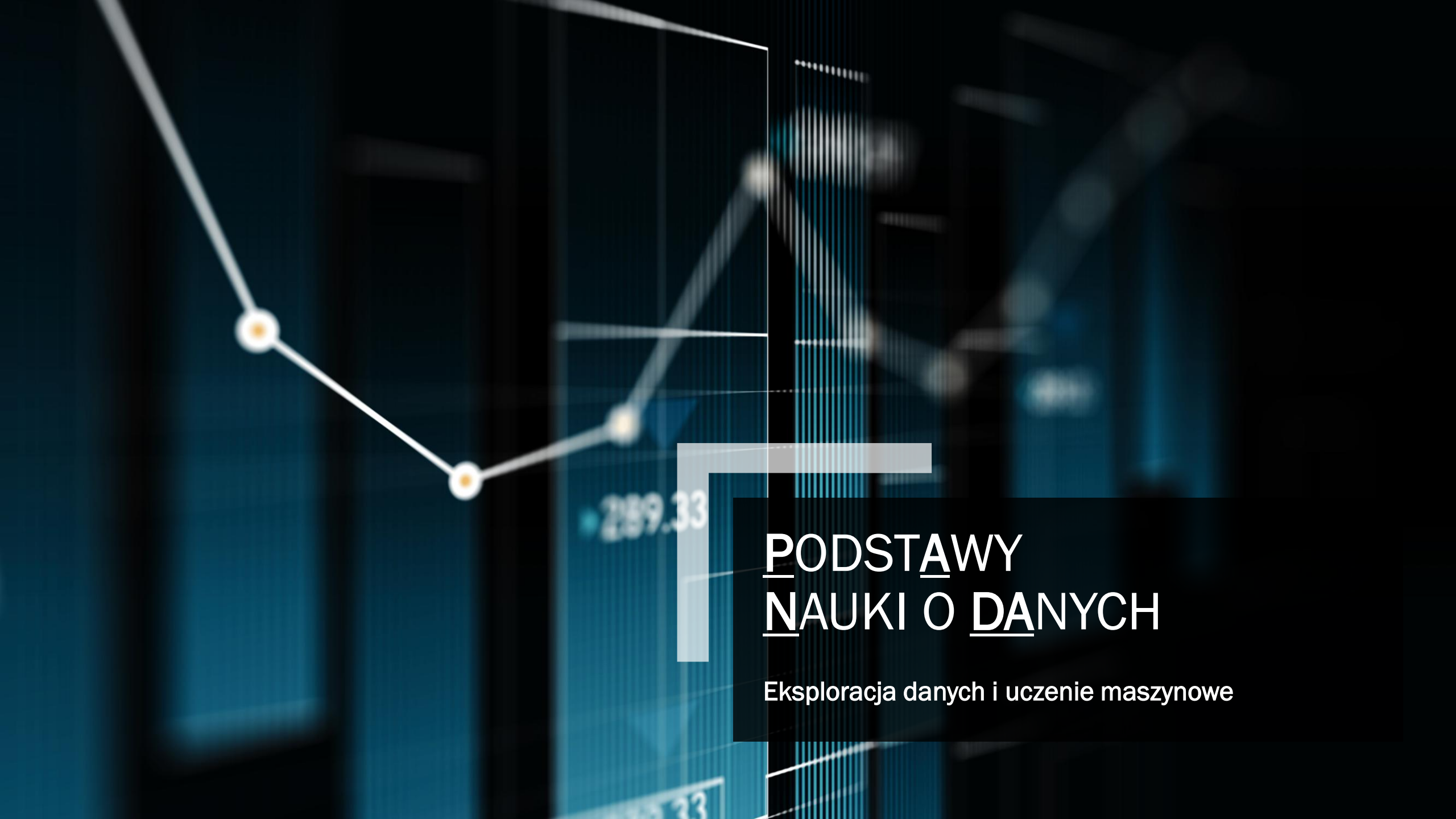




PODSTAWY NAUKI O DANYCH

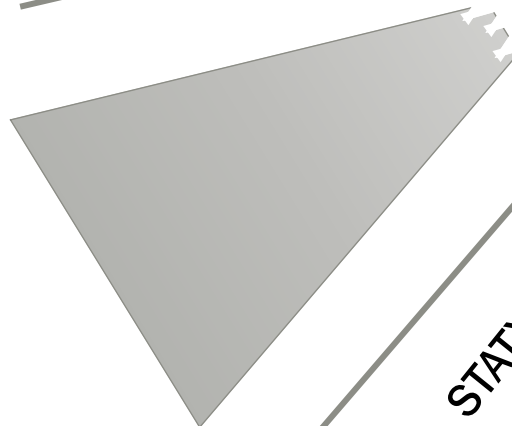
Eksploracja danych i uczenie maszynowe

EKSPLORACJA DANYCH





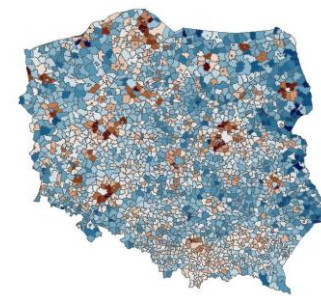
PRAWDOPODOBIENSTWO



STATYSTYKA



próba



2019 North Korean Supreme People's Assembly Election



STATYSTYKA
strategia zbierania danych



cele badawcze
pytania
hipotezy

Dane



EKSPLORACJA
wybór rodzaju modelu

statystyka

dane

- niewielka ilość
- eksperymentalne
- wysokiej jakości
- zbierane czynnie

próba

- losowa
- reprezentatywna

analiza

- pierwotna
- rygorystyczna

eksploracja

dane

- duży wolumen
