**Maksym Malichenko**

**221680**

**ZIISS1-3512IO**

**(Nr. Referatu: 28)**

**tytuł: „Wielki Brat czuwa – czyli prywatność w dobie Internetu”   
podtytuł: „Dane: pozyskiwanie, gromadzenie i analizowanie, wykorzystywanie(konsumowanie)”**

**Uniwersytet Ekonomiczny w Krakowie**

**Kierunek: Informatyka Stosowana**

**Praca wykonana w ramach zajęć z przedmiotu Współczesne Trendy Informatyki**

**Prowadzący: mgr Aleksandra Bojda**

Spis treści

* [Komentarz autora](#komentarz)
* [Dane: pozyskiwanie, gromadzenie i analizowanie, wykorzystywanie](#main)
  + [Czym dane są, skąd się biorą i dokąd zmierzają](#part1)
  + [Pozyskiwanie danych](#part2)
  + [Gromadzenie i analiza danych](#part3)
  + [Wykorzystanie(konsumowanie) danych](#part4)
* [Pytania do referatu](#qa)
* [Literatura, odwołania i cytaty](#literatura)

Komentarz autora  
odnośnie zagadnień które będą poruszone w opracowaniu, a także podanie definicji niektórych pojęć przez autora.

Pojęcie „prywatności” jest bardzo obszerne, więc zdecydowałem skupić się na podtemacie pt. „Jakie dane o nas są dostępne w sieci”.

Od razu chciałbym podać dwie definicji:

1. Dane o nas – wszelkie dane zebrane o nas:
   * + Dane osobowe (np. w definicji Komisji Europejskiej,

<commission.europa.eu/law/law-topic/data-protection/reform/what-personal-data_pl>)

* + - Dane o gustach (muzycznych, politycznych, kulinarnych itp.)
    - Dane o lokalizacji, kontaktach z innymi ludźmi, wydatkach
    - Inne

Wyżej wymienione rodzaje danych są tylko przykładami tego, co z punktu widzenia autora może być postrzegane jako dane. Z punktu widzenia autora wszystkie dane są wrażliwe.

1. Dostępne w sieci – takie które mogą być znalezione w sieci, bądź takie, które były między innymi za jej pomocą pozyskane.

Prywatność będzie utożsamiana z niezbieraniem i nieprzechowywaniem danych, niżej są wymienione wyjątki od tej zasady:

* Dane służące do logowania (login, hasło), bądź PESEL (albo inny identyfikator osobowy) w przypadku dostępu do stron publicznych.
* Dane niezbędne do wykonania usługi (np. dane osobowe podczas zakupów w sieci ),  
  pod warunkiem, iż przestaną być przechowywane zaraz po wykonaniu usługi.

Autor zdaję sobie sprawę, iż taka definicja jest bardzo prymitywna oraz że chronienie prywatności (w znaczeniu zdefiniowanym wyżej) we współczesnych warunkach jest prawie niemożliwe.  
Oprócz tego, chronienie takiego rodzaju prywatności uniemożliwi działanie niektórych serwisów oraz usług sieciowych.  
  
Wszystkie inne definicji prywatności, między innymi definiowane na gruncie prawa, socjologii czy psychologii będą pomijane.  
W opracowaniu zamierzam skupić się na tym, w jaki sposób i jakie dane są pozyskiwane, gromadzone, analizowane i wykorzystywane.  
  
Większość aspektów prawnych będą pomijane.  
Wszelkie aspekty związane ze zbieraniem danych bez użycia sieci będą pomijane.  
Niektóre aspekty związane z

* inżynierią socjalną
* atakami hackerów

z celą dostania się do danych wcześniej nie udostępnianych w sieci będą pomijane.

W opracowaniu zakładam iż dane są zbierane i wykorzystywane przez zbiór podmiotów (korporacje, państwo, osoby fizyczne, biznes z sektora MSP itp.). Dla uogólnienia będę nazywał wszystkie te podmioty „Siecią”, więc może się pojawić np. zdanie „Dane zbierane w Sieci”, co znaczy „dane zbierane przez jeden z podmiotów działających w sieci”.  
Takie uogólnienie zakłada iż dane zebrane w sieci przez jeden z podmiotów są dostępne dla innych. Takie założenie nie odzwierciedla idealnie rzeczywistości, ale ma pod sobą pewne podstawy, które będą omówione w opracowaniu (tzw. „Darknet”, wymiana danymi użytkowników między podmiotami w sieci oraz tzw. „wycieki danych osobowych”).  
  
