Wstęp do Big Data

iBigWorld:
Innovations for Big Data in a Real World

Prof. dr. Dragan Stojanovic, **UNI**
Prof. dr. Natalija Stojanovic, **UNI**

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data definitions

- “Big data odnosi się do zbiorów danych, których rozmiar przekracza możliwości typowych narzędzi oprogramowania bazodanowego do gromadzenia, przechowywania, zarządzania i analizowania.” The McKinsey Global Institute, 2012
- “Big data to dziedzina, która zajmuje się sposobami analizowania, systematycznego wydobywania informacji lub innego postępowania ze zbiorami danych, które są zbyt duże lub złożone, aby mogły być obsługiwane przez tradycyjne aplikacje do przetwarzania danych.” Wikipedia, 2020

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data definitions

- “Big data polega głównie na zbieraniu liczb i wykorzystywaniu tych liczb do przewidywania przyszłości. Im większy masz zestaw danych, tym dokładniejsze będą prognozy dotyczące przyszłości.” Anthony Goldbloom, Kaggle’s founder
- “Big data to zbiory informacji o dużej objętości, dużej zmienności lub dużej różnorodności, które wymagają nowych form przetwarzania w celu wspomagania podejmowanie decyzji, odkrywanie nowych zjawisk oraz optymalizacji procesów.” Gartner

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data przedrostki

- Międzynarodowy Układ Jednostek

kilo (k)	1,000 (3 zeros)
Mega (M)	1,000,000 (6 zeros)
Giga (G)	1,000,000,000 (9 zeros)
Tera (T)	1,000,000,000,000 (12 zeros)
Peta (P)	1,000,000,000,000,000 (15 zeros)
Exa (E)	1,000,000,000,000,000,000 (18 zeros)
Zetta (Z)	1,000,000,000,000,000,000,000 (21 zeros)
Yotta (Y)	1,000,000,000,000,000,000,000,000 (24 zeros)

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data - liczby

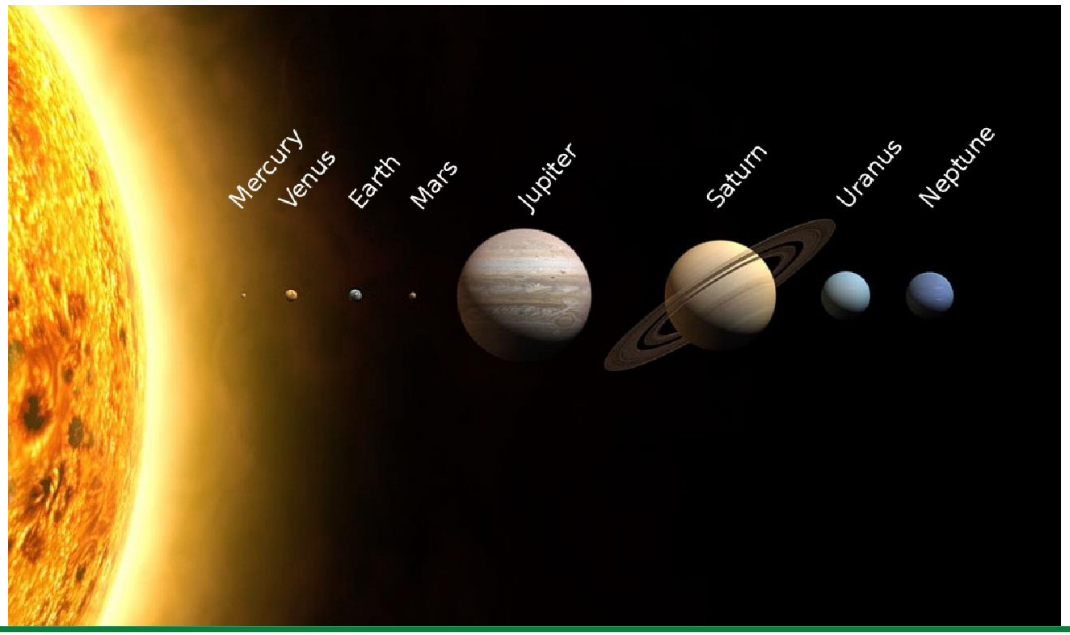
- Ile zgromadzono danych na świecie? Jakie są szacunki?
  - 800 Terabytes, 2000
  - 160 Exabytes, 2006 (1EB = 10^{18}B)
  - 4.5 Zettabytes, 2013 (1ZB = 10^{21}B)
  - 44 Zettabytes by 2020
  - 163 Zettabytes by 2025
- Jak dużo danych generowanych jest w ciągu dnia?
  - 2.5 Exabytes
  - 8 TB, Twitter
  - 50 TB, Facebook
- 90% danych na na całym świecie wygenerowano w ciągu ostatnich 2 lat!

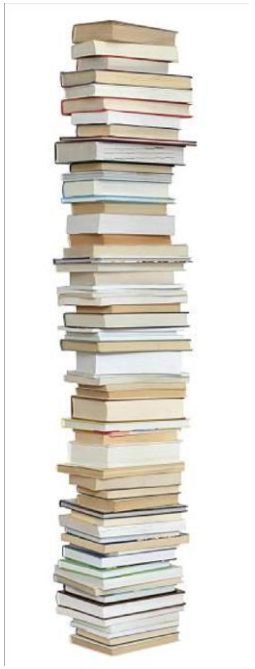
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jak duże są zbiory Big Data?

- Spróbujmy zrobić stos książek zawierający Zettabajt danych
- 1 Zettabyte = stos książek, których wysokość jest PIĘCIOKROTNĄ odległością między Ziemią a Plutonem

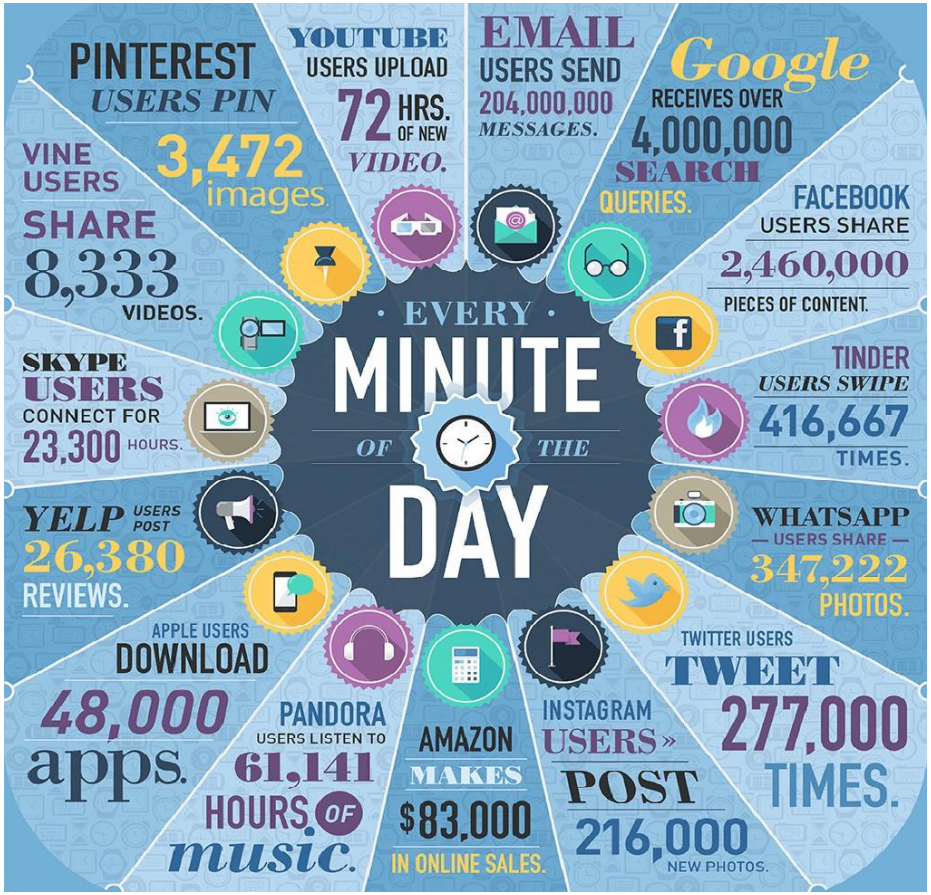


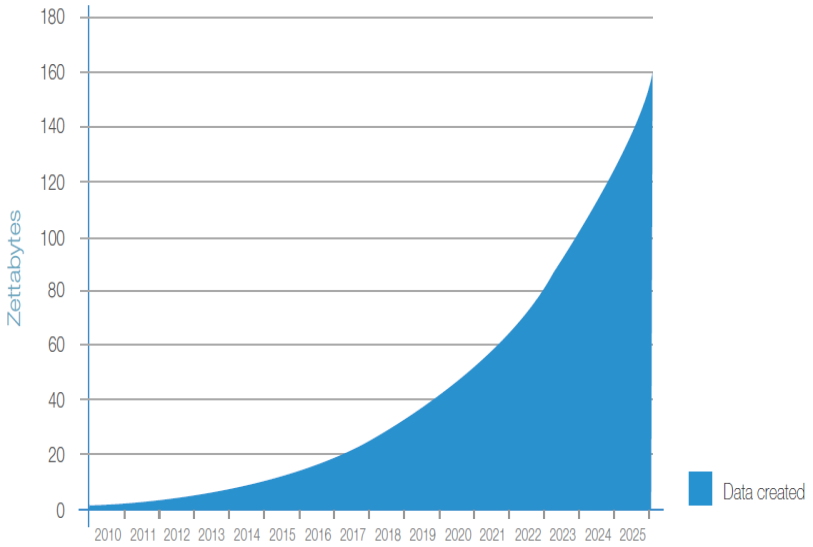


iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Więcej, coraz więcej danych…

