Data Quality Assurance

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Importance of Data Quality

UPFRONT Edited by Nikki Swartz
News, Trends & Analysis

Gartner Warns Firms of 'Dirty Data'

According to Gartner Inc., more than 25 percent of critical data in *Fortune* 1000 companies is flawed.

Speaking at the research and advisory firm's Business Intelligence and Information Management Summit held in Australia in February, Gartner Research Vice President Andreas Bitterer said that poor quality, or "dirty data," is often overlooked by businesses, but it can have a large negative impact on a firm.

"There is not a company on the planet that does not have a data quality problem," Bitterer said. "And where a company does recognize they

quality customer data can cost businesses dearly in terms of higher customer turnover and excessive expenses from customer contact processes

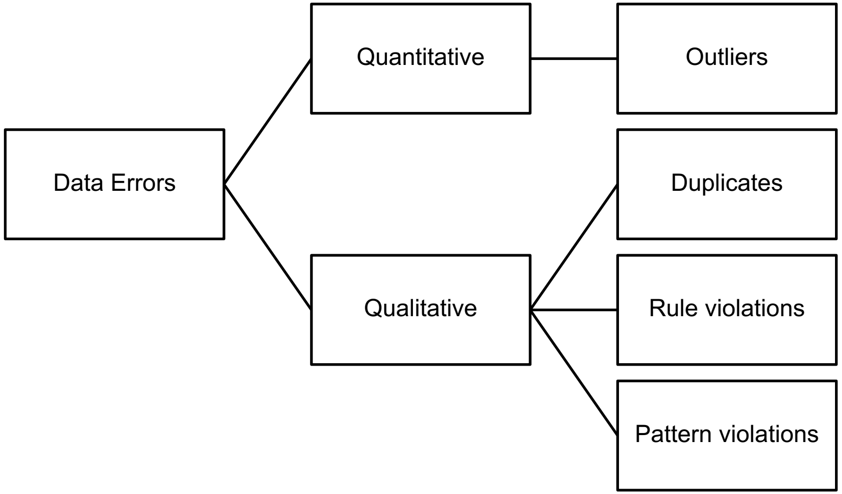
increased sales, lower distribution costs, and better compliance," Bitterer said.

One initiative companies should consider is appointing "data stewards," or people within the company who are responsible for the quality of its information. Firms should also manage information as a corporate asset. Bitterer said businesses also need to invest in technological data quality solutions that can help them profile, cleanse, match, and enrich critical information. Gartner said the market for data quality tools is currently small – $300 million (U.S.) in

Source: [1] Swartz (2007)

What is „dirty“ data?

„We define an error to be a deviation from its ground truth value."



1. **Outliers** include data values that deviate from the distribution of values in a column of a table.
2. **Duplicates** are distinct records that refer to the same real-world entity. If attribute values do not match, this could signify an error.
3. **Rule violations** refer to values that violate any kind of integrity constraints, such as Not Null constraints and Uniqueness constraints.
4. **Pattern violations** refer to values that violate syntactic and semantic constraints, such as alignment, formatting, misspelling, and semantic data types.

Source: [2] Abedjan et al. (2016)

