

# How to handle duplication in large datasets and import scenarios

Andreas Müller, Marcus Döring, Walter G. Berendsohn  
Department of Biodiversity Informatics,  
Botanic Garden and Botanical Museum Berlin-Dahlem



# How to handle duplication ...

same Entity

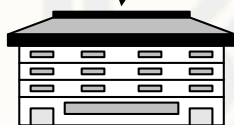
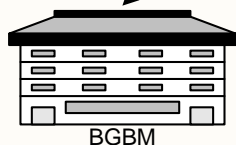
What is a duplicate?

same  
Taxon

physical

digital

1 location:  
same species, 3 specimen



DB1

ID	Genus	Species	Collection	Location
10	Erica	herbacea	BGBM	22,5°/-3,1°
500	Erica	herb.		22,5°/-3,12°
1001	Erica	herbacea	Kew	22,5°/-3,1°
1003				

DB2

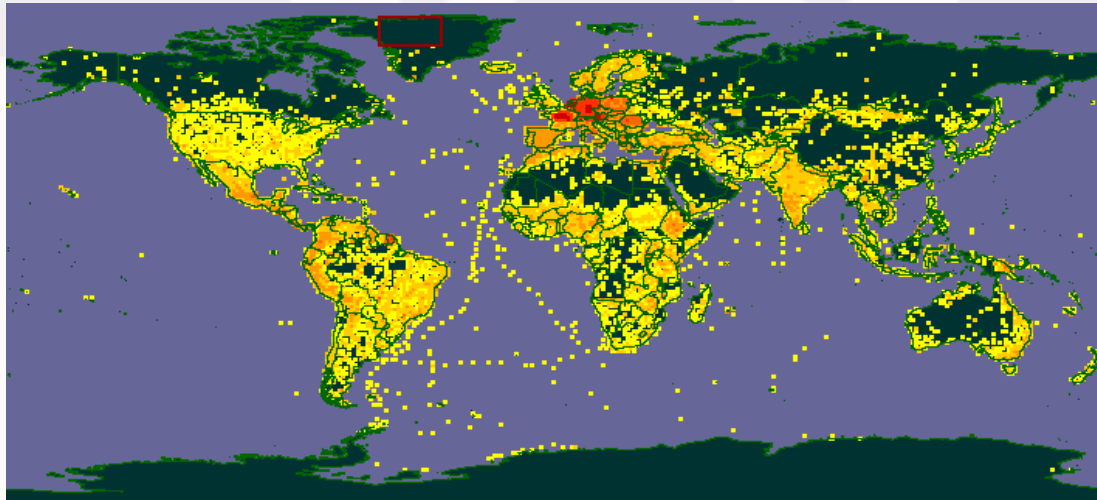
ID	TaxonName	Coll	Loc
20	E. herbacea	BGBM	-
300	Erica herbacea	BGBM	22,5°/-3,1°
1002	Erica herb.	Kew	22,5°/-3,1°
1003			

same: location; collector; taxon; ...  
different: collections; specimen; ...

## Duplicate search in large databases



- GBIF index
  - ~100.000.000 records (specimen, observation, undefined)
  - old GBIF-index:
    - unit, collection, institution, record\_basis, taxon name, collecting\_date, latitude, longitude, locality, ...
  - new GBIF-index:
    - Additional: collector, field number, altitude, ...