- obserwacyjne
- wadliwe
- zbierane biernie

próba

- możliwa
- dogodna

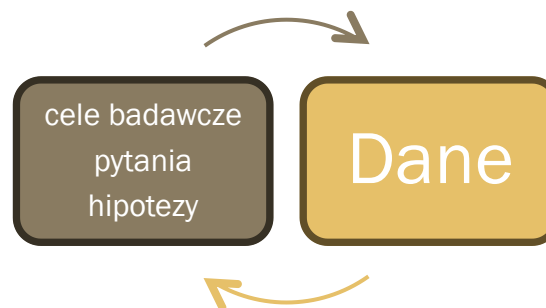
analiza

- wtórna
- swobodna

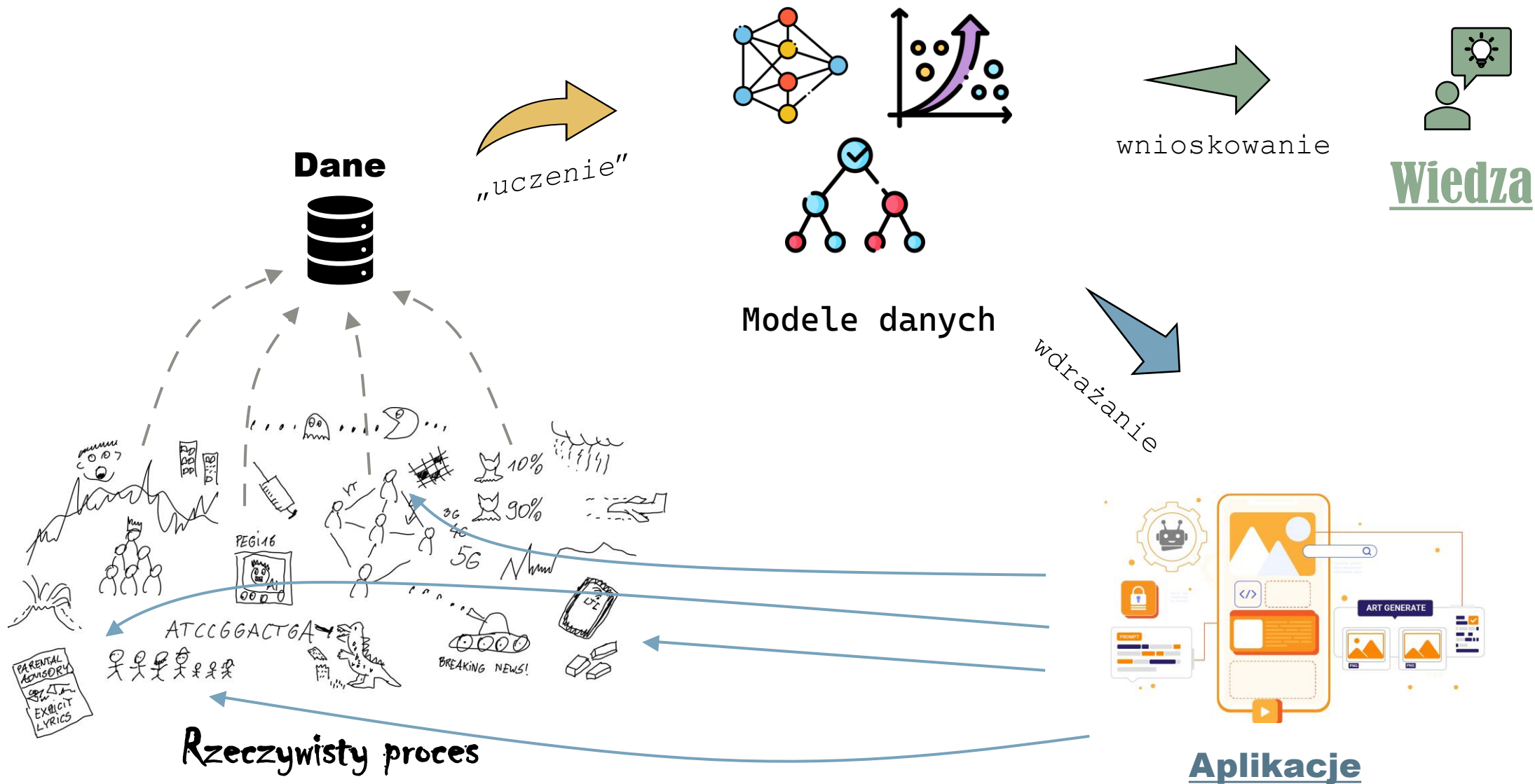
wadliwe dane

- niekompletne
- zaszumione
- odstające
- nieprawdziwe
- niereprezentatywne

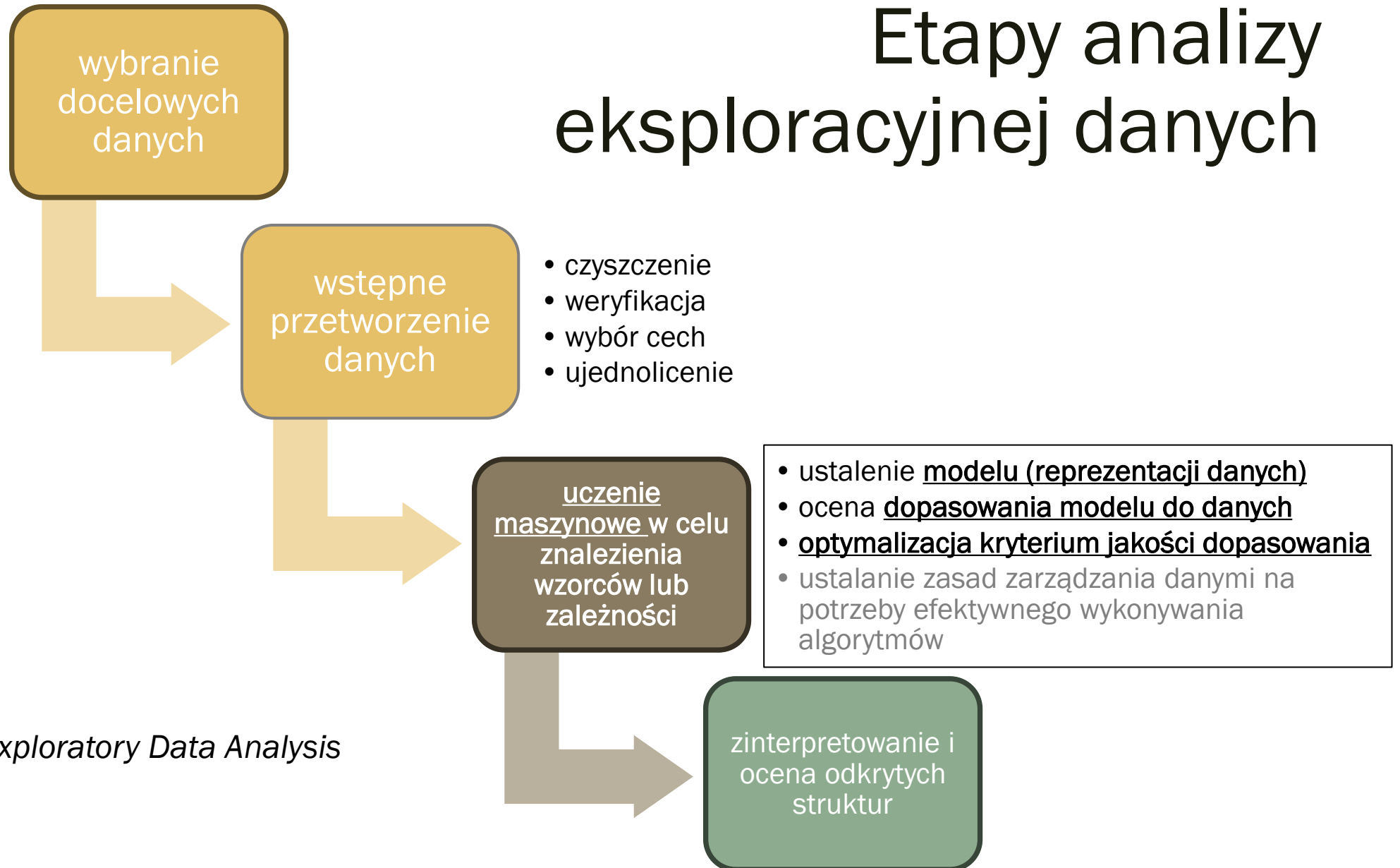
STATYSTYKA
strategia zbierania danych



EKSPLORACJA
wybór rodzaju modelu



Etapy analizy eksploracyjnej danych

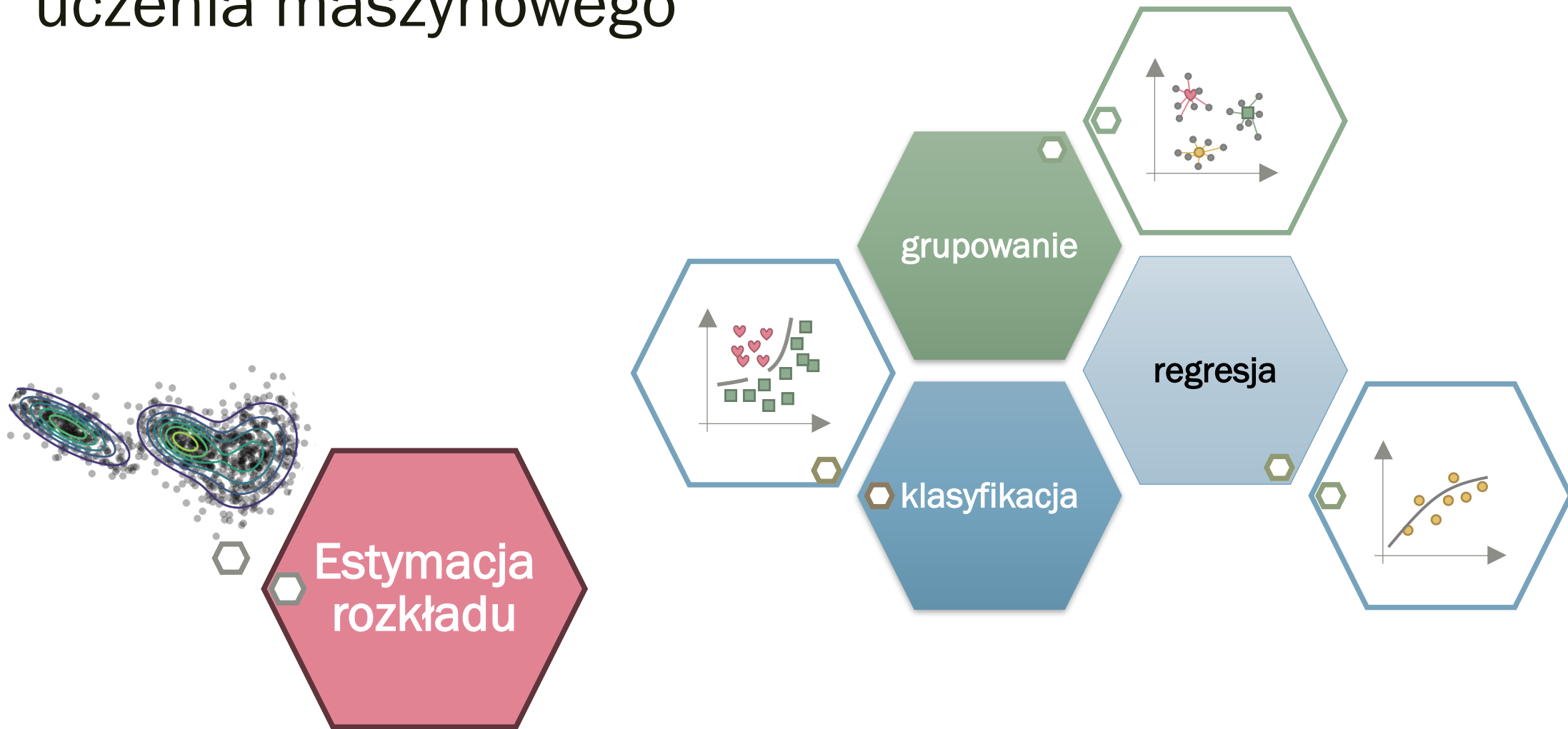


EDA – *Exploratory Data Analysis*

UCZENIE MASZYNOWE



Podstawowe zadania uczenia maszynowego



$$\left\{ (\mathbf{x}_i, y_i) \right\}_{i=1}^N$$

