# WPROWADZENIE DO KDD I DATA MINING

2/20/2008

Wykład 1

Nguyen Hung Son

# Plan wykładu



- Motywacja: Dlaczego data mining?
- Definicje i przykłady zastosowań
- Cele i zagadnienia w data mining
- Przegląd metod w data mining
- Data mining jako proces w projektach KDD

# Motywacja: wielkie bazy danych



- Problem eksplozji danych
  - Narzędzia zbierania danych+rozwój systemów bazodanowych
    →
    gwałtowny wzrost ilości danych zgromadzonych w bazach danych, hurtowniach danych i magazynach danych
  - Np.:
    - N = $10^9$ rekordów w danych astronomicznych,
    - d = $10^2$ ~ $10^3$ atrybutów w systemach diagnozy medycznej

# Motywacje



- „Jesteśmy zatopieni w morzu danych, podczas gdy pragniemy wiedzę”
- PROBLEM: jak wydobyć użyteczne informacje/wiedzy z dużego zbioru danych?
- Rozwiązanie: hurtownia danych + data mining
  - Zbieranie danych (w czasie rzeczywistym)
  - Odkrywanie interesesującej wiedzy (reguł, regularności, wzorców, modeli ...) z dużych zbiorów danych

# Ewolucja w technologii baz danych



- W latach 60-tych:
  - Kolekcja danych, tworzenia baz danych, IMS oraz sieciowe DBMS
- W latach 70-tych:
  - Relacyjny model danych, implementacja relacyjnych DBMS
- W latach 80-tych:
  - RDBMS, zaawansowane modele danych (extended-relational, OO, deductive, ...) oraz aplikacyjno-zorientowane DBMS
- Od 90-tych —obecnie:
  - Data mining, hurtownia danych, multimedialne bazy danych oraz „Web databases”

# (c.d)

## DBMS History

- Late 60's: network (CODASYL) & hierarchical (IMS) DBMS.
  Low-level ``record-at-a-time'' DML, i.e. physical data structures reflected in DML (no data independence)
- 1970: Codd's paper. The most influential paper in DB research.
  Set-at-a-time DML. Data independence. Allows for schema and physical storage structures to change under the covers". Truly important theory, led to "paradigm shift" in thinking and in practice.
  Papadimitriou: "as clear a paradigm shift as we can hope to find in computer science".
  **Turing award**

## DBMS History

- early-to-mid-70's
  - raging debate between the two camps.
  - "great debate" in 1975
- mid 70's
  - 2 full-function (sort of) prototypes
    - Ingres
    - System R
  - Ancestors of essentially all today's commercial systems

## DBMS History

- early 80's
  - commercialization of relational systems
- mid 80's
  - SQL becomes “intergalactic standard”.
    DB2 becomes IBM's flagship product.
    IMS ``sunseted''

## DBMS History

- Today: network & hierarchical essentially dead (though commonly in use!)
  - relational is mainstream, not even sexy
  - SQL (& perhaps RDBMS) too flawed to last in current form.
    semantically flawed in various ways (Date, 1985).
    in an effort to fix it up, standards committees are making a mess
    - design by committee leads to kitchen sink
    - standards body as designers, rather than codifiers
    - leads to wasting time (Sybase) or irrelevance of standard (Informix & IBM shipping SQL3 before standardized)

    various players in research, industry and both scrambling to standardize the "next thing"

# Zastosowanie Data Mining



- Analiza danych i wspomaganie decyzji
  - Analiza i zarządzanie rynkiem
    - marketing, zarządzanie relacjami z klientem, analiza koszyku w transakcjach, segmentacja rynku, ...
  - Analiza i zarządzanie ryzykiem
    - Przewidywanie, zatrzymywanie klientów, kontrola jakości, analiza konkrencji, ulepszenie ubezpieczenie
  - Detekcja oszustw
- Inne zastosowania
  - Text mining (grupa dyskucyjna, poczta, dokumenty) i analiza danych sieciowych (Web mining).
  - Inteligentny system wyszukiwania informacji

# Analiza i zarządzanie rynkiem



- Źródło danych do analizy?
  - Tranzakcje z kart kredytowych, karty stałego klienta, kupony rabatowe, skargi klientów, dane demograficzne
- Docelowe marketing
  - Znaleźć grupy (modele) klientów, którzy chrakteryzują się podobnymi cechami: interest, dochód, sposób spędzania wolnego czasu, ...
- Określenie wzorce czasowe dotyczące zakupu klientów:
  - Np. Propozycja łączenie kont dla małżeńtw itp.
- Krzyżowa analiza rynku
  - Asocjacja (korelacja) między przedażami produktów
  - Predykcja w oparciu o asocjacyjnej informacji

# Analiza i zarządzanie rynkiem(2)



- Profil klienta
  - Klienci jakiego typu będzie kupił dany produkt (clustering lub klasyfikacja)
- Identyfikacja potrzeb klientów
  - Identyfikowanie produktów dla różnych klientów
  - Szukanie czynników, które są atrakcyjne dla nowych klientów
- Informacje podsumujące

# Detekcja oszustw i zarządzanie



- Zastosowanie
  - Szeroko używane w systemach ubezpieczeń zdrowotnych i emerytalnych, w serwisach kart kredytowych, telekomunikacji, ...
- Metody
  - Wykorzystanie danych historycznych do modelowania schematów zachowań oszukańczych, data mining pomaga w wykrywaniu grup podobnych zachowań
- Przykłady:
  - Ubezpieczenie samochodowe: detekcja grup ludzi, którzy wyłudzą piedzqdze z ubezpieczenia
  - Pranie pieniędzy: detekcja podejrzanych transakcji pieniędzy (US Treasury's Financial Crimes Enforcement Network)
  - Ubezpieczenie medyczne: detekcja „profesjonalnych" pacjentów i okręgu doktorów z nimi pracujących, następnie rozszerzyć okrąg podejrzanych pacjentów

# Detekcja oszustw i zarządzanie (c.d.)



- Detekcja niewłaściwych leczeń medycznych
  - Australian Health Insurance Commission wykrył nieprawidłowość w procedurze leczenia (oszczędność 1m AUD rocznie).
- Detekcja oszustw telefonicznych
  - Model rozmów telefonicznych: numer rozmówcy, czas trwania, godzina i dzień tygodnia rozmowy. Analizuje się wzorce, które wykraczają poza normą.
  - British Telecom wykrył dyskretne grupy rozmówców, którzy często ze sobą rozmawiają (przez tel. komórkowe) i złamał wielomilionowe oszustwa.
- Emerytura
  - Analitycy oszacują, że 38% składek emerytalnych kurczy się przez nieuczciwych pracowników.

# Inne zastosowania



- Sport
  - IBM Advanced Scout analizował statystyki z meczów ligi NBA (bloki, asysty, faule) aby zwiększyć poziom gry drużyn New York Knicks oraz Miami Heat
- Astronomia
  - JPL i Palomar Observatory wykryły 22 kwasary za pomocą data mining
- Internet Web Surf-Aid
  - IBM Surf-Aid stosuje algorytmy data mining do analizy logów dostępu do zasobów na stronach komercyjnych do odkrywania preferencji klientów i ich zachowań.
  - Analizuje się efektywność marketing internetowego i ulepsza organizację tych Web site komercyjnych.

# Co to jest data mining



- Data mining = the iterative and interactive process of discovering non-trivial, implicit, previously unknown and potentially useful (interesting) information or patterns from data in large databases

# Co to jest data mining(c.d.)



- Alternatywne nazwy:
  - Czy „data mining” jest właściwą nazwą?
  - *Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.*
- Co nie jest data mining?
  - Inteligentne przetwarzanie zapytań
  - Systemy experskie,
  - experymentalne (małe) programy z ML lub statystyki

# Wzorce = regularność w danych



**DATA MINING:**

the iterative and interactive process of discovering:

- non-trivial,
- implicit,
- previously unknown and
- potentially useful

information or patterns

from data in

**LARGE** databases

**WZORCE MUSZĄ BYĆ INTERESUJĄCE**

**dla pewnej grupy osób:**

- Niebanalne:
- Zrozumiałe - np. muszą być proste
- Oryginalne (nowe, zaskakujące)
- Użyteczne: pasują do nowych danych (z zadowalającym stopniem pewności): miara pewności = ?