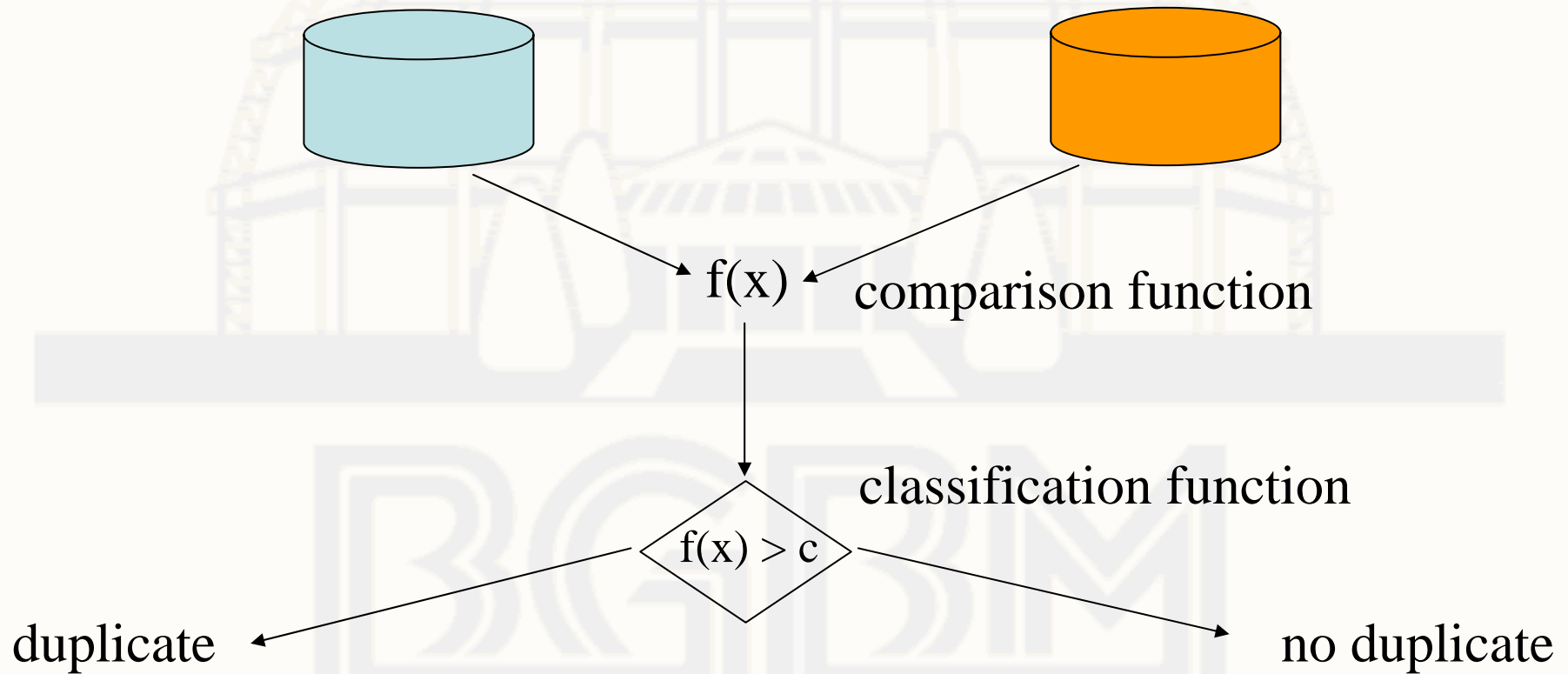
~16.000.000

## Use Cases

- finding (physical) duplicates
- finding additional information for existing entities
  - databases are specialized  $\Rightarrow$  subset of available information
  - different degree of exactness (e.g. longitude DB1:  $22^\circ$  / DB2:  $22,5673^\circ$ )
  - completeness
- avoiding to create duplicates
  - in local databases
  - import into GBIF index
- fuzzy, multiparameter search
  - no need to start with an existing record
  - was under discussion
  - query of type: Name, location, institution, collection date
- no record linkage

