Projekt eksploracji danych

Wykład 1

Jerzy Stefanowski

Instytut Informatyki,
Politechnika Poznańska

Poznań, 2 marca 2020

Skąd się wzieło …?

- Dawno temu, w ...
- Historycznie (..) wymyślił kameralny przedmiot seminaryjny, cyt.
  - Aim: To learn how to deal with real-life data sets (or to win a data mining competition).
  - Scope: Application of data mining algorithms to real-life massive data sets

Jak to było?

Konkurs na rzeczywistych (początkowo symulowanych) danych -> może widzieliście plakaty, np. Sito, DataNinjia,...

oraz mikro-prezentacje studenckie podczas zajęć + tzw. open project's meetings

Lectures

- Invited lecture(s)
- Students' presentations about their solutions in the
- Students' talks on selected topics (time slots: Nov.
- Lectures on demand (video lectures, lecture on a g

Schedule of lectures

11-10-2018	The mining massive data sets challenge [pd
07-11-2018	Launching of the "Data Ninja" competition
14-11-2018	Students' talk: Natural language processing
	Students' talk: Deep networks: TensorFlow
	Student's talk: Building Autoencoders with
21-11-2018	Students' talk: Deep networks: Torch (Jan A
	Students' talk: Deep Networks: Theano (Jar
12-12-2018	Students' talk: Deep reinforcement learning
	Students' talk: Positive-unlabelled learning
	Students' talk: Active learning (Damian Sza
19-12-2018	Students' talk: OpenAi Gym (Paweł Szudro
	Students' talk: Microsoft Azure for Big Dat
	Students' talk: Zero-shot learning (Dawid C
09-01-2019	Students' talk: ML Packages: MLib (Amade
	Students' talk: Multi-label classification (Pa
	Students' talk: Deep networks: TensorFlow
16-01-2019	Presentation of students' reports

The Challenge

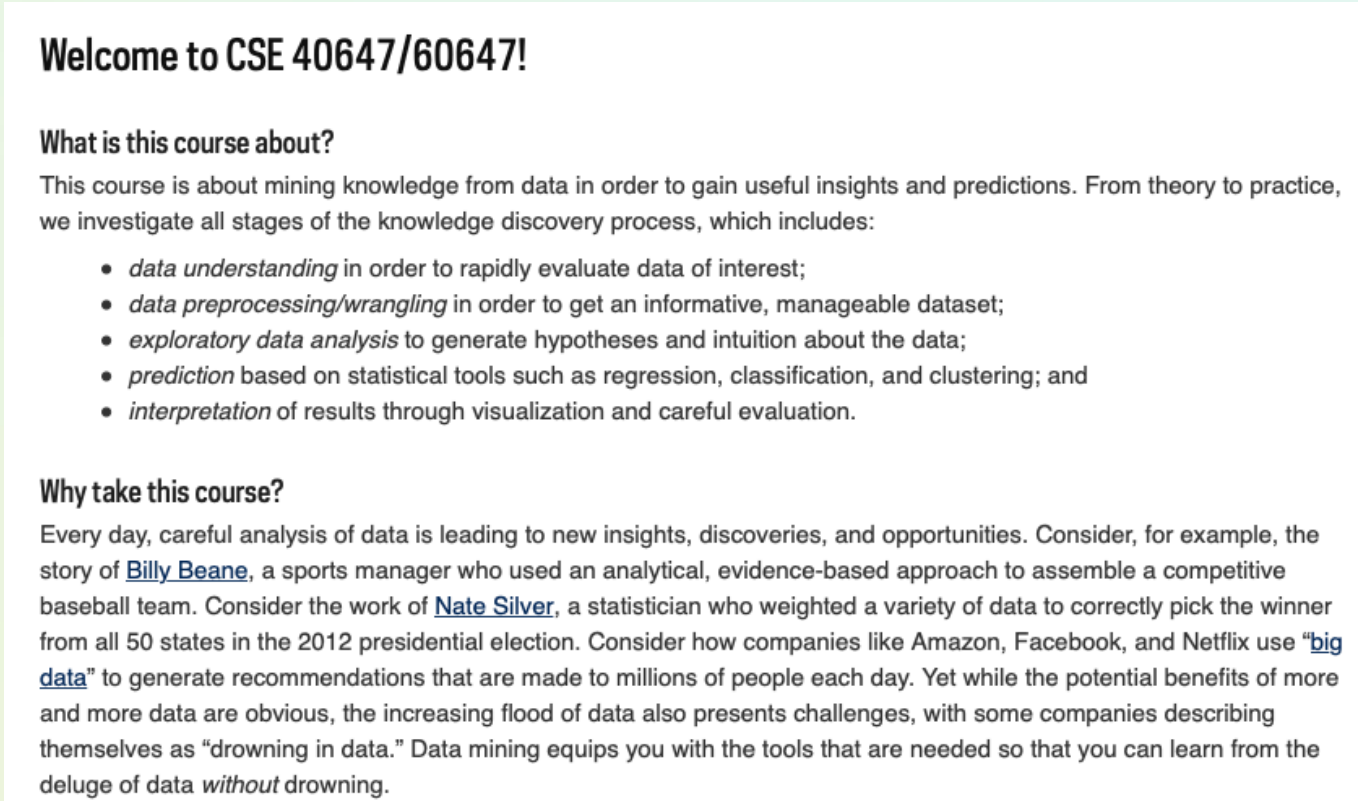
Kilka zagranicznych inspiracji, ...

Course CS6220 /UCLA/: Data mining techniques: project

> “In this project, you will have an opportunity to apply the data mining algorithms and techniques you learned in the class to some real-world problems. You can choose any problem that you are interested in, and formalize it into a data mining task. Then you get some data related to the task and apply some data mining algorithms to your data. Also you need to evaluate and compare your algorithms. And finally, you should submit a report, together with your data and code.”

Univ. Notre Dame
Data mining project

Sprawdź
www.coursera.org
Data Mining Project from
University of Illinois
at Urbana-Champaign



Welcome to CSE 40647/60647!

What is this course about?

This course is about mining knowledge from data in order to gain useful insights and predictions. From theory to practice, we investigate all stages of the knowledge discovery process, which includes:

- *data understanding* in order to rapidly evaluate data of interest;
- *data preprocessing/wrangling* in order to get an informative, manageable dataset;
- *exploratory data analysis* to generate hypotheses and intuition about the data;
- *prediction* based on statistical tools such as regression, classification, and clustering; and
- *interpretation* of results through visualization and careful evaluation.

Why take this course?

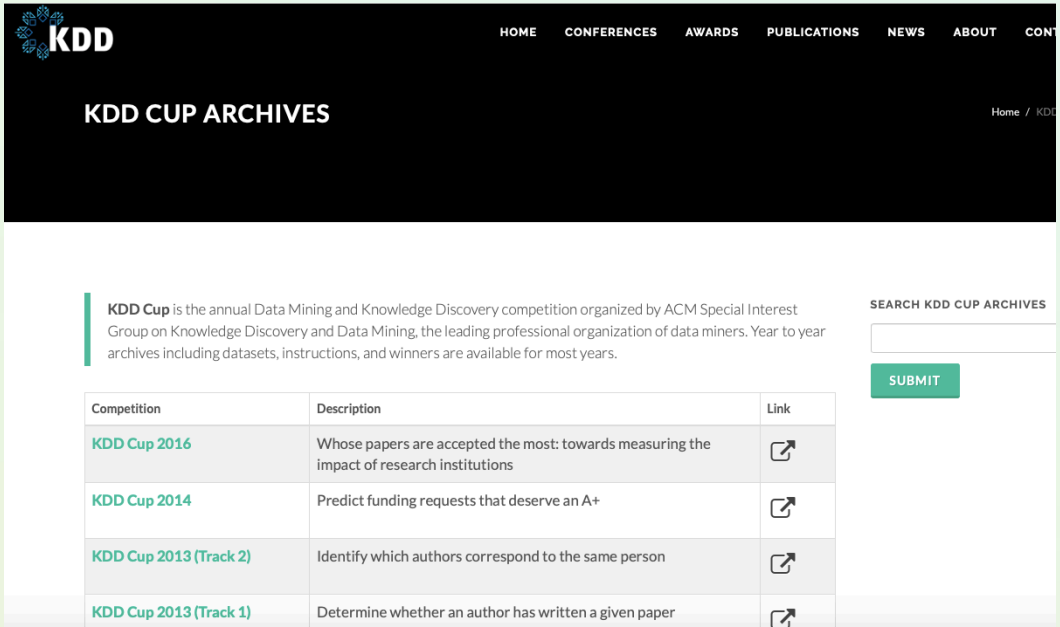
Every day, careful analysis of data is leading to new insights, discoveries, and opportunities. Consider, for example, the story of Billy Beane, a sports manager who used an analytical, evidence-based approach to assemble a competitive baseball team. Consider the work of Nate Silver, a statistician who weighted a variety of data to correctly pick the winner from all 50 states in the 2012 presidential election. Consider how companies like Amazon, Facebook, and Netflix use “big data” to generate recommendations that are made to millions of people each day. Yet while the potential benefits of more and more data are obvious, the increasing flood of data also presents challenges, with some companies describing themselves as “drowning in data.” Data mining equips you with the tools that are needed so that you can learn from the deluge of data *without* drowning.

Poszukiwanie rzeczywistych danych

Tradycja konkursów przy prestiżowych wydarzeniach

- KDD-Cup lub inne konferencje
  - https://www.kdd.org/kdd-cup
  - https://www.kdd.org/kdd2019/kdd-cup
- Int. Data mining Cup (students)
- Netflix challenge
- Kaggle-based challenge
- Inne, ...

Companies have used data science competition as a strategy to bring cultural change or even crowd source their problems to external teams. Netflix in our recent past was one example, where they pioneered this practice by crowdsourcing their recommendation algorithm. Further, data science competition companies, such as Kaggle, Hexagon-ML and others, host competitions either sponsored by companies on their platform or hosted in the companies itself.



W tym roku

Zgodnie z programem:

Ćwiczenia projektowe (p. **mgr K.Miazga**) + wykład (JS i ?)