Dodam też, iż niektóre źródła mogą dostarczać dane kilku osób na raz.   
Np. zdjęcie dwóch kolegów wystawione w sieci przez jednego z nich dostarcza informacji o obu.  
Albo np. wpis o zwycięzcach olimpiady na stronie internetowej szkoły dostarcza informacji odnośnie każdego, kto był wspomniany.  
  
Od razu powiem, że przy wyżej wspomnianych warunkach chronienie prywatności jest prawie niemożliwe, chyba że człowiek nie korzysta z sieci.

Oprócz tego, pracę zamierzam pisać przede wszystkim w oparciu o własne spostrzeżenia i przemyślenia, więc liczba odwołań będzie minimalna.

To powiedziawszy przystępuje do pisania opracowania pod tytułem   
„Dane: pozyskiwanie, gromadzenie, analizowanie i wykorzystywanie”

Wierze, że to w tym zakresie nastąpią największe zmiany w ciągu najbliższych lat, dlatego chcę skupić się właśnie na tym temacie.

**Dane: pozyskiwanie, gromadzenie i analizowanie, wykorzystywanie**

**Czym dane są, skąd się biorą i dokąd zmierzają**

Dane to sygnały (we wszystkich możliwych formach) od pełnego źródła, które całkowicie lub częściowo opisują stan źródła (pomagają odpowiedzieć na pytania: Kto? Co? Jaki? Gdzie? Kiedy? Ile? itp.) . Dane mogą zostać przetworzone na informacje, jeżeli są dostrzegalne i zrozumiałe.

Biorą się z samego faktu istnienia osoby, przedmiotu czy zjawiska. Co ważne, powstają też w wyniku interakcji między nimi. W większości przypadków, są postrzegane przez człowieka (przez wzrok, słuch etc.)

Dane istnieją po to, aby człowiek \*zwierzę? sztuczna inteligencja?\* mógł po ich przeanalizowaniu pozyskać nową informację. Mogą być utrwalane (zapominane bądź zapisywane) oraz przesyłane.

Wyżej podana definicja danych jest uniwersalna, jednakże na potrzeby danego artykułu uważam za stosowne sprecyzować znaczenie tego terminu.

Więc w kontekście prywatności ludźmi w Internecie proponuję taką definicje:

Dane – to fakty i zdarzenia powiązane z użytkownikiem zbierane w Internecie, które po analizie mogą dostarczyć informacji o tym użytkowniku.

W referacie chcę skupić się na czterech najważniejszych procesach powiązanych z danymi:

* pozyskiwaniu
* przechowywaniu
* analizie
* wykorzystaniu

Opis zachodzących zmian w wyżej wymienionych procesach, a także komentarze i wyjaśnienia odnośnie rzeczywistości w której one się odbywają będą stanowiły treść danego referatu.

**Pozyskiwanie danych**

Przede wszystkim warto zaznaczyć, iż dane mogą być pozyskane albo „u źródła” albo „z drugiej ręki”.

Dane pozyskiwane „u źródła” – to takie dane, które dotyczą osoby, od której one były pozyskane.  
  
Żadna nowość iż IT korporacje zbierają o nas najróżniejsze informacje gdzie tylko się da.  
Przyciski w które klikamy, co kupujemy, czego szukamy w wyszukiwarce itp.

*“Facebook can collect data from other browser tabs and pages that children open, and harvest information like which buttons they click on, which terms they search or products they purchase or put in their basket (‘conversions’). There is no reason to store this sort of conversion data, except to fuel the ad delivery system.”*

\*<https://www.theguardian.com/technology/2021/nov/16/facebook-and-instagram-gathering-browsing-data-from-under-18s-study-says>\* 2021/nov/16

Jakby tego było mało dane o wysokości nad poziomem morza, szybkości ruchu czy dane o paleniu też są zbierane.

