# Moduł 11
# Klasyfikacja a Big data

iBigWorld:
Innovations for Big Data in a Real World

Zespół UBB

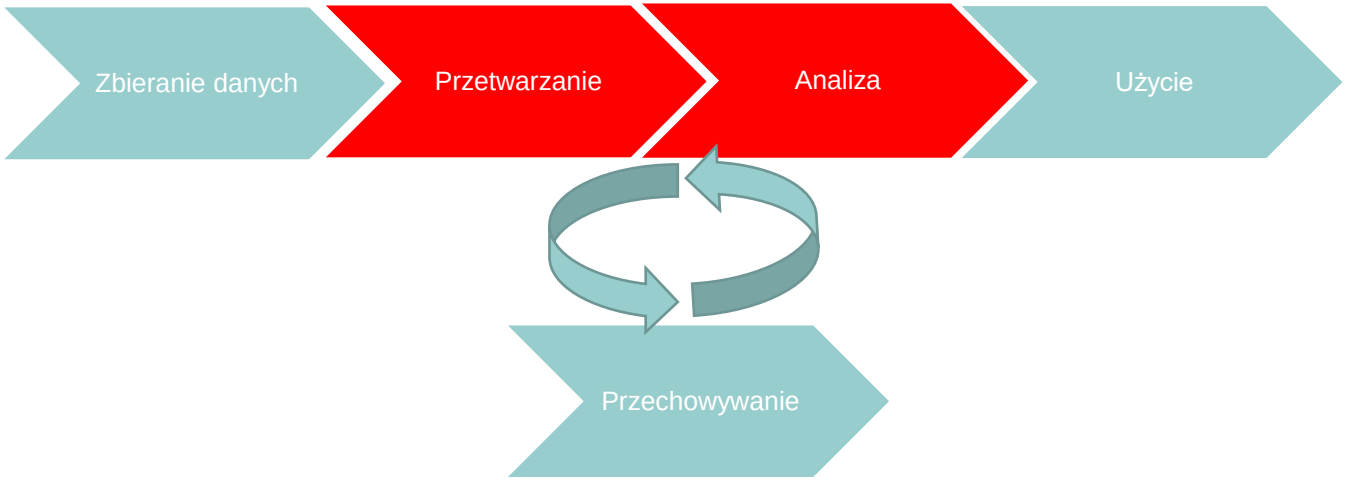


iBigWorld:2020-1-PL01-KA203-082197

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

# Klasyfikacja

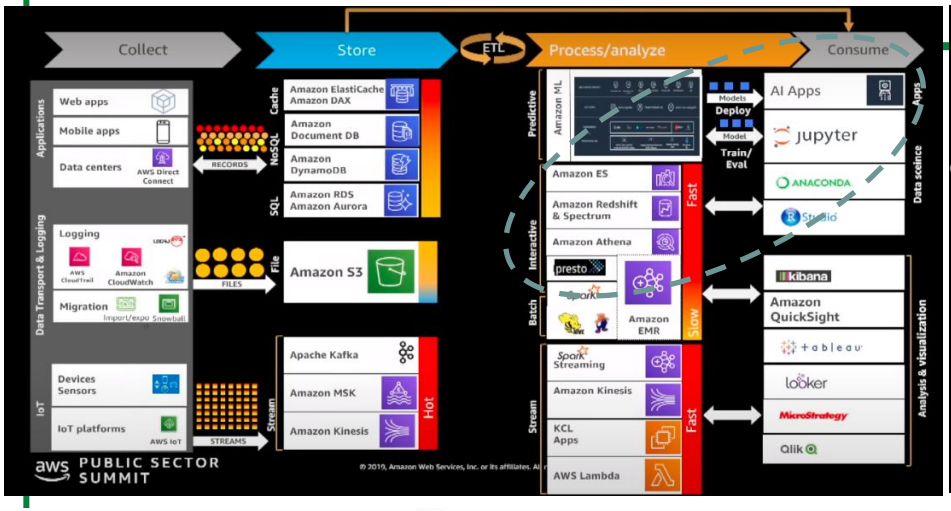
- Potok Big Data po wstępnym przetwarzaniu i eksploracji danych może zostać użyty do budowy modeli klasyfikacyjnych i predykcyjnych

