Data Quality: Where are we on the journey from theory to practice?

Angela Bonifati

University of Lyon 1
Liris – CNRS, France

June 23, 2017

Table of contents

Quality for Big Data

In **Big Data**, quantity is often more emphasized than quality:

- scalable algorithms to compute query answers Q(D) when database D is large
- however, can we trust Q(D) as correct answers?

Quality for Big Data

In **Big Data**, quantity is often more emphasized than quality:

- scalable algorithms to compute query answers Q(D) when database D is large
- however, can we trust Q(D) as correct answers?
- quality is as important as quantity in big data management

Real life is flawed, inaccurate and inconsistent

- More than 25 % of critical data in the world's top companies[1] is flawed
- Pieces of information perceived as being needed for clinical decisions[2] are missing from 13.6% to 81% of the time
- 2% of records in a customer file become obsolete in one month
- Hence, in a customer database[3], 50% of its records may be obsolete and inaccurate within two years.

[1] 'Dirty Data' is a Business Problem, Not an IT Problem, *Gartner*.

[2] D. W. Miller Jr., J. D. Yeast, and R. L. Evans. Missing prenatal records at a birth center: A communication problem quantified. *In AMIA, 2005.*

[3] W. W. Eckerson. Data quality and the bottom line: Achieving business success through a commitment to high quality data. *TR, The Data Warehousing Institute, 2002.*

Cost of poor-quality data

- Statistics shows that "bad data or poor data quality costs US businesses \$600 billion annually"[1]
- "poor data can cost businesses 20%-35% of their operating revenue"[2]
- "poor data across businesses and the government costs the US economy \$3.1 trillion a year"
- for Big Data, the scale of the data quality problem is historically unprecedented.

[1] W. W. Eckerson. Data quality and the bottom line: Achieving business success through a commitment to high quality data. *TR, The Data Warehousing Institute, 2002.*

[2] Wikibon. A comprehensive list of big data statistics, 2012.

Error types: an Employee Dataset T_1

Constraint violation

ID	FN	LN	ROLE	ZIP	ST	SAL
105	Anne	Nash	Emp	85281	NY	110
211	Mark	White	Man	15544	NY	80
386	Mark	Lee	M	85281	AZ	75
215	Anna	Smith Nash	Emp	85283		1

Duplicates

Pattern Violation

Outliers

Query: Find the FN, LN and SAL of distinct employees working in NYC

- The answer is: “Anne Nash 110”, “Mark White 80”

Query: Find the FN, LN and SAL of distinct employees working in NYC

- The answer is: “Anne Nash 110”, “Mark White 80”
- Can we trust this answer?

Query: Find the FN, LN and SAL of distinct employees working in NYC

- The answer is: "Anne Nash 110", "Mark White 80"
- Can we trust this answer?
- If zip code of NYC is 85281, then also "Mark Lee 75" is part of the answer.

Query: Find the FN, LN and SAL of distinct employees working in NYC

- The answer is: "Anne Nash 110", "Mark White 80"
- Can we trust this answer?
- If zip code of NYC is 85281, then also "Mark Lee 75" is part of the answer.
- "Anne Nash" and "Anne Smith Nash" may be the same person (which salary can we trust?)

Foundations of Data Quality: Data Consistency[1]

- Data consistency refers to the validity and integrity of data
- It aims to detect errors typically identified as violations of data dependencies

[1] Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Foundations of Data Quality: Data Consistency[1]

- Data consistency refers to the validity and integrity of data
- It aims to detect errors typically identified as violations of data dependencies
- There are at least two questions associated with data consistency:
  - What data dependencies should we use to detect errors?
  - What repair model do we adopt to fix the errors?

[1] Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Foundations of Data Quality: Data Consistency[1]

- Data consistency refers to the validity and integrity of data
- It aims to detect errors typically identified as violations of data dependencies

[1] Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Foundations of Data Quality: Data Consistency[1]

- Data consistency refers to the validity and integrity of data
- It aims to detect errors typically identified as violations of data dependencies
- There are at least two questions associated with data consistency:
  - What data dependencies should we use to detect errors?
  - What repair model do we adopt to fix the errors?

