An ICN-based data marketplace model based on a game theoretic approach using quality-data discovery and profit optimization

Eunil Seo, *Member, IEEE,* Hyoungshick Kim, *Senior Member, IEEE,* Bhaskar Krishnamachari, *Senior Member, IEEE,* and Erik Elmroth, *Member, IEEE*

Abstract—In the age of data and machine learning, massive amounts of data produced throughout our society can be rapidly delivered to various applications through a broad spectrum of cloud services. However, the spectrum of applications has vastly different data quality requirements and Willingness-To-Pay(WTP), creating a general and complex problem matching consumer quality requirements and budgets with providers' data quality and price. This paper proposes the Information-Centric Networking(ICN)-based data marketplace to foster quality-data trading service to address the challenge above. We embed a WTP mechanism into an ICN-based data broker service running on cloud computing; therefore, a data consumer can request its desired data with a data name and quality requirement. By specifying nominal WTPs, data consumers can acquire data of the desired quality at the range of maximum nominal WTP. At the same time, a data broker can offer data of a suitable quality based on the profit-optimized price and the proposed service quality using ground-truth accuracy trained by data. We demonstrate that the data broker's profit can be almost doubled by using the optimal data size and budget determined by considering the one-leader-multiple-followers Stackelberg game. These results show that a value-added data brokering service can profitably facilitate data trading.

Index Terms—Data marketplace, cloud computing, data discovery, profit maximization, game theory, information-centric network.

1 Introduction

Applications based on Machine Learning (ML) require large amounts of data collected from various sources. In particular, ML model training requires high-quality input data. Because data of many kinds are now generated on massive scales, it is increasingly easy to obtain such large amounts of data from cloud-based data marketplaces. However, these marketplaces face several challenges stemming in large part from variation in the quality of data stored in the giant silos of heterogeneous cloud services and variation in data prices due to the mass production of data. This hinders the development of cost-effective and high-quality ML-based applications. To address these problems and make it easier for customers to obtain data of a quality suitable for their needs, Krishnamachari *et al.* [1] introduced the idea of a marketplace architecture that allows data brokers to buy "raw" data from data providers, apply data analytics, and sell "refined" data to data consumers. Together with the emergence of data service providers (e.g., data brokers). Therefore, in a cloud-based data marketplace, individual datasets can be provided at several different quality levels. Khatri *et al.* [2] suggested that data quality can be evaluated in terms of four dimensions: accuracy, timeliness, completeness, and credibility. However, these dimensions are not directly applicable to the quality of data used for ML model development because the data quality is usually evaluated based on model accuracy [3]. This work focuses on data used as an input for ML-based applications in heterogeneous cloud service domains, so we define data quality based on the accuracy of the trained model [4].

To determine how much a data consumer should be willing to pay for quality data, Niyato *et al.* [5] introduced the willingness-to-pay (WTP) function, $\omega = \omega' \cdot \theta(D)$, where ω' denotes the nominal willingness-to-pay, $\theta(\cdot)$ denotes a non-decreasing function with decreasing marginal accuracy, and D denotes the data size. Thus, ω, the actual willingness-to-pay, increases as the data size D increases. Pedro *et al.* [6] also claim that model accuracy increases with dataset size, and many previous studies have evaluated model quality based solely on the size of the datasets used in their development. However, this work discovers additional factors (outlier removal and data transformation) to improve model accuracy. Our experimental results show that well-refined data can yield greater model accuracy even with comparatively small datasets, reducing data processing costs [7]. Fig. 1 presents an overview of data discovery services in ICN-based cloud computing, with a data broker service acting as a bridge between data consumers and data providers.

The process of acquiring quality data for an ML-based application begins with generating raw data from devices owned by the data provider, who transfers this data to

- *Eunil Seo and Erik Elmroth are with Department of Computing Science, Umeå University, Umeå, Sweden. E-mail: eunil.seo@cs.umu.se, elmroth@cs.umu.se*
- *Hyoungshick Kim is with Department of Software, Sungkyunkwan University, Republic of Korea. E-mail: hyoung@skku.edu*
- *Bhaskar Krishnamachari is with Electrical and Computer Engineering, University of Southern California, Los Angeles, USA. E-mail: bkrishna@usc.edu*

Manuscript received January 19, 2021; revised Feb 25, 2022.

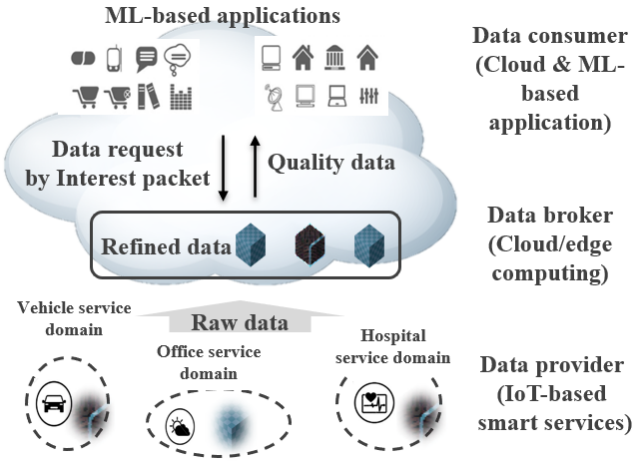


Fig. 1. High level overview of data discovery services in ICN-based cloud computing.

a data broker. The broker then generates multiple refined datasets by processing the raw data and evaluates them by comparing the revenue obtainable from data consumers (i.e., the refined data revenue) to the cost of acquiring data from data providers (i.e., the raw data budget cost). Profit maximization is central to this evaluation; the objective is to identify the dataset quality levels that maximize the broker's profit. To evaluate the profits achievable from refined datasets, we apply a game-theoretic approach [8] involving a non-cooperative competitive data marketplace. The data broker seeks to minimize the data budget cost, while the data providers try to maximize their profit, which derives from the data budget. Additionally, the data broker seeks to maximize its profit by optimizing the data budget, quality, and price. This work highlights the role of the data broker as a value-adding cloud service provider [9] that buys raw data from IoT-based heterogeneous service domains and sells refined data to the cloud and ML-based application developers.

Standard-based data models have emerged from work on large-scale IoT projects [10], [11], [12], and named data discovery was introduced as a component of Information-Centric Networking (ICN) to facilitate the acquisition of required data across heterogeneous service domains [13], [14], [15], [16]. One of the proposed data discovery methods for ICN is Name-Based Access Control (NAC) [17], which enables name-based data discovery services. NAC allows data consumers to obtain necessary data from diverse service domains using data names embedded in an interest packet, as shown in Fig. 1. However, this approach cannot be used for quality data discovery because NAC currently has no quality-related protocol. We, therefore, extend NAC by adding a quality-related requirement. This allows data consumers to discover required data using the data name and to obtain data of the required quality by indicating a nominal WTP. This quality-discovery mechanism would allow data service providers (e.g., data brokers) to offer data of various quality levels at different prices, giving data consumers a wider choice of data meeting their quality requirements.

In the proposed cloud and ICN-based data marketplace, the data consumer can acquire data of the required quality and the data broker service can maximize its profitability by optimizing the service quality and the price of refined data. This work makes the following contributions that enable these outcomes:

- To enable quality data discovery, we embed the WTP mechanism into an ICN running in a data broker. Unlike previously proposed data marketplaces, this allows data consumers to directly demand required-quality data from a data broker using a nominal WTP. Furthermore, we deploy the data broker service in cloud computing to support data discovery transparency between the IoT service domains and the cloud and ML-based applications.
- We formulate one non-cooperative game between a data broker and data providers and one profit optimization problem for a data broker that purchases raw data, improves its quality, and sells the refined data via the cloud. In the context of a one-leader and multiple-follower Stackelberg game, we let data providers (i.e., followers) determine their optimal data size followed by a data budget decided by a data broker (i.e., leader). Then, a data broker determines the optimal service quality, service price, and data budget to maximizes the data broker's profit.
- We introduce the data quality function, a mapping from multiple quality factors to data quality; previously reported utility functions only map data size to data quality and thus allow even low-quality data to achieve 100% accuracy if the dataset is sufficiently large [5]. However, in practice, the accuracy achievable with a given dataset is limited by its quality. Therefore, our utility function limits the maximum achievable accuracy, denoted by Δ, based on the data quality. Furthermore, it allows for the possibility that higher quality data may allow maximum accuracy to be reached more quickly than would be the case with lower quality data.
- We demonstrate how a data broker service running on cloud computing resources can generate datasets of multiple quality levels with different costs and maximize those datasets' profitability using the proposed utility function. We also demonstrate that it is possible to profile each dataset's quality using the data quality function and find the optimal data size, data budget, service quality, and service price for a given use case.

The remainder of this paper is organized as follows. Section 2 presents the system model for the proposed cloud and ICN-based data marketplace using quality-data discovery and a game-theoretic approach. Section 3 formulates the non-cooperative game between stakeholders and defines the utility functions of three stakeholders to solve the profit maximization problem. Section 4 demonstrates how to generate refined datasets with three different quality levels and evaluate their profitability. Section 5 demonstrates the feasibility of data provision based on the proposed quality-data discovery system. Finally, conclusions are presented in Section 7.

2 System Model for ICN-based Data Marketplaces

Data trading is assumed to involve three stakeholders: data providers, a data broker, and data consumers, as shown in Fig. 2.

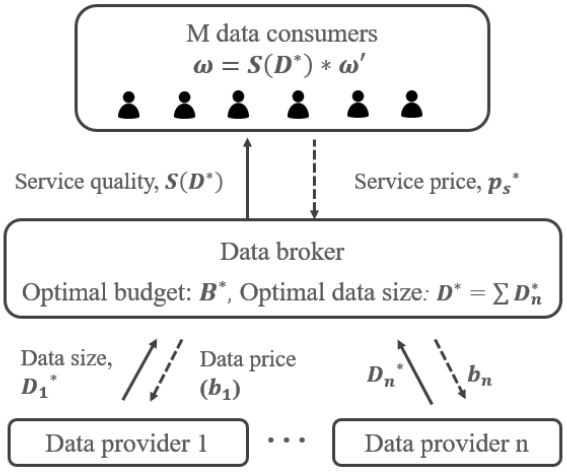


Fig. 2. Systematic view of an ICN-based data marketplace featuring data consumers, data providers, and a data broker.

We specify the interaction between such three stakeholders using the Stackelberg game [8]. The proposed system model involves one competition between data providers, another competition between a data broker and the data providers, and an optimization with respect to the objectives of each stakeholder.

In the first competition, data providers compete by simultaneously bidding for their raw data; such a competition can be expressed as a standard form game in which each data provider submits a bid for its data without knowing the other providers' bids. In the second competition, the data broker and the data providers compete against the budget B and the data size D. Specifically, the data broker tries to optimize its budget B for acquiring raw data while each data provider n tries to optimize its data size D_n to obtain revenue b_n, which is defined by the equation below:

$$b_n = \frac{D_n}{\sum_{n=1}^{N} D_n} B, \tag{1}$$

where b_n represents the revenue of data provider n as a proportion of the data size D_n and N denotes the number of data providers.

In addition to these competitions, this work also aims to optimize a data broker's profit by finding the optimal data budget B^* and the optimal data size D^*, which determines the optimal service quality $S(D^*)$ and service price p_s^*.

We can thus model the interaction between the data broker and the data providers using a hierarchical level one-leader-multiple-followers Stackelberg game in which the leader (the data broker) first suggests its strategy (i.e., the budget B) and then the multiple followers (the data providers) choose their strategies (i.e., their optimal data sizes D_n^*). The data broker can then choose a optimal service price p_s^* and quality $S(D^*)$ according to its strategy B^*, which is determined by the optimal data size D^*, to maximize its profit. Furthermore, each data provider n can plan its optimal strategy D_n^* using the suggested data budget B to maximize its own profit while competing with the other data providers.

Including a maximum nominal WTP (i.e., W') in the broker's utility function extends the optimization of the data broker's profit because the profit depends on the service price p_s and quality $S(D)$ associated with W', which is determined by nominal WTPs (i.e., $\omega' s$) of data consumers. Therefore, this work also defines the relationship between the service price p_s and the maximum nominal WTP W'. Profit optimization based on the service price p_s and the maximum nominal WTP W' is demonstrated in Section 4.4.

The following sections explain the concept of WTP, the service quality and data quality functions, the computation cost, the quality-data discovery mechanism, and the data model. The notation used to define the system model is shown in Table 1.

TABLE 1
Notations used to define the system model.

