

## Tło historyczne.

Oryginalnie termin „eksploracja danych” (Data Mining) był używany przez statystyków do określenia nadużywania danych celem wyciągania niewłaściwych wniosków.

Twierdzenie Bonferroniego orzeka, że jeżeli istnieje zbyt wiele wniosków, które możemy wyciągnąć, część z nich będzie prawdziwa z powodów czysto statystycznych.

Znany przykład: David Rhine, „parapsycholog” z Duke University przeprowadzał w latach 50-tych badania na swoich studentach pod kątem posiadania przez nich zdolności paranormalnych. W tym celu prosił ich o odgadnięcie koloru 10 wylosowanych kart (czarnych lub czerwonych). Około 1/1000 studentów odgadła wszystkie kolory prawidłowo i została okrzyknięta przez Rhine’a jako posiadająca własności paranormalne. Przy kolejnym teście okazywało się jednak, że ich zdolności odgadywania nie odbiegały od przeciętnej. Rhine wysnuł na tej podstawie wniosek, że ludzie tracą zdolności paranormalne w momencie, gdy się o nich dowiadują...

## **Potrzeba matką wynalazku.**

- Problem eksplozji danych:
  - Narzędzia automatycznego zbierania danych i technologia dojrzałych baz danych doprowadziła nas do niesłychanego wzrostu ilości danych gromadzonych w różnych miejscach.
- Toniemy w danych, ale wciąż brakuje nam wiedzy.

## Po co eksplorować dane? Komeracyjny punkt widzenia.

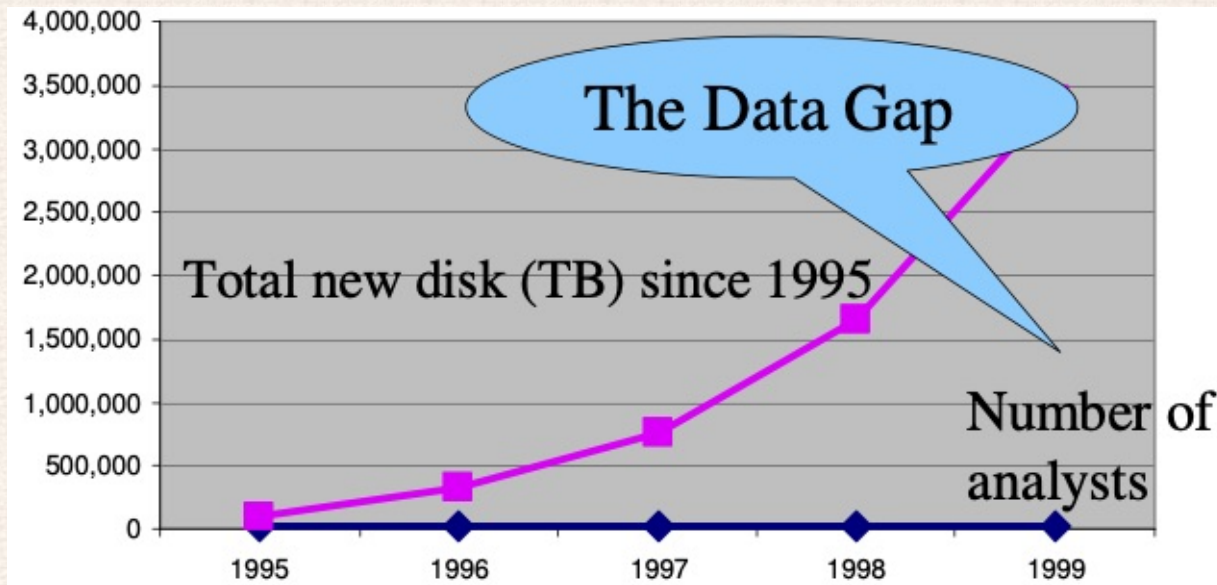
- Mnóstwo danych jest zbieranych i przechowywanych:
  - dane sieci internetowej, e-commerce,
  - zakupy w supermarketach i sieciach handlowych,
  - transakcje kartami kredytowymi.
- Komputery stały się tanie, a ich moc obliczeniowa gwałtownie wzrosła.
- Silna presja konkurencyjności
  - dostarczać lepsze, skrojone pod potrzeby klienta, usługi