Ways to clean data

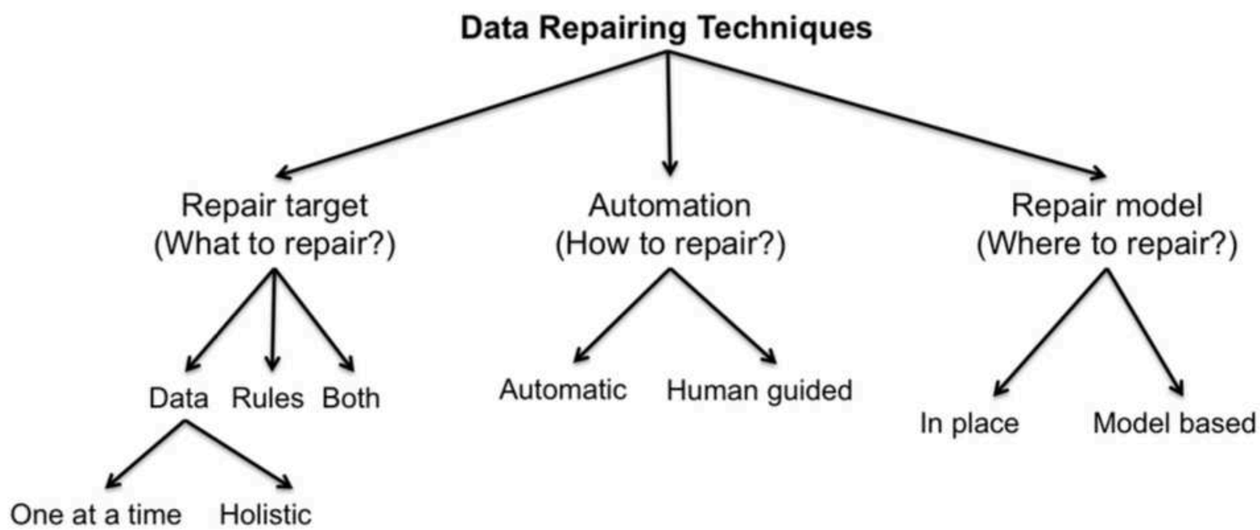


Figure 2: Classification of data repairing techniques.

Source: [3] Chu et al. (2016)

New / Emerging Challenges

TECHNISCHE UNIVERSITÄT DARMSTADT

Scalability

User Engagement

Semi-structured and unstructured data

New Applications for Streaming Data

Growing Privacy and Security Concerns

Source: [3] Chu et al. (2016)

Simplified Data Quality Assurance Process

Error Detection

- Data Linter
- Automating Large-Scale Data Quality Verification

Data Cleaning

- SampleClean, ActiveClean
- HoloClean

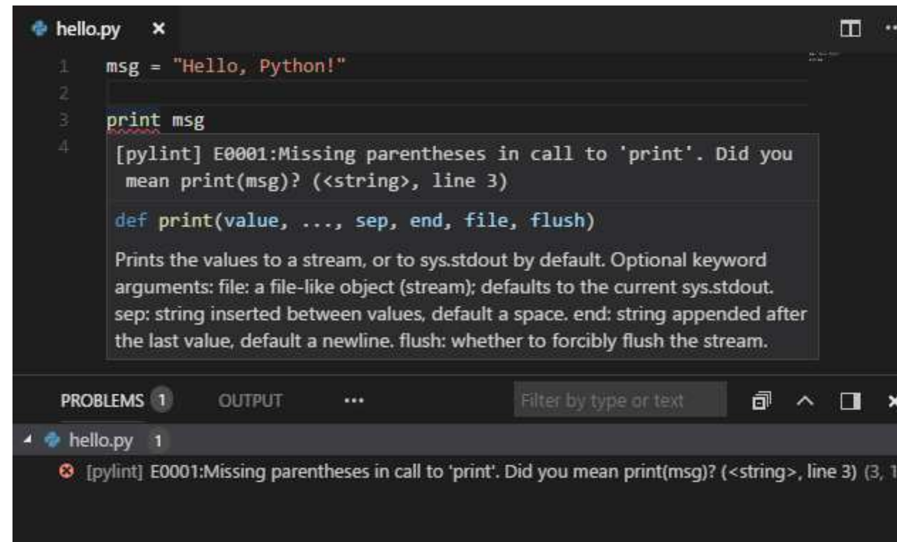
Evaluation

- CleanML

Data Linter 1/3

"[…] cleaning which, even when automated, is a time-consuming and error-prone process of repeated inspection and correction."

Data-linter: "[…] analyzes a user's training data and suggests ways features can be transformed to improve model quality, for a specific model type."



