

Inteligentne systemy informacyjne

Moduł 9

Mieczysław Muraszkiewicz

www.icie.com.pl/lect_pw.htm

Eksploracja danych

szkic

Moduł 9

Tło

Opinie

“The purpose of computing is insight, not numbers.”

Richard Hamming



1916 - 1998

“Knowledge discovery is becoming the most desirable end-product of computing, and that the importance of knowledge acquisition from the available information is second only to endeavors that help protect and preserve our natural environment”

Gio Wiederhold



Komentarz

**Choć dysponujemy
informacjami, to wciąż
brakuje nam ...
wiedzy.**

Terminologia

Eksploracja danych

Ekstrakcja danych

Wydobywanie danych

Archeologia danych

...

Data mining

Definicja

Definicja ED

Tutaj przez eksplorację danych rozumiemy proces automatycznego odkrywania znaczącej, pożytecznej, dotychczas nieznanej i możliwie pełnej wiedzy zawartej w dużych bazach danych, wiedzy ujawniającej ukryte własności badanego przedmiotu.

Wiedza ta przyjmuje postać reguł, prawidłowości, tendencji i korelacji, i jest następnie przedstawiana przygotowanemu do jej spożytkowania użytkownikowi w celu rozwiązania stojących przed nią/nim problemów i podjęcia istotnych decyzji.

Mniej poważna definicja ED



**“Eksploracja
danych polega na
torturowaniu danych
tak długo, aż zaczną
zeczynać”**

Dlaczego ED ?

Odkrytą wiedzę można wykorzystać m.in. do

- **lepszego rozumienia świata, w którym żyjemy.**
- **usprawnienia procesów produkcyjnych, zarządzania, obsługi klientów, marketingu, zmniejszania nadużyć, ograniczenia migracji klientów do konkurentów. A więc łącznie do — zwiększenia przewagi konkurencyjnej.**

Przykłady

Przykład 1

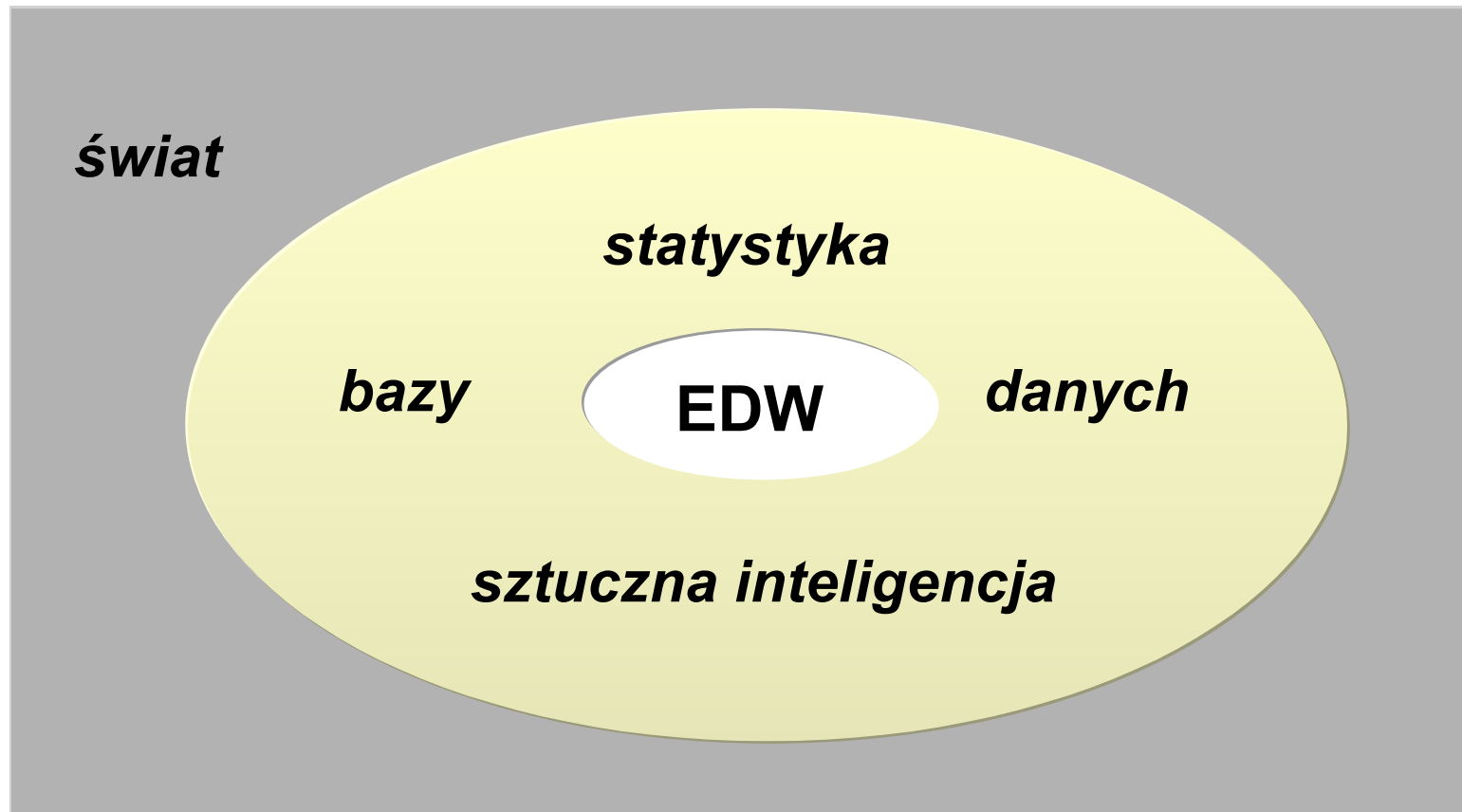
Firma American Express podała, że wykorzystanie technik eksploracji na bazie danych klientów pozwoliło zwiększyć o 10 – 15 % użycie jej kart kredytowych.

