GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

PODSTAWOWE ZAGADNIENIA I TERMINY

Wojciech Waloszek

wowal@eti.pg.gda.pl

Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

GDAŃSK UNIVERSITY OF TECHNOLOGY

Literatura

Larose D.: „Odkrywanie wiedzy z danych”, PWN, 2006

Witten I., Frank E., Hall M.: „Data Mining: Practical Machine Learning Tools and Techniques”, Morgan-Kaufmann, 2011

Hand D., Mannila H., Smyth P.: „Principles of Data Mining”, MIT Press, 2001

Larose D.: „Metody i modele eksploracji danych”, PWN, 2008

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przedmiot – warunki zaliczenia

Wykład + laboratorium = 100%
Wykład = 50%
Laboratorium = 50%

Zaliczenie przedmiotu wymaga zaliczenia wykładu i laboratorium!

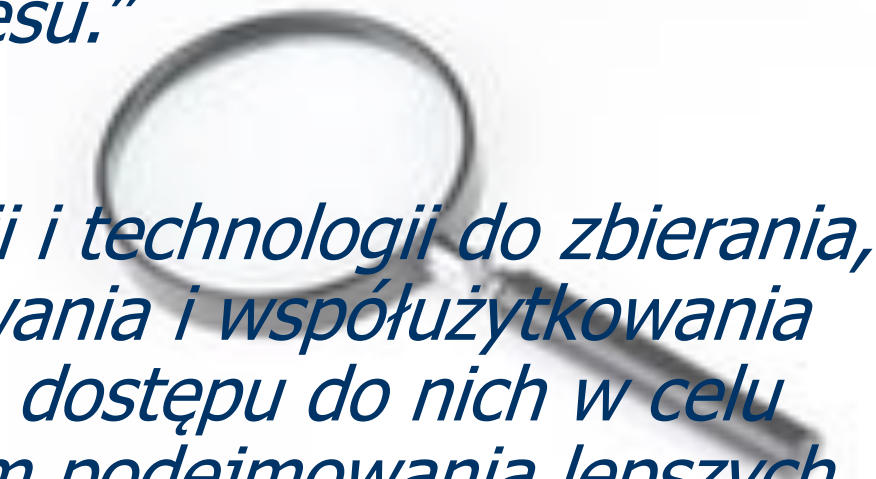
Zasady zaliczenia wykładu: zalicza 51% z kolokwium, które odbędzie się na końcu wykładów (połowa semestru).
Zasady zaliczenia laboratorium: podane na laboratoriach.

Ocena:
<51%, 60%) dostateczna,
<60%, 70%) dostateczna plus,
<70%, 80%) dobra,
<80%, 90%) dobra plus,
<90%, 100%> bardzo dobra,
(100%, ∞) celująca.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych - definicja

- *„Eksploracja i analiza dużej ilości danych w celu odkrycia nieznanych lub ukrytych, ale zrozumiałych informacji oraz użycie ich do podejmowania decyzji biznesowych i ich wdrażania poprzez formułowanie taktycznych i strategicznych inicjatyw oraz ocenę ich sukcesu.”*



- *„Szeroko kategoria aplikacji i technologii do zbierania, przechowywania, analizowania i współużytkowania danych oraz zapewniania dostępu do nich w celu umożliwienia użytkownikom podejmowania lepszych decyzji biznesowych.”*

GDAŃSK UNIVERSITY OF TECHNOLOGY

Rozwiązanie problemów biznesowych technikami eksploracji danych (1)

- „Zbyt dużo danych utrudnia ich analizę i skutkuje zmniejszeniem się ilości przydatnych informacji.”

- **Przewidywanie utraty klientów** – na podstawie historii zakupów oraz danych demograficznych możemy dokonać segmentacji klientów i określić, którzy z nich (i dlaczego) myślą o odejściu do konkurencji.

- **Wykrywanie nadużyć** – bazując na historii użycia karty kredytowej, banki automatycznie oceniają ryzyko, że dana operacja nie jest autoryzowana przez ich posiadacza i w nietypowych przypadkach kontaktują się z klientem w celu ich potwierdzenia.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Rozwiązanie problemów biznesowych technikami eksploracji danych (2)

- **Wykrywanie oszustw i nieprawidłowości** – zbierając dane o typowej aktywności użytkowników i porównując wyniki ich analizy z danymi z monitoringu, można wykryć nietypowe zachowania, w tym takie, które pozostałyby niewykryte przez tradycyjne systemy kontroli.

- **Budowanie skutecznych kampanii marketingowych** – analizując dane demograficzne, można wybrać osoby, które będą prawdopodobnie zainteresowane otrzymywaniem informacji o promocjach i powinny być objęte akcją marketingową.

