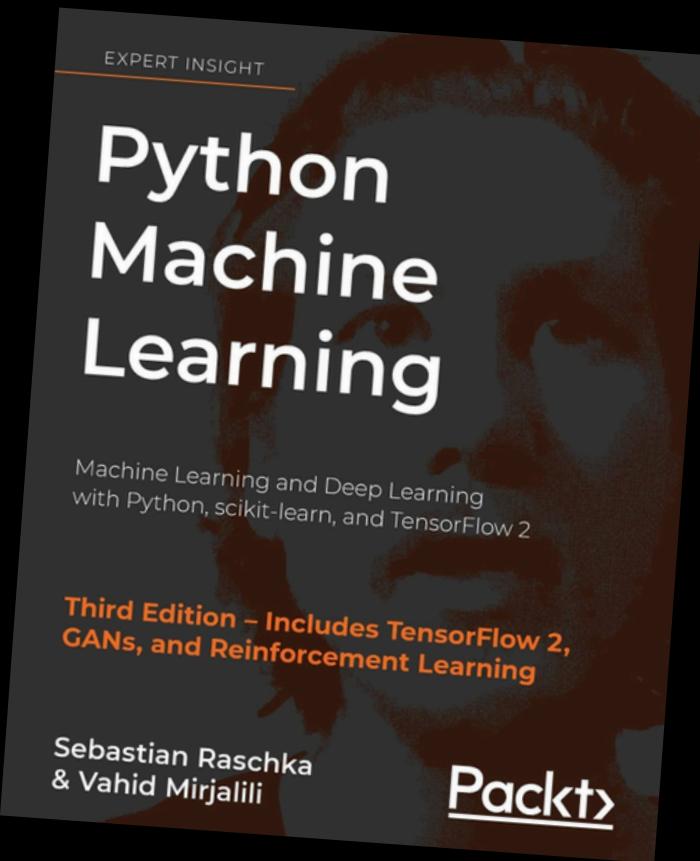
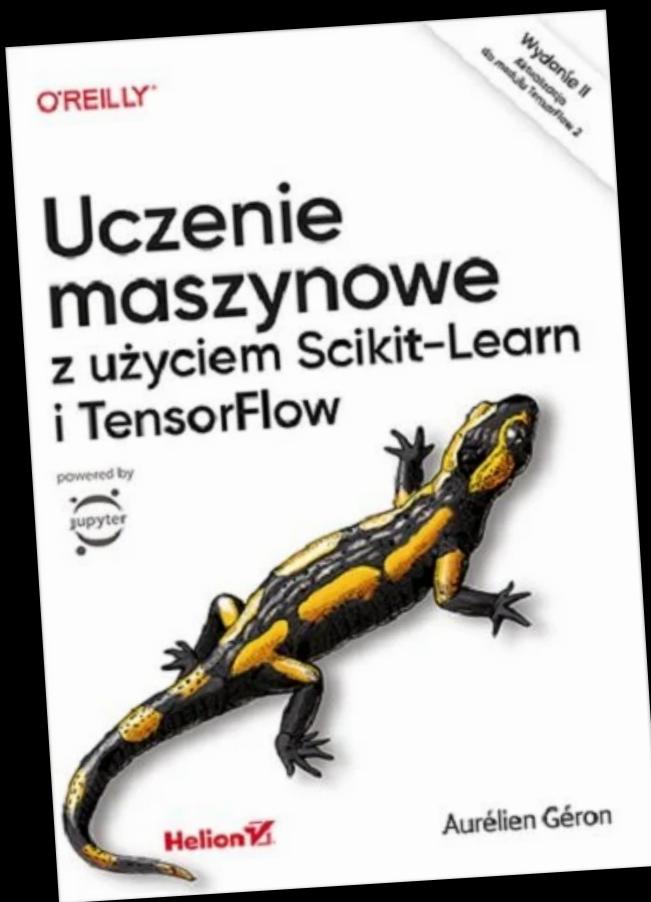




