



# Infrastrukturkonzept für die Integrierte Versorgung MPI-Implementierung

Markus Lamprecht, Sebastian Thiele und Anke Häber

Westsächsische Hochschule Zwickau

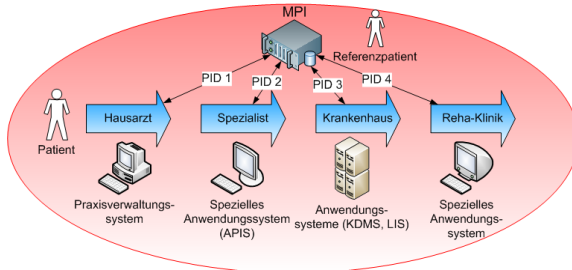
# Motivation

## Eindeutige Patientenidentifikation mit Hilfe eines MPI

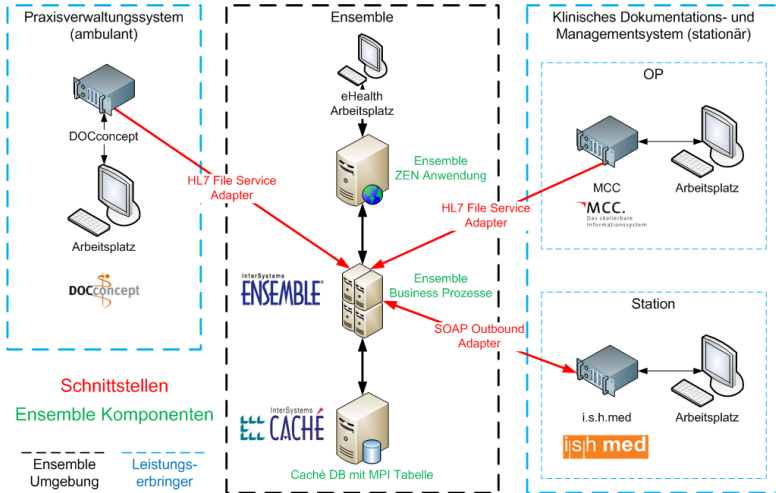
- Schnittstellenvielfalt bei Anwendungssystemen
- Einsatz eines Kommunikationsservers

## automatisierte Dublettenerkennung

- Suchalgorithmen innerhalb eines Kommunikationsservers
- Kombination verschiedener Ähnlichkeitsmaße




# MPI-Infrastruktur



# Dublettenerkennung

## Sorted Neighborhood Algorithmus [Alv97]

1. Schlüsselerzeugung 
2. Sortierung


| # | ID | City       | DateOfBirth | FirstName | Gender | GenerateKey   | LastName   | Street               | Zip   |
|---|----|------------|-------------|-----------|--------|---------------|------------|----------------------|-------|
| 1 | 1  | Oak Creek  | 59428       | Bob       | m      | Cun59428Bm593 | Cunningham | 7957 Oak Court       | 59358 |
| 2 | 2  | Albany     | 46567       | Ralph     | f      | Fro46567Rf499 | Frost      | 9441 Maple Avenue    | 49981 |
| 3 | 3  | Albany     | 46567       | Ralph     | m      | Fro46567Rm499 | Frost      | 9441 Maple Avenue    | 49981 |
| 4 | 4  | Gansevoort | 34776       | Debra     | f      | Huf34776Df636 | Huff       | 6146 Franklin Street | 63638 |
| 5 | 5  | Queensbury | 40382       | Ashley    | f      | Ihr40382Af603 | Ihringer   | 1250 Oak Drive       | 60393 |
| 6 | 6  | Reston     | 42309       | Martin    | f      | Isa42309Mf448 | Isaacs     | 1805 Madison Court   | 44831 |
| 7 | 7  | Reston     | 42309       | Marti     | m      | Isa42309Mm448 | Isaacs     | 1805 Madison Court   | 44831 |
| 8 | 8  | Ukia       | 36901       | Dan       | m      | Jen36901Dm411 | Jenkins    | 251 Second Court     | 41114 |
| 9 | 9  | Ukiah      | 36902       | Dan       | m      | Jen36902Dm411 | Jenkins    | 2515 Second Court    | 41114 |

