# Wprowadzenie do technologii informacyjnej.

## Data mining i jego biznesowe zastosowania

**dr Tomasz Jach**

# Definicje

*„Eksploracja danych polega na torturowaniu danych tak długo, aż zaczną zeznawać.”*

*„Eksploracja danych jest procesem odkrywania znaczących nowych powiązań, wzorców i trendów przez przeszukiwanie danych zgromadzonych w skarbnicach, przy wykorzystaniu metod rozpoznawania wzorców, jak również metod statystycznych i matematycznych”.*

# Po co Data Mining?

**Sytuacja**:

- Eksplozja gromadzenia danych (Miliony rekordów - GB, TB danych)
- Wielowymiarowa struktura danych (Setki zmiennych)
- Złożone zależności występujące w danych

**Problem**:

- Potrzeba wydobycia wiedzy biznesowej z danych
- Standardowe narzędzia nie wystarczają (SQL, Excel, OLAP)

# Możliwości data mining

- **Analiza rzeczywistych danych o zachowaniach i preferencjach klientów**, wynikających z wykorzystywania oferowanych przez przedsiębiorstwo produktów i usług.

- **Łączenie wiedzy i doświadczenia z zakresu IT, analizy danych i data mining z wiedzą z zakresu marketingu, sprzedaży i ekonomii.** Dzięki temu wiemy jakie dane o klientach należy wykorzystać oraz jak je analizować, aby prowadzić skuteczne działania marketingowe i sprzedażowe.

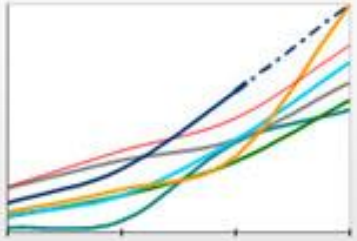
# Dlaczego potrzebujemy eksploracji danych?

Forecasting

Anomalies

Clustering

Influencers



Trends

# W czym tkwi siła Data Mining ?

W data mining wykorzystuje się narzędzia pochodzące z trzech dziedzin:

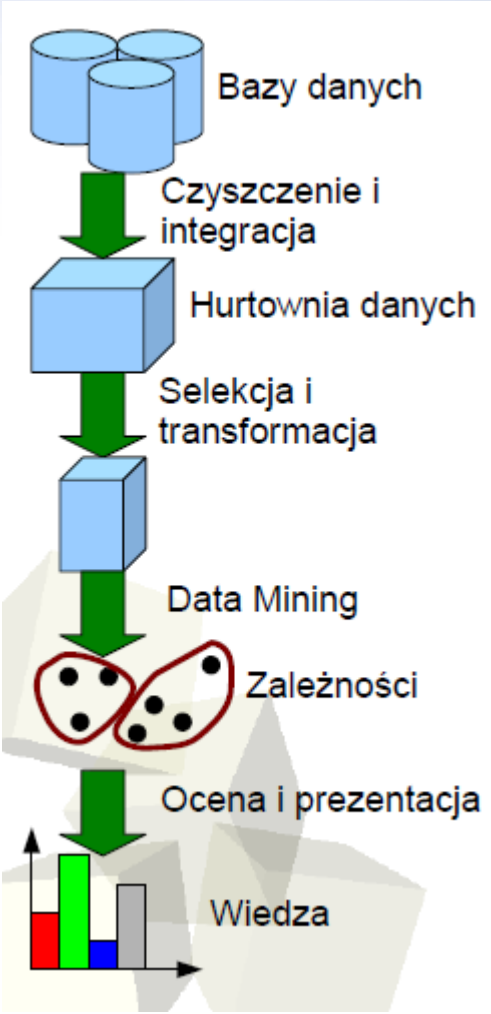
- Technologii bazodanowej (gromadzenie, udostępnianie i przetwarzanie danych),
- Statystyki,
- Uczenia maszyn i sztucznej inteligencji.

W procesie data mining możemy wyróżnić cztery zasadnicze etapy:

- 1. Przygotowanie danych,
- 2. Eksploracyjna analiza danych,
- 3. Właściwa analiza danych (budowa i ocena modelu lub odkrywanie wiedzy),
- 4. Wdrożenie i stosowanie modelu.

# Proces odkrywania wiedzy z BD

- **Czyszczenie danych** – usuwanie szumu i niespójnych danych
- **Integracja danych** – łączenie danych z rożnych źródeł
- **Selekcja danych** – wybór ważnych (dla problemu) danych
- **Transformacja danych** – do postaci odpowiedniej do DM (np. sumowanie, agregacja)
- **DM** – zastosowanie inteligentnych metod do wydobycia zależności, wzorców
- **Ocena zależności** – identyfikacja interesujących zależności ze wszystkich wydobytych
- **Prezentacja wiedzy**



# Metody eksploracji danych

1. **Odkrywanie asocjacji** *(associations):* znajdowanie reguł typu: piwo -> orzeszki

2. **Wzorce sekwencji** *(sequential patterns):* znajdowanie sekwencji dot. np. analiza kliknięć na stronie WWW (co po kolei było klikane). Także **predykcja** przyszłych wydarzeń na podstawie przeszłości.