Wykorzystujemy rzeczywisty problem (ang. YouTube trending)

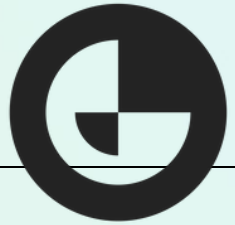
Rzeczywiste dane (lecz b. surowe);

Intensywne wykorzystanie wielu metod z KDD (czyli procesu odkrywania wiedzy):

- Integracja danych i tworzenie wstępnej reprezentacji
- Transformacje, oczyszczenie, wstępne przetwarzanie
- Inżynieria cech (ocena znaczenia, badanie współzależności, redukcja)
- Problem częściowo-nadzorowany (brak etykiet)
- Modele predykcyjne
- Interpretacja / charakterystyka opisu klasy

Dodatkowo analiza szeregów czasowych

Problem dodatkowy – konkurs



OLX GROUP

Data Ninja Otodom challenge

Cel - model (maks. 2 osoby) prognozujący ceny mieszkań i domów na polskim rynku nieruchomości w oparciu o dane pochodzące z serwisu Otodom

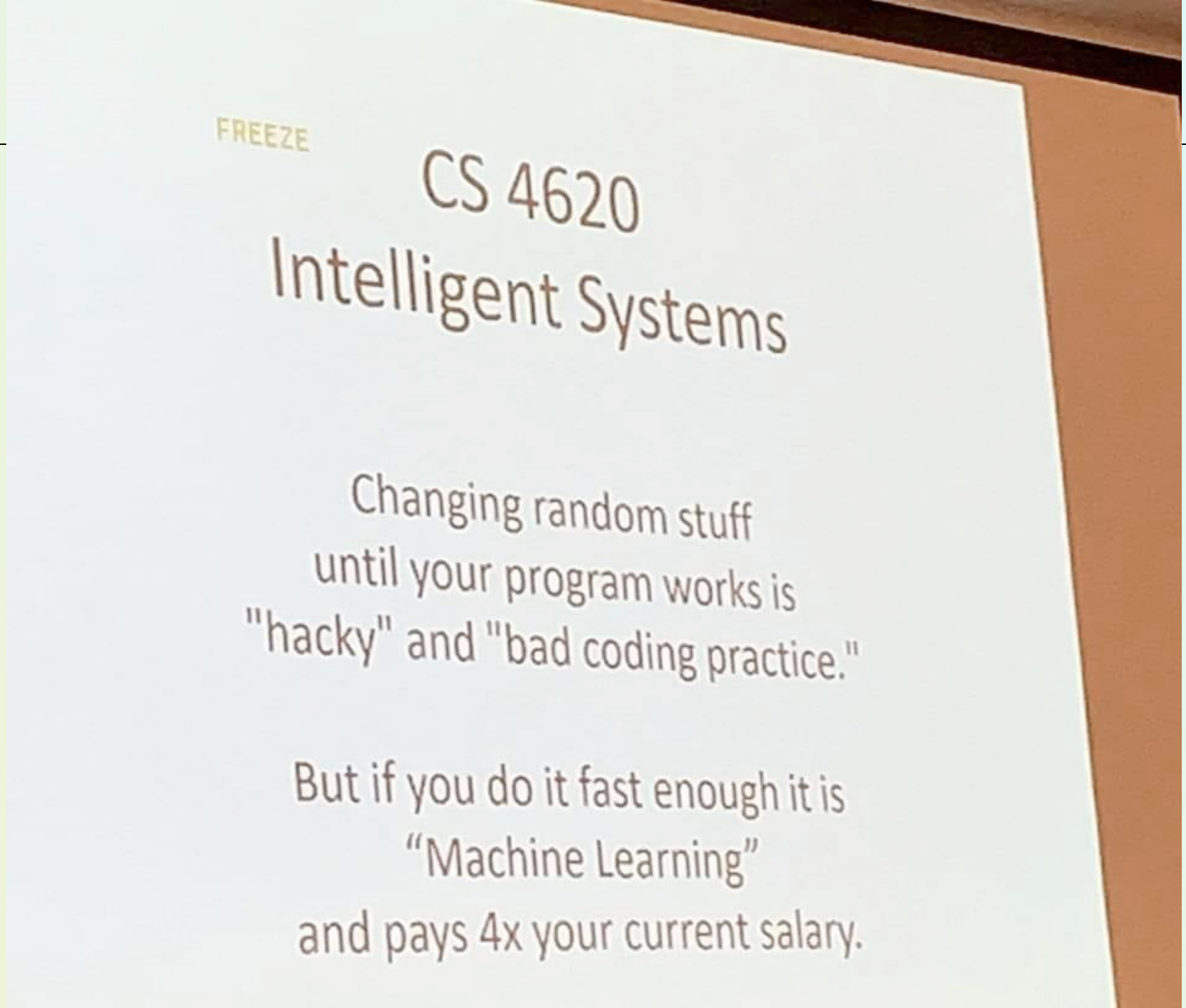
Dane historyczne (opisy, zdjęcia, dane kategoryczne i geolokalizacyjne) na podstawie których stworzyć model, który przewidzi ceny dla nowo dodanych ogłoszeń.

Kryteria - dokładność predykcji modeli - błąd średniokwadratowy (RMSE) oraz jakość rozwiązań dostarczonych w postaci Jupyter Notebook albo R Markdown - proces analizy i transformacji danych, uczenia modelu oraz wytłumaczenia jego predykcji zarówno dla całego zbioru jak i pojedynczych punktów w danych.

Rozwiązanie te zostaną ocenione przez ekspertów z serwisu Grupy OLX /
3 miejsca na podium (nagrody) + staże

Powiązanie z pracą na laboratoriach i zasadami oceny (p. mgr Miazga)

otodom.


FREEZE
CS 4620
Intelligent Systems
Changing random stuff
until your program works is
"hacky" and "bad coding practice."
But if you do it fast enough it is
“Machine Learning”
and pays 4x your current salary.

Mapa pojęć

Sztuczna inteligencja

Uczenie maszynowe

Data Mining
Eksploracja danych

Big data mining

Odkrywanie wiedzy z danych
Process KDD

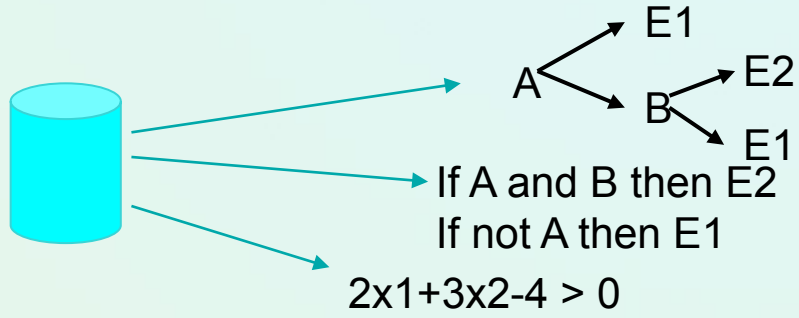
Data science

“You may not know it, but AI is all around you” P.Domingos

Machine learning vs. data mining

- Uczenie maszynowe → (ang. Machine Learning)
- Systemy, które doskonalą swoje działanie na podstawie analizy przykładów działania (nie tylko dane, lecz inne ...)
- *To give computers the capability to learn without being explicitly programmed (def. of Arthur Samuel)*

DARPA LAGR Herminator



- Eksploracja danych → Data mining
  - Poszukiwanie w zgromadzonych danych nieznanych, użytecznych regularności, związków między elementami danych
  - Nacisk na same dane i wyciągnie użytecznych wniosków dla zastosowania!
  - Kluczowy etap w procesie odkrywania wiedzy z danych
  - Obejmuje więcej paradygmatów metod niż tylko ML

Inne rozróżnienia Ml vs. DM

J.Ramon KULeuven:

Both terms overlap, but the emphasis is different.

E.g. pattern mining & association rule mining (description) is usually more associated with data mining than machine learning, while supervised learning (prediction) is more actively researched in the domain of machine learning.

In DM we try to get insight in the data without necessarily getting better at a task (e.g. we detect a community without optimizing for being able to predict community membership).

In theory, machine learning is also possible without data, e.g. in reinforcement learning (where the machine collects the data itself) or self-play in learning to play a game.

Także

Machine Learning is a branch of A.I. with the following issues:
(i) mimic the human or animal ability to learn, and
(ii) matching or exceeding human or animal learning by generalization/ abstraction.

Data mining is referred to business analytics / perspektywa zastosowań

Różnice Ml vs. DM

Za Bernard Marr'em

- DM is looking for patterns that already exist in the data, ML goes beyond what's happened in the past to predict future outcomes.
- In data mining, the 'rules' or patterns are unknown at the start of the process – more exploration in a larger set of models.
- ML the machine is usually given some rules or variables to understand the data and learn / background knowledge / concepts are stronger pre-identified.
- Human interaction - Data mining is a more manual process that relies on human intervention and decision making.
- In ML, the process of 'learning' and refining is automatic. The machine becomes more intelligent by itself.
- Data mining is used on an existing earlier dataset (like a data warehouse) to find patterns /Original data usually raw and dirty/.
- Machine learning, on the other hand, is trained on a 'training' data set, which teaches the computer how to make sense of data, and dataare better prepared (supervised learning)

Knowledge Discovery Process

Knowledge Discovery in Data is the

non-trivial process of identifying

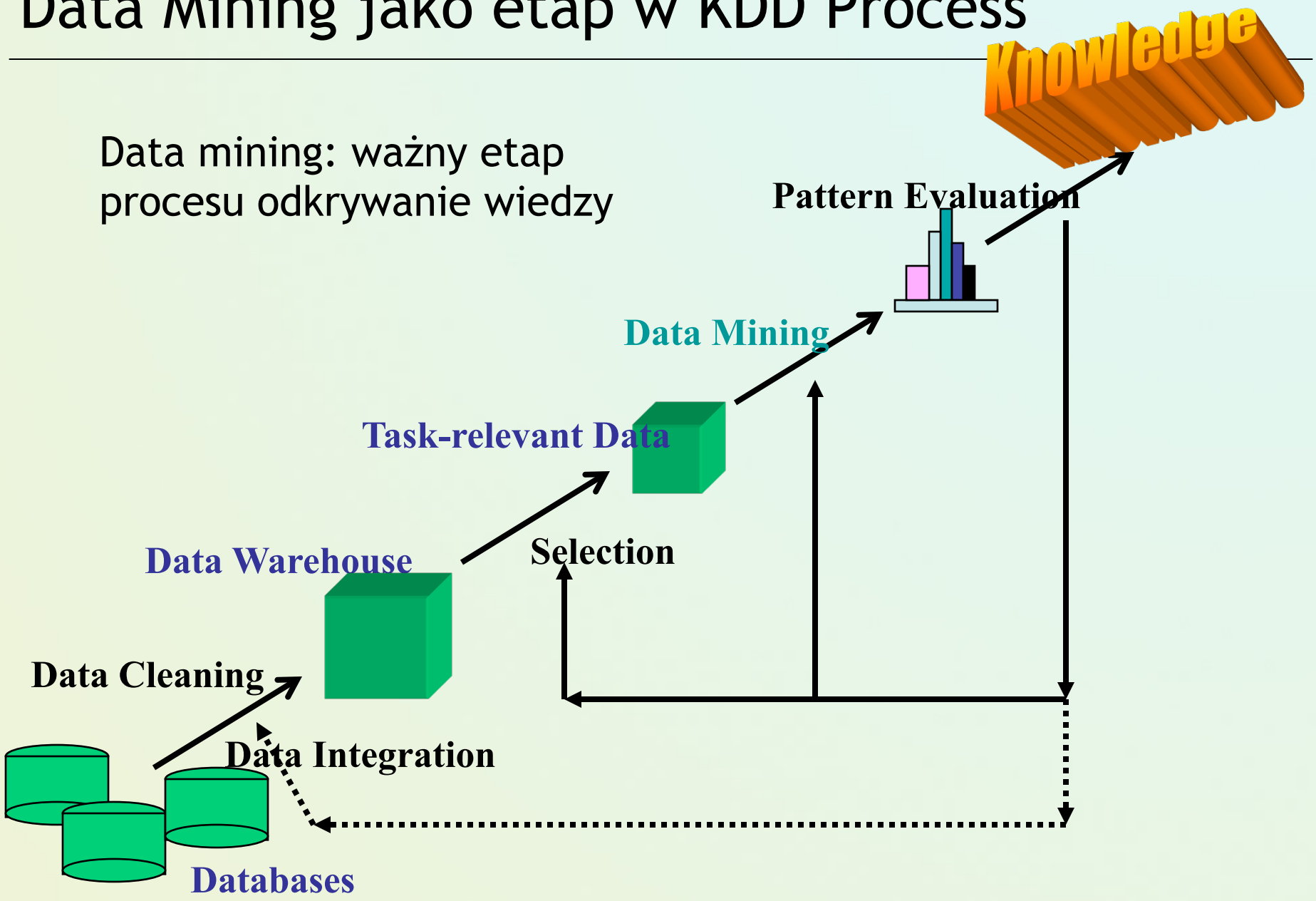
- *valid*
- *novel*
- potentially *useful*
- and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining*, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996.

