Data Curation for ML: Toward a Principled Approach

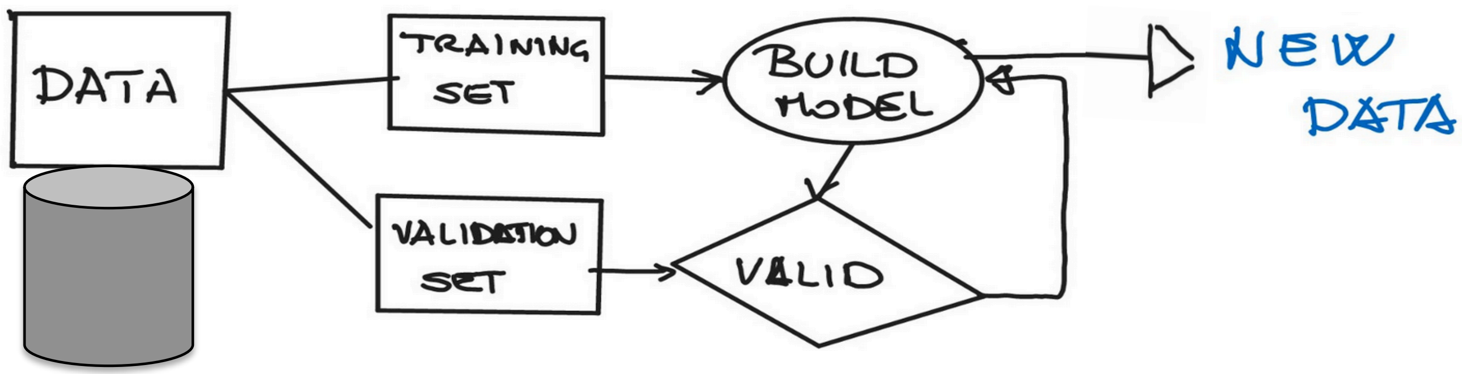
Laure Berti-Equille

laure.berti@ird.fr

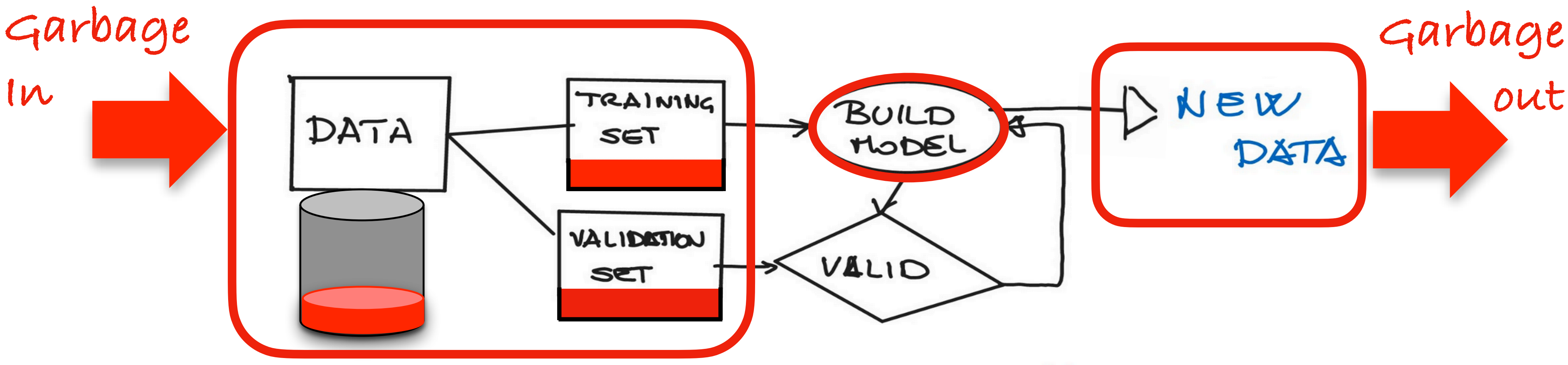
Institut de Recherche pour le Développement
FRANCE

Espace-Dev, IRD, Univ Montpellier, Univ Guyane, Univ La Réunion, Univ Antilles, Univ Nouvelle Calédonie, Montpellier France

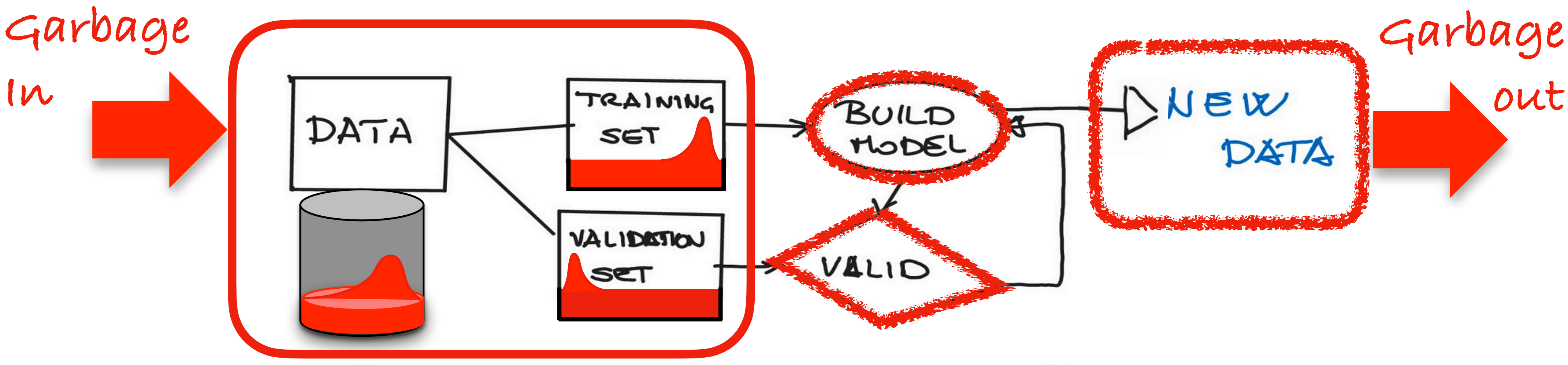
Learning from dirty data is risky



Learning from dirty data is risky



Learning from dirty data is risky



Glitch types and distributions can be very different in the datasets used for training, testing, and validation and they affect accuracy of ML models in different ways.

Two complementary approaches

INTERVENE

How to efficiently fix the data:

- ✦ Detect the anomalies
- ✦ Correct them with minimal cost (domain expert intervention, time, external master data, etc.)
- ✦ Select the repair/preparation strategies that will maximize the ML result quality

MITIGATE

How to reduce the impact of dirty data:

- ✦ Robustify the ML algorithms and apply ML ensembling strategies
- ✦ Use AutoML to find optimal parameter setting
- ✦ Select portions of the data and/or augment the data

Two complementary approaches

INTERVENE

How to efficiently fix the data:

- ✦ Detect the anomalies
- ✦ Correct them with minimal cost (domain expert intervention, time, external master data, etc.)
- ✦ Select the repair/preparation strategies that will maximize the ML result quality

MITIGATE

How to reduce the impact of dirty data:

- ✦ Robustify the ML algorithms and apply ML ensembling strategies
- ✦ Use AutoML to find optimal parameter setting
- ✦ Select portions of the data and/or augment the data

Outline

1. **Detection of data quality problems**
   Profiling data quality

2. **Data cleaning**
   Leveraging the patterns of glitches

3. **Data preparation strategies for ML**
   Learning to clean and prepare the data

Outline

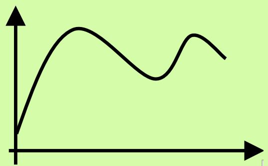
1. **Detection of data quality problems**
   Profiling data quality

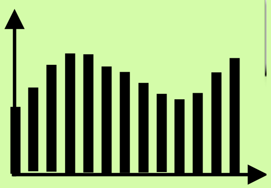
2. **Data cleaning**
   Leveraging the patterns of glitches

3. **Data preparation strategies**
   Learning to clean and prepare the data

Data Quality Problems

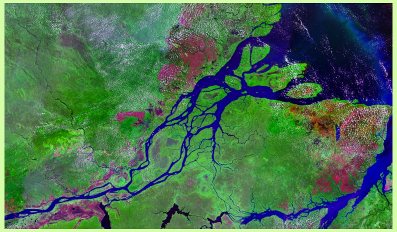
DATA TYPES





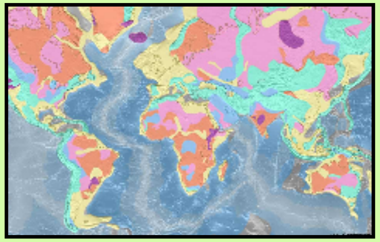
0101010101

ACACGTGT



John Doe

High
Medium
Low



RELATIONSHIPS

- Structural (record)
- Sequential
- Graph-based
- Temporal
- Spatial
- Spatio-Temporal

DATA QUALITY PROBLEMS

TYPE

- Missing Data
- Anomalous Data
- Duplicate Data
- Inconsistent Data
- Obsolete Data
- Incorrect data

CARDINALITY

- Single-Point
- Collection

DETECTION MODE

- Model-based
- Data distribution-based
- Constraint-based
- Pattern-based

Existing approaches for detecting/fixing DQ problems

Declarative

- Data debugging
- Checking data assertions
- Transform

ML-based

Learn from clean data and replace

Declarative Approaches

Checking data assertions and transform

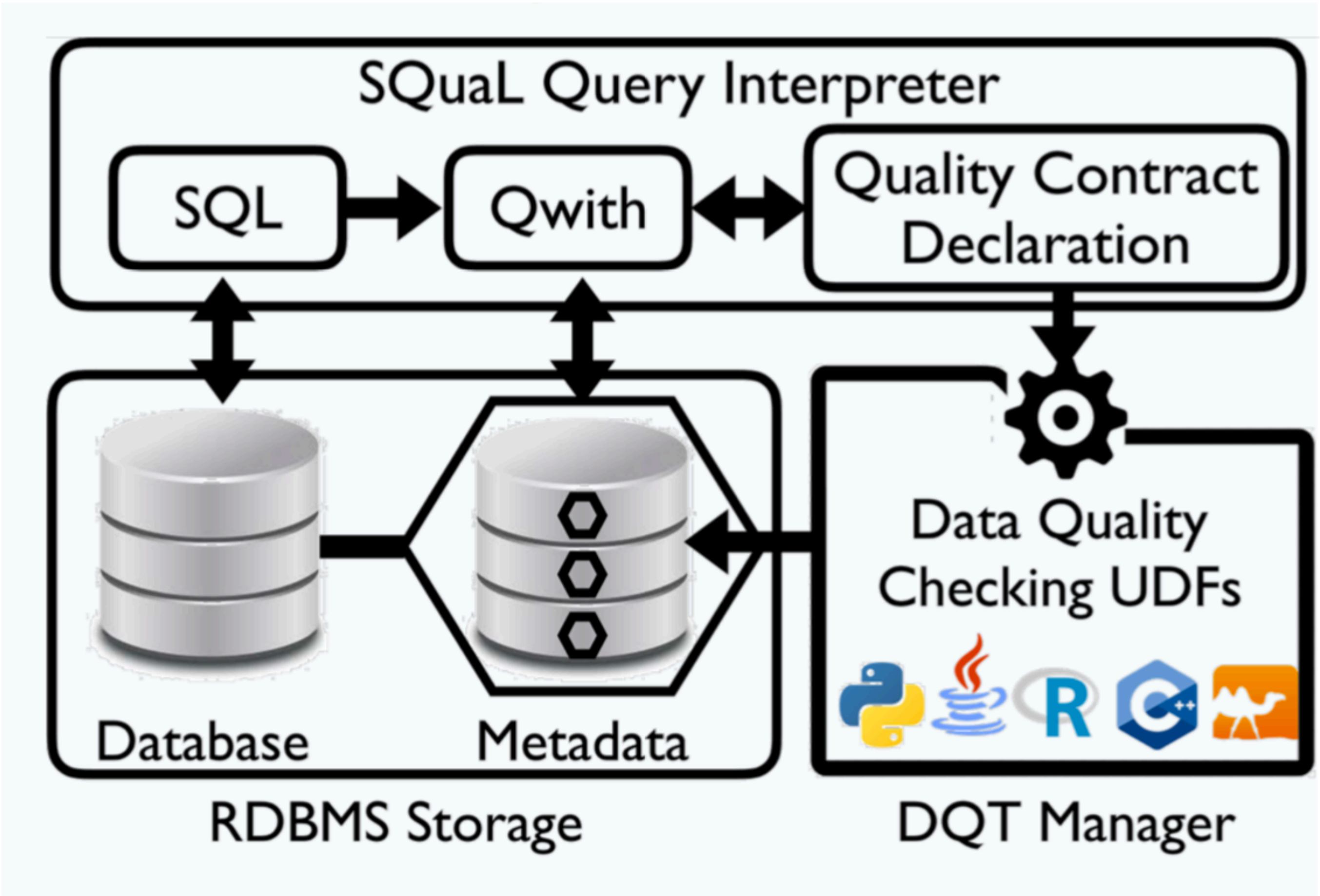
- ***Deequ*** [Schelter et al. ,VLDB 2018] requires cloud infrastructure and manual integration into training and serving systems; dependent on Apache Spark

- ***TensorFlow Data Validation*** (TFDV) [Caveness et al., SIGMOD 2020] integrated with Google TFX difficult to use outside of these platforms

- Lightweight Python-based approaches like ***great_expectations*** (https://greatexpectations.io) or ***hooqu*** (https://github.com/mfcabrera/hooqu) not integrated with the ML development process

Joint work with Ugo Comignani and Noel Novelli (AMU)
[EDBT 2020]

Declarative data profiling with MeSQuaL



https://github.com/ucomignani/MeSQuaL

Joint work with Ugo Comignani and Noel Novelli (AMU)
[EDBT 2020]

MeSQuaL Key Concepts

Flexible declarative data quality profiling with UDFs

1

Data quality dimensions

- completeness: Cell, Row, Column, Table, DB
- uniqueness: Row
- consistency: Cell, Row, Column, Table, DB
- veracity: Cell, Row

Procedural approach with UDFs

2

CONTRACT TYPES

CONTRACT INSTANCES

Declarative approach

3

SELECT …
FROM…
WHERE…
QWITH…

Extended query

Joint work with Ugo Comignani and Noel Novelli (AMU)
[EDBT 2020]

MeSQuaL Examples

DECLARATION

```
CREATE CONTRACTTYPE StatTests (
   autocorrelation BY FUNCTION 'durbinWatsonTest.py' LANGUAGE PYTHON,
   multicollinearity BY FUNCTION 'varInflationFactor.py' LANGUAGE PYTHON,
   heteroscedasticity BY FUNCTION 'BreuschPaganTest.py' LANGUAGE PYTHON,
   KMerrorNormality BY FUNCTION 'KolmogorovSmirnov.py' LANGUAGE PYTHON,
   SWerrorNormality BY FUNCTION 'ShapiroWilkTest.py' LANGUAGE PYTHON);
```

```
CREATE CONTRACT RegressionAssumptions (
   StatTests.autocorrelation > 0
   AND StatTests.autocorrelation < 4
   AND StatTests.multicollinearity <= 4
   AND StatTests.heteroscedasticity < 0.05
   AND StatTests.SWerrorNormality < 0.05);
```

```
CREATE CONTRACTTYPE CheckQDB (
   completeness BY FUNCTION 'completeness.py' LANGUAGE PYTHON,
   uniqueness BY FUNCTION 'uniqueness.py' LANGUAGE PYTHON,
   consistency BY FUNCTION 'consistency.py' LANGUAGE PYTHON,
   outlyingness BY FUNCTION 'outlyingness.py' LANGUAGE PYTHON);
```

```
CREATE CONTRACT CheckBeforeAnalysis (
   RegressionAssumptions
   AND CheckQDB.consistency > 0.9
   AND CheckQDB.outlyingness < 0.2);
```

MANIPULATION

AoT	{ SELECT * FROM ChicagoDataset } QWITH CheckQDB.completeness> 0.95;
	{ SELECT * FROM ChicagoDataset } QWITH CheckBeforeAnalysis AND RegressionAssumptions;
	{ SELECT timestamp, node_id,value_raw,valuehrf FROM ChicagoDataset WHERE ChicagoDataset.sensor = 'o3' } QWITH CheckBeforeAnalysis AND CheckQDB.completeness> 0.95;
MIMIC-III	{ SELECT * FROM Admissions } QWITH CheckQDB.completeness> 0.95;
	{ SELECT * FROM Admissions WHERE Admissions.insurance = 'Private' } QWITH CheckBeforeAnalysis AND CheckQDB.completeness> 0.95;
	{ SELECT gender, dob, admittime FROM Admissions INNER JOIN (SELECT * FROM Patients WHERE dob < '2090-12-12 00:00:00' QWITH CheckQDB.completeness> 0.95) as Pat ON Admissions.subject_id=Pat.subject_id; } QWITH CheckQDB.completeness> 0.95;

Joint work with Ugo Comignani and Noel Novelli (AMU)
[EDBT 2020]

MeSQuaL GUI

https://github.com/ucomignani/MeSQuaL

ML-based Approaches

Learn from clean data and replace/repair

- Pattern enforcement
  - Syntactic patterns (date formatting)
  - Semantic patterns (name/address)
- Value update to satisfy a set of rules, constraints, FDs, CFDs, Denial Constraints (DCs), Matching Dependencies (MDs) with minimal number of changes.
- Value replacement
- Entity resolution