## Po co eksplorować dane? Naukowy punkt widzenia.

- Dane zbierane są w niebotycznej prędkości (częstokroć wielu GB/godzinę)
  - sensory na satelitach,
  - teleskopy,
  - symulacje generujące terabajty danych.
- Tradycyjne techniki nie nadają się do pracy z surowymi danymi.
- Eksploracja danych może pomóc naukowcom:
  - w klasyfikowaniu i dzieleniu danych,
  - w formułowaniu hipotez.

## Eksploracja dużych zbiorów danych – motywacja.

- Często informacja jest „ukryta” wśród danych i nie jest ewidentnie widoczna.
- „Żywi” analitycy mogą potrzebować tygodni, aby odnaleźć użyteczną informację.
- Większość z gromadzonych danych w ogóle nie jest nigdy analizowana



## Czym jest eksploracja danych?

- Eksploracja danych (znajdowanie wiedzy w bazach danych):
  - Efektywna ekstrakcja interesującej (prawdziwej, nietrywialnej, ukrytej, poprzednio nieznannej i potencjalnie użytecznej) informacji oraz wzorców z danych w dużych bazach danych
- Alternatywne nazwy
  - Odkrywanie wiedzy w bazach danych (KDD), ekstrakcja wiedzy, analiza wzorców/danych, archeologia danych, wywiad biznesowy
- Czym nie jest eksploracja danych?
  - (dedukcyjne) wykonywanie zapytań,
  - systemy eksperckie lub małe programy uczenia maszynowego/statystyczne.

## Operowanie na bazach danych vs. eksploracja danych.

- Zapytania
    - Dobrze zdefiniowane
    - SQL
  - Dane
    - Operacyjne
  - Output
    - Precyzyjny
    - Podzbiór bazy danych
- Zapytania
    - Słabo zdefiniowane
    - Brak standardowego języka zapytań
  - Dane
    - Nieoperacyjne
  - Output
    - Rozmyty
    - Nie jest podzbiorem bazy danych

## Obsługa zapytania bazy danych vs. eksploracja danych

- Zapytania w bazach danych (OLTP/OLAP)
  - Ile kanap sprzedała IKEA w pierwszych trzech miesiącach 2020 w Polsce, a ile w Niemczech?
  - Znajdź wszystkich klientów IKEI, którzy kupili kanapę za co najmniej 1000 zł
  - Znajdź wszystkich klientów IKEI, którzy kupili kanapę.
- Zapytania w eksploracji danych
  - W których krajach IKEA odnotowała anomalną sprzedaż kanap? (klasyfikacja)
  - Zidentyfikować wszystkich klientów IKEI kupujących kanapy, którzy mają podobne przyzwyczajenia zakupowe (grupowanie)
  - Jakie inne produkty kupują klienci IKEI, którzy wybrali się do sklepu po kanapę (reguły asocjacyjne)
  - Jakie są najbardziej popularne wzorce zakupowe wśród klientów IKEI po zakupie sofy? (eksploracja sekwencyjna)

## **Dlaczego eksploracja danych? Potencjalne zastosowania**

- Analiza baz danych i wsparcie decyzyjne
  - Analiza rynkowa i management rynku
    - target marketing, management relacji z klientem, analiza koszyka rynkowego, segmentacja rynku itp.
  - Analiza ryzyka i risk management
    - przewidywanie, kontrola jakości, analiza konkurencyjności itp.
- Inne zastosowania
  - Eksploracja tekstu/filtry antyspamowe/klasyfikacja dokumentów
  - Strumieniowa eksploracja danych
  - Eksploracja danych internetowych
  - Biologia i bioinformatyka (analiza danych DNA)
  - Astronomia