*“AppMetrica is taking in some very precise location data. It’s not that uncommon for app analytics to take in latitude/longitude so that growth and marketing teams can see where the app is being used. What is less normal is taking in a user’s altitude, direction, and speed, which together with a timestamp, provides a disturbingly accurate picture of a user’s movements.”*

*“Here are some examples of the segments Crypta generates: The Smokers segment seems to track users who purchase specific smoking products, like e-cigarettes and tobacco. Summer residents uses geolocation data to track which users have dachas (summer homes) and how often they visit them.”*

\*<https://www.confiant.com/news/the-yandex-leak-how-a-russian-search-giant-uses-consumer-data>\* Aug 10, 2023  
(Polecam przeczytania całości artykułu, jest on względnie krótki, a jednocześnie pozwala wyobrazić sobie ilość i rozmaitość danych zbieranych przez IT gigantów. Znacznie ułatwi rozumienie tego, dlaczego temat danych jest obecnie taki ważny)

Przewiduje, iż w kolejnych latach ilość i rozmaitość danych zbieranych przez różne podmioty będzie rosła.

Chciałbym krótko omówić te obszary, w których z mojego punktu widzenia nastąpią największe zmiany.

Pierwszym takim obszarem jest rzeczywistość wirtualna. Obecnie VR okulary może pozwolić sobie prawie każda pracująca osoba w krajach rozwiniętych. Infrastruktura, łatwość wykorzystania oraz oferta dostępnych aplikacji w ciągu ostatnich kilku lat uległy znacznemu ulepszeniu. Nic za tym dziwnego, że co raz więcej osób korzysta z VR.

A graph of a number of users

Description automatically generated

\*<https://www.zippia.com/advice/virtual-reality-statistics/> \* Feb. 27, 2023

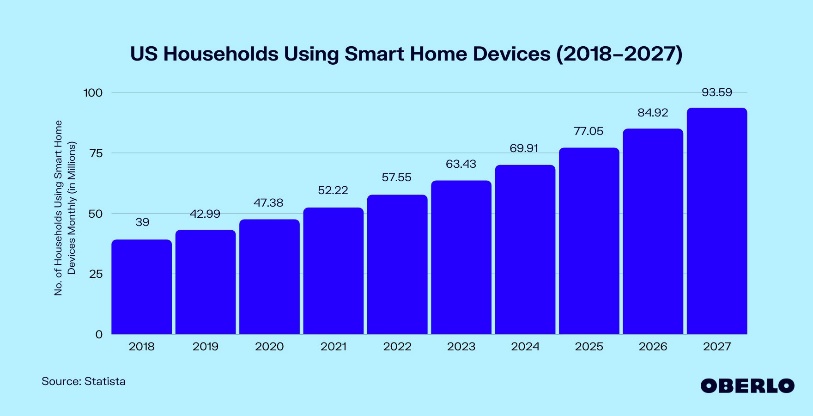
Jakie dane będą mogli zbierać VR urządzenia.

Wzrost człowieka, wyposażenie wnętrza pokoju, leworęczność czy praworęczność, preferencje rozrywkowe itp.  
Biorąc pod uwagę szybkość rozpowszechniania VR, w niedalekiej przyszłości ta technologia może stać się częścią naszej codzienności, i zacząć zbierać o nas dane, zebranie których dotychczas było niemożliwe.

Drugim obszarem wartym uwagi są różne narzędzia które w połączeniu między sobą tworzą systemy domów inteligentnych: różnorodne czujniki, asystenty głosowe (Siri, Alexa itp.), roboty sprzątające, artykuły AGD w wersji „Smart” (przede wszystkim lodówki, ale też pralki, ekspresy do kawy, tostery, czajniki itp.)

One zbierają dane na przykład o naszych preferencjach żywnościowych, o czasie obudzenia czy o wnętrzu mieszkania. W połączeniu między sobą mogą zbierać dane prawie o każdym aspekcie naszego bytu domowego.

Zgodnie z przewidywaniami „Statista” w 2027 roku niecałe 100 mln gospodarstw domowych w Stanach będą korzystali z urządzeń „Smart”.



\*<https://www.oberlo.com/statistics/smart-home-statistics>\*

Kolejnym ważnym obszarem są „Large Language Models”. Mają one bardzo szeroki zakres zastosowań, ale moim zdaniem najważniejszym z punktu widzenia zbierania danych jest:

Wykorzystanie ich jako wyszukiwarek o szerszym (i nieco innym) zakresie możliwości.