Symbol	Definition
ω, ω'	Actual willingness-to-pay, nominal willingness-to-pay
$S_{\mathbb{D}}(D; \Delta,$ $\alpha_1, \alpha_2)$	Service quality function with data size D, maximum achievable acc. Δ, and curve fitting param. (α_1, α_2)
Δ	Maximum achievable accuracy of data
$\mathbb{D}, D$	Dataset supplied by data providers $\sum_{n=1}^{N} \mathbb{D}_n$, data size
$\mathbb{D}_n, D_n$	Dataset of the n^{th} data provider, the data size of $\mathbb{D}_n$
n	Index of individual data provider
N	Total number of the data providers
$U_{\mathbb{D}}$	Data utility function of the dataset $\mathbb{D}$
$\theta(\cdot)$	Non-decreasing function with decreasing marginal acc.
$\mathbb{X}_{i,j}$	Input dataset of i^{th} data entry $\mathbb{D}_i = \{\mathbb{X}_{i,j}, y_i\}$
$x_{i,j}$	Features of an input data entry $\mathbb{X}_{i,j} = \{x_{i,1}, \ldots, x_{i,j}\}$
j	Index of a feature of an input data entry
y_i	Label of the i^{th} data entry $\mathbb{D}_i$
$C_{f,D}$	Computation cost with f and D for data processing
μ	Pre-defined weight parameter for energy consumption
$E_{f,D}$	Energy consumption with inputs of f and D
f	Computation resource (i.e., CPU frequency)
k	Constant related to the hardware architecture
X	Computation workload (i.e., CPU cycles per bit)
$C_{f,D}^{\tau}$	Computation cost to achieve the target accuracy τ
I_{τ}	Optimal number of training iterations to achieve τ

2.1 Willingness-to-pay (WTP) of a data consumer

Data consumers can request data by specifying a quality-related requirement (i.e., a nominal WTP, ω') to a data broker and are supplied with refined data, as shown in Fig. 1. The actual willingness-to-pay (i.e., ω) is related directly to the data size [5]; thus, the larger data size is paid higher. However, the data size is not the only factor affecting model accuracy; properly processed data can improve model accuracy even when using a smaller dataset and fewer training iterations. This work maps WTP to data quality (e.g., Δ) and the data size D both; consequently, the proposed WTP utility function can incorporate additional quality factors such as outlier-removal and time-series data transformation as well as data size. For example, the service quality function $S(\cdot)$ is related to WTP by the following expression:

$$\omega = S_{\mathbb{D}}(D; \Delta, \alpha_1, \alpha_2)^{\,1} \cdot \omega' \tag{2}$$

Here, ω denotes the actual WTP, ω' denotes the nominal WTP, Δ denotes the maximum achievable accuracy determined by the data quality of a dataset $\mathbb{D}$ (i.e., the standard, outlier-removal, and transformed datasets), and α_1 and α_2 are the curve fitting parameters.

When the service quality is 1, the data consumer will pay the maximum price (i.e., ω') for the provided service. $w^{'}$ follows a probability distribution supported on the interval $[0, W']$, where W' denotes the maximum nominal WTP. For simplicity, we assume that ω' follows the uniform distribution [5].

This equation indicates that the data broker's service quality function (i.e., $S(\cdot)$) affects the actual WTP (i.e., ω); therefore, a data consumer will purchase a service if their actual WTP is greater than or equal to the service price, i.e., if $\omega \geq p_s$, as shown in Fig. 2. If a data broker provides a higher quality service (i.e., if $S(\cdot)$ increases), it can expect a higher payment, but it will also incur expensive costs due to buying more data. WTP is a measure of the demand in the data marketplace, and in this work, we assume that the WTP function is known to all stakeholders.

2.2 Service quality function

Previous studies proposed a data utility function that maps data size to model accuracy; in other words, the larger the initial dataset, the greater the model accuracy, as follows:

$$U_{\mathbb{D}}(D; \alpha = [\beta_1, \beta_2]) = 1 - \frac{\beta_2}{1 + \beta_1 \cdot D}, \tag{3}$$

where $\mathbb{D} = \{\mathbb{D}_1, \ldots, \mathbb{D}_n\}$ denotes the dataset provided by the N data providers $\{1, \ldots N\}$, D denotes the data size of the dataset $\mathbb{D}$, $U(\cdot)$ is a non-decreasing function with decreasing marginal accuracy [18], and β_1 and β_2 are curve fitting parameters [19].

We modify this function to incorporate two additional behaviors: 1) the maximum achievable accuracy depends on the quality of the data, and 2) increasing the data quality allows the maximum achievable accuracy to be achieved with fewer training iterations. To measure data quality, we introduce the maximum achievable accuracy, denoted by Δ, and define the **service quality function** with the maximum achievable accuracy (i.e., Δ) and two curve fitting parameters (i.e., α_1 and α_2):

$$S_{\mathbb{D}}(D; \alpha = [\Delta, \alpha_1, \alpha_2]) = \Delta \cdot \left(1 - \frac{\alpha_2}{1 + \alpha_1 \cdot D}\right), \tag{4}$$

where $S(\cdot)$ is a non-decreasing function that increases asymptotically towards Δ denoting the maximum achievable accuracy of the model using the data size D of the datasets $\mathbb{D}$ obtained from the data providers.

The service quality $S(D)$ shown in Fig. 2 is the quality of the dataset $\mathbb{D}$ supplied as a product by the data broker. Unlike the data utility depending on the data size D in Eq. (3), the proposed service quality is associated with two inputs: the maximum achievable accuracy Δ and the size D of the dataset $\mathbb{D}$ sold by all data providers collectively.

To demonstrate the proposed service quality function, we use a well-known sensor dataset [20] consisting of readings from fifty-four sensors deployed in an office. Each sensor is placed in a known location and transmits data on four variables: temperature, humidity, light, and voltage. This dataset has over two million entries with eight features. In Section 4, we specify the service quality of the different datasets with its maximum achievable accuracy Δ and the two curve fitting parameters (i.e., α_1 and α_2). Moreover, the different quality datasets used to train a model are the standard, outlier-removal, and time-series transformed datasets.

2.3 Computation cost for training data

To provide high quality data to consumers, a data broker clarifies and processes raw data. Additionally, it must perform model training using the resulting refined data to evaluate its quality. This work considers two data quality improvement techniques: outlier removal and time-series transformation. The computation cost of applying these methods can be evaluated in terms of energy consumption as follows:

$$C_{f,D} = \mu \cdot E_{f,D}, \tag{5}$$

where μ is a pre-defined weight parameter for energy consumption and $E_{f,D}$ is the energy consumption for a task.

Drawing on the energy consumption equation introduced by Mao et al. [21] - see Eq. (6) - we can define the energy consumption in terms of the CPU frequency f of the broker's hardware and the data size D, as follows:

$$E_{f,D} = k \cdot D \cdot X \cdot f^2, \tag{6}$$

Here, k is a constant related to the hardware architecture and X denotes the computation workload (i.e., the number of CPU cycles per bit).

A data broker generates refined data that has been subjected to outlier removal or time series transformation, uses it for model training, and evaluates the trained model's accuracy to assess the quality of the refined data. The computation cost of processing the raw data depends on the computation resource f, the size of the raw data D, and the previously defined parameters μ, k, and X:

$$C_{f,D} = \mu \cdot k \cdot D \cdot X \cdot f^2, \tag{7}$$

To calculate the computation cost associated with a given target accuracy, τ, the data broker increases the number of training iterations until τ is achieved:

$$C^{\tau}_{f,D} = I_{\tau} \cdot C_{f,D}. \tag{8}$$

where $C_{f,D}$ denotes the computational cost of one training iteration and I_{τ} denotes the number of iterations needed to achieve the target model accuracy.

In general, a model's accuracy will increase with the number of iterations. However, the computational cost of model training significantly exceeds that of data cleaning and processing. Furthermore, the model accuracy exhibits

1. $S_{\mathbb{D}}(D; \Delta, \alpha_1, \alpha_2)$, $S(D; \Delta, \alpha_1, \alpha_2)$, $S(D; \Delta)$ and $S(D)$ are used interchangeably for the simplification.

decreasing marginal utility in that it stops increasing after some number of iterations. In this work, $C^{\tau}_{f,D}$ represents a set of ground-truth values obtained from a simulation.

2.4 Quality-data discovery mechanism and data model

In a named-data Network (NDN) [22], data are named and signed by a data provider, and the assigned names are used extensively in the application and network layers [23]. Therefore, it is possible to access data via the ICN service router using its name instead of a network address (i.e., IP address). We use the ICN system to collect data from various data providers across heterogeneous service domains. Additionally, we propose a quality data discovery process that operates as described below (see also Fig. 3):

1) Data consumers request data by specifying **the data name** and **a quality-related requirement** via a quality-based interest packet, which is an extension of the standard interest packet of an ICN.
2) When an ICN service gateway (ICN-SGW), which interfaces with data consumers [13], receives an "interest packet" with a data name (e.g., "/office/data") and a quality-related requirement (i.e., ω': a nominal WTP), it forwards the "interest packet" to a data broker through an interconnected proxy service instance (i.e., a broker) running on the ICN service router.
3) The data broker preprocesses raw data, evaluates the quality of the refined data, and transmits the refined data together with information on the service quality $S(D)$ and price (e.g., p_s) to the data consumers together with the quality-related coefficients Δ, α_1, and α_2.

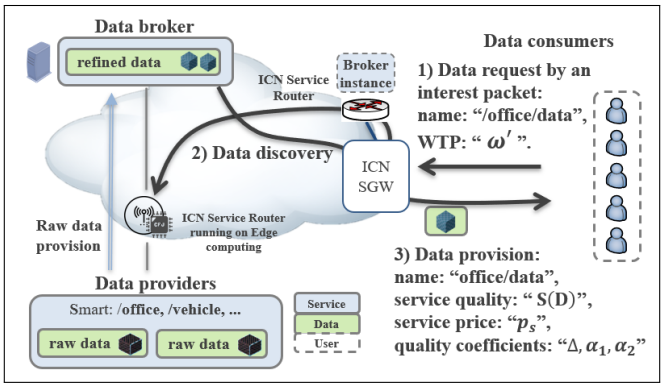


Fig. 3. Proposed quality-data discovery process.

Finally, the data consumer receives target-quality data from the data broker.

The discovery of required-quality data in heterogeneous service domains necessitates the existence of identifiers that are used and understood by both data consumers and data providers to enable data trading in the ICN-based data marketplace. We, therefore, employ hierarchical named identifiers for data and other IoT entities (e.g., devices, services, users, and applications) in a *single identification scheme* [24]. All IoT entities can be named using this identification scheme in a unified way. Using these named identifiers, the data model can express relationships between IoT entities using two attributes: the "service" attribute, which represents a service domain, and the "device" attribute, which represents the name of the device producing the corresponding data. Based on the named identification scheme, we design a data model to enable data discovery based on *data names* and *data quality*. This data model is based on the entity–attribute–value (EAV) model, which is a popular interoperable data model [25]. The EAV data model uses the following attributes to describe the data and its quality:

- name: a named identifier (e.g., "/office/data").
- type: entity type (e.g., sensor data).
- location: measured location, type: (geo:json).
- service: purpose of data, type: named identifier.
- device: device producing data, type: named identifier.
- ω': the nominal WTP.

3 The Non-Cooperative Competition Game and Profit Optimization

We explore the relationship between the service price, service quality, and data market size, which is determined by data quality and the amount that data consumers are willing to pay. Additionally, we demonstrate the feasibility of quality-data trading with direct participation of data consumers.

Data trading can occur in various market structures including monopolistic and oligopolistic markets, single or double-sided auction markets [26], and auction-based marketplaces using non-cooperative competition [27]. This work employs one non-cooperative competition game together with profit optimization based on the various consumer payments in our data marketplace.

The notation used to define the utility functions is shown in Table 2.

TABLE 2
Notations used in the context of profit optimization.

Symbol	Definition
W, W'	Actual maximum WTP, nominal maximum WTP
M	Number of the data consumers
U_n	Utility function of a data provider n
U_R	Utility function of the data broker R
$\mathbb{D}^*, D^*$	Optimal dataset of a data broker, optimal data size
B^*, b^*	Optimal budgets of a data broker and a data provider
p_s, p_s^*	Service price, optimal service price
C^*	Optimal cost, defined as the sum of B^* and $C^{\tau}_{D^*}$

3.1 Game procedure and profit maximization

Fig. 4 illustrates the data trading flow between stakeholders; using the proposed quality-data discovery process, the data consumers can request quality data by specifying a nominal WTP, ω', and can participate in data trading directly, unlike in other proposed systems [5] [28] [29]. Thus, the data consumers play a key role in sustaining data trading by specifying their nominal WTPs, ω'; in effect, the range and amount of nominal WTPs determine the size of the data marketplace and affect all other stakeholders' profits.

