Data Quality

Object Identification - Approximate Joins

Hans - J. Lenz, Freie Universität Berlin

September 2007

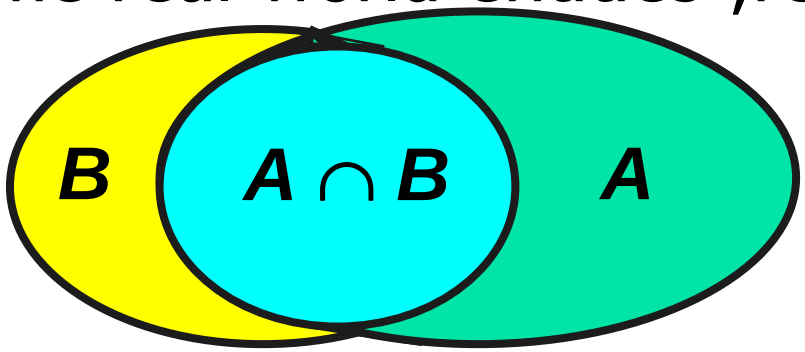
Agenda

- Introduction
- Data Quality
- Classification
- Pre-selection
- Evaluation
- Summary

Source: M. Neiling, H.-J. Lenz, Lecture on „Methods for Administrative Record Census“, Deutschen Statistische Gesellschaft Münster, May 20th, 2005

Object Identification in Databases

- Problem:
  Which elements in given databases refer to the same real-world entities[1],respectively?



- Example:
  Administrative Record Census

1) if global, consistent identifiers are missing

State of the Art

- Two independent development directions:
  - Record Linkage (since the 50ies)
    - Well-proven, robust method
    - Statistical efficiency dominating
    - Mostly used for personal data, e.g. patient information
  - Duplicate-Detection in Databases (since the 80ies)
    - Several methods, e.g. Sorted Neighborhood Method (SNM), Machine Learning,…
    - Performance dominating (SNM, Blocking, Clustering)
- Trend towards convergence since about five years

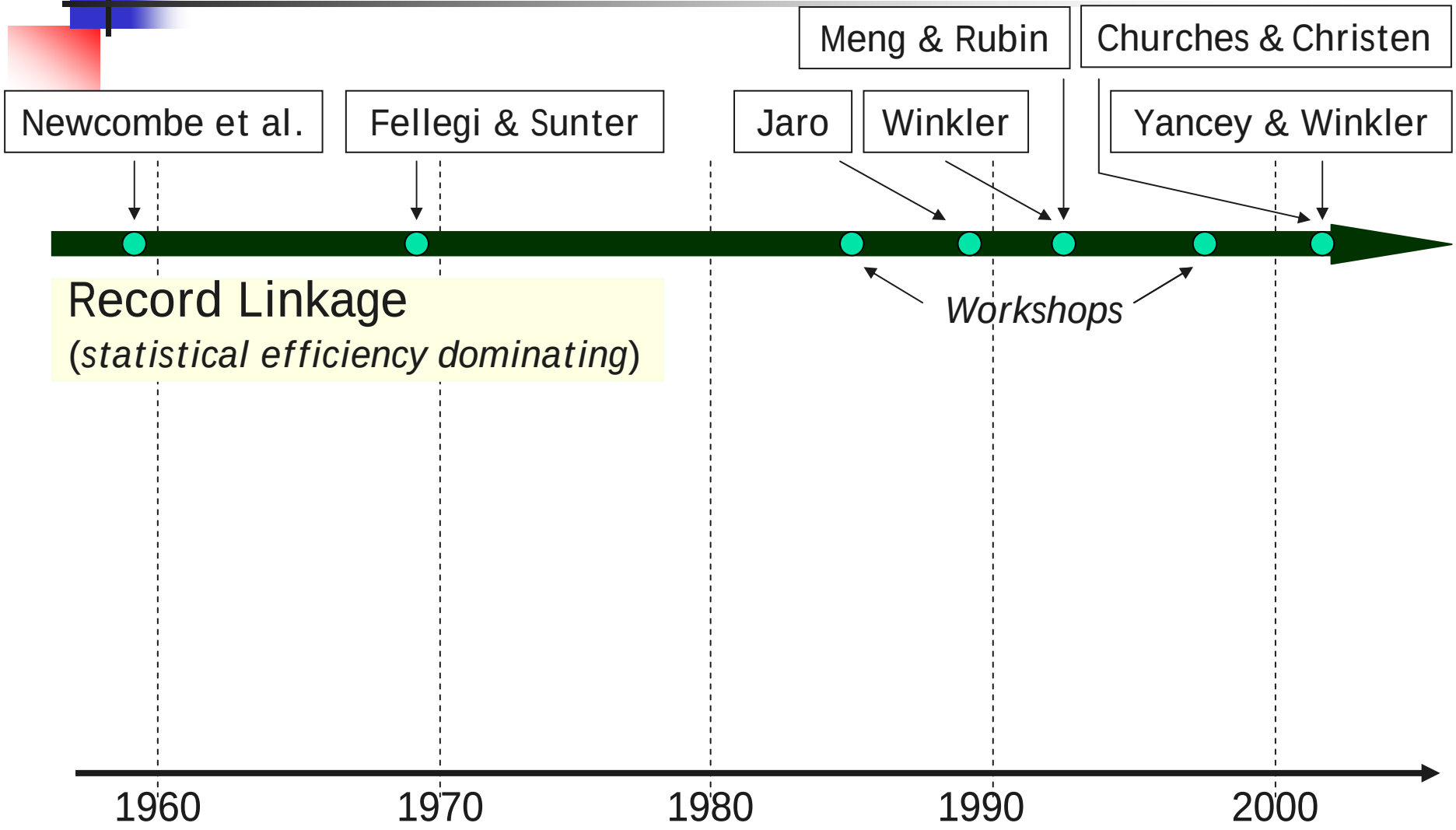
Algorithms

Find Partioning & Compare pairs only within adjacent partition w.r.t. similarity or distance measures, cf. classification

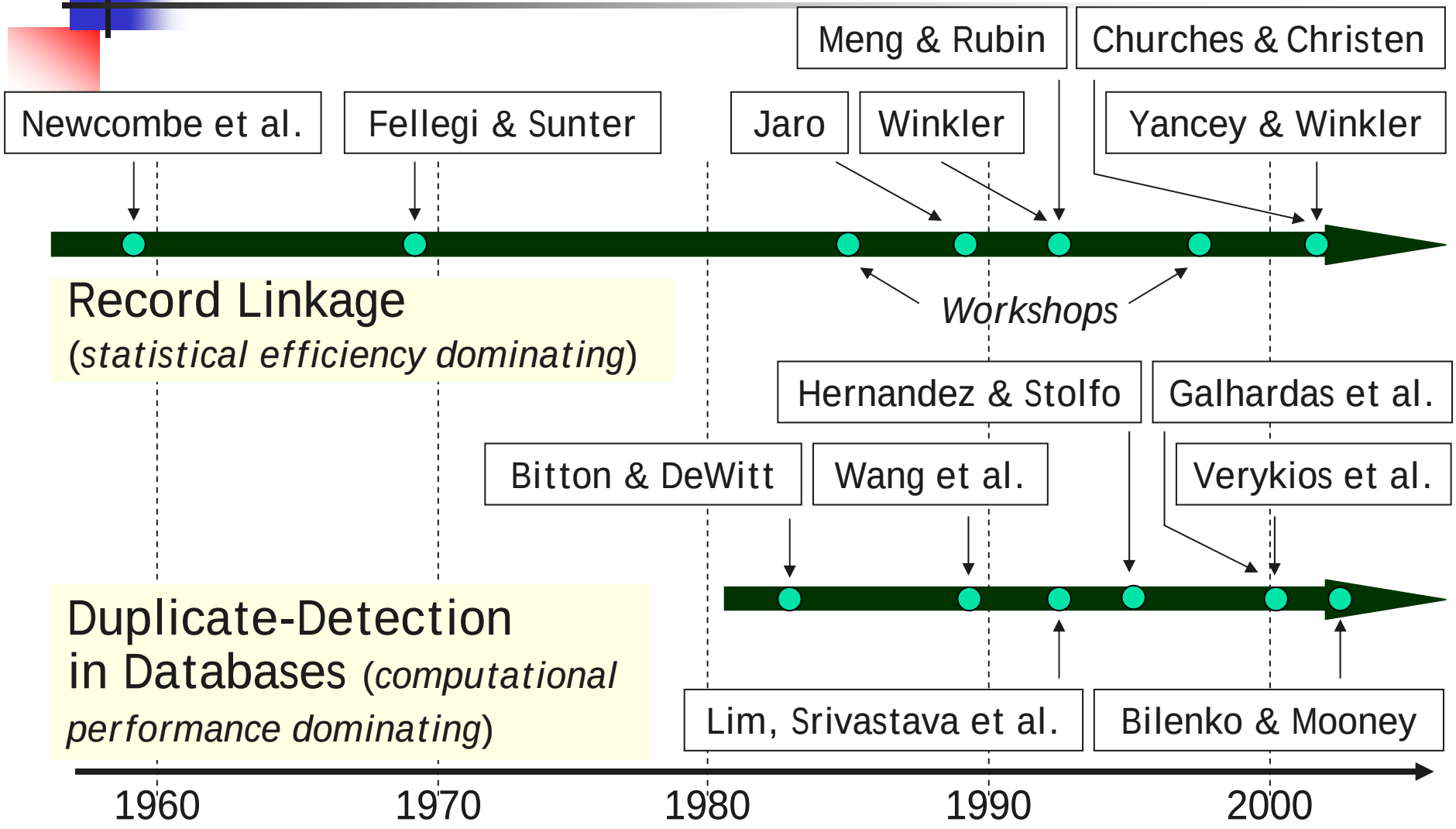
Sorted Neighbourhood Method (SNM) and variants based on sorting, similarity and merging, cf. Hernandez and Stolfo (1998)