- **Ocena ryzyka** – dysponując danymi historycznymi i demograficznymi, można ocenić prawdopodobieństwo spłaty kredytu lub pożyczki przez daną osobę.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Rozwiązanie problemów biznesowych technikami eksploracji danych (3)

- **Przewidywanie sprzedaży** –bazując na danych historycznych, można przewidzieć przyszłe wyniki sprzedaży danych produktów i odpowiednio wcześniej skorygować stan towarów na magazynach.

- **Zrozumienie potrzeb klientów** – segmentacja klientów pozwala ocenić potrzeby każdej z grup i określić czynniki, którymi kieruje się dana grupa klientów, wybierając poszczególne produkty. Możliwe jest też znalezienie czynników, które mają największy wpływ na podejmowane decyzje.

GDAŃSK UNIVERSITY OF TECHNOLOGY

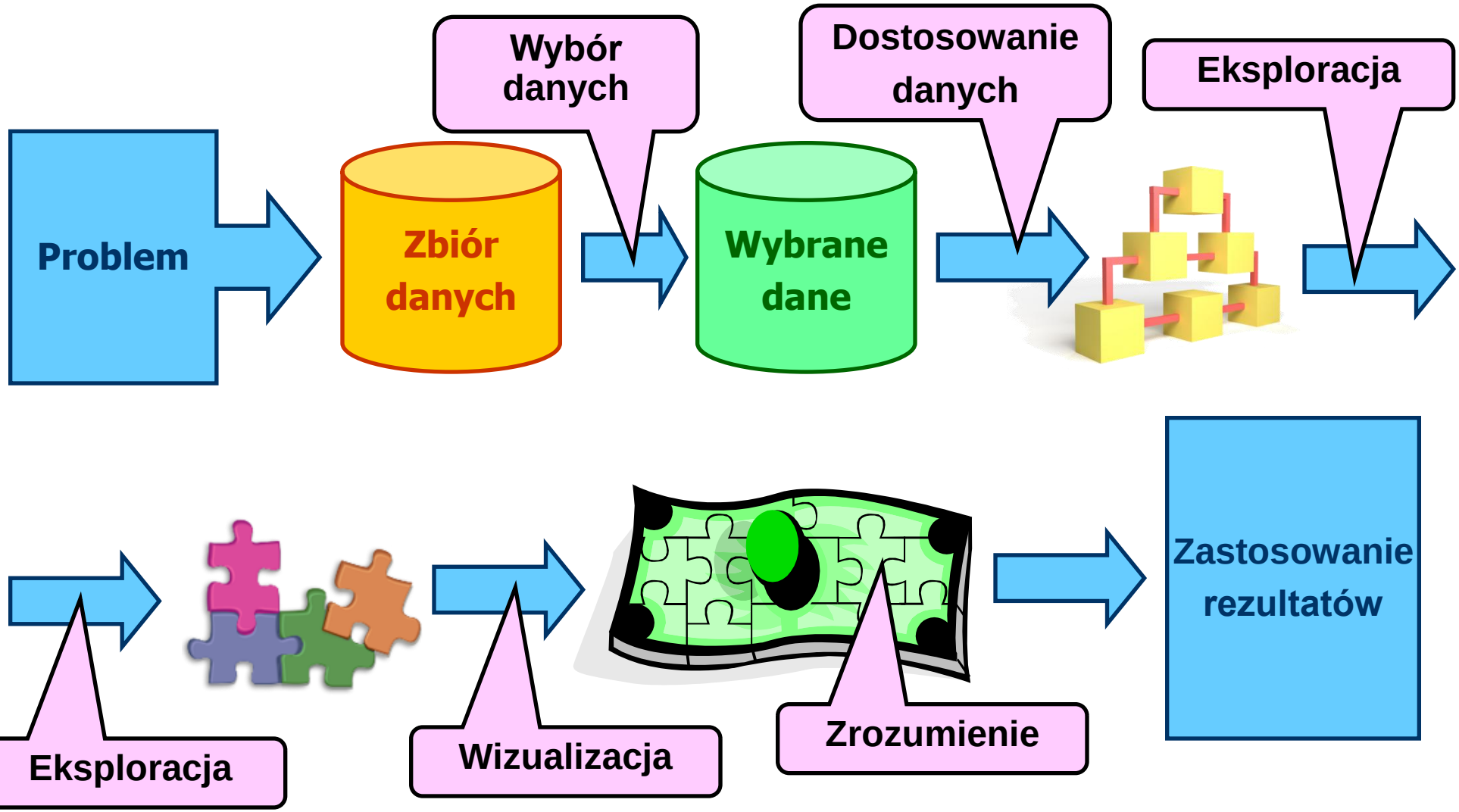
Rozwiązanie problemów biznesowych technikami eksploracji danych (4)

- **Szukanie klientów przynoszących zyski** – na podstawie danych osobowych można przewidzieć, która osoba (i z jakim prawdopodobieństwem) będzie dobrym klientem firmy.