Source: [4] Hynes et al. (2017)

Data Linter 2/3

TECHNISCHE UNIVERSITÄT DARMSTADT

LintDetectors

corresponds to a specific issue to search for, for given model type

DataLinter

engine that applies the LintDetectors to a data set

LintExplorer

presents output to user

Lint Examples:

```
Enum as real: An enum (a categorical value) is encoded
as a real number. Consider converting to an integer and
using an embedding or one-hot vector.

Uncommon sign detector: The data includes some values
that have a different sign (+/−) from the rest of the
data (e.g., −9999), which can affect training. If these
are special markers in the data, consider replacing them
with a more neutral value (e.g., an empty or average
value).
```

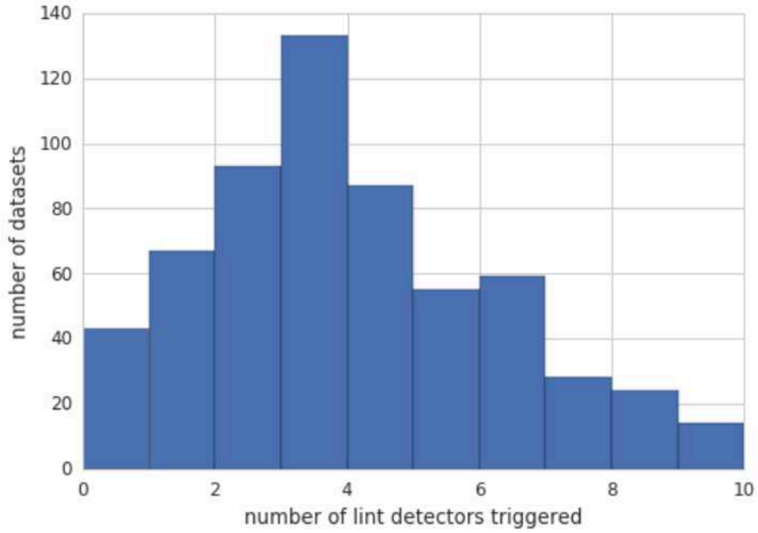
Source: [4] Hynes et al. (2017)

Data Linter 3/3

End-User Evaluation:

- led to a DNN model's precision increasing from 0.48 to 0.59
- after an initial model parameter tuning by engineer
- user was unaware of the benefits of normalizing inputs to a DNN
- so the tool also served as an educational aid

Data Set Evaluation:



Source: [4] Hynes et al. (2017)

Automatic Data Quality Verification 1/5

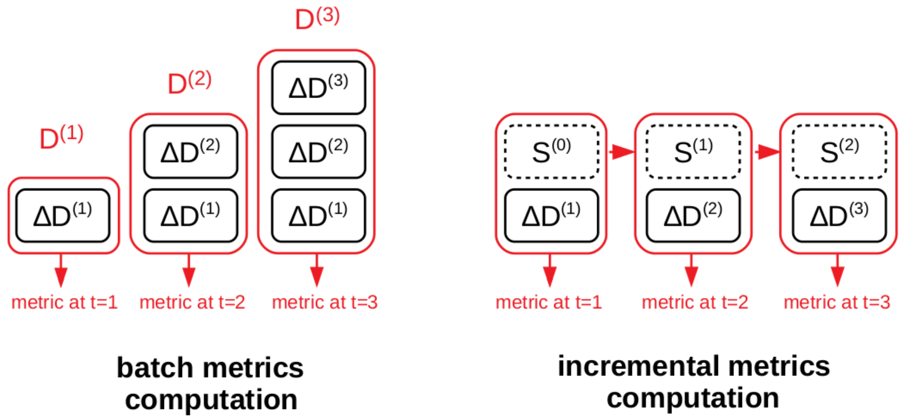
- Declarative API
  - User-defined “unit tests”
  - Combined with custom code

```
val numTitles = callRestService(...)
val maxExpectedPhoneRatio = computeRatio(...)

var checks = Array()

checks += Check(Level.Error)
 .isComplete("customerId", "title",
  "impressionStart", "impressionEnd",
  "deviceType", "priority")
 .isUnique("customerId", "countryResidence",
  "deviceType", "title")
 .hasCountDistinct("title", _ <= numTitles)
 .hasHistogramValues("deviceType",
   _.ratio("phone") <= maxExpectedPhoneRatio)
```

