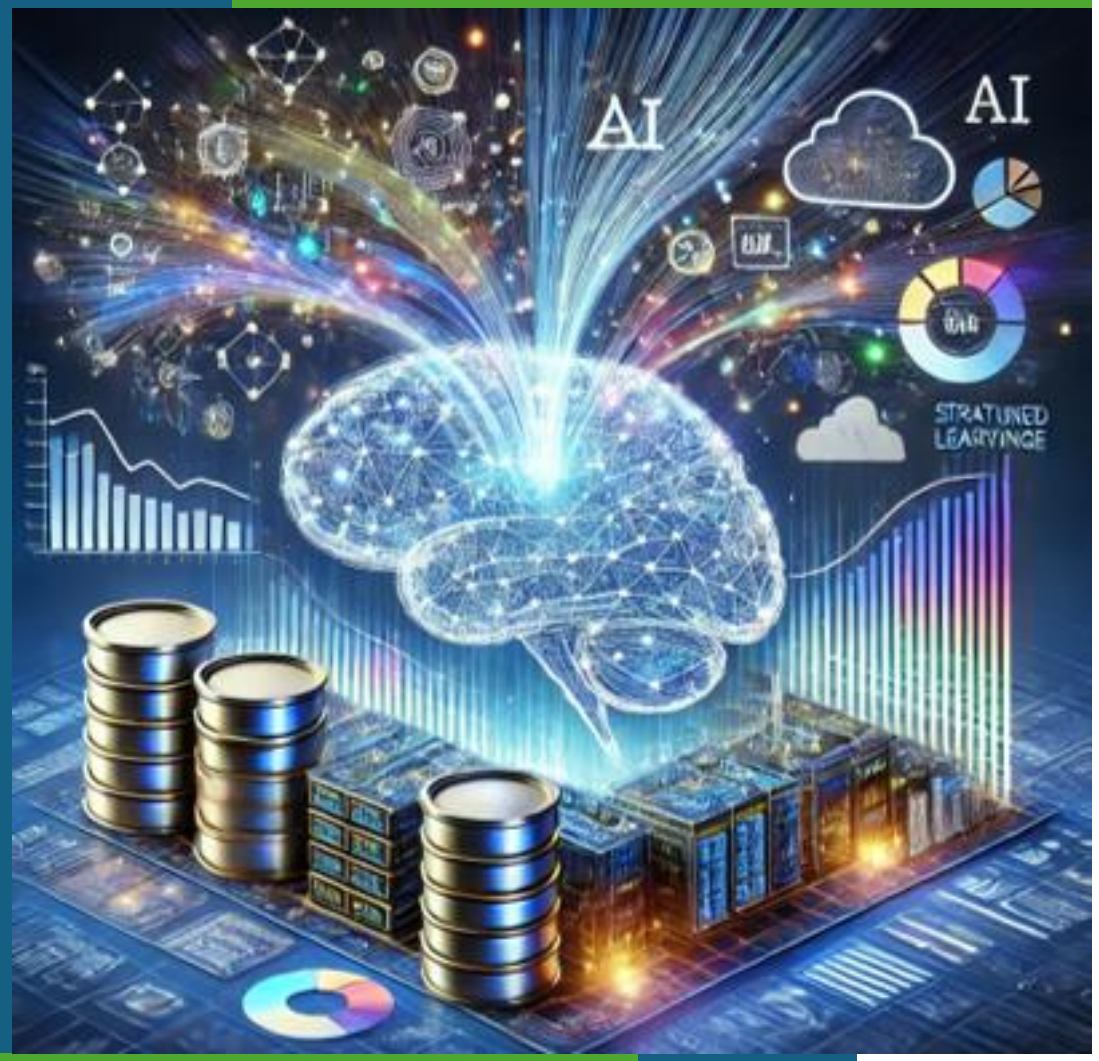
dr inż. Michał Malinowski

https://www.drmalinowski.edu.pl

Temat 6 .

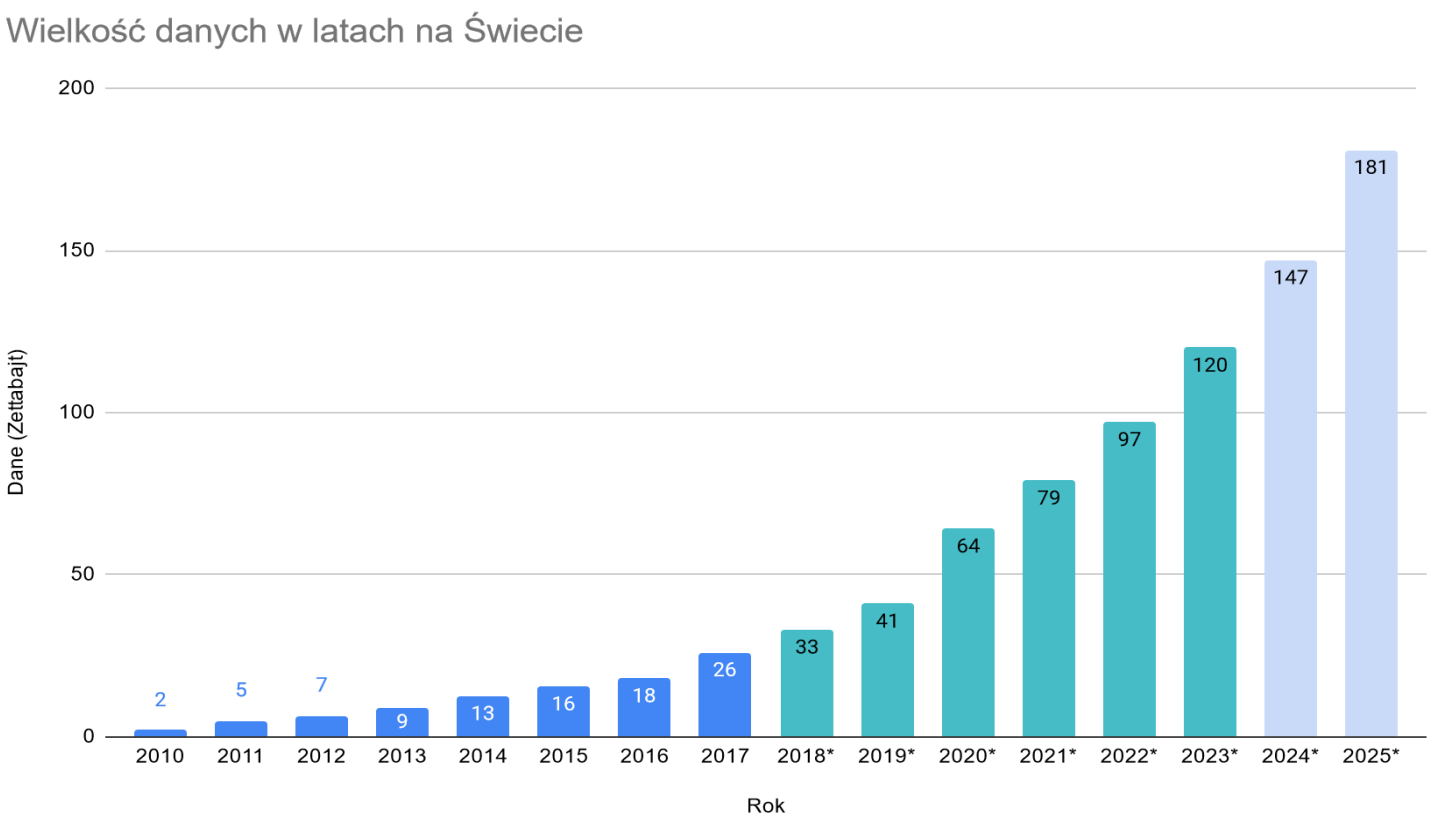
Dane to paliwo – Jakie informacje są kluczowe dla sztucznej inteligencji

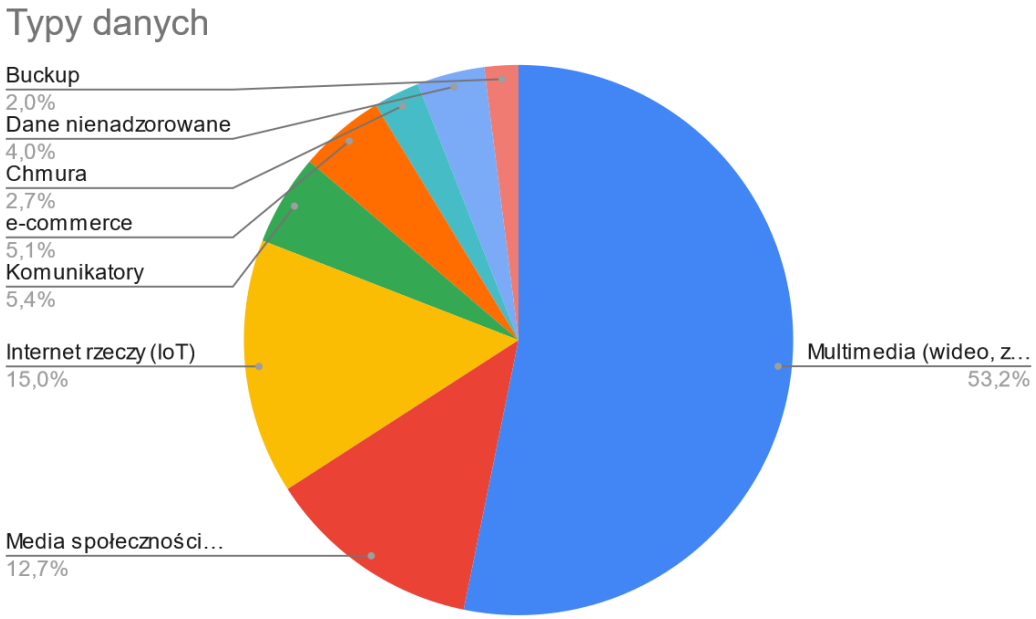
Wprowadzenie do Sztucznej Inteligencji



- Wielkość danych na świecie
- Rodzaje danych wejściowych
- Typy systemów baz danych
- Wyzwania związane z danymi wejściowymi
- Przetwarzanie danych
- Statystyka opisowa
- Procesy ETL i ELT
- Korpus uczący
- Darknet

Wielkość danych na świecie





	Na osobę
2022*	12 TB
2023*	15 TB
2024*	18 TB
2025*	23 TB

Tyle czasu w projekcie AI obejmuje Obrówka danych

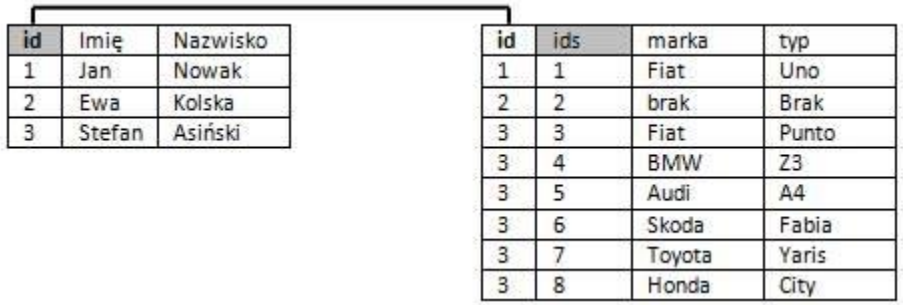
od 50% do 80%



Rodzaje danych wejściowych

Ustrukturyzowane dane

to **dane zorganizowane w tabelach** i bazach danych, które można łatwo przetwarzać przy użyciu SQL lub innych systemów zarządzania danymi.



id	Imię	Nazwisko
1	Jan	Nowak
2	Ewa	Kolska
3	Stefan	Asiński

id	ids	marka	typ
1	1	Fiat	Uno
2	2	brak	Brak
3	3	Fiat	Punto
3	4	BMW	Z3
3	5	Audi	A4
3	6	Skoda	Fabia
3	7	Toyota	Yaris
3	8	Honda	City

✅ Przykład:
✓ **DWH (hurtownie danych)** – magazynowanie dużych ilości informacji.
✓ **DM (Data Mart)** – mniejsze, tematyczne zbiory danych.
✓ **OLTP (Online Transaction Processing)** – systemy przetwarzania transakcji w czasie rzeczywistym.
✓ **ODS (Operational Data Store)** – repozytoria danych operacyjnych.
✓ **CRM (Customer Relationship Management)** – systemy zarządzania relacjami z klientami.
✓ **ERP (Enterprise Resource Planning)** – systemy zarządzania przedsiębiorstwem.

Rodzaje danych wejściowych

Półstrukturyzowane dane

mają **pewną strukturę**, ale nie są przechowywane w tradycyjnych bazach danych. Wymagają dodatkowego przetwarzania, aby mogły być efektywnie analizowane.

✅ Przykład:

✓ CSV (Comma-Separated Values) – pliki tekstowe z danymi tabelarycznymi.
✓ JSON (JavaScript Object Notation) – popularny format danych w API.
✓ XML (Extensible Markup Language) – używany w wymianie danych między systemami.
✓ EXCEL – arkusze kalkulacyjne używane do analizy danych.
✓ log – pliki dzienników systemowych, np. w analizie cyberbezpieczeństwa.
✓ IoT (Internet of Things) – dane z czujników, np. pomiary temperatury.

