LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

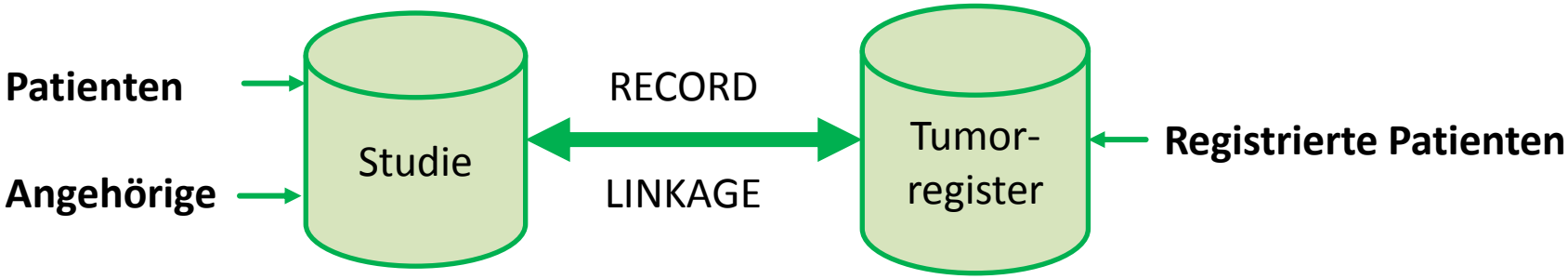
Einfluss von Varianten des anonymen Record Linkage auf Gewichtsverteilung und Klassifikation

Daniel Nasseh
Jürgen Stausberg

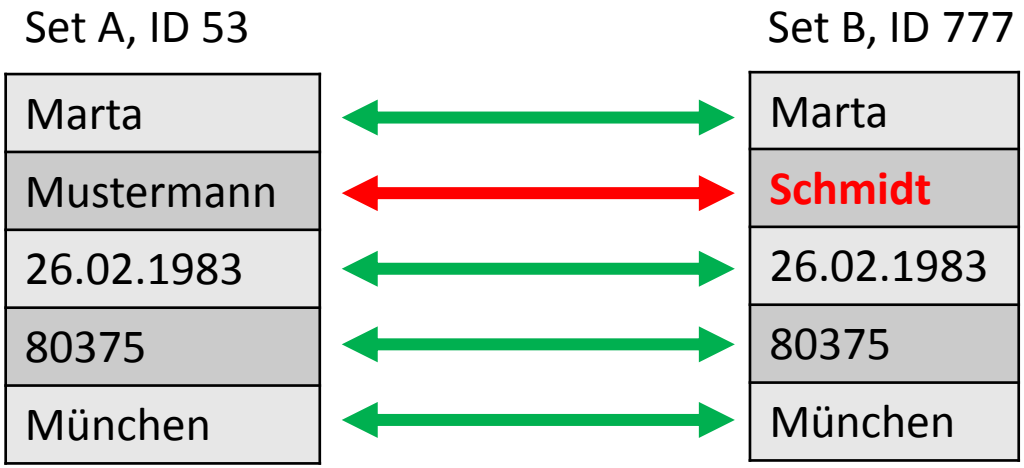
Institut für medizinische **I**nformationsverarbeitung, **B**iometrie und **E**pidemiologie

