BenchCouncil Transactions on Benchmarks, Standards and Evaluations 4 (2024) 100188

Contents lists available at ScienceDirect

BenchCouncil Transactions on Benchmarks, Standards and Evaluations

journal homepage: www.keaipublishing.com/en/journals/benchcouncil-transactions-on-benchmarks-standards-and-evaluations/

Editorial

Fundamental concepts and methodologies in evaluatology

Jianfeng Zhan *

The International Open Benchmark Council, DE, USA
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Evaluatology
Fundamental concepts
Fundamental methodology
Testbed
Experimental platform
Simulation environment

ABSTRACT

While I have authored three articles introducing Evaluatology, a novel discipline that encompasses the science and engineering of evaluation across various domains, I have struggled to fully depict this challenging yet promising field.

This article delves into the fundamental concepts and methodologies within Evaluatology. I aim to provide a complete picture of evaluation problems in Evaluatology based on my proposed fundamental methodology of understanding a thing. In diverse engineering fields, testbeds, experimental platforms, or simulation environments are commonly utilized to evaluate design or implementation decisions. However, a rigorous methodology is often lacking. I propose a rigorous methodology rooted in Evaluatology for testbeds, experimental platforms, or simulation environments.

1. Why am I drafting this article?

I have drafted three articles to present a new discipline named Evaluatology [1–3], among which I coauthored with my colleagues or students on the first article [1]. However, there are three flaws in the previous work [1–3].

First, I failed to draw a complete picture of evaluation problems in Evaluatology. For example, in the first Evaluatology article [1], I focus on the scenario where we can well define an evaluation condition and emphasize how to construct equivalent evaluation conditions where it is almost impossible to achieve in other scenarios. In Section 3.2, I will formally define what is an evaluation condition. In [3], I discussed other essential scenarios. However, I fail to provide a unified methodology framework for different scenarios.

Second, I fail to propose a generalized methodology for evaluating or understanding (in a much broader sense) a thing in the previous work [1,2]. Later, in [3], I proposed a generalized methodology for understanding a thing, which can provide a solid basis for presenting a complete picture of evaluation problems in Evaluatology.

Third, I fail to propose a generalized methodology to define evaluation conditions. The methodology proposed in [1] is limited and only applied to some scenarios. The above reasons justified my motivation for drafting this article.

2. Fundamental concepts in evaluatology

I reuse the concepts in [3] most of the time. Fig. 1 summarizes the primary entities within Evaluatology.

An individual can be defined as "the object described by a given set of properties" [1,3]. A *system* is "a coherent entity comprising interacting or interdependent individuals and/ or systems, regardless of their likeness or diversity, culminating in a unified whole" [1,3–5].

The evaluation subject [1] (in short, subject) is a *thing* that could be an individual or a system. Typical subjects could be "a life, an artifact, an abstract, or even a policy in natural and social sciences".

A *quantity* "embodies any property of a thing whose instances can be compared by ratio or only by order [1,3,6]". The *truth* is "a thing's facts or inherent properties that can be proven true or verified objectively [3]". A *model* is "a streamlined representation of a thing that would otherwise be too intricate to analyze in exhaustive detail" [1,3,7]". A model can manifest as "a physical, mathematical, or other construct" [1,3,7]. As a special model instance, a causal model is "a causal explanation grounded in a model to understand a thing and infer its behavior" [3,8]. Quantity and truth provide partial insights into a thing, while a model gains a full understanding of a thing in a simplified manner.

* Correspondence to: The International Open Benchmark Council, DE, USA.
E-mail address: jianfengzhan.benchcouncil@gmail.com.
URL: https://www.zhanjianfeng.org.

