Technologie Informacyjne

Przygotowanie danych

Adam Krasuski

Szkoła Główna Służby Pożarniczej
Zakład Informatyki i Łączności

December 5, 2016

Dane tabelaryczne

ID	data	# GBA	pożar	obiekt	piętro	powierzchnia
1	10.02.2012	3	tak	blok	IV	40
2	12.12.2012	1	tak	dom	I	120
3	12.10.2010	1	nie	samochód	-	4
4	15.02.2011	5	tak	fabryka	0	1200
5	13.12.2013	2	nie	śmietnik	-	0,4
6	12.11.2012	4	tak	mieszkanie	VII	38
7	17.12.2002	1	tak	hala	0	1210
8	21.02.2001	5	tak	garaż	0	1250

Data i czas

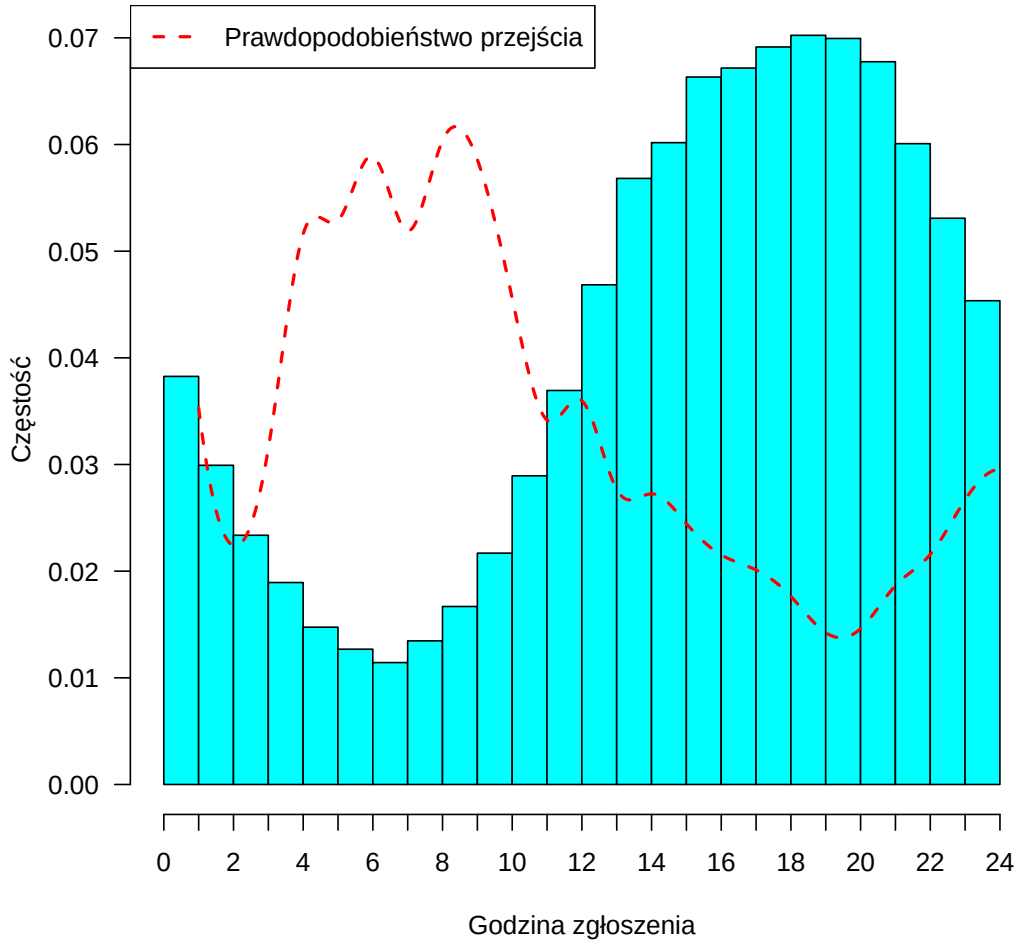
Różne formaty reprezentacji danych:

- 14.12.2012
- 14.12.12
- 12.12.14
- 14 grudnia 2012
- 14 grudnia 2012 20:30

Porównanie pomiędzy poszczególnymi składowymi:

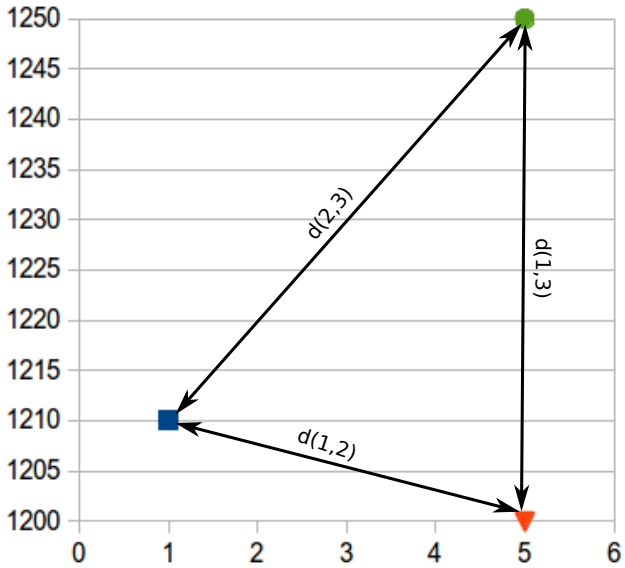
- 14.**12**.2012 – 01.**12**.1990 $\rightarrow$ grudzień
- 14.10.2012 **20**:30 – 01.12.1990 **20**:12 $\rightarrow$ godzina 20-sta

Data i czas


Prawdopodobieństwo przejścia
Częstość
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
2
4
6
8
10
12
14
16
18
20
22
24
Godzina zgłoszenia

Atrybuty numeryczne

ID	# GBA	obiekt	pow.
1	5	fabryka	1200
2	1	hala	1210
3	5	garaż	1250



$$d(1,2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} = 10$$
$$d(1,3) = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} = 50$$
$$d(2,3) = \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2} = 40$$

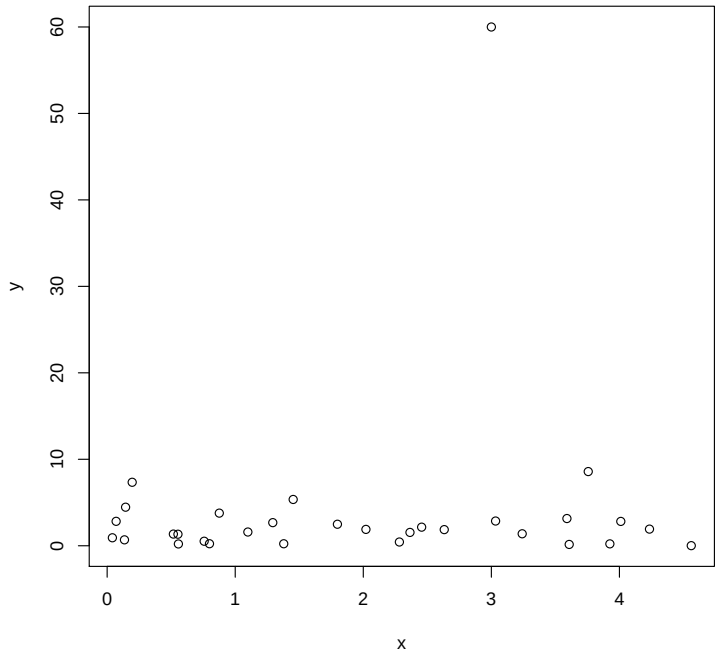
Normalizacja

Wszystkie atrybuty opisujące obiekt mają taką samą wartość minimalną jak i maksymalną – opisane są na tej samej skali.
Przykład: dzielenie przez wartość maksymalną danego atrybutu.

ID	# GBA	# GBA norm.	obiekt	pow.	pow. norm.
1	**5**	1	fabryka	1200	0,96
2	1	0,2	hala	1210	0,97
3	5	1	garaż	**1250**	1

$d\_norm(1,2) = 0,8$
$d\_norm(1,3) = 0,04$
$d\_norm(2,3) = 0,8$

Standaryzacja



$$z = \frac{x - \mu}{\sigma}$$

ID	# GBA	# GBA stand.	obiekt	pow.	pow. stand.
1	5	0,58	fabryka	1200	-0,76
2	1	-1,15	hala	1210	-0,38
3	5	0,58	garaż	1250	1.13

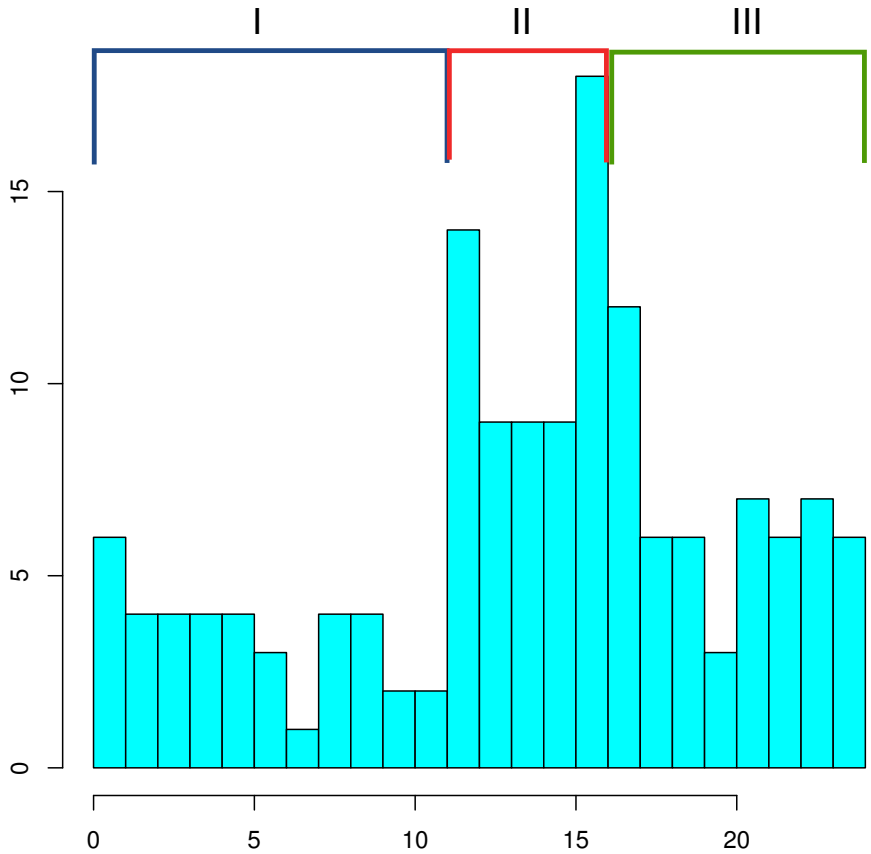
$$d\_stand(1, 2) = 1, 78$$
$$d\_stand(1, 3) = 0, 89$$
$$d\_stand(2, 3) = 2, 29$$

