

POLITECHNIKA WROCŁAWSKA
WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI

METODY ANALIZY I EKSPLORACJI DANYCH

Wykład 1 - Wprowadzenie

DR INŻ. AGATA MIGALSKA



Wykład
1

O MNIE



- Doktorat w dziedzinie informatyki obroniłam w 2018 na PWr
- Pracuję jako Senior Machine Learning Researcher w Alphamoon
- oraz na PWr jako adiunkt.
- Prywatnie uwielbiam góry i wspinaczkę.



KONSULTACJE



C-16 P1.2
Dzień/godzina do
ustalenia



AGATA.MIGALSKA
@PWR.EDU.PL



<https://bit.ly/3SHH1JC>



PYTANIA DO PUBLICZNOŚCI


- Jak się nazywasz?
- Czego oczekujesz od tego kursu? Czego chcesz się nauczyć na tym kursie?
- Czy pracujesz, jeśli tak to gdzie i w jakim charakterze?
- Czy interesuje Cię praca związana z tematyką tego kursu?





ORGANIZACJA KURSU




ZAWARTOŚĆ TEMATYCZNA KURSU


Wprowadzenie 


Wizualizacja
danych 

Statystyki
opisowe w
analizie danych 

Analiza jakości
danych 

 Redukcja
wymiaru 

Klasteryzacja
danych 

Granulacja
danych 

Predykcja 



ZAWARTOŚĆ TEMATYCZNA KURSU C.D.



Metody
eksploracji
danych



Analiza danych
strumieniowych



Planowanie
eksperymentu



Dane
geolokacyjne



Dane tekstowe



Kompresja
danych



Podsumowanie.
Zaliczenie kursu



WARUNKI ZALICZENIA

- Kolokwium zaliczeniowe
 - Zaliczony projekt
 - Ocena końcowa = $0.5 * \text{Ocena z projektu} + 0.5 * \text{Ocena z kolokwium}$
-



LITERATURA

LITERATURA PODSTAWOWA:

1. T. Morzy, Eksploracja danych. Metody i algorytmy, Wydawnictwo Naukowe PWN 2022
2. P. Cichosz, Systemy uczące się, WNT, 2000
3. Hand David, Mannila Heikki, Smyth Padhraic, Eksploracja danych, WNT, Warszawa 2005
4. D. T. Larose, Metody i modele eksploracji danych, Wyd. Nauk. PWN, Warszawa 2008

LITERATURA UZUPEŁNIAJĄCA:

1. M. J. Zaki, M. Wagner Jr.. Data mining and analysis. Fundamental Concepts and Algorithms. Cambridge University Press 2014.
<https://www.webpages.uidaho.edu/~stevel/517/Data%20Mining%20and%20Analysis%20by%20Zaki.pdf>
 2. T. Hastie, R. Tibshirani, J. Friedman. Elements of Statistical Learning. Springer Verlag 2009.
<https://hastie.su.domains/ElemStatLearn/>
 3. G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer 2021. <https://www.statlearning.com>
-



