



Metody Inżynierii Wiedzy

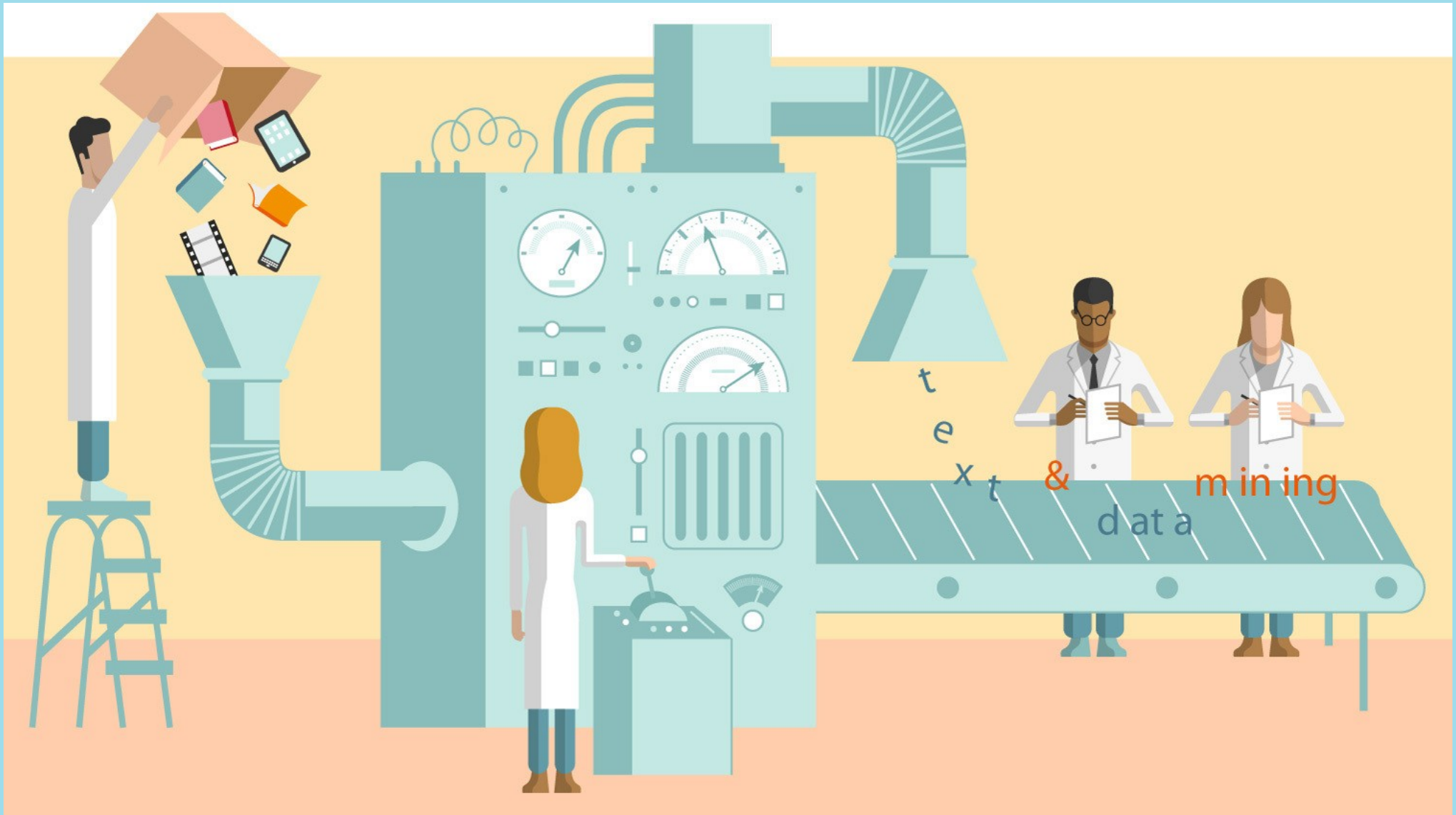
Asocjacyjne Grafowe Struktury Danych

AGDS

Associative Graph Data Structures

Autor: Łukasz Marczak

Data Mining



Eksploracja danych

Wykorzystanie szybkości komputera do znajdowania ukrytych prawidłowości w dużych zbiorach danych

Różne metody eksploracji danych:
metody statystyczne, ewolucyjne, uczenie maszynowe, logika rozmyta, zbiory przybliżone

Eksploracja danych

- Transakcja -zbiór elementów/obiektów
- Pewność(Confidence) –prawdopodobieństwo warunkowe $p(X|Y)$
- Wsparcie(Support) -prawdopodobieństwo wystąpienia wzorca w transakcji $X \cup Y$
- Wzorzec częsty:
wsparcie wzorca $> \text{minsup}(\text{np. } 0.5)$

$7,00 \Rightarrow \textit{Iris - versicolor}$

Algorytm Apriori

1. Krok oczyszczania(*prune step*)

Wyszukanie wzorców częstych (zbiór L_k) na podstawie kandydatów C_k

2. Krok łączenia(*join step*)

Spośród wzorców częstych (L_k) generujemy kombinacje kandydatów C_{k+1} aby znaleźć nowe wzorce częste.