# Asocjacja i charakterystyki



- Przykład reguły asocjacyjnej:
  - *klienci, którzy kupują piwo, często też kupują orzeszki*
- Przykład odkrywania charakterystyk: opis pacjentów chorujących na anginę
  - *pacjenci chorujący na anginę cechują się temperaturą ciała większą niż 37.5 C, bólem gardła, osłabieniem organizmu*

# Przykład zależności w bazach danych



| wiek kierowcy | lat prawo jazdy | kolor pojazdu | poj. silnika | moc | razem szkody |
|---|---|---|---|---|---|
| 42 | 24 | biały | 1610 | 100 | 0 |
| 19 | 1 | czerwony | 650 | 24 | 2500 |
| 28 | 4 | czerwony | 1100 | 40 | 0 |
| 41 | 20 | czarny | 1800 | 130 | 0 |
| 21 | 3 | czerwony | 650 | 24 | 1300 |
| 20 | 1 | niebieski | 650 | 24 | 0 |

- kierowcy, którzy jeżdżą czerwonymi samochodami o pojemności 650 ccm, powodują wypadki drogowe
- kierowcy w wieku powyżej 40 lat jeżdżą samochodami o pojemności większej niż 1600 ccm
- kierowcy, którzy posiadają prawo jazdy dłużej niż 3 lata, nie powodują wypadków

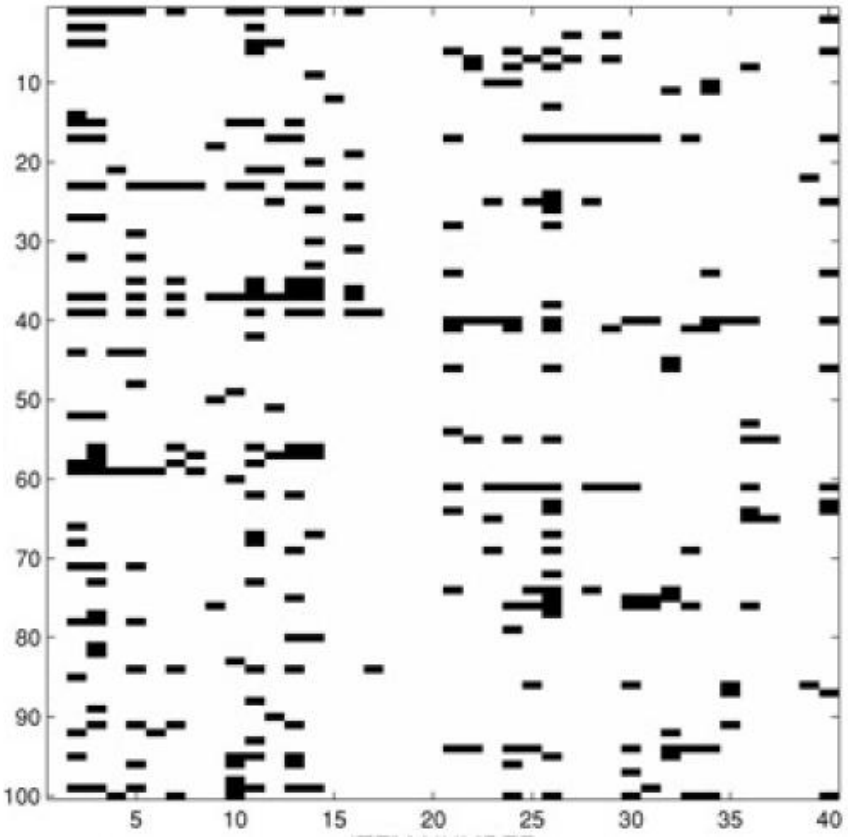
# Przykład zależności (c.d.)



| transakcja | produkt | dzień | cena |
|---|---|---|---|
| 1 | pizza | sobota | 48,40 |
| 1 | mleko | sobota | 2,80 |
| 1 | chleb | sobota | 1,50 |
| 2 | piwo | wtorek | 16,20 |
| 2 | orzeszki | wtorek | 8,50 |
| 3 | chleb | sobota | 1,50 |
| 3 | orzeszki | sobota | 25,50 |
| 3 | piwo | sobota | 32,40 |

- *piwo i orzeszki są zawsze kupowane wspólnie*
- *chleb uczestniczy w transakcjach na kwotę większą niż 50 złotych*

# Dane transakcyjne




- Itemsets
- Transakcja
- Frequent itemsets

# Nie wszystkie wzorce są interesujące!



- Miara „*atrakcyjności*”: Wzorzec jest interesujący jeśli:
  - Jest zrozumiały (wyrażalny w języku naturalnym)
  - Prawdziwy na nowych danych (do pewnego stopia)
  - Potencjalnie użyteczny
  - Oryginalny lub potwierdza pewne hipotezy, które użytkownik chciałby sprawdzić
- Obiektywne i subiektywne miary:
  - Obiektywność: oparta o statystyki i struktury wzorców, e.g., wsparcie, zaufanie, ...
  - Subiektywność: oparta o wiarę użytkownika w dane, e.g., nieoczekiwalność, nowość, czynność, itp.

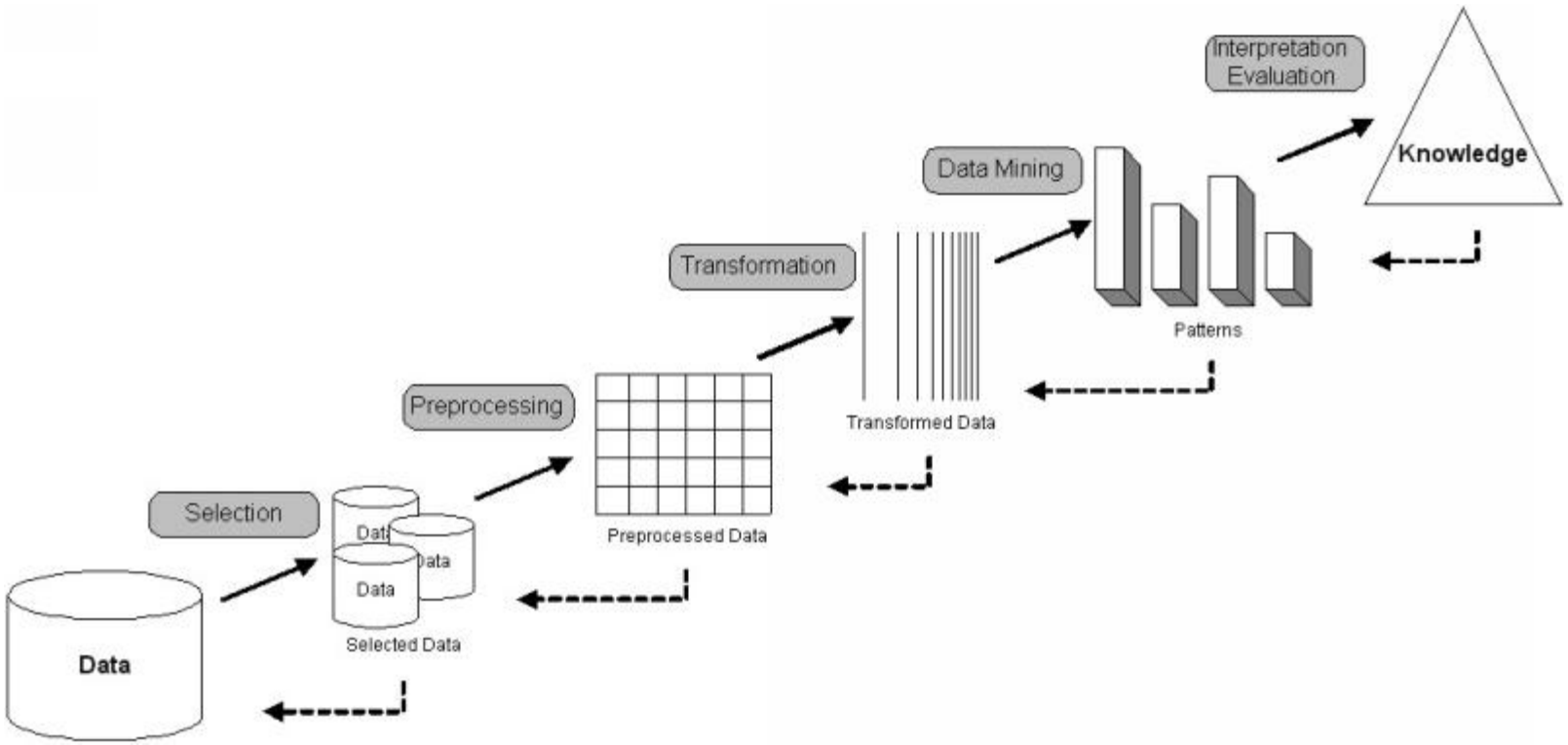
# Problem szukania wzorców?