Kontext: Studie zu familiärem Darmkrebs

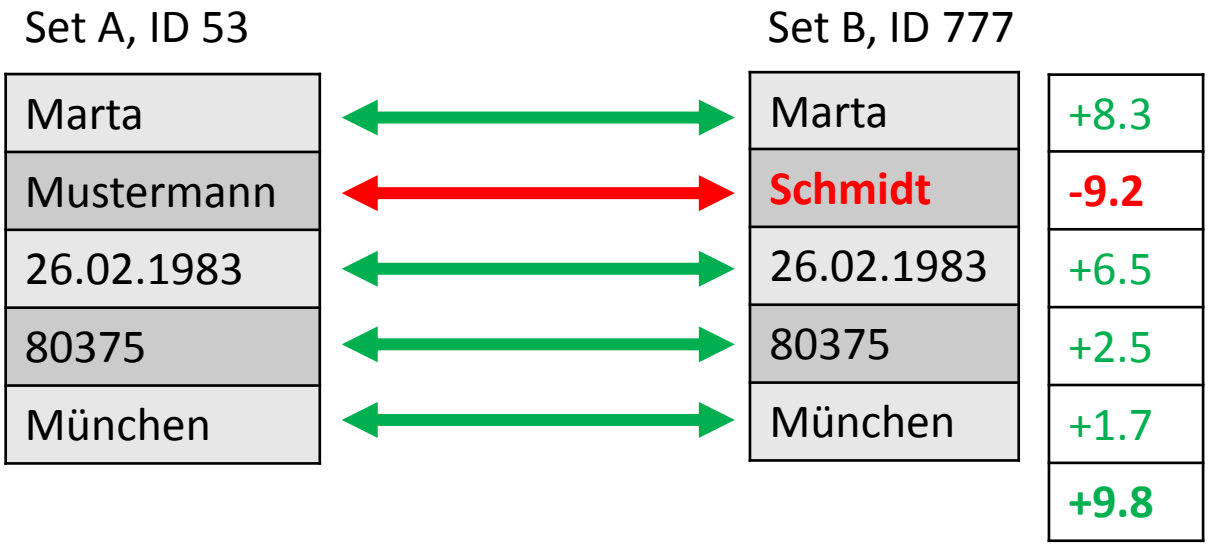
http://www.darmkrebs-familienstudie.de/



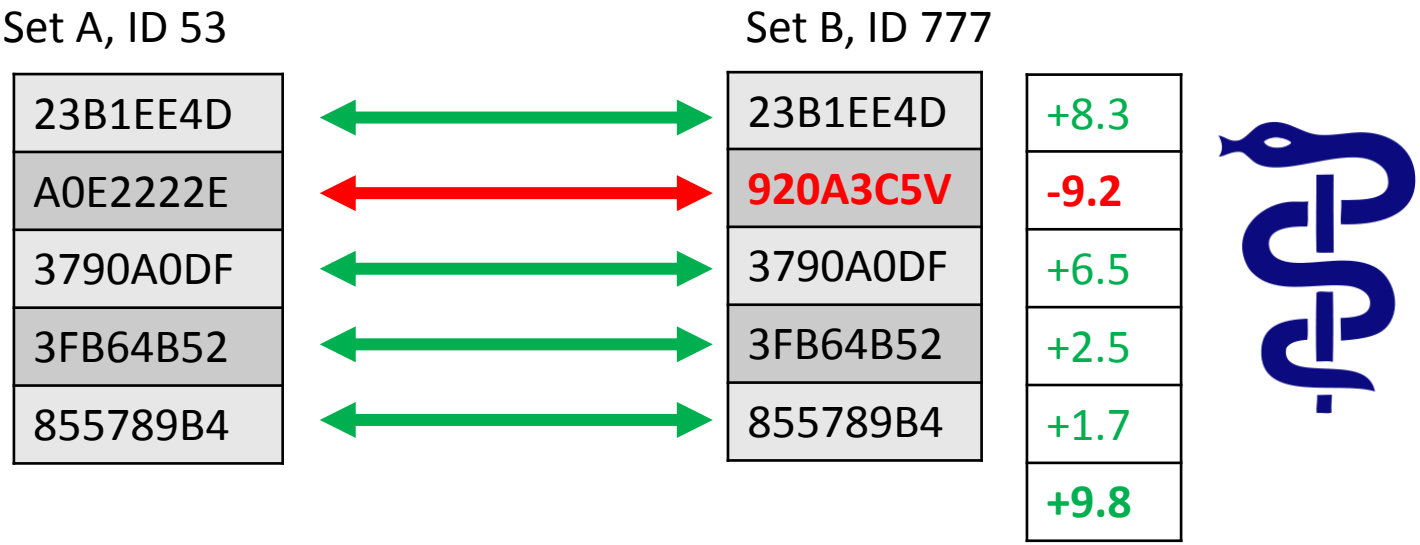
Probabilistisches Record Linkage



Probabilistisches Record Linkage



„Anonymes“ Record Linkage



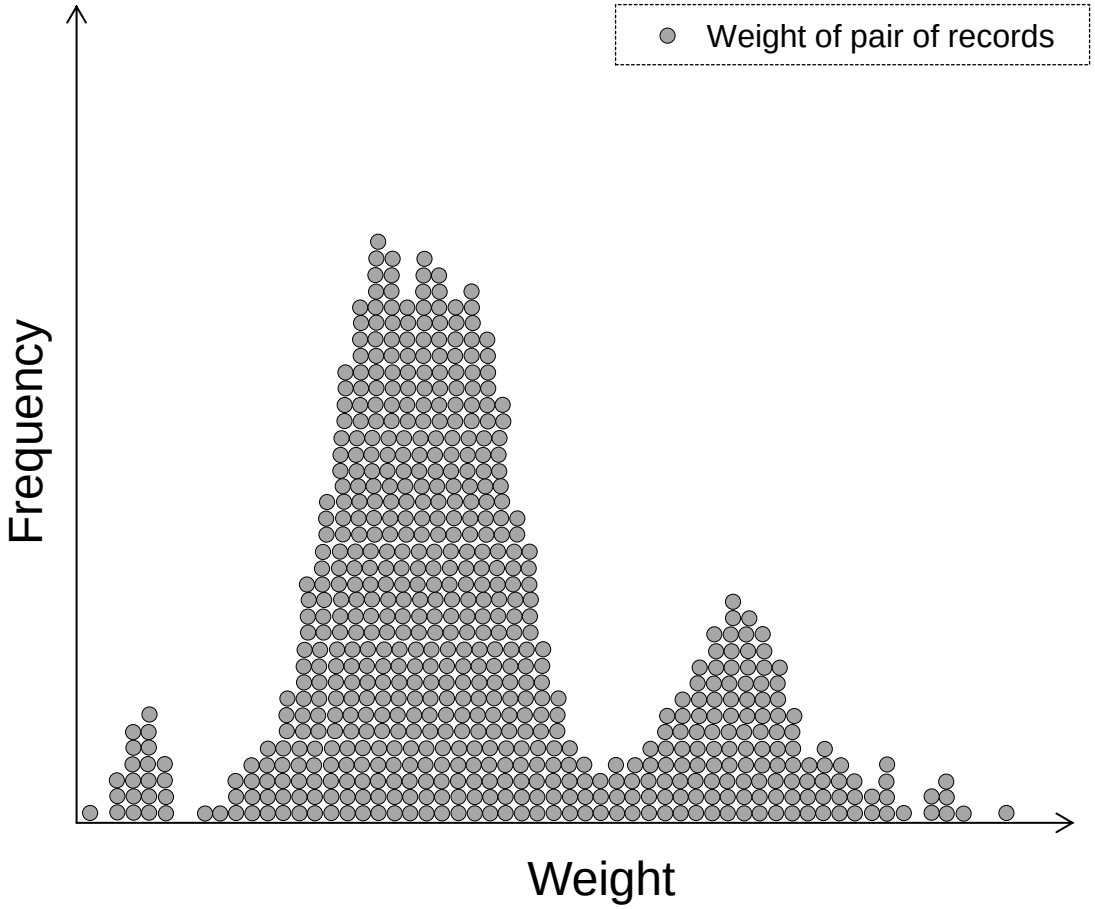
Gewichtsdaten

ID A	ID B	Gewicht
3452	45	54,33
2	294	34,22
36	5833	22,19
277	921	17,33
67	513	8,95
794	23	-5,33
…	…	…

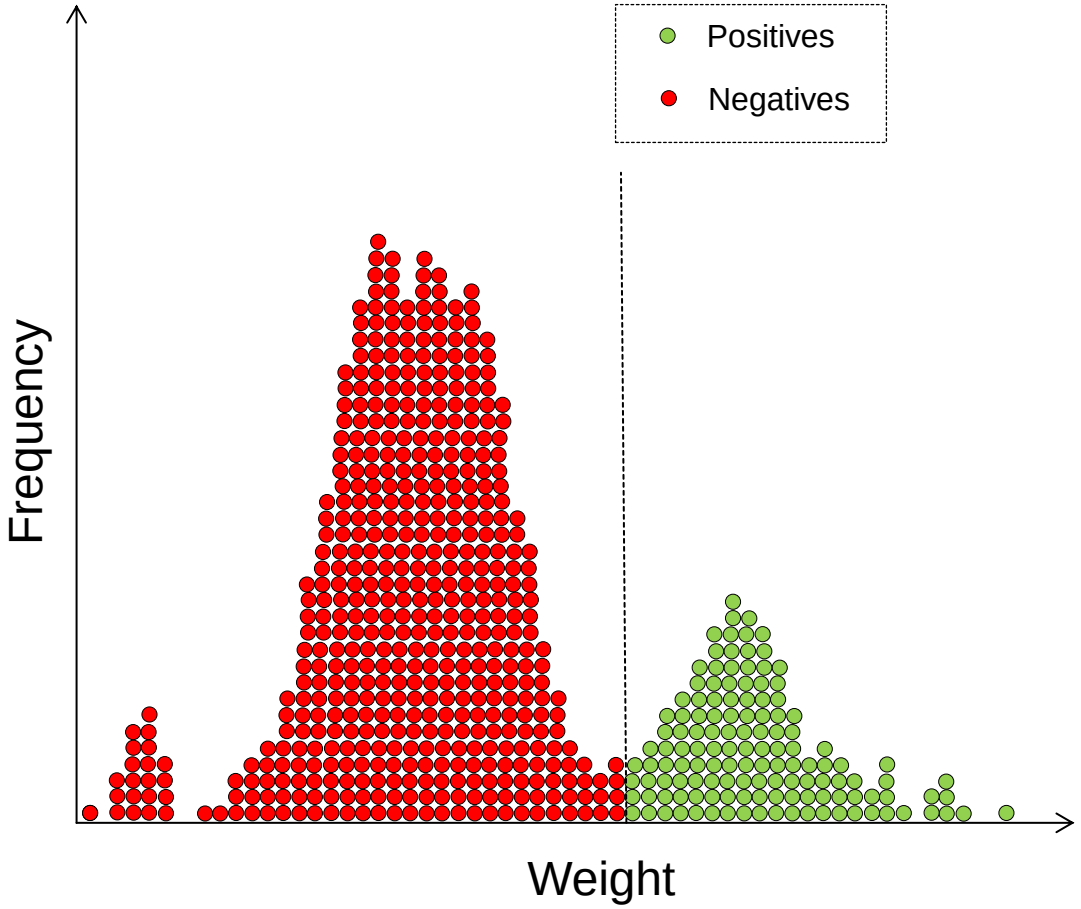
Gewichtsdaten

ID A	ID B	Gewicht
3452	45	54,33
2	294	34,22
36	5833	22,19
277	921	17,33
67	513	8,95
794	23	-5,33
…	…	…

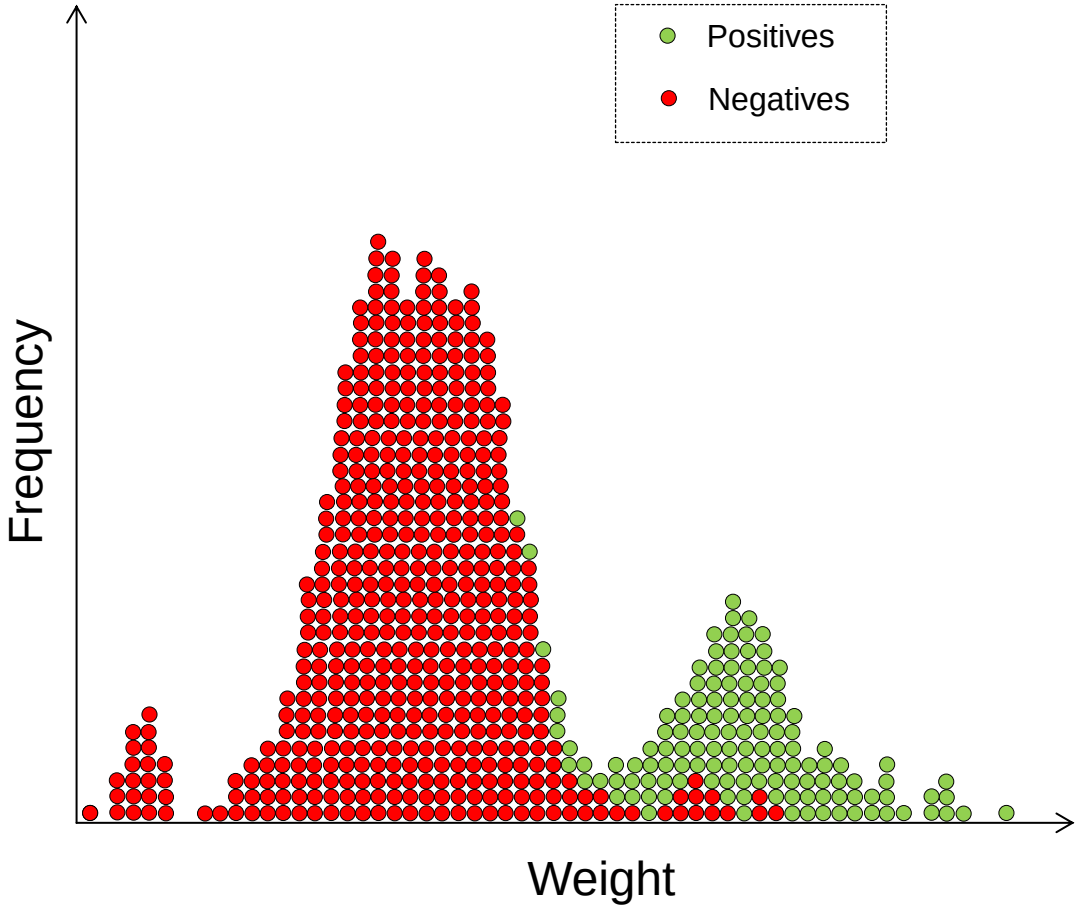
Histogramm der Gewichte paarweiser Vergleiche



