

# Przeszłość odczytana z bazy danych Niepewność odczytu

*Bazy danych dla archeologów*

Prezentacja: **Paweł Borycki**

Instytut Archeologii  
Uniwersytet Warszawski

16 stycznia 2018

<http://pborycki.eu/pdf/yad-vashem.pdf>

# Plan prezentacji

## 1. Wstęp

- Yad Vashem Names Project
- Raport o ofiarach Yad Vashem Names Project

## 2. Uncertain Entity Resolution

- Entity Resolution (ER)
- Miękkie klastrowanie
- Ranked Resolution na Alternujących Drzewach Decyzyjnych

## 3. Przygotowanie danych

- Tagowanie przez historyków-ekspertów
- Cechy podobieństwa rekordów

## 4. Wykonanie i spostrzeżenia

- Wzory danych
- Klasyfikator ADT i jego dokładność
- Jakość systemu (*ItalySet*)

## 5. Podsumowanie

# Yad Vashem Names Project

- Analogowe źródła danych Yad Vashem
  - Filmy, nagrania dźwiękowe, dokumenty, zdjęcia
- Dygitalizacja
  - Ekstrakcja metadanych i tekstów
  - Grafy sieci społecznościowych
    - Techniki OCR
  - Dużo danych nieustrukturyzowanych
  - Różne schematy źródłowe, języki, alfabety
  - Polityczna i historyczna specyfika semantyki
- Cele
  - Ranking spójnych narracji historycznych
    - Generowany automatycznie
    - Osoba, rodzina, społeczność, miejsce
    - Procedura wykonywana przez ekspertów-historyków
  - Jednolity interfejs zintegrowanych danych

# Yad Vashem Names Project

- <https://yvng.yadvashem.org/>

[The Holocaust Remembrance](#) [Digital Collections](#) [Archives](#) [Research](#) [Education & E-learning](#) [Museums](#) [Exhibitions](#) [Righteous](#) [Visiting](#)

VB-110.1

THE CENTRAL DATABASE OF SHOAH VICTIMS' NAMES BETA

DE FR

[Refine Your Search](#) [Search](#) [Advanced Search](#)

121 Results for: [Last / Maiden Name = Frank](#) [First Name = Anne](#)

Page 1 of 3

|   | First Name | Last Name... | Birth Year | Place of Residence ?           | Source                         | Fate b... | Ma |
|---|------------|--------------|------------|--------------------------------|--------------------------------|-----------|----|
| ▶ | Naatje     | Bachrach     | 1889       | 'S Gravenhage, The Netherlands | List of murdered jews from the | murd...   | 📍  |

# Yad Vashem Names Project

- Historia projektu Yad Vashem Names Project
  - 1953 – Założenie Yad Vashem
  - 1954 – Początek rejestracji nazwisk ofiar
    - “Strony Świadectwa”
  - 1957 – 800 000 nazwisk zarejestrowanych
  - 1980. – 14-15 tys. nazwisk rejestrowanych rocznie
  - 1990. – 30 tys. nazwisk rejestrowanych rocznie
  - 1991 – Początek dygitalizacji bazy danych
  - 1998 – 500 000 nazwisk z list deportacyjnych zdigitalizowanych (OCR)
  - 1999 – 1,6 mln “Stron Świadectwa” zdigitalizowanych
    - Szwajcarskie uśpione konta bankowe
  - 2004 – Centralna Baza Danych Nazwisk Ofiar  
Zagłady dostępna w Internecie
    - 14 mln wejść w 2014 roku

# Yad Vashem Names Project

- Historia Yad Vashem Names Project
  - Żydzi i Sprawiedliwi
  - 500 000 źródeł
    - “Strony Świadectwa”
    - Listy transportowe i deportacyjne
    - Listy więźniów obozów i gett
    - Rejestry policyjne
    - Konfiskaty majątku
    - Pamiętniki i wspomnienia (np. Massimo Foa)
  - Problemy językowe
    - Alfabet – hebrajski, łaciński, cyrylica i grecki
    - 30 języków
    - Odszyfrowywanie pisma ręcznego
    - Różna pisownia imion, nazwisk i nazw miejsc
    - Błędy w pisowni (zachowane dla precyzji)
  - Klasy równoważności (abstrakcji)
    - Imiona, Nazwiska, Miejsca, Zawody, Tytuły zawodowe i naukowe, Relacje rodzinne



