

---

# Metody odkrywania wiedzy i maszynowego uczenia w eksploracji danych

---

Marcin S. Szczuka

---

# Wykład 1 - Wprowadzenie

- Eksploracja danych i rola maszynowego uczenia jako narzędzia odkrywania wiedzy.
  - Odkrywanie wiedzy i systemy uczące się.
  - Plan kursu, sprawy techniczne.
-

---

# KDD

Knowledge Discovery in Databases

Odkrywanie wiedzy z danych

&

---

Data Mining

# DM

Eksploracja danych

---

# Główne zagadnienia

- Czym jest KDD & DM?
  - Dlaczego KDD&DM jest potrzebne/przydatne?
  - Proces eksploracji danych.
  - Przegląd technik KDD.
-

---

# Czym jest KDD&DM?

Wieloprzebiegowy i interaktywny proces odkrywania nowych, wartościowych, przydatnych, ogólnych i zrozumiałych wzorców i modeli z

# Wielkich

źródeł danych (baz danych).

---

# Wielkie zbiory danych

- Wielka liczba przypadków  
10<sup>6</sup>-10<sup>9</sup> dla bazy obiektów kosmicznych (astronomia)  
10<sup>6</sup>-10<sup>7</sup> – klienci kompanii telekomunikacyjnej
- Wielka liczba atrybutów (cech, pomiarów, kolumn)  
Setki zmiennych w kartach pacjentów w szpitalu  
Tysiące rodzajów towaru w ofercie dużej firmy

---

# Czym jest KDD&DM

- Nowe: coś o czym nie wiedzieliśmy
  - Wartościowe: rozciąga się na przyszłość
  - Przydatne: możliwa jest reakcja
  - Zrozumiałe: prowadzi do głębszej wiedzy
  - Wieloprzebiegowy (iteracyjny): wiele kroków i wiele powtórzeń
  - Interakcyjny: człowiek jest częścią systemu
-

---

# Cele eksploracji danych

- Przewidywanie
  - Opisywanie
  - Weryfikacja
  - Wykrywanie wyjątków
-



---

# Przewidywanie

- Chcemy przewidzieć rozwój sytuacji w przyszłości na podstawie dotychczasowych przypadków.

*Czy dysponując zapisami sprzedaży z lat poprzednich możemy przewidzieć jakie zapasy magazynowe musimy przygotować na nadchodzący sezon?*

---

---

# Opisywanie

- Dlaczego występują pewne zjawiska?

*Jakie są powody dla których samochody jednego wytwórcy sprzedają się lepiej od bardzo zbliżonych modeli innych producentów?*

---

---

# Weryfikacja

- Wydaje nam się, że występują pewne związki.

*Chcemy sprawdzić czy (i jak) zagrożenie nowotworem zależy od środowiska pracy.*

---

---

# Wykrywanie wyjątków

- W naszej bazie danych mogą się pojawiać zapisy odpowiadające sytuacjom nietypowym.

*Czy jest możliwe zidentyfikowanie tych operacji na kartach kredytowych, które są w rzeczywistości oszustwami?*

---

---

# Dlaczego chcemy eksplorować?

- Mnóstwo danych jest zbierane (i przechowywane w hurtowniach danych)
- Ilość danych jest za duża dla tradycyjnych narzędzi analitycznych.
- Moce obliczeniowe są dostępne i relatywnie tanie.
- Presja konkurencji na rynku:
  - Poznaj lepiej klientów.
  - Poznaj lepiej swój rynek.
- Informacja jest surowcem, wiedza jest towarem.