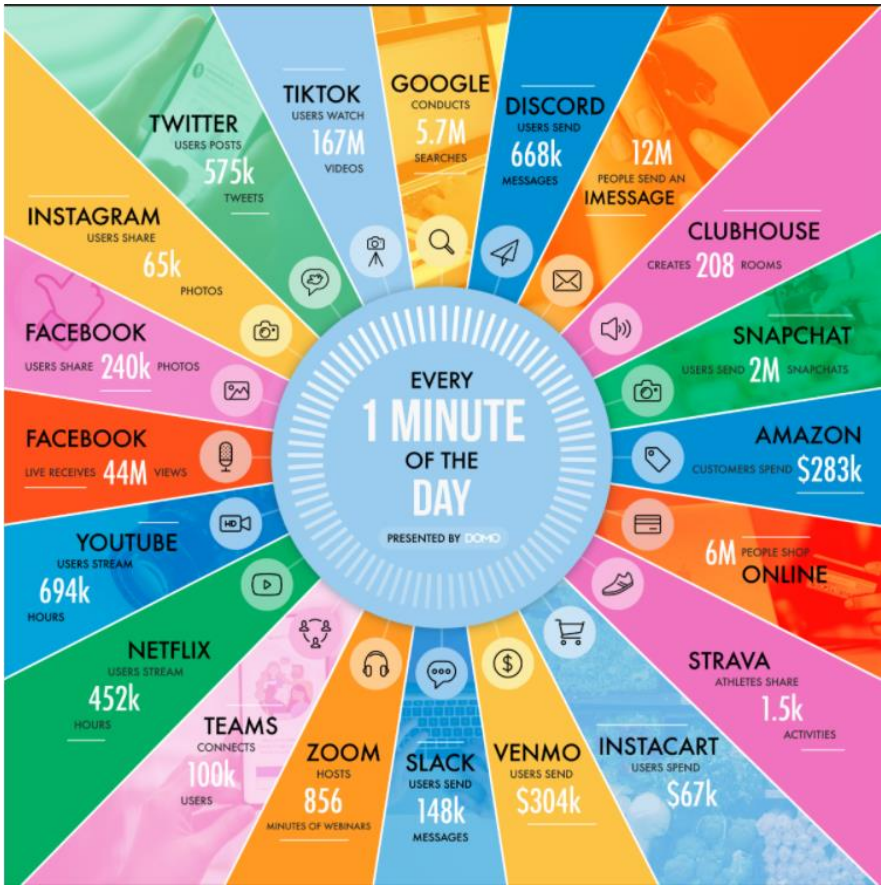




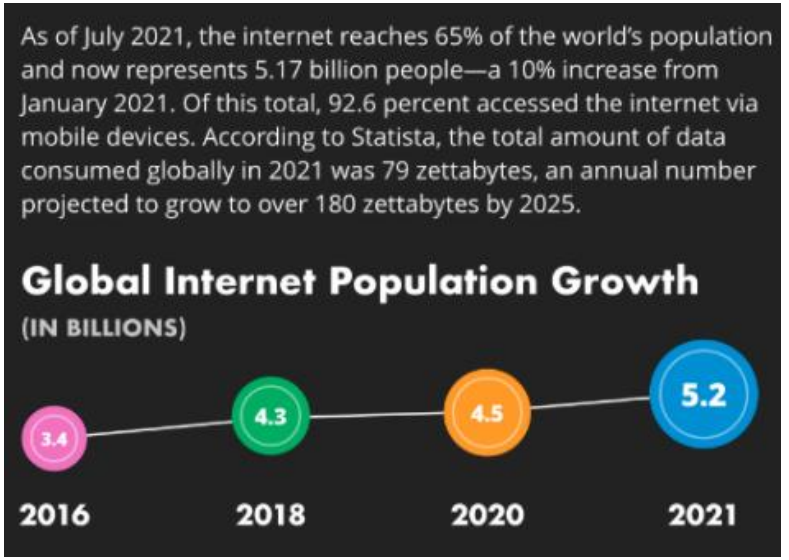
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Dane nigdy nie śpią (*domo.com*)



As of July 2021, the internet reaches 65% of the world's population and now represents 5.17 billion people—a 10% increase from January 2021. Of this total, 92.6 percent accessed the internet via mobile devices. According to Statista, the total amount of data consumed globally in 2021 was 79 zettabytes, an annual number projected to grow to over 180 zettabytes by 2025.

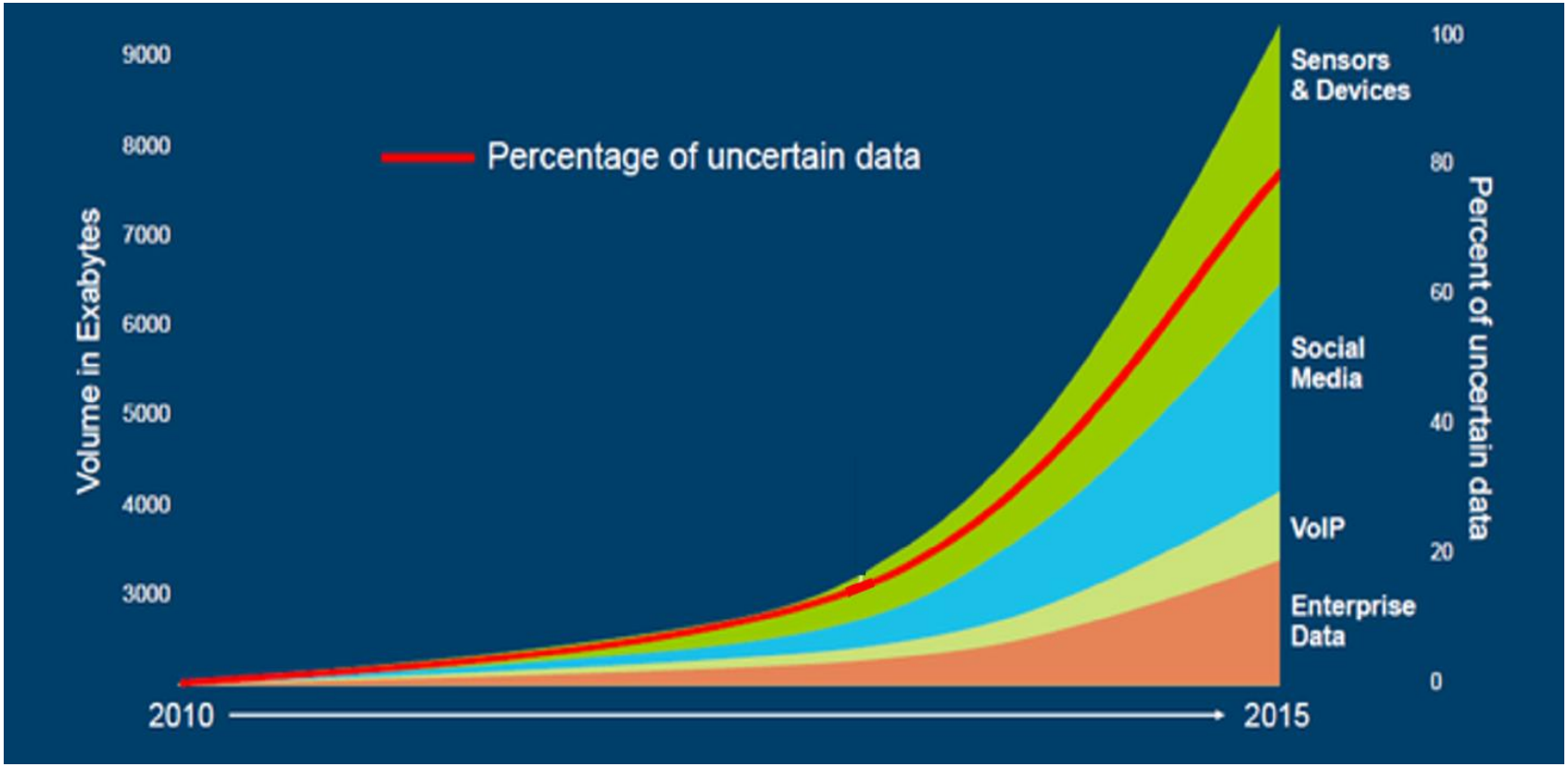


https://www.visualcapitalist.com/from-amazon-to-zoom-what-happens-in-an-internet-minute-in-2021/

