Big Data: nowe wyzwanie dla systemów uczących się

Jacek Koronacki
Instytut Podstaw Informatyki, PAN

Jerzy Stefanowski
Instytut Informatyki, Politechnika Poznańska

Częstochowa, 14 kwietnia 2016

Big Data – dlaczego?

Popularne „buzzwords”:

….

Data mining – przełom lat 80/90

Big Data – obecnie

Coraz częściej rozmawiamy, lecz czy ….

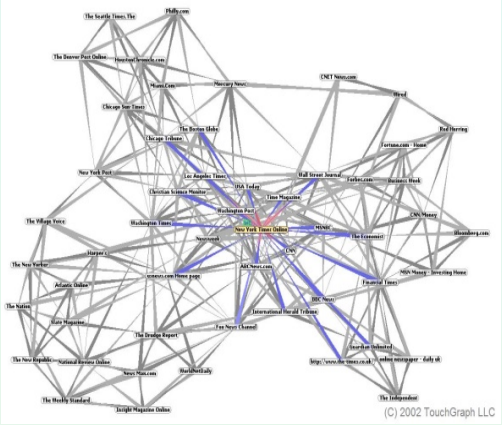
“Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it …” [Dan Ariely]

Spojrzenie badacza ML – zróżnicowane:

… I have been doing “Big Data” for years (though I didn’t call it that and didn’t know that it would, one day, be called that) …

… Is Big Data Analysis a game changer?

Nowe źródła danych



Rozwój technologiczny

Rosnąca wielkość „tradycyjnych” repozytoriów

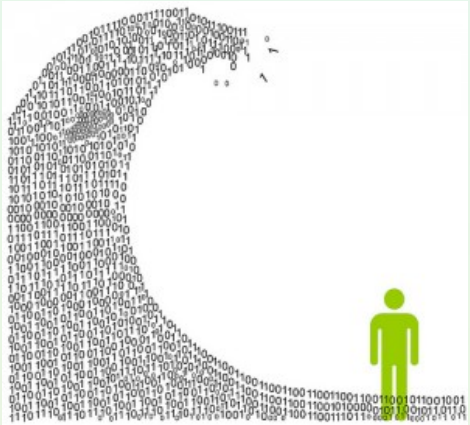
Nowe żródła danych i wzrost złożoności danych:

- Internet, media społecznościowe
- Sieci sensorów, urządzenia mobilne, dane generowane maszynowo

„In many fieds costs of data acquisition are now lower, much lowerr than costs of data analysis”

[Canadian Statistical Sci . Inst 2015]

Wzrost przekonania o roli Big Data i pytania
→ czy następuję rewolucyjna zmiana?



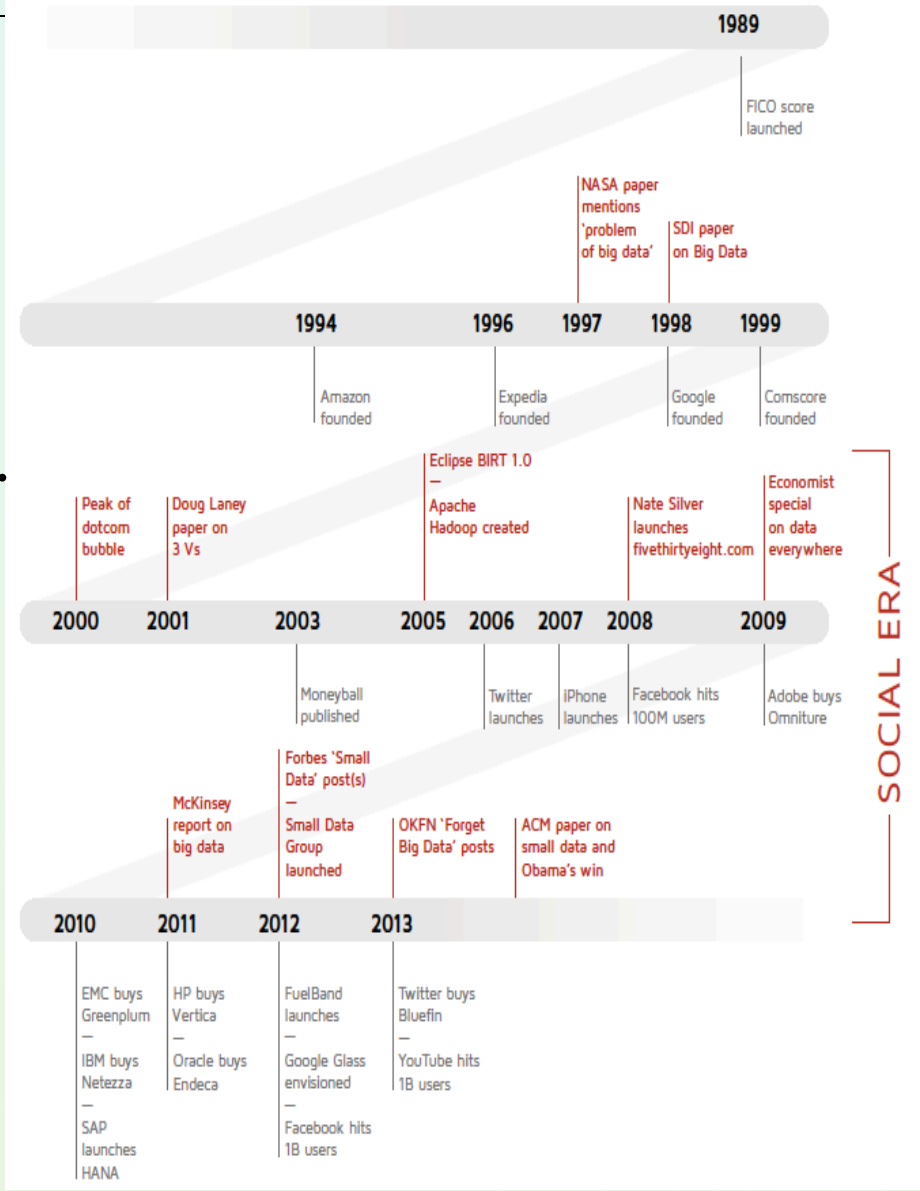
Kilka uwag historycznych

Dyskusja rosnących rozmiarów danych vs. pojęcie Big Data

Massive Data mining, Very Large Databases – inne spojrzenie

"Big Data" – jako termin na amer. konf. pojawia się pod koniec lat 90tych:

- J.Mashey seminarium SGI, 1998
- S. Weiss, N.Indrukhya: Predictive data mining. A practical guide 1998
  - Rozdział 1.1. Big Data. – duża część rozdziału 1 dyskutuje też problemy „Massive data".
- Doświadczenia wielkich projektów badawczych – NASA



Definicje Big Data

"I shall not further attempt to define Big Data, but I know it when I see it" [Bell 2015]

„Zbiór danych, które z uwagi na wielkość i złożoność nie mogą być przetwarzanie tradycyjnymi środkami informatycznymi"

"Big Data" – to dane, których skala, zróżnicowanie i złożoność wymaga nowych technologii i algorytmów w celu odkrycia wartościowej wiedzy … [J.Gama 2015]

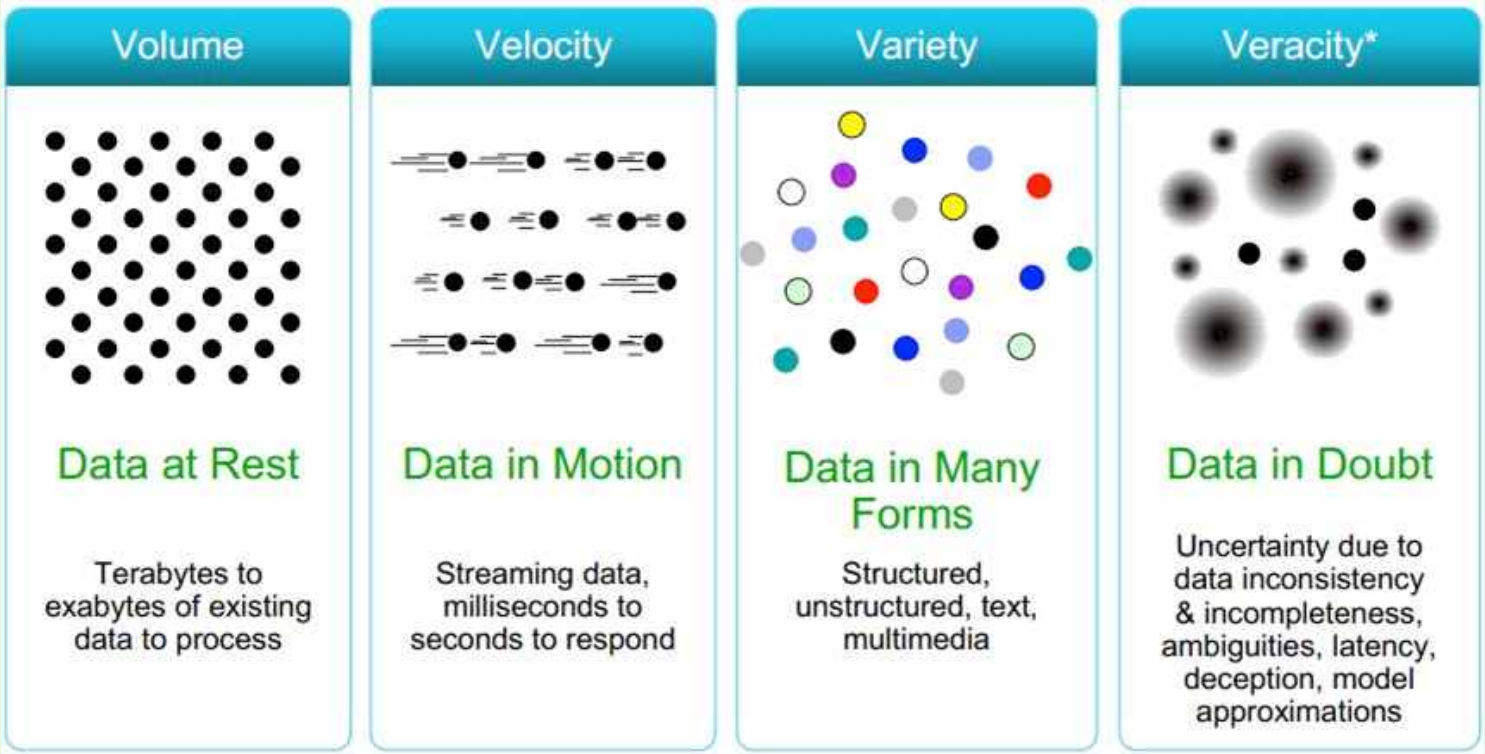
Przegląd różnych definicji:

W.Fan, A.Bifet: Mining Big Data: Current status and forecast to the future. SIGKDD Explorations, vol. 14 (2), 1-15, 2012

A de Mauro, M. Greco, M. Grimaldi: What is Big Data? A consensual definition and a review of key research topics. Proc. 4th Conf. on Integrated Information 2014.

Wiele V – charakterystyka Big Data

3 V → „High volume, velocity and variety” [Doug Laney 2001]

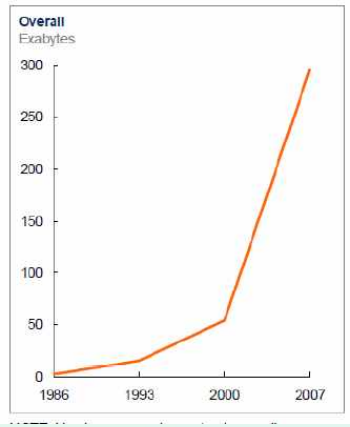


2012 IBM Solutions presentation

Kolejne V’s stopniowo dodawano do definicji (Value, Variability, …)

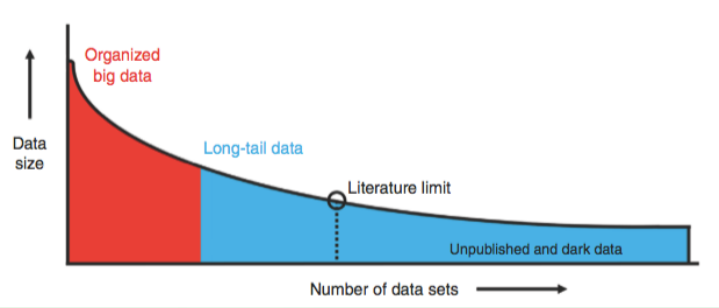
Volume - wielkość

- Big Data Age – wykładniczy wzrost danych elektronicznych
  - Terabytes, petabytes, exabytes, …



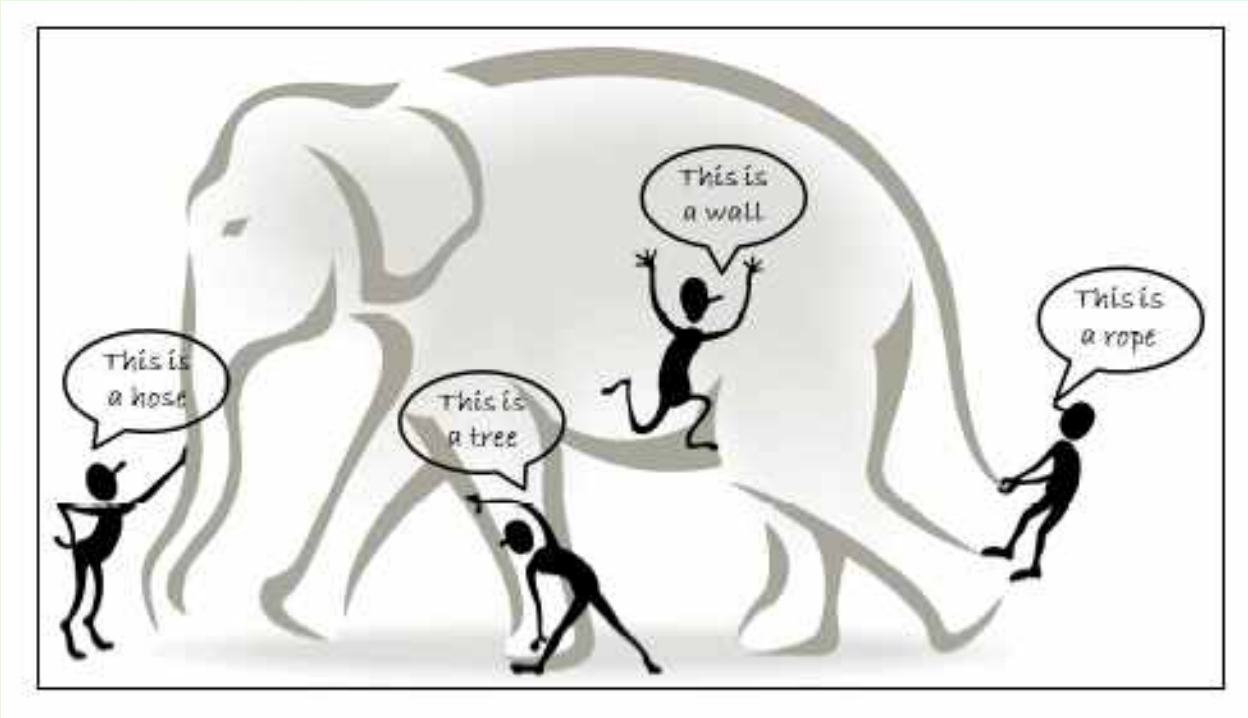
„The size of Big Data is a moving target - what is considered to be Big today will not be so years ahead”

- Wymagania wydajnościowe
  - Ograniczenia na czas przetwarzania, odpowiedzi systemu
  - Trudności z przechowywaniem danych



Nowe wyzwania dla algorytmów eksploracji danych !

Charakterystyka Big Data – połączenie niejednorodnych i złożonych źródeł danych



HACE Theorem: Big Data starts with large-volume, **h**eterogeneous, **a**utonomous data sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data

[Xindong Wu et al. 2013]

Variety → złożone, niejednorodne reprezentacje danych

Różne formaty, typy i struktura danych

- Strukturalne - podstawowe typy danych (relacje baz danych)
- Semi-strukturalne – XML, JSON, Bib
- Teksty, blogi, dokumenty, obrazy, video, audio, sygnały, dane lokalizacyjne, …
- Wiele aplikacji generuje dane w różnych formatach
- Tylko 10% Big Data dobrze ustrukturalizowana

Variety → złożoność danych

Lecz również zróżnicowana wielo-wymiarowość

Trudności przetwarzania takich danych, ich integracji, …

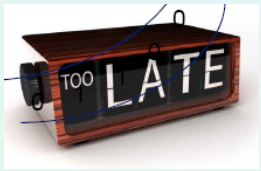
Czy integrować, czy analizować niezależnie w locie?

Variety → zmiany wymiarowości danych

- Niejednorodne, złożone reprezentacje (poza tabela $n \times p$)
- Lecz również (Diverse Dimensionality) → różne liczności atrybutów
  - Przykłady danych medycznych
- Często wysoko wymiarowe (**large p**)
  - Obrazy (statyczne, video)
  - Teksty
  - Sygnały
  - Nietypowe opisy grafów
  - Dane biomedyczne → p większe niż n
  - Genetyka ($p >> n$)
- Wiele problemów dla klasycznych metod (n rośnie przy stałym p) → **Potrzeba nowych algorytmów eksploracji danych !**

Velocity – czas, prędkość i zmienność

- Big Data – generowane z dużą szybkością i wymagają odpowiednio szybkiego przetwarzania
- Online Data Analytics
- Spóźnione decyzje → nieakceptowane straty
- Przykłady
  - e-marketing → oferty we właściwym czasie
  - Monitorowanie stanu zdrowia
  - Decyzje finansowe
  - Wykrywanie defraudacji, zagrożeń
- Wiele problemów → strumienie danych
  - Przyrostowe przetwarzanie
  - Zmienność rozkładów danych → „concept drift”



Potrzeba nowych algorytmów eksploracji danych !