3. **Klasyfikacja** *(classifications):* klasyfikacja danych do grup ze względu na atrybut decyzyjny, np.: klasyfikacja klientów przez bank do grup: dać kredyt / nie dać kredytu

4. **Analiza skupień** *(clustering):* grupowanie danych na wcześniej nieznane klasy, znajdowanie wspólnych cech, np.: zagregowanie statystyk urządzenia sieciowego i określenie jego „zdrowia”.

5. **Podobieństwo szeregów czasowych** *(time-series similarities):* badanie podobieństwa przebiegów czasowych, np. wykresów giełdowych

6. **Wykrywanie odchyleń** *(deviation detection):* znajdowanie anomalii, wyjątków, np.: urządzenie radiowe działające inaczej niż inne, fluktuacje

# Asocjacje

- **Analiza koszykowa** to metoda z zakresu eksploracji danych, tworząca dla zbioru danych zestaw opisujących go przybliżonych reguł typu:

Jeżeli (typ_samochodu='sportowy' i wiek<25) to zwykle (ryzyko='wysokie' i ubezpieczenie='wysokie')

- Każda utworzona reguła opisana jest dwoma miarami **wsparcia** (support) oraz **zaufania** (confidence):

$$wsparcie(A) = w\ ilu\ transakcjach\ występuje\ A$$

$$wsparcie\ względne(A \rightarrow B) = P(A \cup B) = \frac{wsparcie(A \cup B)}{ilość\ wszystkich\ transakcji}$$

$$zaufanie(A \rightarrow B) = P(B|A) = \frac{P(A \cup B)}{P(A)} = \frac{wsparcie(A \cup B)}{wsparcie(A)}$$

# Idea reguł asocjacyjnych

| Nr trans. | Kupione towary | | | | | |
|---|---|---|---|---|---|---|
| 1 | T0 | | T2 | T3 | | T5 |
| 2 | T0 | T1 | T2 | T3 | | |
| 3 | | | T2 | | T4 | T5 |
| 4 | T0 | T1 | T2 | T3 | | |
| 5 | T0 | | T2 | | | T5 |

| Reguła | support | confidence |
|---|---|---|
| T0⇒T2 | 80 % | 100 % |
| T0,T2⇒ T3 | 60 % | 75 % |

Zachodzącą regułę: T0⇒T2 o wsparciu 80% i zaufaniu 100% możemy zinterpretować następująco:

100 % osób, którzy kupili towar T0 kupili również towar T2 a sytuacja ta zachodzi w 80 % wszystkich transakcji.

# Klasyfikacja vs. predykcja

Klasyfikacja

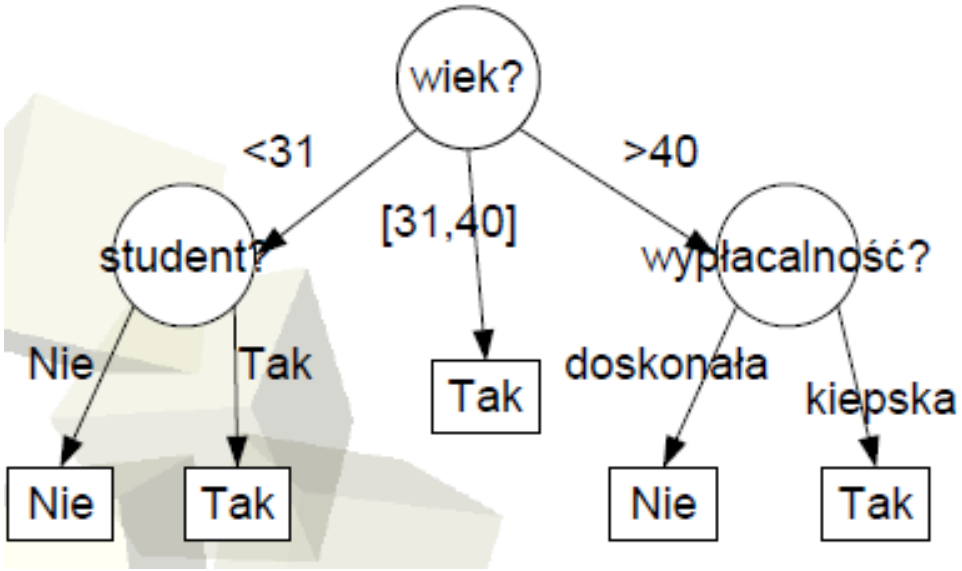
- Drzewa decyzyjne (ID3, CART, C4.5)
- Modele Bayes'a
- Sieci neuronowe (perceptron)
- k-Nearest Neighbours
- Algorytmy Genetyczne
- Zbiory rozmyte i przybliżone