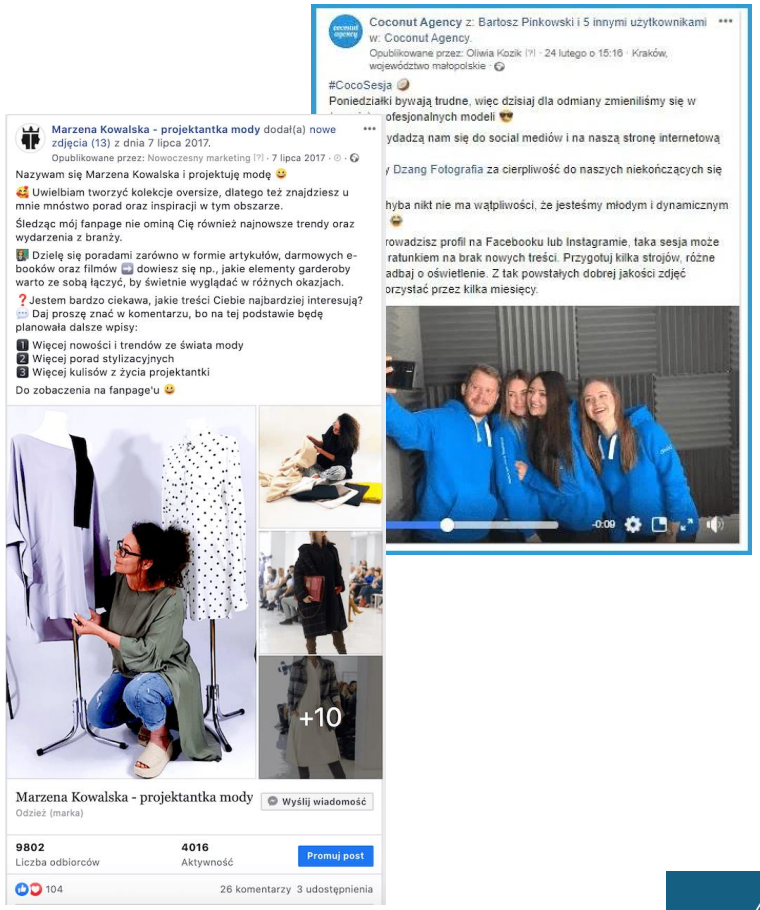

```
{
    "posts": [
        {
            "name": "Web Scraping w Pythonie",
            "url": "http://kamil.kwapisz.pl/web-scraping-python/",
            "author": "Kamil Kwapisz"
        },
        {
            "name": "PEP 8 \u2013 czyli gramatyka dla programisty",
            "url": "http://kamil.kwapisz.pl/gramatyka-dla-programisty/"
            "author": "Kamil Kwapisz"
        }
    ],
    "owner": "Kamil Kwapisz",
    "website": "kamil.kwapisz.pl",
    "lang": "polish"
}
```

```
<?xml version="1.0" encoding="UTF-8"?>
<blog>
    <lang>polish</lang>
    <owner>Kamil Kwapisz</owner>
    <posts>
        <post>
            <author>Kamil Kwapisz</author>
            <name>Web Scraping w Pythonie</name>
            <url>http://kamil.kwapisz.pl/web-scraping-python/</url>
        </post>
        <post>
            <author>Kamil Kwapisz</author>
            <name>PEP 8 – czyli gramatyka dla programisty</name>
            <url>http://kamil.kwapisz.pl/gramatyka-dla-programisty/</url>
        </post>
    </posts>
    <website>kamil.kwapisz.pl</website>
</blog>
```

Rodzaje danych wejściowych

Nieustrukturyzowane dane

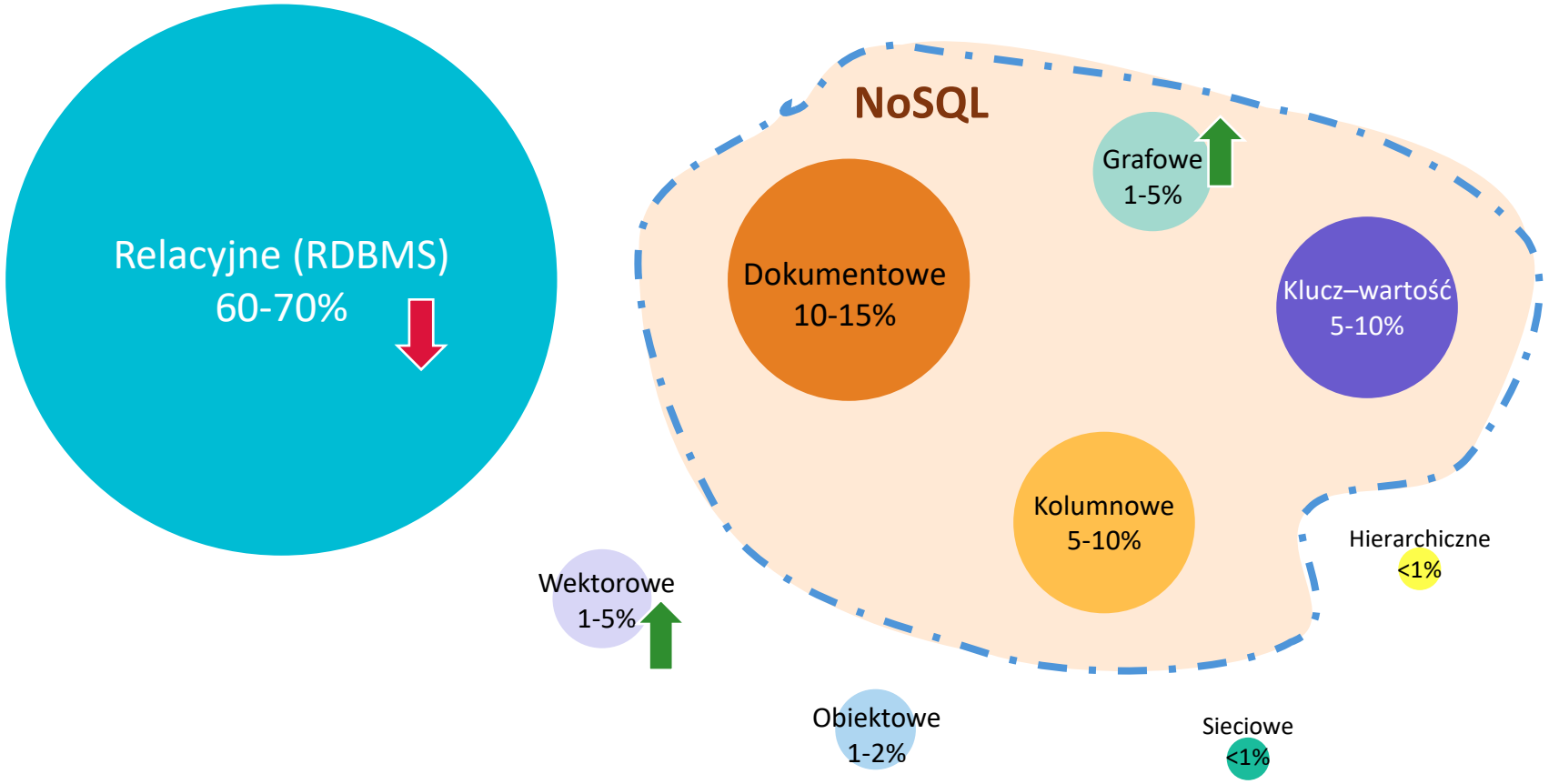
bez określonej struktury, które wymagają zaawansowanego przetwarzania (np. NLP, analiza obrazów).



✅ Przykład:

✓ **Social Media** – posty, komentarze, opinie użytkowników.
✓ **WWW** – strony internetowe, artykuły, dane z wyszukiwarek.
✓ **Word, PDF** – dokumenty tekstowe wykorzystywane w analizie treści.
✓ **JPG** – obrazy, wykorzystywane w rozpoznawaniu obiektów.
✓ **MP4** – pliki wideo, np. do analizy ruchu lub treści audiowizualnych.

Typy systemów baz danych



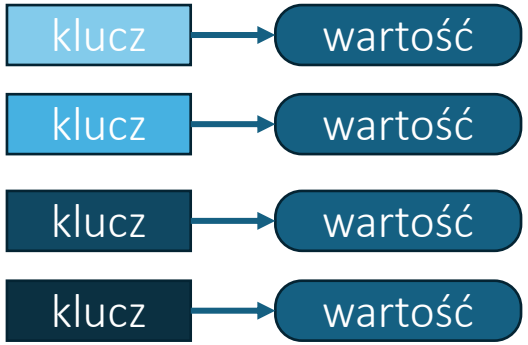
Rodzaje baz NoSQL



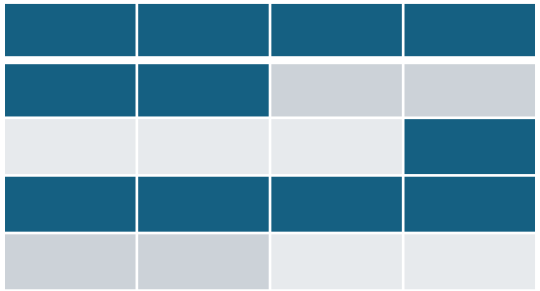
Bazy klucz-wartość
(Key-Value): gdzie wszystko przechowywane jest jako para klucz-wartość.
Przykłady: Redis, Amazon DynamoDB

Bazy dokumentowe
(Document): przechowują informacje jako dokumenty, często w formatach takich jak JSON.
Przykłady: MongoDB, Couchbase.

Bazy par klucz-wartość

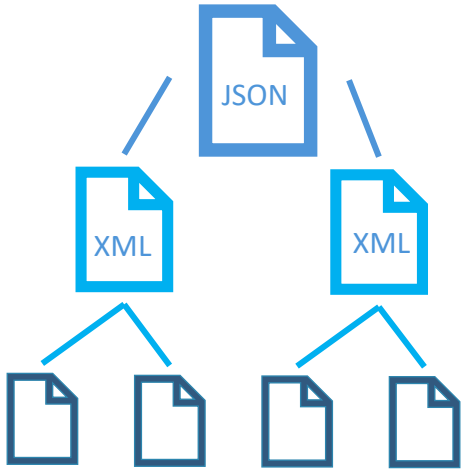


Bazy rodziny kolumn

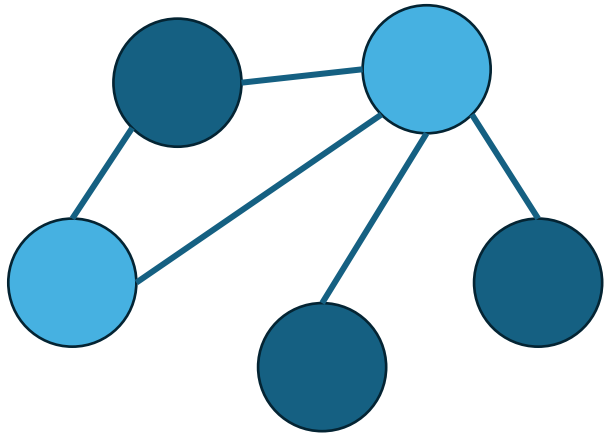


Bazy kolumnowe
(Columnar): dane są przechowywane i wyszukiwane na podstawie kolumn, a nie wierszy.
Przykłady: Apache Cassandra, Google Bigtable.

Bazy dokumentów



Bazy grafowe



Bazy grafowe
(Graph): skoncentrowane na przechowywaniu informacji o węzłach i relacjach między nimi.
Przykład: Neo4j, JanusGraph

Bazy klucz-wartość

Bazy klucz-wartość

to jeden z najprostszych i najbardziej efektywnych modeli przechowywania danych, który polega na skojarzeniu unikalnego klucza z odpowiadającą mu wartością.

Chociaż ich struktura jest wyjątkowo prosta, bazy te stanowią fundament dla wielu zaawansowanych systemów, które obsługują ogromne ilości danych w czasie rzeczywistym.

