Eksploracja danych medycznych
możliwości wsparcia informatycznego

JERZY STEFANOWSKI

Instytut Informatyki
Politechnika Poznańska

ZIM UM 2011

Zawartość

1. Motywacje – dlaczego nowe metody?
2. Rola metod statystycznych
3. Eksploracja danych a systemy uczące
4. Drzewa decyzyjne
5. Indukcja reguł
6. Wybrane zastosowania teorii zbiorów przybliżonych
7. Podsumowanie

1. Motywacje

Zadanie w medycynie → gromadzenie, analiza oraz interpretacja danych

- Nowe techniki → „powódź” dostępnych danych
  - Urządzenia diagnostyczne
  - Rozwój elektronicznych systemów (HSI, medical records ,…)
- Rosnące wymagania wobec systemów e-health

Zbyt duża liczba atrybutów / czynników do rozważenia w podejmowaniu decyzji

Dostępne dane mogą mieć różne znaczenie

Dane mogą być niedokładne, nieprecyzyjne i niepełne

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)

- Inne

Typowe obszary zastosowan

Za R. E. Abdel-Aal „Data mining and medical informatics”

- Screening
- Diagnosis
- Therapy
- Prognosis
- Monitoring
- Biomedical/Biological Analysis
- Epidemiological Studies
- Hospital Management
- Medical Instruction and Training

Analiza ważności atrybutów

- Dane – wielowymiarowe tablice (pacjenci / atrybuty)
- Wiele możliwych podejść:
  - Statystyczna analiza danych
    - Statystyczne miary siły związku.
    - Metody wielowymiarowe (analiza czynnikowa, skalowanie wielowymiarowe, analiza dyskryminacyjna, …)
  - Metody wywodzące się z Uczenia Maszynowego (Machine Learning) i Data Mining (Eksploracja danych)

Statystyka opisowa a wnioskowanie statystyczne

- **Statystyka opisowa** – prezentacja danych w sposób uporządkowany, prosty z wykorzystaniem, np., miar tendencji centralnej i miar rozproszenia. Pomagają one w redukcji dużej liczby danych do zbioru bardziej zwięzłych i "pojemnych" miar.
- **Wnioskowanie statystyczne** – pozwala ustalać prawidłowości i podejmować decyzje na podstawie zredukowanej liczby danych (próby) przy zastosowaniu rachunku prawdopodobieństwa.
- Rachunek prawdopodobieństwa – możliwe jest określenie jak błąd popełnia się uogólniając wyniki z **próby** na całą **zbiorowość**

2. Weryfikacja hipotez statystycznych

- Testy parametryczne → sprawdzenie pewnej hipotezy dotyczącej poziomu nie znanego parametru albo co do postaci rozkładu zmiennej w populacji.
- Na podstawie informacji pochodzącej z próby będziemy podejmować decyzje czy przyjąć albo odrzucić hipotezę.
- Przykłady problemów badawczych dotyczących:
  - wartości badanych zmiennych,
    - np. średni wiek osób chorujących na pewną chorobę wynosi 45 lat.
  - porównania dwóch zbiorowości,
  - skuteczność oddziaływania pewnych bodźców, którym poddawane są te same grupy obiektów,
  - zależności między badanymi zmiennymi,
  - porównania rozkładów zmiennych.

Przykład testu dla 2 zbiorowości (t-Studenta)

Microsoft Excel - T-TEST2.XLS

Plik Edycja Widok Wstaw Format Narzędzia Dane Okno Pomoc Adobe PDF — Wpisz pytanie do Pomocy

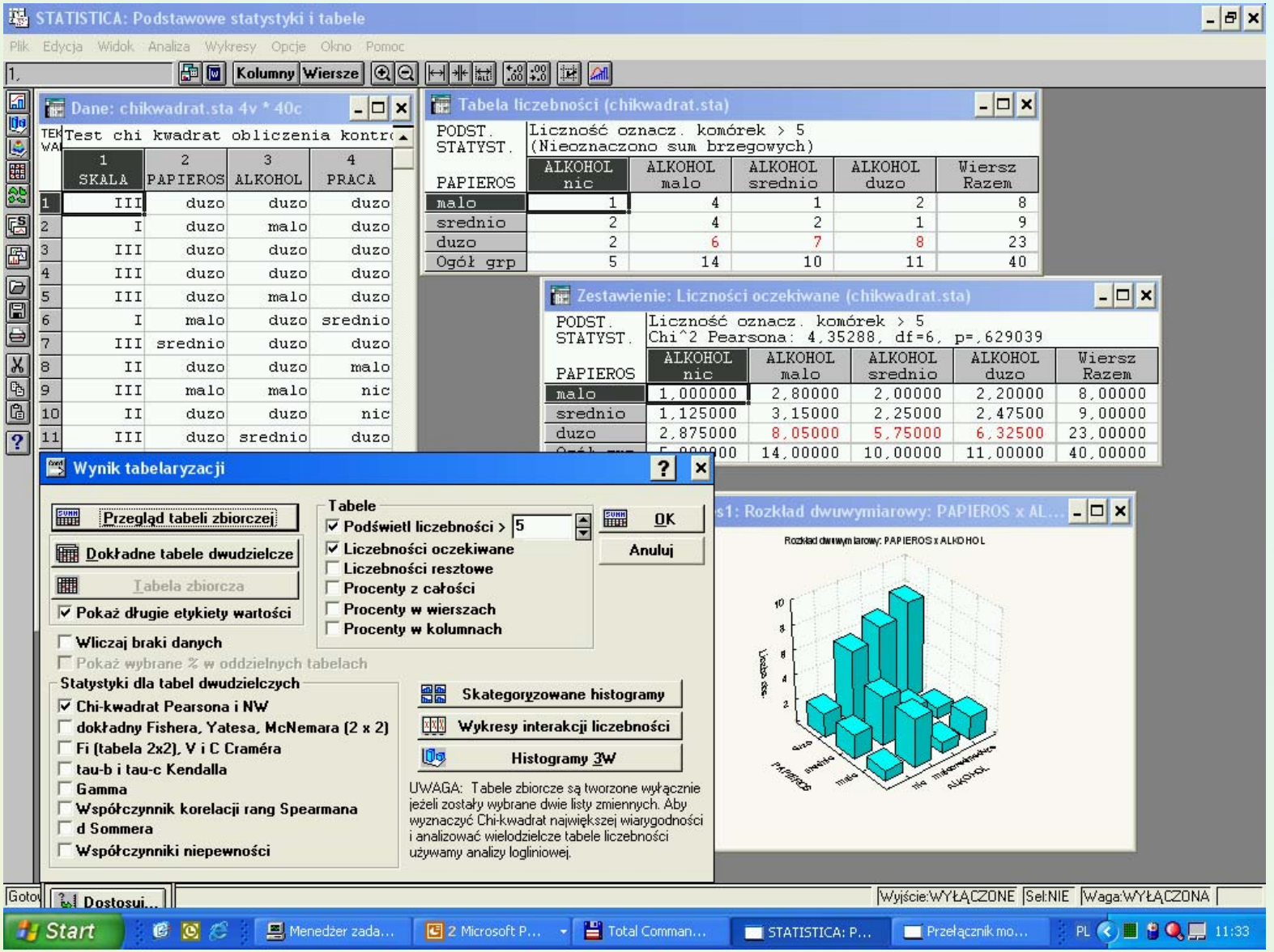
K14 =TEST.T(B10:B31;D10:D29;2;2)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
2	zakłada się równe wariancej - Patrz Oktaba rozdział 7.5														
3	Uwaga są drobne różnice w obliczeniach w stosunku do przykładu z książki Pana Oktaby														
4	Ale wniosek co do odrzucenia hipotezy zerowej Ho jest ten sam - mozna ja odrzucić!														
5															
6	Przykład 7.5.2														
7	Wysokość czaszek ryjówek w miesiącach czerwcu i lipcu														
8						Efekt uzycia opcji z Data Analysis				wykorzystanie dostępnych funkcji					
9		lipiec		czerwiec											
10		6,60		6,5		t-Test: Two-Sample Assuming Equal Variances									
11		6,60		6,5											
12		6,50		6,4			*czerwiec*	*lipiec*							
13		6,50		6,4		Mean	6,34	6,15		t(0.05)	2,02				
14		6,40		6,3		Variance	0,02	0,03		p(T0)	0,000505				
15		6,40		6,2		Observations	22,00	20,00							
16		6,40		6,2		Pooled Variance	0,03								
17		6,40		6,1		Hypothesized Mea	0,00								
18		6,40		6,1		df	40,00		stopnie swobody						
19		6,40		6,1		t Stat	3,78		wartość wyliczona statystyki t -Studenta						
20		6,40		6,1		P(T<=t) one-tail	0,00		test jednostronny						
21		6,30		6,1		t Critical one-tail	1,68		wartośc krytyczna dla 5%						
22		6,30		6,1		P(T<=t) two-tail	0,00		test dwustronny						
23		6,30		6,00		t Critical two-tail	2,02		wartość krytyczna dla 5%						
24		6,30		6,00											
25		6,30		6,00											
26		6,30		6,00		Wniosek Hipotezę mozna odrzucić									
27		6,30		6,00											
28		6,20		5,90											
29		6,10		5,90											
30		6,10													
31		6,00													
32															
33	licznosc	22,00		20,00											
34	srednia	6,34		6,15											
35															

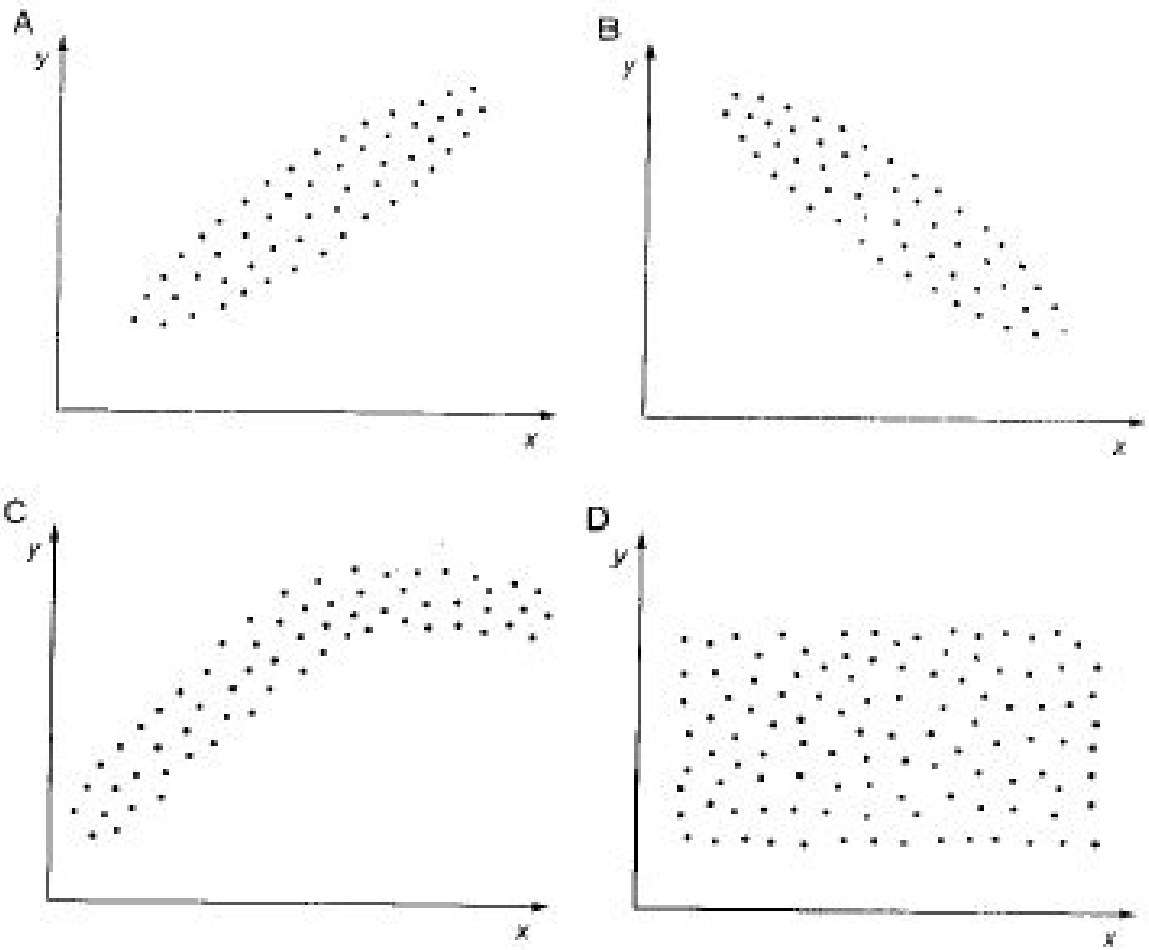
t_Student 1 | t-Stud ver 1 | t-Stud ver2 | t_Student TwoPaired | t_Student TwoPair

Gotowy

Test niezależności zmiennych - Chi-kwadrat

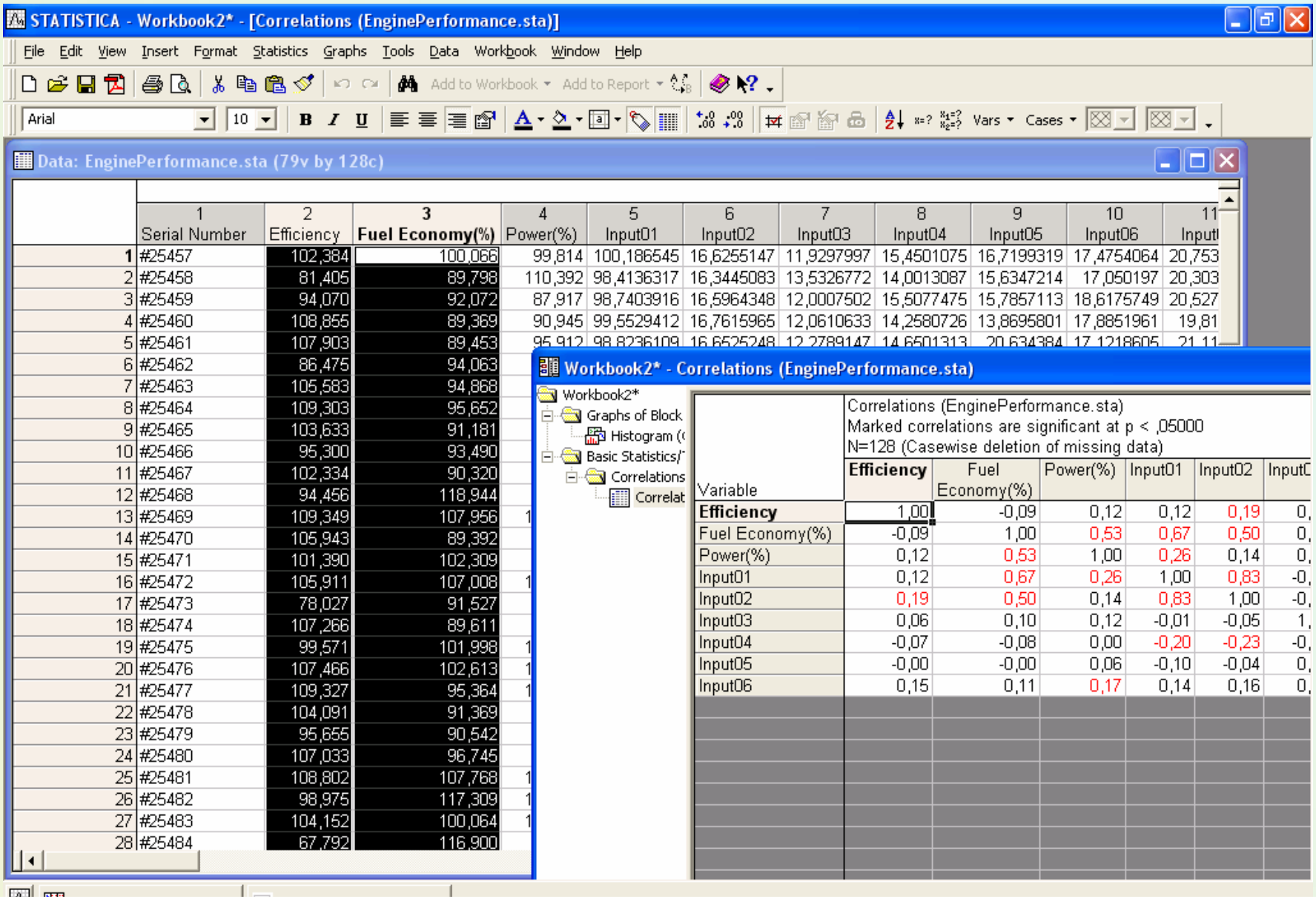


Różne zależności między zmiennymi



Rys. 10.1. Rozkłady empiryczne dwóch zmiennych

Analiza korelacji (liniowej) – pakiet Statistica

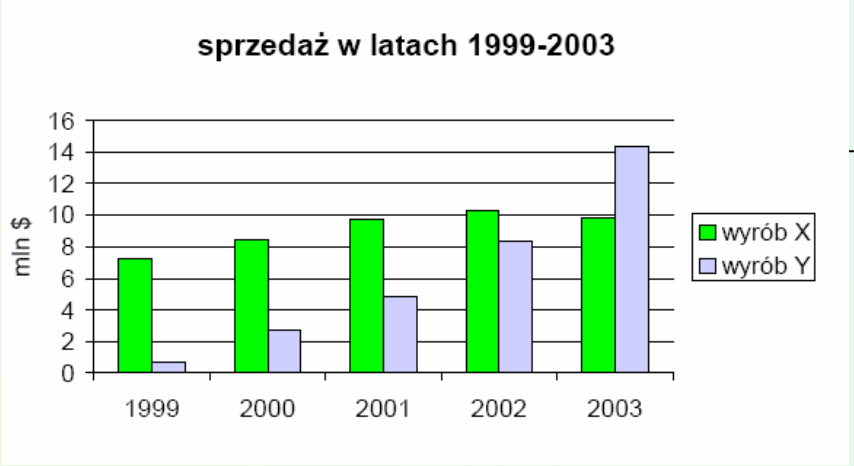


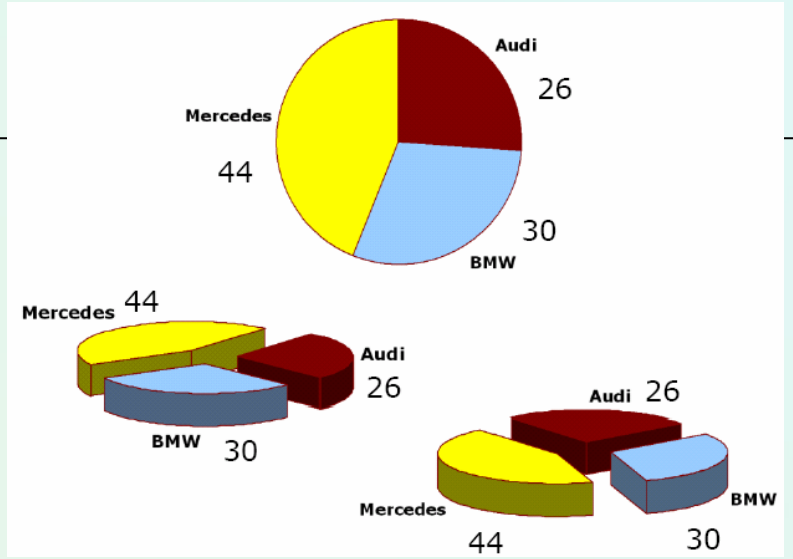
	1 Serial Number	2 Efficiency	3 Fuel Economy(%)	4 Power(%)	5 Input01	6 Input02	7 Input03	8 Input04	9 Input05	10 Input06	11 Input
1	#25457	102,384	100,066	99,814	100,186545	16,6255147	11,9297997	15,4501075	16,7199319	17,4754064	20,753
2	#25458	81,405	89,798	110,392	98,4136317	16,3445083	13,5326772	14,0013087	15,6347214	17,050197	20,303
3	#25459	94,070	92,072	87,917	98,7403916	16,5964348	12,0007502	15,5077475	15,7857113	18,6175749	20,527
4	#25460	108,855	89,369	90,945	99,5529412	16,7615965	12,0610633	14,2580726	13,8695801	17,8851961	19,81
5	#25461	107,903	89,453	95,912	98,8236109	16,6525248	12,2789147	14,6501313	20,634384	17,1218605	21,11
6	#25462	86,475	94,063								
7	#25463	105,583	94,868								
8	#25464	109,303	95,652								
9	#25465	103,633	91,181								
10	#25466	95,300	93,490								
11	#25467	102,334	90,320								
12	#25468	94,456	118,944								
13	#25469	109,349	107,956								
14	#25470	105,943	89,392								
15	#25471	101,390	102,309								
16	#25472	105,911	107,008								
17	#25473	78,027	91,527								
18	#25474	107,266	89,611								
19	#25475	99,571	101,998								
20	#25476	107,466	102,613								
21	#25477	109,327	95,364								
22	#25478	104,091	91,369								
23	#25479	95,655	90,542								
24	#25480	107,033	96,745								
25	#25481	108,802	107,768								
26	#25482	98,975	117,309								
27	#25483	104,152	100,064								
28	#25484	67,792	116,900								

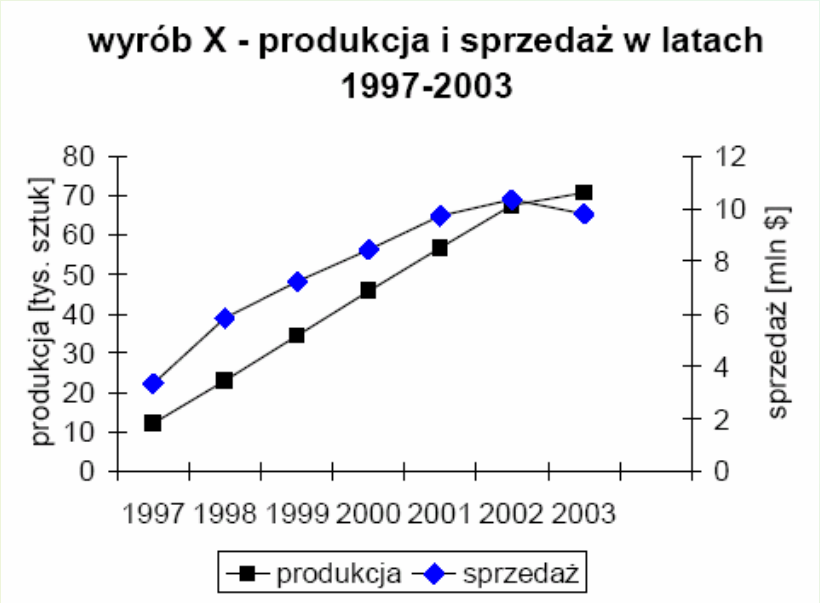
Variable	Efficiency	Fuel Economy(%)	Power(%)	Input01	Input02	Input0
Efficiency	1,00	-0,09	0,12	0,12	0,19	0,
Fuel Economy(%)	-0,09	1,00	0,53	0,67	0,50	0,
Power(%)	0,12	0,53	1,00	0,26	0,14	0,
Input01	0,12	0,67	0,26	1,00	0,83	-0,
Input02	0,19	0,50	0,14	0,83	1,00	-0,
Input03	0,06	0,10	0,12	-0,01	-0,05	1,
Input04	-0,07	-0,08	0,00	-0,20	-0,23	-0,
Input05	-0,00	-0,00	0,06	-0,10	-0,04	0,
Input06	0,15	0,11	0,17	0,14	0,16	0,

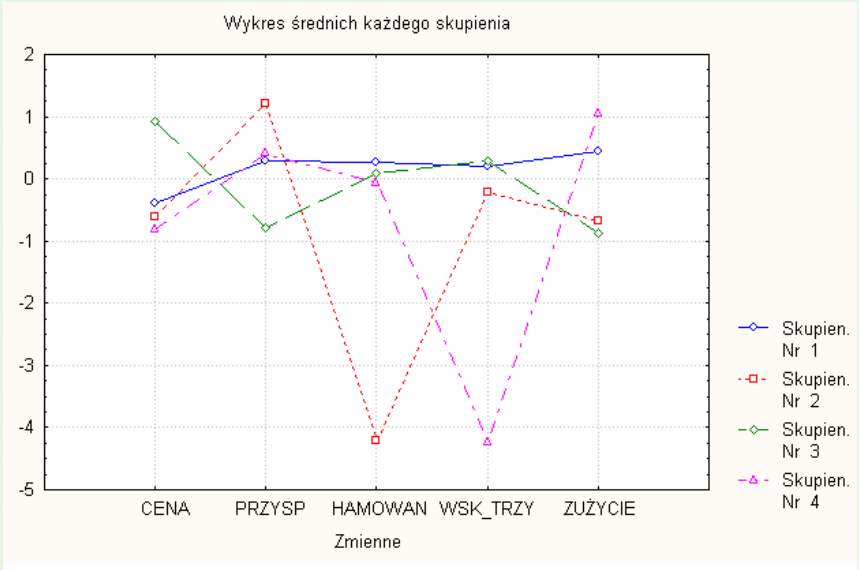
Wizualizacja informacji

- Elementy graficzne → tabele, wykresy, diagramy, rysunki, fotografie, mapy
- UZASADNIENIE
  - ludzie są „wzrokowcami”: 83% przyswajanej wiedzy wynika z pobudzeń wzrokowych
  - człowiek zapamiętuje ok. 43% więcej informacji, jeśli dokument zawiera elementy graficzne
- RÓŻNE FUNKCJE
  - przekazywanie informacji, która jest trudna do przekazania w inny sposób (słowami)
  - pomoc w wyjaśnianiu i wyróżnianiu informacji









- Przykłady różnych wykresów

Inne ciekawe metody

- Statystyka opisowa
  - Miary dynamik zjawisk; metody indeksowe
- Estymacja przedziałowa; przedziały ufności
- Analiza wariancji – ANOVA
- Regresja wielokrotna (także inne zagadnienia)
- Metody nieparametryczne i inne metody analizy danych jakościowych
- Kontrola i poprawa jakości
- Analiza wielowymiarowa
  - Wielowymiarowa analiza wariancji, analiza dyskryminacyjna, metoda głównych składowych i analiza czynnikowa, skalowanie wielowymiarowa
  - Analiza dyskryminacyjna i metody klasyfikowania
  - Analiza skupień
- Model bayesowskie i analiza decyzji
- Metody wyboru prób z populacji
- …

No i jeszcze trochę o software

MINITAB Statistical Software jest archetypem wśród programów statystycznych. Jego pierwsza wersja powstała przed 25 laty. Od 1972 roku był wykorzystywany i cytowany w ponad 300 publikacjach. Jest programem prostym w obsłudze i łatwym do nauczenia.

STATISTICA StatSoft - oferuje bogaty zestaw metod statystycznych i operacji na danych, jest łatwy w obsłudze, ma znakomitą grafikę i narzędzia do tworzenia interaktywnych aplikacji. Lista dostępnych funkcji jest bardzo długa, w jej skład wchodzą między innymi: podstawowe i zaawansowane wielowymiarowe metody statystyczne. Także uproszczone karty kontrolne, analiza przeżycia, sterowanie jakością, analiza procesu, planowanie doświadczeń oraz sieci neuronowe. Wersja Data Miner – metody data mining!

