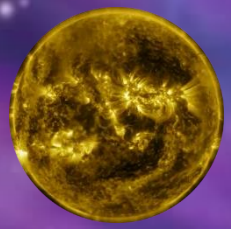
METODY INŻYNIERII WIEDZY

KNOWLEDGE ENGINEERING AND DATA MINING

WSTĘP I TAKSONOMIA



Adrian Horzyk

Akademia Górniczo-Hutnicza

Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Katedra Automatyki i Inżynierii Biomedycznej, Laboratorium Biocybernetyki

30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk

INŻYNIERIA WIEDZY

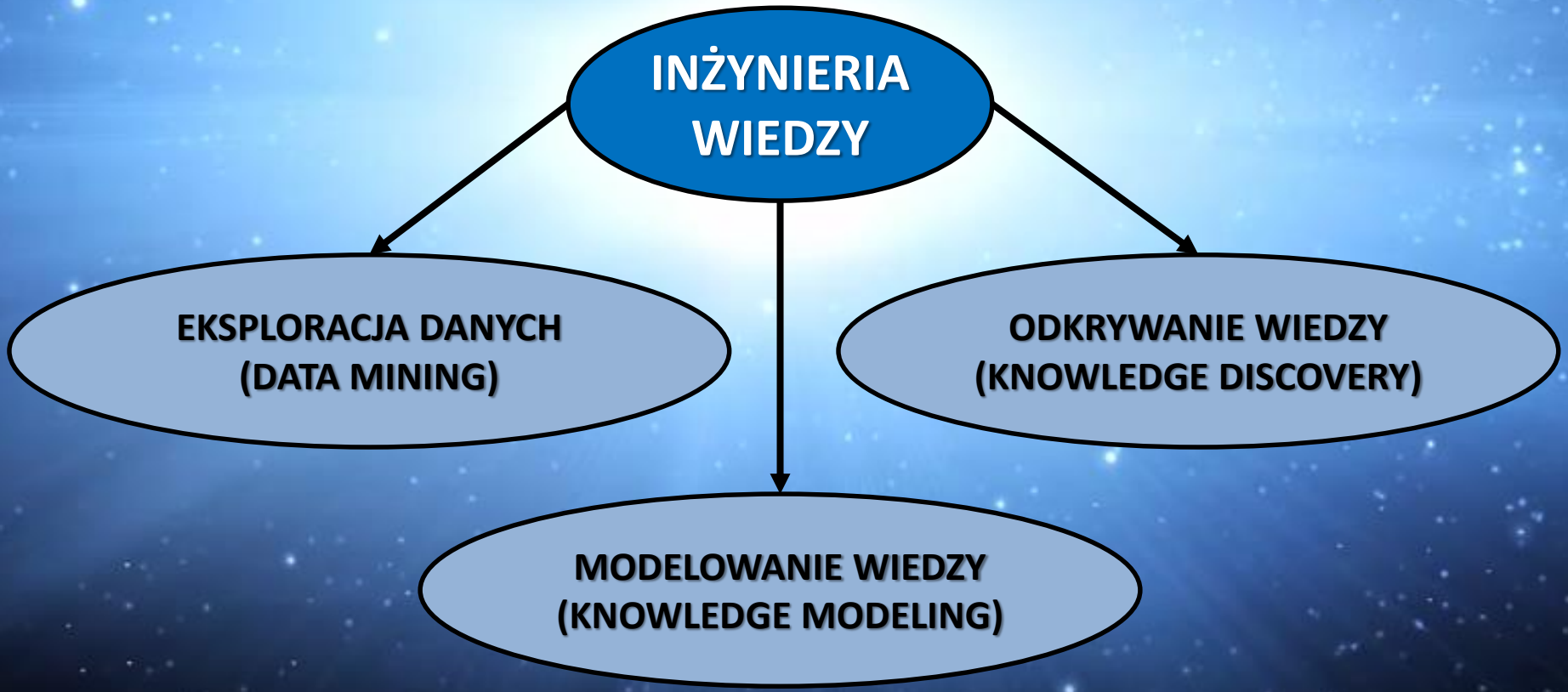
KNOWLEDGE ENGINEERING (KE)

to obszar informatyki zajmujący się metodami eksploracji, reprezentacji i modelowania wiedzy z danych (ich zbiorów, reguł, baz danych) oraz metodami wnioskowania na ich podstawie.