- Szukanie wszystkich wzorców: złożoność obliczeniowa
  - Np. Asocjacja, klasyfikacja, clustering
- Szukanie tylko interesujących wzorców: problem optymalizacji.
  - Różne metody:
    - Szukanie wszystkich wzorców + filtracja.
    - Optymalizacyjny algorytm (lub heurystyka)

# Data Mining: Główny proces w KDD





# Kroki w KDD



- Zrozumienie problemu z danej dziedziny:
  - Wiedzy i główne cele zbadanej dziedziny
- Utworzenie docelową kolekcję danych: (selekcja danych)
- Czyszczenie i wstępne przetwarzanie danych: (czasem stanowi ponad 60% wysiłku!)
- Redukcja i transformacja danych:
  - Znaleźć użyteczne atrybuty, redukcja wymiarów, inna reprezentacja
- Wybór odpowiednich narzędzi data mining
  - klasyfikacja, regrecja, asocjacja, klastrowanie, ...
- Wybór algorytmów
- **Data mining**: szukanie wzorców, modeli.
- Ocena wzorców i prezentacja wyników:
  - Wizualizacja, transformacja, usuwanie zbędnych wzorców ...
- Zastosowanie odkrywanej wiedzy

# Cele w Data Mining



- **Przewidywanie (Prediction)**: To foresee the possible future situation on the basis of previous events.
  *Given sales recordings from previous years can we predict what amount of goods we need to have in stock for the forthcoming season?*
- **Opisywanie (Description)**: What is the reason that some events occur?
  *What are the reasons for the cars of one producer to sell better that equal products of other producers?*
- **(Weryfikacji hipotez) Verification**: We think that some relationship between entities occur.
  *Can we check if (and how) the thread of cancer is related to environmental conditions?*
- **(Wykrywanie wyjątków) Exception detection**: There may be situations (records) in our databases that correspond to something unusual.
  *Is it possible to identify credit card transactions that are in fact frauds?*

# Klasyfikacja systemów Data Mining



- Ze względu na funkcjowalność
  - Opisowe data mining
  - Predykcyjne data mining
  - …
- Inne klasyfikacje:
  - Typ bazy danych do analizy
  - Typ wiedzy do odkrywania
  - Typ używanych metod i technik
  - Typ dziedzin zastosowań

# Funkcjonalności Data Mining



- Opis pojęć: charakteryzacja i dyskryminacja
- Asocjacja: korelacja i przyczynowość
- Klasyfikacja i predykcja
- Clustering (analiza skupień)
- Analiza wyjątków
- Analiza trend i ewolucji
  - Regrecja, analiza sekwencji i okresowości ...
- Inne metody analizy statystycznej

# Klasyfikacja systemów Data mining z różnych punktów widzenia



- **Databases to be mined**
  - Relacyjne bazy danych
  - Hurtownia danych
  - Bazy danych transakcyjnych
  - Zaawansowane DB i magazyny informacji:
    - „Object-oriented" i „object-relational" DBMS
    - Spatial DBMS
    - Szeregi czasowe i temporalne dane
    - Multimedialne i tekstowe bazy danych
    - WWW
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

# Major Issues in Data Mining (1)

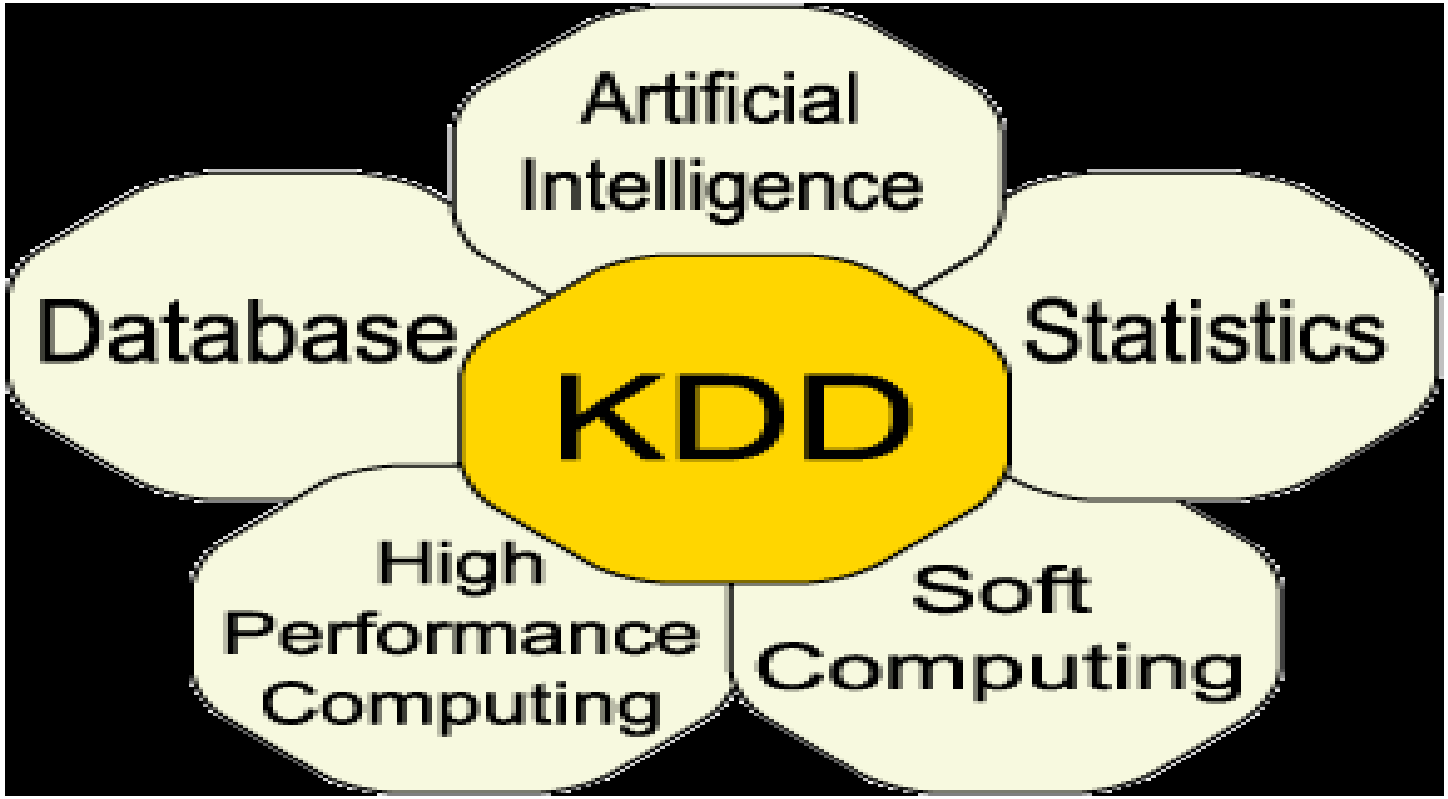


- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- Performance and scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods

# Major Issues in Data Mining (2)

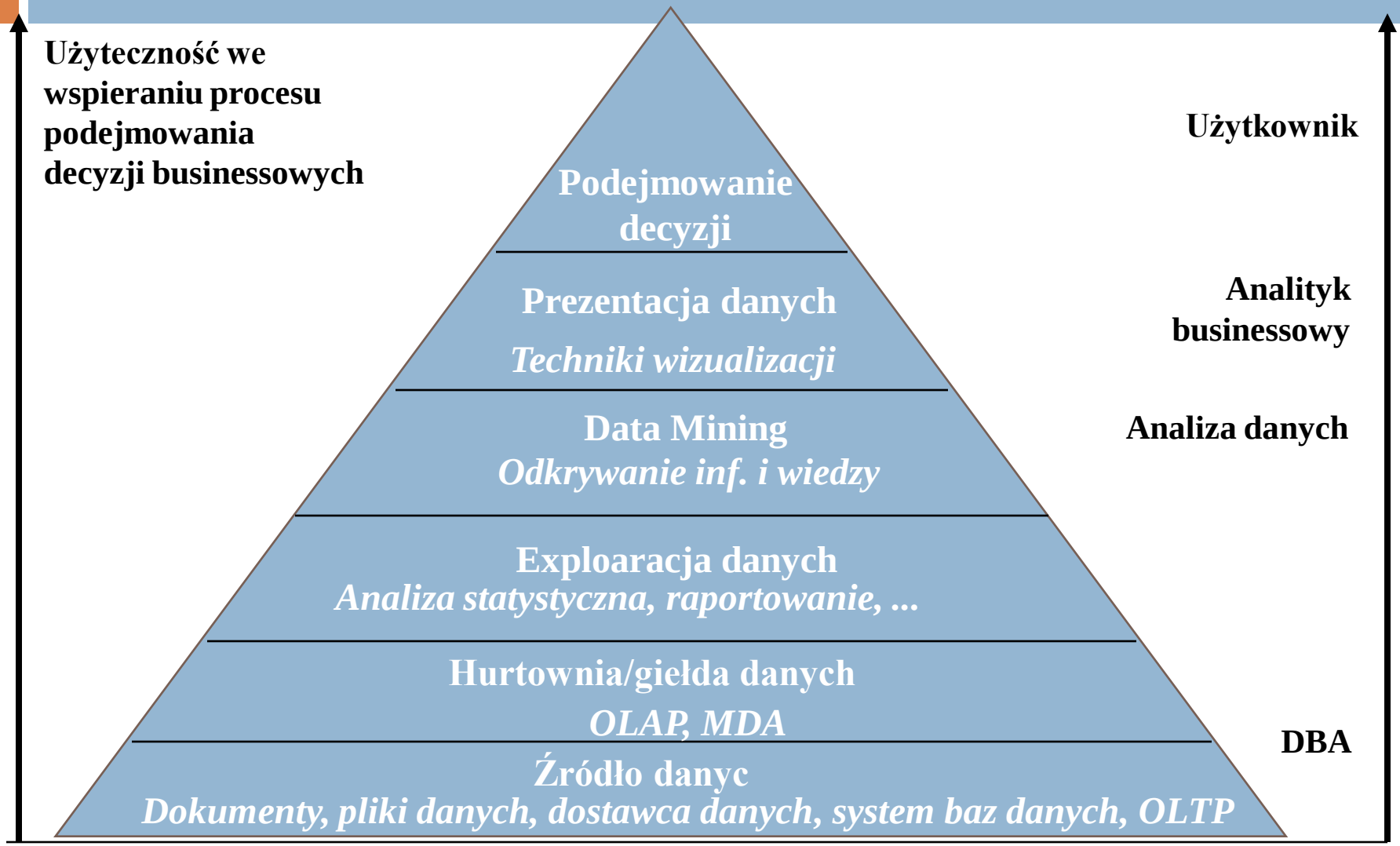


- Issues relating to the diversity of data types
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- Issues related to applications and social impacts
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy


Artificial Intelligence
Database
KDD
Statistics
High Performance Computing
Soft Computing

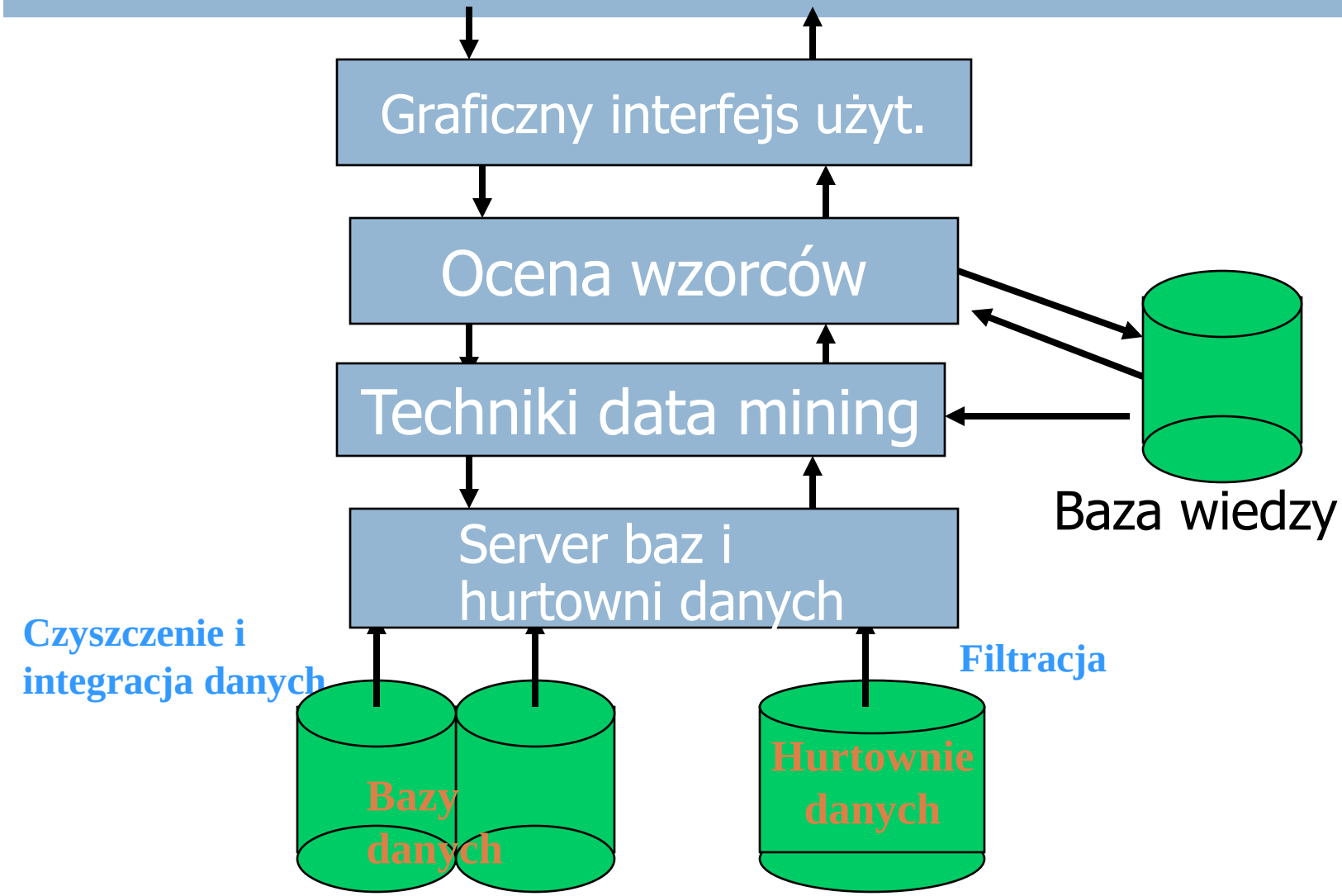
# Data Mining i Business Intelligence





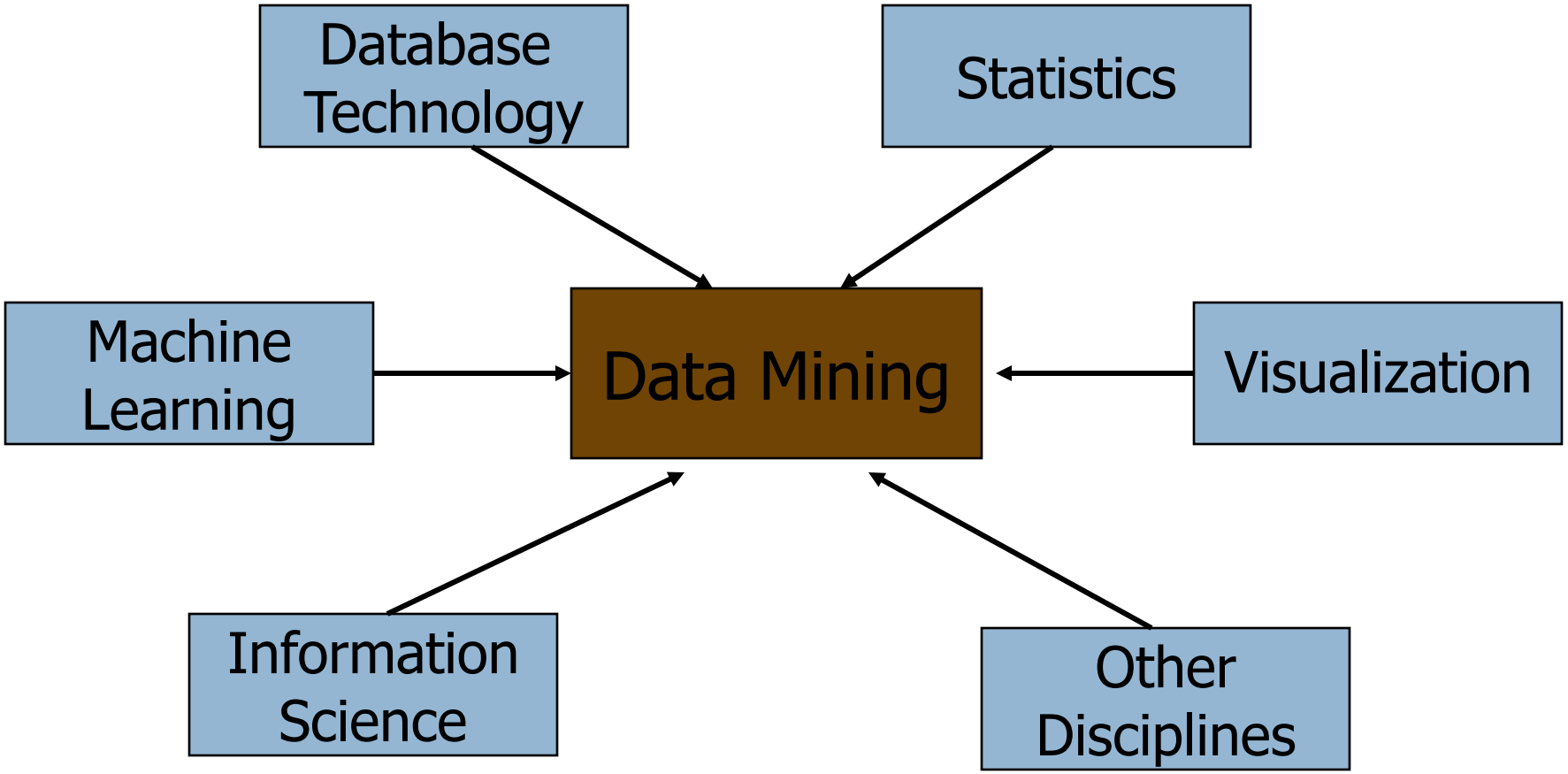
# Architektura typowych systemów Data Mining