Predykcja

- Statystyczna regresja wielowymiarowa, inne rodzaje regresji

# Drzewo decyzyjne

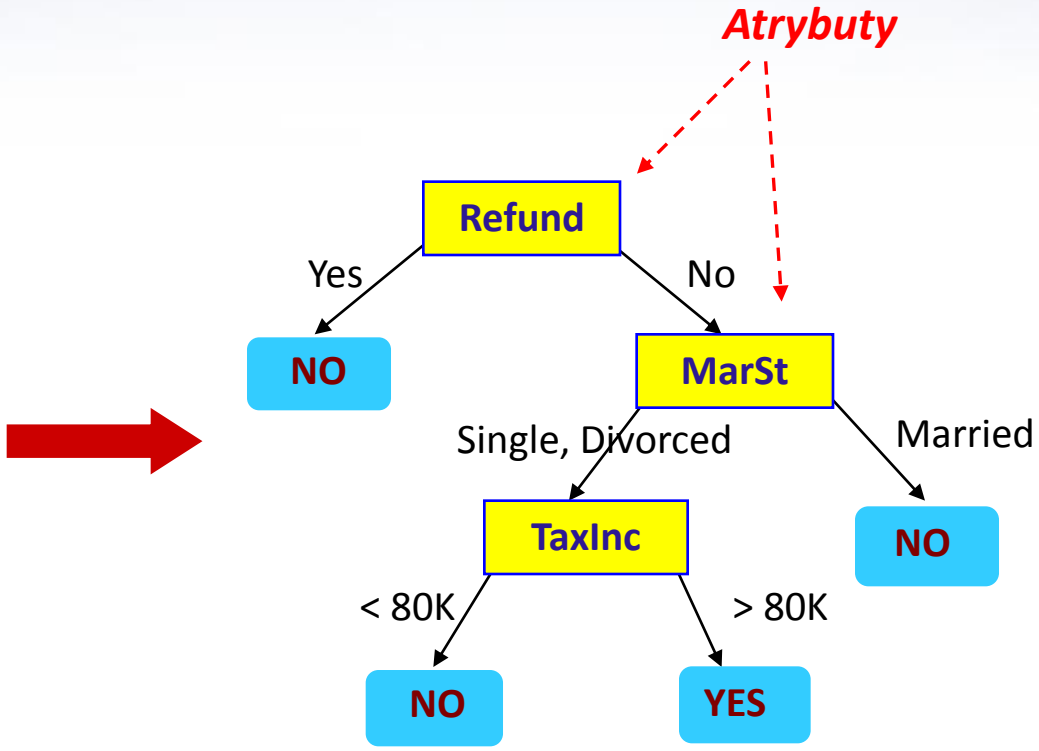
- Drzewa decyzyjne wyznaczają reguły, które pozwalają przypisać obiekty do określonych klas.
- Analiza zbioru obiektów dokonywana jest po kątem przyjętego zestawu atrybutów, a celem analizy jest podział obiektów na jednorodne klasy.
- Podział zbioru ma charakter hierarchicznv.



# Przykład

| | Nominalna | nominalna | ciągła | Klasa |
|---|---|---|---|---|
| *Tid* | **Refund** | **Marital Status** | **Taxable Income** | **Cheat** |
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

