Szybki rozwój technologii informacyjnej spowodował masowy przyrost ilości danych cyfrowych. Są to różnorodne dane pochodzące z wielu źródeł, takich jak: sieć web, poczta elektroniczna, serwisy i sieci społecznościowe (Google+, Facebook, Twitter, fora internetowe, blogi, itp.), sieci mobilne, media, systemy transakcyjne, biblioteki cyfrowe, sieci sensoryczne, itp. Szacuje się, że do 2020 roku dane cyfrowe w Internecie przekroczą ilość 40 zettabajtów. Gdy dane charakteryzują się dużą objętością, różnorodnością, złożonością, zmiennością i szybkim przepływem w czasie, określane są mianem danych typu BD (ang. Big Data).

Gromadzenie i przechowywanie tak ogromnych ilości danych nie jest obecnie zadaniem bardzo trudnym, ponieważ umożliwiają to rozproszone bazy danych i chmury obliczeniowe. Jednak największym wyzwaniem technicznym w analizie i przetwarzaniu danych typu BD jest ich przeszukiwanie, ekstrakcja pożądanych informacji, wizualizacja oraz przetwarzanie w czasie rzeczywistym. Powszechne metody analizy danych oraz techniki informatyczne nie są w stanie przetworzyć tak ogromnych ilości danych w rozsądnym czasie. Jest więc potrzeba tworzenia nowych rozwiązań technologicznych, które będą w stanie sprostać wymaganiom współczesnego świata cyfrowego.

Niniejszy projekt ma na celu znacząco przyczynić się do opracowania zaawansowanych narzędzi do analizy i przetwarzania danych typu BD o specyficznych własnościach. Są to tzw. dane nieujemne, które stanowią pokaźną część wszystkich danych cyfrowych. Zaliczamy do nich np. zbiory obrazów i filmów, sygnały spektralne, dokumenty tekstowe, różne dane biomedyczne, itp. Do przeszukiwania i wydobywania pożądanych informacji z tak ogromnych danych, konieczna staje się redukcja ich objętości, a także ekstrakcja cech znaczeniowych. Do realizacji takich zadań przyjmuje się, że dane mogą być reprezentowane obiektami algebraicznymi, takimi jak wektory, macierze, tensory. Przykładowo, każdy obraz może być wyrażony przez macierz o nieujemnych elementach. Podobnie, dokument tekstowy można reprezentować wektorem częstości występowania słów z danego słownika. W ujęciu geometrycznym, każdy wektor w przestrzeni danych generuje punkt. Zbiór wektorów tworzy zatem przestrzenną chmurę punktów, która dla danych nieujemnych ma kształt wypukłego obiektu geometrycznego. Rozkład punktów w takim obiekcie jest zwykle nieregularny, w którym można zidentyfikować pewne charakterystyczne struktury lub grupy punktów. Rozpoznawanie i modelowanie takich struktur jest przedmiotem wspomnianej ekstrakcji cech. W podejściu algebraicznym, struktury te można opisać modelem matematycznym, reprezentującym analizowane dane. Estymacja parametrów tego modelu nie jest zadaniem łatwym, ale można ją dokonać skalowalnymi algorytmami numerycznymi, które będą badane w niniejszym projekcie. Poszukiwane są algorytmy o szybkiej zbieżności, niskim koszcie obliczeniowym oraz małym zapotrzebowaniem na zasoby pamięci, a przede wszystkim takie, które można łatwo zaimplementować w środowisku obliczeń równoległych i rozproszonych. Eksperymenty będą przeprowadzone na platformach programistycznym przeznaczonych do analizy danych typu BG, np. takich jak Apache Hadoop z modelem MapReduce oraz Apache Spark. Estymowane cechy będą następnie wykorzystywane do wielu innych zadań uczenia maszynowego i sztucznej inteligencji, w tym do grupowania oraz klasyfikacji, np. ekstremalnej (o dużej liczbie klas) lub dynamicznej (z cechami zmieniającymi się w czasie).

W wyniku realizacji projektu powstaną nowe algorytmy, które z pewnością przyczynią się do dalszego rozwoju metod uczenia maszynowego i sztucznej inteligencji. Opracowane narzędzia programistyczne mogą zauważalnie usprawnić współczesne technologie informacyjne, które tak znacząco przyczyniają się do rozwoju cywilizacyjnego.