The name first used by AI, Machine Learning Community in 1989 Workshop at AAAI Conference.

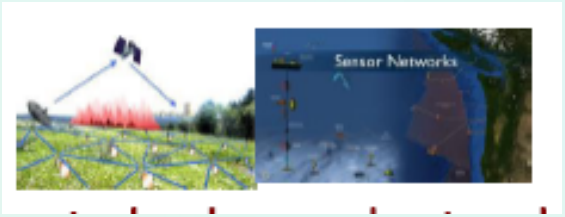
Data Mining jako etap w KDD Process

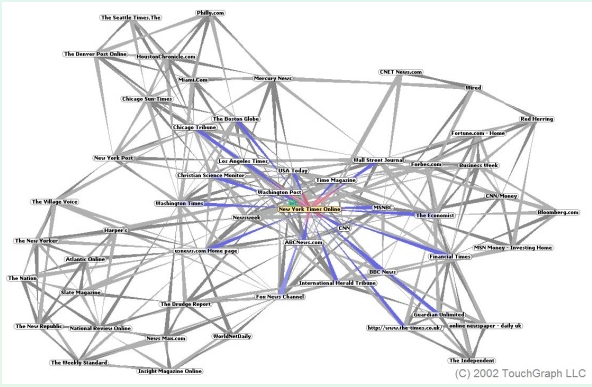
Data mining: ważny etap procesu odkrywanie wiedzy



Nowe źródła danych

Large Survey Telescope





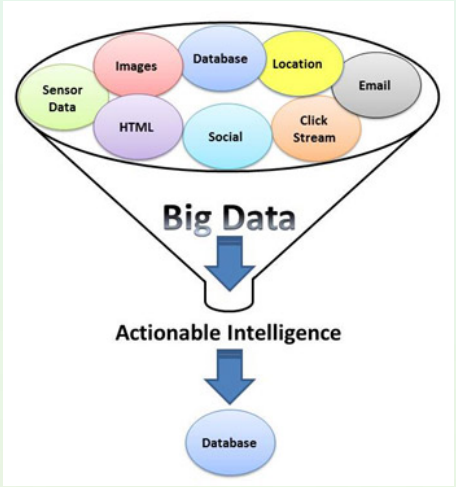
Rozwój technologiczny

Wzrost ilości danych generowanych przez nowe źródła pomiarowe

Rosnąca wielkość „tradycyjnych” repozytoriów:

- Przemysł, gospodarka (DB, DW, ERP, CRM,..)
- Archiwa
- „Open data” z organizacji, rządu, itd.

Wzrost przekonanie o roli Big Data i jej wpływie na społeczeństwo, …



Definicje Big Data

Wielość poglądów

"A collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications." - Wiki

"Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it … [J.Gama 2015]

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

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 ket research topics. Proc. 4th Conf. on Integrated Infromation 2014.

N.Japkowicz, J.Stefanowski: : Big Data Analysis: New Algorithms for a New Society, Springer 2016: rozdział 1.

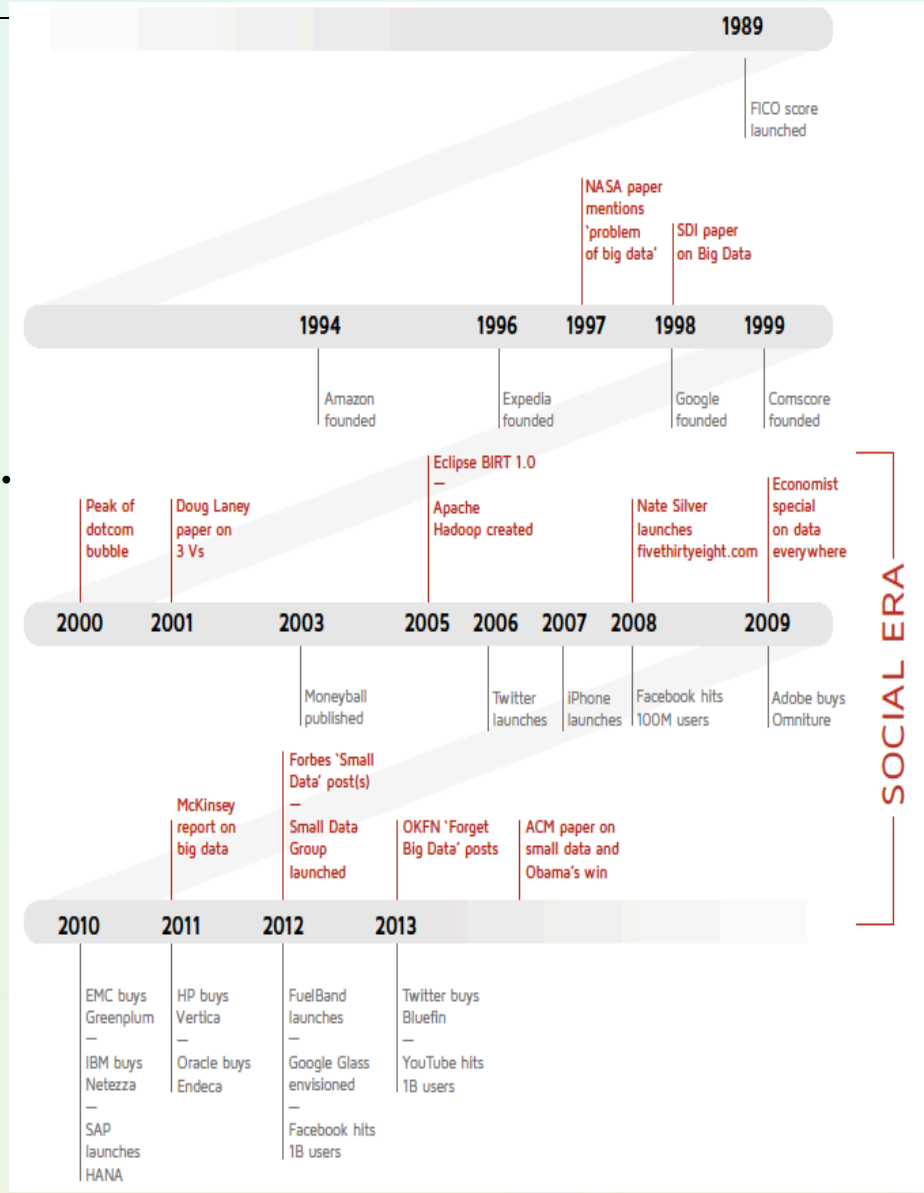
Kilka uwag historycznych

Dyskusja rosnących rozmiarów danych vs. pojęcie Big Data

Massive Data mining, Very Large Databases – inne spojrzenie

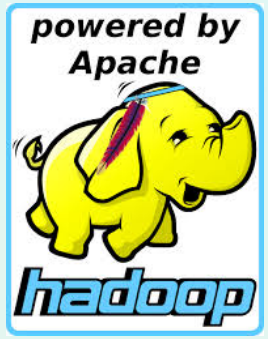
“Big Data” – jako termin na amer. konf. pojawia się pod koniec lat 90tych:

- 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



Rozwiązania obliczeniowe dla Big Data

- 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,…
- Specjalne środowiska programowe
  - Mahout (scalable machine learning)
  - Spark MLBase / Mlib
  - Vowpal Wabbit
  - h2o
  - MOA → SAMOA
  - Domino
  - …



Gdzie jesteśmy?

Sztuczna inteligencja

Uczenie maszynowe

Data Mining
Eksploracja danych

Big data mining

Odkrywanie wiedzy z danych
Process KDD

Data science

Co to jest – Data science – Nauka o danych?

Data Science − może − Science of Data

Czy to nie jest modny “buzzword”?

- Data Science is a field of study which includes everything from Big Data Analytics, Data Mining, Predictive ML Modeling, Data Visualization, Mathematics, and Statistics + intensive programming
  - *Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with different data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, computer science, and information science* [Hayashi et al]
- Historycznie silny związek ze statystyką!

Spojrz https://en.wikipedia.org/wiki/Data_science

- W dużym stopniu powiązane z inżynierskimi umiejętnościami przetwarzania danych (ang. Hacking) i aktualnym oprogramowaniem

Po polsku? − nauka o danych?

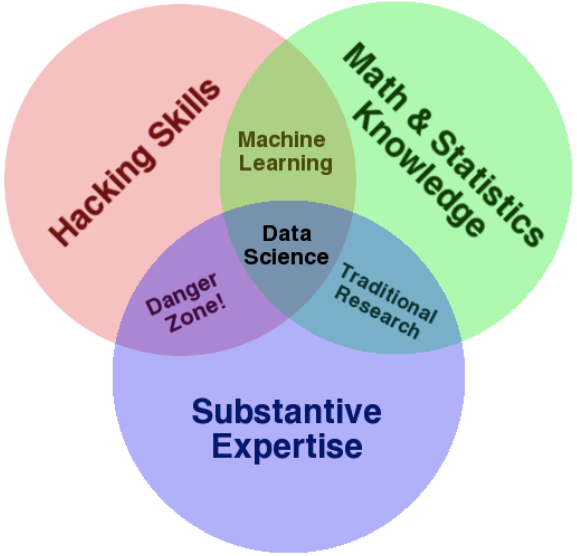


Figure 1-1. Drew Conway's Venn diagram of data science

Data Scientist − the sexiest job of the 21st century

Termin w j. ang. wylansowany około 2012 r. / znany wcześniej

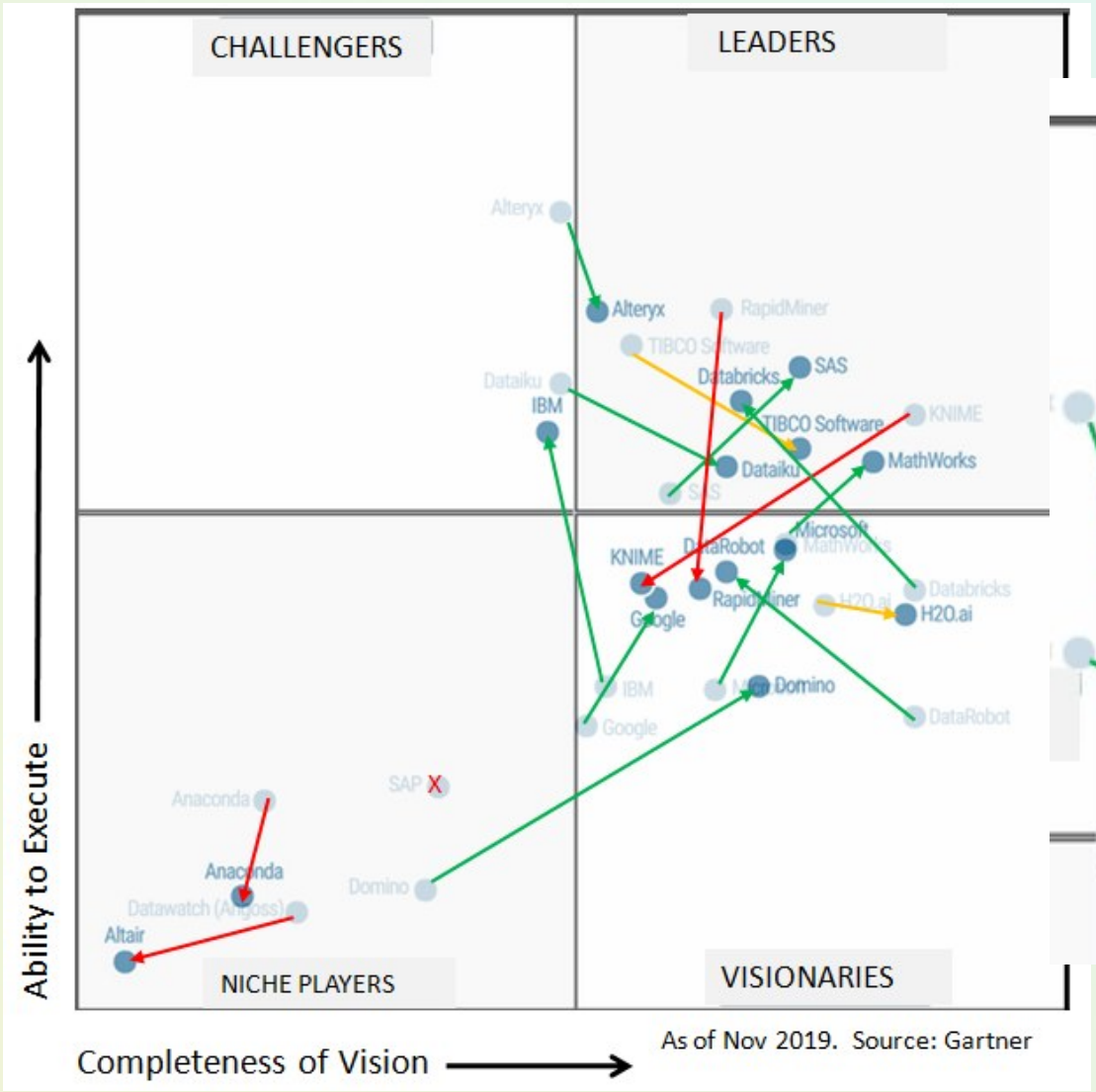
Na ogół zwraca się uwagę, że powinien:

- Wykazać wszechstronne umiejętności poruszania się w niespójnych, różnorodnych zbiorach danych, związane z technologiami, które bardzo dynamicznie ewoluują
- Być zaangażowany w zrozumienie biznesu - potrafi dostrzec szeroki kontekst swoich analiz, wyjść poza schematyczne rozwiązania, przenieść rozwiązanie z innej dziedziny,
- Myśli innowacyjnie i strategicznie, jest ciekawy świata i patrzy na niego przez pryzmat danych z różnorodnych źródeł, rozwiązuje problemy, ma zacięcie hackerskie
- Mięć zdolność do komunikacji - także z wyższą kadrą zarządzającą, wykorzystanie wysokiej jakości wizualizacji, umiejętność wyjaśnienia wykonanych złożonych analiz, pracy w zespole, ...
- Często posiadać kompetencje administratora baz danych służących do przetwarzania dużych zbiorów danych (ang. Big Data) - w spec środowiskach, np. Hadoop czy Spark.

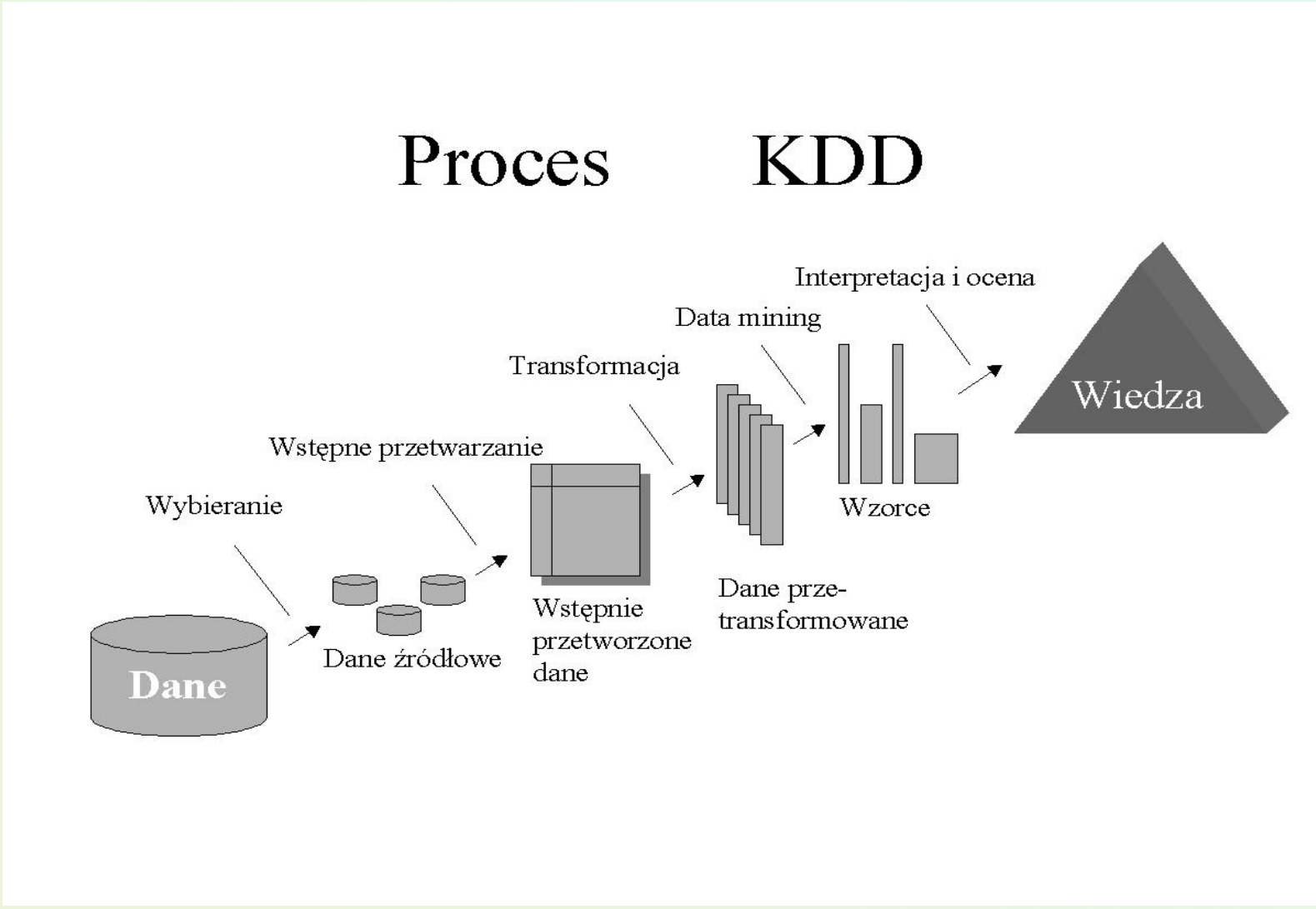
Różni się od tradycyjnego (statystka) analityka danych /inne dane i narzędzia/ choć może dążyć do podobnych wniosków

Lecz : "I think data-scientist is a sexed up term for a statistician" [Nat Silver]

Przegląd oprogramowania / KDNuggets Gartner (2019)

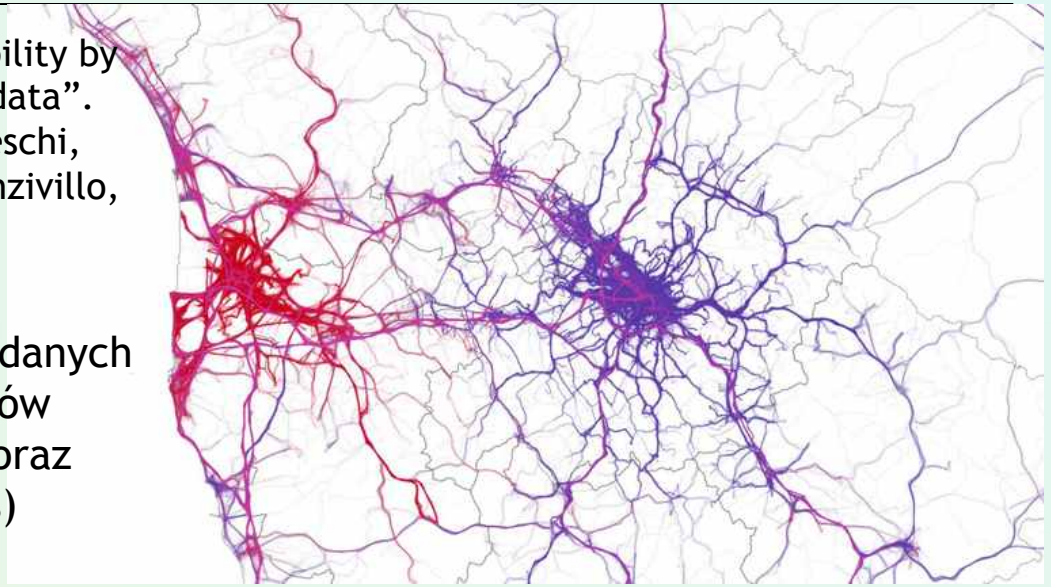


Proces odkrywania wiedzy i etapy początkowe



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



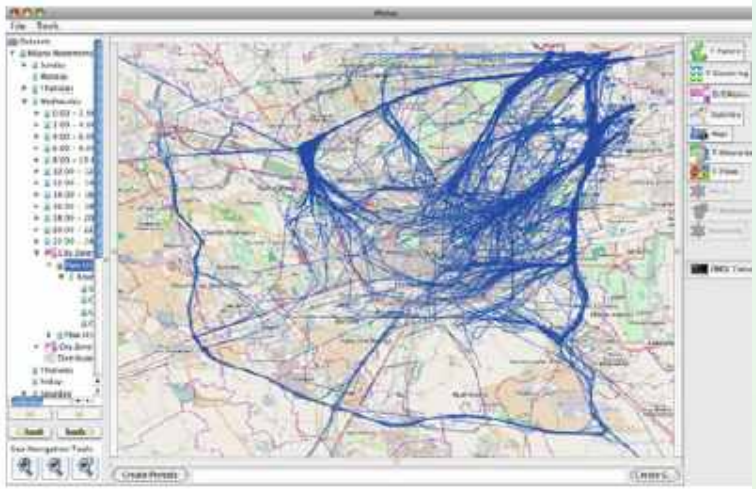


Fig. 17 The result of T-Clustering from the trajectories moving from the center to the North-East area. *Left* The input data set for the clustering algorithm: the trajectories moving from the center to the North-East

Poszukiwanie wzorców zachowań

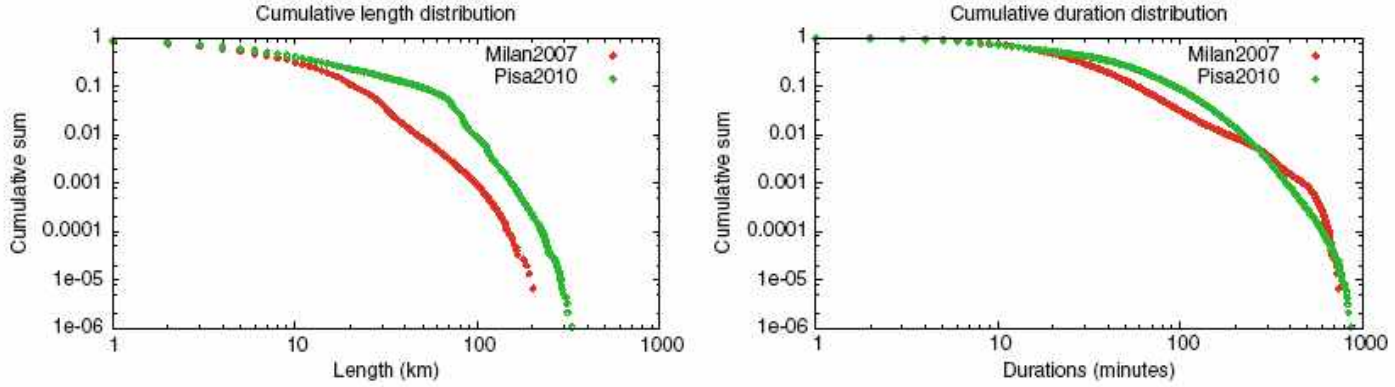
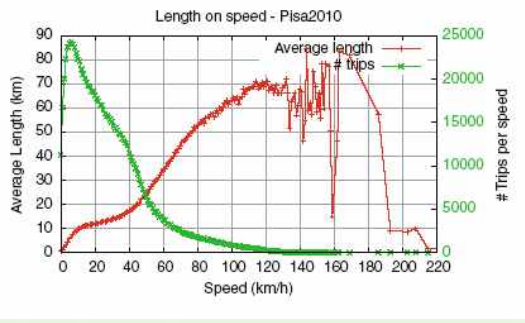