SPSS, SPSS Inc. - bardzo silne zaawansowane narzędzie. Podstawowym elementem pakietu jest moduł SPSS Base zawierający podstawowe statystyki, wielowymiarowe tablice i wykresy. Dodatkowe moduły to Advanced Statistics - zaawansowane funkcje.

SAS, SAS Institute – historycznie podstawowy zestaw modułów przetwarzania danych dla komputerów „mainframe”. Wiele procedur statystycznych. Wsparcie do integracji danych i konstruowania hurtowni danych oraz OLAP.

No i jeszcze dalej o software

STATGRAPHICS Plus Statistical Graphics Corporation - jeden z najpopularniejszych pakietów, jest idealnym narzędziem do celów dydaktycznych oraz do analizy danych o małych i średnich rozmiarach

MATLAB – Statistics Toolbox. Oryginalnie oprogramowanie The MathWorks, Inc. - przeznaczony głównie do realizacji obliczeń numerycznych. Uzupełniony w 1991 roku pakietem SIMULINK, który między innymi przejął funkcję interfejsu, wprowadzając elementy graficznych obiektów oraz okienkowy tryb dialogu z użytkownikiem

S-PLUS MathSoft - przeznaczony do prowadzenia wszechstronnej analizy statystycznej danych i generowania prezentacji graficznych. Zawiera wszystkie opcje języka S, który jest specjalizowany do programowania kompleksowej analizy statystycznej i wizualizacji danych

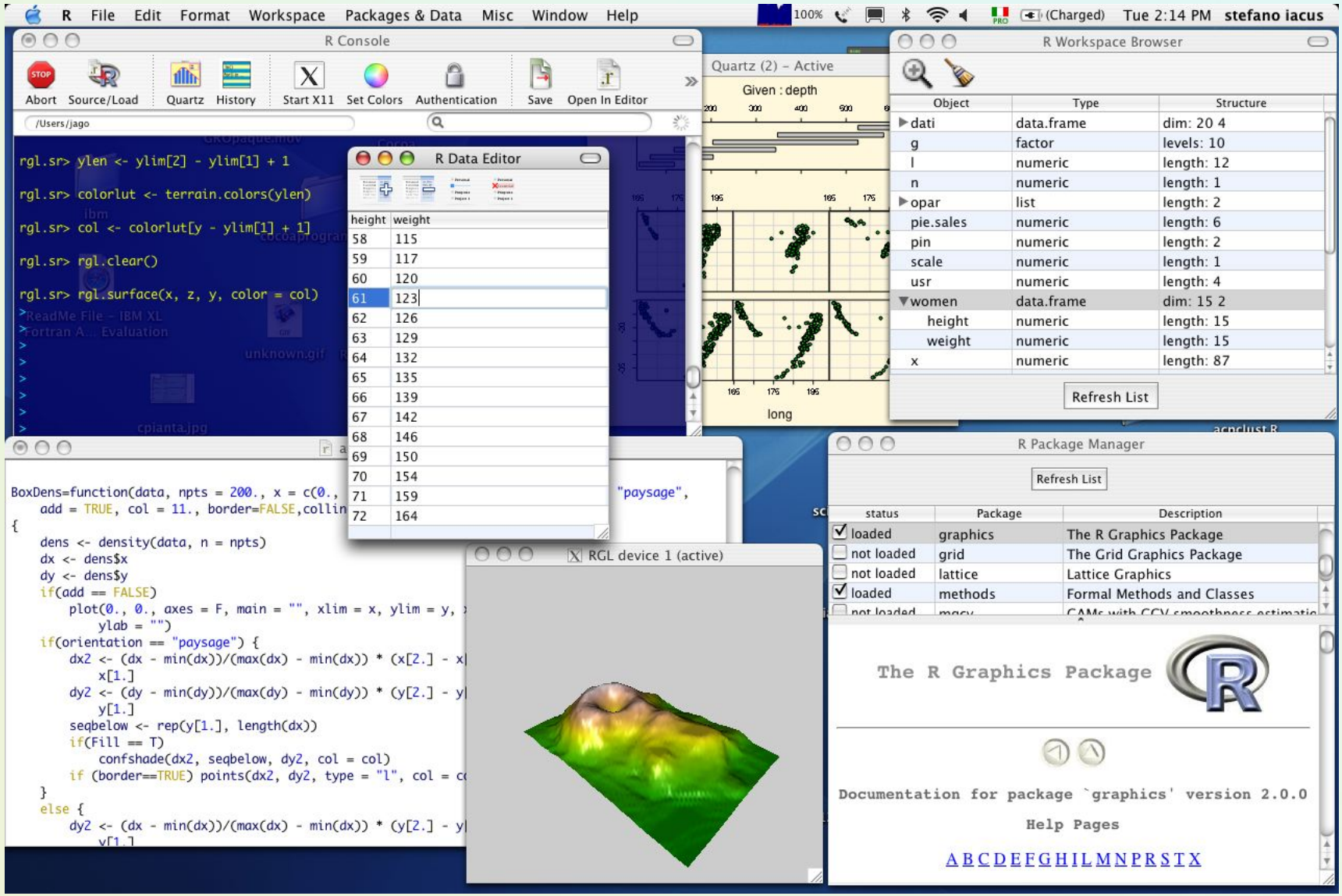
SYSTAT - kolejny pakiet SPSS, Inc.Posiada wiele funkcji przetwarzania danych i prezentacji wyników. Zawiera klasyczne procedury statystyk opisowych, testów nieparametrycznych, korelacji, szeregów czasowych, dostępne są modele liniowe i logliniowe, analizy - skupień, przeżycia, analiza dyskryminacyjna i inne

Sigma Plot 2000 Jandel Scientific Software - program przetwarzania danych, posiada rozbudowaną opcję kreślenia wykresów, ogromne możliwości edytorskie spowodowały, że pakiet wykorzystywany jest do przygotowywania profesjonalnych wizualizacji graficznych

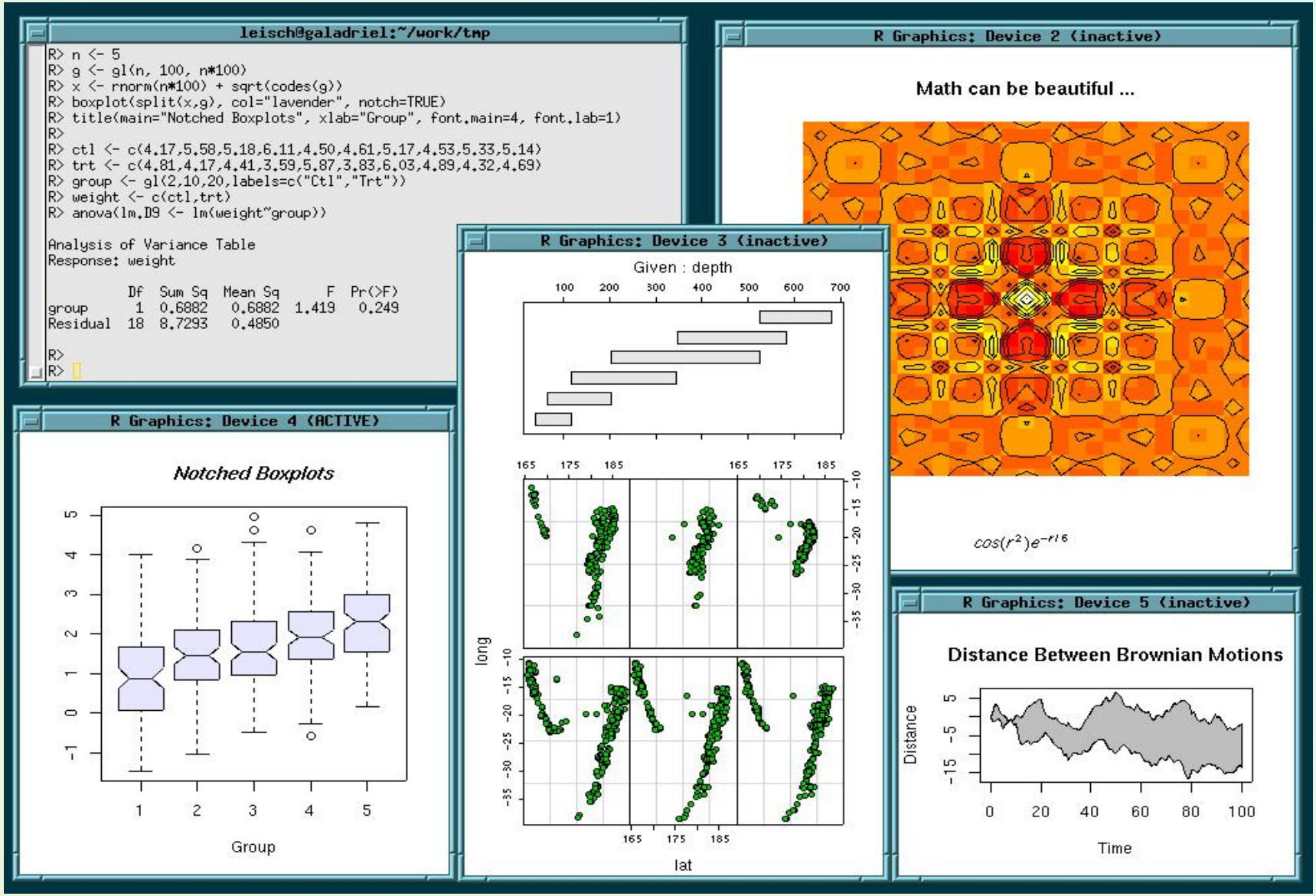
SYSTAT

SIGMAPLOT®

R project


R File Edit Format Workspace Packages & Data Misc Window Help
Tue 2:14 PM stefano iacus
R Console
Abort Source/Load Quartz History Start X11 Set Colors Authentication Save Open In Editor
/Users/jago
rgl.sr> ylen <- ylim[2] - ylim[1] + 1
rgl.sr> colorlut <- terrain.colors(ylen)
rgl.sr> col <- colorlut[y - ylim[1] + 1]
rgl.sr> rgl.clear()
rgl.sr> rgl.surface(x, z, y, color = col)
R Data Editor
height weight
58 115
59 117
60 120
61 123
62 126
63 129
64 132
65 135
66 139
67 142
68 146
69 150
70 154
71 159
72 164
Quartz (2) - Active
Given : depth
long
R Workspace Browser
Object Type Structure
dati data.frame dim: 20 4
g factor levels: 10
l numeric length: 12
n numeric length: 1
opar list length: 2
pie.sales numeric length: 6
pin numeric length: 2
scale numeric length: 1
usr numeric length: 4
women data.frame dim: 15 2
height numeric length: 15
weight numeric length: 15
x numeric length: 87
Refresh List
BoxDens=function(data, npts = 200., x = c(0.,
add = TRUE, col = 11., border=FALSE,collin
"paysage",
{
dens <- density(data, n = npts)
dx <- dens$x
dy <- dens$y
if(add == FALSE)
plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
ylab = "")
if(orientation == "paysage") {
dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
x[1.]
dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
y[1.]
seqbelow <- rep(y[1.], length(dx))
if(Fill == T)
confshade(dx2, seqbelow, dy2, col = col)
if (border==TRUE) points(dx2, dy2, type = "l", col = c
}
else {
dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
RGL device 1 (active)
R Package Manager
Refresh List
status Package Description
loaded graphics The R Graphics Package
not loaded grid The Grid Graphics Package
not loaded lattice Lattice Graphics
loaded methods Formal Methods and Classes
The R Graphics Package
Documentation for package `graphics' version 2.0.0
Help Pages
A B C D E F G H I L M N P R S T X

R project


leisch@galadriel:~/work/tmp
R> n <- 5
R> g <- gl(n, 100, n*100)
R> x <- rnorm(n*100) + sqrt(codes(g))
R> boxplot(split(x,g), col="lavender", notch=TRUE)
R> title(main="Notched Boxplots", xlab="Group", font.main=4, font.lab=1)
R>
R> ctl <- c(4.17,5.58,5.18,6.11,4.50,4.61,5.17,4.53,5.33,5.14)
R> trt <- c(4.81,4.17,4.41,3.59,5.87,3.83,6.03,4.89,4.32,4.69)
R> group <- gl(2,10,20,labels=c("Ctl","Trt"))
R> weight <- c(ctl,trt)
R> anova(lm.D9 <- lm(weight~group))
Analysis of Variance Table
Response: weight
Df Sum Sq Mean Sq F Pr(>F)
group 1 0.6882 0.6882 1.419 0.249
Residual 18 8.7293 0.4850
R>
R>
R Graphics: Device 2 (inactive)
Math can be beautiful ...
$cos(r^2)e^{-r/6}$
R Graphics: Device 3 (inactive)
Given : depth
long
lat
R Graphics: Device 4 (ACTIVE)
Notched Boxplots
Group
R Graphics: Device 5 (inactive)
Distance Between Brownian Motions
Distance
Time

Dlaczego poszukiwac nowych metod?

Za R. E. Abdel-Aal „Data mining and medical informatics”

- Manual hypothesis testing:

  Not practical with large numbers of variables

- User-driven... User specifies variables, functional form and type of interaction:

  User intervention may influence resulting models

- Assumptions on linearity, probability distribution, etc.

  May not be valid

- Datasets collected with statistical analysis in mind

  Not always the case in practice

Rozwój środków obliczeniowych

- Tańsze, pojemniejsze i szybsze pamięci
- Technologicznie można gromadzić TB danych
- Dojrzały stan systemów zarządzania bazami danych
- Rozwój metod (automatycznego) pozyskiwania danych
- Coraz większą moc obliczeniowa

Sztuczna inteligencja

- SI. jest nauk o maszynach realizujących zadania, które wymagają inteligencji wówczas, gdy są wykonywane przez człowieka (Minski)
- SI stanowi dziedzinę informatyki dotyczącą metod i technik wnioskowania symbolicznego przez komputer oraz symbolicznej reprezentacji wiedzy stosowanej podczas takiego wnioskowania (Feigenbaum).
- Sztuczna Inteligencja, to dziedzina informatyki zajmująca się rozwiązywaniem zadań efektywnie „niealgorytmizowalnych” w oparciu o modelowanie wiedzy.

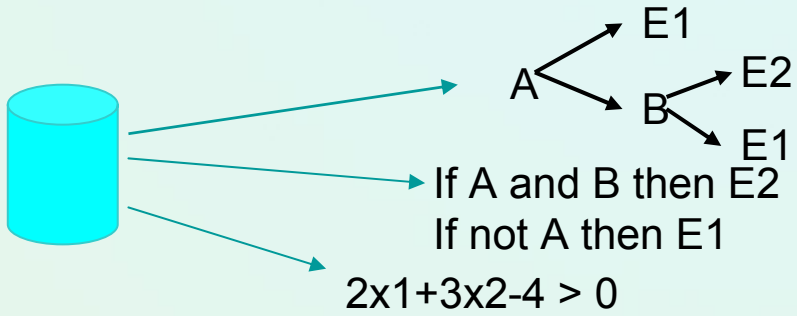
Typowe działy sztucznej inteligencji

- Programy i maszyny grające (teoria gier)
- Systemy eksperckie (ekspertowe)
- Rozpoznawanie języka naturalnego
- Procesy percepcji
- Uczenie maszynowe
- Inteligentne przeszukiwanie danych
- Robotyka
- tzw. sztuczne życie
- Dowodzenie twierdzeń matematycznych

Zaawansowane dziedziny informatyki

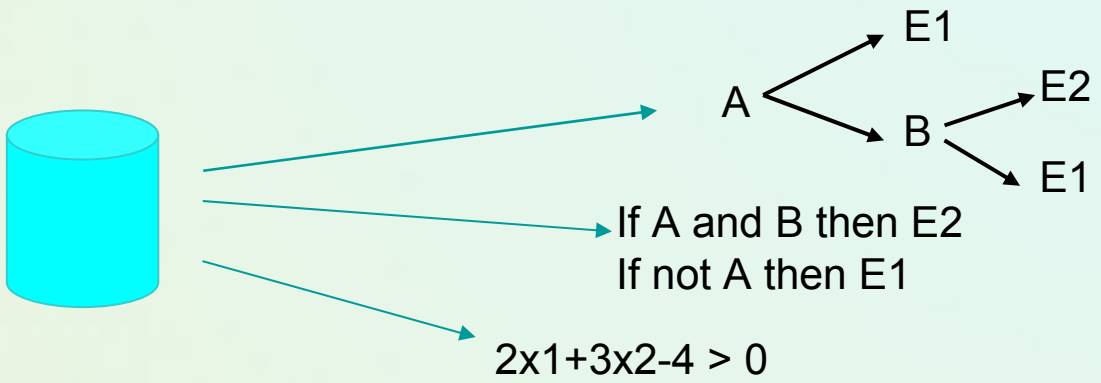
- Uczenie maszynowe → (ang. Machine Learning)
- Systemy, które doskonalą swoje działanie na podstawie

DARPA LAGR Herminator



- Eksploracja danych → Data mining
  - Poszukiwanie w zgromadzonych danych nieznanych, użytecznych regularności, związków między elementami danych.
  - Potencjalnie duże / złożone repozytoria danych
- Odkrywanie wiedzy.

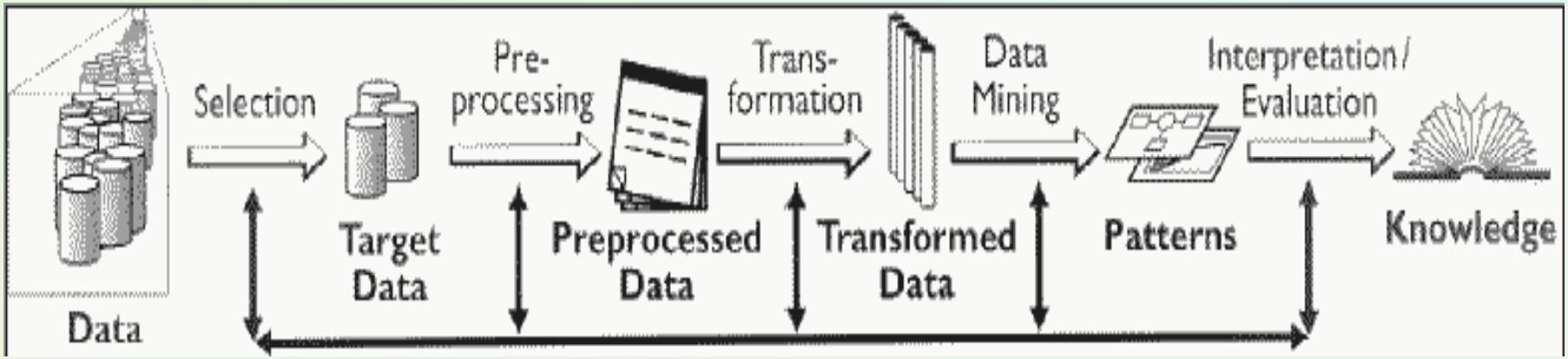
Eksploracja danych



- Data mining (ang.)
  - Poszukiwanie w zgromadzonych danych nieznanych, użytecznych regularności, związków między elementami danych.\
- Eksploracja danych to etap w procesie odkrywania wiedzy.
- Wiedza $\rightarrow$ uporządkowana i formalna reprezentacja odkrytej regularności między elementami danych

Odkrywanie Wiedzy w Bazach Danych

Def: „Nietrywialny proces poszukiwania nowych, użytecznych i zrozumiałych wzorców (regularności) z danych” [Piatetsky]



Przetwarzanie wstępne:

Selekcja – wybór podzbioru przykładów i atrybutów

Pre-processing – wstępne przetwarzanie, czyszczenie danych, standaryzacja

Transformacje – do postaci akceptowanej przez metody eksploracji danych

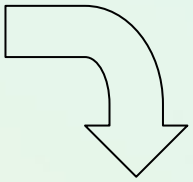
Wiedza klasyfikacyjna

- Problem określania zasad przydziału obiektów do znanych wstępnie klas na podstawie analizy danych o przykładach klasyfikacji.

Wiek	Zawód	dochód	…	Decyzja
21	Prac. fiz.	1220	…	Nie kupi
26	Menedżer	2900	…	Kupuje
44	Inżynier	2600	…	Kupuje
23	Student	1100	…	Kupuje
56	Nauczyciel	1700	…	Nie kupi
…	…	…	…	…
45	Lekarz	2200	…	Nie kupi
25	Student	800	…	Kupuje

Przykłady uczące

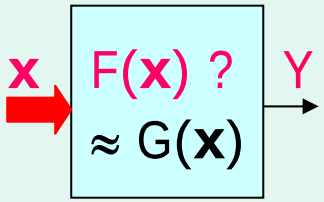
Algorytm eksploracji



Reprezentacja wiedzy:
np. reguły
R1. Jeżeli student to kupuje komputer
R2. Jeżeli dochód > 2400 …

Klasyfikacja nadzorowana - Supervised Learning

- Y=F(**x**): true function (usually not known) for population P
- Przykłady etykietowane <**x**,Y>
- 1. Collect Data: "labeled" training sample drawn from P



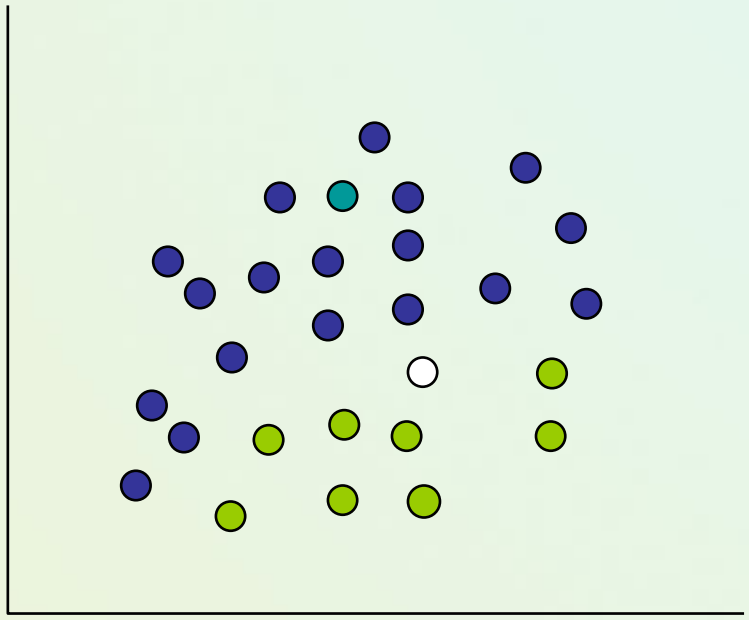
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0 0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0 1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0 0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0 1

- 2. Training: Get G(**x**); model learned from training sample,
  Goal: $E<(F(\mathbf{x})-G(\mathbf{x}))^2> \approx 0$ for future samples drawn from P – Not just data fitting!
- 3. Test/Use:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0 ?

Klasyfikowanie

Odkryj z danych historycznych metody przydziału przykładów do klas. Przykłady uczące - są opisane zbiorem atrybutów i etykietowane



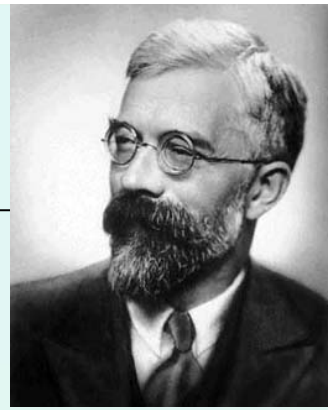
Wiele podejść :
Analiza dyskryminacyjna
Klasyfikatory Bayesowskie,
Sieci neuronowe,
Drzewa i reguły decyzyjne,
K-NN
...

Dane przykłady uczące z klas ●●
Określ przydział dla nowego obiektu ○?