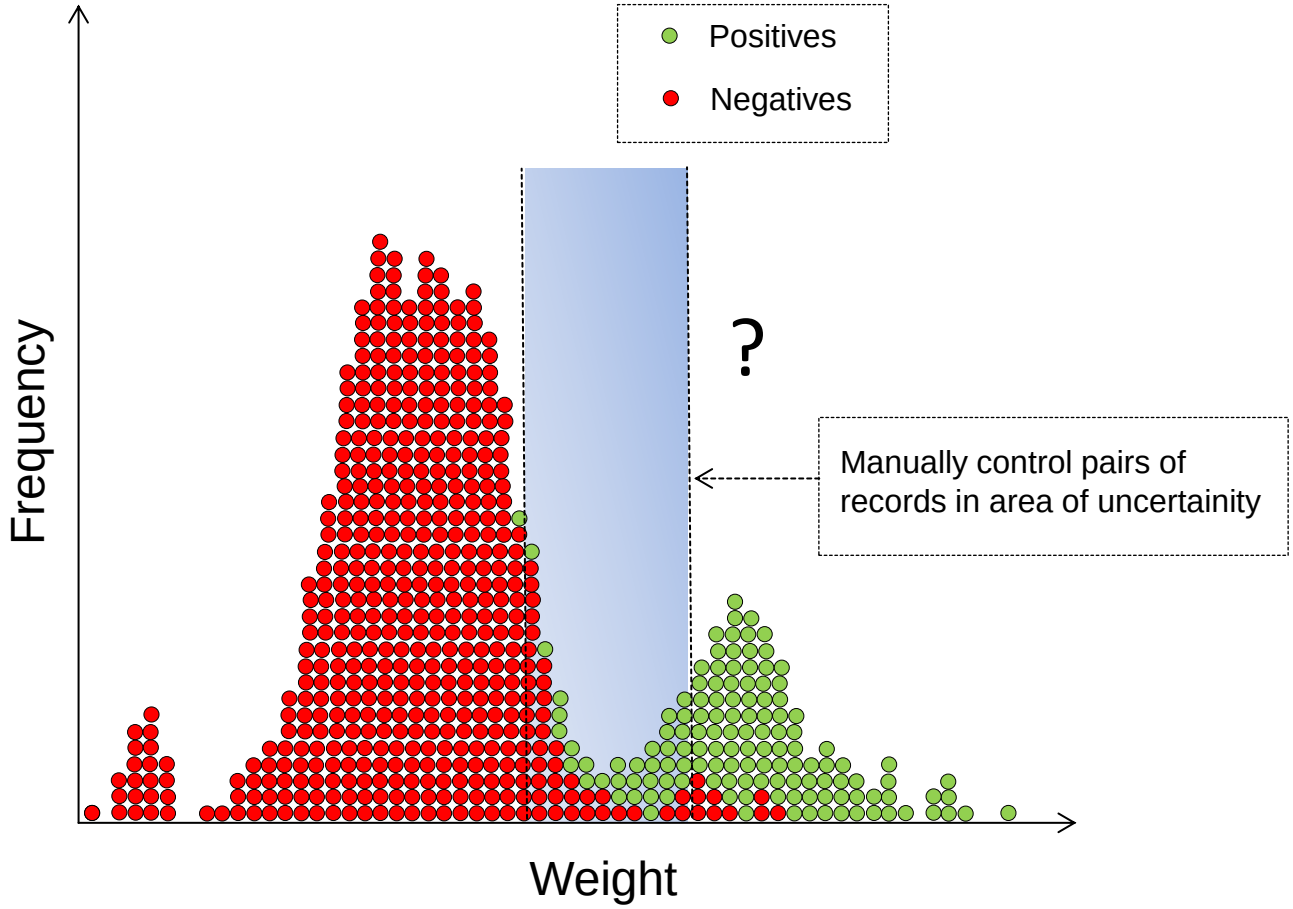
Histogramm der Gewichte paarweiser Vergleiche



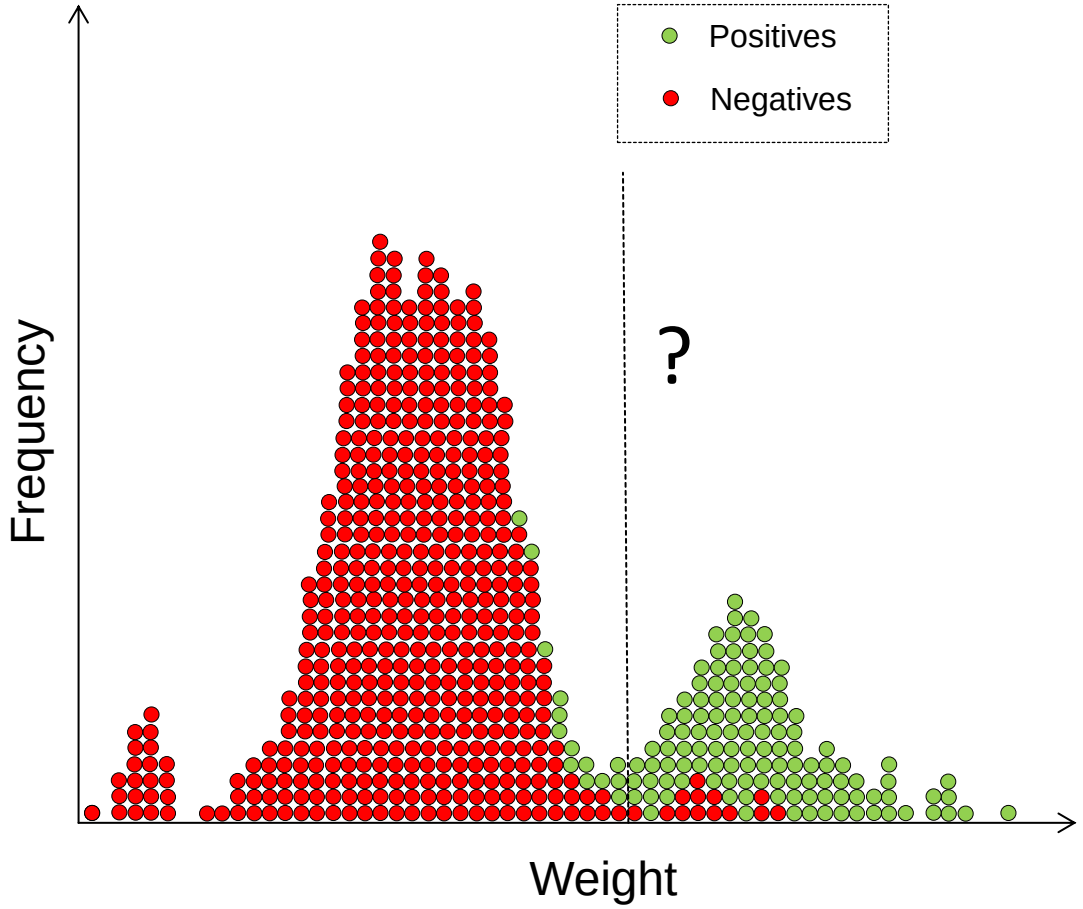
Histogramm der Gewichte paarweiser Vergleiche



