Collaborative Analytics for Data Silos

Jinkyu Kim[1,4], Heonseok Ha[1], Byung-Gon Chun[2,3], Sungroh Yoon[1,3], and Sang K. Cha[1,3]
[1]ECE, [2]CSE, and [3]Big Data Institute, Seoul National University, Seoul 08826, Korea
[4]EECS, UC Berkeley, Berkeley, CA 94720, USA
Correspondence: sryoon@snu.ac.kr

***Abstract*—As a great deal of data has been accumulated in various disciplines, the need for the integrative analysis of separate but relevant data sources is becoming more important. Combining data sources can provide global insight that is otherwise difficult to obtain from individual sources. Because of privacy, regulations, and other issues, many large-scale data repositories remain closed off from the outside, raising what has been termed the *data silo issue*. The huge volume of today's big data often leads to computational challenges, adding another layer of complexity to the solution. In this paper, we propose a novel method called *collaborative analytics by ensemble learning* (CABEL), which attempts to resolve the main hurdles regarding the silo issue: accuracy, privacy, and computational efficiency. CABEL represents the data stored in each silo as a compact aggregate of samples called the *silo signature*. The compact representation provides computational efficiency and privacy preservation but makes it challenging to produce accurate analytics. To resolve this challenge, we formulate the problem of *attribute domain sampling and reconstruction*, and propose a solution called the *Chebyshev subset*. To model collaborative efforts to analyze semantically linked but structurally disconnected databases, CABEL utilizes a new ensemble learning technique termed the *weighted bagging of base classifiers*. We demonstrate the effectiveness of CABEL by testing with a nationwide health-insurance data set containing approximately 4,182,000,000 records collected from the entire population of an Organisation for Economic Co-operation and Development (OECD) country in 2012. In our binary classification tests, CABEL achieved median recall, precision, and F-measure values of 89%, 64%, and 76%, respectively, although only 0.001–0.00001% of the original data was used for model construction, while maintaining data privacy and computational efficiency.**

I. Introduction

The era of big data has been triggered by the constructive combination of sophisticated algorithms, powerful parallel hardware, and massive collections of data. The explosion of data is also driving the demand for new applications of data analytics [1], [2]. In particular, as various types of data are being accumulated in individual databases, the need for the integrative analysis of separate but relevant databases is becoming more important. Despite its potential, integrative analysis remains challenging in practice. Because of privacy, regulations, and other issues, many data repositories cannot be accessed from the outside. This problem is often referred to as the *data silo issue* [3].

To address the privacy issue encountered during data analysis, various techniques have been proposed in the areas of privacy-preserving data publishing (PPDP) [4] and privacy-preserving data mining (PPDM) [5], [6], [7]. These methods aim for effective information sharing and data analytics while protecting sensitive information. Also relevant is distributed data mining [8], [9], which aims to build models by making the best use of available resources in distributed environments. When applying these techniques effectively to handling the data silo issue, however, we see critical concerns remain, such as a limited range of information sharing among secure multi-parties, data-quality degradation, information loss, and increased costs and complexity.

In this paper, we propose a novel method called *collaborative analytics by ensemble learning* (CABEL) to tackle the main challenges involved in today's big-data silo issue: privacy protection and computational efficiency. CABEL introduces the notion of the silo signature, which can compactly and faithfully describe the data stored in individual silos while preserving privacy. To create silo signatures, we introduce the problem of attribute domain sampling and reconstruction, and propose a solution named the *Chebyshev subset*, which is a Chebyshev polynomial-based sample of an attribute domain. CABEL analyzes data sets using only their silo signatures by using a new ensemble learning method called the *weighted bagging of base classifiers* (WBBC). The key idea behind ensemble learning is to reduce the generalization error by training multiple weak models and then combining their outputs. In CABEL, training a weak model corresponds to mining an individual silo, while combining weak models corresponds to integrating the results from individual silos.

For evaluation, we test CABEL with a nationwide health-insurance data set containing approximately 4,182,000,000 records collected from the entire population of an OECD country in 2012. Among various analytics tasks possible by using CABEL, this paper focuses on supervised classification. According to our experiments, CABEL achieved median recall, precision, and F-measure scores exceeding 89%, 64%, and 76%, respectively, despite using only 26 samples per attribute out of the entire set of 4,182,000,000 records (only 0.001–0.00001%) to construct an integrative binary classifier. The running time of CABEL was incomparably faster than the alternatives that require the entire data set. By using the silo signatures rather than the attributes of the original data, privacy was preserved. Our evaluation results confirm that CABEL can effectively address the silo issue.

II. Related Work

A. Privacy Preserving Data Publishing and Mining

PPDP involves publishing data without compromising privacy, preventing the information of a record owner from being linked to records (record linkage), to specific attributes (attribute linkage), or to the table (table linkage) [4]. Methods of protecting privacy from these attacks include k-anonymity [10], l-diversity [11], and δ-presence [12].

To overcome the privacy issue encountered during data analysis, PPDM was proposed [5], [6]. Both PPDP and PPDM consider the trade-off between privacy and data utility, but PPDM mainly aims to construct a data-mining model that allows no access to confidential data. According to surveys [7], [13], three types of PPDM approaches exist: randomized methods, anonymization, and privacy-preserving distributed data mining (PPDDM) [9].

The first category of PPDM relies on randomization for privacy preservation. These methods typically distort data by adding noise to mask the original records and then develop data-mining models based on the distorted data. For instance, Agrawal and Srikant [5] proposed a technique that can estimate the distribution of attributes in the original records even when they are perturbed by noise. Based on a reconstruction procedure, a decision-tree classifier was built whose accuracy was comparable to that of a classifier built with the original data. Discussions of the limitations of randomization and solutions can be found in prior work [14], [15].

The second category of PPDM relies on anonymization, as is also studied in PPDP. Bayardo and Agrawal [16] proposed a k-anonymity-based method that utilizes generalization and suppression to reduce the probability of identifying specific records. Finding the optimal k-anonymization is an NP-hard problem [17], and various heuristics [18], [19], [20] and approximation algorithms [17] have been proposed. Another example is the personalized privacy preservation [21], which relies on assumptions different from those associated with the conventional k-anonymity method to ensure personalized security levels. Additional examples include measuring the information loss during anonymization [16], and reducing the loss in data utility while preserving privacy [22].

PPDDM approaches [9], [7] comprise the third category of PPDM. They gather distributed data using a secure communication protocol to generate aggregated results. The participating parties (*i.e.*, data silos in our context) can be divided into the semi-honest parties that follow the protocol and the malicious parties that do not follow the protocol. The data used by PPDDM can be assumed to have either horizontally or vertically partitioned distributions (see Section III). Secure multiparty computation (SMC) [9] is an approach that can provide a strong level of privacy. To create an analysis model, PPDDM-participating silos can exchange information using an SMC-based protocol. Examples of the specific PPDDM techniques are discussed in Section III.

B. Additional Related Work

Notable examples of collaborative analytics in the medical domain include the Observational Health Data Sciences and Informatics (OHDSI) [23]. It provides an open-source knowledge base for managing and mining health-related information by integrating multiple databases into a standardized structure. OHDSI provides not a specific method for collaborative analytics but an open framework, in which CABEL can be hosted.

In descriptive statistics, there have been two major approaches to describing (non)parametric models [24]: the central tendency (*i.e.*, where the average value lies; *e.g.*, mean, median, and mode) and the variability (*i.e.*, how data values deviate from the central tendency; *e.g.*, range and standard deviation). Based on these, various data summarization approaches have been proposed for solving data-mining problems with reduced computational cost [25], [26].

The idea of weighted bagging has appeared in the literature [27], but there is a fundamental difference between CABEL and the prior work. CABEL generates bootstrap samples of base classifiers instead of data points. For outlier detection, feature bagging [28] was proposed, which also bears some connection to silo-based learning. A fine-grained classifier combination scheme was recently proposed (the independent Bayesian classifier combination [29]), which can adopt the variability of decision-making abilities of base classifiers for different areas in the input space. With heterogeneous types of attribute signatures, we may be able to devise a fine-grained combination suitable for CABEL.

III. Contributions and Advantages of CABEL

Suppose that a silo stores data in a table where each row represents a record of an object and each column represents an attribute of the record. To collect the data for analysis, existing PPDDM approaches extract from each silo a horizontal (row-wise) or vertical (column-wise) subset of the table stored therein. For such approaches to work properly, however, there needs to be an assumption that all the participating silos contain the same set of attributes (for horizontal-partitioning approaches) or objects (for vertical-partitioning approaches). This assumption is often too strong and unrealistic, making it challenging to use existing PPDDM approaches in practice.

In contrast, CABEL does not need such a strong assumption, since CABEL collects the data distributed over multiple silos in the form of silo signatures and analyzes the collected data by ensemble learning. The use of silo signatures not only allows us to compactly and faithfully capture and represent silo contents but also removes the requirement of a common set of objects for the analysis (as needed by vertical-partitioning methods). Furthermore, CABEL needs no common set of attributes over silos, owing to the proposed ensemble-learning scheme. This scheme obtains different sets of attributes from different silos in the form of signature-level base classifiers and then builds a strong classifier by the weighted bagging of the base classifiers, eliminating the need to unify the set of attributes (as needed by horizontal-partitioning methods). Without the strong assumption of data partitioning, we can therefore apply CABEL to a wider spectrum of problems in practice, as compared to existing PPDDM approaches.

