Introduction to Data Cleaning

Helena Galhardas
DEI/IST

References

- No single reference!
- Slides “Data Quality and Data Cleansing” course, Felix Naumann, Winter 2014/15
So far…

- We’ve studied how to perform:
  - String matching efficiently and effectively.
- We’ve seen how string matching is important in data integration.
- Now, we’ll see how string matching is important in data cleaning.

Example (1)

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack Lemmon</td>
<td>430-871-8294</td>
<td>Maple St</td>
</tr>
<tr>
<td>Harrison Ford</td>
<td>292-918-2913</td>
<td>Culver Blvd</td>
</tr>
<tr>
<td>Tom Hanks</td>
<td>234-762-1234</td>
<td>Main St</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table R

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tom Hanks</td>
<td>234-762-1234</td>
<td>Main St</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ton Hanks</td>
<td>234-162-1234</td>
<td>Main Street</td>
</tr>
<tr>
<td>Kevin Spacey</td>
<td>-</td>
<td>Frost Blvd</td>
</tr>
<tr>
<td>Jack Lemon</td>
<td>430-817-8294</td>
<td>Maple Street</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Find records from different datasets that could be the same entity.
Example (2)

```
<country>
  <name>United States of America</name>
  <cities>New York, Los Angeles, Chicago</cities>
  <lakes>
    <name>Lake Michigan</name>
  </lakes>
</country>

<country>
  United States
  <city>New York</city>
  <city>Los Angeles</city>
  <lakes>
    <lake>Lake Michigan</lake>
  </lakes>
</country>
```

and are the same object?

Example (3)


- These two bibliographic references concern the same publication!
The three examples refer to the same problem that is known under different names:
- approximate duplicate detection
- record linkage
- entity resolution
- merge-purge
- data matching …

It is one of the data quality problems addressed by data cleaning

Outline

- Introduction to data cleaning
- Application contexts of data cleaning
- Data quality dimensions
- Taxonomy of data quality problems
- Data quality process
- Main data quality tools
- Real-world examples
**Why Data Cleaning?**

Data in the real world is **dirty**
- **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., occupation=""
- **noisy**: containing errors (spelling, phonetic and typing errors, word transpositions, multiple values in a single free-form field) or outliers
  - e.g., Salary="-10"
- **inconsistent**: containing discrepancies in codes or names (synonyms and nicknames, prefix and suffix variations, abbreviations, truncation and initials)
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., was rating "1,2,3", now rating “A, B, C”
  - e.g., discrepancy between approximate duplicate records

---

**Data Quality Problems (Dirty Data)**

<table>
<thead>
<tr>
<th>CUST</th>
<th>CNr</th>
<th>Name</th>
<th>Birthday</th>
<th>Age</th>
<th>Sex</th>
<th>Phone</th>
<th>ZIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>Costa, Rui</td>
<td>18.2.80</td>
<td>37</td>
<td>m</td>
<td>999999999</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>1235</td>
<td>Rui Costa</td>
<td>18.2.80</td>
<td>27</td>
<td>m</td>
<td>963124568</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

- **Uniqueness**
- **Contradictions**
- **Missing values**
- **Representation**
- **Ref. integrity**
- **Inconsistent values**
- **Duplicates**
Impact of Data Quality Problems

- **Incorrect prices** in inventory retail databases [English 1999]
  - Costs for consumers 2.5 billion $
  - 80% of barcode-scan-errors to the disadvantage of consumer
- **IRS 1992**: almost 100,000 tax refunds not deliverable [English 1999]
- 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. [Strong et al. 1997a]
- **US-Postal Service**: of 100,000 mass-mailings up to 7,000 undeliverable due to incorrect addresses [Pierce 2004]

Why Is Data Dirty?

- **Incomplete data** comes from:
  - non available data value when collected
  - different criteria between the time when the data was collected and when it is analyzed
  - human/hardware/software problems
- **Noisy data** comes from:
  - data collection: faulty instruments
  - data entry: human or computer errors
  - data transmission
- **Inconsistent (and duplicate) data** comes from:
  - Different data sources, so non-uniform naming conventions/data codes
  - Functional dependency and/or referential integrity violation
Application contexts

- **Integrate data** from different sources
  - E.g., populating a DW from different operational data stores or a mediator-based architecture

- **Eliminate errors and duplicates** within a single source
  - E.g., duplicates in a file of customers

- **Migrate data** from a source schema into a different fixed target schema
  - E.g., discontinued application packages

- **Convert poorly structured data** into structured data
  - E.g., processing data collected from the Web

When materializing the integrated data (data warehousing)...

