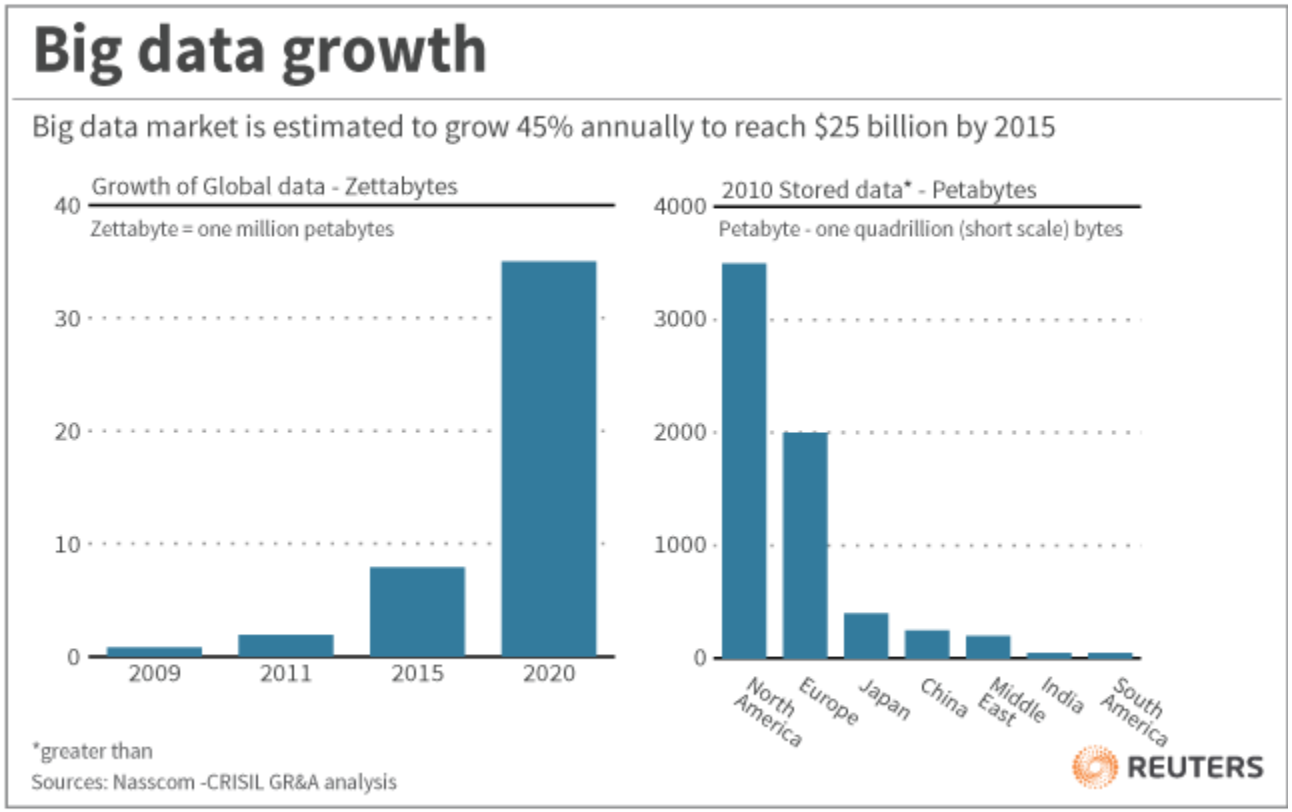
KNOWLEDGE ENGINEERING AND DATA MINING

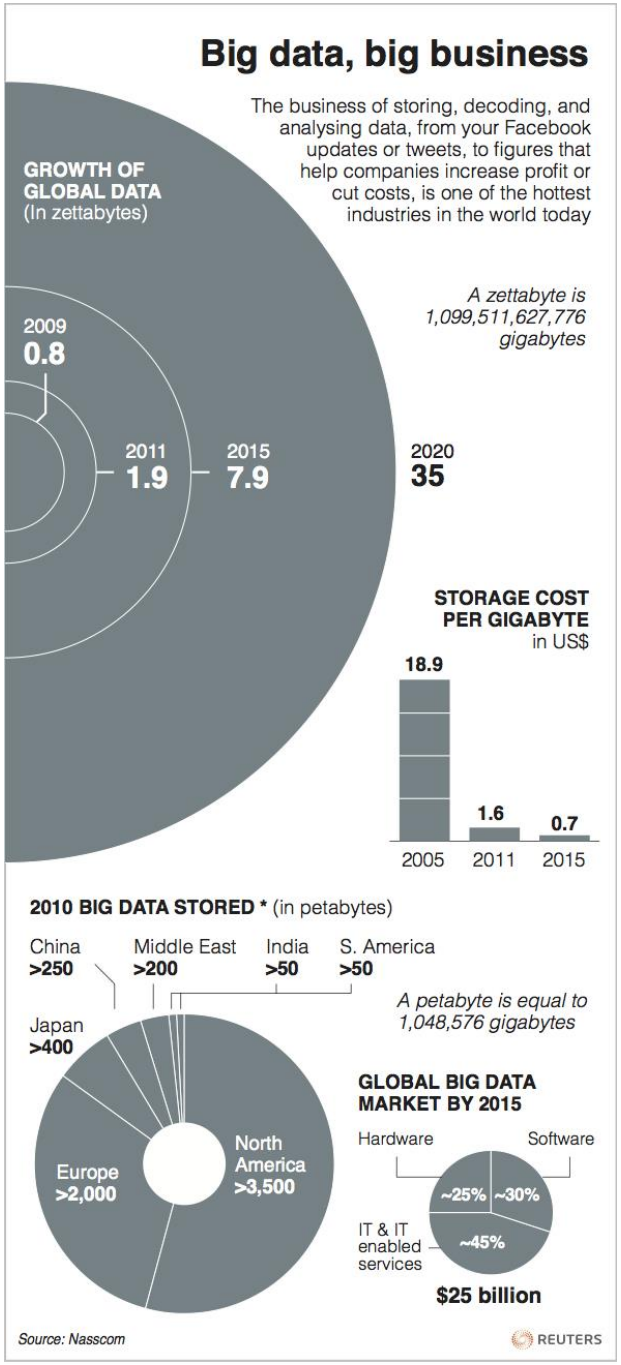
Piotr Świderek, Automatyka i Robotyka
Kraków, 2016

TABLE OF CONTENTS

INTRODUCTION



Reuters graphic/Catherine Trevethan 05/10/12



- Big data business is one of the hottest industries in the world today.