https://doi.org/10.1016/j.tbench.2025.100188

Available online 23 January 2025

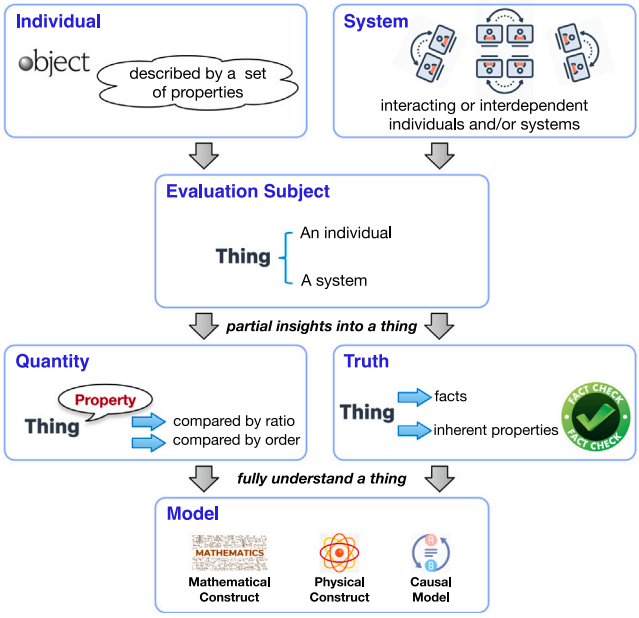

Fig. 1. The primary entities within Evaluatology.

3. Fundamental methodologies in evaluatology

Evaluation is one of the methodologies used to understand a thing. Other methodologies listed in [3] include conjecture, observation, experiment, measurement, and testing.

3.1. The generalized methodology understanding a thing

In my publication [3], I have introduced a generalized methodology for understanding a thing, with the focal point being a concept termed the Self-contained Research System (abbreviated as SRS). As detailed in [3], an SRS must adhere to two key criteria: firstly, it can operate autonomously, and secondly, it should encompass the primary factors that influence the understanding of the thing, known as essential factors. Fig. 2 summarizes the generalized methodology of understanding a thing.

In various contexts, an SRS can be designated differently. For instance, within the realm of evaluation, an SRS may be referred to as a self-contained evaluation system (abbreviated as SES).

To obtain a model of a thing, it is essential to identify and isolate an SRS that is conducive to understanding it. I explained the reason in [3]. If isolating an SRS is unfeasible, external factors may significantly influence understanding a thing. However, once an SRS is identified and isolated, examining the impact of essential factors on the thing becomes viable. The methodology I introduced in [3] is referred to as SRS.

3.2. The relationships among observation, experiment, measurement, testing, and evaluation

If only some of the essential factors are identified within the SRS, I classify it as an observation. Conversely, if all essential factors are identified within the SRS, I classify it as an experiment. The distinction between observation and experiment lies in the presence of hidden or unknown factors that can influence understanding the thing in the former scenario.

Unlike observation and experiments that fully understand the thing, measurements, and testing gain partial insights into a thing by focusing on specific properties, facts, or inherent properties [3]. Measurement is

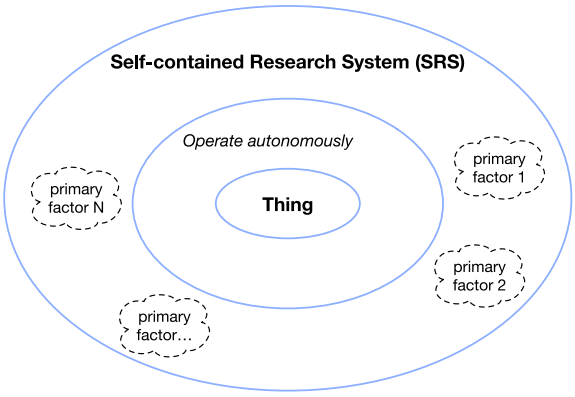

Fig. 2. The Generalized Methodology Understanding a Thing.

"experimentally obtaining one or more values attributed to a quantity of a thing" [3,6]. A test oracle is "a fact or inherent property of a thing and its SRS [3]". Testing is a "verification process to determine whether (1) a thing conforms to the test oracles and/or (2) When a thing operates within an SRS, both the thing and its SRS conform to their test oracles [1,2]".