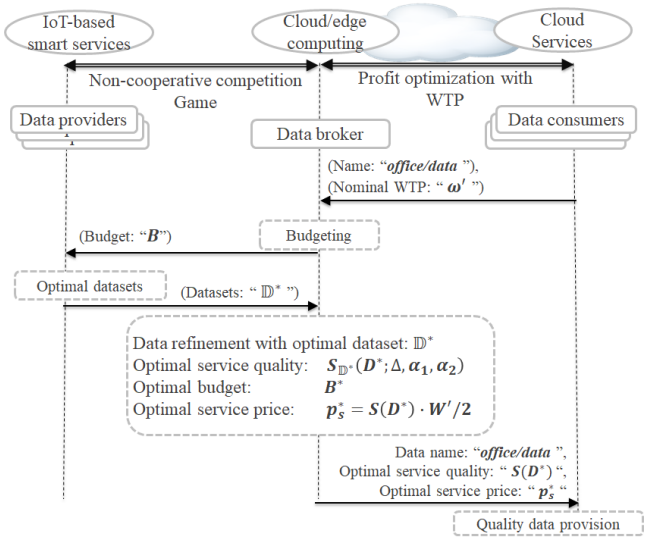


Fig. 4. Non-cooperative competition game between a data broker and data providers, and profit optimization based on WTP between a data broker and data consumers.

This work focuses on the competition between the data providers and a data broker, and profit optimization between a data broker and data consumers, as shown in Fig. 4. A data broker maximizes its profit by discovering an optimal data price p_s^* for the data consumer and finding the data providers' optimal data budget B^*. Based on non-cooperative competition between data providers, the optimized data size D^* can be determined when a data broker decides a data budget B first and informs data providers, as described in Section 3.2. A data broker can then optimize the price of refined data based on this optimal data size and determine the provided service quality, as described in Section 3.3. Finally, it can find the maximized profit by using the optimal budget B^* and the optimal service quality $S(D^*)$ together with the data broker utility function given in Eq. (12). Profit optimization allows the data broker to optimize its profit against the data consumers based on service quality (i.e., refined-data quality) and nominal WTPs. The higher the service quality, the higher the revenue, but higher service quality also necessitates a larger data budget and a greater computation cost for data training.

3.2 Utility of a data provider

Once the data consumers have specified their nominal WTPs, a data broker can predict the revenue from the nominal WTPs and budget to buy data from the data providers. Budgeting is critical for the data broker because data providers will determine the quantity of data they supply based on the given budget. Maximization of the data provider's profit depends strongly on the optimal data size for the data broker. Therefore, providers consider the relationship between their payment (which depends on the data size) and the cost of generating data. We can express the payment that the provider receives for their data in relation to the data broker's overall budget B using the expression $b_n = \frac{D_n}{\sum_{n=1}^{N} D_n} B$, where b_n denotes the data payment and D_n denotes the data size of a data provider n.

We formulate a non-cooperative competition game between a data broker and a data provider in which the broker determines the data budget that maximizes its profit and the data provider determines the amount of data it supplies by maximizing its profit with no cooperation between the two parties. This allows us to define the following utility function for a data provider n:

$$U_n(D_n, \gamma_n) = \underbrace{b_n}_{\text{revenue}} - \underbrace{\gamma_n D_n}_{\text{cost}}, \tag{9}$$

where b_n denotes the data payment from a data broker, γ_n denotes the data generation cost, and D_n denotes the data size to be provided.

In keeping with Theorem 4 of a previously reported approach [28], if all data providers in the market participate ($D_n > 0, \forall n$) and we impose two conditions $n > 1$ and $\sum_{i \neq n} \gamma_i > (n-2)\gamma_n$, the optimal data size becomes:

$$D_n^* = \frac{B(N-1)(\sum_{n=1}^{N} \gamma_n - (n-1)\gamma_n)}{(\sum_{n=1}^{N} \gamma_n)^2}, \forall n, \tag{10}$$

where D_n^* denotes the optimal data size to a data provider n.

We can then sum these sizes and simplify the expression to obtain:

$$D^* = \sum_{n=1}^{N} D_n^* = \frac{B(N-1)}{\sum_{n=1}^{N} \gamma_n}, \tag{11}$$

where D^* denotes the sum of the optimized data sizes sold by all data providers for a data broker.

3.3 Utility of a data broker

The data broker must decide what amount of data to purchase from the data providers $\{1, \ldots, n\}$ and what service price p_s to charge the data customers. The data customers differ in their willingness-to-pay w' and their required data quality. If the willingness-to-pay of a data customer exceeds or equals the data price p_s charged by the data broker, the data customers will buy the required-quality data.

Therefore, when the M data customers request a dataset, we can define the utility of a data broker as their revenue minus their costs. The revenue is the service price paid by the data customers whose willingness-to-pay exceeds or equals the service price specified by the data broker. The costs consist of the budget for purchasing data and computation costs:

$$U_R(p_s, B, D) = \underbrace{p_s \cdot M \cdot Pr(\omega \geq p_s)}_{\text{revenue}} - \underbrace{B - C_D^{\tau}}_{\text{cost}}, \tag{12}$$

Here, p_s denotes the service price offered by a data broker, B denotes the budget for purchasing data, D denotes the data size provided by the M data providers shown in Eq. (11), and ω denotes the actual willingness-to-pay. C_D^{τ} [2] denotes the computation cost to achieve the target accuracy τ using a dataset of size D.

To determine the optimal service price p_s^*, we use the proposed willingness-to-pay $\omega = S(D) \cdot \omega'$, which in turn depends on the service quality $S(D; \Delta, \alpha_1, \alpha_2)$. The inputs

2. C_D^{τ} and C_f^{τ}, D are used interchangeably because f is a constant for a given data broker.

of the service quality are the data size D, the maximum achievable accuracy Δ, and the curve fitting parameters α_1 and α_2, as shown in Eq. (2). In addition, we assume the maximum actual willingness-to-pay W and use the optimal data size D^* supplied by the N data providers. We thus obtain the expression $W = S(D^*) \cdot W'$, where W' denotes the maximum nominal willingness-to-pay. Finally, we can express the utility of the data broker using the optimal data size D^* as follows:

$$\begin{aligned} U_R(p_s, B, D^*) &= p_s MPr(W \geq \omega \geq p_s) - B - C^{\tau}_{D^*}, \\ &= p_s MPr(S(D^*)W' \geq p_s) - B - C^{\tau}_{D^*}. \end{aligned}$$

We assume that the actual willingness-to-pay ω is uniformly distributed over $[0, W]$, where W is greater than or equal to 1. Therefore, we can describe the utility of a data broker as follows:

$$U_R(p_s, B, D^*) = p_s \cdot M \cdot \frac{(S(D^*)W' - p_s)}{W'} - B - C^{\tau}_{D^*}. \tag{13}$$

We can solve for the optimal service price, p_s, by differentiating the utility function of the data broker as follows:

We can solve the optimal service price, p_s, by differentiating the utility function of a data broker, as follows:

$$\frac{\partial U_R(p_s, B, D^*)}{\partial p_s} = M\left(S(D^*) - \frac{2p_s}{W'}\right). \tag{14}$$

The second derivative of $U_R(p_s, B, D^*)$ is $\frac{\partial^2 U_R(p_s, B, D^*)}{\partial p_s^2} = -2M$, which must be less than 0. the optimal (i.e. profit-maximizing) service price is obtained when the first derivative of $U_R(p_s, B, D^*)$ is zero. This price can be expressed as follows:

$$p_s^* = \frac{S(D^*)W'}{2}. \tag{15}$$

According to Eq. (15), the optimal service price depends on service quality and willingness-to-pay, and it is assumed that the latter is uniformly distributed.

To find the optimal budget, we define a modified data broker utility function $U_R(B, D^*)$ by substituting the optimal service price into the original utility function $U_R(p_s, B, D^*)$:

$$U_R(B, D^*) = \frac{MS(D^*)^2(2W' - 1)}{4} - B - C^{\tau}_{D^*}. \tag{16}$$

By differentiating the expressions for D^* (Eq. (11)) and $S(D^*; \Delta)$ (Eq. (4)) with respect to B, we then obtain the variables x_1 and x_2:

$$\begin{aligned} x_1 &= \frac{\partial D^*}{\partial B} = \frac{(N-1)}{\sum_{n=1}^{N} \gamma_n} \\ x_2 &= \frac{\partial S(D^*)}{\partial B} = \frac{\alpha_1 \alpha_2 \Delta x_1}{(1 + \alpha_1 B x_1)^2}. \end{aligned} \tag{17}$$

We can then express the first and second derivatives of the data broker utility function $U_R(B, D^*)$ as follows:

$$\begin{aligned} \frac{\partial U_R(B, D^*)}{\partial B} &= \frac{MW' x_2 S(D^*)}{2} - 1 - x_1 C^{\tau}_{D^*}, \\ \frac{\partial^2 U_R(B, D^*)}{\partial B^2} &= \frac{-MW'\alpha_2 \Delta (\alpha_1 x_1)^2 S(D^*)}{(1 + \alpha_1 B x_1)^3} + \frac{MW' x_2^2}{2}, \\ &= \frac{MW'\alpha_2 (\Delta \alpha_1 x_1)^2 (-2 - 2\alpha_1 B x_1 + 3\alpha_2)}{2(1 + \alpha_1 B x_1)^4}. \end{aligned} \tag{18}$$

To ensure that our utility function converges on a global, rather than local maximization, the second derivative , $\frac{\partial^2 U_R(B, D^*)}{\partial B^2} < 0$, must be negative. Therefore, we must impose the condition $2 + 2\alpha_1 B x_1 > 3\alpha_2$. Under this condition, the optimal solution is as follows:

$$\frac{MW'\alpha_1\alpha_2\Delta^2 x_1 (1 + \alpha_1 B^* x_1 - \alpha_2)}{2(1 + \alpha_1 B^* x_1)^3} = 1 + x_1 C^{\tau}_{D^*},$$

where $(1 + \alpha_1 B^* x_1)$ is the solution of the following cubic equation:

$$t^3 - A_0 t + A_1 = 0.$$

Here, for simplicity, we have substituted $\frac{MW'\alpha_1\alpha_2\Delta^2 x_1}{2(1 + x_1 C^{\tau}_{D^*})}$ with A_0 and $\frac{MW'\alpha_1\alpha_2^2\Delta^2 x_1}{2(1 + x_1 C^{\tau}_{D^*})}$ with A_1.

To ensure that this cubic equation has a root greater than unity, we assume that $\frac{MW'\alpha_1\alpha_2\Delta^2 x_1}{2(1 + x_1 C^{\tau}_{D^*})} > \frac{1}{1 - \alpha_2}$ and $\alpha_2 \leq 2/3.$, and we solve the cubic equation using the trigonometric formula [30] with the assumption that we have three real roots, and cannot have 2 distinct roots greater than 1 because if we had two such roots then we would have two positive values of B, making it impossible to satisfy $\frac{\partial^2 U_R(B, D^*)}{\partial B^2} < 0$ for $B > 0$.

We have three real roots based on the following expression:

$$\begin{aligned} -4A_0^3 + 27A_1^2 = A_0^2\left(27\alpha_2^2 - 4A_0\right) &< A_0^2\left(27\alpha_2^2 - \frac{4}{1 - \alpha_2}\right) \\ &= -A_0^2 \frac{(3\alpha_2 - 2)^2(3\alpha_2 + 1)}{1 - \alpha_2} < 0. \end{aligned}$$

Therefore, we find the largest of the following three roots:

$$\begin{aligned} t_0 &= 2\sqrt{A_2}\cos\left(\frac{1}{3}\arccos(A_3)\right), \\ t_1 &= 2\sqrt{A_2}\cos\left(-\frac{1}{3}\arccos(A_3) + \frac{2\pi}{3}\right), \\ t_2 &= 2\sqrt{A_2}\cos\left(\frac{1}{3}\arccos(A_3) + \frac{2\pi}{3}\right), \end{aligned}$$

where we have substituted $\frac{MW'\alpha_1\alpha_2\Delta^2 x_1}{6(1 + x_1 C^{\tau}_{D^*})}$ with A_2 and $\frac{-3\alpha_2}{2}\sqrt{\frac{6(1 + x_1 C^{\tau}_{D^*})}{MW'\alpha_1\alpha_2\Delta^2 x_1}}$ with A_3 for simplicity.