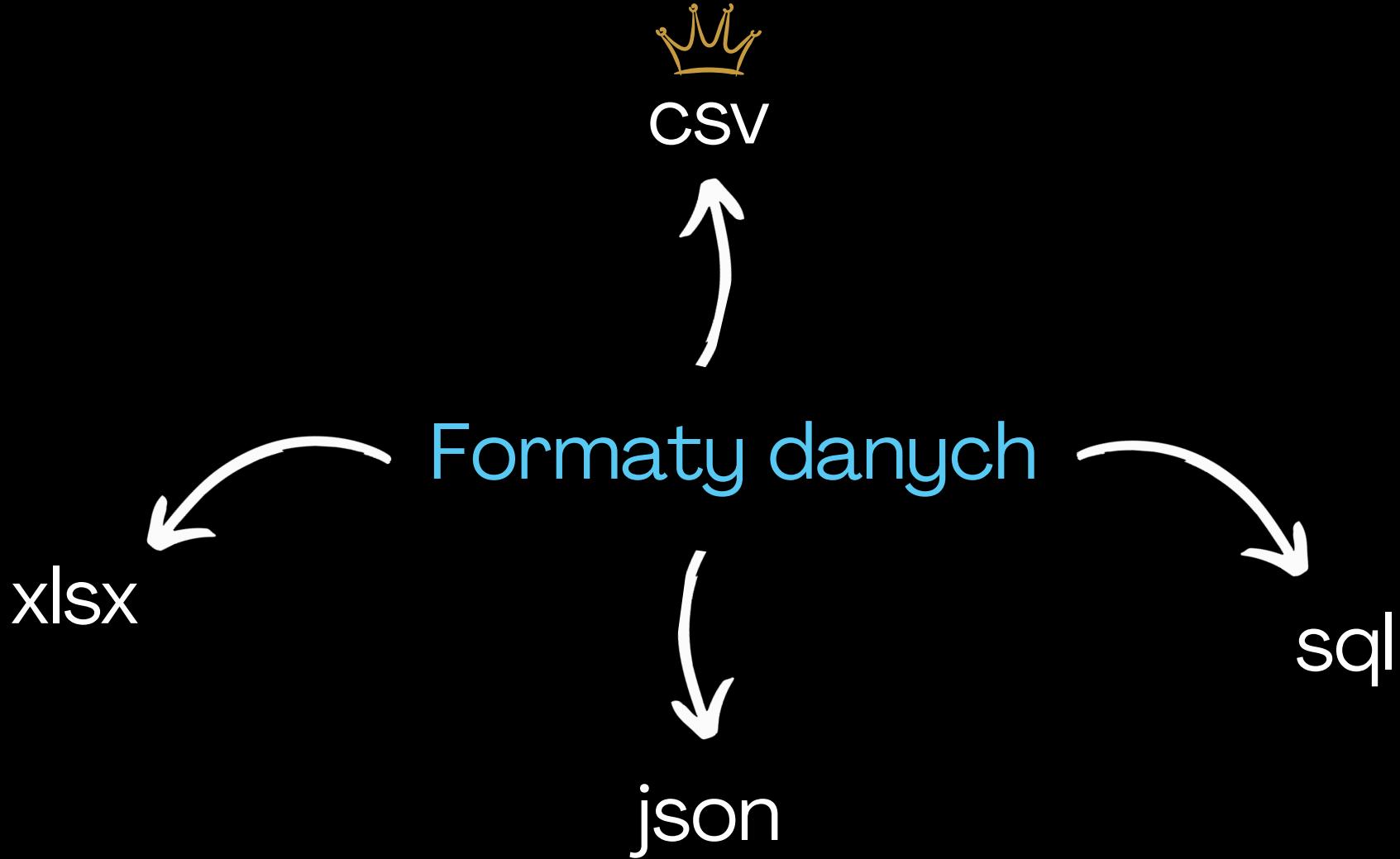
Literatura



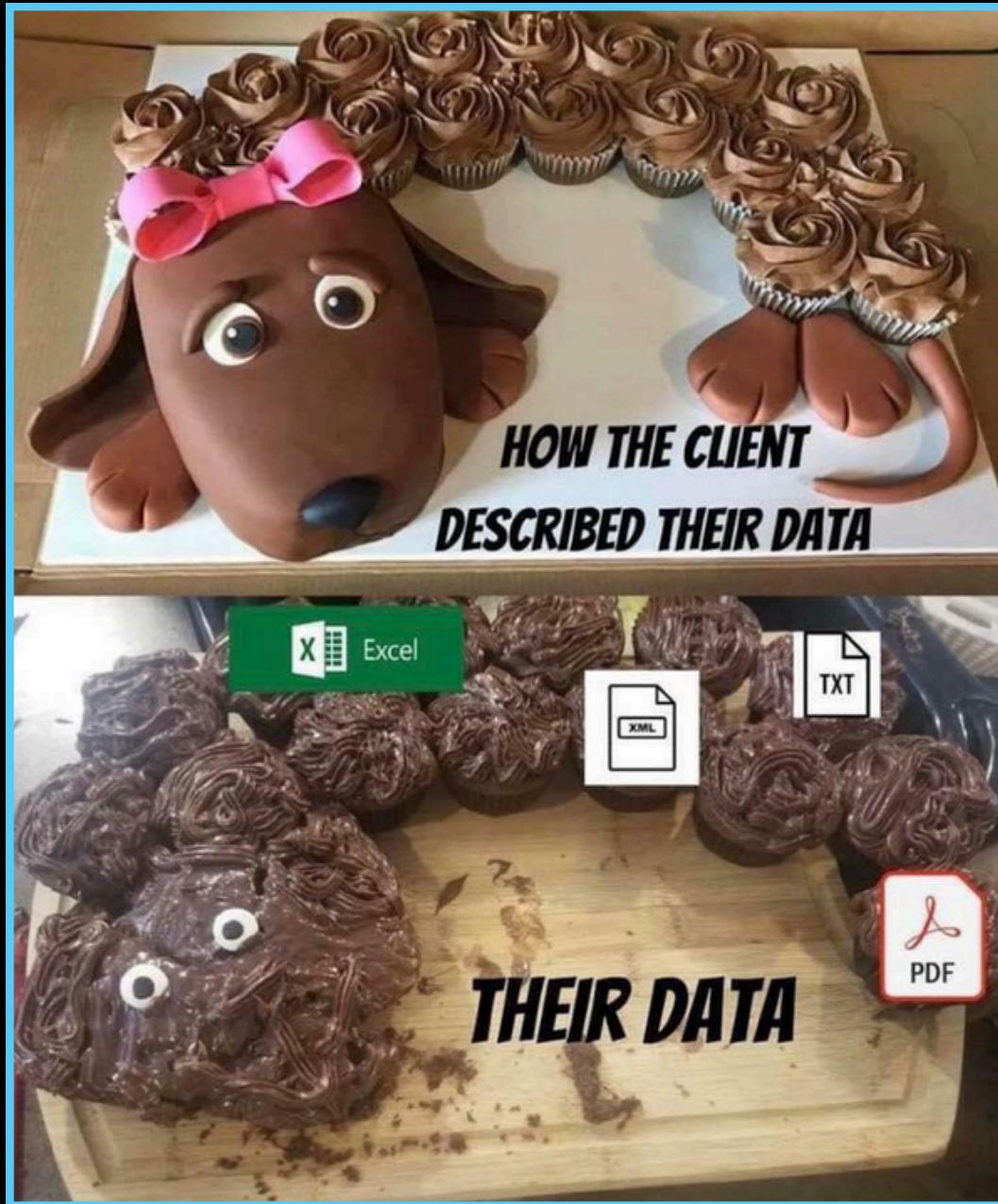


Jakie mogą być formaty danych?





jpg, mp4



Gdzie znaleźć dane?