Przykład 2

Bardzo duża firma handlowa dzięki ekstrakcji potrafiła określić 5-cio procentowy segment tych klientów, którzy charakteryzują się tym, że regularnie udzielają odpowiedzi na różne zapytania firmy. Klienci ci dostarczali 60 % wszystkich odpowiedzi. Dzięki ustaleniu tego faktu firma zwiększyła 12-krotnie stopę odpowiedzi i zmniejszyła koszty opłat pocztowych o 95 %.

Kontekst

Relacja z „innymi”



Odkrywanie wiedzy (KDD)

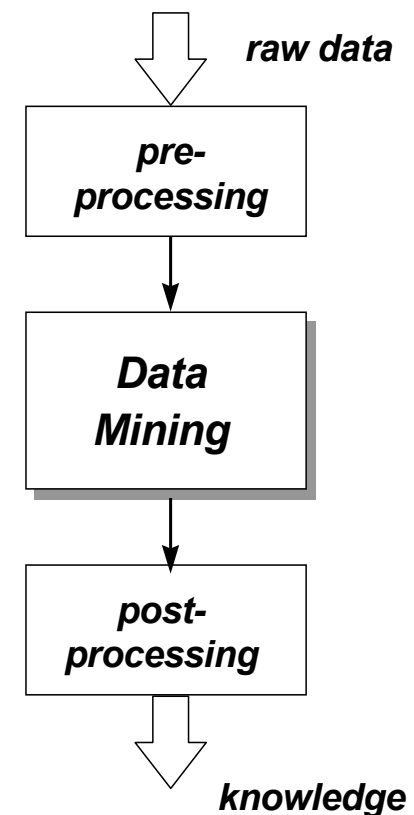
KDD is a multi-step process aimed at identifying valid, novel, potentially useful, and ultimately understandable patterns of data. (Fayyad, et al 1996)

(i) **pre-processing** that includes such operations as data preparation, data selection, and data cleaning;

(ii) **data mining**;

(iii) **post-processing** that comprises, *inter alia*, filtering and evaluation of the data mining results and their proper interpretation.

Knowledge Discovery



Czym ED nie jest ?

- procesem nieodzownie związanym z hurtowniami danych,
- typowym narzędziem analitycznym i środkiem do tworzenia sprawozdań,
- całkowicie zautomatyzowanym procesem,
- łatwym, tanim i szybkim do wdrożenia w organizacji procesem,
- przysłowiowym, wielozadaniowym scyzorykiem armii szwajcarskiej dobrym na wszelkie okazje,
- ...

Techniki eksploracji

Ważniejsze techniki

Najczęściej eksploracja oparta jest na następujących typach działań:

- **klasyfikowanie** *(ang. classification)*
- **regresja** *(ang. regression)*
- **grupowanie** *(ang. clustering)*
- **kojarzenie** *(ang. association)*
- **reguły epizodyczne** *(ang. episode rules)*
- **wizualizacja**

Klasyfikacja

Klasyfikacja jest procesem uczenia się, którego celem jest określenie reguły, która – kiedy już została zaakceptowana – służy do przyporządkowania (zaklasyfikowania) danego pod uwagę elementu do jednej lub więcej wcześniej zdefiniowanych klas (zbiorów).