Trade-off : Accuracy and Computing Cost

Historical Development



Historical Development



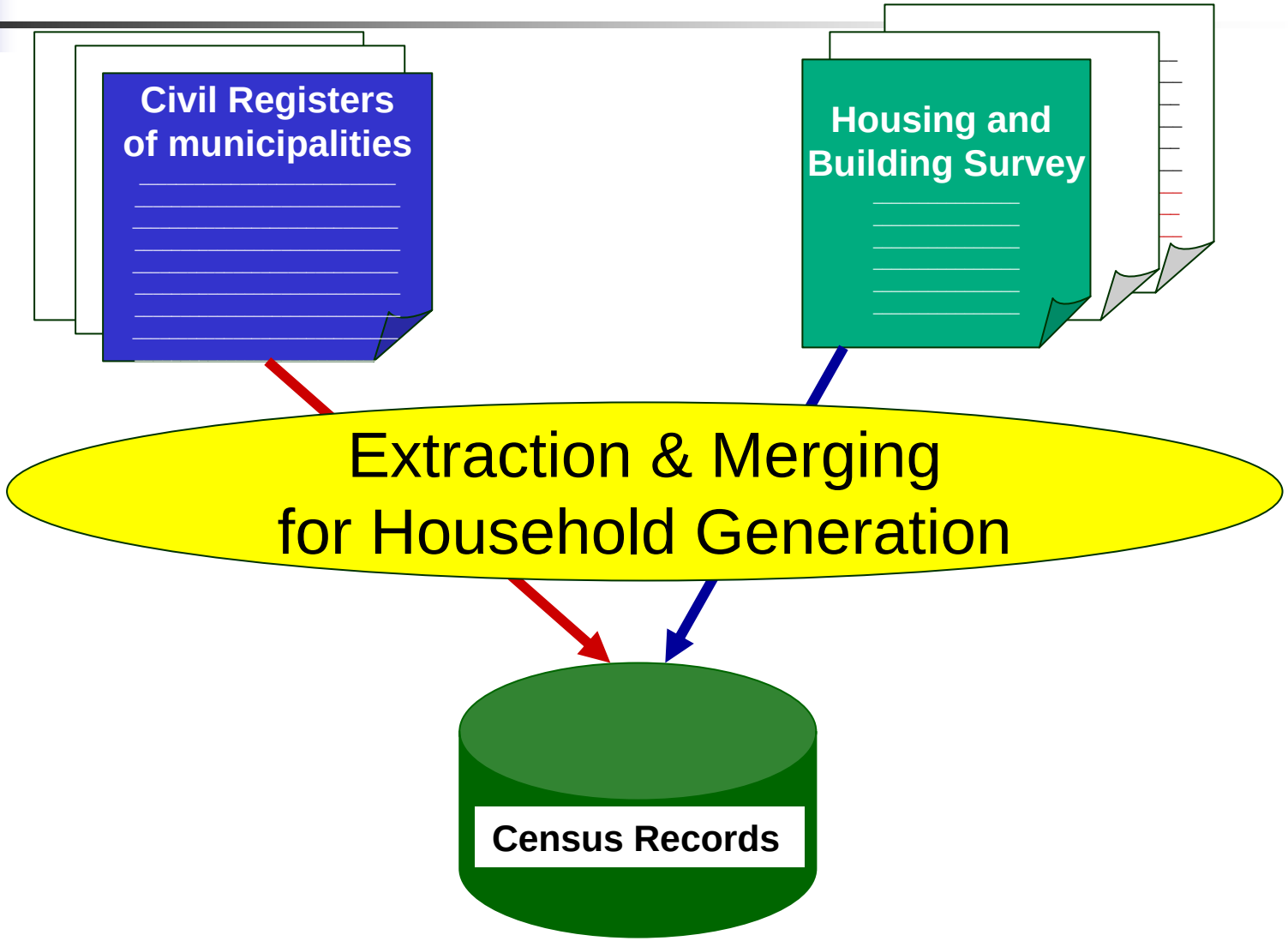
Administrative Record Census

- ARC
- Multiple Databases:

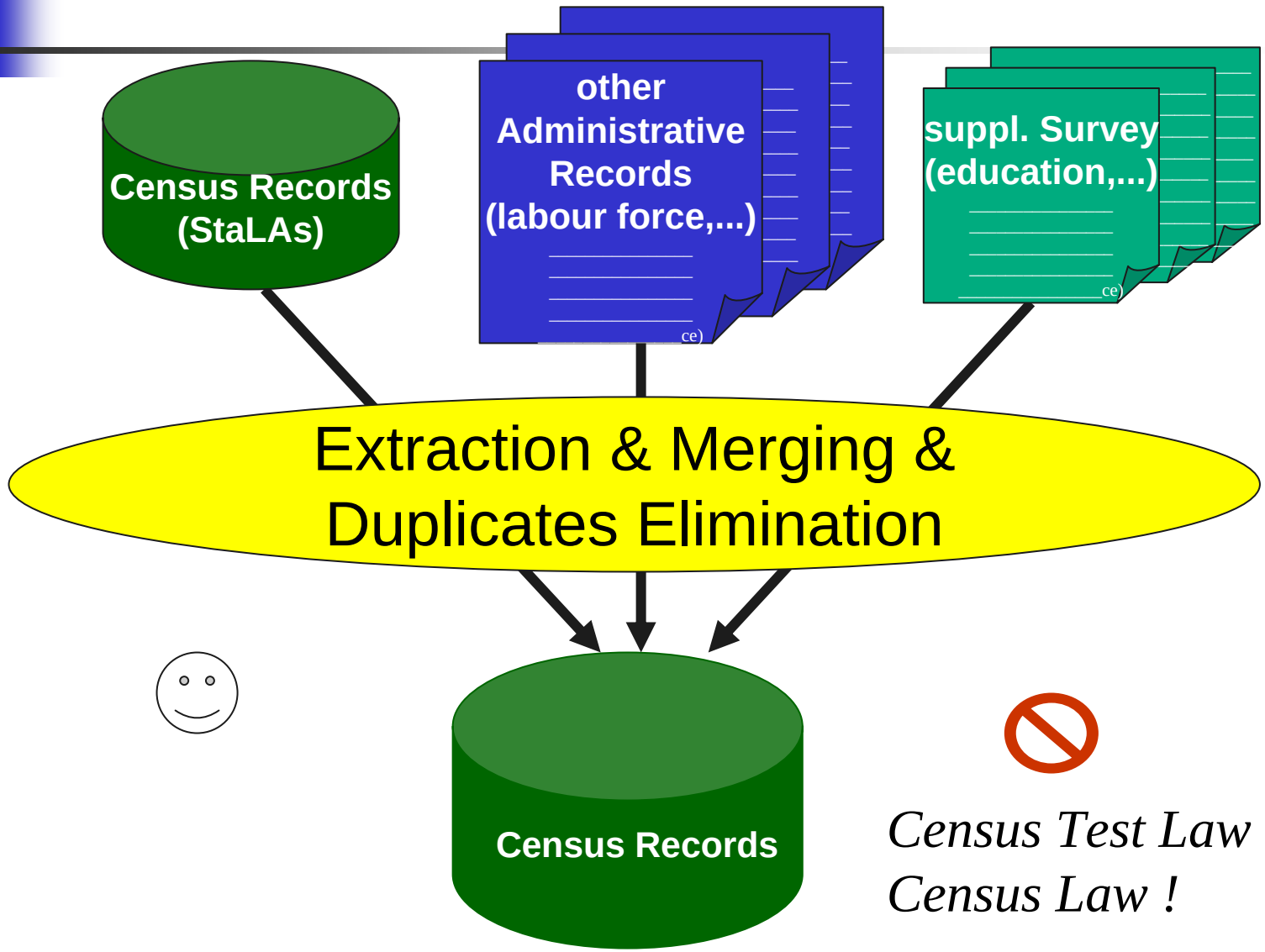
  Civil Register+Housing Register+Bureau of Labour File +...

  - No global primary key stored in *German* registers
  - errors: misprints, obsolete records, duplicates,...
  - null values: missing values
  - missing entities e.g. illegal residents

ARC in Germany – StaLAs view



ARC in Germany – NSI view



Census Test Law and Census Law !

Data Quality – Exa

errors

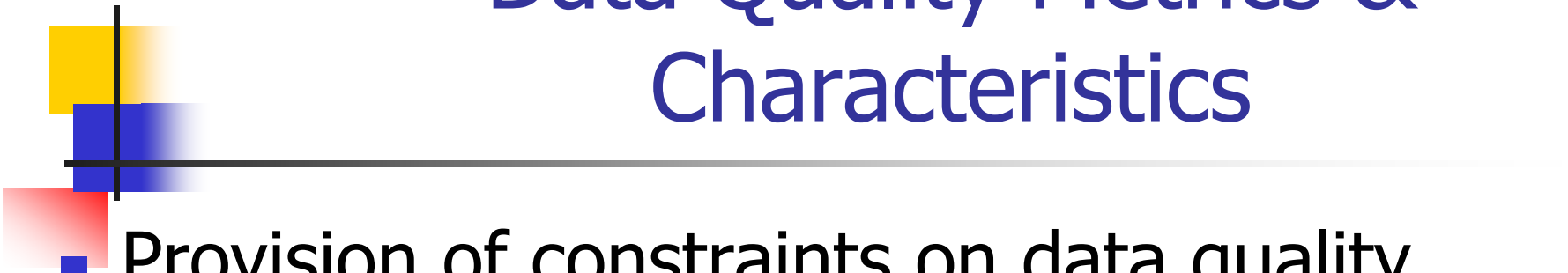
No.	ISBN	Title	Name	Year	Pages
1	0-201-54329-X	An introduction to database systems	Date	**1997**	839
2	0-201-54329-X	An introduction to database systems	Date	**1995**	839
...		...			
		An introduction to spatial database systems	**Gueting**	1994	13
10		An introduction to spatial database systems	**Güting**	1994	13
...		...			
26	0-**210**-14456-5	An introduction to database systems	Date	1977	536
27	0-**201**-14456-5	An Introduction to Database Systems	Date	1977	
...		...			
30		An introduction to database systems	Date		

mistypings

variations

missing values

Data Quality Metrics & Characteristics



- Provision of constraints on data quality
  - Stated by domain experts, or
  - Estimated from samples
- Analysis should be the starting point!
  - Basis for feature selection
  - Main information for pre-selection

Data Quality Metrics & Characteristics

- Provision of constraints on data quality

$$C_k(A_i, Y) \lozenge v, \lozenge \in \{<, \leq, =, \geq, >\}$$

  - Stated by domain experts, or
  - Estimated from samples

- Analysis should be the starting point!
  - Basis for feature selection
  - Main information for pre-selection

Data Quality: Semantic Constraints

Proportion of records with missing values:

$$\text{nulls}(A_i,Y) := \boldsymbol{P}\,(\text{a} \in A_i \subset A \mid \exists Y_i \in Y\colon Y_i(a)=NULL),$$

e.g. nulls(*Addresses*,*BirthDate*) = .07 *)

$a \equiv b$: a and b are duplicates

Number of duplicates between A_i , $A_k \subset A$:

$$\text{duplicates}(A_i, A_k) := |\,\{a \in A_i \mid \exists b \in A_k : a \equiv b \wedge \text{a.ID} < \text{b.ID}\}\,|$$

A:	database table
$A_i \subset A$:	Selection from A
Y:	Attribut set
| ...|:	Set cardinality
v:	value