Gdzie znaleźć dane?

kaggle



Przygotowanie Danych



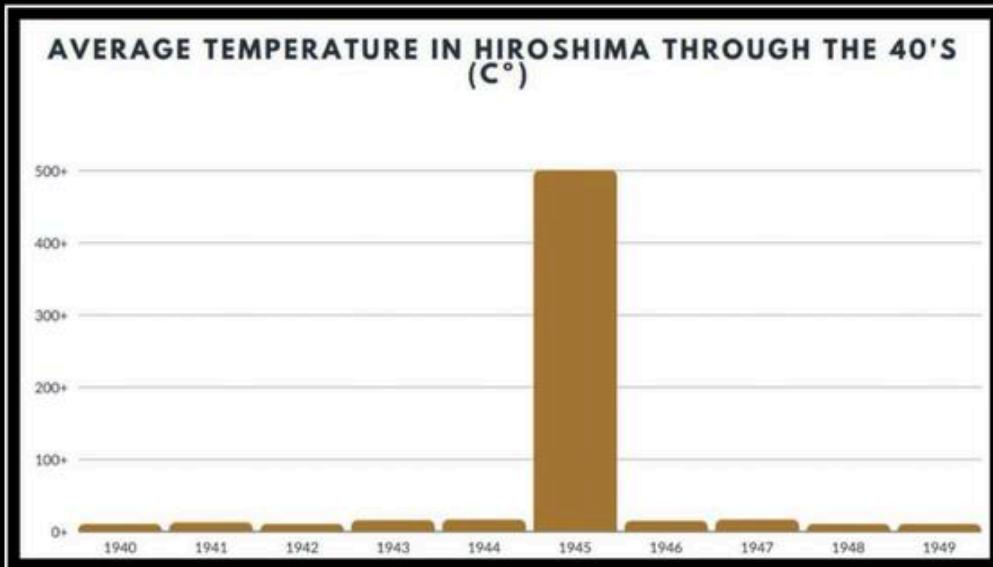
errors

NaN

outliers

inconsistent
data

Przygotowanie Danych



GLOBAL WARMING

<https://knowyourmeme.com/photos/2104681-okbuddyretard>

Etapy czyszczenia danych

- ✓ ogarnięcie NaN-ów
- ✓ poprawa niespójności (np. Male/M/man)
- ✓ usunięcie duplikatów
- ✓ usunięcie zbędnych kolumn
- ✓ poprawa typów danych (float, datetime)
- ✓ ogarnięcie outlierów
- ✓ walidacja logiczna
- ✓ normalizacja/standaryzacja



Przygotowanie Danych

STANDARDYZACJA



średnia = 0

odchylenie = 1

Standaryzacja Danych

Aby sprowadzić dane do rozkładu normalnego:

- Wczytujemy dane,
- Z całego przedziału danych odczytujemy średnią i odchylenie standardowe,
- Wykonujemy standaryzację: $S = (\text{dane} - \text{średnia}) / \text{odchylenie}$.

Aby przywrócić dane do stanu przed standaryzacją:

- Wczytujemy parametry użyte w standaryzacji
- $\text{dane} = S * \text{odchylenie} + \text{średnia}$

Standaryzacja Danych

```
import numpy as np
import numpy.typing as npt

# dane = []
# for _ in range(10):
#     dane.append(np.random.randint(-4,10))

# Generujemy losowe liczby całe z zakresu [-4; 10)
dane: list[int] = [np.random.randint(-4, high=10) for _ in range(10)] # ← Komprehensja list. Analog w komentarzu wyżej
dane_array: npt.NDArray = np.array(dane) # Konwersja na np.NDArray

def standaryzacja(dane: npt.NDArray) → tuple[npt.NDArray, tuple[np.float64, np.float64]]: 1 usage
    srednia: np.float64 = dane.mean()
    odchylenie: np.float64 = dane.std()
    return (dane - srednia) / odchylenie, (srednia, odchylenie)

def przywrocenie_standaryzacji(standaryzowane: npt.NDArray, srednia: np.float64, odchylenie: np.float64) → npt.NDArray:
    return standaryzowane * odchylenie + srednia

print(f"Dane wejściowe: {dane_array}")
standaryzowane, parametry = standaryzacja(dane_array)
print(f"Dane standaryzowane: {standaryzowane}\nŚrednia: {parametry[0]}\nSTD: {parametry[1]}")
przywrocone_dane = przywrocenie_standaryzacji(standaryzowane, *parametry)
print(f"Przywrócone dane: {przywrocone_dane}")
```

