Działanie wielu współczesnych firm w mniejszym lub większym stopniu opiera się na wykorzystaniu systemów informatycznych. Dotyczy to wielu aspektów działania firmy, zarówno jeśli chodzi o obsługę spraw zarządzania przedsiębiorstwem, spraw administracyjnych, księgowych, jak i przetwarzanie zamówień klientów. Poza użyciem gromadzonych danych na potrzeby bieżącego działania firmy, zebrane dane historyczne mogą zostać wykorzystane również do szeroko zakrojonych analiz w celu optymalizacji jej działania, np. zmiany profilu świadczonych usług przez lepsze dopasowanie się do potrzeb klientów czy obserwacji sytuacji wyjątkowych i niebezpiecznych w celu wczesnego ostrzegania przez potencjalnymi zagrożeniami dla zdrowia i życia personelu. Na szczególną uwagę zasługują systemy, które korzystają z metod uczenia maszynowego i wymagają tzw. wysokiej dostępności jak np. systemy monitorowania i analizy zagrożeń, monitoring procesów produkcyjnych czy systemy rekomendacyjne. W przypadku, gdy od wyników predykcji zależy życie lub zdrowie personelu, należy zwrócić szczególną uwagę zarówno na precyzję analizy jak i na stabilność działania metod sztucznej inteligencji w warunkach stresowych – np. częściowego braku danych – co jest jednym z głównych założeń omawianego projektu.

Eksploracja danych pozwala analitykom odkrywać interesujące zależności w danych, dzięki możliwości sprawnego weryfikowania bieżących hipotez i formułowania nowych, poprzez wykonywanie prostych testów i wykorzystywanie ich wyników w kolejnych etapach eksploracji. Niejednokrotnie najtrudniejszym zadaniem stawianym przed analitykami jest zdefiniowanie takiej reprezentacji obiektów opisywanych w zbiorze danych, która w przyszłości będzie najbardziej pomocna przy tworzeniu tzw. modeli predykcyjnych. Niestety, pomimo tego, że znanych jest wiele metod automatycznej selekcji atrybutów przydatnych przy modelowaniu danych, wciąż brakuje narzędzi pozwalających na wybór podzbioru atrybutów odpowiednio sprofilowanego do badanego problemu, który byłby odporny na częściowe braki danych.

Zagadnienie przetwarzania i eksploracji danych oraz odkrywania z nich wiedzy to dziedziny, które wraz ze wzrostem ilości dostępnych danych, są coraz intensywniej rozwijane. Aby dostrzec w danych zrozumiałe dla człowieka pojęcia, konieczny jest często odpowiedni wybór atrybutów opisujących obiekty. Standardowo, w teorii uczenia maszynowego (ang. machine learning) zadanie to jest wykonywane przy wykorzystaniu algorytmów wyboru atrybutów (ang. attribute / feature selection). Ekstrakcja cech (feature extraction) to proces przetwarzania otrzymanych danych, który prowadzi do uzyskania zbioru atrybutów odpowiednio sprofilowanych do analizowanego problemu. Wyróżniamy dwie fazy tego procesu: pierwsza to konstrukcja nowych cech na podstawie pierwotnych danych (czasem określana jako feature engineering), druga to wybór najistotniejszych spośród uzyskanych w ten sposób cech (feature selection). Można wyróżnić trzy główne podejścia do oceny jakości uzyskanych cech: oparte na filtrach (filter methods), oparte o wyniki analizy danych otrzymane dla ocenianych cech (wrapper methods) oraz zagnieżdżone w technikach uczenia się modeli decyzyjnych na podstawie danych (embedded methods). Na przykład, jedną z dziedzin, w której duży nacisk położony jest na metody optymalnego wyznaczania istotnych podzbiorów atrybutów jest teoria zbiorów przybliżonych (ang. rough set theory).

Ekstrakcja cech pozwala nie tylko uprościć reprezentację danych, ale również uzyskać cechy znacznie łatwiej interpretowane zarówno przez użytkowników jak i systemy uczące się. W literaturze przedstawiono wiele sprawdzonych i skutecznych technik i algorytmów wybierania atrybutów, jednak niewiele z nich zwraca uwagę na potrzebę zachowania własności (np. korelacja z decyzją) otrzymanego zbioru atrybutów w sytuacji gdy część z nich będzie niedostępna. W znakomitej większości metody selekcji atrybutów skupiają się na osiągnięciu możliwie małej reprezentacji danych dobrze opisującej badany problem, nie uwzględniając faktu, że w rzeczywistych warunkach nie wszystkie dane muszą być zawsze dostępne na potrzeby analizy.

Najważniejszym spośród podejmowanych w ramach proponowanego projektu wyzwań jest opracowanie podstaw teoretycznych dla metod selekcji atrybutów, które uwzględniałyby możliwość utraty lub niedostępności części danych (cech obiektów) w procesie wyboru atrybutów i zwiększenie dzięki temu stabilności uzyskiwanych wyników. Innymi celami naukowymi stawianymi przed uczestnikami projektu są: dostosowanie miar jakość atrybutów do charakterystyki omawianego problemu, dostosowanie wybranych algorytmów uczenia maszynowego w celu optymalnego wykorzystania własności uzyskanych zbiorów atrybutów, analiza problemu pod kątem dużych danych oraz empiryczna ewaluacja postawionych tez i opracowanych metod. Wartą uwagi jest kompleksowa analiza wpływu opracowywanych metod selekcji atrybutów na jakość predykcji oraz stabilność działania modeli prognostycznych, w przypadku zajścia zmian w otoczeniu oddziałujących na analizowane dane, określanych z ang. jako concept drift.