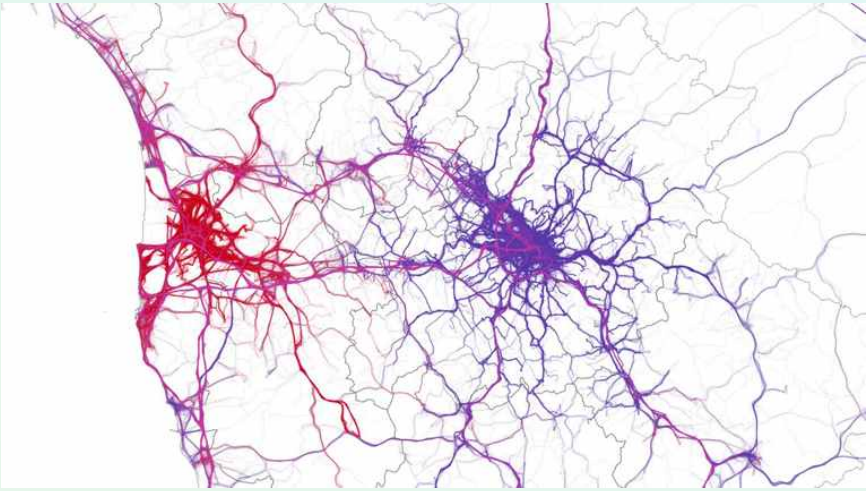
Veracity – jakość, pewność i pochodzenie danych

- Dane pochodzą z wielu źródeł, których jakość jest różna
  - Często odbiega od standardów, zawiera błędy, np. posty w „ social media”
  - Proces łączenia wielu źródeł danych, których się nie kontroluje
- Niepewność dostępnych danych inna niż w klasycznej statystycznej analizie danych
  - Niespójności elementów danych
  - Niekompletności
  - Niejednoznaczności opisów
- Trust / Provenance
  - Poprzednio → staranne wstępne przygotowanie danych, udokumentowanie ich pochodzenia oraz wykonanych operacji przetwarzania
  - Obecnie → trudności w udokumentowaniu drogi pozyskania danych
  - „Trust” → ocena wiarygodności źródeł danych i ich etapów przetwarzanie

Rezultaty eksploracji danych silnie zależą od jakości danych!

Studium przypadku → analiza mobilności ludzi

- "Unveiling the complexity of human mobility by querying and mining massive trajectory data". Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, Roberto Trasarti
- KDD Lab, University of Pisa
- Złożone niejednorodne repozytoria danych → w tym eksploracja b. wielu zapisów (trajektorii) ruchu pojazdów (GPS) oraz pasażerów trasport. (mobile phones)
- Przykładowe pytania i zadania:
  - Odkryj wzorce podróży
  - Gdzie są obszary intensywne ruchu samochodowego? Jak się zmieniają w czasie?
  - Co wpływa na zmiany ruchu?
  - Powiązanie danych z samochodów z danymi demograficznymi.
- M-Atlas → wizualna eksploracja odkrytych wzorców



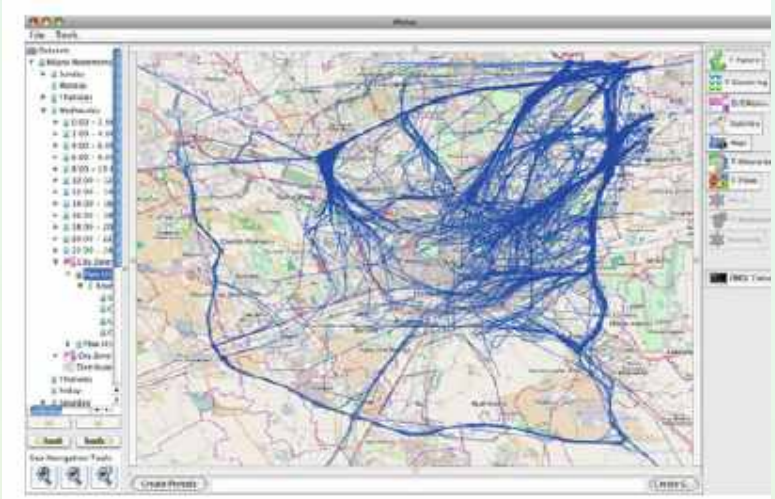


Fig. 17 The result of T-Clustering from the trajectories moving from the center to the North-East area. *Left* The input data set for the clustering algorithm: the trajectories moving from the center to the North-East

Automatyczne tłumaczenie tekstów

Projekt Google Translate

Statystyczna eksploracja olbrzymich repozytoriów danych pozwala na tłumaczenie (nie wszystkich) tekstów lepiej niż tradycyjne modele oparte na wyrafinowanych algorytmach

Dostęp do nowych repozytoriów

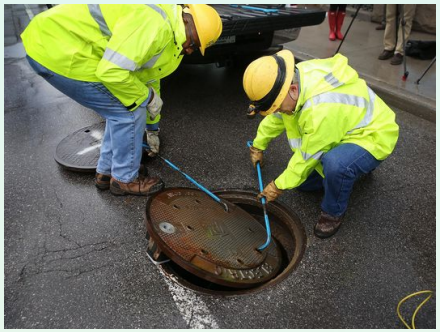
- Angielskie vs. inne języki tłumaczenia różnych publicznie dostępnych tekstów
- UE dokumenty, prawo → tłumaczone na języki krajów europejskich
- Specjalne projekty → digitalizacja książek, tekstów (Google Books)

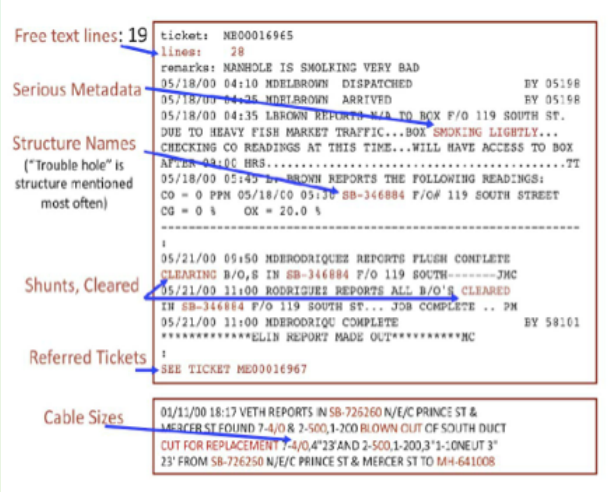
„The bigger the corpus, or body of parallel texts, the better the result”

Halevey A., Norvig P., Pereira F. The unreasonable effectiveness of data. 2009.

Predykcja awarii studzienek sieci elektrycznej

- Podziemna sieć elektryczna Nowego Jorku (90tys mil kabli, ponad 100 lat) – zbyt częste pożary i wybuchy w tunelach i studzienkach – predykcja miejsca potencjalnych awarii
- C. Rudin et al + Con Edison
- Różnorodne repozytoria danych → w tym duże ilości raportów, protokołów (np. trouble tickets – dok. tekstowy)
  - Różnorodność reprezentacji danych
  - Konieczność NLP i analizy tekstów
  - Intensywny pre-processing, ocena wiarygodności i pochodzenia danych
  - Interakcja z ekspertami Con Edison w celu etykietyzacji opisów uszkodzeń
- Algorytmy UM (SVM + ranking learning)
  - Stwórz ranking niebezpieczeństwa studzienek
  - Wytyczne dla ekip kontrolnych





Rudin, C., Passonneau, R., Radeva, A., Jerome, S., Issac, D.: 21st century data miners meet 19-th century electrical cables. IEEE Computer, 103-105, (June 2011).

Inne zastosowania

Popularny (zbyt!) przegląd:

Viktor Mayer-Schönberger, Kenneth Cukier. Big Data: a revolution that will transform how we live, work and think. Eamon, Dolan/ Houghton Mifflin Harcourt (2013) → Polskie tłumaczenie (2014)