Przykład standaryzacji, Jakub Susoł

EDA (Exploratory Data Analysis)



CEL:
zrozumienie rozkładu,
zależności i problemów

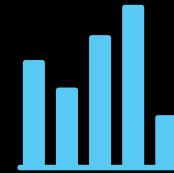


EDA

(Exploratory Data Analysis)



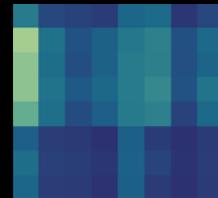
statystyki opisowe



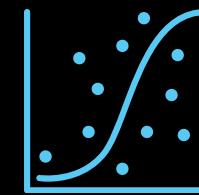
rozkłady, histogramy



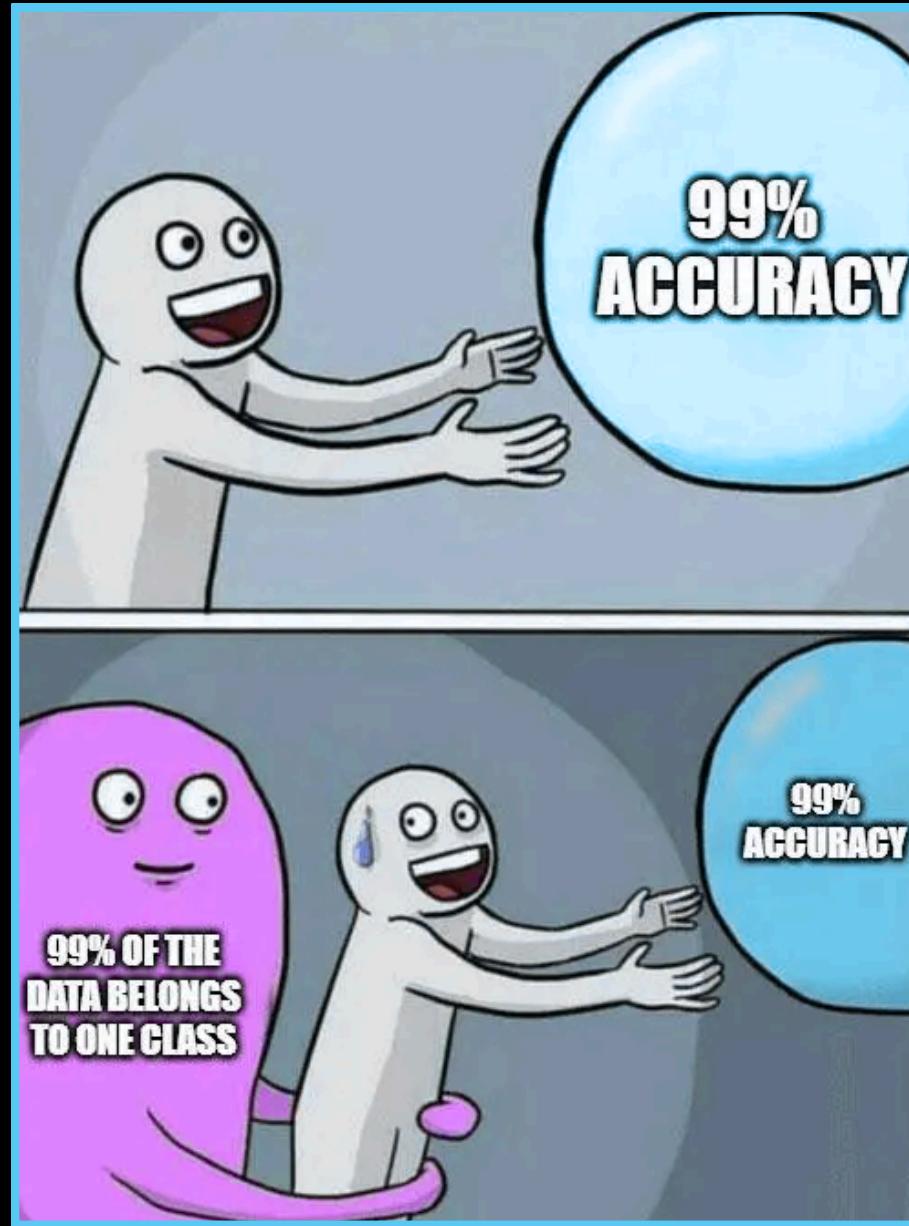
rozkład klas



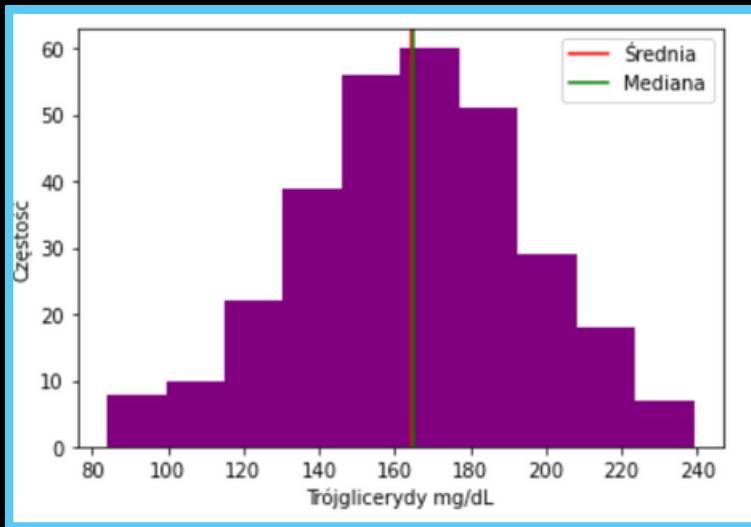
korelacje



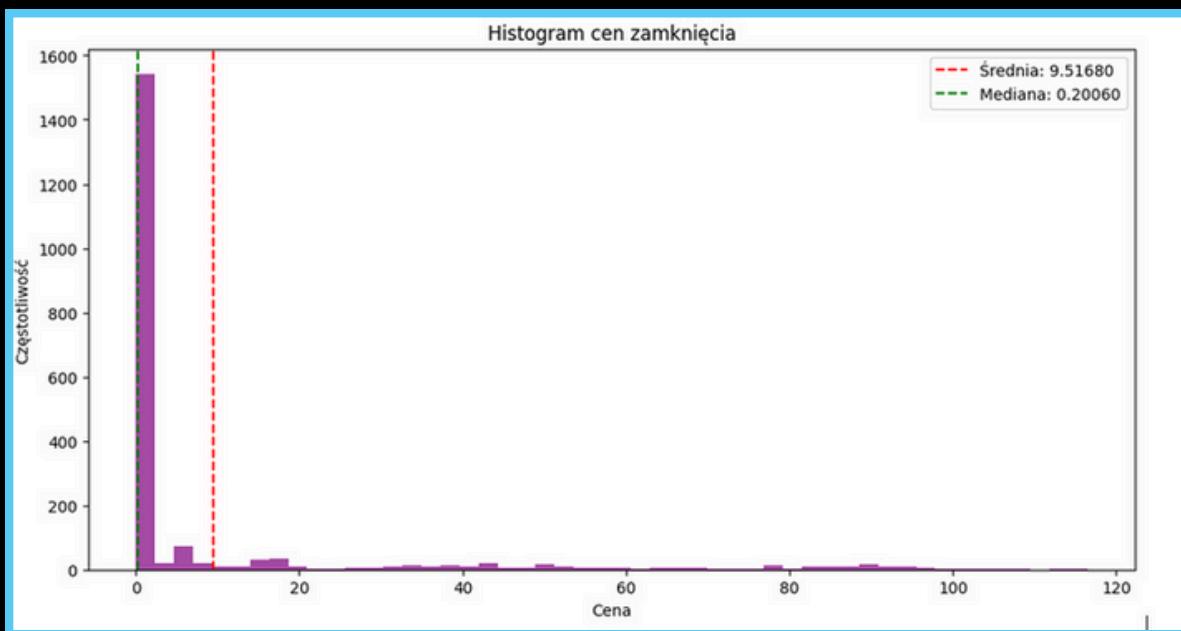
wizualizacje outlierów
(boxplot, scatter plot)