Proces ten korzysta ze zbioru wcześniej poklasyfikowanych przykładów, po to aby określić sposób (model) klasyfikowania całej dostępnej populacji elementów.

Grupowanie

Grupowanie (klasteryzacja) polega na przyporządkowaniu branego pod uwagę elementu do jednej lub wielu grup (klas, zbiorów), przy czym grupy te są wyznaczana przez sam proces grupowania na podstawie analizy danych o wszystkich dostępnych elementach.

Kojarzenie

Kojarzenie polega na odszukiwaniu tych elementów, które wiążą się z zadaniem zdarzeniem lub innym elementem. Algorytmy tu wykorzystywane pozwalają odkrywać reguły typu *jeśli - to*.

Przykład

jeśli : klient kupuje płatki owsiane,
to : w 65 % przypadków klient ten kupi mleko “Łaciate”

Przykładowe zadania

- **Jak rozpoznawać i klasyfikować problemy techniczne (anomalia, awarie), także problemy chronicznie powtarzające się, oraz ujawniać przyczyny anomalii ?**
- **Jak rozpoznawać i klasyfikować alarmy generowane przez sieć ?**

Przykładowe zadania – cd.

- **Jakie są wzorce zachowań użytkowników i jak rozpoznawać połączenia stanowiące nadużycie w stosunku do operatora sieci ?**
- **Jaki jest profil użytkownika i motywacja, które mogą skłonić go do zmiany operatora sieci ?**
- **Jaki jest profil użytkowników, którzy płacą wysokie rachunki ?**
- **Jakiej reakcja użytkowników można się spodziewać na wprowadzenie nowych rodzajów usług czy taryf, uwzględniając różnorodność profili użytkowników ?**

Schemat ED

Schemat ogólny ED

- 1. Zdefiniować problem/zadanie i zanalizować otoczenie.**
- 2. Wybrać zbiór danych do eksploracji i atrybuty.**
- 3. Zdecydować jak przygotować dane do przetwarzania.**
Na przykład: czy wiek reprezentować jako przedział (np. 40-45 lat), czy jako liczbę (np. 40 lat).
- 4. Wybrać algorytm (lub ich kombinację) eksploracji i wykonać program realizujący ten algorytm.**
- 5. Zanalizować wyniki wykonania programu i wybrać te, które uznajemy za rezultat pracy.**
- 6. Przedłożyć wyniki kierownictwu organizacji i zasugerować sposób ich wykorzystania.**