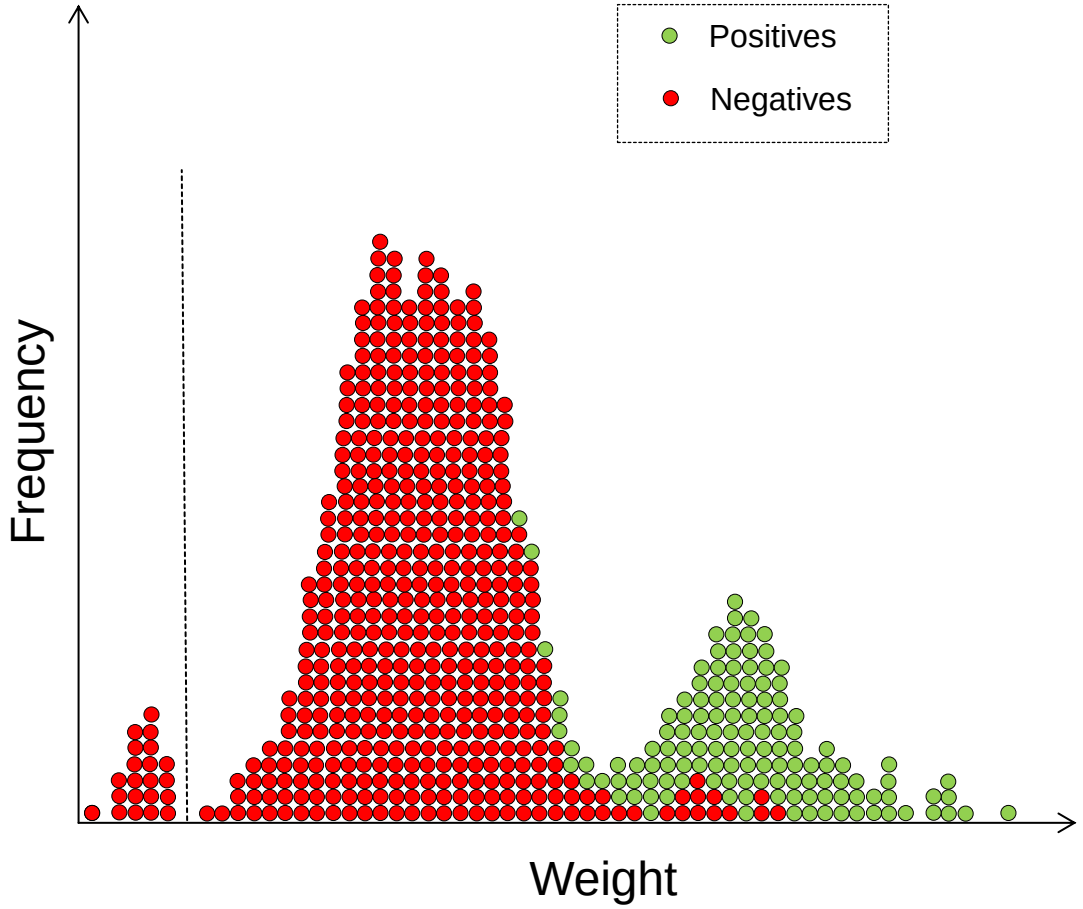
Histogramm der Gewichte paarweiser Vergleiche



Histogramm der Gewichte paarweiser Vergleiche



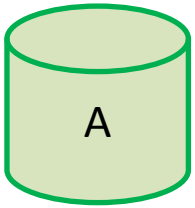
Histogramm der Gewichte paarweiser Vergleiche



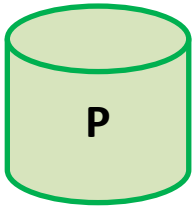
Experimentelles Setup

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Künstliches Datenset



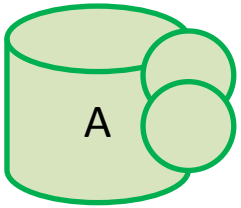
Öffentliches Datenset [*]



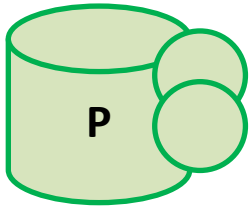
* http://secondstring.cvs.sourceforge.net/viewvc/secondstring/secondstring/.