Dyskretyzacja

ID	# GBA	# GBA dyskr.	obiekt	pow.	pow. dyskr.
1	3	średnia	blok	40	średnia
2	1	mała	dom	120	duża
3	1	mała	samochód	4	średnia
4	5	duża	fabryka	1200	b. duża
5	2	małą	śmietnik	0,4	mała
6	4	średnia	mieszkanie	38	średnia
7	1	mała	hala	1210	b. duża
8	5	duża	garaż	1250	b. duża

GBA: [1 , 2] $\rightarrow$ mała; [3, 4] $\rightarrow$ średnia; [5, ∞) $\rightarrow$ duża;
powierzchnia: [0 , 2] $\rightarrow$ mała; [3, 40] $\rightarrow$ średnia; [41,150] $\rightarrow$ duża; [151, ∞) $\rightarrow$ bardzo duża;

Dyskretyzacja

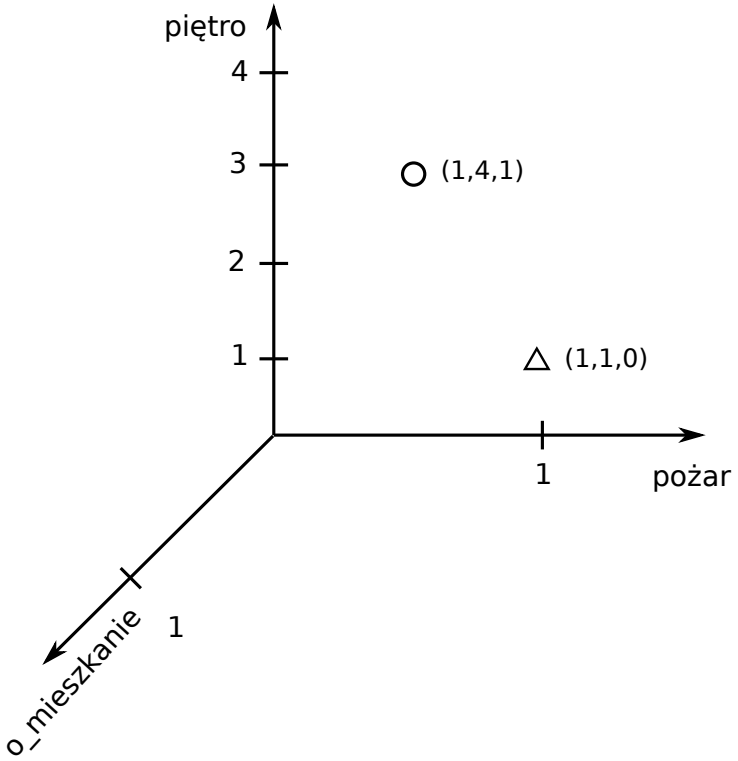


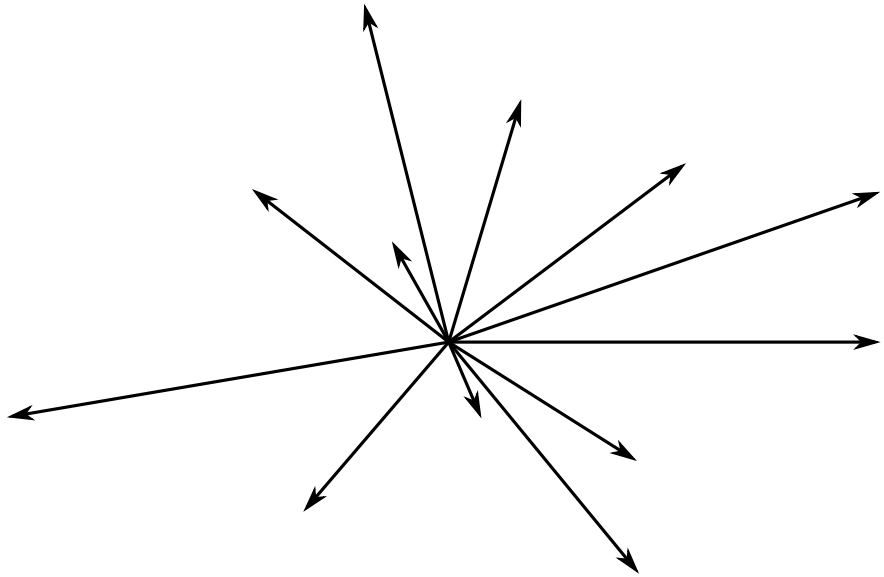
Atrybuty nominalne i porządkowe

ID	pożar	obiekt	piętro
1	tak	mieszkanie	IV
2	tak	dom	I
3	nie	samochód	-
4	tak	fabryka	0
5	nie	śmietnik	-
6	tak	mieszkanie	VII
7	tak	hala	0
8	tak	garaż	0

ID	pożar	o_mieszkanie	o_dom	o_samochód	o_fabryka	o_śmietnik	o_hala	o_garaż	piętro
1	1	1	0	0	0	0	0	0	4
2	1	0	1	0	0	0	0	0	1
3	0	0	0	1	0	0	0	0	-
4	1	0	0	0	1	0	0	0	0
5	0	0	0	0	0	1	0	0	-
6	1	1	0	0	0	0	0	0	7
7	1	0	0	0	0	0	1	0	0
8	1	0	0	0	0	0	0	1	0

Atrybuty nominalne i porządkowe





Puste i błędne wartości

ID	pożar	obiekt	piętro
1	tak	mieszkanie	IV
2	tak	dom	I
3	nie	samochód	
4	tak	fabryka	0
5	nie	śmietnik	brak
6	nie wiem	mieszkanie	VII
7	tak	hala	0
8	tak	garaż	0

- zastąpienie średnią lub najczęściej występującą wartością
- zbudowanie modelu wpisywania wartości
- wartość losowa
- usunięcie wiersza

Dane tekstowe

Przykład:
Po dojechaniu na miejsce zdarzenia stwierdzono że w mieszkaniu na I piętrze doszło do wybuchu wskutek którego wypadło okno wraz z futryną. Po dokładnym rozpoznaniu stwierdzono że w pomieszczeniu łazienki doszło do **pozaru** pralki automatycznej z bliżej nieokreślonych przyczyn.Pralkę ugaszono przy pomocy wody po uprzednim odłączeniu napięcia.W obrębie pralki znaleziono rozerwane opakowanie po dezodorancie które mogło spowodować wybuch i powstanie fali uderzeniowej.Ze względu na niemożliwość jednoznacznego określenia przyczynu powstania eksplozji na miejsce zadysponowano ekipę policji z KP Białołęka której przekazano pomieszczenie wraz z pralką.Żadna z osób znajdujących się w lokalu nie odniosła obrażeń. Po zakończeniu **działan** oddymiono klatkę schodową.Spaleniu uległa pralka automatyczna okopceniu ściany mieszkania.

Reprezentacja dokument-słowo

Term-document-matrix (TDM)

słowo/raport	Raport 1	Raport 2	Raport 3	Raport n
afrykański	1	0	0	0
agresywnie	1	0	0	0
akademik	1	0	0	1
akumulator	0	1	0	0
albert	0	0	1	0
alkoholowy	1	0	0	1
alkomat	1	0	0	0
altanka	0	0	1	0
antywłamaniowy	0	0	1	0
asfalt	0	1	0	0
...	0	0	0	0

Częstość słów

Miary częstości słów:

- **TF** (term frequency) częstość słowa – liczba powtórzeń słowa w dokumencie do liczby wszystkich słów w dokumencie.
- **IDF** (inverse document frequency) odwrotna częstość dokumentu – liczba dokumentów w korpusie (zbiorze) do liczby dokumentów, w którym dane słowo wystąpiło.
- **TF-IDF** = TF x IDF

TDM

słowo/raport	Raport 1	Raport 2	Raport 3	Raport n
afrykański	0.01	0.00	0.04	0.00
agresywnie	0.30	0.03	0.00	0.00
akademik	0.20	0.40	0.00	0.30
akumulator	0.00	0.00	0.00	0.00
albert	0.07	0.00	0.00	0.70
alkoholowy	0.20	0.50	0.00	0.00
alkomat	0.10	0.00	0.00	0.00
altanka	0.00	0.00	0.02	0.01
antywłamaniowy	0.00	0.00	0.00	0.00
asfalt	0.00	0.00	0.05	0.00

Odmiana wyrazów oraz stop lista

Po dojechaniu na miejsce zdarzenia stwierdzono że w mieszkaniu na I piętrze doszło do **wybuchu** wskutek którego wypadło okno wraz z futryną. Po dokładnym rozpoznaniu stwierdzono że w pomieszczeniu łazienki doszło do pożaru **pralki** automatycznej z bliżej nieokreślonych przyczyn. **Pralkę** ugaszono przy pomocy wody po uprzednim odłączeniu napięcia.W obrębie **pralki** znaleziono rozerwane opakowanie po dezodorancie które mogło spowodować **wybuch** i powstanie fali uderzeniowej. Ze względu na niemożliwość jednoznacznego określenia przyczyn powstania eksplozji na miejsce zadysponowano ekipę policji z KP Białołęka której przekazano pomieszczenie wraz z **pralką**. Żadna z osób znajdujących się w lokalu nie odniosła obrażeń. Po zakończeniu działań oddymiono klatkę schodową.Spaleniu uległa **pralka** automatyczna okopceniu ściany mieszkania.