# Raport o ofiarach w Yad Vashem Names Project

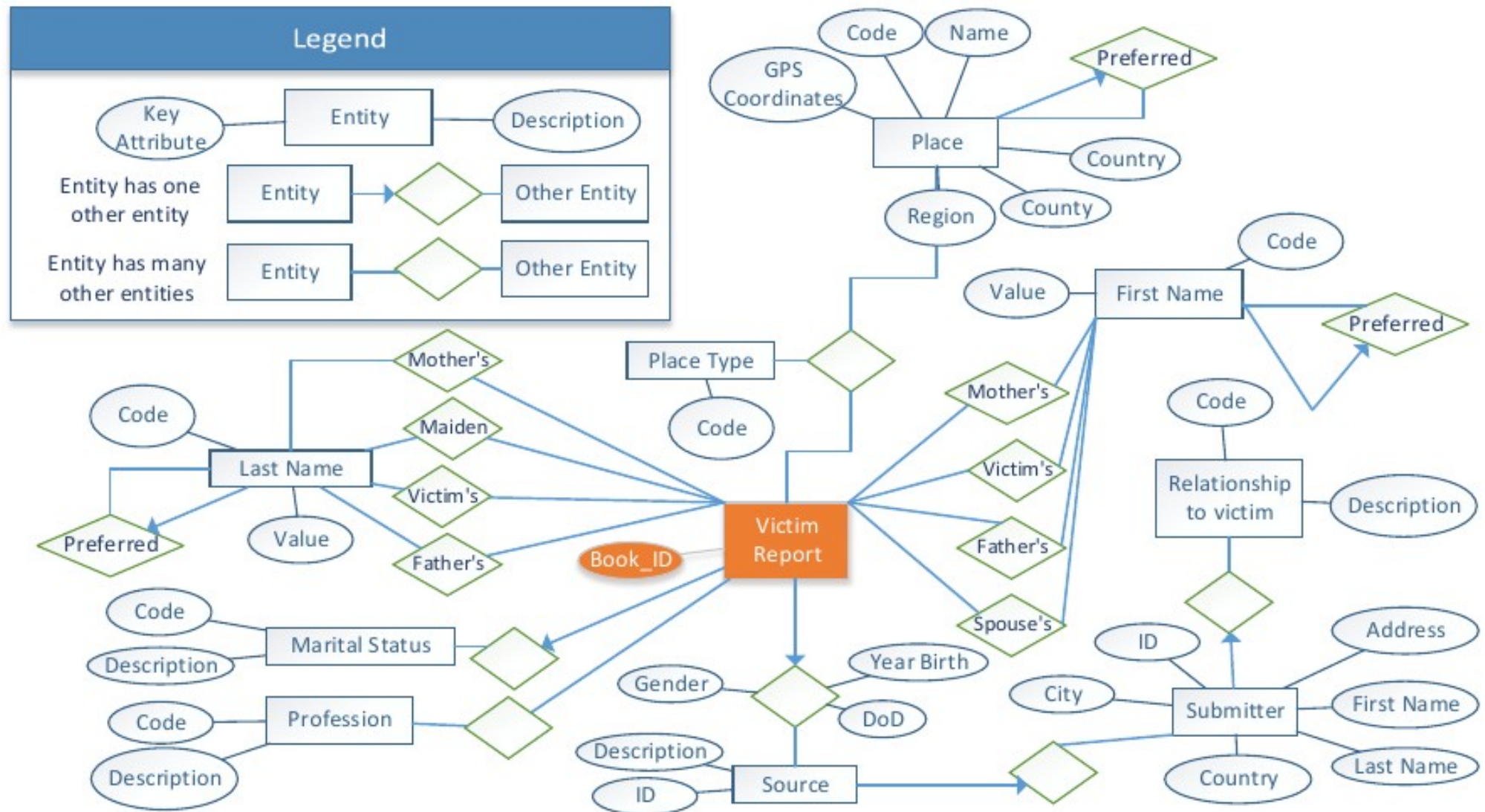


Figure 3: Entity Relationship Diagram of the Names Project Database

# Raport o ofiarach w Yad Vashem Names Project

- Raport o ofiarach w Yad Vashem Names Project
  - 6,5 mln rekordów – raportów o ofiarach
    - W tym 2,1 mln “Stron Świadectwa”
    - 16 656 list ofiar
  - Autorzy “Stron Świadectwa”
    - Brak unikalnego ID autora “Stron Świadectwa”
    - Co najmniej 514 251 różnych autorów



# Raport o ofiarach z rodziny Foa

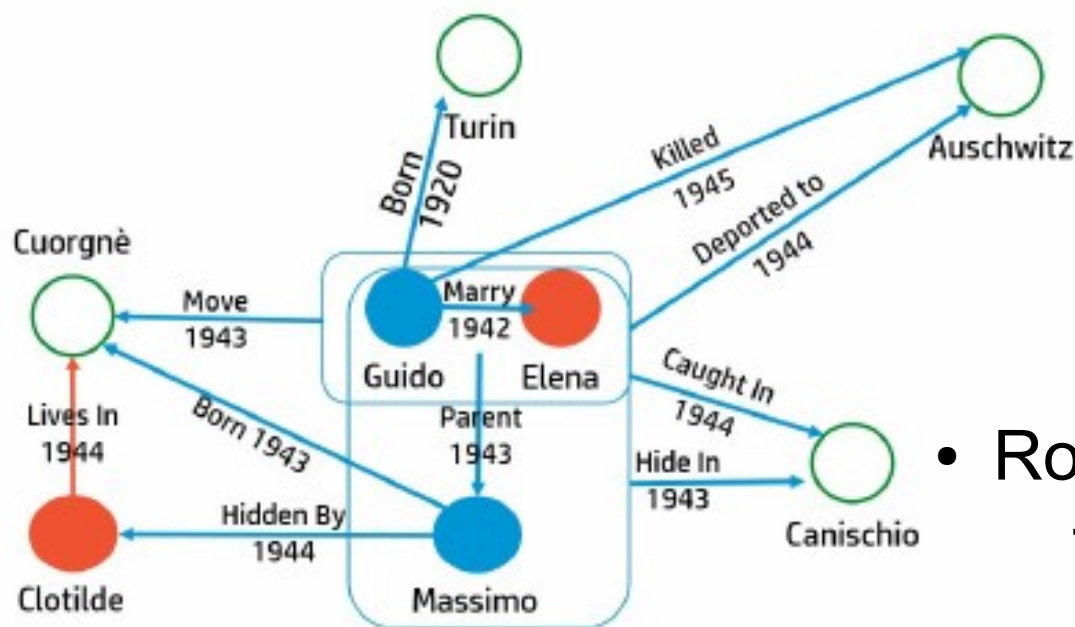


Figure 2: Knowledge graph of Guido Foa



Figure 1: Guido and Massimo Foa, 1944

- Rodzina Foa
  - Guido Foa
    - Ofiara
  - Massimo Foa
    - Uratowany przez Clotilde Boggio
    - Historyk

| BookID  | First | Last | Gender | DOB        | Birth         | Permanent        | War          | Death     | Spouse | Mother | Father |
|---------|-------|------|--------|------------|---------------|------------------|--------------|-----------|--------|--------|--------|
| 1016196 | Guido | Foa  | Male   | 1936-08-02 | Torino, Italy | Torino, Italy    |              |           |        | Estela | Italo  |
| 1059654 | Guido | Foa  | Male   | 1920-11-18 | Torino, Italy | Torino, Italy    |              | Auschwitz | Helena | Olga   | Donato |
| 1028769 | Guido | Foy  | Male   | 1920-11-18 | Turin, Italy  | Canischio, Italy | Turin, Italy |           |        | Olga   | Donato |

Table 1: Three victim reports from the Yad Vashem Names Project DB

# Uncertain Entity Resolution

- Entity resolution (ER)
  - Zadanie integracji danych
  - Tworzenie pojedynczej encji z wielu rekordów danych
    - **Brak wspólnego identyfikatora**
  - Łączenie kont klientów
  - Algorytm blokujący encje dla klastrów – *MFIBlocks*
    - Nakładające się klastry o niezerowym przecięciu
  - Alternujące drzewa decyzyjne (ADT)  
i uczenie maszynowe
    - Przekształcanie klastrów w encje
    - Encja – asocjacja rekordów z rankingiem

| BookID  | First | Last | Gender | DOB        | Birth         | Permanent        | War          | Death     | Spouse | Mother | Father |
|---------|-------|------|--------|------------|---------------|------------------|--------------|-----------|--------|--------|--------|
| 1016196 | Guido | Foa  | Male   | 1936-08-02 | Torino, Italy | Torino, Italy    |              |           |        | Estela | Italo  |
| 1059654 | Guido | Foa  | Male   | 1920-11-18 | Torino, Italy | Torino, Italy    |              | Auschwitz | Helena | Olga   | Donato |
| 1028769 | Guido | Foy  | Male   | 1920-11-18 | Turin, Italy  | Canischio, Italy | Turin, Italy |           |        | Olga   | Donato |

Table 1: Three victim reports from the Yad Vashem Names Project DB

# Uncertain Entity Resolution

- Entity resolution (ER)
  - Porównać wszystkie pary rekordów?
    - Kwadratowa złożoność
  - Blokowanie w klastry
    - Algorytm MFIBlock
    - Porównanie par rekordów tylko w obrębie każdej grupy
    - Liczba porównań zmniejszona o 87-97%
    - Akceptowalny spadek czułości
  - Probabilistyczna baza danych
    - Wiele możliwych wyników ER
    - Relacja Same-as z miarą prawdopodobieństwa

