Introduction to Data Cleaning

Helena Galhardas
DEI/IST

Follow me on LinkedIn for more :
Steve Nouri
https://www.linkedin.com/in/stevenouri/

1

References

- No single reference!
- "Data Quality: Concepts, Methodologies and Techniques", C. Batini and M. Scannapieco, Springer-Verlag, 2006 (Chapts. 1, 2, and 4)
- Slides "Data Quality and Data Cleansing" course, Felix Naumann, Winter 2014/15
- "Foundations of Data Quality Management", W. Fan and F. Geerts, 2012
- Oliveira, P. (2009). "Detecção e correcção de problemas de qualidade de dados: Modelo, Sintaxe e Semântica". PhD thesis, U. do Minho.

2

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

3

Example (1)

Table R

Name	SSN	Addr
Jack Lemmon	430-871-8294	Maple St
Harrison Ford	292-918-2913	Culver Blvd
Tom Hanks	234-762-1234	Main St
…	…	…

Table S

Name	SSN	Addr
Ton Hanks	234-162-1234	Main Street
Kevin Spacey	-	Frost Blvd
Jack Lemon	430-817-8294	Maple Street
…	…	…

- 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>

and

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

are the same object?

Example (3)

P. Bernstein, D. Chiu: Using Semi-Joins to Solve Relational Queries. JACM 28(1): 25-40(1981)

Philip A. Bernstein, Dah-Ming W. Chiu, Using Semi-Joins to Solve Relational Queries, Journal of the ACM (JACM), v.28 n.1, p.25-40, Jan. 1981

- 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

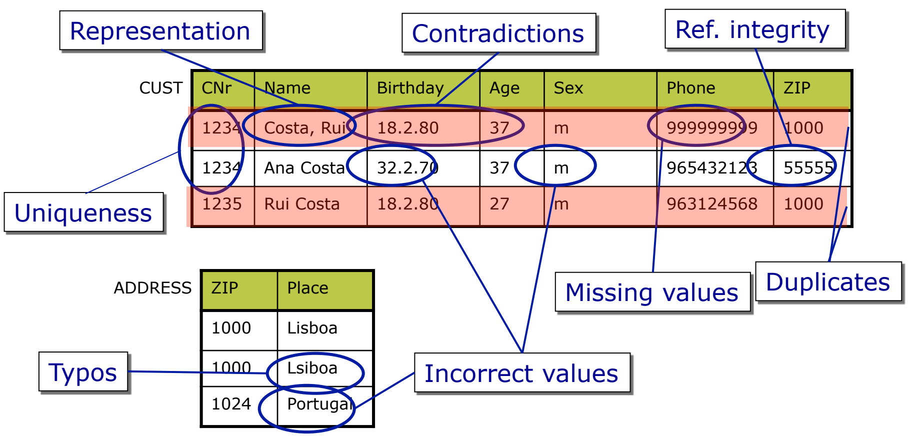
- Introduction to data cleaning
- Application contexts of data cleaning
- Data quality dimensions
- Taxonomy of data quality problems
- Data quality process
- Main data quality tools
- Real-world examples

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)



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]


IRS might be after you — to mail you a check

Incorrect addresses stall nearly 1,500 Tennessee refunds

By BONNA de la CRUZ
Staff Writer

Now that Tilcia L. Menifee knows that she'll be getting $500 in a tax refund from Uncle Sam, she can do some Christmas shopping, she said.

11

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

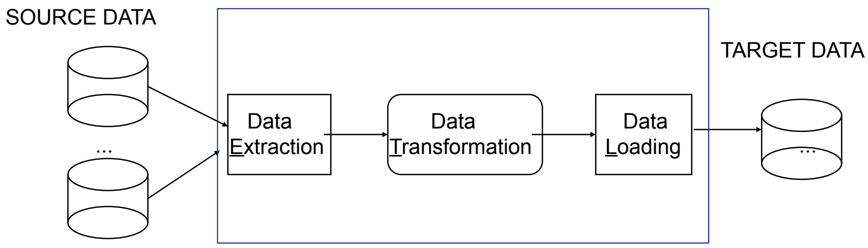
12

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

13

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

14

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…

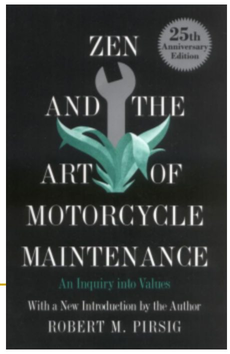
High-quality data!