EXAMPLES

✦ SCARE: Scalable Automatic Repair

✦ On-demand ETL [Yang et al., VLDB'15]

✦ ActiveClean [Krishnan et al., VLDB'16]

✦ HoloClean [Rekatsinas et al., VLDB 2017]

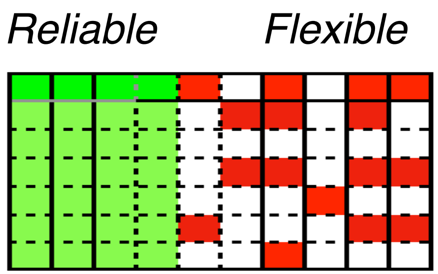
✦ Deep learning for Entity Resolution

✦ Transformers for data prep

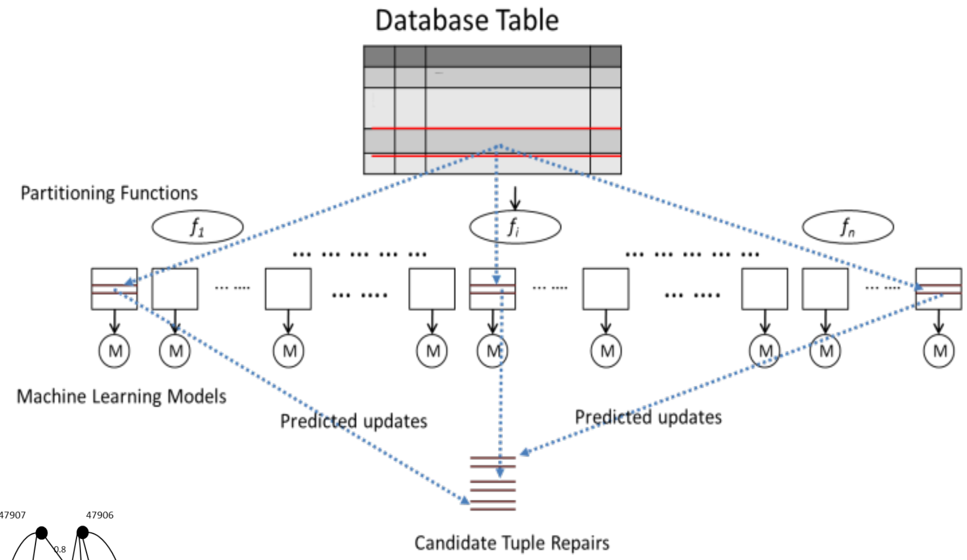
SCARE: SCalable Automatic Repair

[Yakout, Berti-Equille, Elmagarmid, SIGMOD 2013]

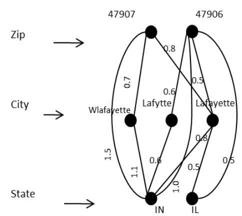
Goal: Find the repair that would maximize the sum of the probabilities of the values co-occurrence (i.e., association strength between predicted and reliable values) under a certain update cost budget.



1. Modeling Dependency and Predicting Updates
2. Data Partitioning
3. Tuple Repair Selection



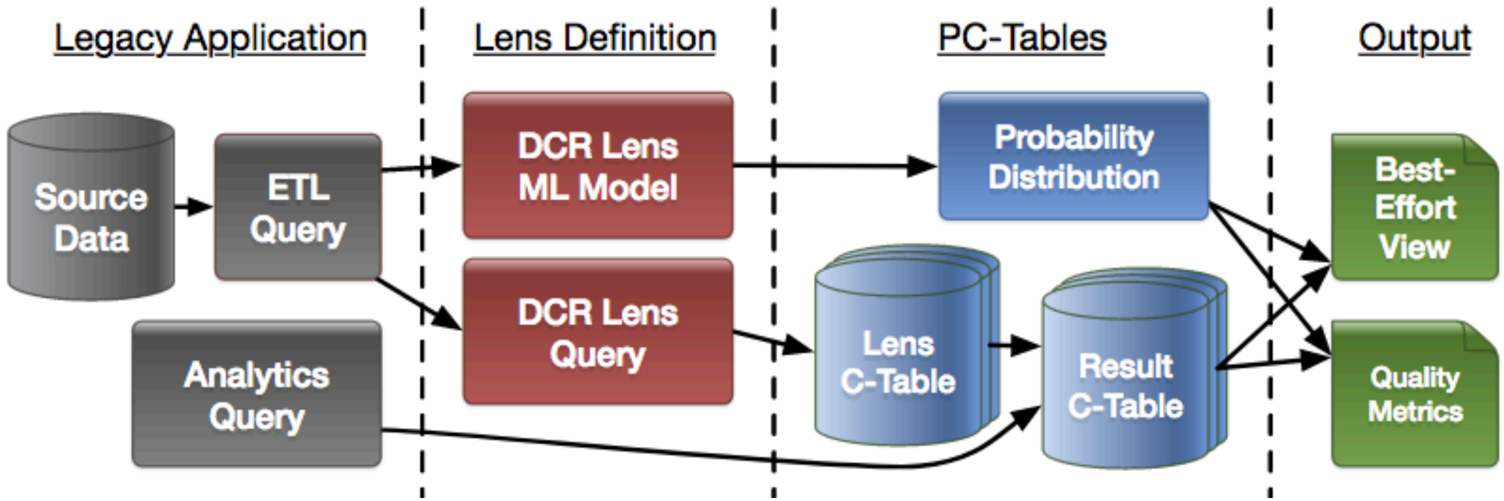
Value predictions for Flexible Attributes E1, E2, E3



On-demand ETL with Lenses

[Yang et al. ,VLDB'15]

Specification of Lens with classifiers from the massive online analysis (MOA) framework for Domain Constraint Repair (DCR).




```
CREATE LENS SaneProduct AS SELECT * FROM Product
  USING DOMAIN_REPAIR( category string NOT NULL,
                       brand    string NOT NULL );
```

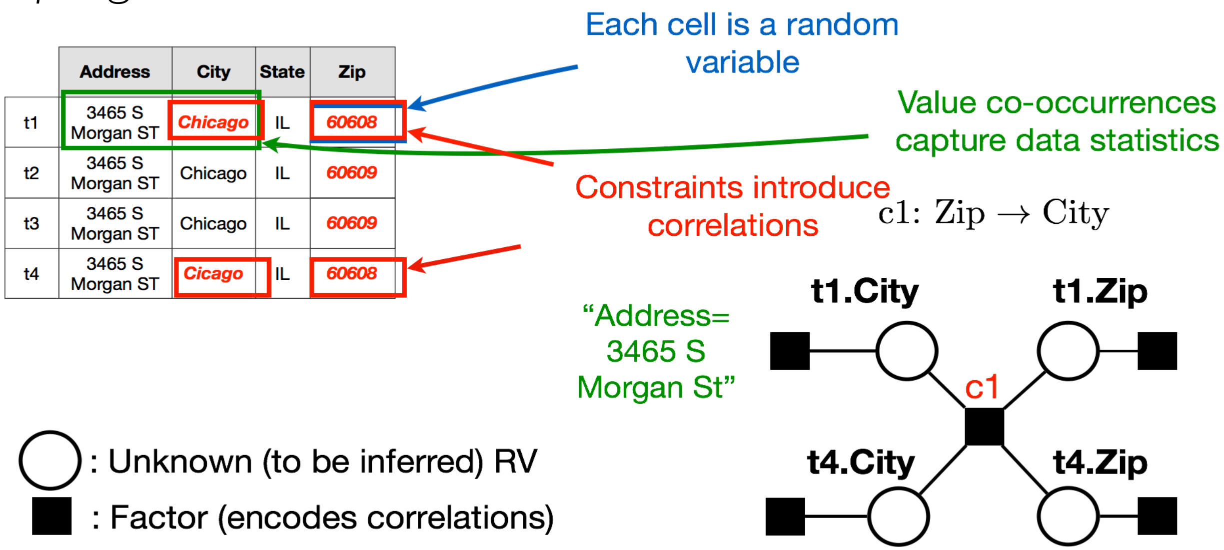
id	name	brand	category
P123	Apple 6s, White	$Var('X', R1)$	phone
P124	Apple 5s, Black	$Var('X', R2)$	phone
P125	Samsung Note2	Samsung	phone
P2345	Sony 60 inches	$Var('X', R4)$	$Var('Y', R4)$
P34234	Dell, Intel 4 core	Dell	laptop
P34235	HP, AMD 2 core	HP	laptop

HoloClean

[Rekatsinas et al., VLDB 2017]

https://github.com/HoloClean/HoloClean

HoloClean generates a factor graph capturing co-occurrences, correlations based on a set of constraints and external evidences. It uses SGD to learn parameters and infer the marginal distribution of unknown variables with Gibbs sampling.



Denial constraints:

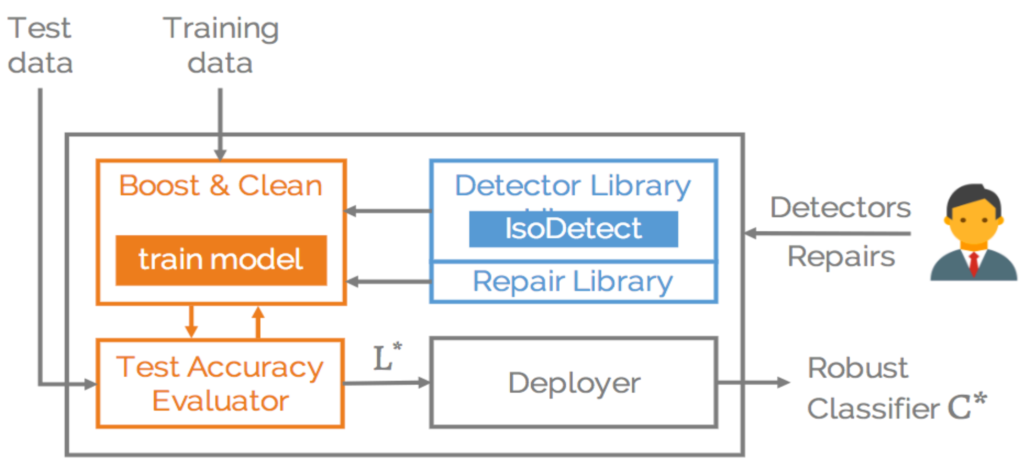
$$\forall t_1, t_2 \in D : \neg(t1[Zip] = t2[Zip] \wedge t1[City] \neq t2[City])$$

$$\forall t_1, t_2 \in D : \neg(t1[Zip] = t2[Zip] \wedge t1[State] \neq t2[State])$$

BoostClean

[Krishnan et al., 2017]

BoostClean selects an ensemble of methods (statistical and logic rules) for error detection and for repair combinations using statistical boosting.



Algorithm : Boost-and-Clean Algorithm

Data: (X, Y)

Initialize $W_i^{(1)} = \frac{1}{N}$

$\mathcal{L}$ generates a set of classifiers $\mathcal{C}\{C^{(0)}, C^{(1)}, ..., C^{(k)}\}$ where $C^{(0)}$ is the base classifier and $C^{(1)}, ..., C^{(k)}$ are derived from the cleaning operations.

for $t \in [1, T]$ **do**

  C_t = Find $C_t \in \mathcal{C}$ that maximizes the weighted accuracy on the test set. ϵ_t = Calculate weighted classification error on the test set $\alpha_t = \ln(\frac{1-\epsilon_t}{\epsilon_t})$ $W_i^{(t+1)} \propto W_i^{(t)} e^{-\alpha_t y_i C_t(x_i)}$: down-weight correct predictions, up-weight incorrectly predictions.

return $C(x) = \text{sign}(\sum_t^T \alpha_t C_t(x))$

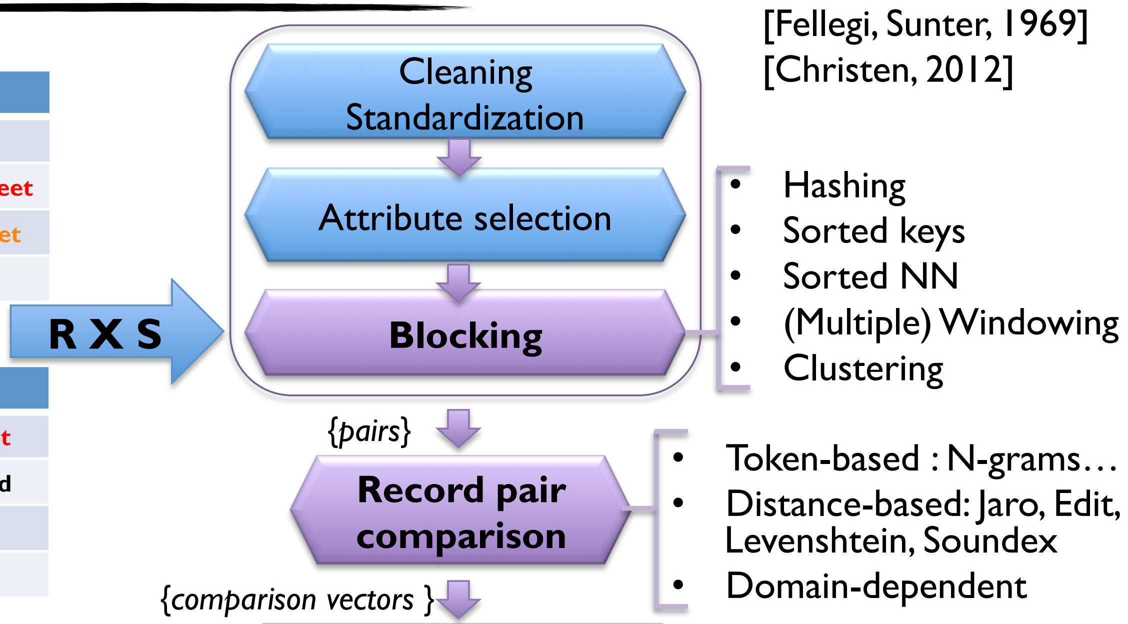
Record Linkage (RL): Generic Workflow

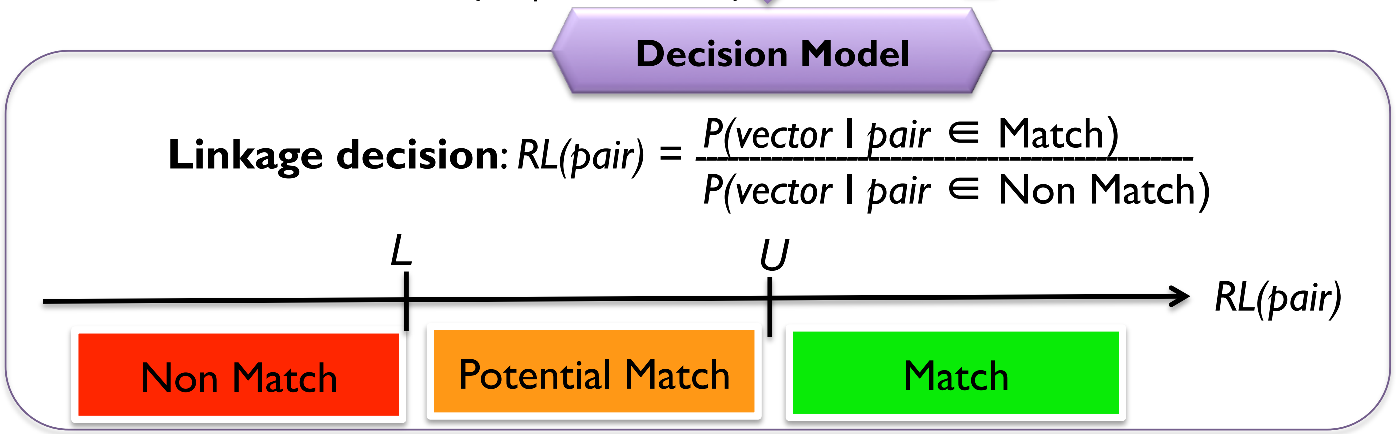
Database R

Name	SSN	Addr
Will Forth	354-564-339	Ada Bd
Jacky Khan	435-232-129	Marple Street
Dom Hack	235-575-689	Main Street
…	…	…

Database S

Name	SSN	Addr
Jack Khan	435-223-129	Marple St
Hans Ford	354-564-339	Clover Bd
Tom Hack	235-557-689	Main St
…	…	…