[1] Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Dependencies for Data Consistency

- Functional Dependencies (FDs) of the kind $A \rightarrow B$, where A and B are attributes of a relation R (e.g. $zip \rightarrow state$ in T_1);
- Conditional Functional Dependencies (CFDs) that extends FDs with pattern tableaux;

Dependencies for Data Consistency

- Functional Dependencies (FDs) of the kind $A \rightarrow B$, where A and B are attributes of a relation R (e.g. *zip* $\rightarrow$ *state* in T_1);
- Conditional Functional Dependencies (CFDs) that extends FDs with pattern tableaux;
- Denial Constraints (DCs) of the kind $\forall \overline{x}\ \neg(\psi(\overline{x}) \wedge \beta(\overline{x}))$, where $\psi(\overline{x})$ is a non-empty conjunction of relational atoms and $\beta(\overline{x})$ a conjunction of built-in predicates $=, \neq, <, >, \leqslant, \geqslant$
- Equality-generating dependencies (EGDs) $\forall \overline{x}\ (\psi(\overline{x}) \rightarrow (x_1 = x_2))$ as a particular case of DCs (and, btw, FDs are a special case of EGDs);

Dependencies for Data Consistency

- Functional Dependencies (FDs) of the kind $A \rightarrow B$, where A and B are attributes of a relation R (e.g. *zip* $\rightarrow$ *state* in T_1);
- Conditional Functional Dependencies (CFDs) that extends FDs with pattern tableaux;
- Denial Constraints (DCs) of the kind $\forall \overline{x}\ \neg(\psi(\overline{x}) \wedge \beta(\overline{x}))$, where $\psi(\overline{x})$ is a non-empty conjunction of relational atoms and $\beta(\overline{x})$ a conjunction of built-in predicates $=, \neq, <, >, \leqslant, \geqslant$
- Equality-generating dependencies (EGDs) $\forall \overline{x}\ (\psi(\overline{x}) \rightarrow (x_1 = x_2))$ as a particular case of DCs (and, btw, FDs are a special case of EGDs);
- Tuple-generating dependencies (TGDs) of the kind $\forall \overline{x}\ (\phi(\overline{x}) \rightarrow \exists \overline{y}\ \psi(\overline{x}, \overline{y}))$ where $\phi(\overline{x})$ and $\psi(\overline{x}, \overline{y})$ are conjunctions of relational atoms over $\overline{x}$ and $\overline{x} \cup \overline{y}$, resp. (subsume inclusion dependencies INDs).

Satisfiability Problem for a Class of Dependencies $\mathcal{C}$

- For a class $\mathcal{C}$ of dependencies and $\phi \in \mathcal{C}$, the satisfiability problem for C is to decide:
  - given a finite set $\Sigma \subseteq \mathcal{C}$ defined on a relational schema R, whether there exists a nonempty finite instance D of R such that $D \models \Sigma$.
  - That is, whether the data quality rules in Σ are consistent themselves.

Implication Problem for a Class of Dependencies $\mathcal{C}$

- For a class $\Sigma \subseteq \mathcal{C}$ of dependencies and $\phi \in \mathcal{C}$, the implication problem for C is to decide:
  - given a finite set $\Sigma \subseteq \mathcal{C}$ and $\phi \in \mathcal{C}$ defined on a relational schema R, whether $\Sigma \models \phi$.
  - That is, whether data quality rules in Σ can be removed to speed up error detection and data repairing.

Complexity of satisfiability and implication analysis

dependencies	**satisfiability**	**implication**
CFDs	NP-complete	coNP-complete
FDs	O(1)	O(n)
CINDs	O(1)	EXPTIME-complete
INDs	O(1)	PSPACE-complete
CFDs + CINDs	undecidable	undecidable
FDs + INDs	O(1)	undecidable

Foundations of Data Quality: Data Consistency[1]

- Data consistency refers to the validity and integrity of data
- It aims to detect errors typically identified as violations of data dependencies

[1]Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Foundations of Data Quality: Data Consistency[1]

- Data consistency refers to the validity and integrity of data
- It aims to detect errors typically identified as violations of data dependencies
- There are at least two questions associated with data consistency:
  - What data dependencies should we use to detect errors?
  - What repair model do we adopt to fix the errors?