# Data Mining: połączenie wielu dyscyplin





# Podstawowe zagadnienia w Data Mining



- Klasyfikacja
- Regresja
- Grupowania (clustering)
- Odkrywanie asocjacji
- Odkrywanie sekwencji
- Odkrywanie charakterystyk
- Wykrywanie zmian i odchyleń

# Klasyfikacja metod data mining



- Względem ich funkcjonalności:
  - Opisowe metody data mining
  - Predykcyjne metody data mining
- Różne perspektywy → różne klasyfikacje
  - Rodzaje baz danych
  - Rodzaje wiedzy do odkrycia
  - Rodzaje technik użytych
  - Rodzaje zastosowań

# Składniki algorytmu Data Mining



- Metoda reprezentacji wiedzy
- Kryteria oceniania wydobywanej wiedzy
- Strategia przeszukiwania

# Reprezentacja wiedzy



- Język (logiczny) użyty do opisywania wydobywanych wzorców
- Eksploracja danych najczęściej wykorzystuje:
  - reguły logiczne
  - drzewa decyzyjne
  - sieci neuronowe (!)

# Metody przeszukiwania



- przeszukiwanie parametrów
- przeszukiwanie modelu

- Trzeba poszukać parametrów lub modeli (z pewnej wybranej rodziny) takich, że maksymalizują kryteria optymalizacyjne

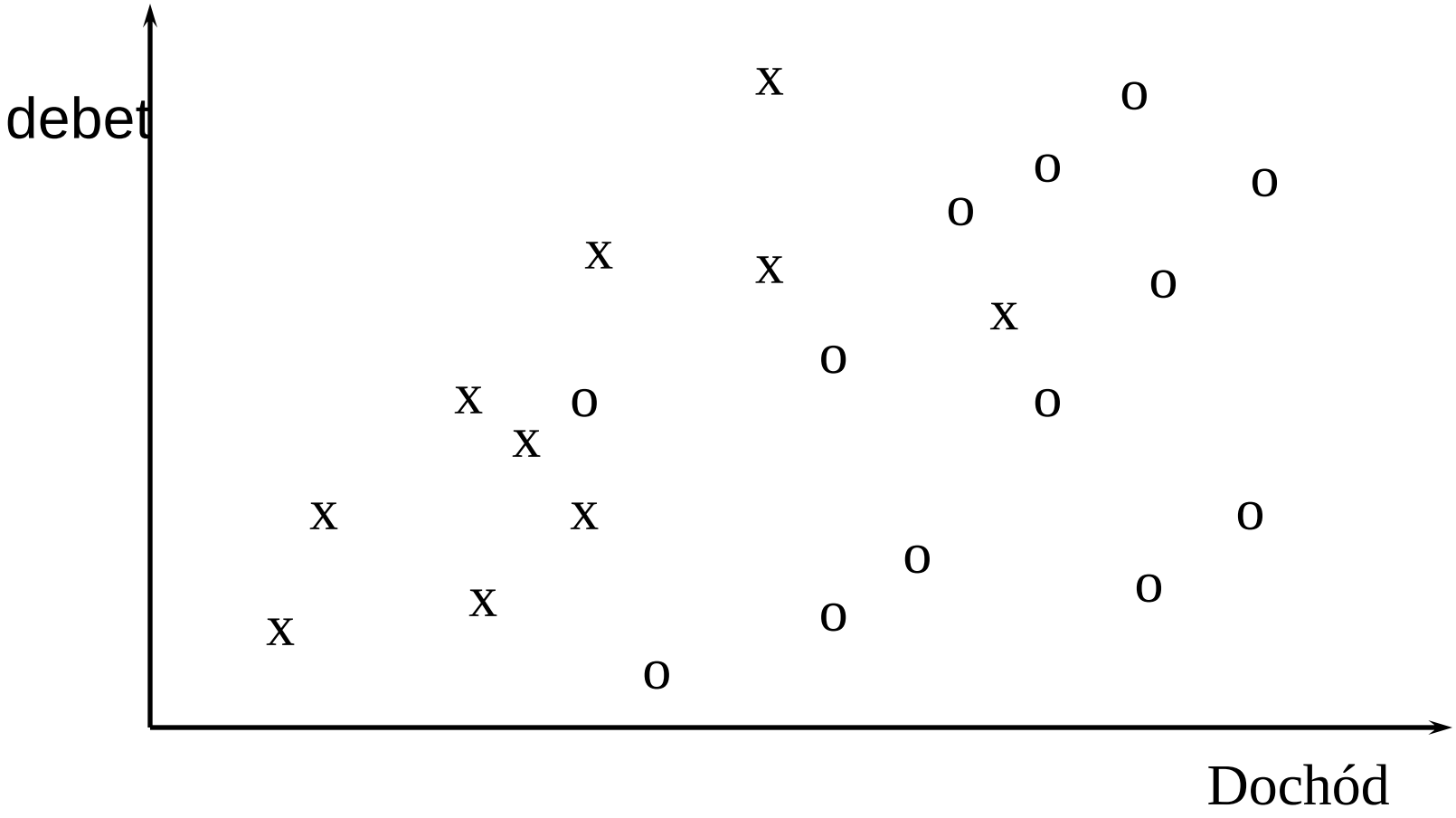
# Popularne metody w Data Mining



- Drzewa decyzyjne, reguły decyzyjne
- Reguły asocjacyjne
- Modele nieliniowe (np. sieci neuronowe)
- Metody oparte o przykładach (CBR, nearest neighbor - wymagają definicji odległości
- Modele probabilistycznej zależności (sieci Bayesowskie) - użycie struktury grafu