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

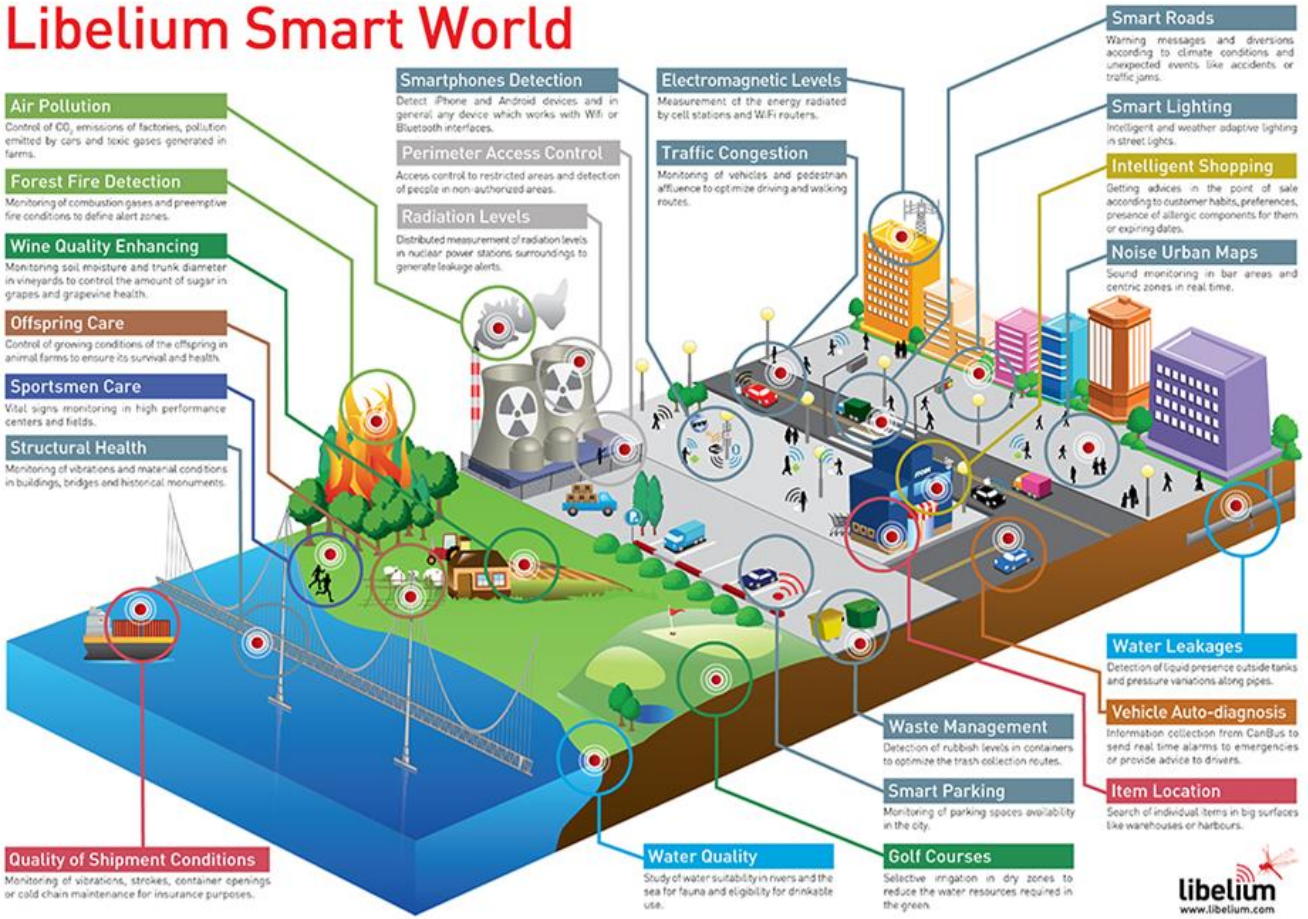
Źródła danych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

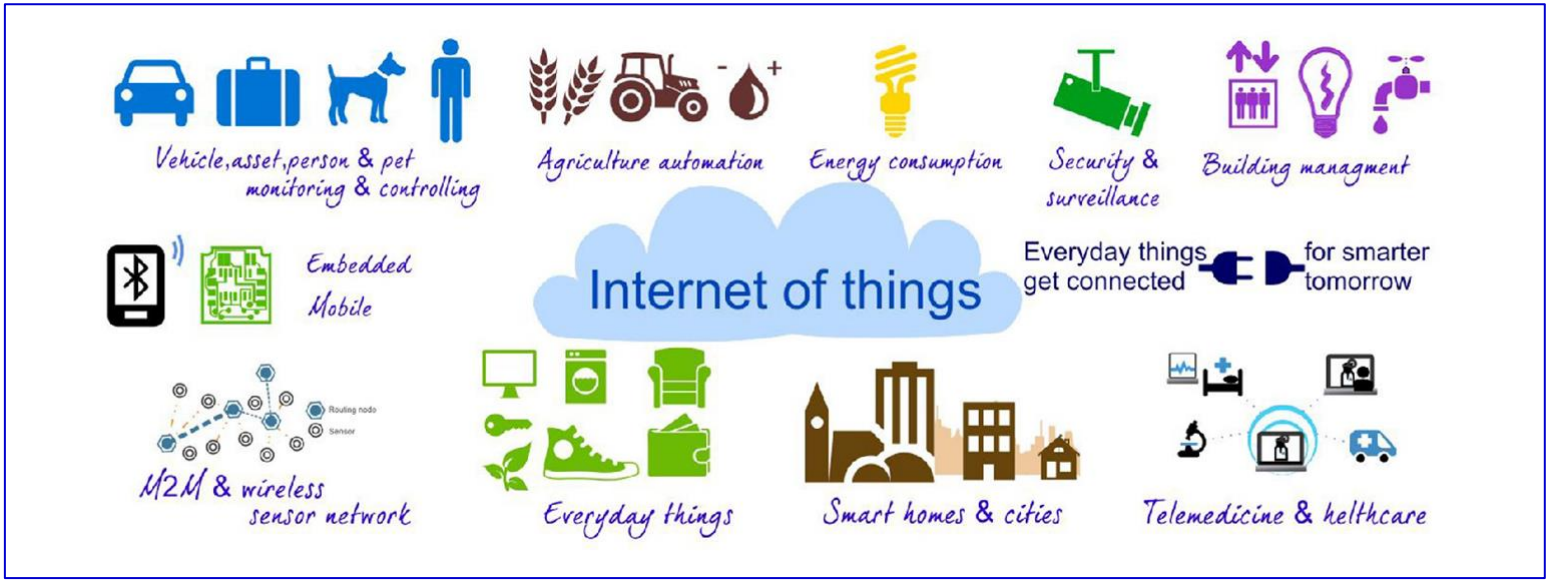
Źródła danych - Inteligentny świat



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

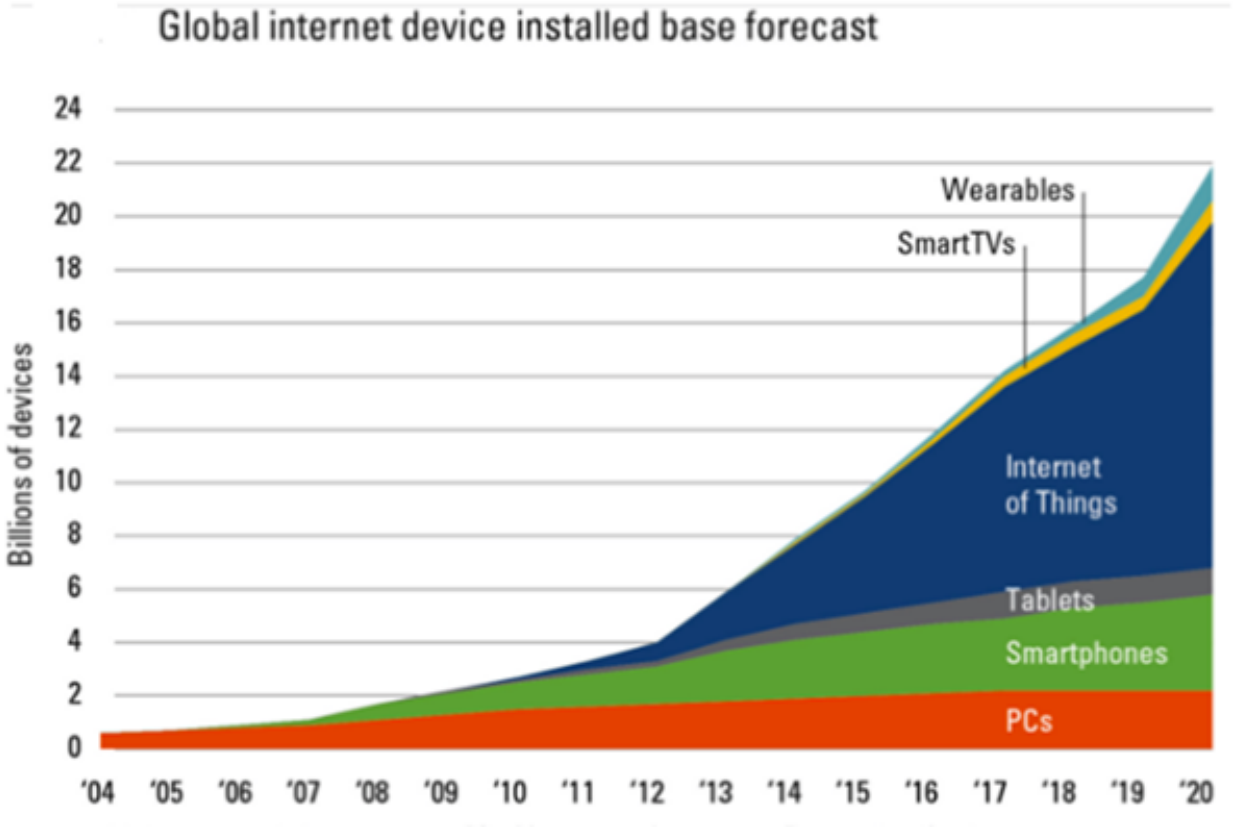
Internet rzeczy (IoT)