**Zbiór treningowy**




**Model: Drzewo decyzyjne**

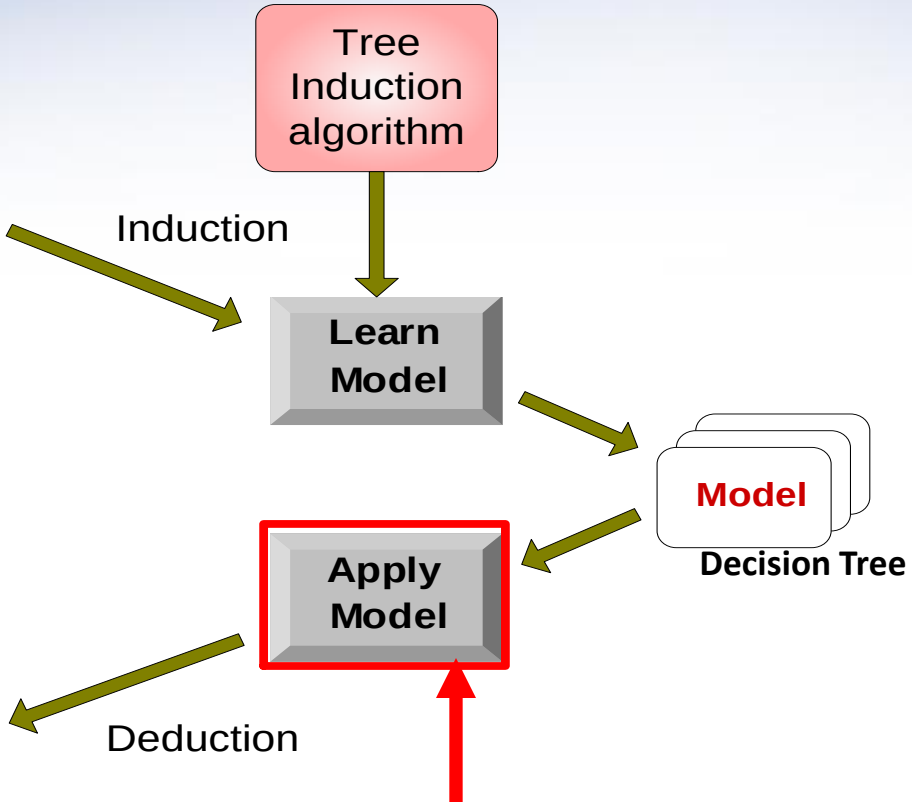
# Zadanie klasyfikacji za pomocą DD

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 1 | Yes | Large | 125K | No |
| 2 | No | Medium | 100K | No |
| 3 | No | Small | 70K | No |
| 4 | Yes | Medium | 120K | No |
| 5 | No | Large | 95K | Yes |
| 6 | No | Medium | 60K | No |
| 7 | Yes | Large | 220K | No |
| 8 | No | Small | 85K | Yes |
| 9 | No | Medium | 75K | No |
| 10 | No | Small | 90K | Yes |

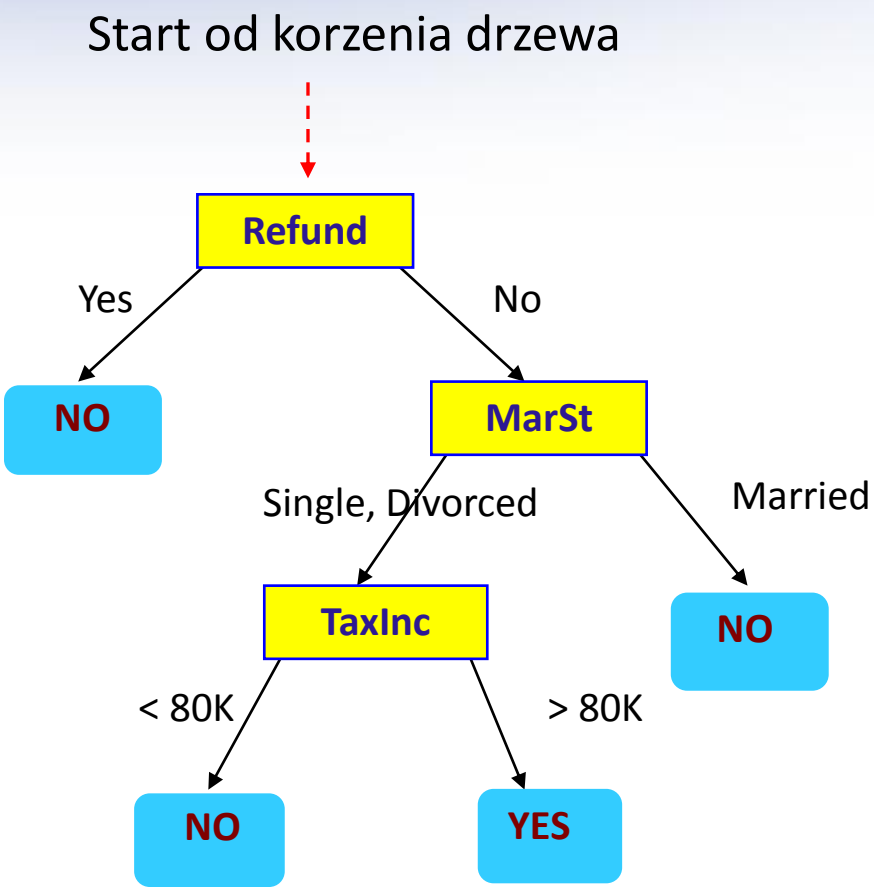
Training Set

| Tid | Attrib1 | Attrib2 | Attrib3 | Class |
|---|---|---|---|---|
| 11 | No | Small | 55K | ? |
| 12 | Yes | Medium | 80K | ? |
| 13 | Yes | Large | 110K | ? |
| 14 | No | Small | 95K | ? |
| 15 | No | Large | 67K | ? |