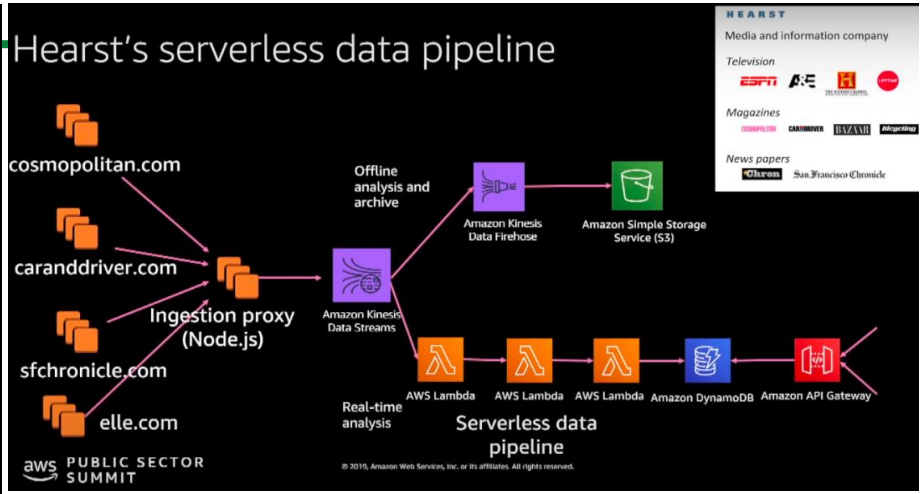


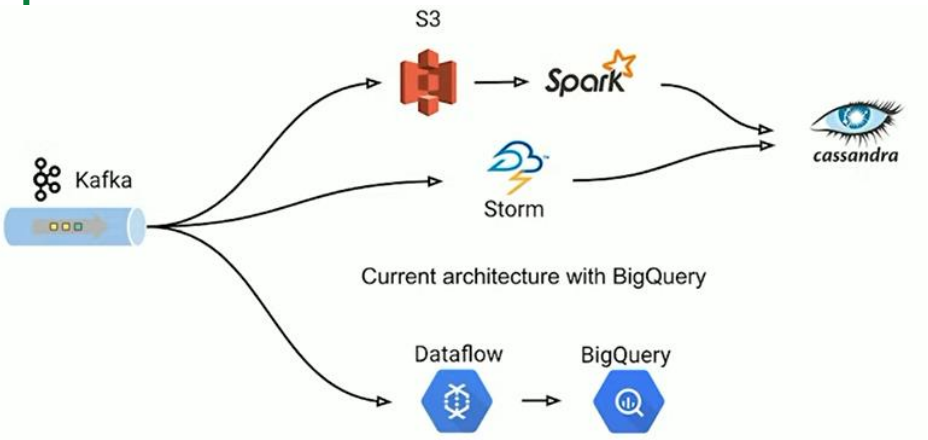


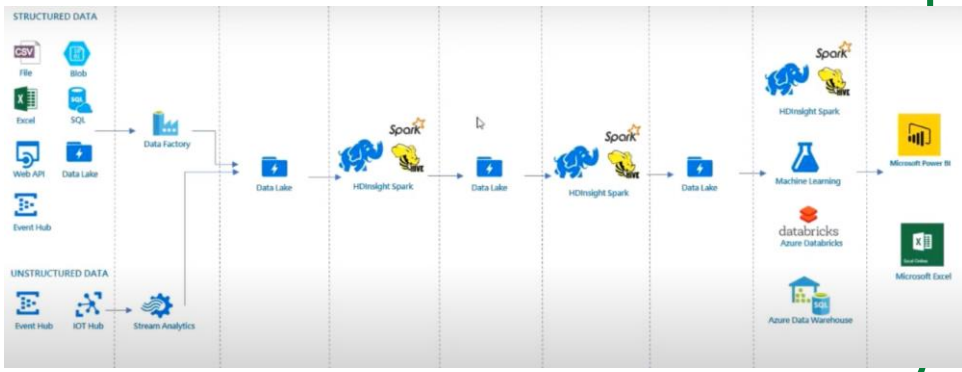

Co-funded by the
Erasmus+ Programme
of the European Union

# Wiele rozwiązań (w zależności od dostawcy)



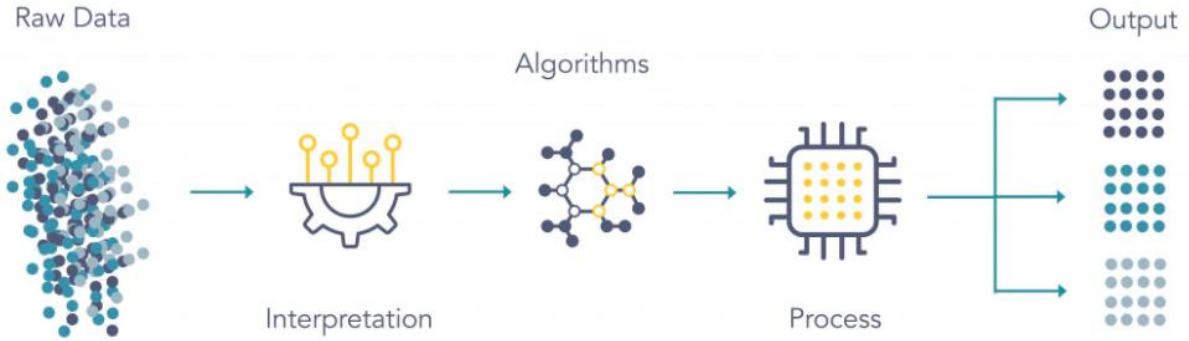




iBigWorld:2020-1-PL01-KA203-082197

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

# Nauczanie maszynowe - podsumowanie

- **podzbiór sztucznej inteligencji i ogólny termin określający, kiedy komputery uczą się na podstawie danych.**
- **opisuje skrzyżowanie informatyki i statystyki, gdzie algorytmy są wykorzystywane do wykonania określonego zadania bez wyraźnego zaprogramowania**
- **rozpoznają wzorce w danych i dokonują przewidywań, gdy pojawią się nowe dane.**



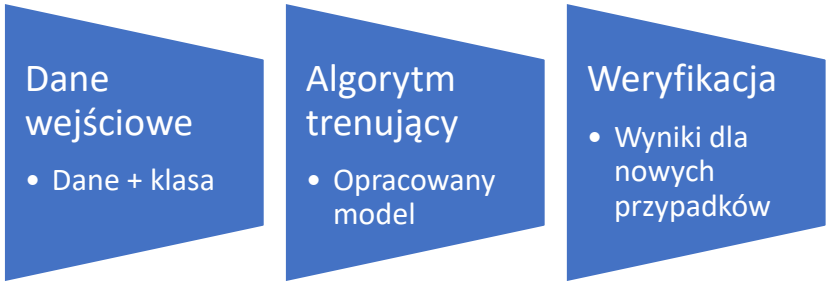

[https://www.logpoint.com/sv/blogg-sv/explained-siemply-machine-learning/]



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

# Nauczanie nadzowowane

- **Uczenie nadzorowane wykorzystuje znaną relację między danymi wejściowymi i wyjściowymi.**