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Światowy sieć - rozwój urządzeń



Global Internet Device Forecast (Source: Market Realist, Gartner, IDC, Strategy Analytics.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dane generowane przez użytkowników

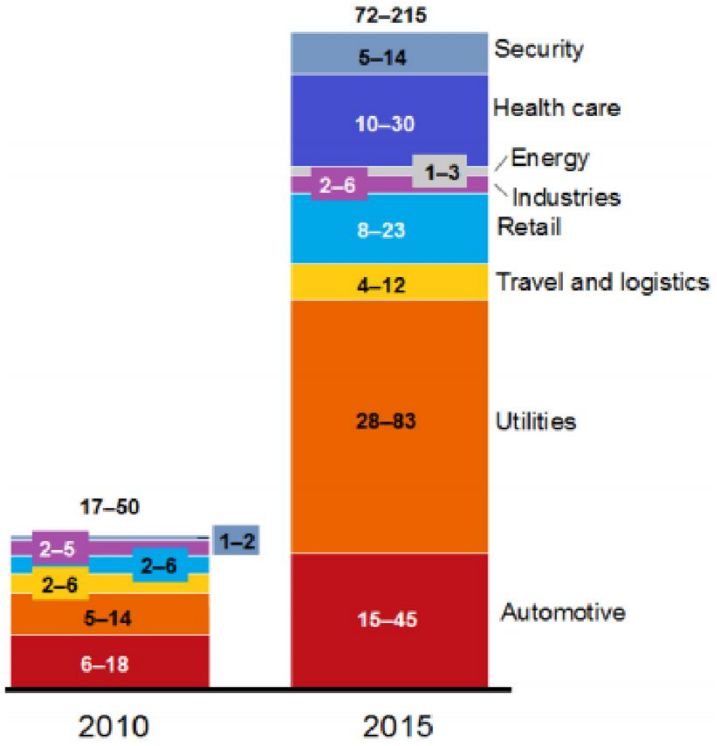
2006

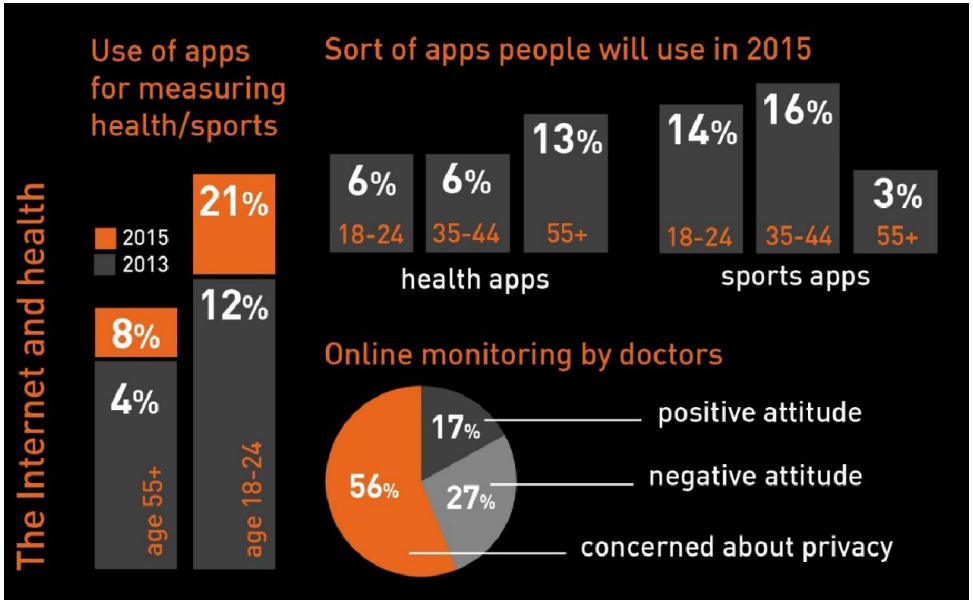
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Big Data w różnych dziedzinach życia

- Dane mają kluczowe znaczenie, niezależnie od tego, czy jesteśmy gigantycznym przedsiębiorstwem, czy indywidualną osobą.

