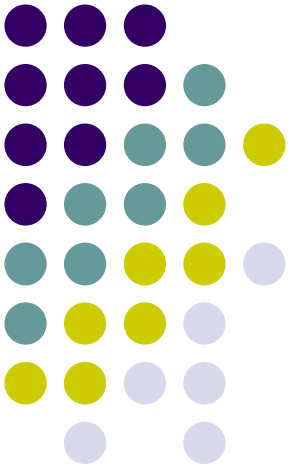
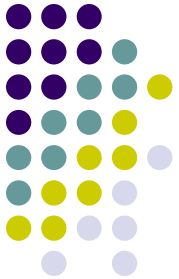
Zastosowanie technologii big data w uczeniu maszynowym

Filip Wójcik
Senior Data Scientist
http://maddatascientist.eu
filip.wojcik@outlook.com
filip.wojcik@ue.wroc.pl

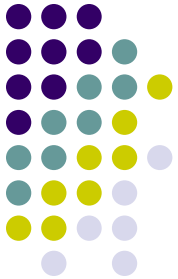


Uniwersytet Ekonomiczny
we Wrocławiu

Agenda



1. Rola i znaczenie technologii big data
2. Znaczenie Big Data dla uczenia maszynowego i przypadki użycia
3. Najpopularniejsze platformy big data