Experimentelles Setup

Künstliches Datenset



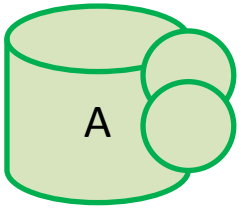
Öffentliches Datenset



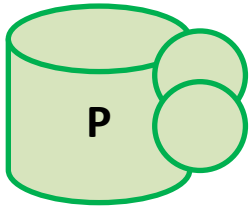
Experimentelles Setup

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Künstliches Datenset



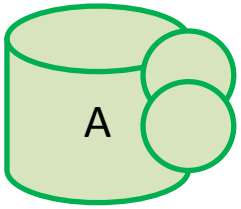
Öffentliches Datenset



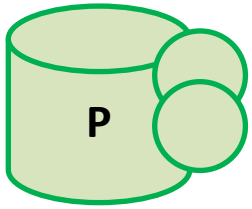
Record Linkage

Experimentelles Setup

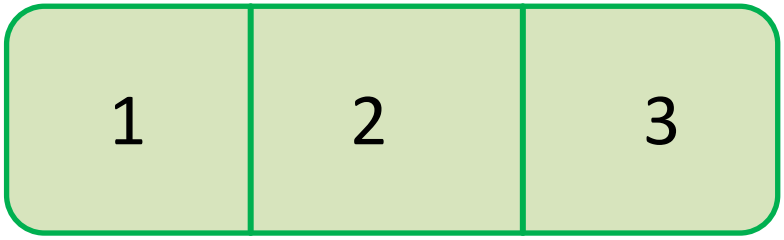
Künstliches Datenset



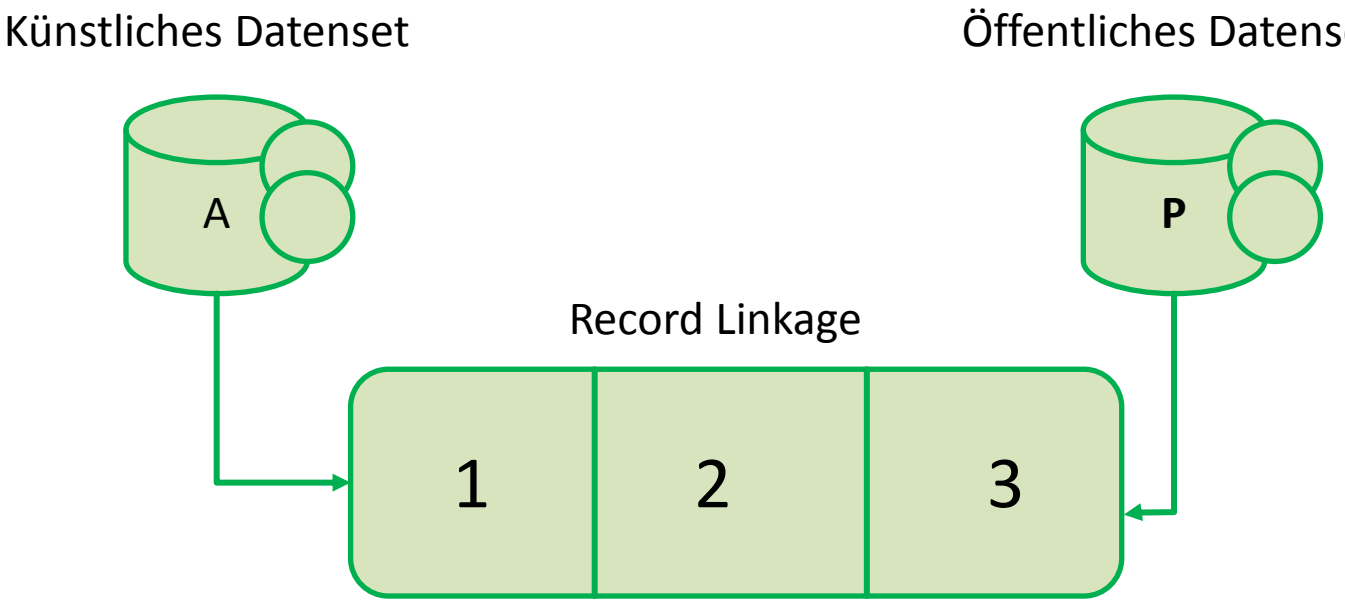
Öffentliches Datenset



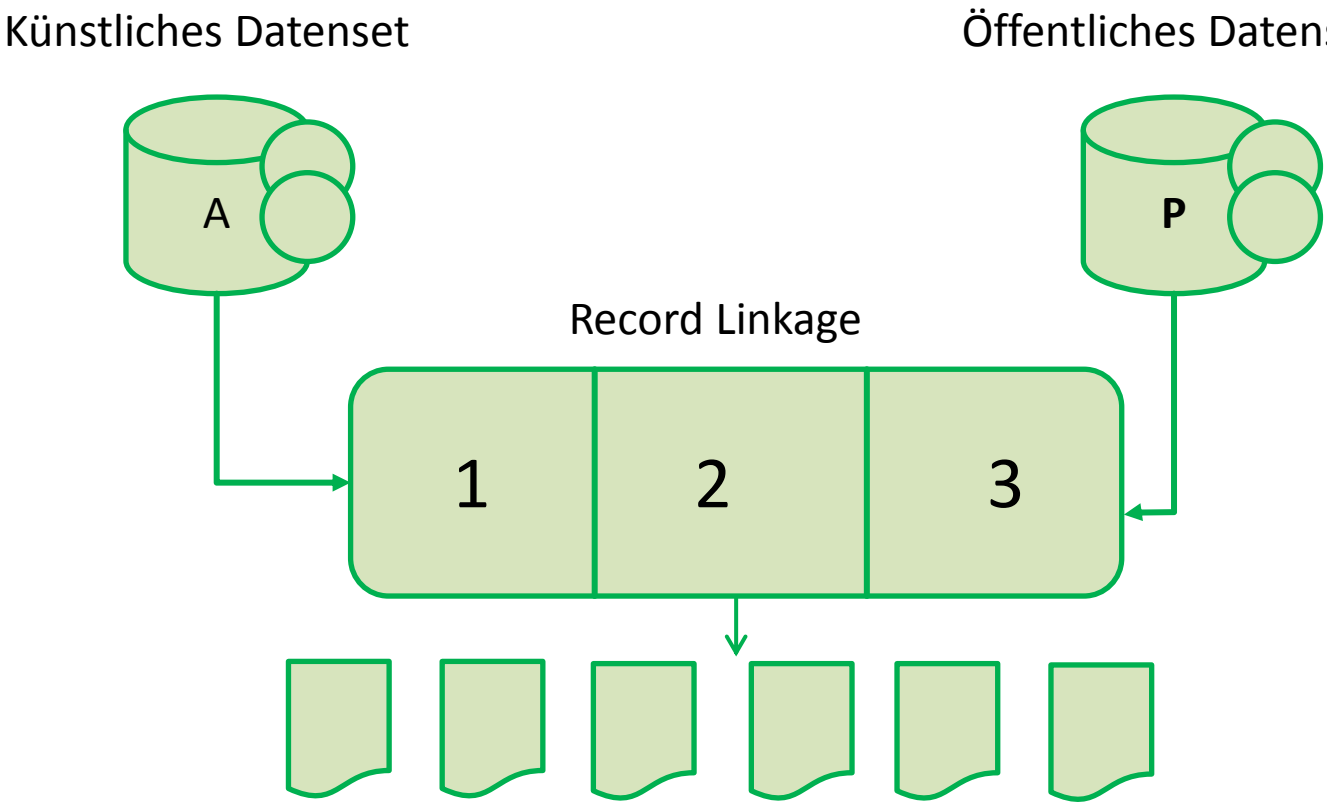
Record Linkage



Experimentelles Setup



Experimentelles Setup



Konfigurationen

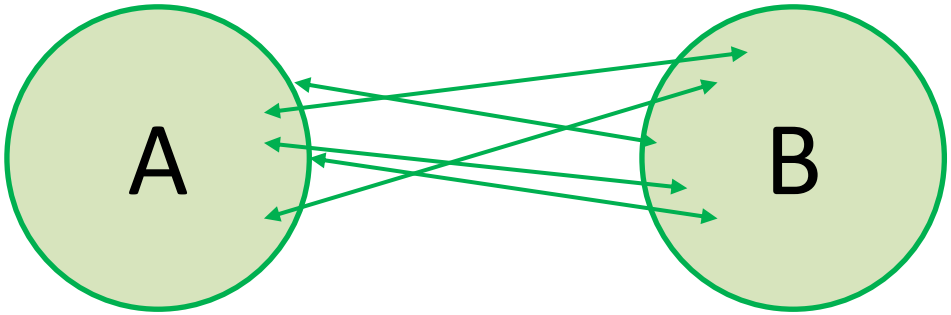
1. Blocking (naiv / mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning

Konfigurationen

1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning

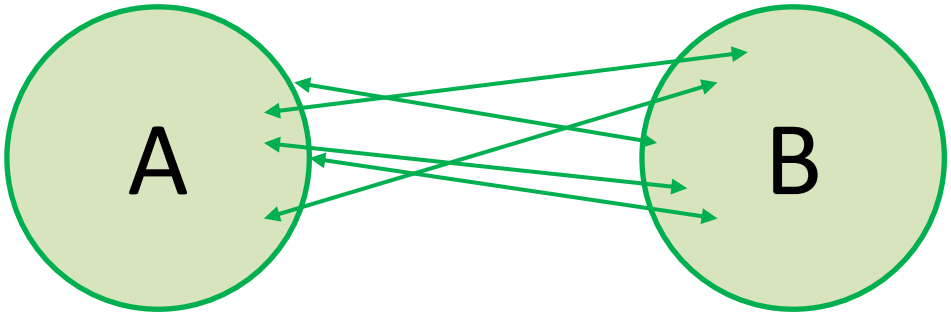
Konfigurationen

1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning



Konfigurationen

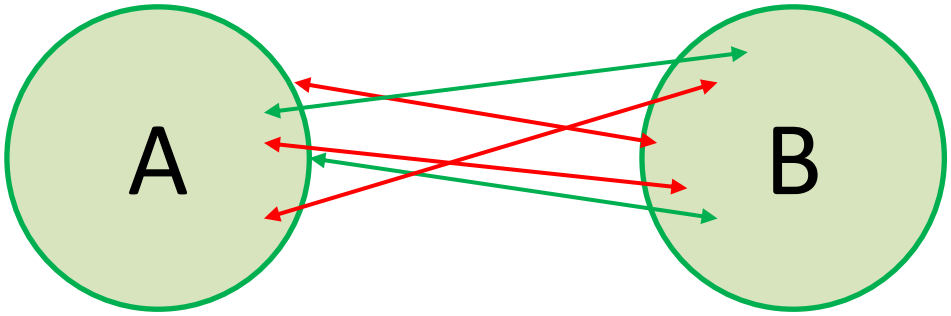
1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning



Bei 10.000*10.000 = 100.000.000 Berechnungen

Konfigurationen

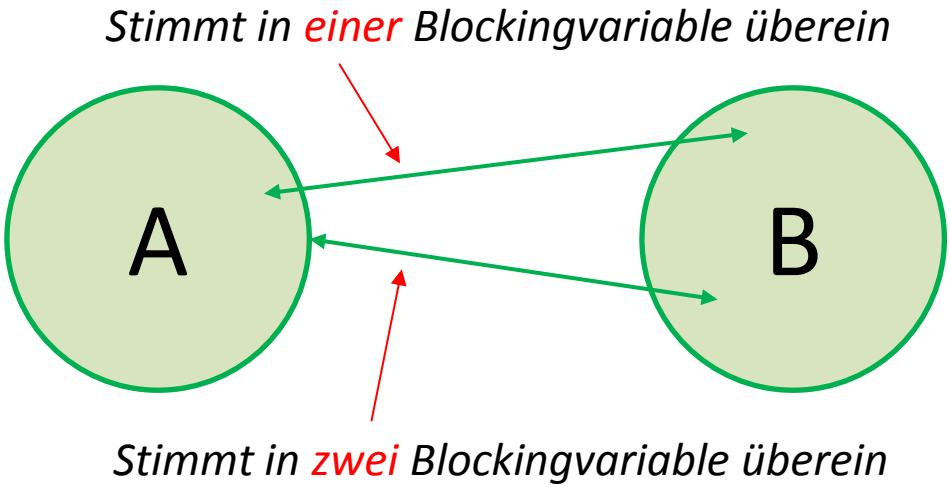
1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning



Record Linkage nur noch auf Vergleichen
die in Blocking-Variablen übereinstimmen!