Klasyfikacja – podejście statystyczne

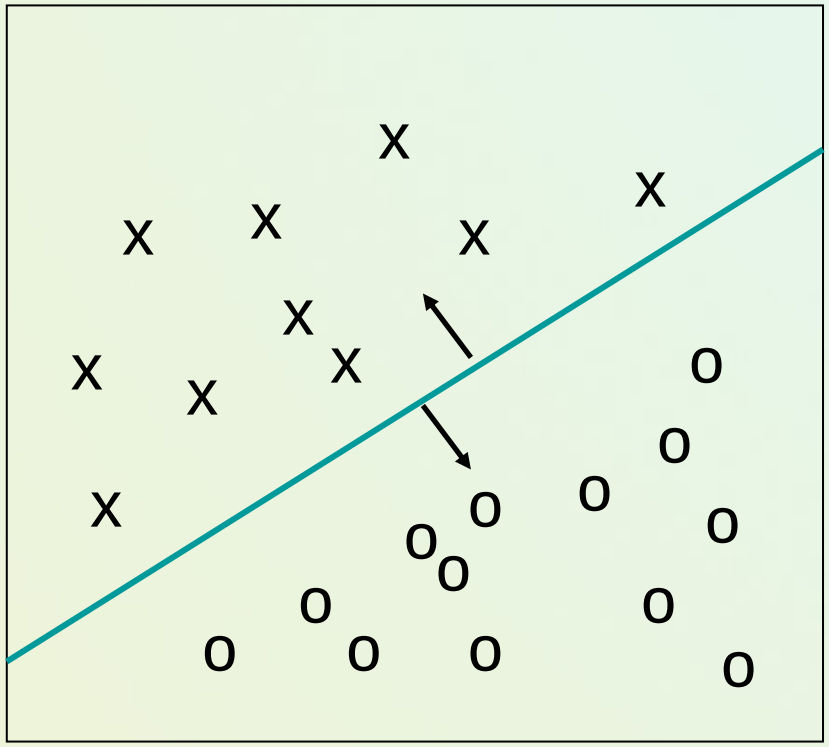
$$D = \left\{ (\mathbf{x}_i, c_i) \,|\, x_i \in R^p, c_i \in \{C_1, \ldots C_k\} \right\}_{i=1}^{N}$$

$$y = f(\mathbf{x}, \mathbf{w})$$



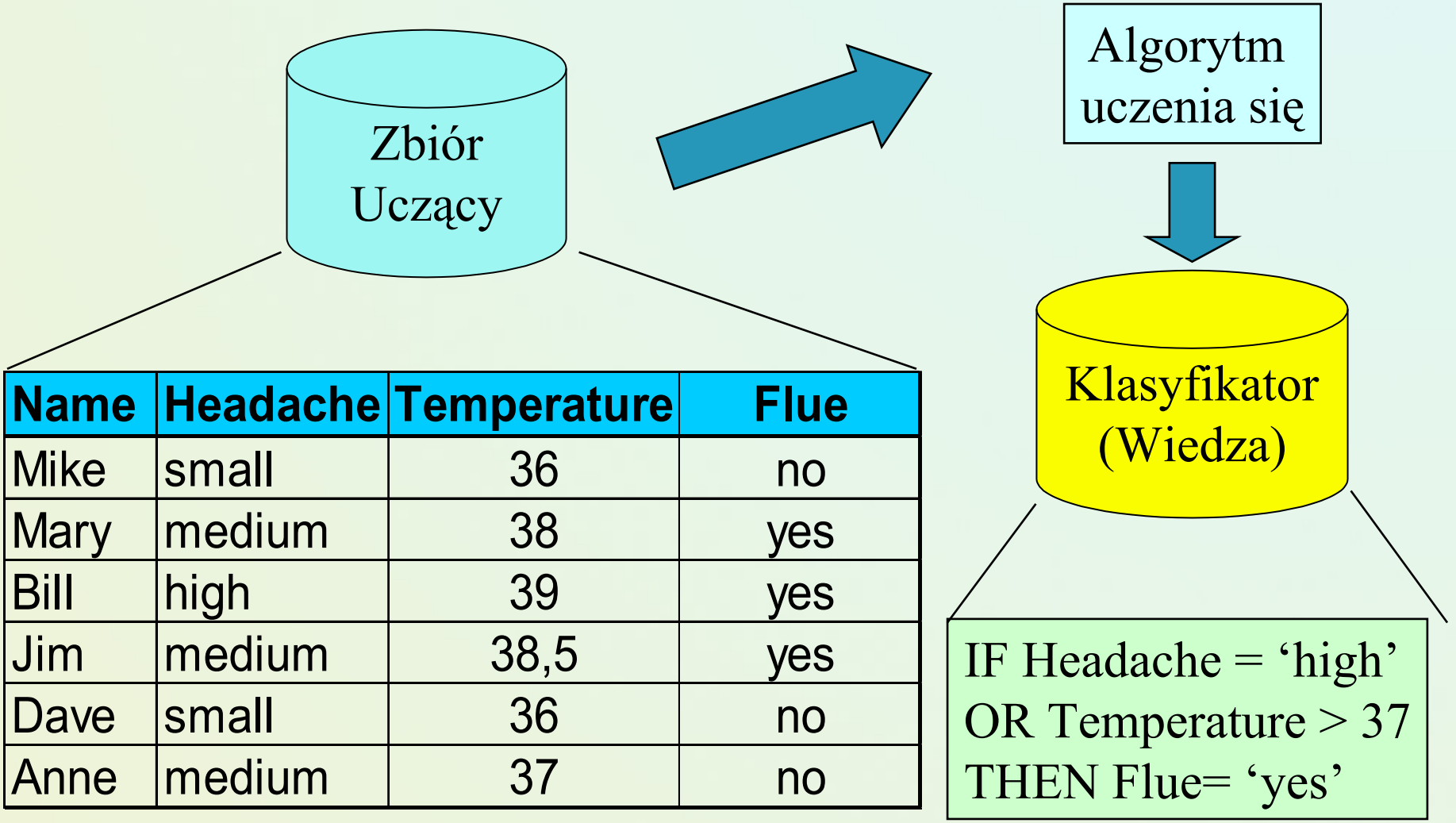
sir Ronald Fisher

$$y = \begin{cases} f(\mathbf{x}_i) > T & \mathbf{x}_i \in C_1 \\ f(\mathbf{x}_i) < T & \mathbf{x}_i \in C_2 \end{cases}$$



- Binarna klasyfikacja (uogólnienie na więcej klas)
- Poszukiwanie przybliżenia „granicy decyzyjnej” – ang. decision boundary
- Obserwacje ponad linią przydziel do klasy ‘x’
- Obserwacje pod linią przydziel do klasy ‘o’
- Przykłady: Fisher-owska analiza dyskryminacyjna, SVM, ANN

Poszukiwanie wiedzy klasyfikacyjnej



Name	Headache	Temperature	Flue
Mike	small	36	no
Mary	medium	38	yes
Bill	high	39	yes
Jim	medium	38,5	yes
Dave	small	36	no
Anne	medium	37	no

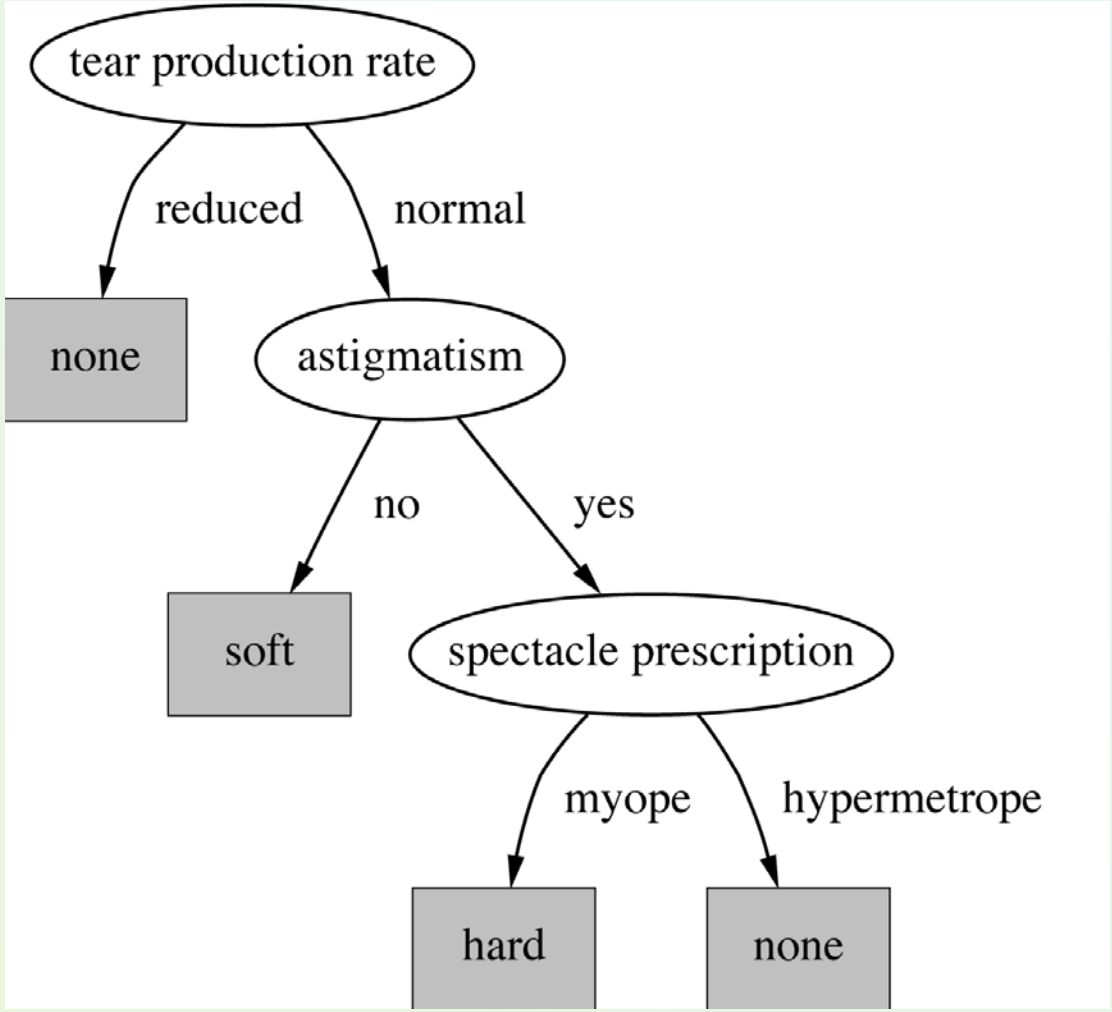
Poszukiwanie wiedzy klasyfikacyjnej (2)

- Uczenie nadzorowane – odkrywanie wiedzy klasyfikacyjnej
  - Przykłady uczące opisane etykietą wskazująca klasę decyzyjną; Na zbiorze uczącym poszukuje się reprezentacji wiedzy klasyfikacyjnej
- **Perspektywy odkrywania wiedzy**
  - Predykcja – przewidywanie przydziału nowych obiektów do klas / reprezentacja wiedzy wykorzystywana jako tzw. klasyfikator
  - Opis klasyfikacji obiektów – wyszukiwanie wzorców charakteryzujących właściwości danych i prezentacja ich użytkownikowi w zrozumiałej formie

Reprezentacja danych – tablica danych

- Tabela (U,A∪{d})
- Przykład tzw. contact lenses / dobór szkieł kontatkowych:
- Atrybuty:
  - age {young, pre-presbyopic, presbyopic}
  - spectacle-prescrip {myope, hypermetrope}
  - astigmatism {no, yes}
  - tear-prod-rate {reduced, normal}
- Decyzja contact-lenses {soft, hard, none}

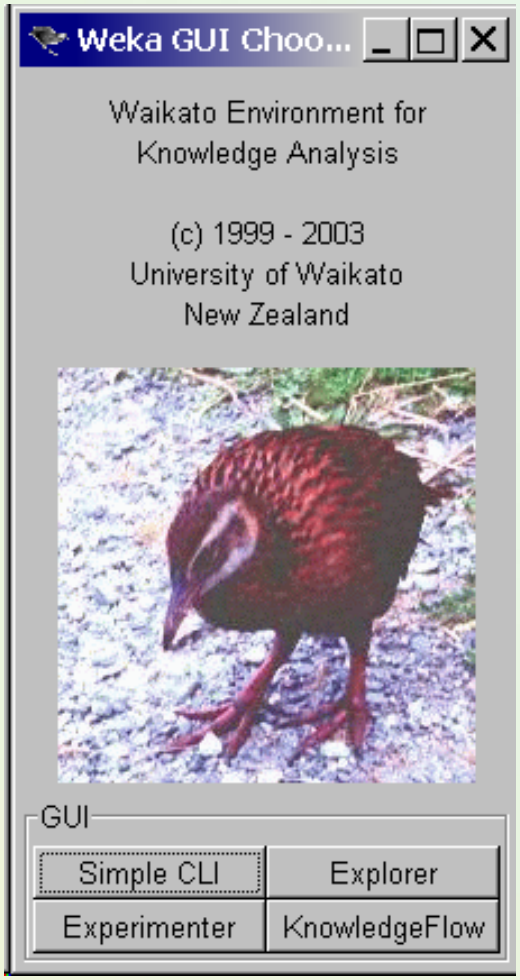
age	specpres	astig	tearprod	contlen
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft
pre-presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	myope	no	reduced	none
presbyopic	myope	no	normal	none
presbyopic	myope	yes	reduced	none
presbyopic	myope	yes	normal	hard
presbyopic	hypermetrope	no	reduced	none
presbyopic	hypermetrope	no	normal	soft
presbyopic	hypermetrope	yes	reduced	none
presbyopic	hypermetrope	yes	normal	none



Decision tree for the contact lens data

Drzewo decyzyjne – drzewo klasyfikacyjne

Weka – software for data mining

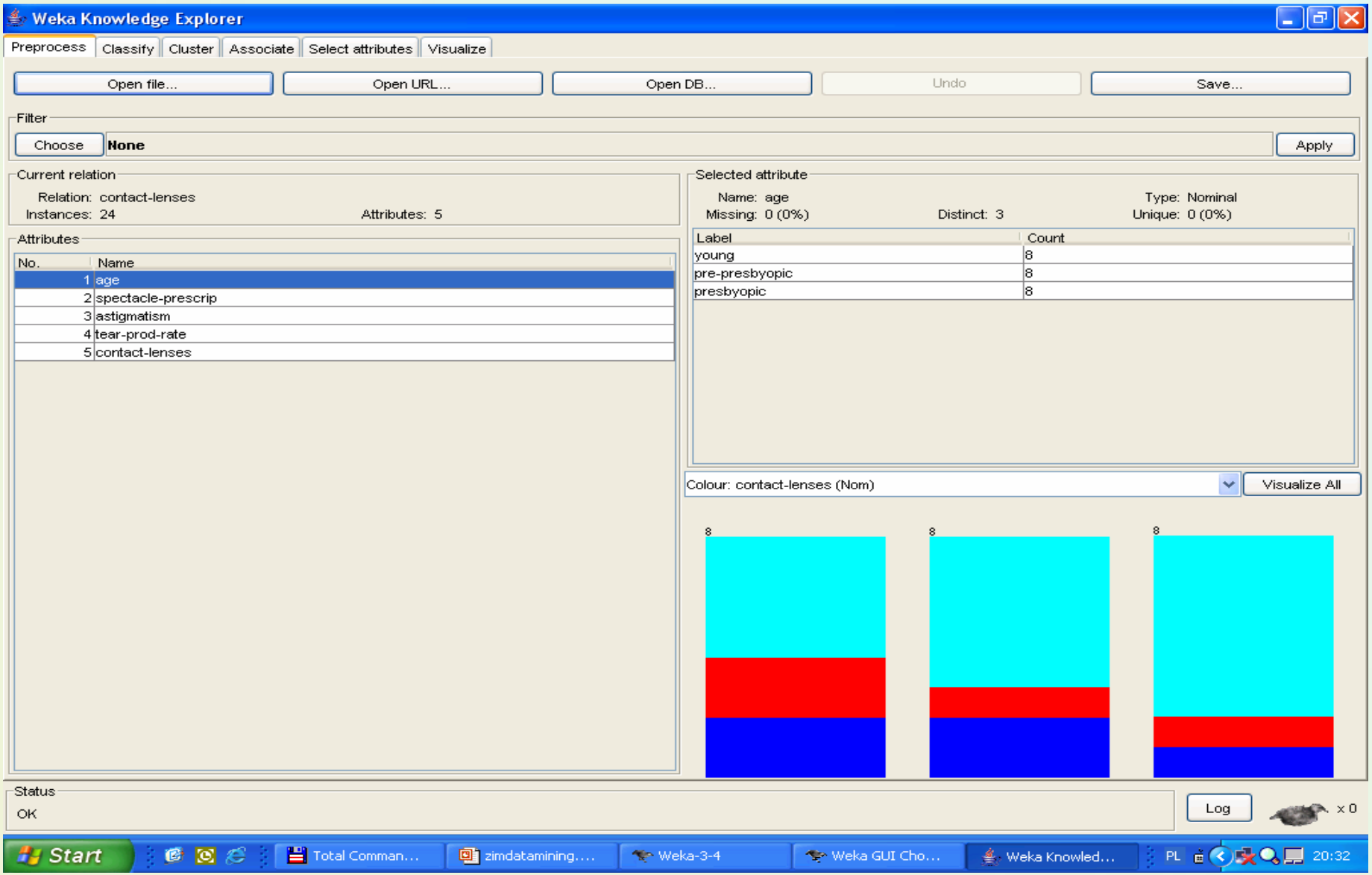


- Waikato Environment for Knowledge Analysis (WEKA); developed by the Department of Computer Science, University of Waikato, New Zealand
- Data mining / Machine learning software written in Java (distributed under the GNU Public License)
- Used for research, education, and applications

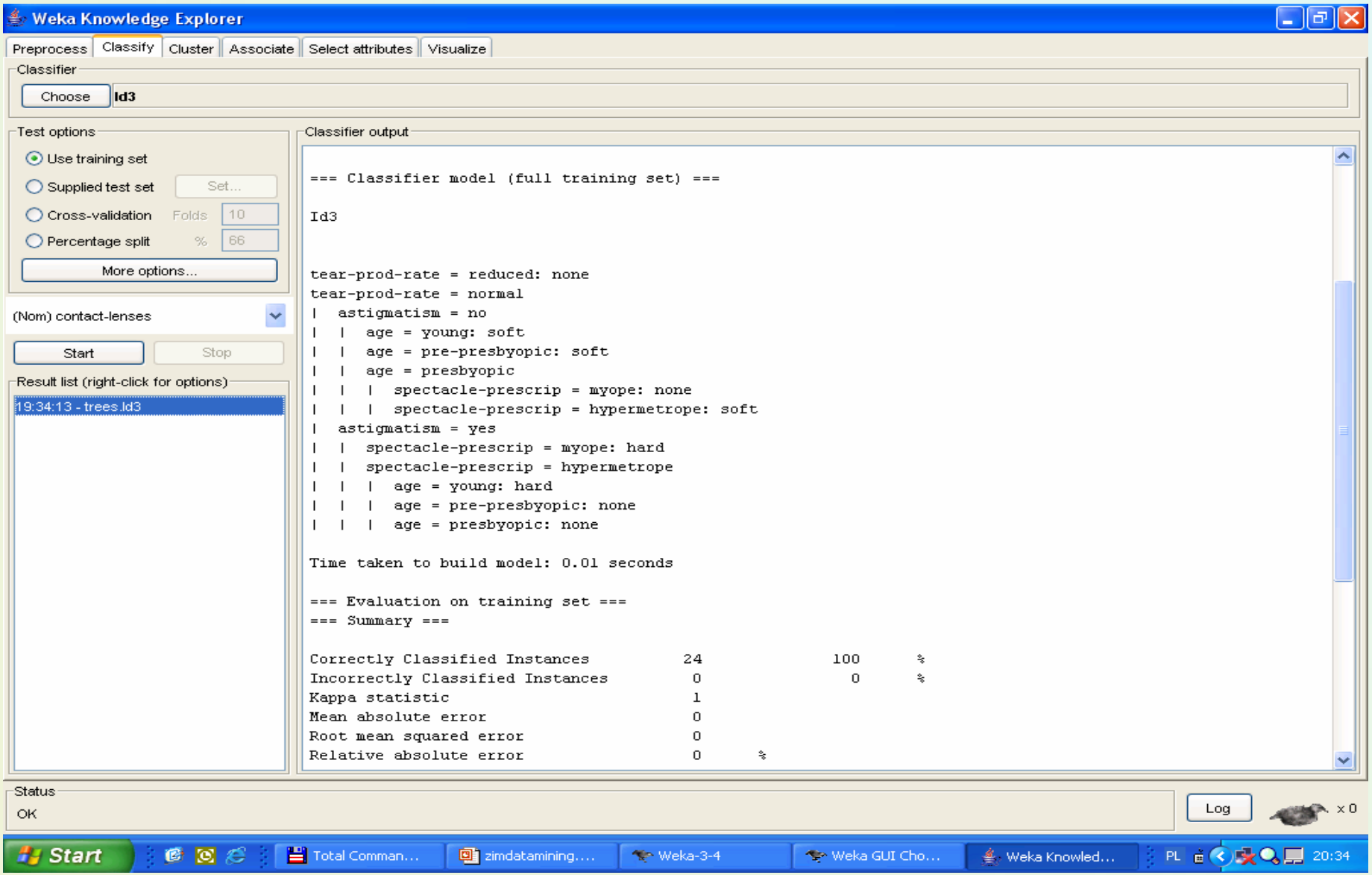
http://www.cs.waikato.ac.nz/ml/weka/

- Ian Witten, Eibe Frank

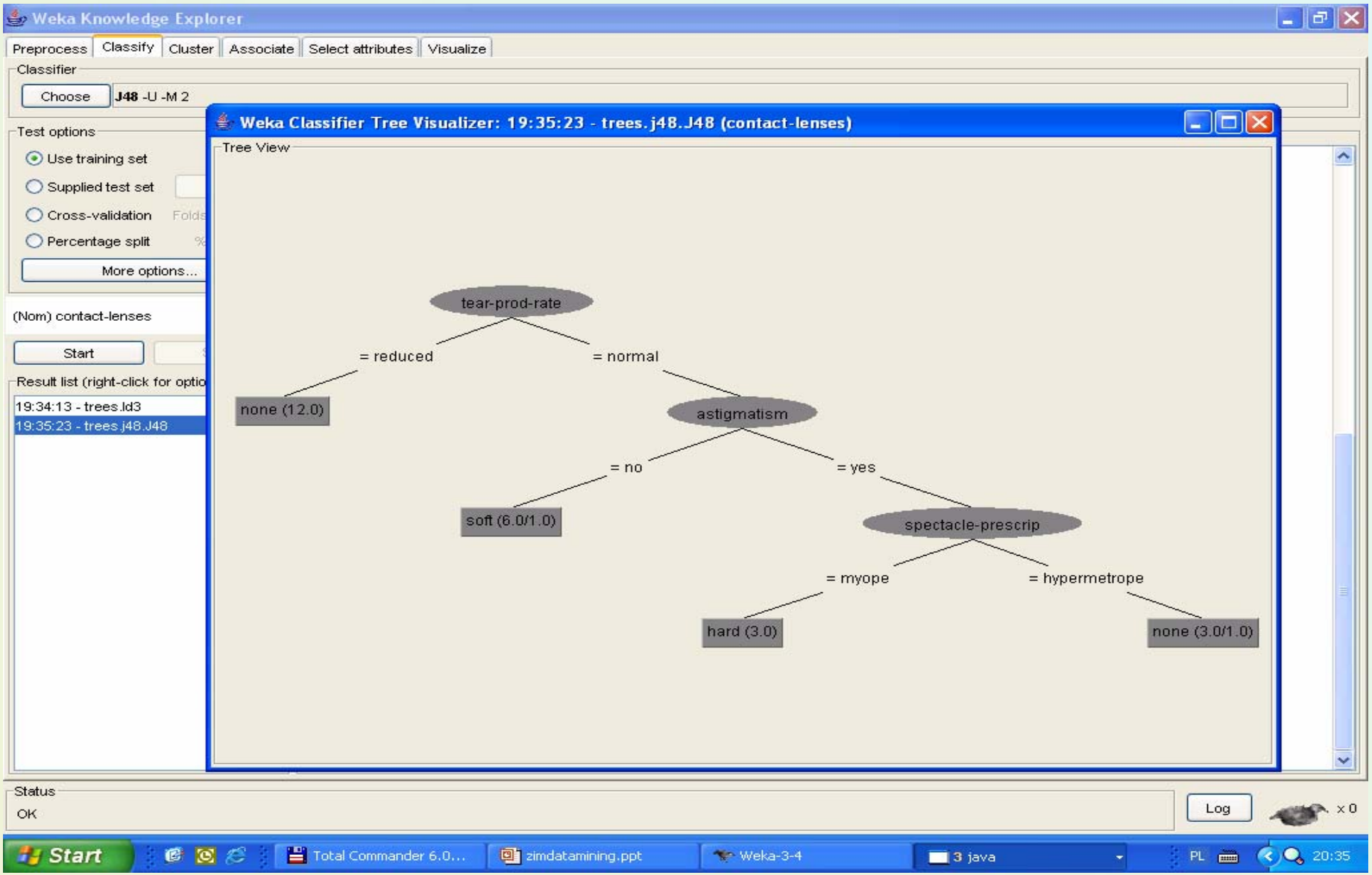
WEKA – analiza pliku contact lenses



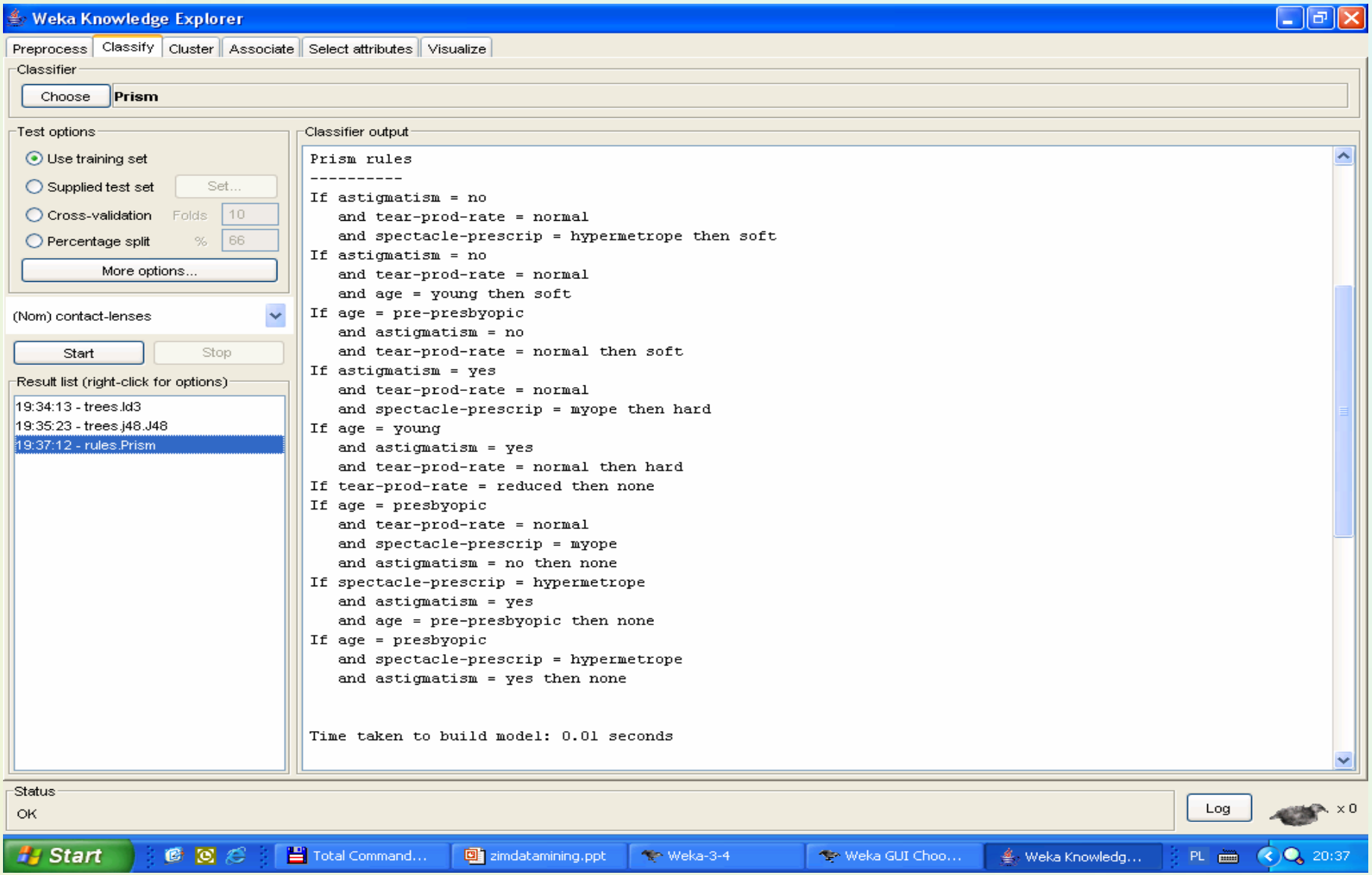
Poszukiwanie drzew ze zbioru przykładów



Widok drzewa



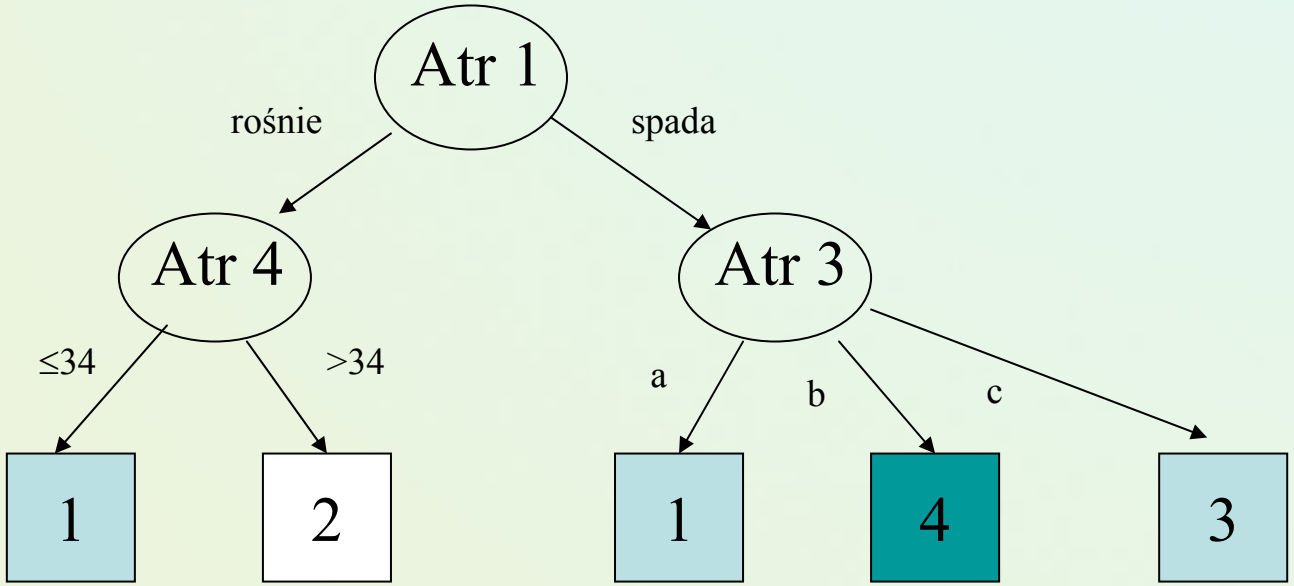
Indukcja reguł – algorytm PRISM



Co to jest drzewo decyzyjne?