Rozkład



Czy jest
normalny?



Jak wielki powinien być zbiór danych?



Krótka odpowiedź brzmi: Tak.

Overfitting w akcji, a.k.a halucynacje GPT

The screenshot shows a dark-themed chat interface with a light blue border. It contains four messages:

- AG: How many letters are there in the word "nineteen"?
- AI: There are eight letters in the word "nineteen".
- AG: are you sure?
- AI: I apologize for my mistake. There are actually nine letters in the word "nineteen". Thank you for pointing that out.

Each message has a small teal square icon with white text ('AG' or AI) and a small circular icon with a white swirl logo to its left. On the right side of each message, there are three small icons: a trash bin, a thumbs up, and a thumbs down.

<https://www.decipherzone.com/blog-detail/chat-gpt-memes>

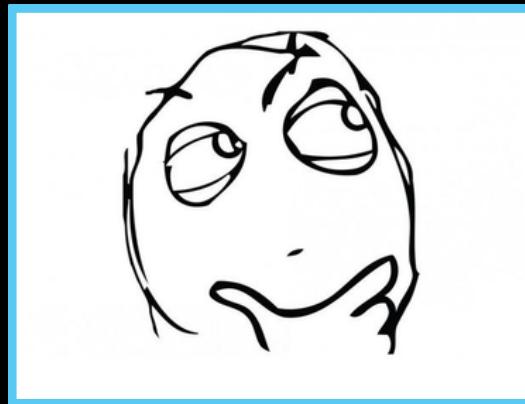


**garbage in,
garbage out**

Amazon's AI Recruiting Tool: seksizm wbudowany w dane



Amazon's AI Recruiting Tool: seksizm wbudowany w dane



Czy to wina modelu czy danych?

Czy można było to naprawić?

Czy rekrutacja przez AI może być w ogóle sprawiedliwa?

COMPAS

sądowy system oceny ryzyka (USA)



używany do oceny czy oskarżony
może trafić na zwolnienie przed procesem

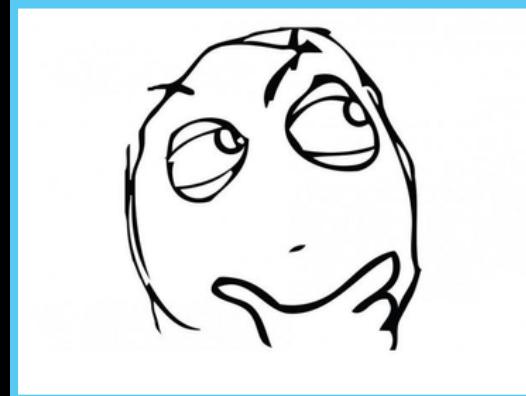
oceniał „ryzyko recydywy”

COMPAS



Czarnoskórych 2x częściej oznaczano jako „wysokiego ryzyka”,
mimo że nie wracali do przestępstwa.

COMPAS



Czy sędzia powinien ufać AI bardziej niż własnemu doświadczeniu?

Czy sprawiedliwość może być statystyczna?

Czy AI mogłoby być bardziej obiektywne niż sędziowie?