- **Wyszukiwanie razem sprzedawanych towarów** – każdy sprzedawca powinien wiedzieć, które towary często kupowane są razem, a które prawie nigdy nie trafiają do tego samego koszyka. Zdobycie tej wiedzy umożliwia historia sprzedaży.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Tworzenie rozwiązań eksploracji danych



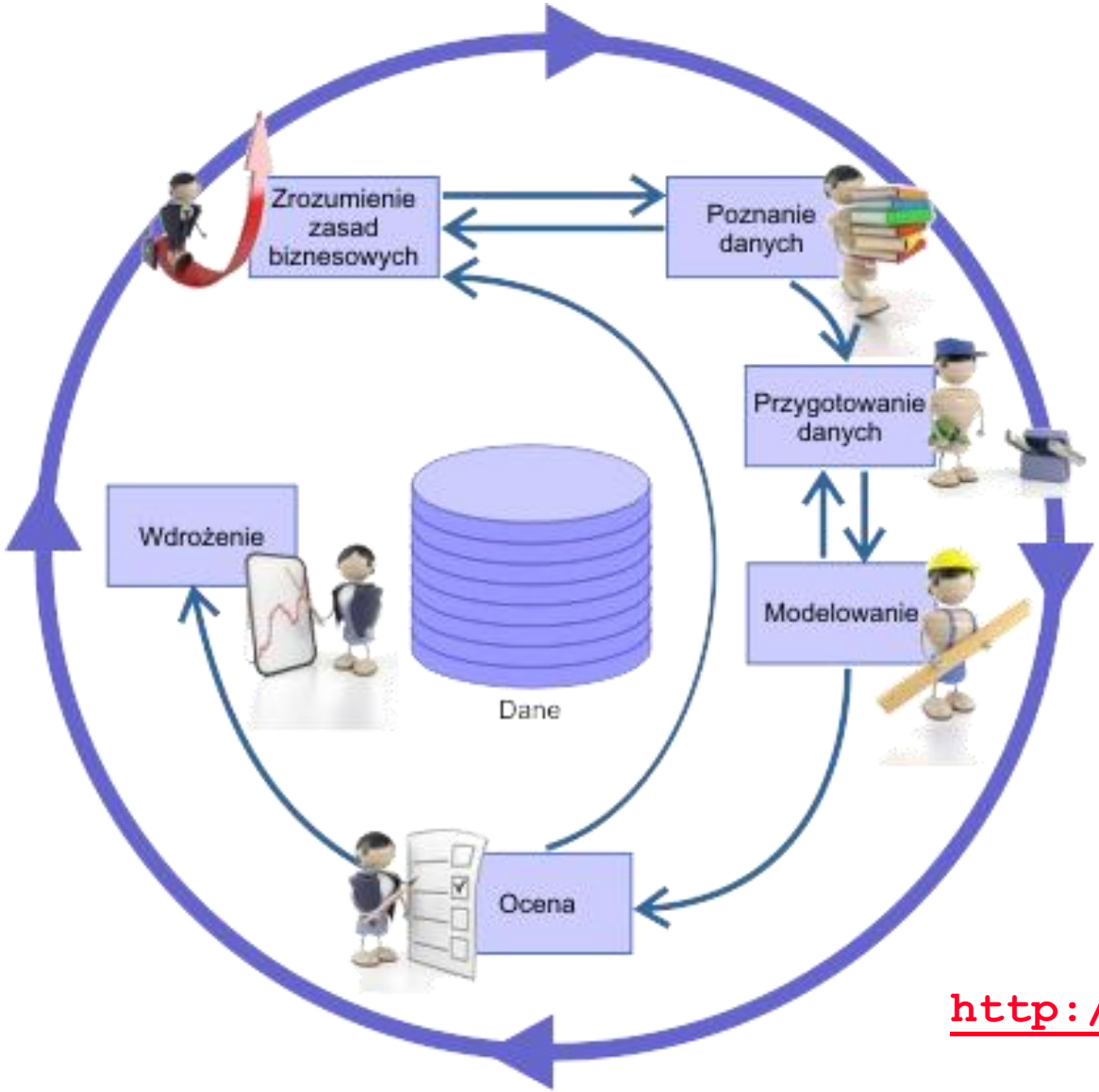
GDAŃSK UNIVERSITY OF TECHNOLOGY

Systemy typu BI

- Ułatwienie podejmowania decyzji na podstawie porównania aktualnych i historycznych danych z przyjętymi celami biznesowymi.
- Intuicyjne i łatwe do zinterpretowania wyniki analiz bieżącej sytuacji firmy.
- Oparte na danych historycznych przewidywania przyszłych sytuacji biznesowych.
- Identyfikowanie trendów i szybkie wykrywanie nieprawidłowości w działaniu firmy.