ETL: Extraction, Transformation and Loading

70% of the time in a data warehousing project is spent with the ETL process
Why is Data Cleaning Important?

Activity of converting source data into target data without errors, duplicates, and inconsistencies, i.e.,

Cleaning and Transforming to get…

High-quality data!

- No quality data, no quality decisions!
  - Quality decisions must be based on good quality data (e.g., duplicate or missing data may cause incorrect or even misleading statistics)

Quality

“Even though quality cannot be defined, you know what it is.”

Robert Pirsig
Outline

- Introduction to data cleaning
- Application contexts of data cleaning
  - **Data quality dimensions**
- Taxonomy of data quality problems
- Data quality process
- Main data quality tools
- Real-world examples

What is Data of Good Quality?

Fitness for use

Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Understandability, Consistency, Concise Representation

179 Dimensions
### Data Quality Dimensions (classical)

#### Accuracy
- Refers to the closeness of values in a database to the true values of the entities that the data in the database represent; if it is not 100% that means that there are errors in data.
- Example: “Jhn” vs. “John”

#### Completeness
- Concerns whether the database has complete information to answer queries.
- Partial knowledge of the records in a table or of the attributes in a record.

#### Currency
- Aims at identifying the current values of entities represented by tuples in a database and to answer queries using those values.
- Example: Residence (Permanent) Address: out-dated vs. up-to-dated

#### Consistency
- Refers to the validity and integrity of data representing real-world entities; if it is violated, leads to discrepancies and conflicts in the data.
- Example: ZIP Code and City inconsistent
**Accuracy**

- **Closeness between a value v and a value v’**, considered as the correct representation of the real-world phenomenon that v aims to represent.
  - Ex: for a person name “John”, v’ = John is correct, v = Jhn is incorrect

**Syntactic accuracy**: closeness of a value v to the elements of the corresponding definition domain D
  - Ex: if v = Jack, even if v’ = John, v is considered syntactically correct, because it is an admissible value in the domain of people names.
  - Measured by means of comparison functions (e.g., edit distance) that evaluate the distance between v and the values of the domain

**Semantic accuracy**: closeness of the value v to the true value v’
  - Measured with a <yes, no> or <correct, not correct> domain
  - Coincides with correctness
  - The corresponding true value has to be known

---

**Granularity of accuracy definition**

- Accuracy may refer to:
  - a single value of a relation attribute
  - an attribute or column
  - a relation
  - the whole database
Metrics for quantifying accuracy

- **Weak accuracy error**
  - Characterizes accuracy errors that do not affect identification of tuples

- **Strong accuracy error**
  - Characterizes accuracy errors that affect identification of tuples

- **Percentage of accurate tuples**
  - Characterizes the fraction of accurate tuples matched with a reference table

Completeness

- “The extent to which data are of sufficient breadth, depth, and scope for the task in hand.”

- Three types:
  - **Schema completeness**: degree to which concepts and their properties are not missing from the schema
  - **Column completeness**: evaluates the missing values for a specific property or column in a table.
  - **Population completeness**: evaluates missing values with respect to a reference population
Completeness of relational data

- The completeness of a table characterizes the extent to which the table represents the real world.
- Can be characterized with respect to:
  - The presence/absence and meaning of null values
    Example: In `Person(name, surname, birthdate, email)`, if `email` is null may indicate the person has no mail (no incompleteness), email exists but is not known (incompleteness), it is not known whether Person has an email (incompleteness may not be the case)
  - Validity of open world assumption (OWA) or closed world assumption (CWA)
    - **OWA**: assumes that in addition to missing values, some tuples representing real-world entities may also be missing
    - **CWA**: assumes the database has collected all the tuples representing real-world entities, but the values of some attributes in those tuples are possible missing

Metrics for quantifying completeness (1)

- Model without null values with OWA
  - Needs a reference relation \( \text{ref}(r) \) for a relation \( r \), that contains all the tuples that satisfy the schema of \( r \)
    \[
    C(r) = \frac{|r|}{|\text{ref}(r)|}
    \]
  - Example: according to a registry of Lisbon municipality, the number of citizens is 2 million. If a company stores data about Lisbon citizens for the purpose of its business and that number is 1,400,000 then \( C(r) = 0.7 \)
Metrics for quantifying completeness (2)