Przestrzeń: Klient	
osoba:1	Nowak
osoba:2	Kowalski
osoba:3	Malinowski

Przestrzeń: Piłkarz	
osoba:1	Smolarek
osoba:2	Boniek
osoba:3	Lato

Przestrzeń: Polityk	
osoba:1	Sikorski
osoba:2	Kuroń
osoba:3	Mazowiecki

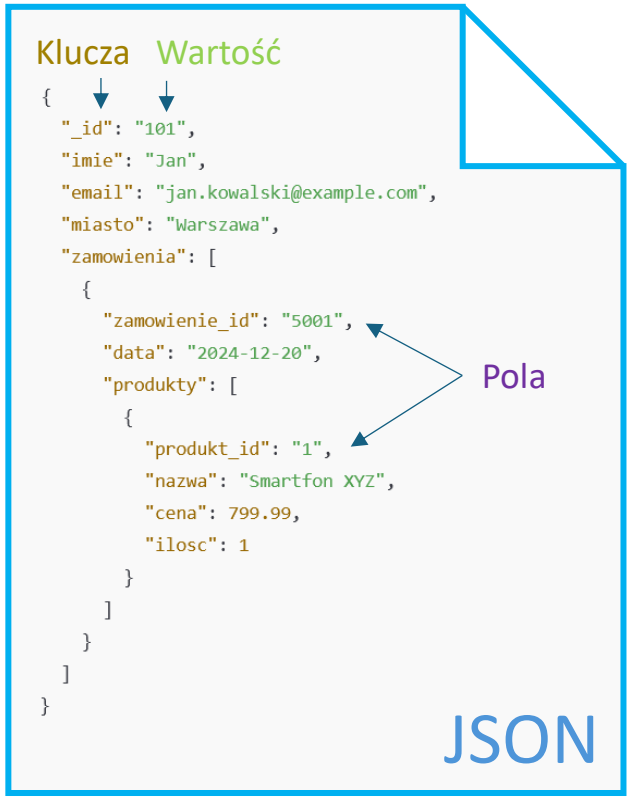


Baza dokumentów (NoSQL)

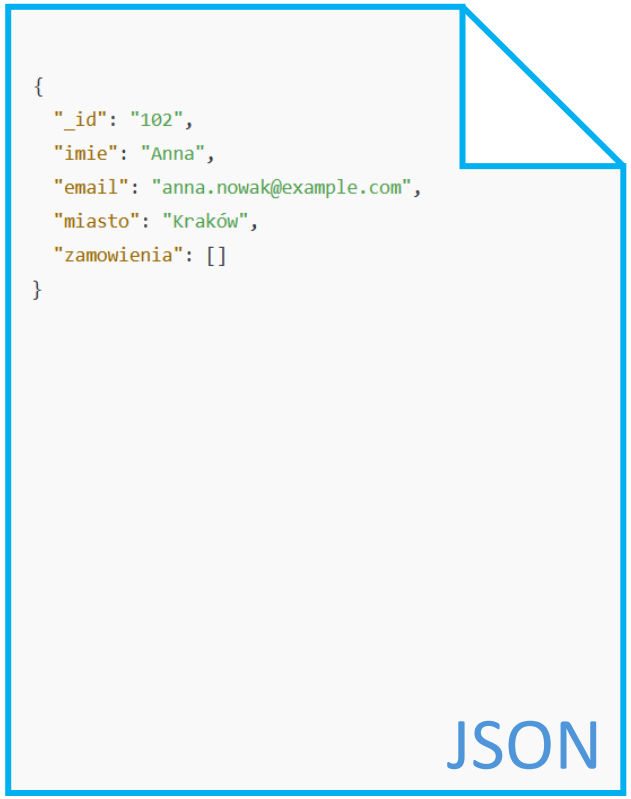
Kolekcja



Dokument



Dokument



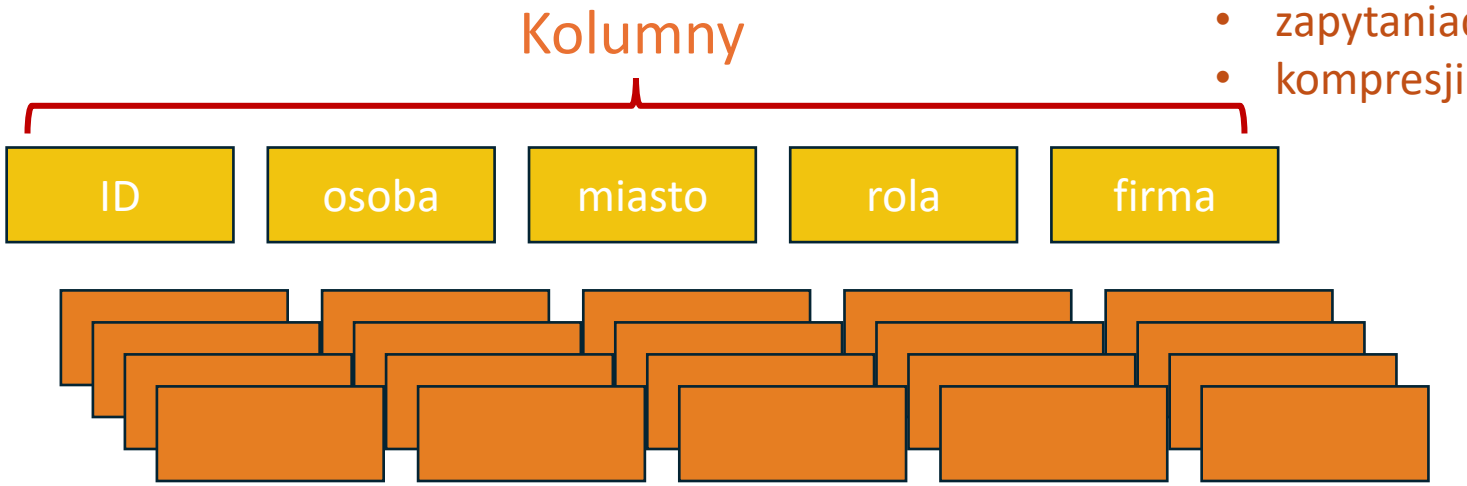
Dokument

Bazy kolumnowe (NoSQL)

Bazy kolumnowe

Doskonale sprawdzają się w:

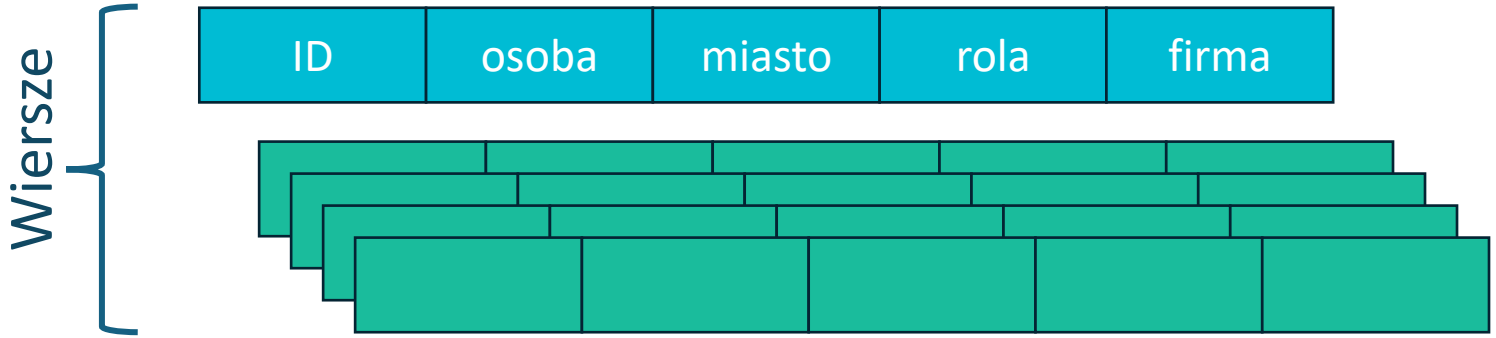
- zapytaniach dotyczących kilku kolumn
- zapytaniach agregujących duże ilości danych
- kompresji danych na poziomie kolumn



Bazy relacyjne

Doskonale sprawdzają się w:

- stopniowym ładowaniu danych
- Przetwarzaniu transakcyjnym
- zapytaniach dotyczących tylko kilku wierszy



Bazy grafowe vs. Bazy relacyjne

Jako kto i gdzie pracuje Kowalski?

Relacyjna baza danych

Osoba

id	nazwisko
1	Nowak
2	Kowalski

Mieszka_w

idosoba	idmiasto	od_roku
1	2	1990
2	1	2011

Firma

id	Nazwa
1	Alfa
2	Beta

Pracuje_w

idosoba	idfirma	jako
1	2	Mechanik
2	1	Księgowy

Miasto

id	nazwa
1	Warszawa
2	Kraków

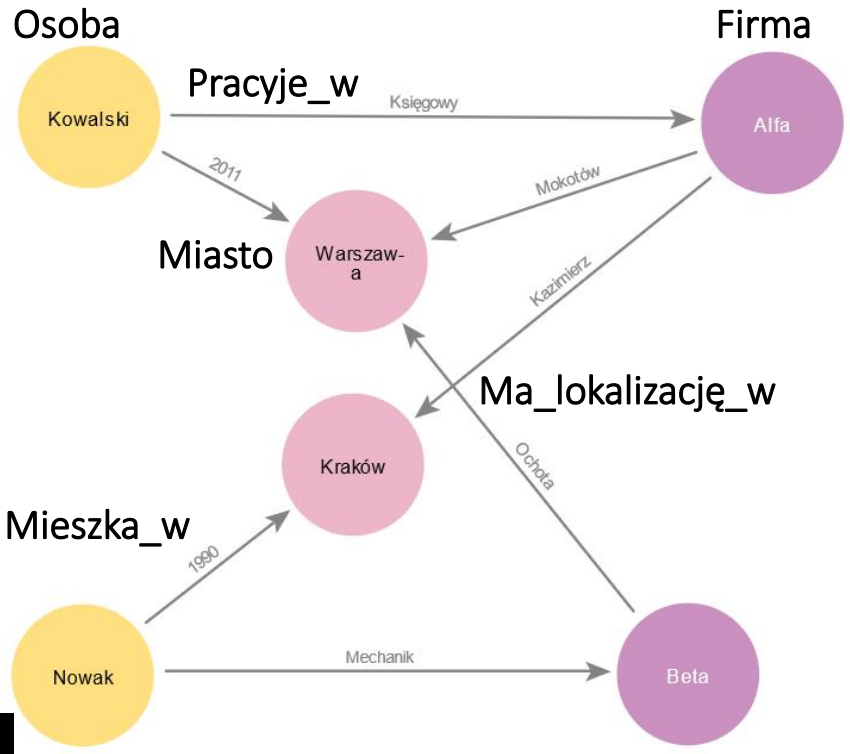
Ma_lokalizacje_w

idfirma	idmiasto	dzielnica
1	1	Mokotów
1	2	Kazimierz
2	1	Ochota

SQL

```
SELECT Firma.nazwa, Pracuje_w.jako FROM Firma
JOIN Pracuje_w ON Pracuje_w.idfirma = Firma.id
JOIN Osoba ON Osoba.id = Pracuje_w.idosoba
WHERE Osoba.nazwisko = "Kowalski"
```

Grafowa baza danych

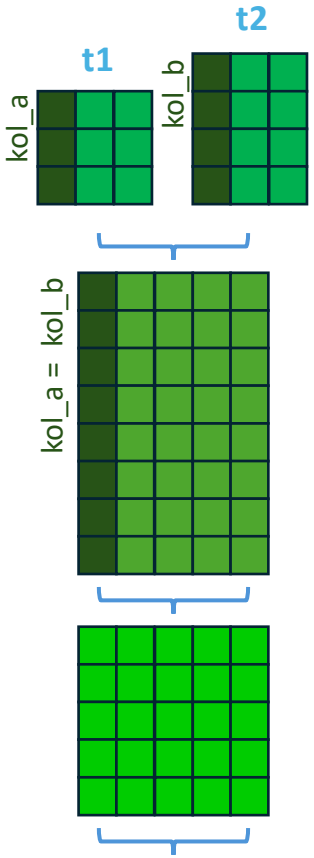


Cypher

```
MATCH (o:Osoba {nazwisko:'Kowalski'})-[p:Pracuje_w]->(f:Firma)
RETURN f.nazwa, p.jako
```