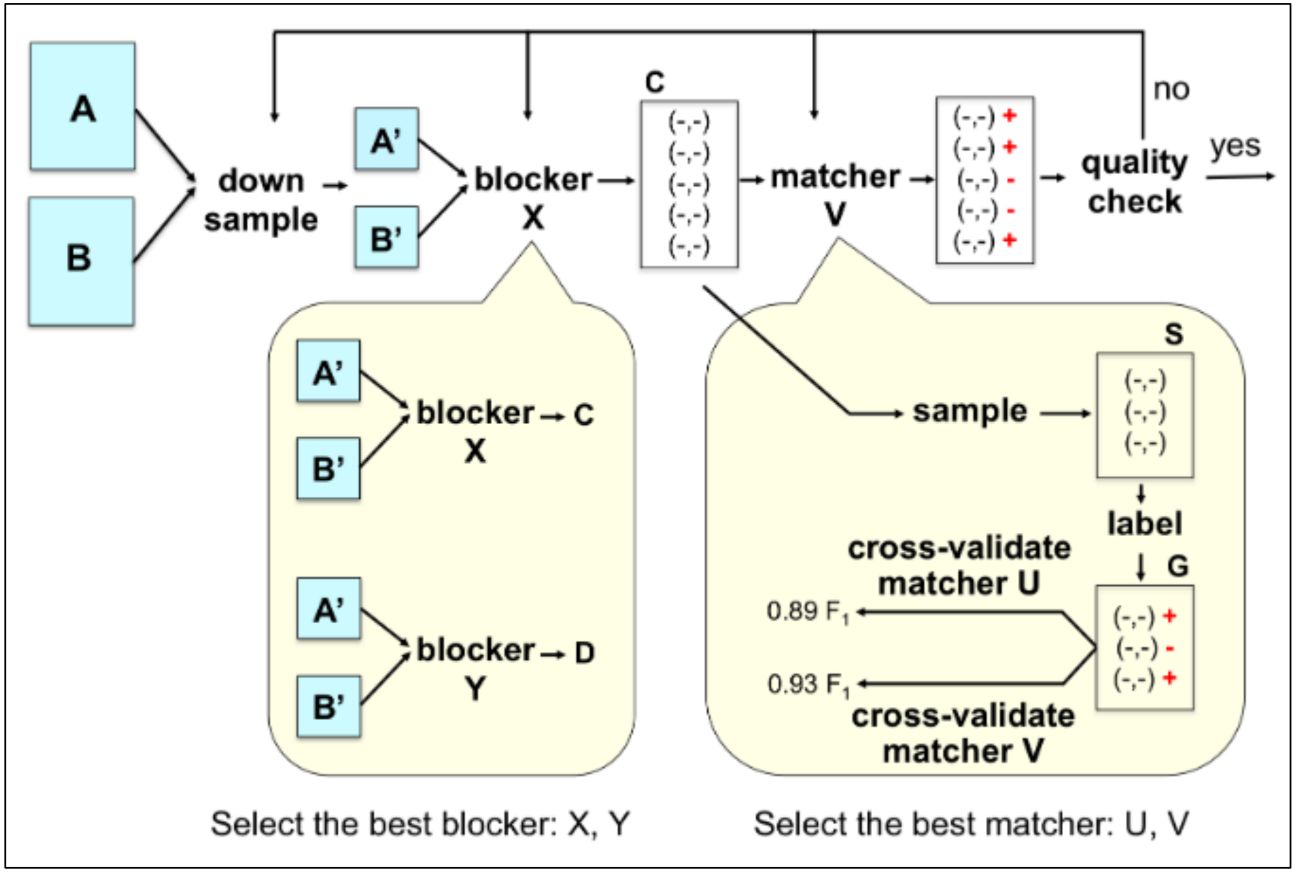




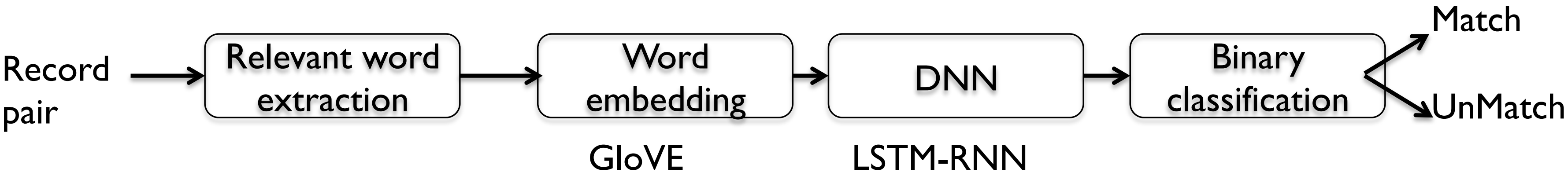
Human-In-The Loop for Entity Matching

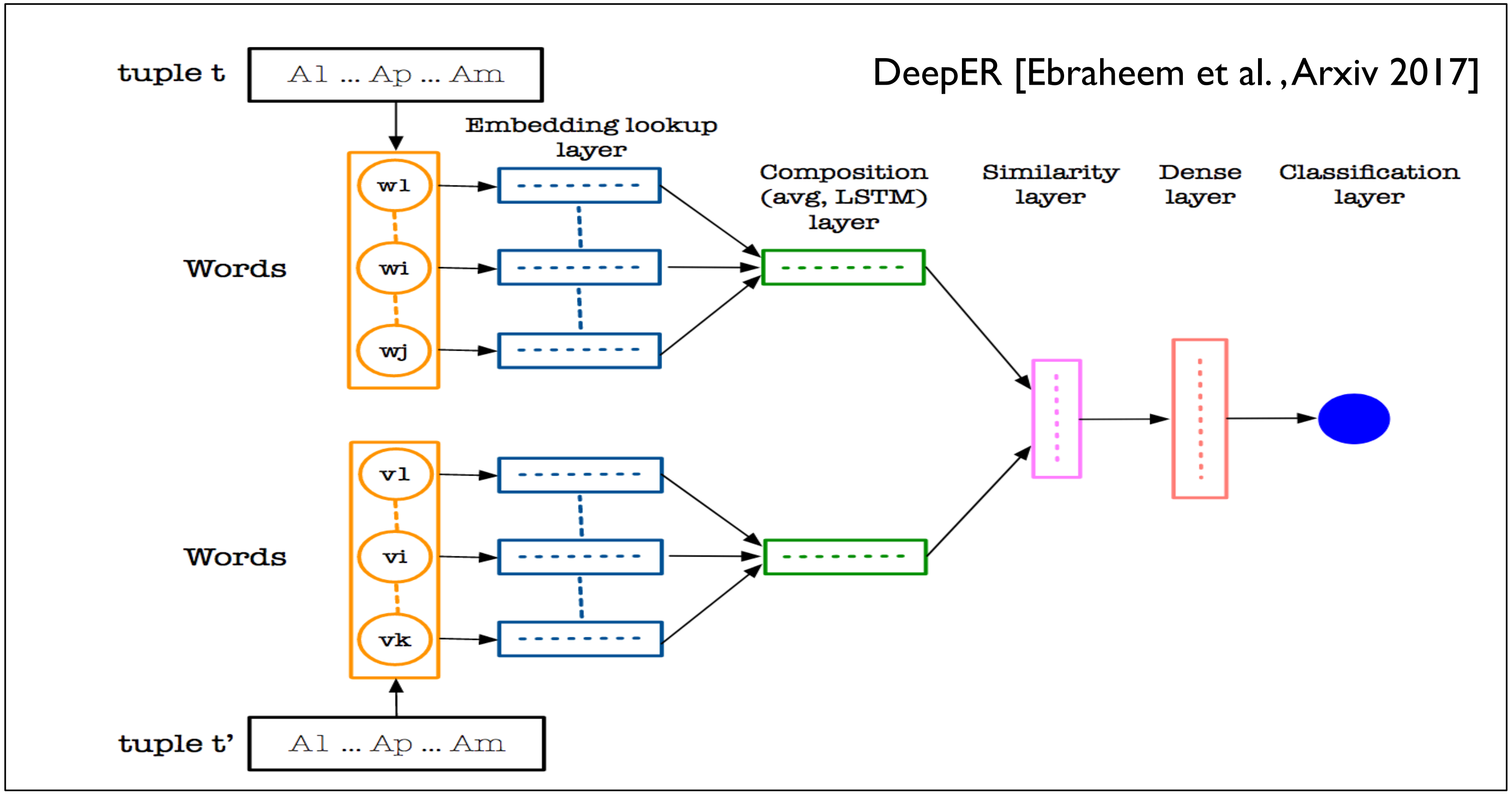
[Doan et al., HILDA@SIGMOD'17]

Magellan project: Lessons learnt for How-to Guide for EM



Deep learning for Entity Resolution





DeepER [Ebraheem et al. ,Arxiv 2017]

Outline

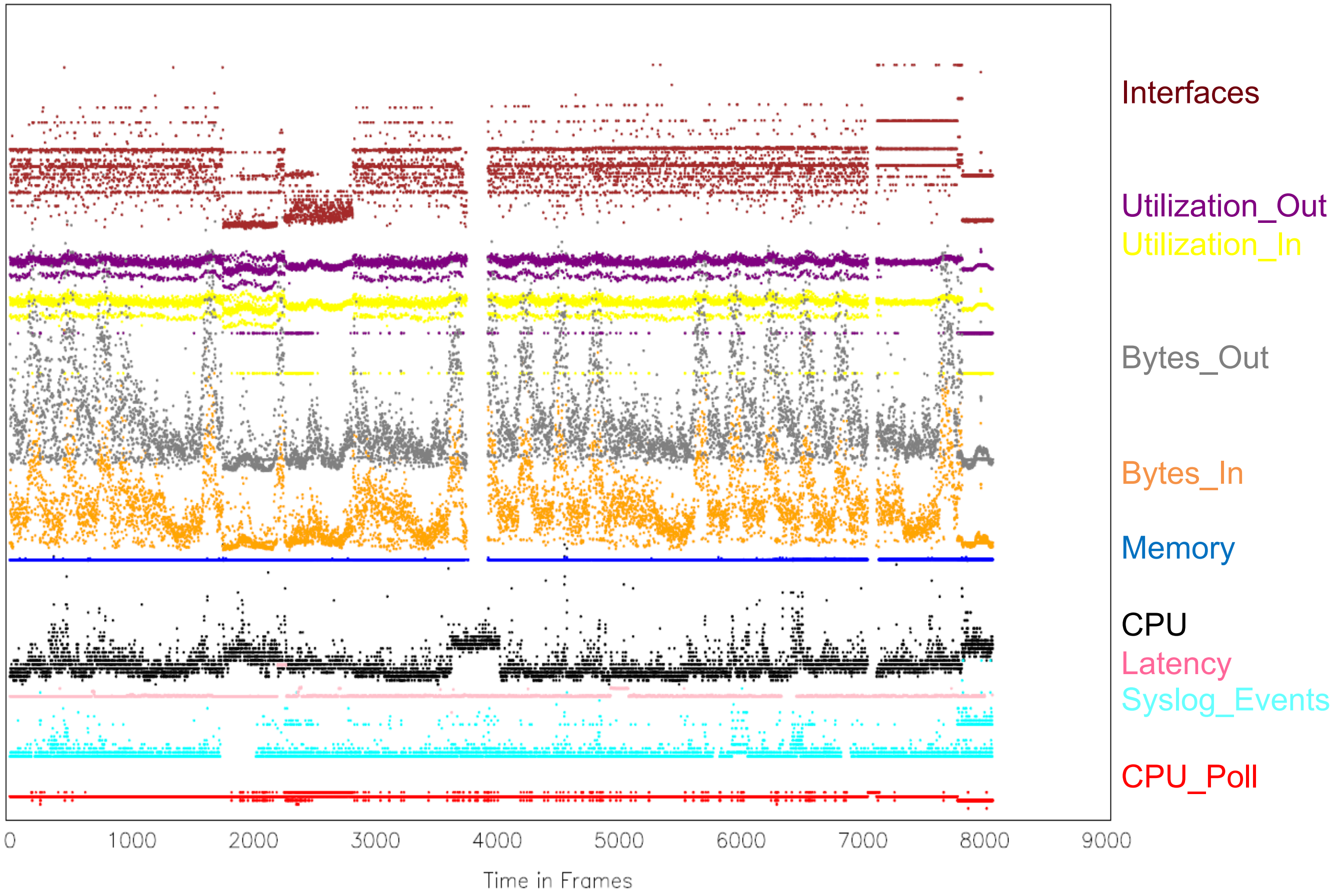
1. **Detection of data quality problems:**
   Profiling data quality

2. **Data cleaning**
   Leveraging the patterns of glitches

3. **Data preparation strategies:**
   Learning to clean and prepare the data

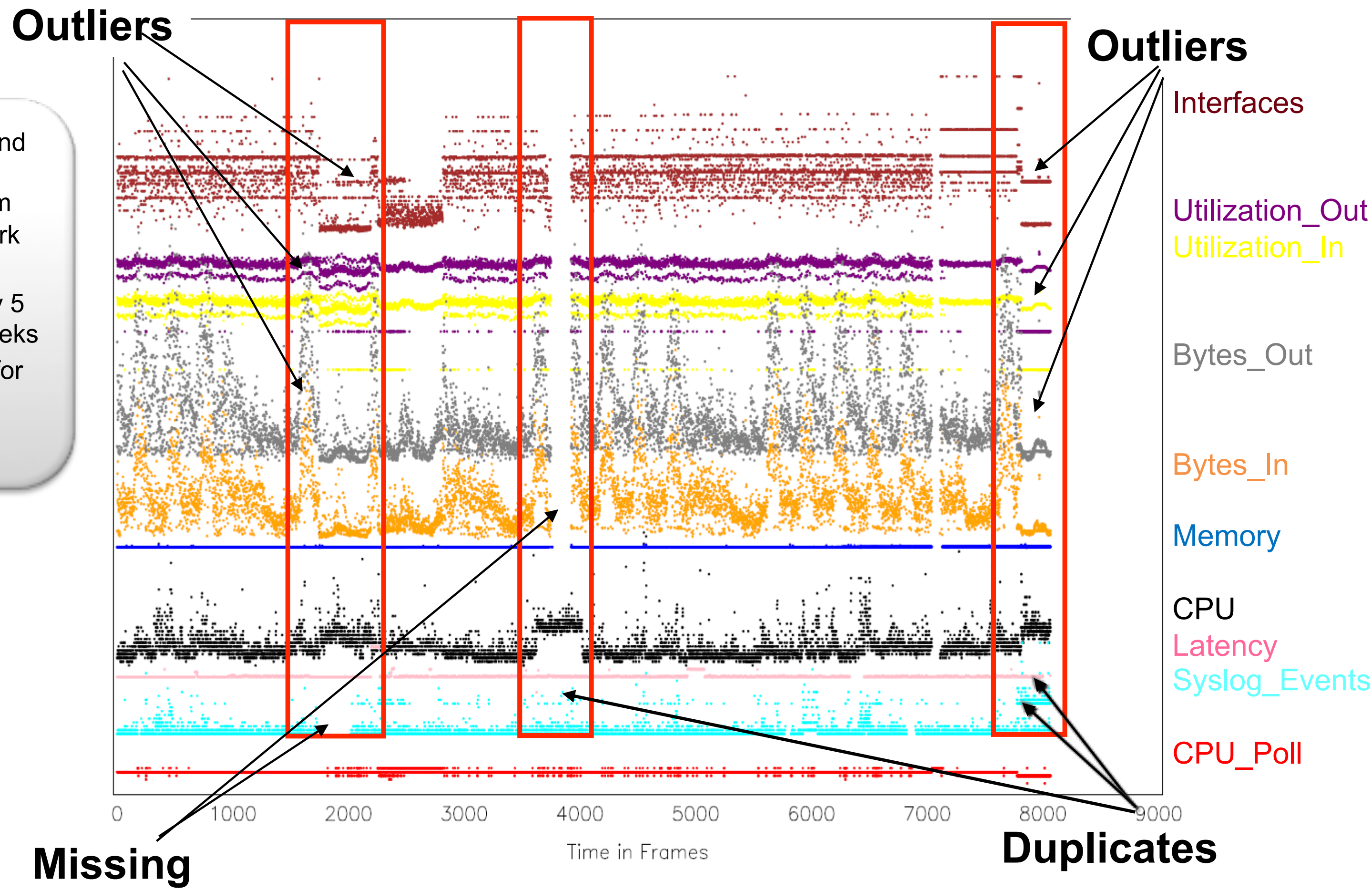
SNMP Data Analysis

- Periodic inbound and outbound traffic measurements from interfaces of network devices
- 10 attributes, every 5 minutes, over 4 weeks
- Axes transformed for plotting