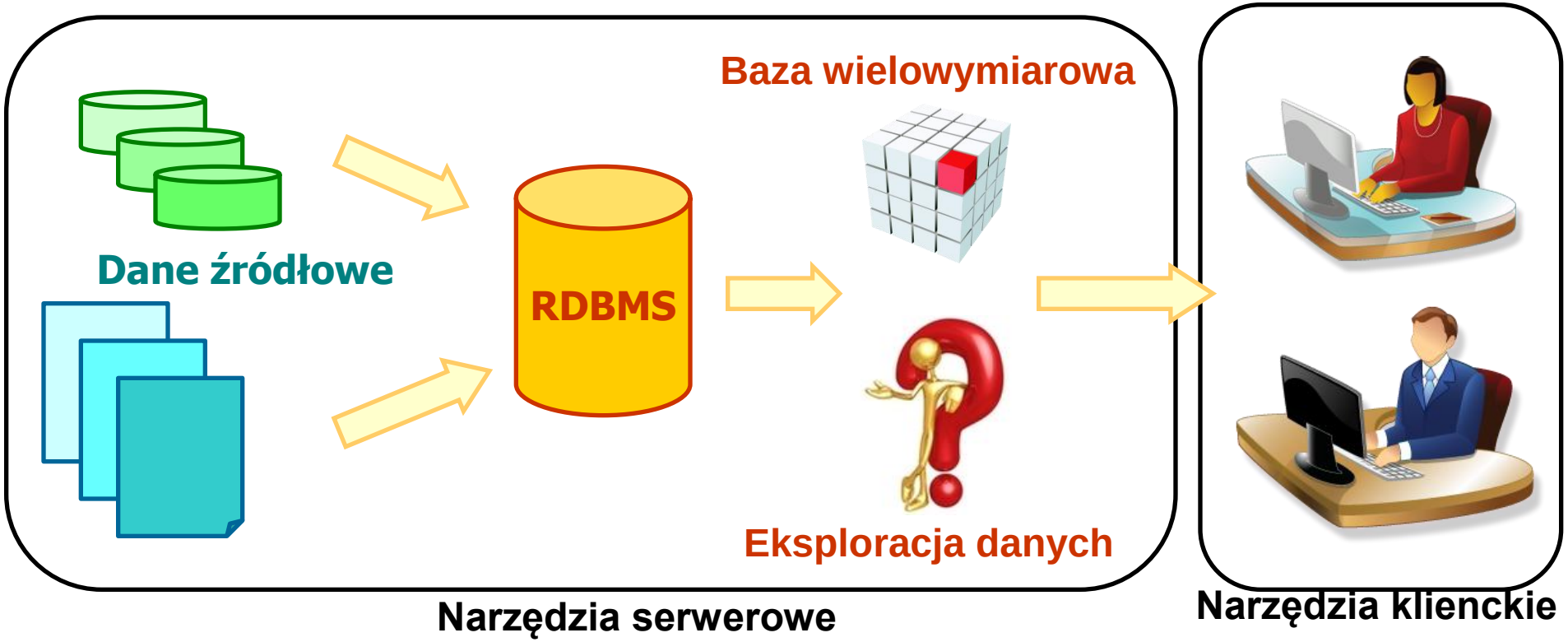
GDAŃSK UNIVERSITY OF TECHNOLOGY

Proces eksploracji danych (*Cross Industry Standard Process for Data Mining*)



http://www.crisp-dm.org

Hurtownie danych a eksploracja danych



„Bazy wielowymiarowe ułatwiają analizowanie danych historycznych, natomiast struktury eksploracji danych umożliwiają przewidywanie przyszłych zdarzeń lub wartości oraz ujawnianie ukrytych informacji w zgromadzonych danych.”

GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych jako dziedzina transdyscyplinarna

Metody eksploracji danych oparte są na szeroko rozumianej sztucznej inteligencji (*Machine learning)*.