- **Model with null values with CWA**: specific definitions for different granularities:
  - **Values**: to capture the presence of null values for some fields of a tuple
  - **Tuple**: to characterize the completeness of a tuple wrt the values of all its fields:
    - Evaluates the % of specified values in the tuple wrt the total number of attributes of the tuple itself
    - Example: `Student(stID, name, surname, vote, examdate)`
    - Equal to 1 for (6754, Mike, Collins, 29, 7/17/2004)
    - Equal to 0.8 for (6578, Julliane, Merrals, NULL, 7/17/2004)

Metrics for quantifying completeness (3)

- **Attribute**: to measure the number of null values of a specific attribute in a relation
  - Evaluates % of specified values in the column corresponding to the attribute wrt the total number of values that should have been specified.
  - Example: For calculating the average of votes in `Student`, a notion of the completeness of `vote` should be useful
- **Relations**: to capture the presence of null values in the whole relation
  - Measures how much info is represented in the relation by evaluating the content of the info actually available wrt the maximum possible content, i.e., without null values.
**Time-related dimensions**

**Currency:** concerns how promptly data are updated
- Example: if the residential address of a person is updated (it corresponds to the address where the person lives) then the currency is high

**Volatility:** characterizes the frequency with which data vary in time
- Example: Birth dates (volatility zero) vs stock quotes (high degree of volatility)

**Timeliness:** expresses how current data are for the task in hand
- Example: The timetable for university courses can be current by containing the most recent data, but it cannot be timely if it is available only after the start of the classes.

**Metrics of time-related dimensions**

- **Last update metadata** for currency
  - Straightforward for data types that change with a fixed frequency

- **Length of time that data remain valid** for volatility

- **Currency + check that data are available before the planned usage time** for timeliness
**Consistency**

- Captures the **violation of semantic rules** defined over a set of data items, where data items can be tuples of relational tables or records in a file
  - **Integrity constraints** in relational data
    - Domain constraints, key definitions, inclusion and functional dependencies

**Other dimensions**

- **Interpretability**: concerns the documentation and metadata that are available to correctly interpret the meaning and properties of data sources
- **Synchronization** between different time series: concerns proper integration of data having different time stamps.
- **Accessibility**: measures the ability of the user to access the data from his/her own culture, physical status/functions, and technologies available.
Outline

- Introduction to data cleaning
- Application contexts of data cleaning
- Data quality dimensions
  - Taxonomy of data quality problems
- Data quality process
- Main data quality tools
- Real-world examples

Taxonomy of data quality problems
[Oliveira 2009]

- Value-level
- Value-set (attribute/column) level
- Record level
- Relation level
- Multiple relations level
Value level

**Missing value**: value not filled in a not null attribute
- Ex: birth date = ‘’

**Syntax violation**: value does not satisfy the syntax rule defined for the attribute
- Ex: zip code = 27655-175; syntactical rule: xxxx-xxx

**Spelling error**
- Ex: city = ‘Lsboa’, instead of ‘Lisbon’

**Domain violation**: value does not belong to the valid domain set
- Ex: age = 240; age: {0, 120}

Value-set and Record levels

**Value-set level**
- **Existence of synonyms**: attribute takes different values, but with the same meaning
  - Ex: emprego = ‘futebolista’; emprego = ‘jogador futebol’
- **Existence of homonyms**: same word used with diff meanings
  - Ex: same name refers to different authors of a publication
- **Uniqueness violation**: unique attribute takes the same value more than once
  - Ex: two clients have the same ID number
- **Integrity constraint violation**
  - Ex: sum of the values of percent attribute is more than 100

**Record level**
- **Integrity constraint violation**
  - Ex: total price of a product is different from price plus taxes
**Relation level**

Heterogeneous data representations: different ways of representing the same real world entity
- Ex: name = ‘John Smith’; name = ‘Smith, John’

Functional dependency violation
- Ex: (2765–175, ‘Estoril’) and (2765–175, ‘Oeiras’)

Existence of approximate duplicates
- Ex: (1, André Fialho, 12634268) and (2, André Pereira Fialho, 12634268)

Integrity constraint violation
- Ex: sum of salaries is superior to the max established

---

**Multiple tables level**

Heterogeneous data representations
- Ex: one table stores meters, another stores inches

Existence of synonyms

Existence of homonyms

Different granularities: same real world entity represented with diff. granularity levels
- Ex: age: {0-30, 31-60, > 60}; age: {0-25, 26-40, 40-65, >65}