$$\left\{ \mathbf{x}_i \right\}_{i=1}^N$$

uczenie nadzorowane
(ang. *supervised learning*)

etykieta
„prawdziwa
odpowieź”

regresja
wyjście ciągłe

klasyfikacja
wyjście
kategorialne

uczenie
półnadzorowane
(ang. *semisupervised
learning*)

uczenie ze
wzmocnieniem
(ang. *reinforcement
learning*)

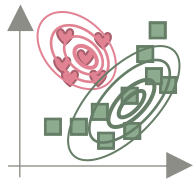
uczenie nienadzorowane
(ang. *unsupervised learning*)

brak etykiet

grupowanie

Uczenie maszynowe

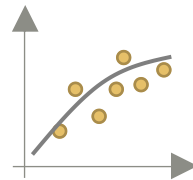
Estymacja rozkładu



- rozkład dedykowany
- estymator jądrowy
- mieszanina rozkładów Gaussa

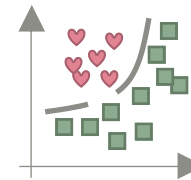
Uczenie nadzorowane

Regresja



- regresja liniowa
- uogólnione modele liniowe
- SVR
- drzewa decyzyjne
- sieci neuronowe

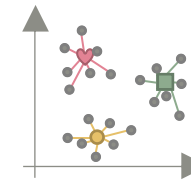
Klasyfikacja



- k-najbliższych sąsiadów
- naiwny Bayes
- regresja logistyczna
- SVM
- drzewa decyzyjne
- lasy losowe
- sieci neuronowe

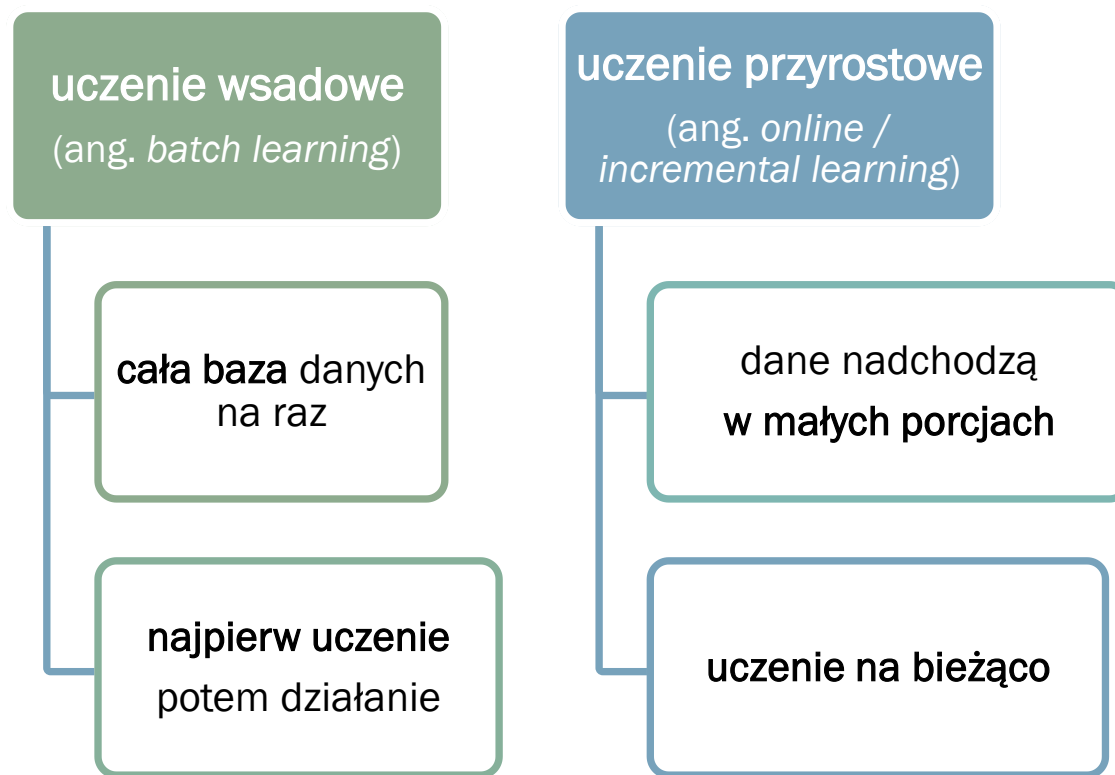
Uczenie nienadzorowane

Grupowanie



- k-średnich
- rozmyte c-średnich
- mieszanina rozkładów Gaussa

Sposoby korzystania z bazy danych



Parametryczne

Ustalona liczba parametrów



- ☐ klasyfikator liniowy
- ☐ regresja liniowa i logistyczna
- ☐ mieszanina rozkładów Gaussa
- ☐ liniowy SVM
- ☐ naiwny klasyfikator Bayesa
- ☐ ...