Lematyzacja

dojechać miejsce zdarzyć stwierdzić mieszkać piętro dojść wybuch wskutek wypaść okno futryna. dokładny rozpoznać stwierdzić łazienka dojść pożar **pralka** automatyczny blisko nieokreślony przyczyna. **pralka** ugasić pomoc woda uprzedni odłączyć napiąć. obrąb **pralka** znaleźć rozerwać opakować dezodorant móc spowodować wybuch powstać fala uderzeniowy. wzgląd niemożliwość jednoznaczny określić powstać eksplozja miejsce zadysponować ekipa policja kp przekazać pomieścić **pralka**. osoba znajdywać siebie lokal odnieść obrazić. zakończyć oddymić klatka schodowy. spalić ulec **pralka** automatyczny okopcieć ściana mieszkać.

Lematyzacja cd.

Lematyzacja - pojęcie to oznacza sprowadzenie grupy wyrazów stanowiących odmianę danego zwrotu do wspólnej postaci, umożliwiającej traktowanie ich wszystkich jako te samo słowo.

poszedł

dopuszczalne w grach

▾ **pójść**

pójść	
występowanie:	Uniwersalny słownik języka polskiego - PWN 2003, 2006, 2008 - S. Dubisz
odmienność:	tak
B	pójdź, pójdźcie, pójdźcież, pójdźmy, pójdźmyż, pójdźże
I	pójdą, pójdę, pójdzie, pójdziecie, pójdziemy, pójdziesz
(+b) i	niepójścia, niepójściach, niepójściami, niepójście, niepójściem, niepójściom, niepójściu, niepójść, pójścia, pójściach, pójściami, pójście, pójściem, pójściom, pójściu, pójść
~	poszedł, poszedłby, poszedłbym, poszedłbyś, poszedłem, poszedłeś, poszedłszy, poszli, poszliby, poszlibyście, poszlibyśmy, poszliście, poszliśmy, poszła, poszłaby, poszłabym, poszłabyś, poszłam, poszłaś, poszło, poszłoby, poszły, poszłyby, poszłybyście, poszłybyśmy, poszłyście, poszłyśmy
aktualizacja:	tevex, 2010-09-09

Podobieństwo semantyczne

dojechać miejsce zdarzyć stwierdzić mieszkać piętro dojść **wybuch** wskutek wypaść okno futryna. dokładny rozpoznać stwierdzić łazienka dojść pożar pralka automatyczny blisko nieokreślony przyczyna. pralka ugasić pomoc woda uprzedni odłączyć napiąć. obrąb pralka znaleźć rozerwać opakować dezodorant móc spowodować **wybuch** powstać fala uderzeniowy. wzgląd niemożliwość jednoznaczny określić powstać **eksplozja** miejsce zadysponować ekipa policja kp przekazać pomieścić pralka.

Analiza ukrytych grup semantycznych

słowo/raport	Raport 1	Raport 2	Raport 3	Raport 4
wybuch	1	0	0	1
samochód	0	1	0	0
eksplozja	1	0	0	0
GBA	0	1	0	0
wyrzut	0	0	1	1
pożar	1	0	1	0
gaśniczy	0	1	0	0

Analiza ukrytych grup semantycznych

Latent Semantic Analysis:

słowo/raport	pojęcie 1	pojęcie 2	pojęcie 3
wybuch	0,25	-0,12	0,03
samochód	-0,11	0,19	-0,05
eksplozja	0,13	-0,70	-0.02
GBA	-0,77	0,22	0,01
wyrzut	0,02	-0,51	0,07
pożar	0,00	0,01	-0.12
gaśniczy	-0,07	0,32	0.01

7,11	0	0
0	2,31	0
0	0,00	0

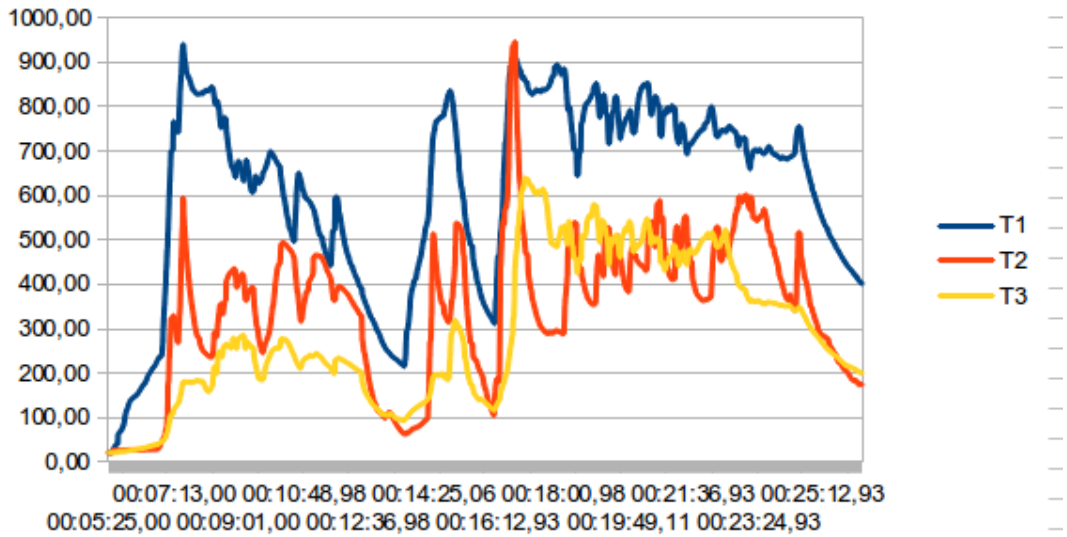
pojęcie/raport	Raport 1	Raport 2	Raport 3	Raport 4
pojęcie 1	0,31	-0,43	0,44	0,21
pojęcie 2	-0,70	0,61	-0,24	-0,33
pojęcie 3	-0,01	0,02	0,02	0,02