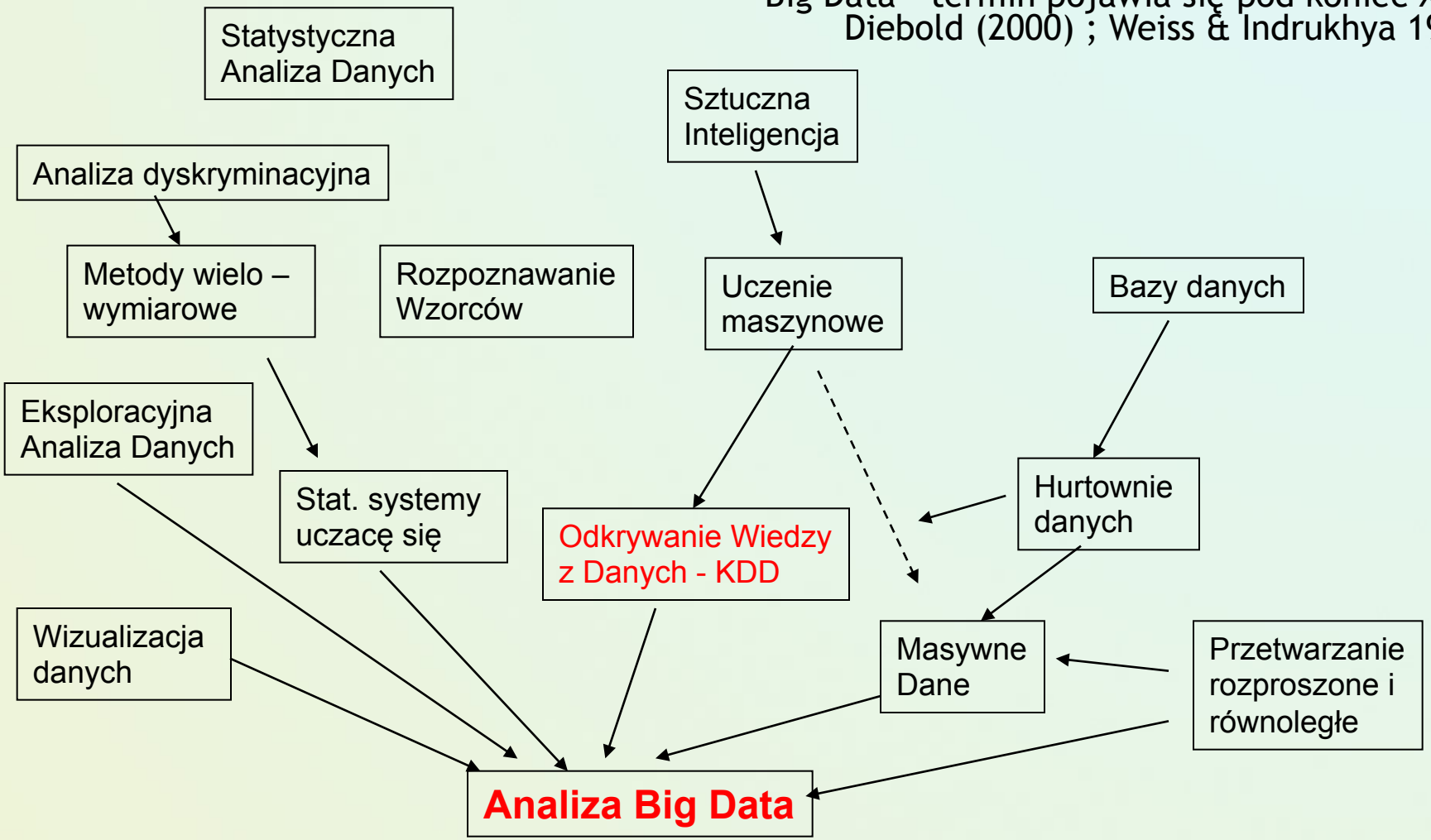
Oraz wiele innych artykułów przeglądowych

> “Big data provides us with more detailed information – often in real time – making it possible to measure properties and patterns of human behavior that formerly could not be measured”
>
> [Gaffield 2015]

> “Big Data, on the other hand, gives us the opportunity not to aggregate the behavior of millions, but instead to take it into consideration at the micro-level”
>
> [Alex Pentland]

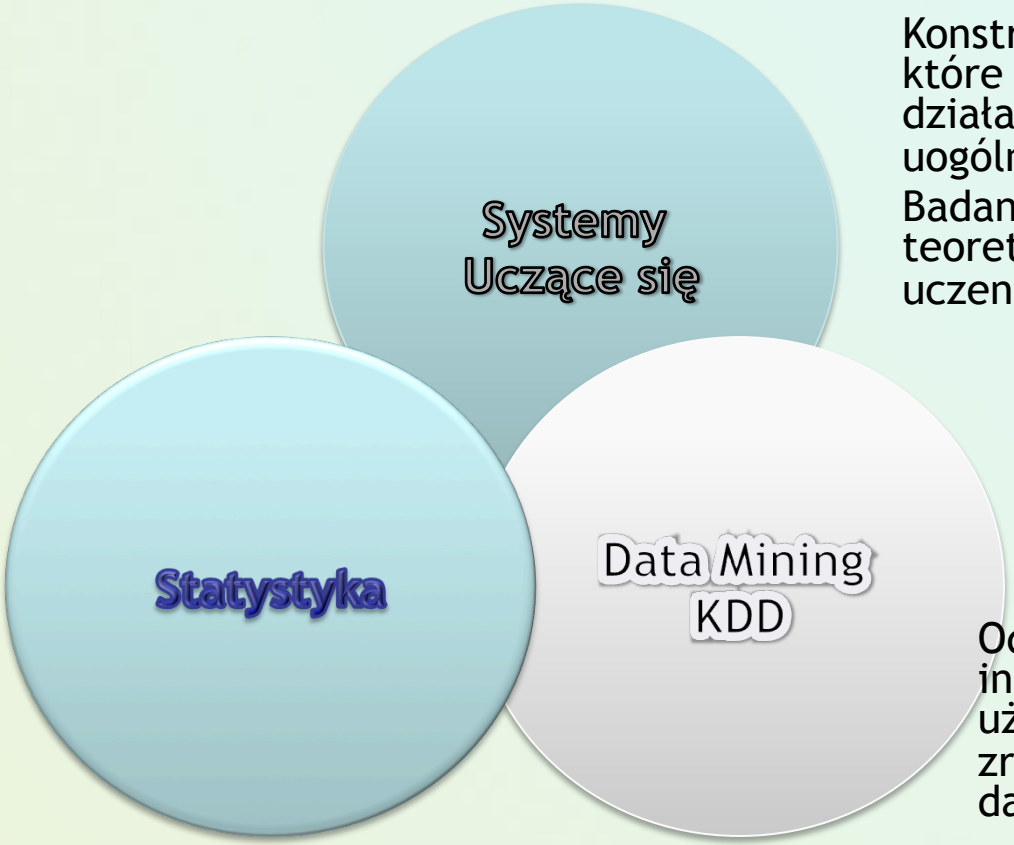
Powiązania dziedzin i …

Big Data – termin pojawia się pod koniec XX w.
Diebold (2000) ; Weiss & Indrukhya 1998



KDD – proces poszukiwania nowych, interesujących, potencjalnie użytecznych (i zrozumiałych) wzorców z danych – G.Piatesky-Shapiro (1989)

3 bliskie dziedziny



Konstruowanie systemów, które doskonalą swoje działanie na podstawie uogólniania doświadczenia.

Badanie podstaw teoretycznych procesu uczenia się:

Odkrywanie nowych, interesujących, potencjalnie użytecznych (i potencjalnie zrozumiałych) wzorców z danych

Powiązanie i różnice między nimi – dyskutowane od lat przez wielu autorów

KDD/Data Mining vs Analiza Big Data

Zagadnienie	Tradycyjne Data Mining	Analiza Big Data
Dostęp do pamięci	Centralna pamięć operacyjna, łatwiejsze wielokrotne operacje odczyt / zapis	Dane często rozproszone Minimalizowanie zapamiętanych elementów i dostępu do nich
Architektura oblicz.	Centralna pojedyncza jednostka (skalowalna)	Rozproszone przetwarzanie Grona (clusters) słabszych komputerów
Dane	Dobra strukturalizacja (rel. DB), jednorodne, statyczne / integracja DW	Zróżnicowane źródła; brak struktury; Zmienne / dynamika i czas
Jakość danych	Dobrze przygotowane lub popr. wiele technik korekcji Udokumentowane pochodzenie Wiarygodne próbkowanie Wyselekcjonowane „dobre dane”	Słaba jakość danych, niepewność i niedokładność; Słabo dokument. pochodzenie i pre-processing; Użyteczne dane mogą być połączone z wieloma bezużytecznymi
Bezpieczeństwo i prywatność	Nie są wymagane Proste metody „anonimizacji”	Krytyczny problem Współdzielenie danych; łączenie danych
Przetwarzanie danych	Klasyczne (batch); może być off-line Brak konieczności próbkowania Prędkość nie tak krytyczna	Możliwość – wymagania on-line; szybkość; Wydajność alg. ma znaczenie Dane nie mieszczą się w pamięci Kompresja i próbkowanie danych
Analiza rezultatów	Rozwinięte metody oceny wyników oraz wizualizacji	Niebezpieczeństwa odkrycia nieznaczących rezultatów Trudności wizualizacji

Więcej w N.Japkowicz, J.Stefanowski: A Machine Learning Perspective on Big Data Analysis (2016)

Big Data – znane metody w innym kontekście

Standard metod Data Mining

- Klasyfikacja nadzorowana
- Regresja / ANN
- Analiza skupień
- Asocjacje (zbiory częste, reguły asocjacyjne)
- Wzorce sekwencji
- Szeregi czasowe
- Wykrywanie anomalii i obserwacji nietypowych
- Statystyka opisowa
- Statystyka wielowymiarowa
- Dekompozycja macierzy (PCA, MDS,…)
- …

Klasyfikacja i predykcja

- Drzewa decyzyjne
- Reguły
- Naive Bayes (+ sieci Bayesowskie)
- K-NN
- Sztuczne sieci neuronowe
- Analiza dyskryminacyjna
- Metoda wektorów wspierających SVM
- Zespoły klasyfikatorów
- …

Spójrz: P. Tan, M. Steinbach, V. Kumar: An Introduction to Data Mining

J.Berman: Principles of Big Data Analysis

Large Scale Machine Learning

- Large *n*, large *p*, large *k* (number of outputs)
  - Czasami inne zależności (large *p*, smaller *n*,*k*)

"The computing resources available do not grow faster than the volume of data"

Needs for linear time learning algorithms

Leon Bottou

Ideal O(*pn*+*kn*)

- Francis Bach i inni:

  Rozwijajmy prostsze metody:
  - Stochastic gradient methods (Robbins and Monro, 1951)
  - **Mixing statistics and optimization**

Spójrz do: F.Bach: Stochastic gradient methods for machine learning 2013

Także prace Michael I. Jordan, np., On statistics, computation and scalability, 2013.