METODY ANALIZY I EKSPLORACJI DANYCH

Wprowadzenie

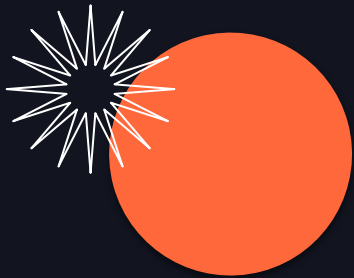
DANE

01

**STRUKTURA
PROJEKTU
ANALITYCZNEGO**

**ANALIZA
DANYCH**

**EKSPLORACJA
DANYCH**



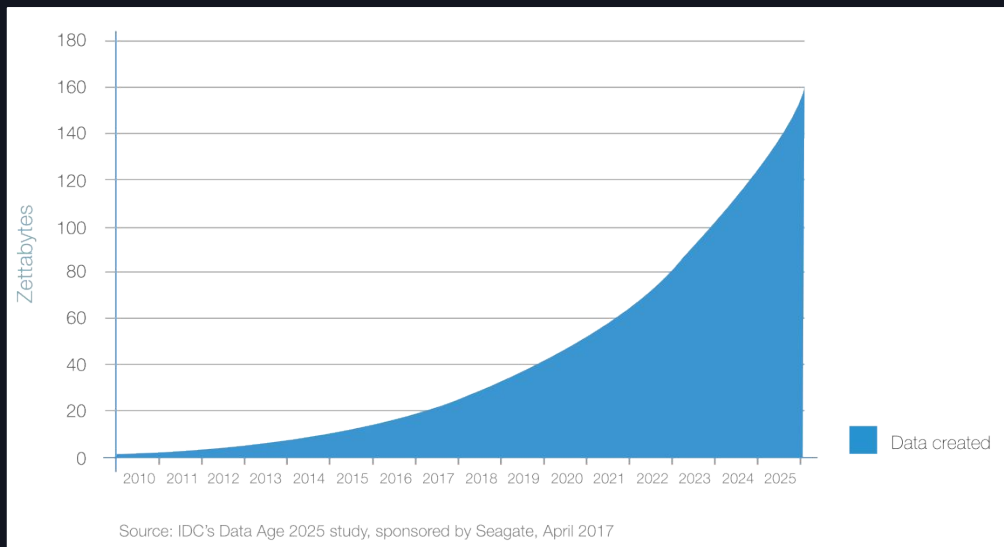
01

DANE

ROCZNA WIELKOŚĆ GLOBALNEJ DATASFERY

Oszacowanie z roku 2017 wg IDC

(1ZB = 10^{21} B)



44ZB

Ilość danych, które miały
powstać w 2020 wg
szacunków z 2017 r.

64ZB

Ilość danych, które
powstały w 2020 wg
szacunków z 2021 r.

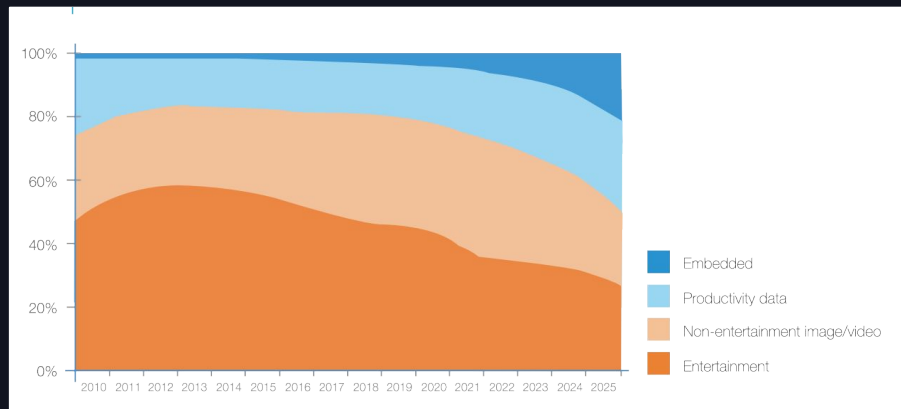
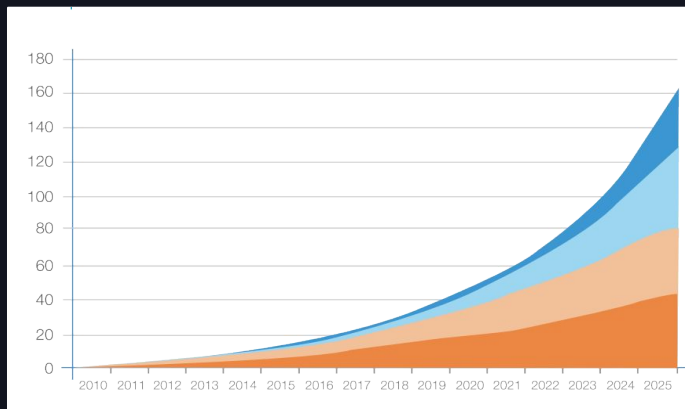
163ZB

Ilość danych, które
powstaną w 2025 wg
szacunków z 2017 r.

ŹRÓDŁA DANYCH

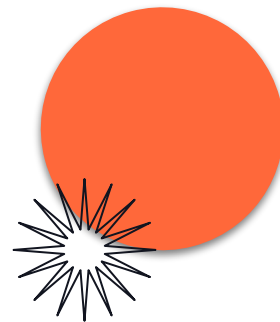
| | |
|--|---|
| PRODUKCYJNE | Pliki, metadane, logi pochodzące z komputerów osobistych, serwerów, telefonów, tabletów. |
| OSADZONE (EMBEDDED) | Dane stworzone przez urządzenia osadzone (embedded) i IoT np. automatyka budynkowa, urządzenia do noszenia, implanty medyczne, itd. |
| ROZRYWKA | Obrazy i video stworzone z myślą o celach rozrywkowych |
| OBRAZY / VIDEO NIE DLA CELÓW ROZRYWKOWYCH | Obrazy i video stworzone z myślą o celach innych niż rozrywkowe, np. kamery ochrony, reklamy |

TWORZENIE DANYCH WG ŹRÓDŁA



Źródło: Badanie "Data Age 2025" przeprowadzone przez IDC, sponsorowane przez Seagate. Kwiecień 2017.

