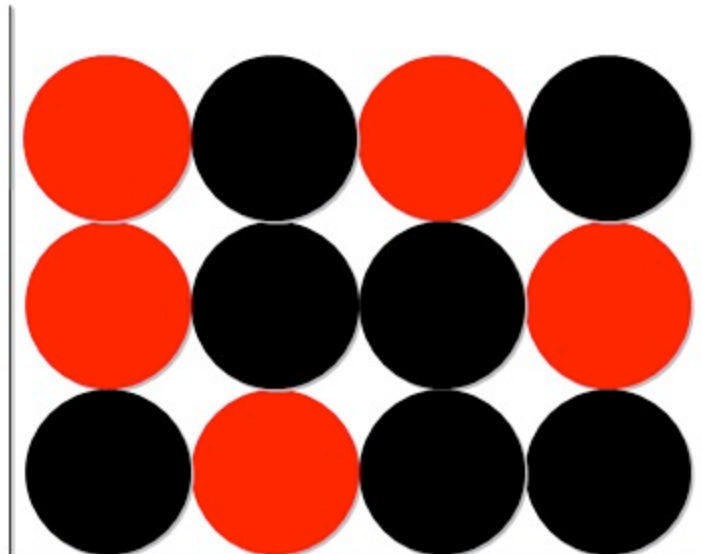


Teoria uczenia się maszyn

- **Teoria prawdopodobieństwa**
 - Dostarcza podstawowego aparatu matematycznego dla opisu algorytmów eksploracji danych.
 - Probabilistyczne metody klasyfikacji i grupowania.
- **Teoria informacji**
 - Definiuje miary jakościowe, wykorzystywane zarówno przy ocenie decyzji, jak i danych (pod względem użyteczności informacji zawartej w cechach).
- **Logika formalna**
 - Stanowi podstawę formułowania reguł decyzyjnych.
 - Metody indukcyjnego programowania logicznego.
- **Statystyka**
 - Narzędzia do statystycznego opisu danych (średnia, korelacja).
 - Transformacje danych.
- **Teoria decyzji**
 - Zasady podejmowania optymalnych decyzji w przypadku braku pewności.



Teoria prawdopodobieństwa – pojęcia podstawowe



Zmienna losowa

- Kula, X

Realizacja zmiennej losowej

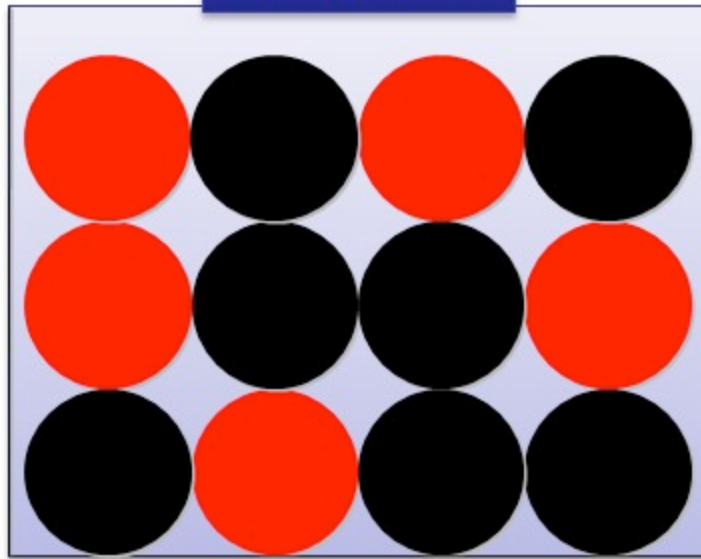
- Czarna kula, $X = b$
- Czerwona kula, $X = r$