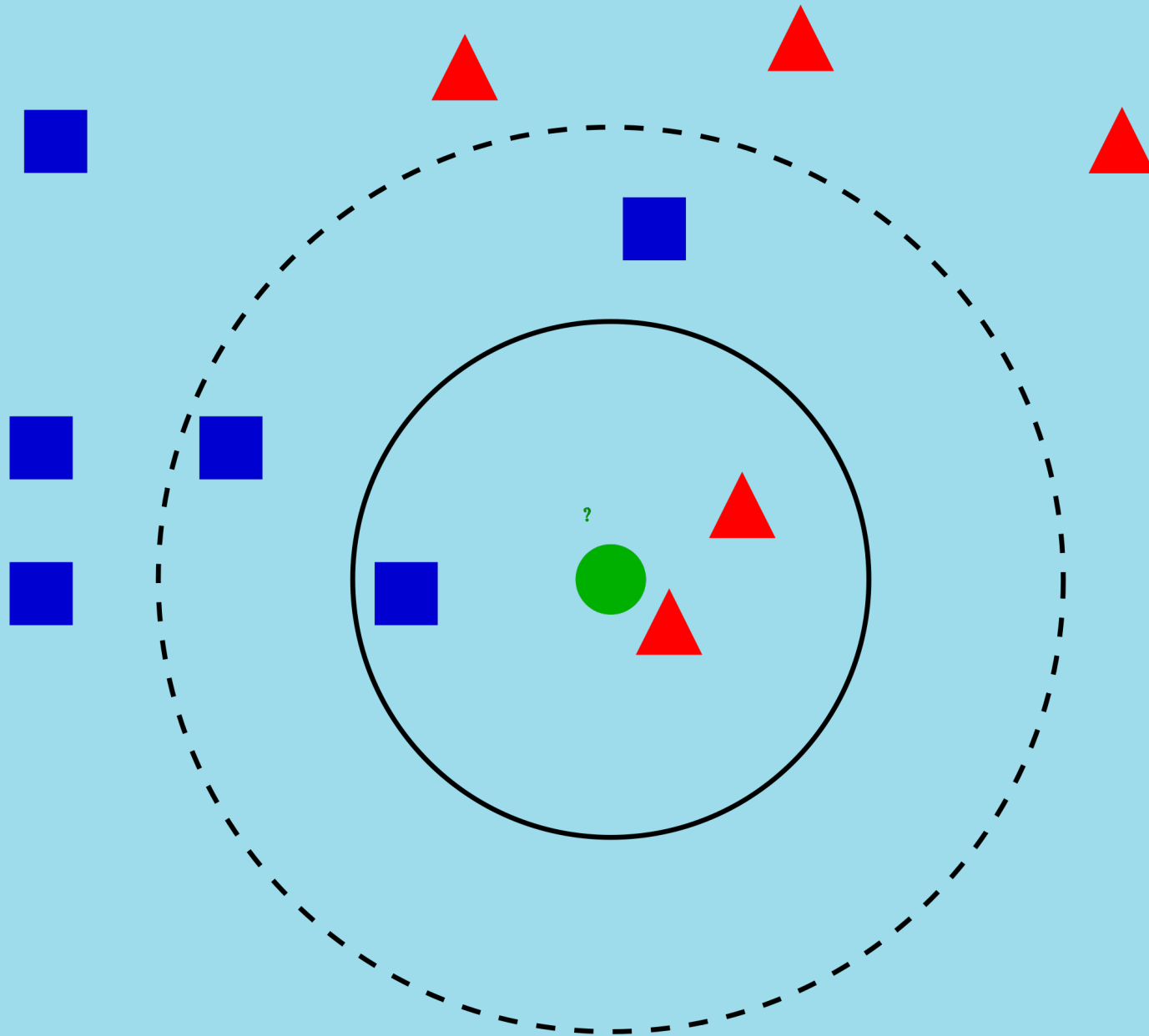
```
int frequencyThreshold = 6;
```

```
Apriori apriori = new Apriori(frequencyThreshold)
    .withProperties(properties)
    .withTransactions(transactions)
    .pruneStep()
    .joinStep()
    .printMostFrequentProperties();
```

Metoda K najbliższych sąsiadów

- Algorytm „leniwy” - nie tworzy wewnętrznej reprezentacji wiedzy
- Zbiór wzorców uczących: pary $\langle x^i, y^i \rangle$, x^i jest postaci wektora: $x^i = \{ x_{i_1}^i, \dots, x_{i_n}^i \}$
- Kandydat jest klasyfikowany na podstawie k sąsiadów

Metoda K najbliższych sąsiadów



Wybór metryki

Odległość Mahalonobisa

$$d_m(\mathbf{x}, \mathbf{y}) := \sqrt{(\mathbf{x} - \mathbf{y})C^{-1}(\mathbf{x} - \mathbf{y})^T}$$