Dane sensoryczne

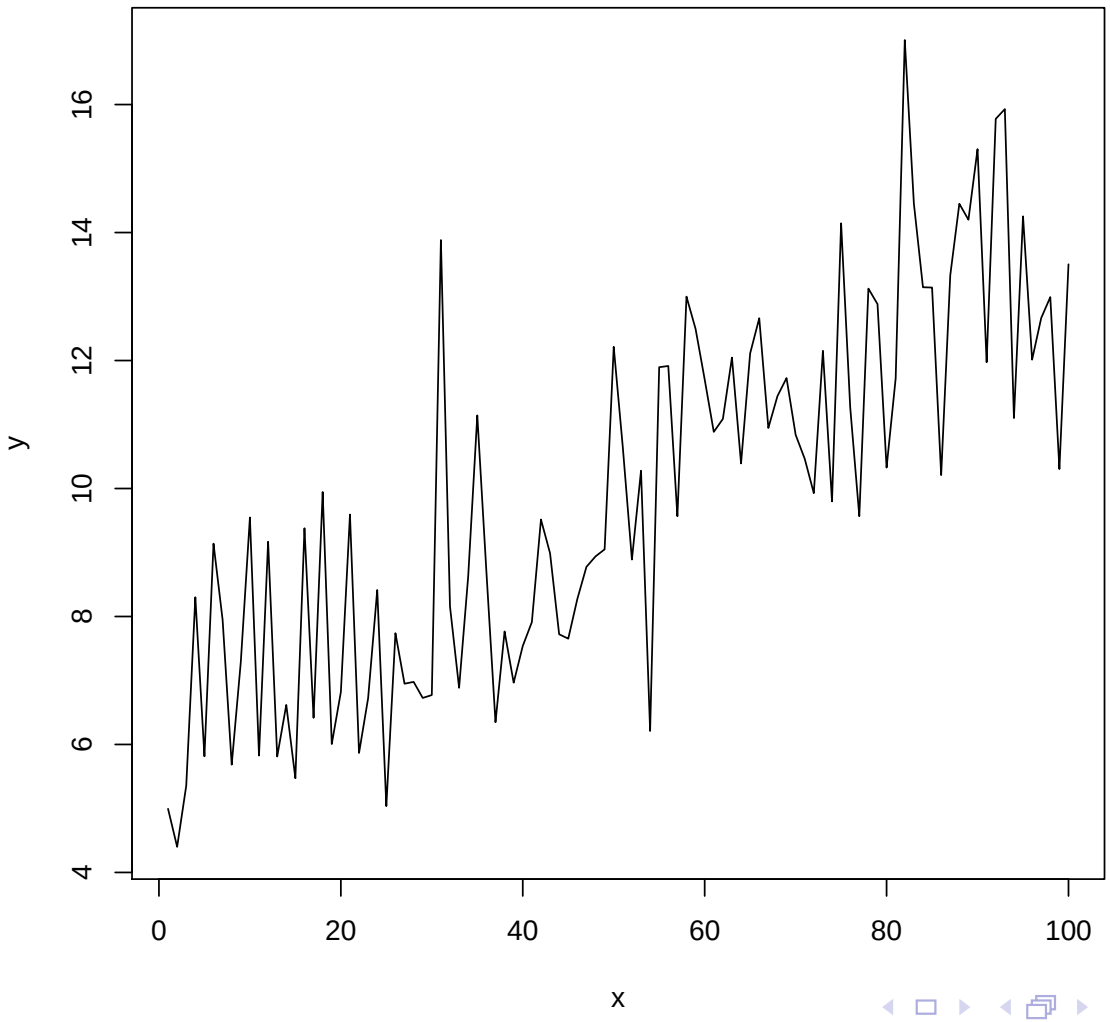
Film

Dane sensoryczne

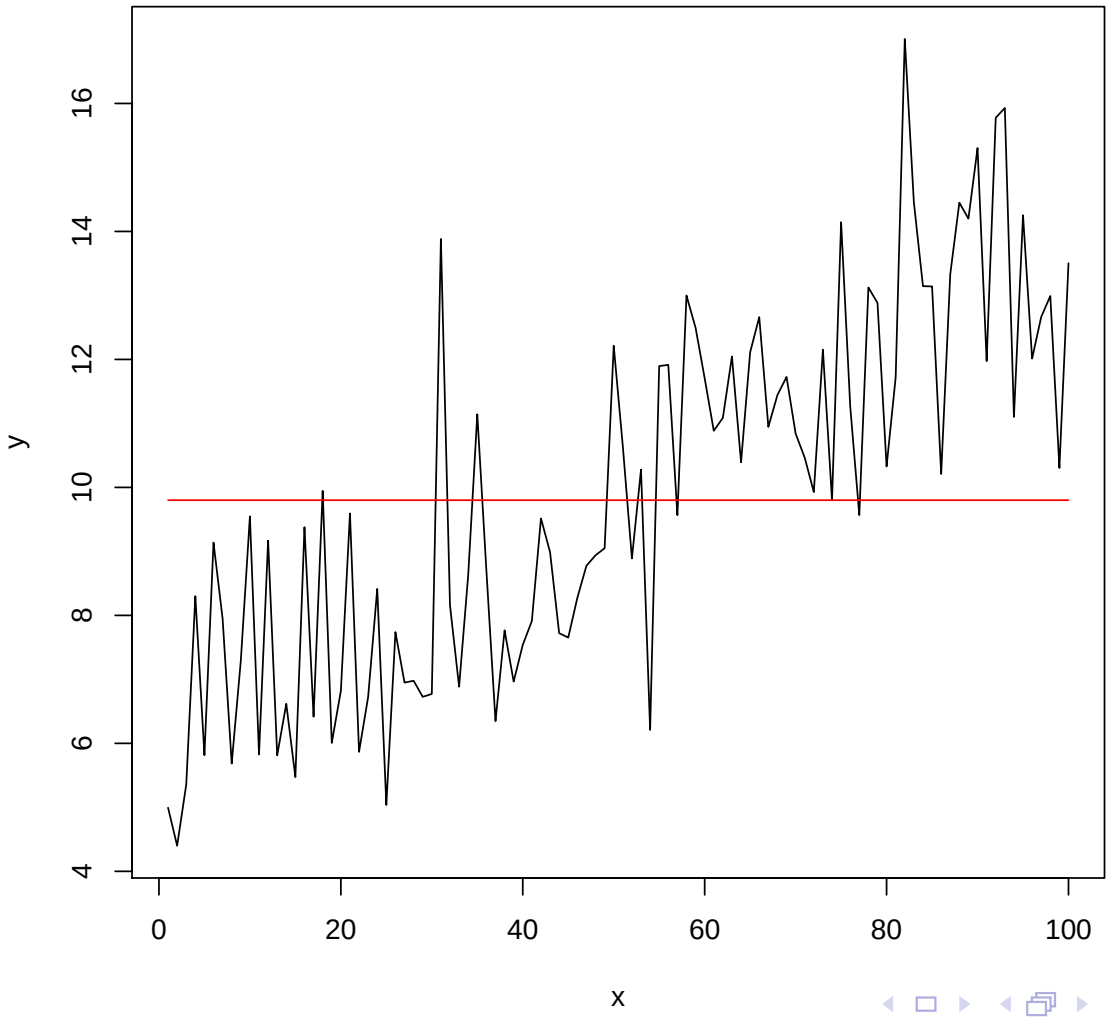
Czas	T1	T2	T3
00:05:23,00	21,10	21,10	21,10
00:05:24,06	21,10	21,10	21,10
00:05:25,00	21,10	21,10	21,10
00:05:26,00	21,10	21,10	21,10
00:05:27,00	21,10	21,10	21,10
00:05:28,00	21,10	21,40	21,10
00:05:29,00	21,10	21,40	21,10
00:05:30,00	21,10	21,40	21,10
00:05:31,00	21,10	21,40	21,10
00:05:32,00	28,70	21,40	21,10
00:05:33,00	31,70	28,00	21,40
00:05:34,00	34,10	27,20	21,40
00:05:35,00	36,70	27,00	21,40
00:05:36,00	39,50	26,80	21,40
00:05:37,00	39,50	26,50	21,60
00:05:38,00	39,50	26,50	21,60
00:05:39,00	39,50	26,50	21,60
00:05:40,00	62,70	26,30	23,20
00:05:41,00	64,80	26,30	23,20
00:05:42,00	66,90	26,30	23,50
00:05:43,00	68,10	26,30	23,70
00:05:44,00	68,80	26,30	23,70
00:05:45,00	69,00	26,10	23,70
00:05:46,00	69,00	25,80	24,00
00:05:47,00	76,70	26,10	24,00
00:05:48,00	79,00	26,30	24,00
00:05:49,00	83,60	26,30	24,00
00:05:50,00	89,00	26,50	24,00
00:05:51,00	97,70	26,80	24,20