Building upon the notions of experiment and observation, I elucidate the evaluation concept. If an experiment or observation engages stakeholders, it falls under the evaluation category, whereas those conducted without stakeholder involvement are classified as natural experiments or observations [3]. As elucidated in the discoveries of [2], evaluation involves "causal inference regarding the impact and value of a subject within an SES customized to fulfill stakeholders' evaluation needs, relying on measurements and/or testing of the SES". Fig. 3 depicts the differences between observation, experiment, and evaluation. Fig. 4 depicts the differences between measurement and testing.

Considered from an alternative angle, the process of evaluating a subject can be depicted as "deliberately imposing an evaluation condition (EC) upon it to establish an SES" [3]. Building on the previous discussion, an EC can be envisioned as the SES from which the subject is removed. We formally delineate an EC as the context that is crucial in guaranteeing independent operation and incorporating the essential factors when applied to the subject.

3.3. Standardized evaluation methodology

As shown in Fig. 5, I summarize a standardized evaluation methodology as follows.

The initial step involves defining and characterizing the thing slated for evaluation, which constitutes the subject. Often, evaluations aim to compare different subjects. Without a clear definition and characterization, it is challenging to ensure that the subjects under investigation can be classified into one category and compared. An integral aspect of this phase is modeling the thing, which includes delineating its structure. In many instances, stakeholders harbor specific requirements for evaluating a component of the thing, such as a branch predictor module within the CPU. Failure to formally define the thing's structure and establish a consensus renders evaluating a designated component futile. Divergent stakeholder perspectives on structures or differing functionalities assigned to the same structure can obfuscate the evaluation process.

The subsequent step involves defining the SES for the specified thing. As per the SES definition, it must fulfill two criteria. Firstly, an SES should have autonomous functionality. Secondly, an SES should encompass the essential factors. Constructing an SES is a complex endeavor that necessitates a trial-and-error approach to attain an optimal or viable SES.

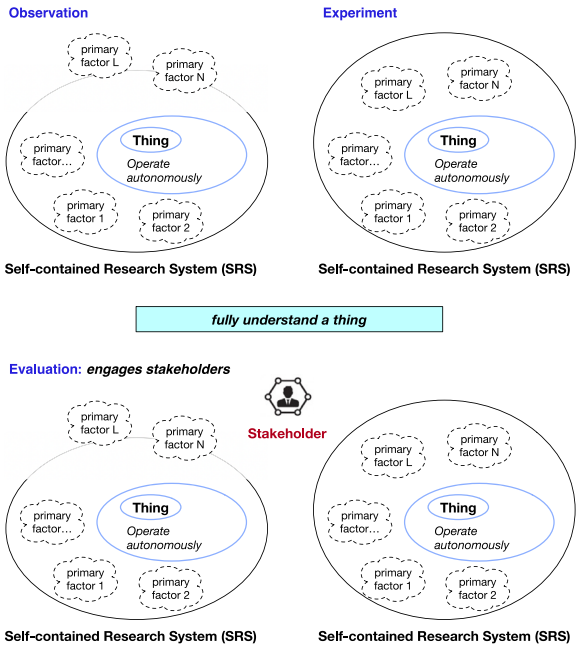

Fig. 3. The Relationships Among Observation, Experiment, and Evaluation.

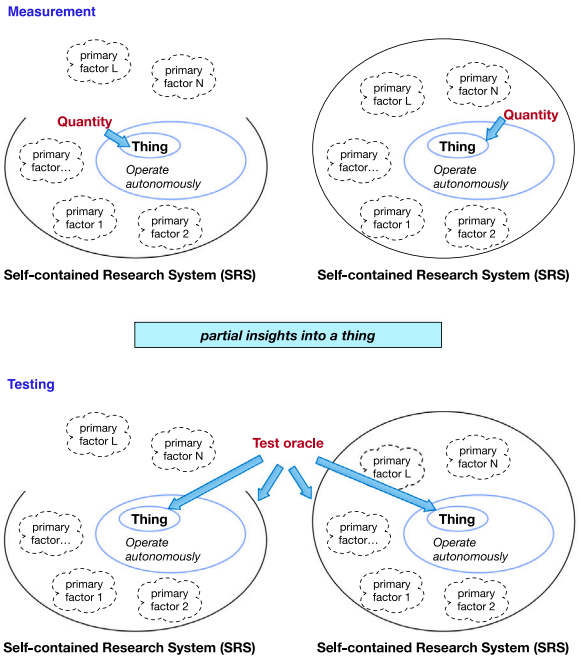

Fig. 4. The Relationship Between Measurement and Testing.