Nieparametryczne

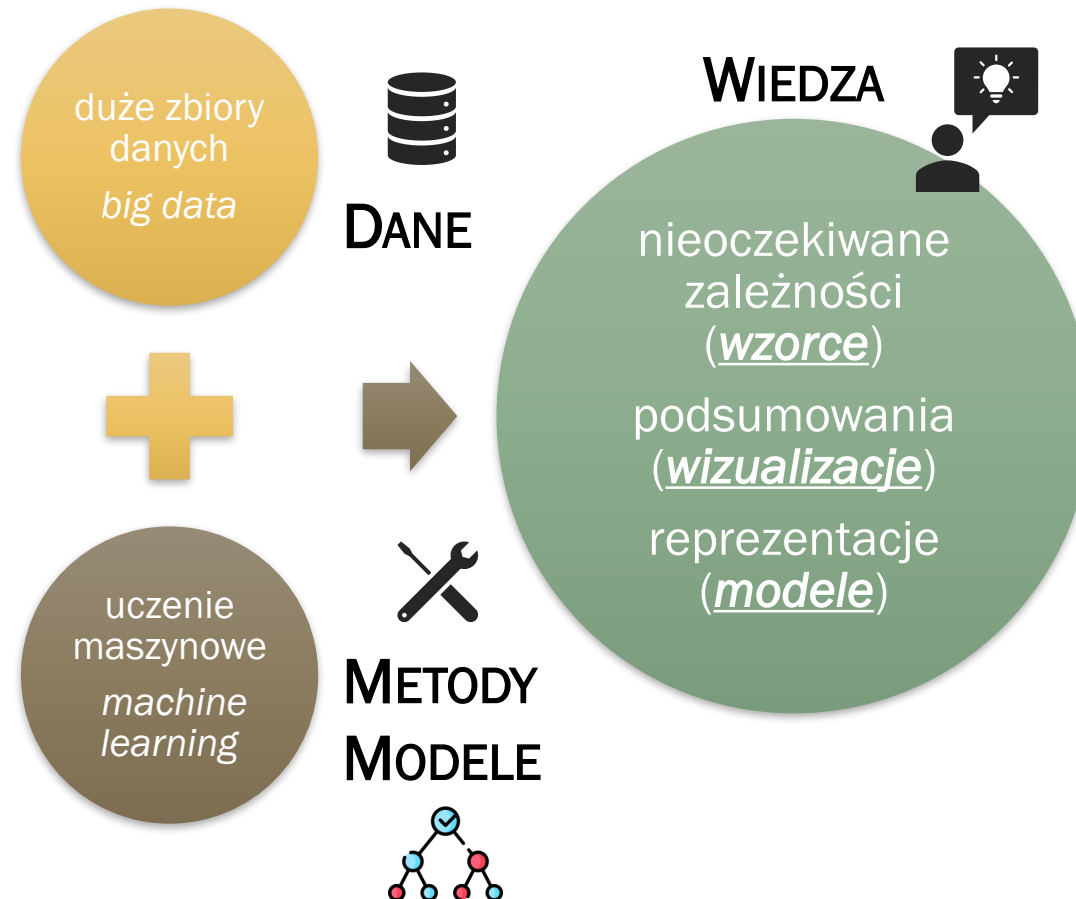
Postać modelu i liczba parametrów zależą od danych



- ☐ k-NN
- ☐ drzewa decyzyjne
- ☐ estymatory jądrowe (KDE)
- ☐ sieci neuronowe
- ☐ ...