- The business of storing, decoding and analyzing data, figures that help companies increase profit or cut costs.

DATA EXPLORATION

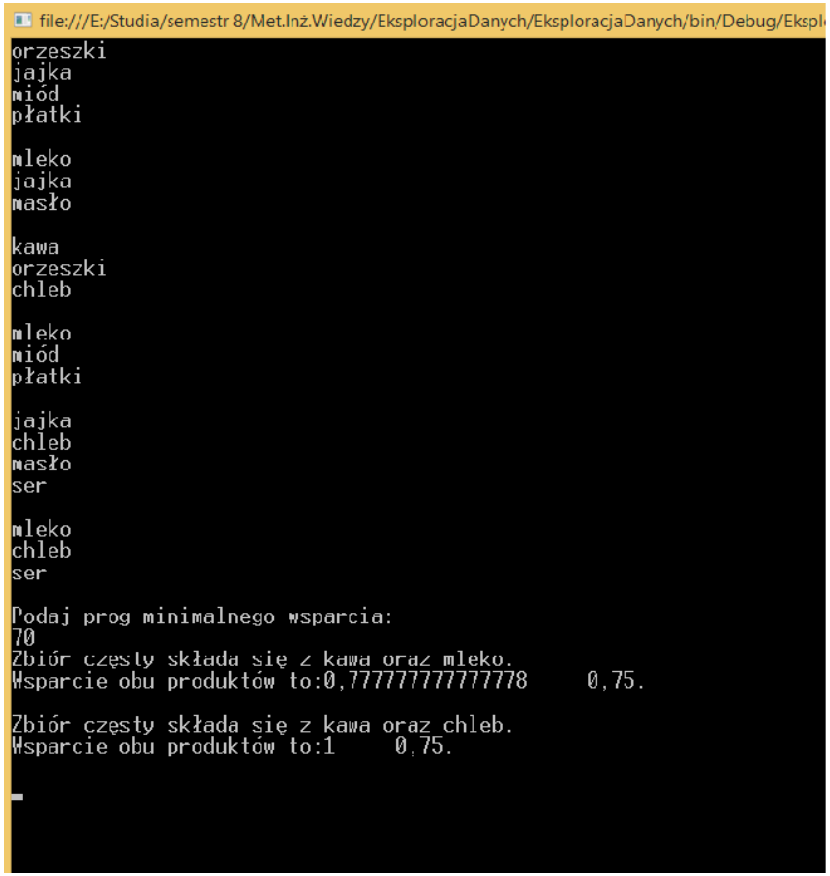
- Data Exploration conduct by searching frequent patterns their subsets or subsequence, which is, according to computationally complex and complex for large collections / databases, requiring use clever algorithms to improve performance.