Poprawa czasu obliczeń wybranych algorytmów

Core methods of statistics / machine learning / mining

- **Querying:** nearest-neighbor **O(N)**, spherical range-search **O(N)**, orthogonal range-search **O(N)**, contingency table
- **Density estimation:** kernel density estimation **O(N^2)**, mixture of Gaussians **O(N)**
- **Regression:** linear regression **O(D^3)**, kernel regression **O(N^2)**, Gaussian process regression **O(N^3)**
- **Classification:** nearest-neighbor classifier **O(N^2)**, nonparametric Bayes classifier **O(N^2)**, support vector machine
- **Dimension reduction:** principal component analysis **O(D^3)**, non-negative matrix factorization, kernel PCA **O(N^3)**, maximum variance unfolding **O(N^3**
- **Outlier detection:** by robust L_2 estimation, by density estimation, by dimension reduction
- **Clustering:** k-means **O(N)**, hierarchical clustering **O(N^3)**, by dimension reduction
- **Time series analysis:** Kalman filter **O(D^3)**, hidden Markov model, trajecto tracking
- **2-sample testing:** n-point correlation **O(N^n)**
- **Cross-match:** bipartite matching **O(N^3)**

Computational complexity using fast algorithms

- **Querying:** nearest-neighbor **O(logN)**, spherical range-search **O(logN)**, orthogonal range-search **O(logN)**, contingency table
- **Density estimation:** kernel density estimation **O(N) or O(1)**, mixture of Gaussians **O(logN)**
- **Regression:** linear regression **O(D) or O(1)**, kernel regression **O(N) or O(1)**, Gaussian process regression **O(N) or O(1)**
- **Classification:** nearest-neighbor classifier **O(N)**, nonparametric Bayes classifier **O(N)**, support vector machine **O(N)**
- **Dimension reduction:** principal component analysis **O(D) or O(1)**, non-negative matrix factorization, kernel PCA **O(N) or O(1)**, maximum variance unfolding **O(N)**
- **Outlier detection:** by robust L_2 estimation, by density estimation, by dimension reduction
- **Clustering:** k-means **O(logN)**, hierarchical clustering **O(NlogN)**, by dimension reduction
- **Time series analysis:** Kalman filter **O(D) or O(1)**, hidden Markov model, trajectory tracking
- **2-sample testing:** n-point correlation **O(N^{logn})**
- **Cross-match:** bipartite matching **O(N) or O(1)**

Za – Alexander Grey: Machine learning on massive data

Drzewa klasyfikacyjne i … (regresji …)

80's

ID3

CART

CHAID

…

C4.5

Oblique trees

90

Ensembles

Mega-induction,
Sampling, splits

Bagging

Boosting

Random Forests

Ivotes

KDD generalizations

end. 90

Data pre-sorting / reduce disc acc.

Approx. representations
Samples, Histograms

Rotation Forests

Gradient Boosted Machines

SLIQ
SPRINT
…

BOAT
CLOUDS
SPIES

Parallel versions

Incremental trees
ID4, ID5R, ITL

Massive incremental tress
Hoeffding trees, **VFDT**

Stream ensembles
Online …/ block based

MOA,
SAMOA

Data streams + drifts
VFDT -> CVFDT; **VFDTc**

Rozwój implementacji w MapReduce, Hadoop / SPARK

Big Data z punktu widzenia uczenia maszynowego

Nowe problemy badawcze

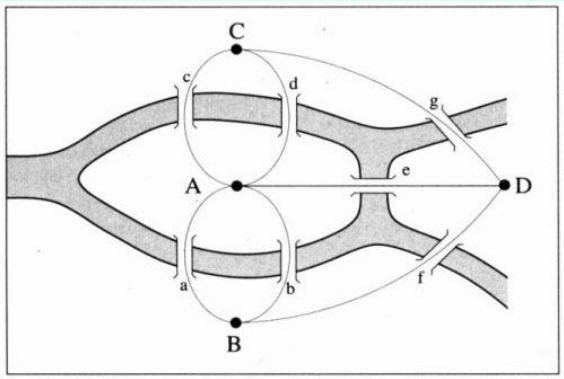
- Analiza grafów
- Social networks
- Integracja lub przetwarzanie on-line różnorodnych reprezentacji danych
- Eksploracja danych strumieniowych
- Analiza danych czasowo-rozproszonych
- Obliczenia mobilne (IoT)
- Wizualizacja danych
- Privacy data mining
- Data trust + provenance
- …

Inne problemy

- Interakcja z ekspertem
- Ocena wiedzy
- Etyka analizy Big Data
- Wpływ na społeczeństwo
- …

Sprawdź: N.Japkowicz and J.Stefanowski (Eds), Big Data Analysis: New Algorithms for a New Society, Springer (2016).

Eksploracja grafów



Grafy – (wierzchołki, krawędzie)

Lata 80

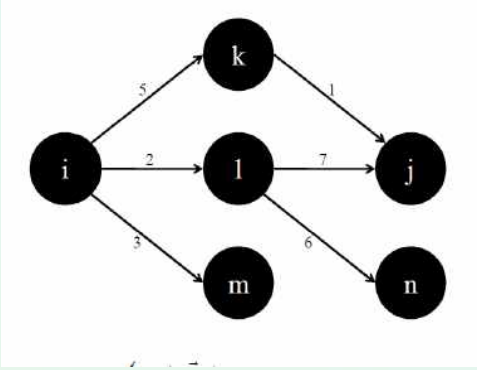
- Kilkadziesiąt wierzchołków

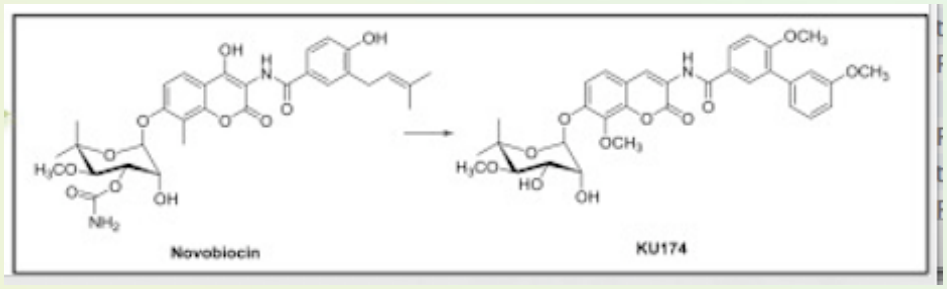
Lata 90

- Setki wierzchołków

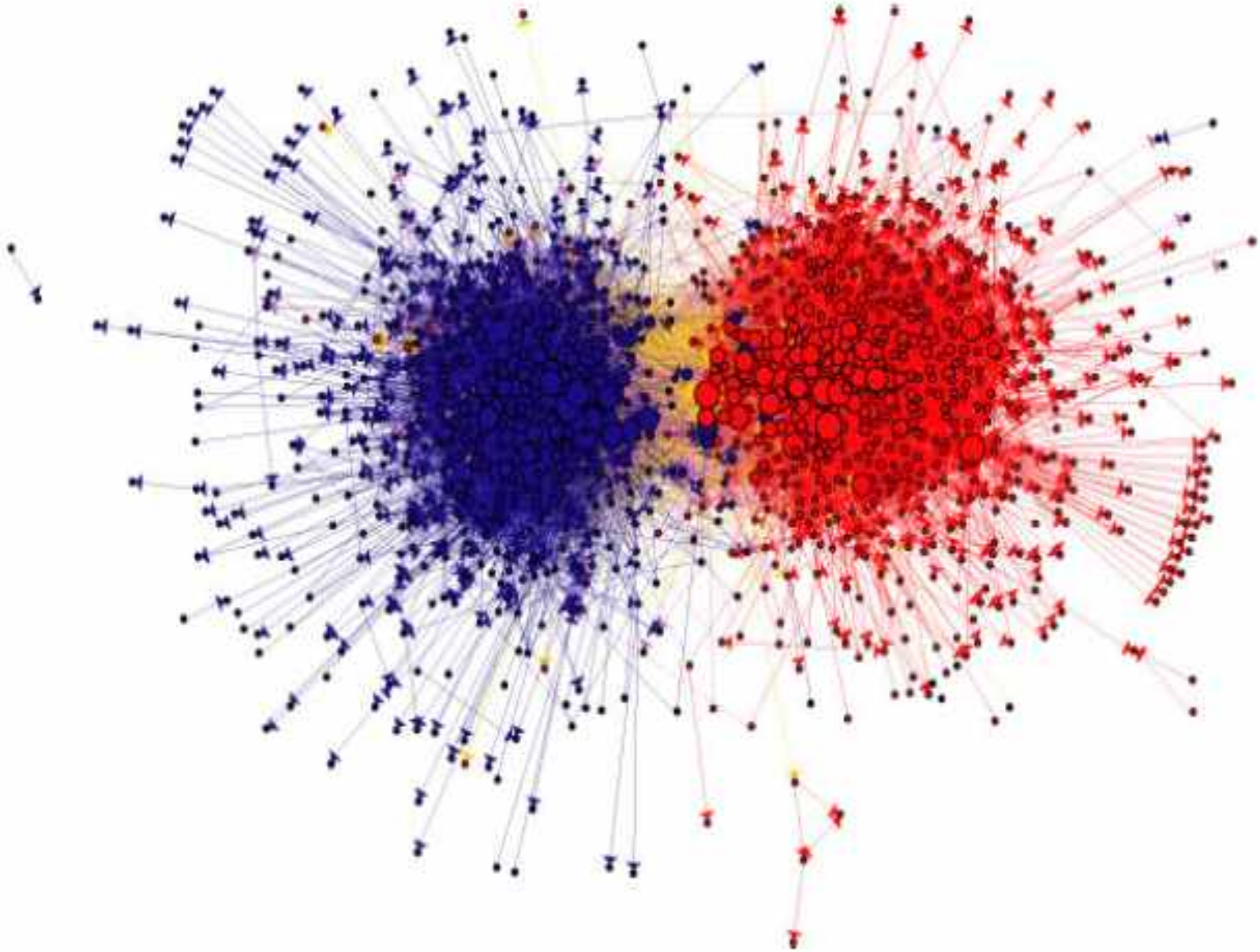
Obecnie (Internent, social networks)