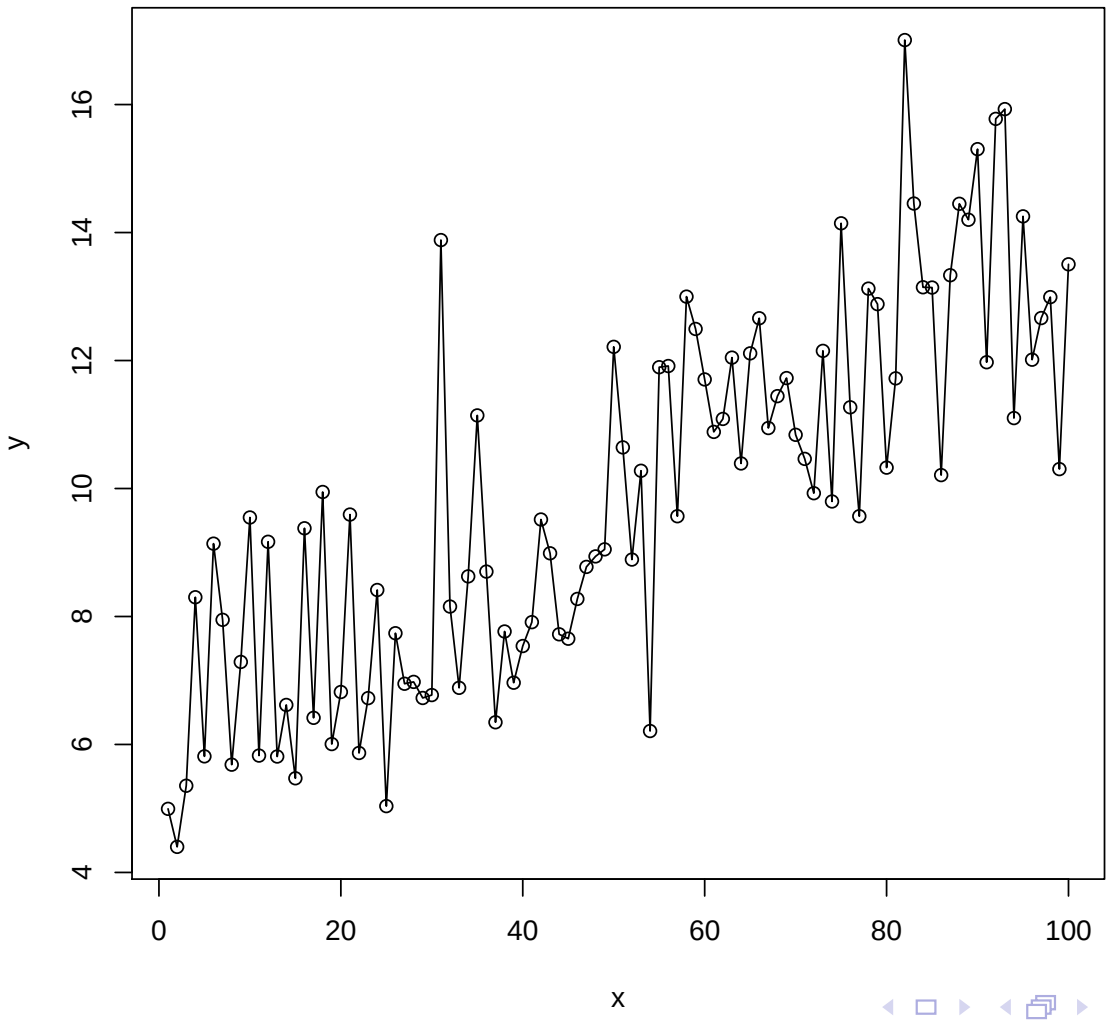
Dane sensoryczne – analiza



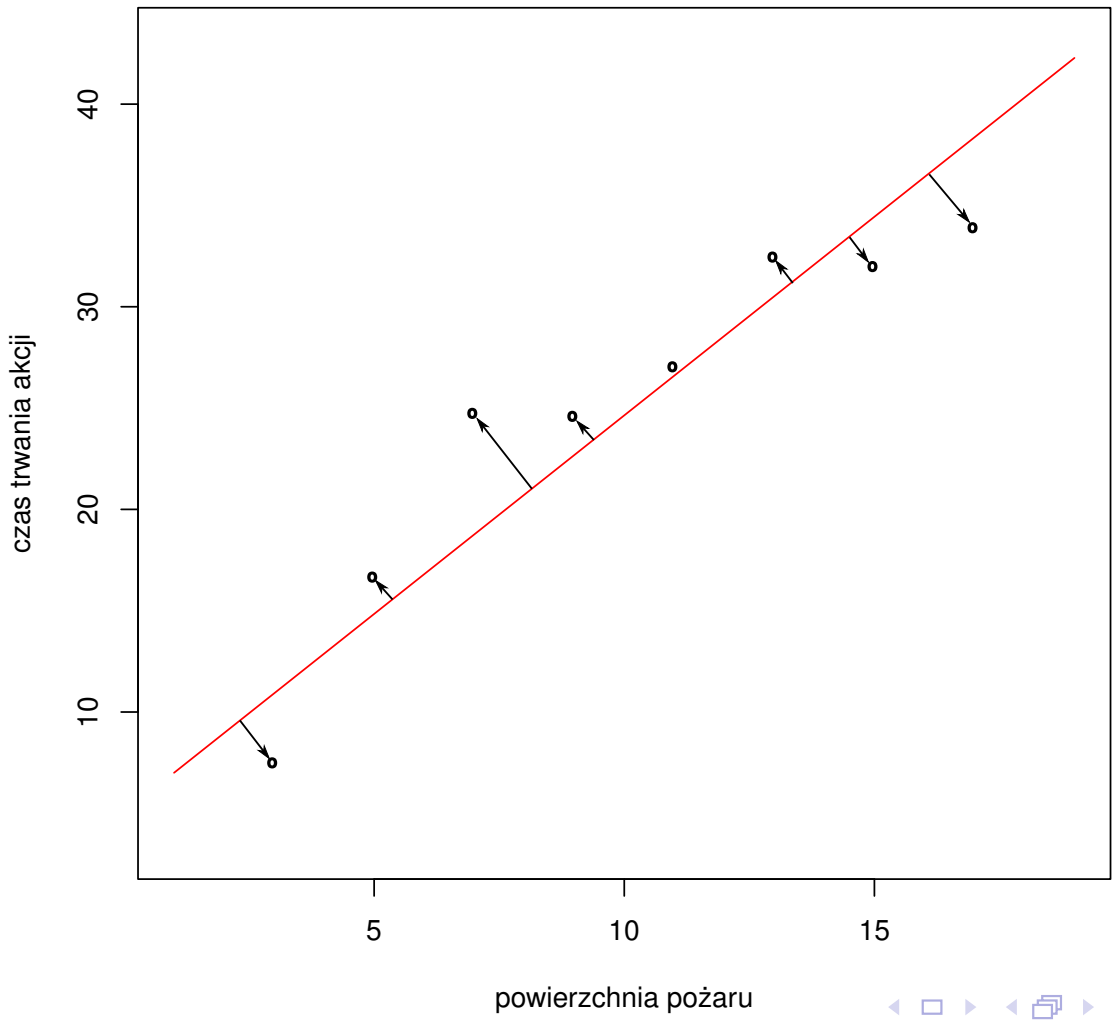
Średnia



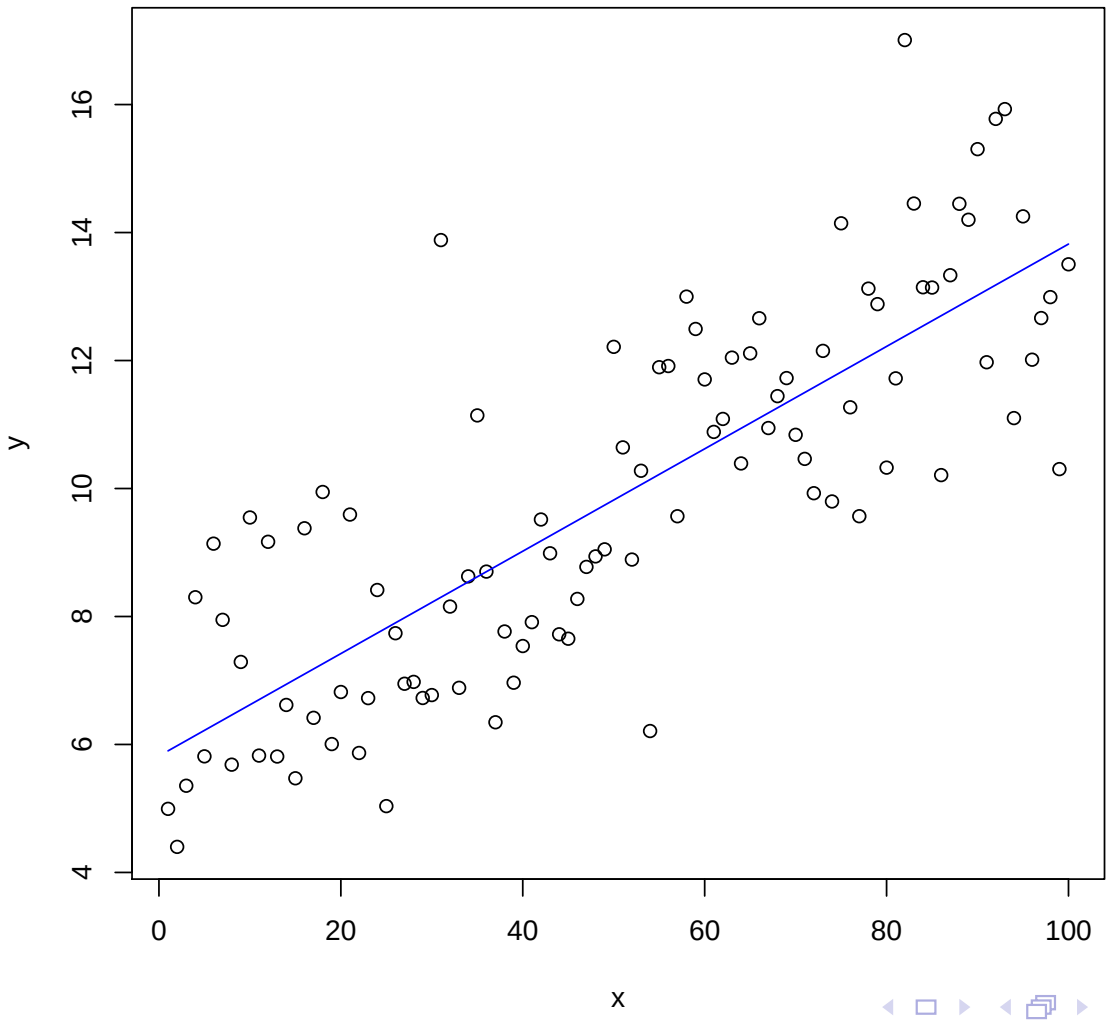
Trend



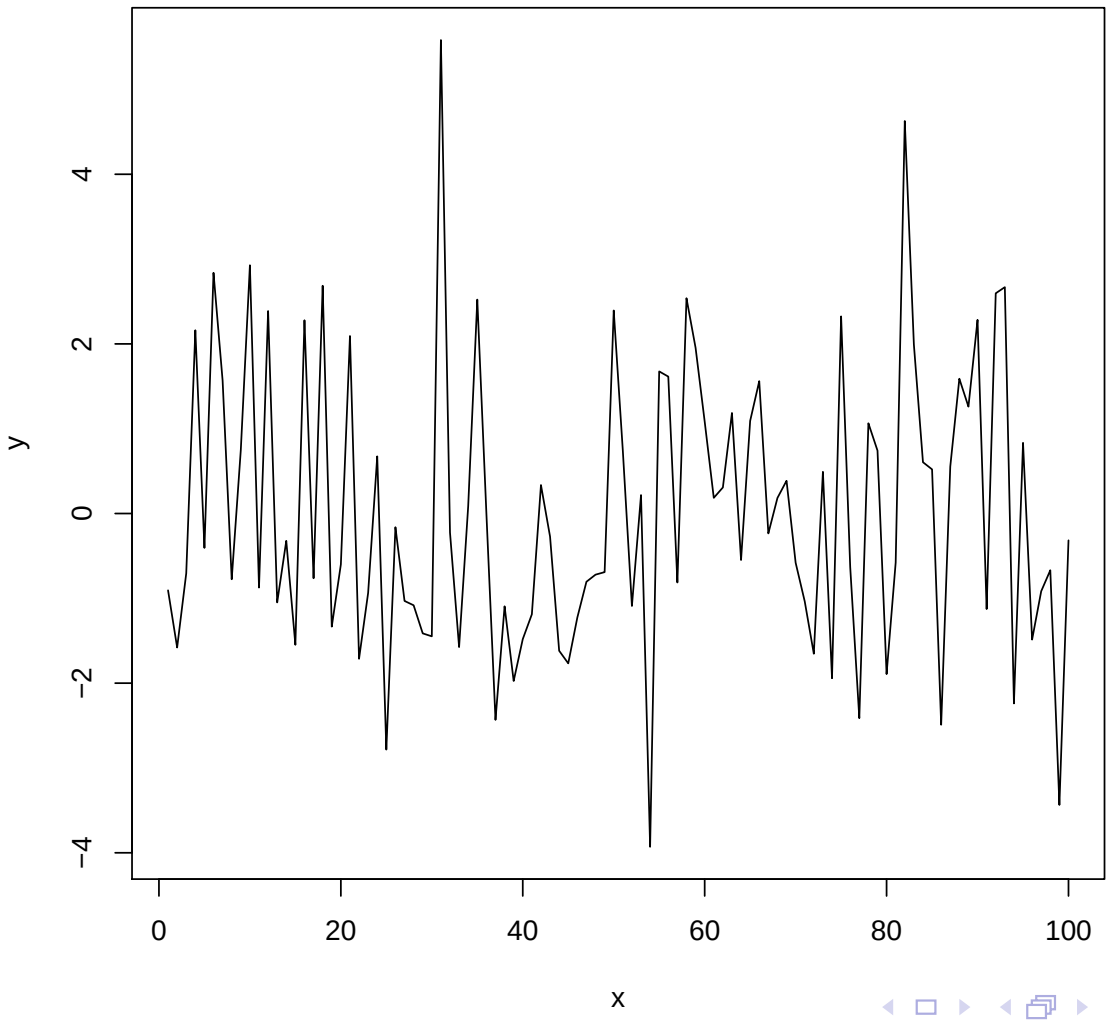
Trend



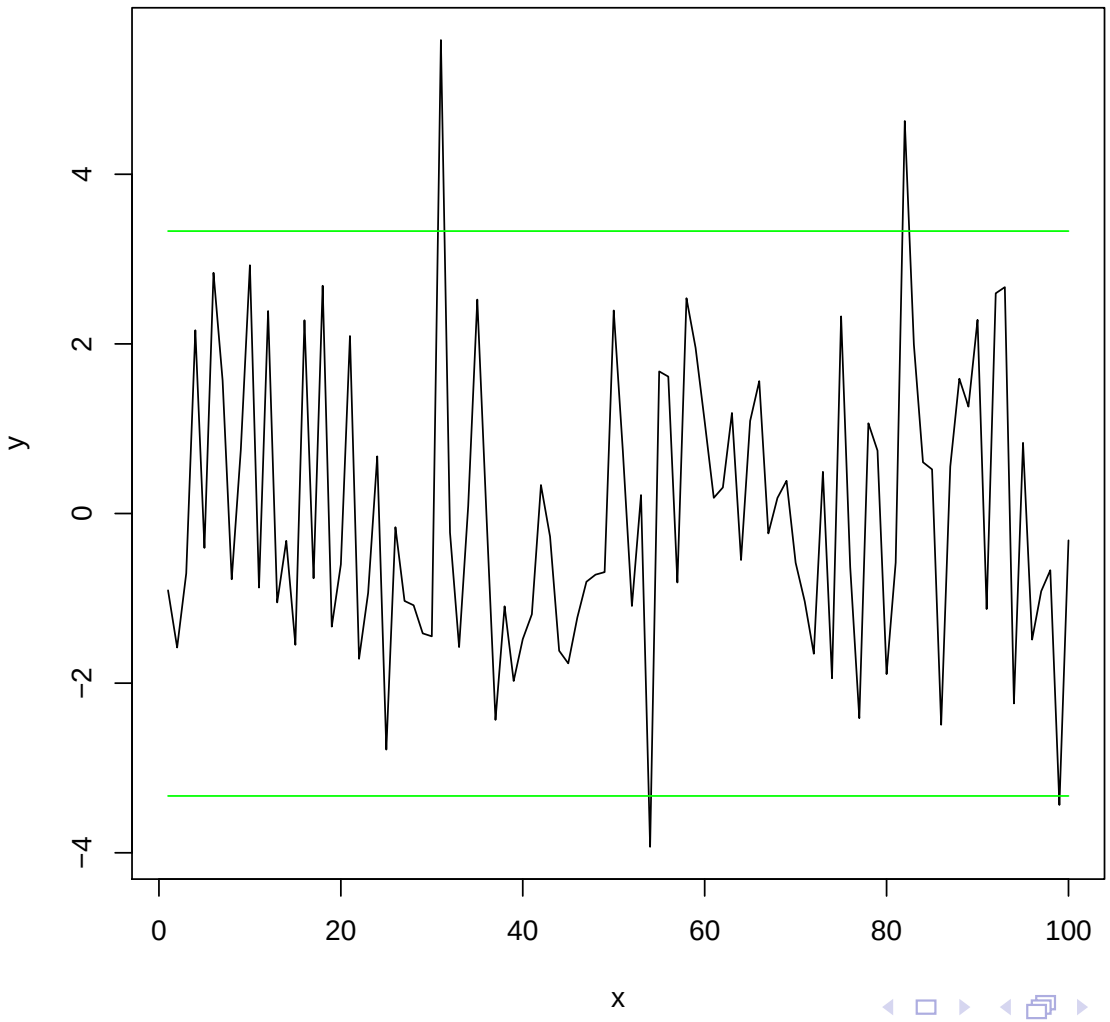
Trend