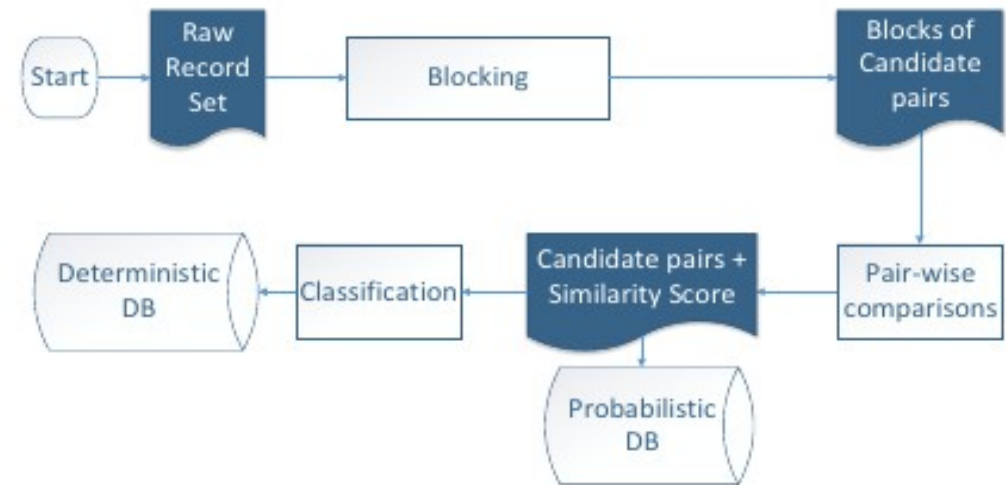


Figure 4: Generic ER Process

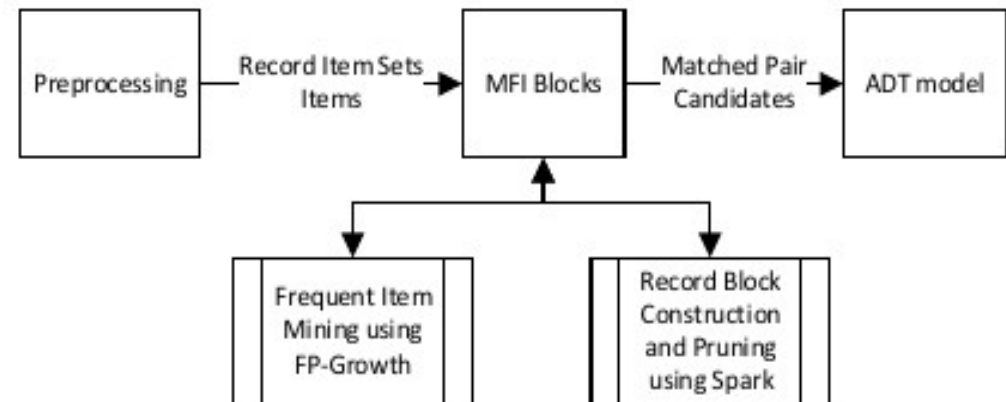
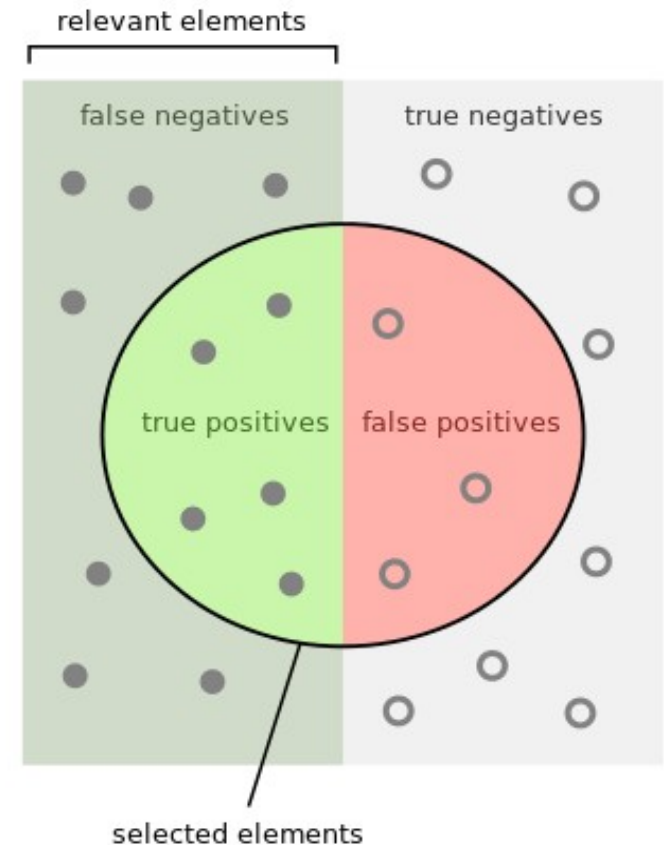


Figure 9: System Architecture

# Miary jakości klasyfikacji

- Jakość klasyfikacji
  - Prawdziwie dodatni (TP)
  - Prawdziwie ujemny (TN)
  - Fałszywie dodatni (FP)
  - Fałszywie ujemny (FN)
- Dokładność (ang. *accuracy*)
  - $\text{acc} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$
- Precyzja (ang. *precision*)
  - $\text{prec} = (\text{TP}) / (\text{TP} + \text{FP})$
- Czułość (ang. *recall*)
  - $\text{rec} = (\text{TP}) / (\text{TP} + \text{FN})$



How many selected items are relevant?

Precision =  $\frac{\text{TP}}{\text{TP} + \text{FP}}$

How many relevant items are selected?

Recall =  $\frac{\text{TP}}{\text{TP} + \text{FN}}$

# Miękkie klastrowanie

- Miękkie klastrowanie (ang. *Soft Clustering*)
  - Miękkie klastry mogą współdzielić rekordy
  - Błędy klastrowania
    - Fałszywie ujemny (FN) – pasujące rekordy w różnych klastrach (niska czułość)
    - Fałszywie dodatni (FP) – niepasujące rekordy w tym samym klastrze (niska precyzja)
  - Cechy algorytmu MFIBlocks
    - Klucz blokowania nie wybierany ręcznie (zwiększa czułość)
    - Dynamiczny, automatyczny and elastyczny wybór klucza blokowania
    - Różne klucze dla różnych bloków, rozmiar bloku zrównoważony
    - Zbiór zwarty
    - Rzadkie (luźne) sąsiedztwo
    - Wiele poziomów granularności dozwolonych dla analizy osób, rodzin i społeczności (miejsc)
    - Dopasowywany rozmiar klastra

# Miękkie klastrowanie

- Maximally Frequent Itemsets (MFI)

| rid     | items                                                                             |
|---------|-----------------------------------------------------------------------------------|
| 1000069 | YB_1927, P1_Lubaczow, P2_Lubaczo, P3_Lwow, P4_Poland, F_Avraham, L_Kesler         |
| 1000069 | P1_Lwow, P2_Lwow, P3_Lwow, P4_Poland, F_Avraham, L_Apoteker, G_0                  |
| 1000069 | P1_Antopol, P2_Kobryn, P3_Polesie, P4_Poland, F_Yitzhak, F_Avraham, L_Postel, G_0 |
| 1000069 | P4_Poland, F_Yitzhak, L_Postel, G_0                                               |

Table 2: Sample records in item-bag form

- Rekord = (rid, itemset)
- **Support** zbioru  $I$  – zbiór rekordów zawierających  $I$ 
  - 2 ostatnie rekordy są supportem  $I = \{F\_Yitzhak, L\_Postel, G\_0\}$
- **Frequent Itemset** – zbiór pól  $I$ , który posiada *support* większy niż stała **minsup**
- Jeśli  $I$  nie należy do innego *frequent set*,  
 $I$  jest **Maximally Frequent Itemset**