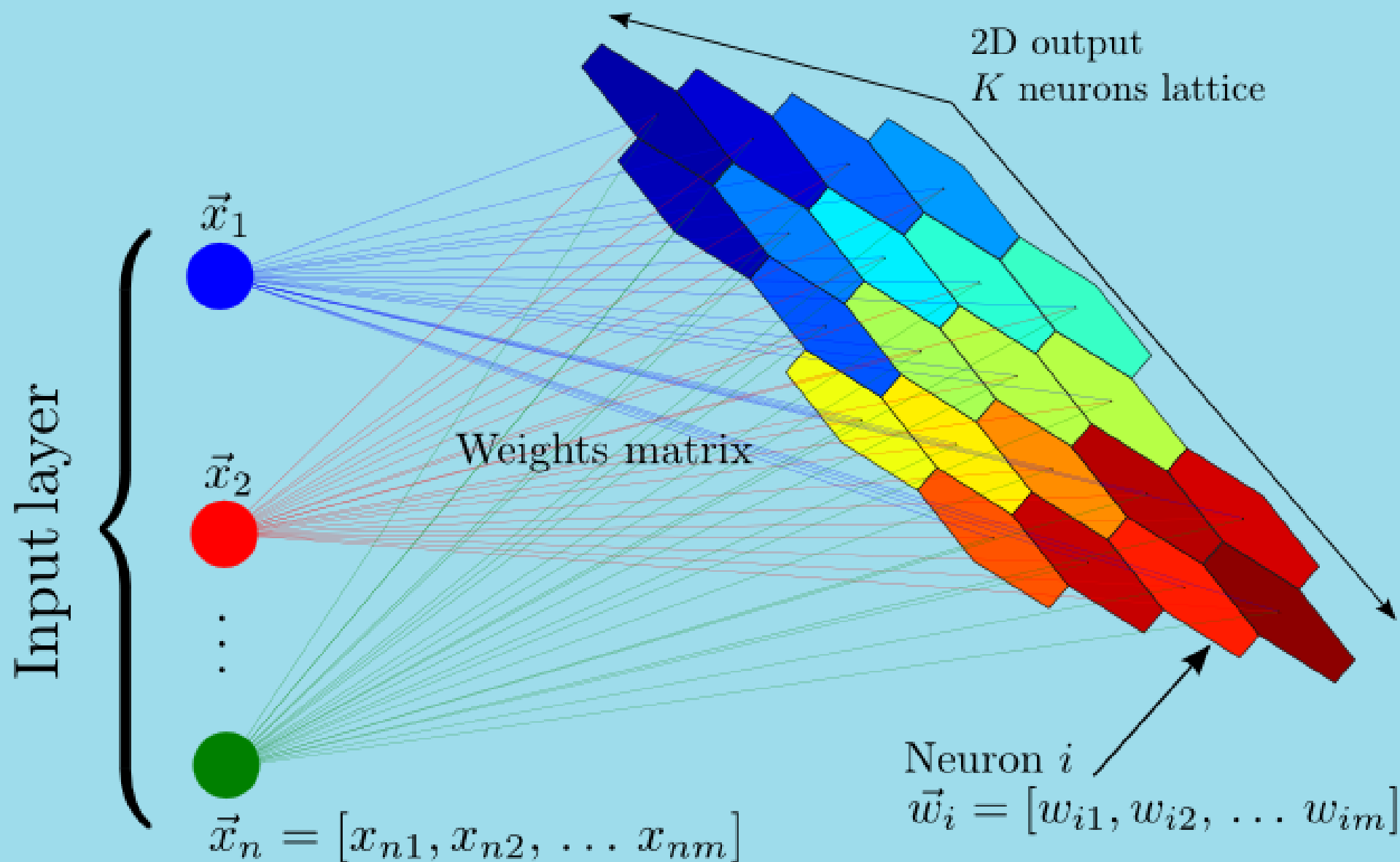
Odległość Minkowskiego

$$L_m(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n |x_i - y_i|^m \right)^{1/m}$$

Odległość Euklidesowa

$$d(A, B) = \sqrt{\sum_{i=1}^n ((x_{iA} - x_{iB})^2)}$$

Samooorganizujące się Mapy - SOM



Samooorganizujące się Mapy - SOM

SOM – sieć neuronowa Kohonena

- Sieć uczona bez nauczyciela
- Nie wymaga założeń, co do rozkładów analizowanych zmiennych
- Oprócz przyporządkowania próbek danych do klas można przeprowadzić wizualizację na płaszczyźnie, łatwiej wtedy wydobyć podobne obiekty

Samooorganizujące się Mapy - SOM

Zalety SOM

- Prosta wizualizacja i interpretacja wyników
- Nie wymaga założeń, co do rozkładów analizowanych zmiennych
- Wykrywa niespodziewane struktury

Samooorganizujące się Mapy - SOM

Wady SOM

- Koszt obliczeniowy(gdy wymiar danych rośnie)
- Dostarczenie poprawnych danych: generacja mapy wymaga aby każdy wzorzec miał wartość we wszystkich wymiarach
- Wrażliwość na „szum” danych

**Praktyczne wykorzystanie
metod inżynierii wiedzy?**



Wyspa tajemnic (2010)

Shutter Island 2 godz. 18 min. **Dramat, Thriller** | USA

★ 8,2/10

436.013 głosów

👁 69 403

chce zobaczyć

▶ zwiastun

reżyser: **Martin Scorsese** obsada: **Leonardo DiCaprio, Mark Ruffalo** dystrybucja: **United International Pictures Sp z o.o.**

Szeryf federalny Daniels stara się dowiedzieć, jakim sposobem z zamkniętej celi w pilnie strzeżonym szpitalu dla chorych psychicznie przestępców zniknęła pacjentka.