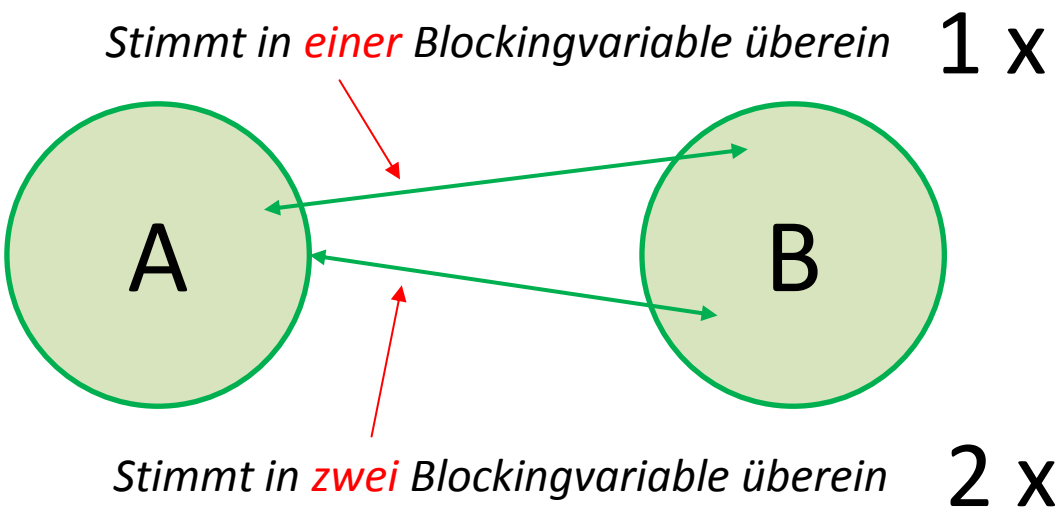
Konfigurationen

1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning



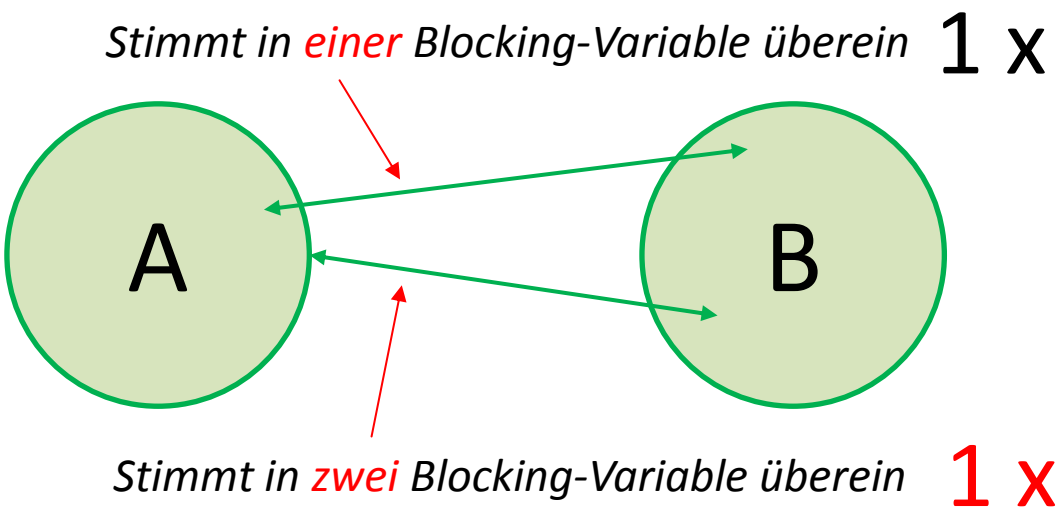
Konfigurationen

1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning



Konfigurationen

1. Blocking (mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning



Konfigurationen

1. Blocking (naiv / mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning

Konfigurationen

1. Blocking (naiv / mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning

ID A	ID B	Gewicht
55	21	+48.8
27	33	+21.4
55	84	+19.3

Konfigurationen

1. Blocking (naiv / mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning

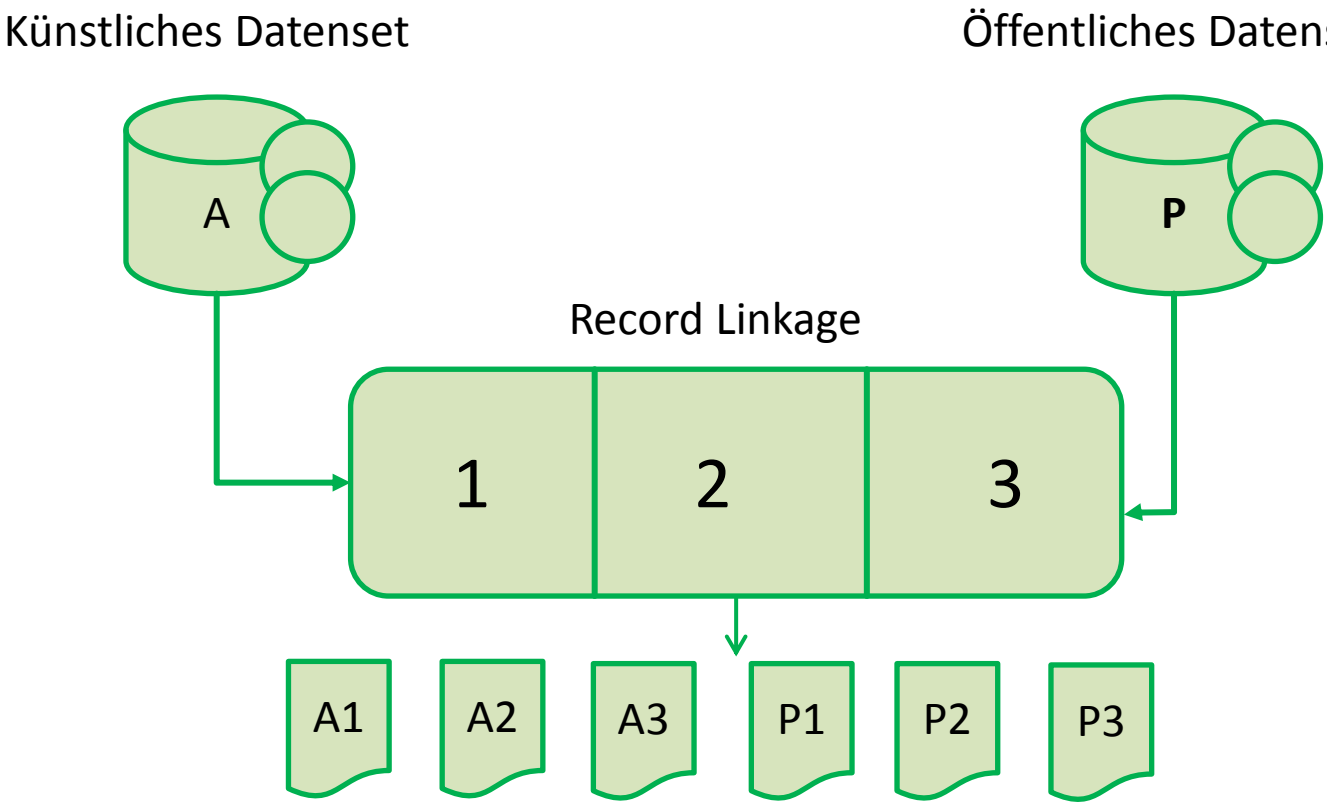
ID A	ID B	Gewicht
55	21	+48.8
27	33	+21.4
55	84	+19.3

