Data Linkage: A Primer

AAPOR Webinar
November 17th, 2022

Joe Sakshaug
Stefan Bender

Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

Content of this session

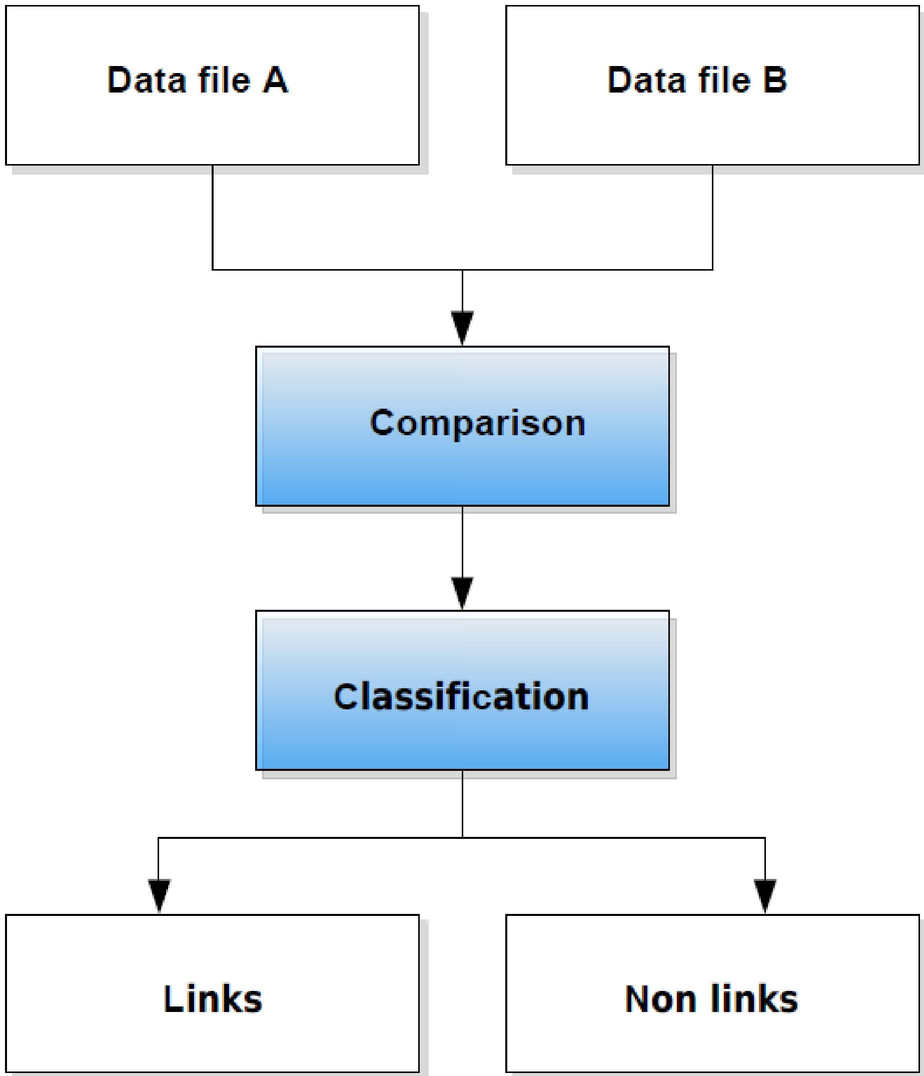
- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

Definition of Record Linkage

- RL is finding records in different data sets that represent the same entity and link them.

- RL is also known as *data matching, entity resolution, object identification, duplicate detection, identity uncertainty, merge-purge.*

The basic record linkage process



5 Main Applications of Record Linkage

1. Merging of two data files
2. Identifying the intersection of the two data sets
3. Updating of data files (with the data row of the other data files)
4. Impute missing data
5. Deduplicate a file

Merging of two data files

- Merging of data files for microanalyses (e.g. survey- or registry data)
- Follow - up of cohorts (e.g. linkage with Cancer registry)
- Retrospective construction of panels
- Merging of panel waves
- Validation of answers in surveys: Comparing individual provided information's with registry data.
- Bias – detection in surveys: Supply data for nonrespondents.
- Supply external data for imputation or weighting of survey data
- Adding contact information to survey samples.

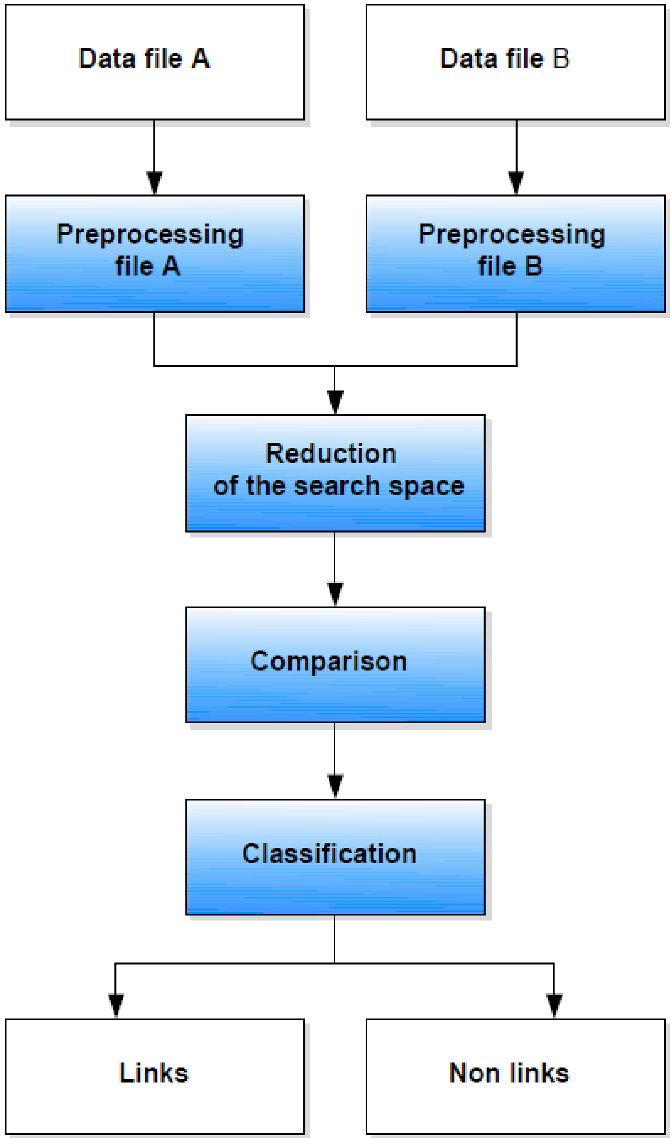
Identifying the intersection of the two data files

- Discovery of undercoverage within a census.
- Discovery of overcoverage and undercoverage in sampling frames.
- Examination of the reidentification risk of micro data files.
- Discovery of underreporting in registries (e.g. linkage with mortality registry).
- Dropping of duplicates as part of data cleansing.

Record Linkage Challenges (Christen 2012)

- Major challenge is that (clean) unique entity identifiers are not available in the databases to be linked.
  - Real world data are dirty (typographical errors and variations, missing and out-of-date values, different coding schemes, etc.)
- Data may require significant amounts of processing and data cleaning prior to linkage
- Scalability
  - Naïve comparison of all record pairs is computationally intensive
  - Remove likely non-matches as efficiently as possible
- No training data in many linkage applications
  - No record pairs with known true match status
- Privacy and confidentiality
  - Personal information, like names and addresses, are commonly required for linking

The extended record linkage process



Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

Identifiers

- Typical identifiers:
  - People: first and last name, address, birth date, sex
  - Establishments / firms: name, legal form, address
- The higher the number of different manifestations of an identifier, the better its suitability for a comparison.
- Complex identifiers should be parsed into its separate components
- Means of getting clean identifiers in the first place

Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

Importance of Preprocessing

“In situations of reasonably high-quality data, preprocessing can yield a greater improvement in matching efficiency than string comparators and ‘optimized parameters’. **In some situations, 90% of the improvement in matching efficiency may be due to preprocessing**.” (Winkler 2009, p. 370)

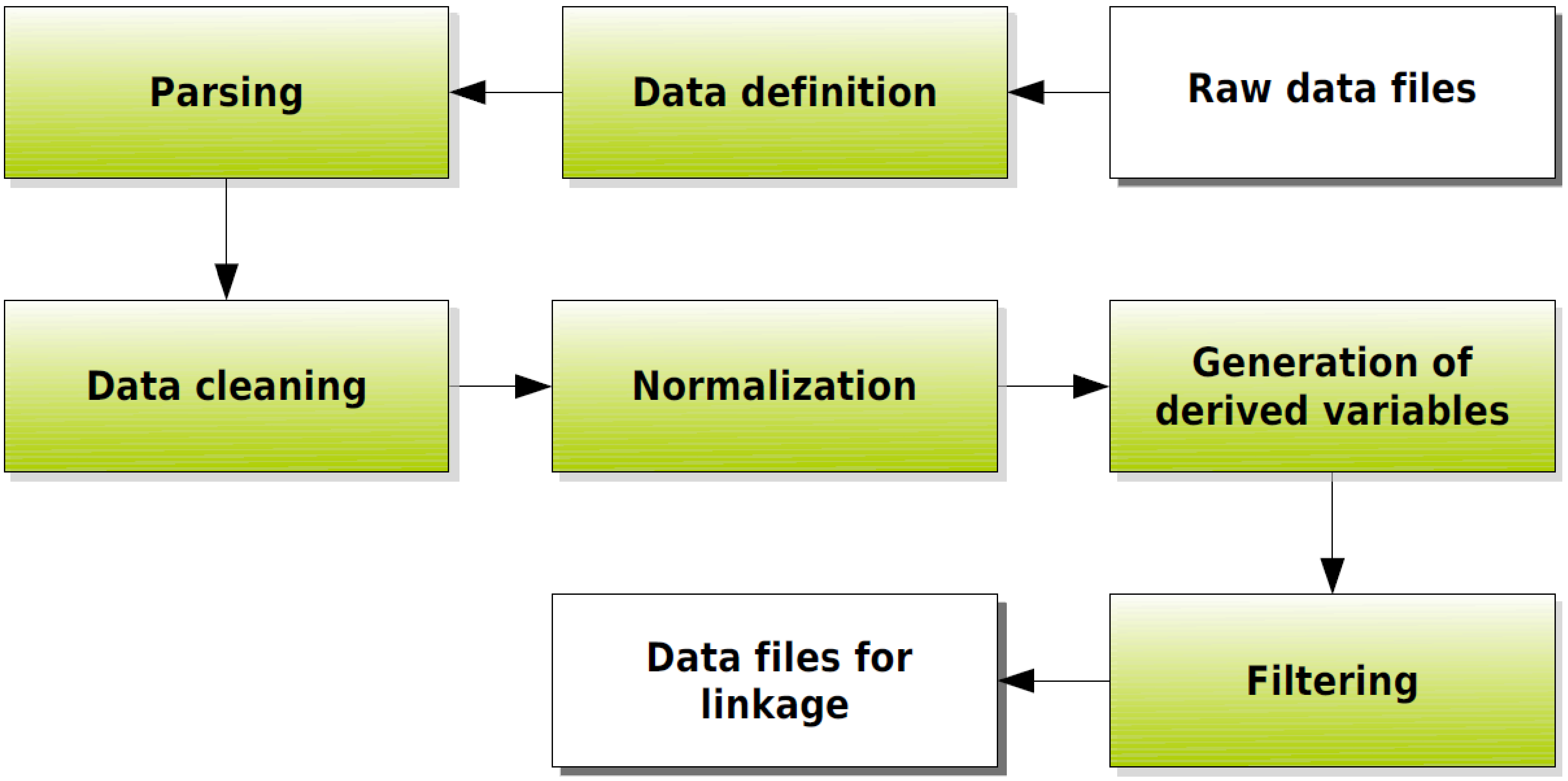
“Inability or lack of time and resources for cleaning up files in preparation of matching are often the main reasons that matching projects fail.” (Winkler 2009, p. 366)

Shares of effort within linkage process

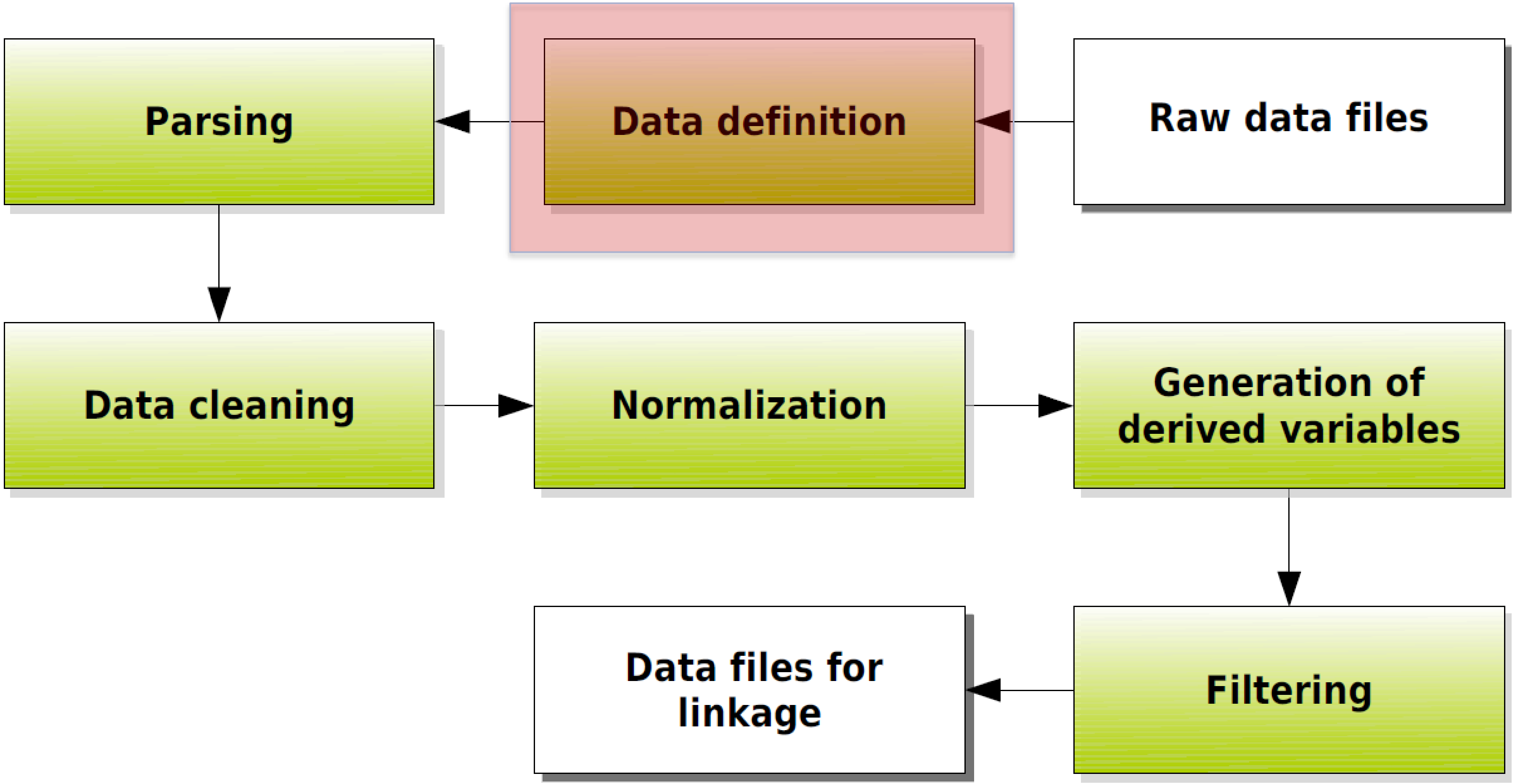
- 5% matching and linking efforts
- 20% checking that the computer matching is correct
- 75% cleaning and parsing the two input files