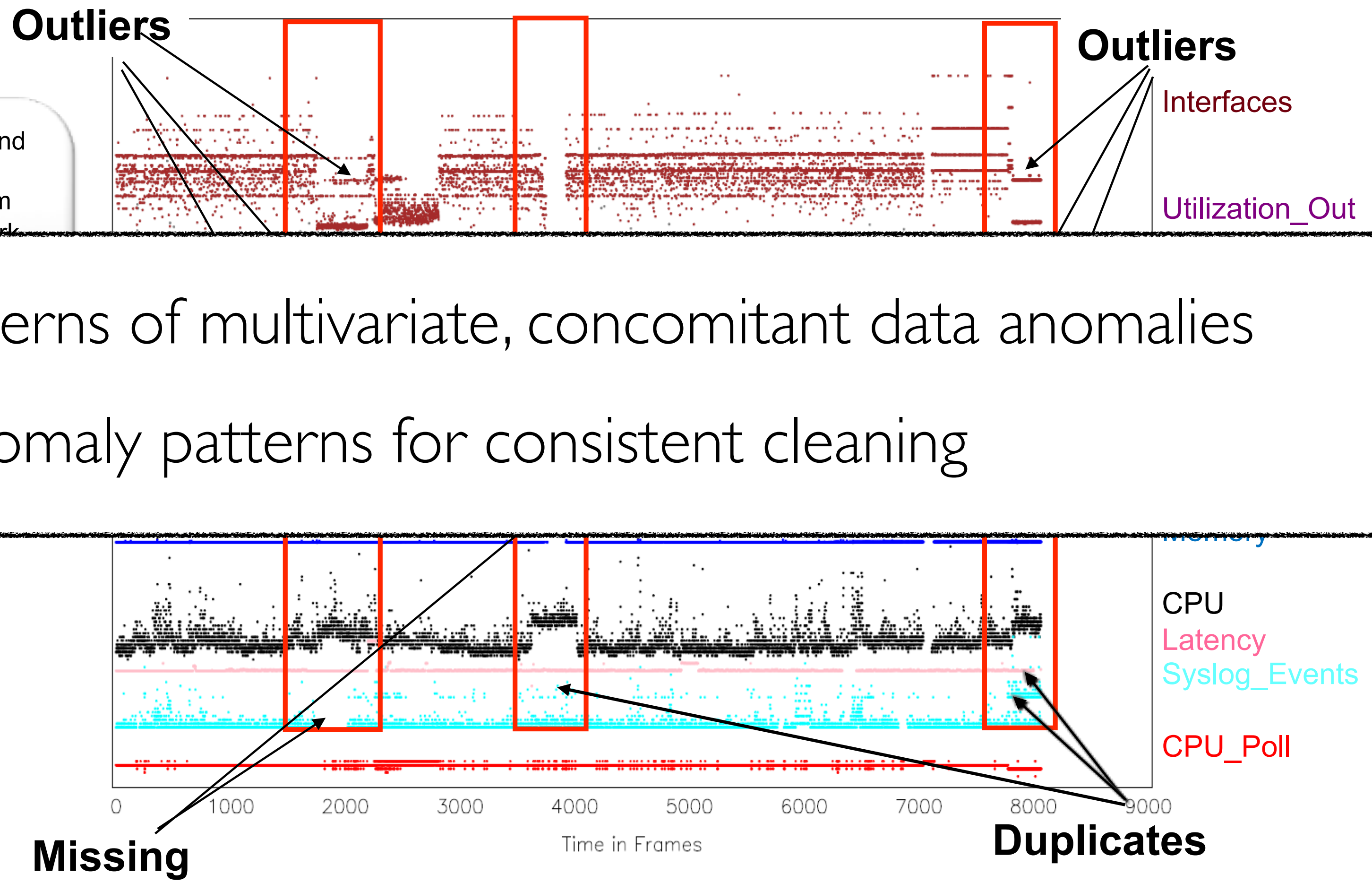


SNMP Data Analysis

- Periodic inbound and outbound traffic measurements from interfaces of network devices
- 10 attributes, every 5 minutes, over 4 weeks
- Axes transformed for plotting



SNMP Data Analysis



- Periodic inbound and outbound traffic measurements from

1. Detect patterns of multivariate, concomitant data anomalies
2. Use the anomaly patterns for consistent cleaning

Understanding Complex Glitch Patterns

Benefits

- A common root cause can generate correlated data errors
- In-depth anomaly analysis could help for:
  - Characterizing anomaly sources, processes, and propagation mechanisms
  - Systematizing data cleaning

Current methods

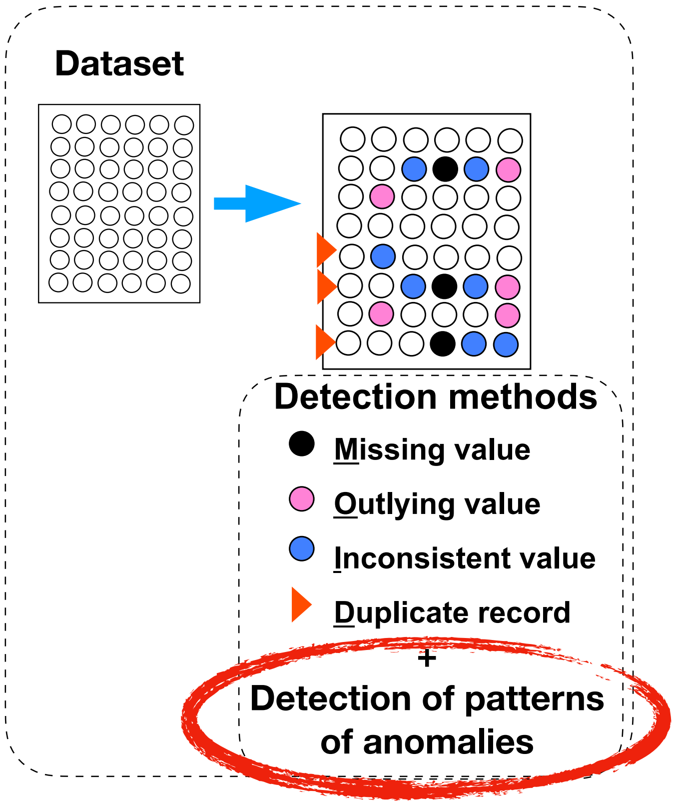
- Make unrealistic assumptions (e.g., MAR)
- Treat glitches in isolation
- Are one-shot approaches (no reiteration between detection and cleaning)

Data cleaning and preprocessing may introduce new errors and distortions.

Joint work with Parni Dasu and Divesh Srivastava (AT&T Lab Research)
[ICDE 2011]

Detection-Exploration-Cleaning Framework

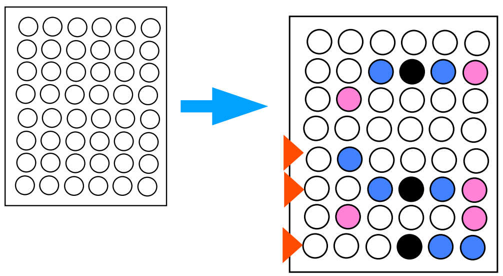
Input:



Detection-Exploration-Cleaning Framework

Input:

Dataset



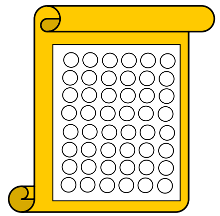
Detection methods

- **Missing value**
- **Outlying value**
- **Inconsistent value**
- **Duplicate record**

+

Detection of patterns of anomalies

Specifications of the ideally preprocessed dataset

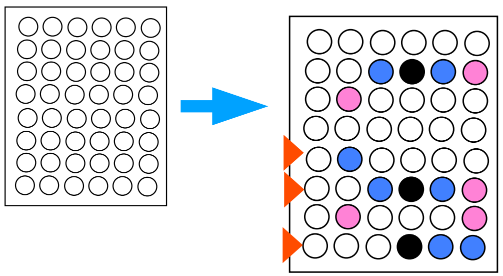


e.g., less than 5% of anomalies, should preserve the median of the original data distributions, etc.

Detection-Exploration-Cleaning Framework

Input:

Dataset



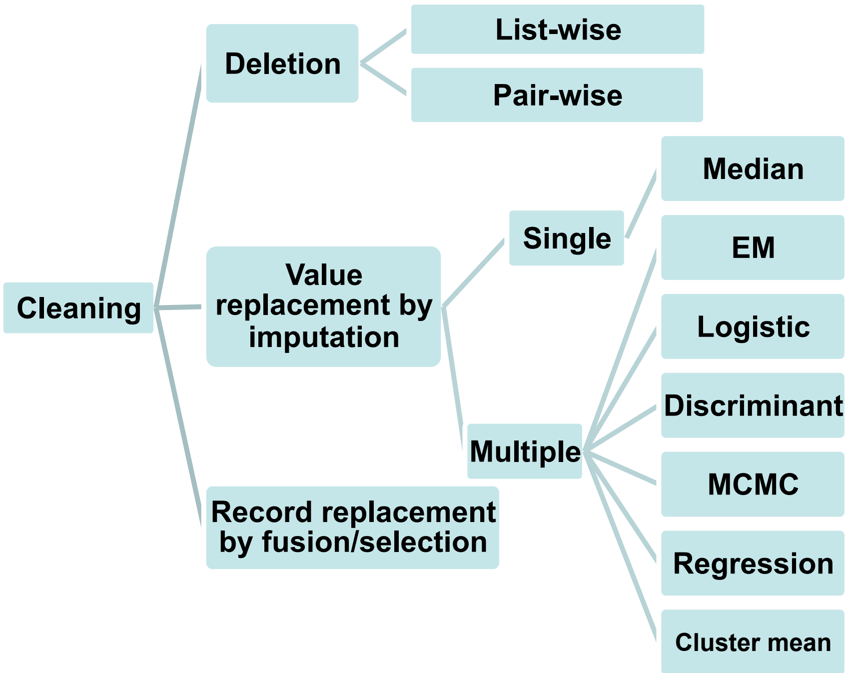
Detection methods

- ● Missing value
- ● Outlying value
- ● Inconsistent value
- ▶ Duplicate record

\+

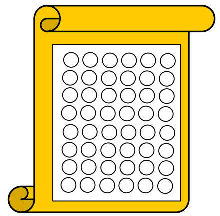
Detection of patterns of anomalies

Candidate cleaning strategies



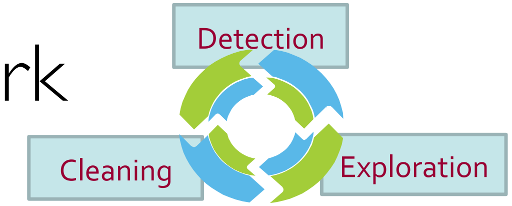
+ Pattern-aware cleaning strategies

Specifications of the ideally preprocessed dataset

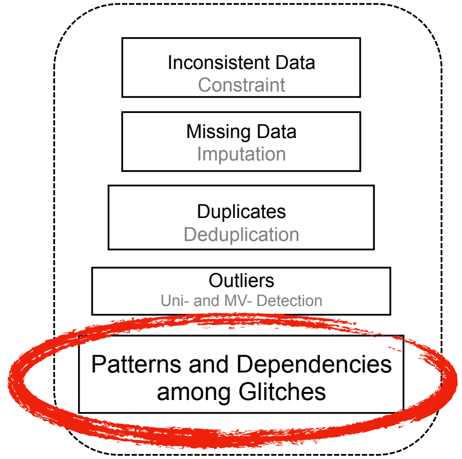


e.g., less than 5% of anomalies, should preserve the median of the original data distributions, etc.

Detection-Exploration-Cleaning Framework



[Berti-Equille, Dasu,Srivastava, ICDE 2011]



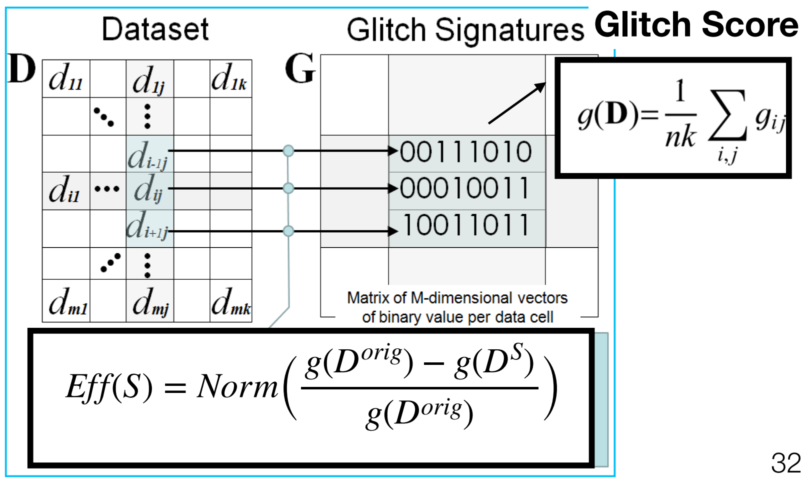
Problem Statement:

Find the quantitative cleaning strategy B composed of M methods among the candidate strategies $\mathbf{S}$ such that its resulting dataset D^B is the closest to an ideal dataset D^* specified from D as

$$D^B = \arg\min_{\{s \in S\}} \left(\mathrm{dist}(D^s, \mathrm{D}^*)\right)$$

$$\text{subject to } Cost(s) \leq \mathrm{U} \text{ and } Eff(s) \geq \Gamma > 0$$

- **dist** is the Kullback-Leibler distance between two data distributions
- **U** is a pre-defined upper bound for the cost of strategy s
- **Γ** is the lower bound of $Eff(s)$, the effectiveness of strategy s



$$g(\mathbf{D}) = \frac{1}{nk}\sum_{i,j} g_{ij}$$

$$Eff(S) = Norm\left(\frac{g(D^{orig}) - g(D^S)}{g(D^{orig})}\right)$$

[ICDE 2011]

Experiments

Real-world and semi-synthetic data

- **EPO Dataset:** 754,075 records, 4 non-key attributes (string, categorical and numerical data)
- **Intel Berkeley Research lab Dataset:** 2,313,682 million readings, 8 attributes (timestamp, sensorID, temperature, light, voltage) collected every 31 seconds from 54 sensors deployed in the between February 28th and April 5th, 2
- **SNMP Dataset:** (8,632 tuples, 11 variables) collected every 5 minutes during one month (timestamps, categorical and numerical values)

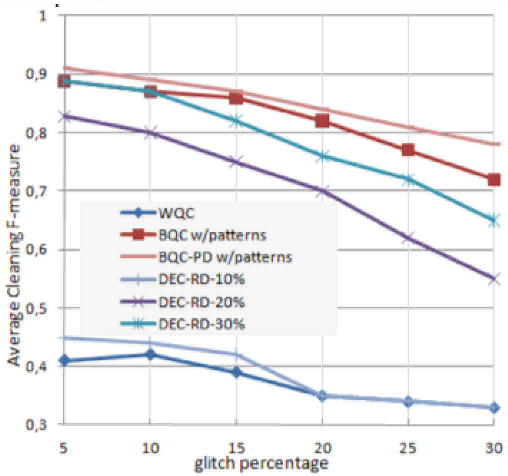
Comparison of various cleaning strategies

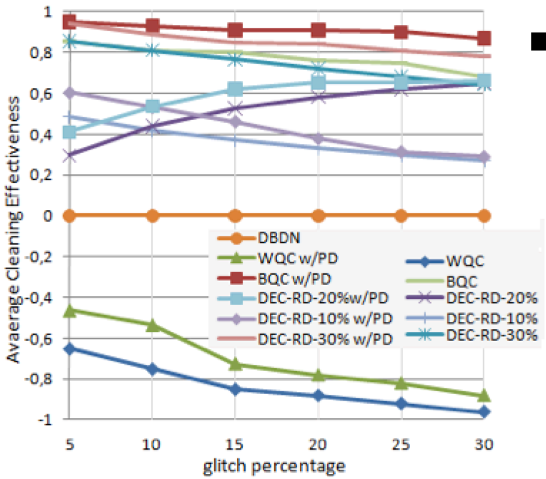
- Cost-based
- Effectiveness-based
- Resource-driven to treat just p% of glitches (DEC-RD)
- Specification-driven to treat a particular glitch type (DEC-SD)
- Pattern-based (DEC-PD)

[ICDE 2011]

Experimental results

SNMP

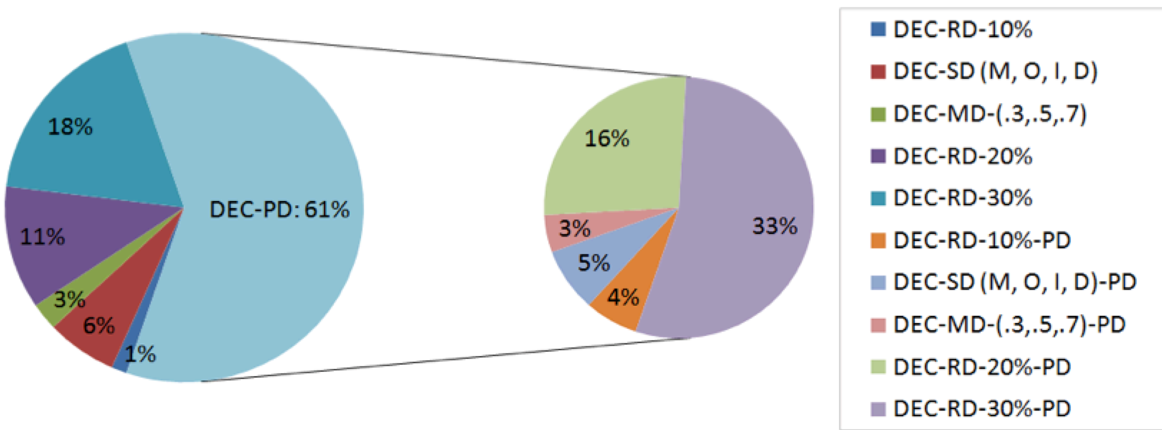




➔ **Pattern discovery always improves the accuracy when glitch percentage increases.**

➔ **Effectiveness is improved by + 8% in average.**

➔ ***61% of the best strategies are pattern-based.***



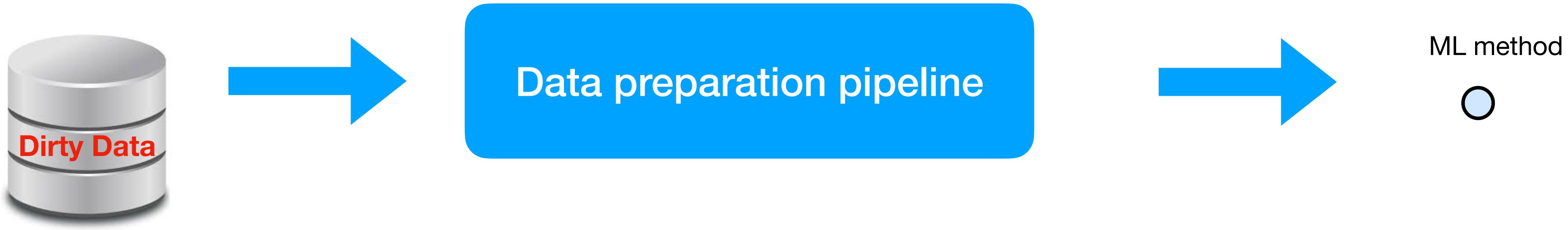
Outline

1. **Detection of data quality problems:**
   Profiling data quality with MeSQuaL