[czytaj recenzje](#)

recenzja **W pułapce mistrza** autor **Bartosz Staszczyszyn**

📊 85% w Twoim guście

Widziałem,



Chcę zobaczyć



Nie interesuje mnie



Szeregowiec Ryan (1998)

Saving Private Ryan 2 godz. 50 min. **Dramat, Wojenny** | USA

★ 8,1/10

538.779 głosów

👁 60 510

chce zobaczyć

▶ zwiastun

reżyser: **Steven Spielberg** obsada: **Tom Hanks, Tom Sizemore** dystrybucja: **ITI Home Video**

W poszukiwaniu zaginionego szeregowca wysłany zostaje doborowy oddział żołnierzy.

📊 85% w Twoim guście

Widziałem,



Chcę zobaczyć



Nie interesuje mnie





Osoby, które możesz znać



Igor Popko

📍 V Liceum Ogólnokształcące im. Augusta Witkowskiego w Krakowie

Lukasz Strączyński i 10 innych wspólnych znajomych

Dodaj znajomego

Usuń



Burcu Gulgen Kozłowska

📍 Pracuje w firmie: Shell

Paulina Zasada i 7 innych wspólnych znajomych

Dodaj znajomego

Usuń



Daniel Żelechowski

📍 Kraków

Piotr Leńczyk i 2 innych wspólnych znajomych

Dodaj znajomego

Usuń



Wojtek Kryściński

📍 Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie

Piotrek Lizończyk i 6 innych wspólnych znajomych

Dodaj znajomego

Usuń



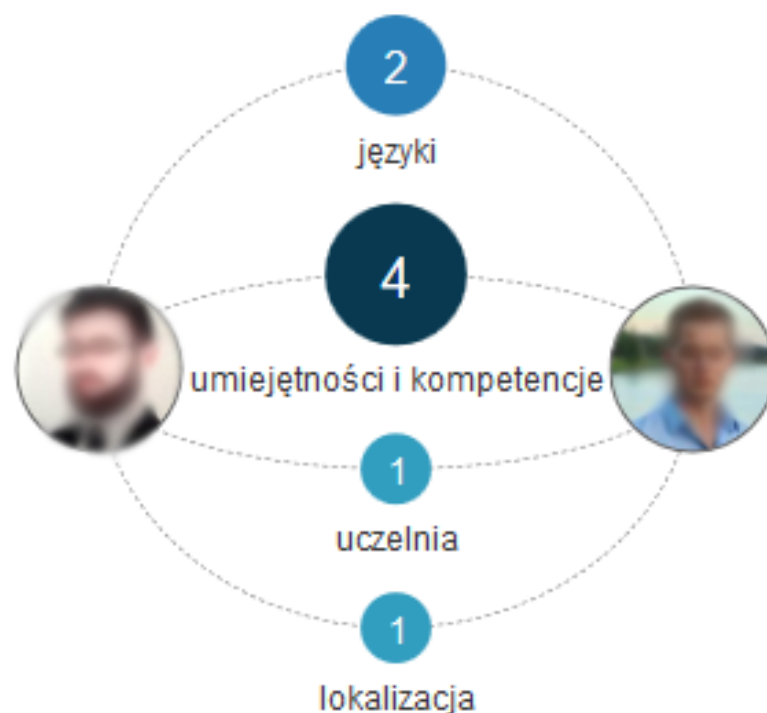
Kamila Niekraszewicz

Madek Kozłowski i Patryk Mieczkowski są wspólnymi znajomymi.

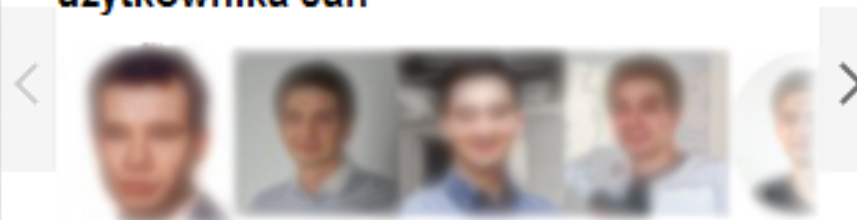
Dodaj znajomego

Usuń

Wspólne z użytkownikiem Jan



Osoby o profilu podobnym do profilu użytkownika Jan



Michał

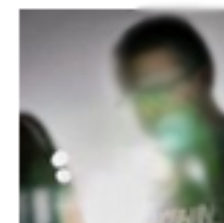
3..