Source: [5] Schelter et al. (2017)

Automatic Data Quality Verification 2/5

- **Declarative**
  - Think about how data should look like
- **Incremental**
  - Support for growing data sets
  - Only needs new data set + state



Source: [5] Schelter et al. (2017)

Automatic Data Quality Verification 3/5

- Actual data quality verification
  - Compute required metrics
- Metrics provided by the tool:
  - Completeness
  - Consistency
  - Statistics
  - → used for consistency metrics

metric	semantic
dimension *completeness*	
`Completeness`	fraction of non-missing values in a column
dimension *consistency*	
`Size`	number of records
`Compliance`	ratio of columns matching predicate
`Uniqueness`	unique value ratio in a column
`Distinctness`	unique row ratio in a column
`ValueRange`	value range verification for a column
`DataType`	data type inference for a column
`Predictability`	predictability of values in a column
statistics (can be used to verify dimension *consistency*)	
`Minimum`	minimal value in a column
`Maximum`	maximal value in a column
`Mean`	mean value in a column
`StandardDeviation`	standard deviation of the value distribution in a column
`CountDistinct`	number of distinct values in a column
`ApproxCountDistinct`	number of distinct values in a column estimated by a hyperloglog sketch [21]
`ApproxQuantile`	approximate quantile of the value in a column [15]
`Correlation`	correlation between two columns
`Entropy`	entropy of the value distribution in a column
`Histogram`	histogram of an optionally binned column
`MutualInformation`	mutual information between two columns

Source: [5] Schelter et al. (2017)

Automatic Data Quality Verification 4/5

TECHNISCHE UNIVERSITÄT DARMSTADT

- Output
  - Fails and successes of constraints
  - “How much” a constraint failed

```
Success("isNonNegative(count)",
 Compliance("count >= 0") == 1.0)),
Failure("isUnique(customerId, countryResidence,
  deviceType, title)",
 Uniqueness("customerId", "countryResidence",
 "deviceType", "title") == 1.0, 0.9967),
```

Source: [5] Schelter et al. (2017)

Automatic Data Quality Verification 5/5

- Learnings
  - Advantages of using a shared data quality library
  - Reuse checks and constraints
  - Reduced manual work on data

Source: [5] Schelter et al. (2017)

Sample Clean

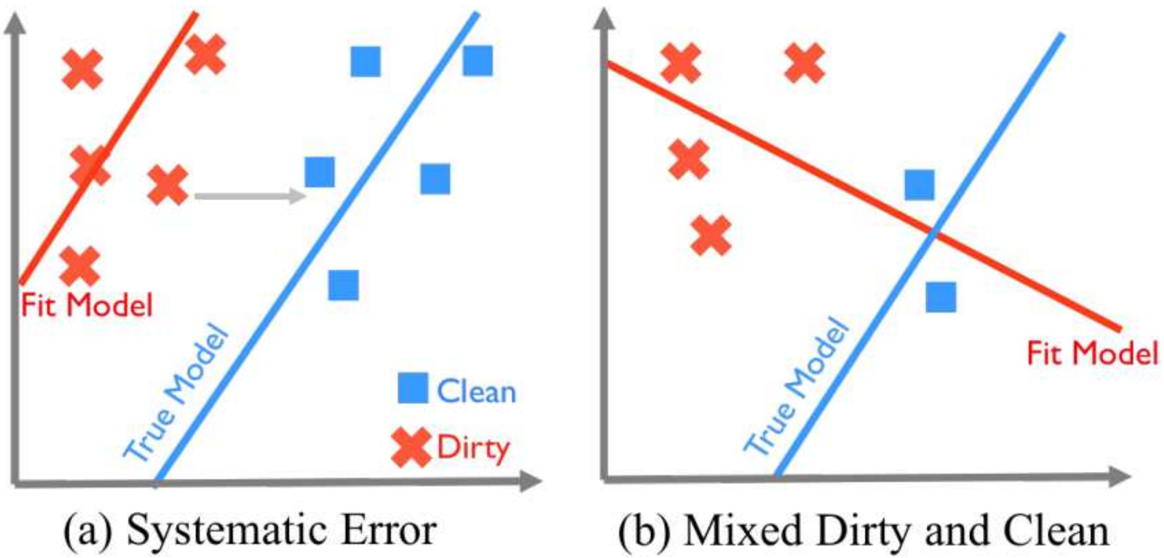
- Two error sources: dirty data and too little data
- Benefits of clean data outweigh error of from using less data