(see Gill 2001, p. 31)

Preprocessing: Workflow



Preprocessing: Workflow



Creating a data definition

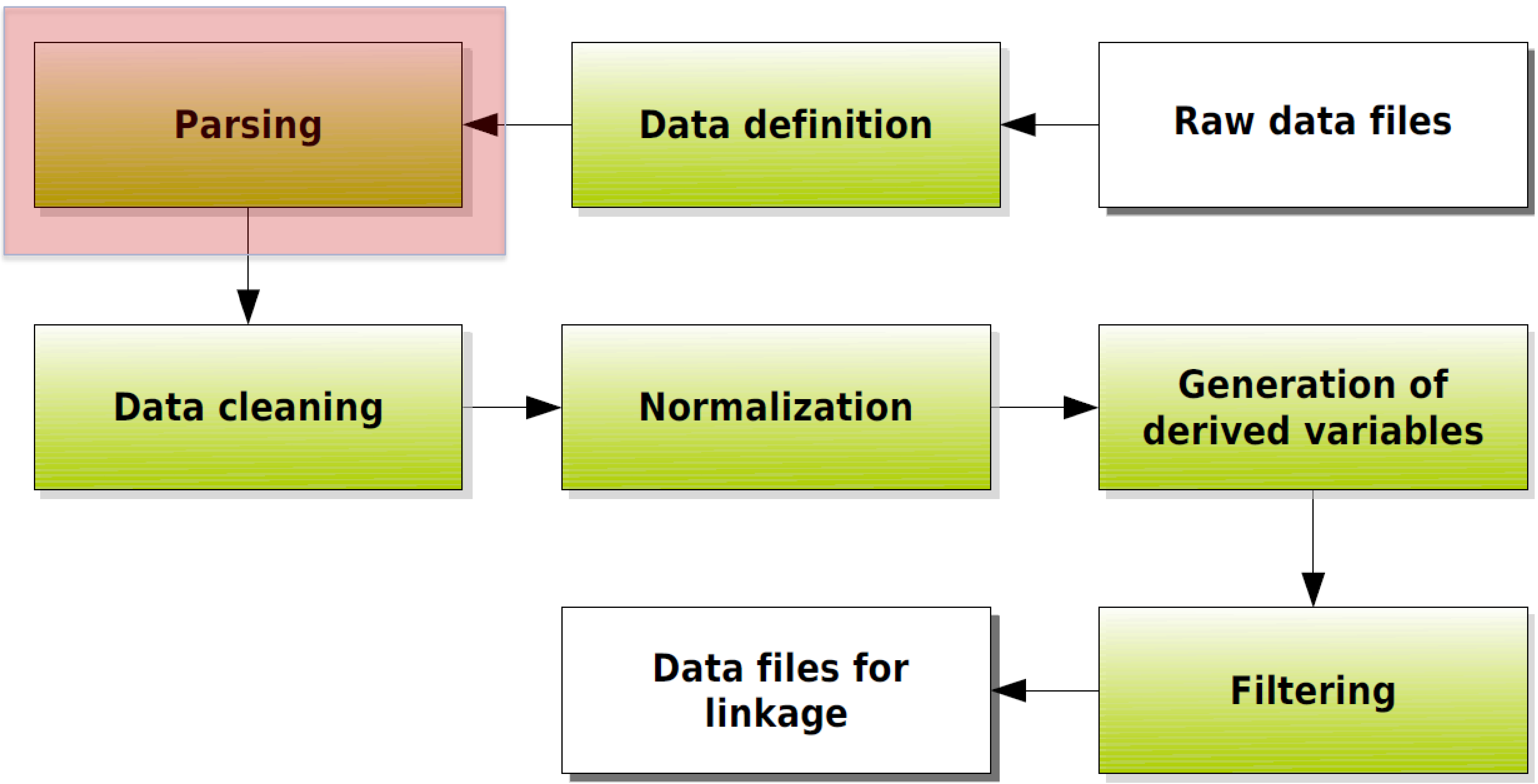
A data definition records attributes for each identifier that are assigned to them conceptually.

It should encompass: variable name, variable type, data type, missing code, code list, variable range, among others…

Example of data definition for sex:

1. Variable name: sex
2. Variable type: categorical, coded
3. Data type: byte
4. Code list: 1 male 2 female 3 not determinable 9 missing

Preprocessing: Workflow



Parsing

- Parsing is the decomposition of a complex variable into single components.
- Subsequently, the single components can be composed to a standard form or can be used as single match variables.
- In simple cases the decomposition takes place through delimiter or through simple regular expressions.
  - Example: field with zip code and place name

Example: Parsing of addresses

Address
39B Lexington Str. 01705 Chicago/ Wheaton

^39B Lexington STR 01705 Chicago / Wheaton$

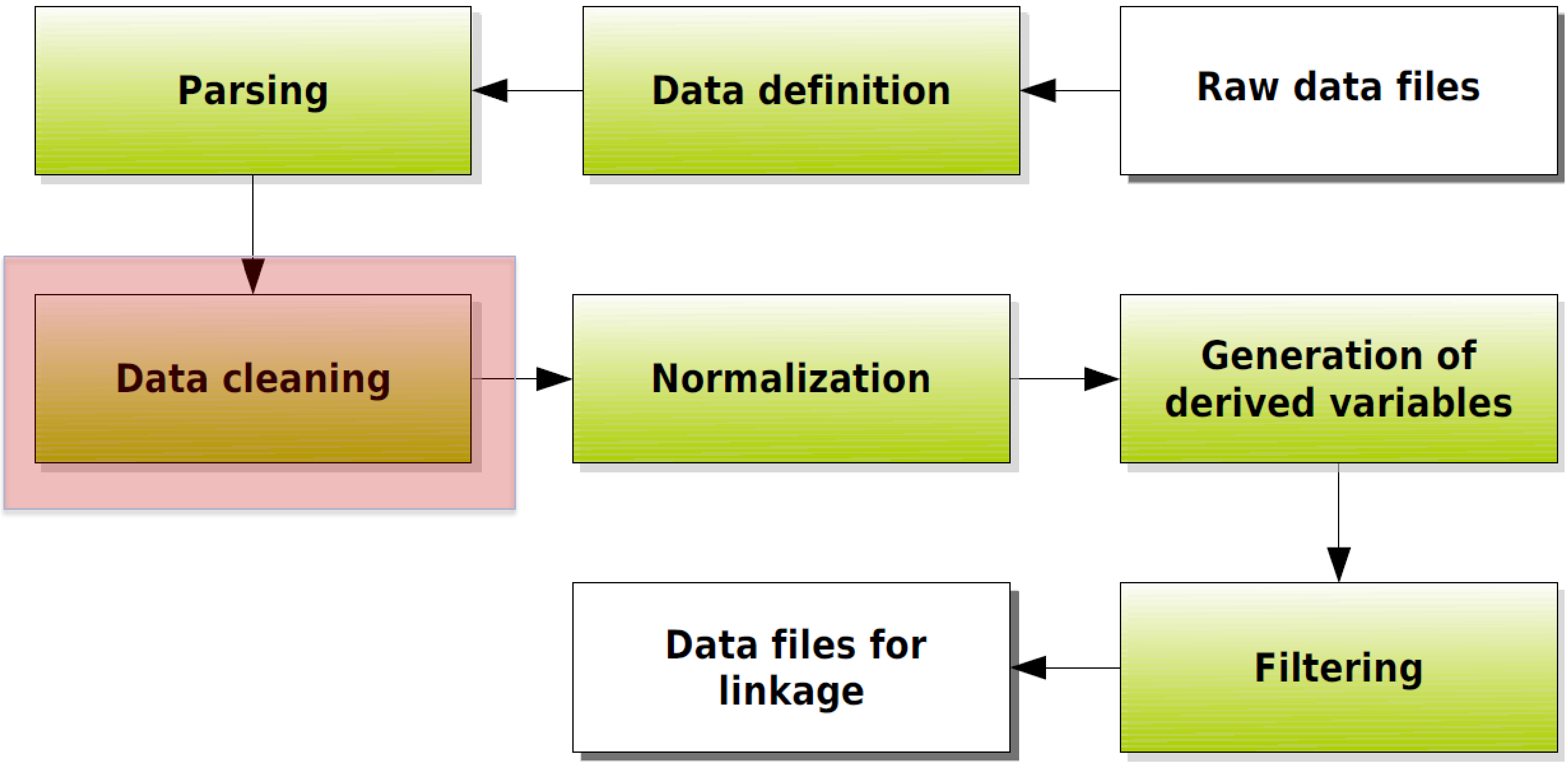
town name	district	zip code	street name	strqual	hnr	hnr.affix
Chicago	Wheaton	01705	Lexington	Street	39	B

Lookup tables for standardization

- Typical are tables for tokens in establishment names, personal names and addresses.

Token	Replacement
str	STR
Street	STR
⋮	⋮
Dr.	DR
Dctr.	DR
Doctor	DR
⋮	⋮
Co	CO
Company	CO
Cmpy	CO
⋮	⋮
sen.	SENIOR
SENIOR	SENIOR
Junior	JUNIOR

Preprocessing: Workflow



Data cleaning: Overview

1. Evaluation of identifiers against data definition
2. Checking plausibility of variable values
3. Checking records for consistency
4. Standardization
5. Deduplication

Checking records for consistency

- Searching for consistency errors: values of least two variables contradict each other, while each value on its own is allowed.
- Examination through formulating and examination of edit-rules (brief: edits)
- For continuous variables mostly equations, for categorical variables mostly if-then rules
  - Ratio edits: $y/x = z$
  - Balance edits: $y + x = z$
  - Consistency edits: if AGE = 15 then STATUS != married
- Alternative: Examination through comparison with lookup tables, which contain pairs (or triples etc.) of variable values, e.g. zip code-place-street

Standardization

- Standardization of different representations of the same information in uncoded categorical variables.
- Without standardization there are more false negatives.
- However, over-standardization leads to risk of false positives.
- Always rule-based: application of a number of replacement rules to the identifier.
- Common implementation of the rules with regular expressions (search & replace).

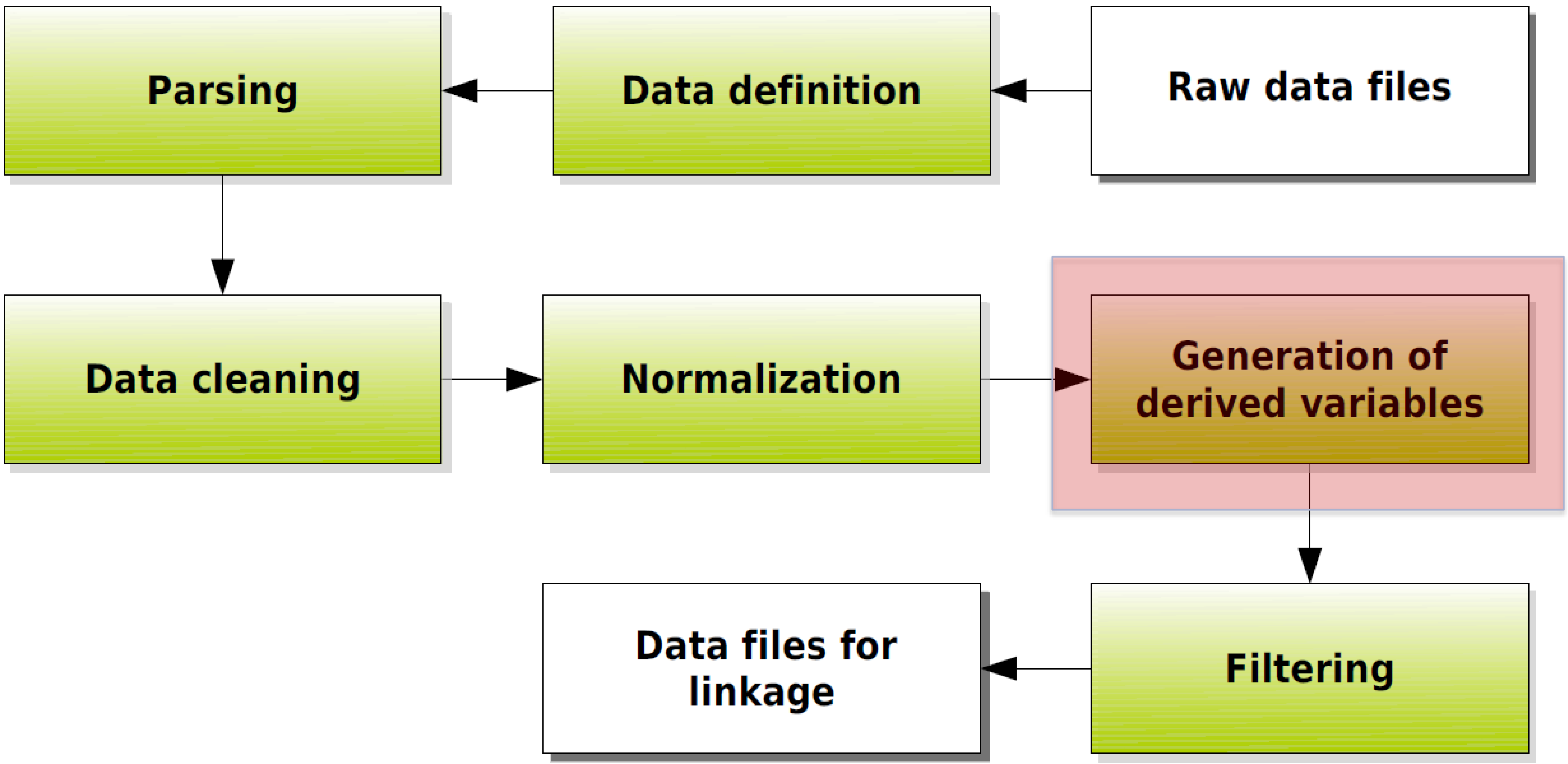
Typical lookup tables for standardization

Entries and their corresponding standard representations:

- Abbreviations in street names (St: Street; Dr: Drive; Blvd: Boulevard)
- Abbreviations and frequent words in establishment names (b.o.: branch office; gen.: general)
- Nickname (Bob: Robert; Jim: James)
- Title (Dr: Doctor)
- Name affixes (v: von; sen: senior)

➔ Highly country and language specific

Preprocessing: Workflow



Generating of derived variables

- Usually used to get appropriate blocking variables
- Typically over-standardized variants of existing identifiers
- Examples:
  - Phonetic codes of first and surname
  - Initial letters of first and surname
  - Truncation of zip codes to 3 or 4 digits

Statistics New Zealand: Standard preprocessing of surname

1. Take **surname**
2. Capitalize
3. Remove spaces
4. Set to missing if surname contains “unknown”
5. Remove any characters other than alphabetic characters
6. Name the resulting field **surname1**
7. Define new variable **initial_surname** = first character of **surname1**
8. Define new variable **soundex_surname** = Soundex code of **surname1** (See Statistics of New Zealand 2006, p.50)

Preprocessing: key take-aways

- Preprocessing is always specific for the concrete application.
  - Example: Establishment vs. individual data
- Expenditure of time for preprocessing often exceeds efforts of the record linkage (comparison, classification).
- Especially with bad data quality preprocessing is the most important factor for the success of linkage projects.
- Budget enough resources for preprocessing.
- Neither is there a universally suitable software for this, nor is there a comprehensive textbook.

Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

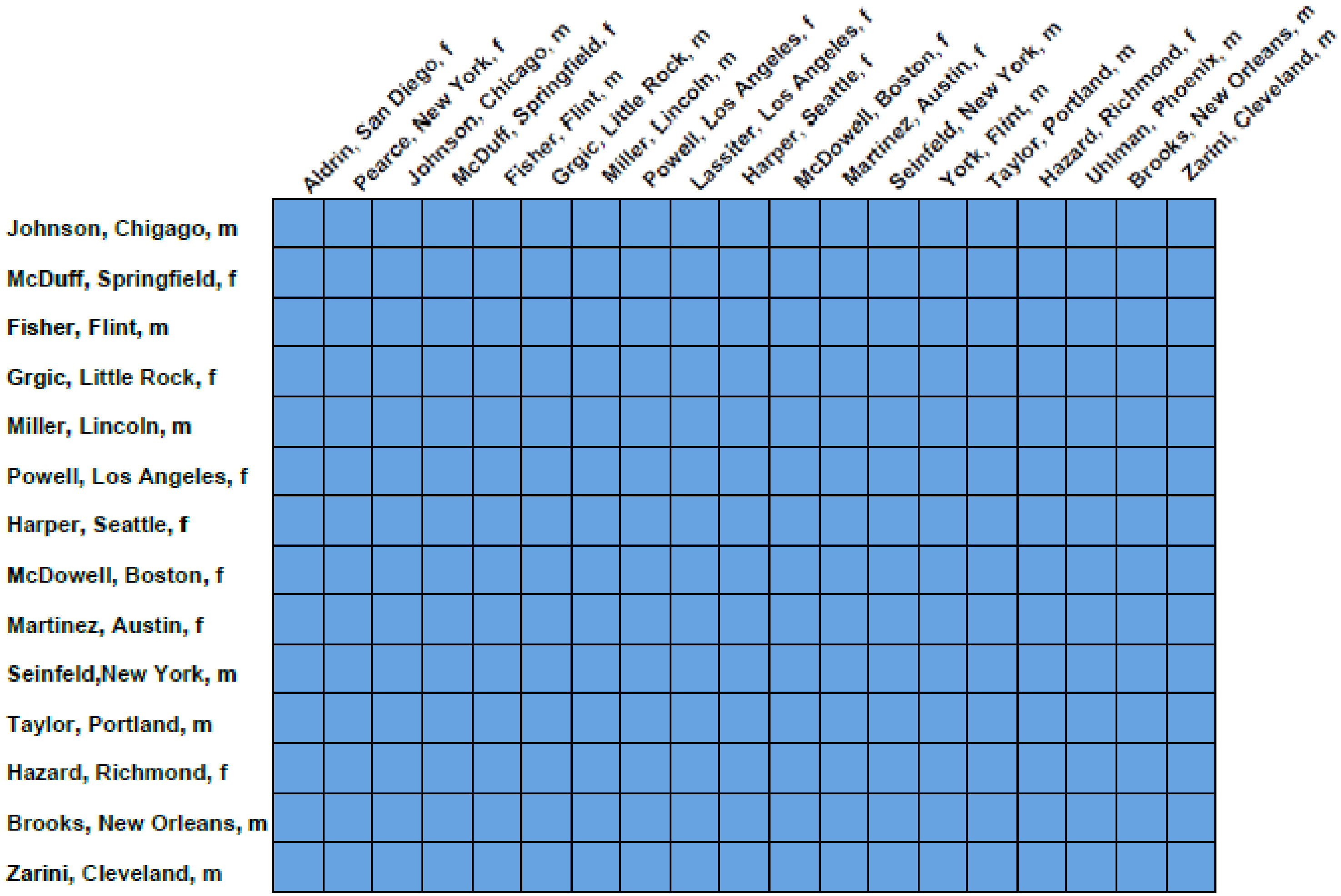
The efficiency problem

- With *n* records in file *A* and *m* records in file *B*, *n x m* pairs have to be compared.
- 100 000 x 100 000 = 10 000 000 000 (10 billion) comparisons
- With 10 000 comparisons per second this takes 278 hours or 11.6 days

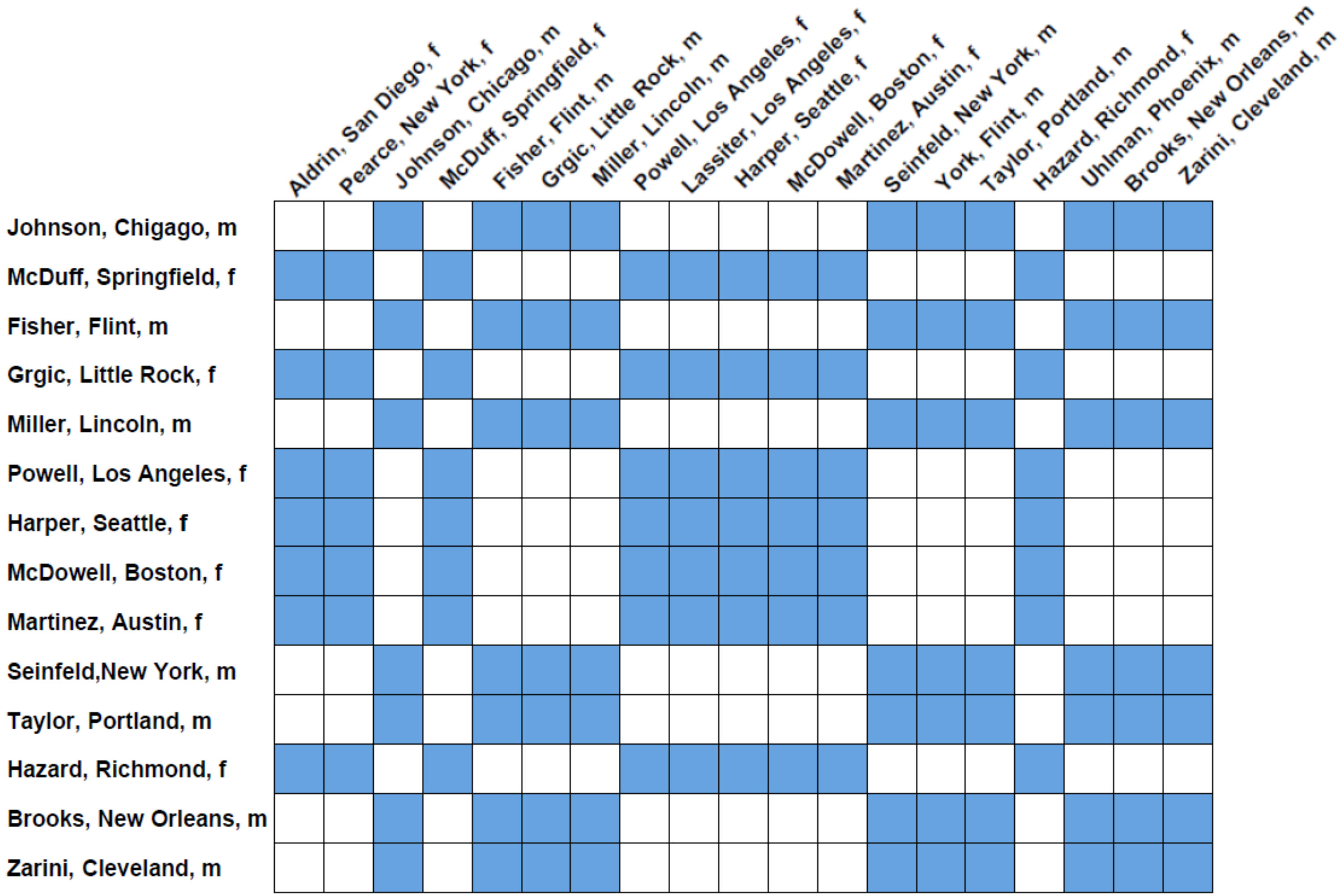
Standard technique: Traditional blocking

- According to its values, a variable partitions both data files into subsets, called blocks or pockets.
- The A- and the B-file are partitioned using the same (blocking) variable.
- Only pairs of records belonging to the same block within a certain file are compared.

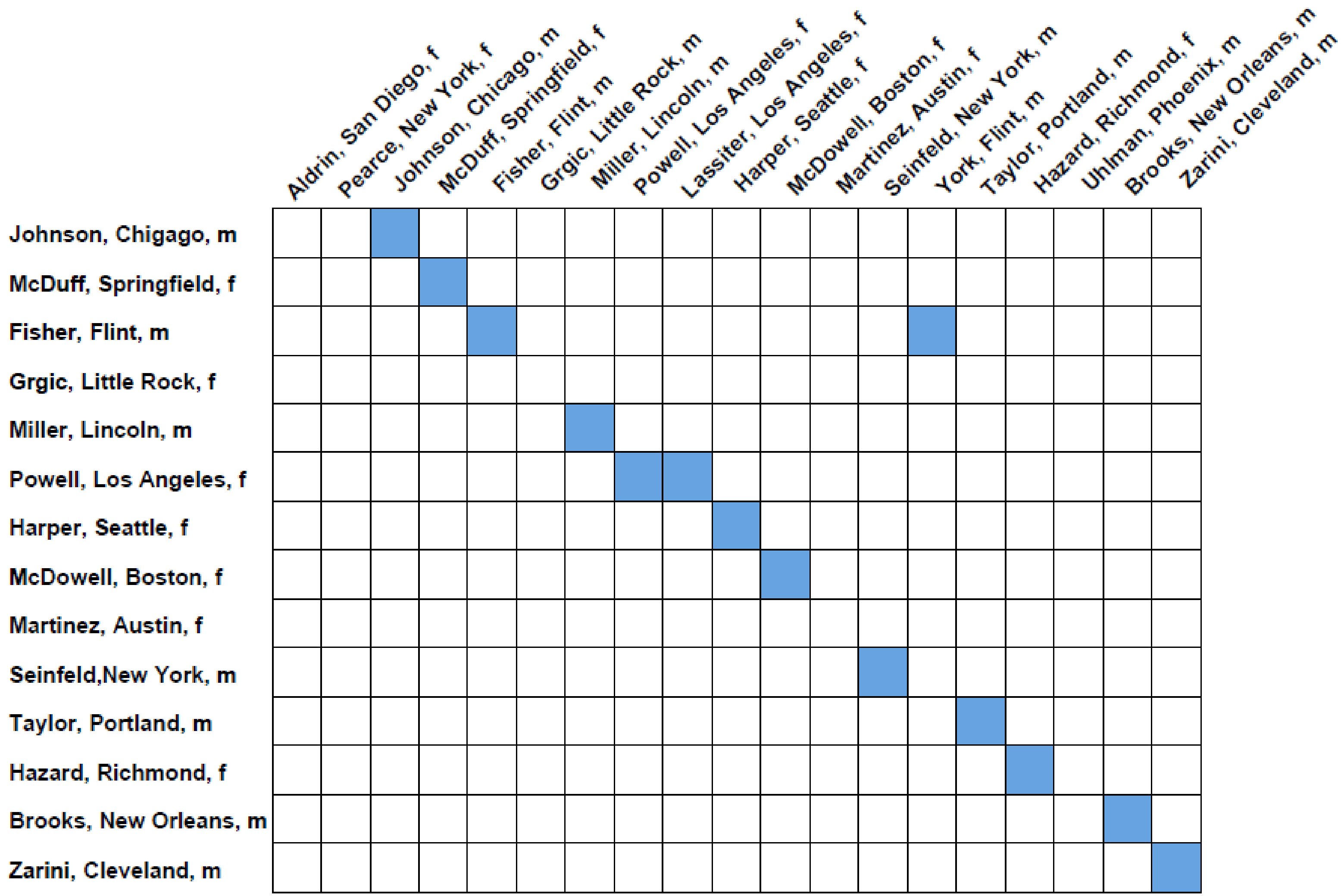
Example: No blocking



Blocking by sex



Blocking by sex and location



Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

String similarities

- Function of a pair of character strings with similarity as function value.
- Common: Standardization of the function value to the interval [0-1] (0: no agreement; 1: complete agreement).
- Variations of the following classifications of string similarity functions are commonly used:
  - Phonetics
  - Edit-distances
  - n-grams
  - Jaro‘s string comparator

Edit-distances: Principle

- An edit-distance between two strings *a* and *b* is the lowest number of permitted edit-operations needed to transfer *a* to *b*
- A certain edit-distance variant is defined by the set of permitted operations.
- For the Levenshtein-distance, for example insertions, deletions and substitutions are allowed
- Common: Normalization using the sum of the length of the strings
- Similarities are obtained by $1 - LD_{norm}$

Levenshtein-distance: Examples

Names	Edit operations	Norm. distance
Neumann Naumann	1 x substitution	1 x 2/14 = 0.14
Maier Meyer	2 x substitutions	2 x 2/10 = 0.40
Mohr Moore	1 x deletion 1 x substitution	2 x 2/9 = 0.44
Acri Ascheri	1 x insertion 3 x deletions	4 x 2/11 = 0.73
Adams Adams	1 x insertion	1 x 2/9 = 0.22

Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

The Fellegi-Sunter Approach: General

- Every pair of records is compared and represented using a vector of components that describe similarity between individual record fields (i.e. identifiers)
  - E.g., “name agrees”, “name disagrees”, “name missing on one or both records”

Principles (I)

- Simply summing up identifier matches cannot be optimal.
- Different identifiers differ in how strongly an agreement is indicative for a link.