TYPY DANYCH



DANE






DANE USTRUKTURYZOWANE

- dane uporządkowane w sposób umożliwiający niezawodną identyfikację poszczególnych stwierdzeń faktów oraz ich wszystkich składników
 - można je przedstawić w formie tabelarycznej, w której mogą być przechowywane w relacyjnej bazie danych
 - liczby, daty i tekst
 - wg Gartner stanowią ok 20% danych w przedsiębiorstwach
 - wymagają niewiele przestrzeni dyskowej
 - istnieją ustandaryzowane sposoby zarządzania i zabezpieczania takich danych
-

DANE NIEUSTRUKTURYZOWANE

- nie mają wstępnie zdefiniowanego modelu danych lub nie są zorganizowane we wstępnie zdefiniowany sposób
 - nie da się w łatwy (nieprzetworzony) sposób przedstawić ich w formie tabelarycznej
 - obrazy, audio, video, emaile, sformatowane pliki tekstowe, arkusze kalkulacyjne
 - wg Gartner stanowią aż 80% danych w przedsiębiorstwach
-

PRZYKŁADY

| | DANE USTRUKTURYZOWANE | DANE NIEUSTRUKTURYZOWANE |
|---|---|--|
|  ECOMMERCE | <ul style="list-style-type: none">• katalog produktów• ceny• dane klienta | <ul style="list-style-type: none">• zachowanie klienta i wzorce wydawania pieniędzy• zadowolenie klienta z usługi |
|  SŁUŻBA ZDROWIA | <ul style="list-style-type: none">• formularze dla pacjentów• dane o ubezpieczeniu• dane do płatności | <ul style="list-style-type: none">• zdjęcia RTG i tomografia komputerowa• notatki z wizyt• zalecenia lekarskie |
|  BANKOWOŚĆ | <ul style="list-style-type: none">• operacje finansowe• dane klienta | <ul style="list-style-type: none">• logi z rozmów telefonicznych• nagrania audio i video z komunikacji pomiędzy klientami a bankiem |

PO CO TO WSZYSTKO?