Referential integrity violation

Existence of approximate duplicates

Integrity constraint violation
Outline

- Introduction to data cleaning
- Application contexts of data cleaning
- Data quality dimensions
- Taxonomy of data quality problems
  - Data quality process
- Main data quality tools
- Real-world examples

Data Quality Process

1. Data Quality Auditing (Assessment)
   - Data Profiling
   - Data Analysis

2. Data Quality Improvement
   - Data Cleaning
   - Data Enrichment
Data quality auditing

- Constituted by:
  - **Data profiling** – analysing data sources to identify data quality problems
  - **Data analysis** – statistical evaluation, logical study and application of data mining algorithms to define data patterns and rules

- **Main goals:**
  - To obtain a definition of the data: metadata collection
  - To check violations to metadata definition
  - To detect other data quality problems that belong to a given taxonomy
  - To supply recommendations in what concerns the data cleaning task

---

Data Profiling

- **Data source discovery**
  - Metadata

- **Schema discovery**
  - Schema matching and mapping
  - Profiling for metadata (keys, foreign keys, data types, …)

- **Data discovery**
  - Column-level: Null-values, domains, patterns, value distributions / histograms
  - Table-level: Data mining, rules
Typical techniques used in data quality auditing

- **Dictionaries of words**: so that attribute values are compared with one or more dictionaries of the domain
  - Ex: wordnet
- **Algorithms to detect functional dependencies and their violations**
- **Algorithms to detect duplicates**
  - String matching for string fields
    - Character-based
    - Token-based
    - Phonetic algorithms
  - Record matching
    - Rule-based
    - Probabilistic
    - ...

Data quality improvement

- **Includes often**:
  - **Data transformation** – set of operations that source data must undergo to fit target schema
  - **Data cleaning** – detecting, removing and correcting dirty data (including approximate duplicate elimination)
  - **Data enrichment** – use of additional information to improve data quality

- **Main goal**:
  - To **correct** the data quality problems detected during the data quality auditing process
Typical techniques used in data cleaning and transformation

- Dictionaries of words
- Libraries of pre-defined cleaning functions
- Machine learning techniques
- Techniques for consolidating approximate duplicates

Methodology for data cleaning

1. Extraction of the individual fields that are relevant
2. Standardization of record fields
3. Correction of data quality problems at value level
   - Missing values, syntax violation, etc
4. Correction of data quality problems at value-set level and record level
   - Synonyms, homonyms, uniqueness violation, integrity constraint violation, etc
5. Correction of data quality problems at relation level
   - Violation of functional dependencies, duplicate elimination, etc
6. Correction of data quality problems problems at multiple relations level
   - Referential integrity violation, duplicate elimination, etc
- User feedback
  - To solve instances of data quality problems not addressed by automatic methods
- Effectiveness of the data cleaning and transformation process must be always measured for a sample of the data set
Data Cleaning Tasks

1. Extraction from sources
   - Technical and syntactic obstacles
2. Transformation
   - Schematic obstacles
3. Standardization
   - Syntactic and semantic obstacles
4. Duplicate detection
   - Similarity functions
   - Algorithms
5. Data fusion / consolidation
   - Semantic obstacles
6. Loading into warehouse / presenting to user

Human Interaction is Needed

- Components to implement
  - Wrappers for technical heterogeneity
  - Schema integration based on correspondences
  - Similarity measure for schema elements
  - Similarity measure for records
- Knobs to turn
  - Thresholds for similarity measures
  - Partition size / window size
- Expert guidance
  - Rule selection / rule specification
  - Schema matching
  - Duplicate detection
  - Data fusion
Outline

- Introduction to data cleaning
- Application contexts of data cleaning
- Data quality dimensions
- Taxonomy of data quality problems
- Data quality process
  - Main data quality tools
- Real-world examples

Existing technology for ensuring data quality

Ad-hoc programs written in a programming language like C or Java or using an RDBMS proprietary language
  - Programs difficult to optimize and maintain

RDBMS mechanisms for guaranteeing integrity constraints
  - Do not address important data instance problems

Data transformation workflow scripts using a data cleaning/profiling tool
Existing technology for ensuring data quality

Ad-hoc programs written in a programming language like C or Java or using an RDBMS proprietary language
- Programs difficult to optimize and maintain

RDBMS mechanisms for guaranteeing integrity constraints
- Do not address important data instance problems

- Data transformation workflow scripts using an data cleaning/profiling tool

Criteria for comparing commercial data quality tools (1)

Debugger:
- Data lineage: data lineage or provenance identifies the set of source data items that produced a given data item
- Breakpoints: breakpoints is an intentional stopping or pausing place in a cleaning program put in place for debugging purposes
- Edit values: the user can edit values during debugging
Criteria for comparing commercial data quality tools (2)

Profiling:

Rules: A rule is a business logic that defines conditions applied to data. They are used to validate the data and to measure data quality.

