Algorithm 1: DLM(r, t_0, κ)

Input: r a set of complete tuples,
t_0 a tuple with incomplete attributes U,
κ the number of tuples with the largest likelihood considered for evaluating a candidate
Output: an imputation t_0' of t_0 with the maximum value likelihood

$t_0' \leftarrow t_0$;
for *each incomplete attribute* $A_q \in U$ *in* t_0 **do**
  for *each* $t_0'[A_q] \in \text{can}(t_0[A_q])$ **do**
    for *each* $t_i \in r$ **do**
      for *each* $A_p \in R$ **do**
        $\mathcal{L}(t_0'[A_q] \mid t_i, A_p) \leftarrow$ likelihood computed by Formula 15;
      $\mathcal{L}(t_0'[A_q] \mid t_i) \leftarrow$ likelihood computed by aggregating $\mathcal{L}(t_0'[A_q] \mid t_i, A_p)$;
    $r(t_0'[A_q]) \leftarrow$ set of tuples $t_i \in r$ with κ-largest likelihoods $\mathcal{L}(t_0'[A_q] \mid t_i)$;
    $\mathcal{L}(t_0'[A_q] \mid r) \leftarrow$ likelihood computed by aggregating $\mathcal{L}(t_0'[A_q] \mid t_i)$ in Formula 12;
  $t_0'[A_q] = \arg\max_{t_0'[A_q] \in \text{can}(t_0[A_q])} \mathcal{L}(t_0'[A_q] \mid r)$;
return t_0'

artificially injecting errors [3], we randomly remove values from various attributes in tuples as incomplete data.

A.5.2 Real Missing Values and Labeled Truth. GPS is a dataset manually collected by carrying a GPS device and walking around the campus. The GPS readings are often temporally unavailable, owing to various reasons such as low battery, energy saving or weak signals. There are 54 incomplete tuples with real missing values on Longitude and Latitude naturally embedded in the dataset. The corresponding timestamp is available as the complete attribute, since the collecting program intends to record the GPS readings in every second. Moreover, as we know exactly the trajectory, truths of the missing values are manually labeled.

A.5.3 Real Missing Values without Ground Truth. Datasets MAM[12] and Adult[13] are two datasets with 13.63% and 7.41% real-world missing values. MAM contains a BI-RADS assessment, which can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. Adult is extracted from census data of the United States in 1994, describing some social information about the citizens. Rather than comparing the imputation accuracy, we study the classification over the datasets without/with various imputation methods. A higher classification accuracy [13] generally indicates the better effectiveness of imputation.

A.5.4 Criteria. For numerical values, to evaluate the imputation accuracy, we compare the imputation t_0' to the corresponding truth $t_0^{\#}$, using the root-mean-square error, $\text{RMS} = \sqrt{\sum_{A \in U} \frac{(t_0'[A] - t_0^{\#}[A])^2}{|U|}}$.

[12] https://sci2s.ugr.es/keel/dataset.php?cod=86
[13] https://sci2s.ugr.es/keel/dataset.php?cod=192

For categorical datasets, e.g., Restaurant, we measure the accuracy on whether an imputation equals exactly the ground truth. Let truth be the set of truth values for incomplete data and found be the set of results returned by imputation algorithms. The accuracy is given by $\text{accuracy} = \frac{|\text{truth} \cap \text{found}|}{|\text{truth}|}$, i.e., the proportion of incomplete values accurately imputed.

A.5.5 Implementation Details. As listed in Table 2, the major competitors of our proposal are as follows.

(1) Neighbor-based kNN Ensemble (kNNE) [8] employs different subsets of complete attributes in the incomplete tuple, to generate a diverse set of NN classifiers for ensemble learning.

(2) Tuple similarity-based MIBOS [31] finds the complete tuples having the maximum number of same values on complete attributes with the incomplete tuple, to impute the missing values.

(3) Gaussian mixture model-based GMM [33] clusters the complete data based on EM algorithm. The incomplete data are classified according to the result of clustering. The complete tuples closest to the incomplete tuple in the same cluster are utilized to impute the missing values.

(4) Clustering-based CMI [35] employs k-Means clustering to divide the dataset (including the instances with missing values) into clusters. The complete tuples in the same cluster are used to impute the incomplete tuples.

(5) Value regression model-based LOESS [5] learns regression models among nearest neighbors to impute missing values, where the incomplete attributes are used as the dependent attributes and all the other complete attributes are regarded as the determinant attributes of the regression models.

(6) Individual regression-based IIM [34] learns individual regression models for each complete tuple, whose determinant and dependent attributes are the complete and incomplete attributes of incomplete tuples, respectively.

(7) Statistics-based ERACER [19] iteratively learns a global relational dependency model, i.e., linear regression for numerical data and relational dependency network for categorical data in our experiments, to infer the probabilistic relationships among attributes.

(8) Matrix completion via convex optimization MC [4] forms a convex relaxation of the matrix completion problem and minimizes the nuclear norm. The convex relaxation is solved using semidefinite programming.

(9) ER [12] finds certain fixes with editing rules defined by value equality. We consider all the complete tuples as master data and discover CFDs from r (by [10]) as editing rules.

(10) HoloClean [24] employs not only the aforesaid discovered CFDs as constraints but also the statistical learning and probabilistic inference to impute missing values.

(11) Similarity constraint-based DD [28] imputes missing values with differential dependencies, discovered by [27].

The imputation candidates of DLM are given by the kNN neighbors as introduced in Section 3.1. Similar to distance models in our DLM, the models in the GMM approach could be offline learned. In order to mitigate the risk of a sub-optimal configuration and insufficient rules in the case of rule-based tools, each approach in comparison has been configured in a best-effort fashion (e.g., by iteratively choosing good parameters or by defining a reasonable set of quality rules).