DANE

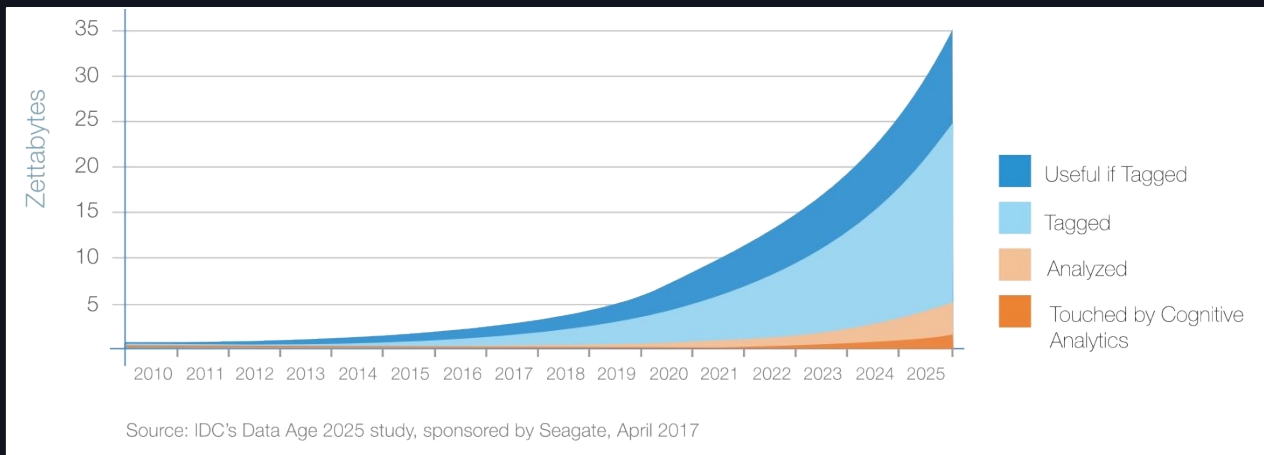


INFORMACJA



WIEDZA

CZERPANIE WIEDZY Z DANYCH NIEUSTRUKTURYZOWANYCH



15%

danych będzie otagowanych
do końca 2025

20%

spośród otagowanych danych
zostanie przeanalizowanych

PRZYSZŁE TRENDY

Zwiększenie automatyzacji w
eksploracji danych

Eksploracja danych
bezpośrednio na
urządzeniach mobilnych

Wzrost eksploracji danych
przestrzennych i
geograficznych

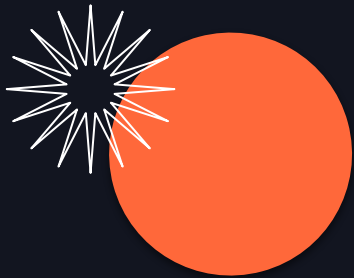


Konsolidacja dostawców
narzędzi eksploracji danych

Wszechobecna eksploracja
danych

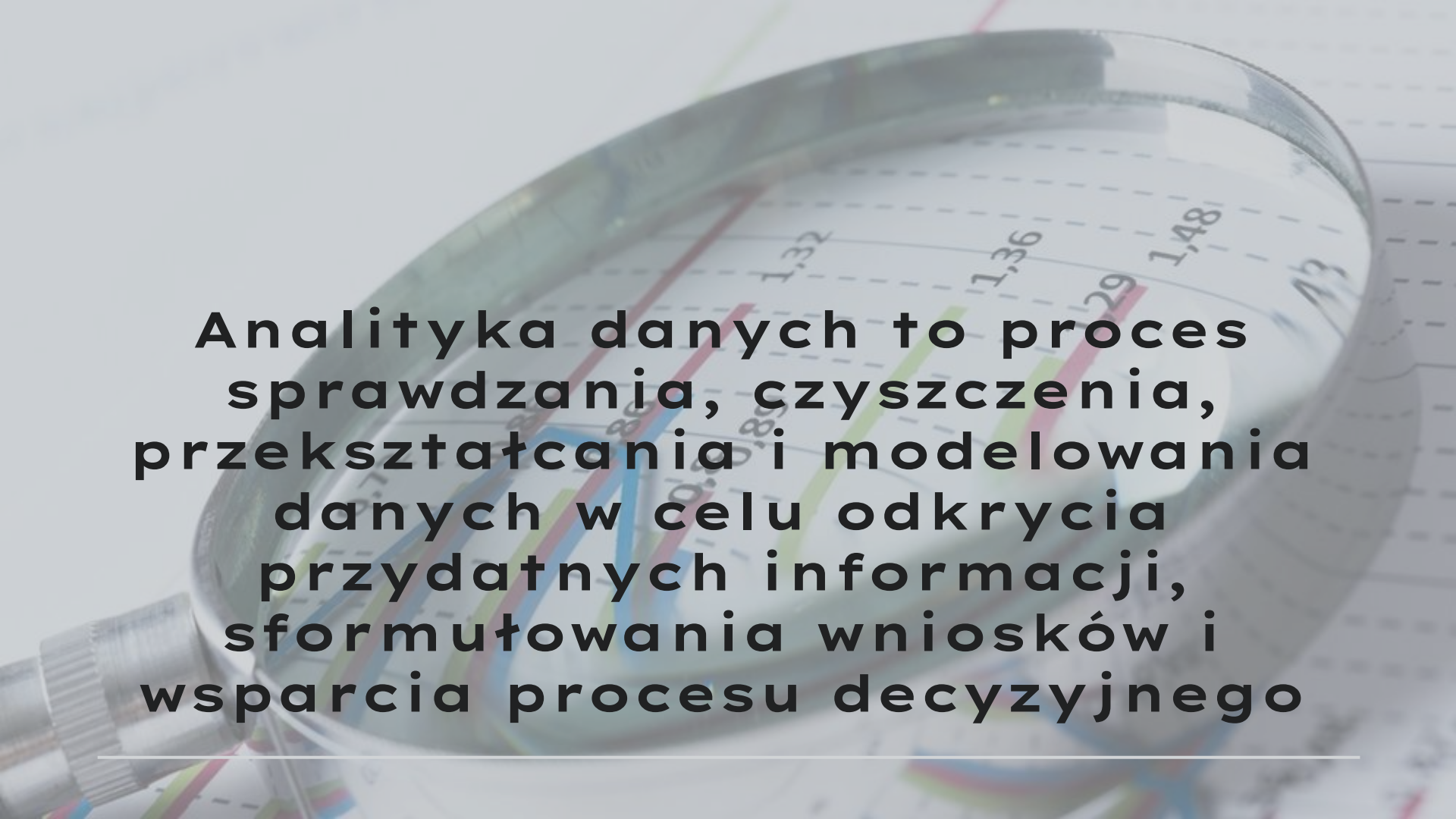
Wzrost eksploracji
multimediów

Dominacja analizy danych w
opiece zdrowotnej i
farmaceutyce



2022

ANALITYKA DANYCH (DATA ANALYTICS)