Mac OS X developer

Nawiąż kontakt



Użytkownicy wyświetlili również



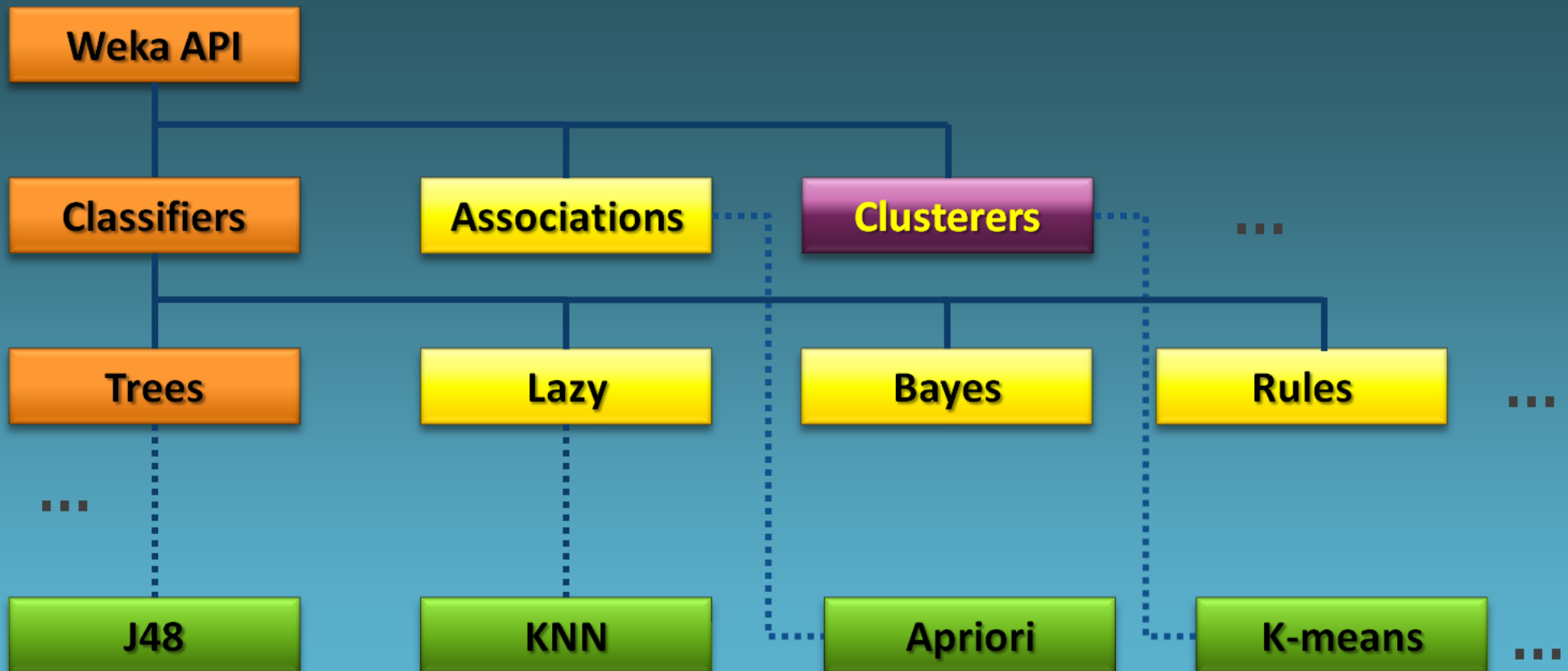
Jakub

Android Guy

WEKA = Waikato Environment for Knowledge Analysis



WEKA API



AGDS

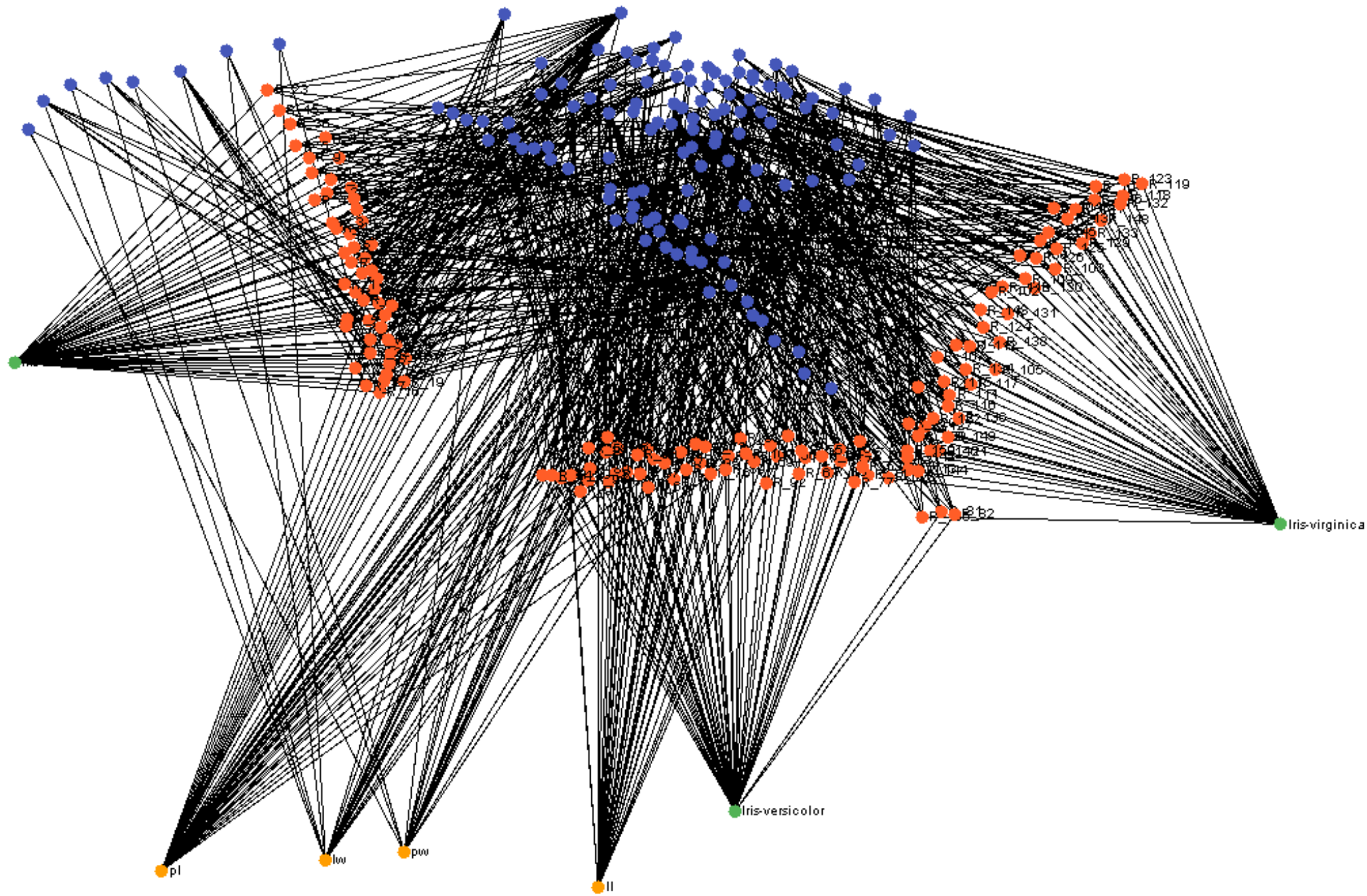
Czym jest AGDS?