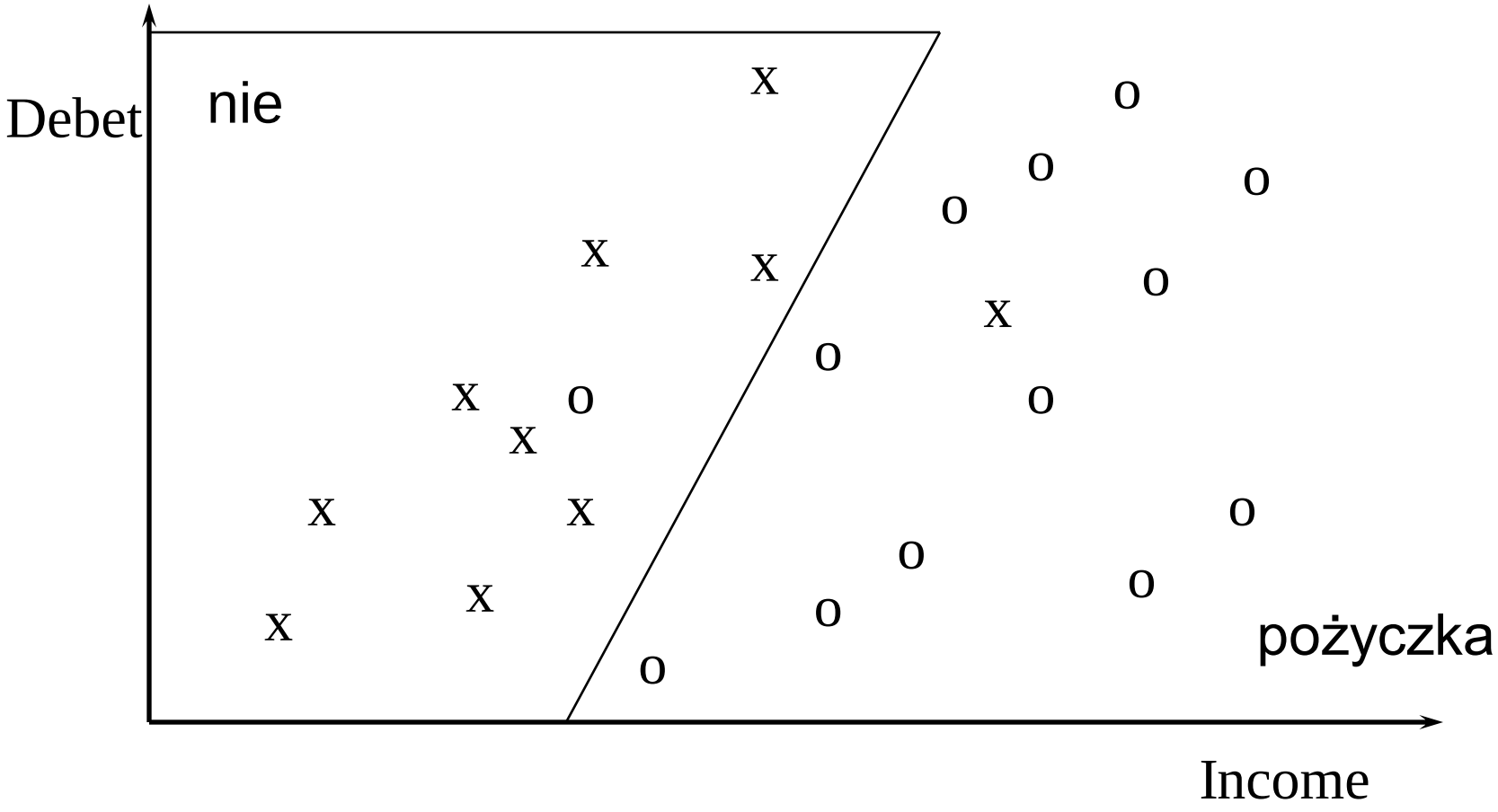
# Przykład: zbiór danych





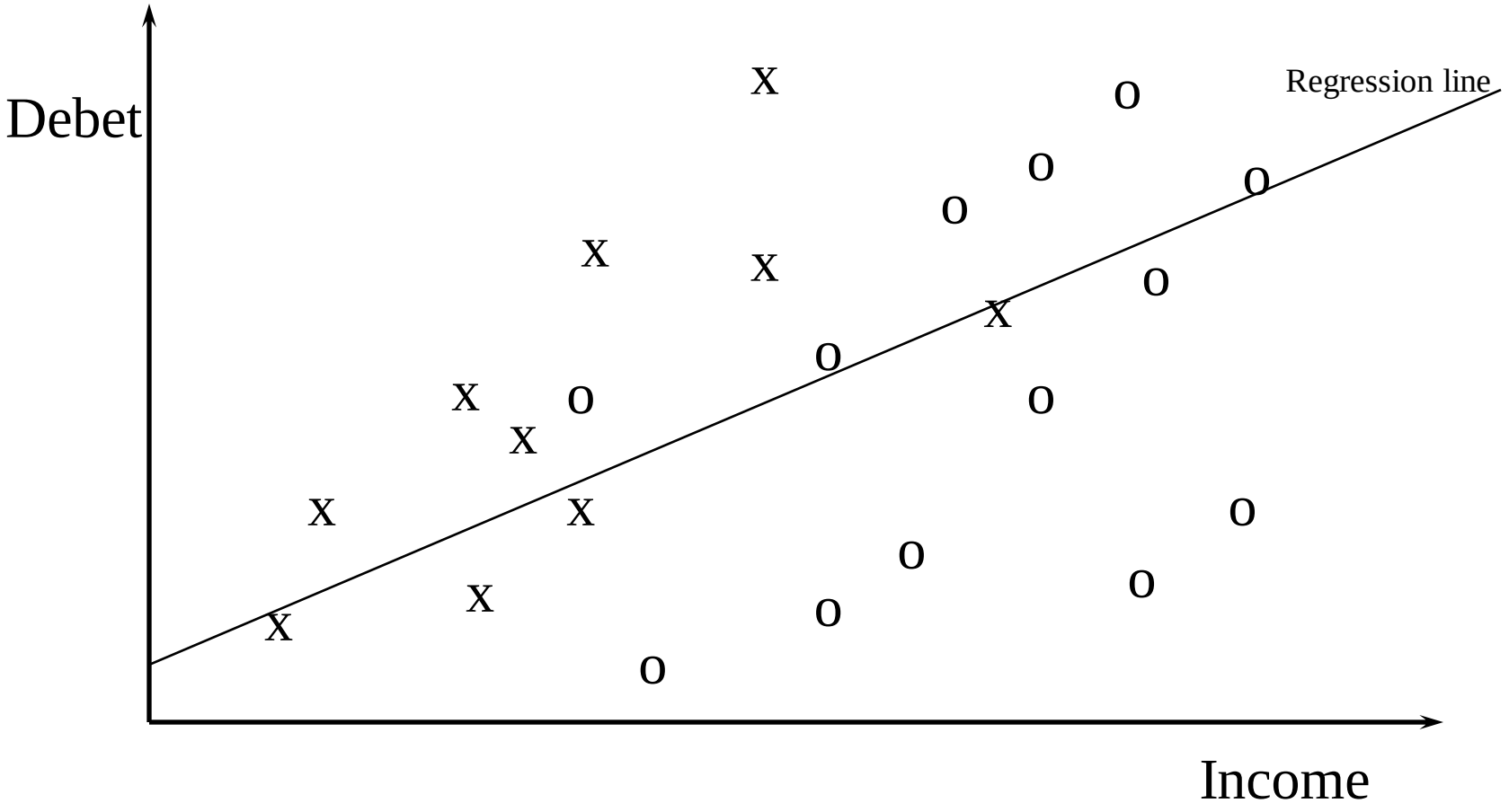
# Liniowa klasyfikacja





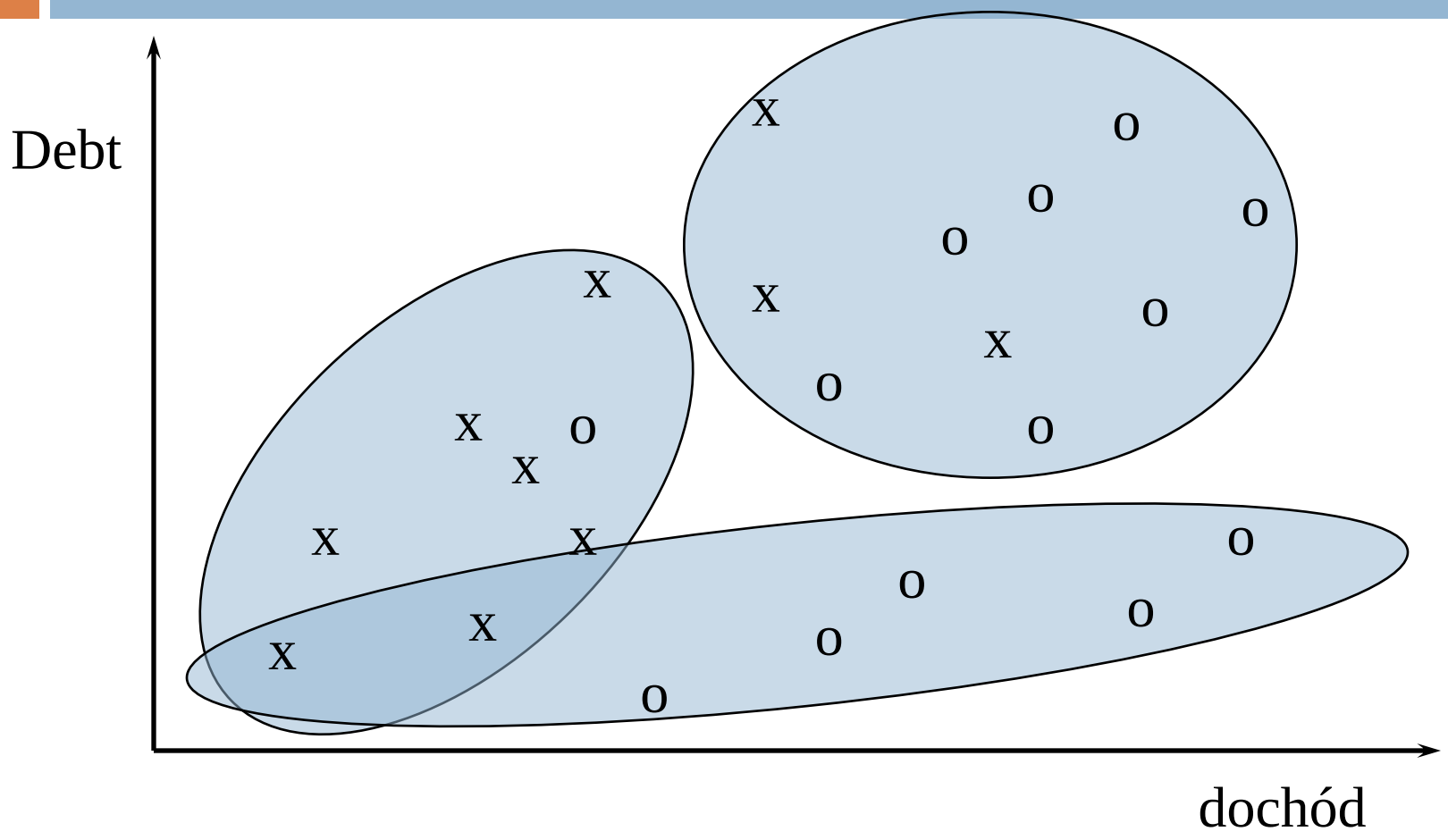