We then define the variable $\theta := \frac{1}{3}\arccos(A_3)$ and note that $\theta \in (\pi/6, \pi/3)$ because it is assumed that $A_3 \in (-1, 0)$. Therefore, we have $0 < \theta < -\theta + \frac{2\pi}{3} < \pi$ and $0 < \theta < \theta + \frac{2\pi}{3} \leq \pi$, which means that $t_0 > t_1$ and $t_0 > t_2$. Thus, the largest root is t_0, and the optimal budget is given by:

$$B^* = \frac{1}{\alpha_1 x_1}\left(2\sqrt{A_2}\cos\left(\frac{1}{3}\arccos(A_3)\right) - 1\right). \tag{19}$$

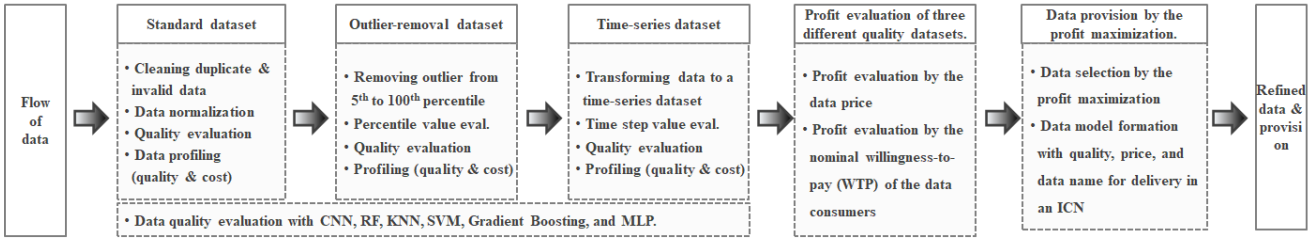


Fig. 5. Flow diagram illustrating the processing of data and the generation and evaluation of refined datasets to maximize the profit of a data broker.

4 USE CASE: SERVICE QUALITY EVALUATION AND PROFIT OPTIMIZATION FOR SENSOR DATA.

This section demonstrates how a data broker can optimize its profit using the raw sensor dataset. The data broker generates three different quality datasets, evaluates the quality of each one, and resolves its profits using the data broker's utility function, $U_R(p_s, B, D)$ in specified in Eq. (13). Finally, a data broker determines and provides the most profitable dataset to the data consumers in the data marketplace.

A data broker can optimize its profit by using the utility function, $U_R(p_s, B, D) = \underbrace{p_s \cdot M \cdot (S(D;\Delta) \cdot W' - p_s)/W'}_{\text{revenue}} - \underbrace{B - C^{\tau}_{D}}_{\text{cost}}$, to maximize revenue and minimizing the data acquisition budget B and the computation resource cost C^{τ}_{D}. The revenue increases with both $S(D;\Delta)$ or W'; however, a data broker can decide to increase the service quality $S(D;\Delta)$ to maximize the revenue because W' is determined by the distribution and amount of nominal WTPs (i.e., $\omega' s$) offered by the data customers. Using the optimal service price p^*_s defined in Eq. (15) and the optimal data size D^* defined in Eq. (11), we can define the maximized revenue, as follows:

$$p^*_s \cdot M \cdot (S(D^*;\Delta) \cdot W' - p^*_s)/W', \tag{20}$$

where M denotes the number of the data consumers, W' denotes the maximum nominal willingness-to-pay, and Δ denotes the maximum achievable accuracy.

With the optimal data size D^* in Eq. (11) and the optimal budget B^* in Eq. (19), we can define the optimal cost to each dataset, as follows:

$$C^* = B^* + C^{\tau}_{D^*}. \tag{21}$$

It is worth noting that the quality of each dataset determines Δ, α_1, and α_2, and that a data broker can use the service quality function, Eq. (4), which depends on these quality-related coefficients (i.e., Δ, α_1, and α_2), to predict their revenue and profit. The service quality function thus allows a data broker to identify the most profitable dataset by resolving these coefficients, as follow:

TABLE 3
Quality-related coefficients and optimized cost for profit prediction.

Δ	α_1	α_2	$C^* = B^* + C^{\tau}_{D^*}$

Most of the data collected by a data broker from heterogeneous service domains will be raw data that has not undergone any cleaning or preparation. Therefore, a data broker performs basic data preparation by cleaning empty entries and removing duplicated and redundant entries. In addition to simple data cleaning, the data broker may apply one or more data refinement techniques to increase the optimal service quality $S(D^*;\Delta)$ based on the optimal dataset $\mathbb{D}^*$. Here we consider two such refinement techniques: outlier removal and time-series transformation, as shown in Fig. 5.

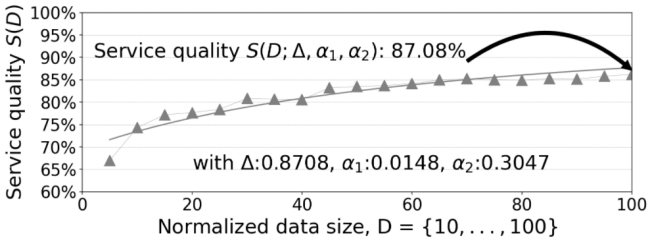
By applying these techniques, three datasets are obtained after data cleaning and preparation: the standard dataset, the outlier-removal dataset, and the time-series transformed dataset. All three can be used directly as an inputs for ML model training. The broker first evaluates the optimal service quality $S(D^*;\Delta)$ of the standard dataset and assesses its profitability $U_R(p^*_s, B^*, D^*)$ based on the optimal service quality and the optimal costs (i.e., B^* and $C^{\tau}_{D^*}$).

After generating the standard dataset, the broker additionally generates the outlier-removal and time-series datasets to improve the optimal service quality of the data. It evaluates the optimal service quality and profitability of these two refined datasets. Each dataset will have different profitability because the quality-related coefficients (i.e., Δ, α_1, and α_2) depend on the quality of the dataset under consideration.

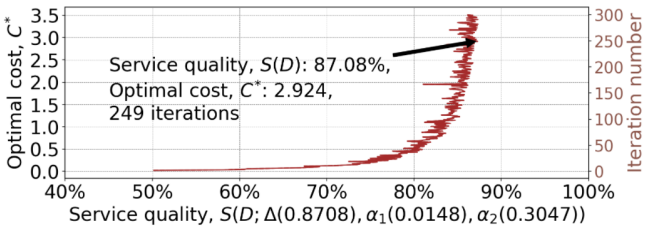
To maximize profit, the data broker evaluates all three datasets and chooses the most profitable one to sell to data consumers. Section 4.1 describes the data cleaning for a standard dataset and illustrates the evaluation of its profitability. Section 4.2 explains how to remove outliers to improve data quality and the profitability evaluation of the resulting dataset. Section 4.3 describes the time-series transformation process and the profit evaluation of the corresponding dataset. Finally, data provision is described in Section 5.

4.1 Standard data generation, quality evaluation, and profiling

The expected price for the raw data is comparatively low because there is no added cost for data processing. The price of refined data is higher due to the costs of model evaluation and data refinement. The greater price of a refined dataset reflects its greater quality, which means a data broker is providing a value-added service. Note that the standard, outlier-removal, and time-series transformed datasets each have a distinct optimal price $p^*_s = \frac{S(D^*)W'}{2}$ and the costs (i.e., B^* and $C^{\tau}_{D^*}$) that depend on their quality. Consequently, the associated profits are also different.



(a) Service quality of the standard dataset as a function of data size.



(b) The optimal cost and number of iterations for the standard dataset.

Fig. 6. The values plotted in (a) are the F1 scores of the CNN model using the standard dataset. The solid line shows the output of the service quality function, Eq. (4) and the corresponding quality-related coefficients (i.e., $\Delta : 0.8708$, $\alpha_1 : 0.0148$, and $\alpha_2 : 0.3047$). The Figure (b) shows the optimal cost and iteration number as a function of the model accuracy. An accuracy of 87.08% requires 2.924 of the optimal cost and 249 iterations.

The standard dataset is obtained by cleaning the raw data obtained from the data providers. This dataset, $\mathbb{D}_i$, consists of the input-output pairs $\{\mathbb{X}_{i,j\in\{1,\ldots,8\}}, y_i\}$, where the input data is $\mathbb{X}_{i,j\in\{1,\ldots,8\}} = \{X_{i,1}, \ldots, X_{i,j}, \ldots, X_{i,8}\}$, and j is the index of the features. y_i is the labeled output value for $\mathbb{X}_{i,j\in\{1,\ldots,8\}}$. The standard dataset contains over $2,210,084$ data entries (i.e., $D = 2,210,084$), each of which has 8 features as shown in Table 4: $X_{i,1}$: date, $X_{i,2}$: time, $X_{i,3}$: sequence number, $X_{i,4}$: sensor ID, $X_{i,5}$: temperature, $X_{i,6}$: humidity, $X_{i,7}$: light, and $X_{i,8}$: voltage.

TABLE 4
A data entry from the sensor dataset.

$X_{i,1}$	$X_{i,2}$	$X_{i,3}$	$X_{i,4}$	$X_{i,5}$	$X_{i,6}$	$X_{i,7}$	$X_{i,8}$

The dataset is a multivariate time-series (MTS) with the following properties:

- $\mathbb{D} = \{\mathbb{D}_1, \ldots, \mathbb{D}_i, \ldots, \mathbb{D}_D\}$, where D is 2,210,084.
- $\mathbb{D}_i = \{\mathbb{X}_{i,j\in\{1,\ldots,8\}}, y_i\}$, where $\mathbb{X}_{i,j\in\{1,\ldots,8\}}$ is the vector of the input data entry, which is the i^{th} data entry measured by the sensor ID (i.e., y_i) taking values between 1 and 50.
- $\mathbb{X}_{i,j\in\{1,\ldots,8\}} = \{X_{i,1}, X_{i,2}, \ldots, X_{i,8}\}$ denotes an input data entry with eight features; the index j thus ranges from 1 to 8.
- Univariate time-series (UTS, e.g., $\mathbb{X}_{i,j\in\{1\}}$) can be obtained from the MTS (e.g., $\mathbb{X}_{i,j\in\{1,\ldots,8\}}$) by extracting time series for individual features: $\{X_{i,1}\}$.

Fig. 7 shows the accuracy of models trained using

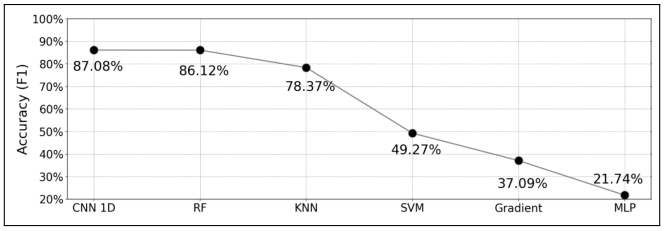


Fig. 7. Model accuracy comparison of six ML algorithms trained using the standard dataset after cleaning.

six different ML algorithms: Convolutional neural network (CNN), Random forest (RF), K-nearest neighbors (KNN), Support Vector Machine (SVM), Gradient Descent, and multilayer perceptron (MLP). The CNN model gives the best fit, with 87.08% accuracy. RF also yielded a high accuracy, 86.12%, and allowed for faster model training than CNN. However, only the CNN algorithm could achieve a very high accuracy of 99.36% using the time-series dataset while also achieving a high accuracy, 87.08%, with the standard dataset; therefore, this algorithm was used exclusively in all subsequent steps.

Fig. 6-(a) shows the ground-truth accuracy as a dotted line and the predicted accuracy as a solid line. The ground-truth accuracy is obtained from a real ML model using the standard dataset, while the predicted accuracy is obtained using the service quality function shown in Eq. (4). In this case, the values of the two curve fitting parameters of the service quality function are $\alpha_1 = 0.0148$ and $\alpha_2 = 0.3047$. The maximum achievable accuracy Δ is 0.8708 (i.e., 87.08%) and depends mainly on the data quality; furthermore, the service quality also increases as the data size increases.