- Pasywna struktura danych
- Umożliwia przechowywanie danych w ich nienaruszonej formie
- Przechowuje posortowane dane względem wszystkich atrybutów równocześnie

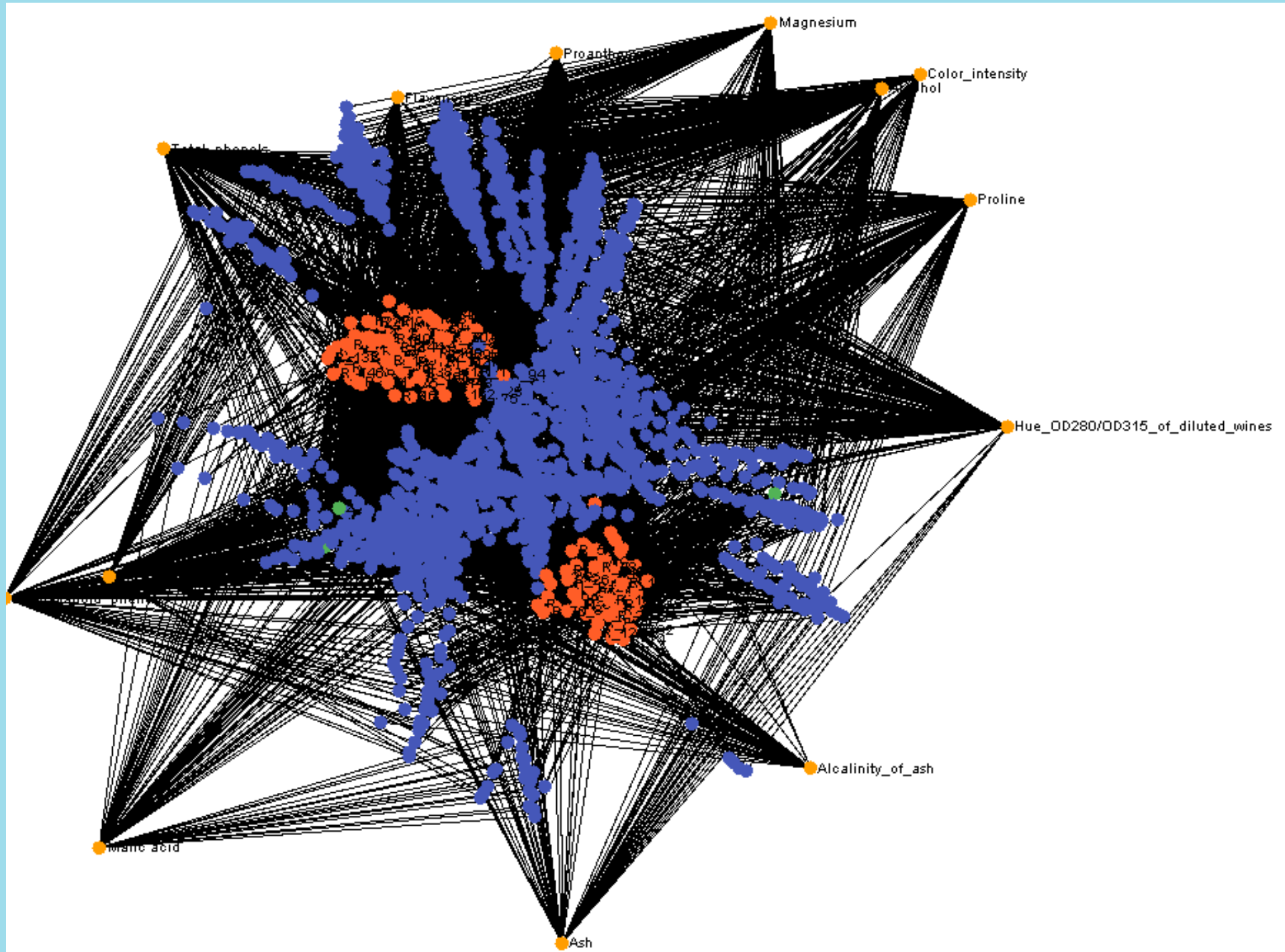
Struktura grafu

- Węzły klas (Iris Setosa, Versicolor, Virginica)
- Węzły atrybutów (długość/szerokość liścia/łodygi)
- Węzły reprezentujące wartości (np. 4.5)
- Węzły reprezentujące rekordy (wiersz tabeli)