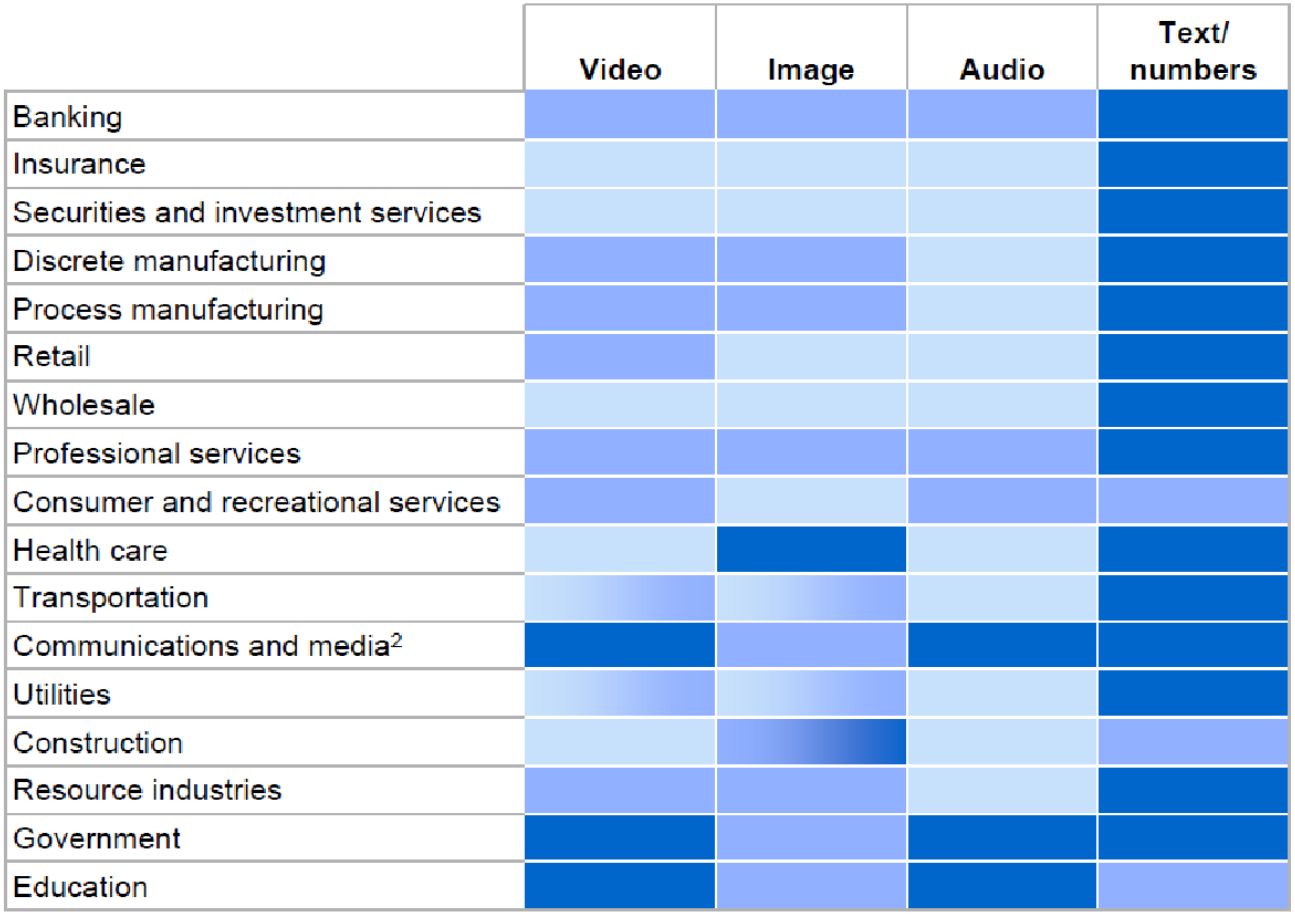




iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data - typy danych

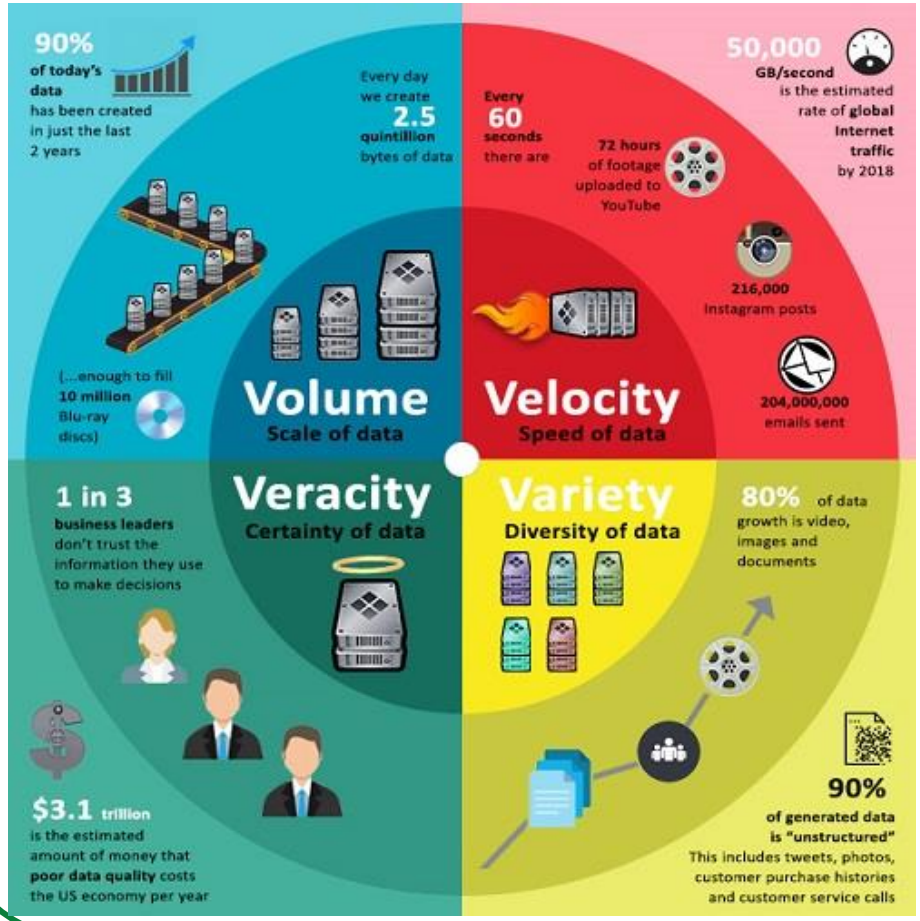


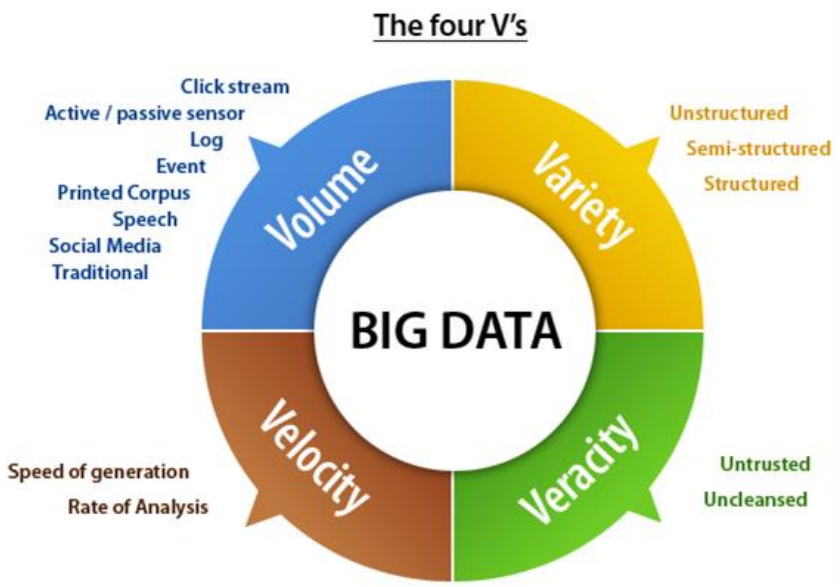
	Video	Image	Audio	Text/ numbers
Banking				
Insurance				
Securities and investment services				
Discrete manufacturing				
Process manufacturing				
Retail				
Wholesale				
Professional services				
Consumer and recreational services				
Health care				
Transportation				
Communications and media[2]				
Utilities				
Construction				
Resource industries				
Government				
Education				

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

4V Big Data + Wartość





iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data: V4+Wartość (Value)

- Rozmiar (Volume)
  - Terabyte (10^{12}), Petabyte (10^{15}), Exabyte (10^{18}), Zettabyte (10^{21})
- Różnorodność (Variety)
  - Structured, semi-structured, unstructured
  - Tekst, obrazy, audio, wideo, Text, image, audio, video, record
- Zmienność (Velocity)
  - Okresowe, bliskie czasu rzeczywistego, czasu rzeczywistego
- Wiarygodność (Veracity)
  - Jakość danych może się znacznie różnić
- Wartość (Value)
  - Big Data może generować ogromne przewagi konkurencyjne

iBigWorld:2020-1-PL01-KA203-082197
Co-funded by the Erasmus+ Programme of the European Union

Dlaczego?

- Ponieważ mamy dane
  - Dane utworzone są już w formie cyfrowej
  - 40% przyrostu danych rocznie
  - 7 lat temu: Internet ma oczy i uszy
- Ponieważ możemy
  - 300 $ za dysk, na którym można przechowywać całą muzykę świata
  - >40 lat prawa Moore'a → duże zasoby obliczeniowe
  - 40 years of databases → pojemność pamięci masowych
  - 68% firm zainwestowało w Big Data w 2018 roku
  - 57 miliardów $ zainwestowanych w Big Data w 2018 roku
- "Software is eating the world" , Marc Andreessen, The Wall Street Journal on August 20, 2011.

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Wielki zderzacz hadronów (LHC) CERN

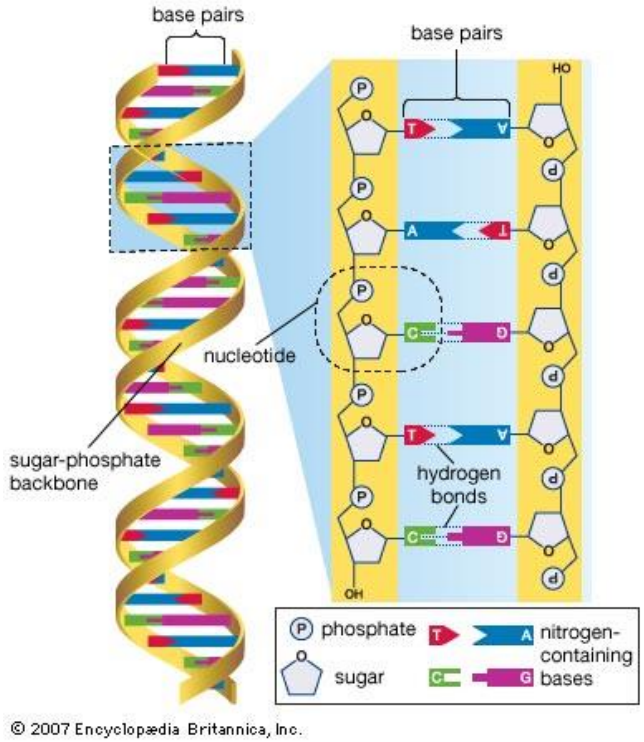
- Zebrał 200 Petabytes danych from 2012, a 1 Petabyte danych jest przetwarzany każdego dnia