*)estimated from a sample of 250.000 addresses

Characteristics: Semantic Keys

- Y is a **semantic key, iff**

$$Y(a) = Y(b) \Longleftrightarrow a \equiv b.$$

$a \equiv b$: a and b are duplicates

- Y is a **semantic diff-key, iff**

$$\text{dist}\big(Y(a), Y(b)\big) \geq \Delta \Longrightarrow a \not\equiv b$$

- Y is a p-**approximate key, iff**

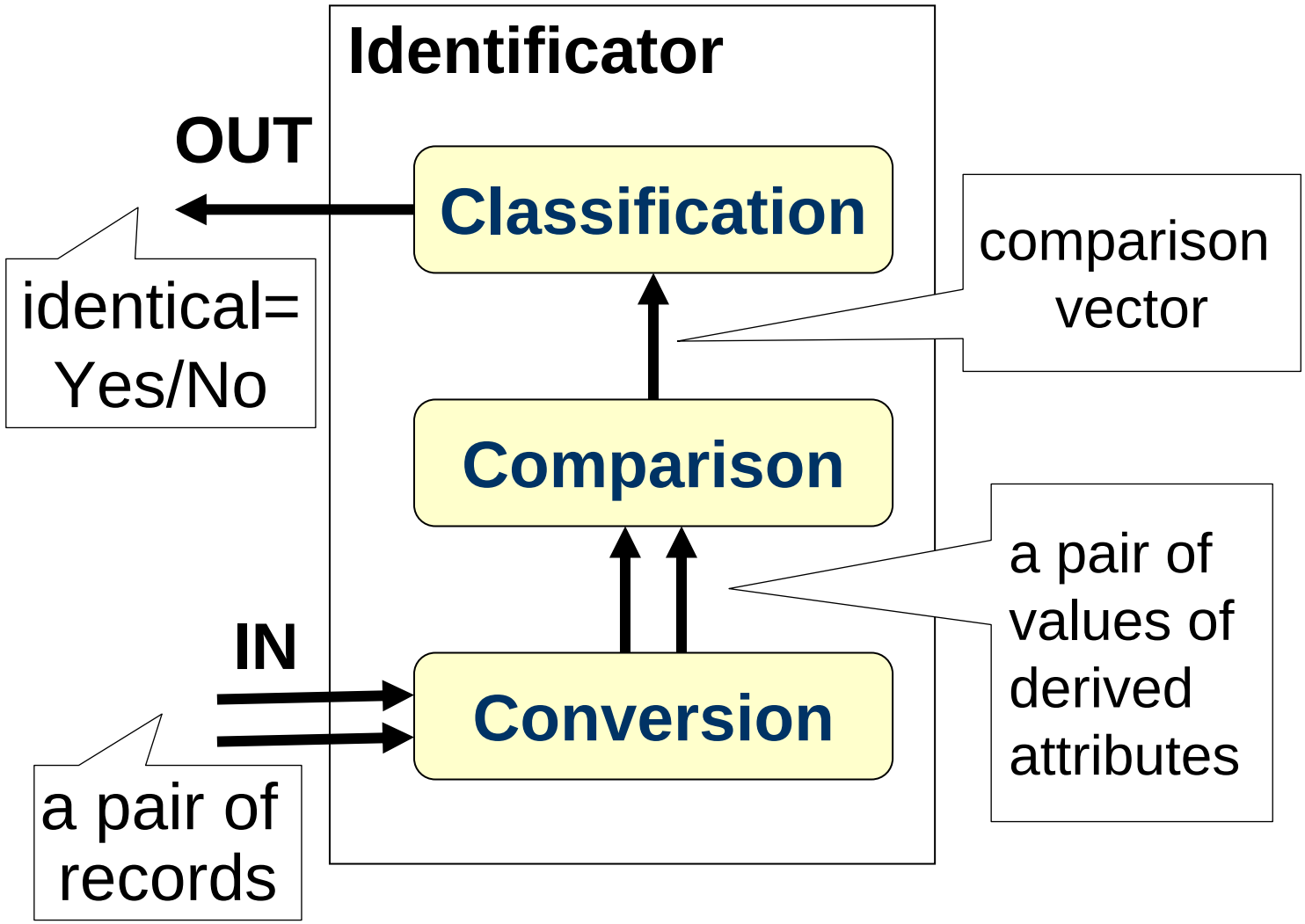
$$\text{accuracy}(Y) := \boldsymbol{P}\,(Y(a) = Y(b) \mid a \equiv b) \geq p \text{ and}$$
$$\text{confidence}(Y) := \boldsymbol{P}\,(\, a \equiv b \mid Y(a) = Y(b)\,) \geq p$$

Three-Step Identification Procedure

1. Conversion
   - Derive identifying information shared by the sources
2. Comparison
   - Apply sophisticated comparison to pairs of records
3. Classification
   - Use decision rule (e.g. induced from samples)

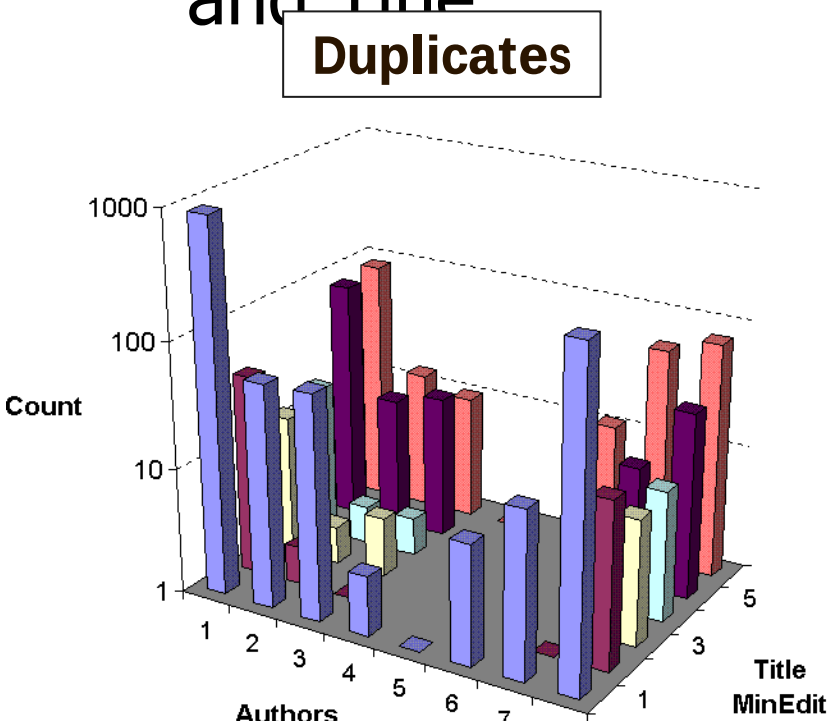
$\Rightarrow$ Apply the whole process to a pre-selection of pairs only!

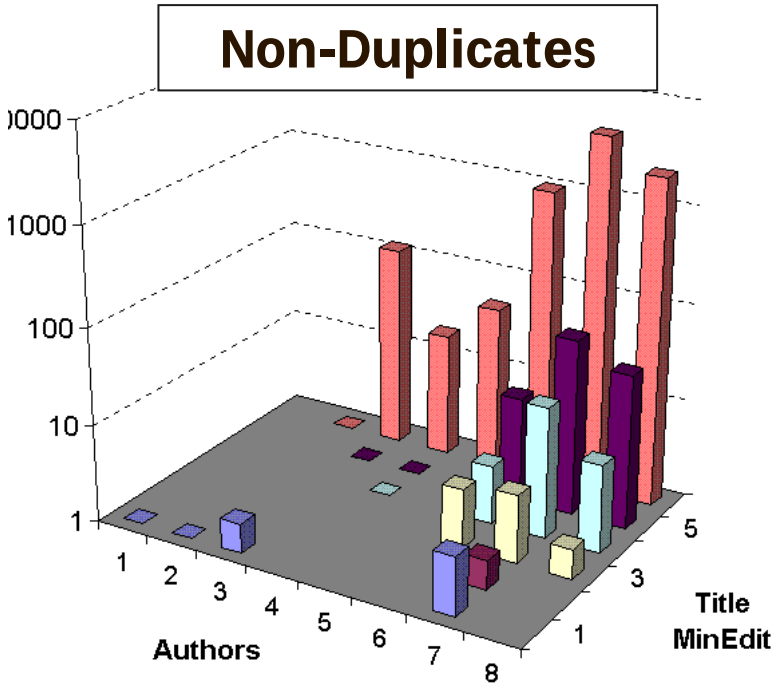
Object Identification



Comparison – Separability Problem

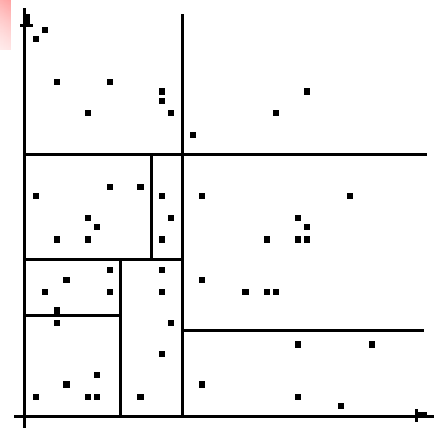
- Comparison shall separate duplicates from non-duplicates
  - Example: Separation of book records by Author and Title



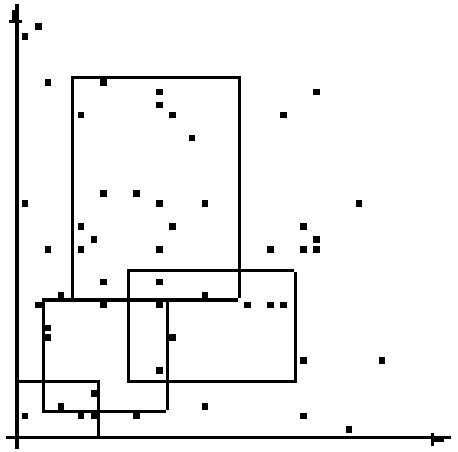


Note: the z-axes are logarithmic scaled

Object Identification: Classification I



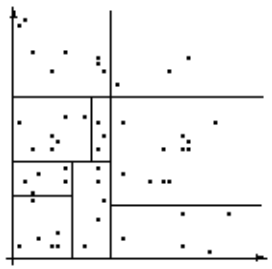
- Decision Tree Induction
  - Divide and cover the comparison space

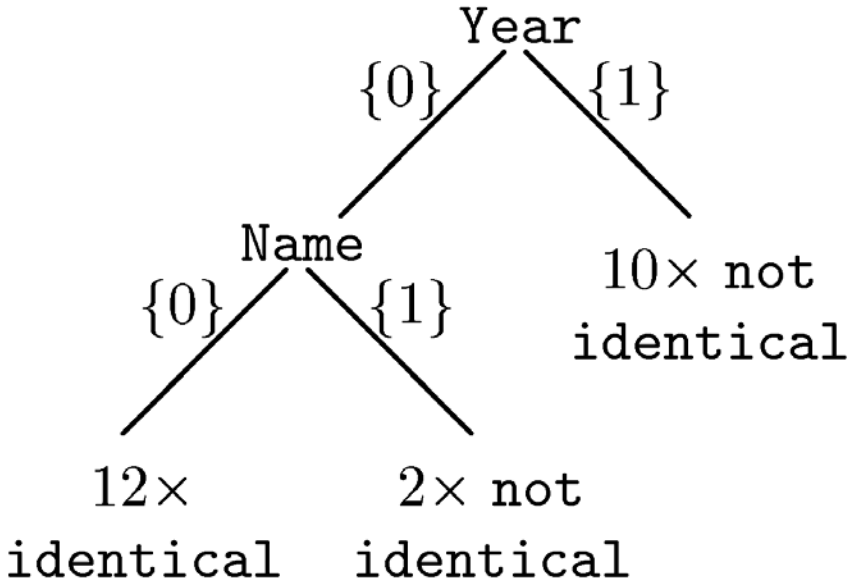


- Association Rule Mining
  - Induction of a classification from a set of possibly contradicting association rules

Classification I:
Decision Tree

- Input: sampled data (here: 24 pairs)





- Output: Partially ordered set of rules,

if($year_1$=$year_2$)**and**(name1=name2)
then identical($book_1$,$book_2$)

Breiman et al.(1984), Quinlan (1986)

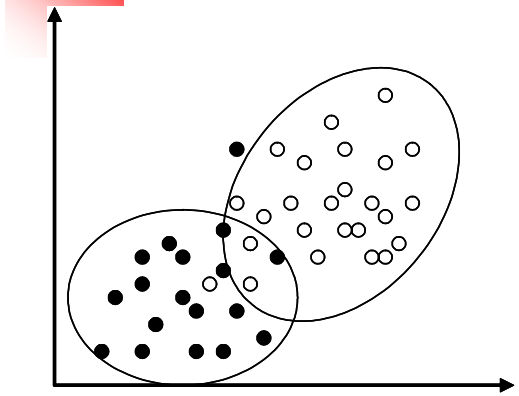
Classification I:
Association Rules

- Input: sampled data (here: 24 pairs)
- Output: set of association rules, e.g.

