Analiza danych niestrukturalnych
Wprowadzenie

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska

Klasyfikacja danych ze względu na strukturę

Struktura danych implikuje metody ich przetwarzania i analizy.

Dane strukturalne

Ściśle określona postać – tabelaryczna lub zgodna z tą postacią, w której można wyodrębnić wiersze (rekordy, obserwacje) i kolumny (pola, cechy, zmienne).
Sprecyzowane: struktura rekordów (jednolita w całym zbiorze), znaczenie atrybutów oraz typy danych.
Pliki baz danych, tekstowe (CSV, TSV), zdefiniowane aplikacji analitycznej (np. SAS).

Dane półstrukturalne

Nazywane też semi-strukturalnymi – bez sztywnego schematu właściwego dla danych strukturalnych, jednak z elementami organizującymi (porządkującymi).
Wykorzystują znaczniki, klucze i metadane, które opisują strukturę informacji.
Pliki XML, JPSON, HTML, logi systemowe, pliki konfiguracyjne.

Dane niestrukturalne

Inaczej nieustrukturyzowane. Charakteryzują się brakiem formalnego schematu opisującego treść i brak jednoznacznej struktury, ale określone formaty zapisu (formaty te definiują jedynie sposób kodowania i przechowywania danych, a nie ich logiczną strukturę czy semantykę).
Stanowią 80% do nawet 90% wszystkich danych wytwarzanych obecnie w praktyce (zarówno w firmach, jak i ogólnie na świecie).

Przykłady danych niestrukturalnych

- Dokumenty tekstowe: TXT, DOCX, PDF – format określa, jak zapisany jest tekst i jego układ, ale nie narzuca znaczenia treści.
- Obrazy: JPEG, PNG, TIFF – format opisuje sposób kodowania pikseli, nie zawartość obrazu.
- Nagrania audio: MP3, WAV – format określa parametry dźwięku, ale nie jego znaczenie.
- Materiały wideo: MP4, AVI – format definiuje sposób zapisu obrazu i dźwięku w czasie.

Szacuje się, że tylko niewielki ułamek (często poniżej 1%) danych niestrukturalnych jest faktycznie analizowany i wykorzystywany do podejmowania decyzji biznesowych.

Aby wydobyć użyteczną informację z danych niestrukturalnych, **konieczne jest zastosowanie zaawansowanych metod analizy**, takich jak przetwarzanie języka naturalnego, analiza obrazu czy rozpoznawanie mowy.