# Miękkie klastrowanie

- Algorytm MFIBlocks

- **MinMaxSup** – maksymalna wartość stałej *minsup* dla bloku rekordów
  - Bloki z *supportem* większym niż *MinMaxSup* są pomijane
- **p** – limit liczby bloków, w których jeden rekord może uczestniczyć

```
Input: Database D; MaxMinSup; p
Result: Set of candidate pairs M
// Set of Covered Records
1 P = ∅ ;
// Set of Candidate Pairs
2 M ← ∅ ;
3 minsup ← MaxMinSup;
4 minTh ← 0 ;
5 while minsup ≥ 2 ∧ P ≠ D do
    // mine MFI from uncovered records
6   MFIs ← MFI(D \ P, minsup);
7   Blocks ← FindSupport(MFIs);
    // Filter blocks
8   Blocks ← Blocks.filter(B → B.size ≤ minsup · p ) ;
    // Set minTh by NG limitation
9   foreach B ∈ Blocks do
10    foreach (ri, rj) ∈ B do
11        CandidatePairs ← (ri, rj);
12        minTh ← UpdateByNG((ri, rj));
13    end
14  end
    // filter blocks violating minTh
15  Blocks ← Blocks.filter(B → Score(B) > minTh ) ;
16  CandidatePairs ← CandidatePairs.filter(pair →
    pair.B ∈ Blocks) ;
17  P ← P ∪ {ri | ∃(ri, rj) ∈ CandidatePairs};
18  P ← P ∪ {rj | ∃(ri, rj) ∈ CandidatePairs};
19  M ← M ∪ CandidatePairs;
20  minsup – – ;
21 end
```

Algorithm 1: Simplified MFIBlocks Algorithm



# Ranked Resolution

- Ranked Resolution
  - Pary rekordów z fazy blokowania – potencjalne duplikaty
  - Parametry pewności
    - Łączna liczba ofiar – **bardziej deterministyczny**
    - Poszukiwanie zabitych krewnych – **bardziej probabilistyczny**
  - Ranking wyników
    - Wartości podobieństwa dla dopasowań obliczone przez klasyfikator
  - Alternujące drzewa decyzyjne (ADT)
    - Węzły *Splitter* i *Prediction*
    - Znak sumy wszystkich węzłów *Prediction* na ścieżce od korzenia do liścia

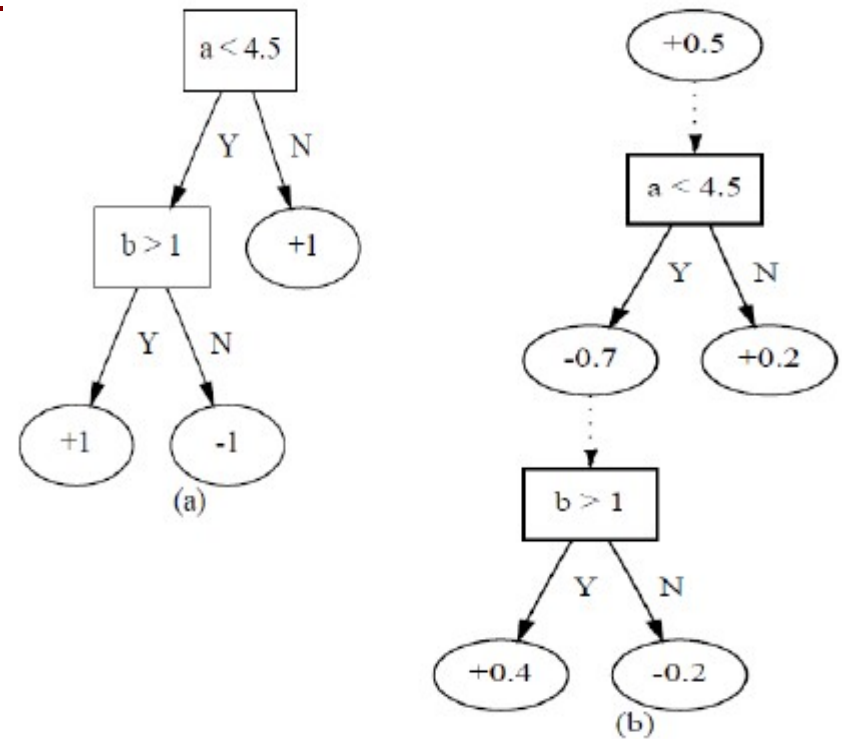


Figure 5: Decision tree standard representation (a) and as an ADTree (b) **12**

$$(a, b) = (3.9, 0.9)$$

$$\text{sign}(+0.5 - 0.7 - 0.2) = -1.$$

# Ranked Resolution

- Ranked Resolution
  - Ogólne alternujące drzewa decyzyjne (GADT)
    - Węzły *Splitter* i *Prediction*
    - Możliwe wiele ścieżek
    - Znak sumy wszystkich węzłów *Prediction* na wszystkich ścieżkach od korzenia do akceptowanych liści
  - ADTs
    - Gęstsze niż standardowe drzewa decyzyjne
  - Pominięcie operacji *sign*
    - Wynik jest miarą ufności
    - Wynik jako ranking a nie decyzja deterministyczna
  - Brakujące wartości
    - Rozpatrywane tylko osiągalne węzły

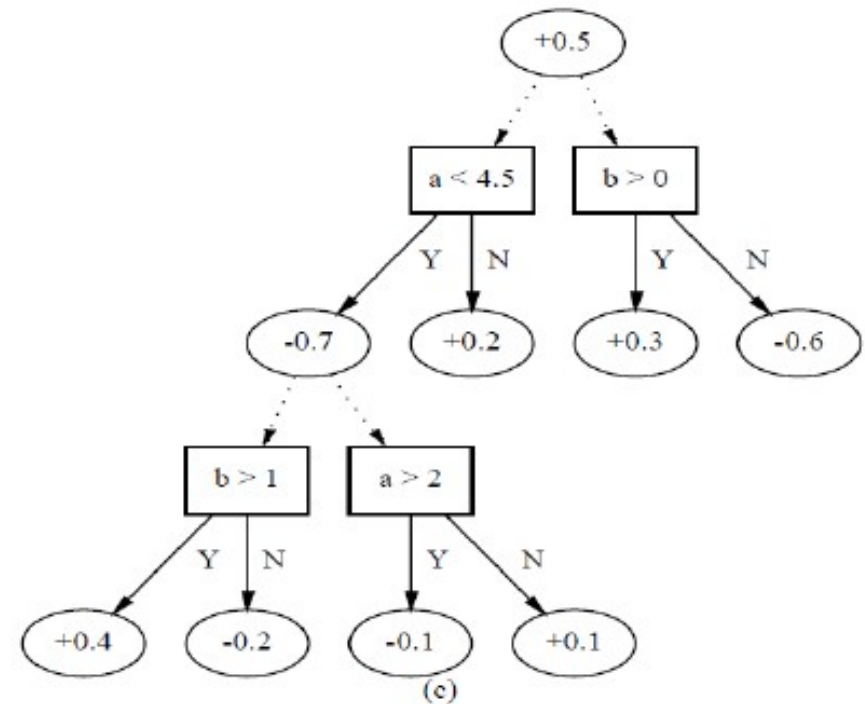


Figure 6: A general alternating tree 12

$$(a, b) = (3.9, 0.9)$$

$$\text{sign}(+0.5 + 0.3 - 0.7 - 0.1 - 0.2) = -1$$

# Przygotowanie danych

- Podzbiory danych
  - *ItalySet*
    - Wszystkie rekordy z “Italy” jako miejscem zamieszkania
    - 9,499 rekordów
  - *RandomSet*
    - Losowy podzbiór pełnego zbioru danych z 6 regionów geograficznych
    - 100,000 rekordów
- Przygotowanie danych
  - Unikalne prefiksy dla pól
    - Imię: Moshe – 'FN\_Moshe'
  - Tagowanie przez historyków
    - Dopasowania posortowane według podobieństwa
    - Podkreślenie różnic
    - Tagi {Tak, Pewnie tak, Może, Pewnie nie, Nie}