Name	Sex	Residence	Date of birth
Tom McDonalds	m	Albuquerque	12/06/1966

Principles (II)

- Assigning appropriate weights to identifiers before summing up would be a better method.
- In order to weight identifiers it must be quantified for each identifier how strongly an agreement indicates a *link*.
- How likely is an agreement within the *matches* compared to within the *non-matches*?

m- and u-probabilities

- The term probabilistic record linkage results from the fact that two conditional probabilities are considered:

$$m_i = P(a_i = b_i \mid M)$$

m-probability for *i*: Probability for agreement of records *a* and *b* for identifier *i* within the matches.

$$u_i = P(a_i = b_i \mid U)$$

u-probability for *i*: Probability for agreement of records *a* and *b* for identifier *i* within the non-matches.

Central likelihood ratio (I)

- The weighting of the identifiers is done by this ratio of likelihoods:

$$\frac{m_i}{u_i} = \frac{P\ (a_i = b_i \mid M)}{P\ (a_i = b_i \mid U)}$$

- The rarer an agreement occurs within the *non-matches* compared to the *matches*, the more strongly does an agreement within the identifier indicate a *link*.
- Therefore, $\frac{m_i}{u_i}$ quantifies how strongly an agreement within an identifier *i* indicates a *link*.

The Fellegi-Sunter Approach: finding links

Given m_i and u_i, link status is determined by considering the likelihood ratio (also known as the **match weight** or **match score**):

$$R = \frac{p(\gamma|M)}{p(\gamma|U)}.$$

Choose thresholds T_1 and T_2:

- Pairs with $R \geq T_1$ are linked
- Pairs with $R \leq T_2$ are not linked
- Pairs with $T_1 > R > T_2$ are sent for **clerical review**

Illustration

- Generally, the identifier *sex* agrees for 99% of the matches; within non-matches agreement usually occurs in 50% of cases.
- In the case of the identifier *surname* typically agrees in about 80% among the matches and in 0.1% among non-matches.

Variable (characteristics)	m-prob	u-prob	m / u	(1-m) / (1-u)
Sex	0.99	0.5	1.98	0.02
Last name	0.80	0.001	800	0.2

- An agreement on *surname* indicates a classification as a link much more strongly than an agreement on *sex*.
- A disagreement on *sex* indicates a classification as a non-link much more strongly than a disagreement on *surname*.

Content of this session

- Introduction to Data/Record Linkage (RL)
- Identifiers
- Preprocessing
- Increasing the Efficiency of the Matching Step (Blocking)
- String comparators
- Probabilistic RL
- Application
- Appendix (RL Software, Literature)

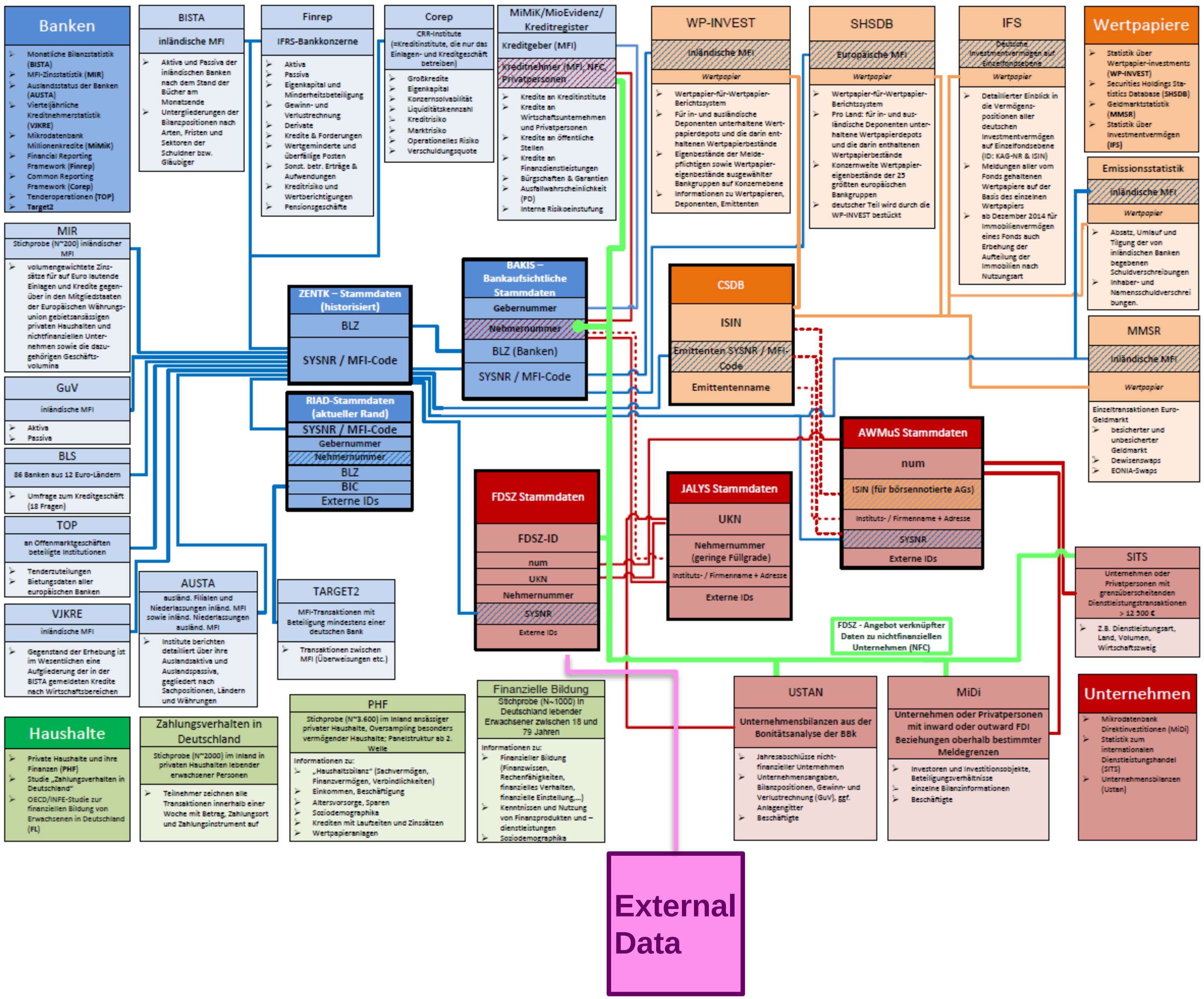
Linking Deutsche Bundesbank Company Data

With a lot of help by Dr. Christopher-Johannes Schild
and Sebastian Seltman

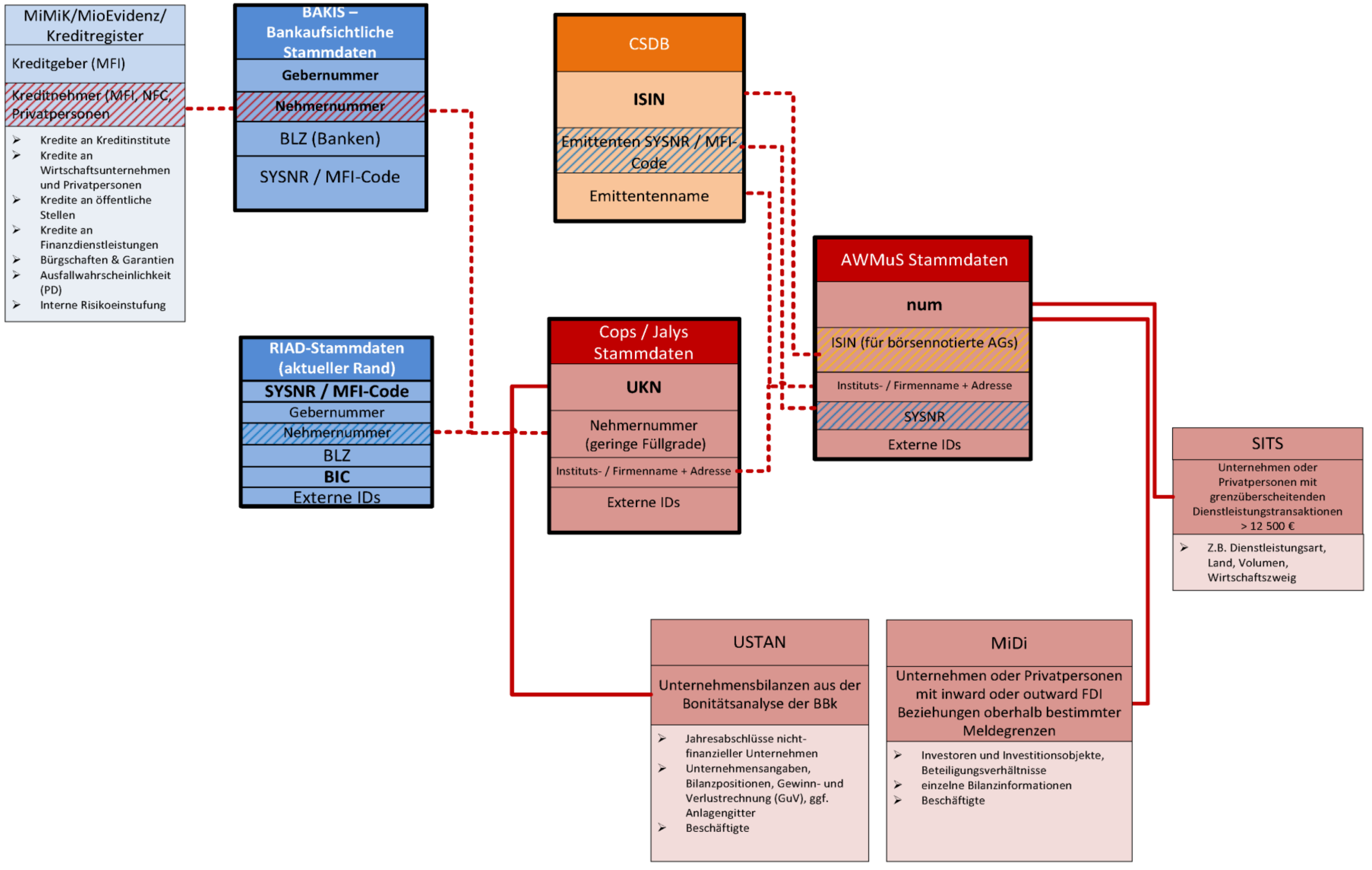
Data/Record Linkage: Goals of Bundesbank

- Improve data quality
- Increase analytical value of data
- More general and flexible Record Linkage System
- Historicized matching tables

Bundesbank's relevant microdata sources and their connections (excerpt)



Bundesbank's relevant microdata sources: Company Data



Company Data I: Microdatabase Direct Investment (MiDi)

- Information on inward foreign direct investments (FDI) as well as outward FDI
  - Granular information on FDI from domestic companies to companies located in other countries and incoming FDI from foreign owned companies to domestic and foreign owned companies
  - Statistical units: reports that contain the investment relationship between the transaction parties
  - Micro data is available as a panel

Company Data II: Corporate Balance Sheets (Ustan)

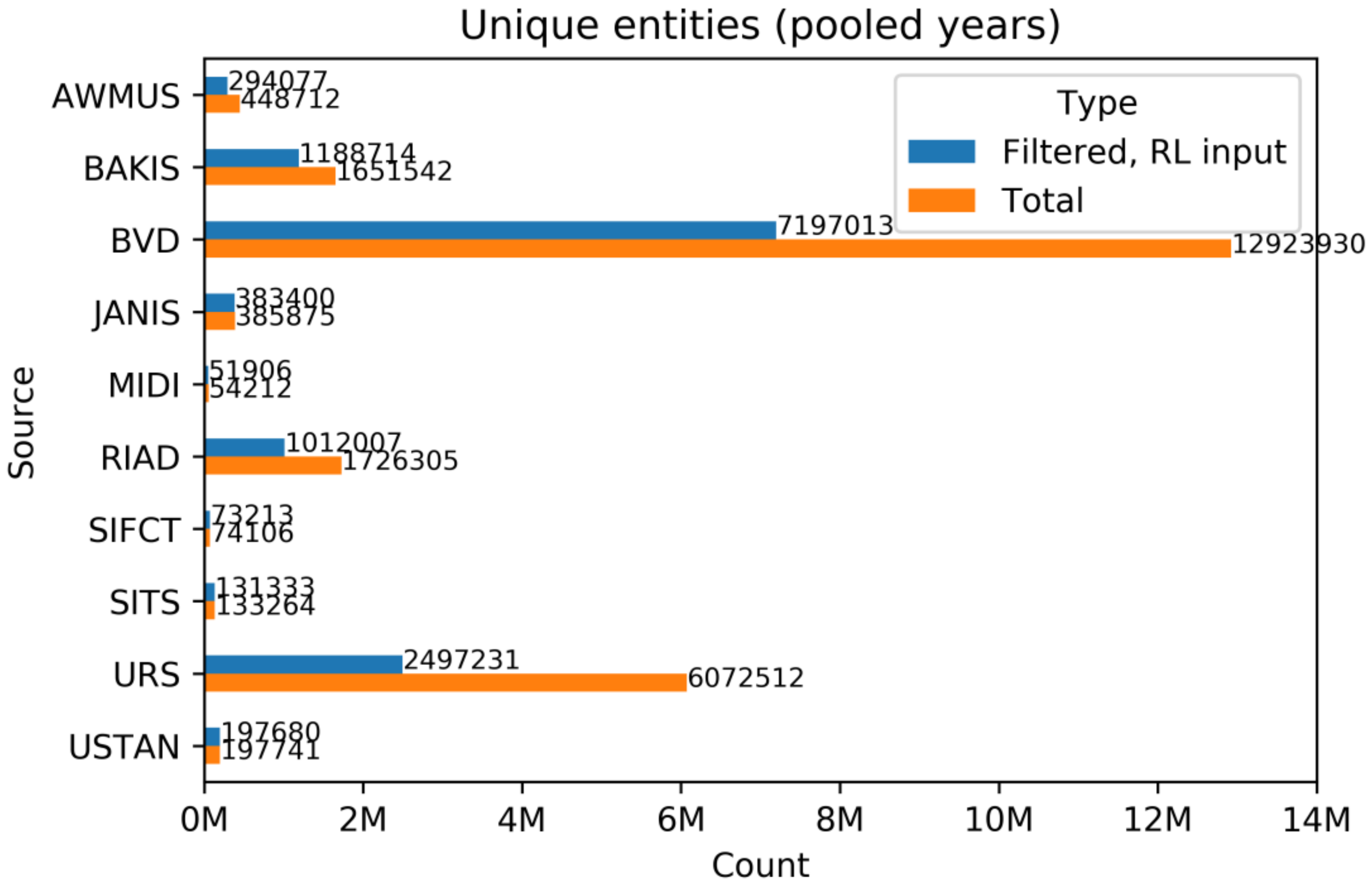
- Originates from Bundesbank’s refinancing activities.
- Information on earnings and financing.
- Detailed information on companies:
  - balance sheets, profit and loss accounts
  - comprehensive income statements („Ergebnisverwendungsrechnung“)
  - statement of changes in fixed assets („Anlagengitter“)
- Bias towards enterprises with higher credit-worthiness

Company Data for the Linkage

In sum there are 15 company data to be linked:

- 7 analytical datasets from Bundesbank covering different time frames between 1987 – 2021.
- 8 master datasets covering different time frames between 1980-2021:
  - 5 from Bundesbank
  - 1 from BvD (reference data complemented by ZEW data)
  - 1 from Global Legal Entity Identifier Foundation (LEI)
  - 1 from the Statistical Agency, the (Statistical) Business Register of Germany

Numbers of companies of sum data data sets



Data linkage

Company data (non financial institutions (NFI)):

There is **no common unique firm identifier** in Germany.