The signature-based data representation used by CABEL is also helpful for privacy preservation. This is because the probability that multiple attributes from a single record appear in a signature decreases exponentially as the number of attributes increases. In terms of avoiding privacy breaches, such as the linkage attacks [4] and the resistance [30], CABEL thus has advantages over the existing alternatives that need all (or most of) the attributes of individual records for analysis.

Previous PPDM approaches mainly have utilized individual machine learning techniques, such as naïve Bayes classification [31], [32], decision tree classification [6], [33], [34], [35], support vector machines (SVMs) [36], [37], association rule

mining [38], [39], [40], and clustering [41], [42]. Although these pioneering approaches were successful with the specific settings they assumed, they overlooked the fact that the data silo problem fits better with, and can be generally formulated as, ensemble learning. To the best of our knowledge, CABEL is the first attempt that utilizes ensemble learning in the context of PPDM. Specifically, CABEL is based on WBBC, a novel ensemble-learning technique that constructs a strong classifier by "classifier" bagging, not by the conventional data bagging approach. WBBC utilizes only silo signatures to create base classifiers for bagging without examining the raw data stored in silos, adding another layer of privacy preservation to CABEL.

In CABEL, the participating data silos do not communicate with each other, but each silo simply delivers its silo signatures to a data miner that performs WBBC-based ensemble learning for classification. This simplified protocol lowers the computational complexity of CABEL significantly, compared with previous PPDDM approaches that require inter-silo communication in an SMC-based protocol for creating an analysis model. Furthermore, when new data silos want to participate in CABEL or other silos have updated their data, we can easily reflect the new data into the classification model by updating it with the new silo signatures in the WBBC procedure. This signature-based approach is also appropriate for parallelization.

When creating attribute signatures that comprise a silo signature, CABEL incorporates the notion of k-anonymity, effectively applying anonymization to each attribute. This differs from the existing approaches that apply k-anonymity to quasi-identifiers. According to Aggarwal [43], the anonymization (such as k-anonymity) in conventional techniques becomes difficult as the dimensionality (*i.e.*, the number of attributes) increases due to the curse of dimensionality. However, CABEL ensures k-anonymity for each attribute using the attribute signature, and for CABEL, the difficulty of enforcing k-anonymity at the entire data level remains independent of the number of dimensions. In CABEL, increasing the number of attribute signatures means considering more base classifiers during the WBBC procedure and thus implies that more data can be used for modeling, increasing the chance of improving the accuracy of the analysis.

IV. Assumptions and Definitions

For the sake of explanation, suppose that an oracle has collected all the data of individual data silos into a single table, which contains m records (*e.g.*, patients) and each record consists of p attributes (*e.g.*, diseases). The table is an $m \times p$ matrix $D = (R, A)$, where R and A are the sets of records and attributes, respectively. The *attribute domain* denoted by V_j is the set of all values appearing in column j of the matrix.

Attributes in a table can be categorized into four groups [44]: (1) *explicit identifier*, a set of attributes clarifying record owners (*e.g.*, name and social security number)—their values should remain closed to the outside; (2) *quasi-identifier* (*e.g.*, age and zip code), a set of attributes whose combined values can potentially identify owners—their specific values should be anonymized for privacy protection; (3) *sensitive attributes* containing sensitive information about individuals (*e.g.*, salary); and (4) *non-sensitive attributes*, which refer to all the other types.

Suppose that we consider L data silos, each of which stores $D_i = (R_i, A_i)$, a random subset of D, where $R_i \subset R$ and $A_i \subset A$ for $1 \leq i \leq L$. Note that $\bigcup_{i=1}^{L} R_i \subseteq R$ and $\bigcup_{i=1}^{L} A_i \subseteq A$. We assume four properties of data silos:

P1 (data confidentiality): Any confidential data in the silo should not be published, and third parties should not be able to reconstruct the confidential records with public or background knowledge [7].

P2 (data anonymization): For publishing data outside the silo, any sensitive information should be protected from being de-identified and linked to a record owner using the released data or public/background knowledge.

P3 (partial observability): Due to the policy or technical issues, only a subset of the stored data is publicly available.

P4 (trustfulness): A data silo may publish intentionally or unintentionally inaccurate records that conflict with other data sources. The information about the degree of trustfulness is often not provided [4].

Based on the above assumptions and properties, we can define collaborative analytics as follows:

Definition 1. *Given L silos having properties **P1–P4** and each storing D_i for $1 \leq i \leq L$,* collaborative analytics *refers to analyzing D_i's individually and then integrating the results with the aim of making the combined result as close as possible to that obtained by analyzing $D = \bigcup_{i=1}^{L} D_i$ as a whole.*

Note that a variety of methods are possible as an example of collaborative analytics; moreover, depending on the analysis method used, Def. 1 can be made more specific.

A. Problem Statement

This paper focuses on supervised classification. Given the whole set of data $D = (R, A)$, suppose that R contains a binary attribute. If D were known in its entirety, we could adopt a binary classifier and train it with D using each row as a training vector. Denote this hypothetical, oracle classifier by mapping $f_D : \mathcal{R} \mapsto \mathcal{Y}$, where $\mathcal{R}$ is the space spanned by the attributes in R except for the label attribute and the binary output space $\mathcal{Y} = \{-1, +1\}$. For measure π, let $\pi(f)$ denote the performance of f measured with respect to π.

Definition 2. *The problem of* collaborative classification *is to find a binary mapping $f : \mathcal{R}' \mapsto \mathcal{Y}$ that minimizes $|\pi(f_D) - \pi(f)|$ for any $\mathbf{x} = (x_1, x_2, \ldots) \in \mathcal{R}'$ by collaborative analytics, where $\mathcal{R}'$ is the space spanned by the attributes in arbitrary $R' \subset R \backslash \{label\}$ and $\mathbf{x}$ is a new test record.*

V. Proposed Methodology

To address the collaborative classification problem described in Def. 2, we propose a novel approach called CABEL. It assumes four major steps for collaborative analytics: data collection, data requesting, data publishing, and data processing (Fig. 1). In the data collection step, each data silo collects data from its own data providers (*e.g.*, patients and medical institutes), and stores it into the storage that is closed off

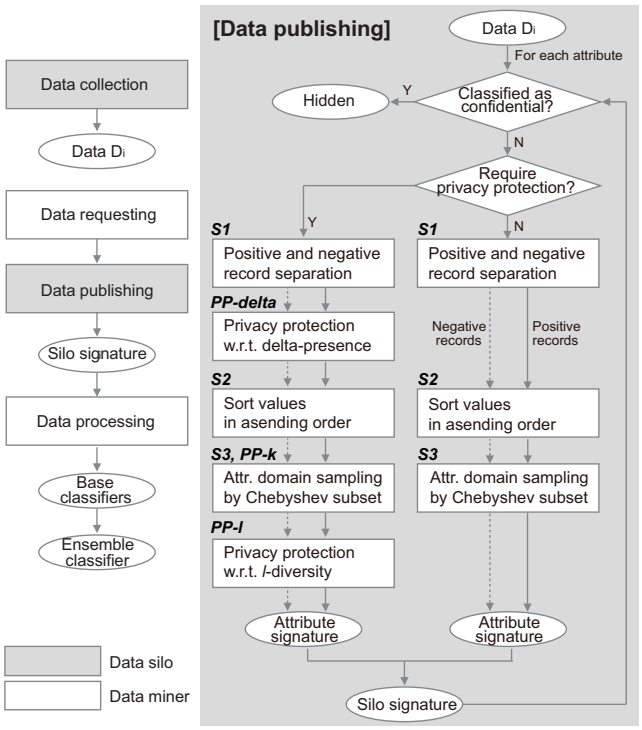

Fig. 1: Major steps in silo-based collaborative analytics (left) and the proposed algorithm to create silo signatures (right).

to the outside parties. In the data requesting step, CABEL requests information on each silo for collaborative analytics. In the data publishing step, each silo releases a compact representation of the table (termed the *silo signature*) as per **P1**–**P4**. CABEL characterizes the data stored in each silo by its signature and analyzes the silo signatures of participating silos on behalf of their data. The size of the signature of a silo is significantly smaller than that of the data stored in the silo, which is helpful for efficiency and privacy but poses a challenge for integrative analysis. In the data processing step, CABEL constructs a classification model to address this challenge. CABEL implements collaborative analytics by ensemble learning: training a weak learner corresponds to analysis inside each silo, while combining weak learners corresponds to integrating the results from participating silos.

A. Creating Silo Signatures

The flowchart in Fig. 1 shows the procedure for creating the silo signature for silo i from the perspective of its owner. To facilitate a further understanding of silo signature creation, we suggest the reader refer to the theory behind it presented in Section V-B (especially Def. 8) along the way.

As defined previously, silo i stores the data table $D_i = (R_i, A_i)$. The table is split into $D_i^+ = (R_i^+, A_i)$ and $D_i^- = (R_i^-, A_i)$, where R_i^+ and R_i^- represent the positive and negative records, respectively.