Fig. 3 Trip length cumulative distribution in log–log scale (*left*), trip duration cumulative distribution in log–log scale (*right*). *Re*
Milano2007 and *green lines* for Pisa2010



Ruch drogowy - zmiany

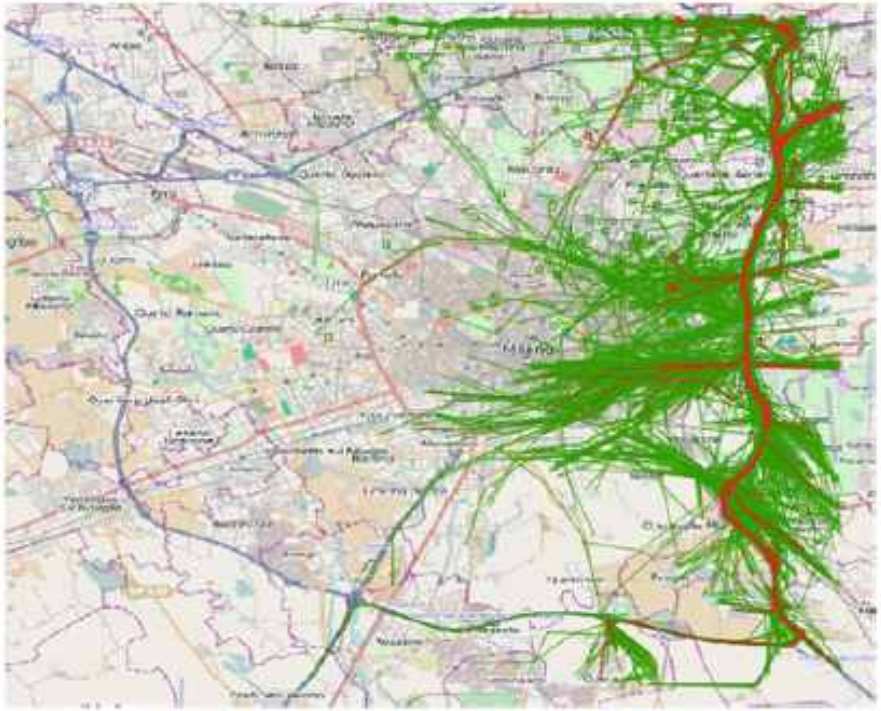


Fig. 12 Classification of new trajectories using a set of specimens from `WednesdaySpecimens`. *Left*, *blue lines* represent the trajectories of a single cluster of Wednesday, April 4, and the *red lines* are the specimens learned for the selected cluster. *Right*, *green lines* are the trajectories of the entire week classified by the same set of specimens. Visual inspection confirms that cluster shape is preserved, albeit the size of the second data set is 7 times larger. Quantitative measures of cluster quality, such as silhouette coefficients, can be easily computed

Grupowanie podobnych mobilności

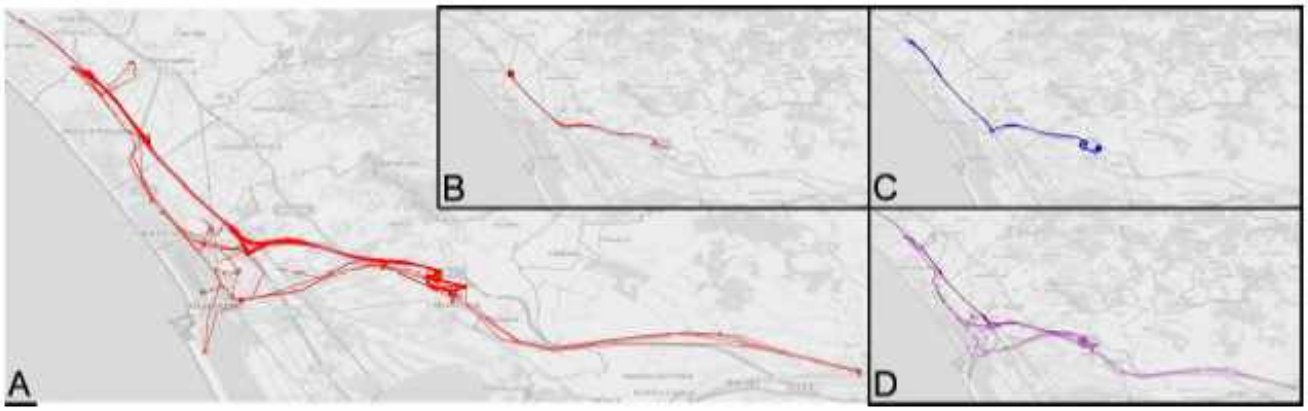
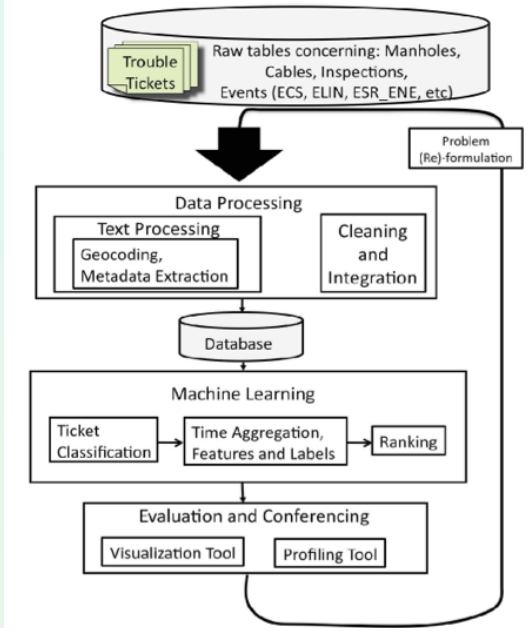


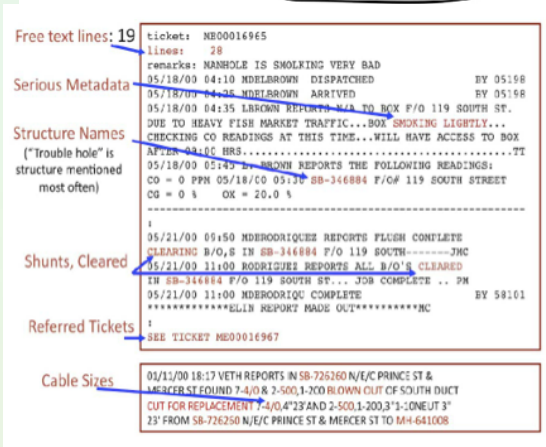
Figure 2: An example of mobility profile extraction: (a) The entire set of trips of a user, (b,c) the two clusters extracted, and (d) the remaining trips which are not periodic.

Fig. 3: *(Left)* The clusters obtained with grid cell size of 2000m. *(Right)* The clusters determined by the level 2 of the Infomap hierarchy for the same grid resolution.

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).

Typowe zadanie w analizie medycznych danych:

- Identyfikacja najważniejszych czynników (atrybutów / cech) dla oceny stanu pacjenta
- Odkrywanie zależności między opisem pacjenta (wartości atrybutów) a decyzją co do pacjenta, np. klasyfikacją pacjentów (diagnozy, sposoby postępowania)
- Wypracowanie zaleceń klinicznych (postępowania)

Highly selective vagotomy (HSV)

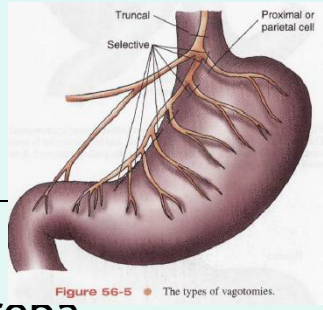


Figure 56-5 The types of vagotomies.

Problem dotyczy analizy stanu zdrowia pacjentów cierpiących na chorobą wrzodową dwunastnicy, leczonych metodą wysoce wybiórczej wagotomi

- Highly selective vagotomy (HSV) - laparoscopic surgery for perforated Duodenal Ulcer Disease – stomach;

Dane dotyczą pacjentów przyjmowanych do szpitala AM P i skierowanych do zabiegu chirurgicznego wysoce wybiórczej wagotomii

- 186 pacjentów × 15 przedopercyjnych atrybutów przypisanych do 4 kategori wyników leczenia (long term result wrt. Visick grading)
- Pytania medyczne:
  - Poszukiwanie istotnych zależności, regularności pomiędzy atrybutami
  - Ustalanie hierarchii ważności atrybutów opisujących pacjentów, także ocena znaczenia atrybutów i ich podzbiorów dla klasyfikacji pacjentów
  - Poszukiwanie tzw. modeli pacjentów charakterystycznych dla klas decyzyjnych
  - Określenie wskazań do przeprowadzania zabiegu chirurgicznego wysoce wybiórczej wagotomii (HSV) dla pacjentów cierpiących na chorobę wrzodową dwunastnicy

Lista atrybutów (w danych trochę inne kody)

a1: Płeć

a2: Wiek [lata]

a3: Okres trwania choroby [lata]

a4: Bolesność uciskowa w nadbrzuszu [T tak/ N nie]

a5: Ból w nadbrzuszu po zjedzeniu posiłku występujący w czasie 1 godziny [N – brak, S – słaby, L – silny].

a6: Występowanie u pacjenta odczucia tzw. „zgagi" czy „odbijania" [tak/ N nie]

a7: Obecność „niszy" na zdjęciu RTG {T tak/ N nie]

a8: Komplikacje wrzodu [N – brak, A – ostry krwotok, M - wielokrotne krwawienie, PE - perforacja, PY - zwężenie odźwiernika

a9: Koncentracja HCL [mmol HCL/100ml]

a10 Objętość wydzielania soku żołądkowego na godzinę [ml/godz]

a11 – Objętość wydzielania soku żołądkowego zalegająca [ml]

a12 – Tzw. BAO – ang. Basic Acid Output [mmol HCL/100ml]

Podobne parametry jak A9, A10, A12 ale po wykonaniu testu histaminowego

a13: Koncentracja HCL [mmol HCL/ 100ml]

a14 – Objętość wydzielania soku żołądkowego na godzinę [ml]

a15 – Tzw. MAO – ang. Maximal Acid Output [mmol HCL/100ml]

Długoterminowy rezultat zabiegu tzw. skali Visick'a,– wyraża ona skuteczność stosowanego leczenia:

1. Znakomity wynik leczenia. (99)
2. Bardzo dobry. (30)
3. Satysfakcjonujący. (29)
4. Niesatysfakcjonujący. (28)

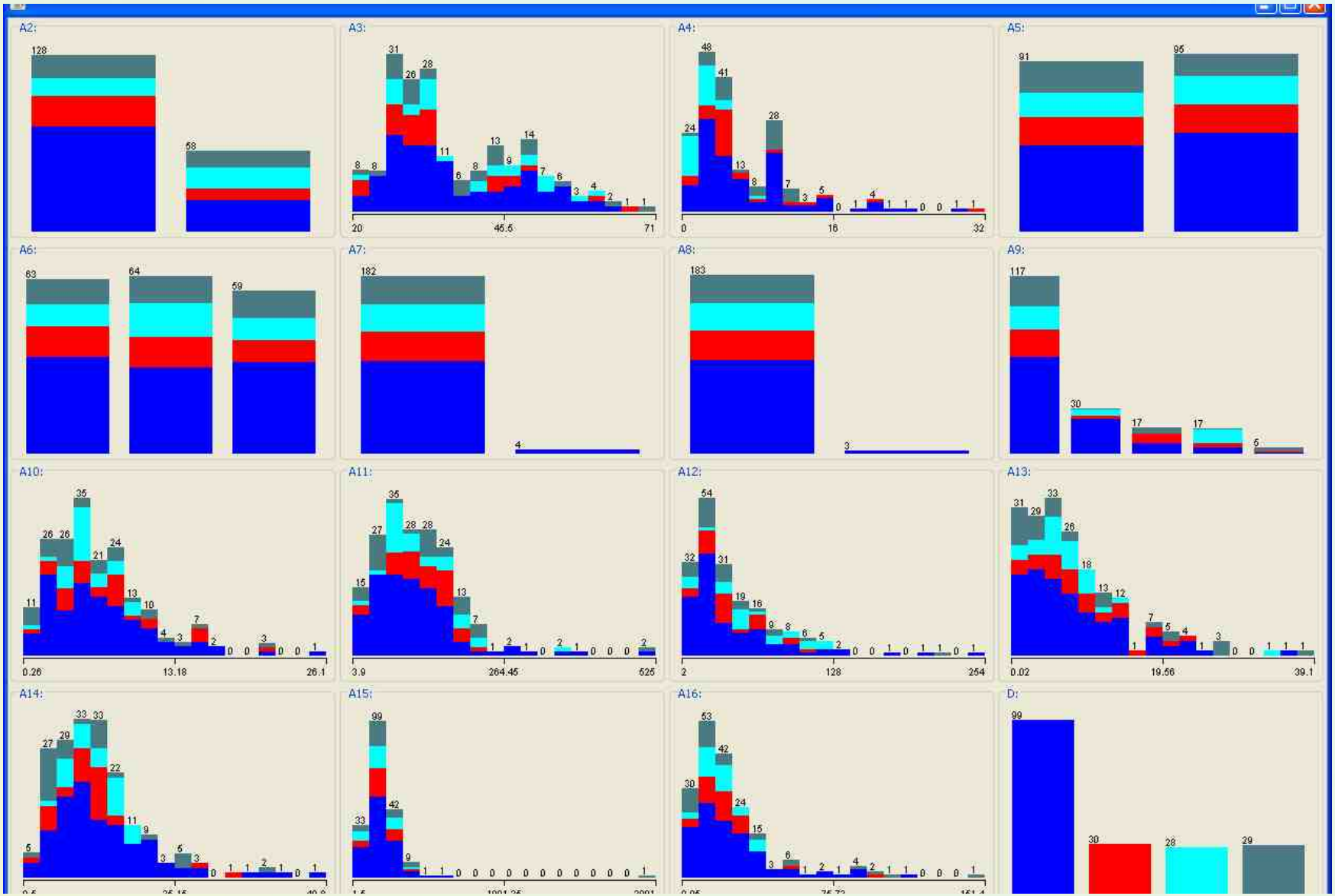
Oceń jakości oryginalnych danych

M169 | fx | 7

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
158	**M**	**45**	**7**	T	S	T	T	**N**	**2,4**	**54**	**22**	**2,7**	**9**	**1**
159	F	32	0,5	N	S	T	T	N	4,8	89	73	5	16,7	
160	**F**	**52**	**0,5**	T	S	T	T	**PE**	**8,92**	**115**	**5**	**8,7**	**15**	**2**
161	F	32	3	N	L	T	T	N	6,7	190	63	11,6	18,8	1
162	**F**	**51**	**1**	N	L	T	T	**PE**	**5,3**	**176**	**101**	**8,7**	**7,4**	**2**
163	M	32	3	T	S	T	T	N	4,4	97	98	9,8	10	1
164	**M**	**47**	**1**	N	N	T	T	**PE**	**5**	**135**	**101**	**9,3**	**11,2**	**2**
165	**F**	**47**	**3**	T	N	T	T	**N**	**5,2**	**78**	**110**	**3,5**	**11,8**	**2**
166	**M**	**51**	**3**	T	S	T	T	**A**	**4,2**	**112**	**45**	**8,85**	**8,2**	**1**
167	**M**	**50**	**9**	T	L	T	T	**N**	**2,52**	**42**	**11**	**1,06**	**5,2**	**1**
168	**m**	**28**	**0.5**	T	N	T	T	**N**	**7.2**	**65**	**42**	**6.4**	**16.4**	**1**
169	F	28	9	T	N	T	T	N	6,7	154	56	12,3	7	20
170	F	30	9	N	S	T	T	N	7,6	123	45	6	12,1	1
171	**M**	**44**	**2**	N	S	T	T	**N**	**5**	**66**	**99**	**6,3**	**10**	**1**
172	F	26	4	N	L	T	T	N	13,7	212	31	27	26	3
173	F	39	10	T	S	T	T	PY	10,6	145	32	12	6,7	2
174	M	30	10	T	L	T	T	N	6,8	143	21	7	5,5	1
175	**M**	**54**	**2**	N	S	T	T	**PE**	**6**	**145**	**118**	**8,7**	**6,8**	**2**
176	F	32	11	T	S	T	T	N	3,9	134	56	6,3	5,7	1
177	M	56	5	N	N	T	T	M	1,5	5,6	25	0,9	5,4	1
178	M	23	4	T	N	T	T	N	13	234	31	28	21	2
179	M	45	8	N	L	T	T	PY	9	185	21	19,6	12	2
180	**F**	**34**	**11**	N	N	T	T	**N**	**3,5**	**167**	**34**	**5,6**	**11,4**	**1**
181	F	33	8	T	S	T	T	N	15,7	155	33	9	15,5	1
182	M	30	4	T	L	T	T	M	8,8	95	30	8,3	13	1
183	M	26	4	N	N	T	T	N	4,5	176	61	7,8	5,6	2
184	F	29	5	N	L	T	T	N	15,4	122	39	18,8	11,7	2
185	M	28	4	N	S	T	T	N	8,4	114	32	9,9	13,4	1
186	M	47	5	T	N	T	T	M	4,6	67	23	3,2	12,1	1

Wykorzystanie raportowania i statystyk opisowych do odkrycia niepoprawnych zapisów atrrybutów –> popraw lub skonsultuj z ekspertem

Ocena atrybutów – wizualna + statystyki opisowe



Badanie zależności atrybutów

Różne metody (miara oceny + metoda poszukiwań) chi2 (WEKA)

Ranked attributes:

44.179 A9:
34.928 A4:
11.147 A11:
3.592 A14:
2.68 A12:
1.816 A15:
1.415 A5:
1.125 A2:
1.05 A16:
0 .85 A3:
0 A8
0 A13
0 A7:
0 A10:
0 A6:
0 A5:

Gain Ratio feature evaluator

Ranked attributes:

0.1461 A4:
0.1326 A9:
0.1246 A14:
0.09646 A11:
0.04641 A15:
0.0015 A3:
0.0064 A10:
0.0052 A13:
0.0044 A16:
0.0032 A12:
0,0008 A2:
0 A8:
0 A7:
0 A6:
0 A5:

Poszukiwanie podzbiorów (np. Consistency Subset Evaluator)
A4,A9, A11, A14, A15 = podobnie RST (z dyskretyzacja)

Podzbiór najistotniejszych atrybutów

Wskazano pięć atrybutów:

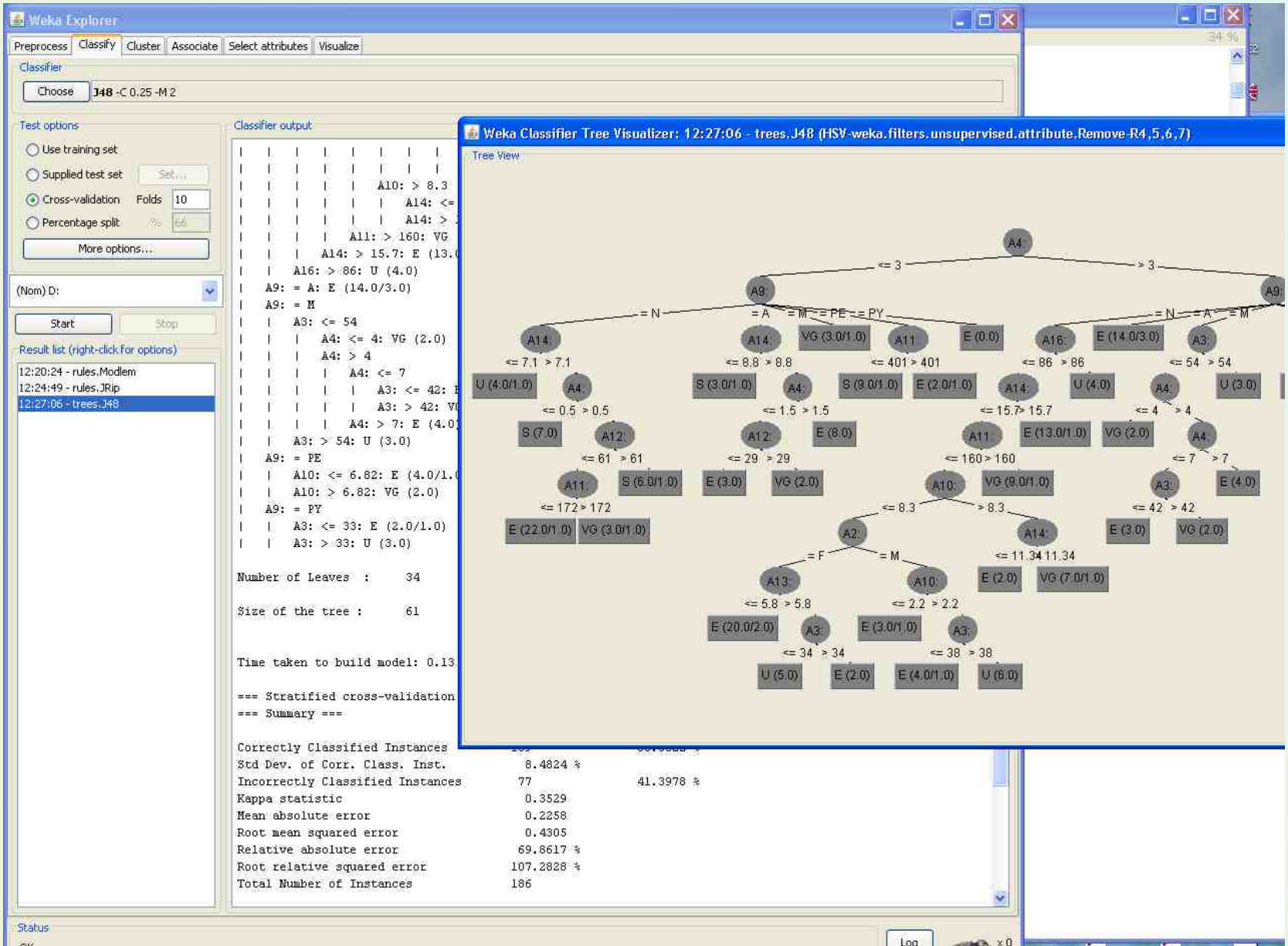
- Okres trwania choroby
- Komplikacje wrzodu
- Objętość wydzielania soku żołądkowego na godzinę
- Koncentracja HCL
- Objętość wydzielania soku żołądkowego na godzinę

4 atrybuty nadmiarowe

Reszta „mało-istotne”

Więcej - Z.Pawlak. K.Słowiński et al. Rough Classification of HSV Patients. Intelligent Decision Support.