Iris (150 rekordów, 4 atrybuty)



Wine(180 rekordów, 13 atrybutów)



Struktura AGDS: Relacyjna Baza Danych

```
905,Mikołaj,Ptak,false,3,Mikroelektronika w technice i medycynie,Kraków,Aleja Niepodległości
906,Paulina,Łukaszewski,false,4,Mikroelektronika w technice i medycynie,Kraków,Zamkowa
907,Zuzanna,Kołodziej,false,2,Informatyka,Kraków,Gliwicka
908,Teresa,Jagodziński,false,5,Informatyka,Poznań,Lwowska
909,Albert,Maliszewski,true,5,Mikroelektronika w technice i medycynie,Szczecin,Szczecińska
910,Mikołaj,Mielczarek,false,5,Mikroelektronika w technice i medycynie,Bydgoszcz,Józefa Poniatowskiego
911,Zofia,Mierzejewski,false,5,Elektrotechnika,Bydgoszcz,Lechicka
912,Ada,Sobczak,true,5,Informatyka,Bydgoszcz,Mikołaja Kopernika
913,Włodzimierz,Łukasiewicz,false,3,Inżynieria Biomedyczna,Kraków,Kamienna
914,Mirosław,Klimczak,false,1,Mikroelektronika w technice i medycynie,Katowice,Jacka Malczewskiego
915,Szymon,Urbańczyk,false,2,Informatyka,Poznań,Hetmańska
916,Fabian,Karwowski,false,1,Informatyka,Lublin,Katedralna
917,Gaweł,Jurkiewicz,false,4,Automatyka i Robotyka,Kielce,Piwna
918,Leszek,Muszyński,false,5,Mikroelektronika w technice i medycynie,Gdańsk,Browarna
919,Sara,Ratajczak,false,1,Automatyka i Robotyka,Poznań,Krakowska
920,Helga,Krajewski,true,5,Mikroelektronika w technice i medycynie,Lublin,Chłodna
921,Agnieszka,Zych,false,4,Automatyka i Robotyka,Szczecin,Gnieźnieńska
922,Gaweł,Rak,false,5,Inżynieria Biomedyczna,Wrocław,Adama Mickiewicza
923,Eliza,Sienkiewicz,true,4,Informatyka,Kielce,Grodzka
924,Agnieszka,Lisowski,false,5,Elektrotechnika,Lublin,Francuska
925,Maria,Świątek,false,1,Elektrotechnika,Gdańsk,1 Maja
926,Tomasz,Małek,false,2,Automatyka i Robotyka,Kielce,Piekna
```

Zalety AGDS

- Posortowane wartości węzłów
- dostęp $O(1)$ do najmniejszego/największego węzła
- Utrwalenie relacji pomiędzy danymi
- Brak powielania danych
- Szybsza analiza danych

Wady AGDS

- Skomplikowany proces dodawania pojedynczego wężła
- Długi proces budowy grafu

Porównanie czasów

Iris – 1500 rekordów

Porównanie czasów

Znajdowanie minimum/maksimum

AGDS	TABELA
7 675 ns	2 916 407 ns

Porównanie czasów

Filtracja względem wartości atrybutu

AGDS	TABELA
329 092 ns	1 486 427 ns

Porównanie czasów

Filtracja względem przedziału
wartości atrybutu

AGDS	TABELA
611 549 ns	2 003 426 ns

Uniwersalność

```
public class ValueNode extends AbstractNode {  
    private double value;
```

```
public class GenericValueNode<T extends Comparable<T>> extends AbstractNode {  
    private T value;
```