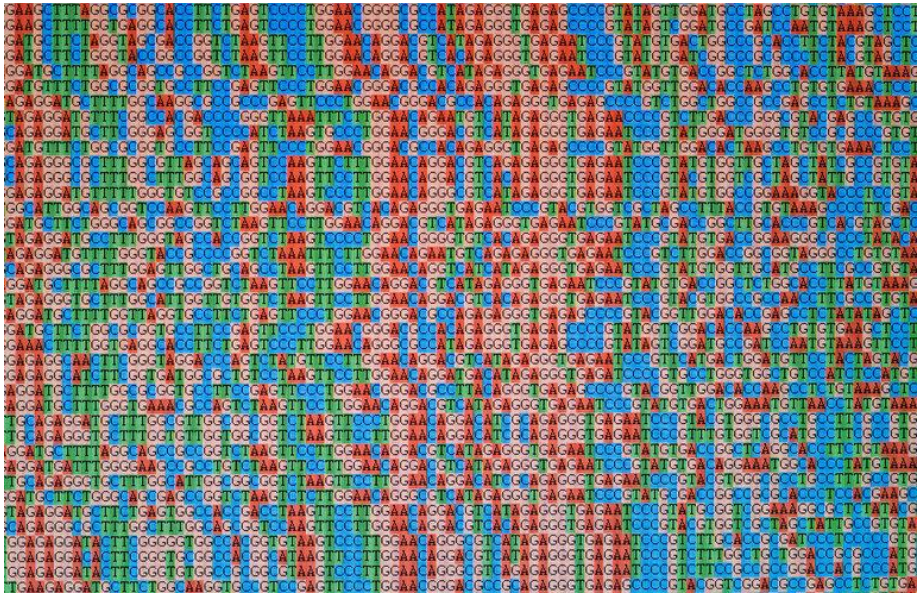
iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Ludzki genom

- DNA pojedynczego osobnika zawiera około 3,2 miliarda par nukleotydów DNA





iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data – Potencjalne możliwości

US health care
- $300 billion value per year
- ~0.7 percent annual productivity growth

Europe public sector administration
- €250 billion value per year
- ~0.5 percent annual productivity growth

Global personal location data
- $100 billion+ revenue for service providers
- Up to $700 billion value to end users

US retail
- 60+% increase in net margin possible
- 0.5–1.0 percent annual productivity growth

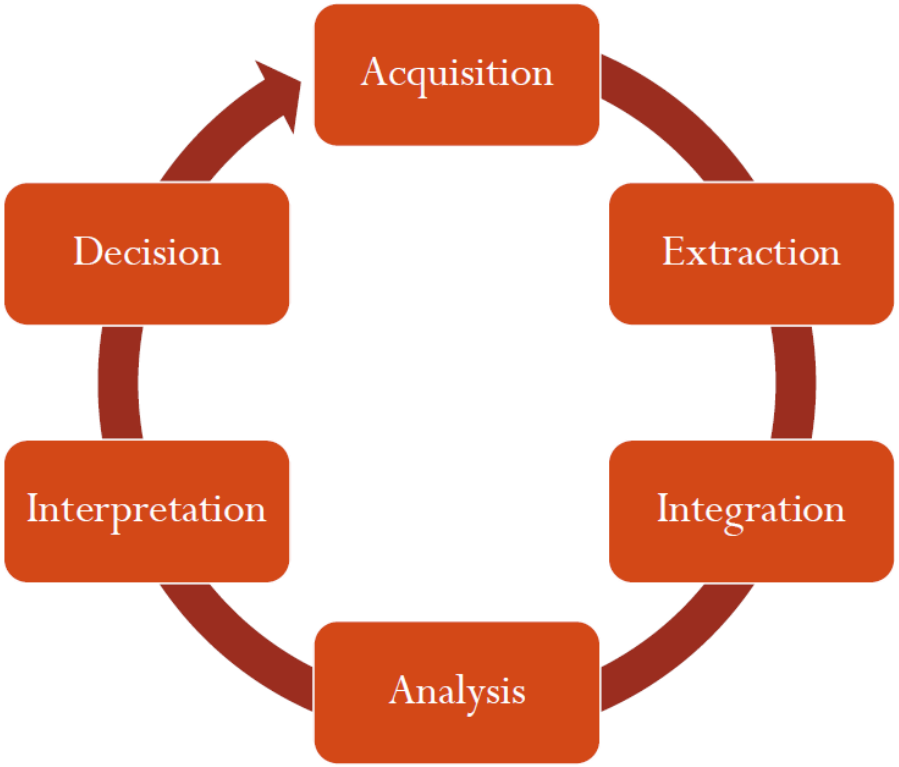
Manufacturing
- Up to 50 percent decrease in product development, assembly costs
- Up to 7 percent reduction in working capital

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Proces Big Data

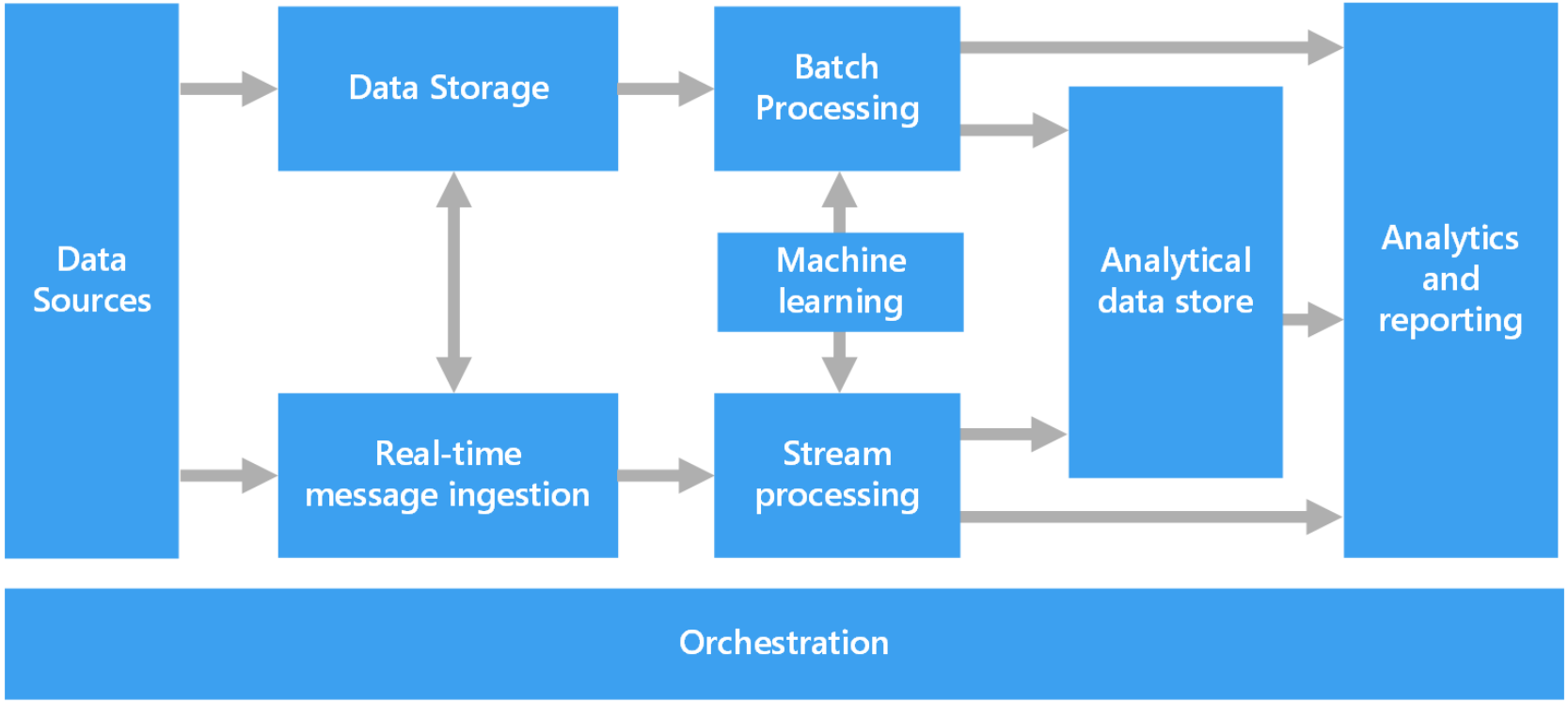
- Podejmuj skuteczne decyzje strategiczne, wykorzystując dostępność przetwarzania, analizy i wizualizacji Big Data