- 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."

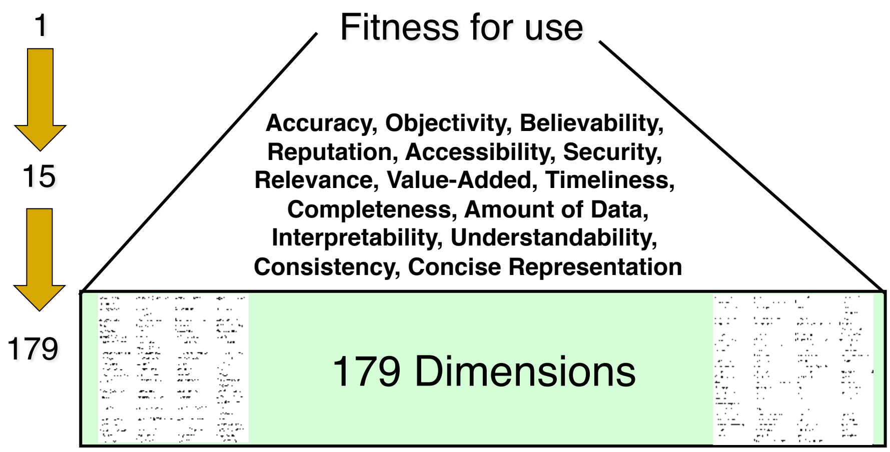
Robert Pirsig



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?



Category	IQ Criteria	TDQM	MBIS	Weikum	DWQ	SCOUG	Chen
Content-related Criteria	Accuracy	Yes	Yes	Yes	Yes	Yes	Yes
	Documentation					Yes	
	Relevancy	Yes	Yes		Yes		Yes
	Value-Added	Yes				Yes	
	Completeness	Yes	Yes	Yes	Yes	Yes	Yes
	Interpretability	Yes			Yes		
Technical Criteria	Timeliness	Yes	Yes	Yes	Yes	Yes	Yes
	Reliability			Yes			
	Latency			Yes			Yes
	Performability			Yes		Yes	
	Response time		Yes	Yes			Yes
	Security	Yes		Yes	Yes		
	Accessibility	Yes	Yes	Yes	Yes	Yes	
	Price		Yes	Yes		Yes	
	Customer Support					Yes	
Intellectual Criteria	Believability	Yes	Yes	Yes	Yes	Yes	
	Reputation	Yes	Yes		Yes		
	Objectivity	Yes					
Instantiation related Criteria	Verifiability			Yes			
	Amount of data	Yes	Yes				Yes
	Understandability	Yes	Yes				
	Concise represent.	Yes					
	Consistent represent.	Yes	Yes	Yes	Yes	Yes	

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

Syntatic 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

Ganularity 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

23

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

24

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

25

Metrics for quantifying completeness (1)

- Model without null values with OWA
  - Needs a reference relation `ref(r)` for a relation `r`, that contains all the tuples that satisfy the schema of r

    `C(r) = |r|/|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

26

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)

27

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.

28

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.

29

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

30

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 availavle.

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

33

Taxonomy of data quality problems [Oliveira 2009]

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

34

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 contraint 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
    - Toke-based
    - Phonetic algorithms
  - Record matching
    - Rule-based
    - Probabilistic
    - ...

Nome	Cod.Postal	Localidade
Maria	2765	Estoril
António	2765	S.João Esoril
José	2780	Oeiras
Andreia	1000	Lisboa
Manuela	2865	Setúbal

Localidade=>Cod.Postal

43

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 enrichement– use of additional information to improve data quality
- Main goal:
  - To correct the data quality problems detected during the data quality auditing process

44

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 CleaningTasks

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)