## Analiza rynku i management rynku.

- Gdzie znajdują się źródła danych do analizy?
  - Transakcje kart kredytowych, lojalnościowych, kuponów promocyjnych, skarg klienckich
- Target marketing
  - Znalezienie grup „modelowych” klientów o takich samych charakterystykach: zainteresowaniach, poziomie zarobków, zwyczajach konsumenckich, wykształceniu, miejscu zamieszkania
- Określenie wzorców zakupowych klientów w czasie
  - Konwersja z samodzielnego do wspólnego rachunku bankowego itp.
- Profilowanie klientów
  - Eksploracja danych pozwala identyfikować które rodzaje klientów co kupują (klasyfikacja i grupowanie)
- Identyfikowanie wymagań klientów
  - Identyfikowanie „najlepszych” produktów dla różnych klientów
  - Używanie technik przewidywania do zidentyfikowania czynników przyciągających klientów
- Dostarczanie raportów podsumowujących
  - Różne wielowymiarowe raporty podsumowujące
  - Informacja o podsumowaniach statystycznych (tendencje danych takie jak np. wariancja)

## **Analiza korporacyjna i risk management**

- Planowanie finansowe i ocena aktywów
  - analiza i prognozowanie przepływów pieniężnych
  - analiza roszczeń warunkowych w celu wyceny aktywów
  - analiza przekrojowa i szeregów czasowych (wskaźnik finansowy, analiza trendu itp.)
- Planowanie zasobów:
  - podsumowanie i porównanie zasobów i wydatków
- Konkurencja:
  - monitorowanie konkurencji i kierunków rynkowych
  - grupowanie klientów w klasy i procedury wyceny opartej na klasach
  - ustalenie strategii cenowej na wysoce konkurencyjnym rynku

## Wykrywanie i management oszustw.

- Zastosowania

- szeroko stosowane w służbie zdrowia, handlu detalicznym, obsłudze kart kredytowych, telekomunikacji itp.

- Podejście

- wykorzystywać dane historyczne do tworzenia modeli nieuczciwych zachowań oraz wykorzystywać eksplorację danych, aby zidentyfikować podobne przypadki

- Przykłady

- ubezpieczenia komunikacyjne: wykrywanie grup ludzi, którzy organizują wypadki aby zarabiać na ubezpieczeniu
  - pranie pieniędzy: wykrywanie podejrzanych transakcji pieniężnych
  - ubezpieczenie medyczne: wykrywanie „pacjentów zawodowych” i współpracujących z nimi lekarzy

## Inne zastosowania.

### *Sporty*

- IBM Advanced Scout przeanalizował statystyki meczów NBA (rzuty zablokowane, asysty i faule) w celu uzyskania przewagi dla New York Knicks i Miami Heat

### *Astronomia*

- JPL i Palomar Observatory odkryły 22 kwazary przy pomocy eksploracji danych

### *Pomoc w surfowaniu po Internecie*

- IBM Surf-Aid stosuje algorytmy eksploracji danych do logów dostępu do sieci WWW aby znaleźć preferencje klienta

## Dalsze przykłady.

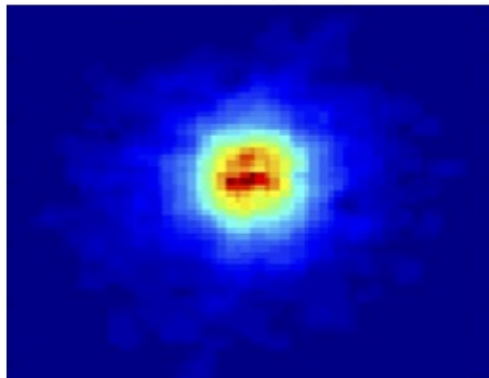
- Drzewa decyzyjne zbudowane na podstawie historii kredytów bankowych pozwalają bankom zdecydować, czy przyznać pożyczkę i według jakiej stopy.
- Wzorce zachowań podróżnych zarządzają sprzedażą zniżek na bilety w samolotach, pokojach hotelowych itp.
- „Pieluchy i piwo”. Umieszczanie artykułów na półkach supermarketów zwiększa ich sprzedaż.
- Skycat i Sloan Sky Survey: grupowanie obiektów nieba według ich poziomów promieniowania w różnych pasmach pozwoliło astronomom na rozróżnianie galaktyk, pobliskich gwiazd i wiele innych rodzajów ciał niebieskich.
- Porównanie genotypu osób z / bez schorzenia pozwoliło na odkrycie zestawów genów, które razem odpowiadają za wiele przypadków cukrzycy.