Fig. 6b shows how the optimal cost and required number of iterations for the CNN model's training using the standard dataset increase as the model's accuracy increases. The optimal cost is required to evaluate the model accuracy and increases in proportion to the number of iterations. The optimal cost C^* for a given model accuracy is calculated using Eq. (8) and increases in proportion to the number of iterations. At the maximum achievable accuracy of 87.08%, C^* is 2.924. The number of iterations required to achieve this accuracy is 249.

The service quality profile of the standard dataset consists of the quality-related coefficients Δ, α_1, and α_2 together with the optimal cost (i.e., C^*), as shown in Table 5.

TABLE 5
Service quality profile of the standard dataset.

Δ	α_1	α_2	C^*
0.8708	0.0148	0.3047	2.924

4.2 Outlier-removal dataset generation, quality evaluation, and profiling

We can use the principle of temporal continuity to detect and remove outliers from the collected sensor data. This approach relies on the fact that the patterns in the data are not expected to change abruptly in the absence of abnormal events if the sensors are working correctly [31]. Sensor data

is widely used for anomaly detection [32] because it is streamed temporal data with characteristics similar to those of personal biodata, mechanical systems diagnosis data, network traffic data, and financial data. Outlier detection has therefore been studied extensively in the context of sensor data [33].

The data broker uses the encoder-decoder architecture depicted in Fig. 8 to remove outliers. The encoder-decoder

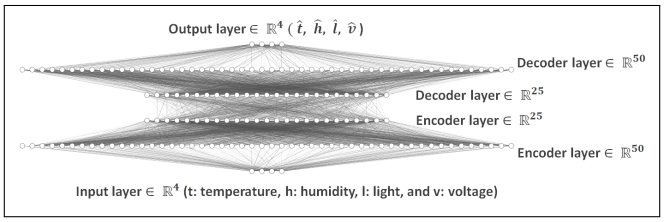


Fig. 8. Encoder-decoder architecture (4D input, 50D & 25D encoders, 25D & 50D decoders, 4D output).

has a four-dimensional input (t: temperature, h: humidity, l: light, and v: voltage), two encoders, two decoders, and a four-dimensional output ($\hat{t}, \hat{h}, \hat{l}, \hat{v}$). The data broker applies the encoder-decoder algorithm to all fifty sensor datasets separately and thus obtains fifty well-trained encoder-decoder models, all with a mean square error (MSE) below 0.1%. Each model represents the character of one of the fifty sensor locations with high accuracy. Fig. 9 shows encoder-decoder models for the first and sixth sensor locations as examples. By adjusting the percentile threshold for outlier

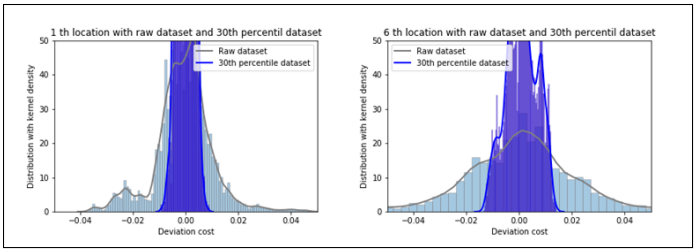


Fig. 9. Distributions of the raw data and the 30^{th} percentile data for the 1^{st} and 6^{th} sensor locations.

removal (e.g., one might remove all values below the 5^{th} percentile and above the 95^{th} percentile from the data), we obtain refined datasets with different shapes. We can then compare the shape similarity of the refined datasets for different locations to that for the standard dataset. The test results revealed that outlier removal at the 30^{th} percentile level yielded better shape similarity than the standard dataset.

Model accuracy varied with the percentile value chosen for outlier removal. Therefore, to find the optimal

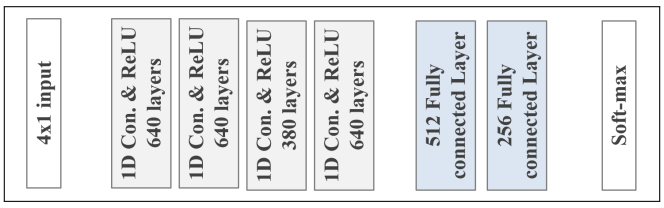


Fig. 10. CNN architecture for training using the outlier removal datasets.

removal percentile, we generated outlier-removal datasets with removal thresholds ranging from the 10^{th} to the 90^{th} percentile, rising in increments of 10. Each of the resulting ten outlier-removal datasets was used to train a one-dimensional CNN with four convolution layers and two dense layers, as shown in Fig. 10. The four convolution layers have 640, 640, 384, and 640 neurons, respectively, while the two fully connected dense layers have 512 and 256 neurons. The total number of parameters to train was 1,386,034.

As shown in Fig. 11, the highest accuracy (94.29%) was achieved using the 30^{th} outlier removal dataset obtained by removing 70% of the data in the standard dataset. It is

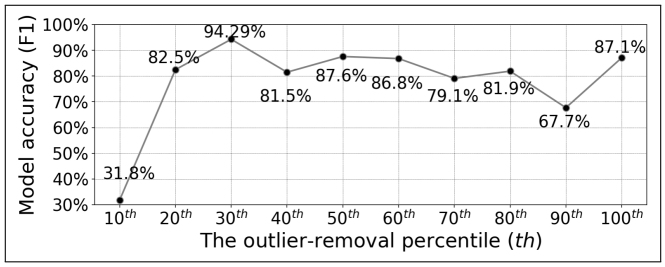


Fig. 11. Model quality using the outlier removal datasets of different removal percentiles.

worth noting that there is no clear relationship between the outlier removal percentile and model accuracy; intuitively, one might expect that the best way of finding the optimal outlier-removal percentile would be through simulation given the potential variation in the quality and characteristics of the data.

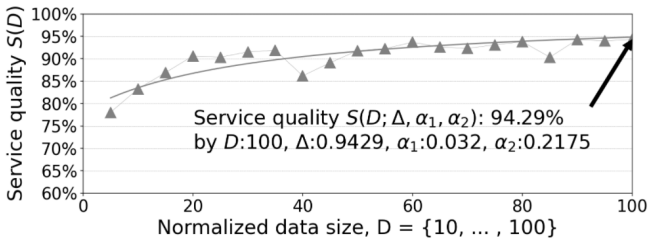
Fig. 12a shows how the service quality achieved using the 30^{th} outlier removal dataset varies with the normalized data size. The maximum achievable accuracy Δ is 0.9429 (i.e., 94.29%) as noted above. The optimized curve fitting parameters for this dataset are $\alpha_1 : 0.0320$ and $\alpha_2 : 0.2175$. For comparative purposes, the maximum achievable accuracy with the standard dataset (i.e., the 100th outlier removal dataset) is 87.08%, and its curve fitting parameters are $\alpha_1 : 0.01480$ and $\alpha_2 : 0.3047$.

Fig. 12b shows how the optimal cost and iteration count of the CNN model trained using the outlier removal dataset vary as the model accuracy increases. As expected, the optimal cost and number of iterations increase with the accuracy. At the maximum achievable accuracy of 94.28%, the optimal cost obtained with Eq. (21) is $C^* = 0.547$ and 156 iterations are required. The outlier removal dataset yields the best (i.e., lowest) optimal cost for CNN model training when compared to the standard and time-series datasets while also yielding greater accuracy than the standard dataset.

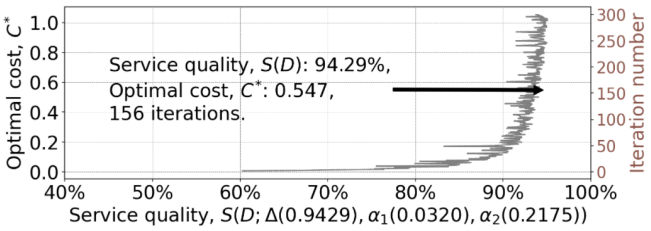
TABLE 6
The outlier removal dataset including the deviation feature.

$X_{i,1}$	$X_{i,2}$	$X_{i,3}$	$X_{i,4}$	$X_{i,5}$	$X_{i,6}$	$X_{i,7}$	$X_{i,8}$	$X_{i,\ dev\_score}$

The addition of a deviation score to the dataset allows data consumers to flexibly train models without losing information from the original data. To obtain the deviation



(a) Service quality of the 30^{th} percentile outlier removal dataset.



(b) Optimal cost and iteration count of the 30^{th} percentile outlier removal dataset.

Fig. 12. The points plotted in (a) are the accuracies of models trained using the 30^{th} percentile outlier removal dataset. The solid line is the output of the service quality function (Eq. (4)) with the parameter values $\Delta : 0.9429$, $\alpha_1 : 0.0320$, and $\alpha_2 : 0.2175$. Figure (b) shows how the optimal cost and required iterations increases with the model accuracy. For an accuracy of 94.29%, the optimal cost is 0.547 and required iterations are 156.

score, we calculate the deviation of the output data from the input data using Eq. (22):

$$X_{i,\ dev\_score} = \frac{\sum_{j=5}^{8} \left(X_{i,j} - \hat{X}_{i,j}\right)^2}{4}. \tag{22}$$

The denominator of 4 is included in this expression because only four features are included in the deviation analysis (i.e., t_i, h_i, l_i, v_i). X_{ij} denotes the actual i^{th} data entry and its j^{th} feature (where j ranges from 5 to 8), and $\hat{X}_{ij}$ denotes the data predicted by the encoder-decoder model for a given sensor location.

We can generate an outlier-removal dataset (See Table 6) for each sensor location (i.e., $y = \{1, \ldots, 50\}$) that includes the deviation feature $X_{i,\ dev\_score}$, as follows:

- We extract the dataset transmitted by a chosen sensor - in this case, the 1^{st} - from the full dataset $\mathbb{D}$. This gives us the sensor-specific dataset $\mathbb{D}_{y=1} = \{\mathbb{D}_{1,y_1=1}, \ldots, \mathbb{D}_{i,y_i=1}, \ldots, \mathbb{D}_{D,y_D=1}\}$, where D denotes the number of data entries and $y = 1$ is the senor ID. This extracted dataset contains $36,911$ data entries (i.e., $D = 36,911$) that are used to train the encoder-decoder.
- $\mathbb{D}_{i,y_i=1}$ consists of the input-output pairs $(\mathbb{X}_{i,j}, 1)$, where i denotes the index of the data entry transmitted by the 1^{st} sensor and j denotes the index of features.
- To calculate the deviation score using Eq. (22), X_{ij} and $\hat{X}_{ij}$ are required. X_{ij} can be obtained using the expression $\{\mathbb{X}_{i,j\in\{5,6,7,8\}}, y_i = 1\}$, and $\mathbb{D}_{i,y_i=1}$ can be represented as $\{X_{i,5} : t_i,\ X_{i,6} : h_i,\ X_{i,7} : l_i,\ X_{i,8} : v_i,\ y_i = 1\}$.

The output data $\hat{X}_{ij}$ predicted by the trained encoder-decoder model has the following form:

- $\hat{\mathbb{D}}_{i,y_i=1}$ consists of the input-output pairs $(\hat{\mathbb{X}}_{i,j}, 1)$, where $\hat{\mathbb{X}}_{i,j}$ denotes $\{\hat{X}_{i,5} : \hat{t}_i,\ \hat{X}_{i,6} : \hat{h}_i,\ \hat{X}_{i,7} : \hat{l}_i,\ \hat{X}_{i,8} : \hat{v}_i\}$ and represents the output data predicted by the encoder-decoder model for the 1^{st} sensor.
- Finally, $X_{i\ dev\_score}$ is obtained using X_{ij} and $\hat{X}_{ij}$, allowing the data broker to generate the outlier-removal dataset as shown in Table 6.

We profile this outlier-removal dataset using the quality-related coefficients Δ, α_1, and α_2 together with the optimal cost (i.e., C^*) as shown in Table 7. The quality-related coefficients are $\Delta : 0.9429$, $\alpha_1 : 0.0320$, and $\alpha_2 : 0.2175$ and the optimal cost C^* calculated using Eq. (21) is determined to be 0.547 based on the optimal data size D^* and budget B^*.

TABLE 7
Service quality profile of an outlier removal dataset.

Δ	α_1	α_2	C^*
0.9429	0.0320	0.2175	0.547

4.3 Time-series data generation, quality evaluation, and profiling

We can transform data by scaling, decomposition, aggregation, feature selection, and data format transformation, giving a wide range of options for improving model accuracy. In this work, the data broker transforms the standard dataset consisting solely of numeric values into a time-series [34]. The one-dimensional numeric values (e.g., temperature, humidity, light, and voltage) are thus transformed into a two-dimensional dataset as shown in Fig. 13.