DEFINITIONS

- Support is the probability that a certain transaction contains calculated with respect to all possible transactions.
- Confidence - is the conditional probability that the transaction containing X also contains Y.
- Mining association rules is to find all the rules X -> Y with specified minimum support s and the specified minimum confidence c:

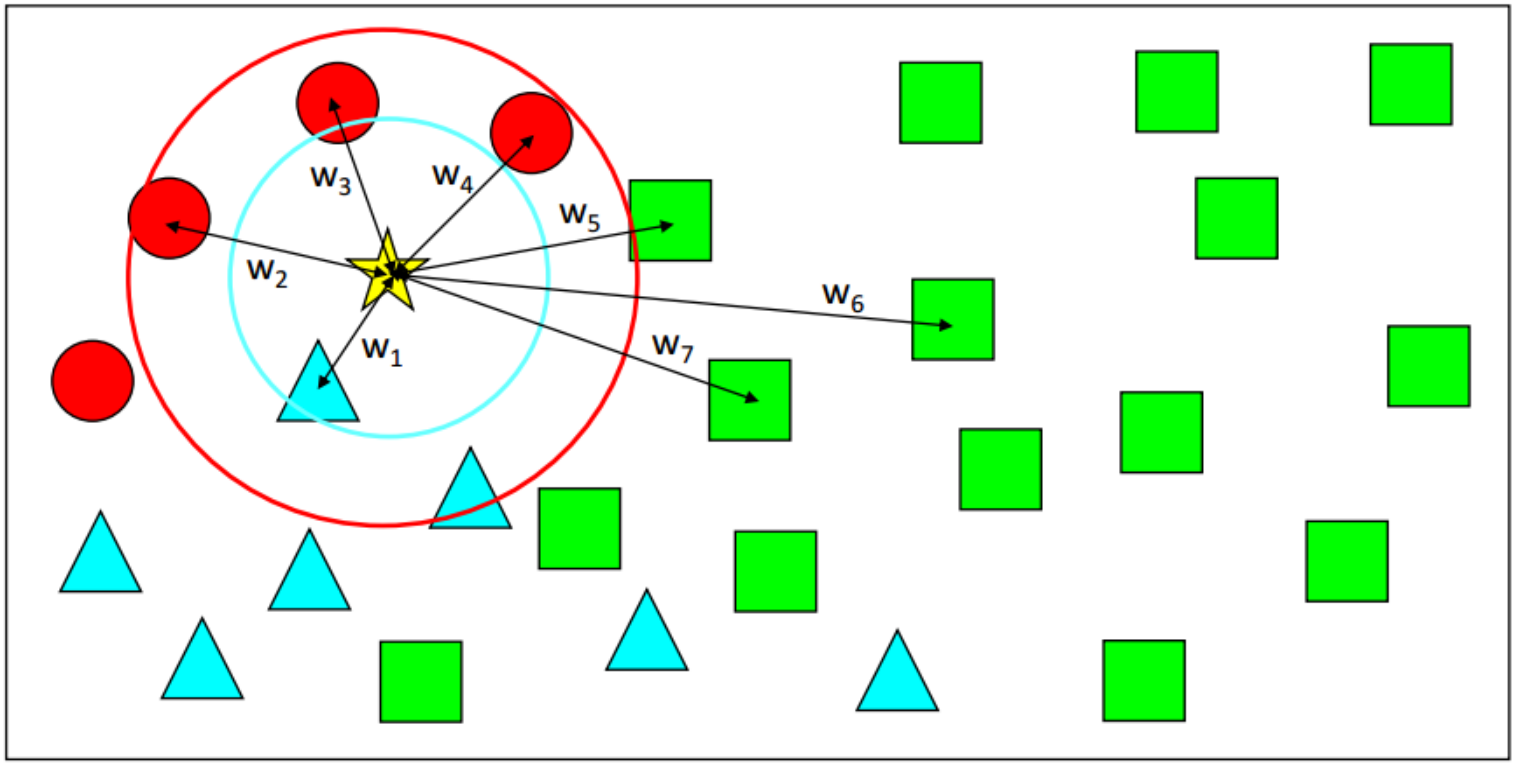
  eg. $s \geq 50\%$, $c \geq 50\%$.