# Prosta regresja liniowa





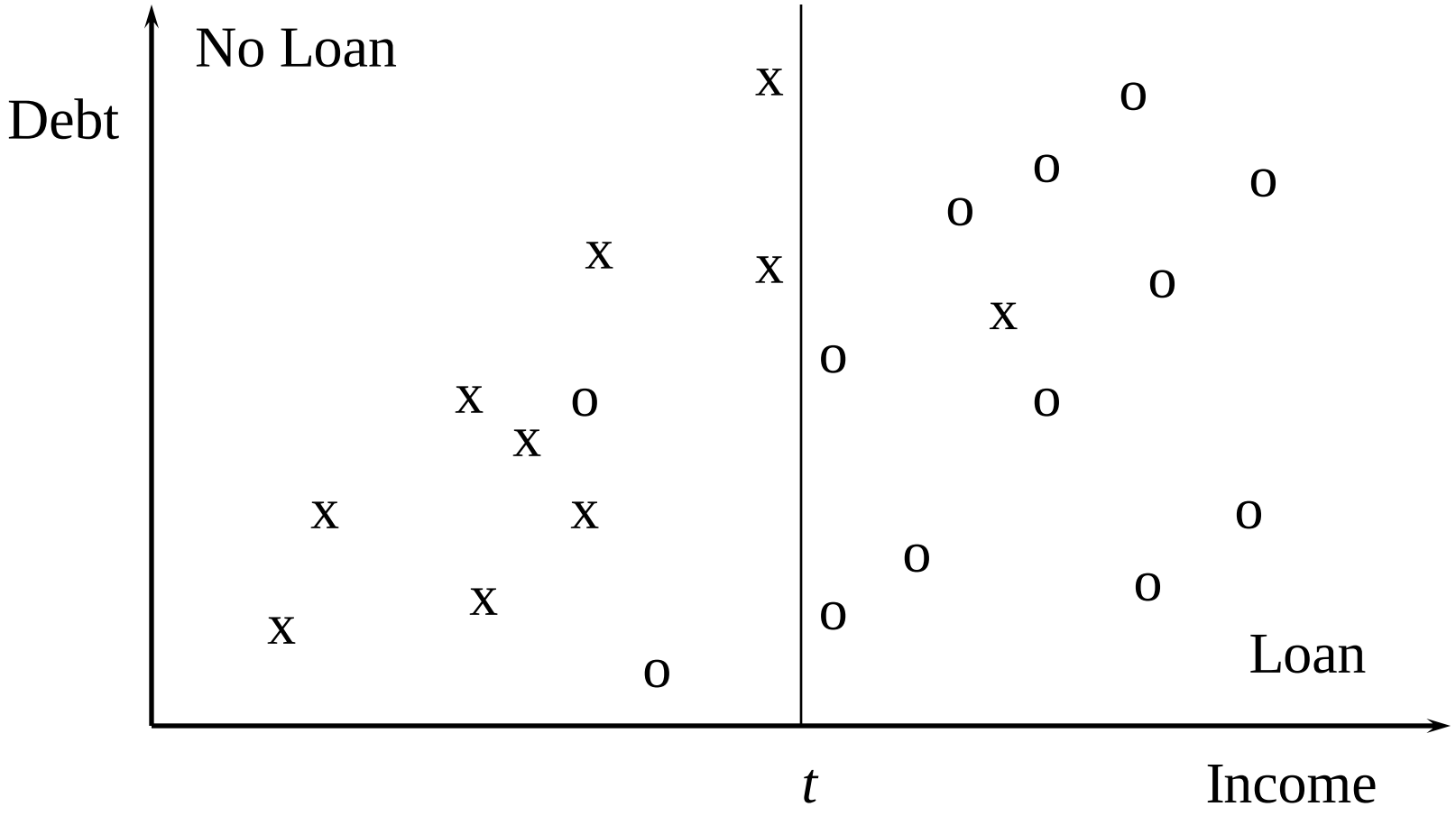
# Clustering





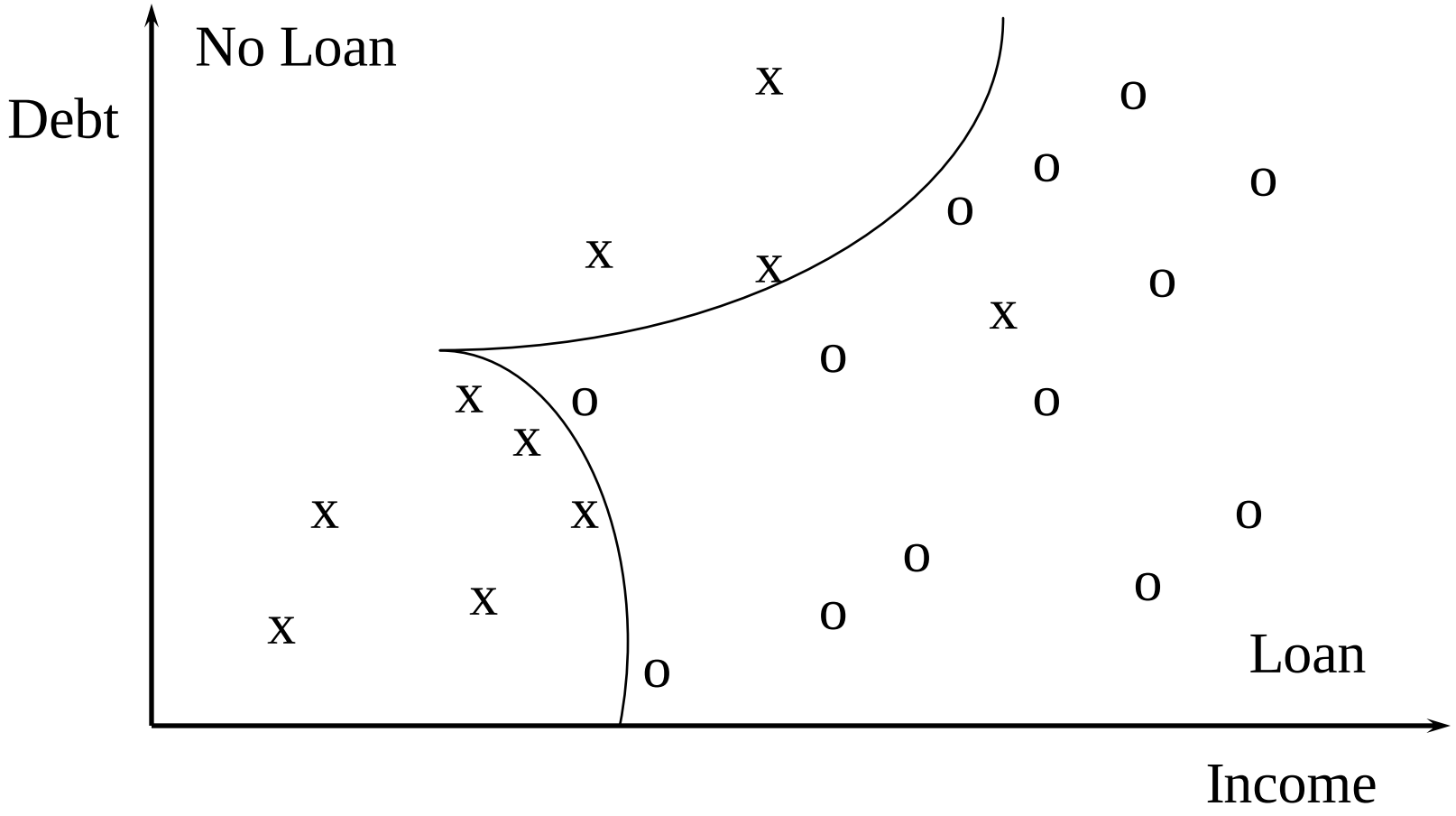
# Pojedyńczy próg (cięcie)