For attribute $j \in A_i$, we first check if j corresponds to an explicit identifier. If so, j is excluded from signature creation (as per property **P1**). Otherwise, we proceed as follows:

Step S1: We create set V_{ij}^+ from the j-th column of D_i^+, which represents the domain of attribute j for R_i^+. We construct V_{ij}^- in the same manner from D_i^-.

Definition 3. *For attribute j, its* positive attribute domain, *denoted by V_{ij}^+, is the set of all the values appearing in the j-column of the matrix (R_i^+, A_i). The* negative attribute domain V_{ij}^- *comes from column j in (R_i^-, A_i).*

Step S2: For each of the two sets V_{ij}^+ and V_{ij}^-, we sort the elements therein with respect to the value of attribute j. The role of this step will become clear in the proof of Theorem 1.

Step S3: By following the procedure explained in Section V-B2, we perform sampling of V_{ij}^+ and V_{ij}^- to obtain two subsets $C_{ij}^+ \subset V_{ij}^+$ and $C_{ij}^- \subset V_{ij}^-$. Each of these two subsets consists of n elements and becomes the ingredients of the silo signature.

Definition 4. *Given a positive integer $n < |R_i^+|, |R_i^-|$ and the attribute domains V_{ij}^+ and V_{ij}^- of attribute j, n-element multisets C_{ij}^+ and C_{ij}^- are the* degree-n Chebyshev subsets *(see Def. 8) of V_{ij}^+ and V_{ij}^-, respectively.*

Normally $n \ll |R_i^+|, |R_i^-|$, and the Chebyshev subsets C_{ij}^+ and C_{ij}^- thus provide a compact representation of the attribute domains V_{ij}^+ and V_{ij}^-, which allows the silo owner to maintain property **P3**. The use of Chebyshev subsets enables us to analyze the data in a silo without publishing the entire data set.

Additionally, if attribute j is either a quasi-identifier or a sensitive attribute, then we perform three types of privacy preservation tasks (denoted as PP-δ, PP-k, and PP-l) to maintain property **P2**. These PP-δ, PP-k, and PP-l steps are intended to preserve privacy in terms of the δ-presence [12], k-anonymity [10], and l-diversity models [11], respectively. More details are presented in Section V-C and Section II.

For each attribute j, we repeat the above steps **S1**–**S3** (and the privacy preservation) to generate the *attribute signature* denoted by $\mathbf{as}_{ij}$. When silo i completes the generation of attribute signatures, silo i finally returns its *silo signature* S_i, which is an aggregate of the attribute signatures.

Definition 5. *In silo i, for each non-confidential attribute j, its* attribute signature $\mathbf{as}_{ij}$ *is a tuple of two Chebyshev subsets of the attribute domains of positive and negative records:*

$$\mathbf{as}_{ij} = (C_{ij}^+, C_{ij}^-). \tag{1}$$

The silo signature *of silo i is then a set of the (non-confidential) attribute signatures in the silo and is denoted by*

$$S_i = \{\mathbf{as}_{ij} |\ \textit{non-confidential attribute } j \in A_i\}. \tag{2}$$

B. Theory of Silo Signature Creation

We assume that a data silo cannot publish its data table in its entirety. For collaborative analytics involving multiple silos, there should be a way to extract the (partial) information from each silo without violating the silo properties **P1**–**P3** and considering the possibility of **P4** (see Fig. 7 regarding **P4**-related results). In response, various approaches have been

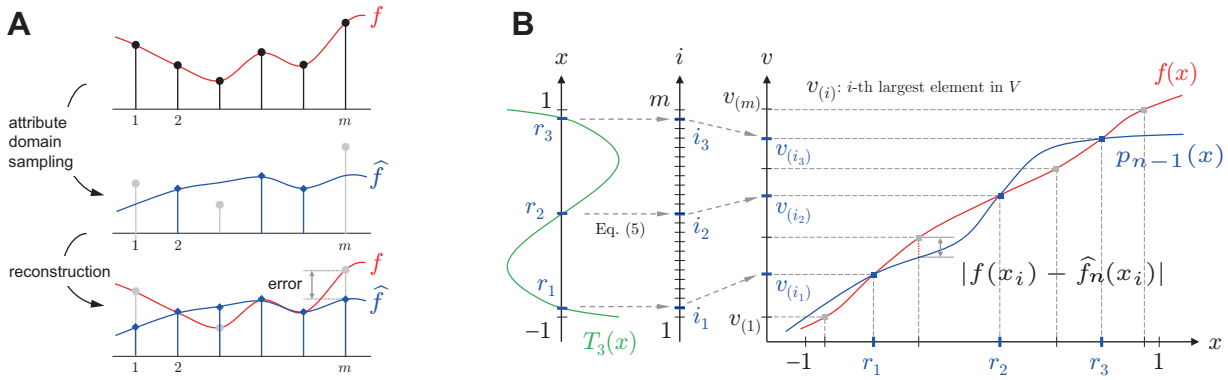

Fig. 2: (a) Concept of the ADSR problem. (b) Finding the degree-n Chebyshev subset $C = \{v_{(i_1)}, \ldots, v_{(i_n)}\}$ of V.

developed in the disciplines of inferential statistics (to understand a population) and descriptive statistics (to summarize a sample from the population).

In the present context, column-wise sampling of silo tables possesses desirable properties. Taking a small number of samples from a large collection reduces computational cost and also helps preserving privacy. Column-wise processing (*i.e.*, sampling from attribute values, not records) provides an additional layer of privacy protection, since no individual's record is included in the sample as a whole.

One important remaining question involves which one to include in the sample. A proper selection of samples will be critical for the quality of analysis. To answer this question, we consider the problem of *attribute domain sampling and reconstruction* (ADSR), as informally shown in Fig. 2(a).

1) Attribute domain sampling and reconstruction: We can regard the attribute values in the silo table as observations from a certain unknown distribution. The silo owner will provide a sample of these observations outside. The outside data miner will want to reconstruct the original distribution as faithfully as possible using only the samples provided by the silo owner. More formally, we define the ADSR problem as follows:

Definition 6. *Let multiset V denote the domain of an attribute in an m-record table. Let V represent a collection of m samples from an unknown function f. Given a positive integer $n < m$, the ADSR problem is to find an n-element multiset $V' \subset V$ such that $\widehat{f}_n$ approximates f optimally w.r.t. a certain criterion, where $\widehat{f}_n$ represents a function derived from V'.*

In CABEL, we address a specific instance of the ADSR problem named the *minimax* ADSR:

Definition 7. *For multiset $V = \{v_1, v_2, \ldots, v_m\}$, assume that $f(x_i) = v_i$ for $i = 1, 2, \ldots, m$. Given a positive integer $n < m$, the minimax ADSR problem is to find an n-element multiset $V' \subset V$ such that the maximum reconstruction error between f and $\widehat{f}_n$ defined below is minimized:*

$$\max_{1 \le i \le m} \left| f(x_i) - \widehat{f}_n(x_i) \right| \tag{3}$$

2) Proposed solution: To address the minimax ADSR problem, we propose a new scheme for sampling the attribute domain, which is named the *Chebyshev subset* (see Fig. 2(b)).

Definition 8. *For multiset $V = \{v_1, v_2, \ldots, v_m\}$, let $v_{(i)}$ denote the i-th largest element in V. Given a positive integer $n < m$, the degree-n* Chebyshev subset *of V is the n-element multiset denoted by*

$$C = \{v_{(i_1)}, v_{(i_2)}, \ldots, v_{(i_n)}\} \tag{4}$$

where

$$i_q = \left\lfloor \frac{(m-1)r_q + (1+m)}{2} + \frac{1}{2} \right\rfloor \tag{5}$$

with $r_q \in [-1, 1]$ being the q-th root of $T_n(x)$, the Chebyshev polynomials of the first kind with degree n [45]:

$$r_q = \cos\left(\frac{2q-1}{2n}\pi\right) \tag{6}$$

on the interval $x \in [-1, 1]$ for $q = 1, 2, \ldots, n$.

The use of the Chebyshev subsets in CABEL is justified by the following theorem:

Theorem 1. *The degree-n Chebyshev subset of V asymptotically solves the minimax ADSR problem in that the maximum error (3) converges to zero as $n \to \infty$.*

Proof. Assume that the attribute domain V is sorted and represent by a vector $\mathbf{v} = (v_1, v_2, \ldots, v_m)$, where $v_i \le v_{i+1}$ for $i = 1, 2, \ldots, m-1$. Note that using $\mathbf{v}$ in lieu of V does not affect the validity of proof. Also assume without loss of generality that the function $f(x)$ underlying V is defined on the interval $x \in [-1, +1]$, and $f(x_i) = v_i$ for $i = 1, 2, \ldots, m$.