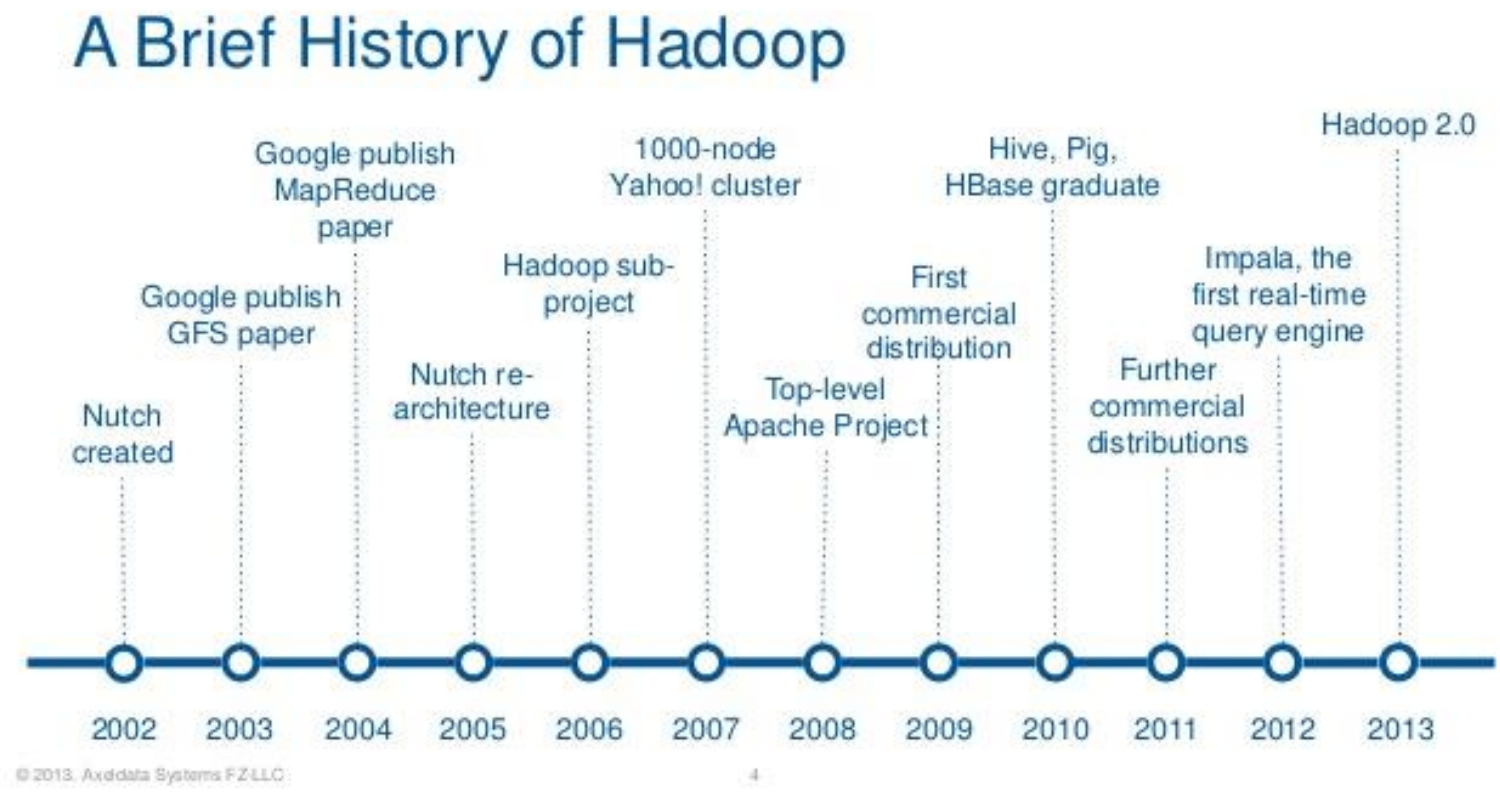
Historia i zarys technologii

ROLA I ZNACZENIE TECHNOLOGII BIG DATA

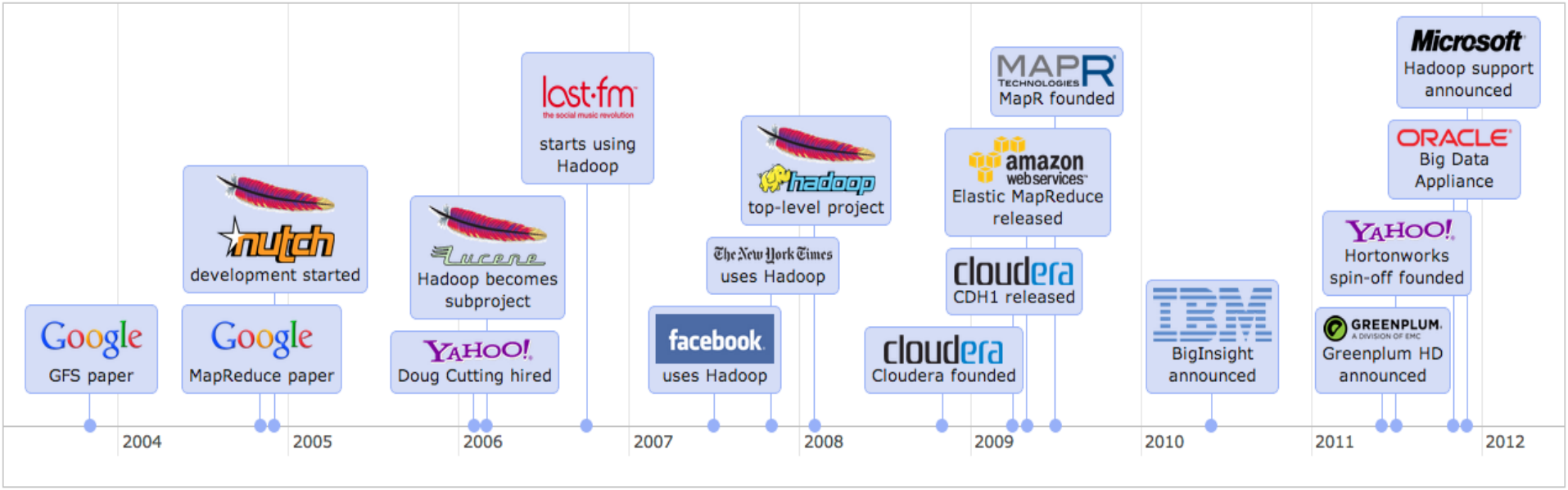
Rola i znaczenie technologii big data 1/6

- Technologie Big data nabrały znaczenia w miarę jak przestrzeń dyskowa zaczęła tanieć
- Złożoność procesów decyzyjnych i ilość produkowanych danych rosła wykładniczo
- Utrzymanie infrastruktury zdolnej do przetwarzania takich zbiorów stawało się coraz droższe
- Rozwój dostawców usług chmurowych (ang. *cloud computing*)