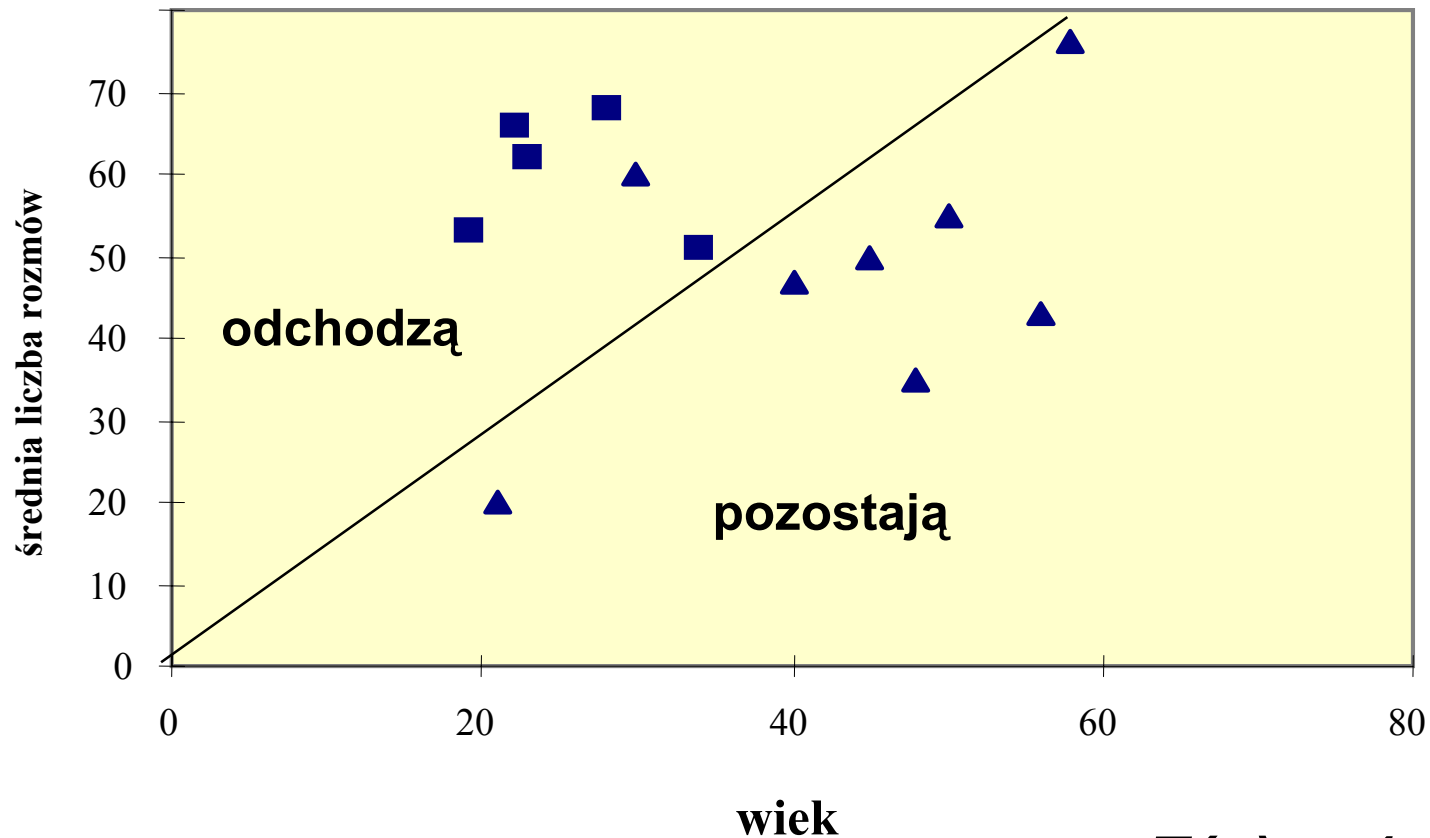
Przykład – Churning

Kierownictwo firmy zostało poinformowane, że nasila się zjawisko przechodzenia jej klientów do firmy konkurencyjnej. Zarząd podjął decyzje o zbadaniu sprawy i ustaleniu przyczyn tego zjawiska. W tym celu rozpoczęto projekt eksploracji danych, którego zadanie brzmiało:

podać charakterystykę klienta, który ma skłonność do zmiany firmy.

ID osoby	wiek	Średnia liczba różnów zamiejscowych /tydzień	Zmiana operatora
1	23	62	Tak
2	40	47	Ne
3	21	20	Ne
4	56	43	Ne
5	45	50	Ne
6	34	51	Tak
7	22	66	Tak
8	19	53	Tak
9	28	68	Tak
10	30	60	Ne
11	58	76	Ne
12	50	69	Ne
13	48	35	Ne

Przykład – cd.



kwadrat - zmienić; prostokąt - pozostał

$$F(x) = 1,3x$$

Realizacja projektów ED

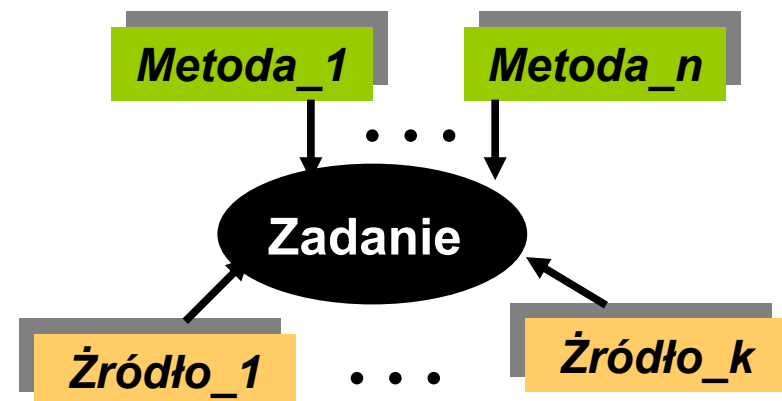
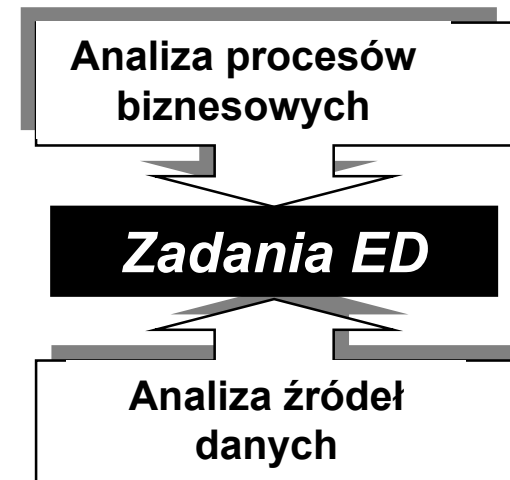
Strategia realizacji