Filters: A filter is used to split the data tuples in different groups. Each group should be validated by a different set of rules.

Criteria for comparing commercial data quality tools (3)

Execution:

User involvement: Support for user interaction in a data cleaning process.

Incremental updates: The ability to incrementally update data targets, instead of rebuilding them from scratch every time.
Commercial Data Cleaning Tools (2014) (1/3)

<table>
<thead>
<tr>
<th>Tools</th>
<th>Debugger</th>
<th>Profiling</th>
<th>Execution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data lineage</td>
<td>Breakpoints</td>
<td>Edit values</td>
</tr>
<tr>
<td>Informatica PowerCenter</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>IBM Information Server</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Talend Open Studio</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Oracle Data Integrator</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>SQL Server Integration Services</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>SAS Data Integration Studio</td>
<td>Y</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Pentaho Data Integration</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Clover ETL</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

Criteria for comparing commercial data quality tools (4)

**Extensibility:**

*Create operators:* the user can define new operators

*Modify operators:* the user can modify standard operators

**User Interface:**

*Drag and drop:* the user can define data quality processes using a drag and drop interface

*Editor:* the user can define and edit data quality processes modeled as workflows using a graphical interface
Commercial Data cleaning tools (2014) (2/3)

<table>
<thead>
<tr>
<th>Tools</th>
<th>Extensibility</th>
<th>User Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Create Operators</td>
<td>Modify Operators</td>
</tr>
<tr>
<td>Informatica PowerCenter</td>
<td>Y (Java)</td>
<td>N</td>
</tr>
<tr>
<td>IBM Information Server</td>
<td>Y (Java)</td>
<td>N</td>
</tr>
<tr>
<td>Talend Open Studio</td>
<td>Y (Java, Groovy)</td>
<td>Y (Java)</td>
</tr>
<tr>
<td>Oracle Data Integrator</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>SQL Server Integration Services</td>
<td>Y (C#, VB)</td>
<td>N</td>
</tr>
<tr>
<td>SAS Data Integration Studio</td>
<td>Y (SAS)</td>
<td>Y (SAS)</td>
</tr>
<tr>
<td>Pentaho Data Integration</td>
<td>Y (Javascript)</td>
<td>N</td>
</tr>
<tr>
<td>Clover ETL</td>
<td>Y (CTI)</td>
<td>N</td>
</tr>
</tbody>
</table>

Criteria for comparing commercial data quality tools (5)

**Scalability:**

- **Grid:** the tool can run a cleaning process on a collection of computer resources from multiple locations

- **Partitioning:** the user can partition the data and run each partition independently (on different CPUs or cores)

- **Pushdown optimization:** the tool translates the transformation logic into SQL queries and sends the SQL queries to the database. The database engine executes the SQL queries to process the transformations

**Others:**

- **Free version:** the tool has a free version
### Commercial Data Cleaning Tools (2014)

#### (3/3)

<table>
<thead>
<tr>
<th>Tools</th>
<th>Scalability</th>
<th>Others</th>
<th>Free version</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatica PowerCenter</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>IBM Information Server</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Talend Open Studio</td>
<td>Y</td>
<td>N</td>
<td>Optional ELT</td>
</tr>
<tr>
<td>Oracle Data Integrator</td>
<td>Y</td>
<td>N</td>
<td>ELT</td>
</tr>
<tr>
<td>SQL Server Integration Services</td>
<td>N</td>
<td>Y</td>
<td>-</td>
</tr>
<tr>
<td>SAS Data Integration Studio</td>
<td>Y</td>
<td>Y</td>
<td>Y (IST)</td>
</tr>
<tr>
<td>Pentaho Data Integration</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Clover ETL</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
</tbody>
</table>

### Research Data cleaning tools (2014)

#### (1/2)

<table>
<thead>
<tr>
<th>Tools</th>
<th>Detection DQ problems</th>
<th>Repair DQ problems</th>
<th>Data Transformations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constraints</td>
<td>Statistical</td>
<td>Search</td>
</tr>
<tr>
<td>Cleanex</td>
<td>QCa</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Lunatic</td>
<td>Eigls</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Nadeef</td>
<td>CFDs, MDs</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Guided data repair</td>
<td>CFDs</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Scare</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Eracer</td>
<td>H</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Continuous data cleaning</td>
<td>FDs</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>
Criteria for comparing research data cleaning tools (1)