Jest to struktura grafu skierowanego z góry na dół:

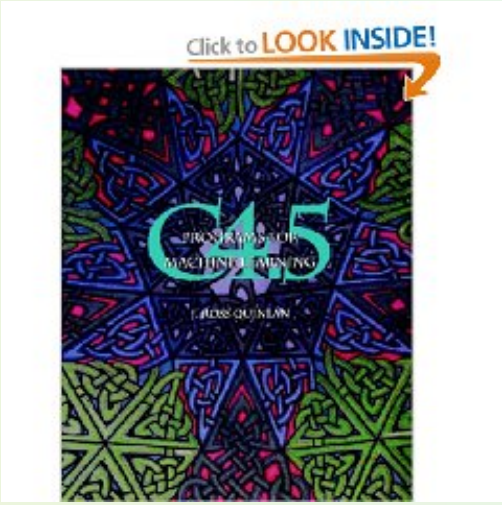
- Węzły reprezentują pytanie o wartości cech
- Z węzłów wychodzą gałęzie które reprezentują wynik pytania
- Liście reprezentują klasy decyzyjne



Metody indukcji drzew decyzyjnych

- Podejście „top-down decision tree generation” obejmuje dwa etapy:
  - Konstrukcja drzewa
    - Na początku wszystkie przykłady w węźle.
    - Rekurencyjnie dziel przykłady w oparciu o wybrane testy na wartościach atrybutu (kryterium wyboru).
  - Upraszczanie drzewa - „Tree pruning”
    - Usuwanie poddrzew, które mogą prowadzić do błędnych decyzji podczas klasyfikacji przypadków testowych („noisy data”, przespecjalizowane opisy …)
  - Przykłady algorytmów: ID3, C4.5, CART,…

J.Ross Quinlan – twórca alg. indukcji drzew

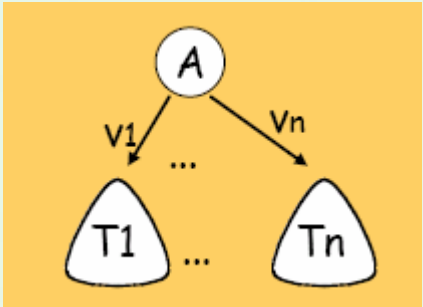


Ross Quinlan completed a PhD in Computer Science at the University of Washington in 1968. He has developed several algorithms used in machine learning and data mining such as ID3, C4.5, FOIL, and more recent commercial systems such as See5 and Cubist. He has held permanent appointments at the University of Sydney, University of Technology Sydney, Rand Corporation, and visiting appointments at Carnegie-Mellon University, MIT, GTE, and Stanford University. He currently heads a small data mining tools company and is an Adjunct Professor at the University of New South Wales. He is a Fellow of the American Association for Artificial Intelligence and the Australian Computer Society.

- Więcej – spójrx http://www.rulequest.com/Personal/
- Także, http://en.wikipedia.org/wiki/Ross_Quinlan

Basic TDIDT algorithm (simplified Quinlan's ID3)

- At start, all training examples *S* are at the root.
- **If** all examples from S belong to the same class Kj
  then label the root with Kj
  else
  - select the „best" attribute A
  - divide S into S1, …, Sn according to values v1, …, vn of attribute A
  - Recursively build subtrees T1, …, Tn for S1, …,Sn

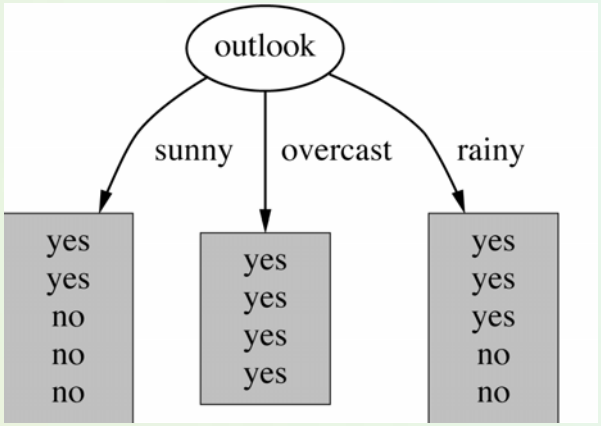


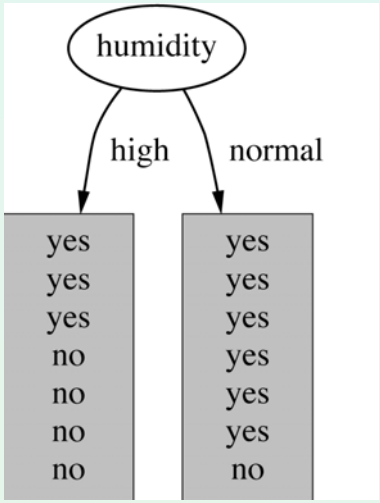
Przykład budowy – Quinlan example „golf”

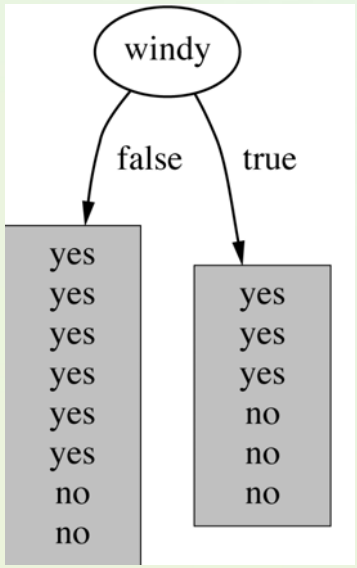
Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

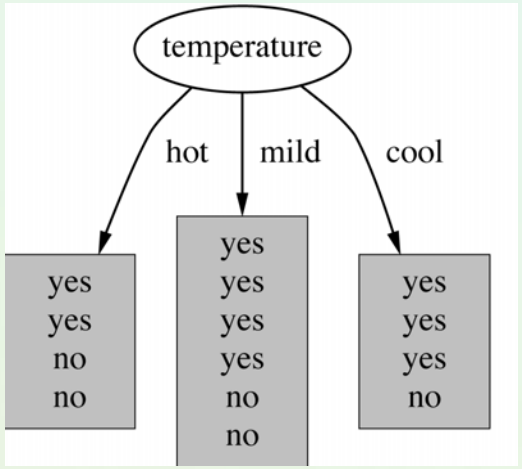
Uproszczona Tabela danych

Który atrybut należy wybrać?



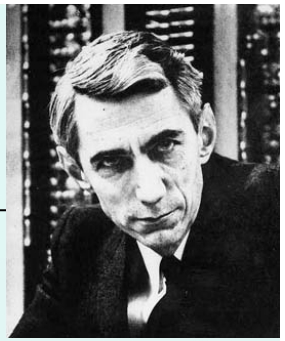






witten&eibe

Entropia informacji



- Entropy information (wprowadził C. Shannon)
  - Given a probability distribution, the info required to predict an event is the distribution's *entropy*
  - Entropy gives the information required in bits (this can involve fractions of bits!)
  - The amount of information, needed to decide if an arbitrary example in *S* belongs to class K_j (p_j - prob. it belongs to K_j).
- Basic formula for computing the entropy for examples in *S*:

$$\text{entropy}(S) = -p_1 \log p_1 - p_2 \log p_2 \ldots - p_n \log p_n$$

- A conditional entropy for splitting examples *S* into subsets *Si* by using an attribute A:

$$entropy(S \mid A) = \sum_{i=1}^{m} \frac{|S_i|}{|S|} \cdot entropy(S)$$

- Choose the attribute with the maximal info gain:

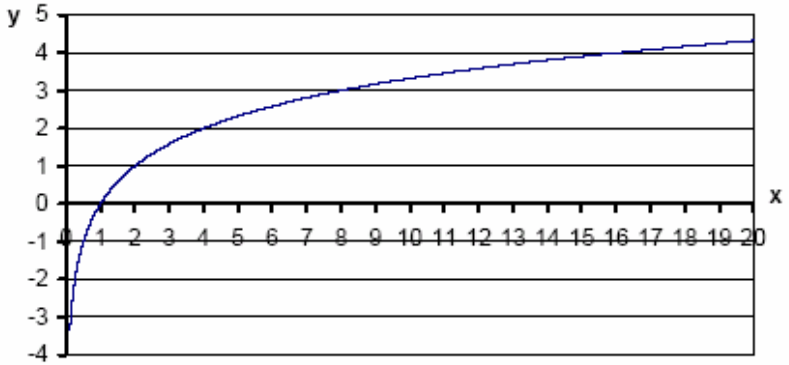
$$entropy(S) - entropy(S \mid A)$$

Funkcja logarytmiczna

$$y = \log_a x$$

a – podstawa logarytmu $x = a^y$

Rozważmy funkcję logarytmiczną dla $a = 2$ (tj. $y = \log_2 x$)



x	1/8	1/4	1/2	1	2	4	8
y	-3	-2	-1	0	1	2	3

- Co zrobić, gdy p=0
- Jak wygląda wykres entropii dla klasyfikacji binarnej

Entropia dla przykładu golf

Nie oceniamy podziału atrybutem, tylko rozkład wartości klas decyzyjnych

Dwie klasy : *yes* and *no*

Z 14 przykładów 9 etykietowanych jako *yes*, reszta jako *no*

$$p_{yes} = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) = 0.41$$

$$p_{no} = -\left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right) = 0.53$$

$$E(S) = p_{yes} + p_{no} = 0.94$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes

Outlook	Temp.	Humidity	Windy	play
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Przykład oceny atrybutu “Outlook”

- “Outlook” = “Sunny”:

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5\log(2/5) - 3/5\log(3/5) = 0.971$$

- “Outlook” = “Overcast”:

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1\log(1) - 0\log(0) = 0$$

Note: log(0) is not defined, but we evaluate 0*log(0) as zero

- “Outlook” = “Rainy”:

$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5\log(3/5) - 2/5\log(2/5) = 0.971$$

- Entropia warunkowa dla podziału wartościami atrybutu

$$\text{info}([3,2],[4,0],[3,2]) = (5/14)\times 0.971 + (4/14)\times 0 + (5/14)\times 0.971$$
$$= 0.693$$

Obliczanie zysku informacyjnego miary entropii

- Information gain:

(information before split) – (information after split)

$$\text{gain("Outlook")} = \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693$$
$$= 0.247$$

- Ostateczne wartości zysku

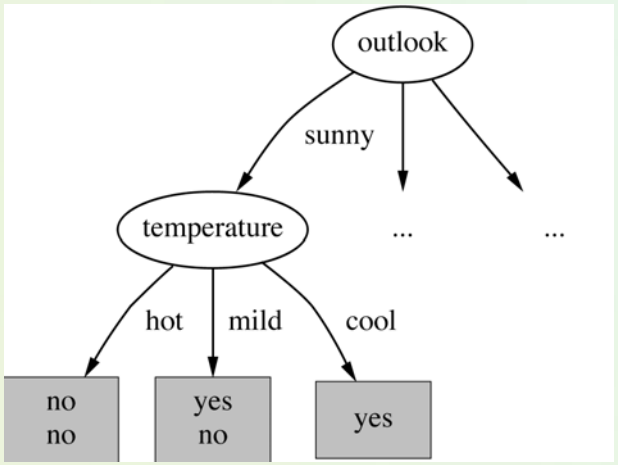
$$\text{gain("Outlook")} = 0.247$$

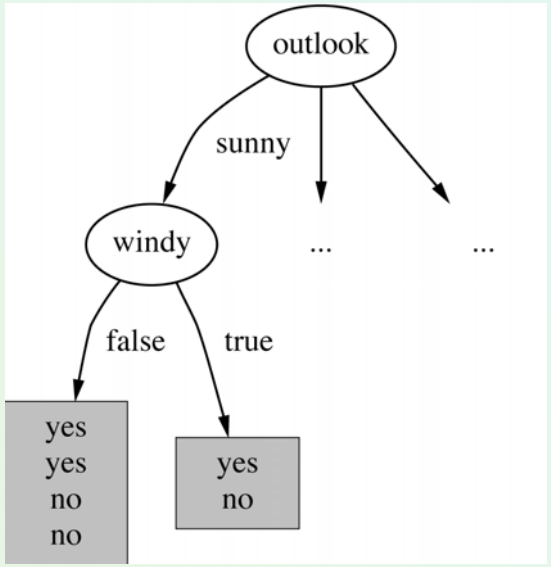
$$\text{gain("Temperature")} = 0.029$$

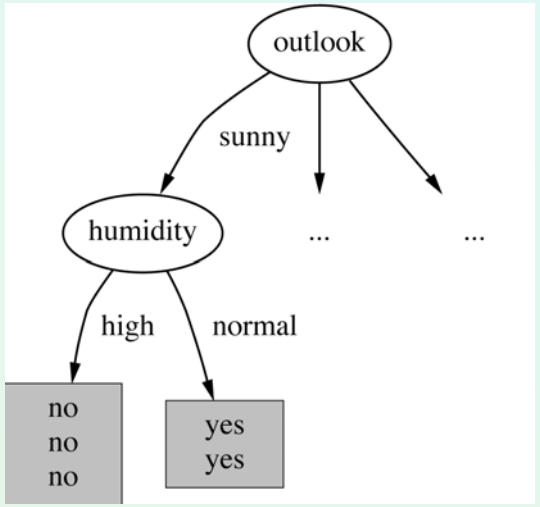
$$\text{gain("Humidity")} = 0.152$$

$$\text{gain("Windy")} = 0.048$$

Dalsze obliczenia





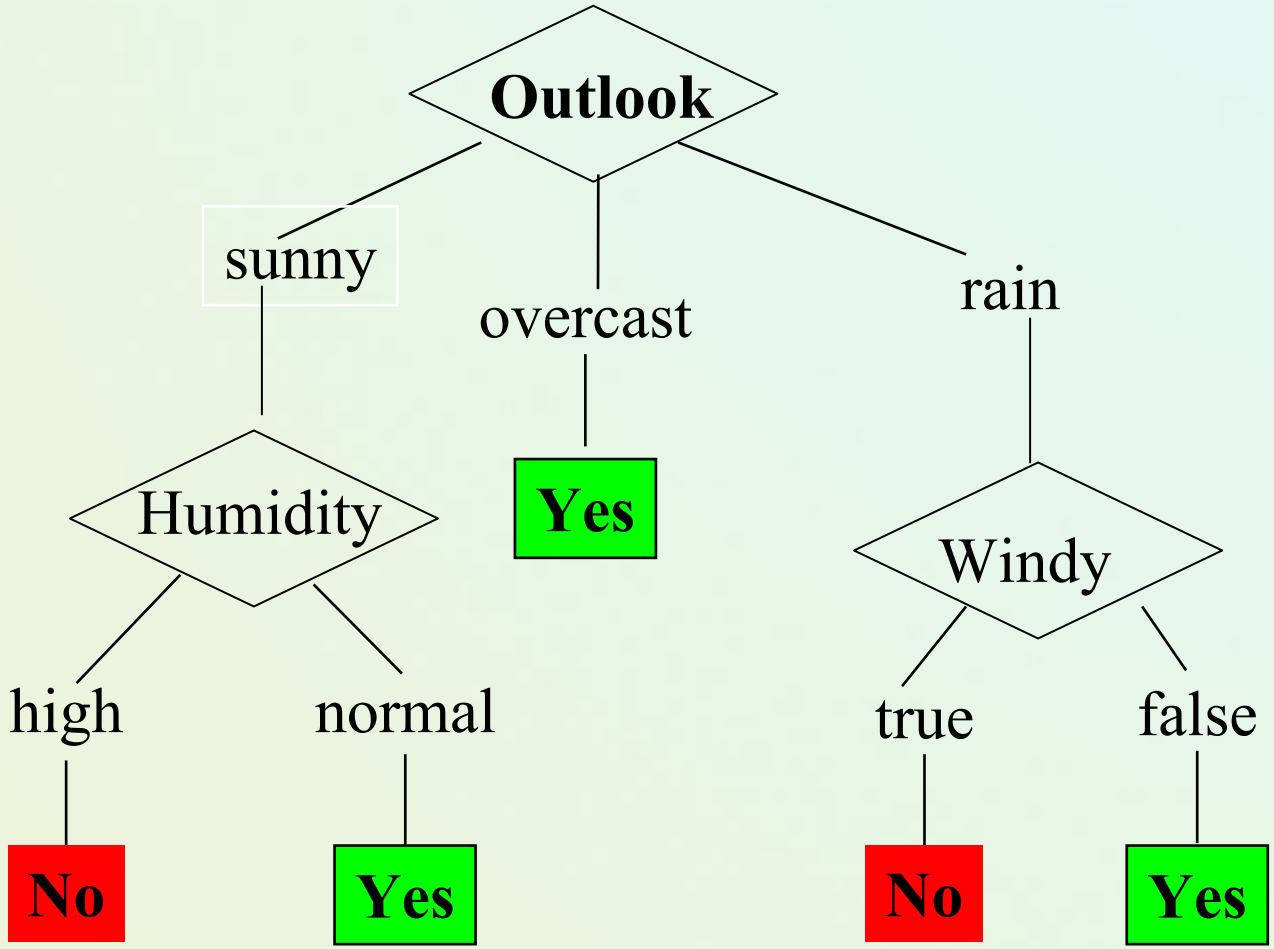


$$gain("Humidity") = 0.971$$

$$gain("Temperature") = 0.571$$

$$gain("Windy") = 0.020$$

Ostateczne drzewo

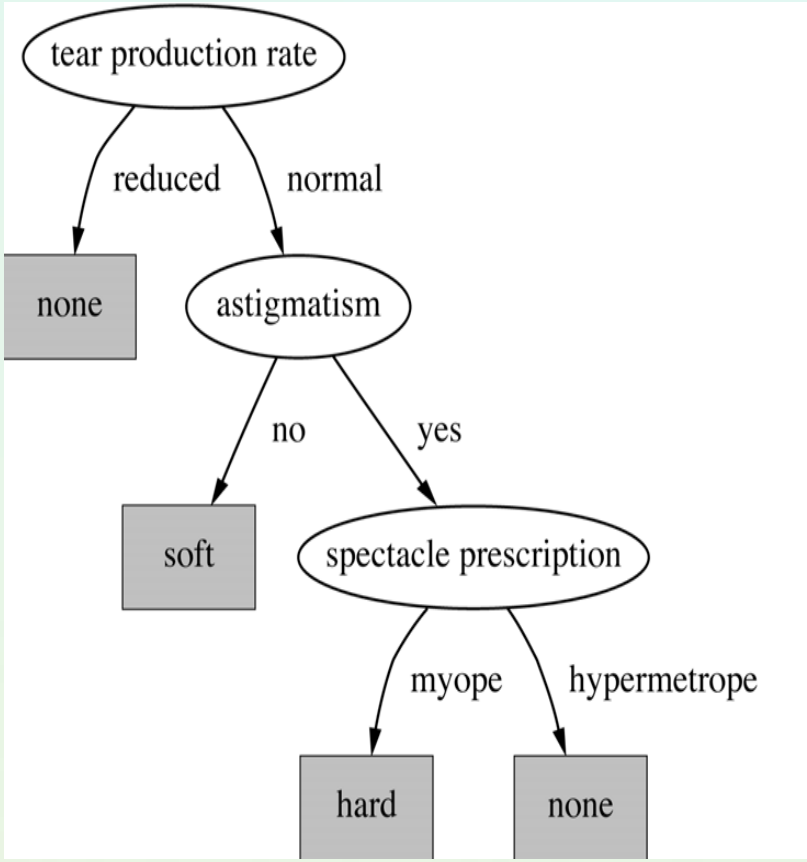


Podstawowe pojęcia w indukcji DT

- Kryterium podziału: *Goodness function*
  - Stosowana, aby wybrać atrybut do tworzenia węzła
  - Różne funkcje są używane:
    - information gain (entropia)
    - Gini index
- Tworzenie gałęzi drzewa:
  - Ustalanie gałęzi, do której przydzielamy podzbiór.
  - binary versus *k*-ary splitting.
- Decyzja kiedy zatrzymać rozbudowę drzewa: impurity measure.
- Tworzenie liści: przypisanie etykiety klasy większościowej.

Wykorzystanie drzewa

- Bezpośrednio:
  - sprawdzaj wartości atrybutu nowego przykładu zaczynając od korzenia do liści
- Pośrednio:
  - zamień strukturę drzewa na zbiór reguł decyzyjnych (upraszczając nadmiarowe warunki)
  - reguły uważa się za czytelniejszą reprezentacje



Ograniczenia w uczeniu się drzew decyzyjnych

Pytania i problemy, np.:

Kiedy należy zaprzestać rozbudowywać drzewa?

aby zapobiec przespecjalizowaniu opisu

duże drzewa są trudne do analizy i zrozumi-enia

Jak uwzględniać atrybuty ilościowe?

Jak uwzględniać atrybuty ze zbyt dużą liczbą wartości w stosunku do dziedzin pozostałych atrybutów?

Jak uwzględniać atrybuty z nieznanymi wartościami?

Jak uwzględniać dane "zaszumione"?

Binary Tree – budowa drzew binarnych

- Drzewa binarne mogę być skuteczniejsze w klasyfikacji nowych faktów
- Podział binarny w węźle drzewa:
  - Atrybuty liczbowe A, reprezentacja w postaci value(A)<x gdzie x jest wartością z dziedziny A.
  - Atrybuty nieliczbowe A, warunek w postaci value(A)$\in$X gdzie X$\subset$domain(A)

Drzewo binarne (Quinlan's C4.5 output)

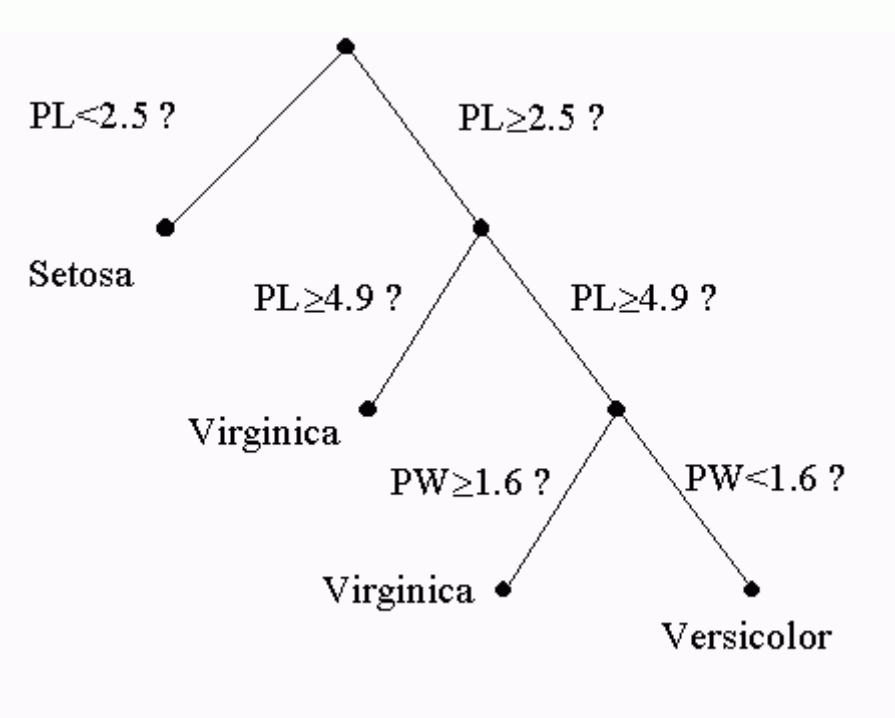
```
Pruned decision tree:

A9 = t:
|   A15 > 228 : + (106.0/3.8)
|   A15 <= 228 :
|   |   A14 <= 102 :
|   |   |   A4 in {l,t}: + (0.0)
|   |   |   A4 = u:
|   |   |   |   A6 in {c,d,cc,i,k,m,q,w,x,e,aa}: + (46.4/3.1)
|   |   |   |   A6 in {j,ff}: - (2.0/1.0)
|   |   |   |   A6 = r: + (0.0)
|   |   |   A4 = y:
|   |   |   |   A6 in {c,i,aa,ff}: - (7.0/3.4)
|   |   |   |   A6 in {d,j,w,x}: + (4.0/1.2)
|   |   |   |   A6 in {cc,k,m,r,q,e}: + (0.0)
|   |   A14 > 102 :
|   |   |   A6 in {j,r}: + (0.0)
|   |   |   A6 in {c,d,k,m,e,aa,ff}:
|   |   |   |   A14 <= 132 : - (4.1/1.2)
|   |   |   |   A14 > 132 :
|   |   |   |   |   A3 <= 1.625 :
|   |   |   |   |   |   A14 <= 292 : - (13.0/1.3)
|   |   |   |   |   |   A14 > 292 :
|   |   |   |   |   |   |   A13 = g: + (2.0/1.0)
|   |   |   |   |   |   |   A13 = s: - (6.0/2.3)
|   |   |   |   |   |   |   A13 = p: - (0.0)
|   |   |   |   |   A3 > 1.625 :
|   |   |   |   |   |   A6 in {k,m}: + (5.0/1.2)
|   |   |   |   |   |   A6 = ff: + (0.0)
|   |   |   |   |   |   A6 in {c,d,e,aa}:
|   |   |   |   |   |   |   A2 <= 32.08 : + (9.5/4.1)
|   |   |   |   |   |   |   A2 > 32.08 : - (8.0/3.5)
|   |   |   A6 in {cc,i,q,w,x}:
|   |   |   |   A8 <= 10.75 : + (36.0/9.3)
|   |   |   |   A8 > 10.75 : - (2.0/1.0)
A9 = f:
|   A4 in {u,y}: - (237.0/17.3)
|   A4 = l: + (2.0/1.0)
|   A4 = t: - (0.0)
```