(R1) **if** both Pages and Year agree
then records refer to the same book
with **confidence** 100% and **support** 58%

Agrawal et al. (1993)

Classification II

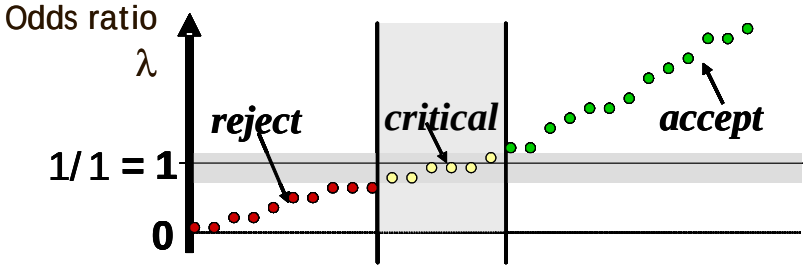


- Proper Bayesian Classifier
  - Estimation of posterior distribution (AutoClass, cf. Cheeseman et al. (1988))

- Record Linkage
  - Based on Likelihood Ratio Test

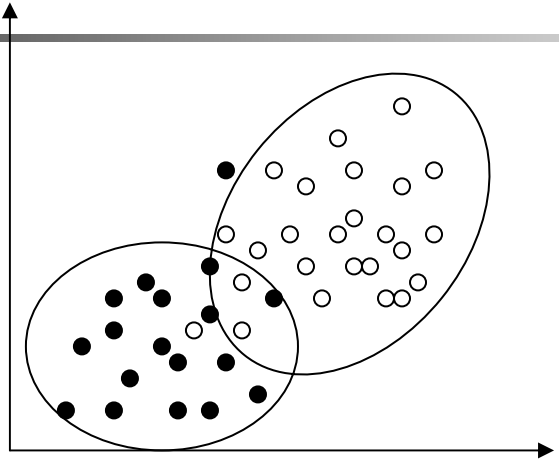
multinomial distributions

$$\lambda(a,b) = \frac{P\{\, f(a,b) \mid (a,b) \text{ is matched} \,\}}{P\{ f(a,b) \mid (a,b) \text{ is not matched} \,\}}$$



  - Estimation of a log-linear model (using EM-algorithm)

Classification II:
Proper Bayesian Classifier



C finite set of classes c

$\pi(c)$ prior distribution on C

$L_x(c)$= Likelihood for c given data x

$P(c \mid x) \propto L_x(c)\ \pi(c)$ posterior distribution on C

select class $c^* = \arg\max P(c \mid x)$

* AutoClass Cheeseman et al. (1988)

Classification II

Record Linkage

- Record Linkage (+ independence assumption)

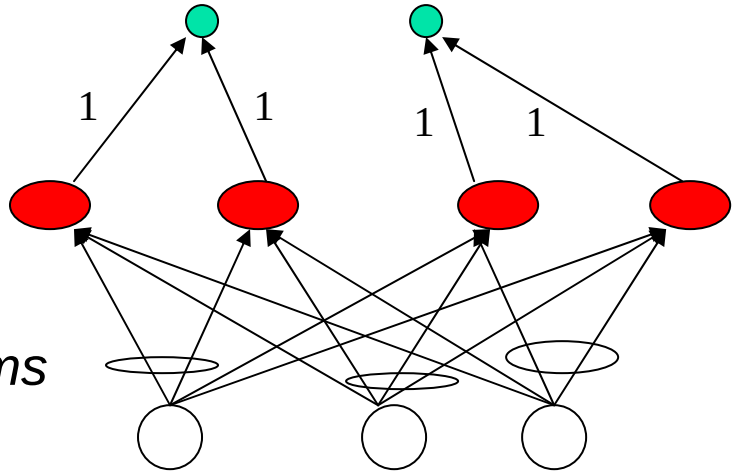
comparison values	ratio λ_i for the i-th attribute				
	ISBN	Title	Name	Year	Pages
0	9/2	10/6	12/6	12/2	7/4
1	1/2	2/4	0/6	0/10	5/8
2	2/8	0/2			

Conditional independence assumption

- λ can be calculated for each comparison case, e.g.
  If only ISBN and Name disagree, we get
  $\lambda(2,0,1,0,0) = \frac{2}{8} \cdot \frac{10}{6} \cdot \frac{0}{6} \cdot \frac{12}{2} \cdot \frac{7}{4} = 0 < 1,$
  class *'not identical'*.

Classification III:
Neuro-Fuzzy Classifier

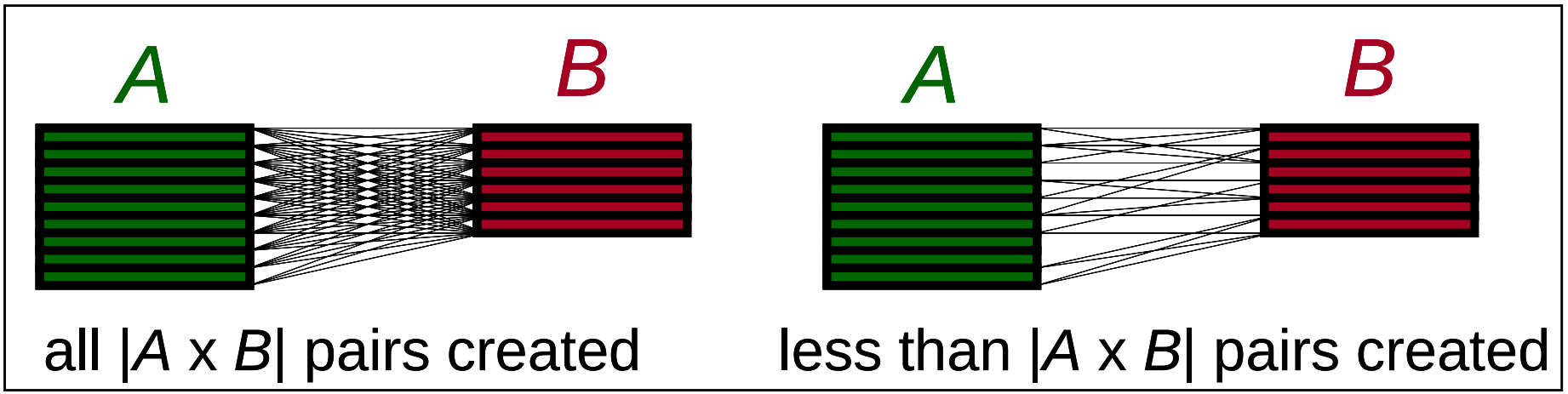
- NEFCLASS (*Nauck, Kruse (1995, 1996))
- 3-Layer Fuzzy Perceptron (***X,R,C***) with
  - input layer ***X***,
  - hidden (linguistic rules) layer ***R***,
  - output (class) layer ***C***



Linguistic terms

Pre-selection of Pairs

- Avoid to build all pairs of records – **O(n²)!**



- Goal: Improve Performance
  - Main cost: loading of records
  - Scalable solution necessary

Pre-selection of Pairs: Example

- Three relational selectors
- Different combinations $s = s_i \cup s_j$ or $s = s_i \cap s_j$
- Intersection (Example from apartment data):

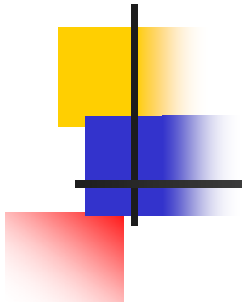
```
SELECT A.*, B.*
  FROM A AS A, A AS B
  WHERE A.District = B.District
    AND A.Size >= B.Size - 1 AND A.Size <= B.Size + 1
    AND A.Rooms >= B.Rooms - 0.5 AND A.Rooms <= B.Rooms + 0.5
```

Selector s_1

Selector s_2

Selector s_3

- Efficient processing by means of "Blocking"



Pre-selection of Pairs

- Pre-selection: Intersection/union of selectors,
  - AND/OR for relational selectors
  - Maximum-metric for metric selectors
  - Sort-Merge in general (e.g. for inverted lists)

  $\Rightarrow$ Parallel execution possible
- Optimization Point of View
  - The more selective the more undetectable duplicates (α-error increases)
  - The less selective the more processing costs

  $\Rightarrow$ Greedy optimization (branch & bound)

Sampling of Pairs

- Samples have to be chosen from pre-selection
- Stratified sampling necessary
  - Strata for selectors
  - Stratum for additional duplicates
  - Possibly, an extra stratum for random pairs
- Selectivity of pre-selection determines
  - Proportion of duplicates in the pre-selection sample
  - Offset α-error rate (undetectable duplicates)