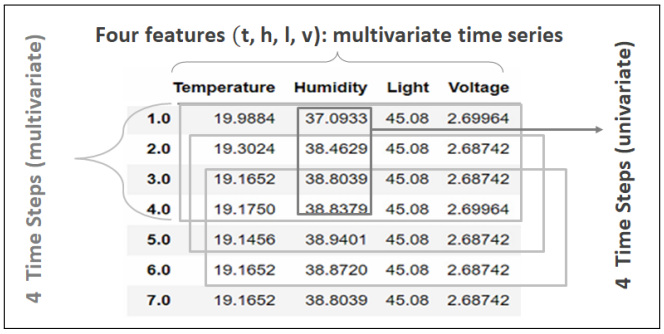
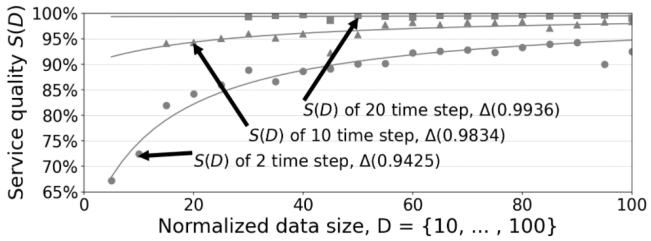


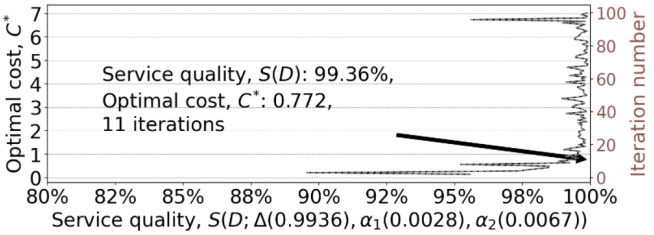
Fig. 13. Transforming one-dimensional data into a multivariate time-series.

Such two-dimensional datasets can be used as inputs for deep learning models such as 2D CNN, RNN, and LSTM models, which are trained by mapping the time-series input to the probability distribution over the class variable values.

Fig. 13 shows the dataset for sensor location#1, which includes 36,911 data entries. To generate a time-series dataset, the data broker specifies a batch size of 50, a time-step of 4, and a feature number of 4 (corresponding to temperature, humidity, light, and voltage). The input data thus becomes



(a) Service quality $S(D)$ after training with three time-series datasets (i.e., 2, 10 and 20 time steps).



(b) The optimal cost and iteration count for the CNN model using the twenty-time-step dataset as a function of the service quality, Eq. (4).

Fig. 14. The points plotted in (a) indicate the actual accuracies of CNN models trained using time-series datasets with time steps of 2, 10, and 20. The solid lines show the corresponding outputs of the service quality function, Eq. (4). The quality-related coefficients of the 20 time-step dataset are $\alpha_1 : 0.0028$, $\alpha_2 : 0.0067$, and $\Delta : 0.9936$. Figure (b) shows that as the service quality increases, the optimal cost and iteration number also increase. When the service quality is 99.36%, the model training optimal cost is 0.772 and the iteration count is 11.

a three-dimensional array consisting of features, time steps, and batches, as shown in Fig. 15. The batch size indicates

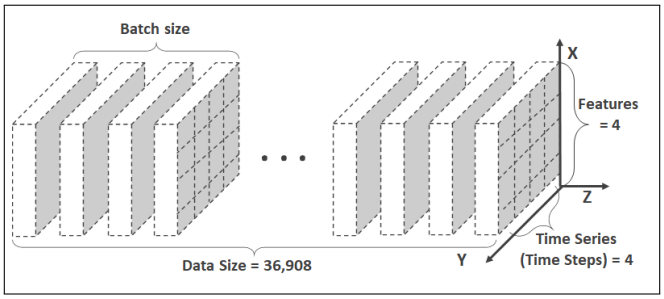


Fig. 15. Input data for deep learning: batch size, time steps, and features.

the amount of two-dimensional input data that each deep learning model consumes before updating its weights.

If the model uses a small batch size, training speed is reduced because larger batch sizes increase the computing time required for backpropagation and weight updating. On the other hand, bigger batch sizes degrade the model's capability to generalize data and require more memory. The batch size thus affects the model accuracy and the training time. The time-step value corresponds to the number of data entries, which are consumed backwards in time by each hidden layer. For illustrative purposes, we use a value of four here, as shown in Fig. 15. Four data entries thus become input data and are used simultaneously; moreover, the input data is converted into a three-dimensional format.

After time-series data transformation, the input data is converted from a one-dimensional format into a two-dimensional dataset that can be used for model training. This is done using a two-dimensional CNN with four convolution, two max-pooling, and four dense layers, as shown in Fig. 16.

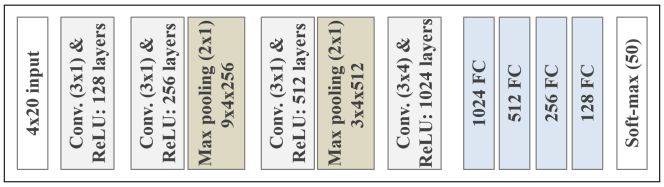


Fig. 16. 2D CNN model architecture for training using the time-series dataset.

The four convolution layers have 128, 256, 512, and 1024 neurons, respectively; the first three have a 3x1 kernel size, while the fourth has a 3x4 kernel. The two max-pooling layers are 2x1, and the four fully connected layers have 1024, 512, 256, and 128. The soft-max provides location predictions based on the time-series transformed sensor datasets. The total number of parameters to train in this case is 8,530,354.

To find the time-step that maximizes model accuracy, we perform time-series transformation of the sensor dataset with time steps between 2 and 30. The data broker feeds these time-series datasets into a CNN and obtains the twenty-eight models shown in Fig.17.

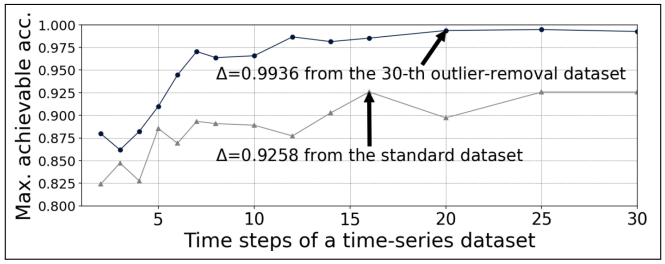
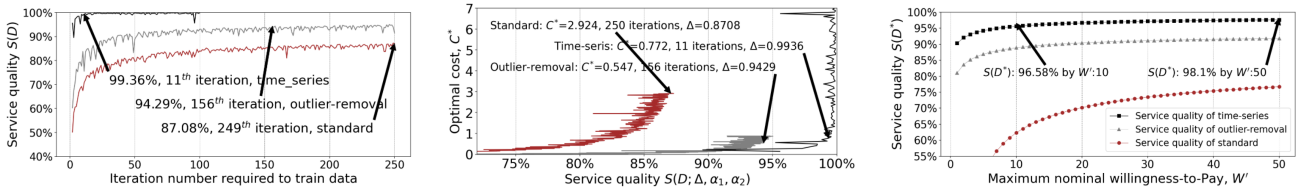


Fig. 17. Maximum achievable accuracy Δ using time-series datasets transformed by the standard dataset and the 30^{th} outlier-removal dataset.

The points plotted in Fig 17 represent the ground-truth model accuracy after training with two different time-series datasets: one obtained by transforming the standard dataset and one by transforming of the 30^{th} outlier-removal dataset. The transformed outlier-removal dataset generally yields higher accuracies, including the best accuracy achieved during this work, 99.36%.

The quality of the time-series dataset clearly depends on the choice of time-step: the greater the number of steps, the greater the accuracy. Fig. 14a shows the accuracy achieved after training with time series having 2, 10, and 20 steps. The twenty time-step dataset delivers the highest accuracy 99.36% with $\alpha_1 : 0.0028$ and $\alpha_2 : 0.0067$. The quality-related coefficients of the two time-step dataset are $\Delta : 0.9425$ (i.e., 94.25%), $\alpha_1 : 0.0731$, and $\alpha_2 : 0.4382$, while those for the ten time-step dataset are $\Delta : 0.9834$, $\alpha_1 : 0.0405$, and $\alpha_2 : 0.1028$.

Fig. 14b shows the optimal cost and iteration count for the CNN model trained with the twenty time-step time-series as the model accuracy increases. When the model

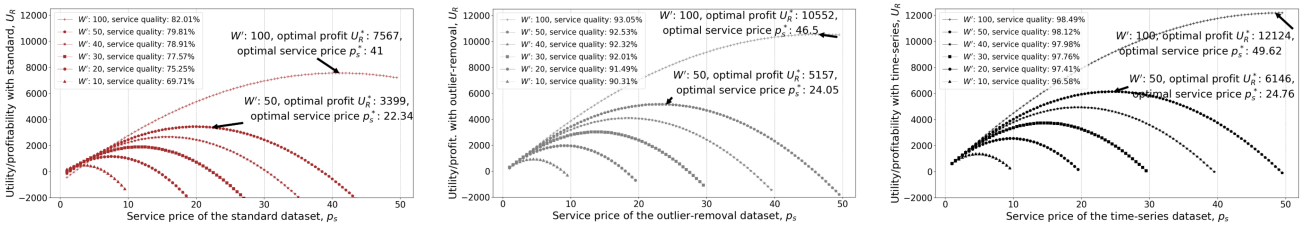


(a) Service quality $S(D)$ of the standard, outlier-removal, and time series datasets as functions of the number of training iterations.

(b) Optimal costs of the three datasets as functions of accuracy.

(c) Service quality of the three datasets as functions of the maximum nominal WTP, W'.

Fig. 18. Service quality and optimal costs of standard, outlier-removal, and time-series datasets.



(a) Profit achieved after training with the standard dataset as a function of the service price p_s and maximum nominal WTP W'.

(b) Profit achieved after training with the outlier-removal dataset as a function of the service price p_s and maximum nominal WTP, W'.

(c) Profit achieved after training with the time-series dataset as a function of the service price p_s and nominal WTP W'.

Fig. 19. Evaluation of the data broker's profit when selling three datasets with different quality levels: the standard, outlier-removal, and time-series datasets. The maximum profit (12124) is achieved with the time series dataset, which has a service price of 49.62 and a service quality of 98.49% when W' is 100.

accuracy reaches its maximum of 99.36%, the optimal cost derived using Eq. (8) is $C^* = 0.772$ and the iteration count is 11. It is worth noting that the twenty time-step dataset requires fewer training iterations than the standard and outlier-removal datasets because it was augmented by grouping data entries into time steps. This allows model training to be performed more quickly and with better accuracy.

The data broker will choose to offer the twenty time-step dataset to prospective consumers due to its superior accuracy. The quality profile of this time-series transformed data is shown in Table 8.

TABLE 8
Service quality profile of a time-series transformed dataset.

Δ	α_1	α_2	C^*
0.9936	0.0028	0.0067	0.772

4.4 Profit evaluation for three different datasets of differing quality

To evaluate the profit obtainable by selling the standard, outlier removal, and time series datasets, the data broker uses the utility function $U_R(p_s^*, B^*, D^*) = p_s^* \cdot M \cdot (S(D^*; \Delta, \alpha_1, \alpha_2) \cdot W' - p_s^*)$ - B^* - $C_{D^*}^{\tau}$ specified by Eq. (13), which has the following parameters: p_s, M, $S(D^*; \Delta, \alpha_1, \alpha_2)$, W', B^*, and $C_{D^*}^{\tau}$:

- $p_s^* = \frac{S(D^*)W'}{2}$ is the service price, which depends on the service quality ($S(D^*)$) and the maximum nominal WTP (i.e., W'). To evaluate the data broker's profit, we let this price be in the range $[0, 100]$.
- M, the number of data consumers, also affects the data broker's revenue. We let $M = 500$.
- W', the maximum nominal WTP, affects the data broker's revenue. We configure $W' = [10, 20, 30, 40, 50, 100]$.
- $S(D^*; \Delta, \alpha_1, \alpha_2)$, the service quality, depends on the optimal data size D^*, the maximum achievable accuracy Δ, and the curve fitting parameters α_1, and α_2. The values of these quantities for the standard, outlier removal, and time series datasets are shown in Tables 5, 7, and 8, respectively.
- D^*, the optimal data size computed using Eq. (11), which depends on B^*, M, γ (which is taken to be 0.7), and the number of data providers (i.e., $N = 15$).
- $C_{D^*}^{\tau}$, the computation cost, which is computed using Eq. (8) and depends on f=300 MHz, k=1, μ=10E-18, and X=433.
- B^*, the data budget, which is computed using Eq. (19) and depends on $M = 500$, $W' = [10, 20, 30, 40, 50, 100]$, D^*, $C_{D^*}^{\tau}$, and the quality-related coefficients Δ, α_1, and α_2.