nazwa	jako
Alfa	Księgowy

SQL - SELECT



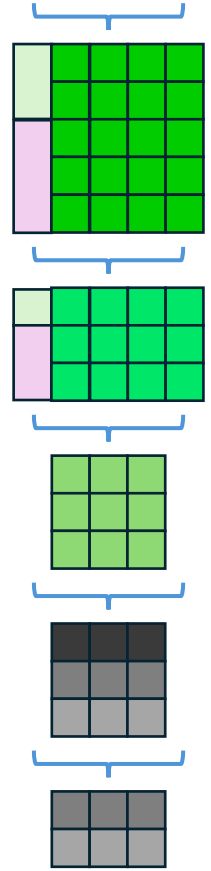
SELECT kol_a, kol_b

FROM t1

JOIN t2

ON t1.kol_a = t2.kol_b

WHERE warunki



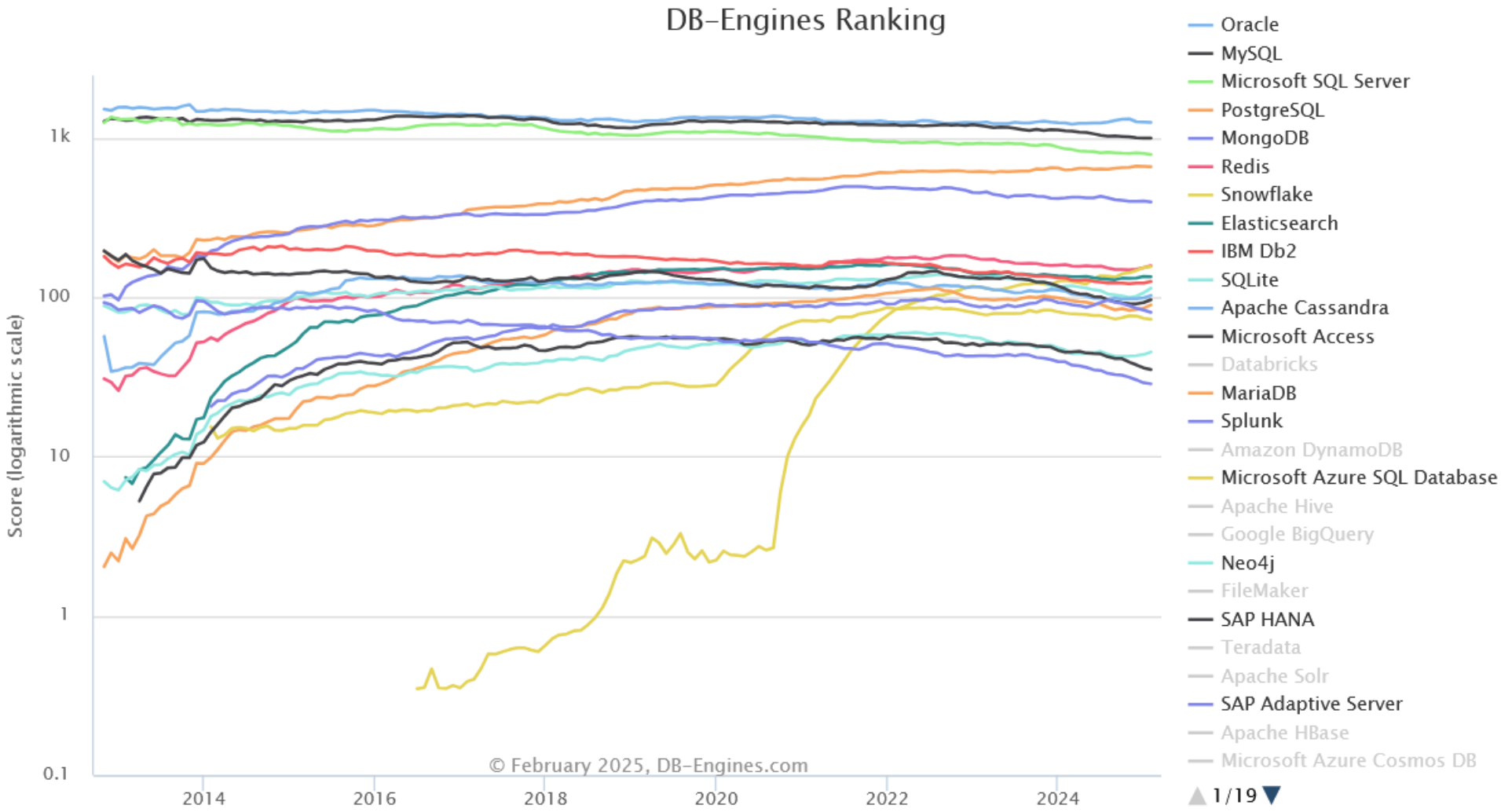
GROUP BY kol

HAVING warunki

ORDER BY kol ASC/DESC

LIMIT liczba

Ranking - Systemy baz danych



DB-Engines Ranking
https://db-engines.com/en/ranking

Zadanie

Treść

Z wykorzystaniem Generator dużych zbiorów danych Mockaroo https://mockaroo.com/ wygeneruj zbiór treningowy.

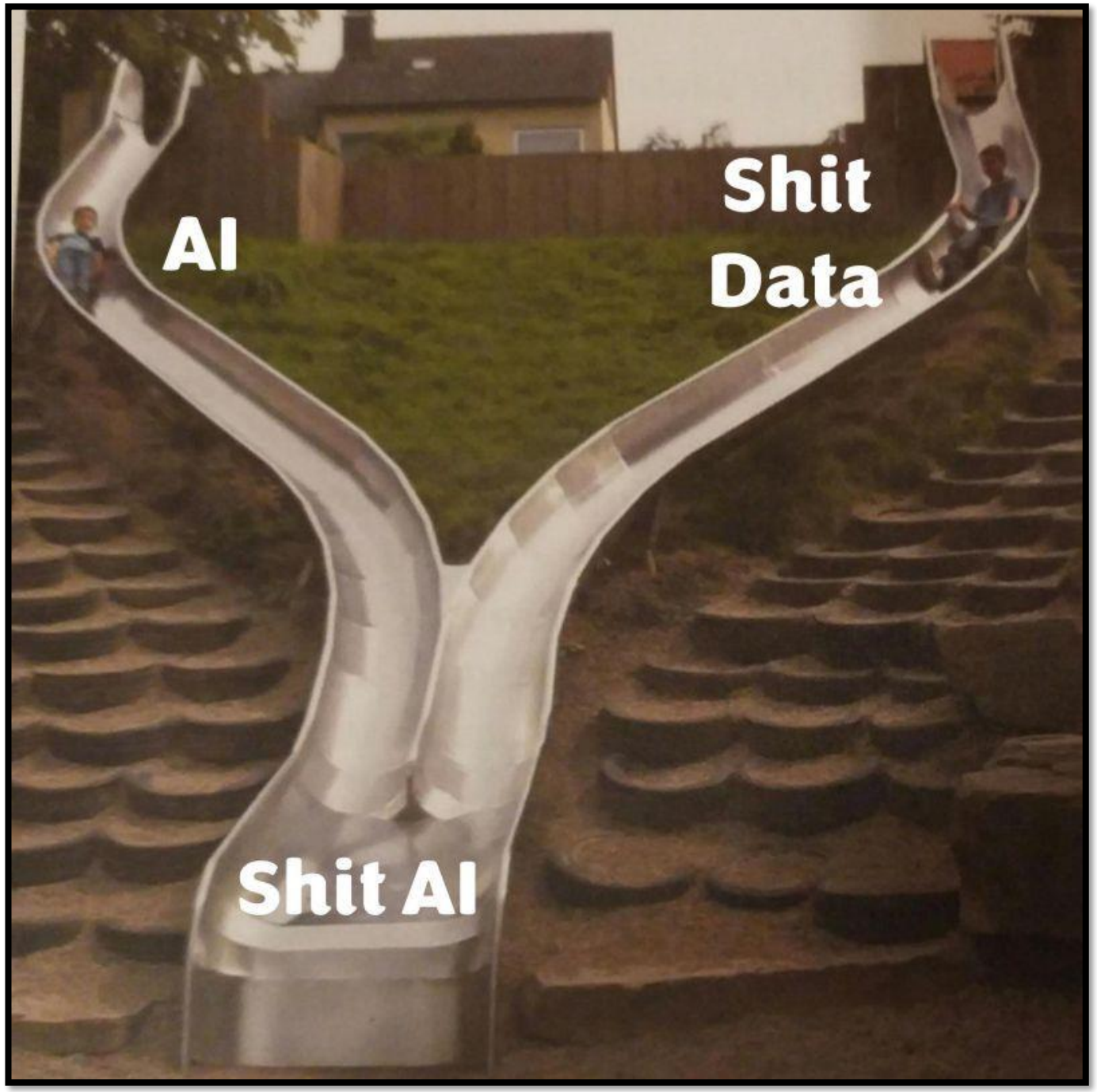
1. Skonfiguruj schemat danych dla predykcji kredytowej i ma zwierać:

- customer_id,
- age,
- income,
- credit_score,
- loan_amount,
- loan_approved,
- e-mail,
- gender

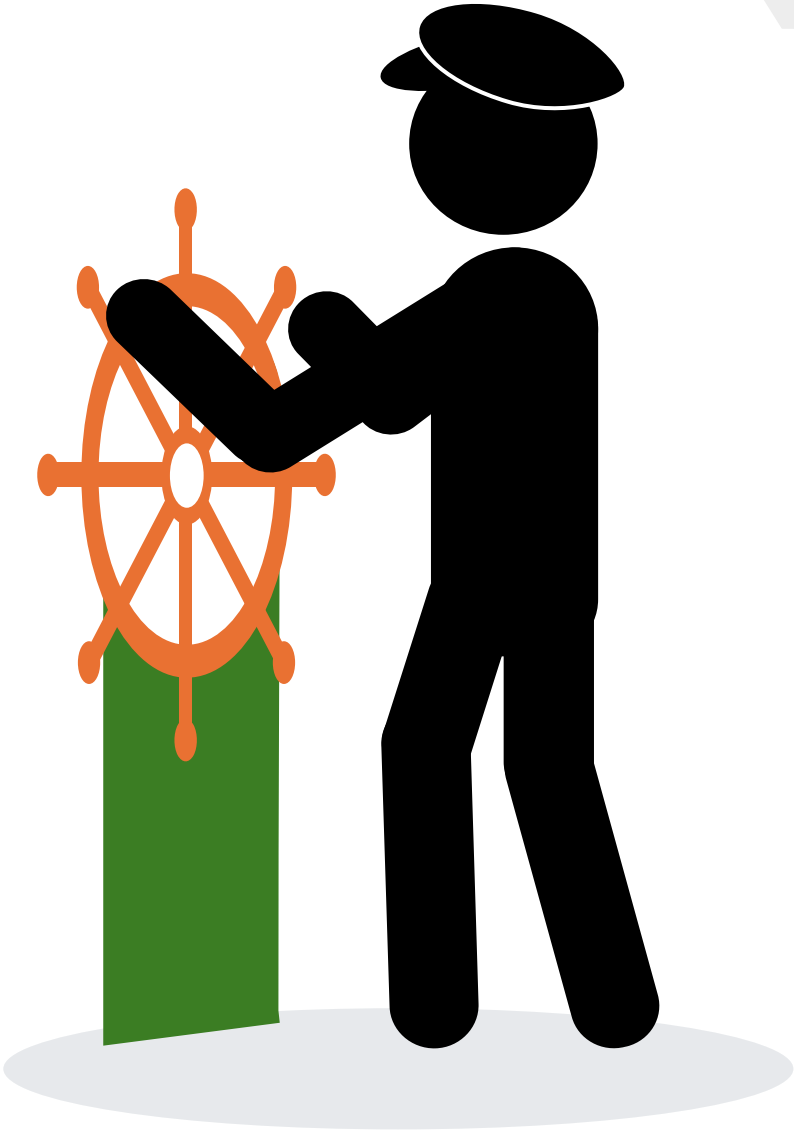
2. Zrób to wykorzystując opcje



3. Wygeneruj 1 000 rekordów i pobierz dane w formacie EXCEL.


AI
Shit
Data
Shit AI

Po ile zwierząt każdego gatunku zabrał Mojżesz do Arki?



Wyzwania związane z danymi wejściowymi

Niekompletne dane (Missing Data)

brakujące wartości mogą zaburzyć wyniki modelu, ponieważ AI może nie mieć wystarczających informacji do podjęcia decyzji.

✅ Przykład:

- Diagnostyka medyczna – w bazie pacjentów brakuje informacji o ciśnieniu krwi, co może wpłynąć na skuteczność modelu przewidującego ryzyko chorób serca.
- Systemy kredytowe – brak informacji o dochodach klienta może prowadzić do błędnej decyzji w ocenie zdolności kredytowej.
- Analiza ankiet – jeśli w badaniu użytkownicy nie wypełnią wszystkich pól, model może mieć zniekształcony obraz rzeczywistości.

Wyzwania związane z danymi wejściowymi

Błędy i szumy w danych (Noisy Data)

to przypadek kiedy dane mogą zawierać błędne wartości, literówki, przypadkowe wpisy, co prowadzi do złych predykcji modelu AI.

✅ Przykład:

- Rozpoznawanie obrazów – w zbiorze danych dla modelu wykrywającego koty znajdują się błędnie oznaczone zdjęcia psów.
- Analiza sentymentu w tekstach – w bazie opinii o produktach część recenzji jest ironiczna, ale model tego nie rozumie.
- Dane finansowe – błędnie zapisane transakcje (np. wartość „10000” zamiast „100.00”) mogą zaburzyć analizę wydatków klientów.

Wyzwania związane z danymi wejściowymi

Wysoka złożoność danych (High Dimensionality)

to sytuacja kiedy duże zbiory danych zawierają wiele cech (zmiennych), co prowadzi do problemu „przekleństwa wymiarowości” i wymaga wydajnego przetwarzania.

✅ Przykład:

- Genomika – analiza DNA obejmuje miliony cech, co sprawia, że modele predykcyjne muszą przetwarzać ogromne ilości informacji.
- Analiza obrazu – w sieciach neuronowych przetwarzających zdjęcia każda pikselowa wartość to osobna cecha, co wymaga dużej mocy obliczeniowej.
- Systemy rekomendacji – platformy streamingowe analizują setki tysięcy filmów, użytkowników i ich preferencji, co prowadzi do wysokiej złożoności obliczeniowej.