Dane, wiedza, modele i predykcje



- Dane gromadzone z różnych przyczyn
- Wydobywanie wiedzy z danych
- Rola modeli

RAMY OGÓLNE

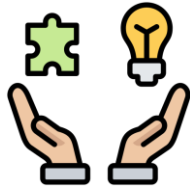


Potok wydobywania wiedzy z danych

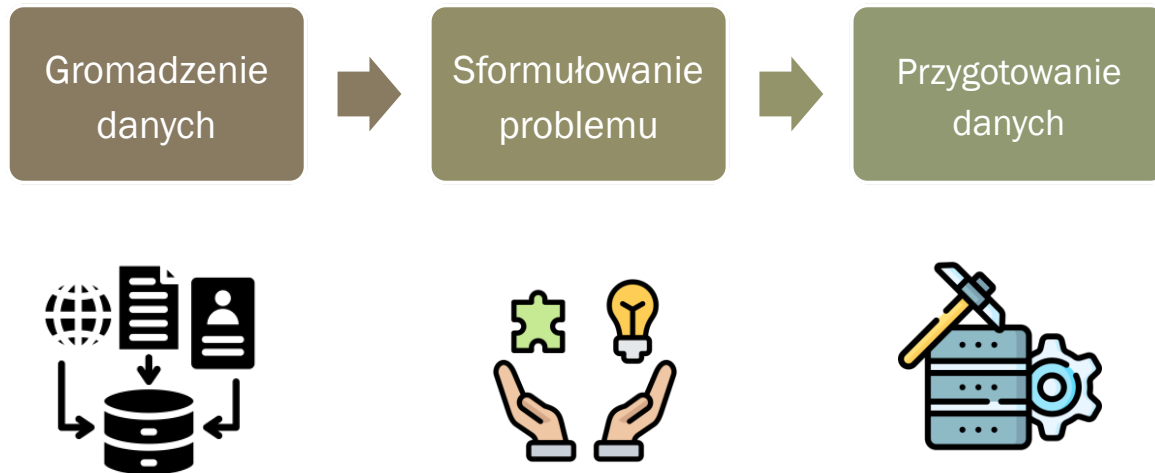
Gromadzenie
danych



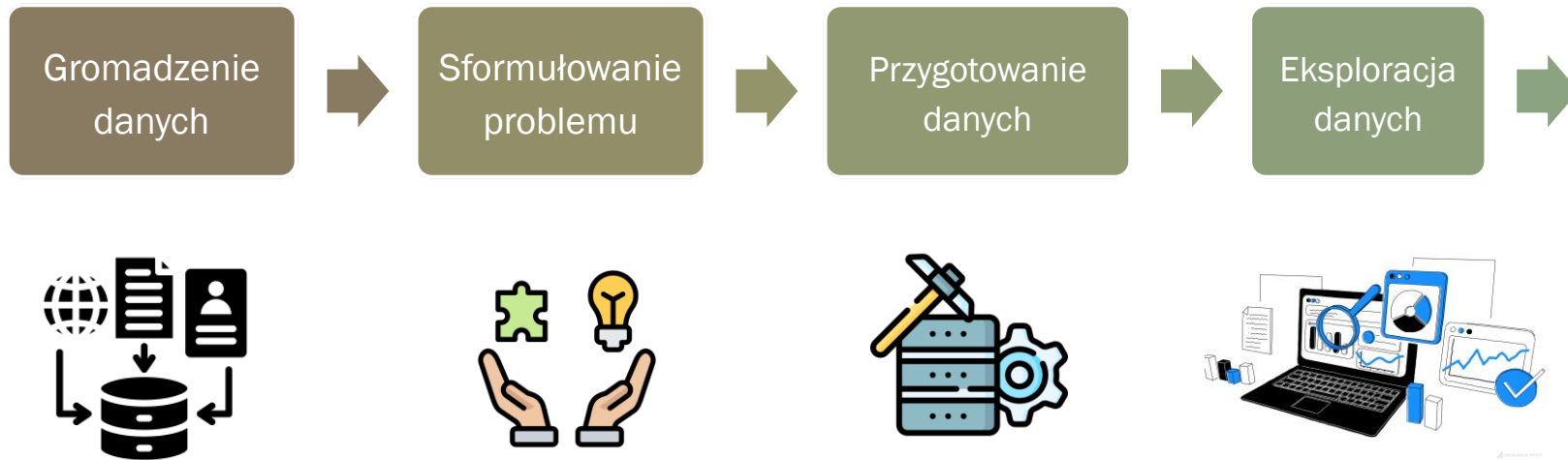
Potok wydobywania wiedzy z danych



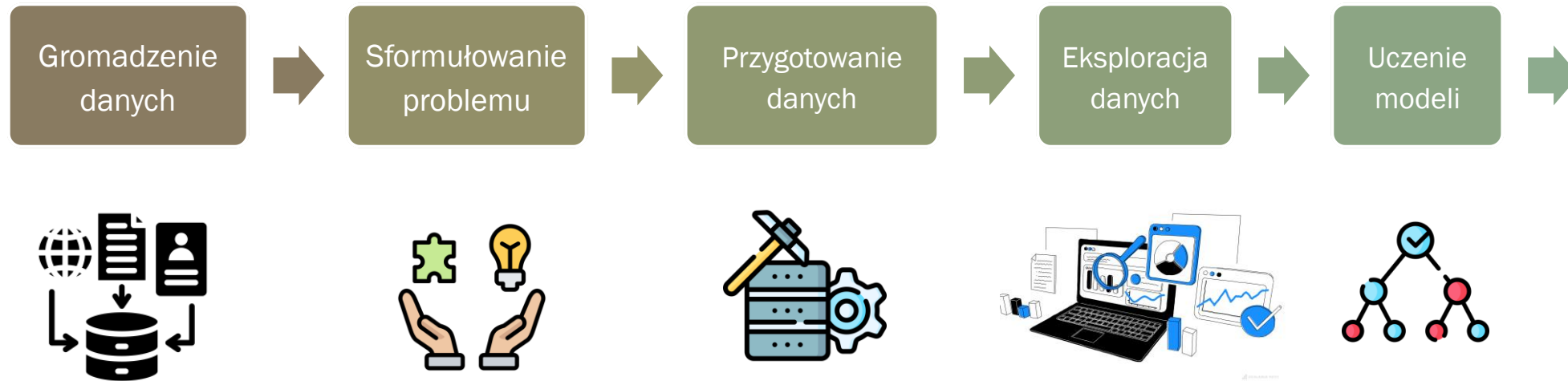
