

W praktycznych sytuacjach analizy danych często nie wszystkie atrybuty danego obiektu udaje się zmierzyć. Przytoczmy tutaj trzy przypadki z którymi się spotkaliśmy, które zmotywowały nas do stworzenia niniejszego projektu.

1. *Dane medyczne*: Innowacyjne badania lekarskie są bardzo kosztowne, czasochłonne, a w niektórych przypadkach kończą się niepowodzeniem ze względu na zły stan zdrowia pacjenta. W rezultacie większość danych badawczych zawiera wiele pustych komórek, podobnie jak w przypadku angiologicznych baz danych. Dlatego ważne jest, aby określić optymalne testy, które pozwoliłyby rozróżnić pacjentów zdrowych od chorych.

2. *Aktywność związków chemicznych*: Bazy danych zawierają tylko niewielką ilość związków o znanej aktywności. Ponadto, nawet jeśli dla jednego białka jest zmierzona jego aktywność to nie mamy informacji o jego aktywności w odniesieniu do innych bardzo podobnych związków chemicznych. Obie te sytuacje mogą być interpretowane jako braki pewnych atrybutów lub ich niejednoznaczności w zbiorze danych, a ich rozwiązanie może pomóc w stworzeniu nowych leków.

3. *Częściowo zniszczone zdjęcia*: Istniejące metody klasyfikacji obrazu zakładają, że analizowane obrazy są dobrej jakości. Oczywiście, w przypadku niektórych zniszczonych zdjęć (zobacz rysunek 1) problem ten można rozwiązać stosując istniejący algorytm wypełniający brakujące piksele. Jednak takie podejście powieli najbardziej podobne regiony obrazu, co w rezultacie może prowadzić do niedokładnych klasyfikacji (w sytuacji duplikowania kluczowych punktów).



Rysunek 1: Odnowienie starej fotografii.

Szeroko rozumianym celem naszego projektu jest poprawienie zrozumienia oraz możliwości pracy z danymi których część atrybutów jest nieznaną. Zauważmy, że umiejętność radzenia sobie z pierwszym i drugim przypadkiem (dającym prawo do wiarygodnej diagnozy medycznej lub komputerowe wspomaganie tworzenie leków, które mają pożądane efekty) może być istotne z punktu widzenia przyszłych pacjentów.