## **Zastosowanie: katalogowanie wyników badań astronomicznych.**

- Cel: przewidzieć klasę (gwiazdę lub galaktykę) obiektów na niebie, szczególnie słabe wizualnie, oparte na teleskopie zdjęcia ankietowe (z Obserwatorium Palomar)
  - 3000 obrazów z  $23\,040 \times 23\,040$  pikseli na obraz
- Podejście:
  - Podziel obraz na segmenty
  - Zmierz atrybuty (cechy) obrazu - 40 z nich na obiekt
  - Modeluj klasę w oparciu o te cechy
  - Można znaleźć 16 nowych kwazarów o wysokim przesunięciu ku czerwieni, niektóre z najdalszych obiektów, które są trudne do znalezienia!

## Klasyfikowanie galaktyk.

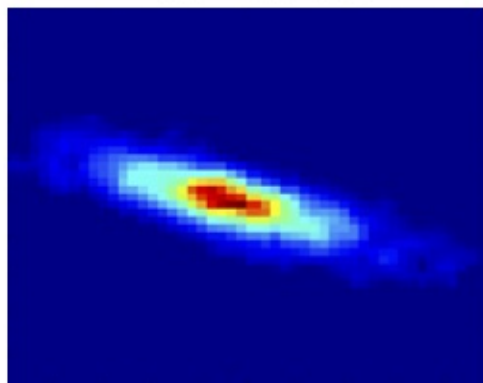
*Early*



### **Class:**

- Stages of Formation

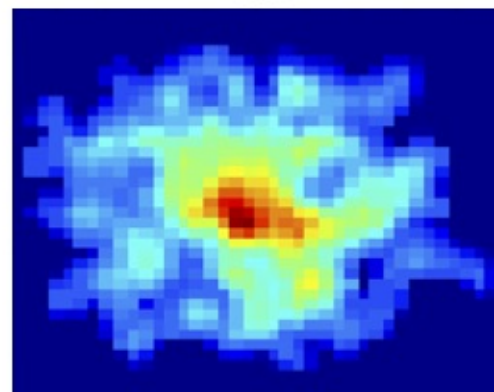
*Intermediate*



### **Attributes:**

- Image features,
- Characteristics of light waves received, etc.

*Late*



### **Data Size:**

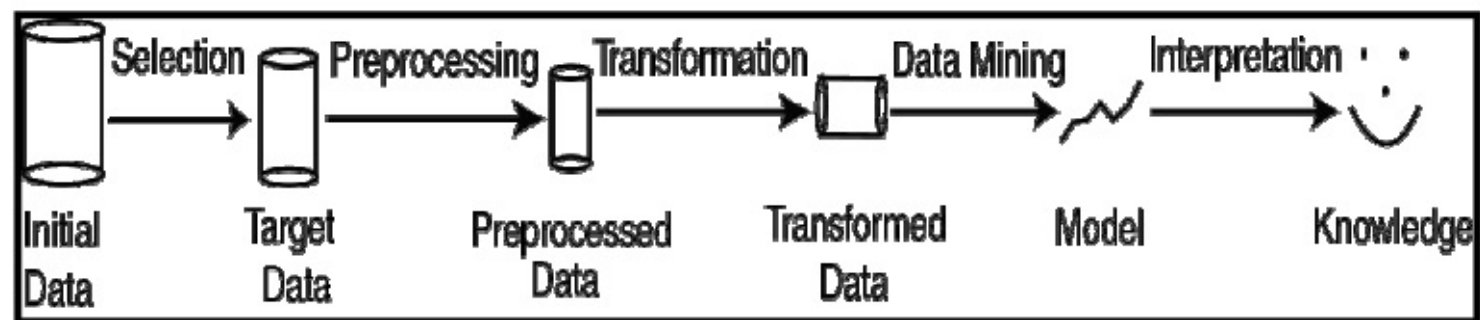
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