EKSPLORACJA WIEDZY Z DANYCH

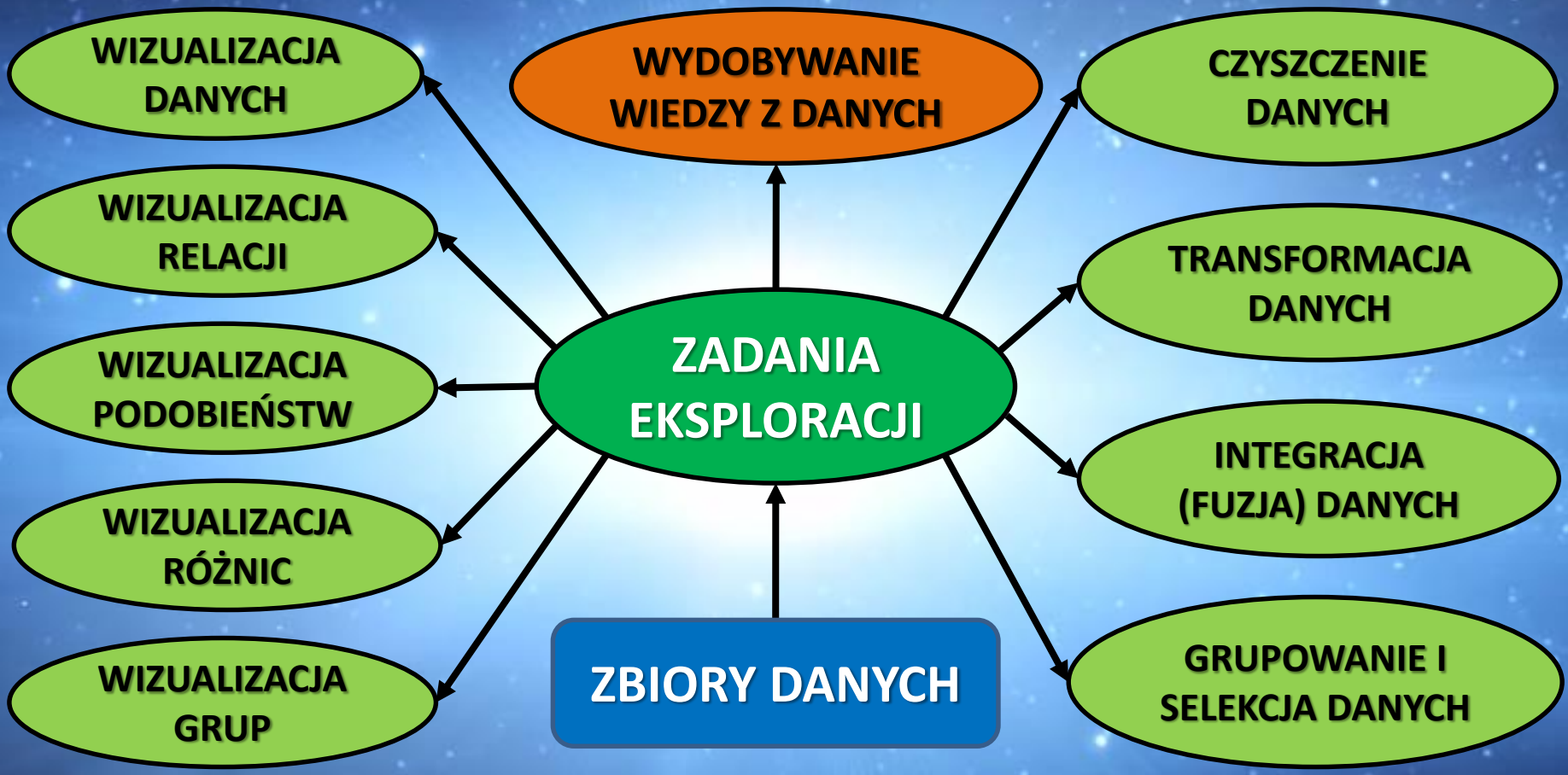
KNOWLEDGE DISCOVERY IN DATABASES (KDD)

to proces odkrywania wiedzy ukrytej w danych lub ich zbiorach (czyli baza danych) polegający na wyszukiwaniu prawidłowości, powtarzalności i zależności (relacji) pomiędzy danymi.