Symboliczne klasyfikatory – J4.8prunned

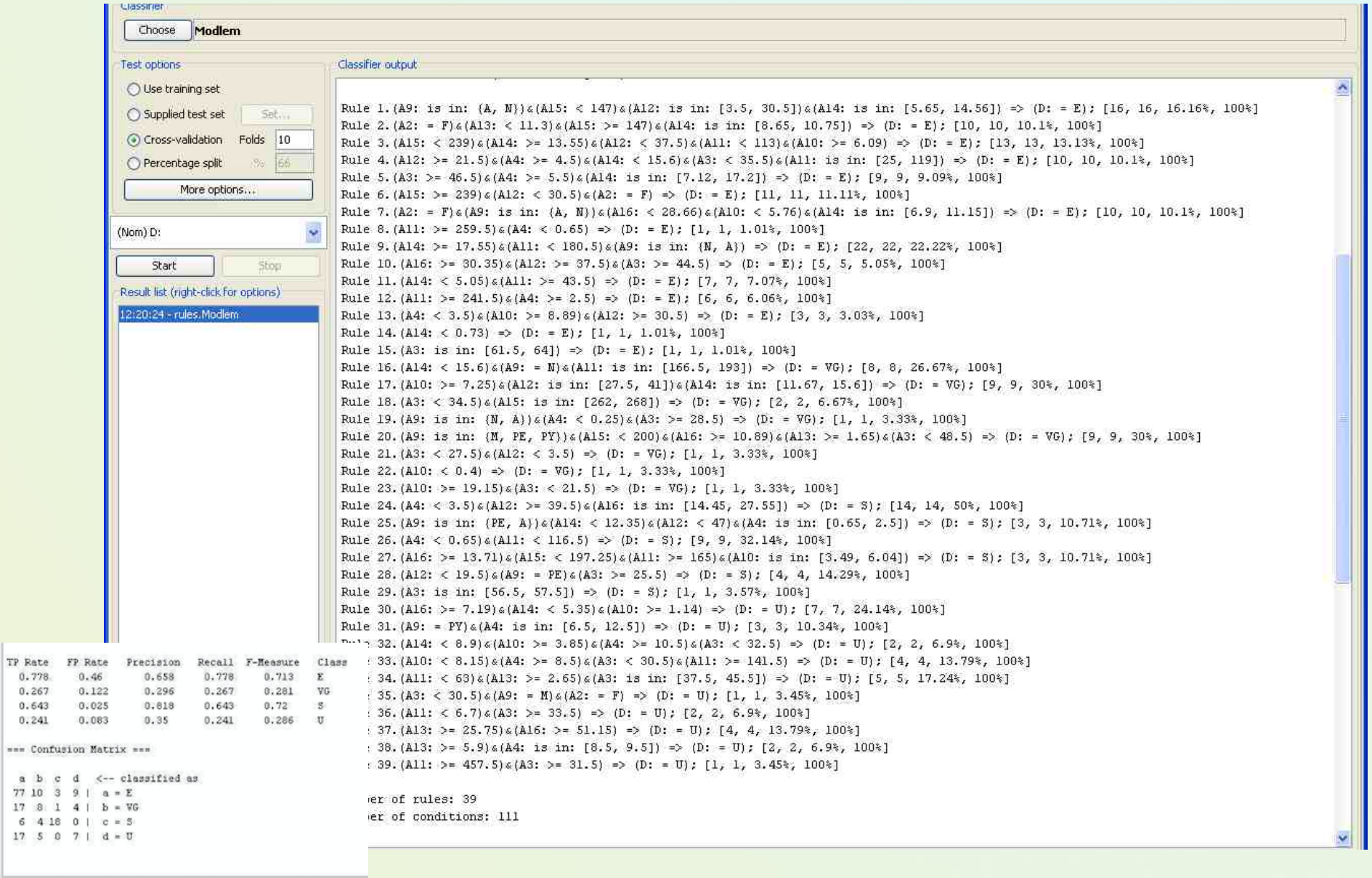


Klasyfikatory – interpretacja (opis pacjenta vs. decyzja)

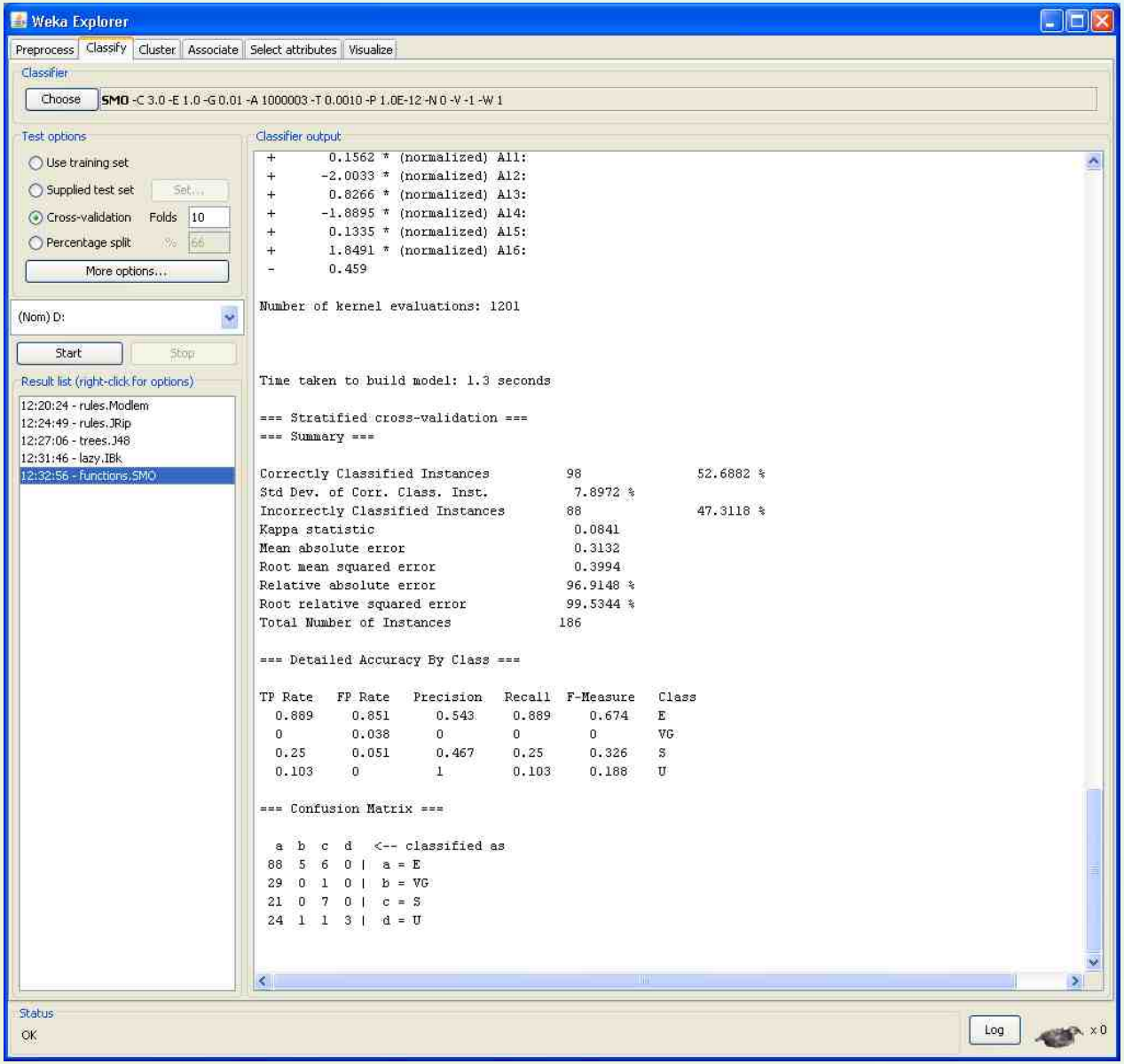
```
JRIP rules:
===========
(A4: <= 3) and (A12: >= 63) and (A15: <= 236) => D:=S (13.0/1.0)
(A4: <= 0.5) and (A11: <= 115) => D:=S (6.0/0.0)
(A10: <= 4.21) and (A10: >= 4.2) => D:=S (3.0/0.0)
(A14: <= 7.1) and (A10: >= 5.8) => D:=U (6.0/1.0)
 => D:=E (158.0/61.0)
Number of Rules : 5
=== Stratified cross-validation ===
Correctly Classified Instances         101               54.3011 %
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.848     0.701      0.579     0.848     0.689    E
  0.033     0.026      0.2       0.033     0.057    VG
  0.5       0.07       0.56      0.5       0.528    S
  0.069     0.057      0.182     0.069     0.1      U
```

Czy reguły RIPPER I PART są łatwo interpretowalne?

Reguły MODLEM (trafność 57%), lecz niezbalansowanie



Próba dostrojenia SVM (podnosimy par. C)



Klasyfikacja danych niezbalanasowanych

Wszystkie klasyfikatory nadmiernie ukierunkowane na pierwsza klasę (Excelent result of treatment)

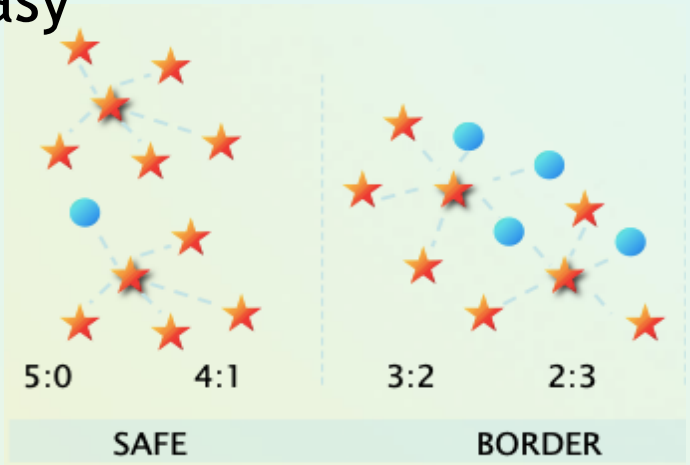
Analiza sąsiedztwa najmniejszej klasy

0% safe

5% bordeline

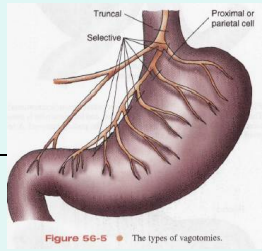
29% rare

66% otliers



Działanie klasyfikatorów (Sensitivity, G-mean, AUC) można polepszyć – pre-processing (SMOTE, SPIDER) lub BRACID około 15%

Highly selective vagotomy (HSV)



Klasyfikatory symboliczne

- LEM2 algorithm → 44 reguł (1- 5 warunki elementarne)
- MODLEM → 39 reguł (1- 6 warunków)
- Predictive performance – accuracy 57% (not the main criterion)

Dla lekarzy ważniejsza jest interpretacja wzorców – rekomendacji postępowania

- Znajdź tzw. characteristic profiles of patients
- Co odróżnia pacjenta z dobrym i złym wynikiem leczenia

Można zastosować podejścia XAI wykorzystujące tzw. Indeks Shapleya (niestety koszty obliczeniowe)

Ocena warunków w LEM2 regułach – class 1 (excelent)