The third step involves acquiring the EC. This can be accomplished by removing the subject under examination from the SES.

Following the establishment of an SES, the fourth step entails analyzing its nature and determining the appropriate evaluation methodologies, which are open issues worth investigating.

Promising evaluation approaches include establishing equivalent EC for the known, well-defined SES [1,9], the causal model approach [8] and the statistical approaches [10].

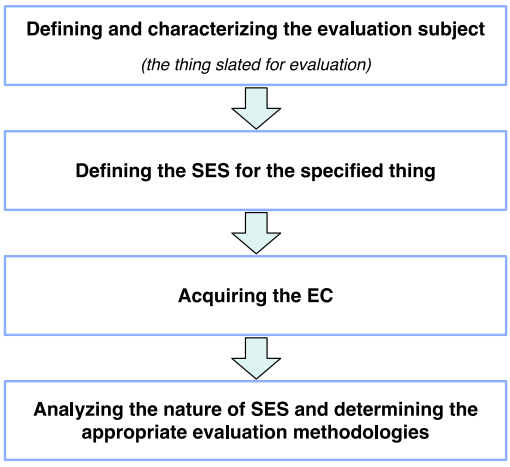

Fig. 5. Standardized Evaluation Methodology.

3.4. The full picture of evaluation problems in evaluatology

My previous research has delved into the diverse natures of SESes [3]. I have slightly adjusted my presentations to demonstrate the full picture of evaluation problems in Evaluatology.

As shown in Fig. 6, the diverse nature of an SES presents various challenges to evaluation [3]. The first kind is when an SES is unknown, e.g., in the case of investigating parallel universal or soul. The second kind is when an SES is only partially known, e.g., when investigating a thing in cosmology and astronomy. The third kind is when an SES is known.

In the third case, there are three different sub-categories. The first subcategory is when an SES is very complex and cannot be well-defined. "When asserting that something is not well-defined, it implies that its structure and functions remain incompletely comprehended. For example, the human body is not well-defined". [3]. The second subcategory is when an SES is known and well-defined but not subject to arbitrary manipulation for different reasons, such as realization limitations, unaffordable costs, unaffordable consequences, or ethical reasons. "If a system can be modeled in a function, arbitrary manipulation entails setting its independent variables to any arbitrary number within its domain. [3]". The third subcategory is when an SES is known, well-defined, and subject to arbitrary manipulation. For example, a computer nearly falls into this category [3].

4. Fundamental methodologies in testbed

Testbeds, experimental platforms, or simulation environments are widely used in different engineering fields to evaluate or test design or implementation decisions. However, they lack a rigorous methodology. In the rest of this article, I use the testbed concept to refer to those systems. When I use the concept of a testbed, it depicts a system that can vary diverse ECs to evaluate design or implementation decisions.

In [1], I introduced a universal evaluation methodology for complex scenarios in collaboration with my colleagues and students. Initially, I outlined the methodology proposed in [1]. Subsequently, I will highlight its limitations. Finally, I will introduce a rigorous methodology built upon my proposed approach in [1].

Fig. 7 illustrates the original universal evaluation methodology in complex scenarios. I previously referred to the complete set of real-world systems utilized for assessing specific subjects as the real-world evaluation system (ES). Nevertheless, the notion of a real-world ES is rather vague. In its place, we have formally articulated the self-contained evaluation system (SES). I have adopted the concept of SES to supplant the real-world ES. Fig. 8 presents the upgraded universal