ZADANIA EKSPLORACJI DANYCH

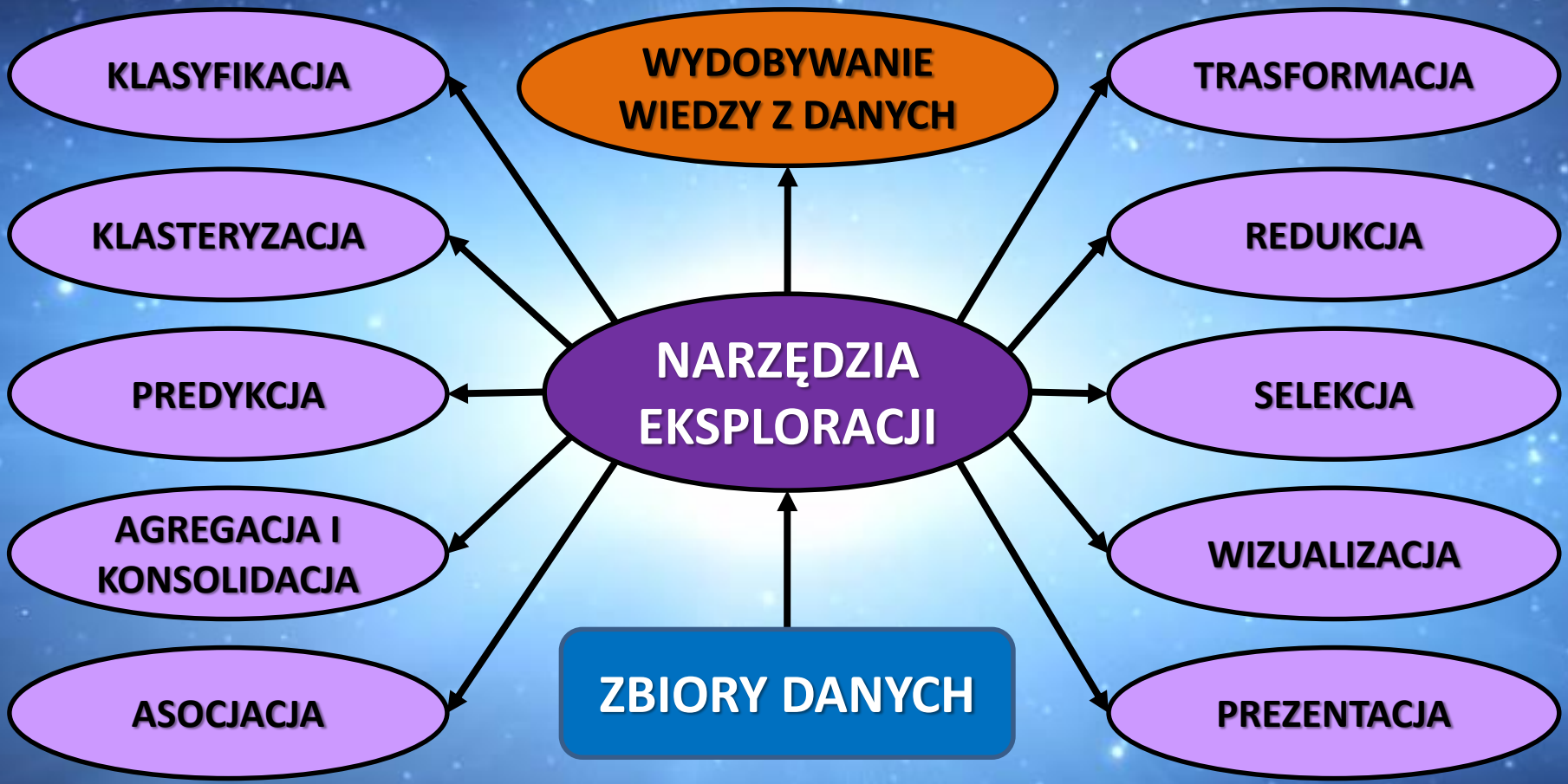
DATA MINING TASKS



Eksploracja danych to zwykle proces wieloetapowy związany z wstępną obróbką danych (czyszczenie, normalizacja, standaryzacja lub inny rodzaj transformacji), porównywaniem, integracją, grupowaniem i selekcją danych oraz wizualizacją danych, ich cech, grup, podobieństw, różnic i zależności (relacji).

NARZĘDZIA EKSPLORACJI DANYCH

DATA MINING METHODS AND TOOLS



Eksploracja danych to zwykle proces wieloetapowy związany z wstępną obróbką danych (czyszczenie, normalizacja, standaryzacja lub inny rodzaj transformacji), porównywaniem, integracją, grupowaniem i selekcją danych oraz wizualizacją danych, ich cech, grup, podobieństw, różnic i zależności (relacji).

KLASYFIKACJA - CLASSIFICATION

- to zadanie przyporządkowania wzorca do pewnej klasy.
- to zadanie rozpoznawania wzorca jako elementu pewnej klasy.

Klasa – to pewna grupa wzorców charakteryzujących się podobnymi cechami/właściwościami dla określających je atrybutów/parametrów.