Fig. 18a illustrates the accuracy of CNN models trained with the standard dataset, the 30^{th} outlier-removal dataset, and the time-series dataset. After training with the time-series dataset, the model accuracy is 11.99% greater than that for the standard dataset. Fig. 18b shows how the optimal cost of each CNN models varies as its accuracy increases. The CNN model trained with time-series data has a higher optimal cost because it has the largest data size, even though it requires fewer training iterations than the other two models. We compare the data broker's profits by considering the quality profiles of the three datasets.

Figs. 18a and b compare the service quality and optimal cost of the standard, outlier-removal, and time-series

datasets. The time-series dataset has the highest accuracy 99.36% and the lowest iteration count 11. The outlier-removal dataset provides a model accuracy of 94.29% and the lowest optimal cost 0.547 due to its small size. The standard dataset provides the lowest accuracy 87.08% and the highest optimal cost 2.924 due to the comparatively long training time.

As the maximum nominal WTP W' increases, the service quality $S(D^*;\Delta)$ converges towards the maximum achievable accuracy Δ for the dataset under consideration. This is because the service quality increases with the size of the dataset D^* in accordance with Eq. (4): $S(D^*) = \Delta \cdot (1 - \frac{\alpha_2}{1+\alpha_1 \cdot D^*})$. Moreover, D^* increases with the budget B^*, in accordance with Eq. (11): $D^* = \frac{B^*(N-1)}{\sum_{n=1}^{N} \gamma_n}$. Consequently, the budget B^* increases with the maximum nominal WTP W' in accordance with Eq. (19).

For example, the service quality for the time-series data is 98.1% when $W' = 50$ is but falls to 96.58% when $W' = 10$, as shown in Fig. 18c. The relationship between consumer payment and service quality is relatively straightforward: higher payments increase the data budget, allowing larger amounts of data to be obtained and thus increasing service quality.

In this context, the service quality is defined by the accuracy achieved using the final refined dataset and is computed using the service utility function, Eq. (4).

After a data broker obtains quality-related coefficients and optimal costs through dataset quality evaluation, the data budget can be determined using Eq. (19). Similarly, the service quality is determined with Eq. (4) and the service price with Eq. (15). Finally, the broker can predict their profits for all datasets using Eq. (13). The profits obtained in the studied case are shown in Table 9. The time-series dataset gives the highest profit (i.e., 6146) with a service quality of 98.1% and an optimal price of 24.76 when the maximum nominal WTP is 50.

TABLE 9
Optimal profitability of three different datasets when $W' = 50$.

Δ	α_1	α_2	C^*	W'	$S(D^*)$	p_s^*	$Profit, U_R$
0.871	0.0148	0.3047	2.924	50	79.8%	22.34	3399
0.943	0.0321	0.547	0.549	50	92.5%	24.05	5157
0.994	0.0028	0.0067	0.772	50	98.1%	24.76	6146

Fig. 19 provides further insight into the relationship between the maximum nominal WTP and the broker's profit. Fundamentally, the greater the service quality and the higher the maximum nominal WTP, the higher the profit. Independently of dataset quality, profitability generally increases with the maximum nominal WTP. However, if the service quality is 87.08% and the maximum nominal WTP is 10, the profit can become negative if the service price is above 7. Using the proposed data broker utility function, a data broker can find optimal prices for datasets of widely varying quality, making it possible to profitably trade in datasets with many different quality levels.

These results also suggest that the best way for the broker to maximize profits and add value is by improving service quality and reducing optimal costs: the time-series data provides the best profit, 6146 and the highest quality, 99.35%, while having a lower optimal cost, 0.772, than the standard data when the maximum nominal WTP is 50, as shown in Fig. 20. It is worth noting that increasing service

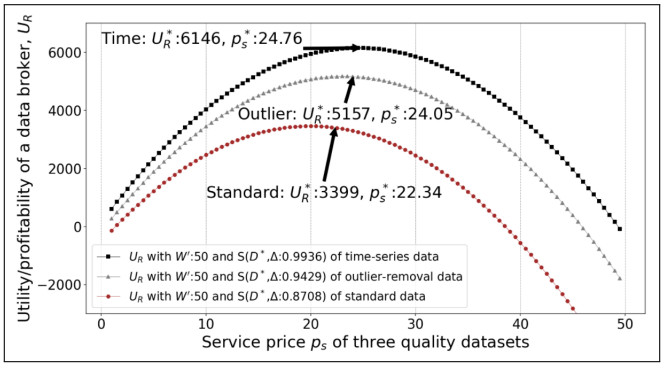


Fig. 20. Profit achieved for the standard, outlier-removal, and time-series datasets with a maximum nominal WTP of 50.

quality by 12.3% (from 87.1% for the standard dataset to 99.4% for the time-series dataset) increases the broker's profit by 2747 (3399 for the standard dataset, and 6146 for the time-series dataset) as shown in Table 9. That is to say, increasing quality by 12.3% almost doubles profit. This strongly suggests that the best way to optimize profit in the data market place is to efficiently refine data.

5 Data Provision

This work demonstrates the potential viability of value-added data brokering services and quality-based data trading for data provisioning. We consider a data broker that generates three datasets of differing quality and evaluates their profitability. Based on this evaluation, the time-series dataset is offered to customers in order to maximize profit, as shown in Fig. 21. It is worth noting that the data broker's

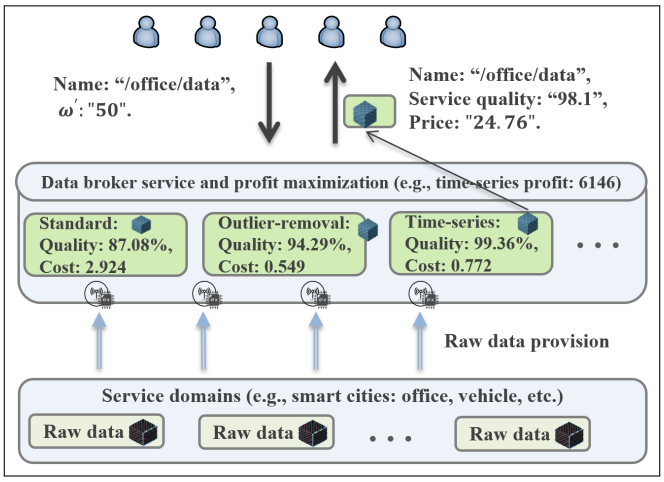


Fig. 21. Quality data discovery and provisioning.

profit is increased by 2747 (rising from 3399 to 6146 when $W' = 50$) by transforming the standard dataset into the time-series dataset. The data broker packets the chosen time-series dataset based on the proposed data model, as shown in Table 10.

TABLE 10
Quality-based data model for data provisioning in a data marketplace.

" name " : " / office / data "	
Type	"sensor data"
Service	"/ smart / office"
Service quality ($S(D^*)$)	"98.1%"
Price (p_s^*)	"24.76"
Quality-related coefficients	{"Δ" : "0.9936"}, {"α_1" : "0.0028"}, {"α_2" : "0.0067"}
Data features	data, time, sequence number, sensor ID, temperature, humidity, light, voltage, **deviation cost**

In our hypothesized data marketplace, data broker services running on cloud computing resources can acquire big data from heterogeneous service domains such as smart cities with various domains including homes, industry, energy, healthcare, travel, transportation, environment, waste management, education, and public safety [35]. After service quality improvement, the data broker can then provide data of varying quality to data consumers based on their quality requirements and willingness to pay.

6 Related Work

The data-as-a-service (DAAS) [36] concept treats data as a valuable good that can be traded in data marketplaces, and Liang *et al.* [37] demonstrated a range of potential data market structures and data pricing models (e.g., monopoly, oligopoly, and intensely competitive markets). Many subsequent studies on this concept have proposed market-based data trading models to maximize the profits of data marketplace stakeholders such as data providers, consumers, and brokers from an economic perspective [28], [29], [38], [39], [40]. These earlier studies focused primarily on the relationship between data providers and consumers when designing market-based data trading models. The novel contribution of this work is that it proposes a data trading model that incorporates both stakeholders and data quality to account for the wide variation in data quality that exists. We additionally demonstrate that increasing the quality of data offered to consumers can significantly increase stakeholders' profits.

Game theory has been used to maximize profits in several studies examining various technologies including Federated Learning [5], [41], [42], [43], [44], Content Delivery Networks [45], [46], [47], and cloud computing [48], [49], [50]. These studies have mainly focused on maximizing profits based on stakeholders' revenues and costs, but none have provided a detailed analysis of optimal pricing based on datasets of differing quality. Therefore, we have studied the relationship between revenues, costs, and data quality, and have demonstrated the optimization of a data broker's profits using data quality as an input variable.

The role of data brokers as intermediaries between data providers and data consumers is becoming increasingly important because these stakeholders will not trade directly if there is no profit for themselves. Hui *et al.* [51] defined the utilities of a data broker (e.g., a roadside unit) and data providers (e.g., a roadside buffer) to optimize the price of data collection services and used a bargaining game to describe the interaction between the data broker and data providers for a sensing service system. On the basis of this analysis, it was demonstrated that a data broker can maximize data service revenues and the profitability of a data service via the brokerage function.

Other studies in this area [5], [28], [39], [52], [53] have similarly defined the primary stakeholders of the data marketplace as data providers, consumers, and brokers. For example, Jang *et al.* [28] used the Stackelberg game model to describe the interaction between service providers (i.e., data broker) and data providers, and demonstrated the optimization of a service provider's profits in an IoT data market. Shen *et al.* [39] also use the Stackelberg game to define the relationship between the data provider, the service provider, and the user, and demonstrated profit optimization for all three participants in the IoT data marketplace. Finally, Ren *et al.* [52] studied the optimization of data purchasing and placement using competitive data trading models featuring a data provider, data consumer, and data broker in the cloud data marketplace.

A central variable in the design of any market-based data trading model is the willingness-to-pay of data consumers. Niyato *et al.* [5] introduced the WTP concept for service consumers to describe trading between a data provider, data consumer, and data broker in an IoT data market, and proposed a model in which the data price is related to the service quality, which in turn depends on the data size. Similarly, Zhang *et al.* [53] developed a model in which an IoT service provider (e.g., data broker) processes and trades data and identified optimal trading decisions for all three players in the IoT market. However, none of these studies analyzed the relationship between the willingness-to-pay of data customers and the quality of the offered data. The results and analysis presented in this work show that the willingness-to-pay affects the quality of the data service, which in turn affects the profits of the market participants. Furthermore, it was demonstrated that the profits of a data broker can be maximized for different levels of willingness to pay.

The WTP scheme proposed herein could be incorporated into a quality discovery protocol by extending an ICN [10], [11], [12]. In this way, the proposed quality discovery packets could enable quality-data discovery across heterogeneous cloud service domains (e.g., smart cities) [13], [14], [15], [16], [17]. This would also allow data consumers to participate directly in data trading by submitting individual specific willingness-to-pay values, unlike other market-based data trading models that have been proposed [5], [28], [29].

7 Conclusion

We extend the ICN concept by modeling an ICN-based marketplace to enable quality-data discovery across heterogeneous cloud service domains (e.g., smart cities) and show that data broker profits can be maximized by embedding a WTP mechanism into an ICN and deploying such a quality-data discovery service on cloud computing resources to acquire and refine big data. Furthermore, embedding a WTP mechanism into an ICN is shown to be a useful way of

allowing data consumers to participate in the data market by indicating their willingness-to-pay and its dependence on data quality. Such a mechanism would be useful in customer-oriented data marketplaces.

The data broker's monetary profit was analyzed as a function of service quality, which in turn depends on data quality, optimal costs, the broker's optimal data budget, and the maximum nominal WTP of customers. The relationship between these factors is described by the proposed data broker utility function. As a result, a data broker can use knowledge of these quantities to identify the most profitable dataset to refine and offer for sale, and find the optimal service price that maximizes profit.