Stratified Sampling of Pairs

A: record table, N: Sample Size

SAME $\subset$ A x A: contains the duplicate pairs in A

N_0: Number of duplicates $N_0 < N$, $N_0 \leq |SAME|$

s(.): Pre-selection $s(A \times A) \subset A \times$

Select a $\in$ A randomly

IF $s(\{a\} \times A) \neq \emptyset$ THEN

Randomly select (a,b) $\in s(\{a\} \times A)$

IF (a,b) $\notin$ P THEN P := P $\cup$ {(a,b)}

UNTIL size(P\SAME) = $N - N_0$ OR

WHILE size(P) < N

Randomly select (a,b) $\in$ SAME

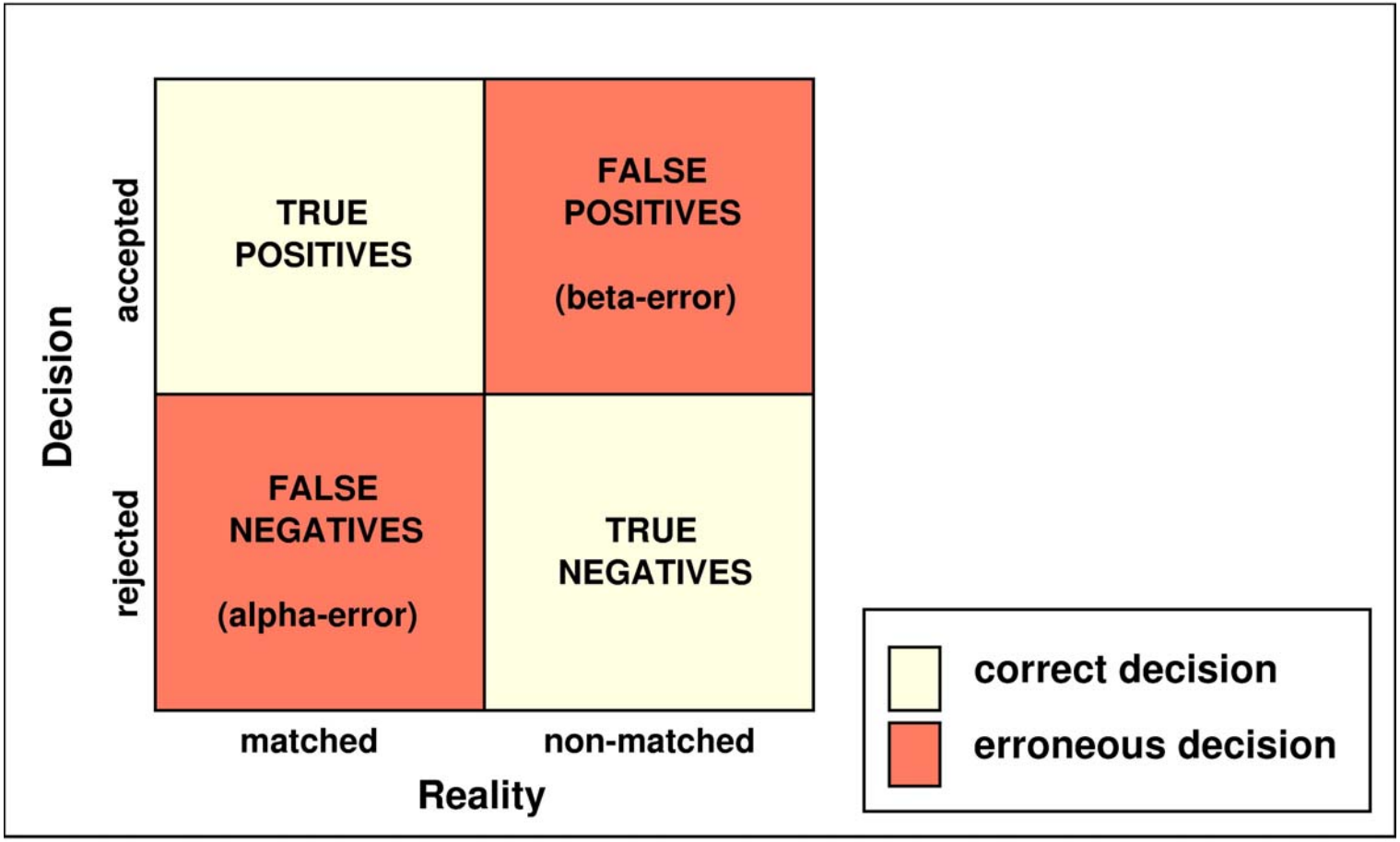
IF (a,b) $\notin$ P THEN P := P $\cup$ {(a,b)}

random pairs from pre-selection

random duplicates

***) strata for selectors are used implicitly

Correctness Assessment for Classifiers



Precision=true pos/accepted; recall=true pos/matched;
Harmonic mean: F = 2 prec x rec/(prec+rec)

Evaluation of Classification Methods

- 3+2 methods tested on a three datasets benchmark
  - Address data, apartment ads, and bibliographic data
  - Different parameters were set for classification models
  - Three sample sizes chosen (12 samples of pairs for each)
  - Samples were split into Learn- & Test-samples

- Address data:
  - 250.000 records, name, address, birth date information
  - Pre-selection chosen: Matching of one Phonetic Code (according to the *'Kölner Phonetic Code'*)
  - 12 attributes derived, up to 17 comparison functions considered, e.g. Minimum-Edit- and Bigram-Distances

Address Data – Comparing Attributes

Nr.	Attributname	Vergleichsfunktion	#Werte	Skala
1	FnameEdit	fctDiscreteMinEditDistance	7	ordinal
2	LnameEdit	fctDiscreteMinEditDistance	7	ordinal
3	ZIP	fctYesNo	2	nominal
4	LNameBigrams	fctPercentageOfMatchingBiGrams	4	ordinal
5	FNameBigrams	1−fctPercentageOfMatchingBiGrams	4	ordinal
6	LNameKoeln	fctHasEqualPhoneticCode	3	nominal
7	FNameKoeln	fctHasEqualPhoneticCode	3	nominal
8	Town	fctYesNoNull	3	nominal
9	Street	fctDiscreteMinEditDistance	7	ordinal
10	HouseNumber	fctHasNonemptyIntersection	3	nominal
11	Year	fctYesNoNull	3	nominal
12	Month	fctYesNoNull	3	nominal
13	Day	fctYesNoNull	3	nominal
14	BirthDate	fctYesNoNull	3	nominal
15	Sex	fctYesNo	2	nominal
16	FullNameBigrams	1−fctPercentageOfMatchingBiGrams	5	ordinal
17	FullNameEdit	fctDiscreteMinEditDistance	7	ordinal

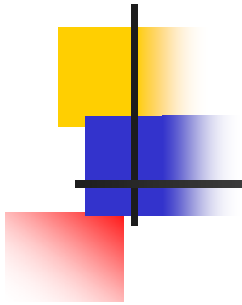
Correctness Results

Address Data

error-prob Classifier*	α %	β %	Hmean (α,β) %
AC: AssoClass	1,1	0,2	0,34
DT: Decision Tree	0,7	0,3	0,42
RL: Record Linkage	1,5	0,3	0,50
BC: AutoClass	1,6	0,1	0,19
NF: NEFClass (Neuro-Fuzzy)	6,6	4,7	5,49

*) the best classification model was selected for each method;

NF is over-fitted



Database „Appartments“

- Appartments
  - N=9.842; n_{Doub}≈2.200; 13 attributes
  - 12 test samples with 10.000 pairs each; n_{Doub}=12%
- Adresses
  - N=250.000; n_{Doub}≈52.000; 13 attributes
  - 12 test samples with 10.000 pairs each; n_{Doub}=21%
- Library
  - N= 10. 000; n_{Doub}≈1.825
  - 12 test samples with 10.000 pairs each

Correctness Results

Appartments

error-prob Classifier	α %	β %	Hmean (α,β) %
AssoClass	0,6	0,1	0,17
Decision Tree	0,1	0,1	0,10
Record Linkage	0,5	0,1	0,17
AutoClass	0,4	0,3	0,34
NEFClass	7,2	4,7	5,69

the best classification model was selected for each method; NF is over-fitted

Database „Library“

- Appartments
  - N=9.842; $n_{Doub} \approx 2.200$; 13 attributes
  - 12 test samples with 10.000 pairs each; $n_{Doub}=12\%$
- Adresses
  - N=250.000; $n_{Doub} \approx 52.000$; 13 attributes
  - 12 test samples with 10.000 pairs each; $n_{Doub}=21\%$
- Library
  - N= 10. 000; $n_{Doub} \approx 1.825$
  - 12 test samples with 10.000 pairs each

Correctness Results

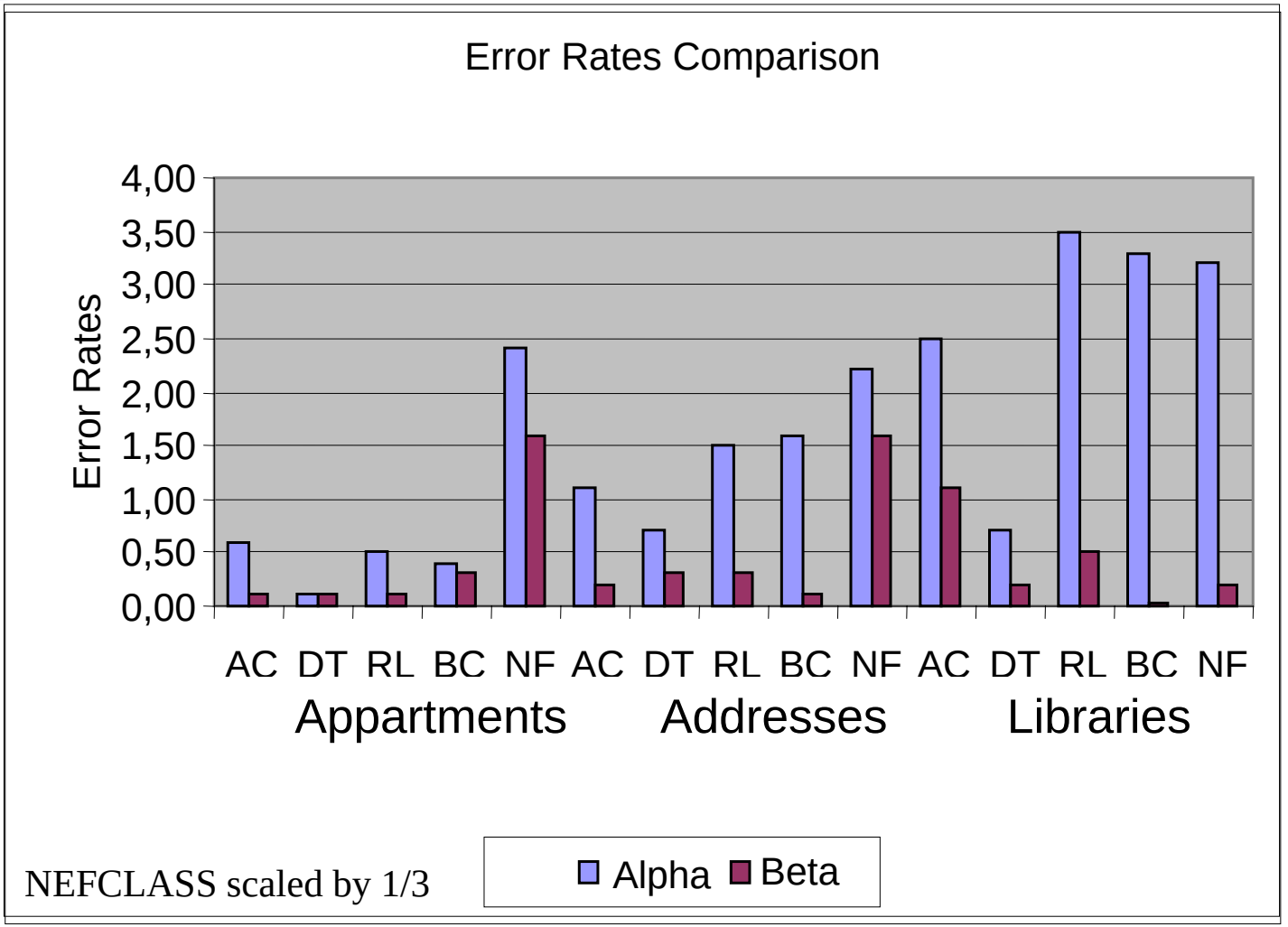
Libraries

error-prob Classifier	α %	β %	Hmean (α,β)%
AssoClass	2,5	1,1	1,53
Decision Tree	0,7	0,2	0,31
Record Linkage	3,5	0,5	0,88
AutoClass	3,3	0,03	0,06
NEFClass	12,7	0,6	1,15