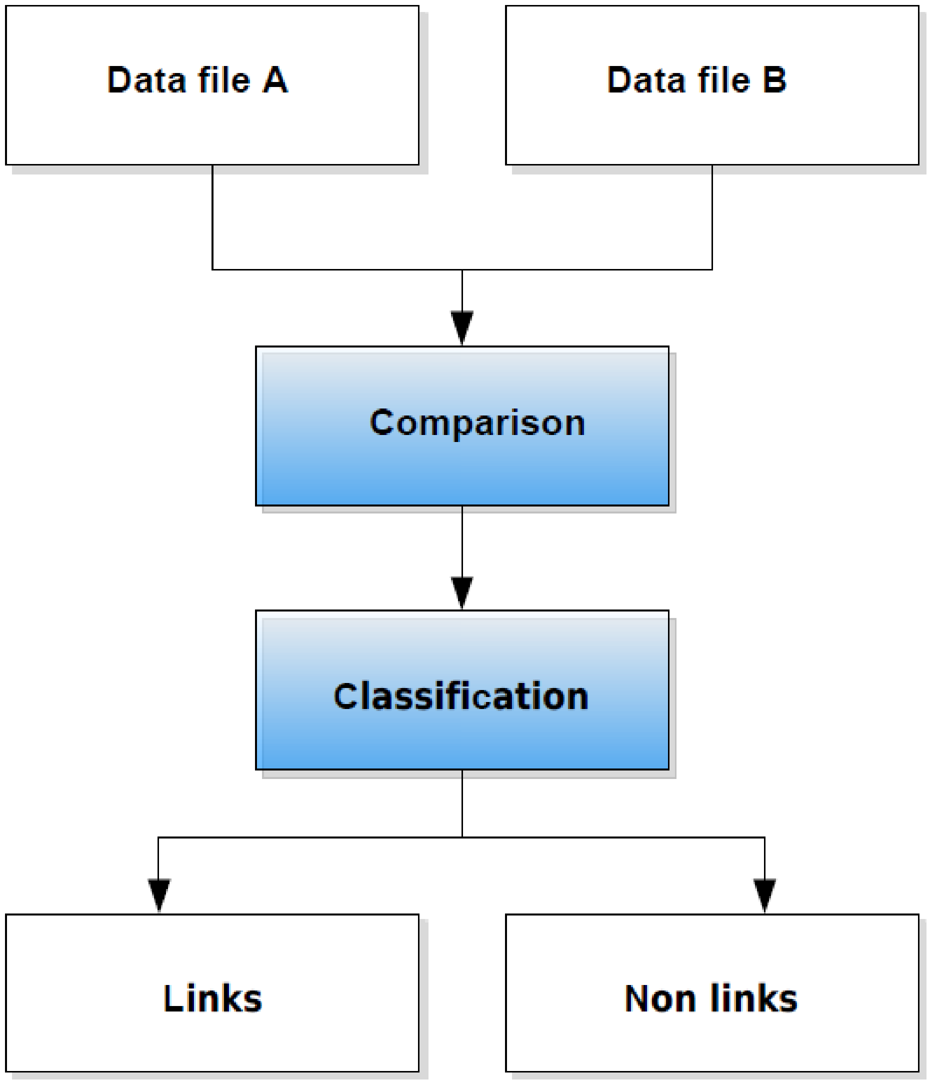
(Company business register-ID not stable)

We have to match firm data…

- … that do not have a common unique identifier / key
- … by using **alternative identifiers** (such as names, addresses, sectors, legal forms)

RDSC has matched several NFI-microdatasets (from Statistics, Banking Supervision and external data) with an advanced machine learning algorithm and generated a **matching table** (with probalistic matching scores)

The basic record linkage process



There is no perfect world

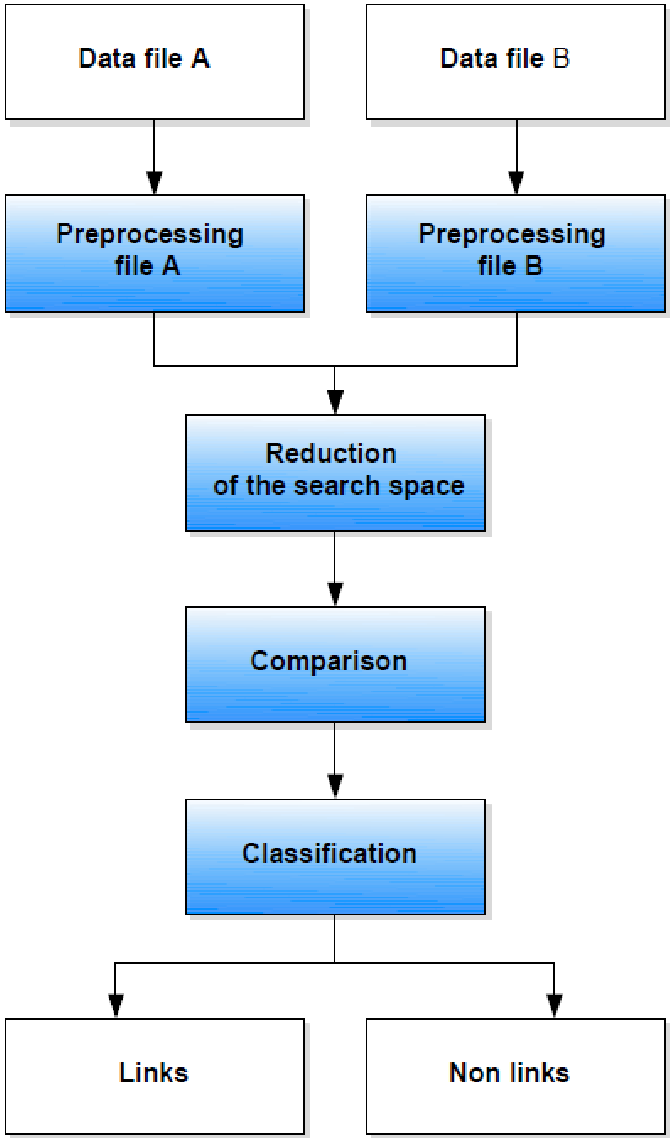
- In a perfect world

True Positive (TP)	
	True Negative (TN)

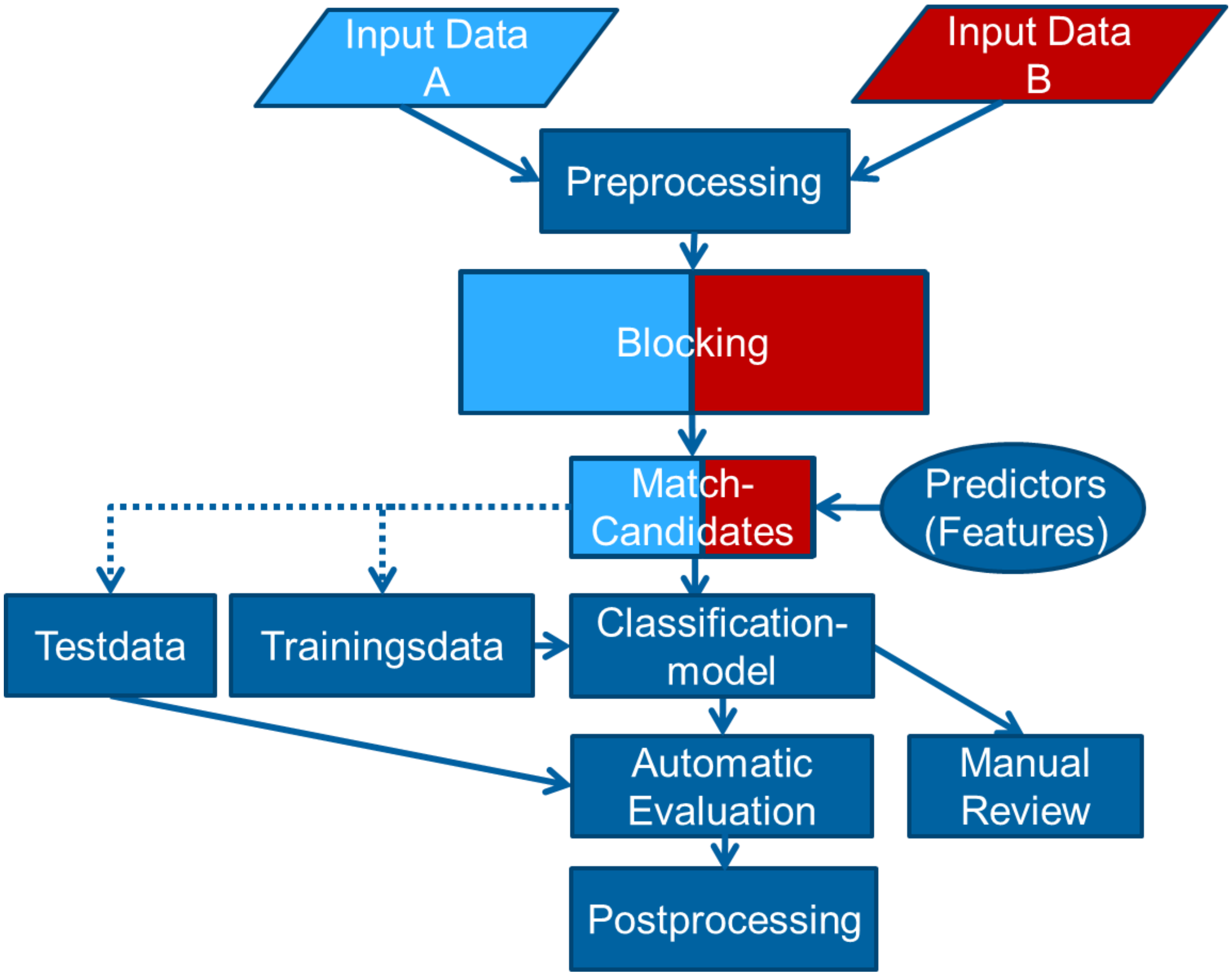
- But we do not live in a perfect world

True Positive (TP)	**False Positive (FP)**
False Negative (FN)	**True Negative (TN)**

The extended record linkage process



Record Linkage Process



THX to Christopher-Johannes Schild

The “Fun” with Company Names: BMW as an example I

- BMW is an abbreviation for *Bayerische Motoren Werke* (…)
- This name is grammatically incorrect (in German, compound words must not contain spaces), which is why the name's grammatically correct form *Bayerische Motorenwerke*
- *Bayerische Motorenwerke* translates into English as *Bavarian Motor Works*.
- The suffix AG, short for *Aktiengesellschaft*, signifies an incorporated entity which is owned by shareholders, thus akin to "Inc." (US) or PLC, "Public Limited Company" (UK).

(source: https://en.wikipedia.org/wiki/BMW)

The “Fun” with Company Names: BMW as an example II

- BMW
- BMW AG
- BMW Aktiengesellschaft
- *Bayerische Motoren Werke*
- *Bayerische Motoren Werke* AG
- *Bayerische Motoren Werke Aktiengesellschaft*
- *Bayerische Motorenwerke*
- *Bayerische Motorenwerke* AG
- *Bayerische Motorenwerke* Aktiengesellschaft

is the same company (and only the German possibilities are shown).

The “Fun” with Company Names: BMW as an example III

- In 2020 Germany had 7,389 Aktiengesellschaften.
- One of them is Bayer AG

- Bayer is clearly not BMW, but what is happening if you compare the names
- Bayer Aktiengesellschaft with BMW Aktiengesellschaft?
- Seems to be really close, right?

The “Fun” with Company Names: BMW as an example IV

- Last Sunday 677 BMW partners were found on https://www.bmw.de/de/fastlane/bmw-partner.html
- It is not always clear, if they use BMW in their names.
- You need a clear definition of what you mean by company, firm, establishment…

Preprocessing: Firm Names as one Example

- Remove known variation in different correct notations,
  - such as standardizing the German word “Gesellschaft” to its most common abbreviation “Ges”
  - and “&,” “+,” “und,” “and” etc to “UND”.
- Replacing German Umlauts “ä,” “ö,” “ü” by their common non-Umlaut replacements “ae,” “oe,” “ue” as well as capitalizing.
- Legal form information is extracted from the firm name field and removed from the firm name.

Doll, Gabor-Toth, Schild (2021): p. 8

Blocking

- Filter variables: 1. cleaned company name,
  2. cleaned company name tokens, truncated,
  3. city, 4. postal Code, 5. street name, 6. NACE (2 digits), 7. telephone, 8. founding year, 9. legal form.
- Combination of these variables comes to a total of 1,130 blocking keys.
- Overall, the blocking procedure reduces the number of comparisons from the order of roughly $N = 10^{13}$ to about $C = 10^{8}$ candidate pairs.