Möbius		Shapley		Banzhaf	
Cond	*Value*	*Cond*	*Value*	*Cond*	*Value*
A6=2	2,34	A6=2	3,85	A6=2	4,01
A9=3	2,31	A4=1	3,41	A4=1	3,57
A4=2	1,89	A4=2	3,16	A4=2	3,08
A4=1	1,58	A9=3	2,59	A9=3	2,72
A3=2	1,27	A3=2	1,65	A3=2	1,88

Möbius		Shapley		Banzhaf	
Cond	*Value*	*Cond*	*Value*	*Cond*	*Value*
A4=1 & A6=2	2,62	A4=1 & A6=2	2,82	A4=1 & A6=2	2,83
A4=1 & A8=1	1,95	A4=1 & A8=1	1,95	A4=1 & A8=1	1,95
A2=2 & A6=2	1,89	A5=2 & A6=1	1,49	A5=2 & A6=1	1,49
A2=2 & A9=3	1,53	A3=3 & A7=2	1,18	A3=3 & A7=2	1,18
A5=2 & A6=1	1,49	A2=2 & A6=2	1,01	A2=2 & A6=2	1,01

Szerszy zakres art: A – age A3, time ; A4 – complications of ulcer; A6 - volume of gastric juice per h; A9 - HCL concentration after histamine; A5 - HCL concentration; A3 - duration of disease

Wartości atrybutów − przedziały dyskretyzacji

HSV –Profile charakterystyczne pacjentów

- Very good result of HSV (class 1)
  - without complications of ulcer or acute haemorrhage from ulcer,
  - medium or small volume of gastric juice per 1 hour (basic secretion),
  - medium volume of gastric juice per 1 hour under histamine,
  - high HCl concentration under histamine
  - / no medium duration of disease
- Satisfactory result of HSV (class 2)
  - long or medium duration of disease,
  - multiple haemorrhages,
  - medium or small volume of gastric juice per 1 hour (basic secretion),
  - medium volume of gastric juice per 1 hour under histamine,
  - medium or low HCl concentration under histamine
- Unsatisfactory result of HSV treatment (class 3)
  - medium or short duration of the disease,
  - perforation of ulcer,
  - high or small volume of gastric juice per 1 hour (basic secretion),
  - high volume of gastric juice per 1 hour under histamine,
  - No low HCl concentration under histamine condition in the rankings
- Bad result of HSV treatment (class 4)
  - Consistent profile
  - + new condition - low HCl concentration under histamine

Inne opisane przykłady procesu KDD

Odkrywanie wiedzy z danych, Larose Daniel (polskie tłumaczenie), PWN, Warszawa, 2006.

np. telecomunication churn problem

Baseball players - sport analytics

Analiza koszyka zakupów

i późniejsza książka pt Metody i modele eksploracji danych, Larose Daniel

Oraz inne źródła − poszukaj samodzielnie

Typowe obszary zastosowań DM

- Bank (analiza ryzyka, kredytowa, niebezpiecznych zachowań – fraud); ubezpieczenia
- Telecom, analiza połączeń, tzw. customer attrition;
- Duże sieci sprzedaży (determine sales trends, develop marketing campaigns, product or customer segmentation)
- Działy HR – poszukiwanie, rekrutacja pracowników oraz przeciwdziałanie odejściom
- Inteligentna produkcja (Industry 4.0), energetyka
- Badania naukowe: astronomia, fizyka, DNA genetyka
- Analiza danych medycznych
- Web, text, and e-commerce
- Smart cities, i ekologia/klimat
- Rolnictwo

Można spojrzeć na Data mining applications w Wikipedia

Rozdziały książki Data Mining / Zaki M., Meira W.

Part I: Exploratory Data Analysis

Chapter 1: Data Mining and Analysis Chapter 2: Numeric Attributes

Chapter 3: Categorical Attributes

Chapter 4: Graph Data

Chapter 5: Kernel Methods

Chapter 6: High-dimensional Data

Chapter 7: Dimensionality Reduction

Part II: Frequent Pattern Mining

Chapter 8: Itemset Mining

Chapter 9: Summarizing Itemsets

Chapter 10: Sequence Mining

Chapter 11: Graph Pattern Mining

Chapter 12: Pattern and Rule Assessment

Part III: Clustering

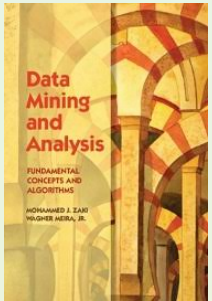
Chapter 13: Representative-based Clustering

Chapter 14: Hierarchical Clustering

Chapter 15: Density-based Clustering

Chapter 16: Spectral and Graph Clustering

Chapter 17: Clustering Validation



Part IV: Classification

Chapter 18: Probabilistic Classification

Chapter 19: Decision Tree Classifier

Chapter 20: Linear Discriminant Analysis

Chapter 21: Support Vector Machines

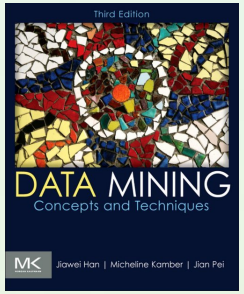
Chapter 22: Classification Assessment

W innych książkach (np. J.Han et al) dodatkowe rozdziały

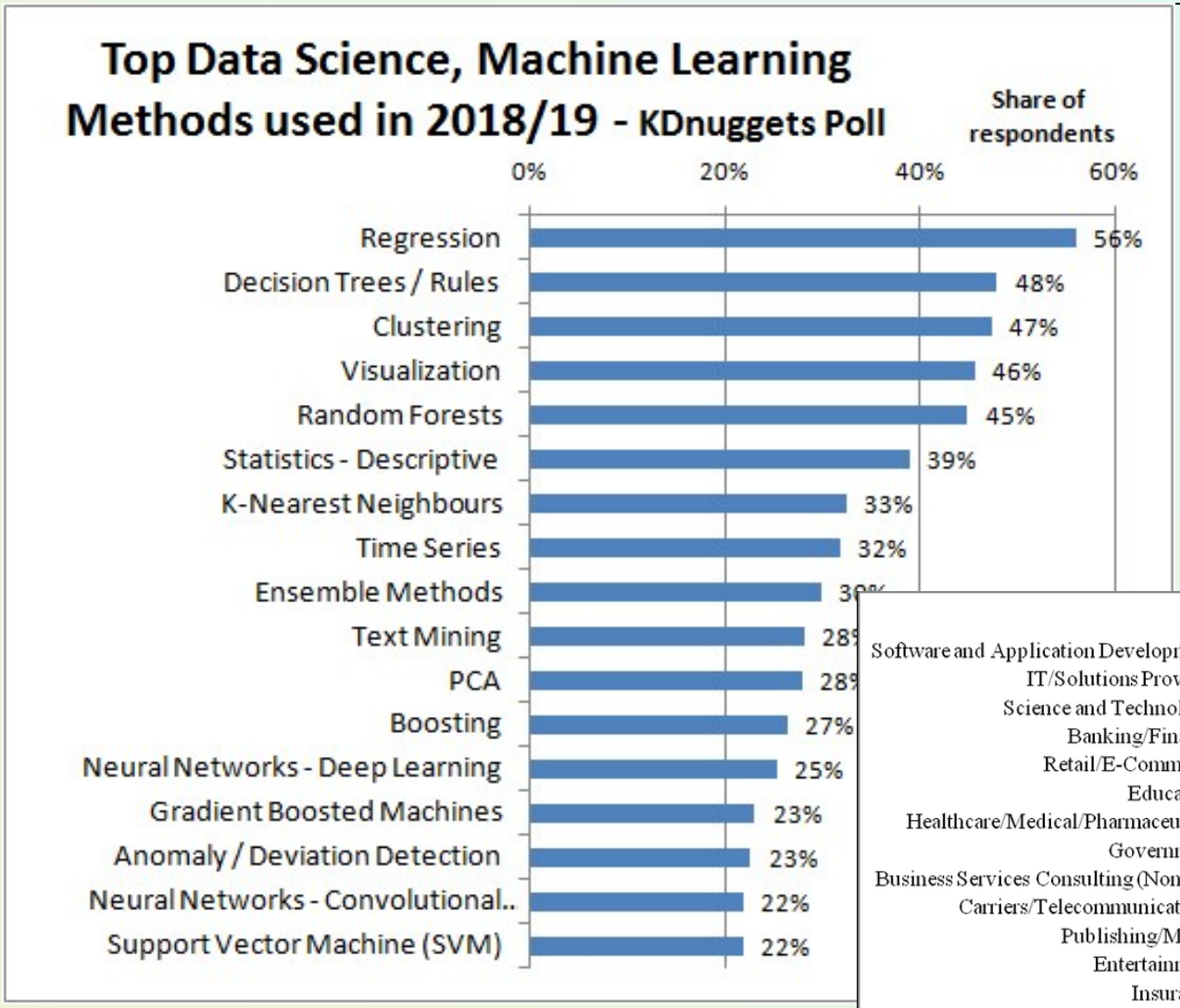
Chapter 4. Data Warehousing and On-Line Analytical Processing (..)

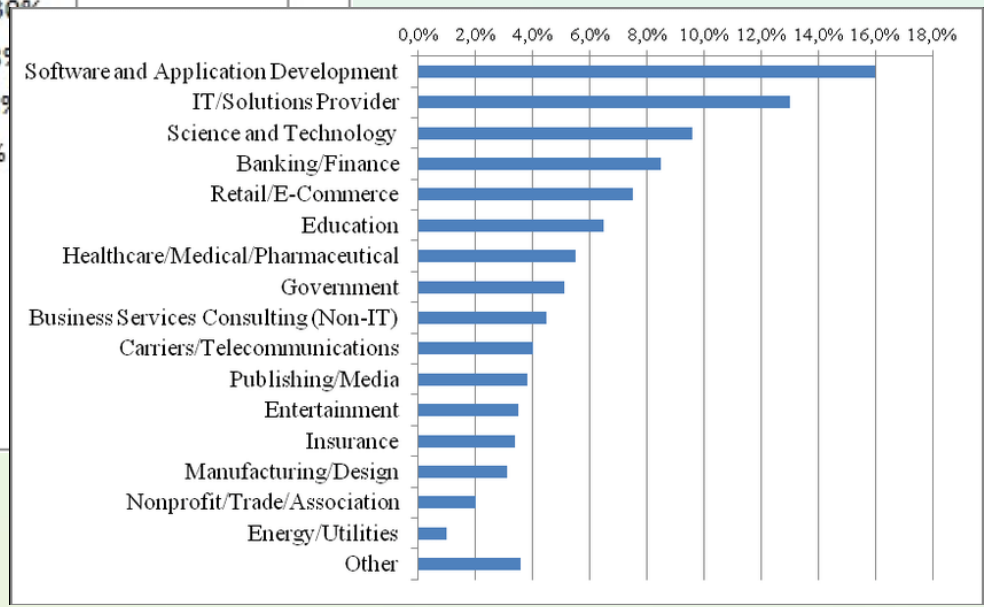
Chapter 12. Outlier Detection

Chapter 13. Trends and Research Frontiers in Data Mining -> Mining Complex Types of Data



Badania opinii nt. Data mining i ML





Mapa dalszych tematów

Data pre-processing oraz inżynieria cech

Imbalanced data

Explainable ML

Semi-supervised approaches

Active learning + Human in the loop

Time series

Data Streams

Visual mining

Może coś jeszcze?

Krótkie prezentacje ochotników (WY!!!)

Zaproszeni goście

I to by było na tyle …

Pytania lub komentarze?

Nie zadawalaj się tym
co usłyszałeś – poszukuj więcej!
Czytaj książki oraz samodzielnie
eksploruj dane!

Kontakt:
Jerzy.Stefanowski@cs.put.poznan.pl
lub www.cs.put.poznan.pl/jstefanowski