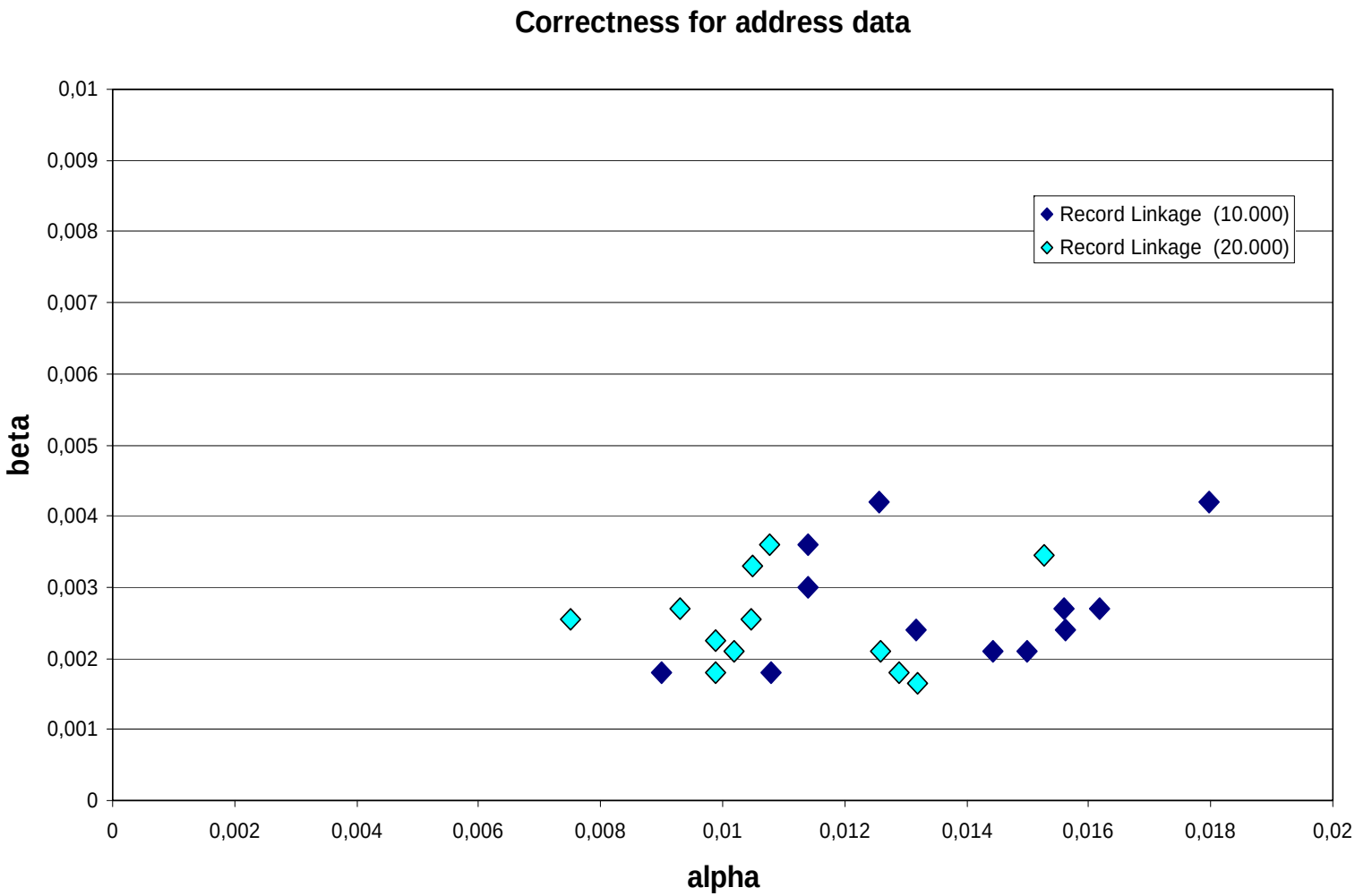
the best classification model was selected for each method; NF is over-fitted