A magnifying glass is positioned over a line graph. The graph features several colored lines (red, green, blue) and data points labeled with numerical values such as 1.32, 1.36, 1.48, 1.29, and 1.13. The background is a light gray grid.

**Analityka danych to proces
sprawdzania, czyszczenia,
przekształcania i modelowania
danych w celu odkrycia
przydatnych informacji,
sformułowania wniosków i
wsparcia procesu decyzyjnego**

RODZAJE ANALITYKI

Dojrzałość analityczna
przedsiębiorstwa

**ANALITYKA
PRESKRYPTYWNA**

Co należy zrobić ?

**ANALITYKA
PREDYKCYJNA**

Co się stanie?

**ANALITYKA
DIAGNOSTYCZNA**

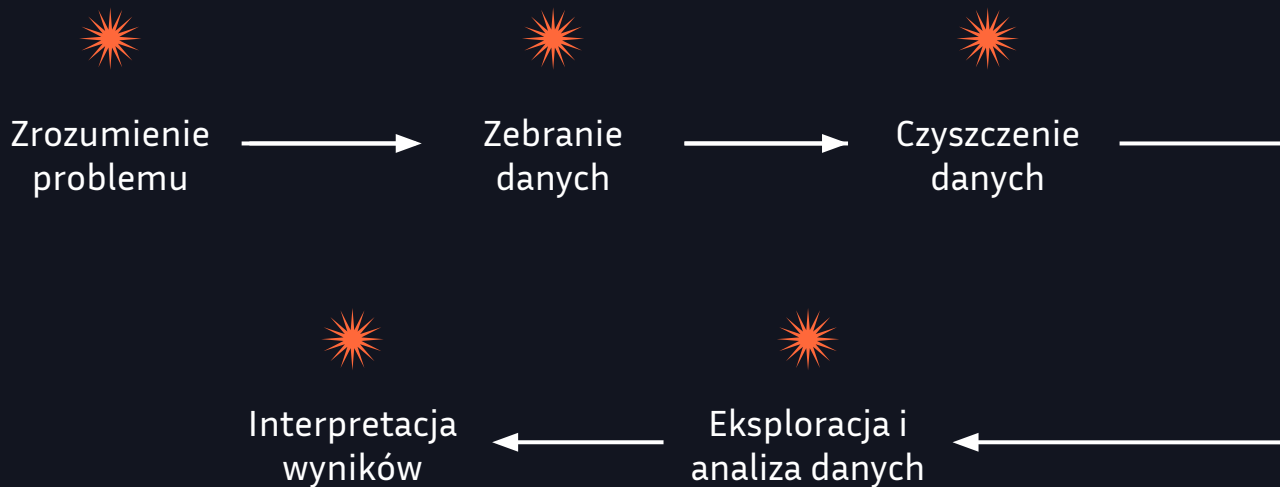
Dlaczego to się stało?

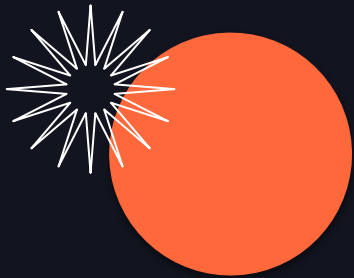
**ANALITYKA
OPISOWA**

Co się stało?