Test Set



# Użycie modelu DD do nowych danych



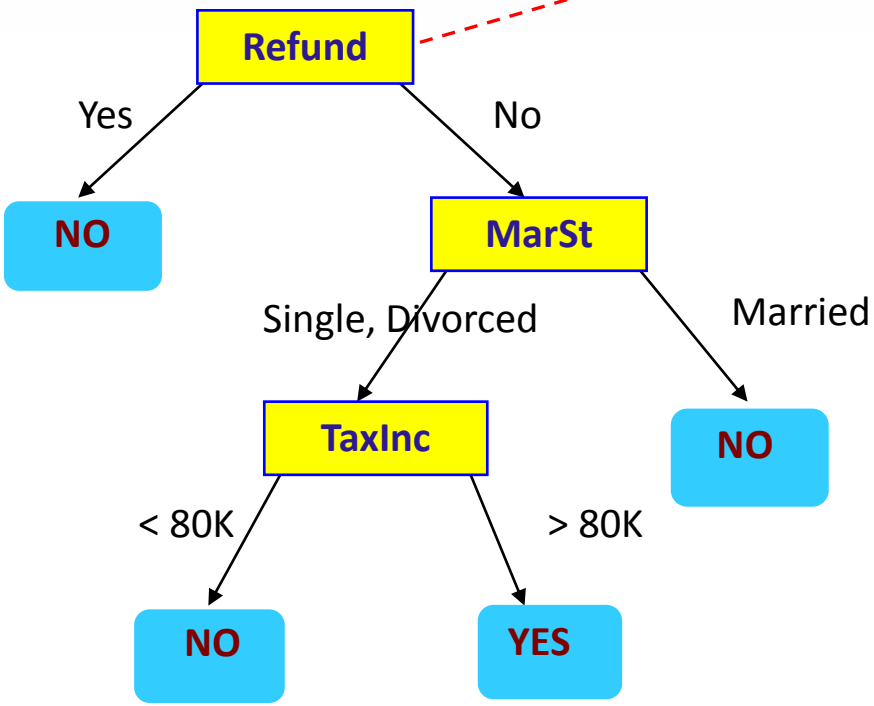

## Nowe dane

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |

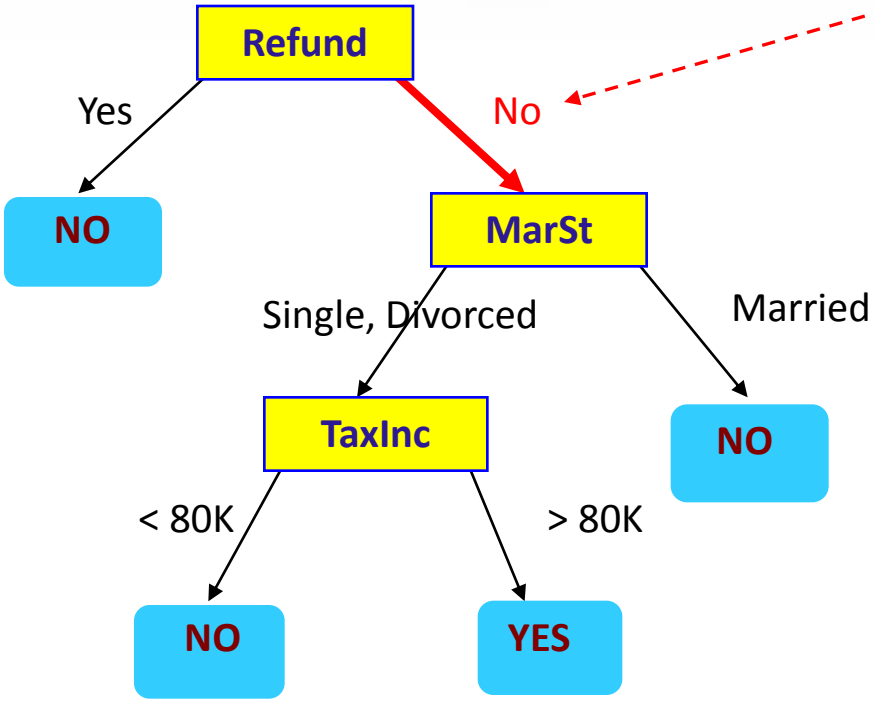
# Użycie modelu DD do nowych danych

## Nowe dane

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |



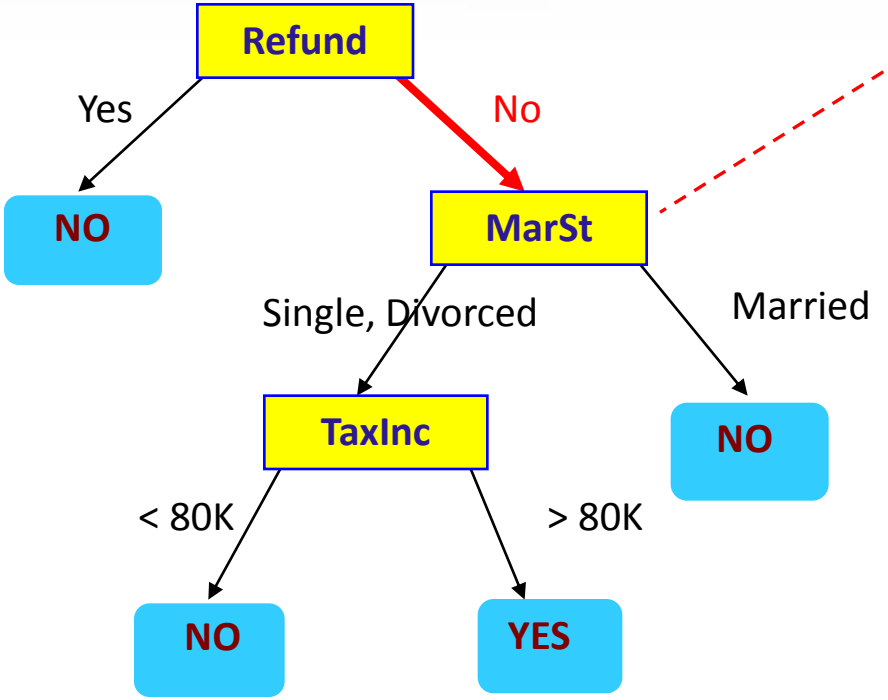
# Użycie modelu DD do nowych danych




## Nowe dane

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |

# Użycie modelu DD do nowych danych



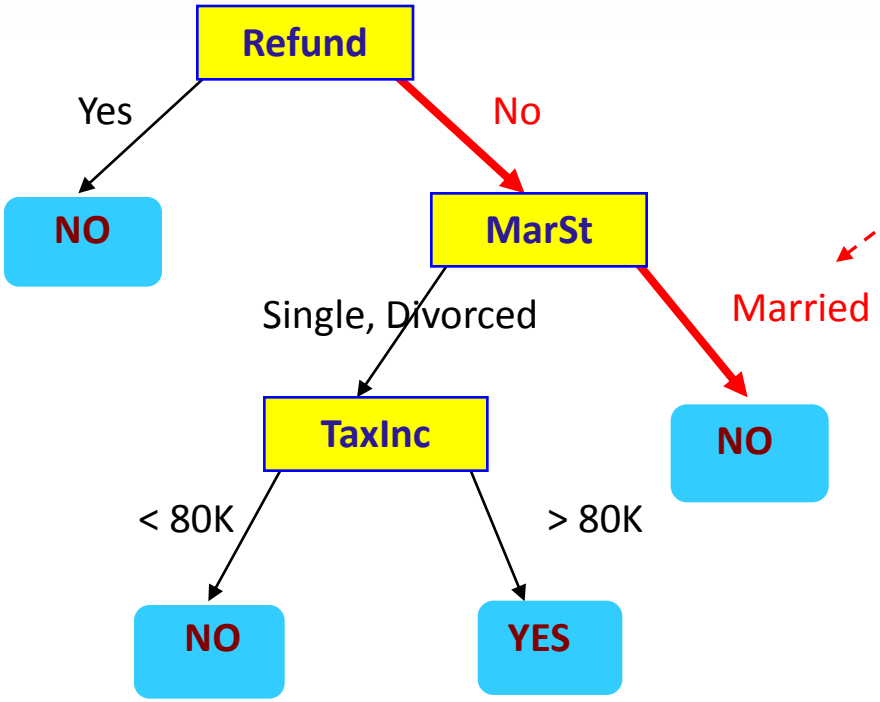