W wyniku klasyfikacji wzorcowi zostaje przyporządkowana pewna klasa, reprezentowana zwykle przez pewną **etykietę klasy**.

Jeśli wzorzec należy równocześnie do kilku klas, wtedy mówimy o zagadnieniu **multiklasyfikacji (multiclass classification)**, np.:

- ser Mozzarella należy do klas: serów, nabiału, produktów spożywczych

Sklasyfikowanie wzorca jako przynależnego do określonej klasy może być rozważane jako proces:

- Binarny / zero-jedynkowy / dyskretny: należy lub nie należy do klasy
- Rozmyty / predyktywny / ciągły: o określonym stopniu przynależności do klasy

REDUKCJA I TRANSFORMACJA DANYCH

Metody redukcji i transformacji danych – mają za zadanie doprowadzić do optymalnej reprezentacji dużych ilości danych, tj. takiej reprezentacji, żeby dane w dalszym ciągu były reprezentatywne dla rozważanego problemu, np. klasyfikacji, czyli umożliwiały poprawną dyskryminację wzorców, tj. rozróżnienie ich według pozostałych po redukcji danych.

Optymalna reprezentacja danych może być osiągnięta na skutek:

- **Redukcji wymiaru danych – czyli usuwania mniej istotnych atrybutów danych, oraz Selekcji atrybutów najistotniejszych pod kątem rozwiązywanego zadania.**
- **Transformacji danych – czyli przekształcenia danych do innej, bardziej oszczędnej lub mniej wymiarowej postaci, która dalej pozwala na ich poprawne rozróżnianie i przetwarzanie, np.:**
  - **metoda analizy głównych składowych (PCA – Principal Component Analysis)**
  - **metoda analizy składowych niezależnych (ICA – Independent Component Analysis)**
- **Agregacji i Asocjacji danych (Aggregate & Associate) – czyli takiej reprezentacji danych, która polega na zagregowaniu reprezentacji takich samych i/lub podobnych danych i ich grup oraz ich odpowiednim do rozwiązywanego zadania powiązaniu w celu przyspieszenia ich przeszukiwania i przetwarzania.**

WIZUALIZACJA I PREZENTACJA
VISUALIZATION & PRESENTATION

to zadania związane z graficzną reprezentacją danych w takiej postaci, żeby zaprezentować dane w taki sposób, aby możliwe było:

- **porównanie liczności danych określonego typu/grupy/zbioru/klasy,**
- **wskazanie zależności (relacji) pomiędzy danymi i ich grupami,**
- **wskazanie minimów, maksimów, średnich, odchyleń i wariancji danych,**
- **wskazanie rozkładów, agregacji, środków ciężkości,**
- **wskazanie podobieństw i różnic pomiędzy danymi i ich grupami,**
- **wskazanie reprezentantów, typowych i nietypowych danych,**
- **wskazanie wzorców lub wartości odstających od przeciętnych (outlier), błędnych, brakujących lub szczególnych,**
- **podział, odfiltrowanie lub selekcja pewnej grupy wzorców,**
- **oceny pokrycia przestrzeni danych i ich reprezentatywności dla zadania,**
- **oceny jakości, zaszumienia, poprawności, dokładności i pełności danych.**

GŁÓWNE ETAPY EKSPLORACJI DANYCH

1. Zrozumienie zadania i zdefiniowanie celu praktycznego eksploracji, czyli przyporządkowanie zadania do grupy: klasyfikacji, grupowania, predykcji lub asocjacji.
2. Przygotowanie bazy danych do analizy poprzez wyselekcjonowanie rekordów z baz danych najlepiej charakteryzujących rozważany problem.
3. Czyszczenie i wstępna transformacja danych poprzez ich normalizację, standaryzację, usuwanie danych odstających, usuwanie lub uzupełnianie niekompletnych wzorców.
4. Transformacja danych z postaci symbolicznej na postać numeryczną poprzez przypisanie im wartości lub rozmywanie (fuzzification) w zależności od stosowanej metody ich dalszego przetwarzania.
5. Redukcja wymiaru danych i selekcja najbardziej znaczących i dyskryminujących cech pozwalających uzyskać najlepsze zdolności uogólniające projektowanego systemu.
6. Wybór techniki i metody eksploracji danych na podstawie możliwości danej metody oraz rodzaju i liczności danych: numeryczne, symboliczne, sekwencyjne...
7. Wybór algorytmu lub aplikacji implementującej wybraną technikę eksploracji danych oraz określenie optymalnych parametrów adaptacji/uczenia wybranej metody (przydatne mogą tutaj być metody ewolucyjne, genetyczne, walidacja krzyżowa).
8. Przeprowadzenie procesu konstrukcji, adaptacji lub uczenia wybraną metodą.
9. Eksploatacja systemu: wnioskowanie, określanie grup, podobieństw, różnic, zależności, następstwa, implikacji.
10. Douczanie systemu na nowych danych lub utrwalanie zebranych wniosków z eksploracji.