*Information is a commodity, knowledge is the product.*

---

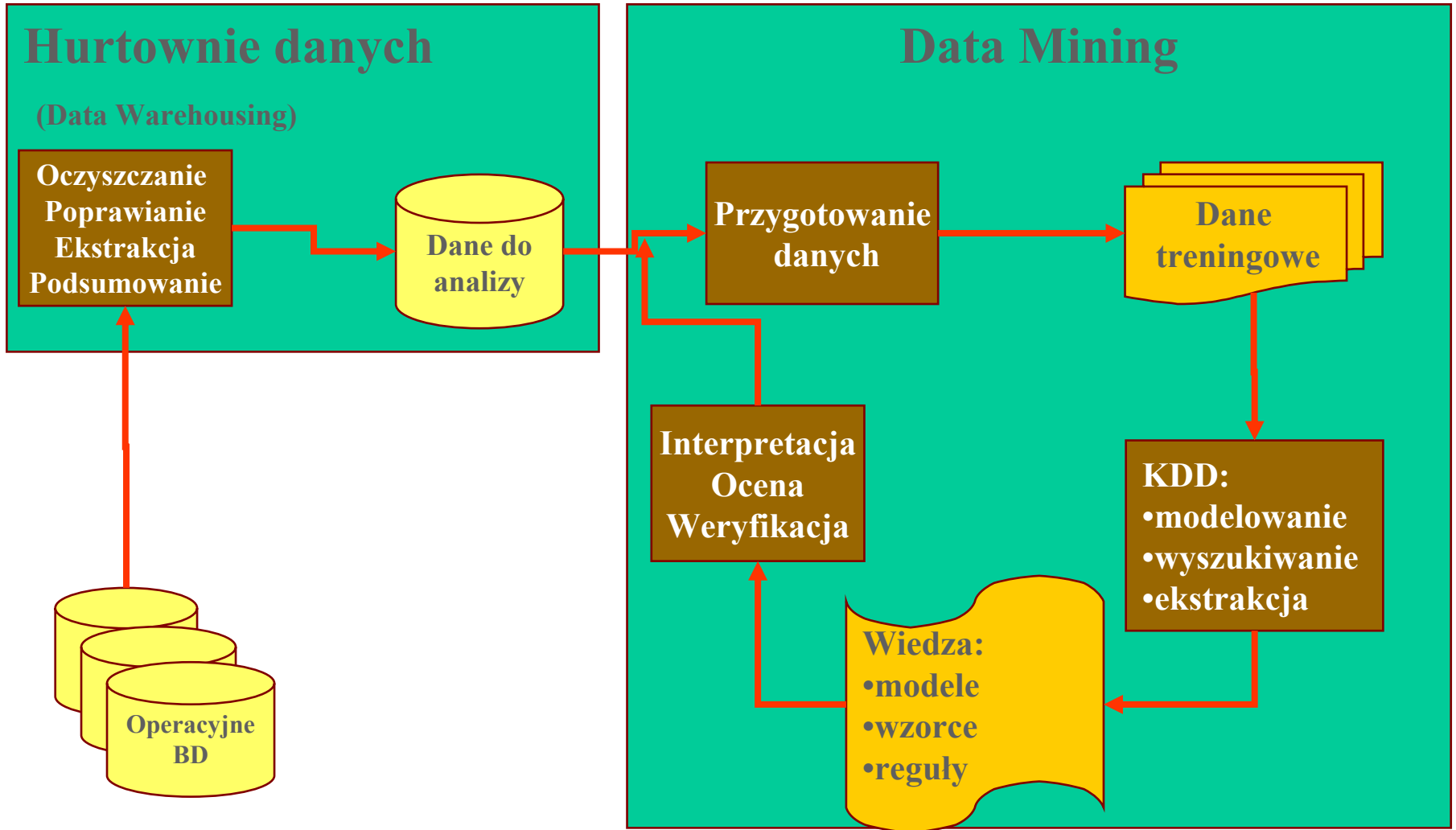
---

# Dlaczego chcemy eksplorować?

Z naukowego punktu widzenia.

- Przepastne źródła danych:
    - Czujniki na satelitach
    - Sekwencje genów
    - Symulacje komputerowe
  - Tradycyjne techniki nie wystarczają.
  - Eksploracja w celu zredukowania, uproszczenia, uogólnienia i wydobycia hipotez naukowych.
-

# Proces eksploracji



---

# Proces eksploracji

- Zrozumienie zadania
    - Posiadana wiedza, oczekiwania użytkownika, cele nadrzędne, otwarte zagadnienia.
  - Stworzenie zbioru danych do analizy
    - Wybranie danych, ocena składników wejściowych, identyfikacja poddziedzin.
  - Przygotowanie danych treningowych
    - Usunięcie szumów, identyfikacja wyjątków, uzupełnienie brakujących elementów.
    - Wybór (tworzenie, wyliczanie) cech, redukcja wymiaru.
-



---

# Proces eksploracji

- Zastosowanie algorytmu(ów) odkrywania wiedzy
    - Asocjacje, wzorce, korelacje, wyjątki, reguły, klastry, etc.
  - Interpretacja, wizualizacja i weryfikacja wiedzy
    - Co jest nowe, nieoczekiwane, nietypowe, powtarzalne?
    - Wykonywanie tylu powtórzeń ile potrzeba dla zapewnienia jakości wyników.
  - Zarządzanie wiedzą
    - Zamknięcie pętli przez włączenie wniosków do układu.
-

---

# Metody odkrywania wiedzy

- Modelowanie predykcyjne (klasyfikacja, regresja)
  - Segmentacja, rozróżnianie, grupowanie (clustering)
  - Modelowanie zależności (modele graficzne, estymacja)
  - Podsumowywanie (asocjacje)
  - Wykrywanie zmian i odchyłeń
-

# Metody c.d.

- Klasyfikacja

- Przypisywanie nowego przypadku (rekordu) do uprzednio zdefiniowanej klasy (klas).
- Silnie związane z uczeniem z nadzorem.