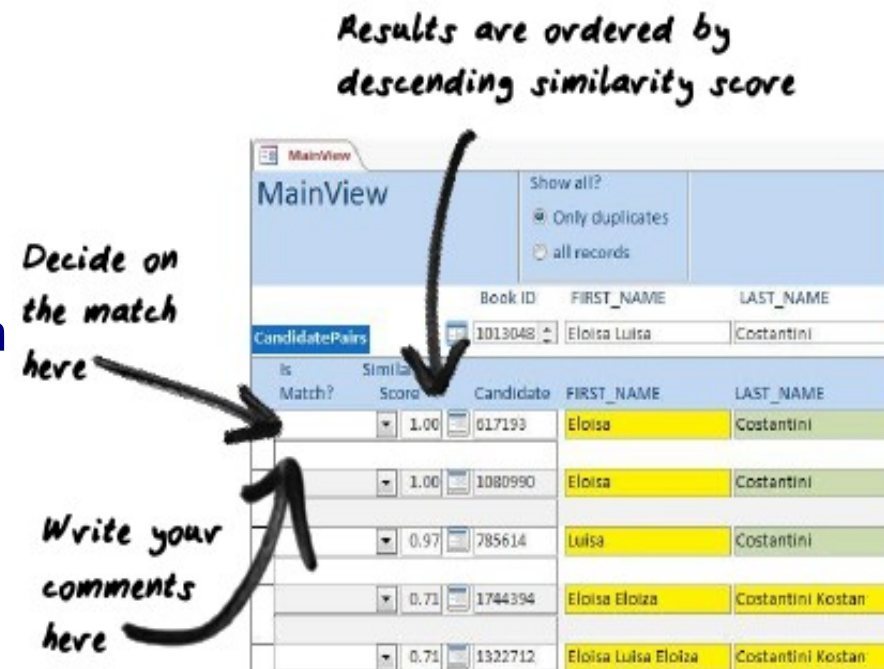


Figure 7: Screen-shot from Tagging Application User Guide

# Przygotowanie danych

- Przygotowanie danych
  - Tagowanie przez historyków-ekspertów
    - Tagi {*Tak*, *Pewnie tak*, *Może*, *Pewnie nie*, *Nie*}
    - *Może* – brak wystarczającej informacji, aby zdecydować, czy para stanowi duplikaty
    - Otagowany zbiór wykorzystany do trenowania klasyfikatora i w eksperymentach
    - Niektóre duplikaty prawdopodobnie pozostają nieotagowane (fałszywie ujemne) ze względu na duży rozmiar zbioru *ItalySet*
    - Fałszywie dodatnie duplikaty do przeanalizowania przez historyków
    - Tagowanie powtarzane

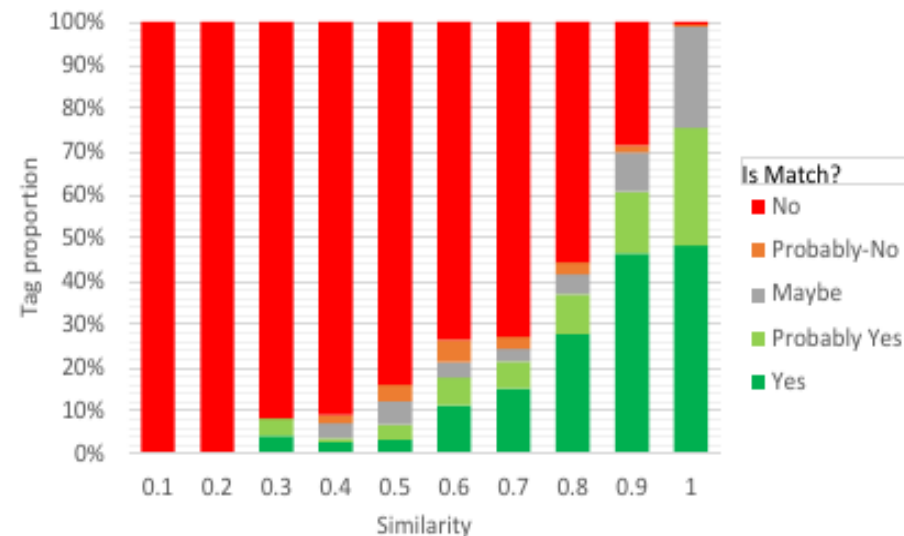


Figure 8: Tag-Similarity Comparison

*ItalySet*

# Przygotowanie danych

- Cechy podobieństwa
  - Dodane dla ułatwienia działania algorytmu ADT
  - Rodzaje porównań
    - **Binarne** – imię, nazwisko, płeć, zawód, inne atrybuty-kategorie
    - **Trinarne** – pola, gdzie mniej niż 5% rekordów ma więcej niż 3 słowa
    - **Niedyskretne** – znormalizowana odległość, współczynnik normalizacji ustalany przez ekspertów
  - 48 zdefiniowanych cech
    - **sameXName** – X: {First, Last, Spouse, Father, Mother, Mother's Maiden, Maiden}, result: {yes, partial, no}
    - **XnameDist** – odległość Jaccarda
    - **BXDist** – komponenty daty urodzenia {D, M, Y} znormalizowane przez maksymalną odległość (31 dla D, 12 dla M, 100 dla Y)
    - **samePlaceXPartY** – X: {Birth, Permanent, War-time, Death}, Y: {City, Country, Region}, result: {yes, no}
    - **PlaceXGeoDistance** – X: {Birth, Permanent, War-time, Death}, odległość w km między dwoma miejscami tego samego typu
    - **sameX** – X: {Source, Gender, Profession}, result: {yes, no}

# Przygotowanie danych

- Cechy podobieństwa
  - Miejsca różnych typów nieporównywane
    - (np. **Birth** i **Permanent**)
  - Wybranie 8-10 cech z 48 zdefiniowanych daje 95% dokładność algorytmu ADT
  - Przykład
    - Różne imiona ojca
    - Odległość Jaccarda pomiędzy imionami ojca – 0,2
    - Brak imienia matki w jednym rekordzie
    - **Podobieństwo:  $-1.3 + -0.25 = -1.55$**
  - **$\leq 0$  – brak dopasowania**
  - **$> 0$  – dopasowanie**

$$f_{sim}(i_1, i_2) = \begin{cases} 0 & type(i_1) \neq type(i_2) \\ jw(i_1, i_2) & type(i_1) = Name \\ 1 - \frac{|i_1 - i_2|}{50} & type(i_1) = Year \\ 1 - \frac{monthDiff(i_1, i_2)}{12} & type(i_1) = Month \\ 1 - \frac{dayDiff(i_1, i_2)}{31} & type(i_1) = Day \\ \text{Max}(0, 1 - \frac{geoDist(i_1, i_2)}{100}) & type(i_1) = Geo \end{cases}$$