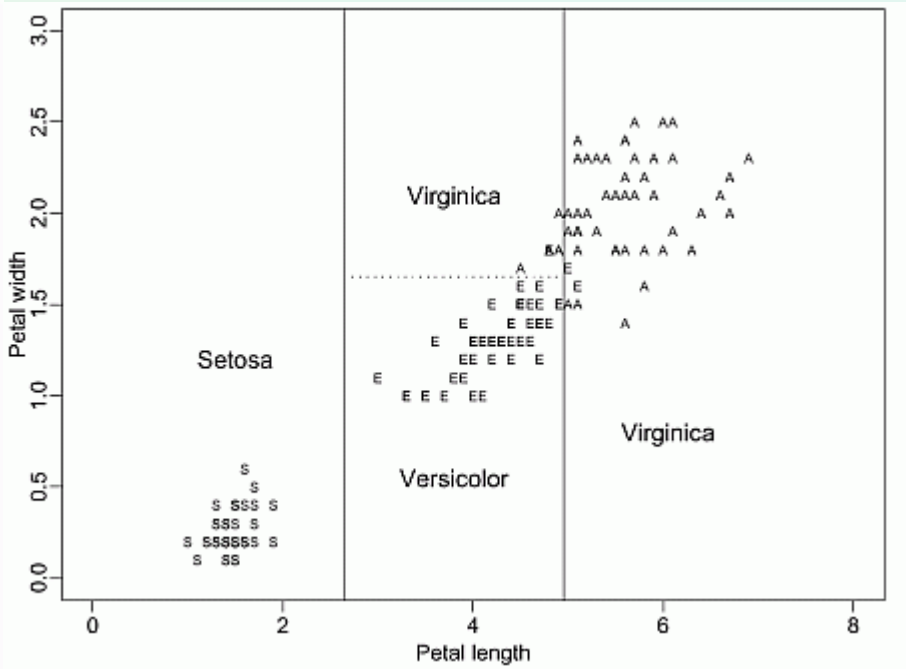
- Crx (Credit Data) UCI ML Repository

Interpretacja graficzna dla atr liczbowych

Hierarchiczny podział na hiper-prostopadłościany

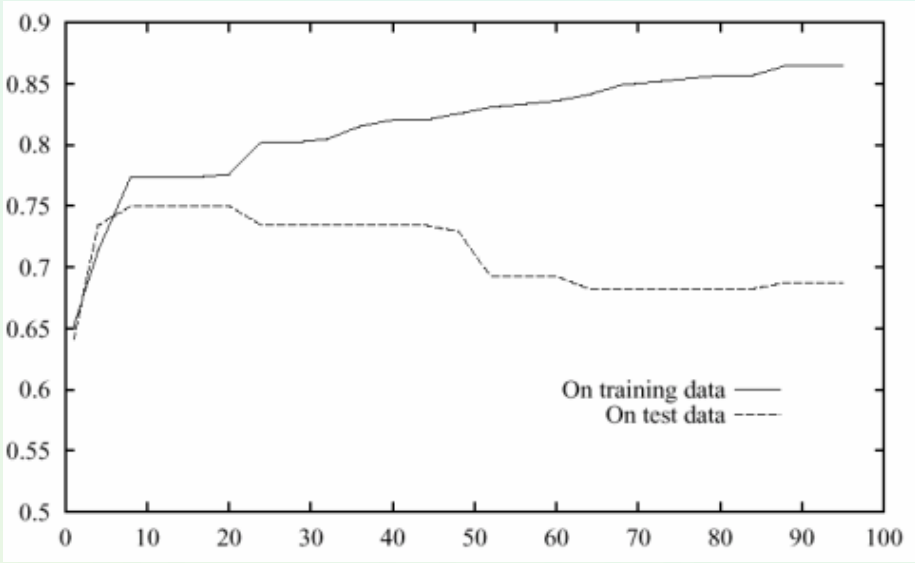
Przykład: Iris flowers data, with 4 features; displayed in 2-D





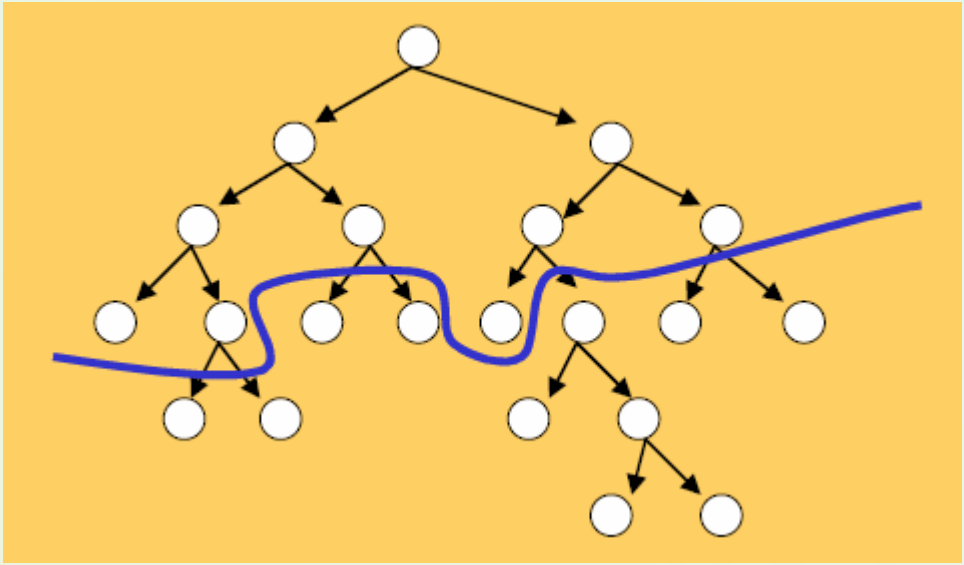
Overfitting the Data – nadmierne dopasowanie do danych uczących

- Podstawowy algorytm ID3 $\rightarrow$ Rozbudowuj gałąź drzewa do pełnego rozróżnienia przykładów
  - Sensowe na spójnych przykładów i celów dokładnego opisu
- Rzeczywiste dane (niespójne, szum informacyjny) oraz cel klasyfikowania przykładów
  - Drzewa mają tendencje do przeuczenia / nadmiernego dopasowania do specyficznych przykładów *overfit* the learning examples
  - Occam razor – zasada brzytwy Occama (z konkurujących drzew wybierz prostsze; ma lepsze własności generealizacyjne)



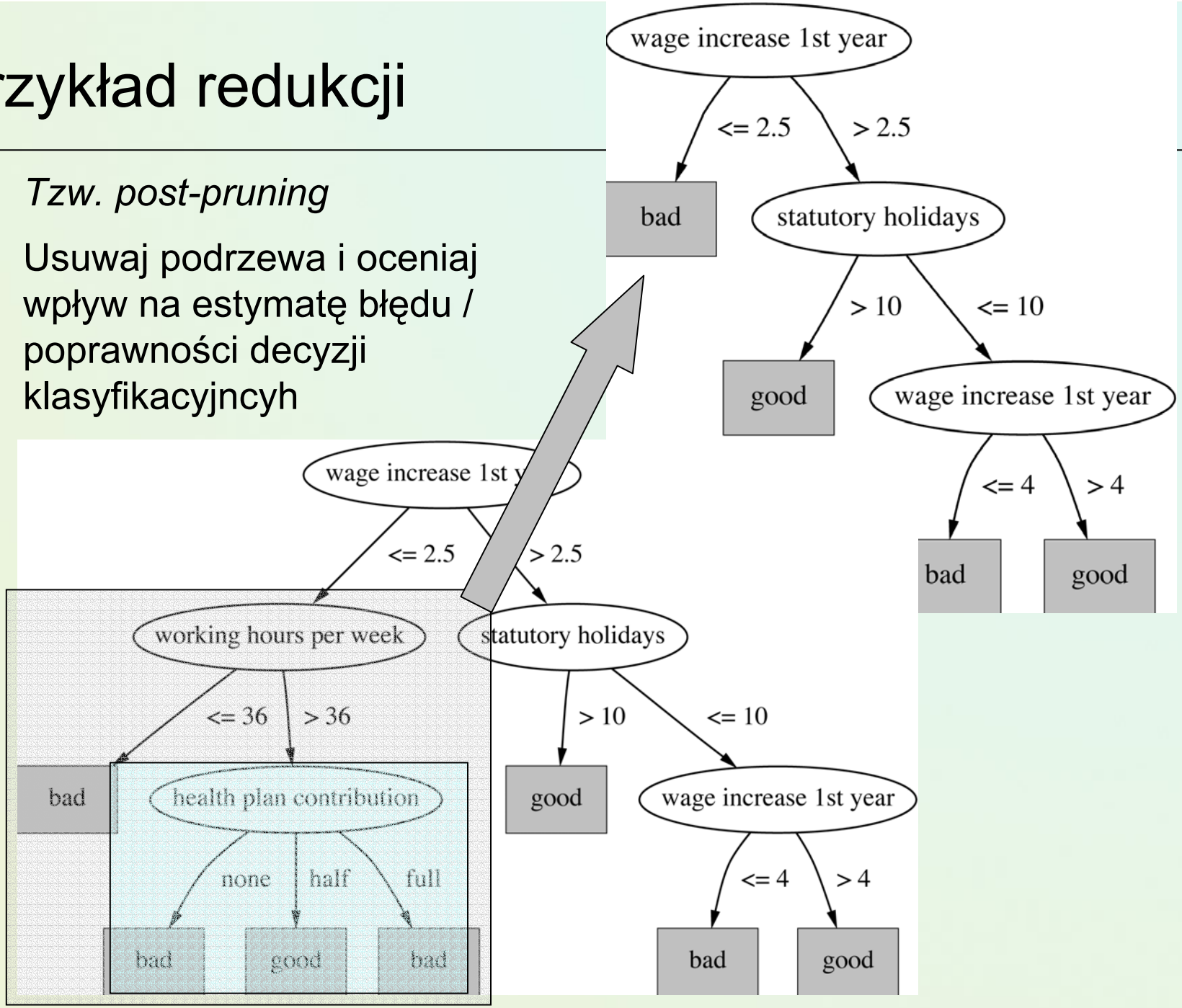
Tree pruning – upraszczanie drzewa

- Tree pruning – mechanizm „walki” z przeuczeniem
- Po uproszczeniu struktury drzewa może wzrosnąć trafność na przykładach testowych!



Przykład redukcji

- *Tzw. post-pruning*
- Usuwaj podrzewa i oceniaj wpływ na estymatę błędu / poprawności decyzji klasyfikacyjncyh



Przykłady zastosowań w medycynie

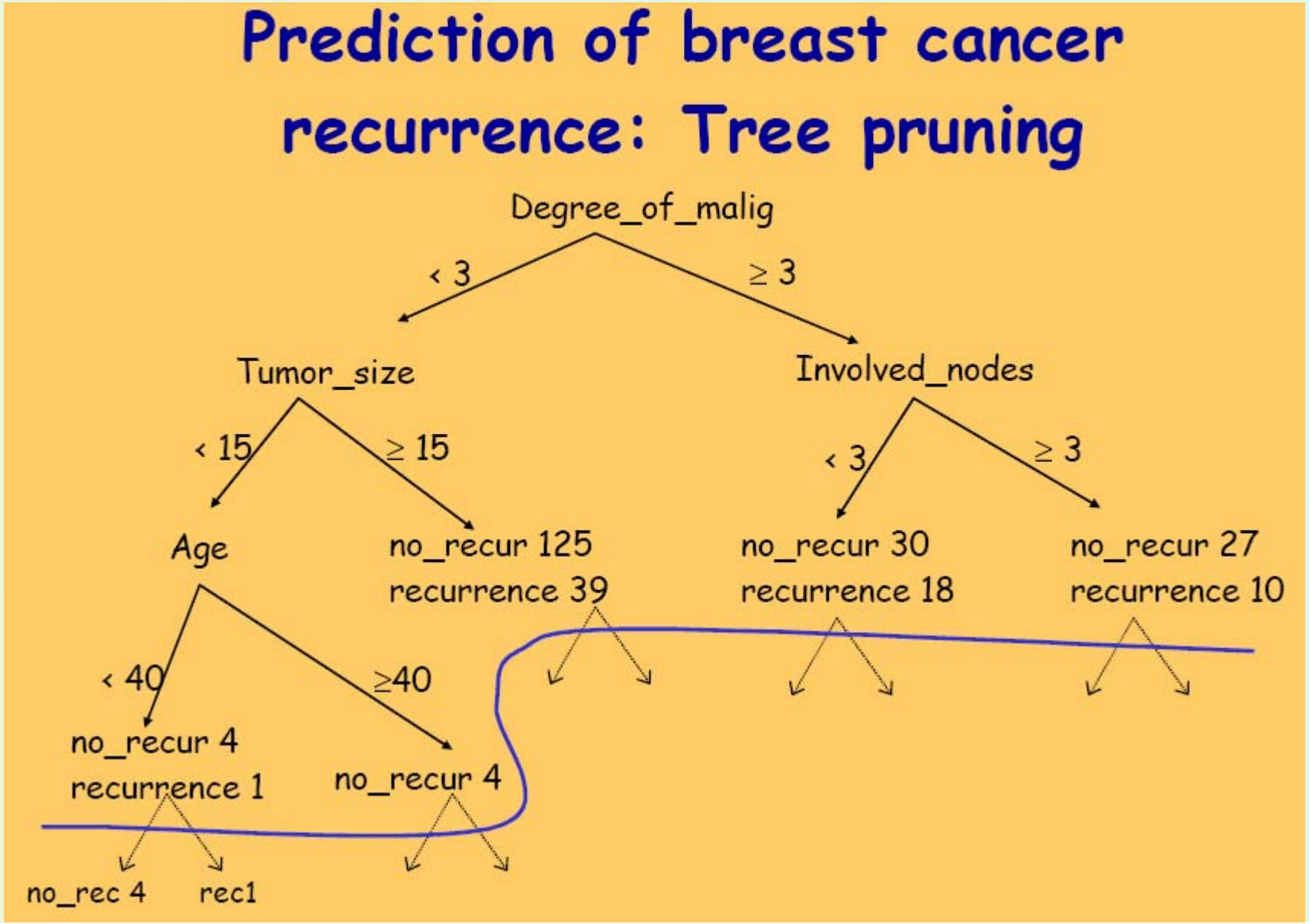
- Wiele przykładów analizy podejmowania decyzji o diagnozowaniu chorób, także terapii, oraz farmacja i budowa związków - leków:
- Przykładowe omówienia:
  - I.Kononenko, I.Bratko, M.Kukar: Application of Machine Learning to Medical Diagnosis. w: Michalski R.S., Bratko I, Kubat M. (red.), Machine learning and data mining, John Wiley & Sons, 1998, s. 389-408.
  - Langley, P., Simon, H. A., Fielded applications of machine learning, w: Michalski R.S., Bratko I, Kubat M. (red.), Machine learning and data mining, John Wiley & Sons, 1998 , s. 113-129.
- Spójrz także na dodatkowe slajdy (włączmy tradycyjny rzutnik ☺…

Przykład drzewa decyzyjnego

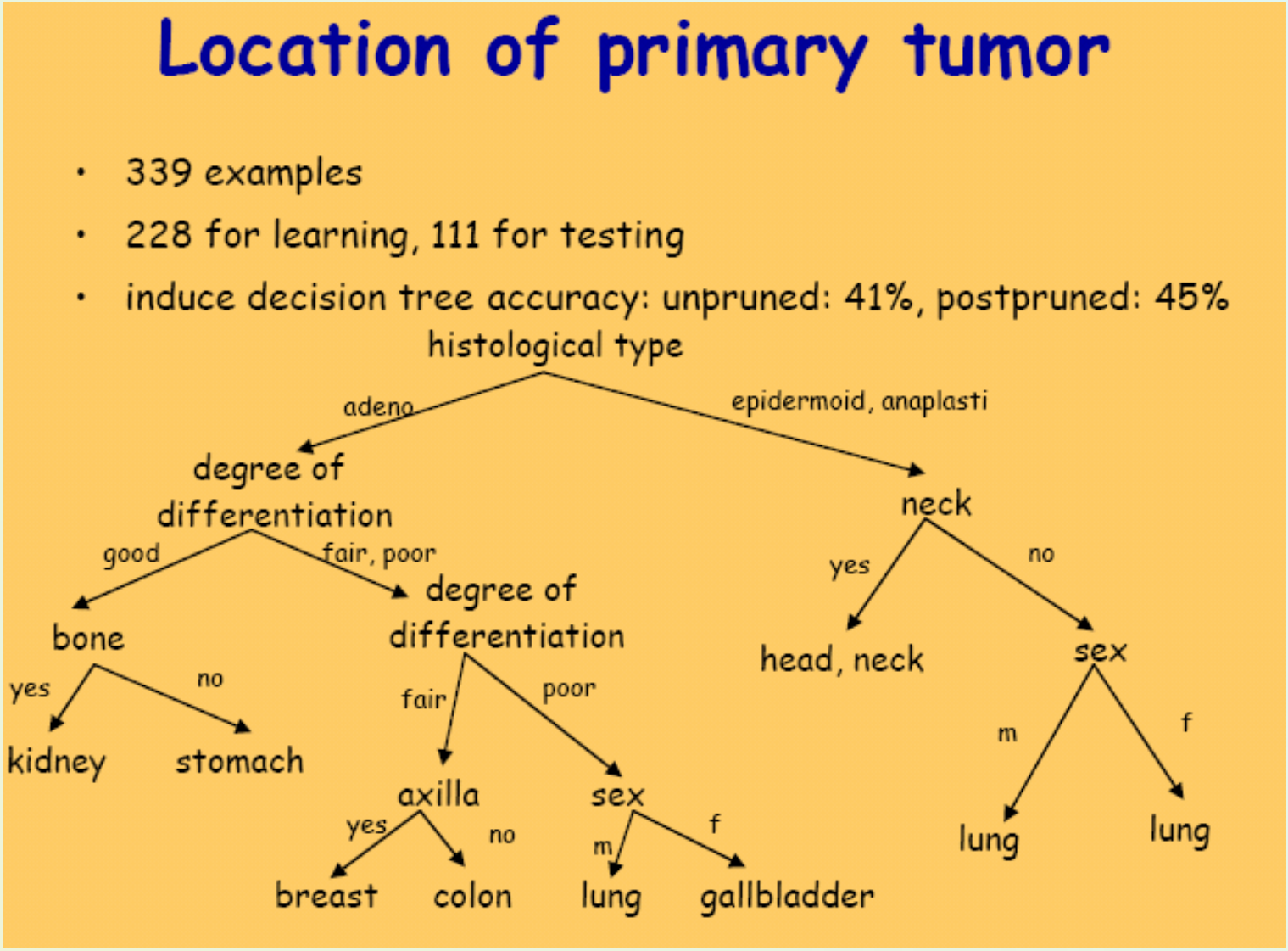
Medicine - Predicting C-Section Risk

- Learned from Medical Records of 1000 Women
- Negative Examples are Cesarean Sections
  - Prior distribution: [833+, 167-] 0.83+, 0.17-
  - *Fetal-Presentation* = 1: [822+, 167-] 0.88+, 0.12-
    - *Previous-C-Section* = 0: [767+, 81-] 0.90+, 0.10-
      - ***Primiparous* = 0: [399+, 13-] 0.97+, 0.03-**
      - ***Primiparous* = 1: [368+, 68-] 0.84+, 0.16-**
        - ***Fetal-Distress* = 0: [334+, 47-] 0.88+, 0.12-**
          - ***Birth-Weight* < 3349 0.95+, 0.05-**
          - ***Birth-Weight* ≥ 3347 0.78+, 0.22-**
        - ***Fetal-Distress* = 1: [34+, 21-] 0.62+, 0.38-**
    - *Previous-C-Section* = 1: [55+, 35-] 0.61+, 0.39-
  - *Fetal-Presentation* = 2: [3+, 29-] 0.11+, 0.89-
  - *Fetal-Presentation* = 3: [8+, 22-] 0.27+, 0.73-

A real life example of tree pruning [U.Lubljana]

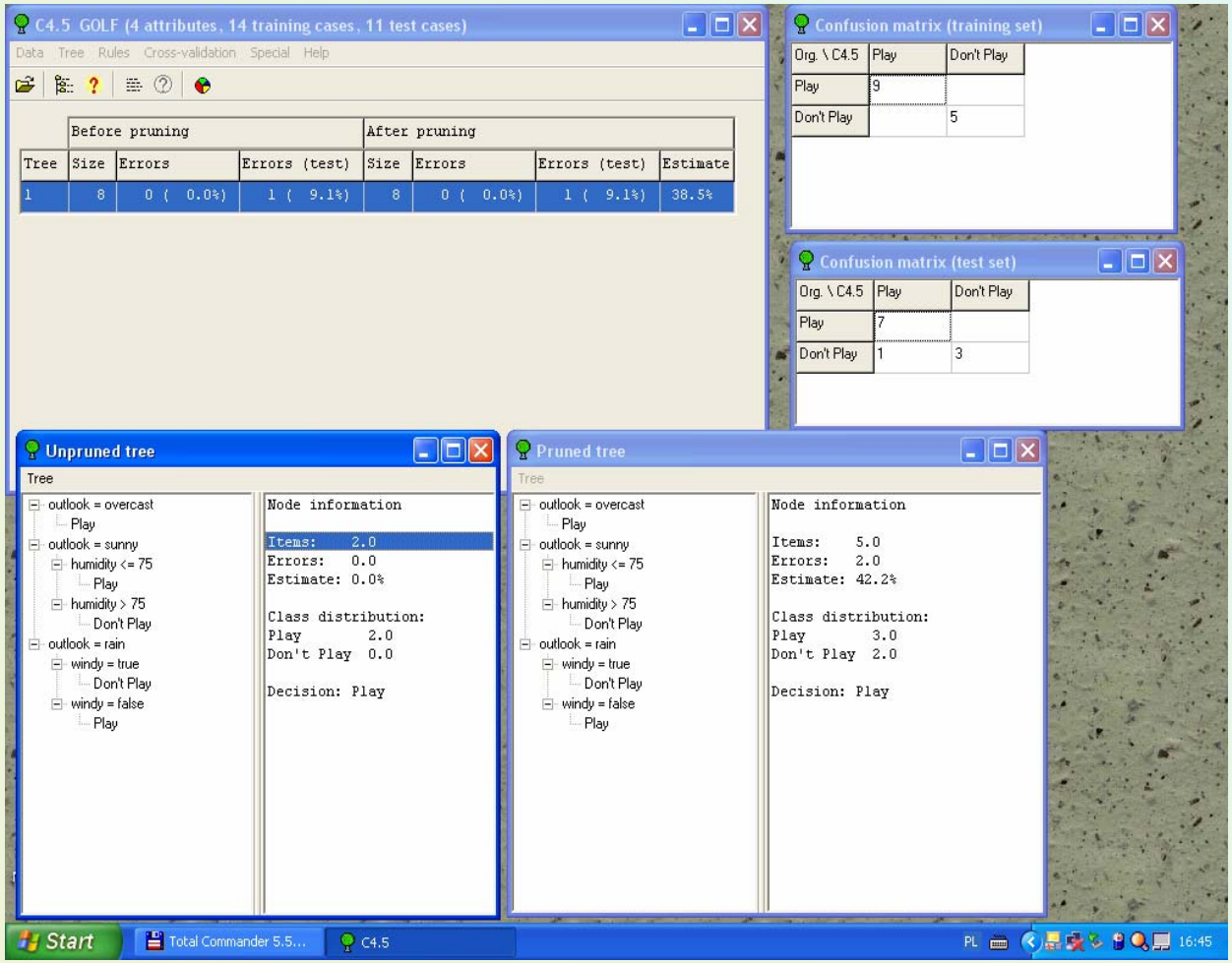


Another example from U.Lubljana

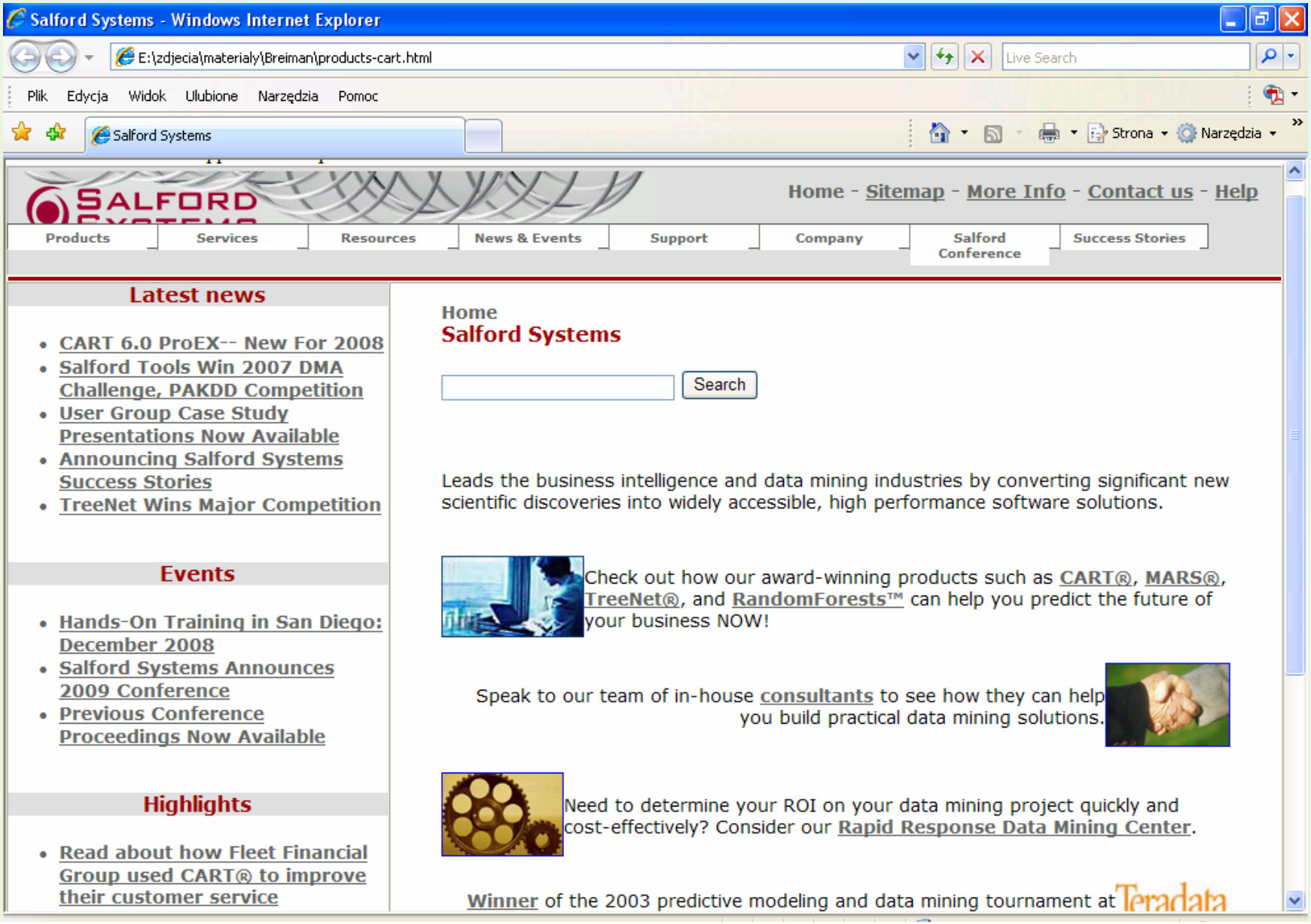


C4.5 Quinlan + PP interfejs

- Opiera się na oryginalnym kodzie J.R. Quinlana



CART from Salford Systems



Analiza dużych bazach danych

- Klasyfikacja rozważana głównie przez:
  - statystyków
  - badaczy z AI –„machine learning”
- Data mining wymaga rozważenia problemów związanych z przetwarzaniem dużych baz danych
  - dotychczasowe zastosowania na małych rozmiarach danych, algorytmy wykorzystują PAO
- Data mining dostarcza rozszerzeń w zakresie:
  - Skalowaności,
  - Przetwarzania równoległego i rozproszonego,
  - Danych o złożonej strukturze,
  - ….