Specyfika danych niestrukturalnych

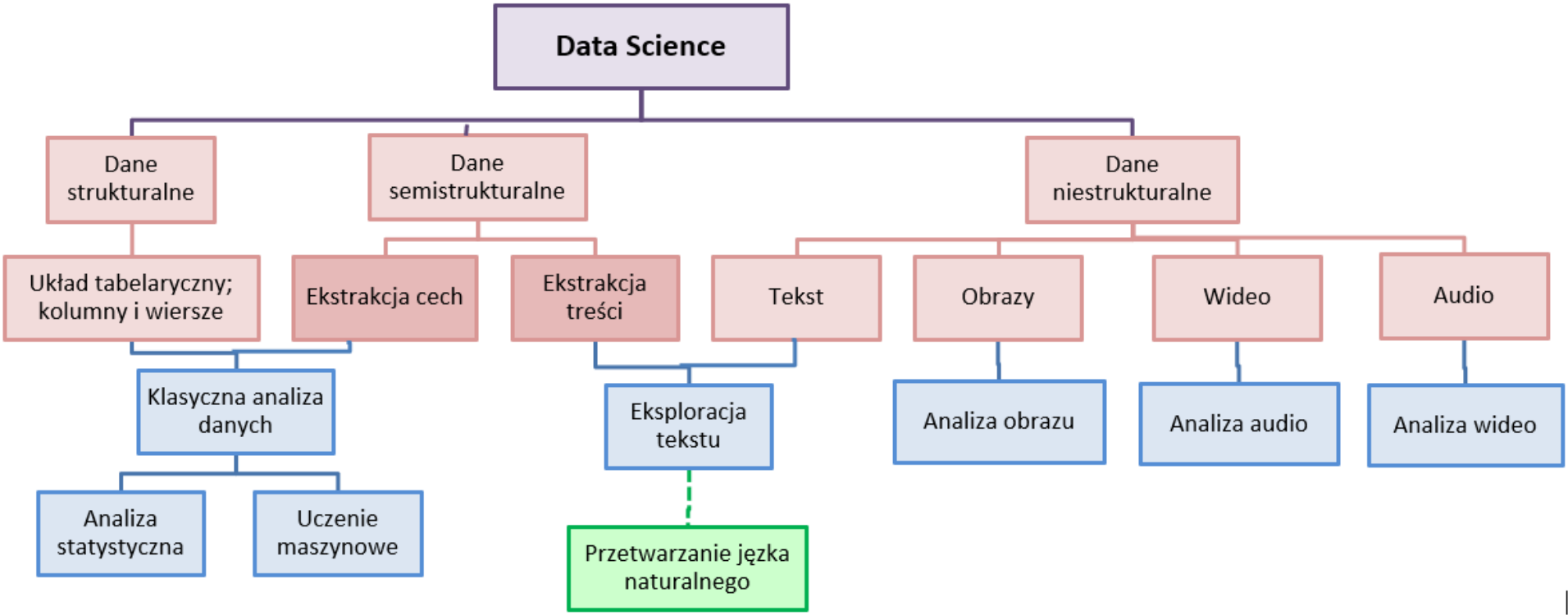
- **Heterogeniczność form i źródeł** →
  Trudności w ujednoliceniu sposobów ich przechowywania, przetwarzania i analizy.
- **Duża skala i dynamiczny przyrost danych** →
  Tworzenie skalowalnych rozwiązań magazynowych oraz aplikacja zaawansowanych algorytmów UM i SI, co implikuje zapotrzebowanie na specjalistyczną infrastrukturę i duże koszty obliczeniowe.
- **Wysoki poziom szumu i redundancji** →
  Stosowanie etapów wstępnego przetwarzania, takich jak filtrowanie, czyszczenie i redukcja danych.
- **Zmienność językowa i kontekstowa** →
  Trudności w jednoznacznej interpretacji i automatycznej analizie (swoistość języka naturalnego).
- **Ograniczenia klasycznej statystyki (i UM)** →
  Orzekształcania do postaci ustrukturyzowanej (ekstrakcja cech, wektoryzacja tekstu, anotowanie danych wizualnych).
- **Interpretowalność wyników** →
  „Czarna skrzynka” i ograniczone zaufanie do wyników
- **Integracja z danymi strukturalnymi** →
  Przekształcenia i harmonizacja reprezentacji danych
- **Bezpieczeństwo i prywatność danych** →
  Ochrona prywatności i zapewnienie zgodności z regulacjami prawnymi