Prawdopodobieństwo

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

Prawdopodobieństwo warunkowe

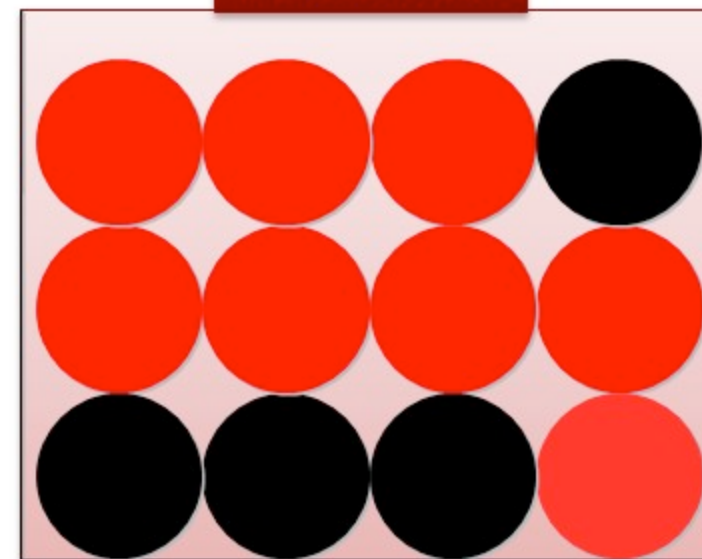
Pudełko nr 1



Zmienna losowa: pudełko, Y

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

Pudełko nr 2



Prawdopodobieństwo wylosowania kuli czerwonej pod warunkiem wylosowania pudełka nr 2:

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

Prawdopodobieństwo *a priori* oraz *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:

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

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

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

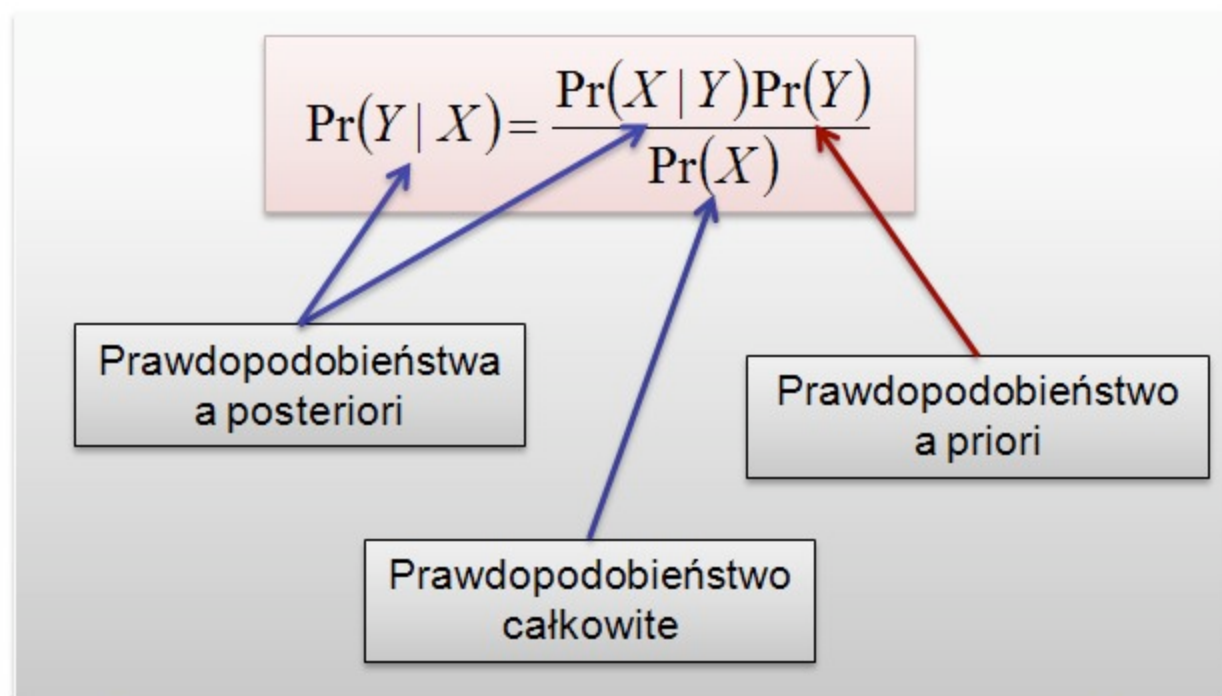
Twierdzenie Bayesa



Thomas Bayes, 1702–1761

Źródło: en.wikipedia.org.