Correctness Results

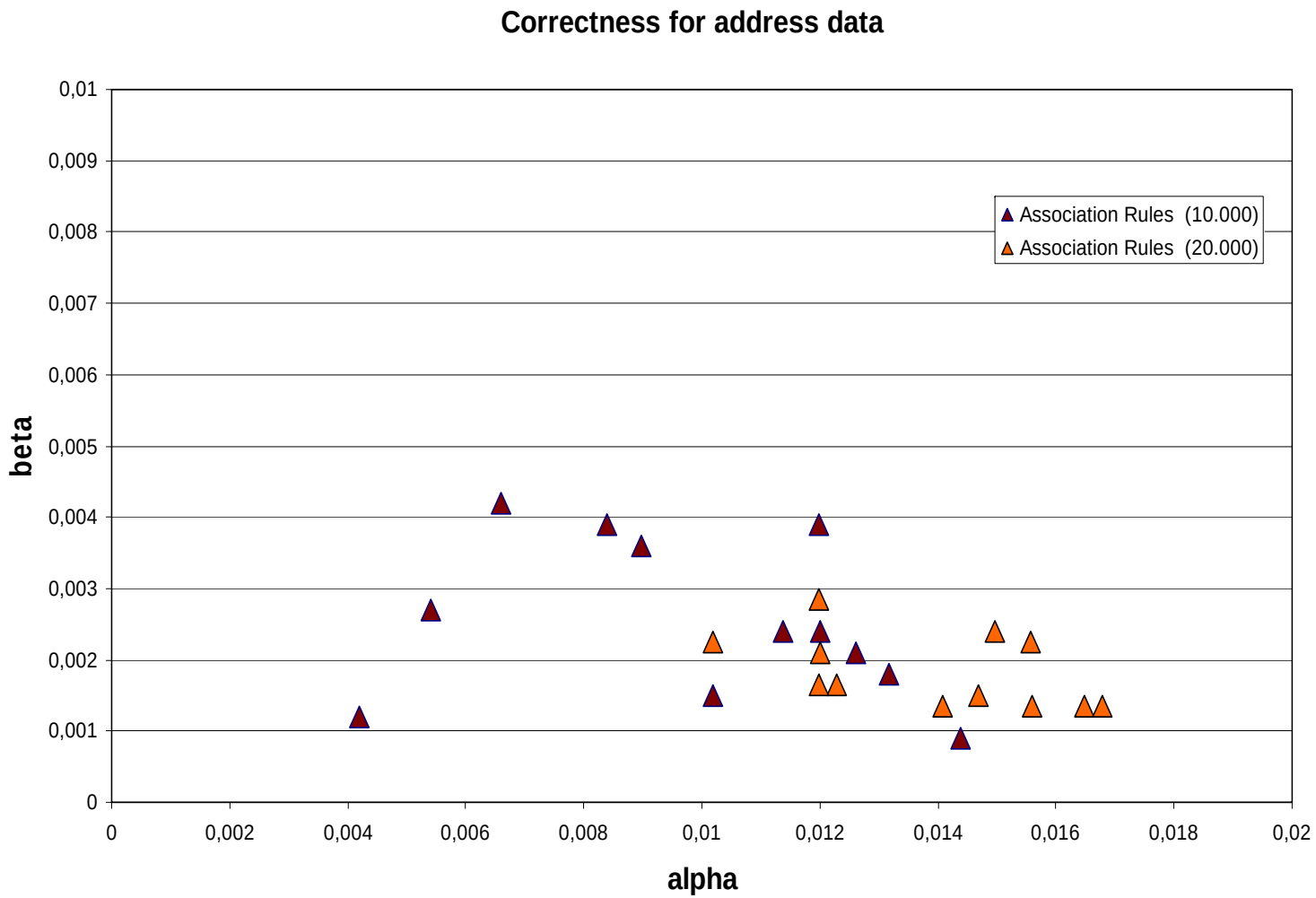
of all data sets



Record Linkage
Effects due to doubling sample size

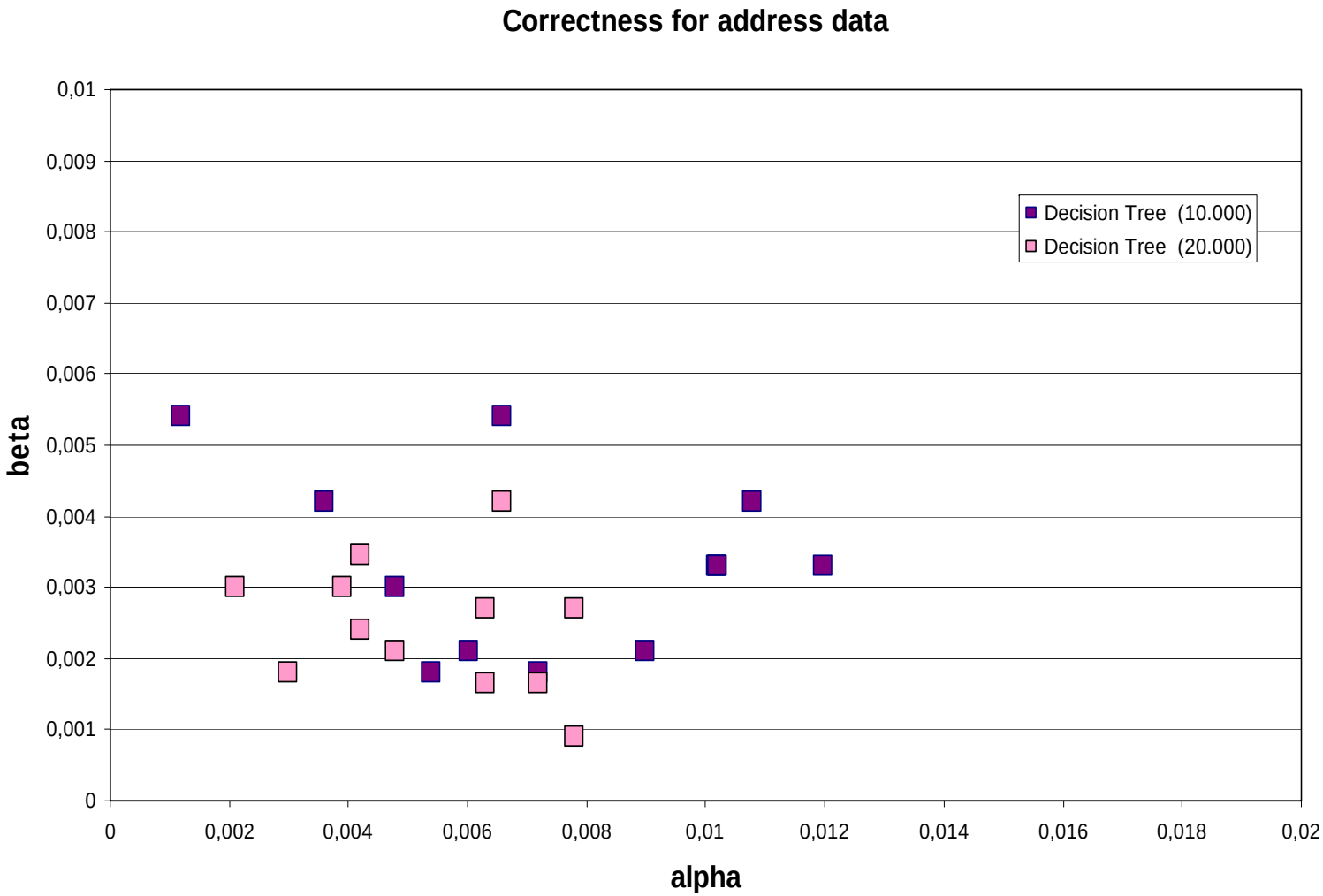


Association Rules
Effects due to doubling sampling size



Decision Tree

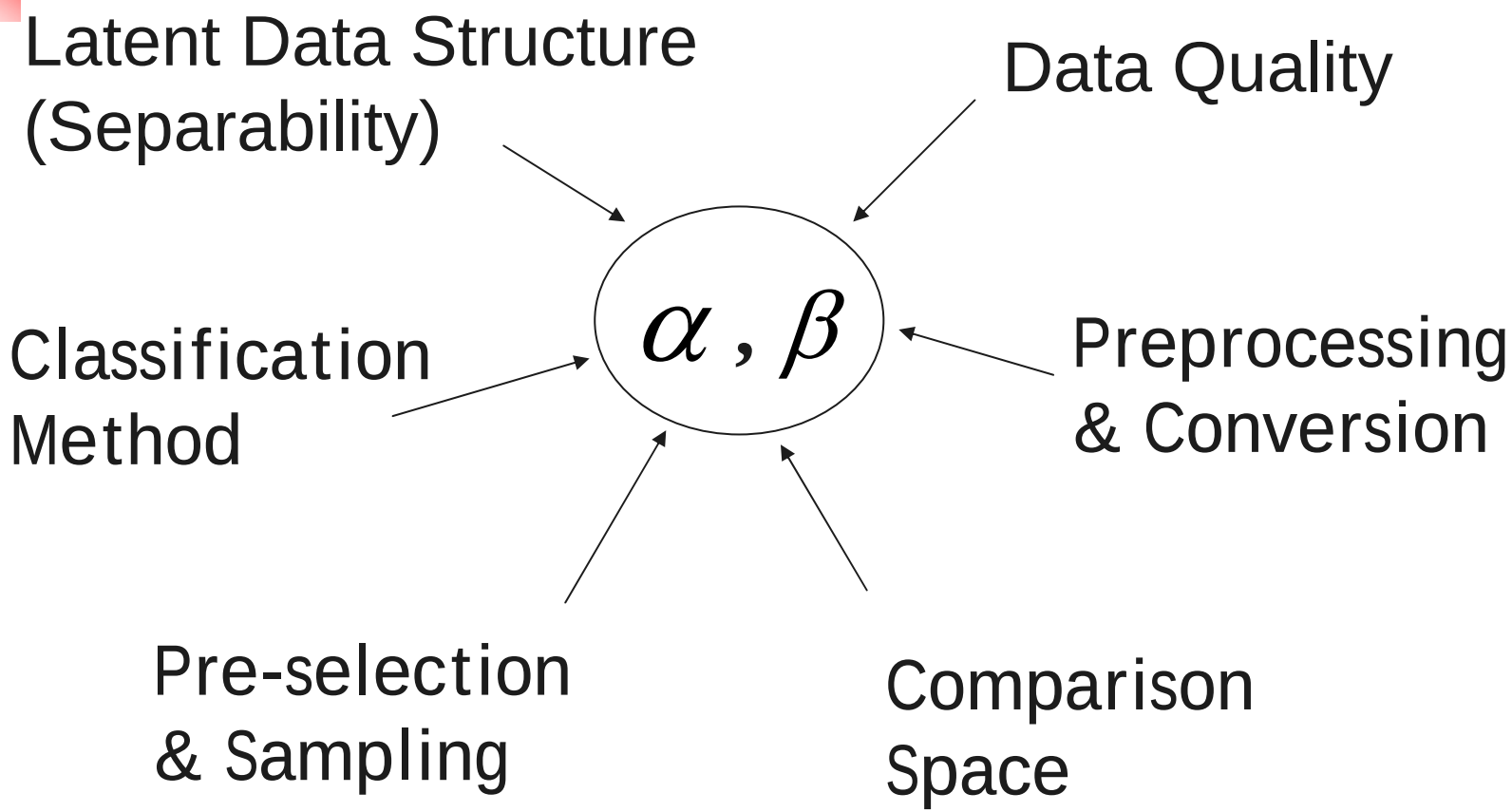
Effects due to doubling sampling size

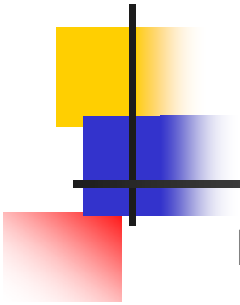


Conclusion

- **Decision Tree** outperformed the others, is robust
- **Record Linkage** well-behaved for correct specified log-linear models and sufficient large samples
- **Association Rules** allows to control (one of) the errors
- From additional study:
  - **Proper Bayes Classifier Autoclass** works also well,
  - **Neuro-Fuzzy Classifier** possibly over parametrized

Summary – Factors





Selected Publications

Neiling, M. (2005): Identification of Real-world objects in multiple databases. To appear in: Procs. of the Annual GfKl Conference 2005, Magdeburg.

Neiling, M. and H.-J. Lenz (2004): *The German Administrative Record Census - An Object Identification Problem.* Allg. Stat. Arch. 88, 259–277.

Neiling, M. (2004): *Identifizierung von Realwelt-Objekten in multiplen Daten-banken.* Dissertation, Brandenburgische Technische Universität Cottbus, 2004.

Neiling, M. and S. Jurk. *The object identification framework.* In Procs. of the ACM Workshop on Data Cleaning, Record Linkage and Object Identification, Washington DC, August 2003.

Neiling, M., S. Jurk, H.-J. Lenz, and F. Naumann. *Object identification quality.* In Procs. of the Intl. Workshop on Data Quality in Cooperative Information Systems (DQCIS2003), Siena, Italy, January 10-11, 2003.

[BM03] M. Bilenko and R. Mooney. Adaptive duplicate detection using learnable string similarity measures. *KDD Conf. 2003*, Washington DC.

[EVE02] M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. Tailor: A record linkage toolbox. *ICDE 2002*, San Jose.

Selected References

[FS69] I.P. Fellegi and A.B. Sunter. A theory of record linkage. *JASA*, 64:1183-1210, 1969.

[HS95] M.A. Hernandez and S.J. Stolfo. The merge/purge problem for large databases. *ACM SIGMOD Conf. 1995*, 127-138,

[GBVR03] L. Gu, R. Baxter et al. Record linkage: Current practice and future directions. Technical Report 03/83, CSIRO Mathematical and Information Sciences, Canberra, Australia, 2003.

[GFS01+b] H. Galhardas, D. Florescu et al. Declarative data cleaning: Language, model and algorithms. *VLDB Conf. 2001*, Orlando.

[VEH00] V.S. Verykios, A.K. Elmagarmid, and E.N Houstis. Automating the approximate record matching process. *J. of Information Sciences*, 126:83--98, 2000.

[Win93] W.E. Winkler. Improved decision rules in the Fellegi-Sunter model of record linkage. The Research Report Series RR-93/12, U.S. Bureau of the Census, 1993.

[Yan02] W. Yancey. Improving parameter estimates for record linkage parameters. In *Proc. of the Section on Survey Research Methodology*. American Statistical Association, 2002.

Forgive me today - tomorrow I may no longer feel guilty !

Some Additional Slides

Data Quality: Semantical Constraints II

Ratio of domain size of an attribute set and the size of A:

$$\text{selectivity}(A_i, Y) := |\{y \in \text{dom}(Y) \mid \exists a \in A_i : y = Y(a)\}| / |A|,$$

e.g. selectivity(*Addresses*,*BirthDate*) = .11,
selectivity(*Addresses*,*FullName*) = .72,

A:	database table
$A_i \subset A$:	Selection from A
Y:	Attribut set
$\|\ldots\|$:	Set cardinality
v:	value

Approximate Keys

- Likelihood that attribute values coincide for duplicates and vice versa, based on
  - accuracy(Y) := $\boldsymbol{P}$ ($Y(a)=Y(b) \mid a \equiv b$),
  - confidence(Y) := $\boldsymbol{P}$ ($a \equiv b \mid Y(a)=Y(b)$),

Definition. Y is a **approximate key** with confidence p, if both accuracy$(Y) \geq p$ and confidence$(Y) \geq p$.

- Examples (address data)
  - accuracy(Year) = .75, confidence(Year) = .98
  - accuracy(Street) = .87, confidence(Street) = .98
  - accuracy(LastName) = .93, confidence(LastName) = .99

Constraints - More Examples

- Δ- accuracy(Year) = .7613, Δ- confidence(Year) = .9217
- Δ- accuracy(Street) = .9579, Δ- confidence(Street) = .9793

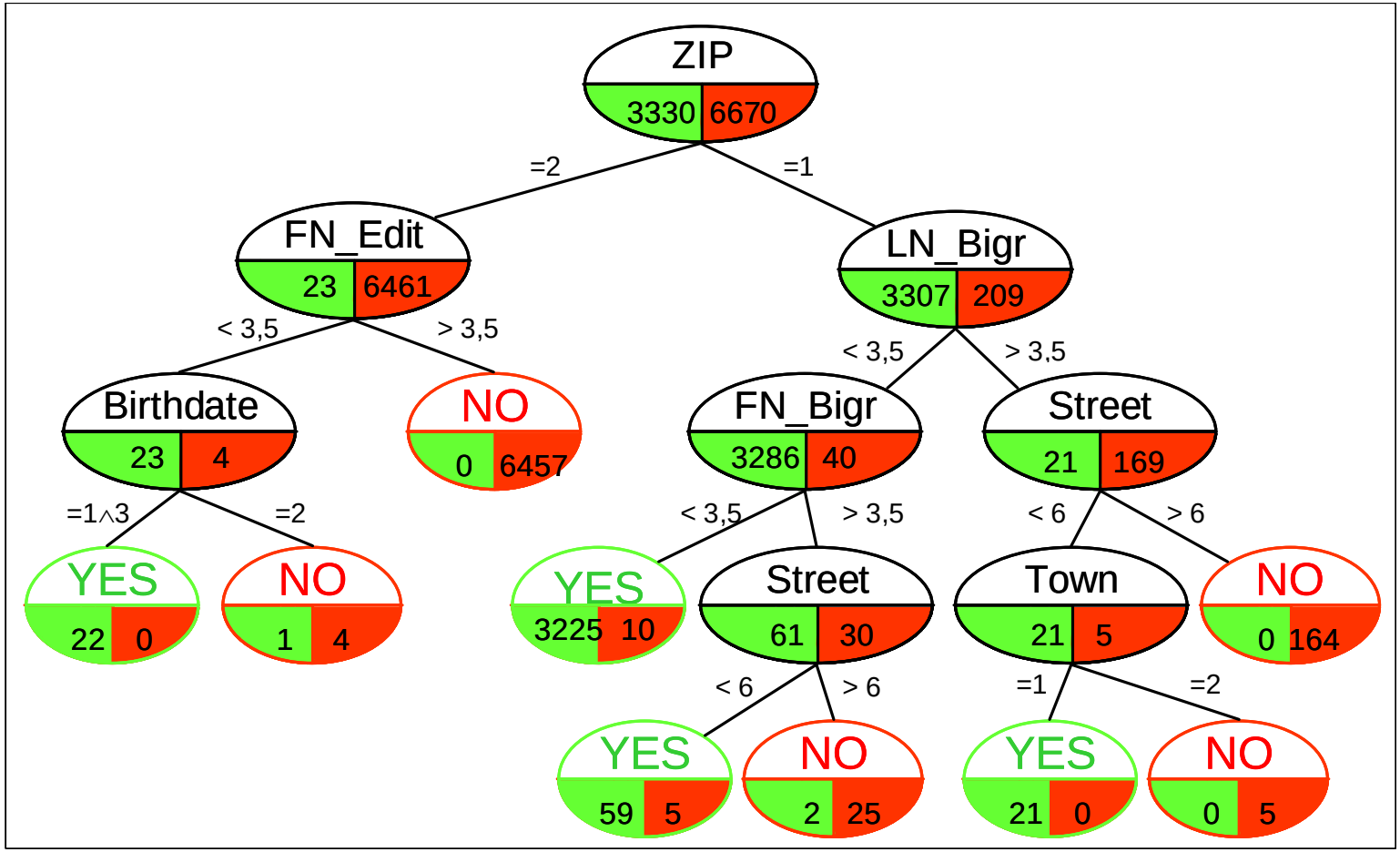
- anti-confidence(Street)= .9983
- anti-confidence(BirthDate)= .9990