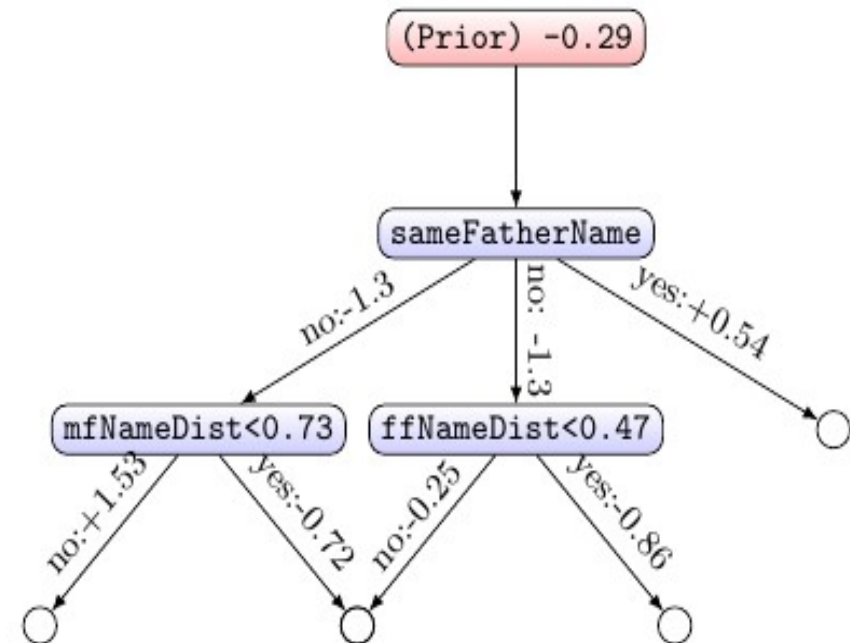


Figure 10: Fragment of Final ADT tree



# Wykonanie i spostrzeżenia

- Ustawienia

- Trening klasyfikatora
  - HP Ultrabook, 4 rdzenie, 8 GB RAM
- *MFIBlocks*
  - Serwer HP DL380, 24 rdzenie, 64 GB RAM
  - 30 GB RAM dla JVM
  - 30 GB RAM dla Apache Spark 1.2.1
  - Algorytm sekwencyjny – tylko 1 z 24 rdzeni używany przez 90% czasu

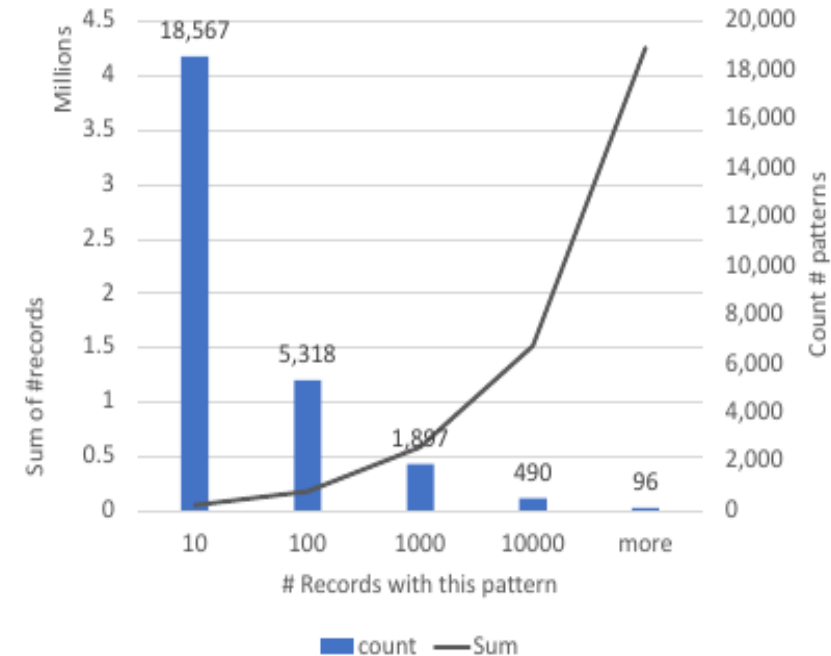


Figure 11: Data Pattern Counts

- Wzory danych

- **Wzór** – zbiór typów pól rekordu
- Liczba wzorów dla całego zbioru danych
  - 4 mln rekordów posiadają 100 wzorów
  - 19 000 wzorów reprezentowanych przez mniej niż 10 rekordów każdy
  - 0,5 mln rekordów posiada najpowszechniejszy wzór (FirstName, LastName, Gender, PermanentPlace)
  - 40 191 rekordów ma pełny wzór ze wszystkimi typami pól

# Wykonanie i spostrzeżenia

- Wzory danych
  - Podobna struktura wzorów w pełnym zbiorze i *RandomSet*
    - Dane dot. miejsc częstsze w *RandomSet*, bo wykorzystane w tworzeniu
  - *ItalySet*
    - Imię ojca ważną częścią tożsamości

| Item Type            | Full Set  |            | 10K Italy Set |            | 100K Set |            |
|----------------------|-----------|------------|---------------|------------|----------|------------|
|                      | Records   | % of Total | Records       | % of Total | Records  | % of Total |
| Last Name            | 6,478,181 | 98%        | 9,402         | 99%        | 96,656   | 97%        |
| First Name           | 6,352,481 | 97%        | 9,402         | 99%        | 97,527   | 98%        |
| Gender               | 5,778,172 | 88%        | 9,223         | 97%        | 86,474   | 86%        |
| DOB                  | 4,185,298 | 64%        | 6,345         | 67%        | 65,005   | 65%        |
| Father's Name        | 3,452,056 | 52%        | 7,392         | 78%        | 49,019   | 49%        |
| Mother's First Name  | 2,639,497 | 40%        | 5,583         | 59%        | 39,816   | 40%        |
| Spouse Name          | 1,759,845 | 27%        | 2,022         | 21%        | 25,964   | 26%        |
| Maiden Name          | 819,618   | 12%        | 1,201         | 13%        | 12,399   | 12%        |
| Mother's Maiden Name | 758,303   | 12%        | 1,954         | 13%        | 11,950   | 12%        |
| Permanent Place      | 4,634,209 | 70%        | 8,348         | 88%        | 83,308   | 83%        |
| Wartime Place        | 3,841,196 | 58%        | 6,806         | 72%        | 70,009   | 70%        |
| Birth Place          | 2,344,618 | 36%        | 8,563         | 90%        | 58,399   | 58%        |
| Death Place          | 2,253,248 | 34%        | 5,675         | 60%        | 42,541   | 43%        |
| Profession           | 2,332,427 | 35%        | 2,559         | 27%        | 27,209   | 27%        |

Table 3: Item Type Prevalence

# Wykonanie i spostrzeżenia

- Liczność zbiorów wartości danych różnych typów
  - **Niska liczność**
    - np. nazwisko
    - **Dobre znaczniki dla dopasowanych par rekordów**
    - **Podstawy klastrów encji**
  - **Wysoka liczność**
    - np. płeć
    - **Pomijane przez algorytm blokujący jako klucze bloków**
    - **Używane przez klasyfikator**
  - **Różnorodność wzorów danych**
    - **Rzadkie współdzielenie wielu typów danych o wysokiej liczności**