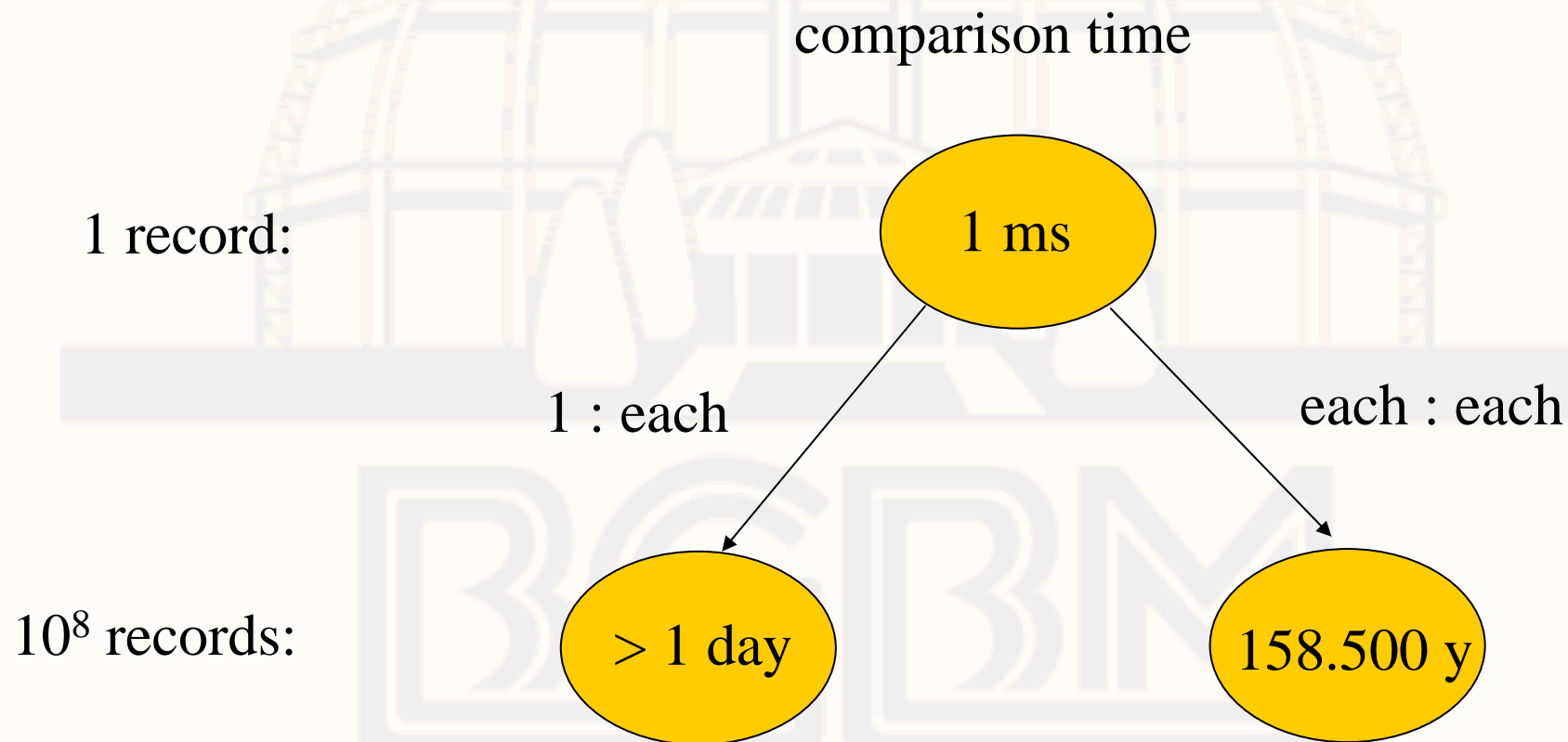
# How to handle duplication ...