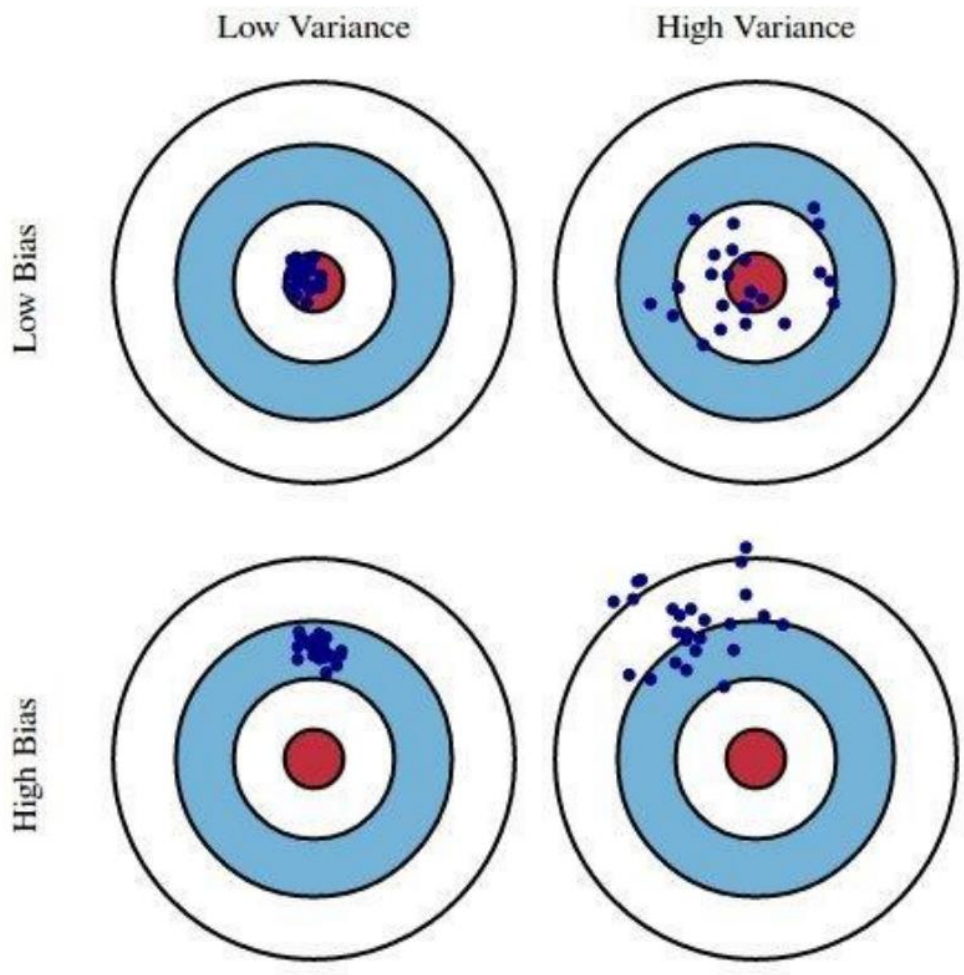
Doll, Gabor-Toth, Schild (2021): p. 9

Classification Model

- **Comparison Features:**
  - A lot of different comparison features are used: name-based, location-based (including geo-references), digits from founding year and the sector code.
- **Groundtruth:**
  - based on common IDs
- **Training / Test Split**
- **Match Prediction**
  - The First Level “base’’-Models are:
    1. random forest
    2. “extreme gradient boosting trees’’-model (XGBoost)
    3. logistic regression
  - Second Level Model takes the first level model scores as features, plus 3 string comparison features.

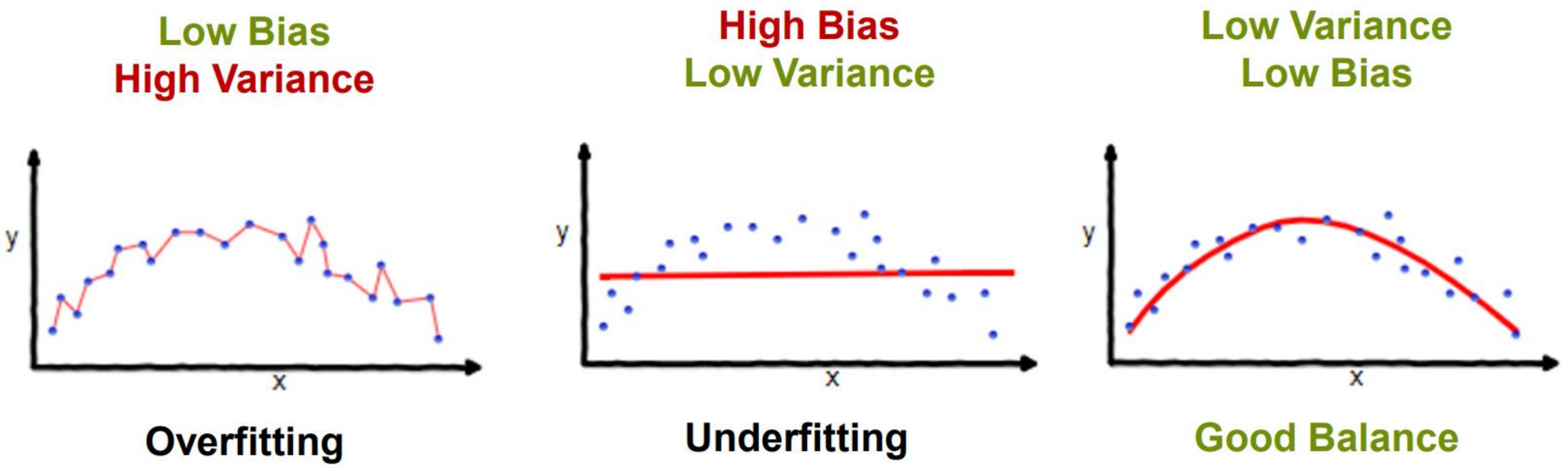
Doll, Gabor-Toth, Schild (2021): pp. 10-14

Bias-Variance Tradeoff: Example I



THX to Sebastian Seltmann

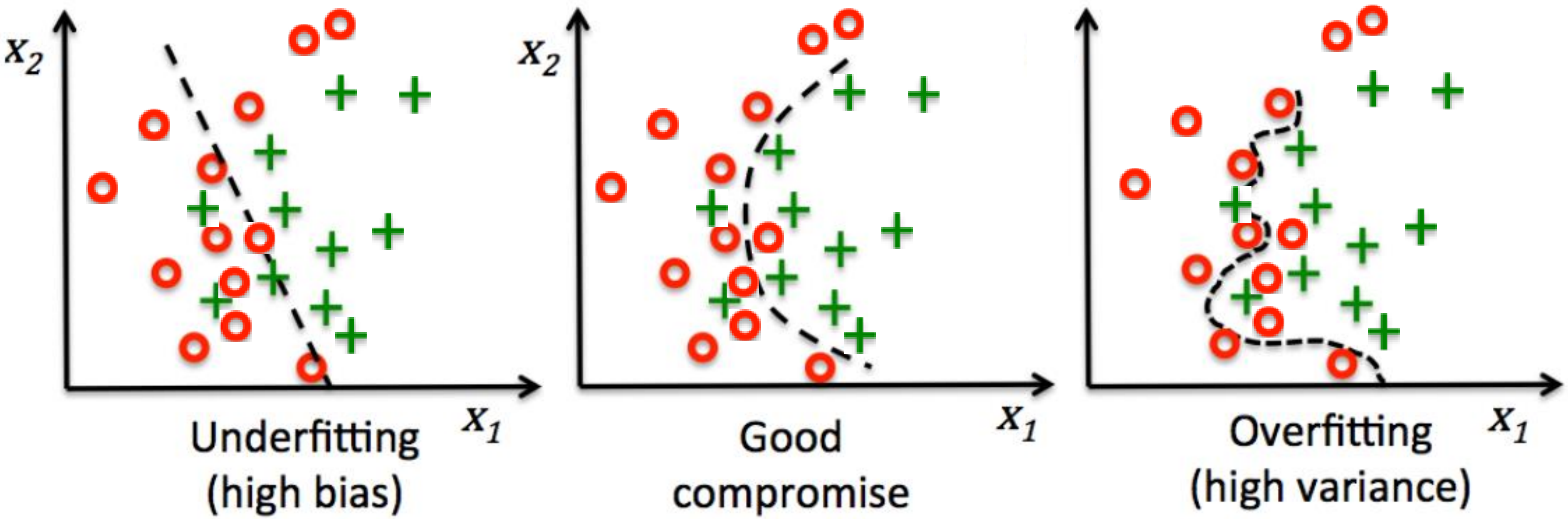
Bias-Variance Tradeoff: Example II



THX to Sebastian Seltmann

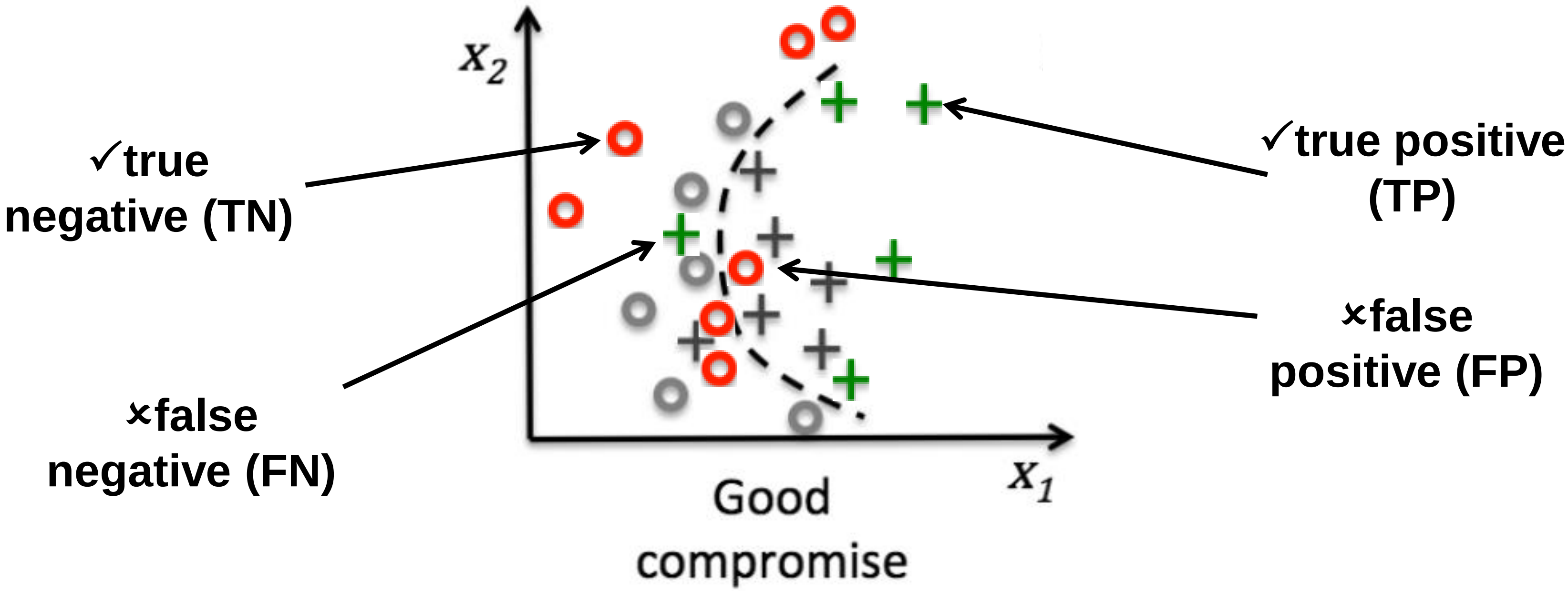
Classification

- Bias vs Overfitting



THX to Christopher-Johannes Schild

Classification



THX to Christopher-Johannes Schild

Model Evaluation Measures: Scoring Functions I

- What is the share of **correctly predicted** cases?
  - Accuracy = (TN + TP) / Total
  - Total = TP+TN+FP+FN

Confusion Matrix

		Actual Data	
		No Default	Default
Predictions	No Default	TN	FN
	Default	FP	TP

- Which share of the **true default** cases is correctly predicted?
  - Sensitivity (or Recall) = TP / (TP + FN)

Confusion Matrix

		Actual Data	
		No Default	Default
Predictions	No Default	TN	FN
	Default	FP	TP

- Which share of the **true non-default** cases is correctly predicted?
  - Specificity = TN / (TN + FP)

Confusion Matrix

		Actual Data	
		No Default	Default
Predictions	No Default	TN	FN
	Default	FP	TP

THX to Sebastian Seltmann

Model Evaluation Measures: Scoring Functions II

Confusion Matrix

		Actual Data	
		No Default	Default
Predictions	No Default	TN	FN
	Default	FP	TP

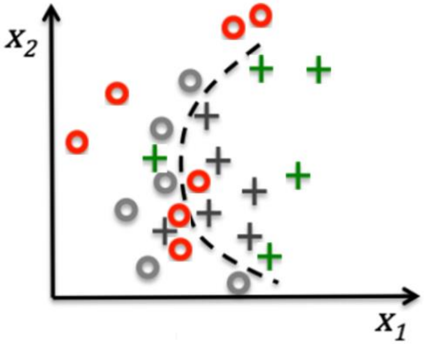
- What is the share of **correct "default" predictions**?
  - Precision = TP / (TP + FP)

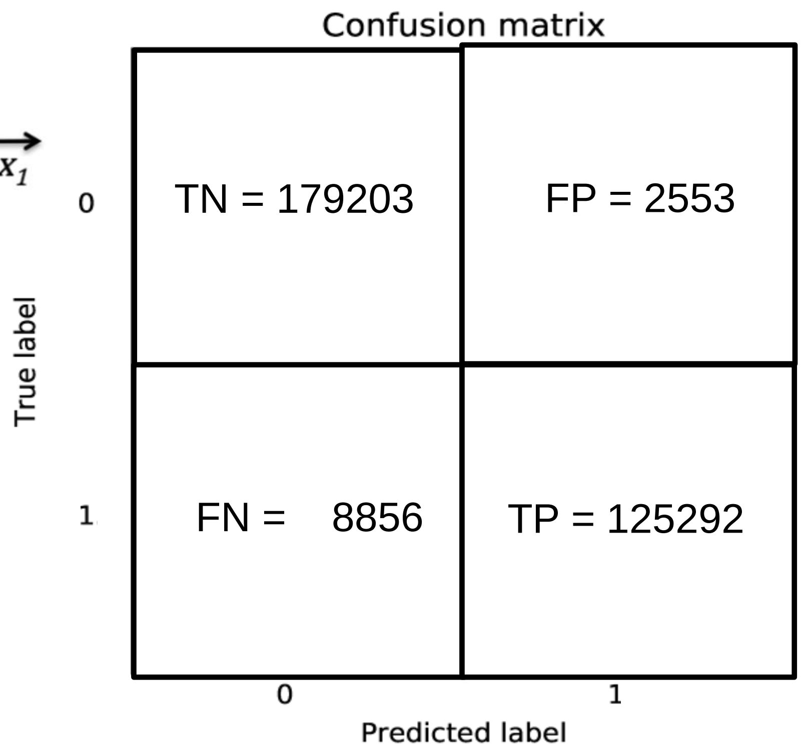
- Can the model identify **true "default"** cases without many **false alarms**?
  - F1-Score = $2 * \frac{Precision * Recall}{Precision + Recall}$
  - F1-Score provides a tradeoff between precision and recall