Złożoność

ETAPY PROJEKTU ANALITYCZNEGO



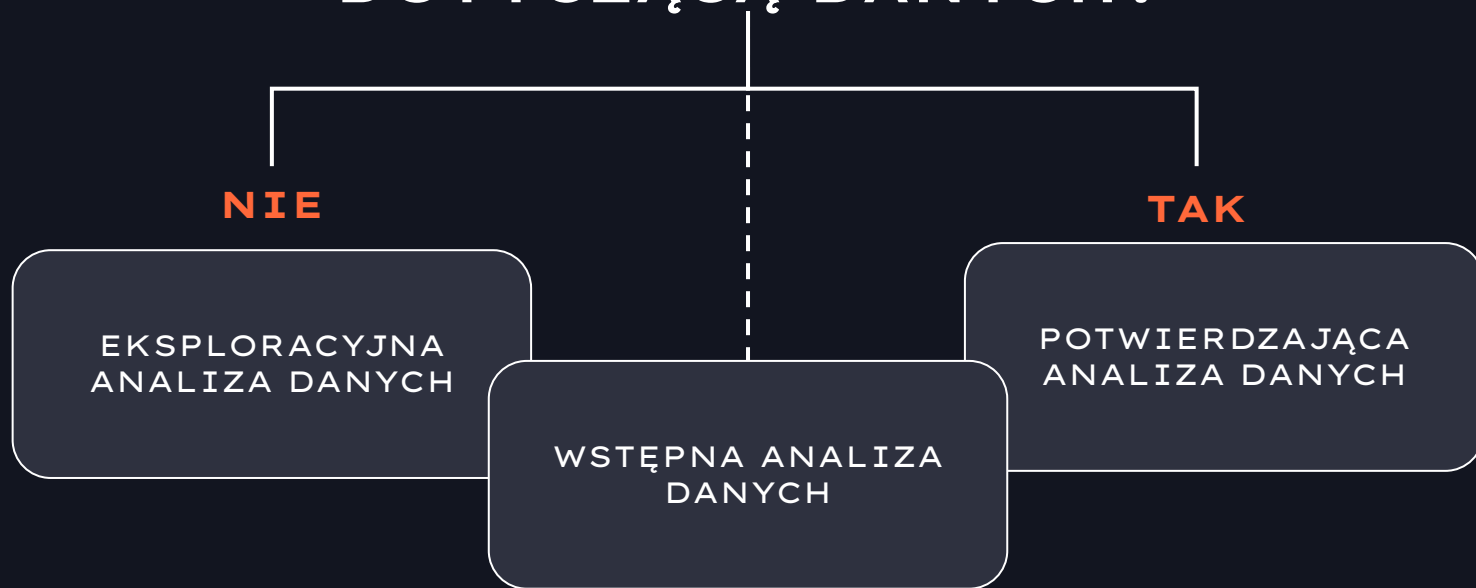


03

ANALIZA DANYCH (DATA ANALYSIS)

ANALIZA DANYCH

CZY MAM HIPOTEZĘ
DOTYCZĄCĄ DANYCH?



WSTĘPNA ANALIZA DANYCH

- Wstępna analiza danych to proces etapów kontroli danych, które należy przeprowadzić po zakończeniu planu badawczego i zebraniu danych, ale przed formalnymi analizami statystycznymi.
 - Celem jest zminimalizowanie ryzyka błędnych lub mylących wyników.
 - Główne kroki:
 - Zidentyfikowanie brakujących lub błędnych danych
 - Ocena, czy pomimo błędów, źródło danych jest wystarczająco dobre, żeby je wykorzystać do przeprowadzenia badania
-

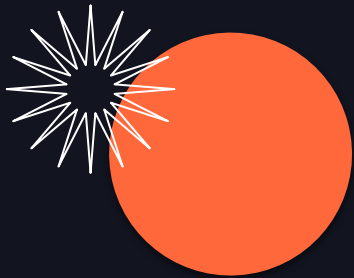
EKSPLORACYJNA ANALIZA DANYCH

- Eksploracyjna analiza danych (ang. Exploratory Data Analysis, EDA) to termin ukuty przez Johna Tukeya (1977).
 - Podejście polegające na graficznej i ilościowej analizie zbiorów danych w celu podsumowania ich głównych cech, często przy użyciu wykresów i innych metod wizualizacji danych.
 - Głównym celem EDA jest:
 - sprawdzenie, co dane mogą nam powiedzieć,
 - sformułować hipotezy do przetestowania.
-

POTWIERDZAJĄCA ANALIZA DANYCH

- Przeprowadzenie testów statystycznych w celu zweryfikowania postawionych hipotez badawczych