[1]Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Repair models[1]

- S-repair: assuming that D is inconsistent but complete, it allows repairs with tuple deletions only;
- C-repair: assuming that D is inconsistent and incomplete, it allows repairs with tuple insertions and deletions;
- CC-repair: looking for a C-repair that is minimal wrt. all possible repairs;
- U-repair: it supports attribute value modifications.

[1] Wenfei Fan: Data Quality: From Theory to Practice. *Sigmod Record, 2015.*

Foundations of Data Quality: Data Deduplication

- Data deduplication (or Record Matching) refers to identifying tuples from one or more relations that refer to the same real-world entity:
  - Given an instance D of R, a set E of entity types, a set X of attributes of R, data deduplication is to determine,
  - for all tuples t,t' in D, and for each entity type e[X], whether t[X] and t'[X] refer to the same entity of type e.

Foundations of Data Quality: Data Deduplication

- Data deduplication (or Record Matching) refers to identifying tuples from one or more relations that refer to the same real-world entity:
  - Given an instance D of R, a set E of entity types, a set X of attributes of R, data deduplication is to determine,
  - for all tuples t,t' in D, and for each entity type e[X], whether t[X] and t'[X] refer to the same entity of type e.
- There are different approaches:
  - rule-based (in this talk), probabilistic,
  - learning-based and distance-based.
- Problem: sources can be unreliable or prone to become dirty after their integration.

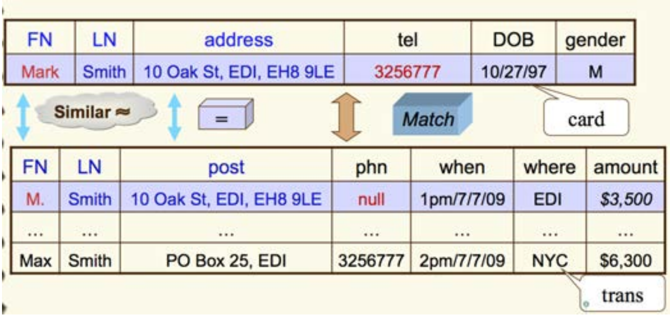
Record matching: an example

FN	LN	address	tel	DOB	gender
Mark	Smith	10 Oak St, EDI, EH8 9LE	3256777	10/27/97	M

the same person?

FN	LN	post	phn	when	where	amount
M.	Smith	10 Oak St, EDI, EH8 9LE	null	1pm/7/7/09	EDI	*$3,500*
...	...	...	...	...	...	...
Max	Smith	PO Box 25, EDI	3256777	2pm/7/7/09	NYC	$6,300

Matching Rules



FN	LN	address	tel	DOB	gender
Mark	Smith	10 Oak St, EDI, EH8 9LE	3256777	10/27/97	M

FN	LN	post	phn	when	where	amount
M.	Smith	10 Oak St, EDI, EH8 9LE	null	1pm/7/7/09	EDI	$3,500
...	...	...	...	...	...	...
Max	Smith	PO Box 25, EDI	3256777	2pm/7/7/09	NYC	$6,300

- IF card[LN, address] = trans[LN, post] AND card[FN] and trans[FN] are similar, THEN identify the two tuples
- In logics: card[LN,address]=trans[LN,post] $\wedge$ card[FN] $\approx$ trans[FN] $\Longrightarrow$ card[X] $\Leftrightarrow$ trans[Y]

Error types: an Employee Dataset T_1 (cont'd)

Constraint violation

ID	FN	LN	ROLE	ZIP	ST	SAL
105	Anne	Nash	Emp	85281	NY	110
211	Mark	White	Man	15544	NY	80
386	Mark	Lee	M	85281	AZ	75
215	Anna	Smith Nash	Emp	85283		1

Duplicates

Pattern Violation

Outliers

Error detection strategies

- Rule-based detection algorithms
- Deduplication
- Pattern verification and enforcement tools
  - Syntactical patterns, such as date formatting
  - Semantical patterns, such as location names
- Quantitative algorithms
  - Statistical outliers

How do existing tools cover the various error types?