- How sure is the model in its predictions?
  - Log Loss = $-\frac{1}{N}\sum_{i=1}^{N}[y_i \log p(y_i) + (1 - y_i)\log(1 - p(y_i))]$
  - Penalizes confident incorrect predictions

THX to Sebastian Seltmann

Evaluation



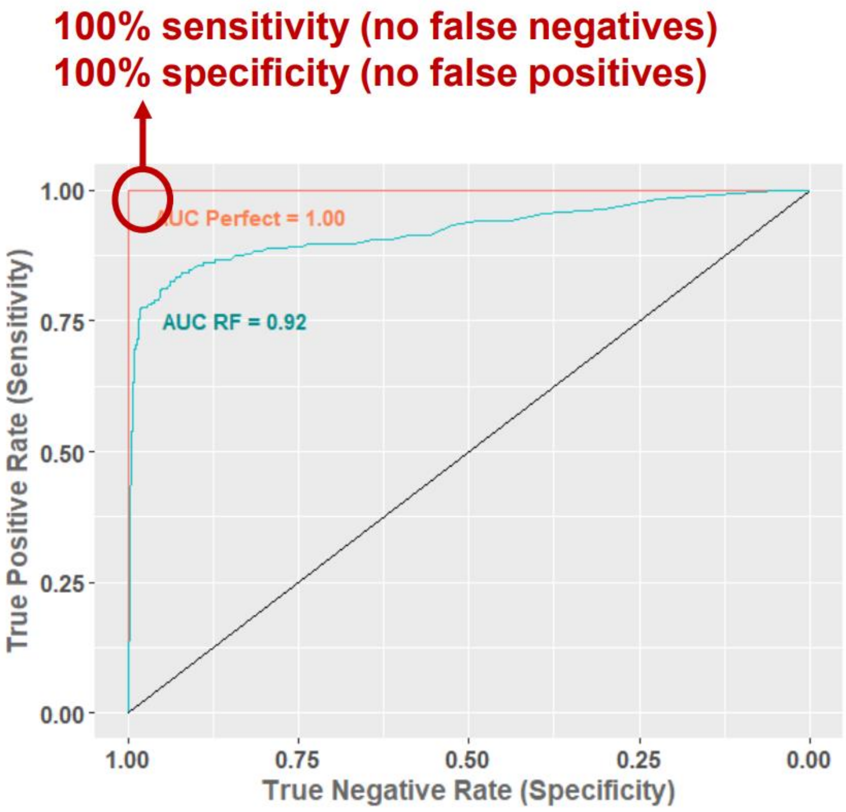


- **Precision** = TP / (TP + FP = **98,0%**
- **Recall / Coverage** = TP / (TP + FN) = **93,3%**

THX to Christopher-Johannes Schild

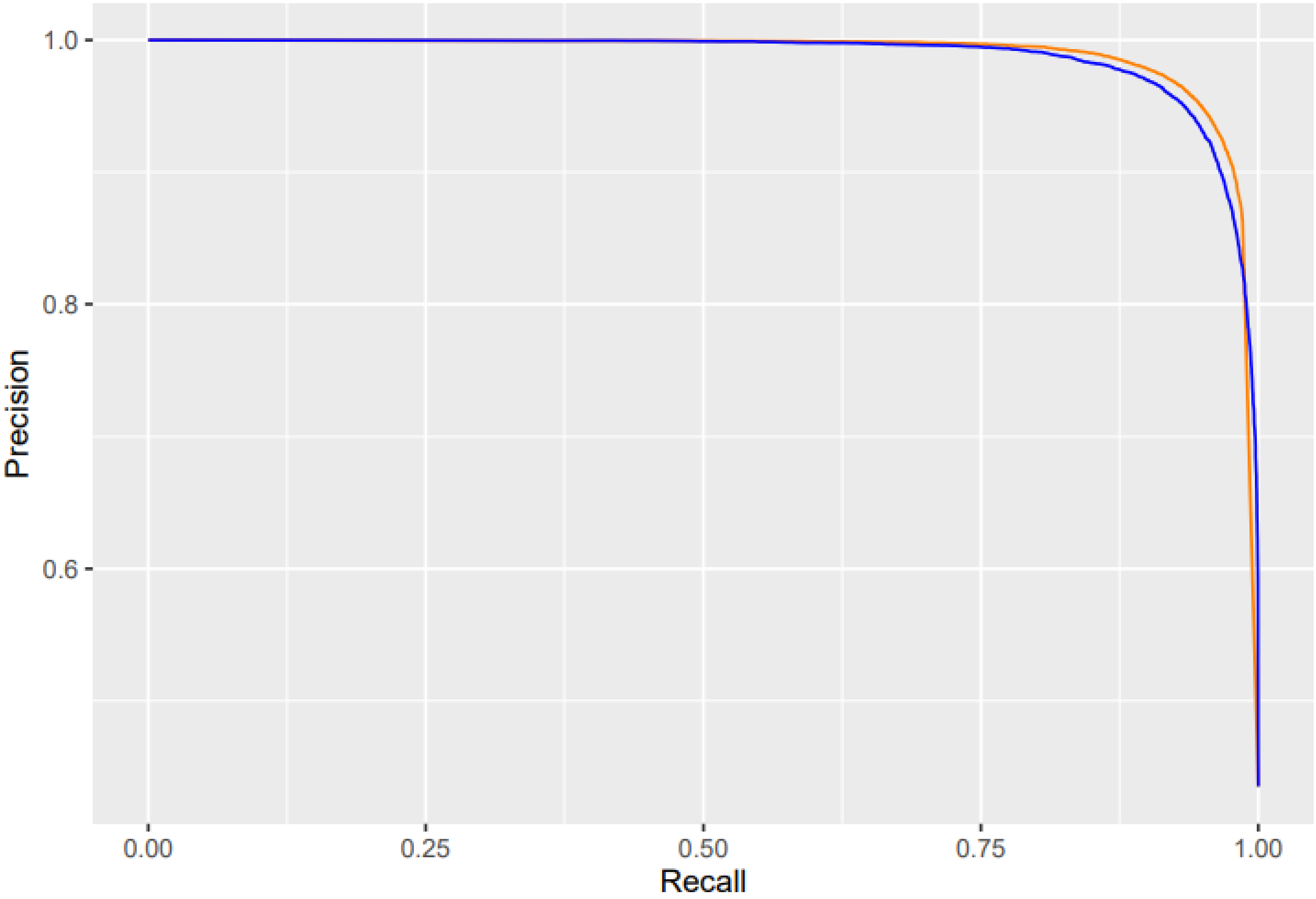
Model Evaluation Measures: ROC Curve

- Receiver Operating Curve (ROC)
  - Plots the **tradeoff** between **Sensitivity and Specificity** for different probability **thresholds**
  - **Decrease threshold**: increase **TP**, but also FP
  - **Increase threshold**: increase **TN**, but also FN
- **AUC** (Area under the Curve)
  - **Probability** that a randomly chosen **positive** case (e.g., “default”) receives from the model a **higher score** (predicted probability) than a randomly chosen **negative** case (e.g., “non-default”)



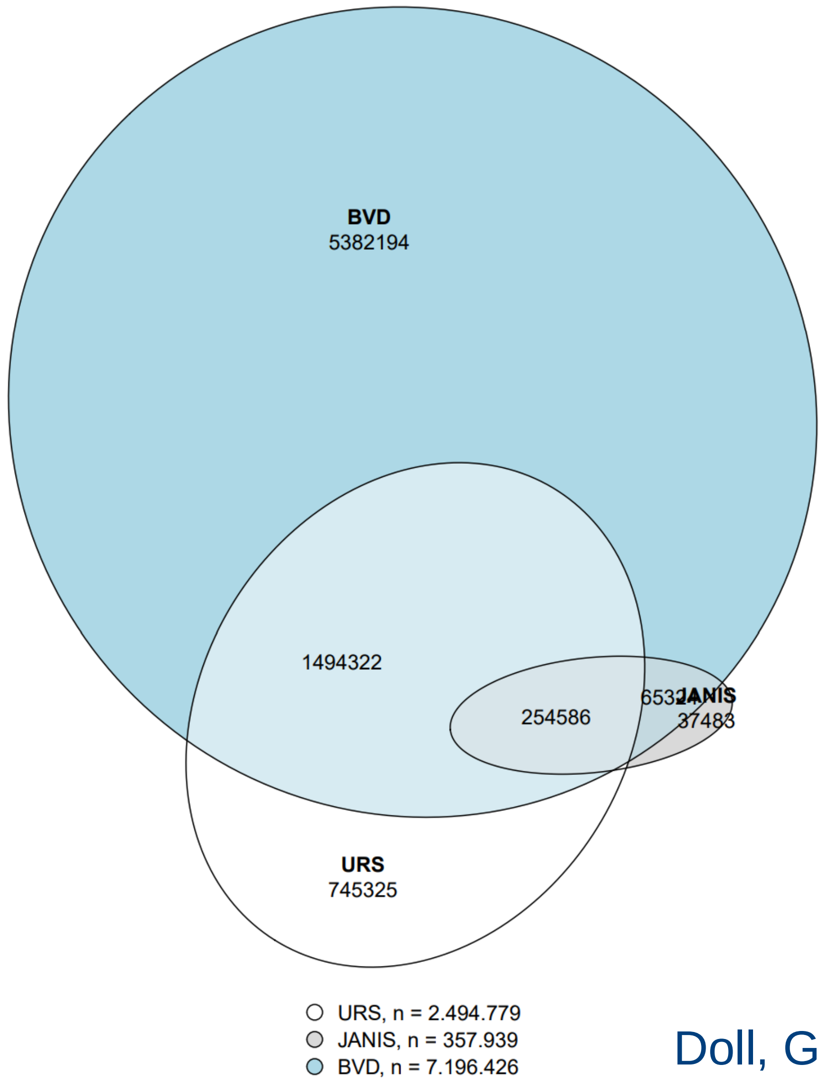
THX to Sebastian Seltmann

Precision / Recall Curves, 1st and 2nd Level Model



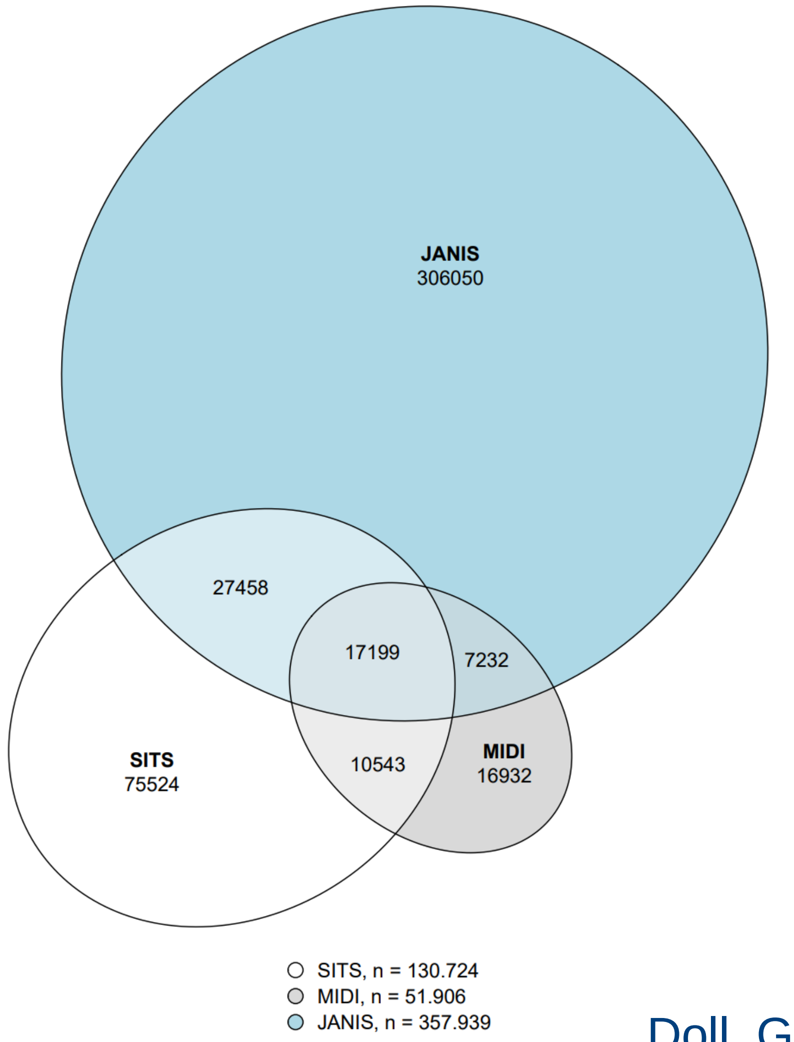
Doll, Gabor-Toth, Schild (2021): p. 16

Matching Overlap of three Master Datasets (BvD, URS and JANIS)



Doll, Gabor-Toth, Schild (2021): p. 26

Matching Overlap of three Bundesbank Datasets (1 Master, 2 Analytical)



Doll, Gabor-Toth, Schild (2021): p. 28

References

1. Hendrik Doll, Eniko Gábor-Tóth, Christopher-Johannes Schild (2021): Linking Deutsche Bundesbank Company Data, Technical Report 2021-05, RDSC, https://www.bundesbank.de/resource/blob/624432/207c774d468e82d76ec19ef6bfa1c8a7/mL/2021-05-company-data.pdf

2. Eniko Gábor-Tóth Christopher-Johannes Schild (2021): Understanding Overlaps between Different Company Data, Technical Report 2021-06, RDSC, https://www.bundesbank.de/resource/blob/624628/4a0ec7a2e5ad7bcfb9274abad8f5a885/mL/2021-06-company-data.pdf

Appendix

(Software, Literature)

Software overview

- Other (free) programs (see Appendix):
  - Big Match
  - GRLS
  - The Link King
  - Link Plus
  - FRIL
  - Open Refine
  - Relais
  - R-Paket „RecordLinkage“
  - TDGen

Freely Extensible Biomedical Record Linkage (FEBRL)

- Project “Parallel Large Scale Techniques for High-Performance record linkage“
- Australian National University (ANU), Department of Computer Science
- Peter Christen
- Project: datamining.anu.edu.au/projects/linkage.html
- Version 0.4.2, 2013
- Download: sourceforge.net/projects/febrl

FEBRL: Features

- Freely available and expandable (open source license): **Python**
- Preprocessing module
- Probabilistic record linkage
- Further classification techniques
- Different blocking techniques
- Many string similarity functions
- Geocoding
- Blindfolded/Privacy Preserving Record Linkage
- Frequency weights

MTB: Basics

- Merge ToolBox (MTB) is a Java application developed by the German RLC
- Current version: 0.742, November 2012
- Free use for academic purposes
- To be found at: http://record-linkage.de/-Downloads--software.htm
- Counseling by the German RLC

MTB: Features

- Probabilistic record linkage
- Many string similarity functions
- Several blocking techniques implemented
- Frequency weights
- 1-1 matching
- Parameter estimation using EM-algorithm
- Array-matching
- Privacy Preserving Record Linkage using Bloom Filters

MTB: Configuration via XML-file

- XML-configuration files allow replicable MTB runs.
- Particularly helpful during testing or if data files have to be divided for size-related reasons
- In the batch-mode configurations can be run successively and automatically.
- After initially creating a configuration, (copies of) the XML-file can be adapted with external editors (e.g. Emacs, Notepad++, WinEdt)

BigMatch (U.S. Census Bureau)

- Useful for matching a very large file against a moderate size file
- Outputs all records from the large file that were stored as probable matches to the same record in the moderate file
- Functions as a preprocessing step to extract smaller files from very large files
  - Smaller files can then be efficiently processed using standard linkage software
- Written in portable *C*, can be run on Linux, Windows, MacIntosh, HP machines running the VMS operating system
- Allows one to run several different blocking criteria
- Developers: William Yancey / Bill Winkler

Generalized Record Linkage System (GRLS)

- Record linkage Software from Statistics Canada
- www.statcan.gc.ca
- Probabilistic record linkage
- Costs: about $ 30.000 one time, annual fee/charge
- Software + Support
- Follow-up software G-Link does not need local DBMS any more.