- **Pod uwagę brane są oznakowane dane:**
  - **Algorytm uczy się odpowiedzi na podstawie danych**
  - **Używa wytrenowanych danych, aby przewidzieć poprawną odpowiedź w nowym przypadku.**
- **Przykłady:**
  - **Regresja: Przewiduje ciągłą wartość liczbową. Jaka będzie temperatura za 1 godzinę?**
  - **Klasyfikacja: Przypisz etykietę. Przykład: „Co to za owoc?”**

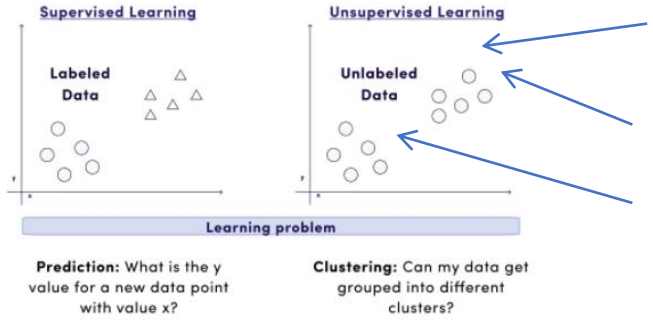





Co-funded by the Erasmus+ Programme of the European Union

# Nauczanie nienadzorowane

- **W tym przypadku staramy się znaleźć pewne prawidłowości w danych. W przeciwieństwie do danych oznaczonych staramy się grupować elementy na podstawie ich charakterystyki (wartości lub terminów)**




| płeć | wiek | waga | wzrost |
|---|---|---|---|
| M | 18 | 60 | 160 |
| M | 30 | 90 | 180 |
| F | 40 | 60 | 159 |

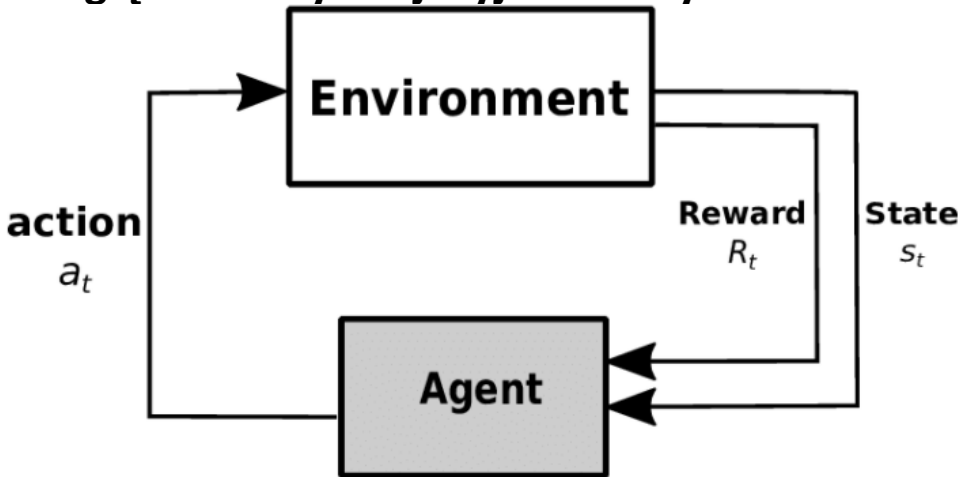
- **Na podstawie kryteriów elementy można grupować (grupować) w różny sposób.**
- **Chodzi o to, aby znaleźć przedstawiciela lub grupę o podobnych wartościach**