We are currently working to extend this approach to additional data types such as image and structure data across heterogeneous application domains for data trading, and to model quality for ML model trading because we expect that both data and ML models will be traded on future data marketplaces.

ACKNOWLEDGMENT

This work was partly supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by Knut and Alice Wallenberg Foundation.

REFERENCES

[1] B. Krishnamachari, J. Power, S. H. Kim, and C. Shahabi, "I3: An iot marketplace for smart communities," in *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 498–499.

[2] V. Khatri and C. V. Brown, "Designing data governance," *Commun. ACM*, vol. 53, no. 1, pp. 148–152, Jan. 2010.

[3] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," *Future Generation Computer Systems*, vol. 89, pp. 548–562, 2018.

[4] A. Richardson and B. F. Aris Filos-Ratsikas, "Rewarding high-quality data via influence functions," *arXiv*, 2019.

[5] D. Niyato, M. A. Alsheikh, P. Wang, D. I. Kim, and Z. Han, "Market model and optimal pricing scheme of big data and internet of things (iot)," in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

[6] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78–87, oct 2012. [Online]. Available: https://doi.org/10.1145/2347736.2347755

[7] K. Walch, "Data preparation for machine learning still requires humans," Mar 2019. [Online]. Available: https://searchenterpriseai.techtarget.com/feature/Data-preparation-for-machine-learning-still-requires-humans

[8] R. Gibbons, "An introduction to applicable game theory," *Journal of Economic Perspectives*, vol. 11, no. 1, pp. 127–149, March 1997.

[9] A. Rieke, H. Yu, D. T. Robinson, and J. V. Hoboken, "Data brokers in an open society," 2016.

[10] S. Guillén, A. Ståhlbröst, and O. Lenz, "Iot european large-scale pilots: Integration, experimentation and testing: Internet of things intelligence evolution," 2017.

[11] O. Vermesan and J. Bacquet, "Cognitive hyperconnected digital transformation," in *Cognitive Hyperconnected Digital Transformation*. River Publisher, 2017, pp. 1–310.

[12] A. Padyab, A. Habibipour, A. Rizk, and A. Ståhlbröst, "Adoption barriers of iot in large scale pilots," *Information*, vol. 11, no. 1, p. 23, Dec 2019.

[13] R. Ravindran, Xuan Liu, A. Chakraborti, Xinwen Zhang, and Guoqiang Wang, "Towards software defined icn based edge-cloud services," pp. 227–235, 2013.

[14] G. Tyson, N. Sastry, I. Rimac, R. Cuevas, and A. Mauthe, "A survey of mobility in information-centric networks: Challenges and research directions," pp. 1–6, 2012.

[15] S. S. Adhatarao, J. Chen, M. Arumaithurai, X. Fu, and K. K. Ramakrishnan, "Comparison of naming schema in icn," pp. 1–6, 2016.

[16] S. Arshad, M. A. Azam, S. H. Ahmed, and J. Loo, "Towards information-centric networking (icn) naming for internet of things (iot): The case of smart campus," 2017.

[17] Z. Zhang, Y. Yu, A. Afanasyev, J. Burke, and L. Zhang, "Nac: Name-based access control in named data networking," pp. 186–187, 2017.

[18] W. E. Armstrong, "The determinateness of the utility function," *The Economic Journal*, vol. 49, no. 195, pp. 453–467, 1939.

[19] T. Strutz, *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*. Wiesbaden, DEU: Vieweg and Teubner, 2010.

[20] S. Madden, "Intel berkeley research lab data," 2003. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html

[21] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[22] L. Zhang, A. Afanasyev, J. Burke, V. Jacobson, k. claffy, P. Crowley, C. Papadopoulos, L. Wang, and B. Zhang, "Named data networking," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 3, pp. 66–73, Jul. 2014.

[23] Y. Yu, A. Afanasyev, D. Clark, k. claffy, V. Jacobson, and L. Zhang, "Schematizing trust in named data networking," pp. 177–186, 2015.

[24] H. Zhang, Z. Wang, C. Scherb, C. Marxer, J. Burke, L. Zhang, and C. Tschudin, "Sharing mhealth data via named data networking," pp. 142–147, 2016.

[25] J. M. Cantera, "Guidelines for the definition of oasc shared data models," *European Union's Horizon 2020 research and innovation programme under grant agreement no 732240*, Mar 2018.

[26] D. Niyato, X. Lu, P. Wang, D. I. Kim, and Z. Han, "Economics of internet of things (iot): An information market approach," 2015.

[27] Y. Jiao, P. Wang, D. Niyato, M. Abu Alsheikh, and S. Feng, "Profit maximization auction and data management in big data markets," p. 6, 2017.

[28] B. Jang, S. Park, J. Lee, and S.-G. Hahn, "Three hierarchical levels of big-data market model over multiple data sources for internet of things," *IEEE Access*, vol. 6, pp. 31 269–31 280, 2018.

[29] H. Oh, S. Park, G. M. Lee, J. K. Choi, and S. Noh, "Competitive data trading model with privacy valuation for multiple stakeholders in iot data markets," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3623–3639, 2020.

[30] R. W. D. Nickalls, "A new approach to solving the cubic: Cardan's solution revealed," *The Mathematical Gazette*, vol. 77, no. 480, pp. 354–359, 1993.

[31] C. C. Aggarwal, *Outlier Analysis*, 2nd ed. Springer Publishing Company, Incorporated, 2016.

[32] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 12, no. 2, pp. 159–170, 2010.

[33] J. Branch, B. Szymanski, C. Giannella, Ran Wolff, and H. Kargupta, "In-network outlier detection in wireless sensor networks," pp. 51–51, 2006.

[34] H. Ismail, G. Forestier, J. Weber, L. Idoumghar, and P. Muller, "Deep learning for time series classification: a review," pp. 917–963, 2019.

[35] O. Elloumi, J.-P. Desbenoit, P. Wetterwald, G. Karagiannis, and J. Heiles, "High level architecture (hla) : release 3.0 : Aioti wg03 - lot standardisation," p. 40, 2017.

[36] D. Machan, "Daas: The new information goldmine," *Wall Street Journal*, p. 3, Aug. 2009. [Online]. Available: https://www.wsj.com/articles/SB125071202052143965

[37] F. Liang, W. Yu, D. An, Q. Yang, X. Fu, and W. Zhao, "A survey on big data market: Pricing, trading and protection," *IEEE Access*, vol. 6, pp. 15 132–15 154, 2018.

[38] M. R. Galbreth, S. T. March, G. D. Scudder, and M. Shor, "A game-theoretic model of e-marketplace participation growth," *J. Manage. Inf. Syst.*, vol. 22, no. 1, pp. 295–319, jan 2003.

[39] B. Shen, Y. Shen, and W. Ji, "Profit optimization in service-oriented data market: A stackelberg game approach," *Future Generation Computer Systems*, vol. 95, pp. 17–25, 2019.

[40] T.-M. C. Jai and N. J. King, "Privacy versus reward: Do loyalty programs increase consumers' willingness to share personal infor-

mation with third-party advertisers and data brokers?" *Journal of Retailing and Consumer Services*, vol. 28, no. C, pp. 296–303, 2016.
[41] H. Yu, Z. Liu, Y. Liu, T. Chen, M. Cong, X. Weng, D. Niyato, and Q. Yang, "A fairness-aware incentive scheme for federated learning," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 393–399.
[42] K. L. Ng, Z. Chen, Z. Liu, H. Yu, Y. Liu, and Q. Yang, "A multi-player game for studying federated learning incentive schemes," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 5279–5281.
[43] H. Yu, Z. Liu, Y. Liu, T. Chen, M. Cong, X. Weng, D. Niyato, and Q. Yang, "A sustainable incentive scheme for federated learning," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 58–69, 2020.
[44] Y. Liu and J. Wei, "Incentives for federated learning: a hypothesis elicitation approach," 2020.
[45] Z. Xiong, S. Feng, D. Niyato, P. Wang, A. Leshem, and Z. Han, "Joint sponsored and edge caching content service market: A game-theoretic approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1166–1181, 2019.
[46] L. Zhang and D. Wang, "Sponsoring content: Motivation and pitfalls for content service providers," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2014, pp. 577–582.
[47] C. Joe-Wong, S. Ha, and M. Chiang, "Sponsoring mobile data: An economic analysis of the impact on users and content providers," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2015, pp. 1499–1507.
[48] D. T. Nguyen, L. B. Le, and V. K. Bhargava, "A market-based framework for multi-resource allocation in fog computing," *IEEE/ACM Transactions on Networking*, vol. 27, no. 3, pp. 1151–1164, 2019.
[49] D. T. Nguyen, L. B. Le, and V. Bhargava, "Price-based resource allocation for edge computing: A market equilibrium approach," *IEEE Transactions on Cloud Computing*, vol. 9, no. 1, pp. 302–317, 2021.
[50] Z. Xiong, S. Feng, W. Wang, D. Niyato, P. Wang, and Z. Han, "Cloud/fog computing resource management and pricing for blockchain networks," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4585–4600, 2019.
[51] Y. Hui, Z. Su, and S. Guo, "Utility based data computing scheme to provide sensing service in internet of things," *IEEE Transactions on Emerging Topics in Computing*, vol. 7, no. 2, pp. 337–348, 2019.
[52] X. Ren, P. London, J. Ziani, and A. Wierman, "Datum: Managing data purchasing and data placement in a geo-distributed data market," *IEEE/ACM Transactions on Networking*, vol. 26, no. 2, pp. 893–905, 2018.
[53] Y. Zhang, Z. Xiong, D. Niyato, P. Wang, and J. Jin, "Joint optimization of information trading in internet of things (iot) market with externalities," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, 2018, pp. 1–6.

Eunil Seo received a B.S. degree from the Department of Information Engineering at Sungkyunkwan University, an M.S. degree from the Department of Computer Science at the University of Southern California (USC), and a Ph.D. degree from the Department of Electrical and Computer Engineering at Sungkyunkwan University in 1998, 2002, and 2019, respectively. In addition to academic research experience, over 16 years, he had a cross-functional career in the industry as a research staff in Samsung Advanced Institute of Technology (SAIT) regarding the Mobile I.P., IPv6, User-Centered Network, Ad-hoc, and sensor network, a member in Network W.G. of ZigBee Alliance, a chair in RIA WG of OMG, and a project manager for the several control related projects of International Thermonuclear Experimental Reactor (ITER). Currently, he is a post-doc fellow in Computing Science at Umeå University. His research interests are resource-efficient ML training using data quality across distributed data, resource-efficient management in Cloud, and traffic and mobility management in SDN.

Hyoungshick Kim is an associated professor in the Department of Software, Sungkyunkwan University. He received a BS degree from the department of Information Engineering at Sungkyunkwan University, a MS degree from the Department of Computer Science at KAIST, and a Ph.D. degree from the Computer Laboratory at University of Cambridge in 1999, 2001 and 2012, respectively. After completing his PhD, he worked as a post-doctoral fellow in the Department of Electrical and Computer Engineering at the University of British Columbia. He previously worked for Samsung Electronics as a senior engineer from 2004 to 2008. His current research interest is focused on usable security and security engineering.

Bhaskar Krishnamachari received the B.E. degree from The Cooper Union for the Advancement of Science and Art, and the M.S. and Ph.D. degrees from Cornell University, all in electrical engineering. He is currently a Professor of electrical and computer engineering with the Viterbi School of Engineering, University of Southern California. He works on the design and analysis of algorithms, protocols, and applications for the Internet of Things and other distributed systems.

Erik Elmroth is Professor in Computing Science at Umeå University. He has been Head and Deputy Head of the Department of Computing Science for 13 years and deputy director for a national supercomputer center for another 13 years. He has established the Umeå University research on distributed systems, addressing autonomous management systems for virtual computing infrastructures such as clouds and edge environments, see http://www.cloudresearch.org. His experience from management and executive groups in large-scale research projects includes highlights such as the 550 m euro Wallenberg AI, Autonomous Systems and Software Program and the Strategic Research Area eSSENCE. He has also been member of the Swedish Research Councils committee for research infrastructure and chair of its expert panel on eScience as well as Chair of Board of the Swedish National Infrastructure for Computing. He has developed two research strategies for the Nordic Council of Ministers. International experiences include a year at NERSC, Lawrence Berkeley National Laboratory, University of California, Berkeley, and one semester at the Massachusetts Institute of Technology (MIT), Cambridge, MA. Elmroth is a lifetime member of the Swedish Royal Academy of Engineering Sciences and vice chair of its division for Information Technology.