- Tysiące, setki tysięcy, miliony





Name	Weighted	Directed	Vertices	Edges
calls	✓	✓	7,786,471	33,292,508
condmat-collab	✓		17,216	110,544
dblp-cite	✓	✓	15,963	344,373
dblp-collab	✓		367,725	2,088,710
disease-g	✓		399	15,634
disease-p	✓		437	81,158
hepth-cite	✓	✓	8,249	335,028
hepth-collab	✓		8,381	40,736
huddle	✓		4,243	997,008
patents-cite	✓	✓	1,461,714	32,418,457
patents-collab	✓		1,162,227	5,448,168
sms	✓	✓	5,016,746	11,598,843

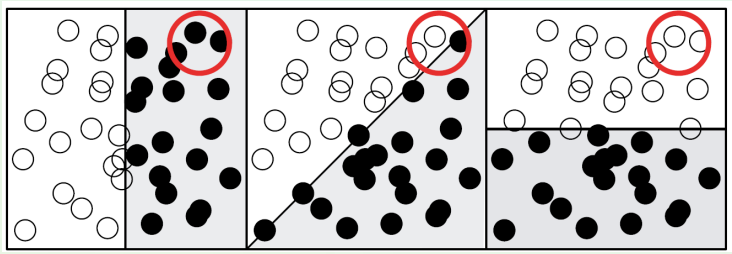


source: Adamic and Glance, The political blogosphere and the 2004 US election: divided they blog, Proceedings of LinkKDD, Chicago, IL, p.36-43, 2005.

Zmienne strumienie danych (data streams)

- Źródła – generują dane w postaci nieprzerwanego strumienia
  - Trudności z przechowywaniem wszystkich elementów danych
  - Konieczność przetwarzania szybkiego strumienia w ograniczonym czasie
  - Przyrostowe działanie algorytmów
- W wielu zastosowaniach zmienność danych (także definicji klas) wraz z upływem czasu
  - Tzw. concept drift
  - Nowe metody oceny klasyfikatorów przyrostowych

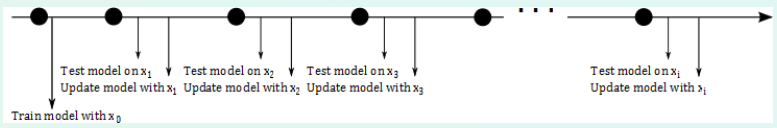
	Statyczne	Strumień
No. odczytów	wielokrotne	pojedyncze
Czas	nie jest krytyczny	ograniczony
Uzycie PAO	b. elastyczne	ograniczone
Rezultaty	dokładne	przybliżone
Przetwarz. Rozproszne	zwykle nie	tak

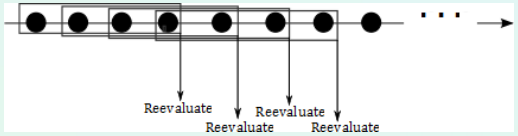


Statyczne algorytmy eksploracji danych nie mogą być stosowane!

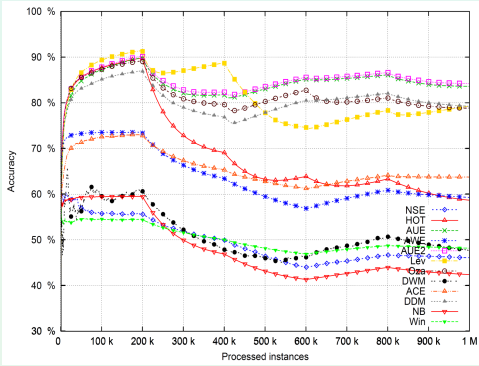
Klasyfikacja zmiennych strumieni

- Zarządzanie pamięcią i “zapominanie”
  → sliding? windows

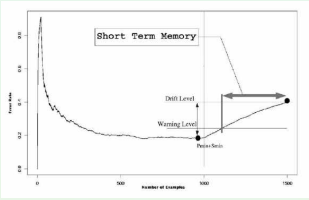




- Inne spojrzenie na ocenę
  - Miary oceny
  - Interleaved test-then-train
    or prequential techniques



- Pełny lub częściowy dostęp do prawdziwych etykiet
  - Częściowo-nadzorowane podejścia (micro-clustering)
  - Strategie aktywnego uczenia się
- Wykrywanie zmian→ drift detection methods
  - Trudniej w nieetykietowanych danych



Spójrz na prace przeglądowe

- J.Gama: Knowledge discovery from data streams. CRC, (2010)
- J.Gama, I.Zliobaite, A.Bifet, M.Pechenizkiy, A.Bouchachia: A survey on concept drift adaptation. ACM Comp. Surveys (2013)
- G.Ditzler, M.Roveri, C.Alippi, R.Polikar: Learning in nonstationary environments – a survey IEEE Comput. Intel. (2015)

Specjalizowane zespoły klasyfikatorów (ensembles)

Block-based ones:

Streaming Ensemble Algorithm (SEA) – Street & Kim 2001

Accuracy Weighted Ensemble (AWE) Wang et al 2003

- BWE - Deckert 2011, Weighted Aging Ensemble (Wozniak et al 2013)

Learn++.NSE – Polikar et al. 2011

Others , e.g., EAE (Jackowski 2014)

Recurring concepts

- CCP – Katakis et al. 2010
- RCD – Goncalvas et al. 2013
- FAE – Diaz et al 2015

Block processing also in some semi-supervised or novel class detection – Masud et al. 2009; Farid et al 2013

On-line (instance based)

WinNow, Weighted Majority Alg. – Littlestone 1988, L & Warmuth 1994

Dynamic Weighted Majority (DWM) – Kolter & Maloof 2003 → AddExp (2005)

On-line bagging and on-line boosting [Oza]

- BagADWIN,
- Leverage bagging – Bifet et al. 2007
- Using in DDD (Minku, Yao)

Hoeffding Option Trees (HOT)

UFFT (Gama at al. 2005)

ADACC – Jaber 2013

Boosting classifiers for drifting concepts – Scholtz & Klinkenberg 2007 + more

Hybrid approaches

ACE – Nishida 2009, OBWE ,

AUE → OAUE [Brzeziński, Stefanowski]

Rozwiązania obliczeniowe

- Przetwarzanie rozproszone i równoległe
  - Hadoop i MapReduce
    → Spec. software Apache (Pig, Hive,Hbase)
  - Spark
- NoSQL bazy danych (Google, Facebook, Amazon,…)
  - Google BigTable
  - Dynamo, Casandra, MongoDB,…
- Nowe środowiska programowe
  - Mahout (scalable machine learning)
  - Spark MLBase / Mlib
  - Vowpal Wabbit
  - h2o
  - MOA → SAMOA
  - …

Decyzje o zrównoleglaniu obliczeń

Nie zawsze konieczne:

Cytat za Yann LeCun (KDnuggets News 14:n05)

"I have seen people insisting on using Hadoop for datasets that could easily fit on a flash drive and could easily be processed on a laptop."

The reality is that most companies or researchers don't need that much data … (Blog)

Decyzja zależna od kontekstu zastosowania

Lecz także od algorytmu

- Many easy tasks on one computer become difficult in a distributed environment
- For example, subsampling is easy on one machine, but may not be in a distributed system [C.Lin 2014]