### 3. Dublettenidentifikation

- paarweiser Attributvergleich mit verschiedenen Ähnlichkeitsmaßen innerhalb eines Sliding Window

# Dublettenerkennung

## Sorted Neighborhood Algorithmus [Alv97]

1. Schlüsselerzeugung 
2. Sortierung


| # | ID | City       | DateOfBirth | FirstName | Gender | GenerateKey   | LastName   | Street               | Zip   |
|---|----|------------|-------------|-----------|--------|---------------|------------|----------------------|-------|
| 1 | 1  | Oak Creek  | 59428       | Bob       | m      | Cun59428Bm593 | Cunningham | 7957 Oak Court       | 59358 |
| 2 | 2  | Albany     | 46567       | Ralph     | f      | Fro46567Rf499 | Frost      | 9441 Maple Avenue    | 49981 |
| 3 | 3  | Albany     | 46567       | Ralph     | m      | Fro46567Rm499 | Frost      | 9441 Maple Avenue    | 49981 |
| 4 | 4  | Gansevoort | 34776       | Debra     | f      | Huf34776Df636 | Huff       | 6146 Franklin Street | 63638 |
| 5 | 5  | Queensbury | 40382       | Ashley    | f      | Ihr40382Af603 | Ihringer   | 1250 Oak Drive       | 60393 |
| 6 | 6  | Reston     | 42309       | Martin    | f      | Isa42309Mf448 | Isaacs     | 1805 Madison Court   | 44831 |
| 7 | 7  | Reston     | 42309       | Marti     | m      | Isa42309Mm448 | Isaacs     | 1805 Madison Court   | 44831 |
| 8 | 8  | Ukia       | 36901       | Dan       | m      | Jen36901Dm411 | Jenkins    | 251 Second Court     | 41114 |
| 9 | 9  | Ukiah      | 36902       | Dan       | m      | Jen36902Dm411 | Jenkins    | 2515 Second Court    | 41114 |

### 3. Dublettenidentifikation

- paarweiser Attributvergleich mit verschiedenen Ähnlichkeitsmaßen innerhalb eines **Sliding Window**

# Dublettenerkennung

## Sorted Neighborhood Algorithmus [Alv97]

1. Schlüsselerzeugung 
2. Sortierung


| # | ID | City       | DateOfBirth | FirstName | Gender | GenerateKey   | LastName   | Street               | Zip   |
|---|----|------------|-------------|-----------|--------|---------------|------------|----------------------|-------|
| 1 | 1  | Oak Creek  | 59428       | Bob       | m      | Cun59428Bm593 | Cunningham | 7957 Oak Court       | 59358 |
| 2 | 2  | Albany     | 46567       | Ralph     | f      | Fro46567Rf499 | Frost      | 9441 Maple Avenue    | 49981 |
| 3 | 3  | Albany     | 46567       | Ralph     | m      | Fro46567Rm499 | Frost      | 9441 Maple Avenue    | 49981 |
| 4 | 4  | Gansevoort | 34776       | Debra     | f      | Huf34776Df636 | Huff       | 6146 Franklin Street | 63638 |
| 5 | 5  | Queensbury | 40382       | Ashley    | f      | Ihr40382Af603 | Ihringer   | 1250 Oak Drive       | 60393 |
| 6 | 6  | Reston     | 42309       | Martin    | f      | Isa42309Mf448 | Isaacs     | 1805 Madison Court   | 44831 |
| 7 | 7  | Reston     | 42309       | Marti     | m      | Isa42309Mm448 | Isaacs     | 1805 Madison Court   | 44831 |
| 8 | 8  | Ukia       | 36901       | Dan       | m      | Jen36901Dm411 | Jenkins    | 251 Second Court     | 41114 |
| 9 | 9  | Ukiah      | 36902       | Dan       | m      | Jen36902Dm411 | Jenkins    | 2515 Second Court    | 41114 |