The Link King

- Kevin Campbell, Washington State Division of Alcohol and Substance Abuse
- www.the-link-king.com
- Version 7.1.21, 2012

Link King: Features

- Probabilistic record linkage
- String similarity functions
- Base SAS license necessary
- Blocking variables cannot be chosen freely

Link Plus

- U.S. Department of Health and Human Services, Centers for Disease Control and Prevention (CDC)
- www.cdc.gov/cancer/npcr/tools/registryplus/lp_tech_info.htm
- Version 2.0, 2007
- Developed for the implementation of the „National Program of Cancer Registries“ (NPCR)

Link Plus: Features

- Probabilistic record linkage
- Edit-distances
- Flexible blocking
- Frequency weights
- 1-1 matching

Fine-Grained Records Integration and Linkage Tool (FRIL)

- Pawel Jurczyk, Department of Mathematics and Computer Science, Emory University
- fril.sourceforge.net
- Version 2.1.5, 2011

FRIL: Features

- Probabilistic record linkage
- String similarity functions
- Different blocking techniques (e.g. sorted neighborhood)
- Parameter estimation using EM-algorithms

Record Linkage At IStat (Relais)

- L’Istituto nazionale di statistica
- www.istat.it/it/strumenti/metodi-e-software/software/relais
- Version 2.2

Relais: Features

- Probabilisitc record linkage
- Rule-based matching
- String similarity functions
- Blocking and Sorted Neighbourhood
- 1-1 Matching
- Parameter estimation using EM-algorithms

R-Package „Record Linkage“

- Record linkage: Detecting duplicate data project; Murat Sariyar/Andreas Borg, IMBEI, Uni Mainz
- r-forge.r-project.org/projects/recordlinkage
- Version 0.4-1, 2012

Record linkage: Features

- Probabilistic record linkage
- String similarity functions
- Blocking
- Parameter estimation using EM-algorithms
- Further classification techniques (machine learning)

Test Data Generator (TDGen)

- Free software developed by the GRLC
- Download of software and English manual from record-linkage.de/-Downloads--software.htm
- Tool to take arbitrary test file and generate a garbeled version of it by introducing simulated errors.
- Flexible control over error insertion probabilities.
- Things to control:
  - Fraction of erroneous rows
  - Number of erroneous columns in each erroneous row (defined manually or by Poisson or uniform distribution)
  - Error types (typographical errors, OCR-type errors, phonetic errors by reverse-modelling their encoding rules and many more)

Literature Review: Record Linkage Overviews

- Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer.
- Elmagarmid, A., Ipeirotis, P. G., and Verykios, V. 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering* 19(1) 1–16.
- Gill, L. E. 2001. *Methods for Automatic Record Matching and Linkage and Their Use in National Statistics*. Norwich: Office of National Statistics.
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer.
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2010. Record linkage. *Wiley Interdisciplinary Reviews: Computational Statistics* 2(5) 535–543.
- Newcombe, H. B. 1988. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press.
- Winkler, W. E. 1995. Matching and record linkage. B. G. Cox et al. (ed.) *Business Survey Methods*. New York: Wiley, pp. 355–384.
- Winkler, W. E. 2004. Methods for evaluating and creating data quality. *Information Systems* 29(7) 531–550.
- Winkler, W. E. 2009. Record linkage. D. Pfeffermann and C.R. Rao (ed.) *Handbook of Statistics 29A, Sample Surveys: Design, Methods and Applications*. Amsterdam: Elsevier, pp. 351–380.

Literature Review: Articles in Reference Books

- Arasu, A. and Domingo-Ferrer, J. 2009. Record matching. L. Liu and M. T. Özsu (ed.) *Encyclopedia of Database Systems*. New York: Springer.
- Domingo-Ferrer, J. 2009. Record linkage. L. Liu and M. T. Özsu (ed.) *Encyclopedia of Database Systems*. New York: Springer.
- Fair, M. E. 2002. Record linkage. L. Breslow (ed.) *Encyclopedia of Public Health*. New York: Macmillan Reference/Gale Group.
- Judson, D. H. 2004. Computerized record linkage and statistical matching. K. Kempf-Leonard (ed.) *Encyclopedia of Social Measurement*. Amsterdam: Elsevier.

Literature Review: Introductions

- Clark, D. E. 2004. Practical introduction to record linkage for injury research. *Injury Prevention* 10(3), 186-191.
- Hassard, T. H. 1986. Writing the book of life: medical record linkage. R. J. Brook et al. (ed.) *The Fascination of Statistics*. New York: Marcel Dekker, pp. 25-46.
- Howe, G. R. 1998. Use of computerized record linkage in cohort studies. *Epidemiologic Reviews* 20(1), 112–121.
- Smith, M. E. 1984. Record linkage: present status and methodology. *Journal of Clinical Computing*13(2–3), 52–69.

Literature Review: Bibliographies and Surveys

- Elmagarmid, A., Ipeirotis, P. G., and Verykios, V. 2007. Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering* 19(1) 1–16.
- Köpcke, H. and Rahm, E. 2010. Frameworks for entity matching: a comparison. *Data & Knowledge Engineering* 69(2) 197–210.
- Machado, C. J. 2004. A literature review of record linkage procedures focusing on infant health outcomes. *Cadernos de Saúde Pública* 20(2) 362–371.
- Winkler, W. E. 2011. Record linkage references (2011Jan15).

Literature Review: Preprocessing

- Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 3: Data Pre-Processing.
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 10: Standardization and Parsing.
- Low, W. L., Lee, M. L., and Ling, T. W. 2001. A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems* 26(8) 585–606.
- Smith, M. E. 1985. Record-keeping and data preparation practices to facilitate record linkage. B. Kilss and W. Alvey (ed.) *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies: 9–10 May 1985; Arlington, VA*. Washington, DC: Department of the Treasury, Internal Revenue Service, Statistics of Income Division, pp. 321–326.
- Winkler, W. E. 1985. Preprocessing of lists and string comparison. B. Kilss and W. Alvey (ed.) *Record Linkage Techniques 1985. Proceedings of the Workshop on Exact Matching Methodologies: 9–10 May 1985; Arlington, VA*. Washington, DC: Department of the Treasury, Internal Revenue Service, Statistics of Income Division, pp. 181–187.

Literature Review: Blocking Methods

- Baxter, R., Christen, P., and Churches, T. 2003. A comparison of fast blocking methods for record linkage. CIMS Technical Report 03/139, CSIRO Mathematics, Informatics and Statistics.

- Christen, P. 2012. A survey of indexing techniques for scalable record linkage and deduplication.

- *IEEE Transactions on Knowledge and Data Engineering* 24(9) 1537–1555.

- Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 4: Indexing.

- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 12: Blocking.

Literature Review: Rule-Based Record Linkage

- Hernández, M. A. and Stolfo, S. S. 1998. Real-world data is dirty: data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2(1) 9–37.
- Low, W. L., Lee, M. L., and Ling, T. W. 2001. A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems* 26(8) 585–606.
- Whang, S. and Garcia-Molina, H. 2010. Entity resolution with evolving rules. *Proceedings of the VLDB Endowment* 3(1) 1326–1337.

Literature Review: Distance-Based Record Linkage

- Arasu, A., Chaudhuri, S., and Kaushik, R. 2008. Transformation-based framework for record matching. *Proceedings of the 24th International Conference on Data Engineering: 7–12 April 2008; Cancún, Mexico*. Los Alamitos, CA: IEEE, pp. 40–49.
- Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 5: Field and Record Comparison.
- Monge, A. E. and Elkan, C. P. 1996. The field-matching problem: algorithms and applications. E. Simoudis, J. Han, and U. Fayyad (ed.) *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining: 2–4 August 1996; Portland*. Menlo Park, CA: AAAI Press, pp. 267–270.
- Whang, S. and Garcia-Molina, H. 2010. Entity resolution with evolving rules. *Proceedings of the VLDB Endowment* 3(1) 1326–1337.

Literature Review: String Similarity Functions

- Bilenko, M. et al. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems* 18(5) 16–23.
- Christen, P. 2006. A comparison of personal name matching: techniques and practical issues. S. Tsumoto et al. (ed.) *Proceedings of the 6th IEEE International Conference on Data Mining - Workshops: 18 December 2006; Hong Kong.* Los Alamitos, CA: IEEE, pp. 290–294.
- Hall, P. A. V. and Dowling, G. R. 1980. Approximate string matching. *ACM Computing Surveys* 12(4), 381–402.
- Stephen, G. A. 1994. *String Searching Algorithms.* Singapore: World Scientific.

Lit. Review: Comparisons of String Similarity Functions

- Bilenko, M. et al. 2003. Adaptive name matching in information integration. *IEEE Intelligent Systems* 18(5) 16–23.
- Christen, P. 2006. A comparison of personal name matching: techniques and practical issues. S. Tsumoto et al. (ed.) *Proceedings of the 6th IEEE International Conference on Data Mining - Workshops: 18–22 December 2006; Hong Kong.* Los Alamitos, CA: IEEE, pp. 290–294.
- Yancey, W. E. 2005. Evaluating string comparator performance for record linkage. Technical Report RSS2005/05. Statistical Research Division, U.S. Census Bureau, Washington, DC.
- Agrawal, R. and Srikant, R. 2002. Searching with numbers. D. Lassner, D. De Roure, and A. Iyengar (ed.) *Proceedings of the 11th International World Wide Web Conference: 7–11 May 2002; Honolulu*. New York: ACM, pp. 420–431.

Literature Review: Probabilistic Record Linkage

- Fellegi, I. P. and Sunter, A. B. 1969. A theory for record linkage. *Journal of the American Statistical Association* 64(328) 1183–1210.
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 8: Record Linkage – Methodology.
- Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84(406) 414–420.

Lit. Review: Adjustment of Weights for String Similarities

- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 13: String Comparator Metrics for Typographical Error.
- Winkler, W. E. 1990. String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 354–359.
- Yancey, W. E. 2005. Evaluating string comparator performance for record linkage. Technical Report RSS2005/05. Statistical Research Division, U.S. Census Bureau, Washington, DC.

Literature Review: Frequency Weights

- Newcombe, H. B. 1988. *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford: Oxford University Press, Chapter 9: Converting 'global' odds to 'specific' odds.
- Winkler, W. E. 1989. Frequency-based matching in the Felligi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 778– 783.
- Yancey, W. E. 2000. Frequency-dependent probability measures for record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, S. 752–757.

Literature Review: Parameter Estimation

- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. Data Quality and Record Linkage Techniques. New York: Springer, Chapter 9: Estimating the Parameters of the Fellegi–Sunter Record Linkage Model.
- Winkler, W. E. 1993. Improved decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 274– 279.
- Winkler, W. E. 1988. Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, pp. 667–671.

Lit. Review: Threshold Selection/Error Rate Selection

- Belin, T. R. and Rubin, D. B. 1995. A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association* 90(430) 694–707.
- Bishop, G. and Khoo, J. 2006. Methodology of evaluating the quality of probabilistic linking. *Proceedings of Statistics Canada Symposium. Methodological Issues in Measuring Population Health: 1–3 November 2006; Gatineau, Canada*. Ottawa: Statistics Canada.
- Winkler, W. E. 1995. Matching and record linkage. B. G. Cox et al. (ed.) *Business Survey Methods*. New York: Wiley, pp. 355–384, Section 20.6: Estimation of Error Rates and Adjustment for Matching Error.
- Winkler, W. E. 2006. Automatically estimating record linkage false match rates. *Proceedings of the Survey Research Methods Section.* American Statistical Association, pp. 3863–3870.

Literature Review: 1-to-1 Assignment

- Jaro, M. A. 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association* 84(406) 414–420.
- Winkler, W. E. 1994. Advanced methods for record linkage. Technical Report RR94/05. Statistical Research Division, U.S. Census Bureau, Washington, DC.

Lit. Review: Bias due to Imperfect Record Linkage

- Chesher, A. and Nesheim, L. 2006. Review of the literature on the statistical properties of linked datasets. Technical Report 3. Department of Trade and Industry, London.
- Lahiri, P. and Larsen, M. D. 2005. Regression analysis with linked data. *Journal of the American Statistical Association* 100(469) 222–230.
- Ridder, G. and Moffitt, R. 2007. The econometrics of data combination. J. J. Heckman and E. E. Leamer (ed.) *Handbook of Econometrics.* Amsterdam: Elsevier, pp. 5469–5547.

Literature Review: Record Linkage Software

- ESSnet Statistical Methodology Project on Integration of Survey & Administrative Data 2011. Deliverable WP3: software tools for integration methodologies.
- Herzog, T. N., Scheuren, F. J., and Winkler, W. E. 2007. *Data Quality and Record Linkage Techniques*. New York: Springer, Chapter 19: Review of Record Linkage Software.

Lit. Review: Privacy-Preserving Record Linkage

- Christen, P. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Heidelberg: Springer, Chapter 8: Privacy Aspects of Data Matching.
- Hall, R. and Fienberg, S. E. 2010. Privacy-preserving record linkage. J. Domingo-Ferrer and E. Magkos (ed.) *Proceedings of the 2010 International Conference on Privacy in Statistical Databases: 22–24 September 2010; Corfu, Greece*. Berlin: Springer, Berlin, pp. 269–283.
- Karakasidis, A. and Verykios, V. S. 2011. Advances in Privacy Preserving Record Linkage. T. Matsuo and T. Fujimoto (ed.) *E-Activity and Intelligent Web Construction: Effects of Social Design*. Hershey, PA: Information Science Reference, Hershey, pp. 22–34.