Naturalne przyspieszenie obliczeń Random Forests / MapReduce

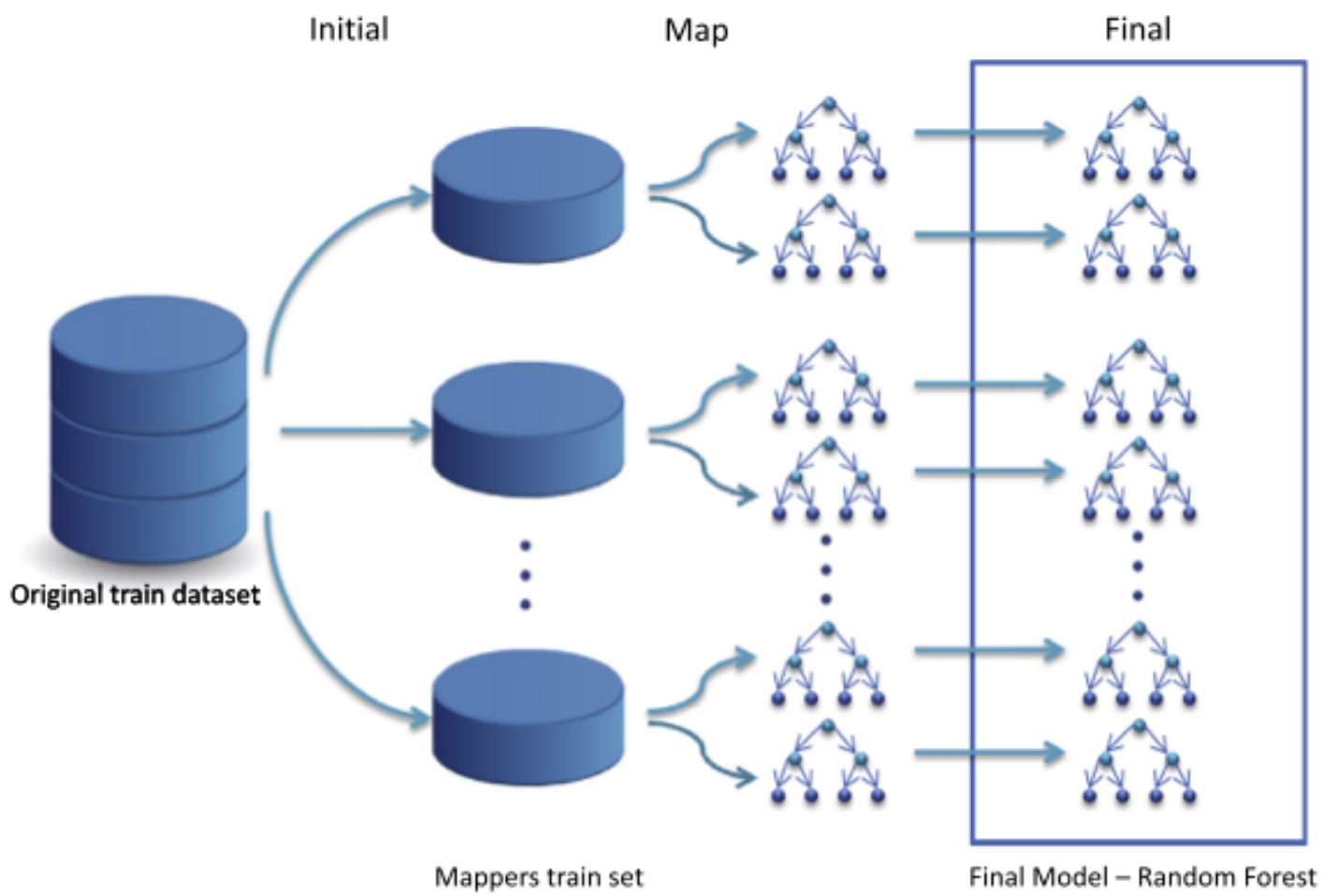
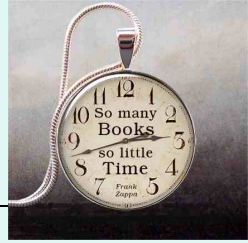


Fig. 3. A flowchart of how the building of the Random Forest is organized in Mahout.

Za S.Rio in. On the use of MapReduce for imbalanced big data using Random Forest 2014

Inne aspekty algorytmów Big Data



Poza przyspieszeniem i zrównoleglaniem działania
(np. jak speed up Deep Learning Networks?)

- Konstruowanie wersji przyrostowych dla rozwiązań statycznych
- Inne metody losowania
  - Bag of little bootstraps [Kleiner in 2012]
- Nowe podejście do wstępnego przetwarzanie danych,
  - np., outliers detection; nowe metody dekompozycji macierzy [M.Jordan i in. 2012]
- Sparse, incomplete, linked data repositories
- Uczenie się z niezbalansowanych danych
- Uczenie sie częściowo nadzorowane
  - Active learning, … [Krempl 2015]
- Rozwój systemów rekomendacyjnych
- Intensywniejsze wykorzystywanie knowledge bases
- Privacy data mining [Matwin]
- …

Big Data z perspektywy **statystycznych systemów uczących się**

- Metody statystyczne stanowią podzbiór ogromnej wagi i niezbywalny metod uczenia maszynowego

- Jakkolwiek statystyczna meta-analiza i np. metody probabilistycznego łączenia rekordów z różnych baz danych są stale rozwijane, to w obszarze Big Data rozwój metod uczenia statystycznego dotyczy głównie analizy danych masywnych z jednego źródła

- Niezbywalność metod uczenia statystycznego wynika z mocy wyjaśnień/rozwiązań, których dostarczają, oraz ich metodologicznej przejrzystości i czystości

Big Data z perspektywy statystycznych systemów uczących się, c.d.

- Statystyk dobrze wie, że potrzebuje powtarzalnych danych i że dane, którymi dysponuje, mogą odpowiedzieć tylko na pewne, dobrze określone pytania – narzuca to specyficzny reżim analizowania tych danych

- Metody statystycznego uczenia dostarczają modeli przyczynowo-skutkowych, gdy to osiągalne (możliwe), i pozwalają zadowolić się podejściem algorytmicznym/predykcyjnym/behawiorystycznym, gdy głębsze poznanie jest nieosiągalne

Big Data z perspektywy statystycznych systemów uczących się, c.d.

- Paradoksalnie, to rozwój technik komputerowych pozwolił statystykom wyjść z Tukeyowskiego więzienia eksploracyjnej analizy danych (EDA), nad bramą którego widniał napis: „Pozwólmy przemówić danym – niech same mówią za siebie”

- W 1979 roku nie tak słynny jak John Tukey, za to bardziej radykalny i też znakomity statystyk, William Eddy napisał:

”The data analytic method denies the existence of "truth"; the only knowledge is empirical.

[...] If we can make without models, I think we should.”

Big Data z perspektywy statystycznych systemów uczących się, c.d.

- Dzisiaj, w powodzi Big Data, niektórzy badacze powiadają właściwie to samo: mając Big Data nie musimy już wyjaśniać, wystarczy znać korelacje, które pozwalają przewidywać

  por. odnośne uwagi w: N.Japkowicz and J.Stefanowski (Eds.), Big Data Analysis: New Algorithms for a New Society, Springer (2016), rozdz. 1 i 2

- Każdy pretekst jest dobry, by opowiedzieć się za głupotą, ale lepiej chcieć pozostać rozumnym i starać się rozumieć (a nie tylko przewidywać)

Big Data z perspektywy statystycznych systemów uczących się, c.d.

Gdy mówimy o metodach statystycznego uczenia w kontekście danych masywnych warto najpierw rozróżnić dwie sytuacje:

- wielkiej liczby obserwacji (*large n*) i małej liczby zmiennych (cech) objaśniających (opisujących każdą obserwację)
- wielkiej liczby cech opisujących każdą obserwację (*large p*) i niekiedy jednocześnie małej liczby obserwacji (*small n large p*)

UWAGA: Od tego slajdu aż do końca omawiania metod statystycznego uczenia ograniczamy się do problemów uczenia pod nadzorem

Big Data z perspektywy statystycznych systemów uczących się, c.d.

W przypadku wielkiej liczby obserwacji (*large n*) i małej liczby zmiennych (cech) objaśniających niezbędne może być podzielenie zadania na mniejsze podzadania, rozwiązanie podzadań i na tej podstawie zbudowanie rozwiązania dla całego zadania

por. odnośne uwagi w: N.Japkowicz and J.Stefanowski (Eds.), Big Data Analysis: New Algorithms for a New Society, Springer (2016), podrozdz. 2.4

Big Data z perspektywy statystycznych systemów uczących się, c.d.

W przypadku zadań o wielkiej liczbie cech i zwłaszcza problemów znanych jako *small n large p problems* można ich omówienie podzielić na trzy części:

- metody oparte na podejściu Monte Carlo
- metody z regularyzacją (karą za złożoność modelu)
- metody bayesowskie

UWAGA: Podział ten jest w pewnym stopniu umowny, ponieważ wymienione części nie są rozłączne – np. kara za złożoność modelu może mieć bayesowski charakter (jak bayesowskie kryterium informacyjne BIC) i mamy wówczas do czynienia z bayesowską regularyzacją. Co więcej, metody różnych typów mają często podobny fundament matematyczny

Big Data z perspektywy statystycznych systemów uczących się, c.d.

Metody oparte na podejściu Monte Carlo, o których tu wspomnimy, to:

- metoda wyboru modelu regresji liniowej (i logistycznej) przez analizę modeli budowanych na losowo wybranych podzbiorach cech objaśniających; p. Mielniczuk i Teisseyre (2011) i (2013)
- metoda selekcji cech objaśniających istotnych dla zadanego problemu klasyfikacji oraz odkrywania współzależności między tymi cechami;
  p. Dramiński i in. (2008) i (2016)

Big Data z perspektywy statystycznych systemów uczących się, c.d.

Metody z regularyzacją (karą za złożoność modelu), to przede wszystkim:

- metoda LASSO i jej rozwinięcia; p. Hastie, Tibshirani i Wainwright (2015); Bühlmann i van de Geer (2011); Pokarowski i Mielniczuk (2015), gdzie LASSO służy do przesiania cech (potem poddanych rankingowi; ostatnim etapem jest wybór modelu z ich zagnieżdżonej rodziny); Bogdan i in. (2015)
- maszyny wektorów podpierających i metody pokrewne; p. np. Tan i in. (2014), gdzie metoda wyboru cech dopuszcza ich ogromną liczbę; Priyadarshini i Agarwal (2015), gdzie mamy propozycję implementacji SVM na platformie Hadoop MapReduce

Big Data z perspektywy statystycznych systemów uczących się, c.d.