Dlaczego to jest tak ważne? Dlatego że użytkownicy zaczynają zadawać pytania, które dotychczas nie były zadawane w wyszukiwarkach tradycyjnych. Możliwość otrzymania prostej i szybkiej odpowiedzi \*bez wysiłku własnego\* na czasem skomplikowane pytania, to coś, czego konwencjonalne wyszukiwarki nie mogły zaoferować. To pozwala na zbieranie danych nieosiągalnych dotychczas, dzięki którym można głębiej zrozumieć zainteresowania i problemy użytkowników. Nić dziwnego, że niektóre wyszukiwarki już mają wbudowane w sobie duże modele językowe (np. Microsoft Bing), a inne przygotowują się do tego.

Z własnego doświadczenia zauważyłem, że co raz więcej osób z mojego otoczenia korzysta z np. „Chat-GPT” w celu wyszukiwania informacji (zdaję sobie sprawę, iż w tym przypadku próba statystyczna nie jest w żaden sposób reprezentatywna).

Nie trzeba zapominać też o smartfonach. To właśnie one w chwili obecnej zbierają o nas najwięcej danych. Mimo tego, nadal nie osiągnęli szczytu swoich możliwości. Dlaczego?  
Z dwóch powodów. Po pierwsze, nie wszystkie dane które powstają w wyniku interakcji z komórką są obecnie zbierane (z powodu braku zainteresowania tymi danymi. Ale to się zmienia, dokładniej omówię to w kolejnej części). Po drugie, nauka się rozwija, a razem z nią ulepszają się stare i powstają nowe sensory. Które z kolei mogą zostać zaimplementowane w nowszych wersjach smartfonów.  
Więc w kolejnych latach czekam na zwiększenie ilości zebranych przez telefony danych. &&&&&&&&

Warto zauważyć, że większość technologii staje się coraz bardziej dostępna z biegiem czasu, dlatego te zmiany dotyczą także najmniej zamożnych grup osób.

Oczywiście, nie są to wszystkie możliwe opcje pozyskiwania danych „u źródła”. Jednak że chciałbym też omówić diametralnie inne podejście. Dane „z drugiej ręki” – to takie, które są pozyskiwane od osób trzecich, tzw. „dostawców danych”. Są tu cztery główne opcję, które jednak nie są sobie sprzeczne i mogą tworzyć rozmaite kombinacje.

Po pierwsze, dane mogą zostać kupione od innej firmy bądź na czarnym rynku (np. Darknet).

Jest to bardzo rozpowszechniana praktyka, zwłaszcza wśród podmiotów gospodarczych, które dopiero zaczynają swoją działalność. Mogą w ten sposób otrzymać przewagę konkurencyjną, bądź, przynajmniej, wyrównać szanse względem konkurencji.

Warto zauważyć, iż danych nie zawsze kupuję się jako towaru, tylko często jako usługę. To znaczy, że sprzedawane są nie rekordy z bazy danych, tylko dane powstałe po ich początkowym przetwarzaniu. Właśnie to podejście jest najczęściej stosowane przez IT korporacje. Temat ten pokrywa się z wykorzystaniem danych, dlatego zostanie omówiony później.

Oprócz tego, dane można kupić nielegalne:

*„Danymi osobowymi najczęściej handlują:*

* *firmy marketingowe,*
* *byli pracownicy banków, firm ubezpieczeniowych, operatorów telekomunikacyjnych,*
* *hakerzy, którzy wykradają dane z komputerów firm oraz osób fizycznych.”*

\*<https://rkrodo.pl/handel-danymi-osobowymi/>\* 21 maja, 2023

Otrzymane w taki sposób dane mogą dotyczyć mas szerokich, wąskiego grona osób czy nawet pojedynczych jednostek. Z reguły są cenniejsze mogą być stosowane do różnych celów, często wykraczających poza strefę interesów gospodarczych.

Po drugie, dane mogą być współdzielone pomiędzy firmami.  
Nie zawsze dochodzi do „sprzedaży” danych, mogą one być wymieniane na podstawie barteru.  
Może też zajść taka sytuacja, iż zebrane dane trafiają od razu do kilku podmiotów gospodarczych (rozpowszechniona sytuacja w przypadku sklepów internetowych).

Po trzecie, dane mogą zostać ujawnione podczas wycieków danych osobowych.