## **Eksploracja danych vs. odkrywanie wiedzy w bazach danych (KDD).**

Odkrywanie wiedzy w bazach danych (KDD – knowledge discovery in databases): proces znajdowania użytecznych informacji i wzorców wśród danych.

Eksploracja danych: użycie algorytmów celem wyekstrachowania informacji i wzorców otrzymanych w procesie KDD.

Eksploracja danych: część procesu KDD.



## Kroki w procesie KDD

- Nauka dziedziny, do którego proces będzie stosowany:
  - odpowiednia wcześniejsza wiedza i cele zastosowania
- Tworzenie docelowego zestawu danych: wybór danych
- Czyszczenie i wstępne przetwarzanie danych: (może zająć 60% wysiłku!)
- Redukcja i transformacja danych:
  - Znajdź przydatne funkcje, zredukuj wymiar/liczbę zmiennych.
- Wybór funkcji eksploracji danych
  - podsumowanie, klasyfikacja, regresja, asocjacja, grupowanie.
- Wybór algorytmu(ów) eksploracji
- Eksploracja danych: wyszukiwanie wzorców zainteresowań
- Ocena wzorców i prezentacja wiedzy
  - wizualizacja, transformacja, usuwanie zbędnych wzorców itp.
- Wykorzystanie odkrytej wiedzy

## **Przykład procesu KDD: analiza logów sieciowych.**

Wybór:

- Wybierz dane logu (daty i lokalizacje) do użycia

Przetwarzanie wstępne:

- Usuń identyfikujące adresy URL
- Usuń logi błędów

Transformacja:

- Podział na sesje (sortowanie i grupowanie)

Eksploracja danych:

- Zidentyfikuj i policz wzorce
- Skonstruuj strukturę danych

Interpretacja/ocena:

- Zidentyfikuj i pokaż często używane sekwencje.