Etap I

1. Identyfikacja procesów podatnych na ED.
2. Wybór metod i narzędzi.
3. Eksperymentalne ED.

Etap II

Realizacja platformy i aplikacje ED.



Narzędzia uniwersalne

Oracle/Darwin

**Oracle/Thinking Machines
Corporation**

Enterprise Miner

SAS

Intelligent Miner

IBM

Mine Set

Silicon Graphics

RD2

Politechnika Poznańska

**oprogramowanie
własne**

Politechnika Warszawska

Spostrzeżenia - 1

Zasadniczym warunkiem powodzenia ED jest udział zlecających prace specjalistów/ekspertów w fazach:

- definiowania zadania,**
- eksperymentów,**
- ewaluacji wyników częściowych.**

Spostrzeżenia - 2

To samo zadanie warto rozwiązywać stosując różne metody eksploracji danych (wyniki mogą być zaskakująco różne !).

Jeśli wybrano już metodę rozwiązania zadania, to należy zabiegać o możliwość prowadzenia eksperymentów na różnych zbiorach danych dotyczących tego zadania.

Spostrzeżenia - 3

Przetwarzanie wstępne i końcowe danych stanowią około 85 % czasu przeznaczzonego na rozwiązywanie zadania.

Spostrzeżenia - 4

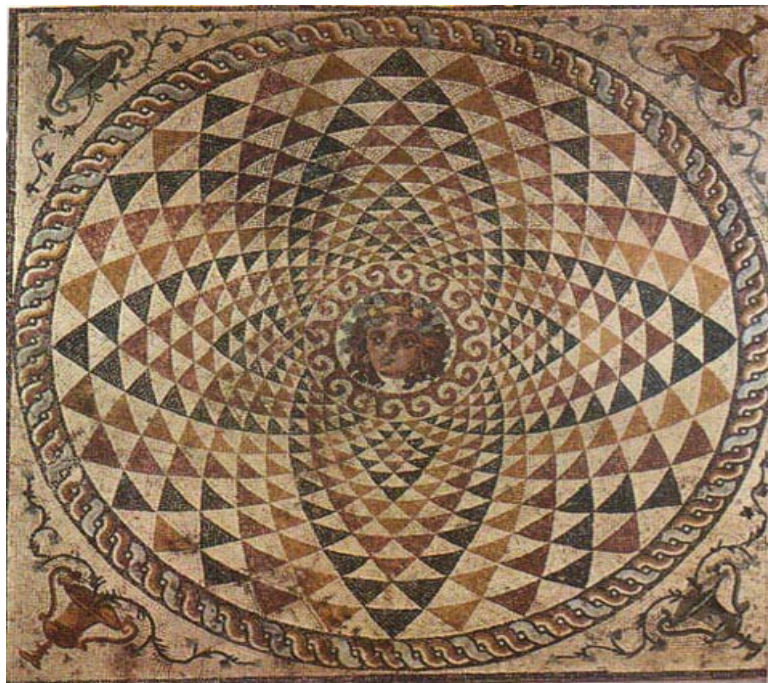
ED jest procesem złożonym, długotrwałym i kosztownym. Opiera się na zaawansowanych metodach, technikach i oprogramowaniu informatycznym. Zazwyczaj ED wymaga eksperymentowania, „dostrajania” i korzystania z kompetentnych konsultantów.

Nowe terytoria

Nowe obszary

- **Integration of DM with information retrieval languages, e.g. SQL;**
- **Standardization efforts, e.g. PMML (Predictive Modeling Markup Language); CRISP (standardized methodology for building Data Mining applications)**
- **Text/Web Data Mining**
 - **retrieval**
 - **documents classification**
 - **documents clustering**
 - **summarization**
 - **automatic indexing**
 - **language recognition**
 - **translation**
 - **...**





www.icie.com.pl/lect_pw.htm

Dziękuję za uwagę