Metody bayesowskie mają bardzo długą historię, ale w kontekście nas interesującym należy wspomnieć przynajmniej o:

- Nurcie wyrastającym z analizy regresji liniowej z bayesowską regularyzacją; p. Browman i Speed (2002) i rozwinięcia tej idei (Bogdan i in.; Frommlet i in.; Chen i Chen)
- Niezwykle ciekawym nieparametrycznym podejściu bayesowskim; p., np., Rasmussen i Williams (2006), gdzie podstawą są procesy gaussowskie, a zatem macierze kowariancji najczęściej o wyrazach będących jądrami gaussowskimi, oraz Yang (2014), gdzie rozwiązuje się zadanie minimaksowe dla ryzyka w L_2; też Tokdar (2011)

Big Data z perspektywy statystycznych systemów uczących się, c.d.

UWAGA: Nieco dłuższe omówienie zasygnalizowanych trzech typów metod uczenia statystycznego oraz szerszy wykaz literatury można zaleźć w oddzielnej prezentacji

Don’t ..., it is not over!

Kilka uwag, jeśli starczy Nam czasu, …

Trochę o mitach Big Data Analysis → odkrywanie prostych współzależności o silniejsze modele

Nie podzielamy tezy Anderson C. (The end of the Theory / Wired Mag. 2008) **„Korelacja jest wystarczająca, nie potrzebujemy modelu przyczynowo-skutkowego"**

- Przykłady negatywne → Google Flue Trends (początkowy sukces – analiza zapytań / szybsza predykcja; zawiodła w 2012, konieczność przebudowy i dostrojenia modelu)

"When to act on a correlation, and when not to", Ritter 2014

- Gra dwóch czynników (1) przekonania o powtarzalności zależności (2) przetargu pomiędzy kosztem braku działania a potencjalnym zyskiem
- "Confidence in a correlation is a function of not only the statistical frequency but also the explanations of what is causing that correlation"

Prace metodologiczne o powiązaniach teorii i danych

"The model shapes the data in trying to best fit it, and the data shapes the model in that it helps us to use models with the appropriate assumptions" [R. Sparks, H.Lenz] A Statistician's insight on Big Data Analysis, 2016

Spójrz także: **Wolfgang Pietsch**: „Big Data? The new science of complexity [2014]

Czy dostęp do wystarczających wielkości?

Mit Big Data – „N = całość / N = All”

- Tylko w nielicznych zastosowaniach dostęp do pełnej zbiorowości
- Nawet popularne media społecznościowe obejmują tylko ograniczoną część ludności
  - Ograniczenia danych Twitter (Tufekci)
- Błędy w zastosowaniu projektu „Boston Street Bump” smartphones (Tim Harford)

„N = całość” - często pobożne życzenie lub iluzja, a nie rzeczywistość!

Więcej dyskusji w:

Boyd, D., Crawford, K.: Critical questions for Big Data. Information, Communication and Society (2012).

Boyd, D., Crawford, K.: Six provocations for Big Data Symposium on the Dynamics of the Internet and Society. Oxford Internet Institute, Sept. 21 (2011).

Harford, T.: Big Data: are we making a big mistake? Financial Times, March 28, 2014.

Tufekci, Z.: Big Data: Pitfalls, methods and concepts for an emergent field. SSRN (March 2013)

Lekcje ze statystyki nie mogą być zapomniane

Mity „More data is better data” lub „Better knowledge can be extracted from large data sets than small ones”

→ nie zawsze poprawne!

Doświadczenia statystyczne nt. reprezentatywności i obciążenia prób

- Niepoprawnie pozyskane próby prowadzą do błędnych predykcji
- Im więcej danych, tym zwiększona szansa na pozorne zależności lub wzorce [Michael Jordan]
  - Spurious Correlations – pozorne zależności
- Nakładające się błędy pomiarów, wielokrotne testowanie, ...

Pomijanie dodatkowych źródeł wiedzy i powierzchowność analiz na podstawie zbyt prostych danych

Analiza stopnia zażyłości osób na podstawie połączeń telefonicznych (Danah Boyd - przykład i krytyka)

Czy Big Data to wielka zmiana?

"Big Data, which started as a technological innovation in distributed computing, **is now a cultural movement by which we continue to discover how humanity interacts with the world** — and each other — at large-scale".

Drew Conway

"Big data" represents **a cultural shift in which more and more decisions are made by algorithms** with transparent logic, operating on documented immutable evidence. I think "Big" refers more to the pervasive **nature of this change** than to any particular amount of data.

Daniel Gillick (Google Senior Resercher)

Analiza Big Data prowadzi do rewolucyjnych I fundamentalnych zmian, gdyż (…):

Viktor Mayer-Schönberger, Kenneth Cukier

Porozmawiajmy o aspektach społecznych

Prywatność i anonimowość = nowe wyzwania

Ludzie są źródłem danych

- *Nie przestrzegają zasad ochrony*
- *Świadomie kuszeni usługami*

Lecz świat Big Data jest bardziej skomplikowany i pełen pułapek

Klasyczne metody anonimizacji są nieskuteczne

- [Barabasi et al. 2008] – ukryta tożsamość zidentyfikowana na podstawie wzorców poruszania się osoby z danych lokalizacyjnych urządzeń mobilnych
- Odkrycie nazwiska osób na podstawie analizy zapytań w logach AOL połączonych z extra danymi [Barbaro, Zeller 2006]
- Poszukanie współzależności rankingów anonimowych użytkowników Netflix z dodatkowymi recenzjami [2008]

Maszyny / algorytmy podejmują decyzję

- Wyniki algorytmów – systemy wspomagania decyzji w odniesieniu do ludzi
  - Decyzje kredytowe, ubezpieczenia, ofert pracy, …
- Czy wiemy dlaczego?
  - Większość algorytmów - „black box” i słaba zdolność wyjaśnienie okoliczności podjęcia decyzji

- W niektórych dziedzinach musimy otrzymać objaśnienie procesu podejmowania decyzji
- Niebezpieczeństwo „false positive” pomyłek
- Czy i kogo zaskarżyć za błędną decyzję?
  - Czy szczegóły stosowanego algorytmu są tajemnicą firmy?

Transparentność analizy Big Data

Zagrożenia stosowania algorytmów wobec ludzi

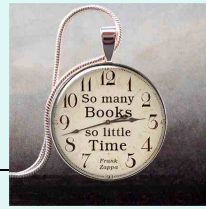
- Wiemy co jutro będziesz robić
  - Profilowanie osób i działania wobec osób przed ich akcją

- Ponadto problem dostępu danych
  - Kto wykorzystuje i do czego moje dane?
  - Kto jest właścicielem danych?
  - Powtórne wykorzystywanie danych.

- Przejrzystość wykorzystywania danych i stosowanych algorytmów
  - Badacze vs firmy komercyjne
  - Lecz kto ma to kontrolować?

"Big Data ethics is huge, messy and personal topic, which cannot be easily settled" [Kord Davis]

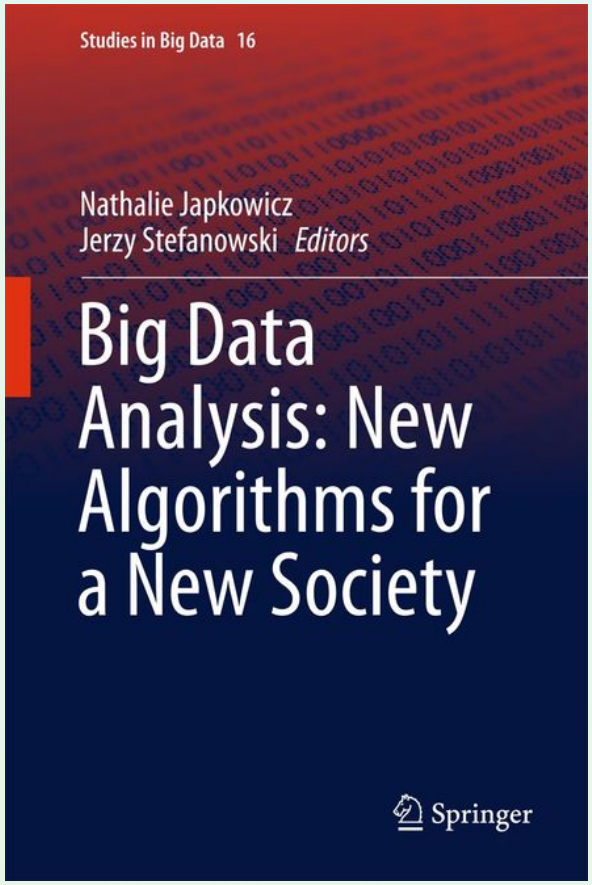
Podsumowując



Więcej w rozdziałach książki

Nathalie Japkowicz, Jerzy Stefanowski: **Big Data Analysis: New Algorithms for a New Society**, Springer 2016:

- A Machine Learning Perspective on Big Data Analysis
- Final Remarks on Big Data Analysis and Its Impact on Society and Science



Dziękujemy za uwagę

Pytania lub komentarze?

Kontakt:
Jacek.Koronacki@ipipan.waw.pl
Jerzy.Stefanowski@cs.put.poznan.pl