PODSTAWOWE POJĘCIA I TERMINOLOGIA

Asocjacja – to proces stowarzyszenia ze sobą dwu lub więcej obserwacji. W najprostszej postaci opisywana jest często przez reguły asocjacyjne. Asocjacje są również postawą działania ludzkiego mózgu, pamięci i inteligencji, więc mogą być reprezentowane przez skomplikowane sieci neuronowe.

Atrybut – to jedna z cech (parametrów) opisujących obiekt za pośrednictwem wartości reprezentujących ten atrybut. Wartości te są określonego typu i mogą posiadać wartości z pewnego zakresu lub zbioru.

Cecha diagnostyczna – deskryptor numeryczny opisujący i charakteryzujący analizowany proces, zwany również atrybutem procesu.

Ekstrakcja cech diagnostycznych – to proces tworzenia atrybutów wejściowych dla modelu eksploracji na podstawie wyników pomiarowych. Proces ten nazywany jest również generacją cech. Proces ten może być powiązany z normalizacją, standaryzacją lub inną transformacją danych, mających na celu uwydatnienie głównych cech modelowanego procesu, które mają istotny wpływ na budowę modelu oraz uzyskiwane wyniki i uogólnienie.

Generalizacja – to zdolność lub właściwość modelu eksploracji danych polagająca na możliwości poprawnego działania (np. przewidywania, klasyfikacji, regresji) modelu na innych danych niż dane uczące.

PODSTAWOWE POJĘCIA I TERMINOLOGIA

Grupowanie (klasteryzacja) – to proces wyszukiwania obiektów zdefiniowanych przy pomocy danych podobnych do siebie.

Klasyfikacja – to proces przyporządkowywania obiektów do określonych klas na podstawie podobieństwa lub innych procesów skojarzeniowych albo w wyniku regresji na podstawie analizy i przetwarzania danych wejściowych, której wynikiem jest wartość odpowiadającą klasie lub stopniu podobieństwa do niej.

Model – to zwykle algorytm lub wzór matematyczny połączony z pewną strukturą lub sposobem reprezentacji przetworzonych danych źródłowych, określany w trakcie procesu uczenia, adaptacji lub konstrukcji.

Obserwacja – to zestaw pomiarów tworzących jeden rekord danych (krotkę).

Predykcja – to wynik procesu regresji lub kojarzenia, w którym otrzymujemy odpowiedź w postaci liczbowej lub innego obiektu.

Redukcja – to proces kompresji stratnej polegający na zmniejszeniu wymiaru wektorów lub macierzy obserwacji poprzez eliminację mało reprezentatywnych lub niekompletnych atrybutów albo w wyniku określania pochodnych reprezentatywnych cech (np. PCA, ICA).

PODSTAWOWE POJĘCIA I TERMINOLOGIA

Adaptacja – to polegający na przedstawieniu danych uczących oraz dobraniu, dopasowaniu lub obliczeniu wartości modelu tak, aby dostosował swoje działanie do określonego zbioru, typu i ew. pożądanych wartości wyjściowych danych uczących.

Uczenie – to proces iteracyjny polegający na wielokrotnym przedstawianiu danych uczących oraz poprawianiu wartości modelu tak, aby dostosował swoje działanie do określonego zbioru, typu i ew. pożądanych wartości wyjściowych danych uczących. Uczenie może być np.: nienadzorowane (bez nauczyciela, *unsupervised*), nadzorowane (z nauczycielem, *supervised*), przez wzmacnianie (*reinforcement*), konkurencyjne (*competitive*), motywowane, Bayesowskie, asocjacyjne…

Testowanie – to proces sprawdzania jakości modelu przeprowadzanym w trakcie procesu uczenia lub adaptacji modelu na zbiorze danych chwilowo wydzielonych i wykluczonych z procesu uczenia (tzw. walidacja np. krzyżowa – *n-fold cross validation*) lub na zbiorze danych testowych całkowicie wykluczonych z procesu uczenia/adaptacji modelu (testowanie właściwe).

Wzorzec – to zestaw lub sekwencja albo inna struktura danych reprezentowanych w postaci wektora, macierzy, sekwencji albo grafu danych stosowana do budowy, adaptacji, uczenia, walidacji i testowania modelu.