### 3. Dublettenidentifikation

- paarweiser Attributvergleich mit verschiedenen Ähnlichkeitsmaßen innerhalb eines **Sliding Window**

# Dublettenerkennung

## Sorted Neighborhood Algorithmus [Alv97]

1. Schlüsselerzeugung 
2. Sortierung


| # | ID | City       | DateOfBirth | FirstName | Gender | GenerateKey   | LastName   | Street               | Zip   |
|---|----|------------|-------------|-----------|--------|---------------|------------|----------------------|-------|
| 1 | 1  | Oak Creek  | 59428       | Bob       | m      | Cun59428Bm593 | Cunningham | 7957 Oak Court       | 59358 |
| 2 | 2  | Albany     | 46567       | Ralph     | f      | Fro46567Rf499 | Frost      | 9441 Maple Avenue    | 49981 |
| 3 | 3  | Albany     | 46567       | Ralph     | m      | Fro46567Rm499 | Frost      | 9441 Maple Avenue    | 49981 |
| 4 | 4  | Gansevoort | 34776       | Debra     | f      | Huf34776Df636 | Huff       | 6146 Franklin Street | 63638 |
| 5 | 5  | Queensbury | 40382       | Ashley    | f      | Ihr40382Af603 | Ihringer   | 1250 Oak Drive       | 60393 |
| 6 | 6  | Reston     | 42309       | Martin    | f      | Isa42309Mf448 | Isaacs     | 1805 Madison Court   | 44831 |
| 7 | 7  | Reston     | 42309       | Marti     | m      | Isa42309Mm448 | Isaacs     | 1805 Madison Court   | 44831 |
| 8 | 8  | Ukia       | 36901       | Dan       | m      | Jen36901Dm411 | Jenkins    | 251 Second Court     | 41114 |
| 9 | 9  | Ukiah      | 36902       | Dan       | m      | Jen36902Dm411 | Jenkins    | 2515 Second Court    | 41114 |

### 3. Dublettenidentifikation

- paarweiser Attributvergleich mit verschiedenen Ähnlichkeitsmaßen innerhalb eines **Sliding Window**

# Dublettenerkennung

## Sorted Neighborhood Algorithmus [Alv97]

1. Schlüsselerzeugung 
2. Sortierung

| # | ID | City       | DateOfBirth | FirstName | Gender | GenerateKey   | LastName   | Street               | Zip   |
|---|----|------------|-------------|-----------|--------|---------------|------------|----------------------|-------|
| 1 | 1  | Oak Creek  | 59428       | Bob       | m      | Cun59428Bm593 | Cunningham | 7957 Oak Court       | 59358 |
| 2 | 2  | Albany     | 46567       | Ralph     | f      | Fro46567Rf499 | Frost      | 9441 Maple Avenue    | 49981 |
| 3 | 3  | Albany     | 46567       | Ralph     | m      | Fro46567Rm499 | Frost      | 9441 Maple Avenue    | 49981 |
| 4 | 4  | Gansevoort | 34776       | Debra     | f      | Huf34776Df636 | Huff       | 6146 Franklin Street | 63638 |
| 5 | 5  | Queensbury | 40382       | Ashley    | f      | Ihr40382Af603 | Ihringer   | 1250 Oak Drive       | 60393 |
| 6 | 6  | Reston     | 42309       | Martin    | f      | Isa42309Mf448 | Isaacs     | 1805 Madison Court   | 44831 |
| 7 | 7  | Reston     | 42309       | Marti     | m      | Isa42309Mm448 | Isaacs     | 1805 Madison Court   | 44831 |
| 8 | 8  | Ukia       | 36901       | Dan       | m      | Jen36901Dm411 | Jenkins    | 251 Second Court     | 41114 |
| 9 | 9  | Ukiah      | 36902       | Dan       | m      | Jen36902Dm411 | Jenkins    | 2515 Second Court    | 41114 |