Wyzwania związane z danymi wejściowymi

Bias w danych (Data Bias)

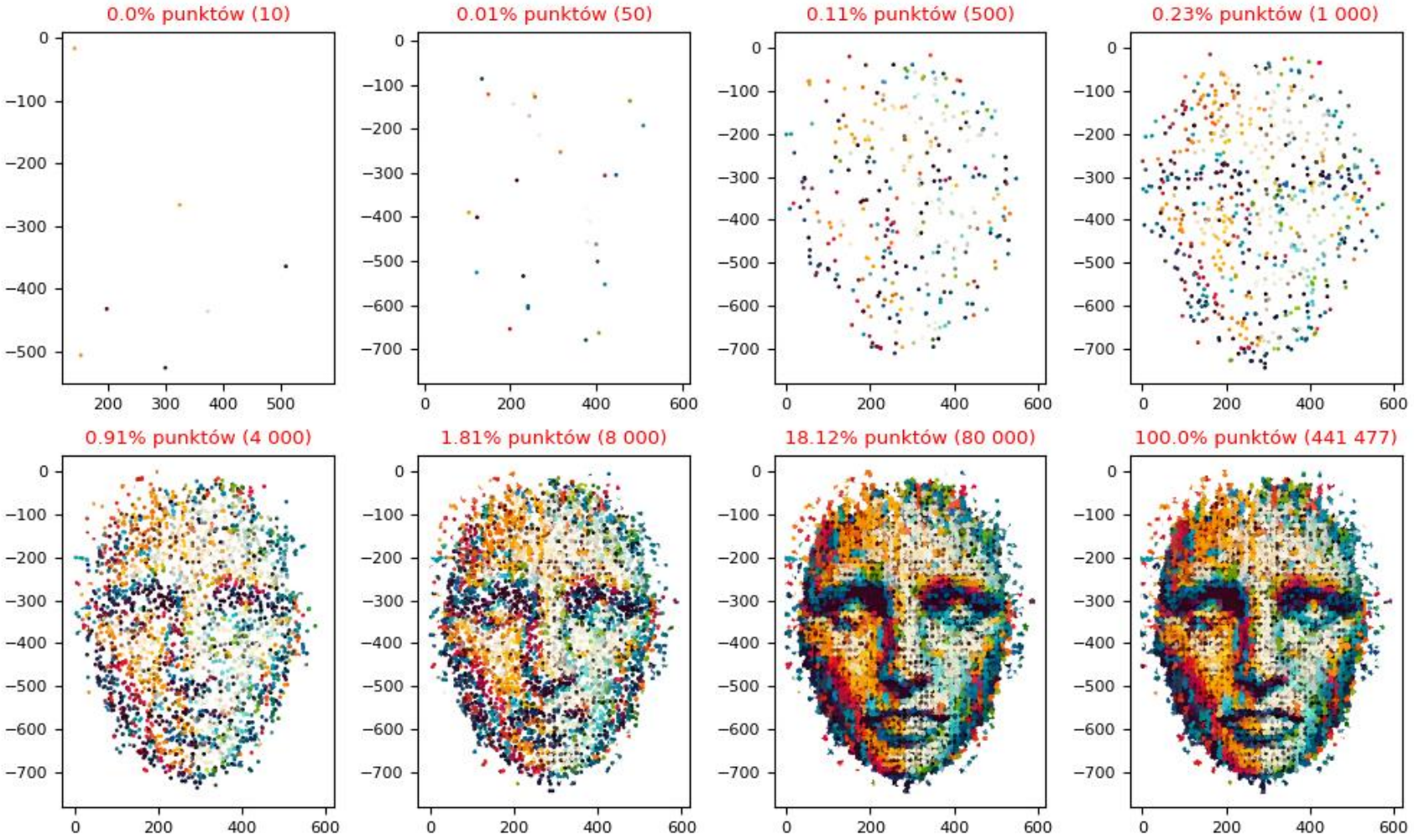
to stan kiedy dane zawierają **uprzedzenia i stronniczość**, które model AI odtwarza, prowadząc do **niesprawiedliwych lub błędnych wyników**.

✅ Przykład:

- **Rekrutacja do pracy** – model AI przeszkolony na historycznych danych faworyzuje mężczyzn, ponieważ w przeszłości większość stanowisk kierowniczych zajmowali mężczyźni.
- **Systemy kredytowe** – bankowy model oceny ryzyka może niesprawiedliwie odrzucać wnioski osób z określonych regionów, jeśli historyczne dane pokazują wyższe ryzyko dla tych lokalizacji.
- **Rozpoznawanie twarzy** – systemy AI mogą mieć niższą dokładność w identyfikacji osób o ciemniejszym kolorze skóry, jeśli zbiór treningowy zawierał głównie zdjęcia osób o jasnej karnacji.

Ile potrzeba danych aby zrozumieć informację?

Dla obrazu wystarczy ~20% danych



Ile potrzeba danych aby zrozumieć informację?



Dla tekstu potrzeba ~90% danych

50% znaków

```
_o__ój_____no__g__ i__or_ac_jny__ w _s__t_i_h
__ka___h__n__z_c___mienił___ajob__z_eko_om_cz_y_n_ ca__m ś_i____. ___wp_y_ę__
_a_w____t_ie_a_pe_t_ __iał_ln___i__os_od_r__ej_ __ p_____c_i_i za__ądz___a, po
k___nik_cję i___r__t_n_._W _____g_l_oś_i__k__je r_z__n_ęt_ od_______y
zn___n__w_ro_____rodu____n___________c___to__rzy_is__an_ _ą_wdr_ż_n_u
_o_____snych ___h_ol__i_. W__ym k___ekś_ie_i_totn__jest_z__z_mi_nie,_w
_aki__posób_IT_______j__n___zros_ ___p__a_czy__ _____ __ __ówne__e__ani__y
_eg__w__y_u.
```

60% znaków

```
R_z__j_te_hn_logii _n__rma_y__ych___osta___ch dekadac_
zna__ąco_zmi_nił_k_ajobra__eko____czny __ __ły__ś__e_ie. I_____y_ę_y n_
w_z_s_kie____e____działal____i__osp_d__cz_j, od____du_c_i _
z_rząd_a_i_,__o_kom_ni__c__ i __rketing._W ___zegó_ności,__r_j_ ro_wi_ięt_
odno__wa_y_z_ac_n___z___ty _r__uk__w_o_c_____re czę___ pr_y_isyw_ne_s_
wdrażan_u___wo_z_snych_te__n_lo_ii__W___m _onte_ści_ ____tn_ j___ z_o__mie____ w
__ki __os_b_IT_w__ywa____a wzr_st gosp_da__zy_i__a__e są__ł_wne mec_anizmy
tego__p__w_.
```

70% znaków

```
Rozw_j t___n_lo__i i_fo_mac_jn_ch w __t_tnic___ekadach_zna_z_c_
zm_enił____j_braz e_o__mic_ny n_ cały_ świecie. _T wpłynę_y
n__w_zystk_e_aspekty__zia_alno_ci gos_odar___j, __ p_odukc_____ar_ąd_ania, po
kom_nik_cj__i ma_ketin_._W _zczeg_lno_ci,_kraje r_z_i_ięt_ o_no_ow_ł___nac_ne
wzro_t_____dukty_ności,_któ_e czę__o prz_pisywa_e_są___r_ża_iu__o_o__esnych
te____lo__i. _ tym kontekś___ i_to__e jest___ozu___ni_,_w ja__ __o_ób_IT
wpływ_ją n_ _zrost _ospodarc_y _ jakie_są g_ówne mec_an__my te____p__w_.
```

80% znaków

```
__zwój_tech__logii i__ormacyj__ch w _sta_n_ch d__a_ach_znacz_co z_ie__ł
krajo_raz e_ono_icz_y na _ałym__wi_cie. IT w_ły_ęły na ___ystkie_aspek_y
_z_ała_ności gosp_d__cz__,_od produkcji i zarz_dza_ia, po komunikację i
marketing. W s_czególności,_kraje roz_inięte od_ot_wały__naczne wzr_sty
pr_d_ktywności, które często przyp_sy_a_e _ą wd__ża_iu nowoczes_ych_t_chnol__ii.
W t_m _on__kście__sto_ne_je_t zrozumien_e, __jak_ spo__b I__wpływ_j_ _a_wzros_
gospodarc_y _ j_kie są głów____echanizm_ tego wpływu_
```

90% znaków

```
Rozwój techn_logii in_o__ac__n__h w ostatnich dekadach zna_zące _mieni_
krajob__z ekonomiczny na całym świec_e. IT wpłynęły na wszystkie_aspekt_
dzi_łalnośc_ g_spodarczej, _d produkcji i _arządzania, _o _omu_i_ację i
mar_eting. W sz_ególności, kraje rozwini_te _dnot_wał_ znaczne
w_rosty_p_oduktywności, które częs_o przypisywane są wdr_żaniu nowoczes_ych
technologii_ W tym kontekście istotne j__t zrozumieni_,_w jaki sposób
IT_wpływa_ą na wzrost gospoda_czy i _akie s_ gł_wne me_hanizmy tego _pływu.
```

100% znaków

```
Rozwój technologii informacyjnych w ostatnich dekadach znacząco zmienił
krajobraz ekonomiczny na całym świecie. IT wpłynęły na wszystkie aspekty
działalności gospodarczej, od produkcji i zarządzania, po komunikację i
marketing. W szczególności, kraje rozwinięte odnotowały znaczne wzrosty
produktywności, które często przypisywane są wdrażaniu nowoczesnych technologii.
W tym kontekście istotne jest zrozumienie, w jaki sposób IT wpływają na wzrost
gospodarczy i jakie są główne mechanizmy tego wpływu.
```

Przetwarzanie danych

Czyszczenie danych – usuwanie błędnych i niekompletnych wartości

Dane mogą zawierać **brakujące wartości**, które uniemożliwiają poprawne działanie algorytmów.
Mogą występować **błędy w zapisie**, np. literówki w tekście, błędne formaty dat, wartości odstające w zbiorach liczbowych.

🛠 **Rozwiązania:**
✓ Usunięcie wierszy lub kolumn z dużą liczbą brakujących wartości.
✓ Wypełnienie braków średnią, medianą lub interpolacją (*imputacja*).
✓ Wykrywanie wartości odstających (*outliers*) i ich usunięcie lub transformacja.

🔷 **Przykład:**
W zbiorze danych o pacjentach niektóre rekordy nie zawierają informacji o ciśnieniu krwi. Można:

- Usunąć te rekordy.
- Uzupełnić brakujące wartości średnią z całego zbioru danych.
- Przewidzieć brakujące wartości za pomocą regresji lub algorytmów uczenia maszynowego.

Przetwarzanie danych

Kodowanie danych kategorycznych – zamiana wartości tekstowych na liczby

Algorytmy uczenia maszynowego wymagają danych numerycznych, więc wartości tekstowe (np. nazwy miast, kategorie produktów) muszą zostać przekonwertowane.

🛠 Rozwiązania:
✓ One-Hot Encoding – tworzenie osobnych kolumn dla każdej kategorii (np. "Miasto_Warszawa", "Miasto_Kraków").
✓ Label Encoding – przypisanie wartości numerycznych do kategorii (np. Warszawa = 0, Kraków = 1).

🔷 Przykład:
Dane o samochodach zawierają kolumnę „Rodzaj paliwa” z wartościami **benzyna, diesel, elektryczny**.

- **One-hot encoding:** Stworzy 3 kolumny: "Benzyna" (1/0/0), "Diesel" (0/1/0), "Elektryczny" (0/0/1).
- **Label encoding:** Zamieni wartości na liczby: Benzyna = 0, Diesel = 1, Elektryczny = 2.

Przetwarzanie danych

Ekstrakcja cech – wybór istotnych informacji

Nadmiar nieistotnych danych może **spowolnić działanie modelu i pogorszyć wyniki**.
Nie wszystkie cechy mają realny wpływ na wynik predykcji.

🛠 **Rozwiązania:**
✓ **Feature Selection** – wybór najważniejszych zmiennych.
✓ **Feature Engineering** – tworzenie nowych cech na podstawie istniejących (np. zamiast "data urodzenia" można użyć "wiek").
✓ **Analiza korelacji** – eliminowanie cech, które są ze sobą silnie skorelowane.

🔷 **Przykład:**
W analizie klientów bankowych **adres e-mail i numer telefonu** nie mają wpływu na decyzję kredytową, ale **historia spłat i dochody** już tak. Warto usunąć nieistotne zmienne, aby uprościć model.

Zadanie



Treść

1. Wybierz najważniejsze cechy w kontekście oceny zdolności kredytowej z wcześniejszego ćwiczenia związanego z generowanie zbioru treningowego.
2. Popraw plik EXCEL usuwając z niego zbędne cechy.
3. O ile procent zmniejszyła się jego pojemność?
4. Zapisz plik w postaci CSV
5. O ile teraz zmniejszyła się jego pojemność?
6. Zamień nazwę pliku na kredyt.csv

Przetwarzanie danych

Normalizacja danych (Data Preprocessing)

Polega na **przekształceniu danych wejściowych** do określonego zakresu wartości, co **przyspiesza proces uczenia** i **poprawia stabilność modelu**..

Dlaczego normalizacja jest ważna?

✅ **Przyspiesza konwergencję** – modele uczą się szybciej, gdy dane mają podobną skalę.

✅ **Zapobiega dominacji cech** – bez normalizacji cechy o dużych wartościach mogą zdominować proces uczenia.

✅ **Poprawia stabilność gradientów** – zmniejsza ryzyko zbyt dużych lub zbyt małych aktualizacji wag (problem zbieżności gradientów).