04

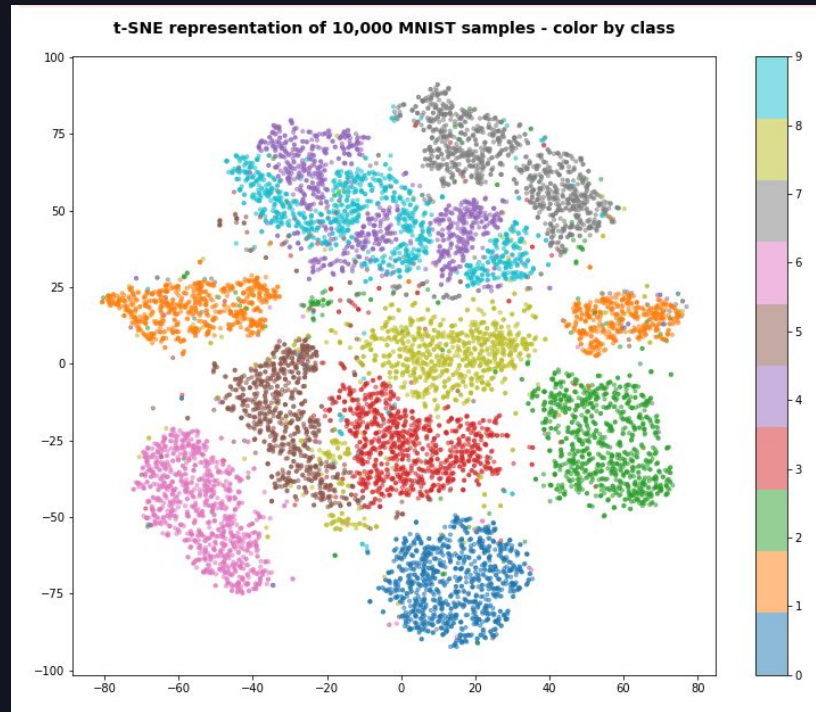
EKSPLORACJA DANYCH (DATA MINING)



**Data Mining to nauka,
sztuka i technologia
eksploracji dużych i
złożonych zbiorów danych
w celu odkrywania
użytecznych wzorców.**

ANALIZA SKUPIEŃ

- Zastosowania:
 - Klastrowanie genów do rodzin
 - Systematyka roślin i zwierząt
 - W tomografii komputerowej służy do rozróżniania typów tkanek
 - Segmentacja konsumentów
 - Analiza sieci społecznych
 - Grupowanie wyników zapytania
 - Segmentacja obrazów
 - ...



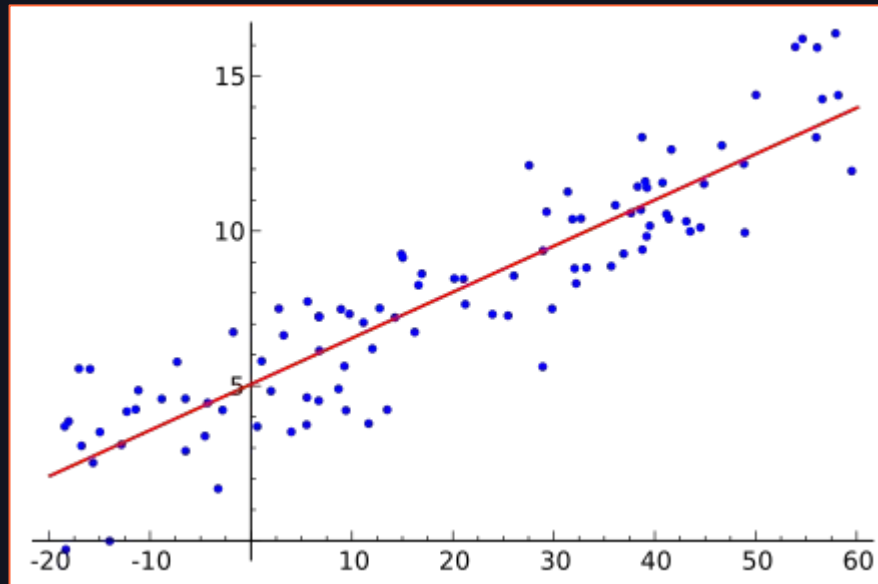
KLASYFIKACJA

- Obszary zastosowań:
 - Klasyfikacja biologiczna
 - Identyfikacja biometryczna
 - Wizja komputerowa
 - Analiza obrazu medycznego i obrazowanie medyczne
 - Rozpoznawanie pisma odręcznego
 - Klasyfikacja dokumentów
 - Rozpoznawanie wzorców
 - Rozpoznawanie mowy
 - ...