## Greedy-Approach



# How to handle duplication ...

## Problems (1): Complexity



# How to handle duplication ...

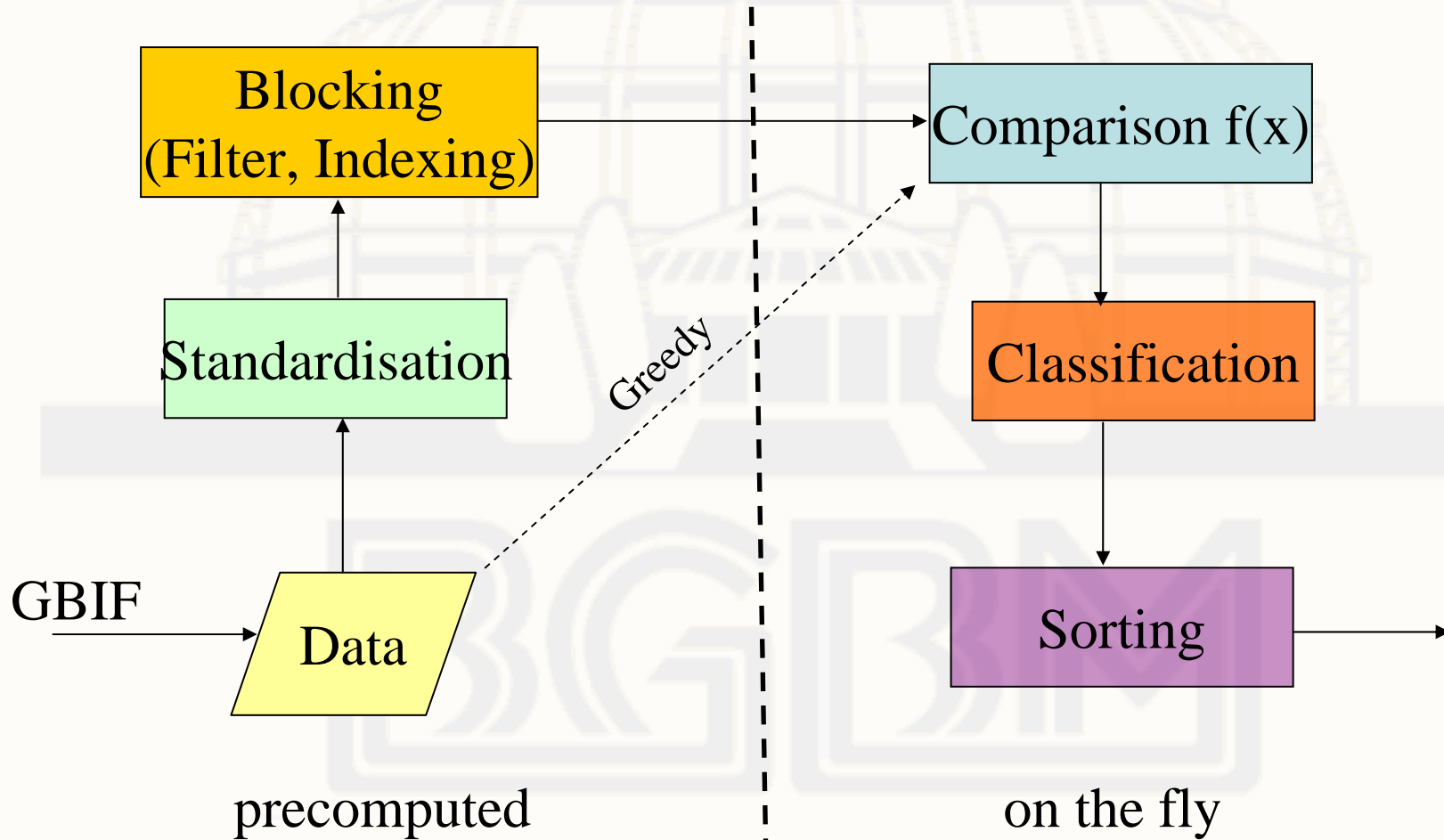
## Problems (2): no standard indexing

- standard database indexes: 1 parameter, exact match
- parameters are
  - missing
  - not specific (e.g. max. name frequency up to  $10^6$ )
  - fuzzy (different spelling, exactness, synonymy, errors in data...)

ID	Genus	Species	Collection	Location
10	Erica	herbacea	BGBM	-
500	Erica	herbacea	BGBM	22,5°/-3,187°
1001	Erica	herb.	Kew	22,5°/-3,1°
1003	Erika	herbazea	BGGGGGGGBM	
9567	Erica	carnea	Keeeeeeew	22,5°/-3,1°

# How to handle duplication ...

## Solution: Adapted record-linkage approach



# How to handle duplication ...

**Data** – Standardisation – Blocking – Comparison – Classification - Sorting

- Data: „Materialized“ View on GBIF index
  - flat table in new database (~5 h)
  - updated regularly

BGBM

# How to handle duplication ...

Data – **Standardisation** – Blocking – Comparison – Classification - Sorting

- Data: „Materialized“ View on GBIF index
  - flat table in new database (~5 h)
  - updated regularly
- Standardisation:
  - GBIF data are widely standardised => not much work to do
    - normalisation on names (reduce fuzzyness)
    - index for locality („geoindex“)

BGBM

# How to handle duplication ...

Data – Standardisation – **Blocking** – Comparison – Classification - Sorting

- Data: „Materialized“ View on GBIF index
  - flat table in new database (~5 h)
  - updated regularly
- Standardisation:
  - GBIF data are widely standardised => not much work to do
    - normalisation for names (reduce fuzzyness)
    - index for locality („geoindex“)
- Blocking: sorted neighborhood, (bi-gram, blocking, ...)