Analiza danych niestrukturalnych ADN w data science

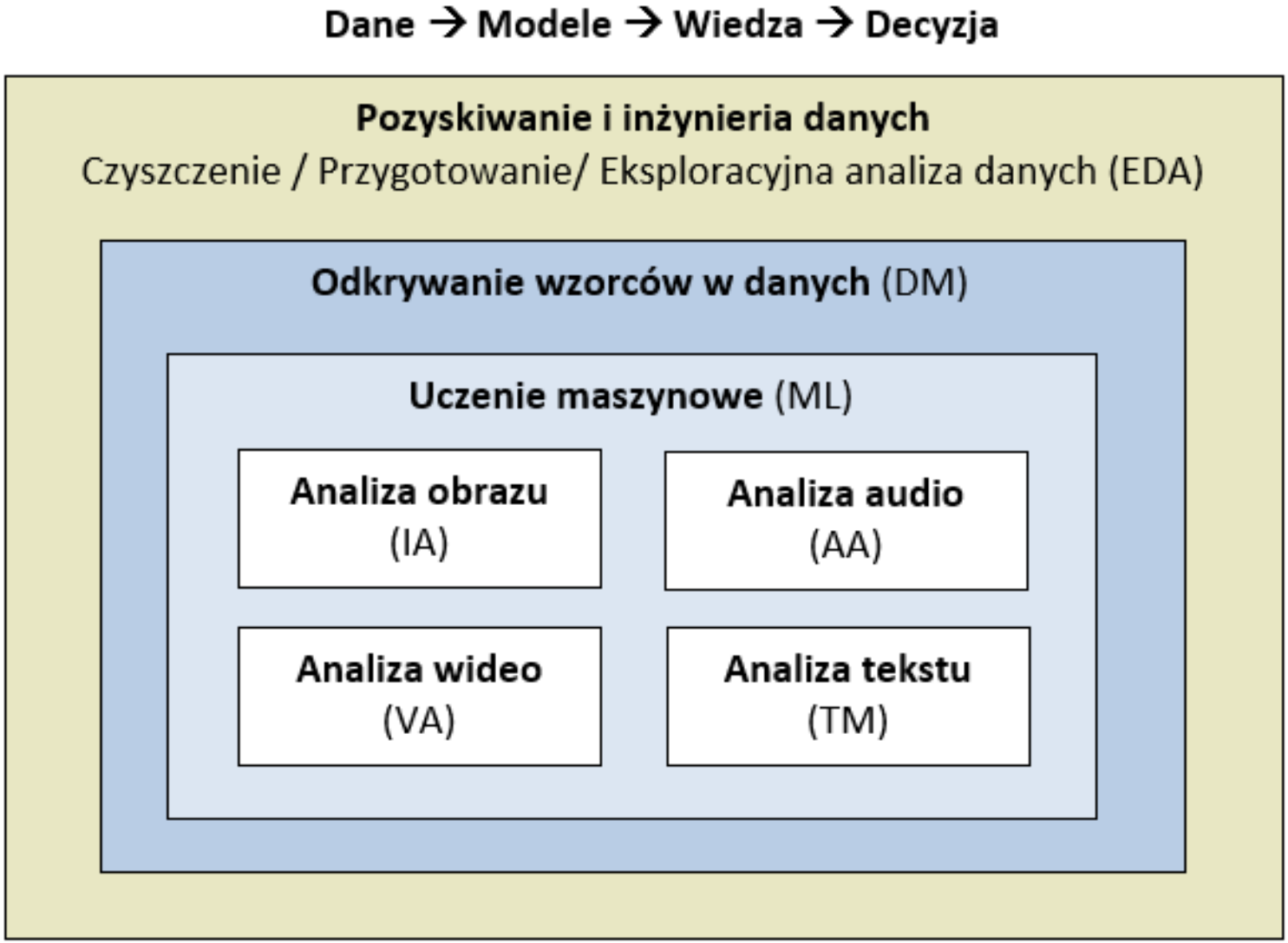
Data Science (DS, nauka o danych, danologia, danetyka)
Dziedzina interdyscyplinarna będąca syntezą zaawansowanej **wiedzy statystycznej** i **matematycznej**, oraz **umiejętności informatycznych** (programowania i zarządzania danymi).
DS + wiedza dziedzinowa = umiejętność interpretacji i komunikowania wyników.
Cel DS: wydobywanie **użytecznej wiedzy z danych** (data-driven knowledge) **niezależnie od ich formy, struktury czy źródła**.

- Ze względu na rodzaj danych (kryterium przedmiotowe) →
  ADN to wyspecjalizowany komponent DS a nie osobnym obszar badawczy.
- Ze względu na stosowane metody (kryterium metodologiczne) →
  ADN wykorzystuje te same metody co analityka danych strukturalnych
- Ze względu na proces analizy (kryterium procesowe – pipeline) →
  ADN różni się od klasycznego DS etapem przygotowania i reprezentacji danych

Umiejscowienie ADN w DS



Schemat koncepcyjny pracy z danymi niestrukturalnymi



Eksploracyjna analiza danych EDA

Exploratory Data Analysis

Badanie i wizualizacja danych; ich poznanie, zrozumienie i podsumowanie:

- identyfikacja głównych cech
- wykrycia anomalii
- formułowanie hipotez

Ważny krok przygotowania danych przed zastosowaniem zaawansowanych technik modelowania, operacje często manualna lub półautomatyczne.

Eksploracja danych DM

Data Mining

Odkrywaniu wzorców, zależności i struktur w danych z wykorzystaniem metod matematycznych i technologii informatycznych (również UM) do badania danych. Te same cele są określane w analizie danych niestrukturalnych, ale dane te muszą być wcześniejszej transformowane do postaci możliwej do eksploracji (czyli ustrukturalizowane).

Uczenie maszynowe ML

Machine Learning

Poddziedzina sztucznej inteligencji; algorytmy analizują zbiory danych, wykrywają w nich prawidłowości i zależności między opisującymi je cechami i na tej podstawie uczą się przewidywać lub podejmować decyzje, bez konieczności bezpośredniego programowania działań przez człowieka.