- Grupowanie pojęciowe, klasteryzacja, klastrowanie (clustering)

- Podział danych na podzbiory (grupy, klastry) takie, że elementy jednego podzbiory posiadają wspólne własności.
  - Silnie związane z uczeniem bez nadzoru.
-

---

# Metody c.d.

- **Asocjacje, reguły asocjacyjne**
    - Identyfikacja zestawów cech które występują razem dla wielu przypadków.
    - Wyszukiwanie powtarzających się wzorców w danych.
    - Związane z zarówno z metodami uczenia się z nadzorem, jak i bez nadzoru.
-

---

# Metody c.d.

- Wyszukiwanie podobieństw
    - Mając zbiór danych i przykład „interesującego” obiektu, konstruujemy zapytanie tak, aby wydobyć z danych zbiór rekordów, które są podobne do naszego prototypu ze względu na wcześniej określone (lub odkryte) kryteria (miarę) podobieństwa.
    - Silne związki z wnioskowaniem aproksymacyjnym i wieloma gałęziami AI.
-

---

# Metody c.d.

## ■ Poszukiwanie odchyleń

- Znajdowanie w bazie danych rekordów, które najbardziej różnią się od pozostałych. Mogą być one traktowane jako zakłócenia (i usuwane) lub jako interesujące przypadki szczególne.
  - Związane z wnioskowaniem aproksymacyjnym i statystycznym, wizualizacją i reprezentacją wiedzy.
-

# Inne pokrewne metody

- Sztuczne sieci neuronowe
- Zbiory rozmyte (Fuzzy Sets)
- Zbiory przybliżone (Rough Sets)
- Analiza szeregów czasowych
- Sieci bayesowskie
- Drzewa decyzyjne
- Programowanie ewolucyjne i algorytmy genetyczne
- Modele Markowa



# Rola “AI” w KDD

- Zarządzanie wiedzą i danymi wyrażonymi w języku naturalnym.
- Wykorzystanie inteligentnych systemów wieloagentowych.
- Metody wnioskowania w sytuacjach niepewnych.
- Techniki reprezentowania i uaktualniania wiedzy.
- Krok w stronę bardziej naturalnej (dla człowieka) reprezentacji wyników.



# Oczekiwania względem KDD

## ■ Skalowalność

- ❑ Efektywne wybieranie próbek.
- ❑ Efektywne wydobywanie danych z bazy.
- ❑ Operowanie raczej na pamięci niż na dysku.
- ❑ Wysoka efektywność obliczeniowa.
- ❑ Modularność.

## ■ Automatyzacja

- ❑ Łatwe w użyciu.
- ❑ Wykorzystuje wiedzę nabytą w poprzednich krokach.

---

# Przykłady zastosowań

- SKICAT – Analiza danych o obiektach kosmicznych.
    - 3 terabajty ( $3 \cdot 10^{12}$  bajtów) obrazów.
  - TASA - Telecom Alarm Sequence Analyser
    - Identyfikacja często pojawiających się alarmów dla strumienia danych o połączeniach.
-

---

# Przykłady zastosowań

- CASSIOPEE – system obsługi błędów
    - Wykorzystywany przez Boeinga w produkcji 737
  - Inteligentne oczyszczanie danych
    - Wykrywanie powtarzających się żądań zasiłku składanych w Welfare Department stanu Washington.
-

---

# Przykłady zastosowań

- PRISM eFraud
    - Pracujący w czasie rzeczywistym system wykrywania oszustw na operacjach on-line dokonywanych za pomocą kart kredytowych. System wykorzystujący sieci neuronowe.
  - Wiele innych (np. IRS)
-

---

# ML

## Machine Learning

---

Uczenie się maszyn

Uczenie maszynowe

Systemy uczące się

---

For it is esteemed a kind of dishonour unto learning to descend to inquiry or meditation upon matters mechanical, except they be such as may be thought secrets, rarities, and special subtilities, which humour of vain supercilious arrogancy is justly derided in Plato.... But the truth is, they be not the highest instances that give the securest information; as may well be expressed in the tale... of the philosopher, that while he gazed upwards to the stars fell into the water; for if he had looked down he might have seen the stars in the water, but looking aloft he could not see the water in the stars. So it cometh often to pass, that mean and small things discover great, better than great can discover the small.