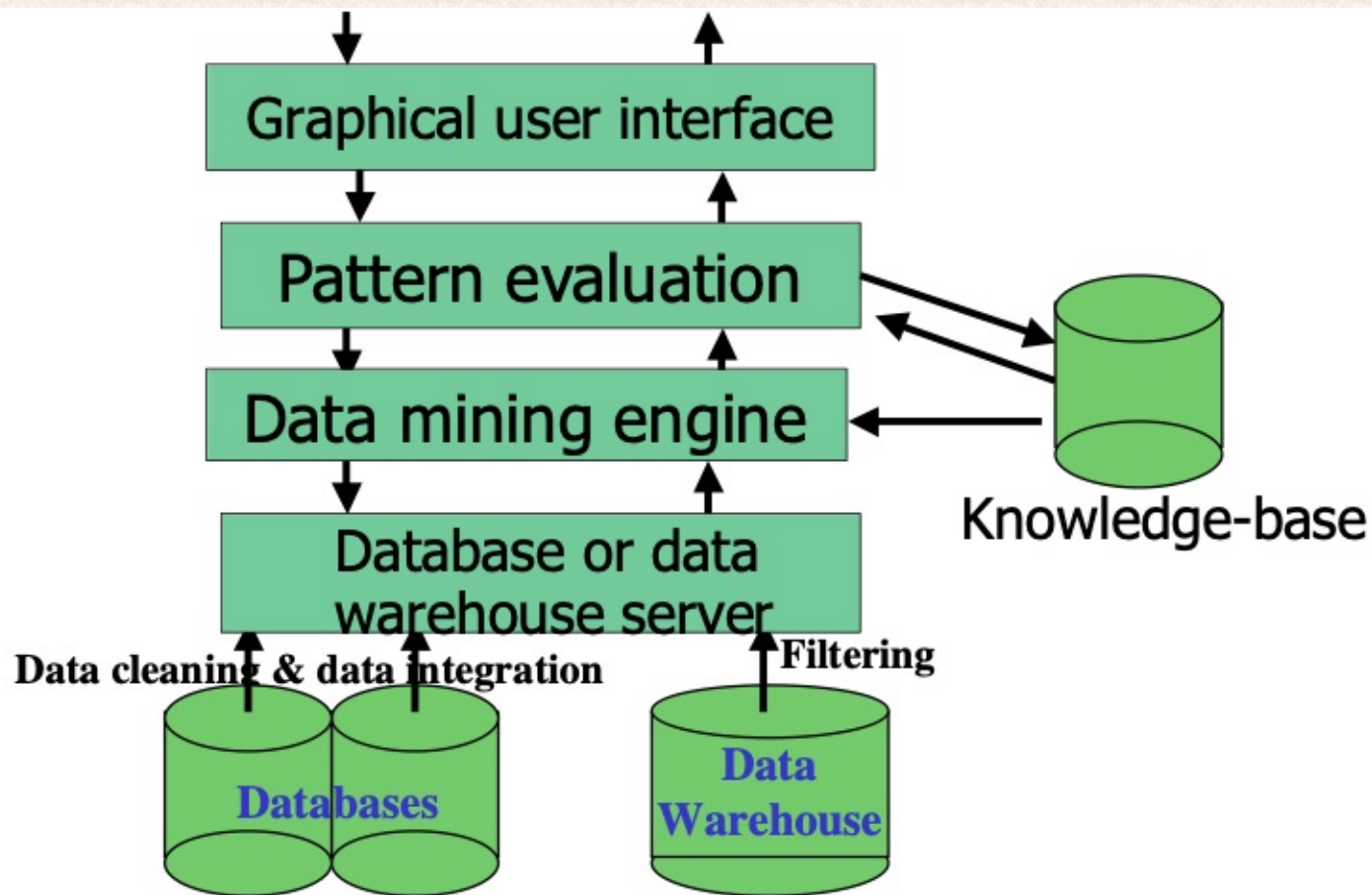
Potencjalne aplikacje użytkownika:

- Przewidywanie zawartości pamięci podręcznej
- Personalizacja

## **Typowe problemy.**

- Interakcja z człowiekiem
- Overfitting
- Wartości odstające
- Interpretacja
- Wizualizacja
- Duże zbiory danych
- Wysoki wymiar
- Dane multimedialne
- Brakujące dane
- Nieistotne dane
- Zakłócone dane