Tools	Debugger: Data lineage	Debugger: Breakpoints	Debugger: Edit values	Profiling: Rules	Profiling: Filters	Execution: User involvement	Execution: Incremental updates
Informatica PowerCenter	Y	Y	Y	Y	Y	N	Y
IBM Information Server	Y	Y	N	Y	Y	N	Y
Talend Open Studio	N	Y	N	Y	Y	N	Y
Oracle Data Integrator	Y	Y	N	Y	Y	N	Y
SQL Server Integration Services	Y	Y	N	Y	N	N	Y
SAS Data Integration Studio	Y	N	N	Y	Y	N	Y
Pentaho Data Integration	N	N	N	Y	N	N	Y
Clover ETL	N	N	Y	Y	Y	N	Y

55

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

56

Commercial Data cleaning tools (2014) (2/3)

Tools	Extensibility: Create Operators	Extensibility: Modify Operators	User Interface: Drag and Drop	User Interface: Grahical Editor
Informatica PowerCenter	Y (Java)	N	Y	Y
IBM Information Server	Y (Java)	N	Y	Y
Talend Open Studio	Y (Java, Groovy)	Y (Java)	Y	Y
Oracle Data Integrator	Y	Y	Y	Y
SQL Server Integration Services	Y (C#, VB)	N	Y	Y
SAS Data Integration Studio	Y (SAS)	Y (SAS)	Y	Y
Pentaho Data Integration	Y (Javascript)	N	Y	Y
Clover ETL	Y (CTL)	N	Y	Y

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)

Tools	Scalability			Others
	Grid	Partitioning	Pushdown Optimization	Free version
Informatica PowerCenter	Y	Y	Y	Y
IBM Information Server	Y	Y	Y	N
Talend Open Studio	Y	N	Optional ELT	Y
Oracle Data Integrator	Y	N	ELT	Y
SQL Server Integration Services	N	Y	-	Y (IST)
SAS Data Integration Studio	Y	Y	Y	Y (IST)
Pentaho Data Integration	Y	Y	N	Y
Clover ETL	Y	Y	N	Y

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

Tools	Detection DQ problems		Repair DQ problems		
	Constraints	Satistical	Search	ML/St	Data Transformations
Cleenex	QCs	N	N	N	Y
Llunatic	Egds	N	Y	N	N
Nadeef	CFDs, MDs	N	Y	N	N
Guided data repair	CFDs	N	Y	Y	N
Scare	N	Y	N	Y	N
Eracer	N	Y	N	Y	N
Continuous data cleaning	FDs	N	Y	Y	N

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

61

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

62

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 treat them indivually (opposed to batch processing)

Extensible: the system allows the user to modify and/or insert new algorithms

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

	User Interface		Others		
Tools	Graphical Interface	User edition	Extensible	Streaming	Scalability
Cleenex	Y	Y	Matching algorithms	N	N
Llunatic	Y	Y	Cost Managers	N	Y
Nadeef	Y	N	Repair algorithms	N	N
Guided data repair	N	Y	N	N	N
Scare	N	N	N	N	Y
Eracer	N	N	N	N	N
Continuous data cleaning	N	N	N	Y	Y

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 Amons
Reporters
MSNBC and NBC News
updated 6:21 p.m. ET Feb. 29, 2008

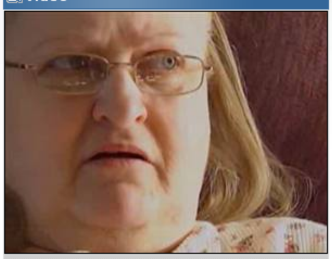
For a dead woman, Laura Todd is awfully articulate.

"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 cleared up the confusion, but the problems keep coming.

Story continues below ↓

Video



Launch

Does this woman look dead to you?

The government says Toni Anderson is dead, but she insists she is very much alive. David MacAnally of NBC affiliate WTHR reports from Muncie, Ind.