Data Quality: Semantical Constraints III

- Example: Berlin Online Apartment Advertisements database (BOA)
- 9,842 cleaned ads from Tagesspiegel & Berliner Morgenpost (from May 18th + 25th, 2002)
- Semantic constraints
  - Reduction of pairs by differentiating keys: 99.876%,
  - Expected number of duplicates

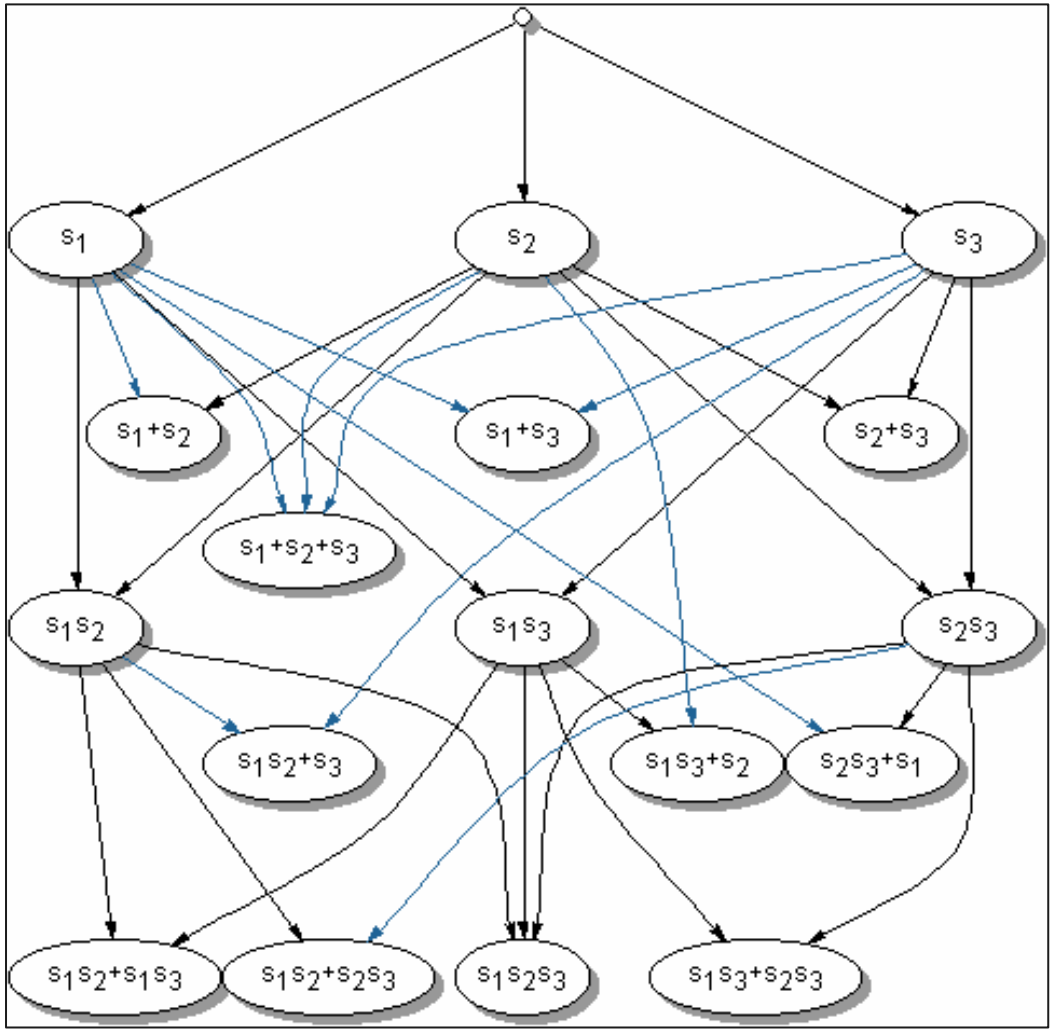
    C_k: $537 < \text{duplicates}(A) < 2.636$ (BOA contains 2.187 duplicates)
  - other C_k's, e.g.

Decision Tree: Example for Addresses

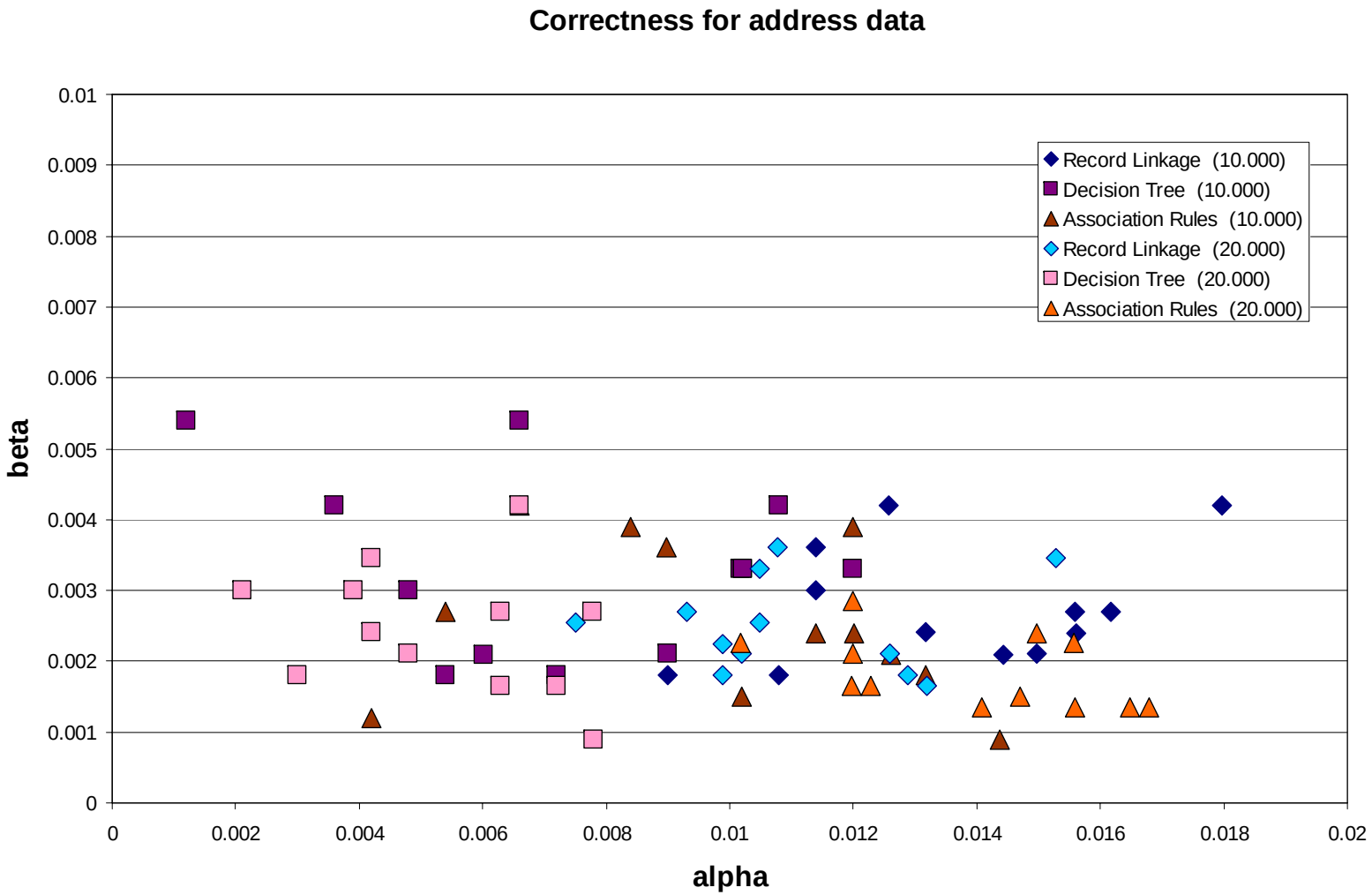


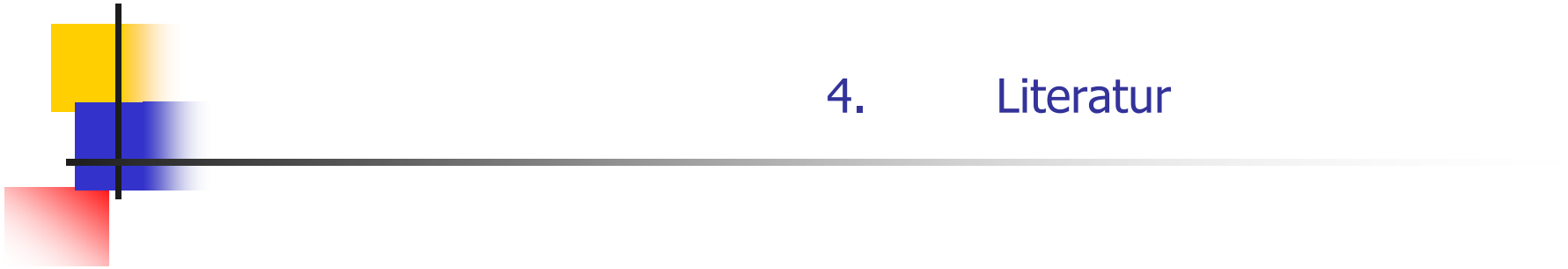
Pre-selection Example

- Three relational selectors
- Exhaustive search



All Results for Address Data





4. Literatur

1. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methods and Techniques. Heidelberg: Springer Verlag (2006)
2. Dombrowski, Erik und Lechtenböger, Jens: Evaluation objektorientierter Ansätze zur Data-Warehouse-Modellierung, Datenbank-Spektrum 15/2005
3. Naumann, Felix: Datenqualität, Informatik-Spektrum_30_1_2007


Simsalabim:
Data Quality
assured!