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

# Nauczanie przez wzmocnieniem

- **W nauczaniu przez wzmacnianie algorytm (zwany także agentem) uczy się metodą prób i błędów, wykorzystując informacje zwrotne do własnych działań. Nagrody i kara działają jako sygnały pożądanego i niepożądanego zachowania.**
- **Przykładem rozwiązania jest mysz w labiryncie. W oparciu o przyjętą strategię może szybciej wyjść z labiryntu.**



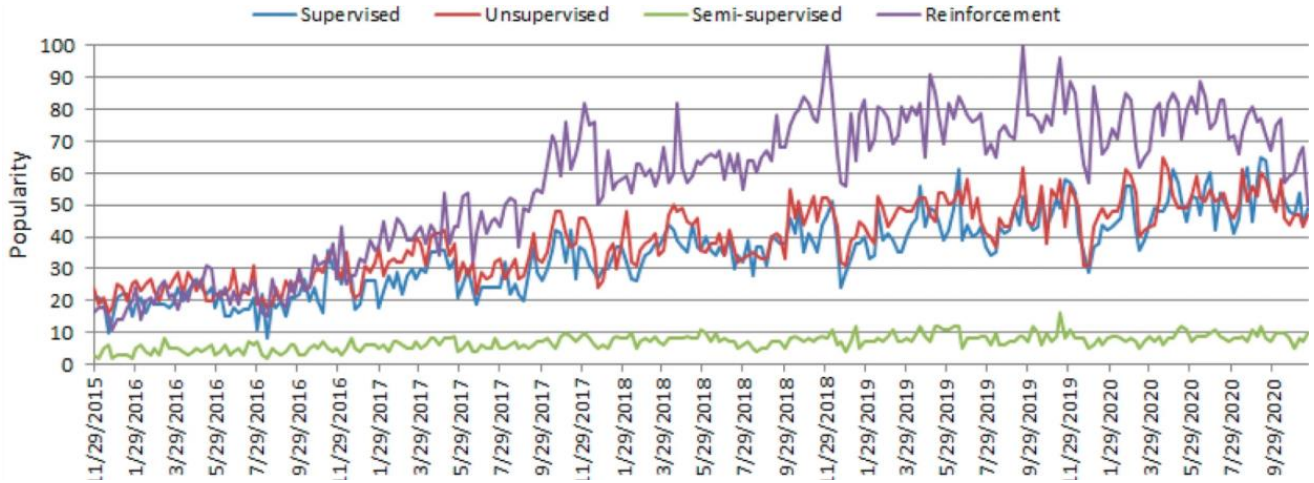

[https://www.researchgate.net/figure/Reinforcement-Learning-Agent-and-Environment_fig2_323867253]



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

# Trendy w uczeniu maszynowym

- Każda grupa nauki znajduje swoje zastosowanie
- W ciągu ostatnich trzech lat zauważono wybuch rozwiązań w zakresie uczenia się przez wzmacnianie
- Nadzorowane i nienadzorowane rozwiązania stale się rozwijają
- Do tej pory wszystkie rodzaje uczenia się są równie popularne