Główne narzędzia to metody statystyczne oraz metody oparte na teorii informacji i logice matematycznej

GDAŃSK UNIVERSITY OF TECHNOLOGY

Pożądana forma danych wejściowych

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	**800**	**32**	**wyższe**	**tak**	**tak**
S	**1200**	**35**	**średnie**	**tak**	**tak**
S	**700**	**26**	**podstawowe**	**nie**	**nie**
M	**600**	**45**	**wyższe**	**nie**	**tak**
M	**650**	**38**	**średnie**	**tak**	**tak**
S	**900**	**28**	**wyższe**	**nie**	**nie**
S	**1100**	**65**	**średnie**	**tak**	**nie**
M	**500**	**22**	**średnie**	**nie**	**nie**
S	**800**	**43**	**podstawowe**	**tak**	**nie**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Wstępne przygotowanie danych

Przygotowanie danych wejściowych w formie macierzy *n* x *p* jest często zadaniem trudnym i pracochłonnym, wymagającym:

- przeanalizowania istniejących źródeł,
- identyfikacji danych błędnych i brakujących,
- przeprowadzenia procesu czyszczenia
- selekcji, integracji i transformacji danych,
- denormalizacji schematu

Uwaga na GIGO: Garbage In Garbage Out

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dane wejściowe

Wartość

Atrybut (zmienna)

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	**800**	**32**	**wyższe**	**tak**	**tak**
S	**1200**	**35**	**średnie**	**tak**	**tak**
S	**700**	**26**	**podstawowe**	**nie**	**nie**
M	**600**	**45**	**wyższe**	**nie**	**tak**
M	**650**	**38**	**średnie**	**tak**	**tak**
S	**900**	**28**	**wyższe**	**nie**	**nie**
S	**1100**	**65**	**średnie**	**tak**	**nie**
M	**500**	**22**	**średnie**	**nie**	**nie**
S	**800**	**43**	**podstawowe**	**tak**	**nie**

Przykład

Zbiór uczący

GDAŃSK UNIVERSITY OF TECHNOLOGY

Rodzaje atrybutów

- **Nominalne**
  - **Zbiór wartości dyskretnych**

Sam.
tak
nie

- **Porządkowe**
  - **Zbiór wartości dyskretnych z relacją porządku**

Wykształcenie
wyższe
średnie
podstawowe

- **Numeryczne**
  - **Ciągłe z zakresu liczb rzeczywistych (przedziałowe, ilorazowe(?))**

Wiek
28
35
26

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dane wejściowe (2)

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	**800**	**32**	**wyższe**	**tak**	**tak**
S	**1200**	**35**	**średnie**	**tak**	**tak**
S	**700**	**26**	**podstawowe**	**nie**	**nie**
M	**600**	**45**	**wyższe**	**nie**	**tak**
M	**650**	**38**	**średnie**	**tak**	**tak**
S	**900**	**28**	**wyższe**	**nie**	**nie**
S	**1100**	**65**	**średnie**	**tak**	**nie**
M	**500**	**22**	**średnie**	**nie**	**nie**
S	**800**	**43**	**podstawowe**	**tak**	**nie**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przygotowanie danych — uwagi

Specyficzne zagadnienia przygotowania danych dla procesu eksploracji obejmują m.in.:

- ocenę sposobu pozyskania i obciążenia próbki,
- ocenę reprezentacji wszystkich interesujących klas przykładów,
- ocenę wpływu denormalizacji na otrzymane wnioski,
- korektę typu atrybutów na potrzeby odpowiedniego modelu (np. dyskretyzacja atrybutów ciągłych),
- skalowanie danych
- selekcję atrybutów pod kątem wpływu na przebieg procesu eksploracji,
- selekcję i uzupełnianie przykładów

Zadania eksploracji danych

EDA — *exploratory data analysis*

Zestaw interaktywnych i wizualnych technik analizy zbioru danych bez bardziej precyzyjnego wskazania celu eksploracji. Często uznawany za element etapu wstępnej analizy danych. Główne narzędzia wizualizacji:

- histogramy,
- wykresy pudełkowe (*box plot*),
- wykresy punktowe (rozproszone, *scatter plot*),
- mapy konturowe (*contour plot*),
- macierze wykresów punktowych,
- skalowanie wielowymiarowe

Zadania eksploracji danych (2)

Predykcja: klasyfikacja i regresja

Zestaw technik nakierowanych na predykcję wartości jednego z atrybutów na podstawie wartości innych. Terminu klasyfikacja używamy, gdy przewidujemy wartość atrybutu nominalnego, regresji używamy dla przewidzenia wartości atrybutów numerycznych.

Deskrypcja: klasteryzacja i segmentacja

Zestaw technik nakierowanych na opis zbioru danych jako całości poprzez wyznaczenie globalnych charakterystyk oraz podział na przykładów na grupy.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zadania eksploracji danych (3)

Odkrywanie zależności: asocjacje

Zestaw technik nakierowanych na odkrywanie znaczących zależności pomiędzy wartościami różnych atrybutów bez wskazania konkretnego wyróżnionego atrybutu (wzorce).