Let $p_{n-1}(x)$ denote a polynomial of degree at most $(n-1)$ that interpolates $f(x)$ at n distinct points in the interval. Then, by the interpolation theory, for each x in the interval there exists $\xi \in [-1, 1]$ such that

$$f(x) - \widehat{f}_n(x) = \frac{f^{(n)}(\xi)}{n!} \prod_{q=1}^{n} (x - r_q) \tag{7}$$

where $\widehat{f}_n(x) \triangleq p_{n-1}(x)$, and the function f is by construction n times continuously differentiable on the interval $[-1, 1]$. To solve the minimax ADSR problem, we thus minimize

$$\max_{x \in [-1,1]} \left| \prod_{q=1}^{n} (x - r_q) \right| \tag{8}$$

since the other terms in (7) are independent of the locations of interpolation points.

The product in (8) is a monic polynomial of degree n, and it is known that the maximum absolute value of any such polynomial is bounded below by 2^{1-n} and that the bound is attained by $(2^{1-n})T_n$, the monic Chebyshev polynomials of the first kind with degree n.

Since the interpolation nodes x_i's are the roots of the T_n, the interpolation error satisfies

$$\max_{x\in[-1,1]}\left|f(x)-\widehat{f}_n(x)\right| \leq \frac{1}{2^{n-1}n!}\max_{\xi\in[-1,1]}\left|f^{(n)}(\xi)\right| \quad (9)$$

and therefore $\lim_{n\to\infty}\max_{1\leq i\leq m}\left|f(x_i)-\widehat{f}_n(x_i)\right|=0$.

The Chebyshev nodes r_q's are defined in the interval $[-1,1]$ as (6). To construct the Chebyshev subset defined in (4), we make indices ranging between 1 and m by the affine transformation followed by rounding (5). □

C. Privacy Protection

During the silo signature creation, three types of well-known privacy models (see Section II) are incorporated.

1) PP-δ against table-linkage attack (δ-presence): To protect privacy in terms of δ-presence [12], we need to bound the probability of an individual being included in a published data set within range $\delta=(\delta_{\min},\delta_{\max})$. The PP-$\delta$ step extracts R_i^*, a random subset of R_i, from each silo i with data $D_i=(R_i,A_i)$, and then derives the signature S_i from $D_i^*=(R_i^*,A_i)$.

By definition [12], δ-presence then holds for D_i^* as long as we keep the ratio $|R_i^*|/|R_i|$ in the range defined by δ, *i.e.*, $\delta_{\min}\leq|R_i^*|/|R_i|\leq\delta_{\max}$. CABEL further performs column-wise sampling of D_i^* to create the silo signature, which thus possesses a higher degree of δ-presence than D_i^* has.

The parameters $\delta_{\min}$ and $\delta_{\max}$ define the level of trade-offs between the utility and privacy of the anonymized data and are determined depending on the privacy conditions of the application and prior beliefs [12].

2) PP-l against attribute-linkage attack (l-diversity): To preserve privacy in terms of l-diversity [11], we need to ensure that every equivalence class in a table has at least l distinct values for any sensitive attribute. For example, for medical silos, a combination of quasi-identifiers should not be linked to whether or not the record owner has a certain disease.

For a specific disease, we assume that the positive class represents the individuals diagnosed with that disease and the negative class represents the others. Then, from the perspective of the positive class, l-diversity holds ($l\geq 2$) in silo i if

$$\left[\min(C_{ij}^+)\geq\min(C_{ij}^-)\right]\wedge\left[\max(C_{ij}^+)\leq\max(C_{ij}^-)\right] \quad (10)$$

for any attribute j, where C_{ij}^+ and C_{ij}^- are the components of the attribute signature defined in (1). This is because fulfilling Eq. (10) makes it impossible to distinguish the positive and negative classes simply by examining C_{ij}^+ and C_{ij}^-.

The PP-l step modifies C_{ij}^+ in such a way that for each $v\in C_{ij}^+$,

$$v=\begin{cases}\min(C_{ij}^-) & \text{if } v<\min(C_{ij}^-)\\ \max(C_{ij}^-) & \text{if } v>\max(C_{ij}^-)\end{cases} \quad (11)$$

CABEL aggregates attribute signatures into a single silo signature, providing a stronger notion of privacy than l-diversity.

3) PP-k against record-linkage attack (k-anonymity): To ensure k-anonymity [10], a published record should be indistinguishable from at least $k-1$ other records. Utilizing the Chebyshev subset C of an attribute domain V can be considered as a generalization process, in which V is partitioned into n intervals, and individual values of attributes are replaced with a value representing one of the intervals. That is, creating silo signatures inherently ensures privacy protection in terms of k-anonymity.

In CABEL, a silo does not release its table directly but returns its silo signature. However, for the sake of explanation in terms of k-anonymity, suppose that silo i releases its data $D_i=(R_i^+\cup R_i^-,A_i)$ after transforming it by replacing the individual values of attribute j with one of the n values in the degree-n Chebyshev subsets C_{ij}^+ or C_{ij}^-. Let us denote this transformed data D_i^*. It is then evident that D_i^* possesses k-anonymity with average $k=\min\left\{\left\lfloor|R_i^+|/n\right\rfloor,\left\lfloor|R_i^-|/n\right\rfloor\right\}$.

In reality, CABEL utilizes silo signatures, each of which consists of tuples of n-element Chebyshev subsets, where normally $n\ll|R_i|$. This implies large k, thus ensuring a strong level of privacy protection in terms of k-anonymity.

D. Ensemble Learning on Silo Signatures

For silo-based collaborative analytics, there exist challenges (denoted as **C1**–**C4** below) that must be addressed.

C1: *Different silos store different tables but may share common attributes. For the same attribute, there can thus be multiple (normally different) signatures from different silos.*

This challenge is related to the *covariate shift* [46] problem. When combining multiple signatures for a common attribute, CABEL thus considers the potential shifts among the distributions underlying the signatures, as detailed in Section V-D3.

If there are a sufficient number of different signatures for each attribute, then we could naturally use bagging because different signatures imply different training samples. However, in reality, we face the following challenge:

C2: *The number of data silos is often limited, and the number of different signatures for an attribute is typically insufficient for training a learner.*

In response to **C2**, we consider all the attribute signatures from the participating silos as a data pool. This gives the effect of increasing the training samples, but there still remains a challenge if we use this data pool as it is, as follows:

C3: *Different attributes have different types (e.g., quantitative vs. qualitative) and ranges (e.g.,* `age` *in* $[0,120]$ *vs.* `glucose(mmmol/L)` *in* $[3,8]$*), which are thus difficult to mix for learning.*

To resolve **C3**, we derive a classifier $\mathcal{M}_j$ from each attribute signature j and then represent the signature not by the raw values but by the decision $d_j\in\{-1,+1\}$ from the classifier.

It is then logical to employ the idea of ensemble learning to combine the outcomes of these 'base' classifiers to make an integrative decision. However, there comes a new challenge with regard to devising a scheme for combining these base classifiers.

C4: *Each base classifier $\mathcal{M}_j$ makes its decision based on a single attribute j (i.e., a 'scalar' classifier), but a new object* $\mathbf{x}$

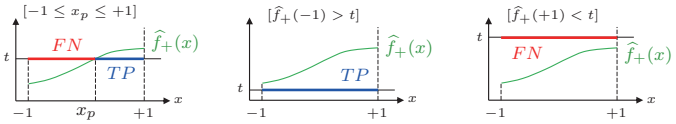

Fig. 3: Definitions of $TP(t)$ and $FN(t)$.

for testing the combined learner consists of multiple attributes and needs a 'vector' classifier.

A simple solution to **C4** would be to let each $\mathcal{M}_j$ classify $\mathbf{x}$ with respect to attribute j only and then to combine the result by an ensemble-learning technique. However, this solution would revive the challenge **C2**.

To address all of these challenges, we propose a new ensemble-learning method termed the *weighted bagging of base classifiers* (WBBC).

1) Generating base classifiers from signatures: Recall that $\mathbf{as}_{ij} = (C_{ij}^+, C_{ij}^-)$ denotes the signature of attribute j in silo i. Using the signatures, we train a weak binary-classifier $\mathcal{M}_{ij}$ that returns decision $d_{ij} \in \mathcal{Y} = \{-1, +1\}$. To reduce overfitting, we use a simple model as the base classifier, as is usually the case in ensemble learning. The base classifier $\mathcal{M}_{ij}$ is thus defined as

$$d_{ij} = \mathcal{M}_{ij}(x) = \text{sign}(x - t_{ij}) \tag{12}$$

where x is an unseen object, t_{ij} is a threshold, and $\text{sign}(\cdot)$ is the operator that returns ± 1 according to the sign of its argument. We assume that a value in the attribute domain of j is a scalar, and this also holds for x.

Training $\mathcal{M}_{ij}$ then corresponds to determining the value of t_{ij}. To this end, the reconstruction concept stated in Section V-B comes into play. Specifically, we first reconstruct two functions f^+ and f^- that underlie the Chebyshev subsets C_{ij}^+ and C_{ij}^-, by using two functions $\widehat{f}_+$ and $\widehat{f}_-$, respectively. Recall that such estimates $\widehat{f}_+$ and $\widehat{f}_-$ are polynomials defined in the interval $[-1, 1]$. We then estimate the accuracy[1] of the classification based on $\widehat{f}_+$ and $\widehat{f}_-$. We finally set t_{ij} to the value that maximizes this estimated accuracy α_{ij}:

$$t_{ij}^* = \underset{t}{\text{argmax}} \{\alpha_{ij}(t)\} \tag{13}$$

$$= \underset{t}{\text{argmax}} \left\{ \frac{TP(t) + TN(t)}{TP(t) + FN(t) + TN(t) + FP(t)} \right\} \tag{14}$$

$$= \underset{t}{\text{argmax}} \{TP(t) + TN(t)\} \tag{15}$$

where the terms TP/FN/TN/FP depend on t and are defined by the lengths of segments in the interval $[-1, 1]$.