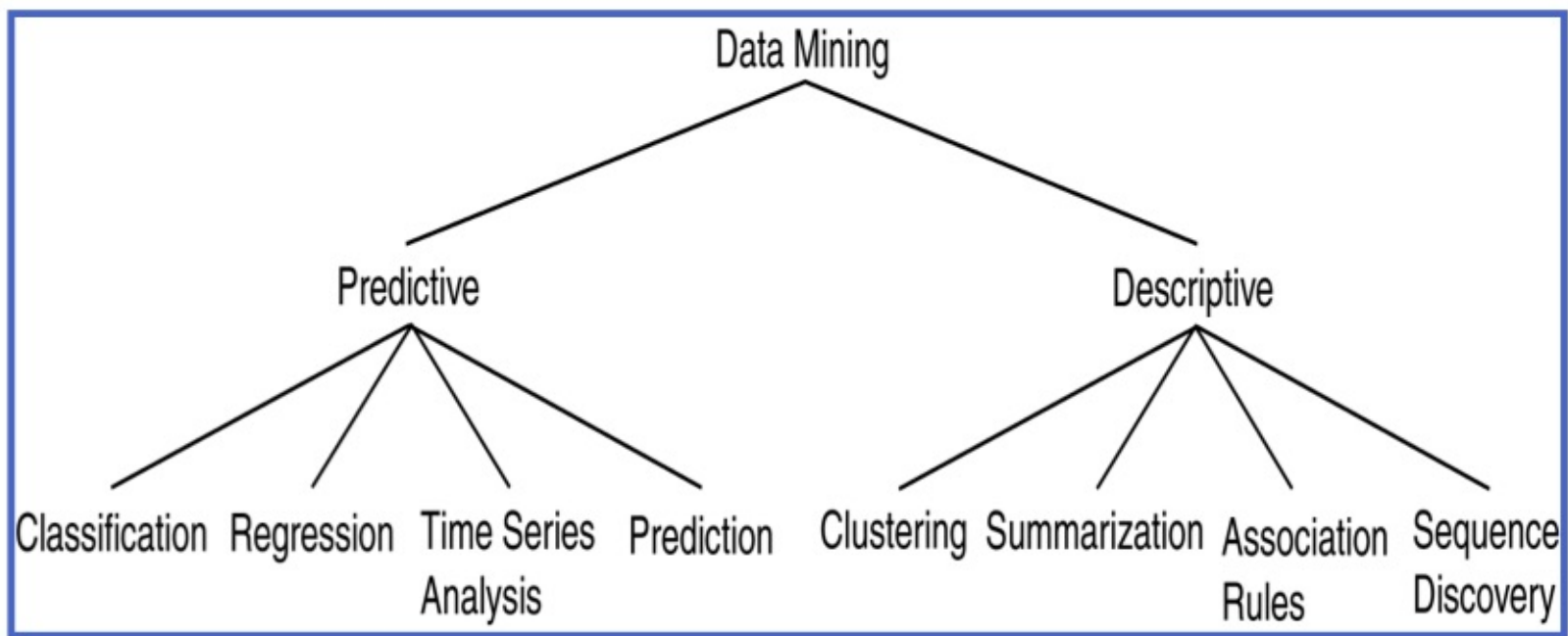
## Architektura typowego systemu eksploracji danych.



## **Eksploracja danych: na jakiego rodzaju danych?**

- Relacyjne bazy danych
- Magazyn danych
- Transakcyjne bazy danych
- Zaawansowane bazy danych i repozytoria informacji
  - Bazy danych obiektowe i obiektowo-relacyjne
  - Dane przestrzenne i czasowe
  - Dane szeregów czasowych i dane strumieniowe
  - Bazy tekstowe i multimedialne
  - Heterogeniczne i starsze bazy danych
  - WWW

## Modele i zadania eksploracji danych.



## **Podstawowe zadania eksploracji danych.**

**Klasyfikacja** odwzorowuje dane na predefiniowane grupy lub klasy

- Nadzorowana nauka
- Rozpoznawanie wzorców
- Prognoza

**Regresja** służy do odwzorowania elementu danych na zmienną prognozowaną o wartości rzeczywistej.

**Klastrowanie** grupuje podobne dane w klastry.

- Uczenie się bez nadzoru
- Segmentacja
- Partycjonowanie

**Podsumowanie** mapuje dane na podzbiory ze skojarzonymi uproszczonymi opisami.

- Charakterystyka
- Uogólnienie

**Analiza połączeń** ujawnia relacje między danymi.

- Analiza powinowactwa
- Reguły asocjacji
- Analiza sekwencyjna określa wzorce sekwencyjne.

## Przykład: analiza szeregów czasowych.

- Przykład: giełda papierów wartościowych
- Przewidywanie przyszłych wycen
- Określanie podobnych wzorców w czasie
- Klasyfikacja zachowań