Rola i znaczenie technologii big data 2/6



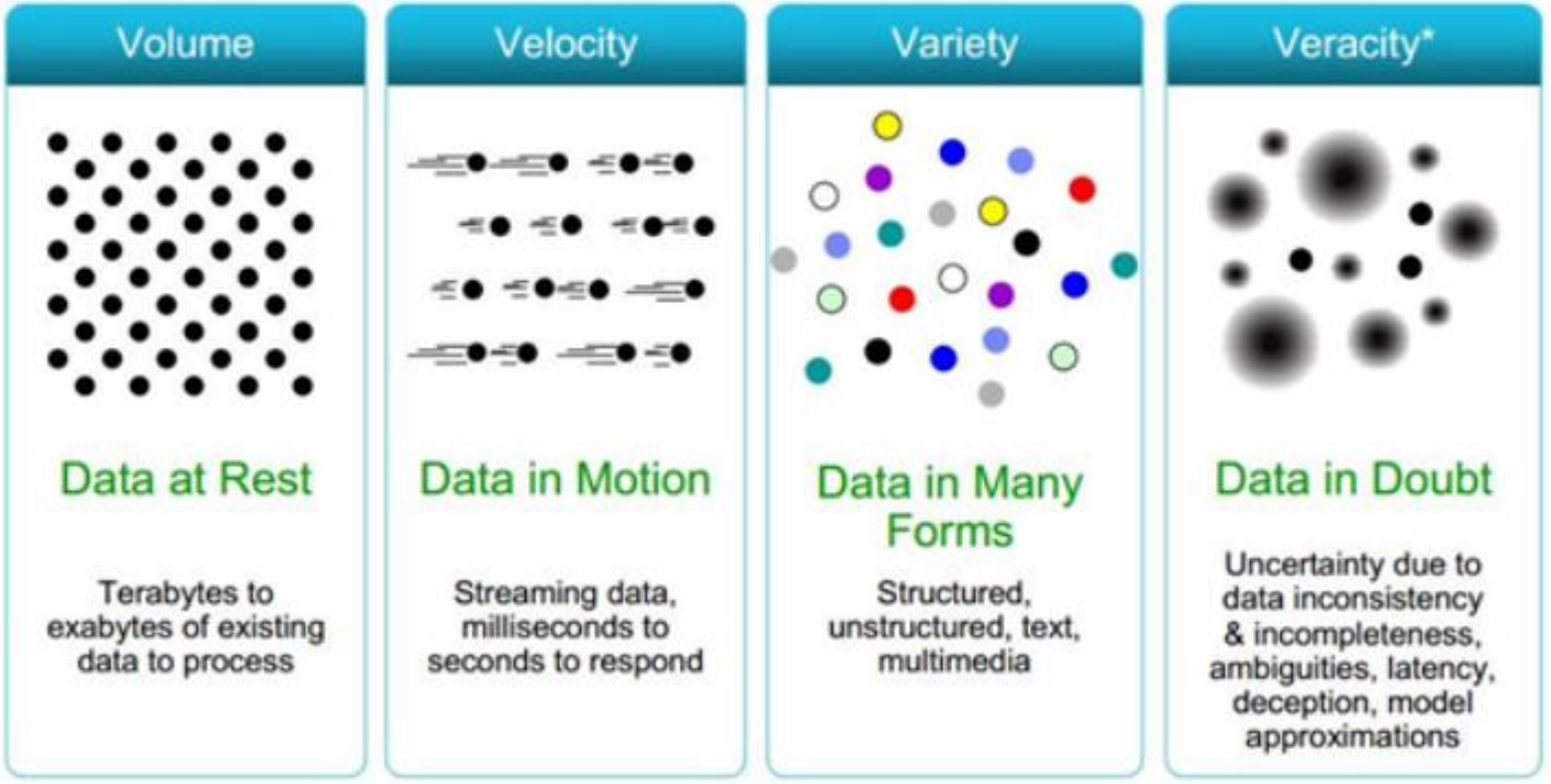
Rola i znaczenie technologii big data 3/6



Source: http://andego.hu/files/2013/01/timeline2.png

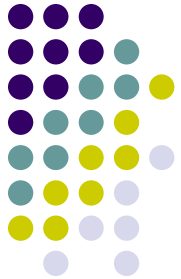
Rola i znaczenie technologii big data 4/6