Konwergencja odnosi się do **momentu, w którym model osiąga stabilny poziom wydajności**, a jego **błąd (loss) przestaje się znacząco zmniejszać** podczas kolejnych iteracji treningu.
Oznacza to, że model znalazł **optymalne (lub bliskie optymalnemu) wartości wag**, które najlepiej dopasowują się do danych treningowych.

Przetwarzanie danych

Metody normalizacji danych

1 Standardizacja (z-score normalization)

Przekształca dane, tak aby miały **średnią (μ) równą 0** i **odchylenie standardowe (σ) równe 1**.

$$z = \frac{x - \mu}{\sigma}$$

Gdzie:
- x – wartość cechy,
- μ – średnia arytmetyczna,
- σ – odchylenie standardowe.

Zastosowanie
✓ Regresja liniowa i logistyczna
✓ Modele oparte na odległościach (SVM, KNN)
✓ Analiza PCA (redukcja wymiarów)

Przetwarzanie danych

Metody normalizacji danych

2 Min-max scaling (normalizacja do przedziału [0,1])

Skaluje dane do określonego zakresu, najczęściej [0,1]

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Gdzie:

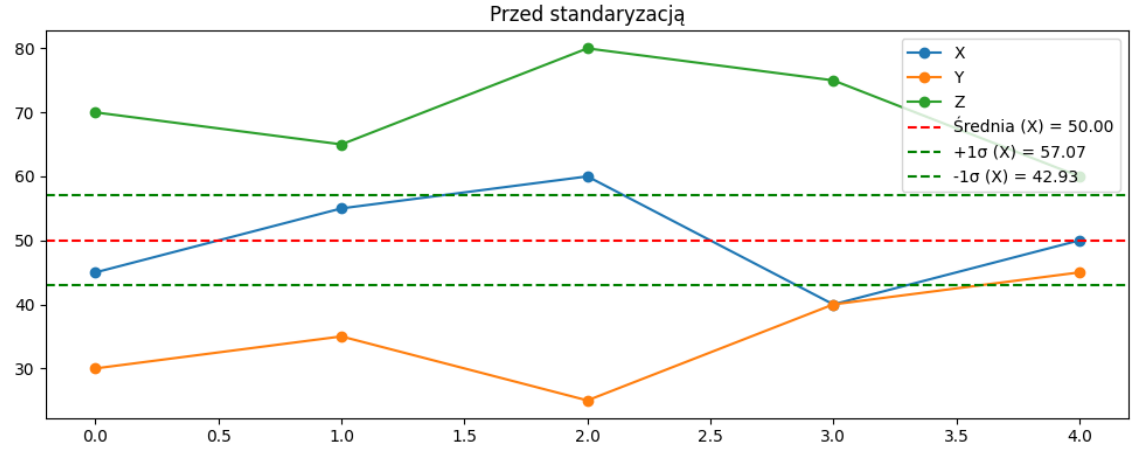
- x – oryginalna wartość,
- x_{min} , x_{max}– minimalna i maksymalna wartość w zbiorze danych,
- σ – odchylenie standardowe.

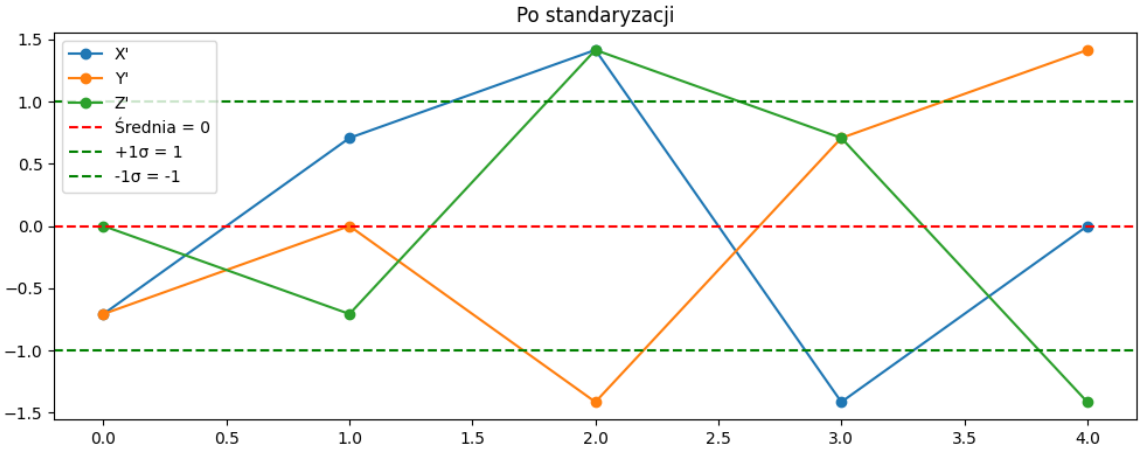
Zastosowanie

✓ Sieci neuronowe (MLP, CNN, RNN)
✓ Uczenie głębokie (Deep Learning)
✓ Modele wrażliwe na skalę, np. gradient descent

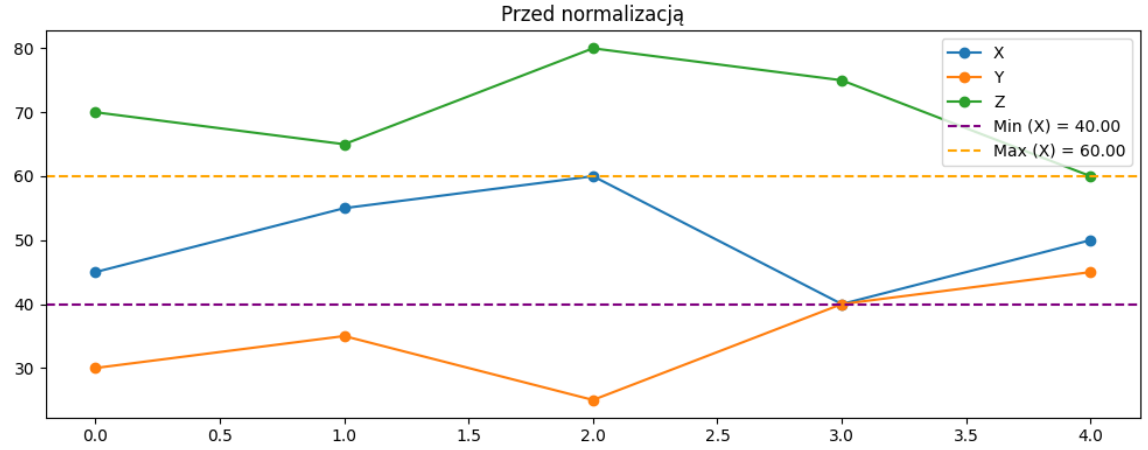
Przetwarzanie danych

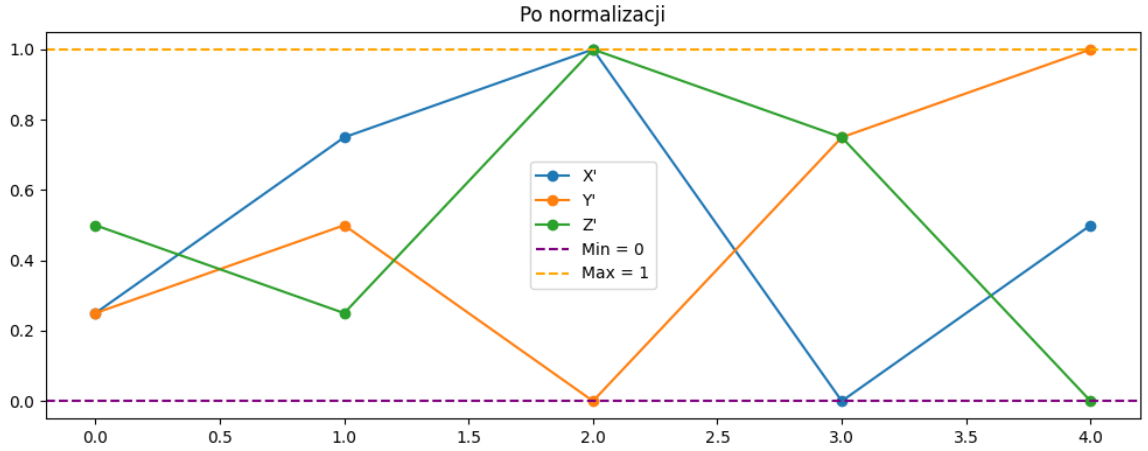
Standaryzacja





Normalizacja





Zadanie



Treść

W Google Colab https://colab.research.google.com/ dokonaj normalizacji przygotowanego wcześniej pliku CSV ze zbiorem treningowym.

1. Prześlij plik kredyt.csv do Google colab
2. Wyświetl w postaci tabeli dane z pliku kredyt.csv dane są rozdzielone znakiem ";„ a znak dziesiętny to „,”
3. Zrób wykres gdzie na osi X jest 'customer_id' a na Y 'age', 'loan_amount' i 'credit_score’
4. Zrób wykres gdzie na osi X jest 'customer_id' a na Y znormalizowane 'age', 'loan_amount' i 'credit_score’

https://colab.research.google.com/drive/11WPnC50kFW6ylAWR6XYzHN_k3uedRFt1?usp=sharing

Statystyka opisowa

Statystyka opisowa

to **zbiór metod analizy danych**, które pozwalają na **zrozumienie ich rozkładu, zależności i charakterystyk**.

Jest kluczowa w sztucznej inteligencji, ponieważ umożliwia **eksplorację danych przed ich wykorzystaniem w modelach AI**.

Zastosowanie statystyki opisowej w AI

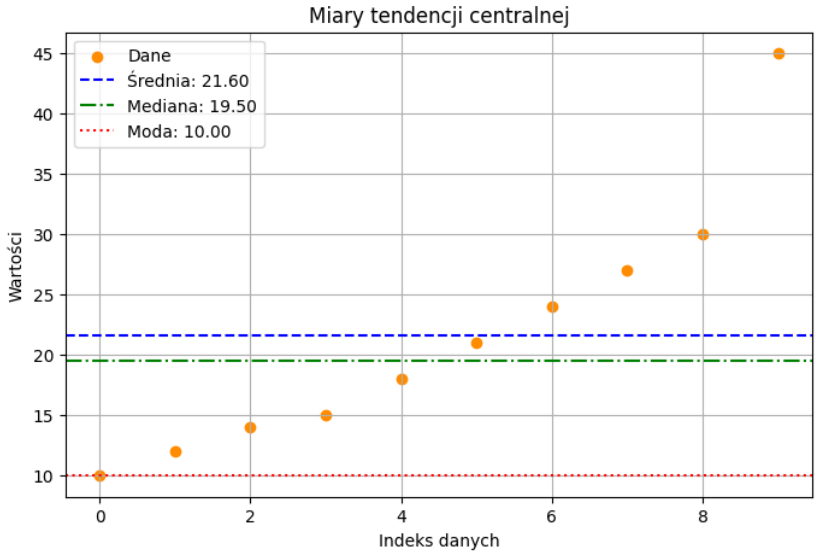
- **Eksploracja danych (EDA)** – sprawdzanie, jakie są średnie, odchylenia i zależności między danymi.
- **Wykrywanie anomalii** – identyfikacja wartości odstających (np. wykrywanie oszustw bankowych).
- **Przygotowanie danych do modeli AI** – np. normalizacja, transformacja zmiennych.
- **Analiza wyników modeli** – ocena skuteczności predykcji na podstawie rozkładu błędów.

Statystyka opisowa

Miary tendencji centralnej (opisują „środek” danych)

✓ **Średnia (Mean)** – wartość przeciętna w zbiorze danych.
✓ **Mediana (Median)** – środkowa wartość, odporna na wartości odstające.
✓ **Moda (Mode)** – najczęściej występująca wartość w zbiorze danych.

📌 *Przykład:* W analizie dochodów średnia może być zawyżona przez kilku miliarderów, dlatego mediana lepiej odzwierciedli typowy dochód.

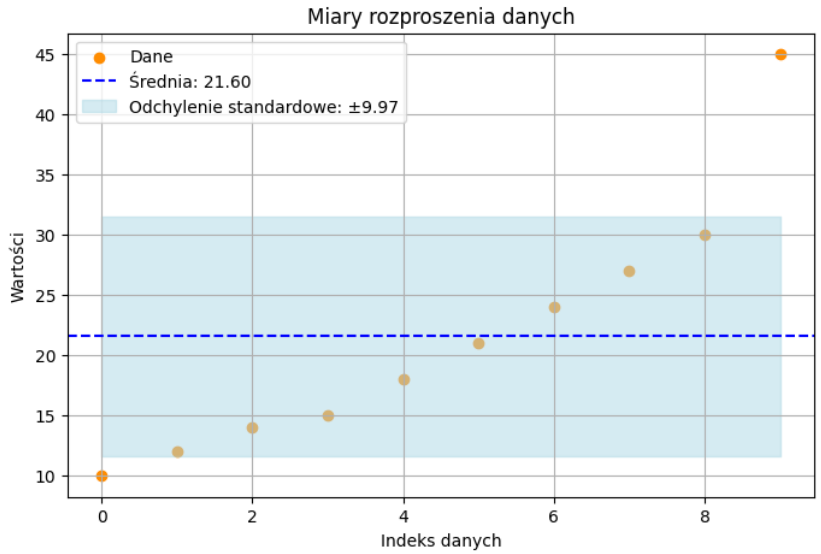


Statystyka opisowa

Miary rozproszenia ***(opisują, jak bardzo dane są zróżnicowane)***

✓ **Odchylenie standardowe (Standard Deviation, σ)** – pokazuje, jak bardzo wartości odbiegają od średniej.
✓ **Wariancja (Variance, σ^2)** – kwadrat odchylenia standardowego, mierzy zmienność w zbiorze danych.
✓ **Rozstęp (Range)** – różnica między wartością maksymalną a minimalną.