Modele uczenia maszynowego

- uczenie **nadzorowane**: (Supervised Learning); model uczy się na wielowymiarowych etykietowanych danych historycznych (obecny nadzór, nauczyciel) w celu przewidzenia odpowiedzi dla nowych danych
- **nienadzorowane** (Supervised Learning); model analizuje wielowymiarowe dane bez nadzoru/nauczyciela, szukając w nich ukrytych wzorców, struktur, anomalii lub relacji
- uczenie **ze wzmocnieniem** (Reinforcement Learning); model oparty na interakcji z otoczeniem, który przypomina naukę poprzez metodę prób i błędów (nauka przez interakcje i doświadczenie)

Efektem końcowym jest model matematyczny - zestaw reguł i wag, który pozwala komputerowi przetwarzać dane wejściowe w celu wygenerowania konkretnej decyzji lub informacji.

Analiza obrazu IA

Image Analysis

Cyfrowe przetwarzanie obrazów. Obraz jest czyszczony i poprawiany poprzez zastosowanie filtrów i operacji na pikselach. Następnie algorytmy rozpoznają, co jest na obrazie (rozpoznają i intepretują dostarczone dane obrazowe).

Techniki IA wg poziomu zaawansowania

- Poziom 1. Podstawowe rozpoznanie co jest na obrazie
  - ✓ Klasyfikacja i rozpoznawanie (Image Classification and Recognition)
  - ✓ Rozpoznawanie znaków i tekstu (OCR)
  - ✓ Ekstrakcja cech (Feature Extraction)
- Poziom 2. Poziom średni – lokalizacja obiektów
  - ✓ Detekcja obiektów (Object Detection)
  - ✓ Rozpoznawanie punktów kluczowych (Keypoint Detection)
  - ✓ Wyszukiwanie przez podobieństwa (Image Retrieval)
- Poziom 3. Poziom zaawansowany – lokalizacja obiektów
  - ✓ Segmentacja semantyczna i instancji (Segmentation)
  - ✓ Estymacja głębi (Depth Estimation)
  - ✓ Rekonstrukcja (Inpainting)

Analiza audio AA

Audio Analysis

Automatyczne wydobywanie informacji, struktury i znaczenia z sygnałów dźwiękowych. W ujęciu technicznym fala akustyczna (dane niestrukturalne) jest przekształcana w strukturalne dane liczbowe lub tekstowe.

Techniki AA wg poziomu zaawansowania

- Poziom1. Podstawowe przetwarzanie i ekstrakcja (Low-level)
  - ✓ Analiza cech spektralnych (Spectral Feature Extraction)
  - ✓ VAD (Voice Activity Detection)
- Poziom 2. Identyfikacja i dopasowywanie (Pattern Matching)
  - ✓ Fingerprinting (Acoustic Fingerprinting)
  - ✓ Klasyfikacja zdarzeń dźwiękowych (Audio Tagging)
- Poziom 3: Analiza biometryczna i semantyczna (High-level Analysis)
  - ✓ Identyfikacja mówcy (Speaker Identification
  - ✓ Diaryzacja mówców (Speaker Diarization)
  - ✓ Rozpoznawanie emocji (Speech Emotion Recognition)
- Poziom 4: Złożone systemy generatywne i separacyjne (Expert-level); systemy kognitywne
  - ✓ Separacja źródeł (Source Separation)
  - ✓ Przetwarzanie sygnałów mowy (Speech Signal Processing: ASR. TTS)