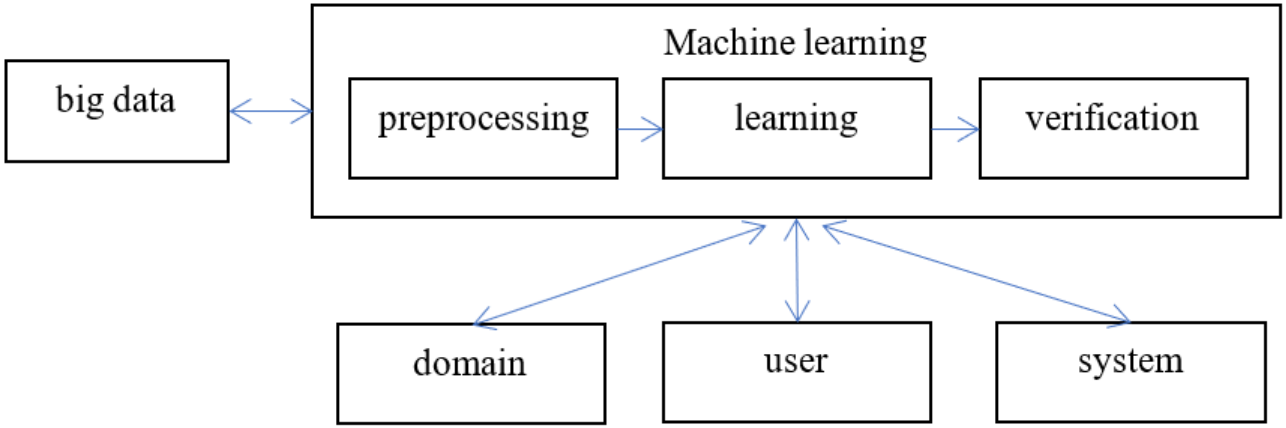

[Machine Learning: Algorithms, Real-World Applications and Research Directions. Iqbal H. Sarker. *SN Computer Science* **volume 2**, Article number: 160 (2021) ]



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

# Uczenie maszynowe i Big Data

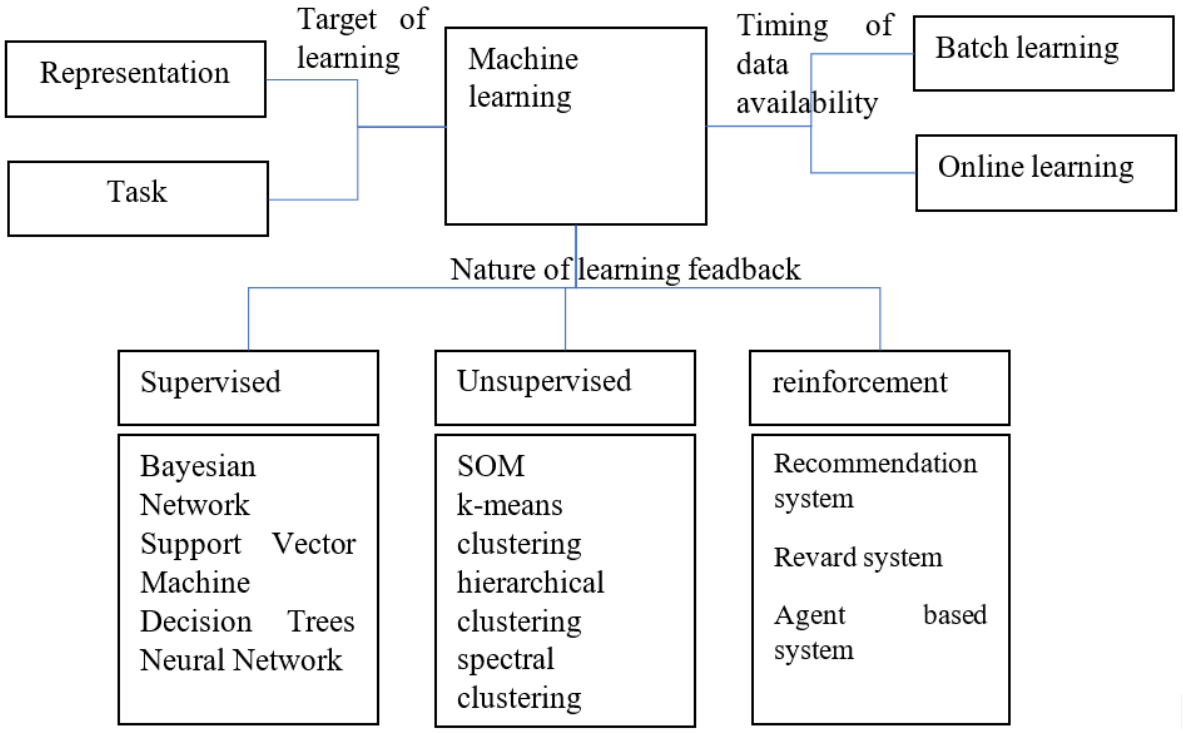
- Uczenie maszynowe wiąże się z Big Data poprzez charakterystykę użytkownika (user), domeny (domain) i systemów (system).
- Big Data może być przetwarzana przez ML, a jego wynik będzie nową informacją-częścią Big Data.
- Wynik może być wykorzystany przez użytkownika w określony sposób, który daje dodatkowe tło (wiedza dziedzinowa, informacje zwrotne lub doświadczenie).
- Architektura systemu może wpływać na sposób przetwarzania i może prowadzić do modyfikacji istniejącej już architektury.