📌 *Przykład:* W systemach AI analizujących ceny nieruchomości duża wariancja może sugerować różnice w lokalizacjach i metrażach.

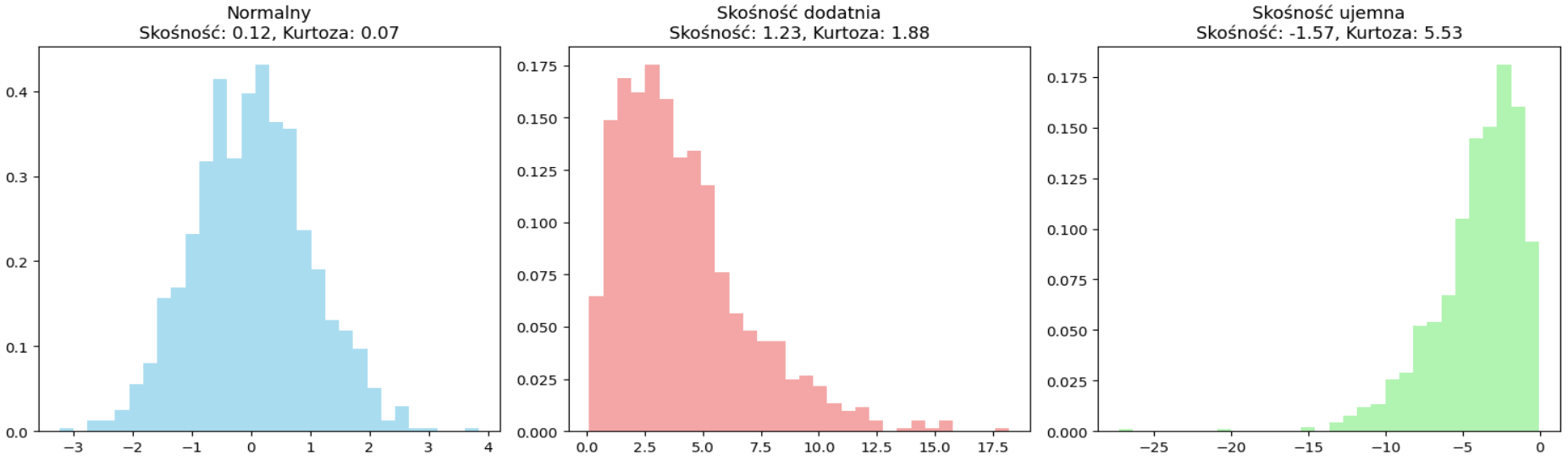


Statystyka opisowa

Miary kształtu rozkładu *(opisują asymetrię i koncentrację danych)*

✓ **Skośność (Skewness)** – określa, czy dane są bardziej przesunięte w prawo (długi ogon po prawej) czy w lewo.
✓ **Kurtoza (Kurtosis)** – mierzy, jak bardzo dane są skoncentrowane wokół średniej.

📌 *Przykład:* W modelach AI analizujących czas spędzany przez użytkowników na stronie wysoka skośność może oznaczać, że większość użytkowników opuszcza stronę szybko, ale nieliczni zostają długo.



Zadanie



Treść

W Google Colab https://colab.research.google.com/ dokonaj oszacowania dla wcześniej przygotowanego pliku kredyt.csv.

1. Oszacuj Średnią, Mediana, Moda dla pól 'age', 'loan_amount' i 'credit_score’ dane są rozdzielone znakiem ";„ a znak dziesiętny to „,”
2. Oszacuj Odchylenie standardowe, Wariancja, Rozstęp dla pól 'age', 'loan_amount' i 'credit_score’
3. Oszacuj Skośność, Kurtoza dla pól 'age', 'loan_amount' i 'credit_score’
4. Co oznaczają oszacowane wyniki ? Napisz odpowiedź po polsku

https://colab.research.google.com/drive/11WPnC50kFW6ylAWR6XYzHN_k3uedRFt1?usp=sharing

Procesy ETL i ELT

ETL i ELT

Procesy **ETL (Extract, Transform, Load)** i **ELT (Extract, Load, Transform)** są kluczowe dla **przetwarzania danych wejściowych w AI**.
Pozwalają na **gromadzenie, czyszczenie i organizowanie danych**, które następnie są wykorzystywane w modelach sztucznej inteligencji.

🛠 Jak ETL i ELT pomagają w AI?

- 🔹 **Oczyszczanie danych** – usuwanie błędów i niekompletnych wartości.
- 🔹 **Standaryzacja i normalizacja** – ujednolicanie formatu danych wejściowych.
- 🔹 **Obsługa Big Data** – ELT umożliwia analizę ogromnych zbiorów danych w czasie rzeczywistym.
- 🔹 **Łączenie różnych źródeł** – integracja danych z systemów IoT, baz SQL, API.

Procesy ETL i ELT

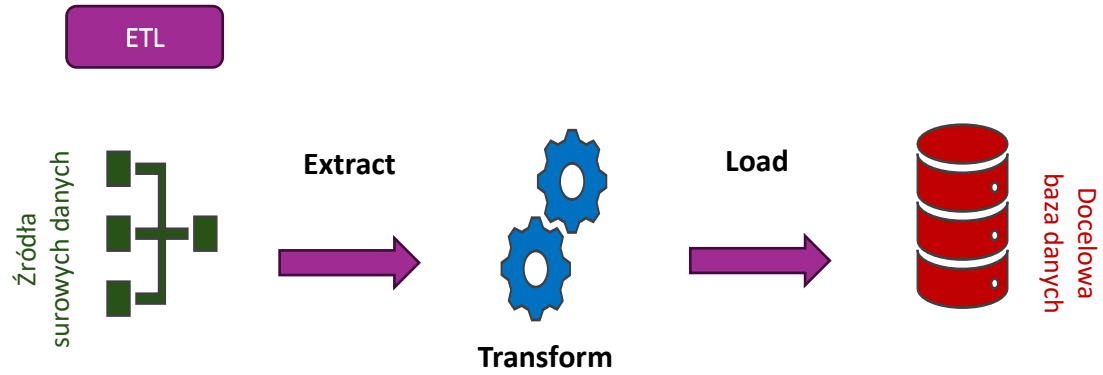
ETL (Extract, Transform, Load)

to tradycyjny proces przetwarzania danych, w którym:

1. **Extract (Ekstrakcja)** – pobieranie danych z różnych źródeł (bazy danych, API, IoT).
2. **Transform (Transformacja)** – czyszczenie, normalizacja, standaryzacja danych przed ich zapisaniem.
3. **Load (Załadowanie)** – zapisanie gotowych danych do hurtowni danych lub systemu AI.

✅ **Zastosowanie w AI:**

✓ Modele klasyczne ML (np. regresja, drzewa decyzyjne).
✓ Przetwarzanie ustrukturyzowanych danych (CRM, ERP).
✓ Analityka predykcyjna i raportowanie biznesowe.



Procesy ETL i ELT

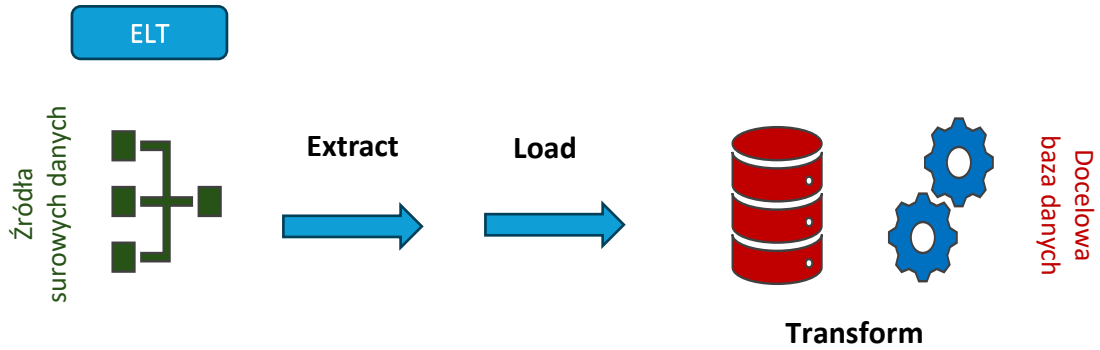
ELT (Extract, Load, Transform)

to nowoczesne podejście do przetwarzania danych, wykorzystywane w big data i chmurze.

1. **Extract (Ekstrakcja)** – pobranie danych do systemu (np. hurtowni danych, jeziora danych).
2. **Load (Załadowanie)** – zapisanie surowych danych bezpośrednio do systemu.
3. **Transform (Transformacja)** – przekształcenie danych w miejscu docelowym (np. za pomocą SQL lub narzędzi AI).

✅ Zastosowanie w AI:

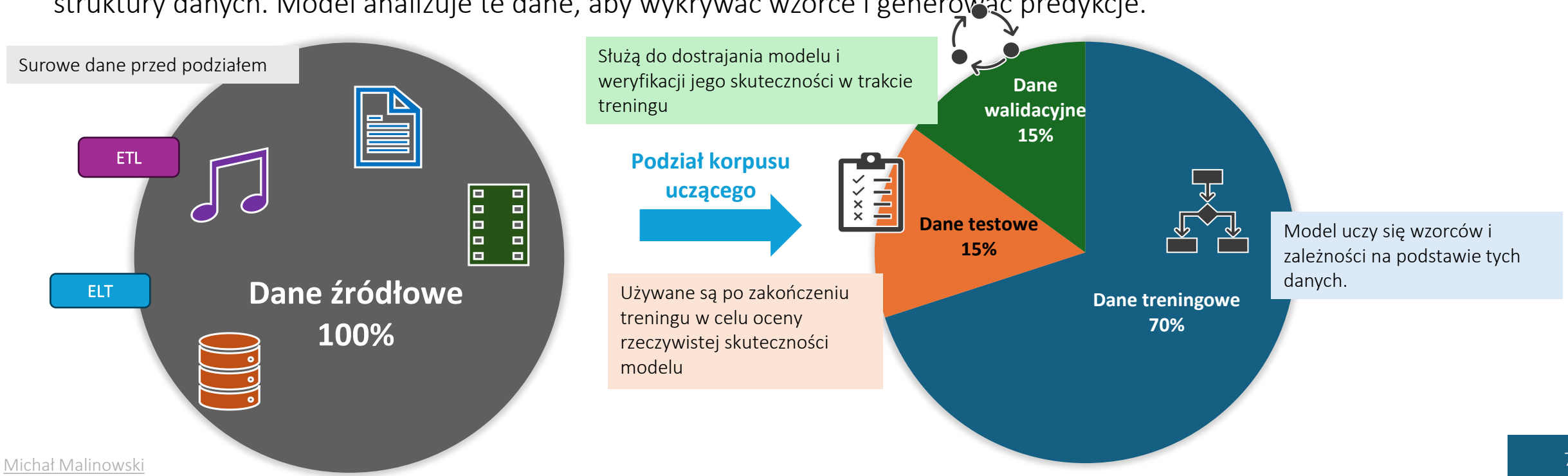
✓ Modele deep learning i big data (np. analiza obrazów, NLP).
✓ Przetwarzanie danych w chmurze (AWS, Google BigQuery).
✓ Systemy analizy w czasie rzeczywistym.



Korpus uczący

Korpus uczący

to **zbiór danych** wykorzystywany do trenowania modeli AI. Może zawierać **teksty, obrazy, dźwięki, liczby** oraz inne struktury danych. Model analizuje te dane, aby wykrywać wzorce i generować predykcje.