Analiza szeregów czasowych i trendów

Zestaw technik nakierowanych konkretnie na analizę danych oznaczonych czasowo bądź zapisów łańcuchów zdarzeń. Często nie wyróżniany i zaliczany do predykcji.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Klasyfikacja

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. **Wyróżniamy pewien atrybut**
2. **Na postawie jego wartości wyróżniamy *klasy***
3. **Budujemy modele określające zasady przynależności do wyróżnionych klas w zależności od wartości pozostałych atrybutów**

GDAŃSK UNIVERSITY OF TECHNOLOGY

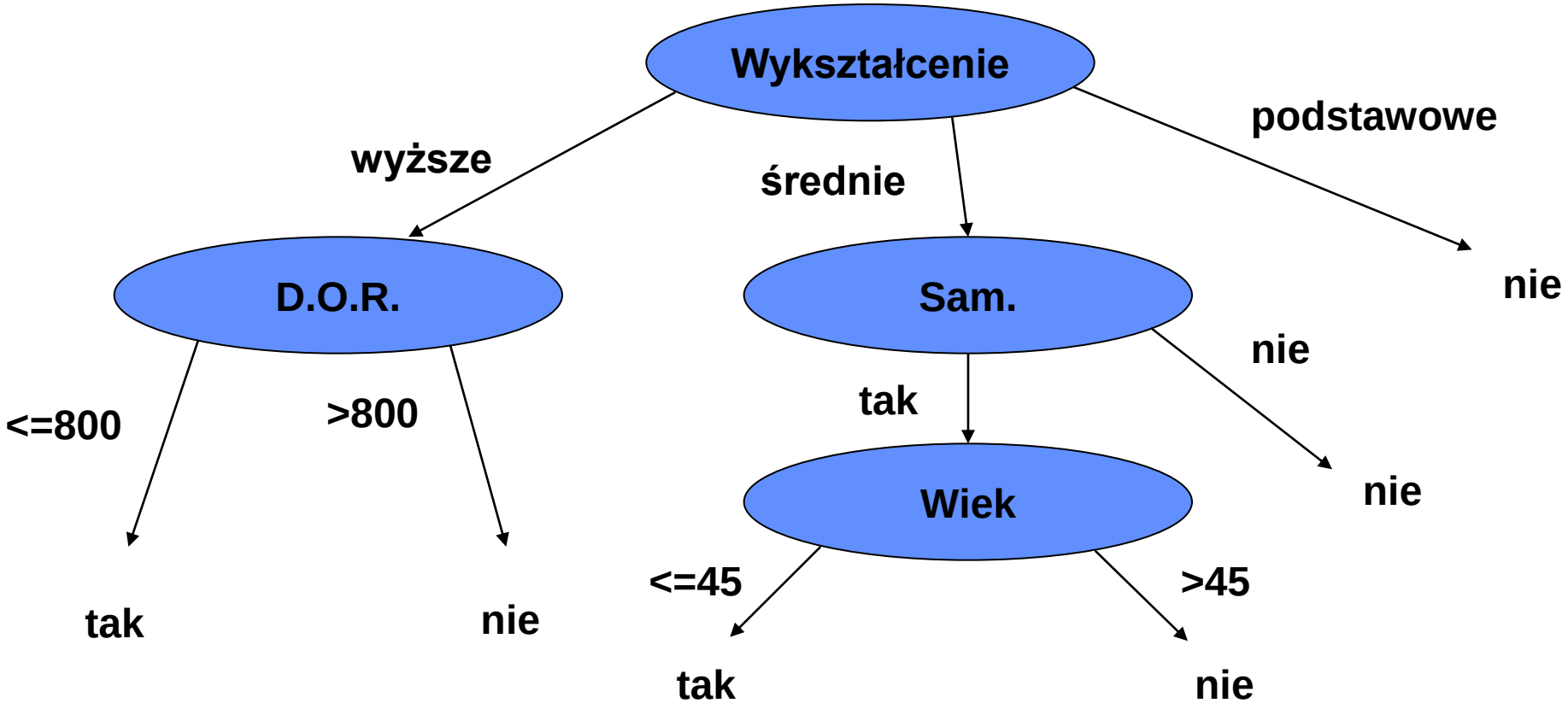
Klasyfikator decyzyjny - reguły

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

```
if Wykształcenie=podstawowe then Z.K.=nie
if Sam.=tak and Wiek<=45 then Z.K.=tak
if S.C.=S then Z.K.=nie
if D.O.R.<=500 then Z.K.=nie else Z.K.=tak
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

Klasyfikator decyzyjny – drzewa



GDAŃSK UNIVERSITY OF TECHNOLOGY

Asocjacje

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. **Nie wyróżniamy żadnego atrybutu**
2. **Szukamy zależności pomiędzy wartościami atrybutów w ramach przykładów**

GDAŃSK UNIVERSITY OF TECHNOLOGY

Reguły asocjacyjne