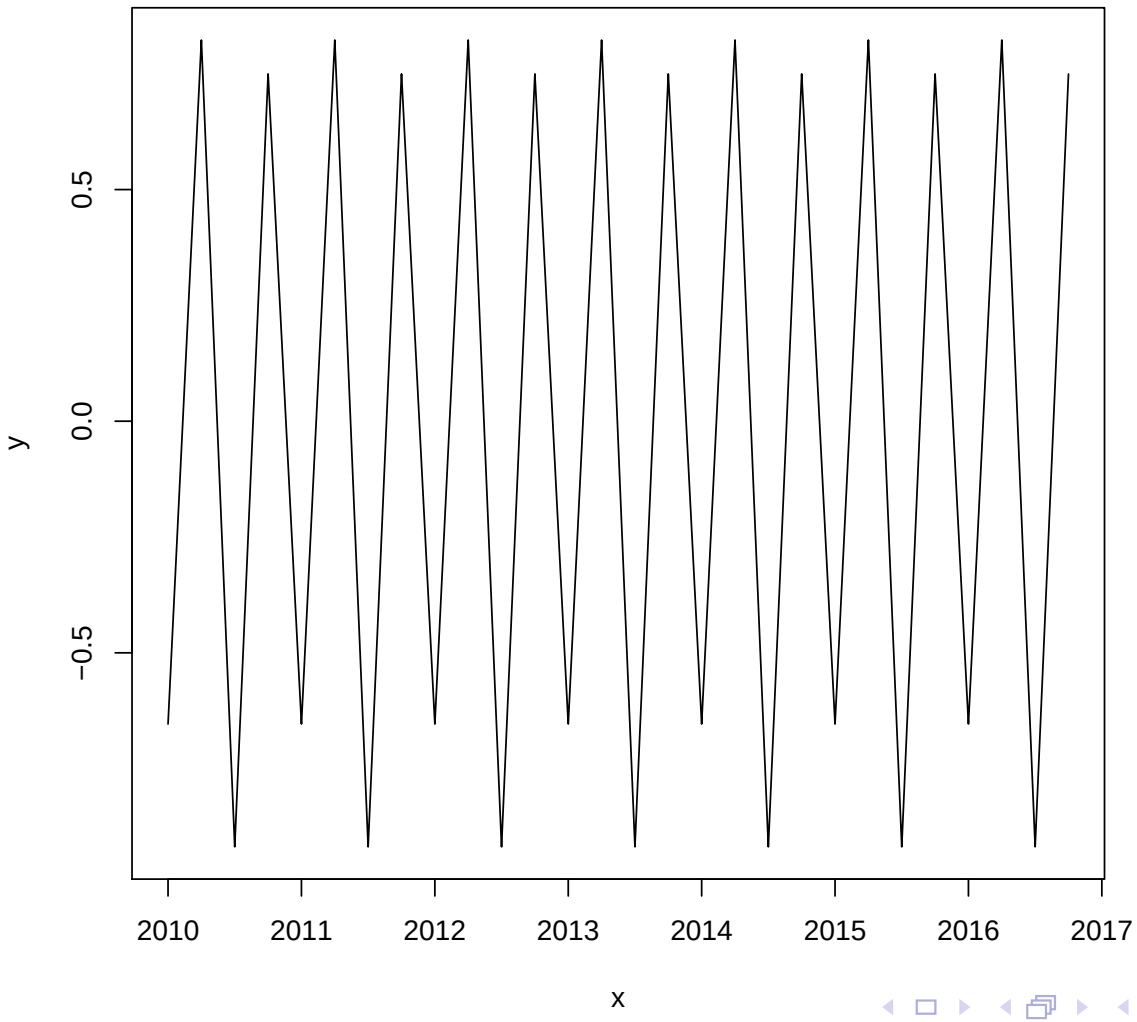
Wartości odstające



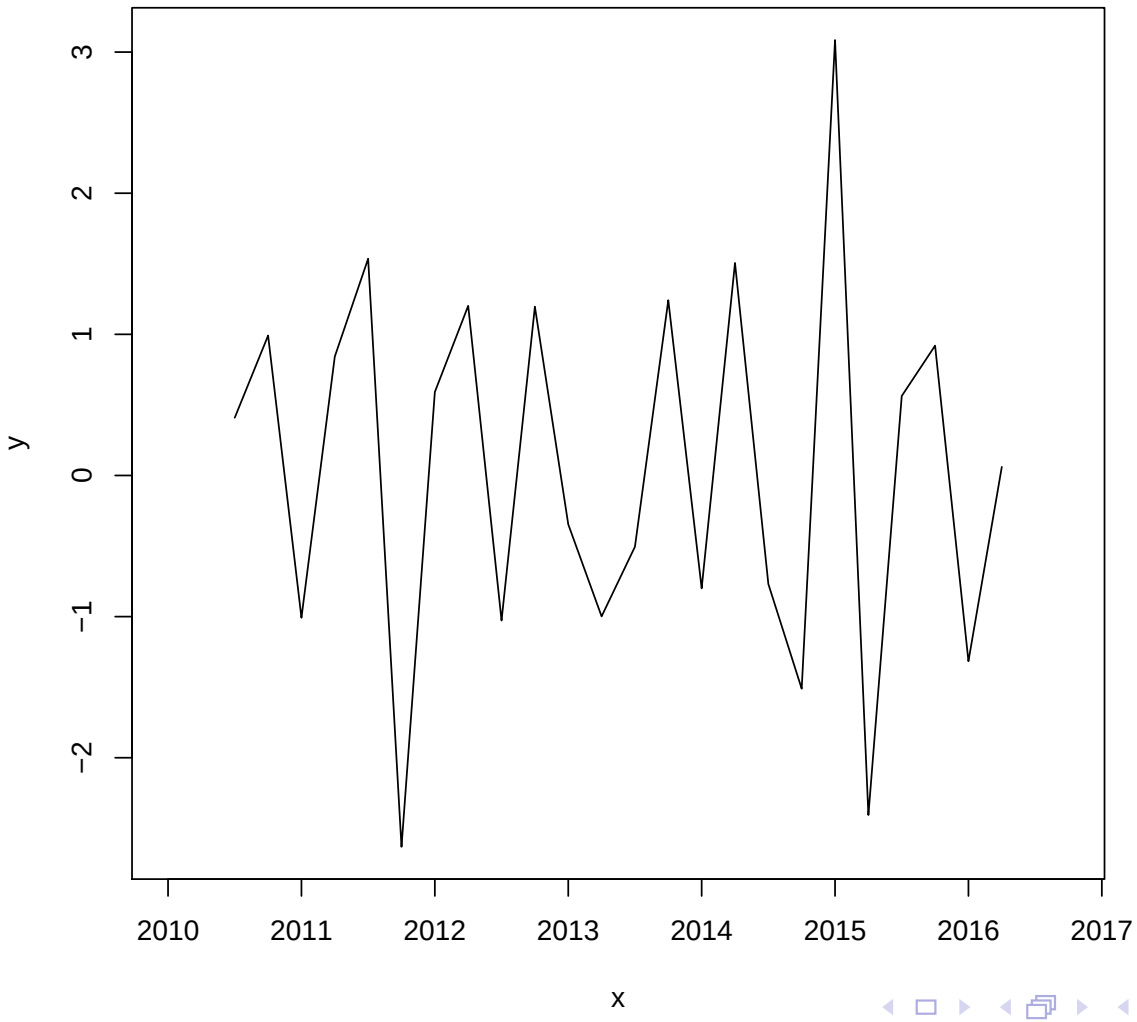
Wartości odstające

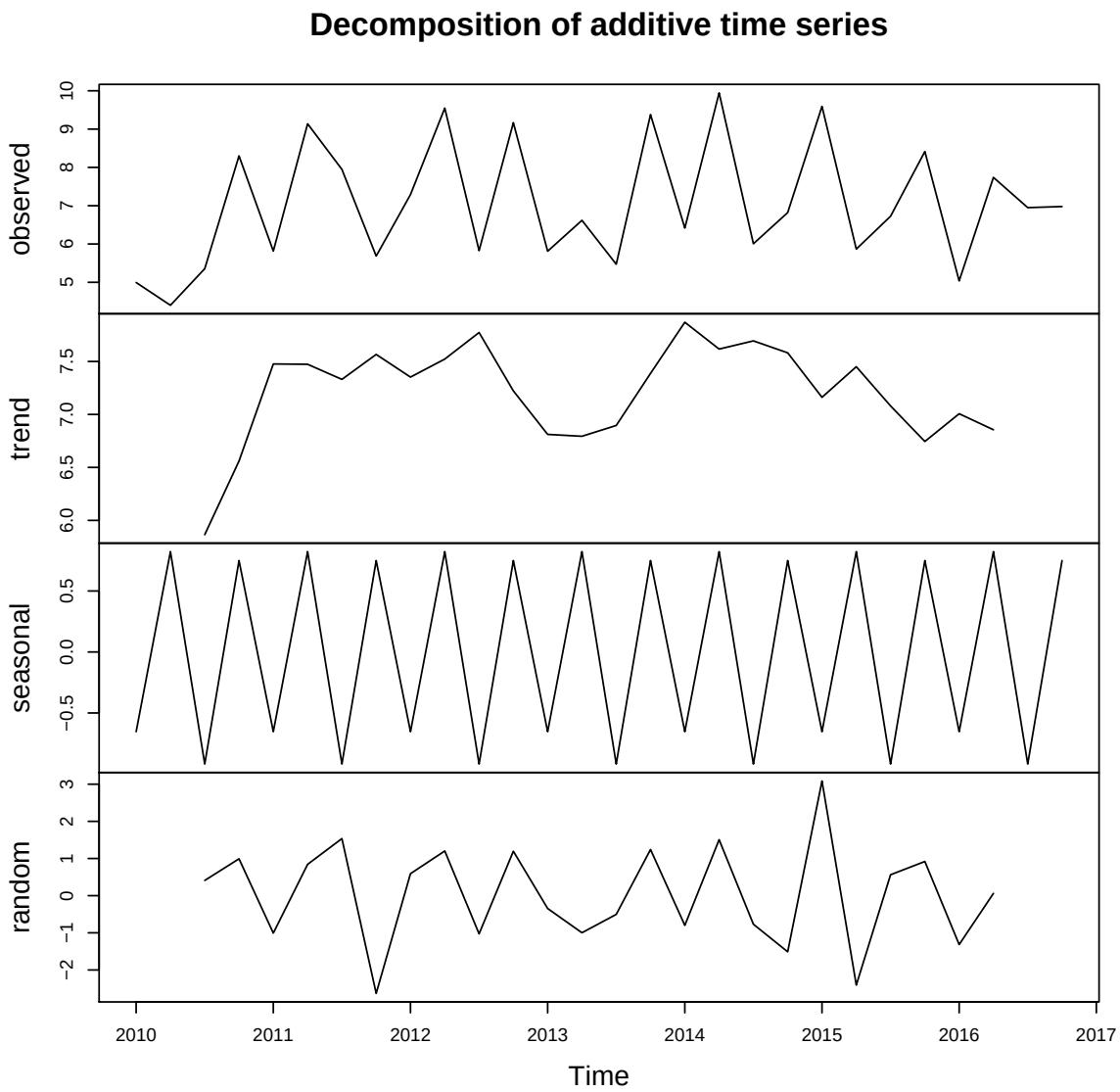


Wahania sezonowe



Szum

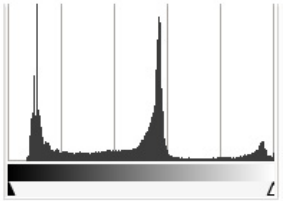


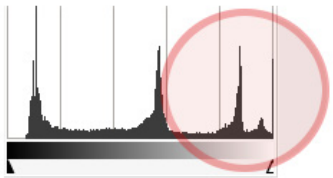

Decomposition of additive time series