Drzewa decyzyjne a data mining ?

- Wybór drzew decyzyjnych:
  - Względnie szybkie algorytmy indukcji drzew, z możliwością modyfikacji dla dużych zbiorów danych
  - Możliwość transformacji w reguły klasyfikacyjne
  - Jawna reprezentacja wiedzy w postaci symbolicznej wygodnej do interpretacji przez człowieka
  - Łatwość realizacji klasyfikowania nowych przypadków
  - Osiągają wysoką trafność klasyfikowania, porównywalną z innymi metodami
  - Odporność na niedoskonałe dane
- Zastosowania do problemów z bardzo dużą liczbą przykładów opisanych setkami atrybutów (akceptowalne czasy odpowiedzi)

Indukcja reguł decyzyjnych

- Podstawowa idea - reguły poszukuje się bezpośrednio z danych
  - potencjalnie większa zrozumiałość wiedzy
  - ale więcej różnych podejść
    - — opisy rzeczywistych danych minimalnym zbiorem reguł dyskryminujących/ klasyfikujących
    - — poszukiwanie bardziej wyczerpujących zbiorów reguł o dobrych własnościach interpretacyjnych
  - więcej parametrów do sterowania w metodach regułowych

Indukcja reguł metodą generowania kolejnych pokryć

Sequential covering (*Xj* klasa; *A* atrybuty; *E* przykłady, τ próg akceptacji);
begin
    R := ∅; {zbiór poszukiwanych reguł}
    r := *learn-one-rule*(klasa *Xj*; atrybuty *A*; przykłady *E*)
    while *evaluate*(*r*,*E*) > τ **do**
    begin
      R := *R*∪*r*;
      E := *E*\[*R*]; {usuń przykłady pozytywne pokryte przez *R*}
      r := *learn-one-rule*(klasa *Xj*; atrybuty *A*; przykłady *E*);
    end;
return *R*
end.

Ryszard Michalski

- Funkcja *learn-one-rule* dla danego zbioru przykładów znajduje jedną regułę pokrywającą możliwie jak najwięcej przykładów pozytywnych i jak najmniej negatywnych.

MODLEM – algorytm indukcji reguł

- MODLEM [Stefanowski 98] generuje minimalny zbiór reguł pokrywających zbiór przykładów uczących
- Właściwości – możliwość wykorzystanie teorii zbiorów przybliżonych do analizy sprzecznych przykładów; bezpośrednie przetwarzanie atrybutów liczbowych (bez wstępnej dyskretyzacji) oraz nieznanych wartości atrybutów.
- Strategia wykorzystywania reguł $\rightarrow$ LERS [Grzymała 94]

obj.	*a*1	*a*2	*a*3	*a*4	*D*
*x*1	m	2.0	1	a	C1
*x*2	f	2.5	1	b	C2
*x*3	m	1.5	3	c	C1
*x*4	f	2.3	2	c	C1
*x*5	f	1.4	2	a	C2
*x*6	m	3.2	2	c	C2
*x*7	m	1.9	2	b	C1
*x*8	f	2.0	3	a	C2

if (*a*1 = m) *and* (*a*2 $\leq$ 2.6) *then* (*D* = C1) {*x*1,*x*3,*x*7}

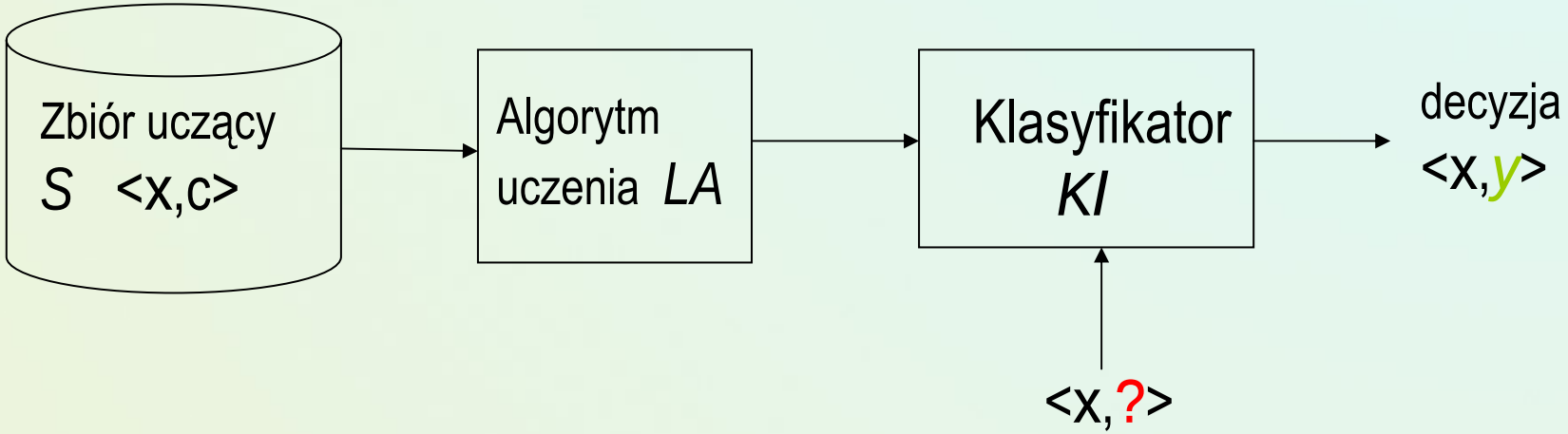
if (*a*2 $\in$ [1.45, 2.4]) *and* (*a*3 $\leq$ 2) *then* (*D* = C1) {*x*1,*x*4,*x*7}

if (*a*2 $\geq$ 2.4) *then* (*D* = C2) {*x*2,*x*6}

if (*a*1 = f) *and* (*a*2 $\leq$ 2.15) *then* (*D* = C2) {*x*5,*x*8}

Predykcja nowych faktów - klasyfikatory

- Poszukiwanie reprezentacji wiedzy o przydziale obiektów do klas na podstawie opisu obiektów za pomocą wartości atrybutów (zbiór uczący).
- Predykcja klasyfikacji nowych obiektów (zbiór testowy)



Przykłady $S = \{\langle \mathbf{x}_1, c_1 \rangle, \langle \mathbf{x}_2, c_2 \rangle, \cdots; \langle \mathbf{x}_n, c_n \rangle\}$
$\mathbf{x}_i = \langle x_{i1}, x_{i2}, \ldots, x_{im} \rangle$ opisywane przez m atrybutów
Atrybuty różnego typu
c_i – etykieta jednej z klas $\{C_1, \ldots, C_K\}$

Miara oceny, np:
trafność klasyfikowania

$$\eta = \frac{N_c}{N_t}$$

Eksperymentalna ocena
→ *Cross validation*

Trafność klasyfikowania

- Użyj przykładów testowych nie wykorzystanych w fazie indukcji klasyfikatora:
  - N_t – liczba przykładów testowych
  - N_c – liczba poprawnie sklasyfikowanych przykładów testowych
- Trafność klasyfikowania (classification accuracy):

$$\eta = \frac{N_c}{N_t}$$

- Alternatywnie błąd klasyfikowania.

$$\varepsilon = \frac{N_t - N_c}{N_t}$$

Inne możliwości analizy:

- macierz pomyłek (ang. confusion matrix),
- koszty pomyłek i klasyfikacja binarna,
- miary Sensitivity i Specificity / krzywa ROC

Macierz pomyłek

- Analiza pomyłek w przydziale do różnych klas przy pomocy tzw. macierz pomyłek (ang. *confusion matrix*)
- Macierz $r \times r$, gdzie wiersze odpowiadają poprawnym klasom decyzyjnym, a kolumny decyzjom przewidywanym przez klasyfikator; na przecięciu wiersza *i* oraz kolumny *j* - liczba przykładów *n-ij* należących oryginalnie do klasy *i*-tej, a zaliczonej do klasy *j*-tej

Przykład:

	Przewidywane klasy decyzyjne		
Oryginalne klasy	K_1	K_2	K_3
K_1	**50**	**0**	**0**
K_2	**0**	**48**	**2**
K_3	**0**	**4**	**46**

Klasyfikacja binarna

- Niektóre zastosowania → jedna z klas posiada szczególne znaczenie, np. diagnozowanie poważnej choroby. Problem → klasyfikacja binarna.

Oryginalne klasy	**Przewidywane klasy decyzyjne**	
	Pozytywna	**Negatywna**
Pozytywna	***TP***	***FN***
Negatywna	***FP***	***TN***

- Nazewnictwo (inspirowane medycznie):
  - *TP* (ang. *true positive*) – liczba poprawnie sklasyfikowanych przykładów z wybranej klasy (ang. *hit*),
  - *FN* (ang. *false negative*) – liczba błędnie sklasyfikowanych przykładów z tej klasy, tj. decyzja negatywna podczas gdy przykład w rzeczywistości jest pozytywny (błąd pominięcia - z ang. *miss*),
  - *TN* (ang. *true negative*) – liczba przykładów poprawnie nie przydzielonych do wybranej klasy (poprawnie odrzuconych z ang. *correct rejection*),
  - *FP* (ang. *false positive*) – liczba przykładów błędnie przydzielonych do wybranej klasy, podczas gdy w rzeczywistości do niej nie należą (ang.

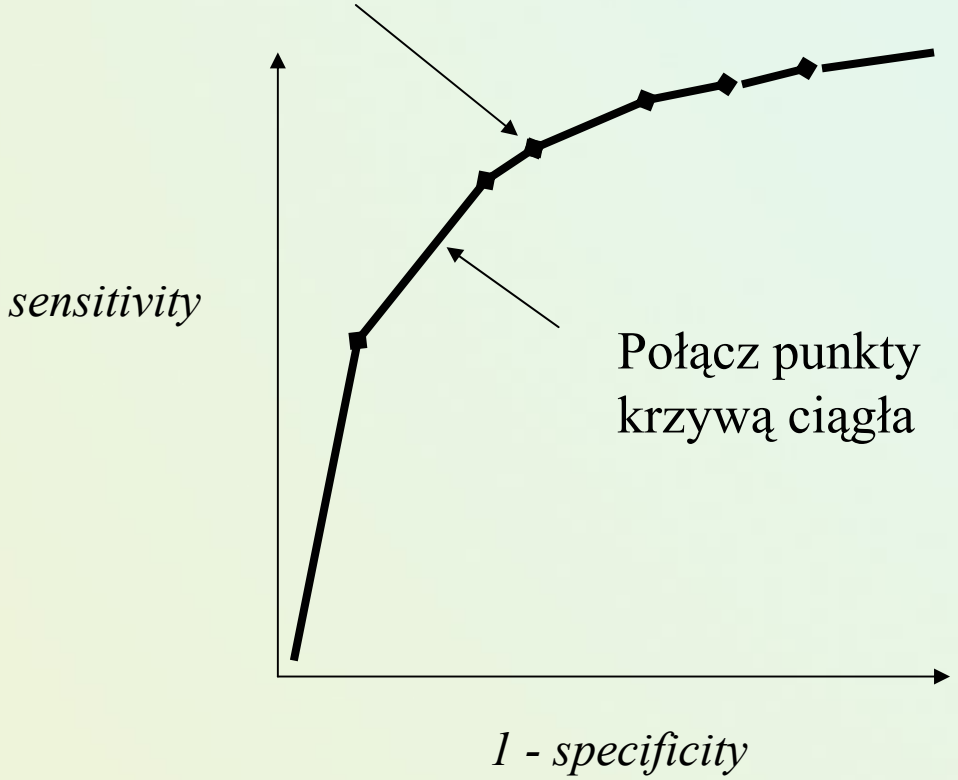
Miary stosowane w analizie klasyfikacji binarnej

- Dodatkowe miary oceny rozpoznawania wybranej klasy:
  - Wrażliwość / czułość (ang. *sensitivity*) = *TP* / (*TP*+*FN*),
  - Specyficzność (ang. *specificity*) = *TN* / (*FP*+*TN*).
- Inne miary:
  - *False-positive rate* = *FP* / (*FP*+*TN*), czyli 1 – specyficzność.
- Wnikliwszą analizę działania klasyfikatorów binarnych dokonuje się w oparciu o analizę krzywej ROC, ang. *Receiver Operating Characteristic*).

Oryginalne klasy	**Przewidywane klasy decyzyjne**	
	Pozytywna	**Negatywna**
Pozytywna	***TP***	***FN***
Negatywna	***FP***	***TN***

Krzywa ROC - analiza

Algorytm może być parametryzowany, i w rezultacie otrzymuje się serie punktów odpowiadających doborowi parametrów



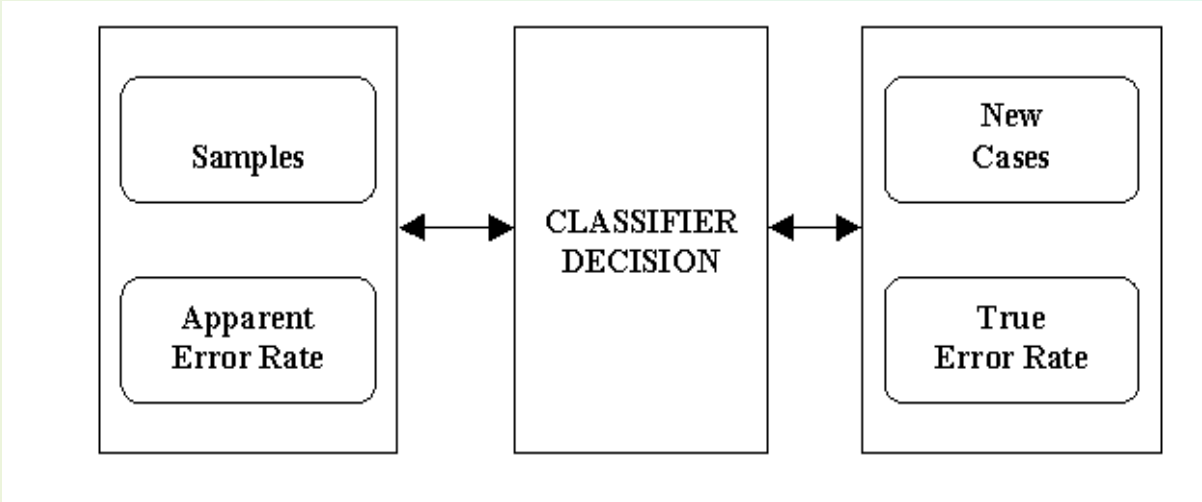
Wykres nazywany ‘krzywą’ ROC.

Jak szacować wiarygodnie ?

- Zależy od perspektywy użycia wiedzy:
  - Predykcja klasyfikacji albo opisowa
- Ocena na zbiorze uczącym nie jest wiarygodna jeśli rozważamy predykcję nowych faktów!
  - Nowe obserwacje najprawdopodobniej nie będą takie same jak dane uczące!
  - Choć zasada reprezentatywności próbki uczącej …
- Problem przeuczenia (ang. overfiting)
  - Nadmierne dopasowanie do specyfiki danych uczących powiązane jest najczęściej z utratą zdolności uogólniania (ang. generalization) i predykcji nowych faktów!

Podejście empiryczne

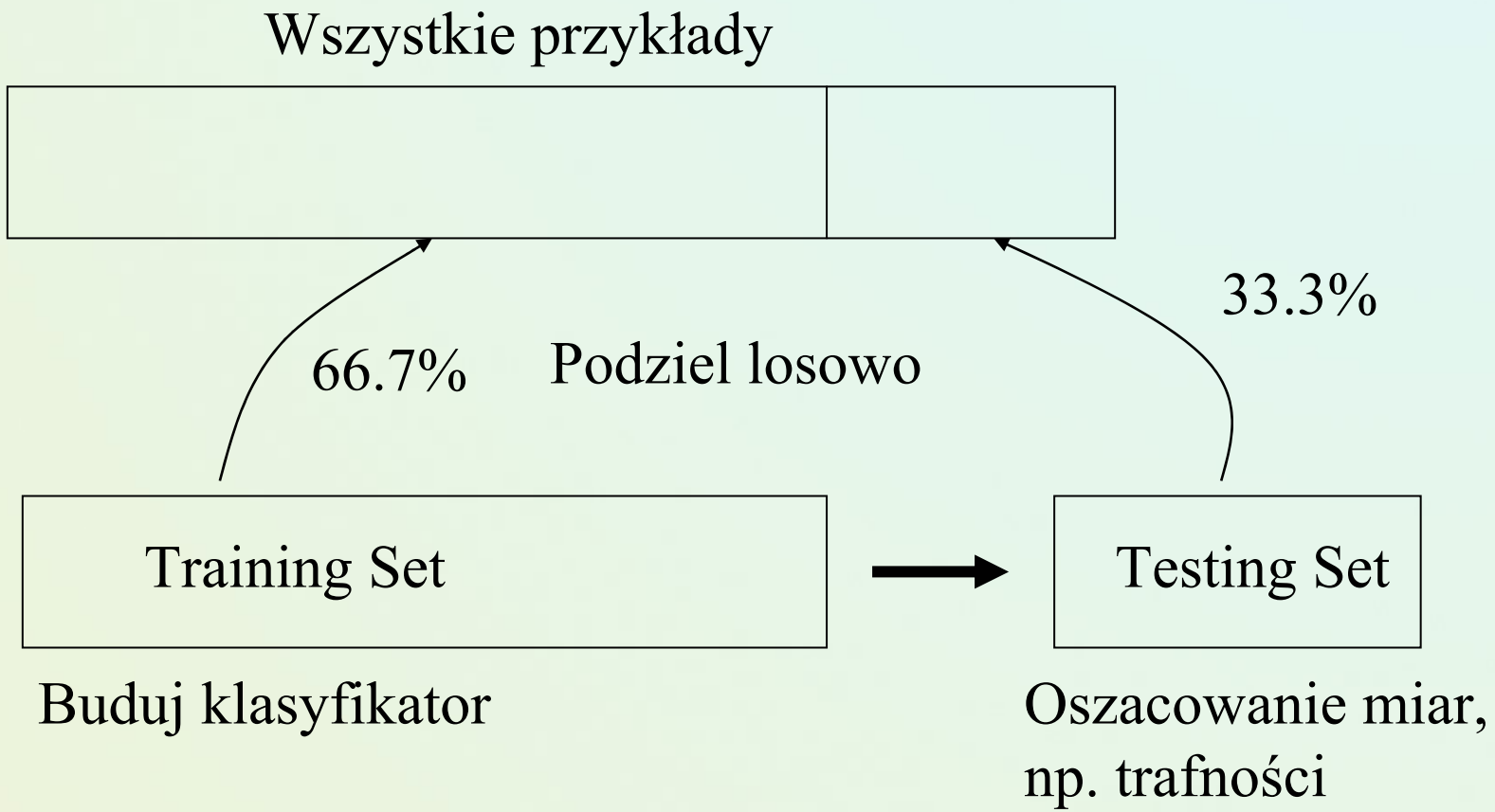
- Zasada „Train and test”
- Gdy nie ma podziału zadanego przez nauczyciela, to wykorzystaj losowe podziały
- Nadal pytanie jak szacować wiarygodnie?



Empiryczne metody estymacji

- **Techniki podziału: „hold-out”**
  - Użyj dwóch niezależnych zbiorów: uczącego (2/3), testowego (1/3)
  - Jednokrotny podział losowy stosuje się dla dużych zbiorów (hold-out)
- **„Cross-validation” - Ocena krzyżowa**
  - Podziel losowo dane w *k* podzbiorów (równomierne lub warstwowe)
  - Użyj *k-1* podzbiorów jako części uczącej i pozostałej jako testującej (*k*-fold cross-validation).
  - Oblicz wynik średni.
  - Stosowane dla danych o średnich rozmiarach (najczęściej k = 10) Uwaga opcja losowania warstwowego (ang. stratified sampling**).**
- **Leaving-one-out**
  - Dla małych rozmiarów danych.
  - „Leaving-one-out” jest szczególnym przypadkiem, dla którego liczba iteracji jest równa liczbie przykładów

Jednokrotny podział (hold-out) – duża liczba przykładów (> tysięcy)



Inne metody klasyfikacyjne

- sztuczne sieci neuronowe
- klasyfikacja bayesowska
- analiza dyskryminacyjna (statystyczna)
- metody k-najbliższych sąsiadów
- metoda wektorów wspierających (SVM)
- algorytmy genetyczne
- metody oparte na logice matematycznej (ILP)
- ...

Klasyfikacja Bayesowska

- Probabilistic learning: Pozwala na obliczanie prawdopodobieństw związanych z hipotezami; efektywne podejście do wielu problemów praktycznych.
- Incremental: Każdy nowy przykład może zmienić oszacowanie prawdopodobieństw; Możliwość uwzględniania prawdopodobieństw a'priori.
- Predykcja probabilistyczna: Predykcja wielu hipotez, "ważonych" za pomocą prawdopodobieństw.
- Standard: Pomimo, że w pewnych przypadkach metody bayesowskie są nieatrakcyjne obliczeniowo, tp mogą dostarczać standard optymalnych decyzji, z którym można porównywać inne metody.

Twierdzenie Bayes'a

- Dla zbioru uczącego *D, prawdopodobieństwo posteriori hipotezy h, P(h|D)* wynika z twierdzenia Bayes'a:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- MAP (maximum posteriori) hypothesis:

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Praktyka: założenie znajomości prawdo-podobieństw *a priori*, duże koszty obliczeniowe.
- Przykład: System d'Dombal $\rightarrow$ spójrz na dodatkowe slajdy.

Zbiór uczący

- Przykład Quinlan'a (*Play a game*).

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Naiwny klasyfikator bayesowski (II)

- Faza uczenia $\rightarrow$ obliczanie prawdopodobieństw warunkowych

Outlook	P	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	P	N
high	3/9	4/5
normal	6/9	1/5
Windy		
true	3/9	3/5
false	6/9	2/5

Analiza nowego przypadku

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

- Nowa sytua

Likelihood of the two classes

For “yes” = $2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$

For “no” = $3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$

Conversion into a probability by normalization:

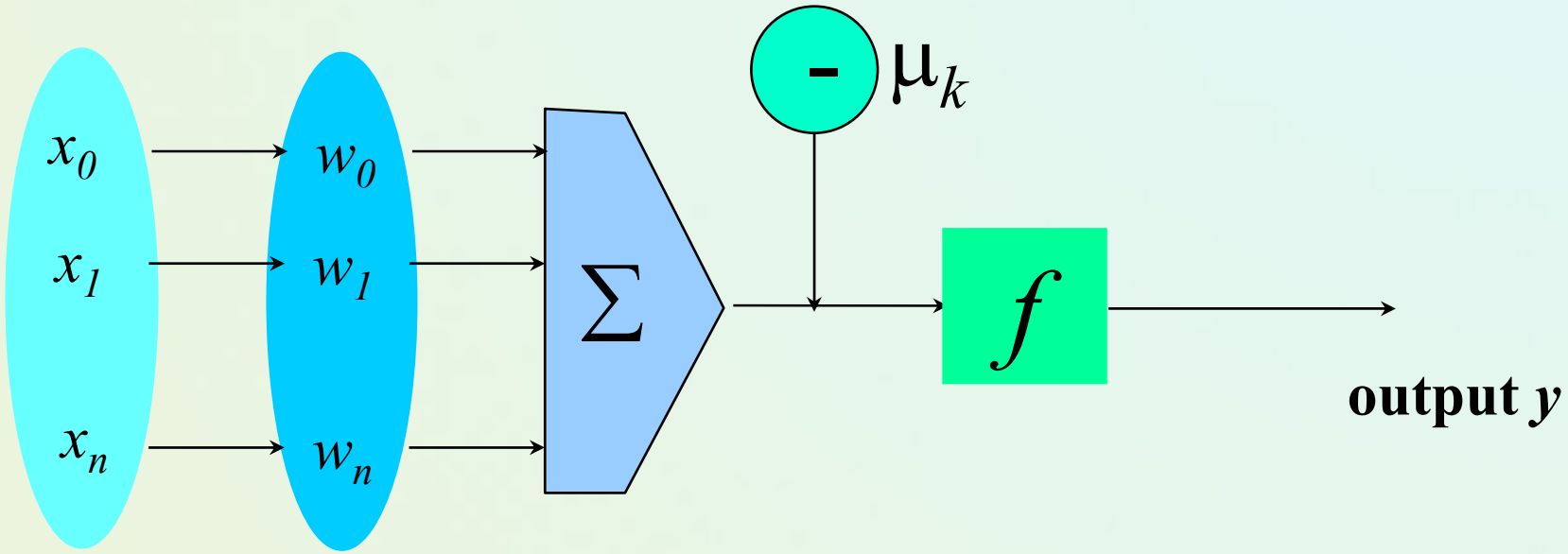
P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

Sieci Neuronowe

- Zalety
  - Często wysoka trafność klasyfikacji
  - Odporne na „zaszumione” dane
  - Wyjście może być dyskretne, liczbą rzeczywistą, lub wektorem dyskretnym / liczbowym
  - Po nauczeniu, względnie szybki klasyfikator
- Krytycyzm
  - Długi i złożony proces uczenia
  - Pozyskana wiedza niemożliwa do wyjaśnienia (wagi).
  - Trudno uwzględniać wiedzę dziedzinową