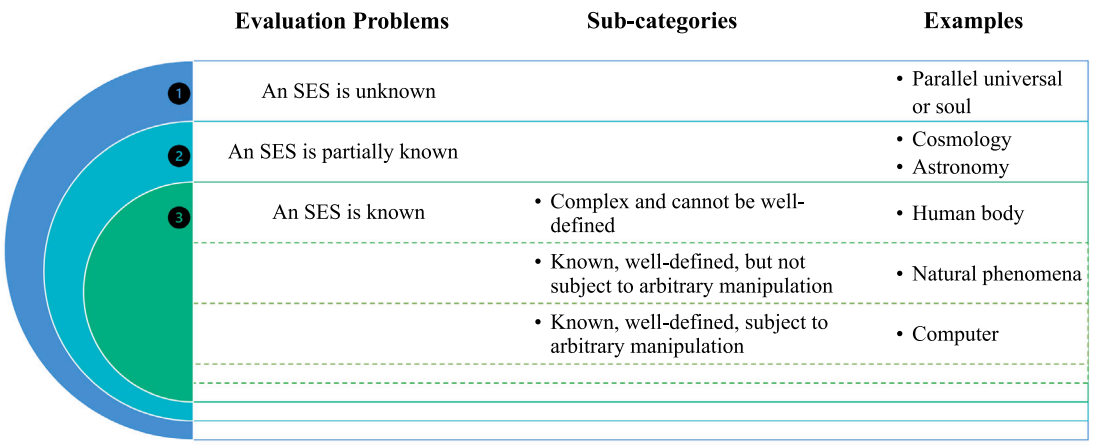

Fig. 6. The Full Picture of Evaluation Problems in Evaluatology.

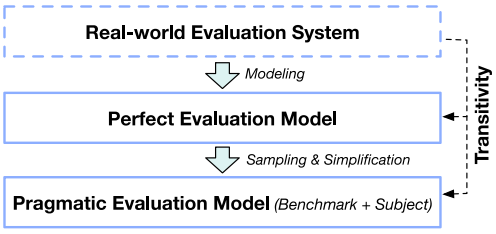

Fig. 7. The original universal evaluation methodology in complex scenarios [1], with the permission of the authors.

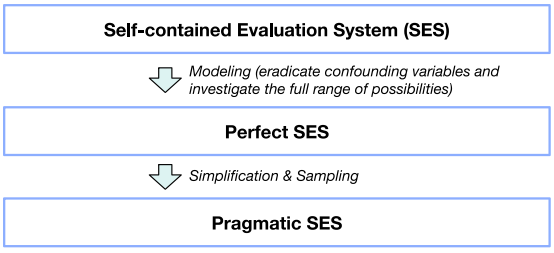

Fig. 8. The upgraded universal evaluation methodology in complex scenarios, based on the concept of the Self-contained Evaluation System (SES).

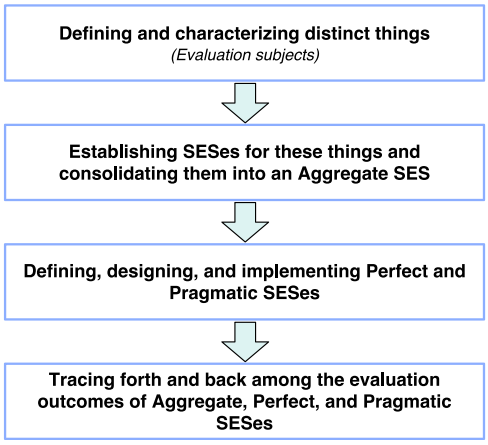

Fig. 9. A standardized evaluation methodology for a testbed.

evaluation methodology in complex scenario, based on the concept of SES.

An SES encounters various hurdles like that of a real-world ES outlined in [1]. Initially, dealing with confounding variables within an SES presents a significant obstacle. Completely eradicating these confounding variables is frequently challenging, if not unattainable.

Furthermore, manipulating the SES proves to be daunting, rendering the establishment of controlled environments for subject evaluation nearly impossible.

Moreover, it is crucial to recognize that SES, irrespective of its characteristics, tends to exhibit a predisposition towards specific groupings. Instead, it should be subject to arbitrary manipulation.

I propose the concept of a Perfect SES that could replicate the SES with the highest level of fidelity. In theory, a Perfect SES would possess three characteristics that enhance the evaluation of subjects.