# How to handle duplication ...

Data – Standardisation – **Blocking** – Comparison – Classification - Sorting

- Sorted neighborhood
  - multi-attribute index
  - multi-pass index

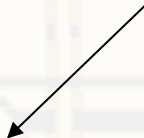
ID	Gen.	Species	Collection	Location
10	Erica	herbacea	BGBM	-
...				
500	Erica	herbacea	-	22,5°/-3,1°
...				
1001	Erica	herb.	Kew	22,5°/-3,1°
1002	Erika	herbazea	BGBM	-
...				
9567	Erica	carnea	Kew	22,5°/-3,1°

# How to handle duplication ...

Data – Standardisation – **Blocking** – Comparison – Classification - Sorting

- Sorted neighborhood
  - Multi-attribute index
  - Multi-pass index

sorted



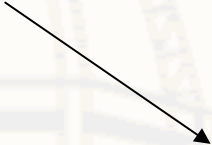
ID	Gen.	Species	Collection	Location	Index1	Index2
9567	Erica	carnea	Kew	22,5°/-3,1°	EricCarnKew_ 022357	Kew_022357EriHer
...					...	
500	Erica	herbacea	-	22,5°/-3,1°	EricHerb____ 022357	____022357EriHer
10	Erica	herbacea	BGBM	-	EricHerbBgbm_____	Bgbm_____ EriHer
...					...	
1001	Erica	herb.	Kew	22,5°/-3,1°	EricHerbKew_ 022357	Kew_022357EriHer
...					...	
1002	Erika	herbazea	BGBM	-	ErikHerbBgbm_____	Bgbm_____ EriHer