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

# Taksonomia uczenia maszynowego dla Big Data




[Lina Zhoua,∗, Shimei Pana, Jianwu Wanga, Athanasios V. Vasilakosb. Machine learning on big data: Opportunities and challenges. Neurocomputing 237 (2017) 350–361]



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

# Cel nauki

- Nauka może być ukierunkowana na rozwiązanie konkretnego zadania lub nakierowana na znalezienie konkretnych funkcji, które można wykorzystać. W tym kontekście ML można podzielić na uczenie reprezentatywne i uczenie zadaniowe.
- Nauka reprezentacji wyszukuje nowe reprezentacje danych, które ułatwiają znajdowanie znaczących informacji dla klasyfikatorów lub predyktorów. Często jest to taki sposób, który rozplątuje podstawowe czynniki zmienności, wychwytuje rozkład a posteriori leżących u podstaw eksploracji, stosując szacowanie gęstości i redukcję wymiarowości.
- Nauka zadaniowa ma zwykle określony cel (pożądane rezultaty) i może być podzielona na trzy typy: klasyfikacja, regresja i grupowanie. W przypadku problemu z klasyfikacją techniki ML tworzą model, który przypisuje dane wejściowe do jednej lub więcej predefiniowanych klas. Problem regresji różni się od klasyfikacji w kontekście wyjścia, którym jest wartość ciągła. Na koniec klasteryzacja, która tworzy grupy danych.



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

# Dostępność danych w Big Data

- Big Data są zwykle dostępne do uczenia maszynowego w postaci ogromnych zbiorów danych gromadzonych przez znaczny okres czasu lub jako strumień danych (np. operacje, teksty, głos lub obrazy).
- Na podstawie czasu przetwarzania można je podzielić na uczenie wsadowe i uczenie online.
- W przypadku uczenia wsadowego modele są generowane z wykorzystaniem całych danych uczących, natomiast w przypadku uczenia się online modele uczenia się są stale aktualizowane na podstawie każdej nowej porcji danych.
- Model ML dla fragmentów danych w uczeniu wsadowym zakłada jego niezależną charakterystykę i prawdopodobieństwo rozkładu dla każdej klasy. Jednak ta cecha zwykle nie jest spełniana przez rzeczywiste dane, dlatego wymagane jest dodatkowe przetwarzanie wstępne.
- Dane strumieniowe zwykle nie zawierają żadnych założeń statystycznych dotyczących danych.
- Podejście online jest stosowane, gdy dane są stale generowane i są gorące (ich użyteczność szybko maleje z czasem).Warto zauważyć, że każdy algorytm uczenia maszynowego można kategoryzować w wielu wymiarach (np. klasyczne drzewa decyzyjne to nadzorowane algorytmy uczenia wsadowego).

iBigWorld:2020-1-PL01-KA203-082197

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

# Typy danych w Big Data

- *Strukturalny:* ma dobrze zdefiniowaną strukturę, jest zgodny z modelem danych według standardowej kolejności np. : nazwiska, daty, adresy, numery kart kredytowych, informacje o giełdzie, geolokalizacja.
- *Nieustrukturyzowane:* nie ma zdefiniowanego z góry formatu lub organizacji danych np. : dane z czujników, wiadomości e-mail, wpisy na blogach, wiki i dokumenty edytora tekstu, pliki PDF, pliki audio, filmy, obrazy, prezentacje, strony internetowe i wiele innych rodzajów dokumentów biznesowych można uznać za dane nieustrukturyzowane.
- *Częściowo ustrukturyzowane: to* dane częściowo ustrukturyzowane nie są przechowywane w relacyjnej bazie danych np. : dokumenty HTML, XML, JSON, bazach danych NoSQL itp.
- *Metadane:* podstawowa różnica między „danymi" a „metadanymi" polega na tym, że te dane są po prostu materiałem, który można klasyfikować, mierzyć, a nawet dokumentować coś związanego z właściwościami danych organizacji: autor, rozmiar pliku, data wygenerowana przez dokument, słowa kluczowe do zdefiniowanego dokumentu itp.

iBigWorld:2020-1-PL01-KA203-082197

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

# Kontekst użytkownika w ML Big Data

- Z systemów ML w kontekście Big Data korzysta wielu ekspertów dziedzinowych, badaczy i programistów.
- Wykorzystując klasyczne podejście użytkownicy dobierają metody i podejście ML na każdym z etapów uczenia się: zbierania danych, uczenia się i weryfikacji.
- Wyniki zazwyczaj były prezentowane użytkownikowi końcowemu jako dane, wykres lub odpowiedź na pytanie dotyczące domeny.
- Informacja zwrotna dla użytkownika końcowego wiązała się zwykle z wynikami i stanowiła podstawę danych wejściowych do kolejnej iteracji eksperckiej.
- W dzisiejszych czasach użytkownicy końcowi są włączani w ten proces na wielu etapach. Początkowo użytkownik był źródłem etykiet danych (jako ekspert dziedzinowy), teraz wykorzystując przejrzystość w projektowaniu systemu uczącego, pozwala lepiej zrozumieć nie tylko dane, ale ich kontekst, który okazał się dostarczać lepsze etykiety/feedback.
- Dodatkowo zaangażowanie użytkowników w ML pomaga poprawić doświadczenie z danymi i zauważyć pewne relacje, które w innym przypadku mogłyby być traktowane jako nieistotne w przypadku eksperta od danych.