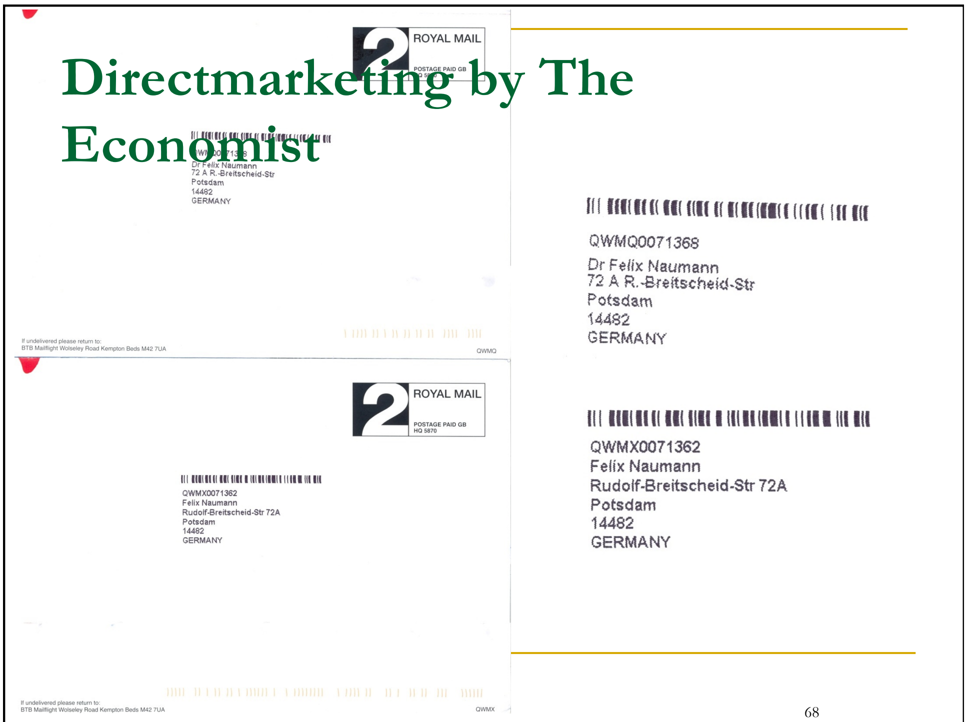
NBC News Channel

Google searches for Britney Spears

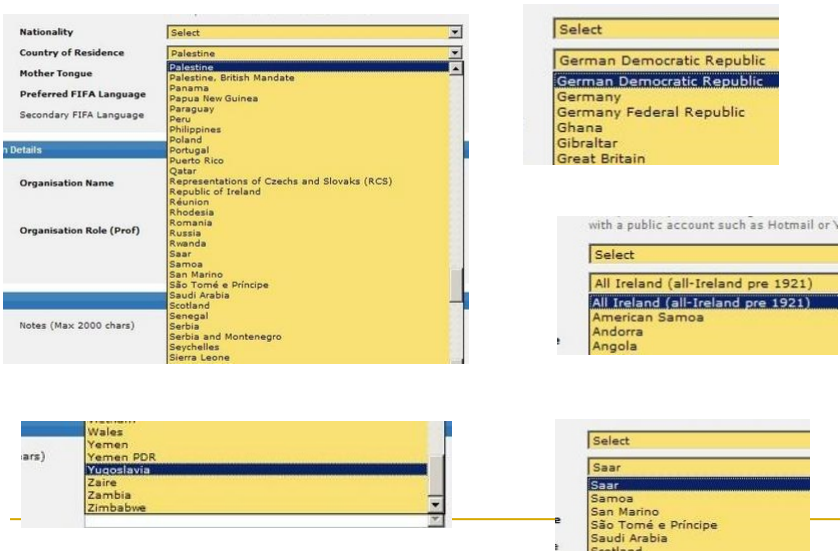
488941 britney spears	29 britent spears	9 brinttany spears	5 brney spears	3 britiy spears	2 brirreny spears
40134 brittany spears	29 brittnany spears	9 britanay spears	5 broitney spears	3 britmeny spears	2 brirtany spears
36315 brittney spears	29 britttany spears	9 britinany spears	5 brotny spears	3 britneeey spears	2 brirttany spears
24342 britany spears	29 btiney spears	9 britn spears	5 bruteny spears	3 britnehy spears	2 brirttney spears
7331 britny spears	26 birttney spears	9 britnew spears	5 btiyney spears	3 britnely spears	2 britain spears
6633 briteny spears	26 breitney spears	9 britneyn spears	5 btrittney spears	3 britnesy spears	2 britane spears
2696 britteny spears	26 brinity spears	9 britrney spears	5 gritney spears	3 britnetty spears	2 britaneny spears
1807 briney spears	26 britenay spears	9 brtiny spears	5 spritney spears	3 britnex spears	2 britania spears
1635 brittny spears	26 britneyt spears	9 brtittney spears	4 bittny spears	3 britneyxxx spears	2 britann spears
1479 brintey spears	26 brittan spears	9 brtny spears	4 bnritney spears	3 britnity spears	2 britanna spears
1479 britanny spears	26 brittne spears	9 brytny spears	4 brandy spears	3 britntey spears	2 britannie spears
1338 britiny spears	26 btittany spears	9 rbitney spears	4 brbritney spears	3 britnyey spears	2 britannt spears
1211 britnet spears	24 beitney spears	8 birtiny spears	4 breatiny spears	3 britterny spears	2 britannu spears
1096 britiney spears	24 birteny spears	8 bithney spears	4 breetney spears	3 brittneey spears	2 britanyl spears
991 britaney spears	24 brightney spears	8 brattany spears	4 bretiney spears	3 brittnney spears	2 britanyt spears
991 britnay spears	24 brintiny spears	8 breitny spears	4 brfitney spears	3 brittnyey spears	2 briteeny spears
811 brithney spears	24 britanty spears	8 breteny spears	4 briattany spears	3 brityen spears	2 britenany spears
811 brtiney spears	24 britenny spears	8 brightny spears	4 brieteny spears	3 briytney spears	2 britenet spears
664 birtney spears	24 britini spears	8 brintay spears	4 briety spears	3 brltney spears	2 briteniy spears