Typowe problemy z danymi, motywujące do używania technologii Big Data



Source: http://blog.aziksa.com/2013/10/big-data-introduction/

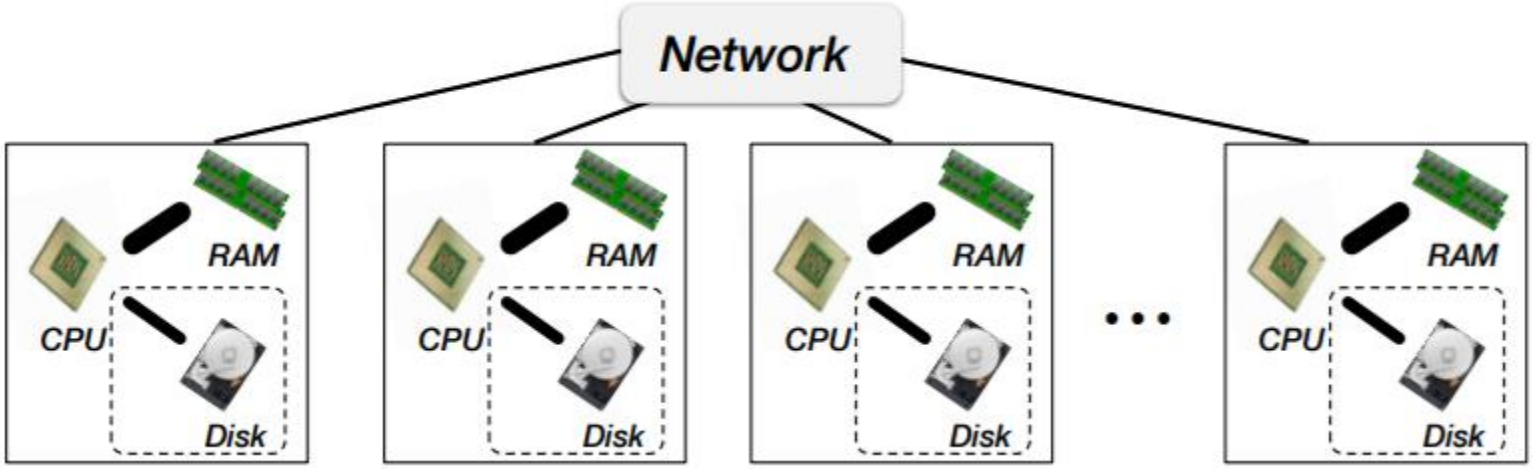
Rola i znaczenie technologii big data 5/6

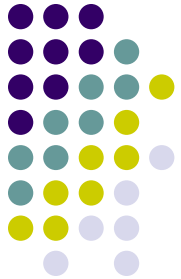


- Algorytm Map Reduce od Google pozwolił zrównoleglić obliczenia na tzw. *commodity hardware* czyli maszyny powszechnego użytku
- Dzięki temu możliwe stało się budowanie lokalnych klastrów z danymi wewnątrz firm i organizacji
- Takie skalowanie też ma swoje limity – stąd operatorzy chmurowi ☺

Rola i znaczenie technologii big data 6/6

Koordynacja operacji na klastrach Big Data za pomocą przesyłu sieciowego. Operacje są wykonywane lokalnie

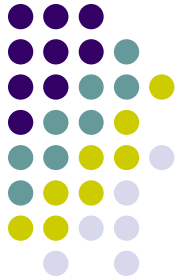




Dlaczego uczenie maszynowe i big data często przedstawiane są razem?

ZNACZENIE BIG DATA DLA UCZENIA MASZYNOWEGO

Znaczenie Big Data dla uczenia maszynowego 1/5

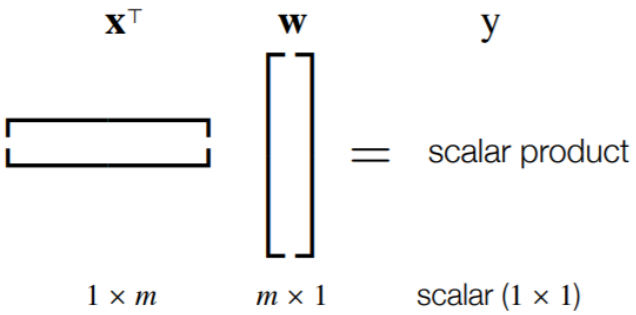


- Operacje uczenia maszynowego zazwyczaj są bardzo kosztowne obliczeniowo
- Nierzadko potrzebują wyliczyć własności danych w oparciu o cały zbiór (np. Entropia Shannona)
- Obok statystycznej analizy danych podstawą uczenia maszynowego są operacje na wektorach i macierzach – przy dużych zbiorach danych to nie zdaje egzaminu