Analityka wideo VA

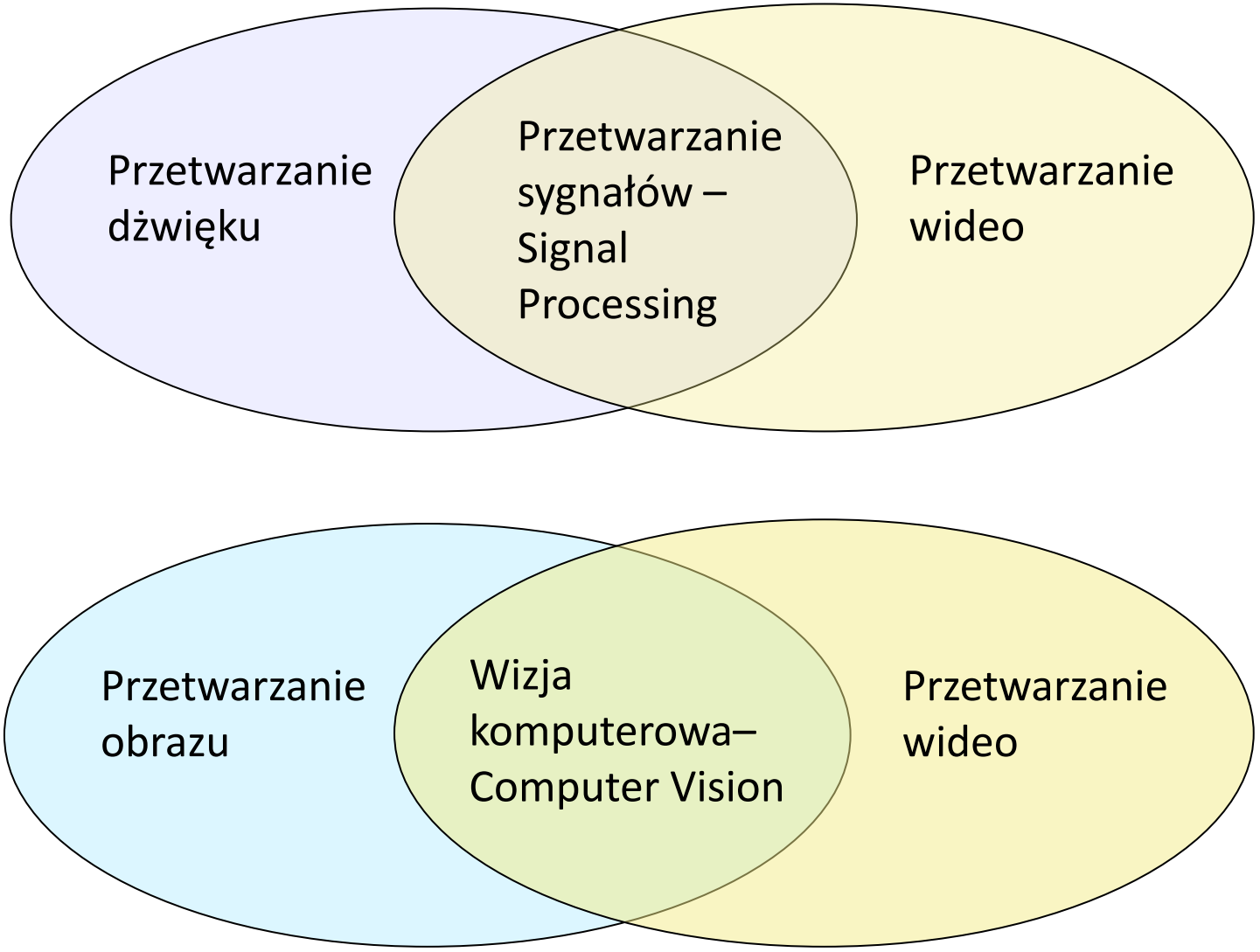
Video Analytics (Spatio-temporal Analysis)

Automatyczne wydobywanie informacji, struktury i znaczenia z nagrań wideo lub strumieni na żywo. Analiza-przestrzenno-czasowa (Spatio-temporal Analysis) bada zmiany zachodzące w obrazie zarówno w przestrzeni (gdzie jest obiekt), jak i w czasie (co robi) - śledzenie zmian między klatkami.
VA to automatyczna analiza wizyjna w czasie rzeczywistym.

Techniki VA wg poziomu zaawansowania

- Poziom 1. Bazowy (przetwarzanie niskopoziomowe)
  - ✓ Detekcja ruchu i tła (Motion & Direction Detection)
- Poziom 2. Operacyjny (średniozaawansowany)
  - ✓ Śledzenie obiektów (Object Tracking):
- Poziom 3. Analityczny (wysokozaawansowany)
  - ✓ Mapy cieplne (Heatmaps).
  - ✓ Liczenie obiektów (People/Vehicle Counting)/ Analiza gęstości tłumu (Crowd Density Estimation).
  - ✓ Rozpoznawanie zachowań i zdarzeń (Behavior Analysis).

Przenikanie obszarów ADN w zakresie przetwarzania danych



Analiza tekstu TM

Text Mining

Proces automatycznego wydobywania wysokiej jakości informacji, ukrytych wzorców i trendów z dużych zbiorów nieustrukturyzowanych danych tekstowych. W przeciwieństwie do tradycyjnej eksploracji danych (data mining), która operuje na uporządkowanych bazach danych, text mining zajmuje się językiem naturalnym – m.in. artykułami, e-mailami czy wpisami w mediach społecznościowych

We wszystkich swoich obszarach TM korzysta z NLP na etapie przygotowania i reprezentacji tekstu. NLP dostarcza przetworzone, ustrukturyzowane lub zakodowane znaczenie tekstu, a TM wykorzystuje je do statystycznej analizy, wykrywania wzorców i klasyfikacji.