DATA EXPLORATION

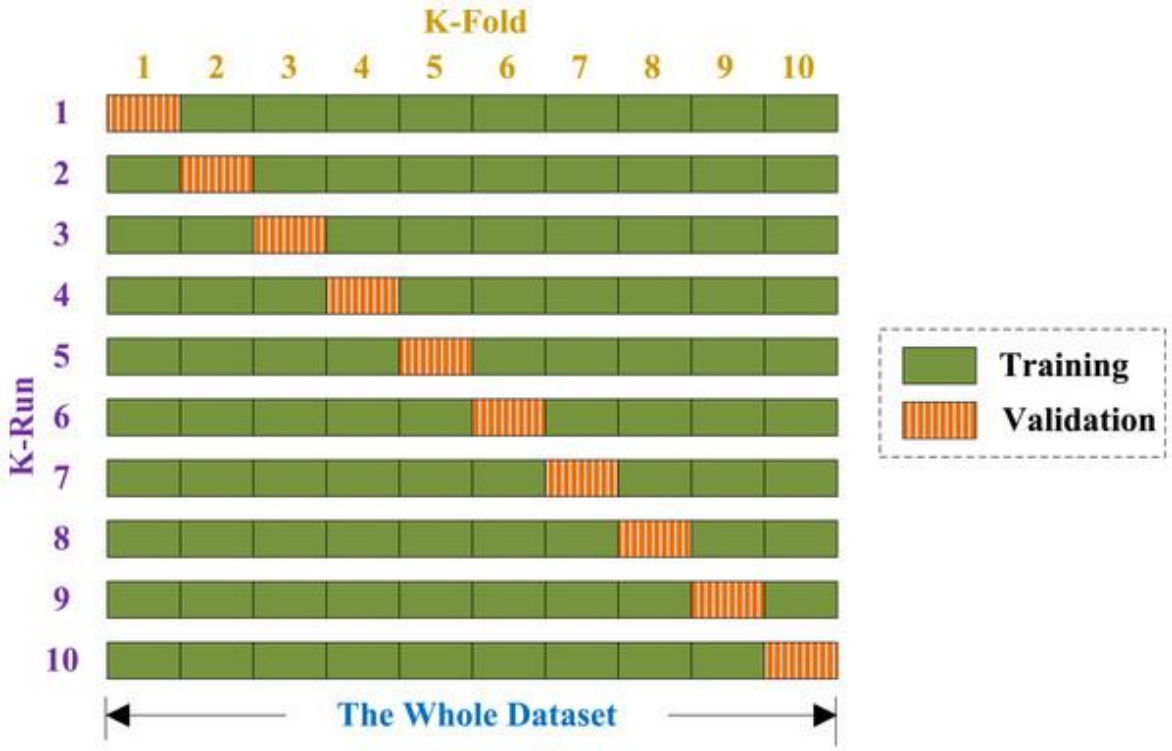


- Run file EksploracjaDanych\bin\Release\EksploracjaDanych.exe
- Print the dataset
- Enter the support and confidence threshold

KNN WITH CROSS-VALIDATION



KNN WITH CROSS-VALIDATION



https://peerj.com/articles/1251/

KNN WITH CROSS-VALIDATION

Dataset:	Records:	Classes:	Parameters:
IrisDataAll	150	3	4
Wine	178	3	13

How to run. Enter file:
\KNN_CrossWalidation\KNN_CrossWalidation\bin\Release.KNN_CrossWalidation.exe