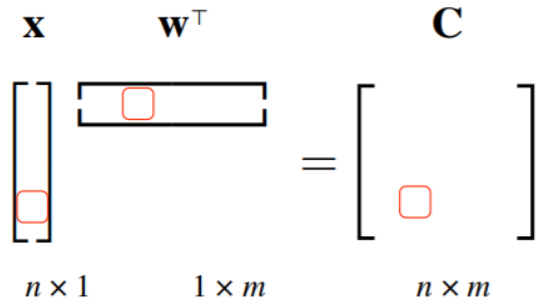
Znaczenie Big Data dla uczenia maszynowego 2/5

Problematyczne operacje algebry liniowej

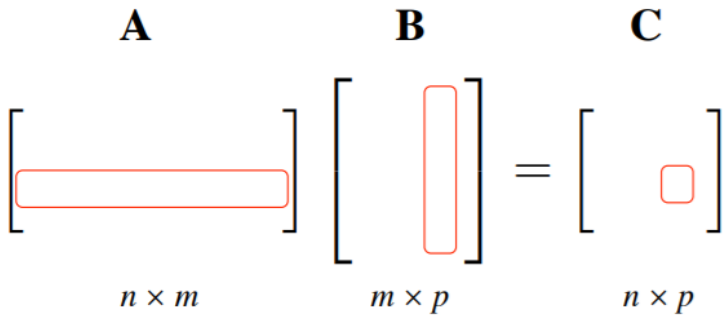
mnożenie dużych wektorów i macierzy



Iloczyn wektorowy



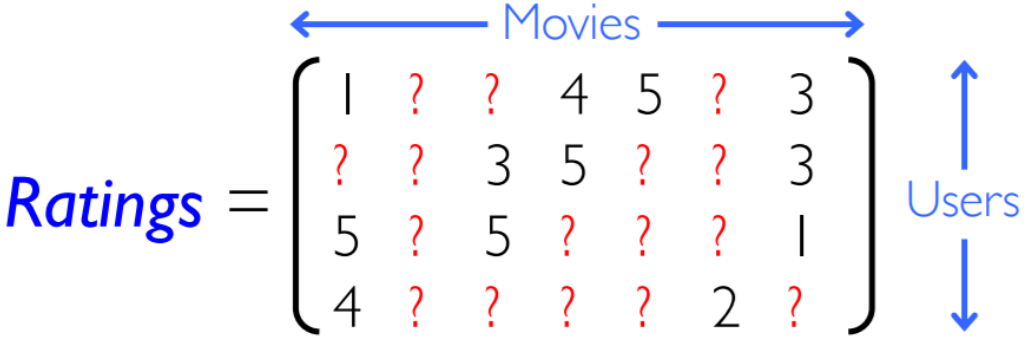
Produkt diadyczny (ang. *outer product*)

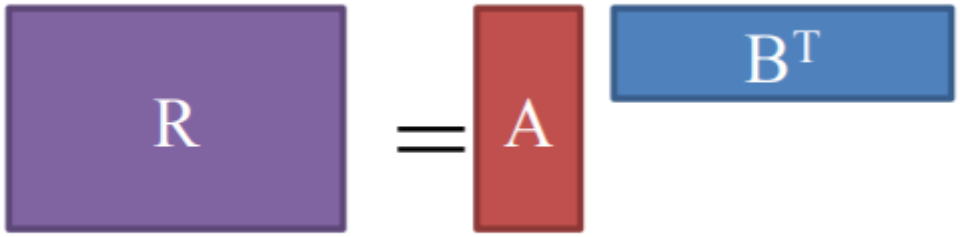


Mnożenie macierzy

Znaczenie Big Data dla uczenia maszynowego 3/5

Przykładem mogą być **systemy rekomendacyjne** oparte na rozkładzie macierzy (SVD, NNMF). Wymagają rozkładu **całej** macierzy na składowe.