---

Francis Bacon, *The Advancement of Learning*

---

# Czym jest ML?

Na nasze potrzeby:

*Zbiór metod i algorytmów, które poprawnie rozszerzone i połączone z innymi metodami dają nam narzędzie do zajmowania się problemami takimi jak:*

---

---

# Zadania ML (przykłady)

- Nauczyć się grać w jakąś grę (np. szachy)
  - Pomóc postawić diagnozę na podstawie zmierzonych symptomów.
  - Nauczyć się znajdować właściwą drogę w nieznanym otoczeniu.
  - Znaleźć zależność funkcyjną między dwoma obserwacjami.
-



---

# Zadania ML (przykłady)

- Zaklasyfikować strony WWW do uprzednio wskazanych kategorii.
  - Przybliżyć nieznaną funkcję na podstawie przykładów (obserwacji).
-

# Motywacja do ML

- W różnych rzeczywistych zastosowaniach jest niezwykle trudno (a czasem się nie da) znaleźć najlepsze rozwiązanie „ręcznie”.
- Złożone problemy często mają jedynie częściowe (jeśli w ogóle jakieś) modele numeryczne.
- Zbiory danych są zbyt duże by człowiek był je w stanie ogarnąć na poziomie szczegółowym.

---

# Dane osobowe

- Imię i nazwisko - Machine Learning
  - Urodzony – Nie do końca jasne, gdzieś między 1965 i 1968 w Europie i USA. Nazywany ML od początku lat 70.
  - Rodzice – jak każdy sukces, ma wielu: Michalski, Larson, Mitchell, Tecuci, Saitta, Carbonell, Quinlan, ....
-

---

# Dane osobowe c.d.

- Egzamin dojrzałości – 1983 przez publikację:

*Machine Learning: An Artificial Intelligence Approach*, R. S. Michalski, J. Carbonell, T. Mitchell (Eds.), TIOGA Publishing Co., Palo Alto

oraz 1st ICML

1986 – pierwszy numer

Machine Learning Journal , Kluwer AP

---

---

# Taksonomia ML

- Wiedza deklaratywna vs. proceduralna (wiedza vs. zdolność/możliwość)
  - Metody pozyskiwania wiedzy:
    - Przez bezpośrednie wstawienie
    - Przez obserwację i odkrywanie (bez nadzoru)
    - Z przykładów (z nadzorem)
    - W oparciu o pytania
    - Uczenie ze wzmocnieniem
-

---

# Taksonomia ML

- Metody reprezentowania wiedzy:
    - reguły
    - drzewa decyzyjne
    - klauzule logiki zdaniowej
    - rozkłady częstości i prawdopodobieństwa
    - modele parametryczne
    - funkcje przejścia w automatach skończonych
    - ....
-

---

# Nauka jako wyszukiwanie

Dysponując przestrzenią możliwych hipotez (rozwiązań) znajdź, w sposób efektywny, najlepszą z nich z uwzględnieniem zadanych kryteriów. Jedno z pierwszych sformułowań zadań ML. Motywacja dla badania związków z technikami optymalizacji.

---

---

# Nota bene

Większość technik, które zostaną przedstawione w trakcie tego wykładu jest oparta na ogólnej koncepcji **wnioskowania indukcyjnego**

---



# Uczenie się a statystyka

- Statystyka – oryginalnie zajmowała się testowaniem hipotez, estymacją błędu itp. W ostatnich czasach to podejście bardzo się zmienia.
- ML – zajmuje się tworzeniem hipotez z wykorzystaniem dedykowanego „języka”.

Statystyka dostarcza wielu bardzo dobrych narzędzi, ale należy ich używać mądrze i z umiarem.

---

# Sprawy techniczne

- Literatura
  - Profil wykładów
  - Zaliczenia etc.
-

---

# Literatura

- Cichosz P., Systemy uczące się, WNT, Warszawa, 2000
  - Mitchell T.M., Machine Learning, McGraw-Hill, 1997,
  - Berry M.J.A, Linoff G. Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management (wydanie 2), Wiley Computer Publishing, 2004
  - Witten I., Frank E., Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999
-

# Plan wykładów (luźny)

- Zagadnienia, które chcę przedstawić:
  - Dane, pierwszy kontakt
  - Uczenie się maszynowe, podstawowe pojęcia
  - Tworzenie drzew decyzyjnych
  - Tworzenie i używanie reguł decyzyjnych
  - Metody bayesowskie i wnioskowanie probabilistyczne.
  - Grupowanie pojęciowe, klastrowanie
  - Reguły asocjacyjne i tematy pokrewne
  - Wnioskowanie oparte o przykłady

---

# Terminy, kontakt, zaliczenia

- Wykład – co poniedziałek, 9:40, aula C
- Konsultacje – czwartki 10-12,  
Wydział MIM UW, Banacha 2, pok. 1240,  
tel. 5544124, [szczuka@mimuw.edu.pl](mailto:szczuka@mimuw.edu.pl)

- Materiały:

<http://www.mimuw.edu.pl/~szczuka/mme/>

- Zaliczenia:
    - Zaliczenie na stopień
    - Punkty za frekwencję – max. 20
    - Dla wybrańców – zaliczenie przez prezentację
    - Pisemny sprawdzian na końcu zajęć (nie egzamin)
-