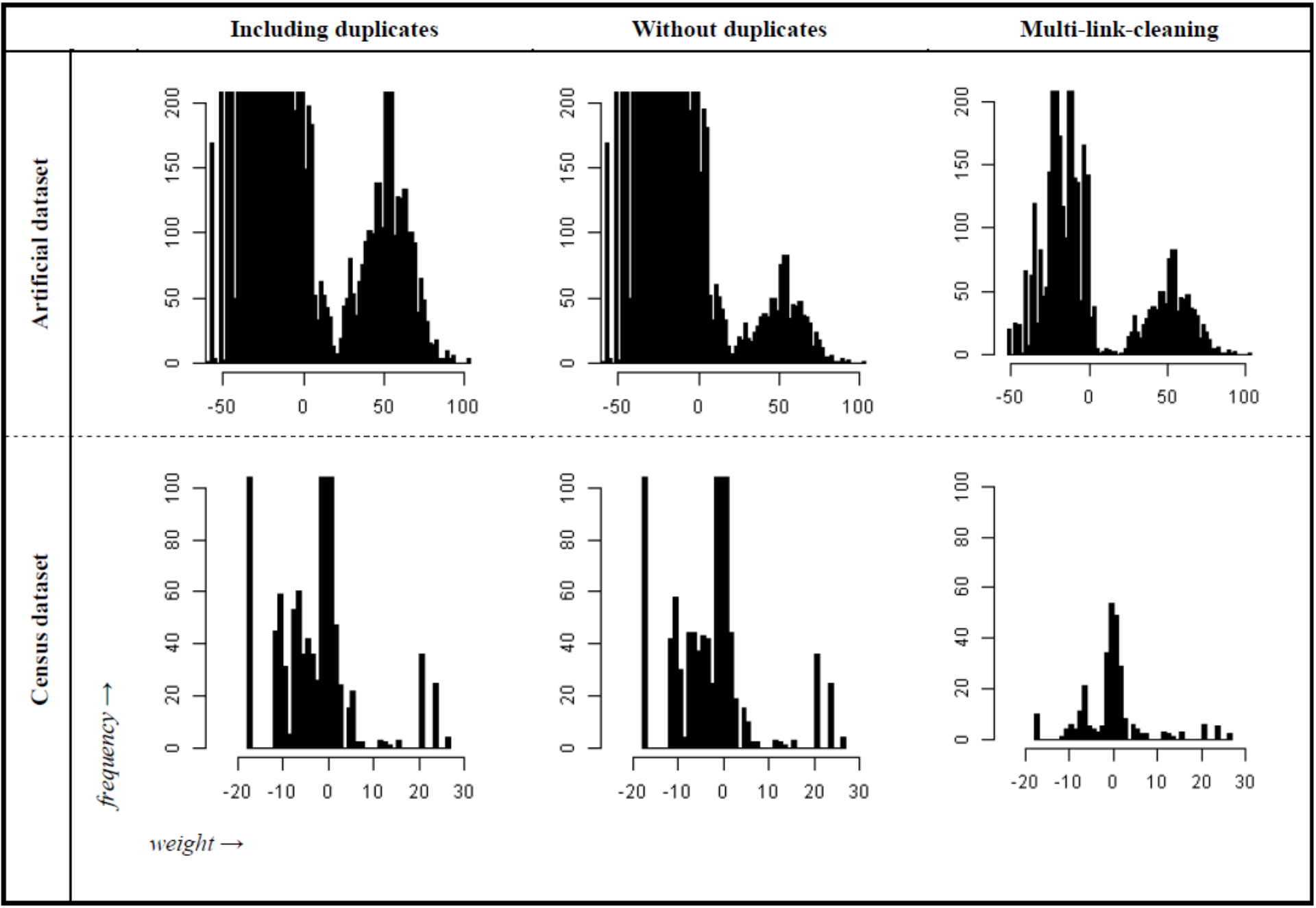
Konfigurationen

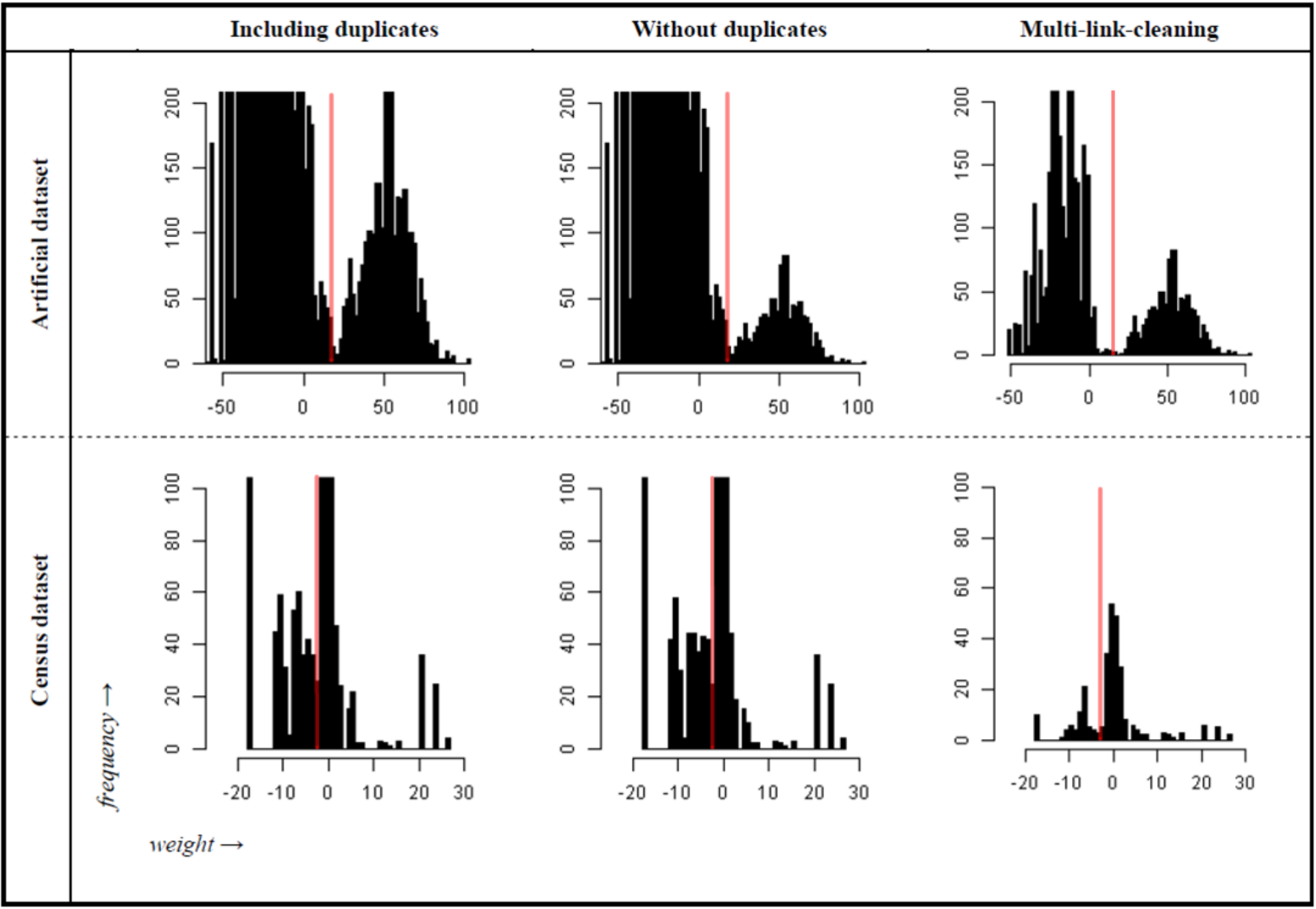
1. Blocking (naiv / mit mehrfach auftretenden Links)
2. Blocking (mit „unique“ Links)
3. Multi-Link-Cleaning