3. Dublettenidentifikation

- paarweiser Attributvergleich mit verschiedenen Ähnlichkeitsmaßen innerhalb eines **Sliding Window**

## Dublettenerkennung II

### Ähnlichkeitsmaße

- Damerau-Levenshtein Algorithmus [Fre64]
- Jaro-Winkler Distanz [Mat95]
- Boolesche und Stringlängenvergleiche

```
Vergleiche Isaacs mit Isaacs
Vergleiche Martin mit Marti
Vergleiche f mit m
Vergleiche 1805 Madison Court mit 1805 Madison Court
Vergleiche 42309 mit 42309
Distanzliste: 0|1|0|1|.96664|1|0|1|0|0|1|0|1|0|0
Ermittelte Gewichte:
0.3|0.2|1|0.96664|1|0.05|0|0.05|0.1|0.05|0.1|0|0.05|0.05
Endergebnis: +4.22 von 4.50 Punkten
Matchresult: +93.70 %
```

## Ergebnisse

- Zuverlässige Dublettenerkennung beim Vergleich von Testdaten
- Vergleich von Datensätzen mit Sliding Window = 2

| # | ID | City       | DateOfBirth | FirstName | Gender | GenerateKey   | LastName   | Street               | Zip   |
|---|----|------------|-------------|-----------|--------|---------------|------------|----------------------|-------|
| 1 | 1  | Oak Creek  | 59428       | Bob       | m      | Cun59428Bm593 | Cunningham | 7957 Oak Court       | 59358 |
| 2 | 2  | Albany     | 46567       | Ralph     | f      | Fro46567Rf499 | Frost      | 9441 Maple Avenue    | 49981 |
| 3 | 3  | Albany     | 46567       | Ralph     | m      | Fro46567Rm499 | Frost      | 9441 Maple Avenue    | 49981 |
| 4 | 4  | Gansevoort | 34776       | Debra     | f      | Huf34776Df636 | Huff       | 6146 Franklin Street | 63638 |
| 5 | 5  | Queensbury | 40382       | Ashley    | f      | Ihr40382Af603 | Ihringer   | 1250 Oak Drive       | 60393 |
| 6 | 6  | Reston     | 42309       | Martin    | f      | Isa42309Mf448 | Isaacs     | 1805 Madison Court   | 44831 |
| 7 | 7  | Reston     | 42309       | Marti     | m      | Isa42309Mm448 | Isaacs     | 1805 Madison Court   | 44831 |
| 8 | 8  | Ukia       | 36901       | Dan       | m      | Jen36901Dm411 | Jenkins    | 251 Second Court     | 41114 |
| 9 | 9  | Ukiah      | 36902       | Dan       | m      | Jen36902Dm411 | Jenkins    | 2515 Second Court    | 41114 |

- Schranke ergibt ab 85% Dublette (mehr Fehlerklassen geplant)



## Vielen Dank für Ihre Aufmerksamkeit!

Westsächsische Hochschule Zwickau

Dr.-Friedrichs-Ring 2A

D-08056 Zwickau

Telefon: +49 375 536 0

[markus.lamprecht@fh-zwickau.de](mailto:markus.lamprecht@fh-zwickau.de)

[sebastian.thiele@fh-zwickau.de](mailto:sebastian.thiele@fh-zwickau.de)

[anke.haeber@fh-zwickau.de](mailto:anke.haeber@fh-zwickau.de)

## Literatur



ALVARO MONGE UND CHARLES ELKAN: *An Efficient Domain-Independent Algorithm for Detecting Approximately Duplicate Database Records.*

Proceedings of the Workshop on Research Issues on Data Mining and Knowledge Discovery, 1997.



FRED DAMERAU: *A technique for computer detection and correction of spelling errors.*

Communications of the ACM.7 (Nr. 3), 1964.



MATTHEW A. JARO: *Probabilistic linkage of large public health data file.*

Statistics in Medicine 14 (5-7) 491-8, 1995.

<http://www.census.gov/srd/papers/pdf/rrs2006-02.pdf>.