Fig. 3 depicts the definitions of $TP(t)$ and $FN(t)$:

$$TP(t) = \begin{cases} 1 - x_p & \text{if } -1 \le x_p \le +1 \\ 2 & \text{if } x_p < -1 \text{ or } \widehat{f}_+(-1) > t \\ 0 & \text{if } x_p > +1 \text{ or } \widehat{f}_+(+1) < t \end{cases} \tag{16}$$

$$FN(t) = 1 - (-1) - TP(t) = 2 - TP(t) \tag{17}$$

[1]Defined as $(TP + TN)/(TP + TN + FP + FN)$, where $TP(TN)$ means the number of true positives (negatives), while $FP(FN)$ denotes that of false positives (negatives).

where x_p denotes the point at which $\widehat{f}_+(x) = t$. In the same manner, we can define $TN(t)$ and $FP(t)$.

Note that the above explanation is for building a classifier from the signature of a numerical attribute. We can build a classifier for a categorical attribute in a similar way.

2) Handling the covariate shift problem: We consider the possible differences between the distributions from which different signatures of a common attribute were created. We utilize the conditional probability averaging (CPA) [46], where the binary classification of object x is performed by averaging the probabilities of x being a positive or negative object over different distributions from which x can originate and then assigning x to the class that has the larger average probability.

Let x_j be the value of attribute j of a record. The probability of $\mathbf{x}$ belonging to the positive class is estimated as follows.

$$Pr\{+1|x_j\} = \frac{1}{N_j} \sum_{i=1}^{N_j} Pr\{\mathcal{M}_{ij}(x_j) = +1\} \tag{18}$$

$$\approx \frac{1}{N_j} \sum_{i=1}^{N_j} \alpha_{ij}(t_{ij}^*) \tag{19}$$

where N_j is the number of silos whose table contains attribute j, and $\alpha_{ij}(t_{ij}^*)$ is the accuracy of $\mathcal{M}_{ij}$ with its optimal threshold t_{ij}^*, as defined in Section V-D1.

Only for the attribute j that exists in multiple silos and thus has multiple signatures, we update the base classifier $\mathcal{M}_{ij}$:

$$\mathcal{M}_{ij} = \begin{cases} +1 & \text{if } Pr\{+1|x_j\} > Pr\{-1|x_j\} \\ -1 & \text{otherwise} \end{cases} \tag{20}$$

where $Pr\{-1|x_j\} = 1 - Pr\{+1|x_j\}$.

3) Weighted bagging of base classifiers (WBBC):

Training Phase:

1) Derive the first-level base classifier $\mathcal{M}_{ij}$ from each $\mathbf{as}_{ij}$ as explained in Section V-D1. Let N denote the total number of base classifiers constructed.
2) Calculate the probability weight w_{ij} of $\mathcal{M}_{ij}$ as follows:
$$w_{ij} = \frac{\alpha_{ij}(t_{ij}^*)}{\sum_{i,j} \alpha_{ij}(t_{ij}^*)} \tag{21}$$
3) Generate B bootstrap samples of N classifiers with probability weights w_{ij}. In other words, there are B bags of classifiers, and each bag contains N base classifiers. The probability of the base classifier $\mathcal{M}_{ij}$ being sampled is proportional to w_{ij}.

Testing Phase: $\mathbf{x} = (x_1, \ldots, x_j, \ldots)$ denotes a new object.

1) Derive a second-level base classifier $\mathcal{B}_b$ from bag b of the first-level classifiers, for $b = 1, \ldots, B$. To this end, let each $\mathcal{M}_{ij}$ in the bag b classify $\mathbf{x}$ w.r.t. x_j and combine the result by voting:
$$\mathcal{B}_b(\mathbf{x}) = \text{sign}\left(\sum_{\forall \mathcal{M}_{i,j} \in b} \mathcal{M}_{ij}(\mathbf{x}) \right) \tag{22}$$
where $\mathcal{M}_{ij}(\mathbf{x}) \triangleq \mathcal{M}_{ij}(x_j)$.
2) Determine $d_{\mathbf{x}}$, the class of $\mathbf{x}$, by combining the decisions from all second-level base classifiers by voting:

$$d_{\mathbf{x}} = \mathcal{H}(\mathbf{x}) = \text{sign}\left(\sum_{b=1}^{B} \mathcal{B}_b(\mathbf{x})\right) \tag{23}$$

where $\mathcal{H}$ denotes the final classifier WBBC returns.

E. Remarks

The worst-case time complexity to generate signatures is determined by the sorting operation and is thus $O(m \log m)$ per attribute, where m is the number of records. The ensemble-learning step takes $O(nN)+O(N^2)+O(B)$, where n, N, and B represent the signature size, the number of base classifiers, and the number of classifier bags, respectively. Even for large data, their values remain small, and the time demand of CABEL is often negligible (*e.g.*, analyzing the four billion record insurance data took less than 10 minutes with CABEL).

VI. Experimental Results

A. Experiment Setup

Under a research contract with the national health-insurance agency of an OECD country, we obtained three large-scale data tables that contain a total of 4,182,000,000 records collected from the *entire population* of the country in 2012. The three tables acquired contain outpatient prescriptions (OP: 1,352,200,000 records), clinical statements (CS: 2,114,800,000 records), and medical expense statements (MES: 715,000,000 records), respectively. Table I provides more details.

Each of the records stored in these tables includes various attributes and binary classification labels (*i.e.*, diabetes or not) and is thus ideal for supervised classification studies. In our experiments, we focused on three common types of diseases (tuberculosis, diabetes mellitus, and pneumonia).

To process over four-billion records, we utilized a grid that consists of 64 data nodes in the Hadoop environment. Each node has an 8-core Intel i7 processor with a 1TB HDD and 16GB memory. Signature-based ensemble learning was carried out in the Matlab environment. For comparison with CABEL, we also prepared the implementations of a deep belief network (DBN) [47], logistic regression [48], naïve Bayes [48], AdaBoost [49], and random forests [50], [51].[2] Unless otherwise stated, we used 10-fold cross validation for the training of the classifiers used.

1) Creating silo-based analytics scenarios: Note that each of the three tables (OP, CS, and MES) was available in its entirety. Nonetheless, for the sake of assessing the performance of CABEL, we simulated a three-silo scenario by assuming that each of these three tables was separately stored in a silo that abides by the four silo properties **P1**–**P4** stated in Section IV. To test CABEL, we thus used only the silo signatures derived from these tables instead of the whole data.

B. Proof-of-Concept Experiments

Fig. 4(a) shows how the approximation error (measured in terms of the normalized mean squared error (NMSE) averaged over the attributes in a silo) varies for different values of n,

[2]We used Matlab Statistics and Machine Learning Toolbox™ for AdaBoost and naïve Bayes and the Matlab Deep Learning Toolbox for DBN.

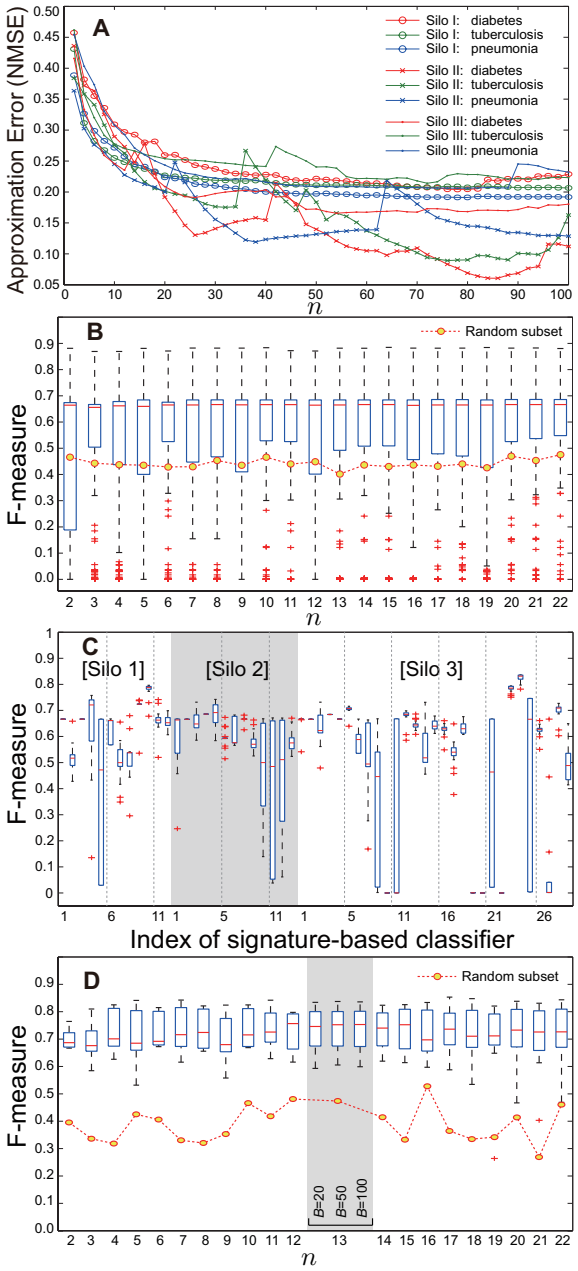

Fig. 4: Proof-of-concept experiments. (a) Approximation error for degree-n Chebyshev subsets. (b and c) F-measure distribution of base classifiers for different n. (d) Classification performance of CABEL for different values of n and B.