## Nowe dane

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |

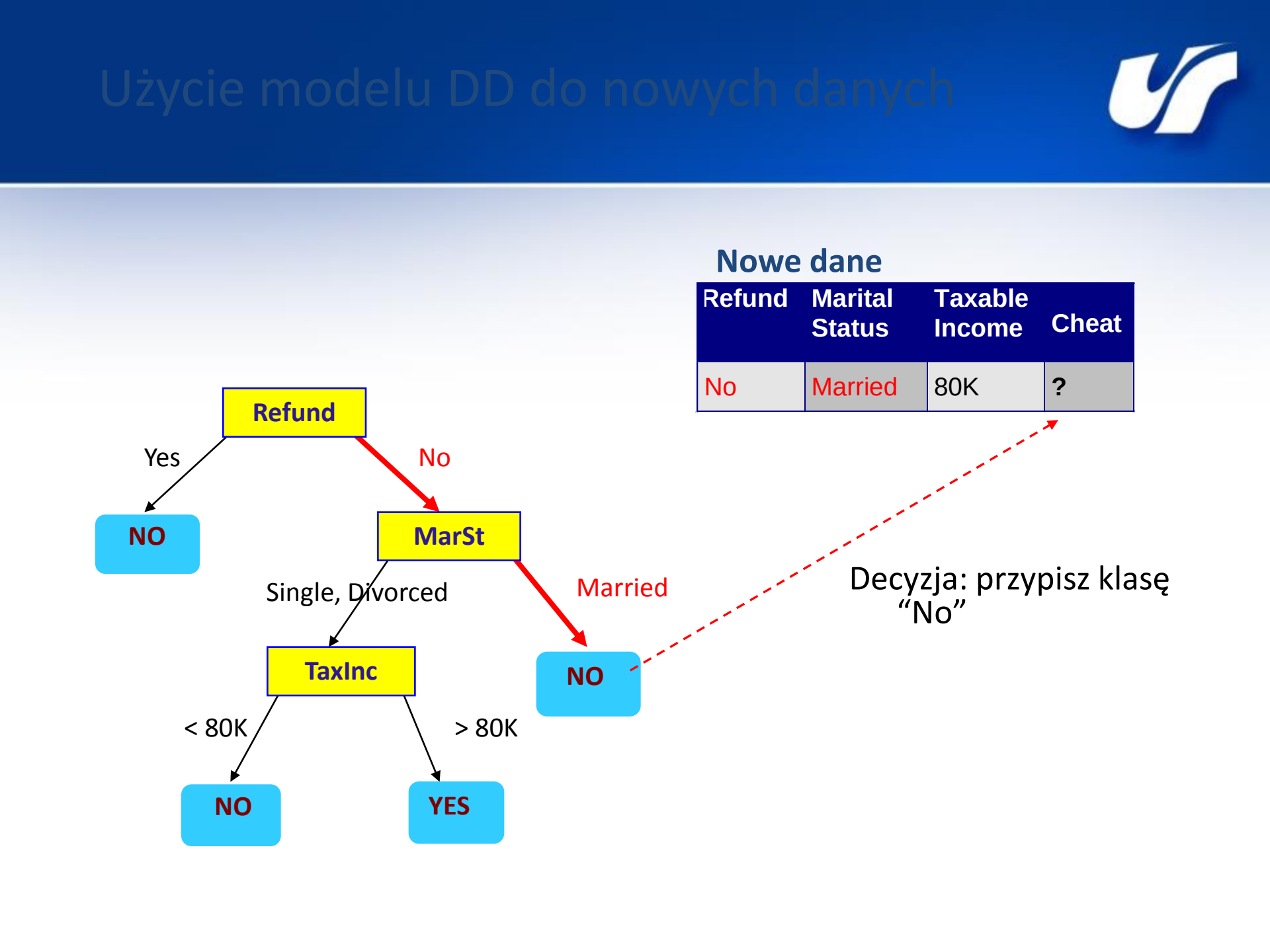
# Użycie modelu DD do nowych danych

## Nowe dane

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |



# Użycie modelu DD do nowych danych



**Nowe dane**

| Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|
| No | Married | 80K | ? |



Decyzja: przypisz klasę "No"

# Zalety i wady drzew decyzyjnych

Do **zalet** drzew decyzyjnych zaliczyć można :

- drzewa decyzyjne mogą reprezentować dowolnie złożone pojęcia, których definicje można wyrazić w zależności od atrybutów,
- przynoszą korzystne efekty przy analizie dużych zbiorów danych,
- czytelna forma graficzna - jeżeli tylko drzewo nie jest zbyt złożone,
- możliwość przejścia od grafu do opisu poprzez reguły.

Do **wad** drzew zalicza się :

- nie stwarzają możliwości aktualizacji (dodania) nowych danych,
- zjawisko przeuczenia drzewa.

# Czym jest grupowanie danych ?

Grupowanie polega na podziale niejednorodnej grupy obiektów (klientów) na grupy.

Wszystkie osoby znajdujące się w tej samej grupie uważane są za podobne do siebie, osoby znajdujące się w różnych grupach są różne.