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura Big Data

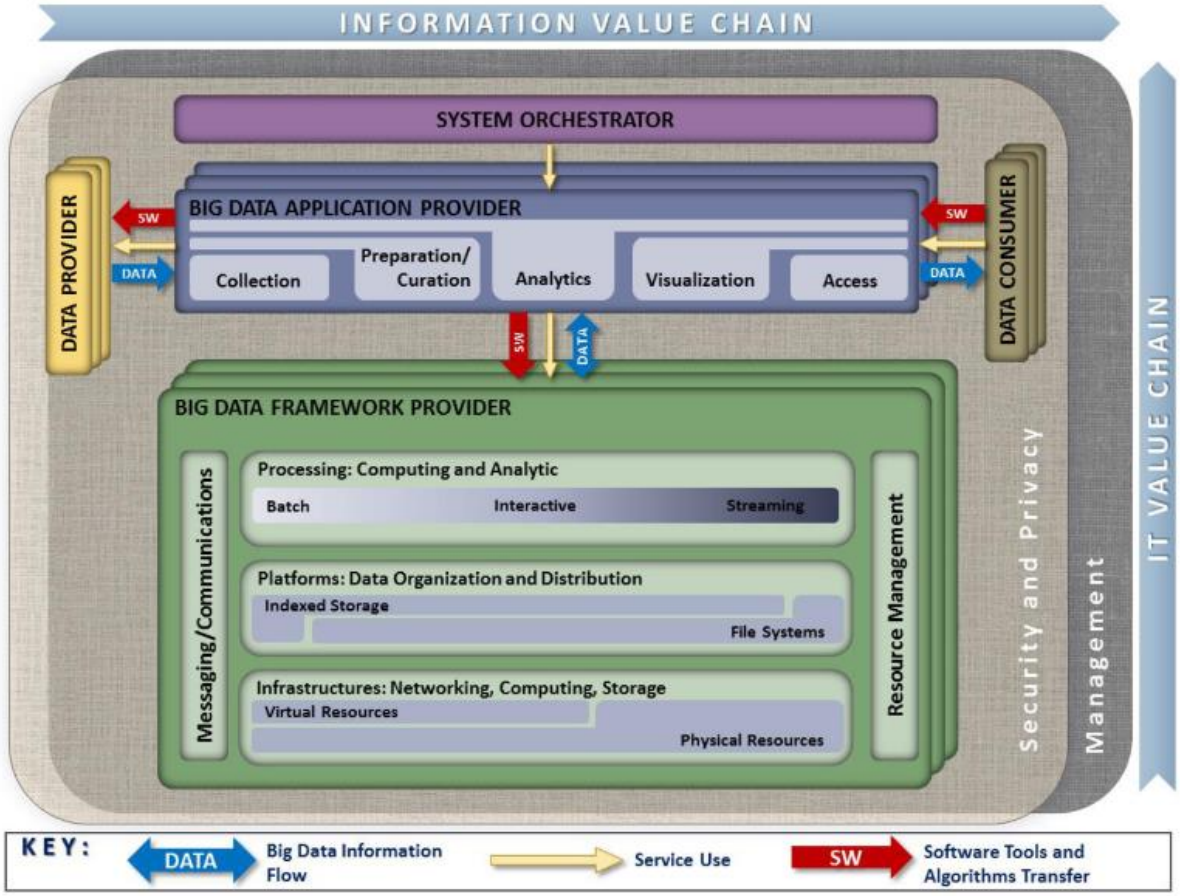


https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Architektura referencyjna NIST Big Data



https://doi.org/10.6028/NIST.SP.1500-6r2

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Środowisko Big Data 2018

http://samiora.blogspot.com/2018/07/big-data-eco-system-and-landscape-2018.html

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Dystrybucja i replikacja

- Architektura rozproszona
  - Wspólne rozwiązanie architektoniczne do przetwarzania Big Data: klaster podstawowych zasobów sprzętowych, również w chmurze
  - **Skaluj w poziomie, a nie w górę (w pionie)!**
  - Wyzwania: elastyczność i przetwarzanie danych na krawędziach sieci
- Przetwarzanie rozproszone
  - Model nie współdzielony
  - Nowe paradygmaty programistyczne, np. programowanie funkcjonalne
- Replikacja zasobów
  - Dobrze znane rozwiązanie pozwalające na osiągnięcie odporności na uszkodzenia
  - Spójność ostateczna (twierdzenie CAP)

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Jak implementować Big Data

- Tradycyjny sposób: korzystanie z klastra serwerów w siedzibie
  - Węzły obliczeniowe są przechowywane w szafach rackowych
  - 8-64 węzłów obliczeniowych w szafie
- Może być wiele szaf węzłów obliczeniowych
  - Węzły w szafie połączone są w sieć, zwykle gigabitowym Ethernetem
  - Szafy są połączone na innym poziomie sieci lub przełącznikiem
  - Przepustowość komunikacji wewnątrz szaf jest zwykle znacznie większa niż w przypadku komunikacji między szafami
- Rozważyć:
  - Konieczność zarządzania infrastrukturą sprzętową i
  - Platformy przetwarzania (pozyskiwanie, instalacja, konfiguracja, ...)

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Gdzie przetwarzać Big Data

- Ale centra danych w chmurze znajdują się w rdzeniu sieci
- Główne wyzwania:
  - Przenieść dane do chmury
    - Opóźnienie nie jest zerowe (z powodu prędkości światła)!
    - Mniejszy problem: przepustowość sieci
  - Bezpieczeństwo i prywatność danych

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

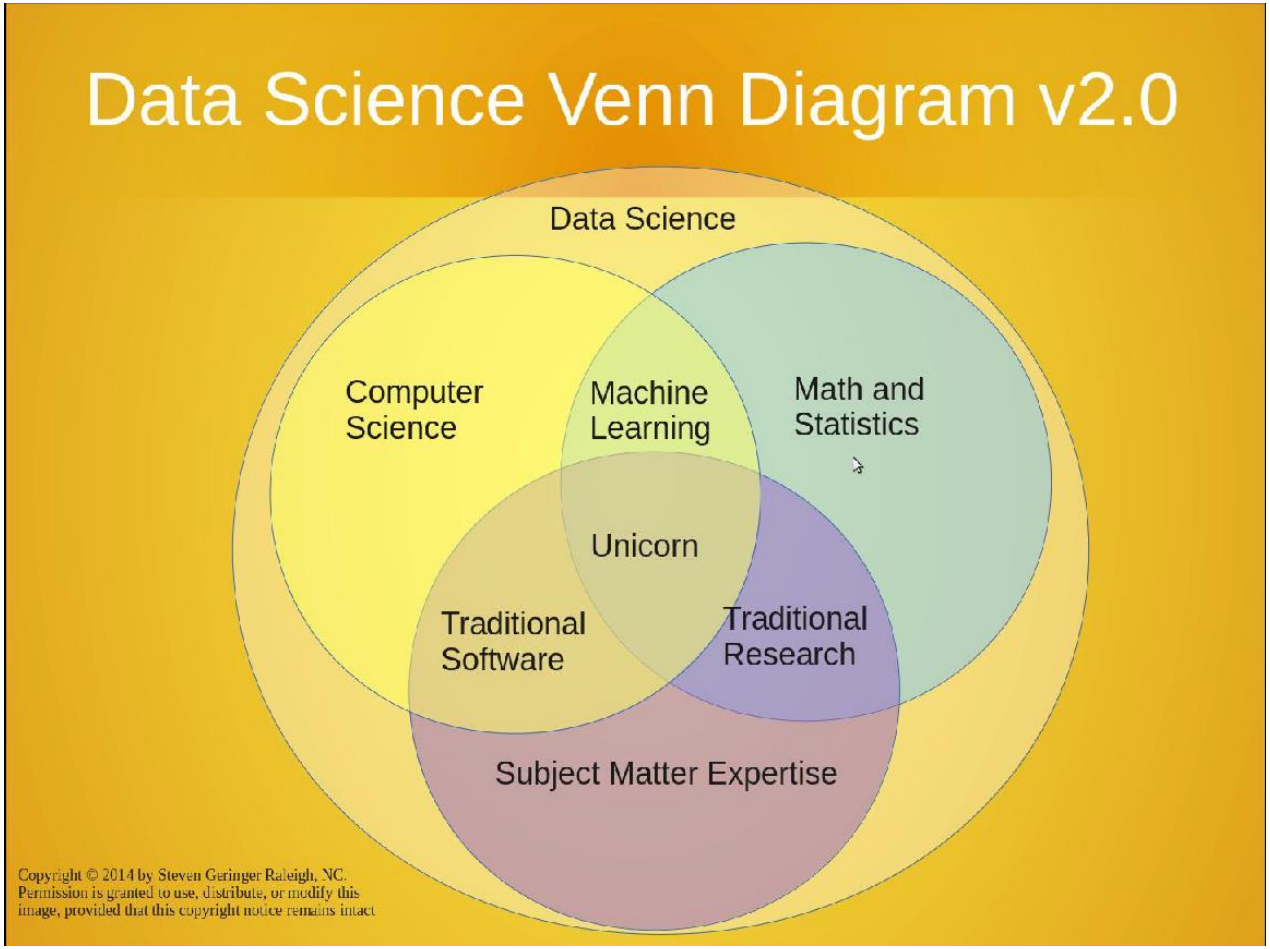
Nauka o danych

- Interdyscyplinarna dziedzina, która próbuje wyodrębnić wzorce, spostrzeżenia i wiedzę z ustrukturyzowanych, częściowo ustrukturyzowanych i nieustrukturyzowanych danych
- Eksploracja danych
- Odkrywanie wiedzy w bazach danych
- Nauczanie maszynowe
- Rozpoznawanie wzorców
- Analiza predykcyjna
- Analiza Big Data