the degree of Chebyshev subsets, or equivalently, the length of the attribute signature elements C_{ij}^{+} and C_{ij}^{-} (Def. 4). Recall from Section V-D1 that the Chebyshev subsets C_{ij}^{+} and C_{ij}^{-} are used to construct the estimates $\widehat{f}_{+}$ and $\widehat{f}_{-}$, respectively. For each n, shown in the plot is the averaged NMSE between $\widehat{f}_{+}$ and f_{+} (interpolated from all of the values of the attribute domain) for each of the diseases (diabetes, tuberculosis, and pneumonia) for each silo (nine curves in total).

The results shown in Fig. 4(a) empirically prove our proposal in two aspects. First, the approximation error decreases

TABLE I: Three health-insurance data sets collected from the entire population of an OECD country in 2012.

Silo	Contents	Size	# records	# attr.	# diabetes [freq. (%)]	# tuberculosis [freq. (%)]	# pneumonia [freq. (%)]
I	OP	190 GB	1,352,200,000	12	48,982,245 [3.62]	1,265,284 [0.09]	14,993,422 [1.11]
II	CS	261 GB	2,114,800,000	13	53,515,800 [2.53]	3,030,613 [0.14]	10,951,364 [0.52]
III	MES	165 GB	715,000,000	29	14,643,615 [2.04]	412,671 [0.06]	3,515,413 [0.49]
Total		616 GB	4,182,000,000	54	117,141,660 [2.80]	4,708,568 [0.11]	29,460,199 [0.70]

Abbreviations: OP, outpatient prescription; CS, clinical statement; MES, medical expense statement.

as we increase the value of n from 2 to 100. Second, the approximation error is kept moderately low (*i.e.*, below 0.5 but not extremely low), implying that the base classifier derived from attribute signatures (denoted by $\mathcal{M}$'s in Section V-D) will be indeed "weak" (*i.e.*, slightly better than random guessing), which is desirable for reducing overfitting in ensemble learning. Simple models are better to avoid overfitting but often suffer from high bias. The generalization of a machine learner is affected by both bias and variance. Our use of ensemble learning is justified by reducing bias by combining multiple tables from silos and analyzing them integratively.

To verify the second aspect further, we evaluated the performance of signature-based classifiers with different values of n. Fig. 4(b) shows the F-measure distribution of all 54 base classifiers for different n values, while Fig. 4(c) shows the F-measure distribution of individual base classifiers for different n values. For comparison, the F-measure distribution of a random subset is also shown for each n. As expected, base classifiers performed slightly better than random guessing for all $n > 2$. The variability of the performance and the existence of outliers justify our bagging-based combination approach, since bagging is useful when base classifiers are unstable [52].

Lastly, the F-measure performance of CABEL is shown in Fig. 4(d) for different n values. For the entire range of n values tested, CABEL maintained its performance, demonstrating the effect of ensemble learning. That is, unlike the result shown in Fig. 4(a), where the choice of n affected the quality of approximation by the Chebyshev subset, the value of n did not make a significant difference in the integrative classification result. Nonetheless, we observed that n in a certain range produced the highest median F-measure; thus, we set $n = 13$ as the default value in our experiment. For $n = 13$, we tested three different values of B (the number of bootstrap samples; see Section V-D3) and observed little difference in performance. We used $B = 50$ as the default value.

C. Classification Performance Comparison

As shown in Fig. 5, we measured the classification performance of CABEL in terms of recall, precision, and F-measure. For comparison, we also tested DBN, logistic regression, naïve Bayes, AdaBoost, random forests, and random classification (*i.e.*, coin tossing). For CABEL, in addition to WBBC, we implemented two alternatives: a voting scheme (all first-level base classifiers vote for a final decision without proceeding to the second level) and the unweighted version of WBBC.

To train CABEL, only $2n = 26$ points (13 for positive and 13 for negative examples) per attribute were used (only 10^{-3}–10^{-5}% of the data). In contrast, we used 250,000 of the total points per attribute for training the other classifiers.

To extract data to train CABEL, no separate anonymization was used because using silo signatures has the effect of anonymization. According to the three types of privacy-preserving mechanisms CABEL provides (*i.e.*, PP-k, PP-l, and PP-δ described in Section V-C), the anonymization level CABEL effected in this specific experiment was $31,744 \leq k \leq 4,116,600$ for k-anonymity, $l = 2$ for l-diversity, and $10^{-5} \leq \delta \leq 10^{-7}$ for δ-presence. Based on these numbers, to train the other alternatives, we applied three types of anonymization to the sampled data using a public tool [53]: δ-presence ($\delta = 10^{-4}$), l-diversity ($l = 2$), and k-anonymity ($k = 2$, 100, and 1000).

Even though a tiny fraction of samples was used for training, CABEL outperformed the alternatives in terms of the median F-measure regardless of k. Specifically, the median F-measure of CABEL was 14%, 20–43%, 17–43%, 17–43%, 17–31%, and 53% higher than those of DBN, logistic regression, naïve Bayes, AdaBoost, random forests, and random classification (RC), respectively. Furthermore, the alternatives to CABEL suffered from performance degradation when k was increased to ensure higher levels of anonymization. Note that using $k > 1000$ (in order to match the anonymization level of CABEL) caused the competing methods to show unacceptably low performance levels (results not shown).

The average runtimes for training CABEL, naïve Bayes, and RC were only a few minutes, whereas the other alternatives took significantly more time (on the order of hours).

D. Effect of Collaborative Analytics

To test our assumption that collaborative analytics involving multiple silos can lead to improved learning performance, we evaluated the performance of CABEL by using different numbers and combinations of silos for training CABEL.

Fig. 6(a) shows the receiver operating characteristic (ROC) curves of CABEL trained by one, two, and three silos (for the one- or two-silo case, the average values over three alternative combinations are shown). The combination scheme used in training CABEL was voting, as explained in Section VI-C. Fig. 6(b) and (c) shows the ROC curves when the unweighted bagging of base classifiers and WBBC were used, respectively.

In all the cases tested, using more silos resulted in higher performance in area-under-curve (AUC) values. The AUC value was nearly 8% points higher for the three-silo training as compared to the one-silo case, as shown in Fig. 6(c).

Fig. 6(d) further compares the distribution of F-measure for the one-, two-, and three-silo cases. The combining scheme used was the weighted bagging of base classifiers. This figure illustrates the effect of bagging, which is known to reduce the variance of base classifiers [52]. With multiple silos, the

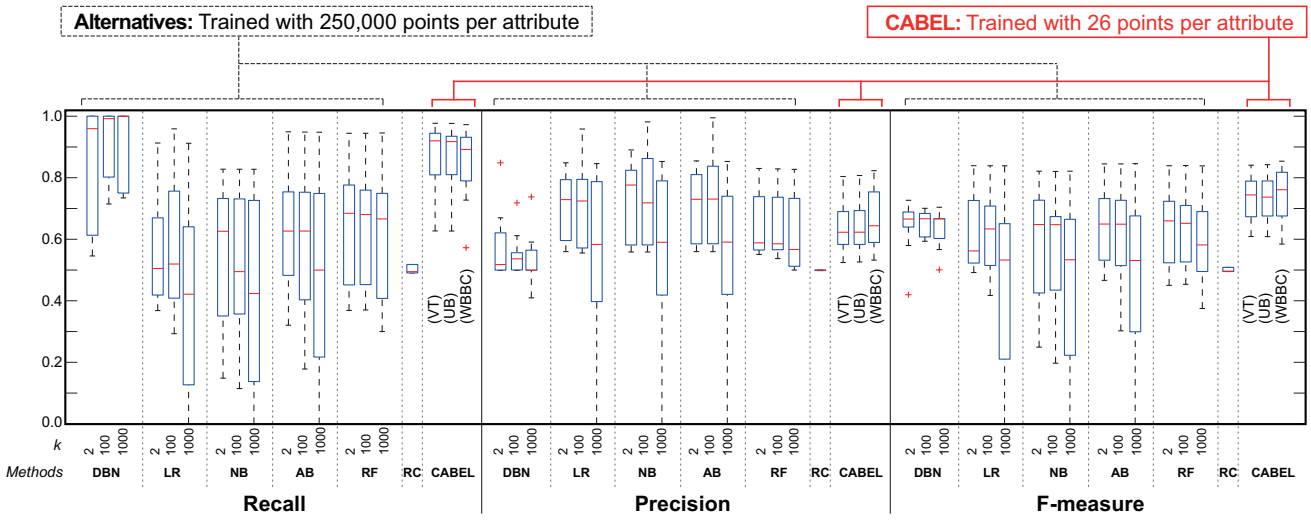

Fig. 5: Performance comparison: CABEL with WBBC option is the proposed method. Abbreviations: DBN, deep belief network; LR, logistic regression; NB, naïve Bayes; AB, AdaBoost; RF, random forest; RC, random classification; VT, voting; UB, unweighted bagging of base classifiers; WBBC, weighted bagging of base classifiers. Parameters for alternatives: $k = 2$, 100, and 1000 for k-anonymity; $l = 2$ for l-diversity; $\delta = 10^{-4}$ for δ-presence. Parameters for CABEL: $n = 13$; $B = 50$.