“*A user in a low-level hacking forum on Saturday published the phone numbers and personal data of hundreds of millions of Facebook users for free.*

*It includes phone numbers, full names, locations, email addresses, and biographical* „

\*<https://www.businessinsider.com/stolen-data-of-533-million-facebook-users-leaked-online-2021-4?IR=T> \* Apr 3, 2021

Takie wycieki danych nie są rzadkością, i odbywają się względnie często. Przyczyną tego jest niedbałość o bezpiekę danych zarówno przez użytkowników jak i przez firmy.

Wyżej opisane zjawiska (sprzedaż, wymiana i wycieki danych) sprawiają, że dane, które pojawiły się w sieci, zostają dostępne wszystkim. Dlatego w referacie pomijam temat pod tytułem „kto ma dostęp do danych”.

Dane potrafią generować olbrzymie pieniądze. Duże korporacje już zarabiają na tym, ale większość podmiotów gospodarczych póki co nie przystąpiła do tego. Dlatego w przyszłości spodziewam się większego udziału małych i średnich firm w zbieraniu i sprzedaży/współdzieleniu danych. Platformy do wymiany danymi już powstają. W przyszłości najprawdopodobniej zwiększą skale i poszerzą zakres swojej działalności.

**Gromadzenie i analiza danych**

Skoro temat przechowywania danych jest osobnym obszernym tematem, omówione zostaną tylko niektóre obszary tego zagadnienia.

Z powodu tego że gromadzenie i analiza danych są między sobą mocno powiązane, będę rozpatrywał te 2 procesy łącznie.  
  
Ostatnio dużo się mówi o „Big Data”, a rozmaite kursy oraz oferty pracy „data analyst” często są widoczne w ogłoszeniach w sieci. To świadczy, że kapitalizacja rynku oraz ilość osób w nim zatrudnionych rośnie. To z kolei zwiększy popyt na nowe narzędzia, techniki oraz podejścia do analizy danych. Który prędzej czy później doprowadzi do powstania/wynalezienia owych.   
Oprócz tego, większa ilość siły roboczej znaczy, że więcej danych będzie można przeanalizować.  
  
I tu zaczyna się kolejny temat warty omówienia: dane które wcześniej były „wyrzucane” obecnie zostają zbierane i analizowane. Dlaczego? Dlatego że nie ma bezużytecznych danych – zawsze da się pozyskać z nich jakąś informację. Oprócz tego, mogą pomóc w zinterpretowaniu danych z innych źródeł. Często są przechowywane z nadzieją że przydadzą się później.  
  
Ale skoro musimy płacić za przechowywanie danych musimy się zastanowić czy to się opłaca. Oczywiście, odpowiedź na to pytanie każda firma musi udzielić samodzielne. Jednak, jak się okazuję, cena pamięci z roku na rok się zmniejsza.

A graph of a graph showing the cost of computer memory storage

Description automatically generated

\*<https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage> \*   
2023-03-01

Tak, na przykład, w ciągu ostatnich 9 lat cena na pamięć dyskową spadła trzykrotnie (z 45$ do 14.3$ za TB). Najprawdopodobniej taki trend się zachowa, więc pamięć i przechowywanie danych będą co raz tańsze. Co znaczy, że co raz większa ilość danych będzie możliwa i opłacalna do przechowywania.

Warto też przypomnieć iż w analizie danych od dawna się stosuję sztucznej inteligencji. Skoro przyszłość AI nie jest tematem tego wypracowania powiem tylko, że popularność zagadnienia doprowadzi do zwiększenia ilości specjalistów oraz narzędzi w tym obszarze. Co swoją drogą przyczyni się do rozwoju AI i zwiększeniu efektywności jego użycia w analizie danych.

**Wykorzystanie(konsumowanie) danych**

Jeżeli dane są zebrane i przeanalizowane, czas zadać pytanie, jakie korzyści możemy z tego czerpać.  
Odpowiedź na to pytanie przede wszystkim zależy od tego, kim jesteśmy.  
  
W przypadku gdy jesteśmy małą/średnią firmą bezpośrednio produkującą towary bądź oferującą usługi najważniejsze dla nas pytanie brzmi: „jak ulepszyć swój wynik finansowy”. Jeden za sposobów na to: przewidzieć zapotrzebowanie na własny produkt. Na przykład, salon fryzjerski czy restauracja chcieliby wiedzieć, w jakie dni i w jakie godziny mogą się spodziewać większej ilości klientów. Oprócz tego, chcieliby wiedzieć jakie pozycje w ofercie są najbardziej popularne oraz kim jest ich przeciętny klient. Przeanalizowanie danych umożliwia otrzymanie takiej informacji, a więc pomaga w prowadzeniu biznesu.  
Jakie dane mogą w tym celu pomóc? Na przykład dane o tym, kiedy osoby w okolicy kończą prace, czy mają przerwę obiadową, preferencje żywnościowe itp.