observed
trend
seasonal
random
Time
2010
2011
2012
2013
2014
2015
2016
2017

Dane multimedialne

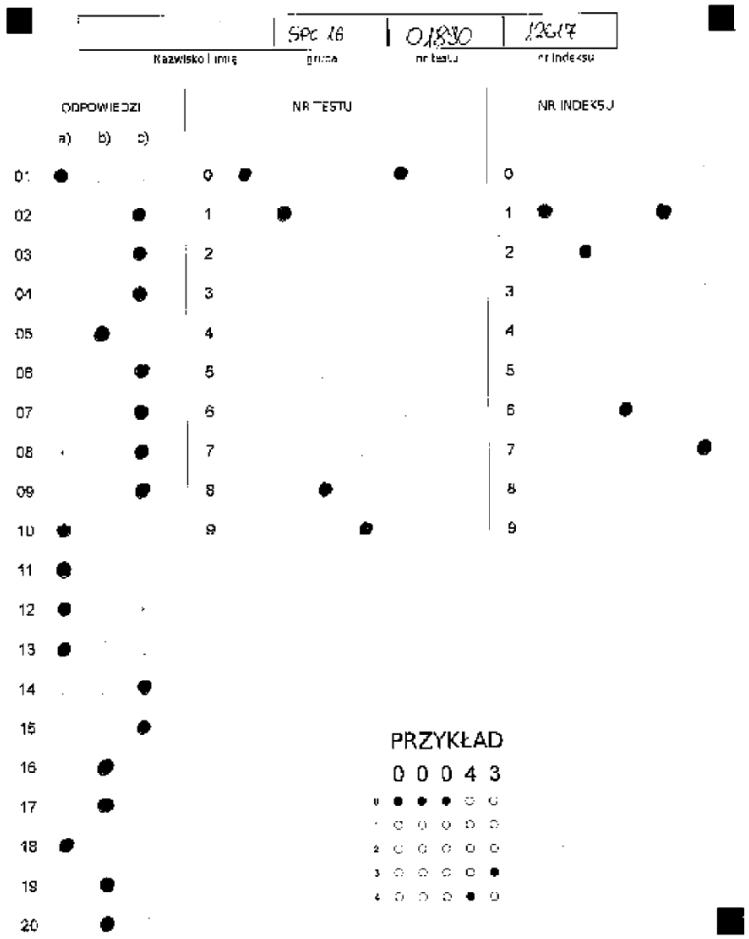
Multimedia (łac. multum + medium) – media, stanowiące połączenie kilku różnych form przekazu informacji np. tekstu, dźwięku, grafiki, animacji, wideo.