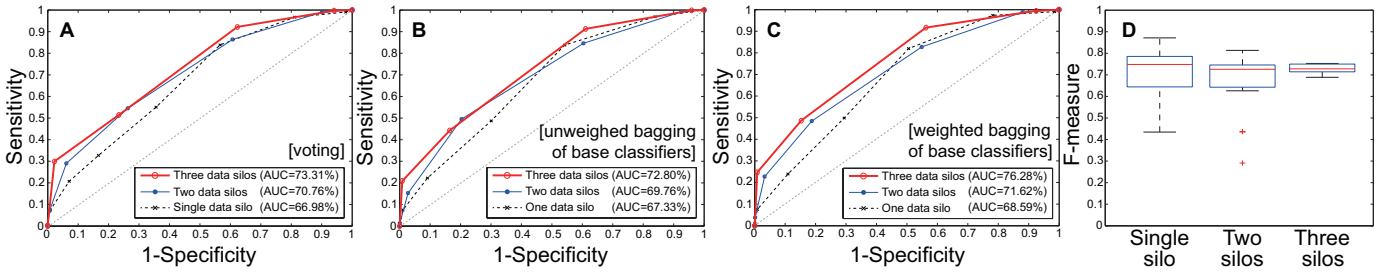

Fig. 6: (a-c) Multi-silo analytics improves ROC performance. (d) Multi-silo analytics enhances F-measure values.

distribution of F-measure became noticeably narrower and the median F-measure level improved slightly.

E. Effect of Varying Trustfulness Property: Label Noise

Recall from Section IV that some data silos may provide inaccurate information either by mistake or deliberately (the silo property **P4**). A robust data-mining model for collaborative analytics would yield consistent performance, even when some data silos are distrustful. To assess the robustness of CABEL, we measured the classification performance with varying degrees of trustfulness of data silos.

To simulate a distrustful data silo, we define the degree of distrustfulness as the fraction of intentionally mislabeled samples in the silo signature. For instance, 50% distrustfulness corresponds to a scenario in which half of the positive and negative samples are mislabeled.

Fig. 7 shows the F-measure of CABEL measured when we vary the degree of distrustfulness from 0% to 100%. We can observe that the F-measure remains reasonably high until the degree of distrustfulness exceeds approximately 50%. This result suggests that the proposed technique for collaborative analytics will work even when some of the data silos fail to provide perfectly accurate information.

TABLE II: Median F-measure of CABEL versus noise

Noise (%)	0	5	10	15	20	25	30	35	40	45
Silo I	.81	.80	.78	.75	.73	.71	.69	.68	.66	.65
Silo II	.69	.69	.68	.67	.67	.67	.67	.67	.67	.67
Silo III	.77	.75	.76	.75	.73	.71	.67	.67	.65	.65
Average	.75	.75	.74	.72	.71	.69	.68	.67	.66	.66

F. Robustness against Noisy Signatures: Attribute Noise

The proposed signature concept is based on the sorted interpolation by the Chebyshev polynomials, which are known to be robust in the presence of outliers, white noise, and impulsive noise, as rank-order statistics [54]. It may still suffer from aliasing (*i.e.*, the loss of high-frequency components) when degree n is too small [55]. As shown in Fig. 4(a), using smaller n values thus tended to increase the error at the individual signature level. However, the ensemble learning-based combination of base classifiers made the F-measure of CABEL practically independent of n, as shown in Fig. 4(d).

We performed additional experiments to assess the robustness of CABEL against signature-level noise. Table II lists the F-measure variation of CABEL as different levels of noise are added to the attribute signatures. The noise level is defined as

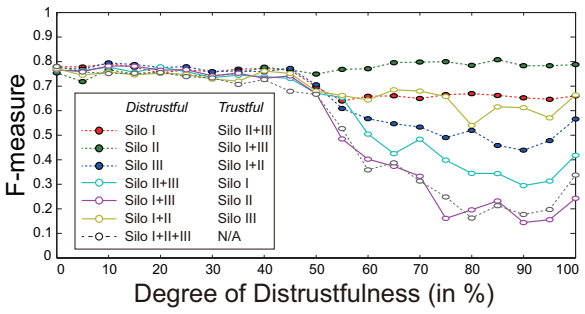

Fig. 7: Effect of distrustfulness on CABEL's performance.

TABLE III: Effects of covariate shift consideration on AUC

Group 1	Group 2	no consideration		covariate shift considered		
		Grp 1 only	Grp 2 only	CPA	JPA	NRC
Top-tier	Others	.75	.82	.82	.81	.81
Third-tier	Others	.78	.80	.81	.79	.81
Second-tier	Others	.70	.81	.79	.79	.81
First-tier	Others	.77	.80	.81	.81	.82
Male	Female	.81	.82	.81	.79	.80
Average		.76	.81	.81	.80	.81

Abbreviations: CPA, conditional probability averaging; JPA, joint probability averaging; NRC, normalizing ranges of Chebyshev subsets. [data: Silo III]

the fraction of deliberately swapped attribute values (between positive and negative classes) before signature generation. With up to 20% noise, CABEL maintained median F-measure over 0.7, and even with 45% noise, the F-measure was greater than 0.66, showing the robustness of the proposed method against noisy signatures.

G. Effective Handling of Covariate Shifts by CABEL

As explained in Section V-D2, CABEL examines multiple signatures of the same attribute from different silos and adjusts the corresponding base classifiers by CPA. This is to alleviate the covariate shift problem (*e.g.*, the same attribute may have different distributions for top-tier and second-tier hospitals).

To observe the effectiveness of this adjustment, we compared the area-under-curve (AUC) values obtained by CABEL with and without considering the covariate-shift problem, as listed in Table III. In the experiments, we divided the Silo III (MES) data into various pairs of groups and ran CABEL on each group with and without considering covariate shifts for binary classification. For comparison, we tested two alternatives to CPA: the joint probability averaging (JPA) method [46] and the method of normalizing the ranges of Chebyshev subsets (NRC) for different signatures.

As expected, considering covariate shifts was effective for improving the performance. CPA and NRC slightly outperformed JPA. Unlike CPA, however, both JPA and NRC require additional information from each silo (*e.g.*, the number of records). Therefore, we decided to use CPA in CABEL.

VII. Discussion

The ADSR problem bears some resemblance to downsampling/upsampling and reconstruction in signal processing. The process of creating signatures corresponds to downsampling, while estimating the underlying function from signatures is linked to upsampling and reconstruction. Nonetheless, ADSR requires variable inter-sample distances, which makes the application of signal processing techniques challenging. The recent advances in compressed sensing may open up new possibilities for ADSR and for creating silo signatures as an alternative to the Chebyshev subsets used in CABEL.

The major difference between CABEL and conventional ensemble learning comes from the fact that CABEL needs to sample base classifiers, not data points. This makes any theoretical analysis of the proposed weighted bagging of base classifiers challenging. In terms of ensemble-learning, it would be intriguing to devise a boosting-based combination approach in CABEL. To this end, we need to sample base classifiers sequentially, penalizing poorly performing learners. However, given the fact that the base learners derived from attribute signatures could be unstable, as presented in Section VI-B, bagging-based methods may still be better than boosting, which is known to be susceptible to outliers and noise.

Along with devising a boosting-based scheme for CABEL, it would be worth trying a decision tree as the base classifier. In the literature, AdaBoost combined with decision trees as the weak learners is often referred to as the best out-of-the-box classifier [51]. Alternatively, we could make each base classifier return not only the sign but also the magnitude (*i.e.*, the margin from the threshold) of its decision such that the confidence in classification is represented by the margin.

We may consider using different types/sizes of signatures for different silos and devise a fine-grained combination scheme. However, it remains to be seen whether such optimizations would lead to performance gain in general. For the health-insurance data, we tried different values of n for different attributes, only to see negligible performance boosts.

Based on the property **P4** (trustfulness) in Section IV, CABEL considers the participating silos as malicious parties [7] rather than semi-honest parties. We tested and presented in Section VI-E the robustness of CABEL with the existence of distrustful silos. To proceed one step further, it would be interesting to incorporate into CABEL the interplay between related and competing data silos, for example, from a game-theoretic perspective [56].