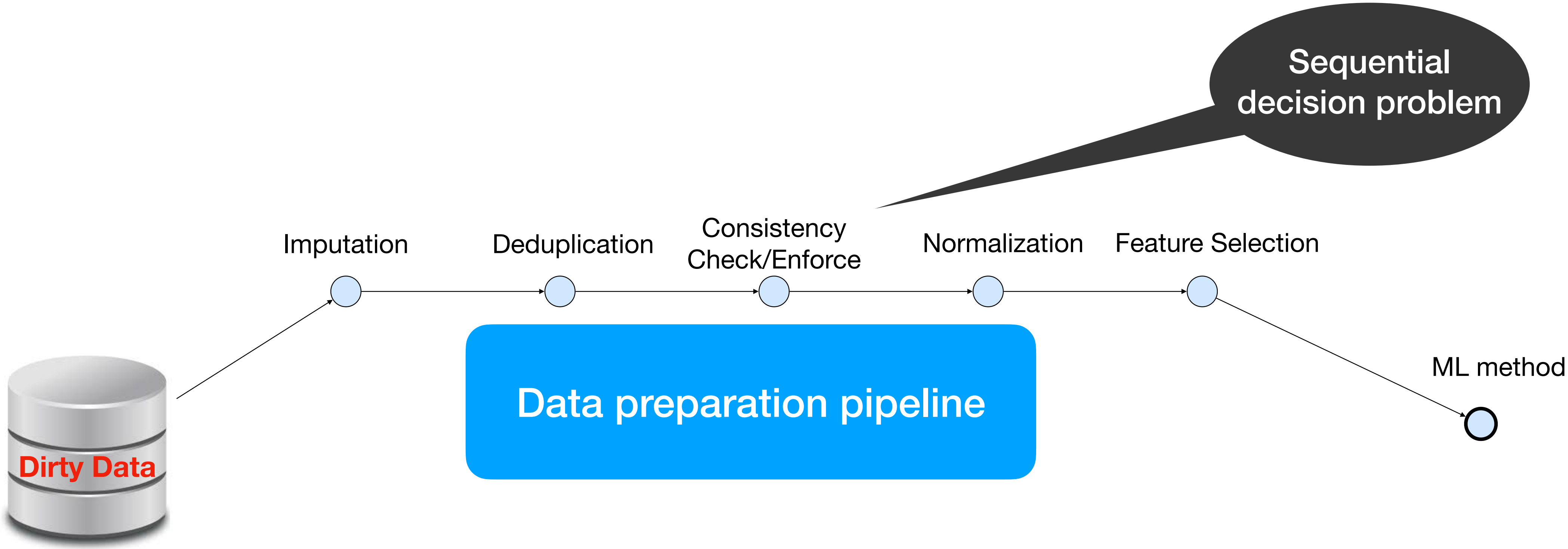
2. **Data cleaning**
   Leveraging the patterns of glitches