664 brintney spears	24 britnwy spears	8 brinttey spears	4 briitny spears	3 broteny spears	2 britenys spears
664 briteney spears	24 brittni spears	8 briotney spears	4 briittany spears	3 brtaney spears	2 britianey spears
601 bitney spears	24 brittnie spears	8 britanys spears	4 brinie spears	3 brtiiany spears	2 britin spears
601 brinty spears	21 biritney spears	8 britley spears	4 brinteney spears	3 brtinay spears	2 britinary spears
544 brittaney spears	21 birtany spears	8 britneyb spears	4 brintne spears	3 brtinney spears	2 britmy spears
544 brittnay spears	21 biteny spears	8 britnrey spears	4 britaby spears	3 brtitany spears	2 britnaney spears
364 britey spears	21 bratney spears	8 britnty spears	4 britaey spears	3 brtiteny spears	2 britnat spears
364 brittiny spears	21 britani spears	8 brittner spears	4 britainey spears	3 brtnet spears	2 britnbey spears
329 brtney spears	21 britanie spears	8 brottany spears	4 britinie spears	3 brytiny spears	2 britndy spears
269 bretney spears	21 briteany spears	7 baritney spears	4 britinney spears	3 btney spears	2 britneh spears
269 britneys spears	21 brittay spears	7 birntey spears	4 britmney spears	3 drittney spears	2 britneney spears
244 britne spears	21 brittinay spears	7 biteney spears	4 britnear spears	3 pretney spears	2 britney6 spears
244 brytney spears	21 brtany spears	7 bitiny spears	4 britnel spears	3 rbritney spears	2 britneye spears
220 breatney spears	21 brtiany spears	7 breateny spears	4 britneuy spears	2 barittany spears	2 britneyh spears
220 britiany spears	19 birney spears	7 brianty spears	4 britnewy spears	2 bbbritney spears	2 britneym spears
199 britnney spears	19 brirtney spears	7 brintye spears	4 britnmey spears	2 bbitney spears	2 britneyyy spears
163 britnry spears	19 britnaey spears	7 britianny spears	4 brittaby spears	2 bbritny spears	2 britnhey spears
147 breatny spears	19 britnee spears	7 britly spears	4 brittery spear		
147 brittiney spears	19 britony spears	7 britnej spears	4 britthey spea		
147 britty spears	19 brittanty spears	7 britneyu spears	4 brittnaey spears		

Source: http://www.google.com/jobs/britney.html

Directmarketing by The Economist



FIFA registration form (2010)



German Umlaute

Next lecture

- Data Matching

Follow me on LinkedIn for more :
Steve Nouri
https://www.linkedin.com/in/stevenouri/