**Detection:**

**Constraints** – use of rules or/and conditions
- EGDs - equality generating dependencies
- QCs - quality constraints
- CFDs - Conditional functional dependencies
- MDs - Matching dependencies

**Statistical** – dirty tuples are detected based on simple statistics or in complex data analysis

---

Criteria for comparing research data cleaning tools (1)

**Repair:**

**Search:** The system explores the space of possible clean tables and heuristically selects the best table

**ML/St:** The system uses machine learning and/or statistical models to infer data values or to prune the search

**Data transformations:** The system models the data cleaning process as a data transformation graph
Criteria for comparing research data cleaning tools (3)

**User Interface:**
- **Graphical Interface:** the system provides a visualizing tool and menus to interact
- **User edition:** the system allows the user to edit data values

**Others:**
- **Scalability:** the system execution time grows linearly with the number of input tuples
- **Streaming:** the system receives tuples and processes each of them individually (opposed to batch processing)
- **Extensible:** the system allows the user to modify and/or insert new algorithms

Research Data cleaning tools (2014) (1/2)

<table>
<thead>
<tr>
<th>Tools</th>
<th>User Interface</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Graphical Interface</td>
<td>User edition</td>
</tr>
<tr>
<td>Cleenex</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Lunatic</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Nadeef</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>Guided data repair</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>Scare</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Eracer</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Continuous data cleaning</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>
Outline

- Introduction to data cleaning
- Application contexts of data cleaning
- Data quality dimensions
- Taxonomy of data quality problems
- Data quality process
- Main data quality tools
  - Real-world examples

Death by Typo

‘Resurrected,’ but still wallowing in red tape
Government records incorrectly kill off thousands, and there’s no easy fix

By Alex Johnson and Nancy Amosa
Reporters
NBC News
updated 6:31 p.m. ET Feb. 29, 2008

For a dead woman, Laura Todd is awfully articulate.

“I don’t think people realize how difficult it is to be dead when you’re not,” said Todd, who is very much alive and kicking in Nashville, Tenn., even though the federal government has said otherwise for many years.

Todd’s struggle started eight years ago with a typo in government records. The government has reassured her numerous times that it has cleaned up the confusion, but the problems keep coming.

Story continues below...
<table>
<thead>
<tr>
<th>Rank</th>
<th>Query</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Britney Spears</td>
<td>48,981</td>
</tr>
<tr>
<td>2</td>
<td>Britney spears</td>
<td>40,134</td>
</tr>
<tr>
<td>3</td>
<td>Brittny spears</td>
<td>36,515</td>
</tr>
<tr>
<td>4</td>
<td>Brittny Spears</td>
<td>24,342</td>
</tr>
<tr>
<td>5</td>
<td>Brittny Spears</td>
<td>7,331</td>
</tr>
<tr>
<td>6</td>
<td>Brittny Spears</td>
<td>5,933</td>
</tr>
<tr>
<td>7</td>
<td>Brittny Spears</td>
<td>2,994</td>
</tr>
<tr>
<td>8</td>
<td>Brittny Spears</td>
<td>1,987</td>
</tr>
<tr>
<td>9</td>
<td>Brittny Spears</td>
<td>5,933</td>
</tr>
<tr>
<td>10</td>
<td>Brittny Spears</td>
<td>1,478</td>
</tr>
<tr>
<td>11</td>
<td>Brittny Spears</td>
<td>1,427</td>
</tr>
<tr>
<td>12</td>
<td>Brittny Spears</td>
<td>1,318</td>
</tr>
<tr>
<td>13</td>
<td>Brittny Spears</td>
<td>1,261</td>
</tr>
<tr>
<td>14</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>15</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>16</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>17</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>18</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>19</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>20</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>21</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>22</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>23</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>24</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>25</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>26</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>27</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>28</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>29</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
<tr>
<td>30</td>
<td>Brittny Spears</td>
<td>990</td>
</tr>
</tbody>
</table>

Source: [http://www.google.com/jobs/britney.html](http://www.google.com/jobs/britney.html)

---

**Directmarketing by The Economist**
FIFA registration form (2010)

German Umlaute

Search Results for 'dessloch'

• Stefan Dickschock
• Stefan Deslouch
Next lecture

- Data Matching