  → Only use a clean sample

Source: [6] Krishnan et al. (2015)

Simpson‘s Paradox

- Another problem: training on partially cleaned data



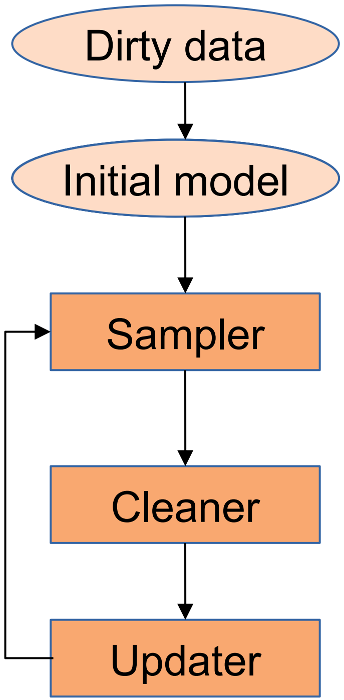
Source: [7] Krishnan et al. (2016)

Active Clean 1/2

- Extends Sample Clean
- Prevent the effects of partially cleaned data
- Use samples of cleaned data and integrate it into training of model

Source: [7] Krishnan et al. (2016)

Active Clean 2/2



1) Train on dirty data for initial model
2) Select sample records
3) Clean sample
4) Update weights of model (using cleaned sample)

Source: [7] Krishnan et al. (2016)

Holo Clean 1/4

- Two tasks of data cleaning
  - 1) Error detection → automation works fine
  - 2) Data cleaning → automation fails

Source: [8] Rekatsinas et al. (2017)

Holo Clean 2/4

- Qualitative data repairing
  - Integrity constraints
  - External information
- Quantitative Data repairing
  - Statistical methods

Source: [8] Rekatsinas et al. (2017)

Holo Clean 3/4

- Using them separately yields bad results
- Issue addressed by Holo Clean
  - Bad automation for data repairing
  - Solution: combine quantitative and qualitative data repairing

Source: [8] Rekatsinas et al. (2017)

Holo Clean 4/4

TECHNISCHE UNIVERSITÄT DARMSTADT

Dataset (τ)	Metric	HoloClean	Holistic	KATARA	SCARE
Hospital (0.5)	Prec.	**1.0**	0.517	0.983	0.667
	Rec.	**0.713**	0.376	0.235	0.534
	F1	**0.832**	0.435	0.379	0.593
Flights (0.3)	Prec.	**0.887**	0.0	n/a	0.569
	Rec.	**0.669**	0.0	n/a	0.057
	F1	**0.763**	0.0^{*}	n/a	0.104
Food (0.5)	Prec.	**0.769**	0.142	1.0	0.0
	Rec.	**0.798**	0.679	0.310	0.0
	F1	**0.783**	0.235	0.473	0.0^{+}
Physicians (0.7)	Prec.	**0.927**	0.521	0.0	0.0
	Rec.	**0.878**	0.504	0.0	0.0
	F1	**0.897**	0.512	$0.0^{\#}$	0.0^{+}

Source: [8] Rekatsinas et al. (2017)

CleanML 1/3

„**ML Community** has been focusing on understanding the impact of noises to ML models “

„**DB Community** has been focssing on understanding the fundamental process of data cleaning“

- Most of the real-world applications these problems do not occur on their own
- Common practice: ***Data cleaning followed by ML model training***
- Need of study the impact of cleaning on ML models
- Construct benchmarks to evaluate the impact