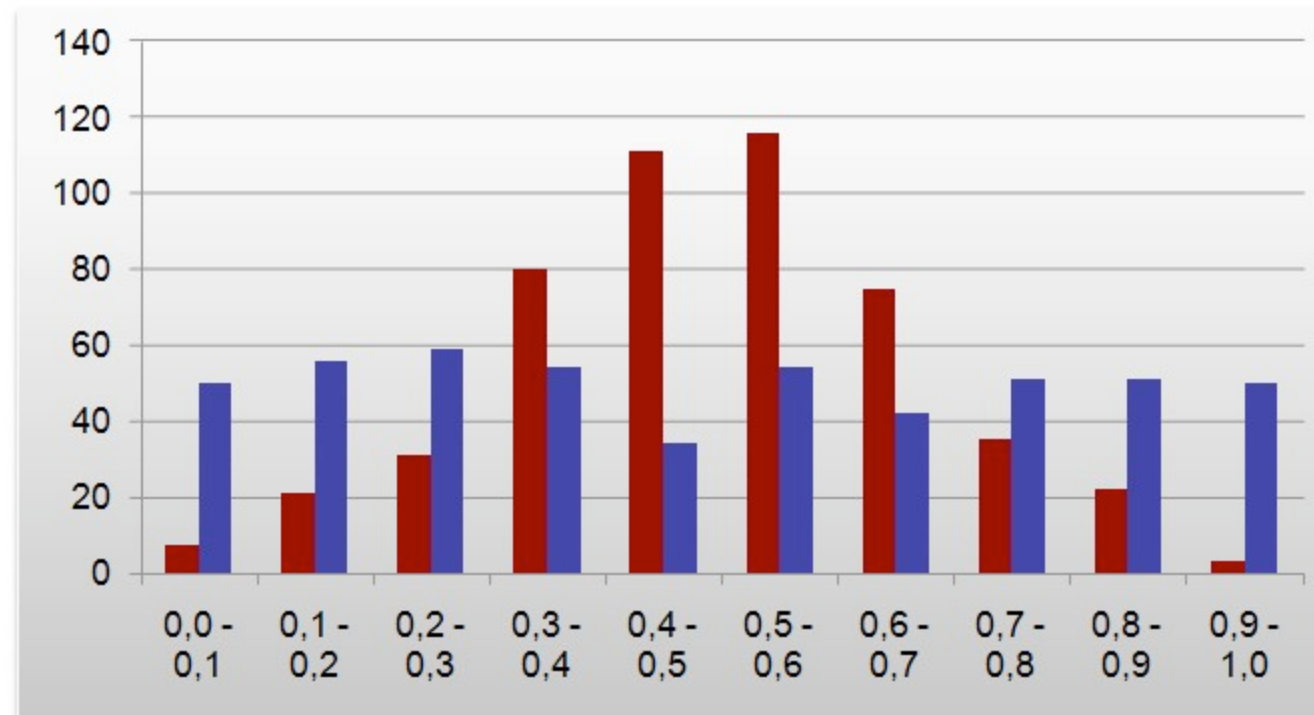
Grafika dostępna na zasadach GFDL



$$\begin{aligned}\Pr(Y = 1 | X = b) &= \frac{\Pr(X = b | Y = 1)\Pr(Y = 1)}{\Pr(X = b)} \\ &= \frac{\frac{7}{12} \cdot 0,4}{0,43} \approx 0,54\end{aligned}$$

Rozkład prawdopodobieństwa

Liczba wylosowanych wartości w przedziałach



Rozkład jednorodny

Przykład
Wyniki rzutu kostką

Rozkład normalny

Przykład
Oceny z egzaminu

Wartości zmiennej losowej



Teoria decyzji

Uczenie maszynowe pozwala **wnioskować**,
np. o przynależności wektorów danych do klas

Jak na podstawie informacji uzyskanej na drodze automatycznego
wnioskowania **podejmować optymalne decyzje** dotyczące obiektów
stanowiących przedmiot badania?

Potencjalne problemy przy podejmowaniu decyzji:

Różne typy błędów mogą pociągać
za sobą różne konsekwencje.