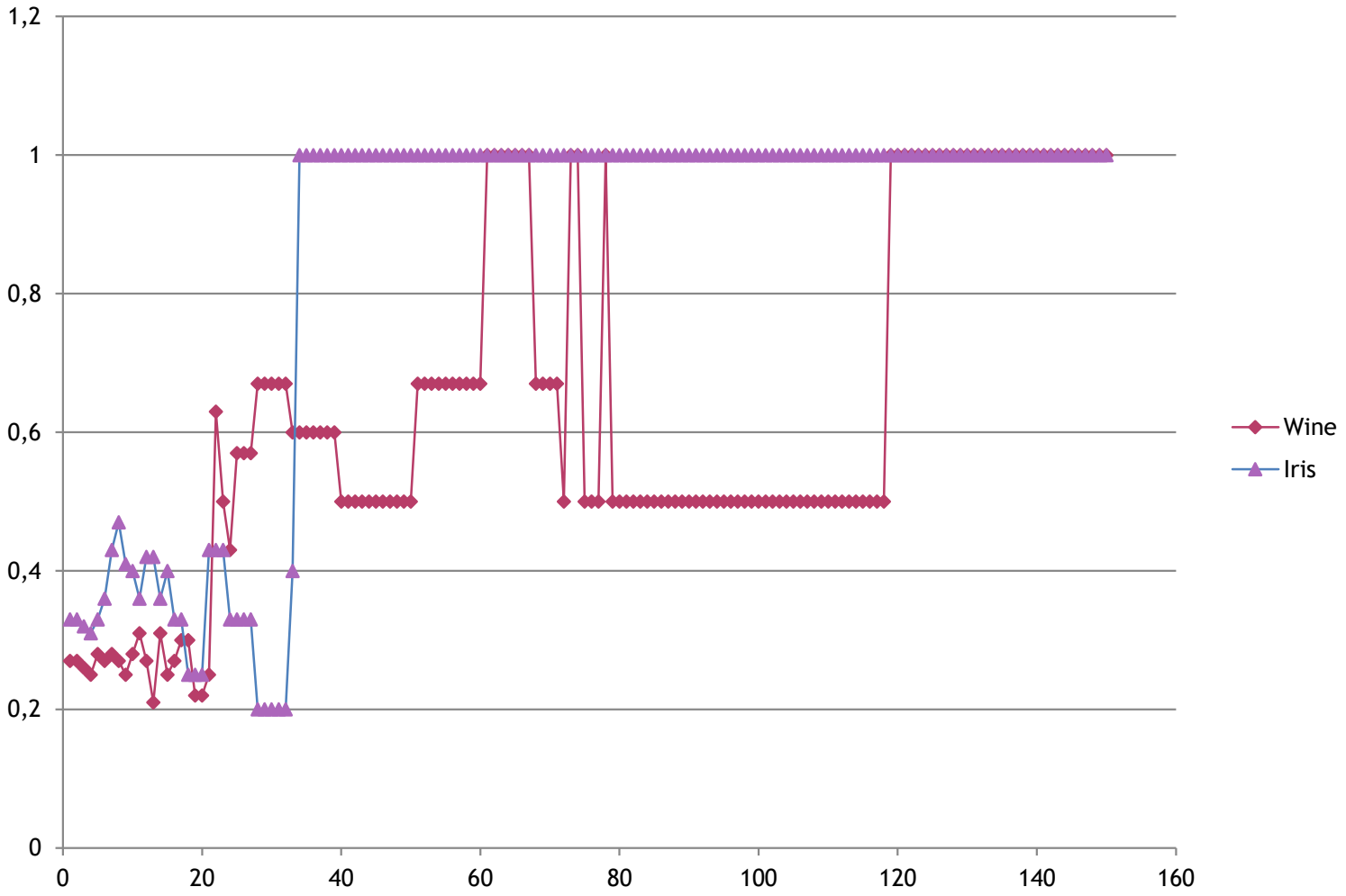
KNN WITH CROSS-VALIDATION

```
Poprawność dobranych elementów to: 0,269662921348315
Dzielimy zbiór na 2 grup.
Poprawność dobranych elementów to: 0,269662921348315
Dzielimy zbiór na 3 grup.
Poprawność dobranych elementów to: 0,254237288135593
Dzielimy zbiór na 4 grup.
Poprawność dobranych elementów to: 0,25
Dzielimy zbiór na 5 grup.
Poprawność dobranych elementów to: 0,277777777777778
Dzielimy zbiór na 6 grup.
Poprawność dobranych elementów to: 0,266666666666667
Dzielimy zbiór na 7 grup.
Poprawność dobranych elementów to: 0,28
Dzielimy zbiór na 8 grup.
Poprawność dobranych elementów to: 0,272727272727273
Dzielimy zbiór na 9 grup.
Poprawność dobranych elementów to: 0,25
Dzielimy zbiór na 10 grup.
Poprawność dobranych elementów to: 0,277777777777778
Dzielimy zbiór na 11 grup.
Poprawność dobranych elementów to: 0,3125
Dzielimy zbiór na 12 grup.
Poprawność dobranych elementów to: 0,266666666666667
Dzielimy zbiór na 13 grup.
Poprawność dobranych elementów to: 0,214285714285714
Dzielimy zbiór na 14 grup.
Poprawność dobranych elementów to: 0,307692307692308
Dzielimy zbiór na 15 grup.
Poprawność dobranych elementów to: 0,25
Dzielimy zbiór na 16 grup.
Poprawność dobranych elementów to: 0,272727272727273
Dzielimy zbiór na 17 grup.
Poprawność dobranych elementów to: 0,3
Dzielimy zbiór na 18 grup.
Poprawność dobranych elementów to: 0,3
Dzielimy zbiór na 19 grup.
Poprawność dobranych elementów to: 0,222222222222222
Dzielimy zbiór na 20 grup.
Poprawność dobranych elementów to: 0,222222222222222
Dzielimy zbiór na 21 grup.
Poprawność dobranych elementów to: 0,25
Dzielimy zbiór na 22 grup.
Poprawność dobranych elementów to: 0,625
Dzielimy zbiór na 23 grup.
Poprawność dobranych elementów to: 0,5
Dzielimy zbiór na 24 grup.
Poprawność dobranych elementów to: 0,428571428571429
Dzielimy zbiór na 25 grup.
Poprawność dobranych elementów to: 0,571428571428571
Dzielimy zbiór na 26 grup.
Poprawność dobranych elementów to: 0,571428571428571
Dzielimy zbiór na 27 grup.
Poprawność dobranych elementów to: 0,571428571428571
Dzielimy zbiór na 28 grup.
Poprawność dobranych elementów to: 0,666666666666667
```