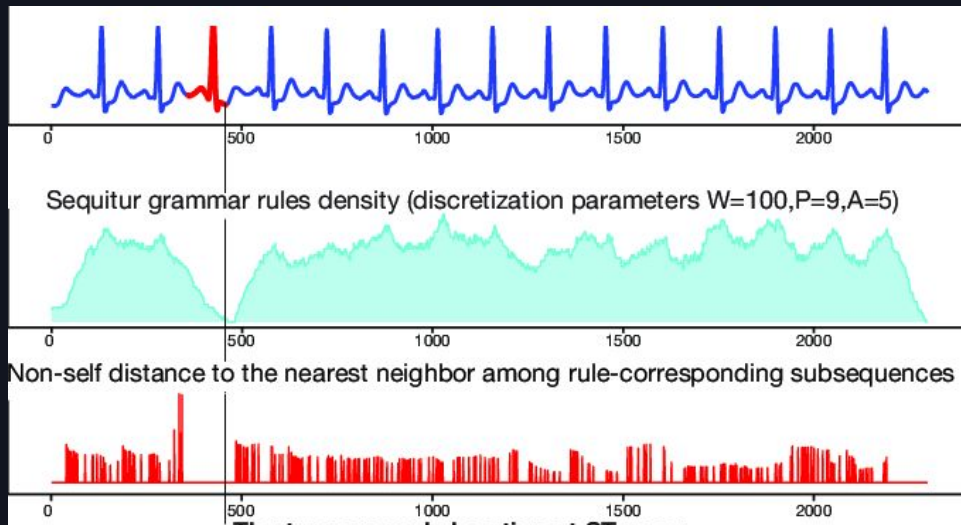
REGRESJA

- Obszary zastosowań:
 - Prognoza popytu
 - Szacowanie kwoty roszczeń z tyt. ubezpieczenia
 - Prognoza rynku i cen
 - Analiza relacji między zmiennymi np. wpływ leków na ciśnienie pacjenta
 - Analiza danych z ankiet
 - Przewidywanie zachowania konsumentów (np. wydana kwota)
 - ...



WYKRYWANIE ANOMALII

- Zastosowania:
 - Wykrywanie włamań w cyberbezpieczeństwie
 - Wykrywanie oszustw
 - Wykrywanie usterek
 - Wykrywanie zdarzeń w sieciach czujników
 - Wykrywanie defektów w obrazach za pomocą wizji maszynowej
 - Diagnostyka medyczna
 - Egzekwowanie prawa



EKSPLORACJA WZORCÓW SEKWENCYJNYCH

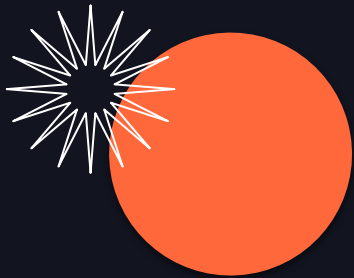
- Zastosowania:
 - Sekwencje zakupowe klientów
 - Leczenie medyczne
 - Katastrofy naturalne (np. trzęsienia ziemi)
 - Procesy naukowe i inżynierskie
 - Ceny akcji / surowców
 - Sekwencja kliknięć na stronie
 - Sekwencje operacji wykonywanych w programie komputerowym
 - Sekwencje biologiczne np. DNA

| SID | Sequence |
|-----|---|
| 1 | $\langle \{a, b\}, \{c\}, \{f, g\}, \{g\}, \{e\} \rangle$ |
| 2 | $\langle \{a, d\}, \{c\}, \{b\}, \{a, b, e, f\} \rangle$ |
| 3 | $\langle \{a\}, \{b\}, \{f, g\}, \{e\} \rangle$ |
| 4 | $\langle \{b\}, \{f, g\} \rangle$ |

NAUKA REGUŁ ASOCJACYJNYCH

- Zastosowania:
 - Analiza koszyka rynkowego np. {ziemniaki, cebula} -> {burger}
 - Eksploracja wykorzystywania sieci Web
 - Bioinformatyka
- W przeciwieństwie do eksploracji wzorców sekwencyjnych, nie uwzględnia kolejności elementów ani w ramach transakcji, ani między transakcjami.





05

**CZY PRZEŻYŁ(A)BYŚ
KATASTROFĘ TITANICA?**

THANKS!

**DZIĘKUJĘ
ZA UWAGĘ**