Potok wydobywania wiedzy z danych



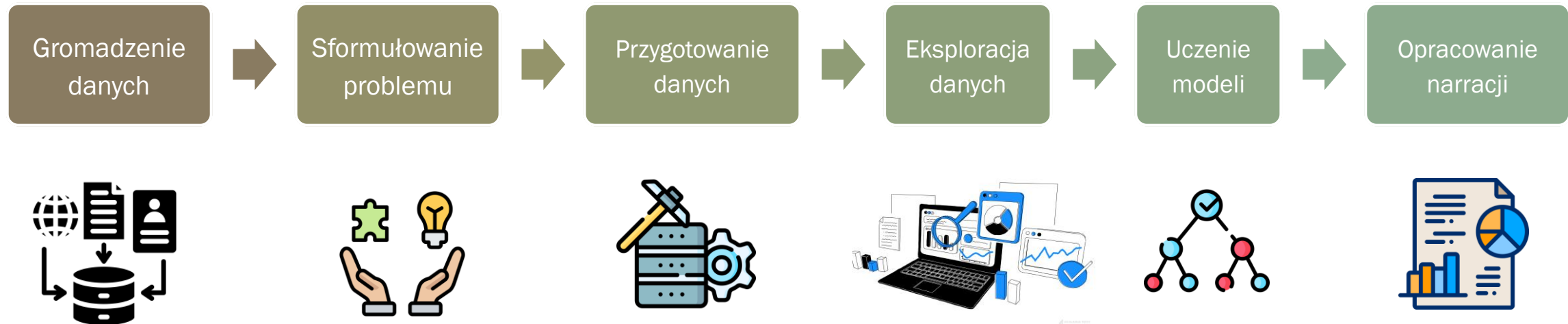
Potok wydobywania wiedzy z danych



Potok wydobywania wiedzy z danych



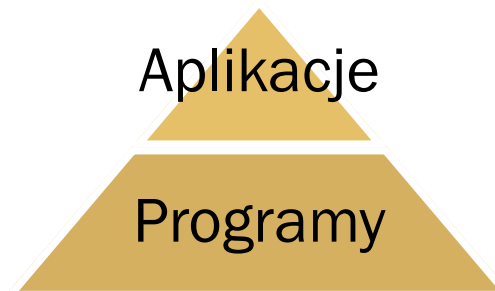
Potok wydobywania wiedzy z danych



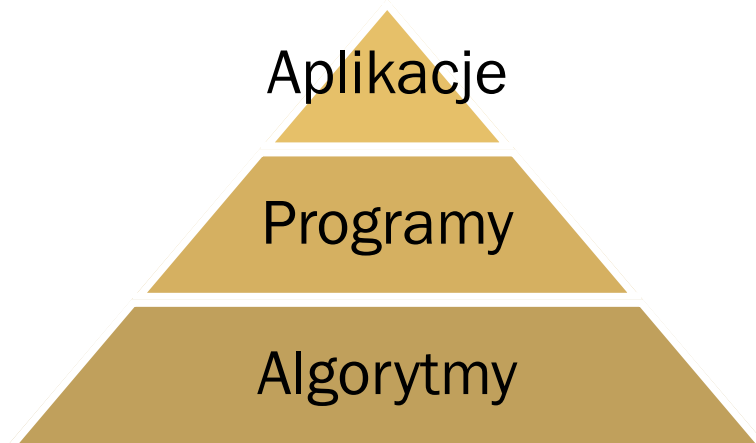
Skąd się biorą aplikacje?



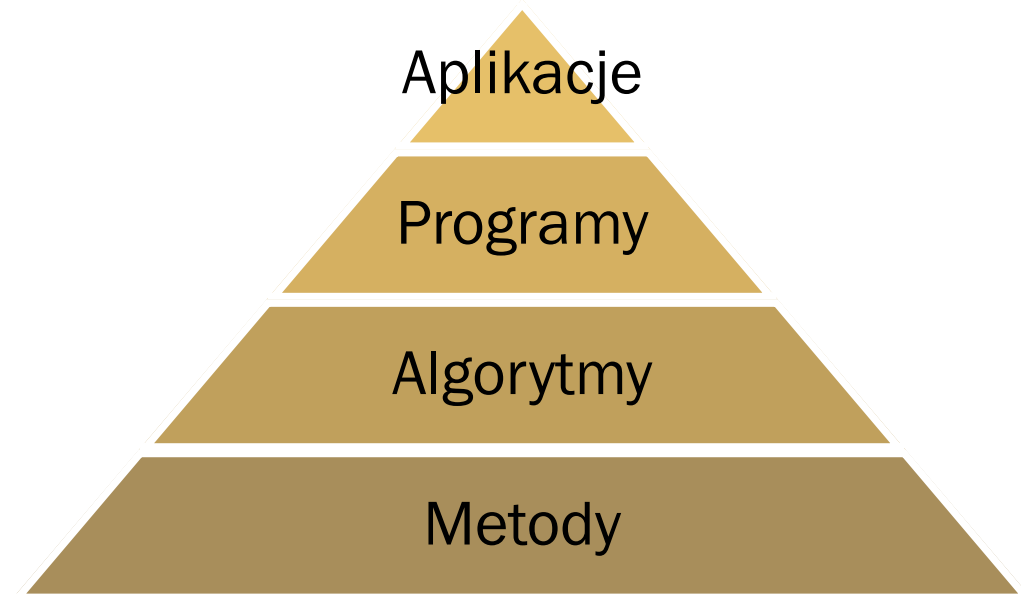
Skąd się biorą aplikacje?