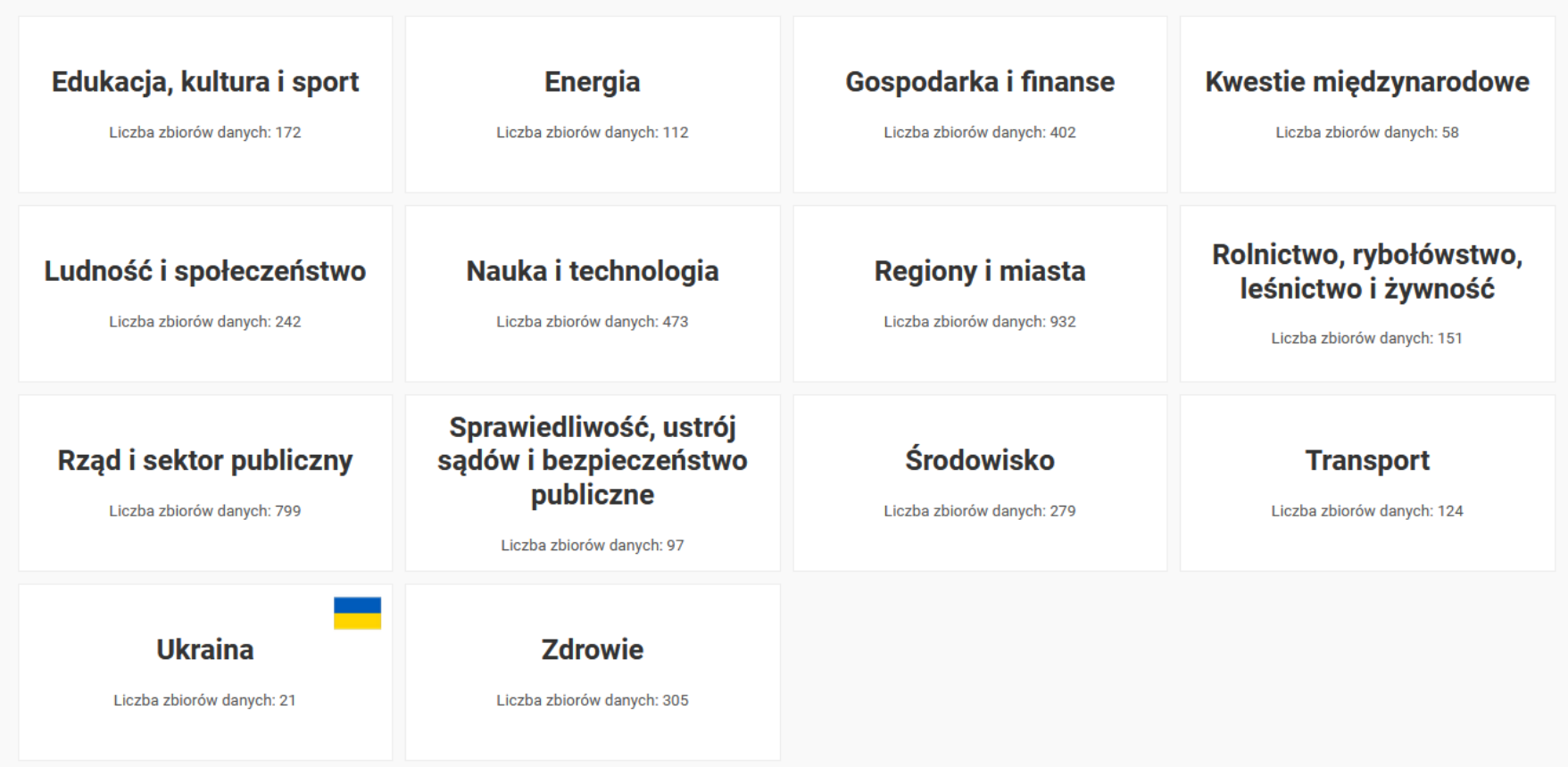


Korpus uczący

Otwarte dane

to informacje udostępniane przez instytucje publiczne, które mogą być swobodnie używane, modyfikowane i udostępniane przez każdego. Dane te są dostępne w formatach umożliwiających ich łatwe przetwarzanie i ponowne wykorzystanie.

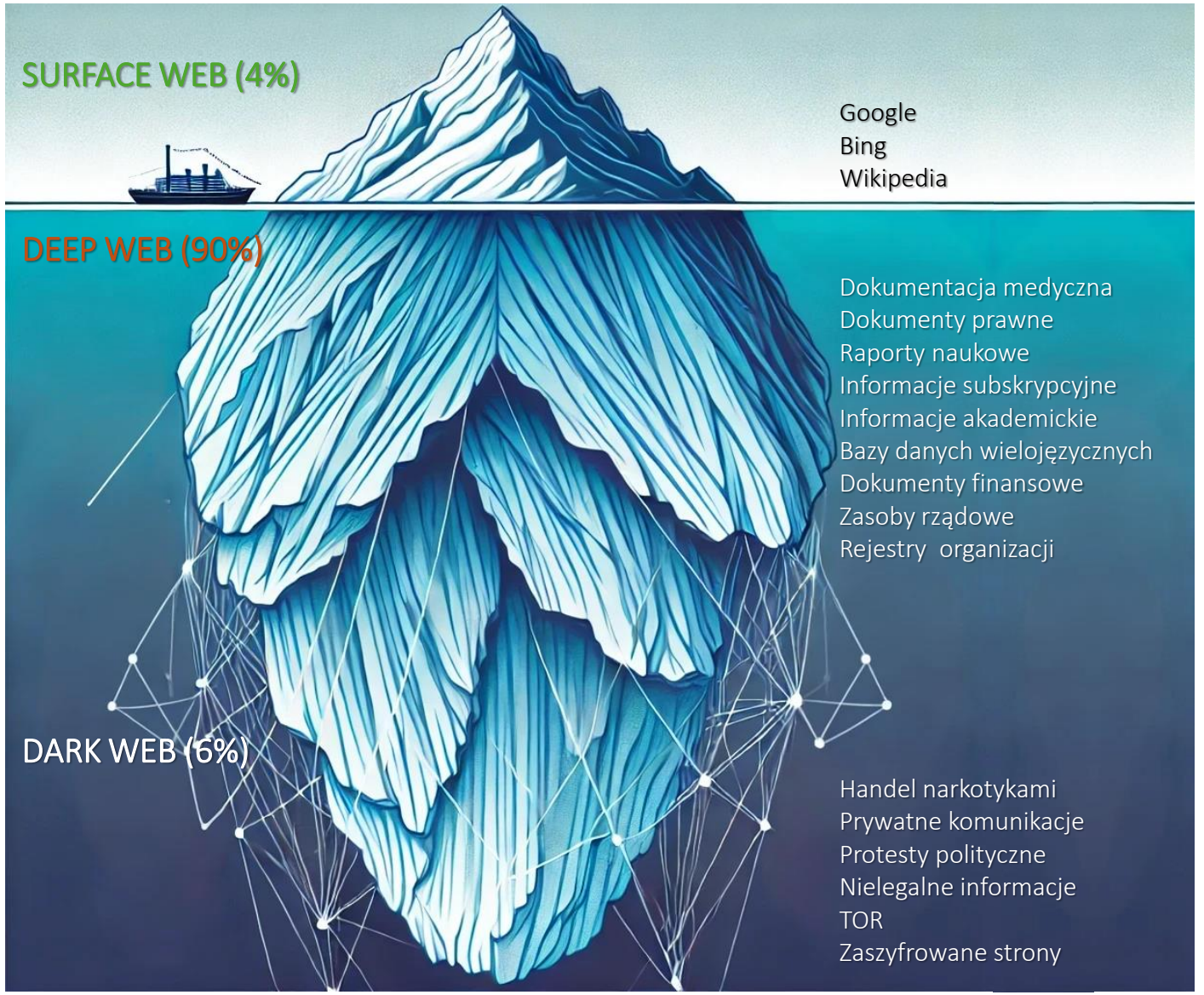
https://dane.gov.pl/pl



Korpus uczący

Obszar zastosowania	Nazwa korpusu	Opis	Link
Ogólne zbiory danych	**Kaggle Datasets**	Różne dziedziny: NLP, CV, finanse, zdrowie	https://www.kaggle.com/datasets
	Hugging Face Datasets	Korpusy do NLP, CV, multimodalnych modeli	https://huggingface.co/datasets
	UCI Machine Learning	Popularne zbiory danych do klasycznych problemów ML	https://archive.ics.uci.edu/ml/index.php
Przetwarzanie języka naturalnego (NLP)	**Wikipedia Dumps**	Kopia treści Wikipedii	https://dumps.wikimedia.org/
	SQuAD	Zbiór pytań i odpowiedzi dla chatbotów i AI	https://rajpurkar.github.io/SQuAD-explorer/
Analiza obrazów (Computer Vision, CV)	**CIFAR-10 / CIFAR-100**	Małe obrazy (32x32 px) do klasyfikacji	https://www.cs.toronto.edu/~kriz/cifar.html
	COCO	Obrazy z opisami i segmentacją obiektów	https://cocodataset.org/
Analiza dźwięku i mowy	**LibriSpeech**	Duży zbiór nagrań mowy	https://www.openslr.org/12/
	VoxCeleb	Nagrania głosowe znanych osób	https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

Darknet


SURFACE WEB (4%)
Google
Bing
Wikipedia
DEEP WEB (90%)
Dokumentacja medyczna
Dokumenty prawne
Raporty naukowe
Informacje subskrypcyjne
Informacje akademickie
Bazy danych wielojęzycznych
Dokumenty finansowe
Zasoby rządowe
Rejestry organizacji
DARK WEB (6%)
Handel narkotykami
Prywatne komunikacje
Protesty polityczne
Nielegalne informacje
TOR
Zaszyfrowane strony

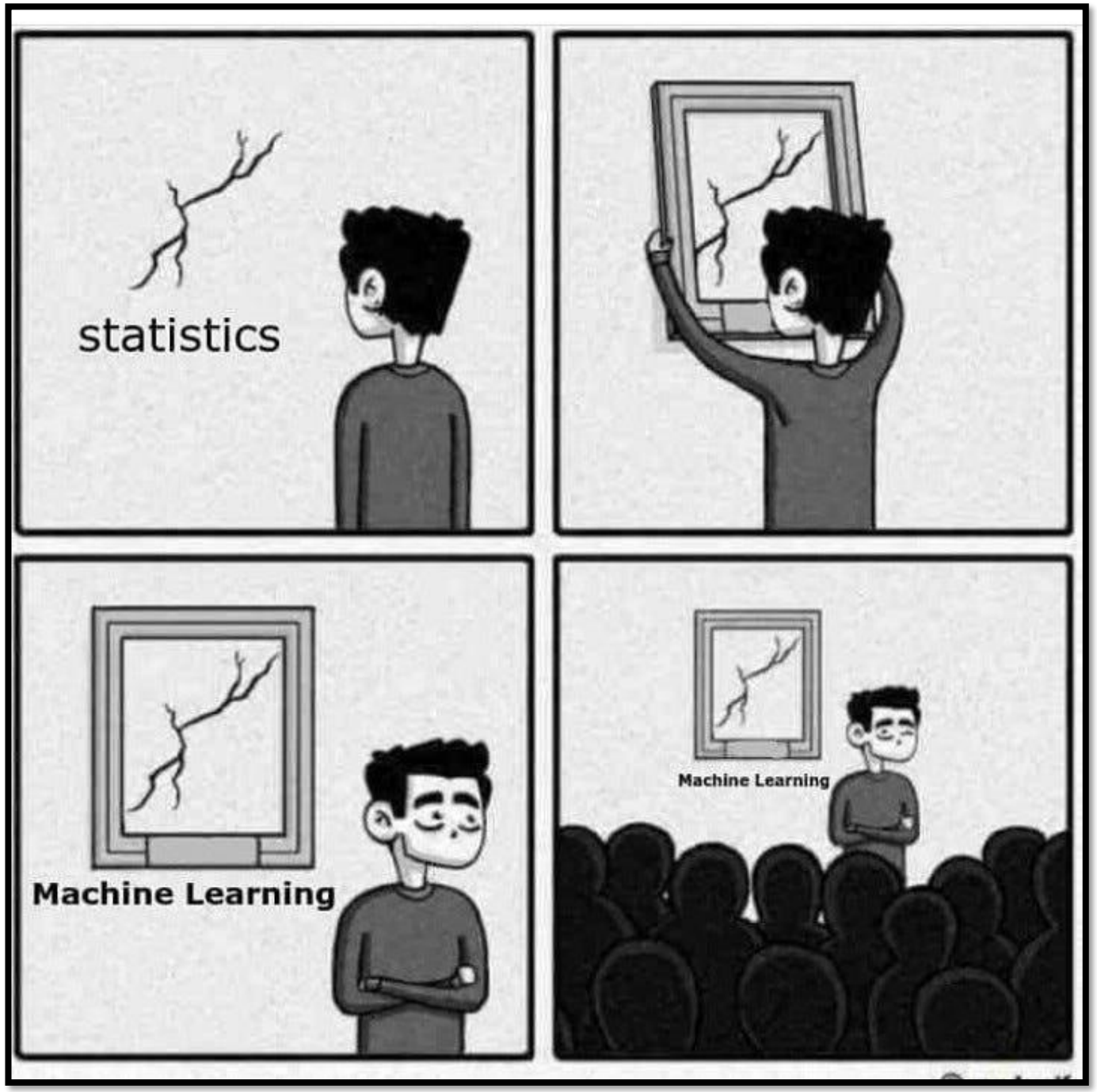
Zadanie



Treść

Skorzystaj z Otwartych Danych https://dane.gov.pl/pl i zbuduj zbiór treningowy.

1. Pobierz plik CSV ze strony https://dane.gov.pl/en/dataset/540/resource/64621,pracownicy-policji-stan-etatowy-i-stan-zatrudnienia-wg-suzb-w-jednostkach-organizacyjnych-policji/table
2. Nazwij plik policja.csv
3. Dokonaj Transformacji tego plik w EXCEL'u aby miał podstawową postać formatu CSV
4. Prześlij plik do Google colab https://colab.research.google.com/
5. Przedstaw dane w postaci tabelarycznej z przesłanego pliku policja.csv (o ile plik będzie dobrze transformowany)


statistics
Machine Learning
Machine Learning