	DBoost	DC-Clean	OpenRefine	Trifacta	Pentaho	KNIME	Katara	Tamr
Pattern violations			✔	✔	✔	✔	✔	
Constraint violations		✔						
Outliers	✔							
Duplicates								✔

Comparative analysis of DQ tools on real datasets[1]

- Previous studies focused on synthetic datasets or real-world datasets with artificially injected errors
- However, the effectiveness of these tools on real-world data 'in the wild' is unclear
- Real data often contains multiple errors (duplicates plus IC violation etc.)
- All tools assume considerable human involvement, which is costly
- A comparative analysis of the above tools on various real datasets is carried out:
  - What is the precision and recall of each tool?
  - How many errors in the data sets are detectable by applying all the tools combined?
  - Is there a strategy to minimize human effort by leveraging the interactions among the tools?

[1] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, Nan Tang: Detecting Data Errors: Where are we and what needs to be done? *PVLDB'16*.

Towards real data

Dataset	# columns	# rows	# rows ground truth	Errors
MIT VPF	42	24K	13k (partial)	6.7%
Merck	**61**	2262	2262	19.7%
Animal	14	60k	60k	0.1%
Rayyan Bib	11	**1M**	1k (partial)	**35%**
BlackOak	12	94k	**94k**	34%

	MIT VPF	Merck	Animal	Rayyan Bib	BlackOak
Pattern violations	✔	✔	✔	✔	✔
Constraint violations	✔	✔	✔	✔	✔
Outliers	✔	✔		✔	✔
Duplicates	✔				✔

Lessons learned

- The conclusion of the previous study was that there is no single dominant tool.
- Various tools worked well on different data sets.
- A holistic composite strategy must be used in any practical environment.
- However, the combined overall recall is well less than 100% (even with ad-hoc cleaning service and enrichment process).
- Thus, need to develop new ways of finding data errors that can be spotted by humans.
- Cons: no real scientific data (except for Animal).

CNRS Mastodons MedClean (2016-ongoing) [1]

Title

Nettoyage et transformation virtuels des grandes masses de donneés médicales et de sciences du vivant

Partners

- **Liris**, University of Lyon 1 (A. Bonifati, E. Coquery, M. S. Hacid, R. Thion)
- **Limos**, Blaise Pascal University (F. Toumani, M. Bouet, R. Ciucanu)
- **Lipade**, Paris Descartes University (S. Benbernou, I. Ileana, M. Ouziri. S. Sahiri)
- **HEGP** (A. Burgun, A. S. Janot, B. Rance)
- **Institut Cochin**, Inserm & INSB CNRS (P. Bourdoncle, T. Guilbert, A. Trautman)

[1] https://liris.cnrs.fr/medclean/wiki/doku.php

Ongoing research objectives

Collection and annotation of datasets

- Two activities to be carried out in parallel: Clinical Data (HEGP), Biological Data (INSB).
- Complementary notions of data quality needed.
- Use-case driven understanding of the quality problems (upon image metadata for biological data, queries on clinical data).
- Real datasets (even though with confidentiality issues)
- For more details, please attend Bastien's talk in the afternoon.

Conclusions and Future Directions

Data Quality for Scientific Data

- **Data quality**: design of quality-aware algorithms for scientific datasets.
- **Lack of ground truth**: several open problems out there! cleaning is unfeasible

State-of-the-art and directions of research

- Existing large-scale data cleaning methods for relational databases, entity resolution for graphs....
- Combinations of data formats and additional error types: are we ready?