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

# Kontekst domeny w ML Big Data

- Znajomość domeny umożliwia ML odkrywanie nowych wzorców, których nie da się wykryć na podstawie samych zestawów danych.
- Analizowane zbiory danych są zawsze ograniczone, a ich ilość nie zawsze pozwala odkryć wszystkie przydatne wzorce.
- W niektórych dziedzinach uzyskanie danych jest kosztowne ze względu na ich różnorodność i specyficzne wymagania.
- Dlatego wiedza dziedzinowa ekspertów może poprawić odporność uzyskanych wzorców i uwzględnić wymagany poziom ogólności.
- Istnieje kilka sposobów na włączenie wiedzy dziedzinowej do Big Data dla aplikacji ML [10]:
  - przygotowywanie przykładów szkoleniowych;
  - generowanie hipotez lub przestrzeni hipotez;
  - modyfikowanie celu wyszukiwania;
  - i rozszerzenie wyszukiwania.
- Uzyskane wzorce można wykorzystać do poszerzenia wiedzy dziedzinowej



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

# Kontekst systemu w ML Big Data

- Wykonano wiele prac w ML w celu dostosowania oddzielnych algorytmów do charakterystyki Big Data.

| Zrównoleglenie | cel | Techniki | Badania |
|---|---|---|---|
| nie | | optymizacja | [6][7][8] |
| tak | Zorientowany na dane | mapReduce | BN[9][10], DT[3], TM[11], GP[12][13][14][15] |
| | | Grafy rozproszone | GA[16] |
| | | inne | SVM[2],NN[17],GP[1,4] |
| | Zorientowany na model | wielowątkowość | SVM[2] |
| | | MPI/OpenMP | NN[5],TM[14] |
| | | GPU | NN[5][17][18] |
| | | inne | SVM [19], NN [20], GP [1][4] |



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

# ML BIG Data - rozwiązania

- Klasyczne metody ML mają na celu przetwarzanie danych za pomocą algorytmu sekwencyjnego, który nie zawiera cech równoległości (powolne dla BiG Data).
- Pierwsze podejście polegało na opracowaniu nowych algorytmów sekwencyjnych, ale o znacznie większej złożoności czasowo-przestrzennej. Algorytm z tej kategorii jest generalnie nieskomplikowany, dzięki czemu możliwe jest jego zastosowanie dla Big Data.
- Istnieje granica, do której może dojść pojedynczy algorytm sekwencyjny, dlatego cecha równoległości jest obecna w większości metod ML. Jest to szczególnie ważne w przypadku dużych rozmiarów i funkcji dużych zbiorów danych. Równoległość algorytmów ML zawiera:
  - równoległość danych, czyli partycjonowanie danych wejściowych w pionie, poziomie lub na mniejsze, łatwe do zarządzania fragmenty (np. przez Hadoop), a następnie przetwarzanie ich równolegle i na koniec ich łączenie;
  - równoległość modelu/parametru: gdzie tworzone są różne algorytmy uczenia się na podstawie modelu/parametrów, a następnie obliczane są jednocześnie w blokach,
  - model mieszany stosuje się tam dwie ww. metody [1–4].

iBigWorld:2020-1-PL01-KA203-082197

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

# Rozwiązanie ML w Big Data

- Poprzez zrównoleglenie w modelach ML może osiągnąć skalowalność wystarczającą do radzenia sobie z Big Data. Skalowalność można osiągnąć w praktyce dwoma podejściami:.
- Użyć warstwy oprogramowania pośredniczącego, która implementuje istniejące procedury uczenia się na platformach Big Data, takich jak Hadoop i Spark. Warstwa dostarcza prymitywów (prostych operacji), które można zastosować w celu wykonania zadań uczenia się. Takie podejście pozwala analizować różne podejścia (algorytmy uczenia się) w jednym środowisku. Przykładem oprogramowania pośredniczącego big data są Spark MLlib [12] i Mahout [13], które są reprezentatywne dla projektów open-source, które obsługują wiele skalowalnych wersji algorytmów wspólnego uczenia się: klasyfikacja, regresja, klastrowanie, filtrowanie kolaboracyjne i redukcja wymiarowości.
- Wdrażać indywidualne algorytmy uczenia się uruchamiane na platformie Big Data. Takie podejście jest zaimplementowane na szczycie silnika Big Data. Bezpośrednie podejście może zwiększyć jego efektywność, jednak wymaga znacznych umiejętności programistycznych.