# How to handle duplication ...

Data – Standardisation – **Blocking** – Comparison – Classification - Sorting

- Sorted neighborhood
- Multi-attribute index
- Multi-pass index

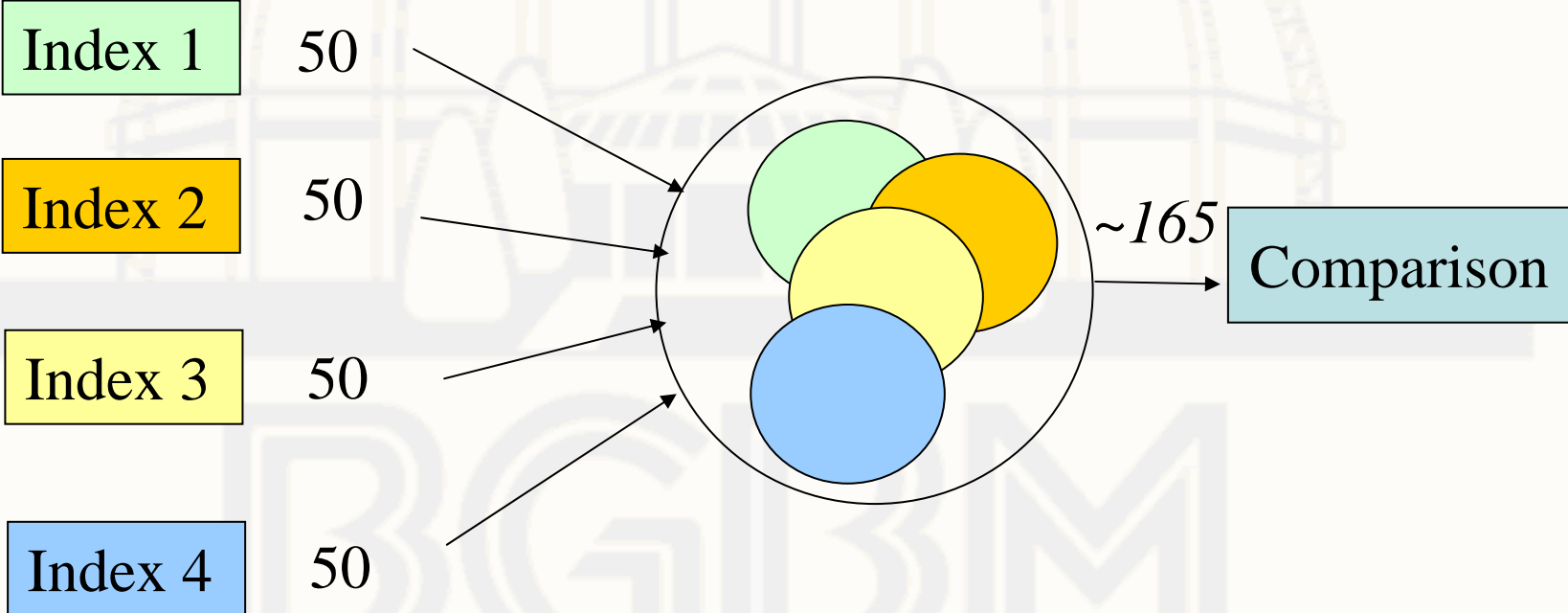
sorted



ID	Gen.	Species	Collection	Location	Index1	Index2
500	Erica	herbacea	-	22,5°/-3,1°	EricHerb____022357	____022357EriHer
...					...	...
10	Erica	herbacea	BGBM	-	EricHerbBgbm_____	Bgbm_____EriHer
1002	Erika	herbazea	BGBM	-	ErikHerbBgbm_____	Bgbm_____EriHer
...					...	...
1001	Erica	herb.	Kew	22,5°/-3,1°	EricHerbKew_022357	Kew_022357EriHer
9567	Erica	carnea	Kew	22,5°/-3,1°	EricCarnKew_022357	Kew_022357EriHer
...					...	

# How to handle duplication ...

Data – Standardisation – **Blocking** ⇒ **Comparison** – Classification - Sorting



# How to handle duplication ...

Data – Standardisation - Indexing - **Comparison** – Classification - Sorting

- Attribute level
  - probabilistic: probability of equality
  - different probability functions for each attribute-type
    - String comparison functions: winkler, editdistance, jaro, seqmatch
    - Numeric functions: distance, tolerance
    - Date functions
  - flexibel for optimization (e.g. by using frequency tables)
- Record level
  - add (weighted) probabilities for each attribute

# How to handle duplication ...

Data – Standardisation - Indexing - Comparison – **Classification** - Sorting

- Data: „Materialized“ View on GBIF index,
  - flat table in new database (~5 h)
  - regular update with new data
    - no inkonsistencies !!
- Standardisation:
  - GBIF data are widely standardised => not much work to do
    - normalization for names (v, w -> v), not for result
    - index for locality ( 2 numbers -> 1 number; „geographic index“)
- Indexing: blocking, sorted neighborhood, bi-gram
- Comparison:
  - attribute level
    - probabalistic approach
  - recordset level
- **Classification: by probability, number**

# How to handle duplication ...

## Data – Standardisation - Indexing - Comparison – Classification - **Sorting**

- Data: „Materialized“ View on GBIF index,
  - flat table in new database (~5 h)
  - regular update with new data
    - no inkonsistencies !!
- Standardisation:
  - GBIF data are widely standardised => not much work to do
    - normalization for names (v, w -> v), not for result
    - index for locality ( 2 numbers -> 1 number; „geographic index“)
- Indexing: blocking, sorted neighborhood, bi-gram
- Comparison:
  - attribute level
    - probabalistic approach
  - recordset level
- Classification: probability percentage, number
- **Sorting**

## Implementation

- Febri („Freely Extensible Biomedical Record Linkage“)
  - language: Python
  - probabilistic record linkage ("fuzzy" matching)
- Adapted for duplicate finding and fuzzy, multi-parameter search
- Future Work
  - new GBIF index
  - optimization
  - web-service (XML-RPC)

Thank you ...

?

Questions