Additional future work includes the following: First, CABEL currently focuses on binary classification, but future versions will be extended to multiclass classification and regression by devising base classifiers capable of such tasks. Second, CABEL currently does not protect the information pertaining to which types of attributes are stored in a silo, but it is possible to hide such information by utilizing an SMC-based protocol [57], [9]. Last, by storing the meta-data (*i.e.*, sorted attribute data) from each silo signature, we can update silo signatures rapidly, which suggests that we may formulate CABEL as an online learning problem effectively to reflect the updated or new data from silos into collaborative analytics.

VIII. Conclusion

We have proposed a new method called CABEL for collaborative analytics of data silos. For the effective sampling of silo data while preserving privacy and computational efficiency, we introduced the *silo signatures*. To gain theoretical insights into

the signatures, we formulated the ADSR problem and then proposed a solution called the *Chebyshev subset* that led to effective silo signatures. Combining the information obtained from silo signatures was carried out through a new ensemble-learning method called WBBC. We tested CABEL with the large-scale real data that contained over four billion records and confirmed the effectiveness of CABEL. To the best of our knowledge, this work is the first attempt to apply collaborative analytics to nationwide health-insurance data. We anticipate that CABEL plays a key role in addressing the silo issue, which is becoming more important in today's big-data era.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea grants funded by the Korean Government (Ministry of Science, ICT and Future Planning) [No. 2011-0009963], by the ICT R&D program of MSIP/ITP [14-824-09-014, Basic Software Research in Human-level Lifelong Machine Learning (ML Center)], by Samsung Scholarship, by SK Hynix, by Samsung Electronics, and by SAP Labs Korea.

REFERENCES

[1] S. LaValle *et al.*, "Big data, analytics and the path from insights to value," *MIT Sloan Management Review*, vol. 21, 2013.
[2] C. Lynch, "Big data: How do your data grow?" *Nature*, vol. 455, no. 7209, pp. 28–29, 2008.
[3] R. Jain, "Out-of-the-box data engineering events in heterogeneous data environments," in *IEEE ICDE*, 2003, pp. 8–21.
[4] B. Fung *et al.*, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys*, vol. 42, no. 4, p. 14, 2010.
[5] R. Agrawal and R. Srikant, "Privacy-preserving data mining," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 439–450, 2000.
[6] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology 2000*. Springer, 2000, pp. 36–54.
[7] C. C. Aggarwal and S. Y. Philip, *A general survey of privacy-preserving data mining models and algorithms*. Springer, 2008.
[8] B.-H. Park and H. Kargupta, "Distributed data mining: Algorithms, systems, and applications," 2002.
[9] C. Clifton *et al.*, "Tools for privacy preserving distributed data mining," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 28–34, 2002.
[10] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.
[11] A. Machanavajjhala *et al.*, "l-diversity: Privacy beyond k-anonymity," *ACM TKDD*, vol. 1, no. 1, p. 3, 2007.
[12] M. E. Nergiz *et al.*, "Hiding the presence of individuals from shared databases," in *ACM SIGMOD*. ACM, 2007, pp. 665–676.
[13] K. Saranya *et al.*, "A survey on privacy preserving data mining," in *ICECS*. IEEE, 2015, pp. 1740–1744.
[14] H. Kargupta *et al.*, "On the privacy preserving properties of random data perturbation techniques," in *ICDM*. IEEE, 2003, pp. 99–106.
[15] Z. Huang *et al.*, "Deriving private information from randomized data," in *ACM SIGMOD*. ACM, 2005, pp. 37–48.
[16] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in *IEEE ICDE*. IEEE, 2005, pp. 217–228.
[17] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *ACM SIGMOD-SIGACT-SIGART*. ACM, 2004, pp. 223–228.
[18] K. LeFevre *et al.*, "Incognito: Efficient full-domain k-anonymity," in *ACM SIGMOD*. ACM, 2005, pp. 49–60.
[19] ——, "Mondrian multidimensional k-anonymity," in *IEEE ICDE*. IEEE, 2006, pp. 25–25.
[20] F. Kohlmayer *et al.*, "Flash: efficient, stable and optimal k-anonymity," in *PASSAT and SocialCom*. IEEE, 2012, pp. 708–717.
[21] X. Xiao and Y. Tao, "Personalized privacy preservation," in *ACM SIGMOD*. ACM, 2006, pp. 229–240.
[22] D. Kifer and J. Gehrke, "Injecting utility into anonymized datasets," in *ACM SIGMOD*. ACM, 2006, pp. 217–228.
[23] R. D. Boyce *et al.*, "Bridging islands of information to establish an integrated knowledge base of drugs and health outcomes of interest," *Drug Safety*, vol. 37, no. 8, pp. 557–567, 2014.
[24] D. J. Sheskin, *Handbook of parametric and nonparametric statistical procedures*. CRC Press, 2003.
[25] W. Fan *et al.*, "A general framework for accurate and fast regression by data summarization in random decision trees," in *ACM SIGKDD*. ACM, 2006, pp. 136–146.
[26] P. Mitra *et al.*, "Density-based multiscale data condensation," *IEEE TPAMI*, vol. 24, no. 6, pp. 734–747, 2002.
[27] A. D. Shieh and D. F. Kamm, "Ensembles of one class support vector machines," in *Multiple Classifier Systems*. Springer, 2009, pp. 181–190.
[28] A. Lazarevic and V. Kumar, "Feature bagging for outlier detection," in *ACM SIGKDD*. ACM, 2005, pp. 157–166.
[29] E. Simpson *et al.*, "Dynamic bayesian combination of multiple imperfect classifiers," in *Decision Making and Imperfection*. Springer, 2013, pp. 1–35.
[30] V. S. Verykios *et al.*, "State-of-the-art in privacy preserving data mining," *ACM SIGMOD Record*, vol. 33, no. 1, pp. 50–57, 2004.
[31] J. Vaidya and C. Clifton, "Privacy preserving naıve bayes classifier for vertically partitioned data," in *SDM*. SIAM, 2004, pp. 522–526.
[32] J. Vaidya *et al.*, "Privacy-preserving naive bayes classification," *VLDB*, vol. 17, no. 4, pp. 879–898, 2008.
[33] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*. Australian Computer Society, Inc., 2002, pp. 1–8.
[34] J. Vaidya and C. Clifton, "Privacy-preserving decision trees over vertically partitioned data," in *Data and Applications Security XIX*. Springer, 2005, pp. 139–152.
[35] F. Emekçi *et al.*, "Privacy preserving decision tree learning over multiple parties," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 348–361, 2007.
[36] H. Yu *et al.*, "Privacy-preserving svm classification on vertically partitioned data," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2006, pp. 647–656.
[37] O. L. Mangasarian and E. W. Wild, "Privacy-preserving classification of horizontally partitioned data via random kernels." in *DMIN*, 2008, pp. 473–479.
[38] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE TKDE*, no. 9, pp. 1026–1037, 2004.
[39] F. Zhang *et al.*, "Privacy-preserving two-party distributed association rules mining on horizontally partitioned data," in *CloudCom-Asia*. IEEE, 2013, pp. 633–640.
[40] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *ACM SIGKDD*. ACM, 2002, pp. 639–644.
[41] ——, "Privacy-preserving k-means clustering over vertically partitioned data," in *ACM SIGKDD*. ACM, 2003, pp. 206–215.
[42] X. Lin *et al.*, "Privacy-preserving clustering with distributed em mixture modeling," *Knowledge and information systems*, vol. 8, no. 1, pp. 68–81, 2005.
[43] C. C. Aggarwal, "On k-anonymity and the curse of dimensionality," in *VLDB*. VLDB Endowment, 2005, pp. 901–909.
[44] L. Burnett *et al.*, "The "GeneTrustee": a universal identification system that ensures privacy and confidentiality for human genetic databases." *Journal of Law and Medicine*, vol. 10, no. 4, pp. 506–513, 2003.
[45] E. Kreyszig, *Advanced engineering mathematics*. John Wiley & Sons, 2010.
[46] W. Fan and I. Davidson, "On sample selection bias and its efficient correction via model averaging and unlabeled examples." in *SDM*. SIAM, 2007, pp. 320–331.
[47] G. Hinton *et al.*, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
[48] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 4, no. 4.
[49] Y. Freund *et al.*, "Experiments with a new boosting algorithm," in *ICML*, vol. 96, 1996, pp. 148–156.
[50] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.
[51] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
[52] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine learning*, vol. 36, no. 1-2, pp. 105–139, 1999.
[53] F. Prasser *et al.*, "Arx-a comprehensive tool for anonymizing biomedical data," in *AMIA Annual Symposium*, vol. 2014. American Medical Informatics Association, 2014, p. 984.
[54] P. J. Huber, *Robust statistics*. Springer, 2011.
[55] J. C. Mason and D. C. Handscomb, *Chebyshev polynomials*. CRC Press, 2002.
[56] H. Kargupta *et al.*, "Multi-party, privacy-preserving distributed data mining using a game theoretic framework," in *PKDD*. Springer, 2007, pp. 523–531.
[57] W. Du and M. J. Atallah, "Secure multi-party computation problems and their applications: a review and open problems," in *Proceedings of the 2001 workshop on New security paradigms*. ACM, 2001, pp. 13–22.