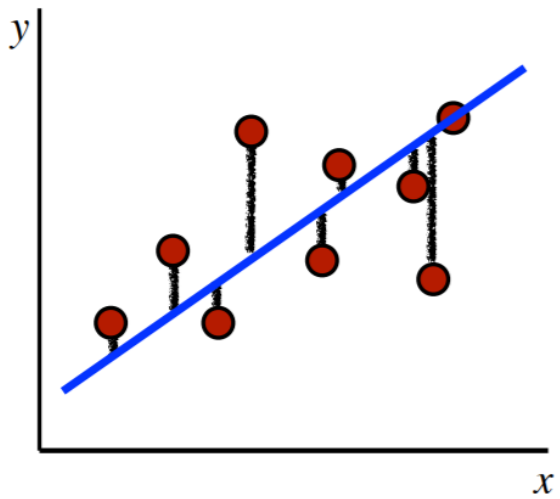




Znaczenie Big Data dla uczenia maszynowego 4/5

Kolejny przykład – **klasyczna regresja liniowa.** Wymaga mnożenie przez siebie macierzy wag *w* oraz macierzy zmiennych egzogenicznych *x*

$$y \approx \hat{y} = w_0 + w^T X \qquad w = (X^T X)^{-1} X^T y$$



Kilka operacji matematycznych o dużej złożoności:

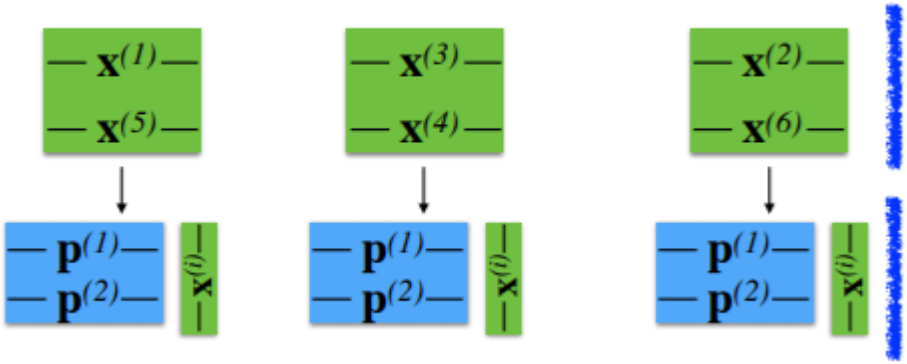
- Iloczyn macierzy $O(nd^2)$
- Odwracanie macierzy (włącznie z liczeniem wyznacznika): $O(d^3)$

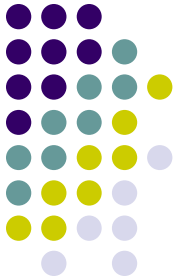
Znaczenie Big Data dla uczenia maszynowego 5/5

Ostatni przykład – rozkład macierzy na składowe główne (**PCA**) obejmujący wyliczanie najważniejszych jej komponentów w oparciu o macierz kowariancji.

$$\mathbf{Z} = \mathbf{X}\mathbf{P}$$

$$\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}^{(i)}\mathbf{x}^{(i)}$$

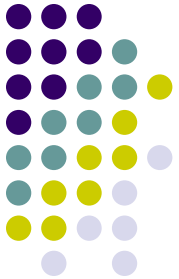




Jakie platformy chmurowe Big Data są najpopularniejsze i dlaczego?

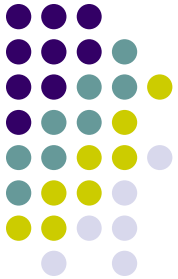
CHMUROWE PLATFORMY BIG DATA

Chmurowe platformy Big Data 1/3



- Okazało się, że założenie o tzw. *commodity hardware* nie do końca jest takie idealne
- Można mieć infrastrukturę złożoną z powszechnych komponentów, ale ktoś musi je utrzymywać w fizycznej lokalizacji
- Stąd rosnąca popularność dostawców platform chmurowych
- Oferowane są w połączeniu z usługami uczenia maszynowego i sztucznej inteligencji (AIAS –*AI as a service*)