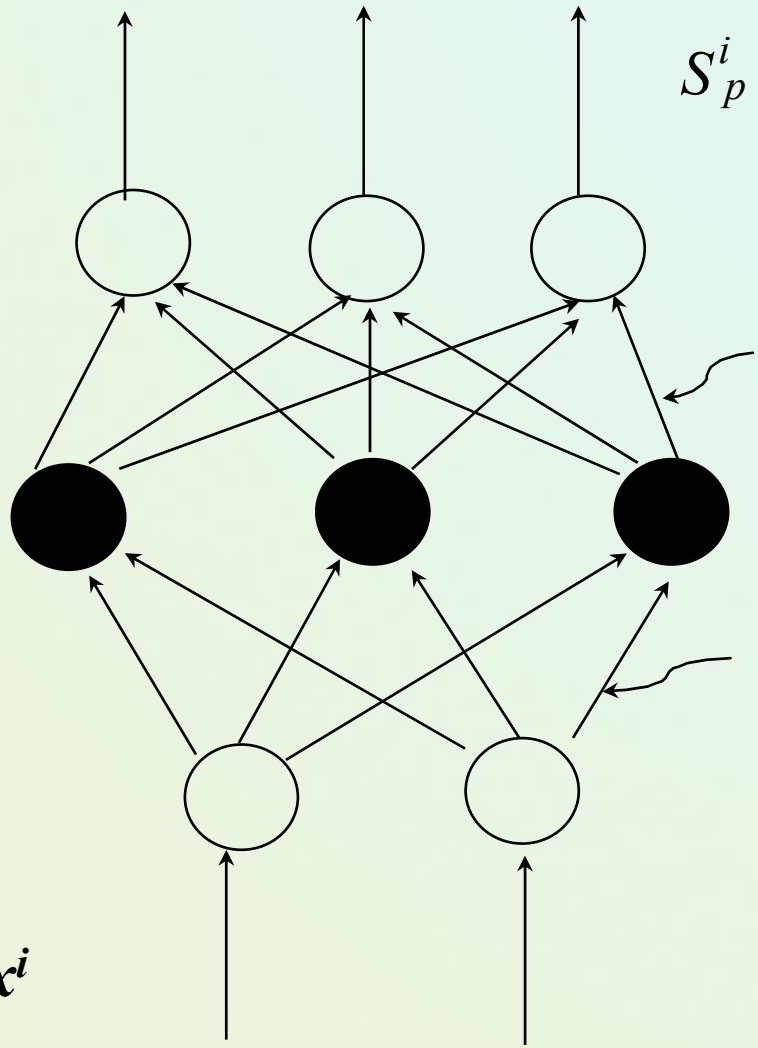
Sztuczny neuron



Input vector x **weight vector w** **weighted sum** **Activation function**

- n-wymiarowy wektor x jest odwzorowywany w zmienną y za pomocą iloczynu skalarnego i nieliniowej funkcji przejścia

Multi Layer Perceptron



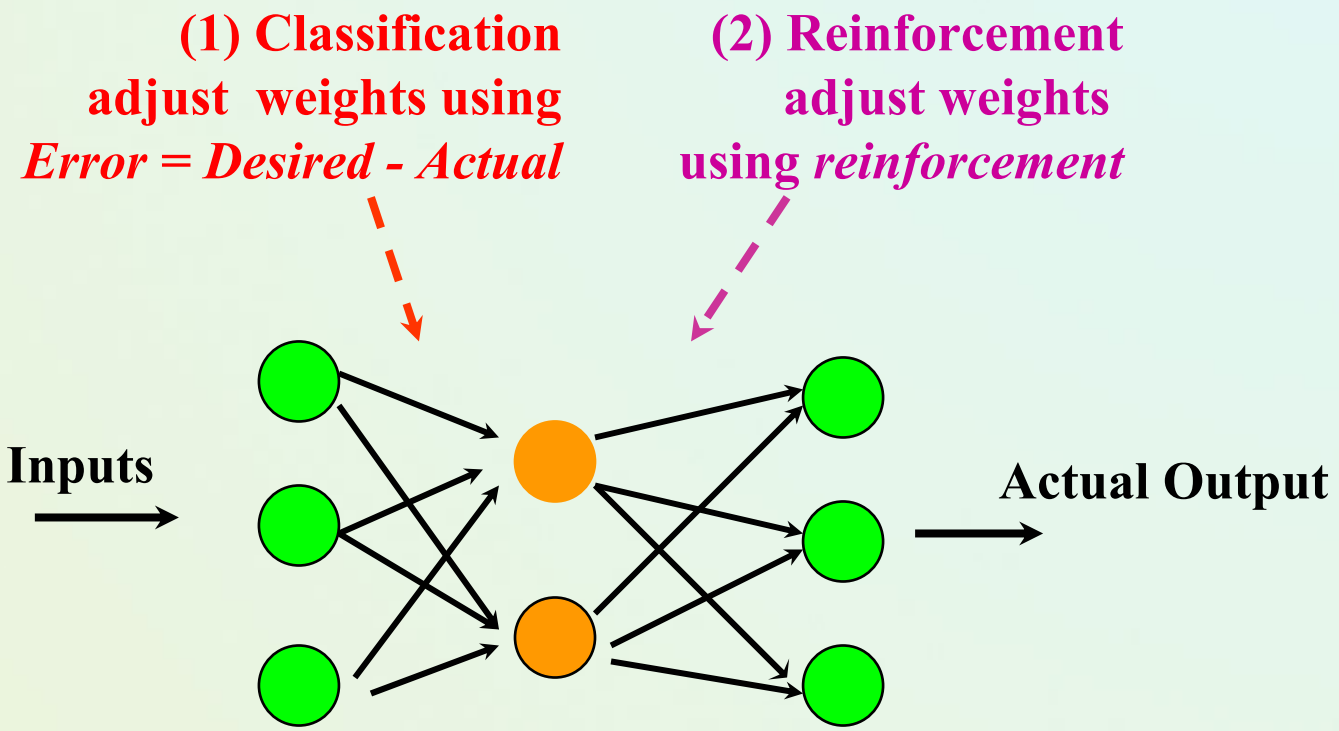
$$S_p^i = \sigma\left(\sum_{m=1}^{h} \alpha^m v_p^m\right)$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

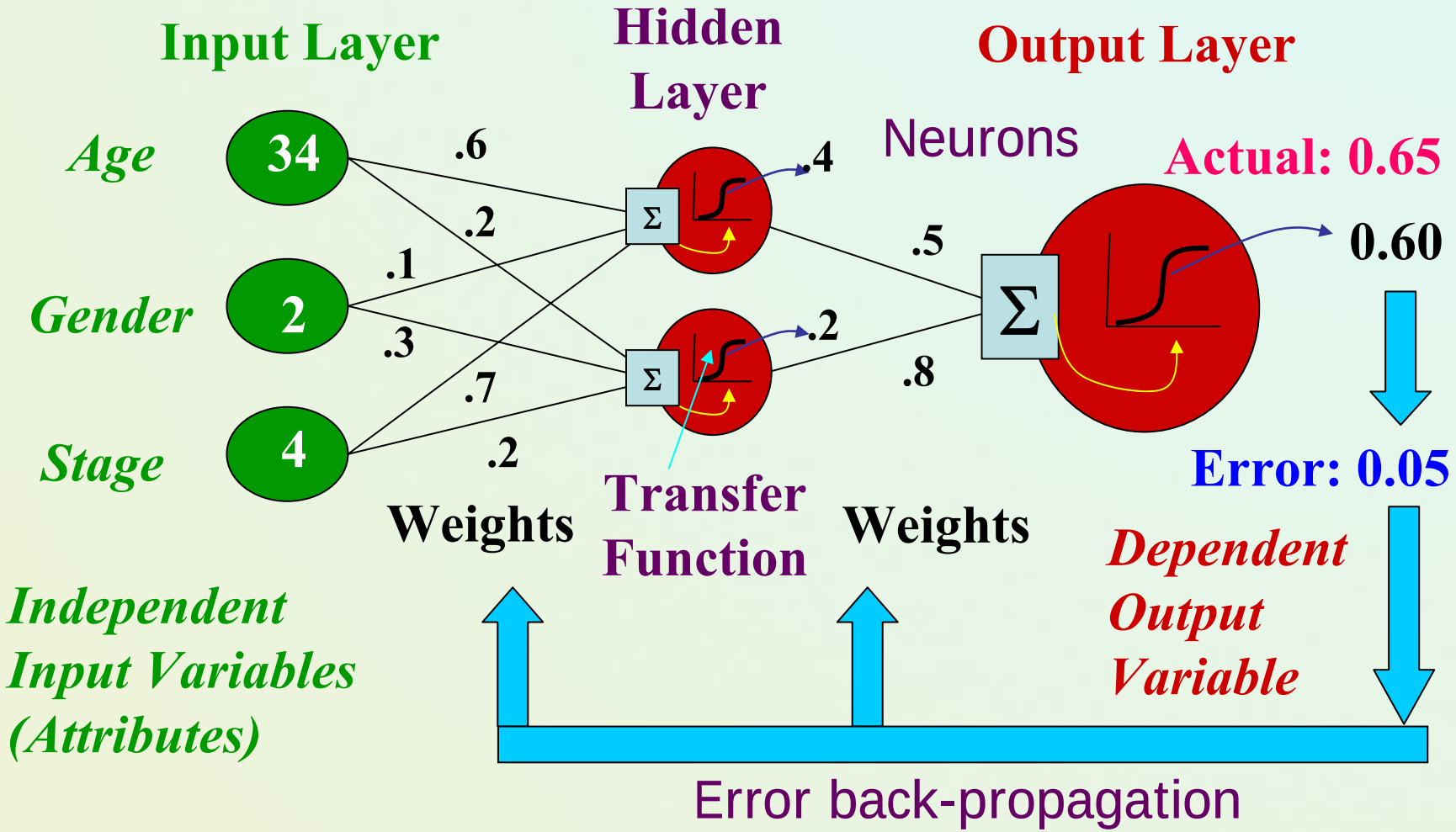
$$\alpha^m = f\left(\sum_{l=1}^{n} (x_l^i w_l^m) - r^m\right)$$

$$f(x) ::= \delta(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

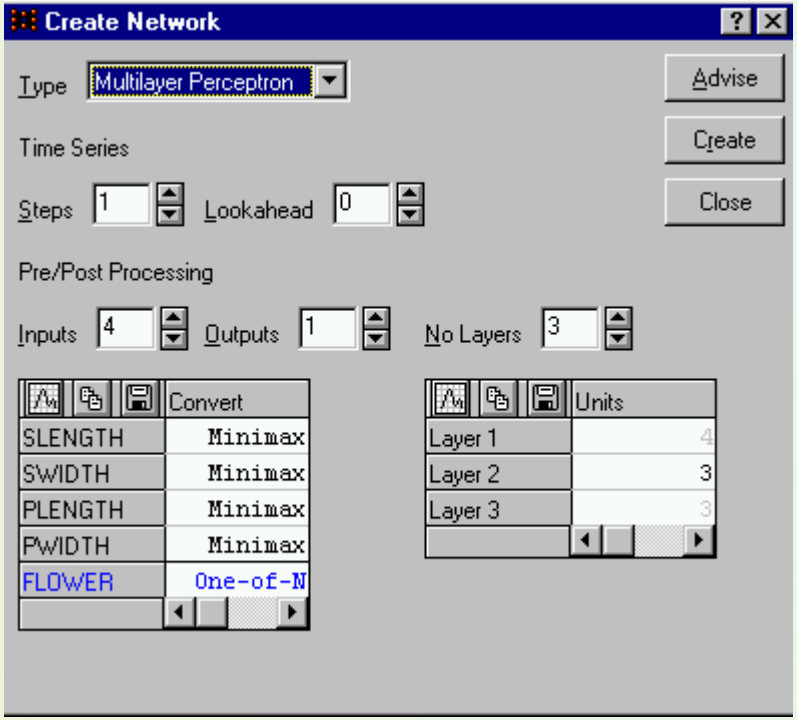
Paradygmaty Uczenia się (I) - Sieć Warstwowa

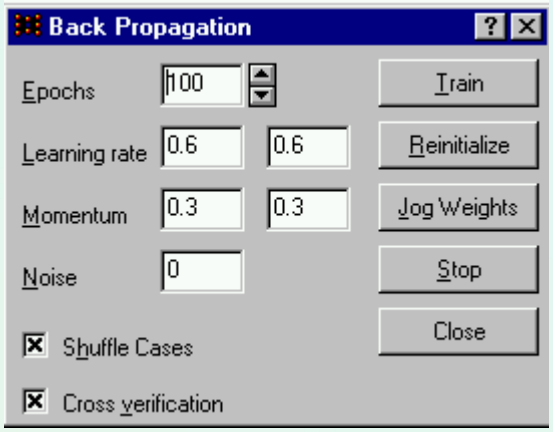


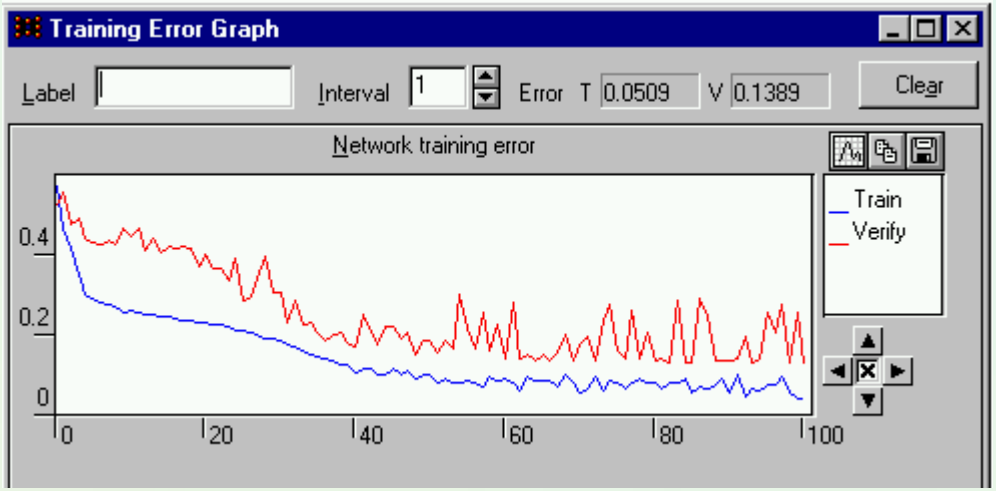
The Neural Network (NN) Approach

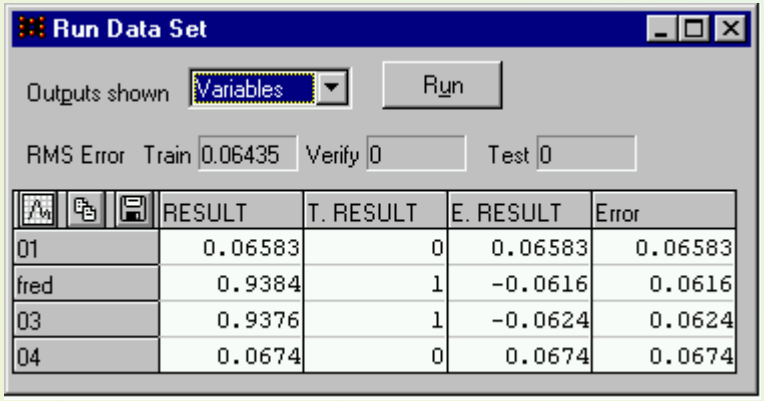


STATISTICA



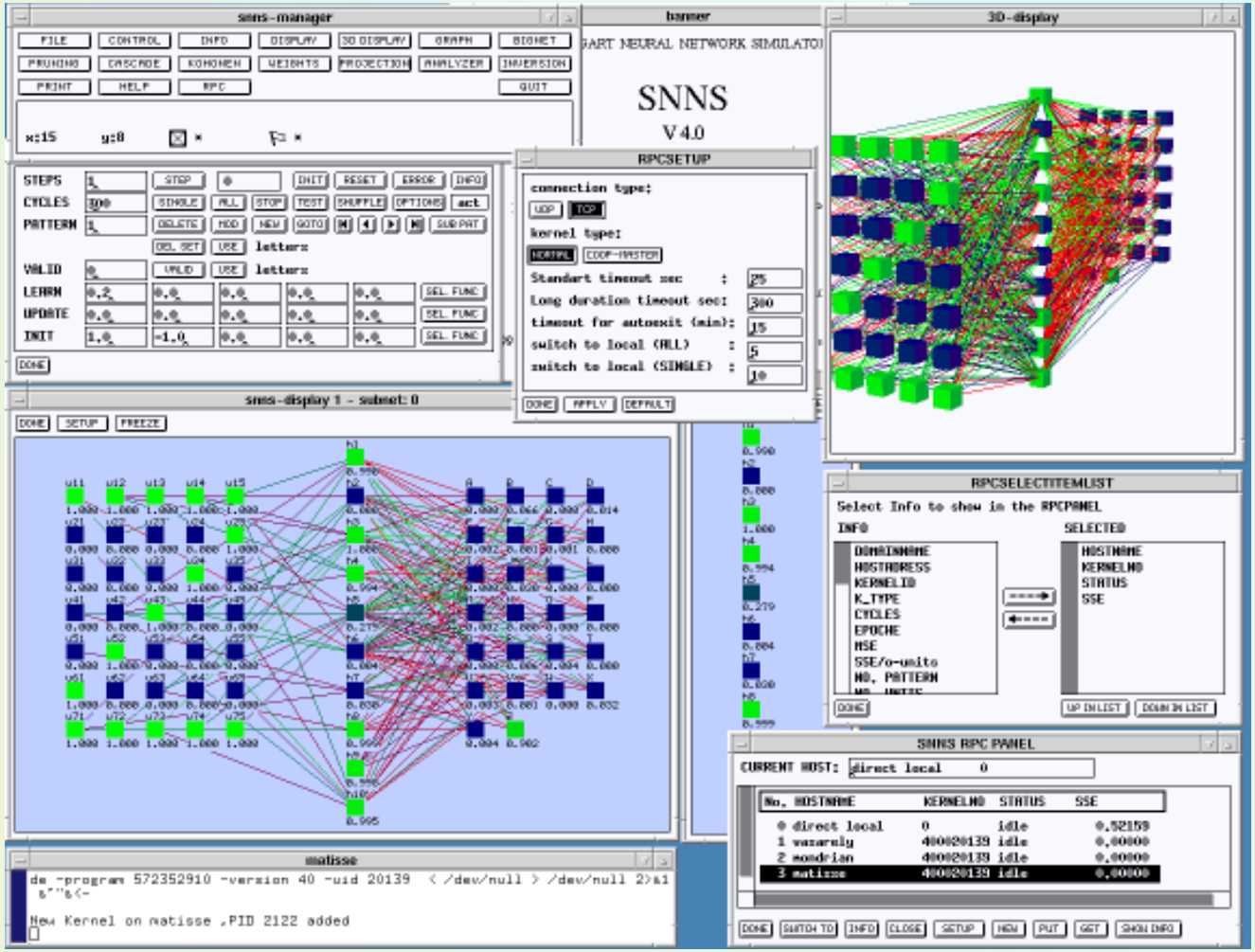






- Tworzenie i uruchomienie sieci

Przykład pakietu SNNS



- *Stuttgart Neural Network Simulator*

Przykład: Klasyfikacja elektroencefalogramu

- **Zadanie:** klasyfikacja elektroencefalogramu do jednej z trzech klas:
  1. normal (*norm*)
  2. shizophrenic (*shiz*)
  3. obsessive compulsive disorder (*ocd*)

Dane wejściowe:

- postać oryginalna danych: wynik badania elektroencefalograficznego zaklasyfikowany przez neurologów do jednej z trzech klas: 19 kanałów próbkowanych z częstotliwości 128 Hz, pacjent w spoczynku

Przetwarzanie wstępne:

- usunięcie artefaktów (np. zakłóceń wynikłych z poruszeń pacjenta).
- usunięcie składowych o wysokich częstotliwościach (np. szumów) poprzez obróbkę filtrem dolnoprzepustowym (30 Hz),..
- z 19 sygnałów do dalszej obróbki wybrany kanał Cz (pobierany ze szczytu głowy) $\rightarrow$ wcześniejsze badania wykazały największą przydatność do celów diagnostycznych (w przybliżeniu sygnał dostarczany tym kanałem jest liniową kombinacją części innych sygnałów) .
- Wybór okna czasowego (pierwsze 250 sekund) – stan największego uspokojenia pacjenta
- ….

Klasyfikacja elektroencefalogramu

- **Ekstrakcja cech**: Metoda: autoregresji (AR); polega na przybliżenia przebiegu przez rekurencyjna zależność:

$$s_t = \sum_{j=1}^{N} \alpha_j s_{t-j} + \varepsilon_t$$

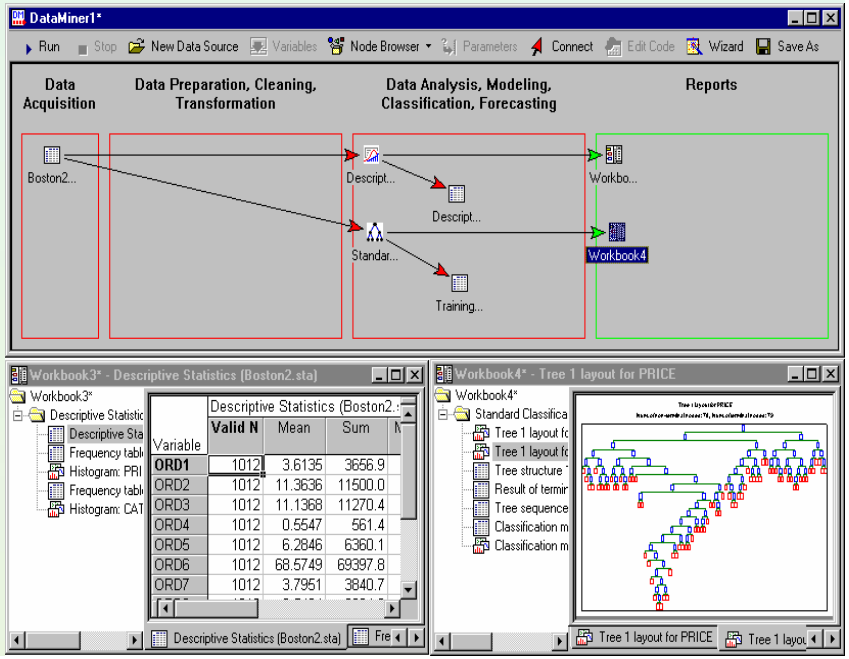
- *st* - sygnał w chwili *t* , ε pewna zmienna losowa o rozkładzie normalnym
- Szukanymi parametrami opisującymi przebieg s αj dla *j*=1..N; oblicza się specjalnymi metodami - są atrybutami podawanymi na wejścia sieci. Współczynniki mają podobną interpretację jak elementy widma częstotliwości otrzymanego np. przy pomocy FFT, lecz są bardziej odporne na szumy.

EEG – konstrukcja sieci neuronowej

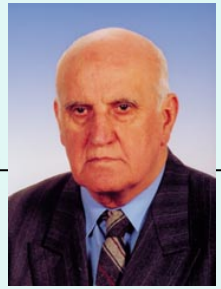
- Topologia: dwuwarstwowy perceptron o architekturze 8-15-3 (8 neuronów w warstwie wejściowej, 15 w warstwie ukrytej i 3 w wyjściowej)
- Algorytm uczący: klasyczny *backpropagation*
- Trafność klasyfikowania: 91%
  - Więcej w: Tsoi A.C., So D.S.C, Sergejew A.: Classification of Electroencephalogram using Artificial Neural Networks. Technical Report, University of Queensland, Australia, March 1993
  - Przejrzyj inne książki, aby poznać więcej zastosowań, np. R.Tadeusiewicza.

Dostępne narzędzia, implementacje metod

- Open source lub freeware
  - **WEKA** – Waikato
  - MOA – Datastreams Waikato
  - Rapid Miner (YALE)
  - Orange Lubljana
  - KNIME Konstanz
  - MLC++ Stanford
  - SSNN Stutgart
  - ROSE i inne (PP Poznań)
- Systemy komercyjne
  - SAS Institute: Enterprise Miner
  - Statistica Data Miner
  - SPSS : Clementine
  - Oracle 9i Miner
  - IBM: QUEST and Intelligent Miner
  - Silicon Graphics: MineSet
  - Review 10 Top Data Mining Tools Edler

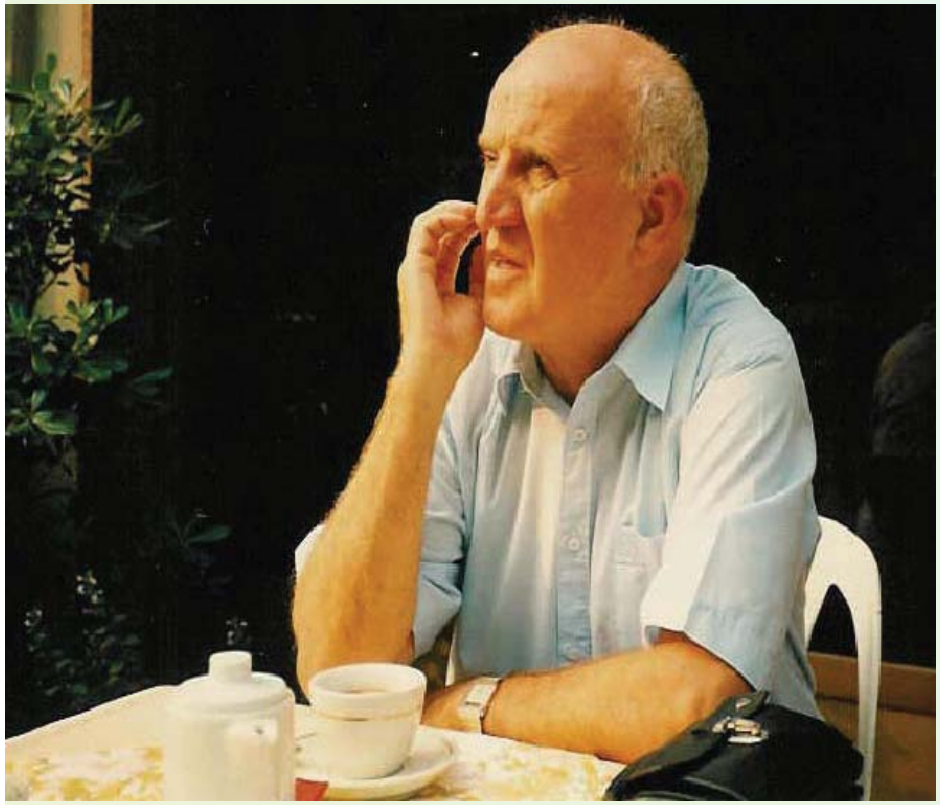


Rough Sets – Zbiory przybliżone



- Zainteresowanie niedoskonałościami i niespójnościami w danych.
  - Inconsistency – niespójność w opisie obiektów nie zawsze jest wynikiem błędów lub „szumu” informacyjnego
  - W niektórych zastosowaniach pożądana jest identyfikacja pewnej części wiedzy pozyskanej z dostępnych danych i oddzielenie jej od części możliwej lub niepewnej.
  - Istnieją różne typy i semantyki niespójności w danych.
- **Rough sets theory** – teoria zbiorów przybliżonych (Z.Pawlak) i jej uogólnienia (np. R.Słowiński, A.Skowron, Y.Yao, S.Greco,…) dostarcza podstaw teoretycznych do analizy niespójnych danych.