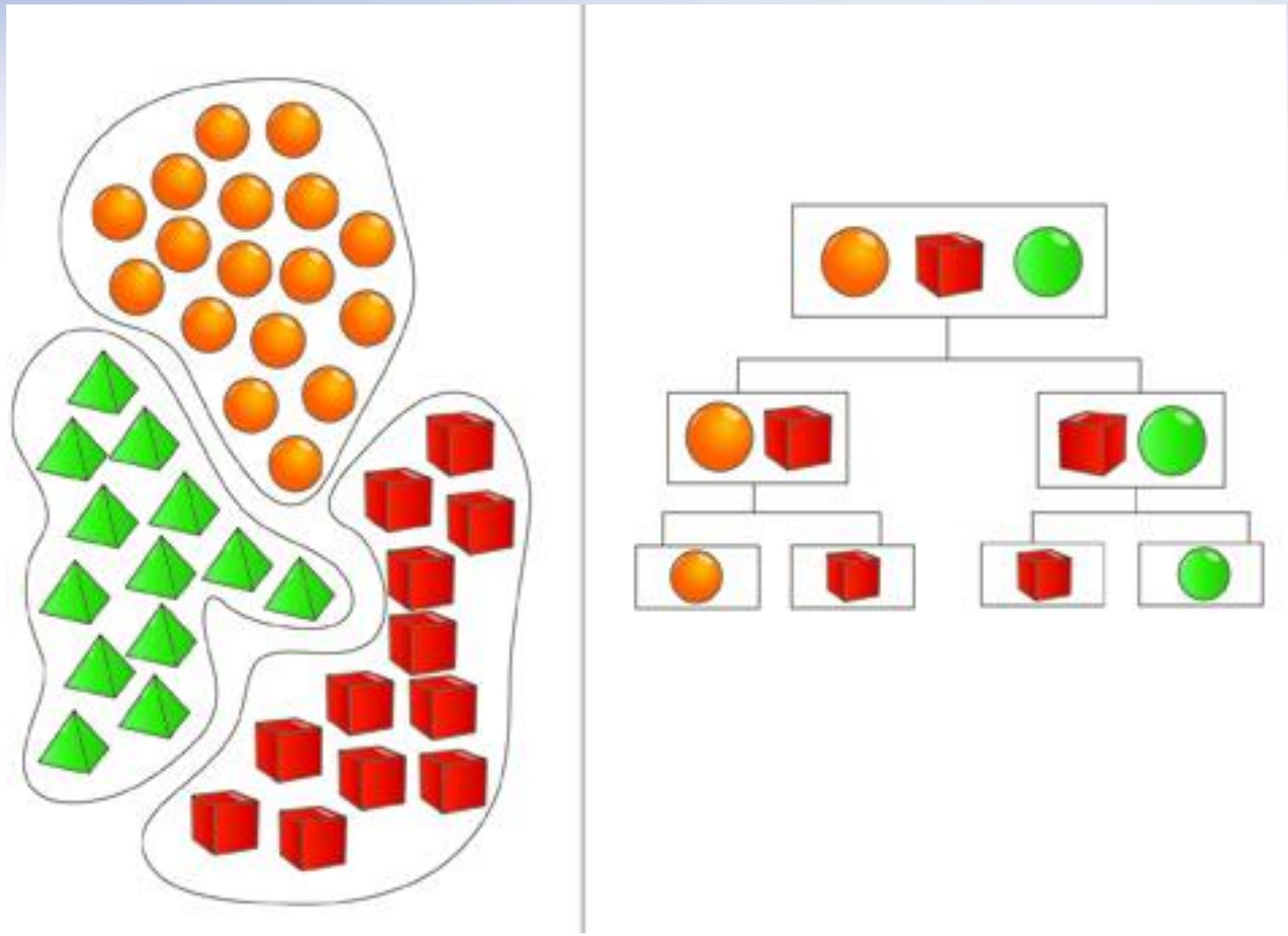
Dzięki tego typu podziałowi nie musimy już określać tylu strategii, ilu mamy klientów. Wystarczy jeśli dany sposób postępowania przypiszemy do całej grupy (segmentu) podobnych osób.

Najbardziej popularnymi metodami stosowanymi do segmentacji są metody **analizy skupień** oraz **samoorganizujące się mapy Kohonena** (SOM).

# Cel analizy skupień

- Celem **analizy skupień** (cluster analysis) jest wyodrębnienie ze zbioru danych obiektów, które byłyby podobne do siebie, i łączenie ich w grupy.
- W wyniku działania tej analizy z jednego niejednorodnego zbioru danych otrzymujemy grupę kilku jednorodnych zbiorów.
- Obiekty znajdujące się w tym samym zbiorze uznawane są za „podobne do siebie”, obiekty z różnych zbiorów traktowane są jako „niepodobne”.
- Pojęcie analizy skupień obejmuje szereg algorytmów klasyfikacji. Do najważniejszych należy zaliczyć metodę k-średnich oraz AHC.

# Metody niehierarchiczne oraz hierarchiczne



# Grupowanie – segmentacja danych

- Klasyczne metody niehierarchiczne (K-Means, k-Medoids)
- Metody hierarchiczne (AHC, Agnes, BIRCH, CURE, Chameleon)
- Metody oparte na gęstości (DBSCAN, OPTICS, DENCLUE)
- Metody gridowe (STING, WaveCluster, CLIQUE)
- Metody oparte na modelu (Podejście statystyczne, sieci neuronowe)
- Analiza odchyleń (ang. outlier analysis)

# Moc segmentacji danych

- Zaletami segmentacji jest możliwość patrzenia na naszych klientów z nieco innej perspektywy, niejako z lotu ptaka.
- Taka perspektywa może umożliwić lepsze zrozumienie ich cech i zachowań. Dzięki temu mamy możliwość spójnie i precyzyjne określić grupy klientów, którymi jesteśmy zainteresowani.
- Wyniki przeprowadzonej segmentacji mogą również zasugerować działania, jakie należy podjąć w stosunku do określonych grup.