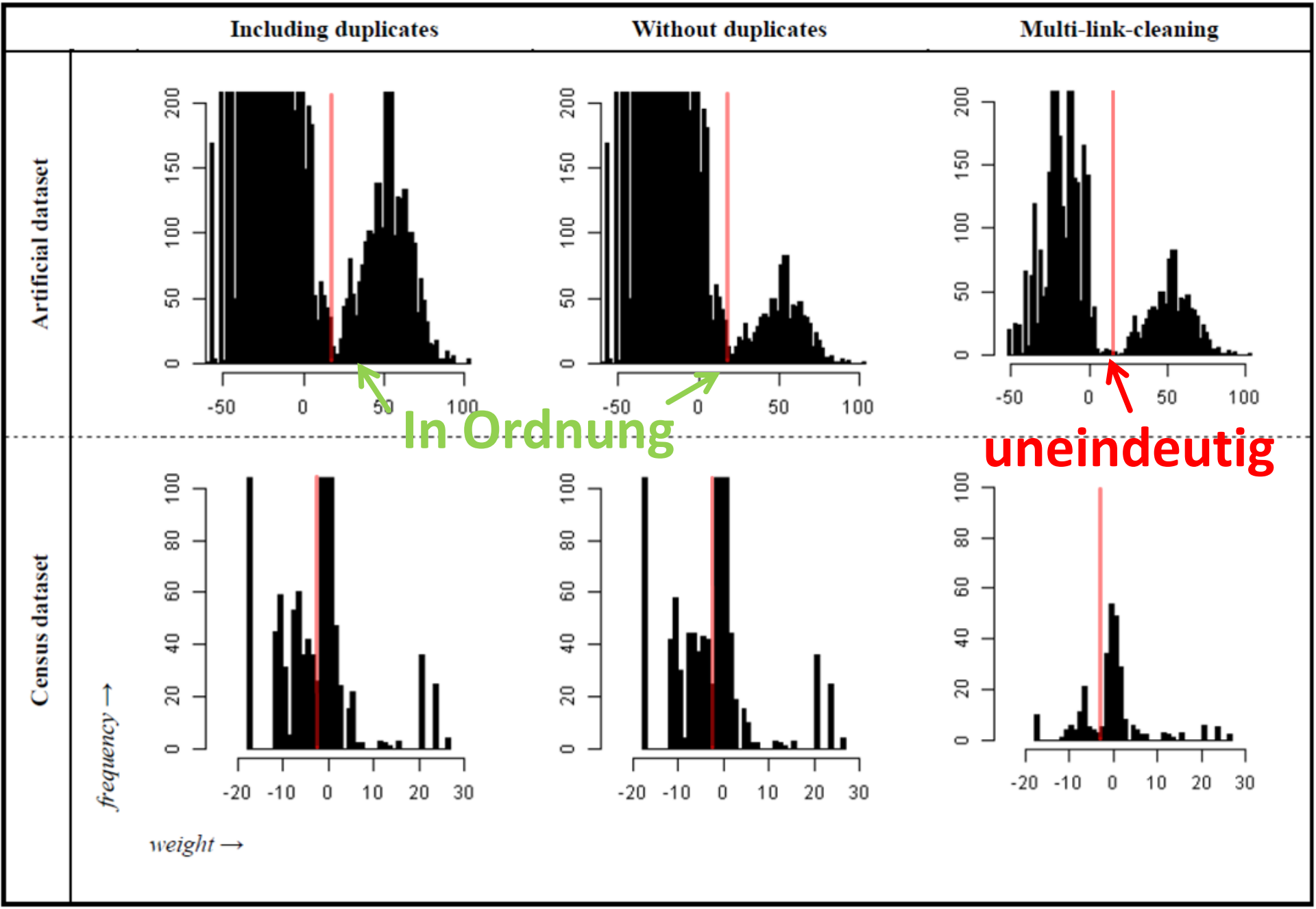
ID A	ID B	Gewicht
55	21	+48.8
27	33	+21.4
~~55~~	~~84~~	~~+19.3~~

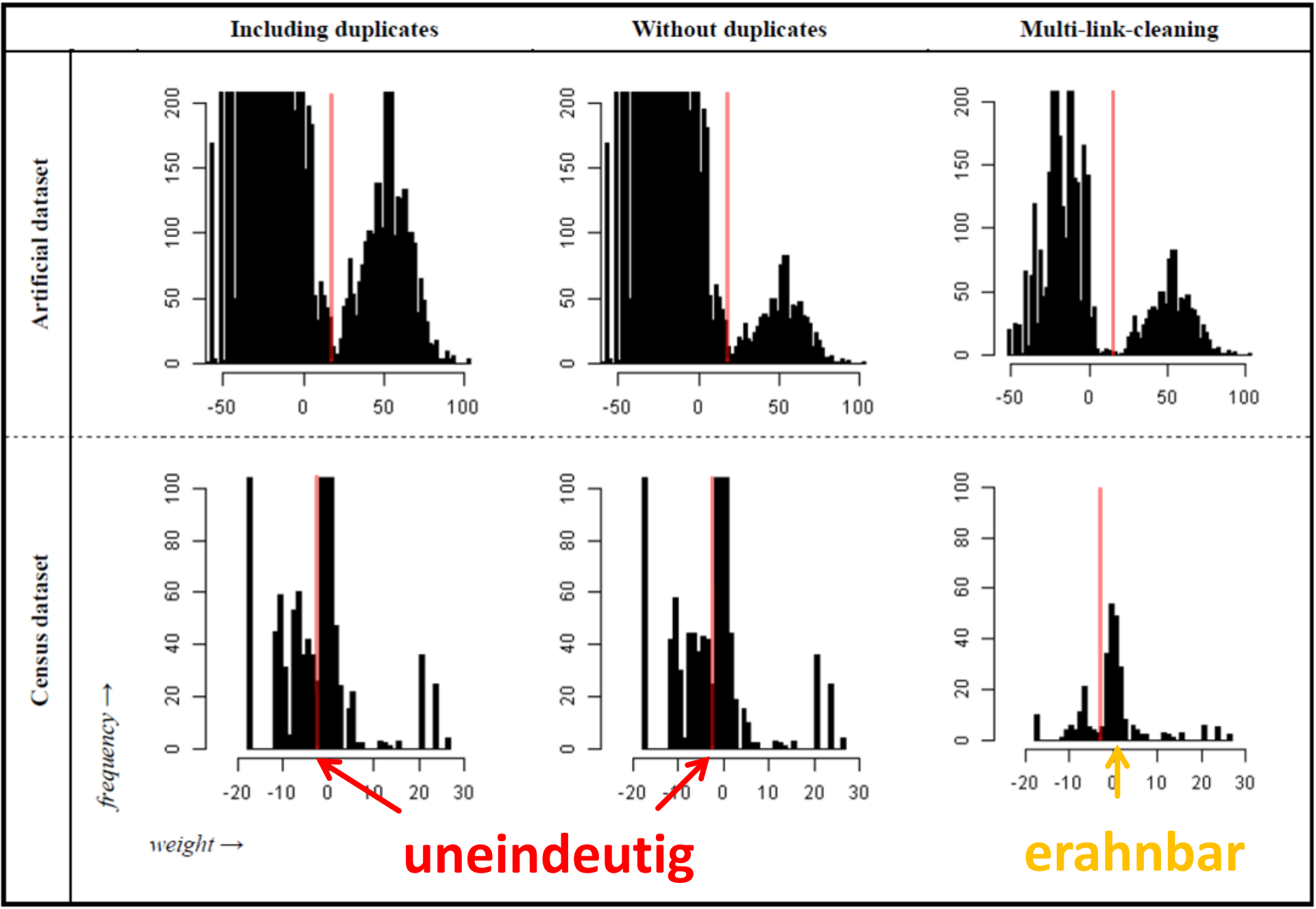
Experimentelles Setup




Including duplicates
Without duplicates
Multi-link-cleaning
Artificial dataset
Census dataset
frequency →
weight →
0
50
100
150
200
-50
0
50
100
0
20
40
60
80
100
-20
-10
0
10
20
30


Including duplicates
Without duplicates
Multi-link-cleaning
Artificial dataset
Census dataset
frequency →
weight →


Including duplicates
Without duplicates
Multi-link-cleaning
Artificial dataset
Census dataset
In Ordnung
uneindeutig
frequency →
weight →


Including duplicates
Without duplicates
Multi-link-cleaning
Artificial dataset
Census dataset
frequency →
weight →
uneindeutig
erahnbar

Haupterkenntnisse

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Bei Verwendung von unüberwachten Klassifikationssystemen ist die Gefahr einer kompletten Fehlklassifikation vorhanden
- Klassifikation ist nicht automatisierbar
- Es kann hilfreich sein verschiedene Konfigurationen zu verwenden um die korrekte Schranke zu ermitteln
- Bei schmutzigen oder kleinen Datensets bieten sich eher überwachte Klassifikationssysteme an (Boosting, Bagging).

Haupterkenntnisse

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Bei Verwendung von unüberwachten Klassifikationssystemen ist die Gefahr einer kompletten Fehlklassifikation vorhanden
- Klassifikation ist nicht automatisierbar
- Es kann hilfreich sein verschiedene Konfigurationen zu verwenden um die korrekte Schranke zu ermitteln
- Bei schmutzigen oder kleinen Datensets bieten sich eher überwachte Klassifikationssysteme an (Boosting, Bagging).

Haupterkenntnisse

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Bei Verwendung von unüberwachten Klassifikationssystemen ist die Gefahr einer kompletten Fehlklassifikation vorhanden
- Klassifikation ist nicht automatisierbar
- Es kann hilfreich sein verschiedene Konfigurationen zu verwenden um die korrekte Schranke zu ermitteln
- Bei schmutzigen oder kleinen Datensets bieten sich eher überwachte Klassifikationssysteme an (Boosting, Bagging).

Haupterkenntnisse

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

- Bei Verwendung von unüberwachten Klassifikationssystemen ist die Gefahr einer kompletten Fehlklassifikation vorhanden
- Klassifikation ist nicht automatisierbar
- Es kann hilfreich sein verschiedene Konfigurationen zu verwenden um die korrekte Schranke zu ermitteln
- Bei schmutzigen oder kleinen Datensets bieten sich eher überwachte Klassifikationssysteme an (Boosting, Bagging). *=> Basieren auf Trainingssets*

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Vielen Dank für Ihre Aufmerksamkeit!

Fragen?