Aplikacja okienkowa

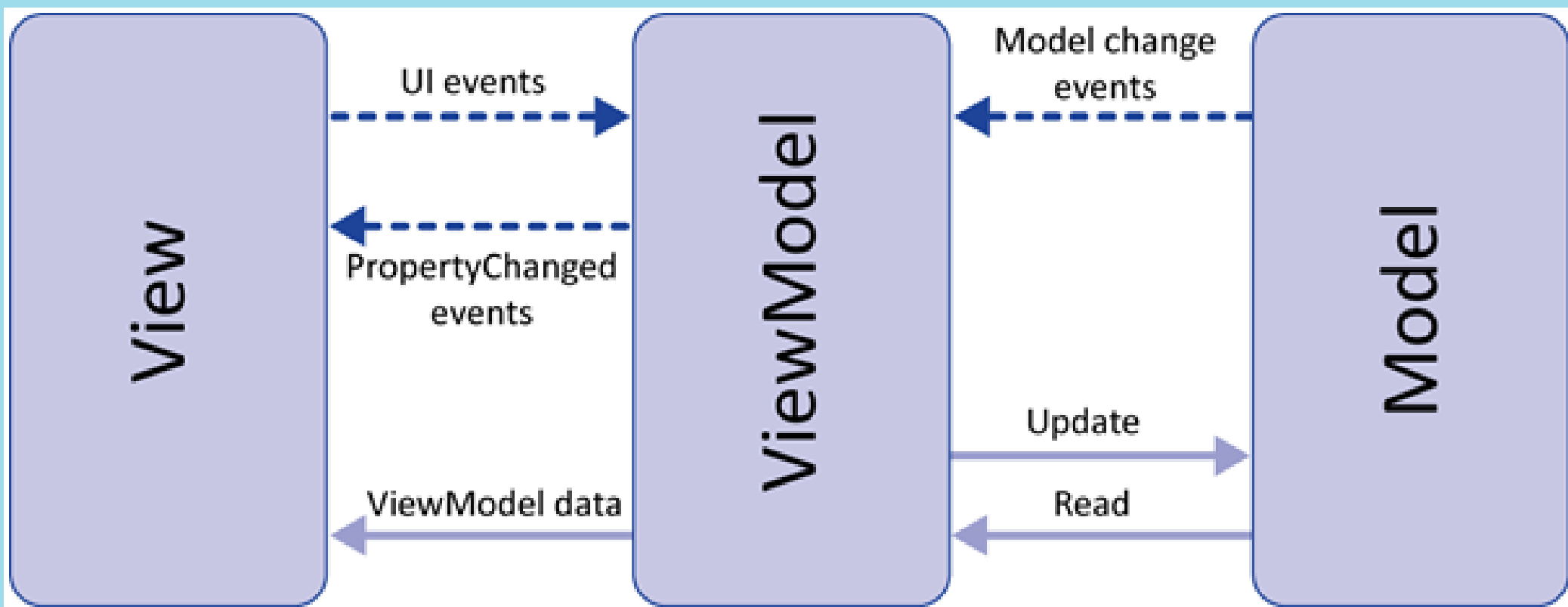
- Java
- Biblioteka Swing(graficzny interfejs użytkownika)

Obiektowa reprezentacja danych

- Elastyczność w dodawaniu struktur o różnej budowie

```
/**  
 * ReadStrategy tells DataReader how to save line as object  
 *  
 * @param <DESTINATION>  
 */  
public interface ReadStrategy<DESTINATION> {  
    DESTINATION createNewRow(String line);  
}
```

Wzorzec projektowy MVVM



Klasyfikacja węzłów

Methods of Knowledge Engineering

Graph **Classify** Similarity Correlation Filter Other

[2, 0.5, 1, 3]
[4, 2, 0.5, 6]

Add

Classify

Insert new Leaf params

3	2	1	2
---	---	---	---

Add

Klasyfikacja węzłów (Iris)

Methods of Knowledge Engineering

Graph Classify Similarity Correlation Filter Other

[0.2, 1.3, 3, 6]
[0.2, , 3.5,]
[, 1.7, , 4.4]

Add

Classify

Items classified to

ll	lw	pl	pw	Class name
0.3	1.3	2.3	4.5	Iris-setosa
0.2	1.5	3.1	4.6	Iris-setosa
0.2	1.4	3.5	5.1	Iris-setosa
0.2	1.3	3.2	4.7	Iris-setosa

OK

Węzły podobne do wybranych(Wine)

Methods of Knowledge Engineering

Graph Classify **Similarity** Correlation Filter Other

Find most similar Reload

similarity result:

Color_intensity	Hue_OD280/OD315_of_diluted_wines	Proline	Class name
16.8	95.0	1280.0	1
15.2	112.0	1450.0	1
16.2	117.0	795.0	1
16.0	126.0	780.0	1

OK

Korelacja między węzłami

The screenshot shows the 'Methods of Knowledge Engineering' application window. The 'Correlation' tab is selected in the top menu. A list of nodes is displayed on the left, each with a label and a vector of five values. Node R_93 is highlighted in green. A 'Compute' button is located above the list. A dialog box is open in the center, displaying the result of a correlation calculation.

Methods of Knowledge Engineering

Graph Classify Similarity **Correlation** Filter Other

Compute

R_88 : [1.3,2.3,4.4,6.3]
R_89 : [1.3,3.0,4.1,5.6]
R_90 : [1.3,2.5,4.0,5.5]
R_91 : [1.2,2.6,4.4,5.5]
R_92 : [1.4,3.0,4.6,6.1]
R_93 : [1.2,2.6,4.0,5.8]
R_94 : [1.0,2.3,3.3,5.0]
R_95 : [1.3,2.7,4.2,5.6]
R_96 : [1.2,3.0,4.2,5.7]
R_97 : [1.3,2.9,4.2,5.7]
R_98 : [1.3,2.9,4.3,6.2]
R_99 : [1.1,2.5,3.0,5.1]
R_100 : [1.3,2.8,4.1,5.7]
R_101 : [2.5,3.3,6.0,6.3]

Result is 0.9071702674358986

OK

KONIEC