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

if Wykształcenie=średnie and Z.K.=tak then Sam.=tak

if Wykształcenie=podstawowe then S.C.=S

GDAŃSK UNIVERSITY OF TECHNOLOGY

Klasteryzacja

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

1. Polega na odnajdywaniu grup „podobnych” do siebie przykładów

2. Grupy te nazywamy *klastrami*

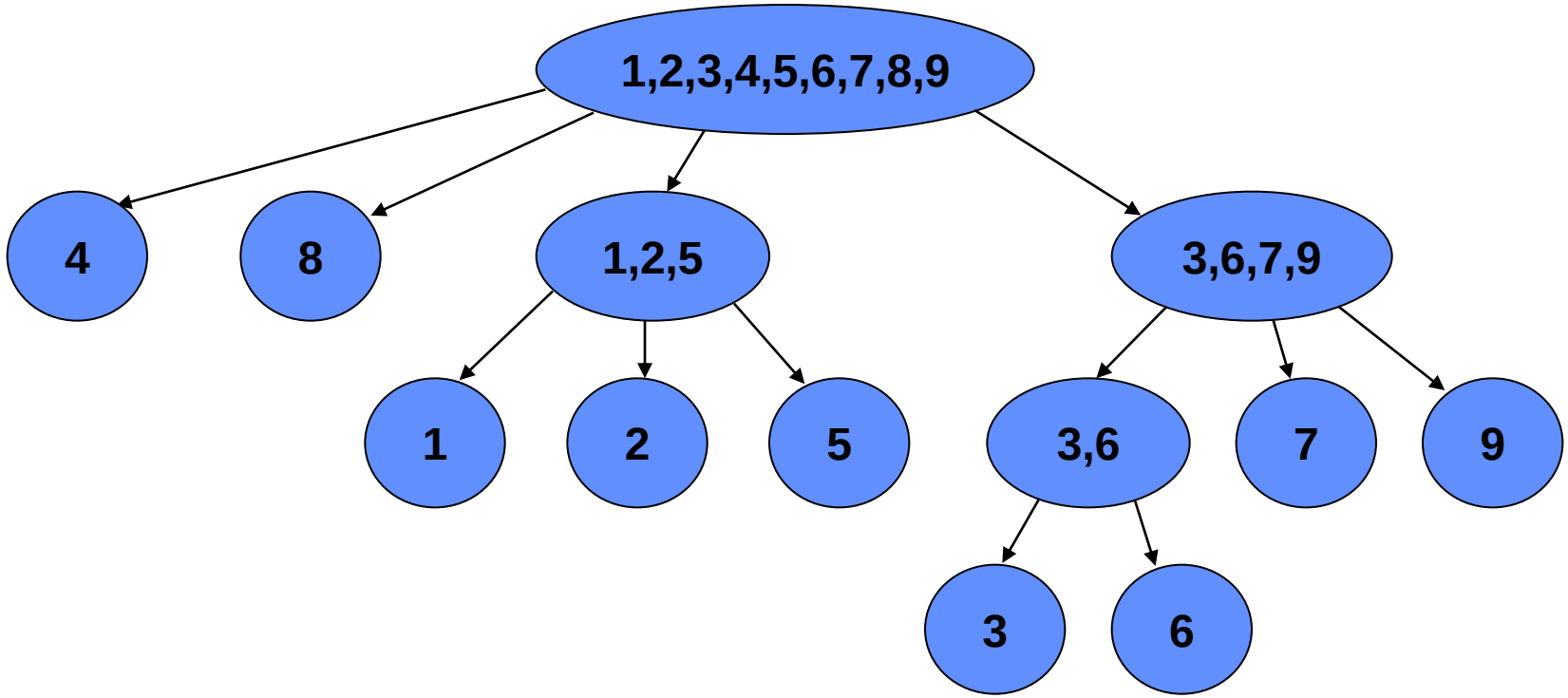
GDAŃSK UNIVERSITY OF TECHNOLOGY

Klasteryzacja - przykład

S.C.	D.O.R.	Wiek	Wykształcenie	Sam.	Z.K.
S	800	32	wyższe	tak	tak
S	1200	35	średnie	tak	tak
S	700	26	podstawowe	nie	nie
M	600	45	wyższe	nie	tak
M	650	38	średnie	tak	tak
S	900	28	wyższe	nie	nie
S	1100	65	średnie	tak	nie
M	500	22	średnie	nie	nie
S	800	43	podstawowe	tak	nie

GDAŃSK UNIVERSITY OF TECHNOLOGY

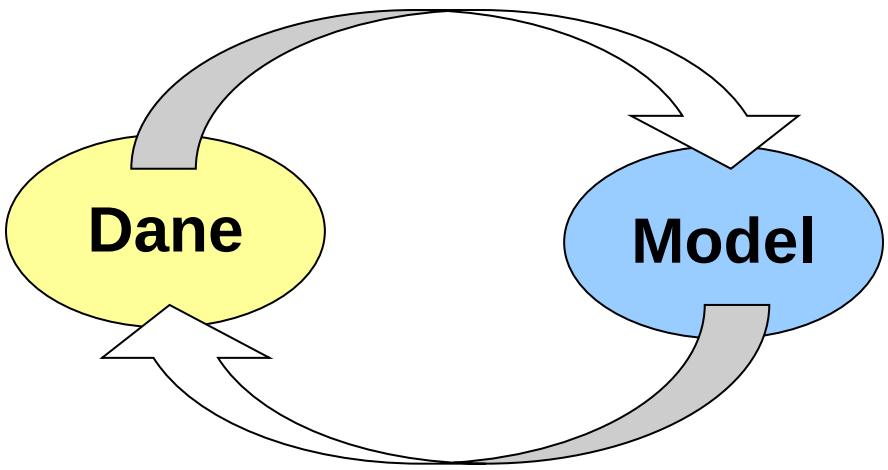
Klasteryzacja hierarchiczna

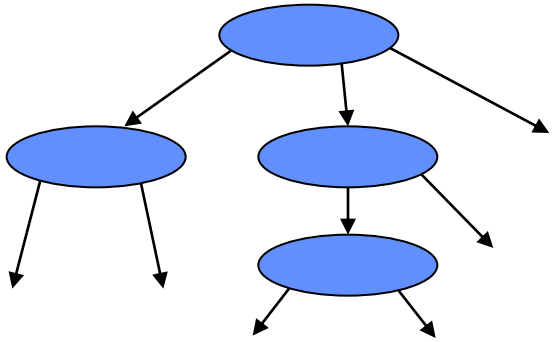


GDAŃSK UNIVERSITY OF TECHNOLOGY

Tworzenie modelu danych

Proces eksploracji danych można rozumieć jako dopasowywanie modelu do danych





GDAŃSK UNIVERSITY OF TECHNOLOGY

Model danych – komentarz

„Wszystkie modele są złe,

ale niektóre użyteczne”

(prz. George Box)

Wyniki są nieodwracalnie obciążone błędami

- Skończone i niepełne zbiory trenujące,