Checking correctness for different K.

```
Podaj wartość K.
15
Dzielimy zbiór na 15 grup.
Poprawność dobranych elementów to: 0,4
```

KNN WITH CROSS-VALIDATION



KNN WITH CROSS-VALIDATION

Summary:

- Guess correctness increase with an increasing K value.
- Increasing K need more time to show results.
- KNN algorithm is $O(n^2)$

SELF-ORGANIZING MAP

- It provide a way of representing multidimensional data in much lower dimensional spaces - usually one or two dimensions.
- One of the most interesting aspects of SOMs is that they learn to classify data *without supervision*.

SELF-ORGANIZING MAP

```
x = 0 y = 0
Dystans: 0,89 0,09 0,45 0,28
x = 0 y = 1
Dystans: 0,48 0,84 0,2 0,03
x = 0 y = 2
Dystans: 0,75 0,59 0,94 0,14
x = 0 y = 3
Dystans: 0,5 0,33 0,69 0,89
----------------------------------------
x = 1 y = 0
Dystans: 0,25 0,08 0,8 0,64
x = 1 y = 1
Dystans: 1 0,19 0,55 0,39
x = 1 y = 2
Dystans: 0,1 0,94 0,3 0,66
x = 1 y = 3
Dystans: 0,85 0,69 0,05 0,24
----------------------------------------
x = 2 y = 0
Dystans: 0,6 0,96 0,16 0,99
x = 2 y = 1
Dystans: 0,35 0,55 0,91 0,26
x = 2 y = 2
Dystans: 0,1 0,3 0,66 0,01
x = 2 y = 3
Dystans: 0,21 0,57 0,4 0,76
----------------------------------------
x = 3 y = 0
Dystans: 0,96 0,32 0,15 0,87
x = 3 y = 1
Dystans: 0,71 0,07 0,26 0,62
x = 3 y = 2
Dystans: 0,46 0,17 0,01 0,37
x = 3 y = 3
Dystans: 0,21 0,92 0,76 0,12
----------------------------------------
```