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Diagram analizy danych



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

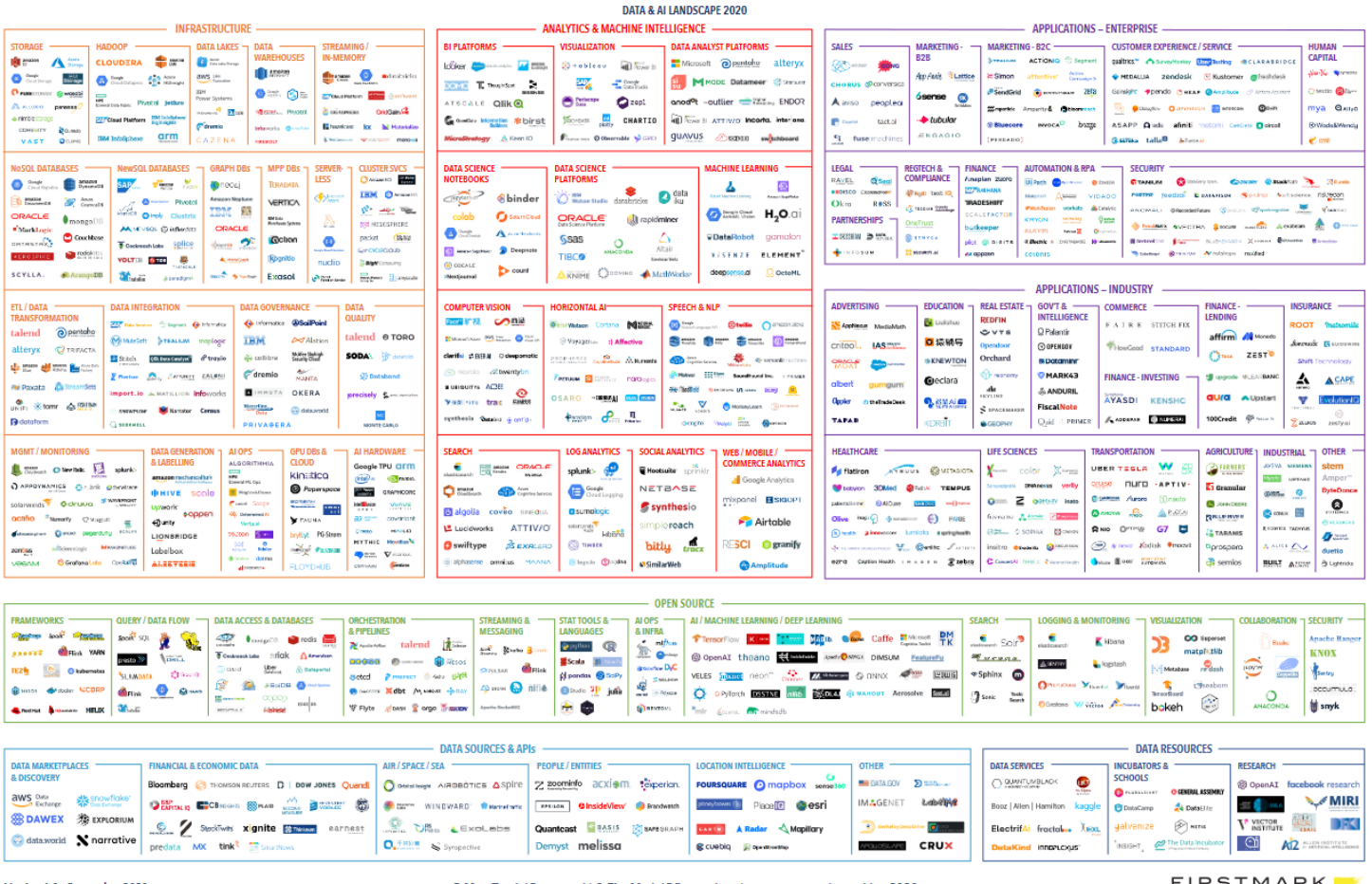
Analiza danych

- Liczba zastosowań dla analizy danych wzrosła o 650% od 2012 roku
- Według amerykańskiego Biura Statystyki Pracy do 2026 r. zostanie utworzonych około 11,5 miliona miejsc pracy
- Ponadto stanowisko Data Scientist należy do najlepszych wschodzących miejsc pracy w Linkedin
- Wszystkie statystyki wskazują na rosnące zapotrzebowanie na naukowców zajmujących się danymi

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union

Big Data i AI środowisko 2020



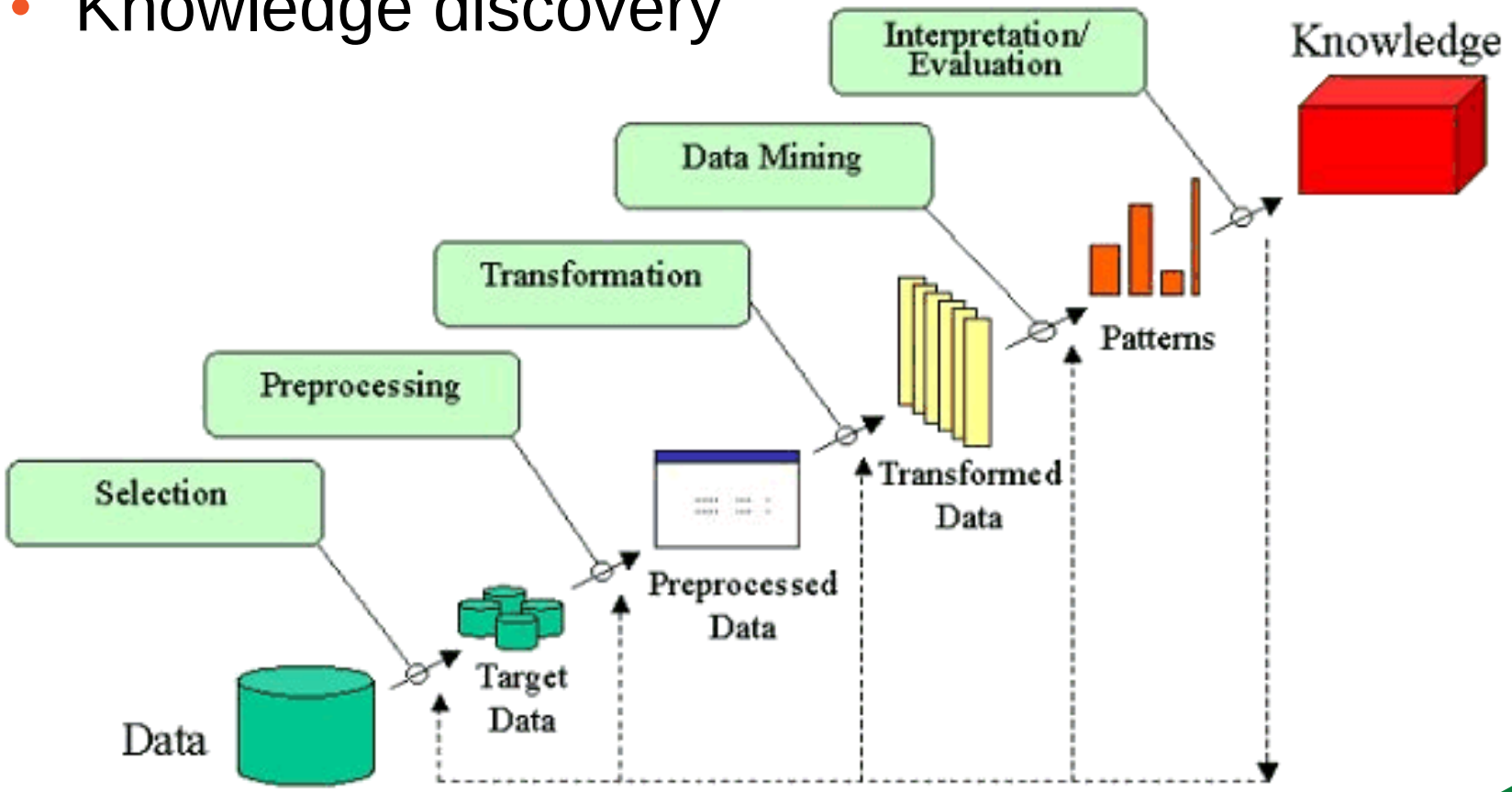
iBigWorld:2020-1-PL01-KA203-082197

https://mattturck.com/data2020/

Co-funded by the Erasmus+ Programme of the European Union

Proces przetwarzania danych

- Knowledge discovery



iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Prywatnośc i bezpieczeństwo danych

- Prywatność danych
- Bezpieczeństwo danych
- Integralność danych
- Analiza pod kątem bezpieczeństwa

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the
Erasmus+ Programme
of the European Union

Bibliografia

- Nathan Marz, James Warren, Big Data Principles and best practices of scalable real-time data systems. Manning Publications Co., 2015.
- Martin Kleppmann, Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly Media, 2017.
- John D. Kelleher, Brendan Tierney, Data Science, The MIT Press, 2018
- Joel Grus, Data Science from Scratch: First Principles with Python, 2nd Edition, O'Reilly Media, 2019

iBigWorld:2020-1-PL01-KA203-082197

Co-funded by the Erasmus+ Programme of the European Union