| Item Type     | 10K Italy Set |          | 100K Set |          |
|---------------|---------------|----------|----------|----------|
|               | Items         | Rec/Item | Items    | Rec/Item |
| Last Name     | 1,495         | 6        | 20,541   | 5        |
| First Name    | 775           | 12       | 1,544    | 63       |
| Gender        | 2             | 4,612    | 2        | 43,237   |
| Maiden Name   | 505           | 2        | 5,019    | 2        |
| MMN           | 496           | 4        | 4,731    | 3        |
| MFN           | 382           | 15       | 825      | 48       |
| Profession    | 355           | 7        | 2,115    | 13       |
| Spouse Name   | 430           | 5        | 955      | 27       |
| Father's Name | 372           | 20       | 777      | 63       |
| Birth Day     | 31            | 127      | 31       | 738      |
| Birth Month   | 12            | 332      | 12       | 2,068    |
| Birth Year    | 98            | 64       | 119      | 541      |
| Birth City    | 670           | 12       | 9,519    | 6        |
| Birth County  | 278           | 19       | 1,112    | 44       |
| Birth Region  | 91            | 86       | 151      | 223      |
| Birth Country | 33            | 259      | 52       | 1,116    |
| War City      | 290           | 23       | 6,976    | 8        |
| War County    | 114           | 30       | 1,715    | 25       |
| War Region    | 54            | 116      | 195      | 218      |
| War Country   | 20            | 338      | 45       | 1,556    |
| Perm. City    | 226           | 36       | 8,328    | 10       |
| Perm. County  | 73            | 70       | 1,051    | 73       |
| Perm. Region  | 19            | 425      | 81       | 800      |
| Perm. Country | 1             | 8,348    | 23       | 3,622    |
| Death City    | 209           | 26       | 3,802    | 11       |
| Death County  | 87            | 7        | 924      | 13       |
| Death Region  | 53            | 19       | 238      | 80       |
| Death Country | 18            | 310      | 45       | 945      |

# Klasyfikator ADT

- Dokładność klasyfikatora ADT
  - Tagi: *{Tak, Może, Nie}*
  - Czy zaliczyć pary z tagiem *może* do zbioru treningowego?
    - 611 par z tagiem *może* z 10 017 otagowanych par
  - Autor *MV* dostarczył 1400 z 9499 rekordów w *ItalySet*
    - Większość autorów dostarcza 1-5 “Stron Świadectwa”
    - 2 autorów, w tym MV, dostarczyło ponad 1000 “Stron Świadectwa”
    - Unikalny wzór wszystkich rekordów MV:  
(FirstName, LastName, FatherName, BirthPlace, DeathPlace)
    - 3,183 par zawierających rekordy MV z 10,017 par
    - Ryzyko zbytniego dopasowanie modelu do zbioru *ItalySet*

| Condition             | N      | Accuracy |
|-----------------------|--------|----------|
| Maybe:=No             | 10,016 | 94.2%    |
| Maybe values omitted  | 9,406  | 96.4%    |
| Identify Maybe values | 10,016 | 95.1%    |

Table 5: Classifier Quality - *Maybe* values

| Condition  | N     | Accuracy |
|------------|-------|----------|
| With MV    | 9,406 | 96.5%    |
| Without MV | 6,833 | 94.2%    |

Table 6: Classifier Quality - *Maybe* values

# Klasyfikator ADT

## • Dokładność klasyfikatora ADT

- Autor **MV** dostarczył 1400 z 9499 rekordów w *ItalySet*
  - **Pominięcie rekordów MV** powoduje, że model:
  - **Kładzie mniejszy nacisk na imię ojca (FFN)**
  - **Kładzie większy nacisk na imię (FN)**
  - **Lepiej reprezentuje zachowanie podzbioru losowego**

|                    |                      |        |
|--------------------|----------------------|--------|
| -0.289             |                      |        |
| (1)sameFFN = no:   | -1.314               |        |
|                    | (6)MFNdists < 0.728: | -0.718 |
|                    | (6)MFNdists ≥ 0.728: | 1.528  |
|                    | (8)FFNdists < 0.471: | -0.863 |
|                    | (8)FFNdists ≥ 0.471: | -0.247 |
| (1)sameFFN ≠ no:   | 0.539                |        |
| (2)sameFN = no:    | -1.475               |        |
|                    | (7)FNdist < 0.728:   | -0.371 |
|                    | (7)FNdist ≥ 0.728:   | 1.241  |
| (2)sameFN ≠ no:    | 0.791                |        |
|                    | (5)FFNdists < 0.609: | -0.734 |
|                    | (5)FFNdists ≥ 0.609: | 0.757  |
|                    | (10)SNdist < 0.738:  | -0.982 |
|                    | (10)SNdist ≥ 0.738:  | 0.786  |
| (3)B3dist < 1.5:   | 1.142                |        |
| (3)B3dist ≥ 1.5:   | -0.29                |        |
| (4)LNdist < 0.671: | -1.454               |        |
| (4)LNdist ≥ 0.671: | -0.078               |        |
|                    | (9)MNdist < 0.606:   | -0.978 |
|                    | (9)MNdist ≥ 0.606:   | 1.58   |

Table 7: Full dataset ADT model

|                    |                      |        |
|--------------------|----------------------|--------|
| -0.132             |                      |        |
| (1)sameFN = no:    | -1.461               |        |
|                    | (9)FFNdists < 0.539: | -0.486 |
|                    | (9)FFNdists ≥ 0.539: | 0.433  |
| (1)sameFN ≠ no:    | 0.61                 |        |
|                    | (2)sameFFN = no:     | -1.296 |
|                    | (2)sameFFN ≠ no:     | 1.171  |
|                    | (5)MFNDists < 0.708: | -0.962 |
|                    | (5)MFNDists ≥ 0.708: | 0.557  |
|                    | (7)DPGeoDist < 223:  | -0.532 |
|                    | (7)DPGeoDist ≥ 223:  | 1.088  |
| (3)B3dist < 4.5:   | 0.674                |        |
|                    | (10)B3dist < 1.5:    | 0.431  |
|                    | (10)B3dist ≥ 1.5:    | -0.379 |
| (3)B3dist ≥ 4.5:   | -0.906               |        |
| (4)LNdist < 0.671: | -1.292               |        |
| (4)LNdist ≥ 0.671: | -0.1                 |        |
|                    | (6)MNdist < 0.606:   | -0.938 |
|                    | (6)MNdist ≥ 0.606:   | 1.516  |
| (8)SNdist < 0.663: | -0.765               |        |
| (8)SNdist ≥ 0.663: | 0.978                |        |

Table 8: ADT model without MV records

# Jakość systemu – *ItalySet*

- Jakość wyników otrzymanych dla *ItalySet*
  - Parametry konfiguracji
    - **Neighborhood Growth (NG)** – rozmiar dozwolonego przecięcia pomiędzy klastrami
    - **MaxMinSup** – początkowe minimalny rozmiar supportu konieczny do uznania za *Frequent Itemset*, potem *minsup* zmniejsza się w każdej iteracji
    - **Expert Item Similarity (ExpertSim)** – obliczanie odległości dla dat, nazw i miejsc z wykorzystaniem wiedzy eksperckiej, a nie tylko na podstawie odległości geograficznej lub Podobieństwa Jaccarda
    - **Same Source Discard (SameSrc)** – pomijanie par, jeśli oba rekordy pochodzą z tego samego źródła
    - **Classification (Cls)** – klasyfikator ADT filtruje dopasowania o niskiej ufności
  - **NG i MaxMinSup** dopasowane w celu otrzymania optymalnych rezultatów