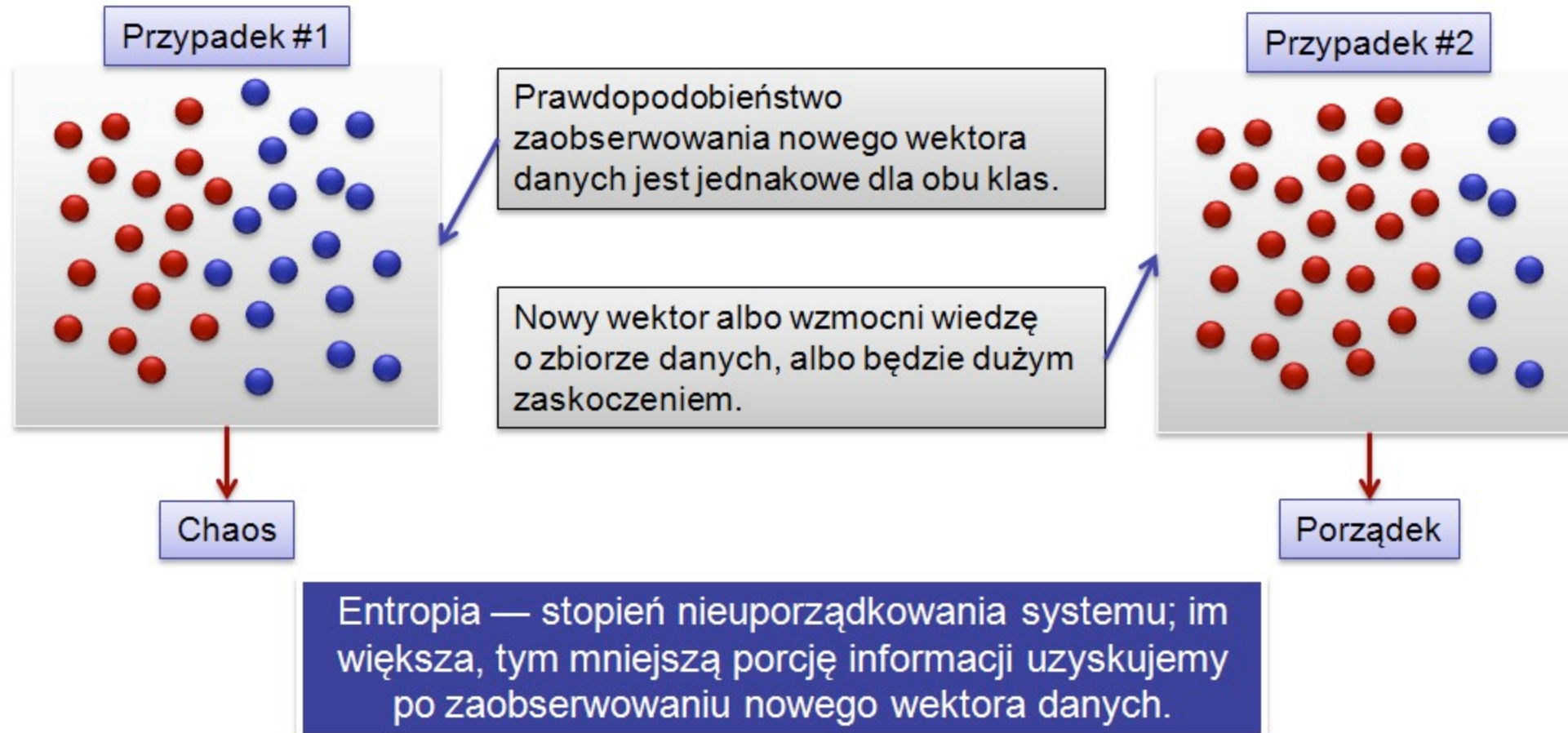
Różne, niezależne cechy obiektów mogą
prowadzić do odmiennych wniosków.

Mała próba wektorów danych może być
niewystarczająca, aby predykcja była wiarygodna.



Teoria informacji

Wektory danych w dwuwymiarowej przestrzeni cech z podziałem na klasy





W kolejnych wykładach...

- Metody uczenia oparte na wnioskowaniu indukcyjnym
 - Konstruowanie drzew i reguł decyzyjnych
- Klasyfikatory liniowe
 - Uogólnienie klasyfikatorów liniowych na przypadki klas separowanych nieliniowo
- Klasyfikacja w trybie uczenia nienadzorowanego
 - Podstawowe algorytmy grupowania
- Wymiar danych
 - Metody selekcji cech znaczących
- Ocena jakości klasyfikacji
 - Minimalizacja błędu rzeczywistego
 - Uczenie uwzględniające koszt błędów
- Uczenie nienadzorowane c.d.
 - Grupowanie rozmyte
 - Metody nienadzorowanej selekcji cech
- Transformacje danych





Literatura

W języku polskim

- Cichosz P., *Systemy uczące się*, WNT 2000
- Korancki J., Ćwik J., *Statystyczne systemy uczące się*, WNT 2006
- Koronacki J., Mielniczuk J., *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT 2006
- Handl D., Mannila H., Smyth P., *Eksploracja danych*, Chądzyńska A. (tłum.), WNT 2005

W języku angielskim

- Witten I.H., Frank E., *Data Mining. Practical Machine Learning Tools and Techniques* (2nd Ed.), Morgan Kaufman 2005
- Bishop C.M., *Pattern Recognition and Machine Learning*, Springer 2006