- Run file EksploracjaDanych\bin\Release\EksploracjaDanych.exe

  SOM-kohonen\SOM-kohonen\bin\Release\SOM-kohonen.exe
- Network with random weights

SELF-ORGANIZING MAP

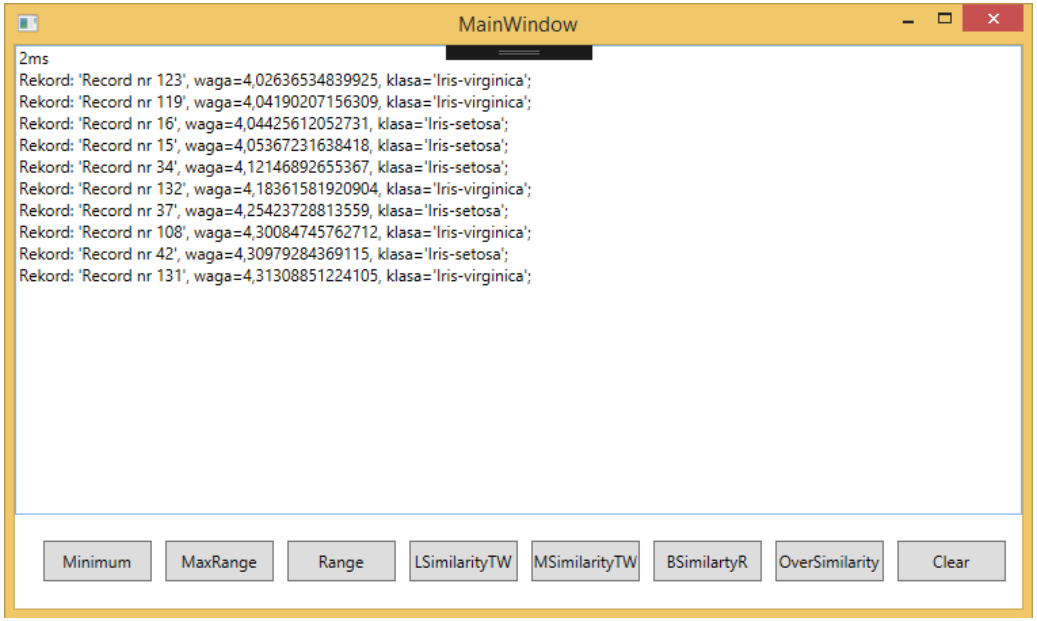
```
-----------------------------------------
--------------------------------
x = 0 y = 0
Dystans: 4,68 2,47 1,86 0,48
x = 0 y = 1
Dystans: 4,69 2,43 1,99 0,53
x = 0 y = 2
Dystans: 4,72 2,41 2,18 0,6
x = 0 y = 3
Dystans: 4,76 2,41 2,4 0,69
-----------------------------------------
x = 1 y = 0
Dystans: 4,61 2,42 1,63 0,4
x = 1 y = 1
Dystans: 4,62 2,4 1,74 0,44
x = 1 y = 2
Dystans: 4,66 2,4 1,93 0,51
x = 1 y = 3
Dystans: 4,72 2,41 2,18 0,6
-----------------------------------------
x = 2 y = 0
Dystans: 4,57 2,39 1,48 0,35
x = 2 y = 1
Dystans: 4,58 2,38 1,56 0,38
x = 2 y = 2
Dystans: 4,62 2,4 1,74 0,44
x = 2 y = 3
Dystans: 4,69 2,43 1,99 0,53
-----------------------------------------
x = 3 y = 0
Dystans: 4,56 2,38 1,42 0,33
x = 3 y = 1
Dystans: 4,57 2,39 1,48 0,35
x = 3 y = 2
Dystans: 4,61 2,42 1,63 0,4
x = 3 y = 3
Dystans: 4,68 2,47 1,86 0,48
-----------------------------------------
```

- Weights in network after learning.