3. **Data preparation strategies**
   Learning to clean and prepare the data

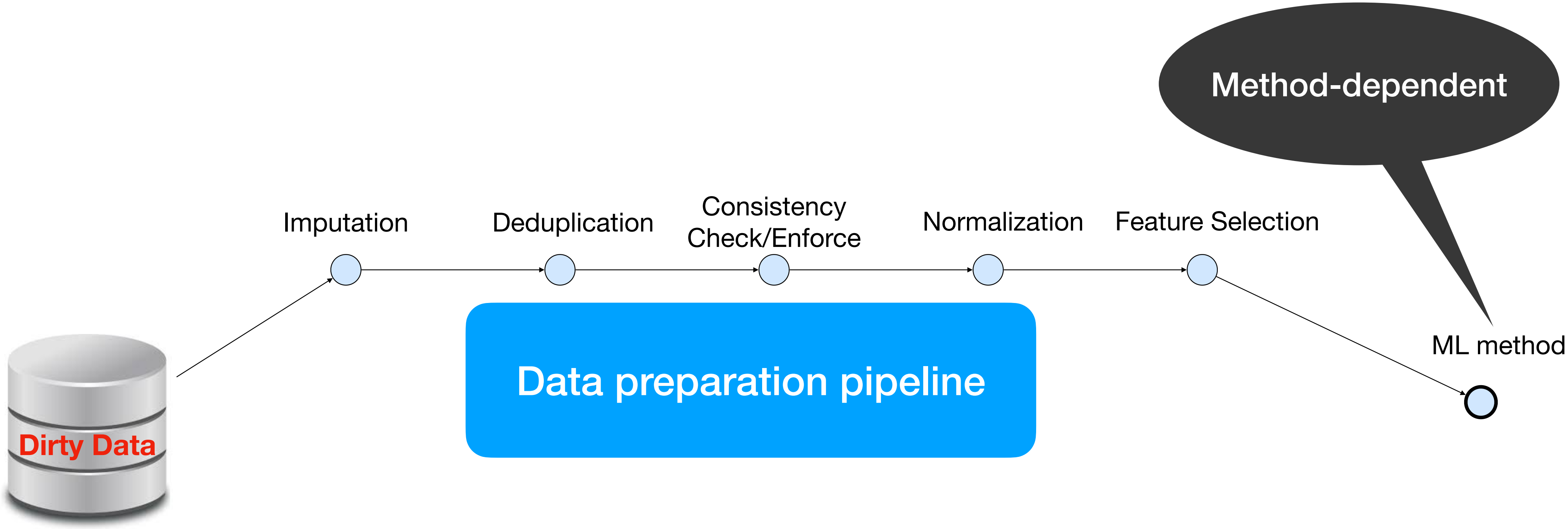
Data preprocessing is challenging



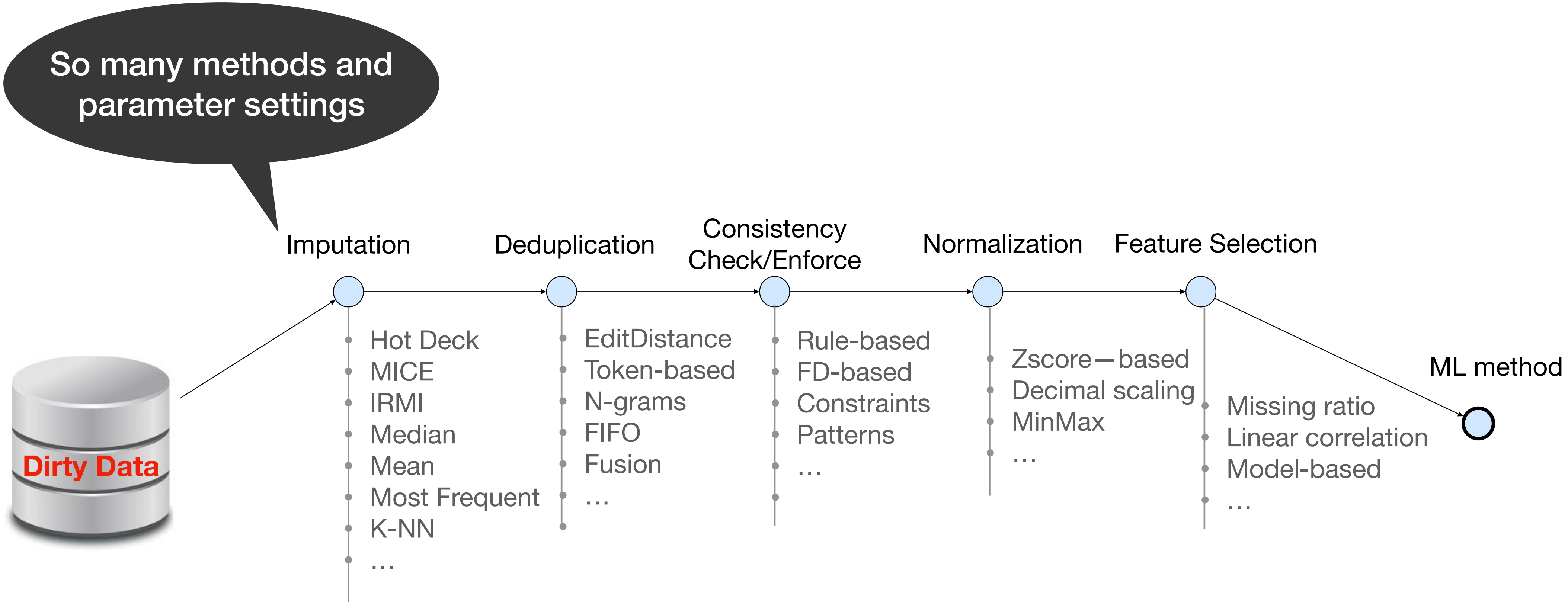
Data preprocessing is challenging



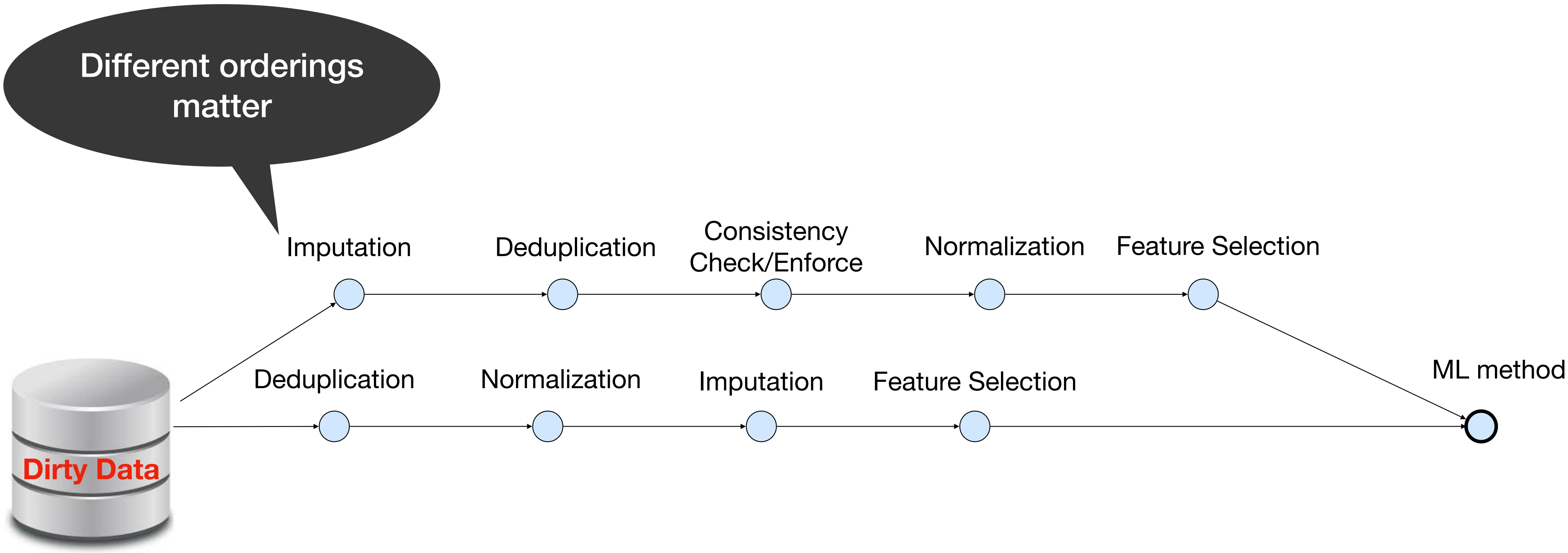
Data preprocessing is challenging



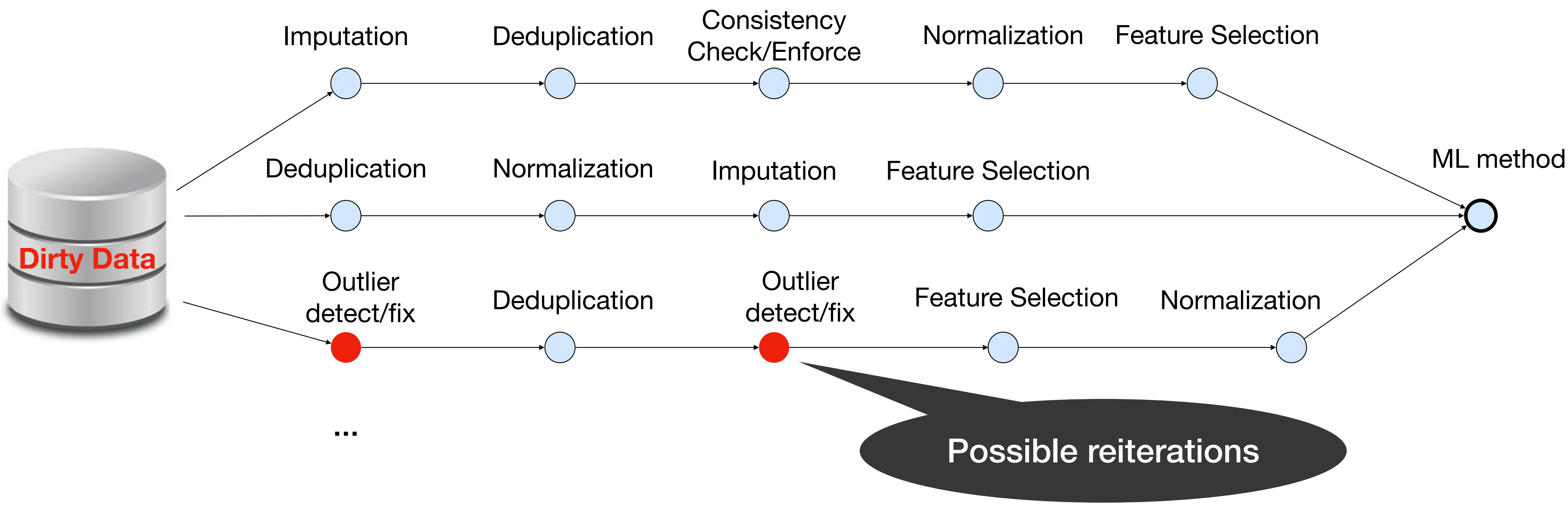
Data preprocessing is challenging



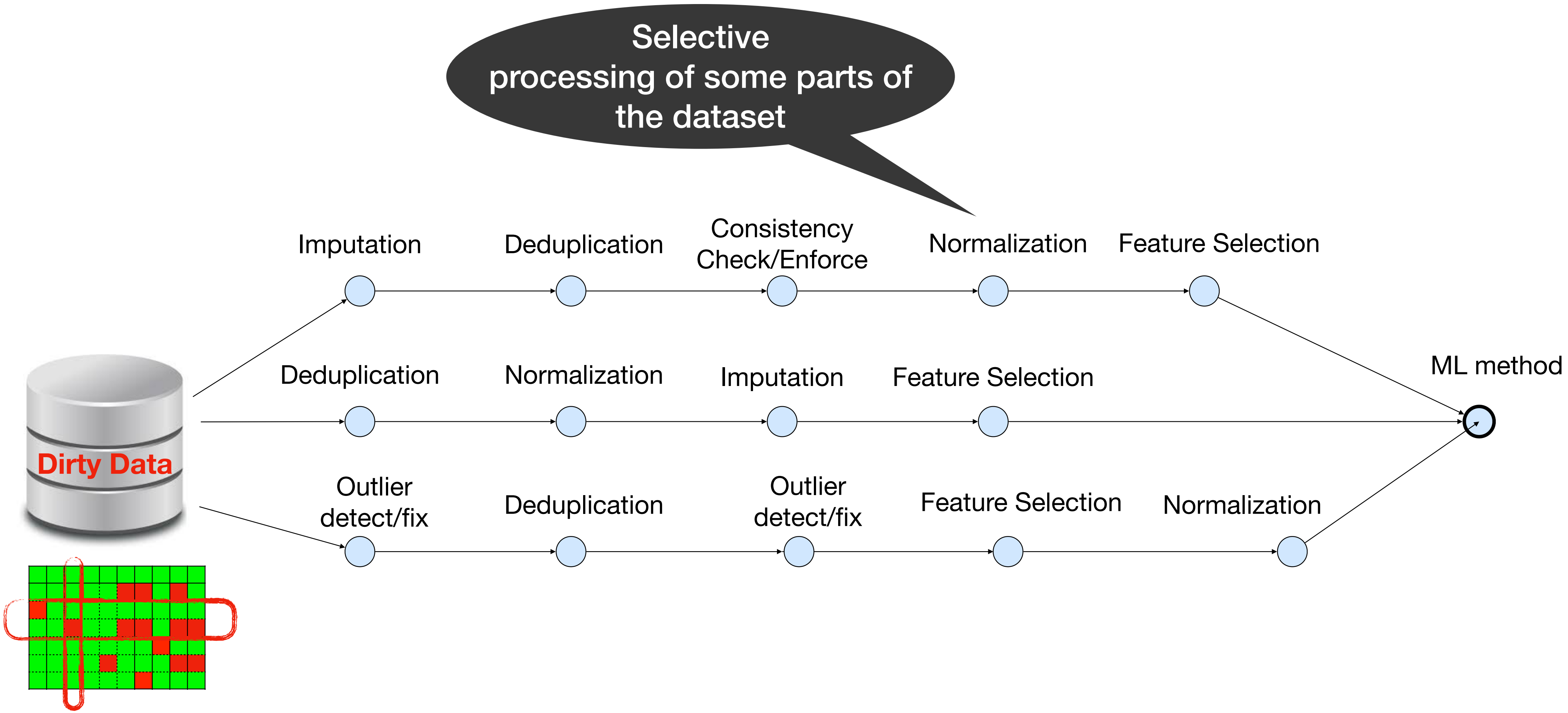
Data preprocessing is challenging



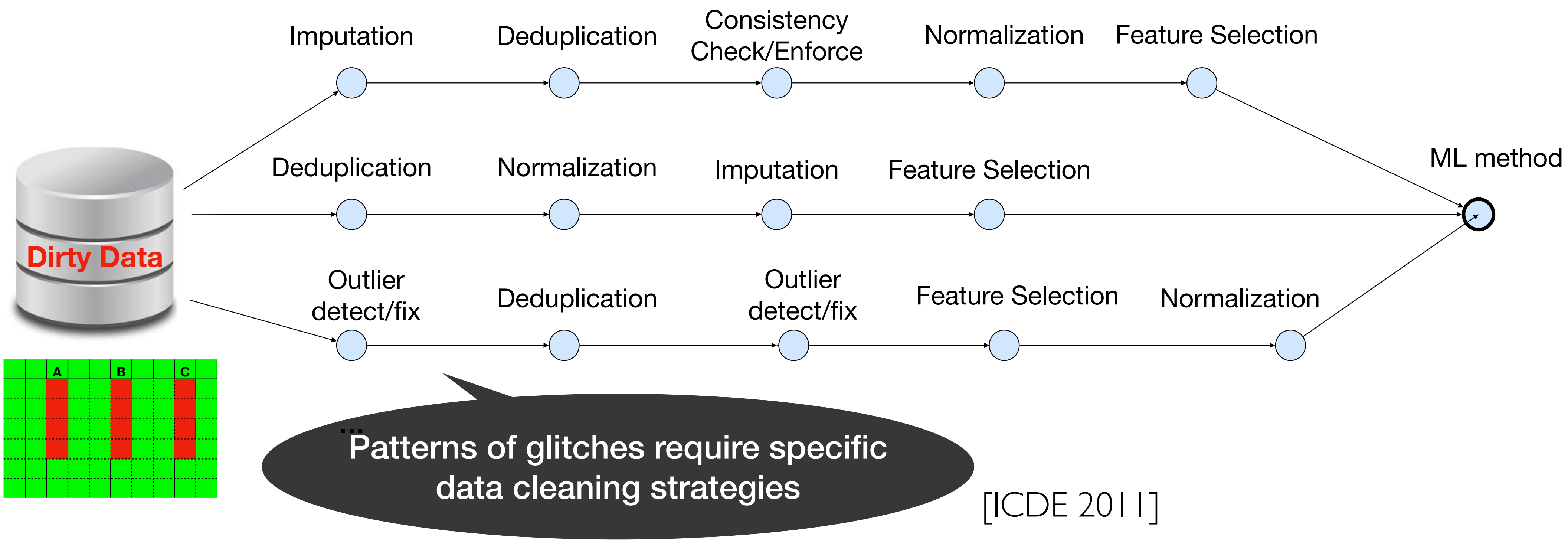
Data preprocessing is challenging



Data preprocessing is challenging

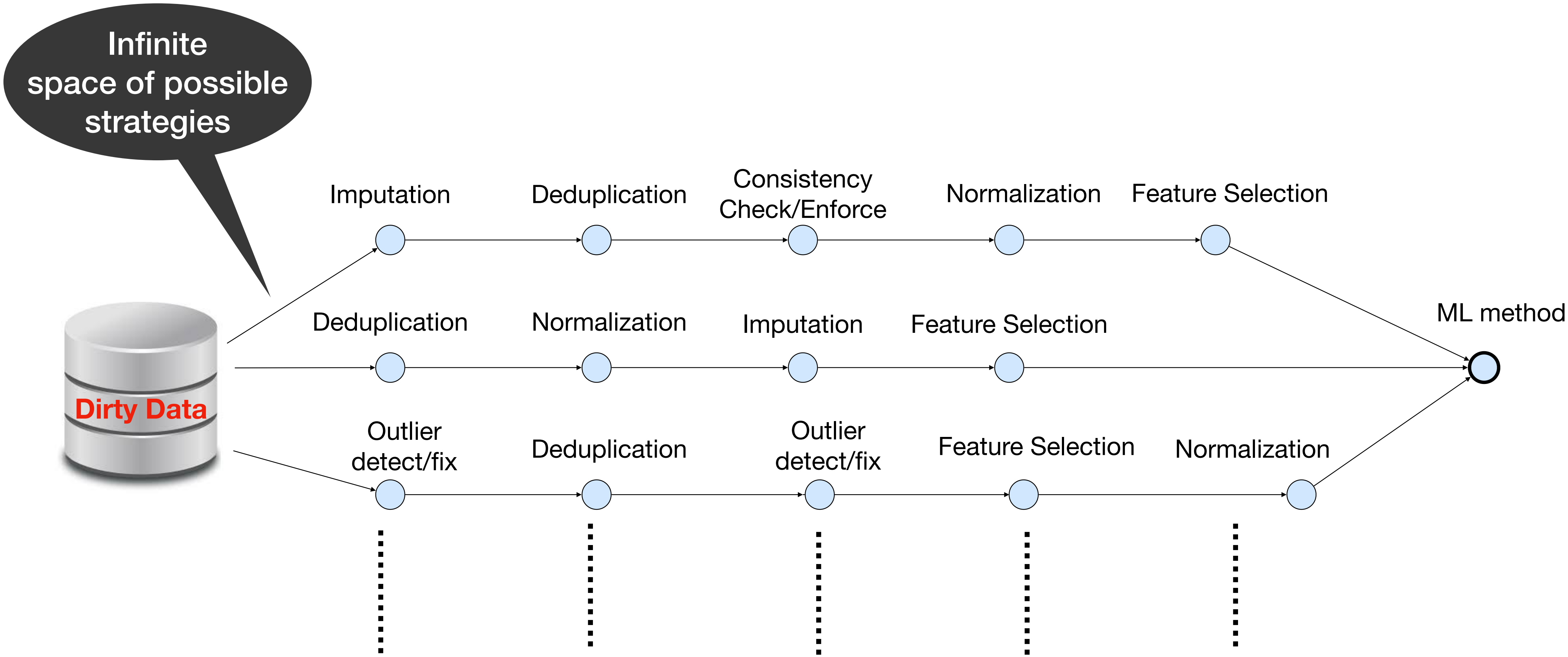


Data preprocessing is challenging



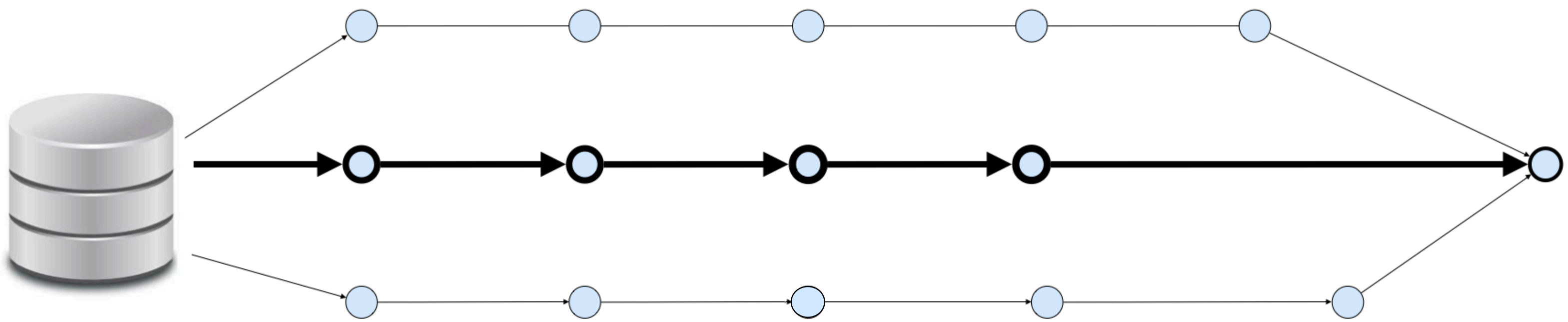
[ICDE 2011]

Data preprocessing is challenging



Optimization Problem

Can we help the user in composing the data preparation pipeline that maximizes the quality performance of the ML method ?



Optimization Problem

Can we help the user in composing the data preparation pipeline that maximizes the quality performance of the ML method ?

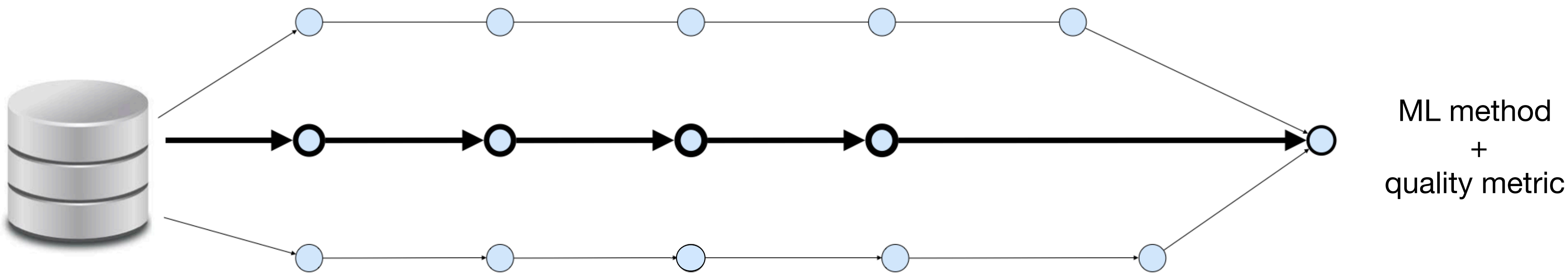
No training example for "good" data cleaning

Metric-dependent

No model a priori

AutoML-like approach

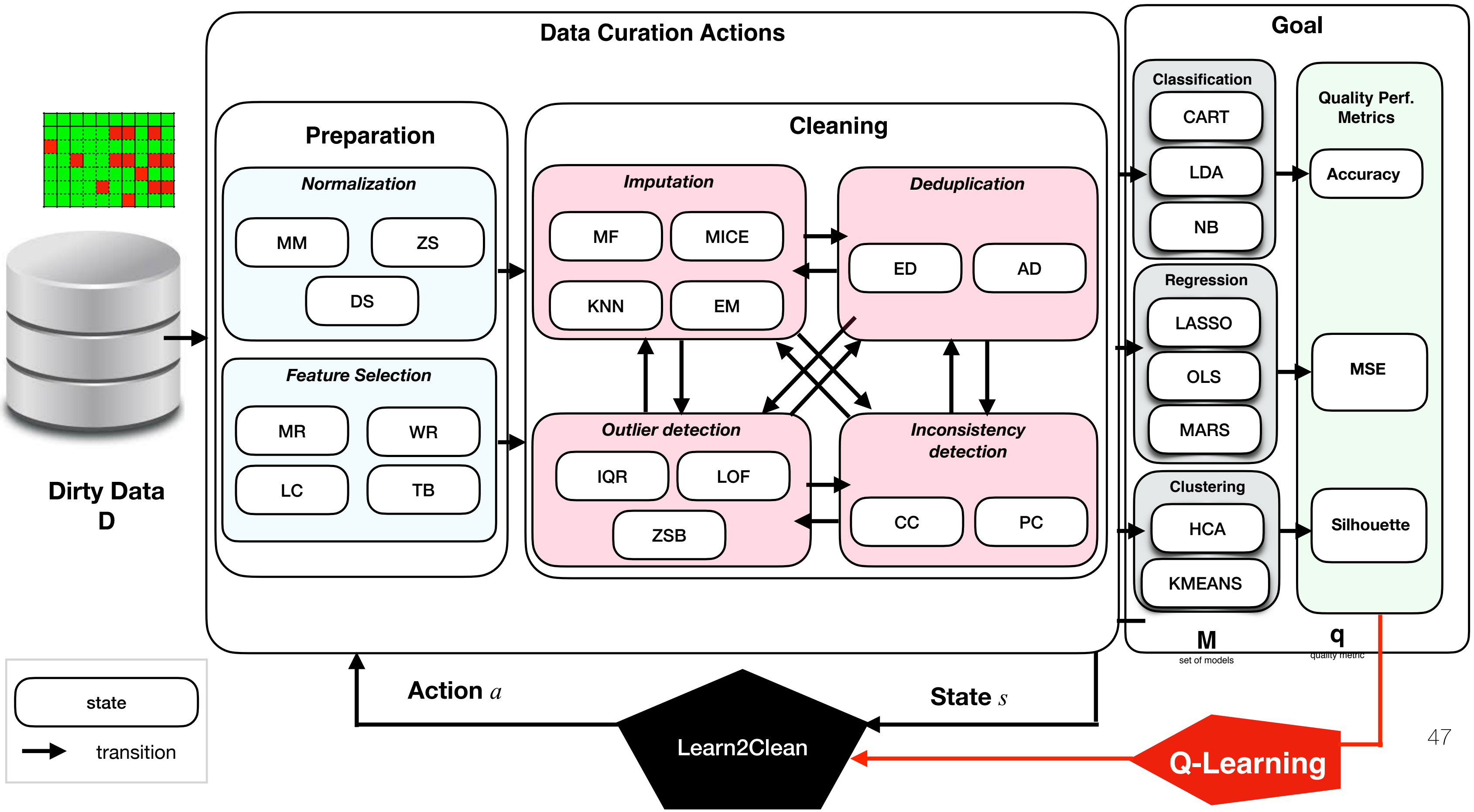
Human-In-The Loop



[The WebConf 2019]

First Solution: Learn2Clean

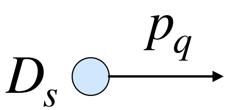
https://github.com/LaureBerti/Learn2Clean



AutoML-like approach for data curation

Reinforcement Learning Framework

Markov Decision Process State Action Transition Reward **Learn2Clean**



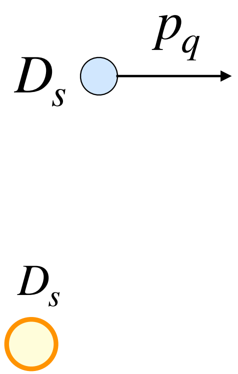
[The WebConf 2019]

Reinforcement Learning Framework

Markov Decision Process

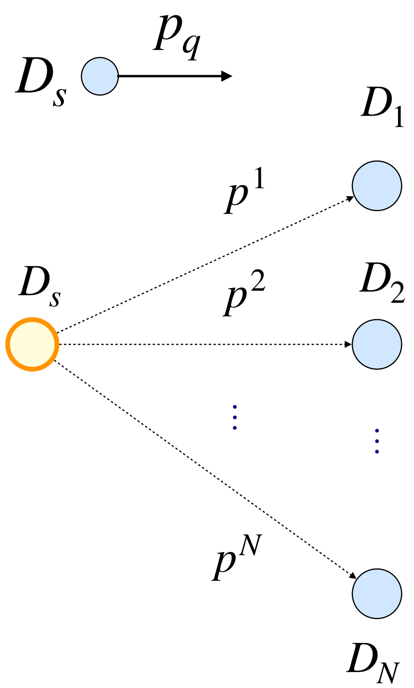
State | Action | Transition | Reward

Learn2Clean



Reinforcement Learning Framework

Markov Decision Process | State | **Action** | Transition | Reward | **Learn2Clean**

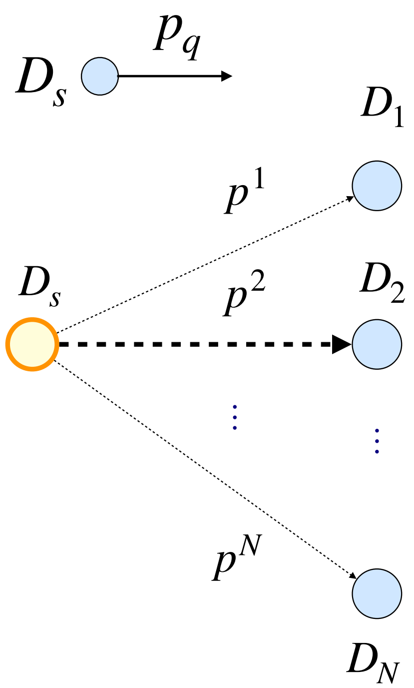


A

[The WebConf 2019]

Reinforcement Learning Framework

Markov Decision Process | State | **Action** | Transition | Reward | Learn2Clean



A

- imputation: MICE, EM, KNN, MF
- normalization: DS, MM, ZS
- feature selection: MR, WR, LC, TB
- outlier detect/fix: ZSB, LOF, IQR
- consistency check/fix: CC, PC
- duplicate detect/fix: AD, ED
- LASSO or OLS or MARS for regression
  HCA or KMEANS for clustering
  CART or LDA or NB for classification

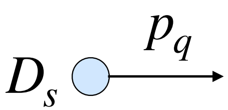
[The WebConf 2019]

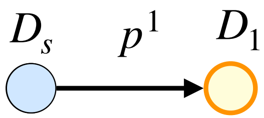
Reinforcement Learning Framework

Markov Decision Process

State | Action | **Transition** | Reward

Learn2Clean





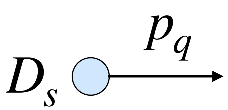
[The WebConf 2019]