Chmurowe platformy Big Data 2/3



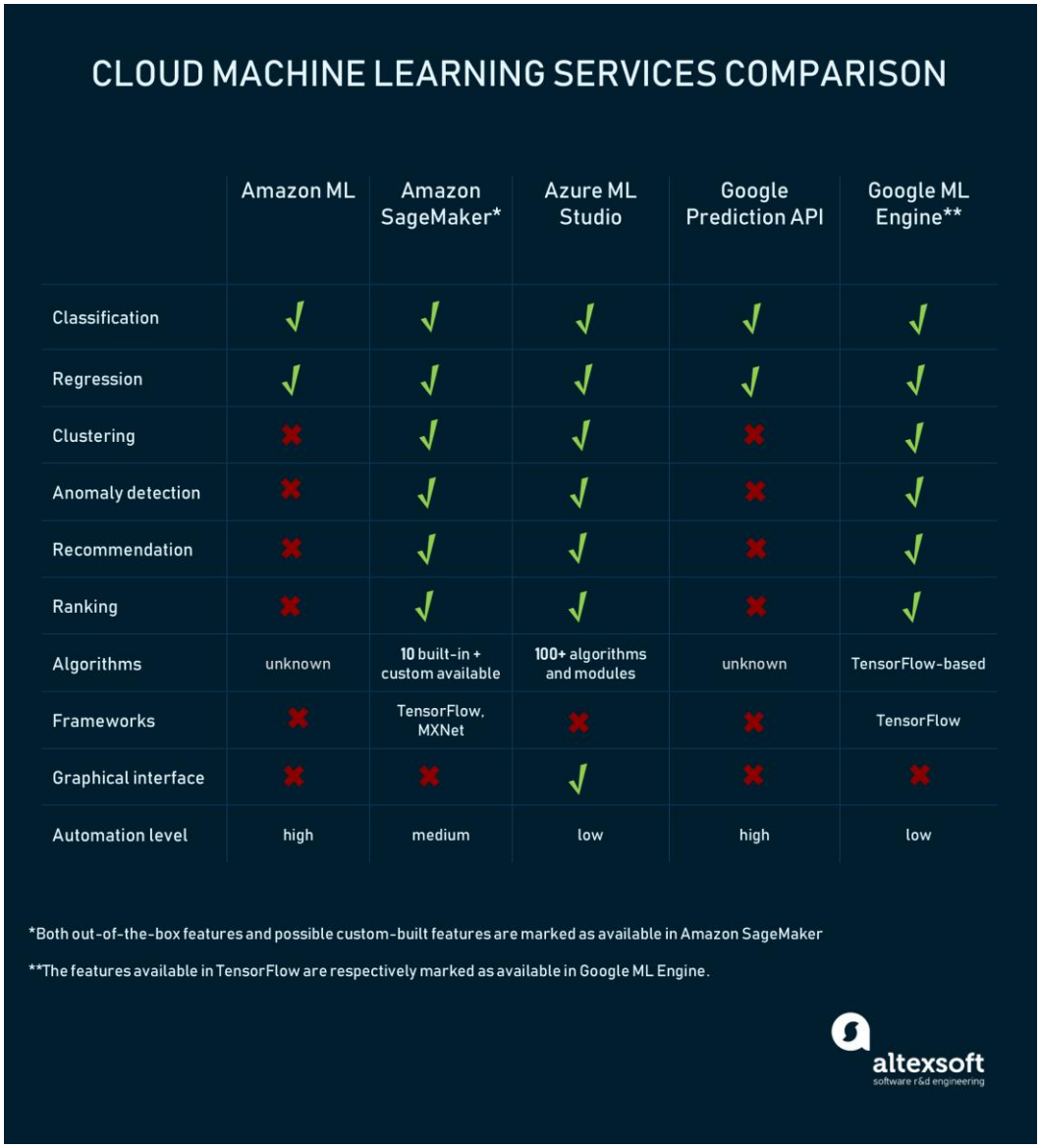
- Microsoft Azure jest jedną z najdynamiczniej rozwijających się platform
- MS oferuje rozwiązania własne oraz Open Source
- Chmura ma charakter „umiarkowanie zaawansowany” – w miarę prosty interfejs użytkownika i prosta konfiguracja

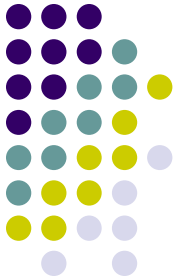
- AWS jest największą platformą chmurową
- Najczęściej wybierane rozwiązanie przez startupy
- Uważane za niskopoziomowe i wymagające dużej konfiguracji
- Oferowane są produkty zarówno własne jak i open source

- Google stawia głównie na własne rozwiązania, nawet jeśli są po prostu odmianą istniejących technologii open source
- Najprostsze rozwiązanie w konfiguracji i zarządzaniu

- Słynna platforma AI, która nauczyła się grać w GO oraz gry komputerowe
- Uważana za najbardziej zaawansowaną sztuczną inteligencję oferowaną jako usługa
- Jednocześnie najdroższa ze wszystkich

Chmurowe platformy Big Data 3/3





Dziękuję za uwagę