Jeżeli natomiast jesteśmy firmą marketingową bądź zamierzamy wprowadzić na rynek nowy projekt potrzebujemy informacji o tym, czy i kto byłby zainteresowany naszym produktem, w jaki sposób najskuteczniej możemy go promować.   
W tym przypadku, przydadzą się dane o: zainteresowaniach, dobrobycie finansowym, ulubionych celebrytach przyszłych konsumentów, popularności podobnych produktów itp.

Partia polityczna bądź organizacja społeczna? Najprawdopodobniej jesteśmy zainteresowani sterowaniem opinii publicznej. Żeby móc to zrobić musimy wiedzieć, co chcą od nas usłyszeć odbiorcy. Żeby o tym się dowiedzieć warto zbadać kim oni są. W tym celu pomogą nam dane o:   
zamożności, wieku, wyznaniu religijnym, aktywności na portalach społecznościowych, oglądane treści na „YouTube” itp.

Wyżej opisane przypadki są do siebie dość podobne: korzystamy z danych, żeby dostać odpowiedzi na pytania, które nas interesują. Jednocześnie korzysta się raczej z danych zebranych przez kogoś innego.  
Z kolei IT korporacje głównie wykorzystują dane w zupełnie inny sposób. „Przerabiają” ich, wyciągając informacje i nowe dane (zwykle bardziej wartościowe i użyteczne) które później sprzedają. Np. po przeanalizowaniu danych o lokalizacji i aktywności fizycznej można wywnioskować, że osoba uwielbia turystykę pieszą. To, z kolei, cenne dane, które mogą być sprzedane sklepu ze sprzętem sportowym.

Mimo tego, że temat konsumowania danych przez państwo bądź osoby fizyczne jest zjawiskiem względnie rzadkim, nadal jest warty omówienia. W przypadku państwa odpowiedź jest względnie prosta: używa ona danych do zapełnienia bezpieczeństwa i kontroli obywateli.  
Jednak w przypadku osób fizycznych udzielenie odpowiedzi jest znacznie trudniejszym zadaniem. Mogą oni używać danych do różnych oszustw, spraw kryminalnych, postępów w karierze czy do „stalkowania” ludzi. Warto zaznaczyć, że to nie jest wyczerpująca lista możliwych czynności, lecz niektóre przykłady.  
  
Według mojej opinii, wykorzystanie danych to obszar w którym niewiele się zmieni w przyszłości. Oczywiście, wolumen i rozmaitość danych się powiększą, więcej osób i organizacji przystąpią do ich konsumowania. Jednak wszystkie możliwe obszary wykorzystania danych już zostały „odkryte”.

Pytania:

1. Jakie urządzenie zbiera o nas najwięcej danych?

* Smartfon.

1. Czym są dane pozyskane z „drugiej ręki”?

* To są dane, które nie były otrzymane bezpośrednio od podmiotu, którego one dotyczą.

1. Czym wykorzystanie danych przez IT gigantów różni się od konsumowania danych mniejszymi organizacjami?

* IT korporacje dążą do sprzedania przetworzonych danych i informacji, podczas gdy mniejsze organizacje korzystają z danych żeby znaleźć rozwiązania swoich problemów.

**Literatura, odwołania i cytaty**

1. \*<https://www.theguardian.com/technology/2021/nov/16/facebook-and-instagram-gathering-browsing-data-from-under-18s-study-says>\* 2021/nov/16
2. \*<https://www.confiant.com/news/the-yandex-leak-how-a-russian-search-giant-uses-consumer-data>\* Aug 10, 2023
3. \*<https://www.zippia.com/advice/virtual-reality-statistics/> \* Feb. 27, 2023
4. \*<https://www.oberlo.com/statistics/smart-home-statistics>\*
5. \*<https://rkrodo.pl/handel-danymi-osobowymi/>\* 21 maja, 2023
6. \*<https://www.businessinsider.com/stolen-data-of-533-million-facebook-users-leaked-online-2021-4?IR=T> \* Apr 3, 2021
7. \*<https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage> \*   
   2023-03-01