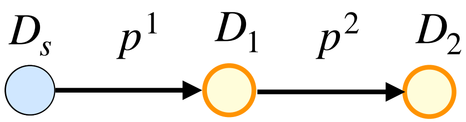
Reinforcement Learning Framework

Markov Decision Process

State | Action | **Transition** | Reward

Learn2Clean





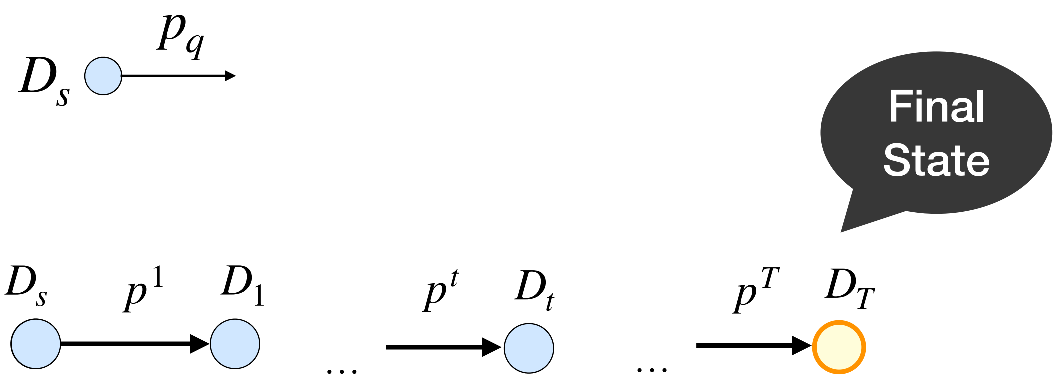
[The WebConf 2019]

Reinforcement Learning Framework

Markov Decision Process

State | Action | **Transition** | Reward

Learn2Clean



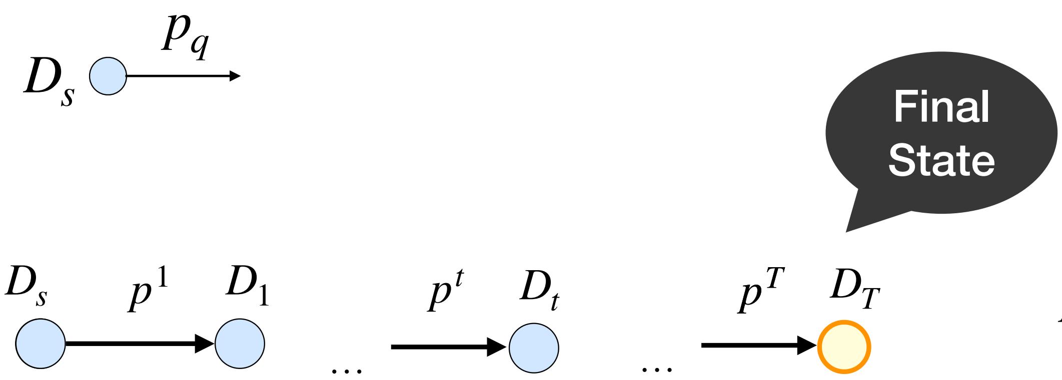
LASSO or OLS or MARS for regression
HCA or KMEANS for clustering
CART or LDA or NB for classification

Reinforcement Learning Framework

Markov Decision Process

State | Action | Transition | **Reward**

Learn2Clean



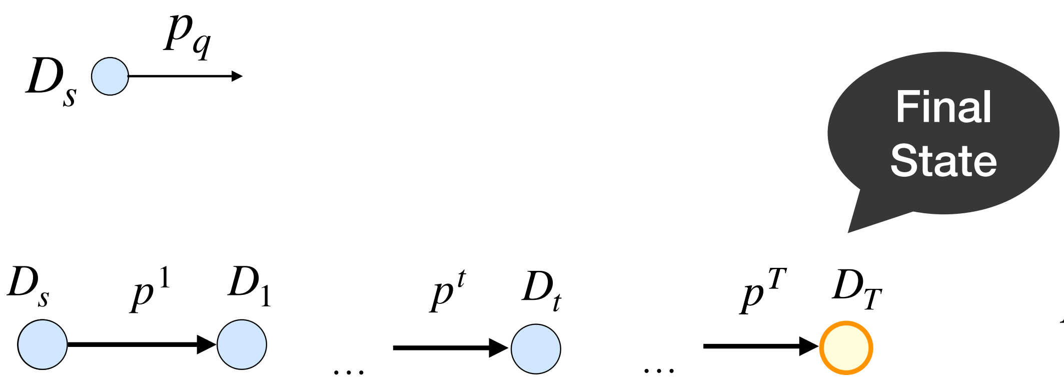
deterministic

$R =$

MICE	EM	KNN	MF	DS	MM	ZS	MR	WR	LC	TB	ZSB	LOF	IQR	CC	PC	AD	ED	LASSO
−1	−1	−1	−1	0	0	0	−1	0	0	0	0	0	0	−1	−1	0	0	100
−1	−1	−1	−1	0	0	0	−1	0	0	0	0	0	0	−1	−1	0	0	100
−1	−1	−1	−1	0	0	0	−1	0	0	0	0	0	0	−1	−1	0	0	100
−1	−1	−1	−1	0	0	0	−1	0	0	0	0	0	0	−1	−1	0	0	100
−1	−1	−1	−1	−1	−1	−1	0	0	0	0	0	0	0	−1	−1	0	0	−1
−1	−1	−1	−1	−1	−1	−1	0	0	0	0	0	0	0	−1	−1	0	0	−1
0	0	0	0	−1	−1	−1	0	0	0	0	0	0	0	−1	−1	0	0	−1
0	0	0	0	0	0	0	−1	−1	−1	−1	0	0	0	−1	−1	0	0	−1
0	0	0	0	0	0	0	−1	−1	−1	−1	0	0	0	−1	−1	0	0	−1
0	0	0	0	0	0	0	−1	−1	−1	−1	0	0	0	−1	−1	0	0	−1
0	0	0	0	0	0	0	−1	−1	−1	−1	0	0	0	−1	−1	0	0	−1
0	0	0	0	−1	−1	−1	0	0	0	0	−1	−1	−1	−1	−1	0	0	100
−1	−1	−1	−1	−1	−1	−1	0	0	0	0	−1	−1	−1	−1	−1	0	0	100
0	0	0	0	−1	−1	−1	0	0	0	0	−1	−1	−1	−1	−1	0	0	100
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	0	−1	−1	0	0	100
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	0	0	0	−1	−1	0	0	100
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	−1	−1	100
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	−1	−1	100
−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

Reinforcement Learning Framework

Markov Decision Process | State | Action | Transition | **Reward** | Learn2Clean



deterministic

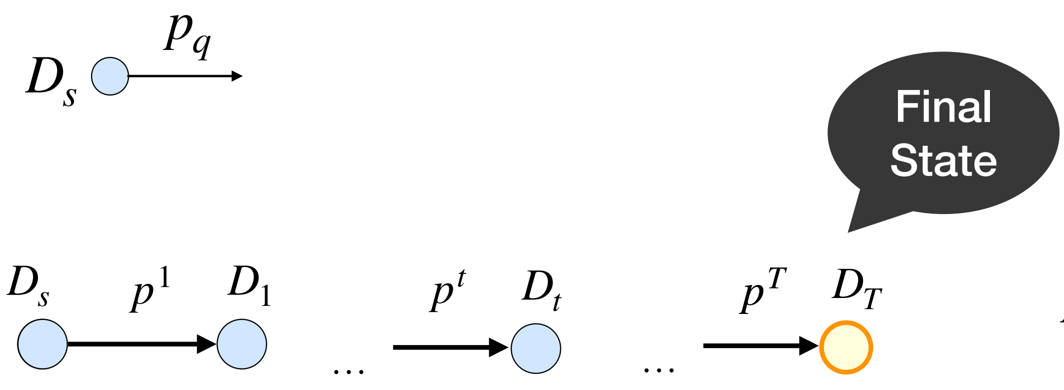
MICE EM KNN MF DS MM ZS MR WR LC TB ZSB LOF IQR CC PC AD ED LASSO

$$R = \begin{bmatrix}
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 100 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 100 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{bmatrix}$$

LASSO or OLS or MARS for regression → MSE
HCA or KMEANS for clustering → Silhouette
CART or LDA or NB for classification → Accuracy

Reinforcement Learning Framework

Markov Decision Process — State | Action | Transition | **Reward** — Learn2Clean



deterministic

MICE EM KNN MF DS MM ZS MR WR LC TB ZSB LOF IQR CC PC AD ED LASSO

$$R = \begin{bmatrix}
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & -1 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 100 \\
0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & -1 & -1 & 0 & 0 & 100 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 100 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 100 \\
-1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{bmatrix}$$

LASSO or OLS or MARS for regression → MSE
HCA or KMEANS for clustering → Silhouette
CART or LDA or NB for classification → Accuracy

Quality metric

Learn2Clean selects the sequence of preprocessing actions that maximizes the quality metric (or minimizes the error)

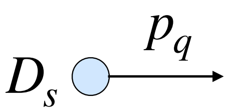
[The WebConf 2019]

Reinforcement Learning Framework

Markov Decision Process

State | Action | Transition | **Reward**

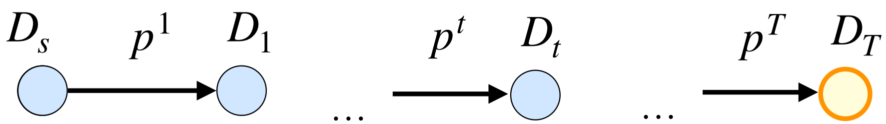
Learn2Clean



Policy

Softmax action selection

$$\pi = P(a \mid s) = \frac{e^{Q(s,a)/k}}{\sum_j e^{Q(s,a_j)/k}}$$



Q-table

Value iteration update

$$Q^{\pi}(s,a) \leftarrow (1-\alpha)\,.\,Q(s,a) + \alpha\,.\,\overbrace{\left(R(s,a) + \gamma\,.\,\max_{a'} Q(s',a')\right)}^{\text{learned value}}$$

new value — $Q^{\pi}(s,a)$; learning rate — α; old value — $Q(s,a)$; reward — $R(s,a)$; discount factor — γ; optimal future value — $\max_{a'} Q(s',a')$

[The WebConf 2019]

Experiments

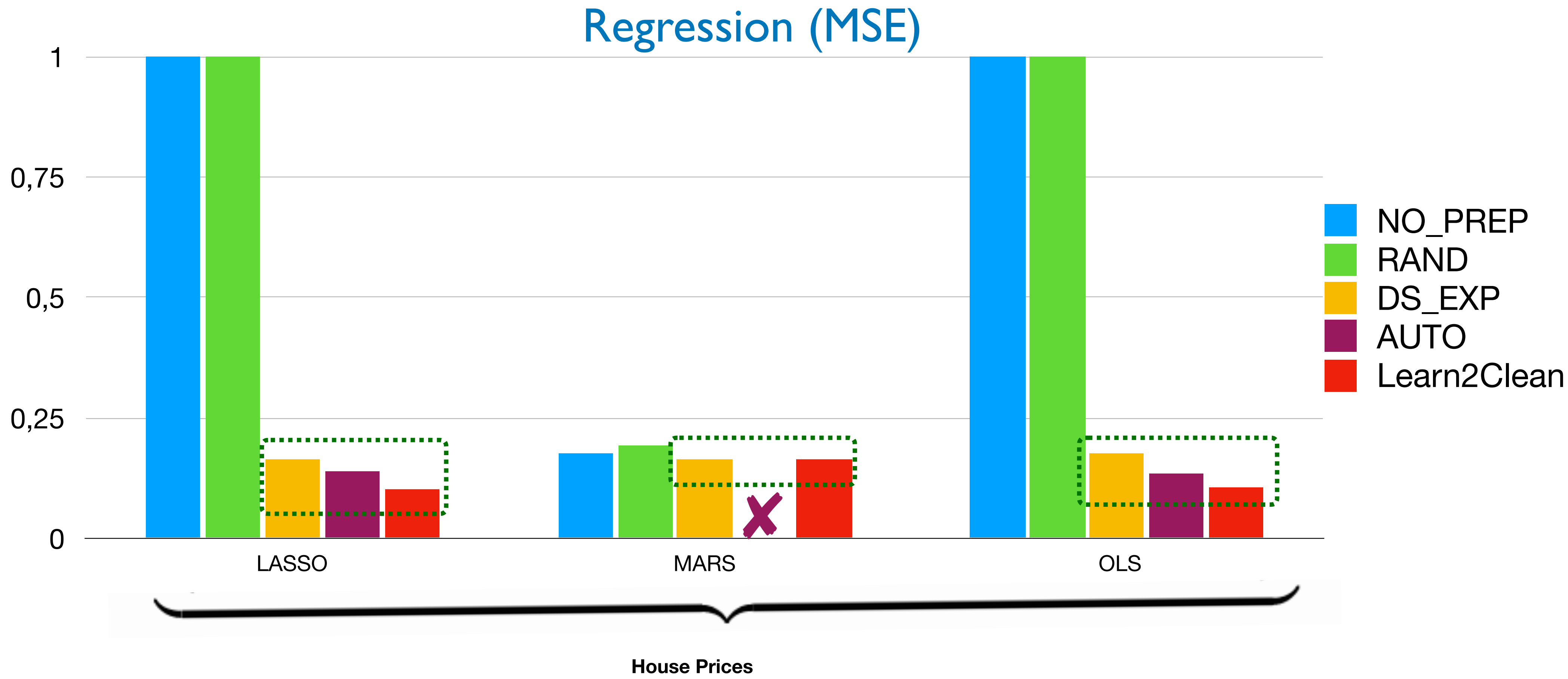
Datasets

Name	# Att.	# Rows	Clustering	Regression	Classification
House Prices	81	1.46k	✓	✓	✓
Google Playstore Users	5	64.3k	✓		
Google Playstore Apps	13	10.8k	✓		✓

Evaluation : Silhouette for Clustering
MSE for Regression
Accuracy for Classification

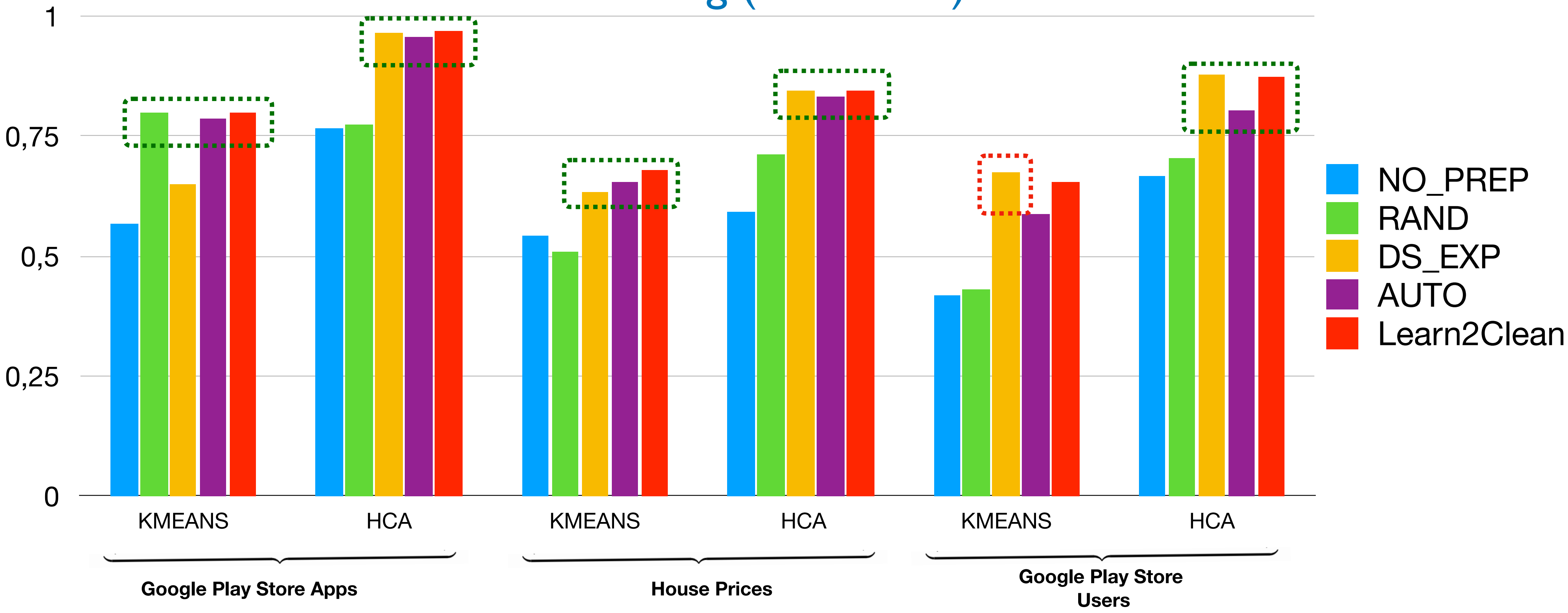
[The WebConf 2019]