Histogramy

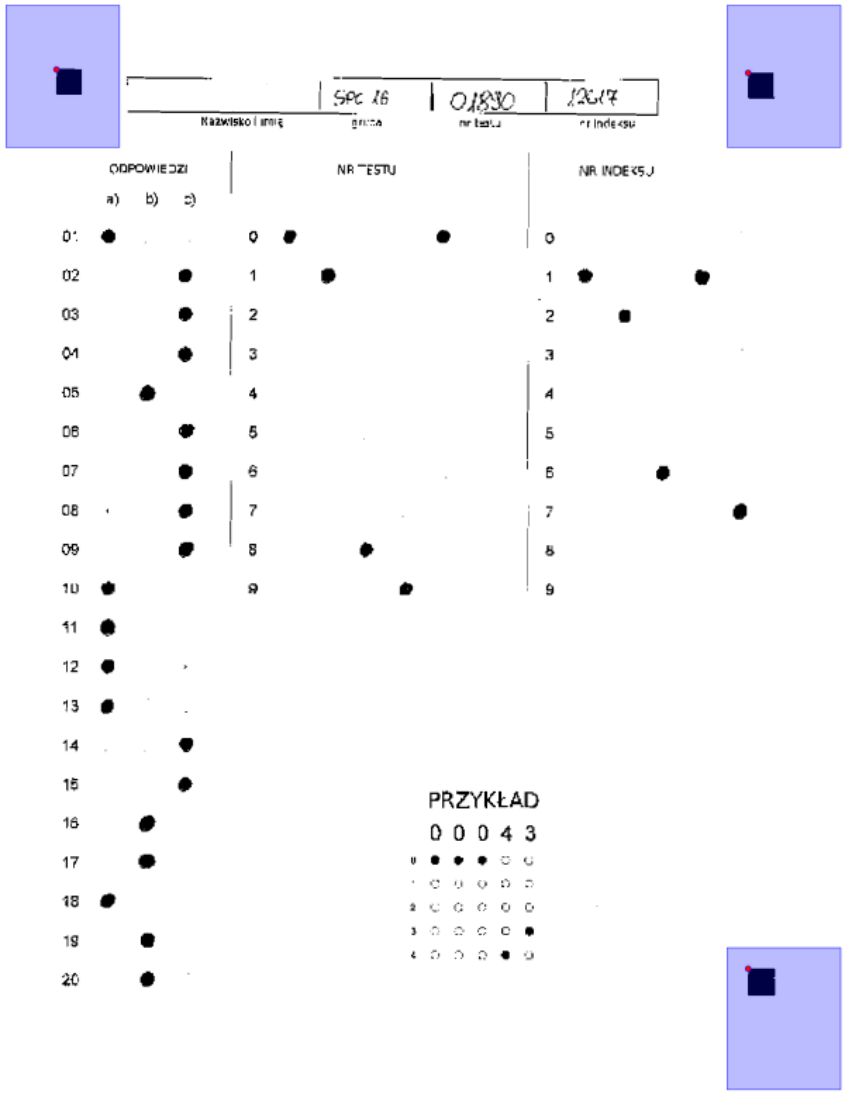




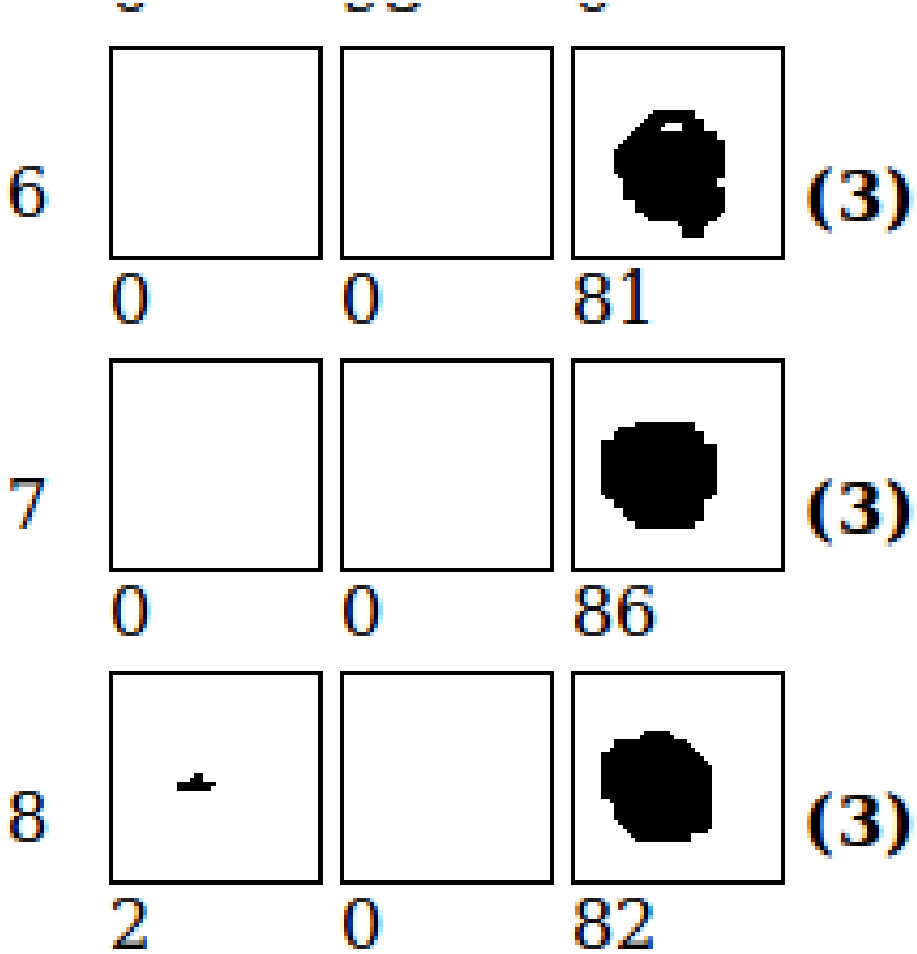
Dopasowanie do wzorca



Dopasowanie do wzorca

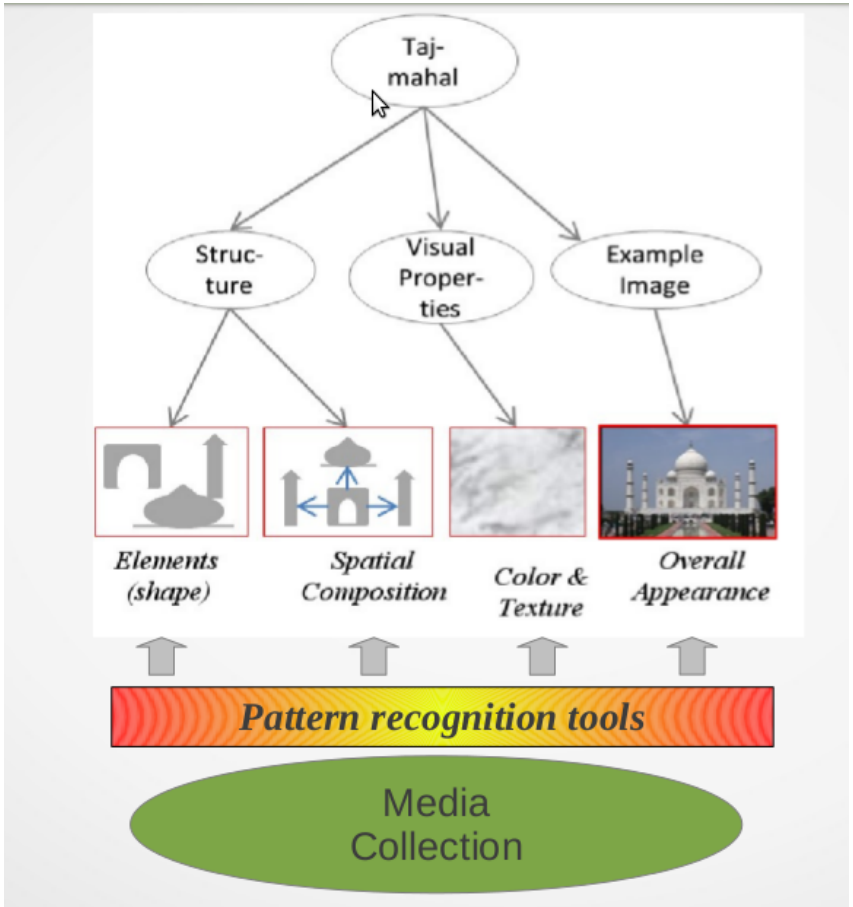


Dopasowanie do wzorca



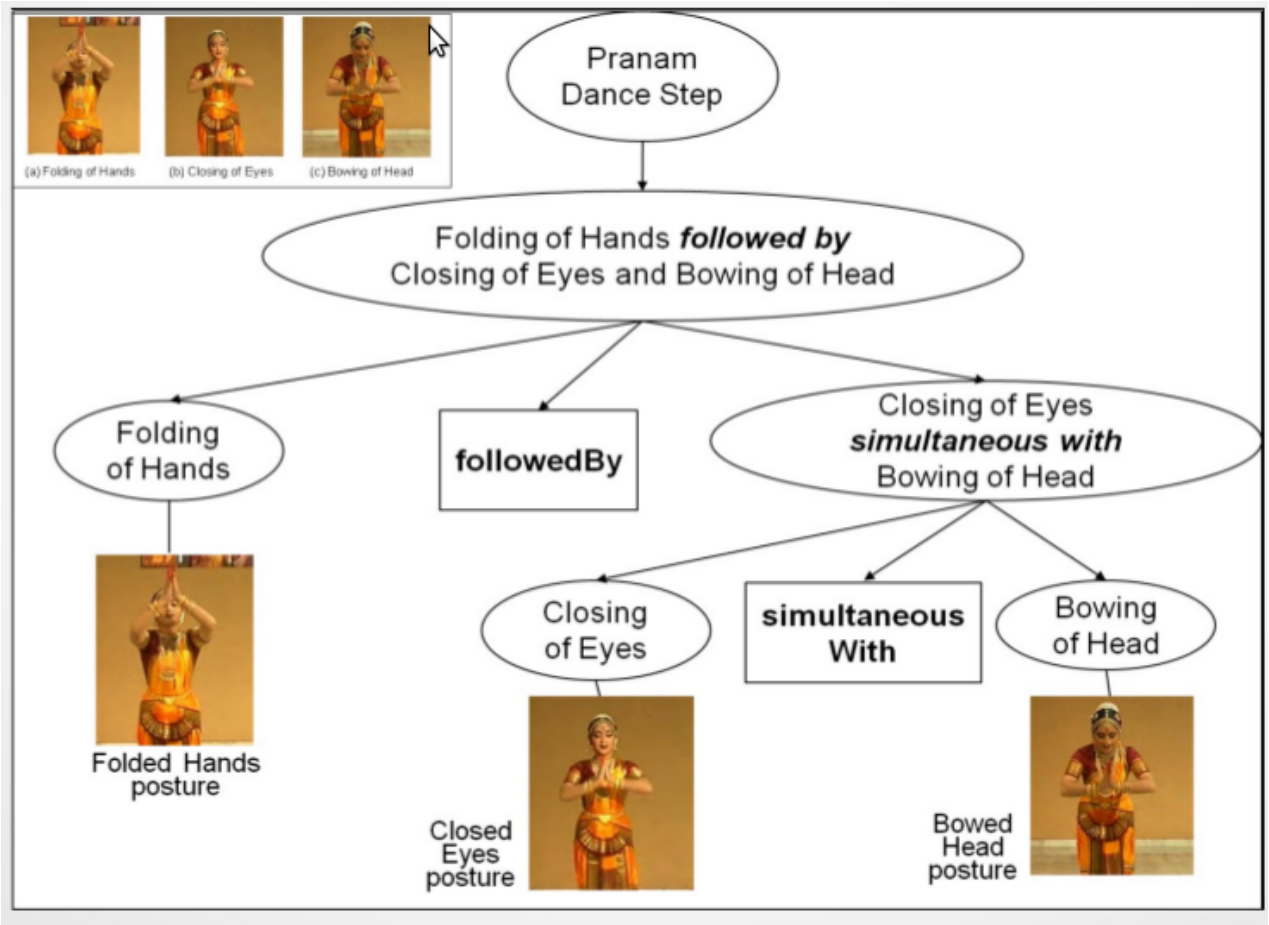
źródło: Y. Fang, K. Wang, J. Cheng, H. Lu: A Real-Time Hand Gesture Recognition Method. 2007.

Analiza z użyciem ontologii



Źródło: dzięki uprzejmości Hiranmay Gosh

Analiza z użyciem ontologii cd.



Źródło: dzięki uprzejmości Hiranmay Gosh