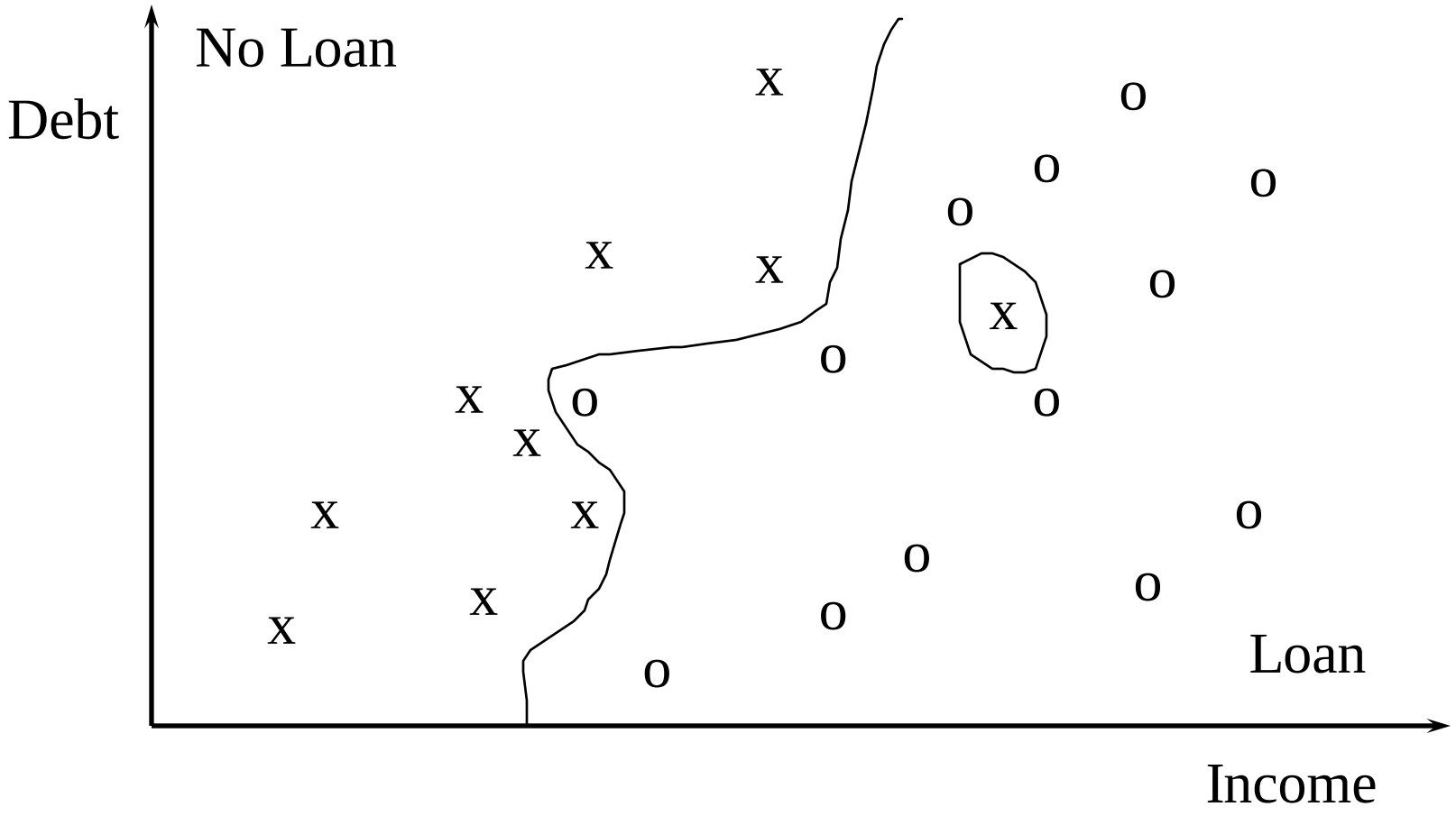
# Nieliniowy klasyfikator





# Najbliższy sąsiad





# Dziedziny zastosowań dla algorytmów



- Każda technika pasuje tylko do pewnych problemów
- Trudność polega na znalezieniu właściwego sformułowania problemu (dobre pytania)
- Nie ma jeszcze żadnego kryterium, potrzebne jest wyczucie eksperta!!!

# Np. drzewa decyzyjne



| pasują do | nie pasują do |
| --- | --- |
| przestrzeni wielo-wymiarowej | danych, w których podział jest wyznaczony przez wielomian drugiego rzędu |
| danych opisanych atrybutami różnych typów | |

# Główne problemy w Data Mining



- Metodologia i interakcja z użytkownikiem
  - Odkrywanie wiedzy różnych typów z danych
  - Interakcja podczas odkrywania na różnych poziomach abstrakcji
  - Prezentacja z wykorzystanie wiedzy dziedzinowej
  - Język zapytań (komunikacja) z systemami data mining
  - Opisywanie i wizualizacja wyników
  - Dane zaszumione i dane niekompletne
  - Ocena wzorców: problem oceniania atrakcyjność wzorca
- Osiągi i skalowalność
  - Efektywność i skalowalność algorytmów data mining
  - Metody przetwarzań równoległych, współbieżnych i  inkrementalnych

# Główne problemy w Data Mining(2)



- Różnorodność typów danych
  - Obsługa relacyjnych i złożonych typów danych
  - Odkrywania wiedzy z różnorodnych baz danych i globalnego systemu informacji (np. WWW)
- Zastosowania
  - Zastosowanie odkrywanej wiedzy:
    - Narzędzia data mining dla poszczególnych dziedzin
    - Inteligentny system zapytań
    - Sterowanie procesem i podejmowanie decyzji
  - Problem integracji odkrywanej wiedzy z istniejącą wiedzą
  - Chronienie bezpieczeństwa danych, integralność i prywatność.

# Bibliografia o KDD




- ***Data Mining: Concepts and Techniques***. J. Han and M. Kamber. Morgan Kaufmann, 2000.
- ***Knowledge Discovery in Databases***. G. Piatetsky-Shapiro and W. J. Frawley. AAAI/MIT Press, 1991.
- ***Data Mining Techniques: for Marketing, Sales and Customer Support***. M. Berry, G. Linoff (Wiley)
- ***Advances in Knowledge Discovery and Data Mining***. U.S. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, AAAI/MIT Press, 1996.
- ***Rough Sets in Knowledge Discovery I & II***. L. Polkowski, A. Skowron (Springer)

# KDD w internecie



- Konferencje i czasopisma:
  - Data mining and KDD (SIGKDD member CDROM):
    - Conference proceedings: KDD, and others, such as PKDD, PAKDD, etc.
    - Journal: Data Mining and Knowledge Discovery
  - Database field (SIGMOD member CD ROM):
    - Conference proceedings: ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, DASFAA
    - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIIS, etc.
  - AI and Machine Learning:
    - Conference proceedings: Machine learning, AAAI, IJCAI, etc.
    - Journals: Machine Learning, Artificial Intelligence, etc.
  - Statistics:
    - Conference proceedings: Joint Stat. Meeting, etc.
    - Journals: Annals of statistics, etc.
  - Visualization:
    - Conference proceedings: CHI, etc.
    - Journals: IEEE Trans. visualization and computer graphics, etc.
- System WEKA: www.cs.waikato.ac.nz/ml/weka
- Knowledge Discovery Nuggtes: http://www.kdnuggets.com
- Dr K. Thearling http://www.thearling.com
- The Data Mine http://cs.bham.ac.uk/~anp/TheDataMine.html