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

# Przypadki użycia do rozważenia

| | Przypadki użycia (link to danych) |
|---|---|
| 1 | W jakim kraju na podstawie sytuacji Covid-19? |
| 2 | Analiza medyczna na podstawie danych globalnych Global Health Data Exchange (http://ghdx.healthdata.org/) (liczba, etykiety) |
| 3 | Segmentacja naczyń (zdjęcia) (https://www.idiap.ch/software/bob/docs/bob/bob.db.drive/stable/index.html) |
| 4 | Numer domu Google Street View (SVHN) Zbiór danych (https://github.com/aditya9211/SVHN-CNN) (zdjęcia) |
| 5 | Physionet (sygnały) (physionet.org) |
| 6 | Przetwarzanie języka naturalnego (teksty) i aplikacja do akcji |
| 7 | Klasyfikacja transformatorów mocy na podstawie infradźwięków |
| 8 | Wykrywanie pojazdów drogowych w oparciu o czujniki na drodze |
| 9 | Analiza codziennych czynności na podstawie danych z czujników (telefon komórkowy) |



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

# Referencje


[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Sys-tems," CoRR, 2016.
[2] Y. You, H. Fu, S.L. Song, A. Randles, D. Kerbyson, A. Marquez, et al. Scaling support vector machines on modern HPC platforms J. Parallel Distrib. Comput., 76 (2015), pp. 16-31
[3] B. Panda, J.S. Herbach, S. Basu, R.J. Bayardo PLANET: massively parallel learning of tree ensembles with MapReduce Proc. VLDB Endow., 2 (2009), pp. 1426-1437
[4] E. Xing, Q. Ho, W. Dai, J.-K. Kim, J. Wei, S. Lee, et al. Petuum: a new platform for distributed machine learning on Big data
[5]R. Collobert, K. Kavukcuoglu, and C. Farabet, Torch7: A Matlab-like En-vironment for Machine Learning, in: Proceedings of the Neural Information Pro-cessing Systems (NIPS) Workshop on BigLearn, 2011
[6]T.Yang, Q.Lin, R.Jin, Big data analytics: Optimization and randomization, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 2327–2327.
[7] W. Xu, Towards Optimal one pass large scale learning with averaged sto-chastic gradient descent, 2011.
[8] L. Bottou, Large-Scale Machine Learning with Stochastic Gradient Descent, in: Proceedings of COMPSTAT, 2010, pp. 177–186.
[9] J. Wang, Y. Tang, M. Nguyen, I. Altintas A Scalable data Science workflow approach for Big data Bayesian network learning Proc. 2014 IEEE/ACM Int. Symp. Big Data Comput. (2014), pp. 16-25
[10] K. Yue, H. Wu, X. Fu, J. Xu, Z. Yin, W. Liu A data-intensive approach for discovering user similarities in social behav-ioral interactions based on the bayesian network Neurocomputing, 219 (2017), pp. 364-375
[11] A. Kumar, A. Beutel, Q. Ho, E.P. Xing, Fugue: Slow-Worker-Agnostic Dis-tributed Learning for Big Models on Big Data, in: Proceedings of the 17th Inter-national Conference on Artificial Intelligence and Statistics (AISTATS), Reykja-vik, Iceland, 2014, pp. 531–539.
[12]K. Sankar, H. Karau Fast Data Processing with Spark (Second ed.), Packt Publishing (2015)
[13]S. Owen, R. Anil, T. Dunning, E. Friedman Mahout in Action Manning Publications Co. (2011)
[14]C.T. Chu, S.K. Kim, Y.A. Lin, Y. Yu, G.R. Bradski, A.Y. Ng, et al. Map-reduce for machine learning on multicore. NIPS (2006), pp. 281-288
[15] A.K.Ghoting, R.E.Pednault, B.Reinwald, V.Sindhwani, S.Tatikonda, Y.Tian, et al., SystemML: Declarative machine learning on MapReduce, in: Pro-ceedings of the 27th International Conference on Data Engineering (ICDE), 2011.
[16] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, J.M. Hellerstein Distributed GraphLab: a framework for machine learning and data mining in the cloud
[17]Theano Development Team, Theano: A Python framework for fast compu-tation of mathematical expression. Available: arXiv:1605.02688.
[18] Y.Jia, E.Shelhamer, J.Donahue, S.Karayev, J.Long, R.Girshick, et al., Caffe: Convolutional Architecture for Fast Feature Embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, 2014.
[19] J.-x. Dong, A. Krzyzak, C.Y. Suen Fast SVM training algorithm with decomposition on very large data sets IEEE Trans. Pattern Anal. Mach. Intell., 27 (2005), pp. 603-618
[20] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, et al., Large scale distributed deep networks, in: Proceedings of the Neural Information Processing Systems, Lake Tahoe, Nevada, United States, 2012, pp. 1232–1240.





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