# Jakość systemu – *ItalySet*

- Jakość wyników w *ItalySet*
  - Rodzina Capelluto
    - Elsa Capelluto (age 11)
    - Giulia Capelluto (age 13)
    - Alberto Capelluto (age 5)
  - Wspólne wartości pól
    - LastName, FatherName, MotherName, PermanentPlace
  - Fałszywie dodatni jako osoba
    - Można uniknąć stosując parametr SameSrc (ich ciotka)
  - Prawdziwie dodatni jako rodzina
    - Zadanie Entity Resolution dot. rodzin

| FIRST_NAME | LAST_NAME | DOB        | Gender | Father | Mother           |
|------------|-----------|------------|--------|--------|------------------|
| Elsa       | Capelluto | 04 10 1933 | Female | Khaim  | Zimbul Schiarkon |

| FIRST_NAME | LAST_NAME | DOB        | GENDER | Father | Mother           |
|------------|-----------|------------|--------|--------|------------------|
| Giulia     | Capelluto | 24 11 1931 | Female | Khaim  | Zimbul Schiarkon |

|                                                                                     |         |           |            |      |                        |
|-------------------------------------------------------------------------------------|---------|-----------|------------|------|------------------------|
|  | Alberto | Capelluto | 04 01 1939 | Male | Khaim Zimbul Schiarkon |
|-------------------------------------------------------------------------------------|---------|-----------|------------|------|------------------------|

Figure 13: Meaningful False Positives - The Capelluto Children



Figure 14: Capelluto Family Picture

# Jakość systemu – *ItalySet*

- Jakość wyników w *ItalySet*

- Współczynnik F-1

- $F1 = 2 * (\text{Precyzja} * \text{Czułość}) / (\text{Precyzja} + \text{Czułość})$

- Inne konfiguracje Neighborhood Growth (NG) i MaxMinSup

- Najwyższy współczynnik F-1 dla:

- NG = 3.5 dla MaxMinSup = 4

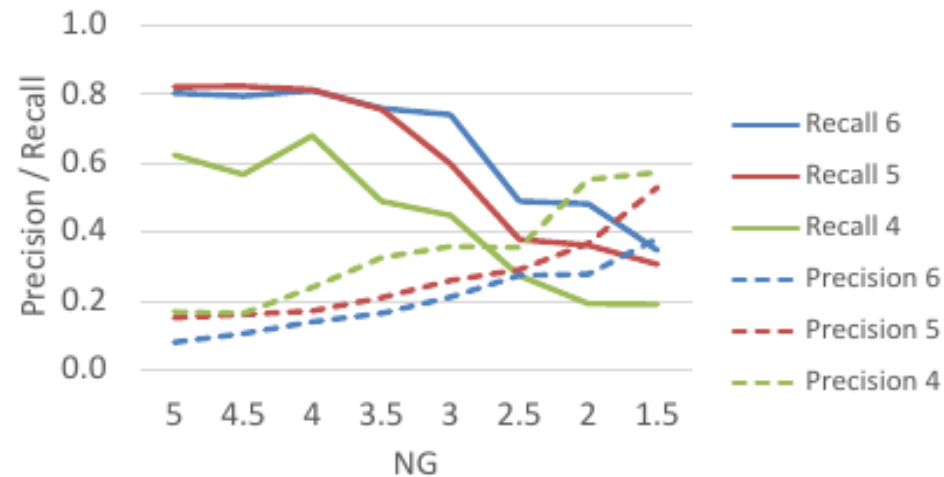
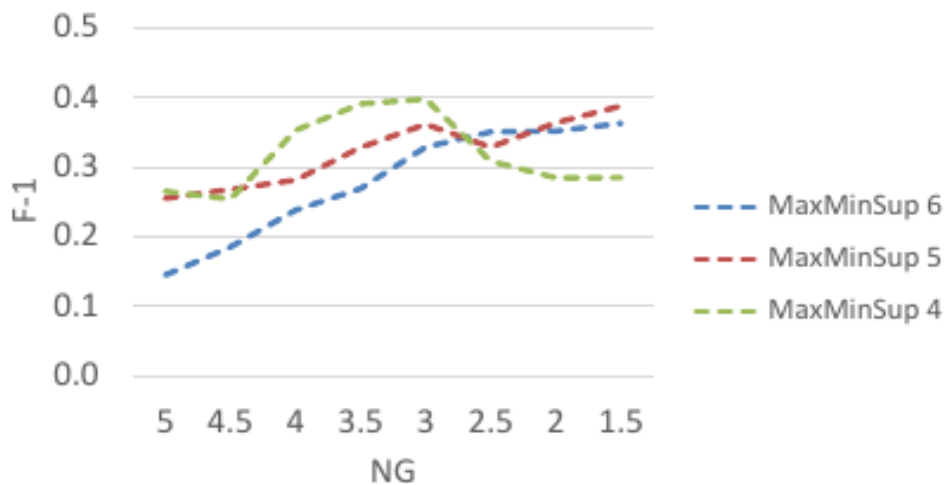
- NG = 3.0 dla MaxMinSup = 5 lub 6

- Czułość ważniejsza od Precyzji

- Wyniki filtrowane za pomocą parametrów classifier i sameSrc

- Maksymalizacja Czułości z małą stratą Precyzji:

- NG = 3, 3,5 lub 4; MaxMinSup = 5 – optymalna konfiguracja



# Jakość systemu – *ItalySet*

- Jakość wyników w *ItalySet*
  - Czułość ważniejsza od Precyzji
    - NG = 3, 3,5 lub 4; MaxMinSup = 5 – optymalna konfiguracja
  - Negatywny wpływ ExpertSim na F-1
    - Algorytm MFIBlocks oparty na monotoniczności zbiorów
    - Właściwość tracona przy zastosowaniu niemonotonicznej miary podobieństwa zaproponowanej przez historyków-ekspertów
  - Autorzy algorytmu szacują, że 94% of Fałszywie Dodatnich dopasowań algorytmu to dopasowania Prawdziwie Dodatnie błędnie otagowane przez historyków-ekspertów z Yad Vashem

| Condition                   | Recall | Precision | F-1   |
|-----------------------------|--------|-----------|-------|
| Base                        | 0.770  | 0.172     | 0.279 |
| Expert Weighting            | 0.886  | 0.151     | 0.256 |
| ExpertSim                   | 0.872  | 0.118     | 0.205 |
| <i>SameSrc</i>              | 0.691  | 0.241     | 0.350 |
| <i>Cls</i>                  | 0.850  | 0.204     | 0.328 |
| <i>SameSrc</i> + <i>Cls</i> | 0.660  | 0.325     | 0.427 |

→ Zwiększenie Czułości

→ Zwiększenie Precyzji i F-1

Table 9: Quality under Varying Conditions

# Podsumowanie

- Podsumowanie

- Analiza prawdziwych danych historycznych
- *MFIBlocks* jest dobrze skalowalny pod względem rozmiaru i złożoności zbioru danych Yad Vashem
- Obsługa tagu *może* musi zostać zbadana z przeanalizowaniem 3 alternatyw
- Uncertain Entity Resolution może być wykorzystany do każdego problemu danych historycznych pochodzących z wielu źródeł
- Jak możemy wykorzystać bezpośrednią i pośrednią wiedzę o różnicowaniu źródeł rekordów?
- Czy możemy efektywnie przeprowadzić entity resolution na różnych poziomach granularności?

# **Badania**

## **Multi-Source Uncertain Entity Resolution at Yad Vashem Transforming Holocaust Victim Reports into People**

**Tomer Sagi, Omer Barkol, Ruth Bergman**  
Hewlett Packard Labs, Guttwirt Industrial Park, Haifa, Israel

**Avigdor Gal**  
Technion Israel Institute of Technology, Haifa, Israel

**Alexander Avram**  
Yad Vashem, Jerusalem, Israel

# **Przeszłość odczytana z bazy danych**

## **Niepewność odczytu**

**Dziękuję.**