Zdzisław I. Pawlak (1926-2006)



- Najbardziej znany na świecie polski naukowiec z zakresu informatyki
- Brał udział w konstrukcji polskich komputerów (latach 50-60)
- Pierwsze prace polskie prace naukowe publikowane w USA (1953)
- Logiczne podstawy informatyki
- Teoria automatów
- Maszyny bezadresowe
- Języki wyszukiwania informacji
- Teoria zbiorów przybliżonych

Rough Sets (2)

- Opiera się na badaniu relacji zachodzących między opisami obiektów.
- Podstawowe relacje: nierozróżnialność, podobieństwo, dominacja
- Jeśli opisy obiektów za pomocą atrybutów są niespójne (sprzeczne), to tworzy się tzw. przybliżenia zbiorów (klas obiektów).
- Rough set – zbiór przybliżonych jest parą przybliżeń zbioru!
- Może być użyta dla oceny przydatności atrybutów dla przybliżenia klasyfikacji, selekcji atrybutów, redukcji tablicy danych, zależności między zbiorami atrybutów, generacji wzorców z danych (reguł, templates, reguł asocjacyjnych), itd.
- Oraz wiele innych …

Przybliżenie klasyfikacji obiektów

U	*Headache*	*Temp.*	*Flu*
U1	Yes	Normal	No
U2	Yes	High	Yes
U3	Yes	Very-high	Yes
U4	No	Normal	No
U5	*No*	*High*	*No*
U6	***No***	***Very-high***	***Yes***
U7	*No*	*High*	*Yes*
U8	***No***	***Very-high***	***No***

Klasy relacji nierozróżnialności (zbiory elementarne) dla zbioru atrybutów $R = \{Headache, Temp.\}$:
$\{u1\}, \{u2\}, \{u3\}, \{u4\}, \{u5, u7\}, \{u6, u8\}$.

$X1 = \{u \mid Flu(u) = yes\}$

$= \{u2, u3, u6, u7\}$

$\underline{R}X1 = \{u2, u3\}$

$\overline{R}X1 = \{u2, u3, u6, u7, \mathbf{u8, u5}\}$

$X2 = \{u \mid Flu(u) = no\}$

$= \{u1, u4, u5, u8\}$

$\underline{R}X2 = \{u1, u4\}$

$\overline{R}X2 = \{u1, u4, u5, u8, \mathbf{u7, u6}\}$

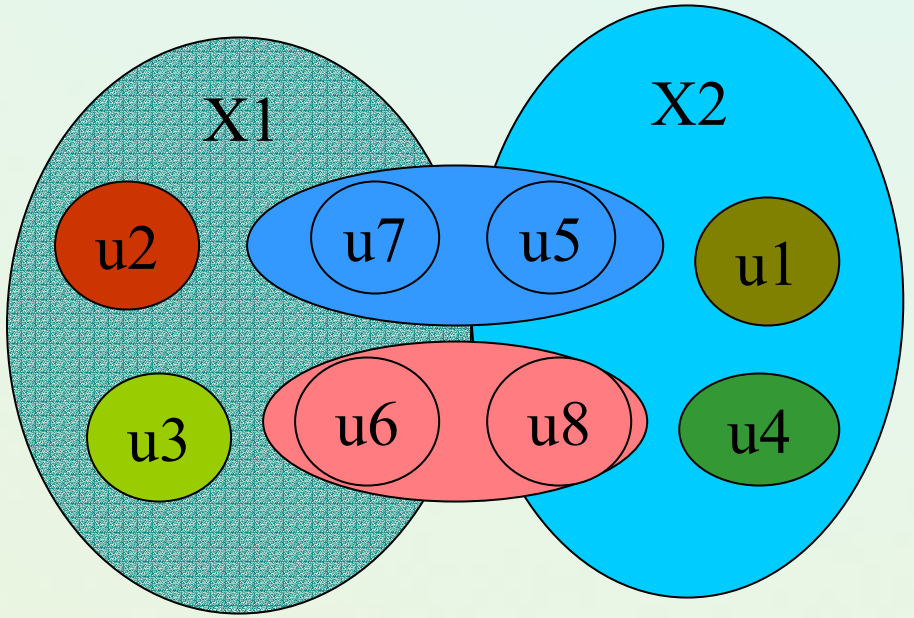
Przybliżenie klas decyzyjnych *X*1 oraz *X*2

R = {Headache, Temp.}
U/R = { {u1}, {u2}, {u3}, {u4}, {u5, u7}, {u6, u8}}

X1 = {u | Flu(u) = yes} = *{u2,u3,u6,u7}*
X2 = {u | Flu(u) = no} = *{u1,u4,u5,u8}*

$\underline{RX1}$ = *{u2, u3}*
$\overline{RX1}$ = *{u2, u3, u6, u7, **u8, u5**}*

$\underline{RX2}$ = *{u1, u4}*
$\overline{RX2}$ = *{u1, u4, u5, u8, **u7, u6**}*



Przykład redukcji tablicy decyzyjnej

U	*Headache*	*Muscle pain*	*Temp.*	*Flu*
U1	Yes	Yes	Normal	No
U2	Yes	Yes	High	Yes
U3	Yes	Yes	Very-high	Yes
U4	No	Yes	Normal	No
U5	No	No	High	No
U6	No	Yes	Very-high	Yes

CORE **= {Headache,Temp}** $\cap$
{MusclePain, Temp}
= {Temp}

Reduct1 **= {Muscle-pain,Temp.}**

U	*Muscle pain*	*Temp.*	*Flu*
U1,U4	Yes	Normal	No
U2	Yes	High	Yes
U3,U6	Yes	Very-high	Yes
U5	No	High	No

Reduct2 **= {Headache, Temp.}**

U	*Headache*	*Temp.*	*Flu*
U1	Yes	Norlmal	No
U2	Yes	High	Yes
U3	Yes	Very-high	Yes
U4	No	Normal	No
U5	No	High	No
U6	No	Very-high	Yes

LEM2 – An Example (1)

U	*Headache*	*Nausea*	*Temp.*	*Flu*
x1	no	no	normal	No
x2	yes	no	high	Yes
x3	yes	yes	high	Yes
x4	yes	no	normal	No
x5	no	no	high	No
x6	no	no	high	Yes

IND: {x1}, {x2}, {x3}, {x4}, {x5,x6}
YES: lower appr. {x2,x3}
upper {x2,x3,x5,x6}
NO: lower approx. {x1,x4}
upper {x1,x4,x5,x6}

Reguły (pewne) dla dolnego przybliżenia klasy **(Flue=Yes)** – przykłady pozytywne {x2,x3}

(headache,yes)	{x2,x3+ ; x4-}
(nausea,no)	{x2+ ; x1,x4,x5,x6-}
(nausea,yes)	{x3+ }
(temperature,high)	{x2,x3+ ; x5,x6-}

Wybierz t1 (headache,yes), ale {x2,x3+ ; x4-} $\not\subset$ {x2,x3}, poszukuj następnego warunku; wybierz (temperature,high),
Teraz t1$\cap$t2= {x2,x3+ ; x4-} $\cap$ {x2,x3+ ; x5,x6-} $\subseteq$ {x2,x3}
reguła **(headache=yes) $\cap$ (temperature=high) $\rightarrow$(Flue=Yes)** pokrywa wszystkie przykłady pozytywne.

LEM2 – An Example (2)

U	*Headache*	*Nausea*	*Temp.*	*Flu*
x1	no	no	normal	No
x2	yes	no	high	Yes
x3	yes	yes	high	Yes
x4	yes	no	normal	No
x5	no	no	high	No
x6	no	no	high	Yes

IND: {x1}, {x2}, {x3}, {x4}, {x5,x6}
YES: lower appr. {x2,x3}
upper {x2,x3,x5,x6}
NO: lower approx. {x1,x4}
upper {x1,x4,x5,x6}

Reguły pewne **(Flue=No)**: Dolne przybliżenie {x1,x4}

(headache,no}	{x1+; x5,x6-}
(headache,yes)	{x4+ ; x2,x3-}
(nausea,no)	{x1,x4+;x2,x5,x6-}
(temperature,normal)	{x1,x4+ ; ∅}

Wybierz t_1 (temperature,normal),
t1= {x1,x4+ ; ∅-} $\subseteq$ {x1,x4}
Reguła **(temperature=normal) →(Flue=No)** pokrywa wszystkie przykłady pozytywne.

Przykładowe zastosowania teorii zbiorów przybliżonych w analizie medycznych danych.

- Klasyfikacja pacjentów z chorobą wrzodową dwunastnicy leczonych metodą wysoce wybiórczej wagotomii.
- Wieloetapowe leczenie ostrego zapalenia trzustki płukaniem otrzewnej.
- Leczenie kamicy moczowej za pomocą ESWL.
- Analiza leczenia pacjentów z obrażeniami wielo-narządowymi przyjmowanych w izbie przyjęć.
- Weryfikacja diagnozowania uszkodzeń stawu kolanowego na podstawie MRI.
- Rokowanie przeżycia przy leczeniu raka piersi.
- Wspomaganie decyzji w przypadku bólów brzucha u dzieci w izbie przyjęć.
- …

- Więcej w: Z.Pawlak, K.Słowiński, J.Stefanowski: Teoria zbiorów przybliżonych w analizie danych medycznych, 2002.

Leczenie pacjentów z chorobą wrzodową dwunastnicy

- Celem było określenie wskazań do przeprowadzania zabiegu chirurgicznego wysoce wybiórczej wagotomii (*HSV*) dla pacjentów cierpiących na chorobę wrzodową dwunastnicy.
- Dane: Pacjenci opisani są za pomocą 11 atrybutów i przydzieleni do 4 klas wyrażających skuteczność zabiegu chirurgicznego. Klasy w sensie liczności są niezrównoważone (silna przewaga grupy wyleczonych pacjentów). Atrybuty są różnego typu (nominalnego, porządkowego jak i liczbowego).

Leczenie pacjentów z chorobą wrzodową dwunastnicy

Aspekty metodologiczne:

- Postawione zadania:
  - ocena znaczenia atrybutów i ich podzbiorów dla przybliżenia klasyfikacji pacjentów;
  - poszukiwanie tzw. modeli pacjentów charakterystycznych dla klas decyzyjnych;
  - poszukiwanie reprezentacji zależności pomiędzy wartościami wybranych atrybutów warunkowych a decyzyjnym w postaci reguł decyzyjnych.
- Zastosowano teorię zbiorów przybliżonych oraz metody indukcji reguł.
- Na podstawie rezultatów analizy danych oraz wiedzy o leczeniu sformułowano algorytmy kliniczne.

Analiza MRI zdjęć uszkodzeń stawu kolanowego

- **Sformułowanie problemu**: Problem dotyczył weryfikacji klinicznej diagnoz pewnej formy uszkodzenia stawu kolanowego (*anterior cruciate ligament* – ACL) postawionych na podstawie symptomów zaobserwowanych na zdjęciach wykonanych techniką magnetycznego rezonansu jądrowego (MRI). Procedura chirurgiczna artroskopii umożliwia bezpośrednią weryfikacje uszkodzenia stawu, lecz jest techniką inwazyjną i wykonana niepoprawnie może prowadzić do komplikacji.

Dane i ich charakterystyka:

- Wybrano grupę 140 pacjentów, dla których wykonano zarówno standardowe procedury diagnostyczne i badanie za pomocą MRI jak i artroskopię. Dla 100 pacjentów wykluczono, a dla 40 potwierdzono występowanie uszkodzenia ACL.
- Przy współpracy z ekspertami medycznymi oraz na podstawie wstępnej analizy statystycznej wyselekcjonowano zbiór 6 atrybutów opisujących pacjentów. Większość atrybutów wynikała z pomiarów przeprowadzonych na zdjęciu MRI (PCL index; pomiary X i Y). Pozostałe (wiek, płeć, strona ciała)

Przetwarzanie wstępne - dyskretyzacja

Table 1. Comparison of various discretizations

Evaluation measures	*Expert*	*MI*	*ME*	*MD*
Rough sets results:				
accuracy of class 1	0.653	0.822	0.951	1.0
accuracy of class 2	0.843	0.922	0.980	1.0
quality of classification	0.879	0.943	0.987	1.0
no. of attributes in a core	6	6	4	4
number of reducts	1	1	2	1
LEM2 results:				
number of rules	23	20	16	20
total classification accuracy	85%	88.86%	92.86%	90%
sensitivity of Class 1	67.2%	80%	90%	87.5%
average rule strength (obj.)	9.65	8.57	10.62	4.93
average rule length	3.67	2.67	2.47	2.73

AGE (years): [min,16.5), [16.5,35), [35,max]
X (mm): [min,8.5), [8.5,11.75), [11.75,14.5), [14.5,max]
Y (mm): [min,2.75), [2.75,3.75), [3.75,4.75) [4.75,max]
PCLINDEX: [min,3.225), [3.225,3.71), [3.71,4.125), [4.125,4.535), [4.535,max]

Przybliżenia klasyfikacji

- Jakość przybliżenia = 0.9429

Table 2. Rough approximations of the patients' classification

class	number of patients	cardinality of lower approx.	cardinality of upper approx.	accuracy of class
1	40	39	41	0.9512
2	100	99	101	0.9802

Znaczenie atrybutów

Table 3. Significance of attributes of approximating the patients' classification

Attribute	AGE	SEX	SIDE	X	Y	PCLINDEX
Significance	0.057	0.043	0.021	0.00	0.00	0.071

- Rdzeń:
- PCLINDEX,AGE, SEX, SIDE

Minimalny zbiór reguł

- MODLEM

rule 1. *if* (PCLINDEX $<$ 3.225) *then* Class1 [26, 65%]
rule 2. *if* (AGE=[16.5,35))$\wedge$(PCLINDEX=[3.225,3.71)) *then* Class1 [6, 15%]
rule 3. *if* (SEX=MALE) $\wedge$ (SIDE=RIGHT) $\wedge$ (PCLINDEX=[3.225,3.71)) *then* Class1 [3, 7.5%]
rule 4. *if* (AGE=[16.5,35)) $\wedge$ (PCLINDEX=[3.71,4.125)) $\wedge$ (X$\geq$14.5) *then* Class1 [2, 5%]
rule 5. *if* (X=[8.5,11.75)) $\wedge$ (PCLINDEX=[4.125,4.535)) $\wedge$ (SEX=MALE) *then* Class1 [1, 2.5%]
rule 6. *if* (X=[8.5,11.75)) $\wedge$ (PCLINDEX=[3.225,3.71)) $\wedge$ (AGE$\geq$ 35) *then* Class1 [2, 5%]
rule 7. *if* (PCLINDEX=[3.71,4.125)) $\wedge$ (X=[8.5,11.75)) $\wedge$ (SEX=1) *then* Class1 [1, 2.5%]
rule 8. *if* (PCLINDEX$\geq$4.535) *then* Class2 [75, 75%]
rule 9. *if* (SEX=FEMALE) $\wedge$ (PCLINDEX=[4.125,4.535)) *then* Class2 [10,10%]
rule 10. *if* (PCLINDEX=[3.71,4.125)) $\wedge$ (AGE$\geq$ 35) *then* Class2 [6,6%]
rule 11. *if* (X=[11.75,14.5)) $\wedge$ (Y=[2.75,3.75)) $\wedge$ (SEX=FEMALE) *then* Class2 [8, 8%]
rule 12. *if* (SIDE=LEFT) $\wedge$ (X=[11.75,14.5)) $\wedge$ (Y=[2.75,3.75)) *then* Class2 [7, 7%]
rule 13. *if* (PCLINDEX=[3.225,3.71)) $\wedge$ (AGE$\geq$ 35) $\wedge$ (SEX=MALE) *then* Class2 [2, 2%]
rule 14. *if* (AGE$<$16.5) *then* Class2 [14, 14%]
rule 15.*if* (PCLINDEX=[3.225,3.71))$\wedge$(Y=[3.75,4.75))$\wedge$(AGE$\geq$ 35)$\wedge$(SIDE=LEFT) *then* Class2 [1,1%]

Satysfakcjonujący zbiór reguł (EXPLORE)

- Rule. Strength ≥ 10% Klasy

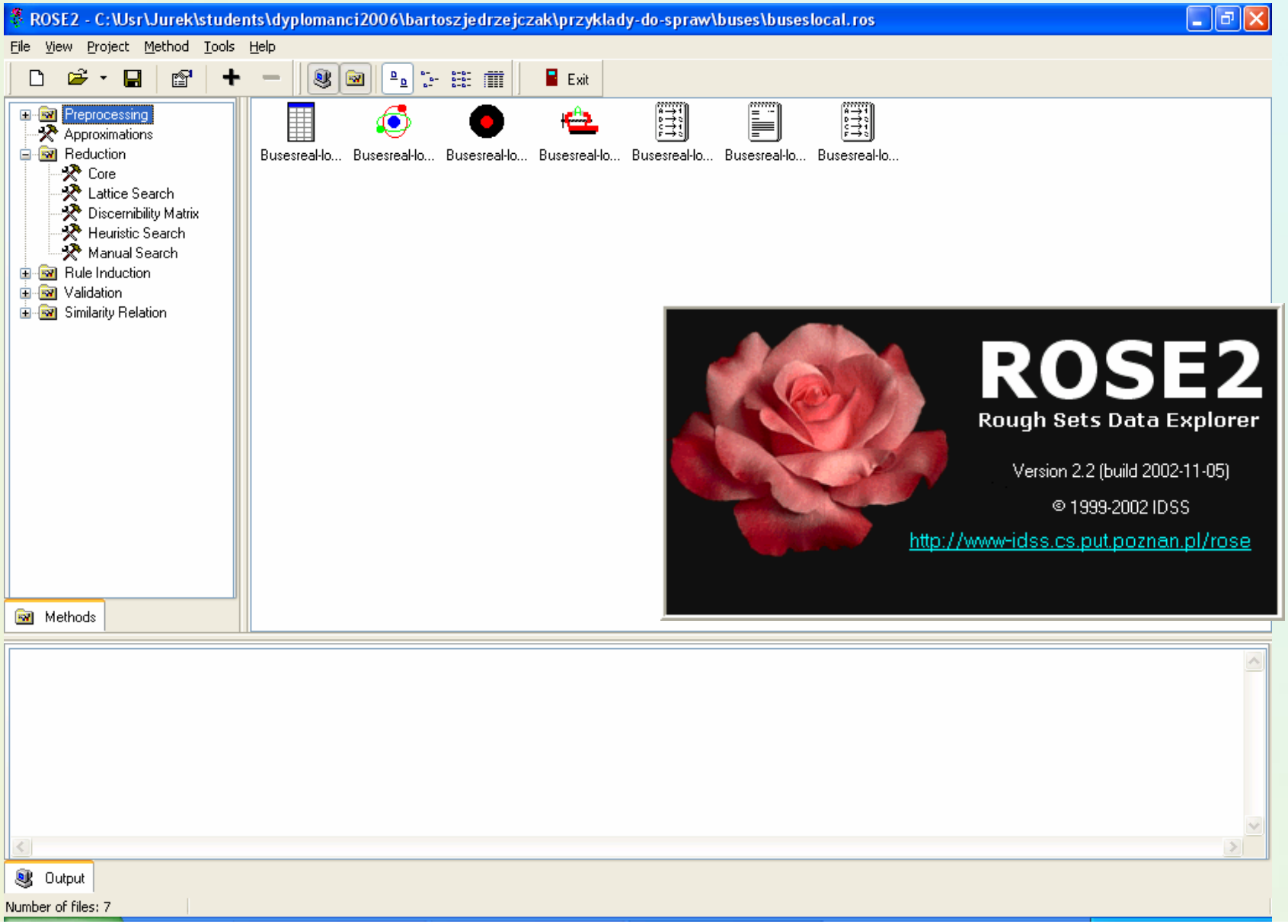
rule 1. *if* (X < 8.5) *then* Class1 [4, 10%]
rule 2. *if* (PCLINDEX < 3.225) *then* Class1 [26, 65%]
rule 3. *if* (AGE=[16.5,35)) ∧ (Y ≥ 4.75) *then* Class1 [6, 15%]
rule 4. *if* (AGE=[16.5,35)) ∧ (PCLINDEX = [3.225,3.71)) *then* Class1 [6, 15%]
rule 5. *if* (SEX = MALE) ∧ (Y ≥ 4.75) *then* Class1 [9, 22.5%]
rule 6. *if* (SIDE = RIGHT) ∧ (Y ≥ 4.75) *then* Class1 [7, 17.5%]
rule 7. *if* (X = [8.5,11.75)) ∧ (Y = [3.75,4.75)) *then* Class1 [9, 22.5%]
rule 8. *if* (X = [11.75,14.5)) ∧ (Y ≥ 4.75) *then* Class1 [5, 12.5%]
rule 9. *if* (AGE=[16.5,35)) ∧ (X = [8.5,11.75)) ∧ (Y =[2.75,3.75)) *then* Class1 [5, 12.5%]
rule 10. *if* (AGE=[16.5,35)) ∧ (X = [11.75,14.5)) ∧ (Y =[3.75,4.75)) *then* Class1 [4, 10%]
rule 11. *if* (SEX = MALE) ∧ (X = [8.5,11.75)) ∧ (Y =[2.75,3.75) *then* Class1 [6, 15%]
rule 12. *if* (SEX = MALE) ∧ (X = [8.5,11.75)) ∧ (PCLINDEX =[3.225,3.71)) *then* Class1 [4, 10%]
rule 13. *if* (SEX = MALE) ∧ (Y = [2.75,3.75)) ∧ (PCLINDEX= [3.225,3.71)) *then* Class1 [4, 10%]
rule 14. *if* (SIDE = RIGHT) ∧ (X = [8.5,11.75)) ∧ (Y = [2.75,3.75)) *then* Class1 [5, 12.5%]
rule 15. *if* (AGE<16.5) *then* Class2 [14, 14%]
rule 16. *if* (PCLINDEX ≥ 4.535) *then* Class2 [75, 75%]
rule 17. *if* (AGE≥ 35) ∧ (X ≥ 14.5) *then* Class2 [16, 16%]
rule 18. *if* (AGE≥ 35) ∧ (Y < 2.75) *then* Class2 [14, 14%]
rule 19. *if* (SEX = FEMALE) ∧ (X ≥ 14.5) *then* Class2 [28, 28%]
rule 20. *if* (SEX = FEMALE) ∧ (PCLINDEX = [4.125,4.535)) *then* Class2 [10, 10%]
rule 21. *if* (SIDE = RIGHT) ∧ (Y < 2.75) *then* Class2 [19, 19%]
rule 22. *if* (SIDE = LEFT) ∧ (X ≥ 14.5) *then* Class2 [24, 24%]
rule 23. *if* (X ≥ 14.5) ∧ (Y = [2.75,3.75)) *then* Class2 [29, 29%]

Ocena zdolności klasyfikacyjnych

Table 4. Classification performance of decision rules

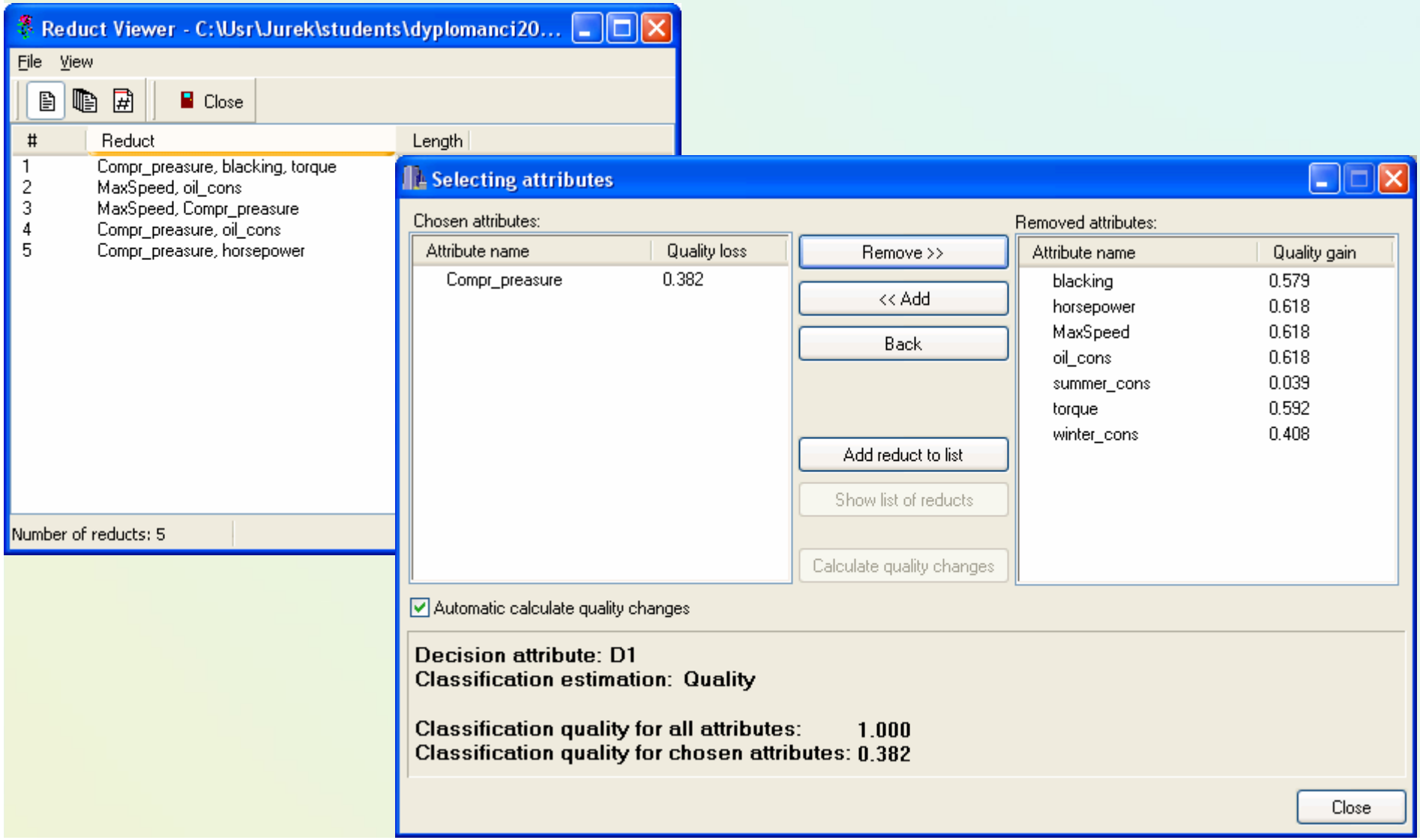
Rule set	Overall	Class 1		Class 2	
	accuracy	sensitivity	specificity	sensitivity	specificity
minimum	92.86%	90%	94%	94%	90%
satisfactory	93.57%	85%	97%	97%	85%

Oprogramowanie do TZP → ROSE (IDSS)



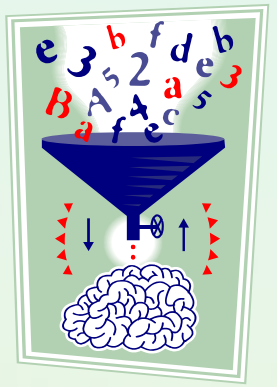
Rose → redukcja

- Redukt vs. significance for classification



Podsumowanie:

- Nowe technologie dostarczają wielu danych
  - data mining pomaga odnaleźć interesujące zależności
- Data Mining i Uczenie Maszynowe
- Wiele prac, ale
  nie oczekujmy masowych „automatycznych”
  zastosowań ?
- Podstawowe zadania:
  - Klasyfikacja nadzorowana, grupowanie, …
- Omówione metody:
  - Drzewa i reguły, klasyfikacja bayesowska,
    sztuczne sieci neuronowe,
    teoria zbiorów przybliżonych (mniej ML bardziej …)



Czy są jakieś pytania?

I to by było na tyle …

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