- Błędy w zbiorach trenujących,
- Uproszczenia w stosowanych algorytmach,
- Brak znajomości semantyki przetwarzanych danych

GDAŃSK UNIVERSITY OF TECHNOLOGY

Błąd nadmiernego dopasowania

Nazwisko	Wiek	Wykształcenie	Sam.	Z.K.
Abacki	32	wyższe	tak	tak
Babacka	35	średnie	tak	tak
Cabacki	26	podstawowe	nie	nie
Dabacka	45	wyższe	nie	tak
Ebacki	38	średnie	tak	tak
Fabacki	28	wyższe	nie	nie
Gabacka	65	średnie	tak	nie
Habacka	22	średnie	nie	nie
Ibacki	43	podstawowe	tak	nie

```
if Nazwisko=Abacki then Z.K.=tak
if Nazwisko=Babacka then Z.K.=tak
if Nazwisko=Cabacki then Z.K.=nie
  ...
```

GDAŃSK UNIVERSITY OF TECHNOLOGY

Obrona

Zwiększanie właściwości predykcyjnych:

- Podział zbioru trenującego na część trenującą i testującą

Uogólnianie modelu:

- Przycinanie drzew decyzyjnych,
- Przycinanie reguł,
- Łączenie klastrów

GDAŃSK UNIVERSITY OF TECHNOLOGY

Techniki oceny wyników

Analiza wyników:

- Macierze błędu,
- Krzywe ROC,
- Wykres wzrostu (przyrostu)

Badanie siły predykcyjnej modelu:

- *m n-fold cross-validation with stratification*,
- 0,368 bootstrap
- leave-one-out

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

EKSPLORACYJNA ANALIZA DANYCH (EDA)
I PRZEKSZTAŁCANIE DANYCH