Source: [9] Li et al. (2019)

CleanML 2/3

Table 15: Benchmark Results(Organized by Query)

Q1

Q1(E=Inconsistencies)

R	P	S	N
R1	14.29% (8)	85.71% (48)	0%(0)
R2	25.0% (2)	75.0% (6)	0%(0)
R3	37.5% (3)	62.5% (5)	0%(0)

Q1(E=Duplicates)

R	P	S	N
R1	17.86% (10)	71.43% (40)	10.71% (6)
R2	12.5% (1)	62.5% (5)	25.0% (2)
R3	25.0% (2)	50.0% (4)	25.0% (2)

Q1(E=Mislabels)

R	P	S	N
R1	59.52% (75)	26.19% (33)	14.29% (18)
R2	61.11% (11)	27.78% (5)	11.11% (2)
R3	61.11% (11)	27.78% (5)	11.11% (2)

Q1(E=Outliers)

R	P	S	N
R1	31.55% (265)	57.02% (479)	11.43% (96)
R2	33.33% (40)	60% (72)	6.67% (8)
R3	30% (3)	70% (7)	0% (0)

Q1(E=Missing Values)

	P	S	N
R1	61.51% (155)	34.92% (88)	3.57% (9)
R2	50.00% (18)	50.00% (18)	0.00% (0)
R3	50.00% (3)	50.00% (3)	0.00% (0)

R1: *How does cleaning some type of error using a detection method and a repair method affect a ML model for a given dataset?*

R2: *How does cleaning some type of error using a detection method and a repair method affect the best ML model for a given dataset?*

R3: *How does the best cleaning method affect the predictive performance of the best model for a given dataset?*

Source: [9] Li et al. (2019)

CleanML 3/3

Conclusions:

- Data cleaning does not necessarily improve the quality of downstream ML models
- Impacts depend on:
  - Errors and their distribution in datasets
  - correctness of cleaning algorithms
  - structure of ML model
- **Model selection** and **cleaning algorithm selection** can increase robustness of impacts → No best solution!

Source: [9] Li et al. (2019)

Conclusion

– Data Quality Assurance is a substantial part of building machine learning models
– and hence it must be integrated into the development pipeline
– Data Quality Assurance is a field of continuous research and development in the upcoming years
– New techniques of Data Cleaning are on their way

Sources

[1] N. Swartz. Gartner warns firms of ‘dirty data’. *Information Management Journal*, 41(3), 2007.

[2] Abedjan, Ziawasch, et al. "Detecting data errors: Where are we and what needs to be done?." *Proceedings of the VLDB Endowment* 9.12 (2016): 993-1004.

[3] Chu, Xu, et al. "Data cleaning: Overview and emerging challenges." *Proceedings of the 2016 International Conference on Management of Data*. 2016.

[4] Hynes, Nick, D. Sculley, and Michael Terry. "The data linter: Lightweight, automated sanity checking for ml data sets." *NIPS MLSys Workshop*. 2017.

[5] Schelter, Sebastian, et al. "Automating large-scale data quality verification." *Proceedings of the VLDB Endowment* 11.12 (2018): 1781-1794.

[6] Krishnan, Sanjay, et al. "SampleClean: Fast and Reliable Analytics on Dirty Data." *IEEE Data Eng. Bull.* 38.3 (2015): 59-75.

[7] Krishnan, Sanjay, et al. "Activeclean: An interactive data cleaning framework for modern machine learning." *Proceedings of the 2016 International Conference on Management of Data.* 2016.

[8] Rekatsinas, Theodoros, et al. "Holoclean: Holistic data repairs with probabilistic inference." *arXiv preprint arXiv:*1702.00820 (2017).

Sources

[9] Li, Peng, et al. "CleanML: A Benchmark for Joint Data Cleaning and Machine Learning [Experiments and Analysis]." *arXiv preprint arXiv*:1904.09483 (2019).

Acknowledgements & License

– Images are either by the authors of these slides, attributed where they are used, or licensed under Pixabay