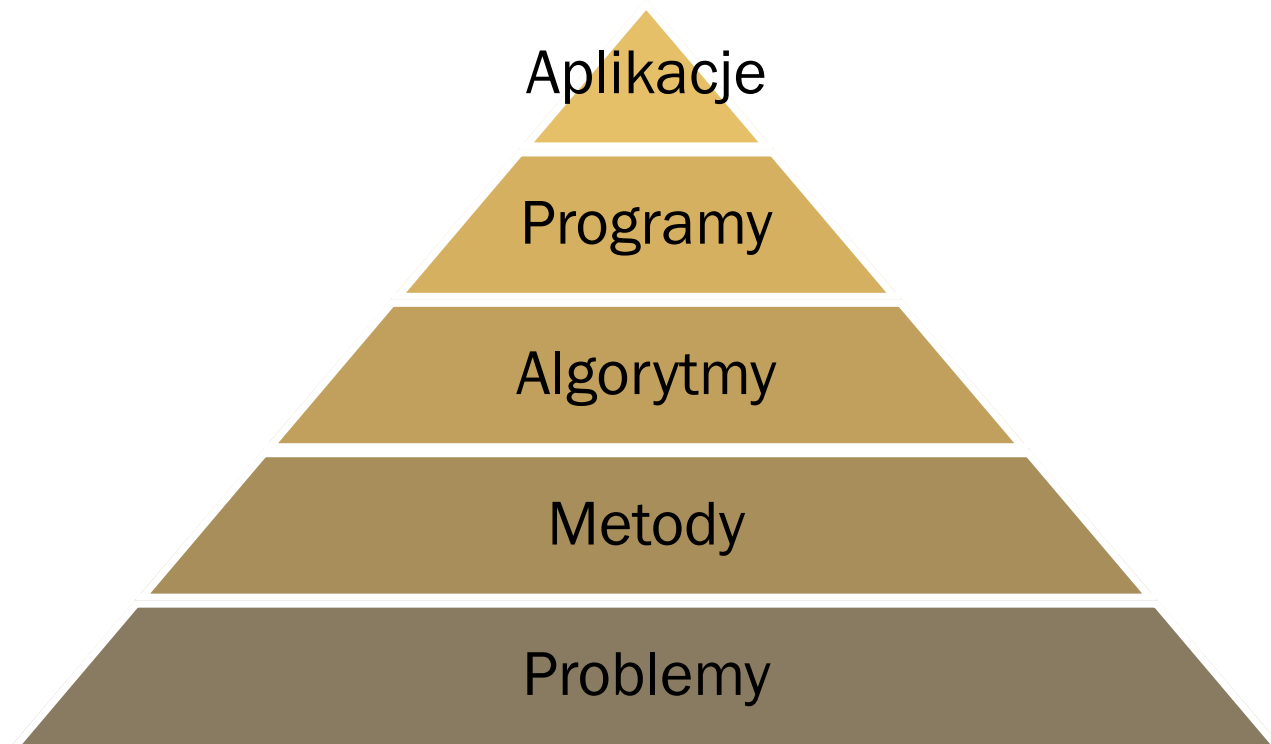
Skąd się biorą aplikacje?



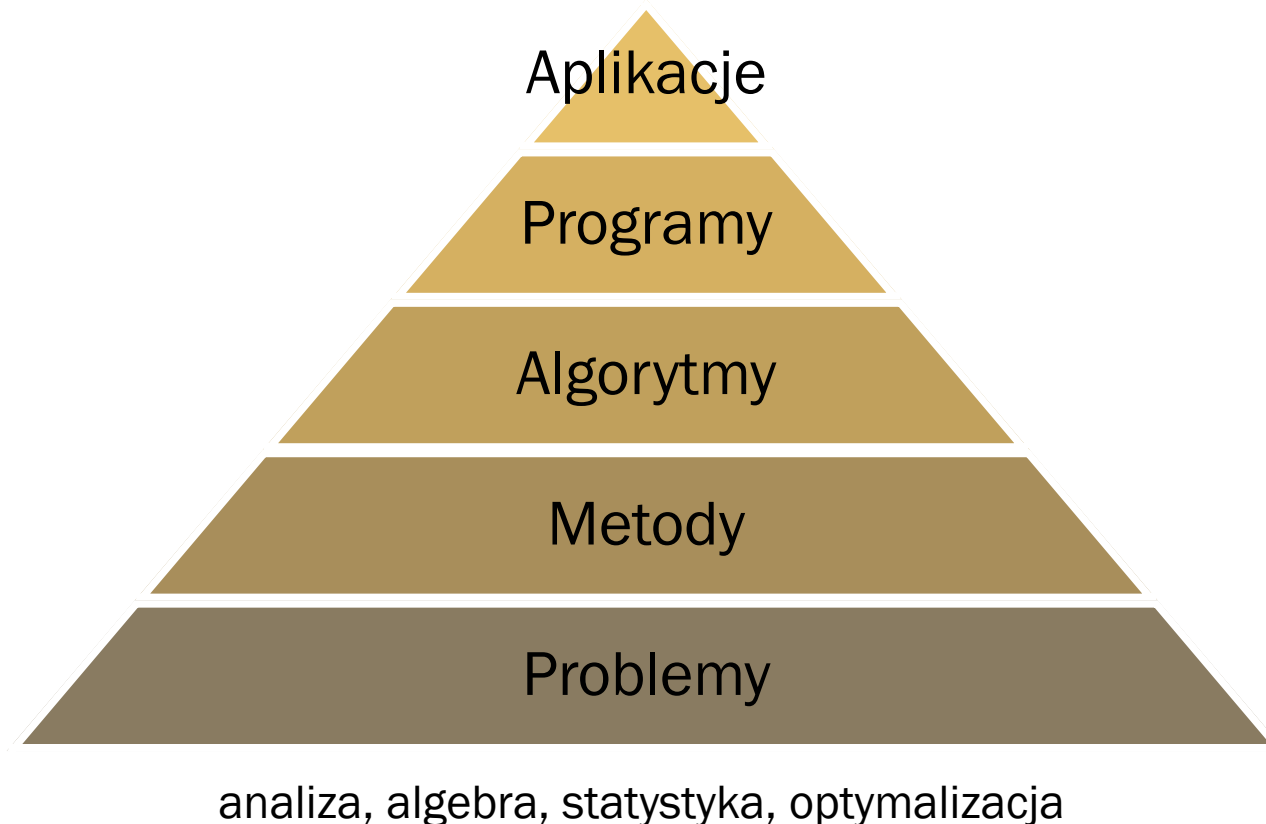
Skąd się biorą aplikacje?



Skąd się biorą aplikacje?



Skąd się biorą aplikacje?





THE END