Experimental Results

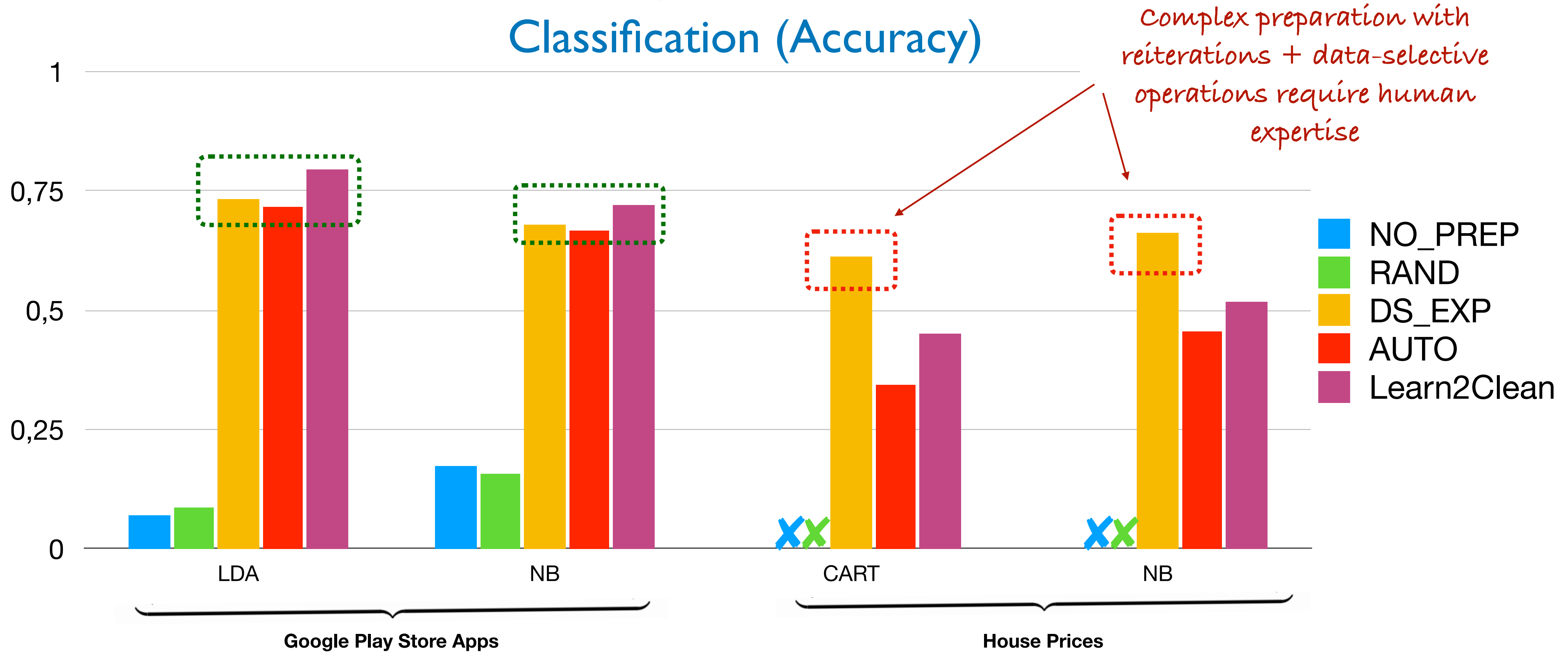


Experimental Results

Clustering (Silhouette)

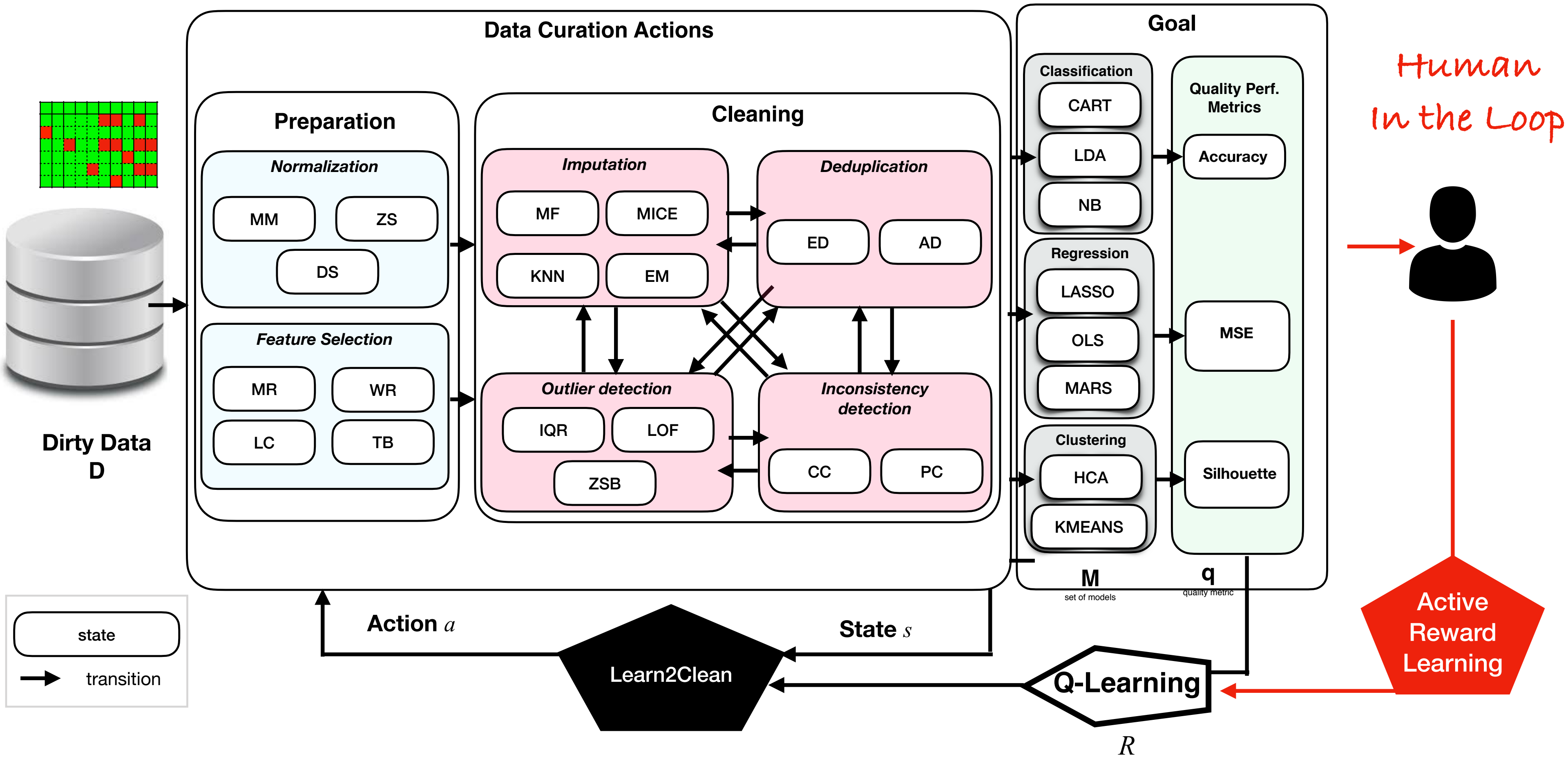


Experimental Results



[HILDA@SIGMOD2019]

HIL with Active Reward Learning

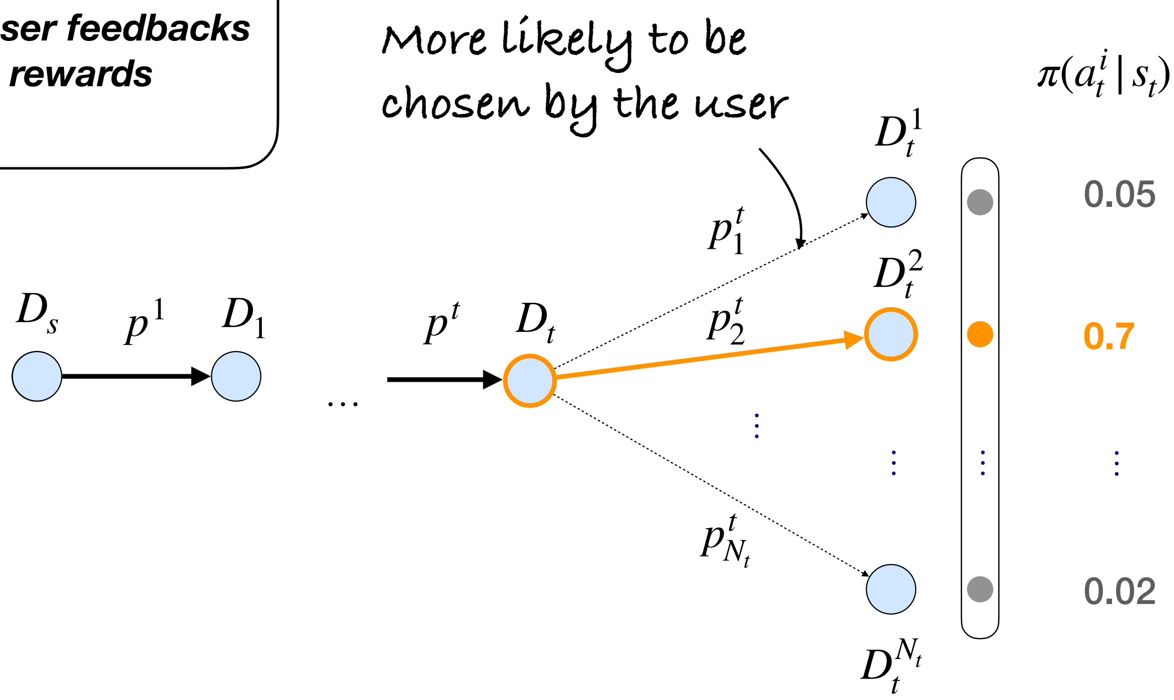


[HILDA@SIGMOD2019]

Learn2Clean
+
HIL

Active Reward Learning

Goal: ***learn from user feedbacks to adapt the rewards***

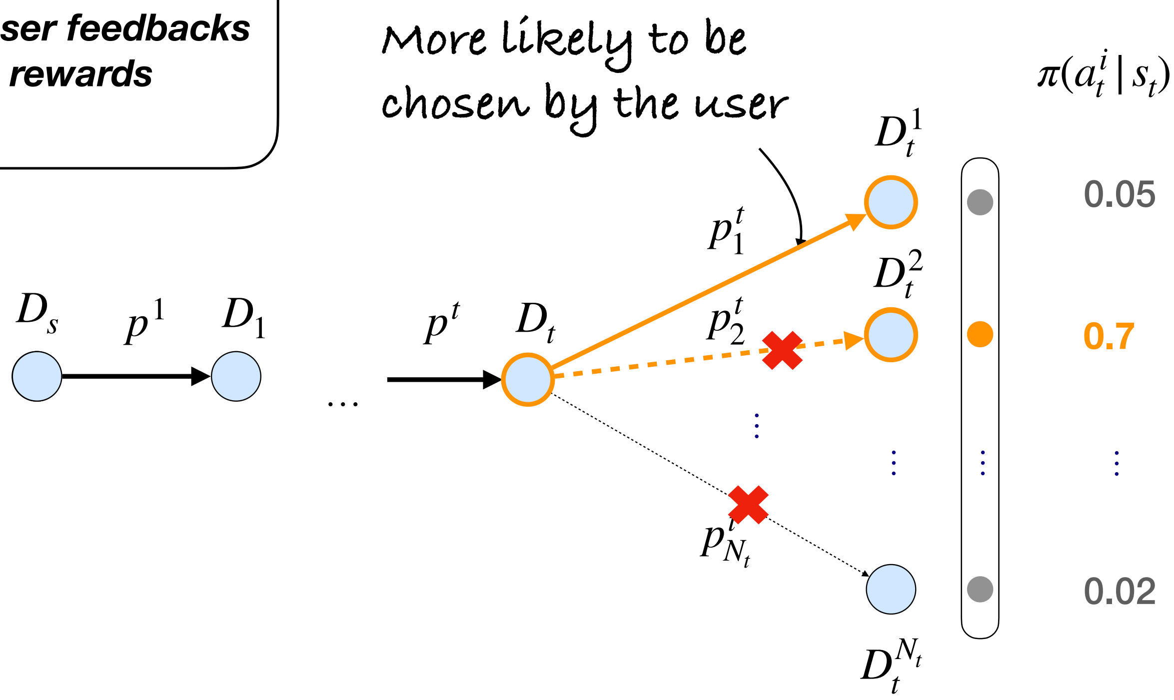


[HILDA@SIGMOD2019]

Learn2Clean
+
HIL

Active Reward Learning

Goal: ***learn from user feedbacks to adapt the rewards***

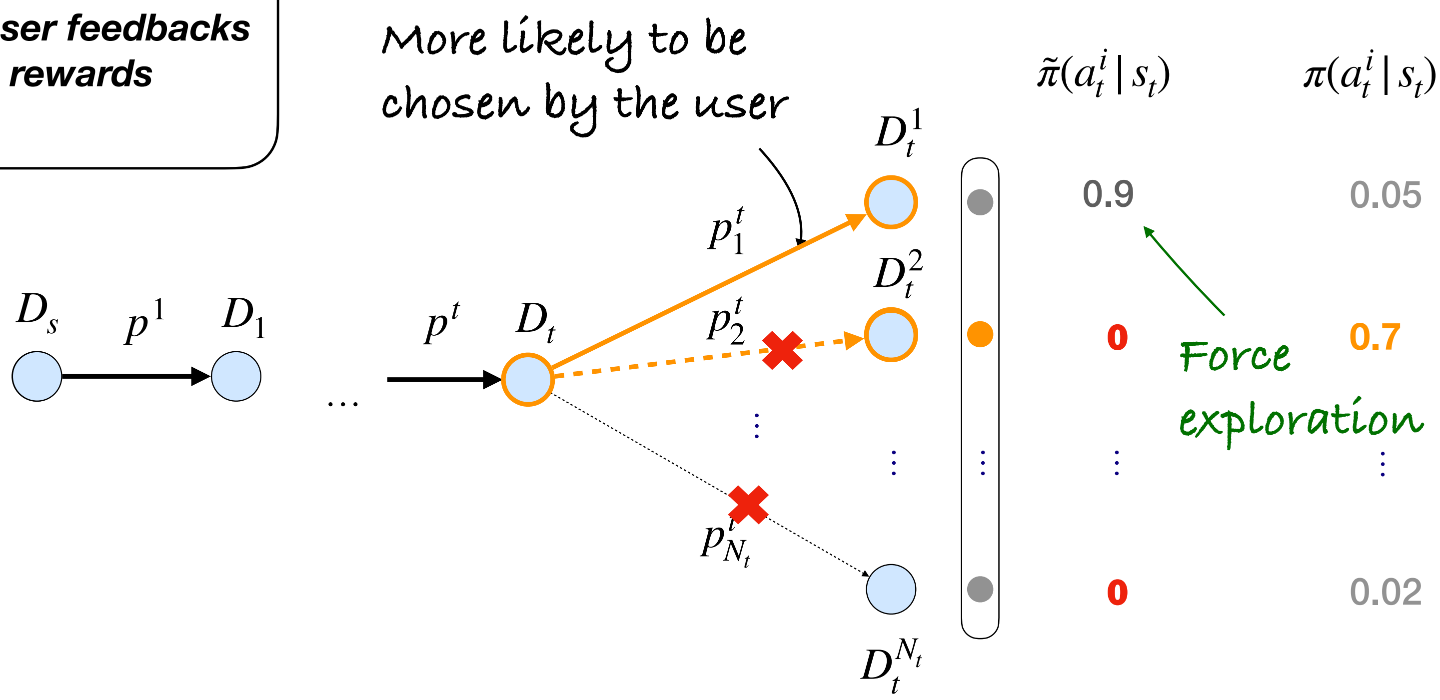


[HILDA@SIGMOD2019]

Learn2Clean
+
HIL

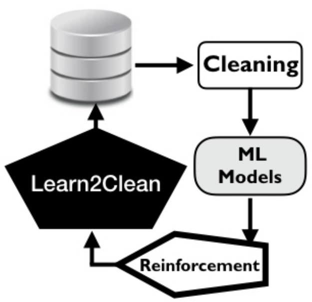
Active Reward Learning

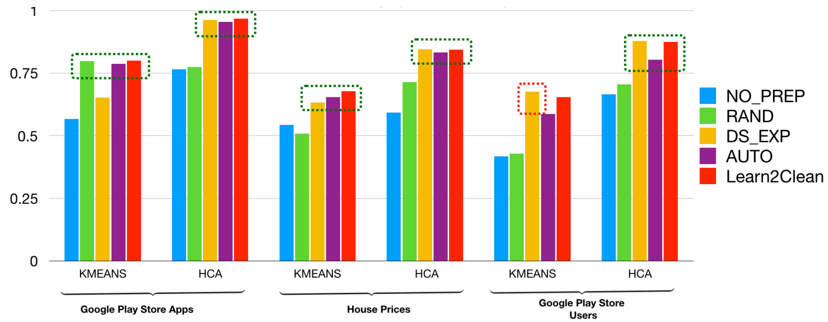
Goal: ***learn from user feedbacks to adapt the rewards***



Ongoing work

- New version of Learn2Clean with deep RL agents
- Combine AutoML, AutoCuration, and HIL
- Learn better reward functions
- Extend the library of ML and data preparation methods
- Extend experiments with more intricate data glitches and various glitch distributions





Code: https://github.com/LaureBerti/Learn2Clean

Concluding Remarks

- ML crucially needs principled data curation and preparation, adequate tooling, and user assistance
- The impact of data preprocessing variability is largely underestimated in ML
- Many data preprocessing tasks require seamless integration of Human-in-the-Loop and automated ML-based solutions
- Perfect timing for many R&D opportunities:
  - Manage and orchestrate human/machine resources
  - Challenge and transfer research ideas to operational and very large-scale contexts

Thank you!