ASSOCIATED GRAPH DATA STRUCTURE (AGDS)

- A passive data structure, which make more faster operation for example: filtering by value, attribute and searching similar group of elements.
- AGDS doesn’t contain duplicates or excess data.

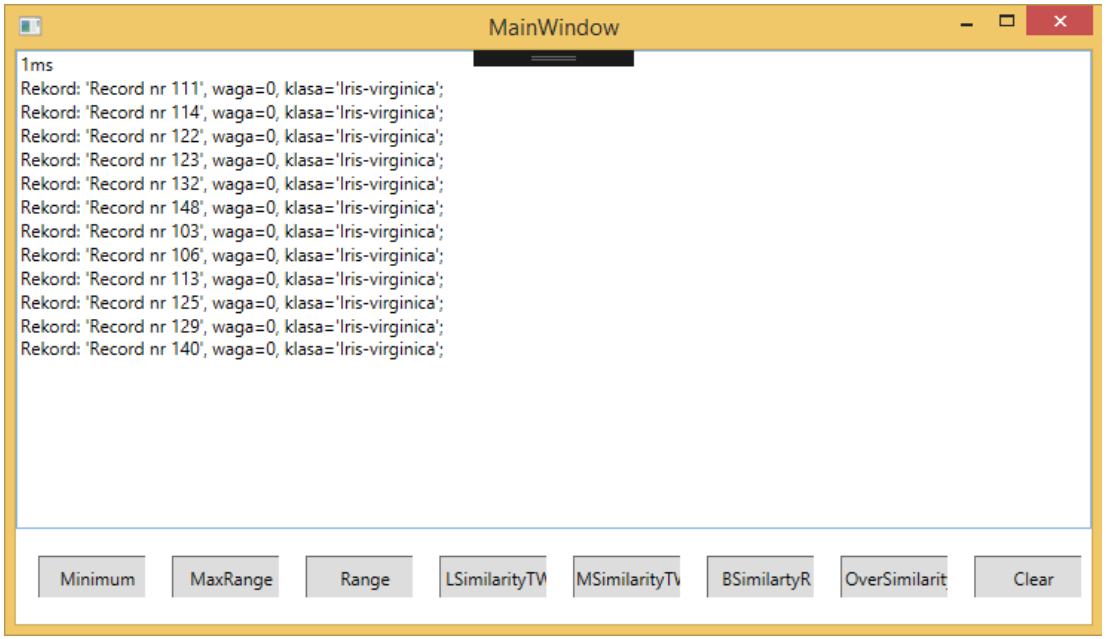
ASSOCIATED GRAPH DATA STRUCTURE (AGDS)



Searching least similar elements.

Run the file:
AGDS\WpfApplication1\bin\Release\WpfApplication1.exe
Run the test:
tabele-porownanieAGDS\Tablele-AGDS\Tablele-AGDS\bin\Debug.
Tablele-AGDS.exe

ASSOCIATED GRAPH DATA STRUCTURE (AGDS)



Searching range of value between 2-2,1 for attribute „petal-width”.

COMPARING THE TIME FOR AGDS AND TABLE

- Tested database: Iris Data
- Number of recors: 150
- Number of columns: 4

COMPARING THE TIME FOR AGDS AND TABLE

Measure executing 10000-times seven types of functions

1. Find Attribute Minimum
2. Find Attribute Maximum Range
3. Find Attribute Range
4. Find Least Similar Elements for two elemants
5. Find Most Similar Elements for two elemants
6. Find Elements under similarity rate
7. Find Elements over similarity rate for two patterns

COMPARING THE TIME FOR AGDS AND TABLE

Function nr	AGDS (time [ms])	Table (time [ms])
1	19	32
2	55	67
3	28	39
4	315	670
5	315	700
6	120	215
7	215	290

COMPARING THE TIME FOR AGDS AND TABLE

Observations:

1. The AGDS makes easier searching min, max, range values, bacause of using sorted lists for attributes.
2. Tests have shown a correct AGDS function implementation.
3. Execution time grows more slowly than in the tables.

SUMMARY

- Because of very fast extensive database, effectively handling Data will be the challenge of the next years.
- Data mining and machine learning will be developed faster and faster.