First, a Perfect SES would streamline manipulation, enabling a free setting of diverse EC. This adaptability would empower researchers to delve into multiple scenarios and appraise subjects under varying conditions, enriching the evaluation process in both depth and breadth.

Secondly, a Perfect SES could effectively eradicate confounding variables. By isolating and controlling variables of interest, researchers could attain more precise insights into how specific variables influence the subjects under scrutiny.

Moreover, a Perfect SES would possess the capacity to comprehensively investigate and grasp the full range of possibilities within an SES.

The nature of the Perfect SES, which involves accommodating extensive populations of ECs along with numerous independent variables, may result in substantial evaluation costs. Yet, to tackle this issue, it is crucial to introduce a Pragmatic SES that streamlines the Perfect SES through two essential modifications.

To reduce evaluation costs associated with multiple independent variables, it is crucial to identify and focus on those variables that substantially impact evaluation outcomes. By identifying and ranking these crucial variables, researchers can streamline evaluations, utilize resources more effectively, and exclude or regulate insignificant variables, reducing complexity and costs. It is important to note that simplification in creating a Pragmatic SES will likely decrease its accuracy.

Moreover, employing sampling techniques can efficiently handle extensive populations of ECs. Instead of assessing every possible scenario, researchers can choose representative samples that encompass the population's diversity and breadth. This method ensures a more manageable evaluation process while maintaining adequate coverage and representation.

In its essence, a testbed functions as a pragmatic SES, offering support for evaluating various categories of things. As shown in Fig. 9, I propose a standardized evaluation methodology for a testbed as

outlined below:

The initial phase involves defining and characterizing distinct things (subjects). Next, it entails establishing SESes for these things and consolidating them into an Aggregate SES. Subsequently, the process includes defining, designing, and implementing Perfect and Pragmatic SESes. Finally, it necessitates tracing forth and back among the evaluation outcomes of Aggregate, Perfect, and Pragmatic SESes.

Acknowledgments

I am very grateful to Dr. Wanling Gao for preparing all the figures and to Dr. Lei Wang and Dr. Wanling Gao for the discussions.

References

[1] J. Zhan, L. Wang, W. Gao, H. Li, C. Wang, Y. Huang, Y. Li, Z. Yang, G. Kang, C. Luo, H. Ye, S. Dai, Z. Zhang, Evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. 4 (1) (2024) 100162.
[2] J. Zhan, A short summary of evaluatology: The science and engineering of evaluation, BenchCouncil Trans. Benchmarks, Stand. Eval. (2024) 100175.
[3] J. Zhan, Five axioms of things, BenchCouncil Trans. Benchmarks, Stand. Eval. (2024) 100184.
[4] System. https://www.merriam-webster.com/dictionary/system, Accessed: February 6, 2024.
[5] A. Backlund, The definition of system, Kybernetes 29 (4) (2000) 444–451.
[6] I. BiPM, I. IFCC, I. IUPAC, O. ISO, The international vocabulary of metrology—basic and general concepts and associated terms (VIM), JCGM 200 (2012) 2012.
[7] H.D. Young, R.A. Freedman, L.A. Ford, University Physics with Modern Physics, 2020.
[8] J. Pearl, D. Mackenzie, The Book of Why: the New Science of Cause and Effect, Basic books, 2018.
[9] C. Wang, L. Wang, W. Gao, Y. Yang, Y. Zhou, J. Zhan, Achieving consistent and comparable CPU evaluation outcomes, 2024, arXiv preprint arXiv:2411.08494.
[10] G.W. Imbens, D.B. Rubin, Causal Inference in Statistics, Social, and Biomedical Sciences, Cambridge University Press, 2015.

Dr. Jianfeng Zhan is a Full Professor at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), and University of Chinese Academy of Sciences (UCAS), the director of the Research Center for Distributed Systems, ICT, CAS. He received his B.E. in Civil Engineering and M.Sc. in Solid Mechanics from Southwest Jiaotong University in 1996 and 1999 and his Ph.D. in Computer Software and Theory from the Institute of Software, CAS, and UCAS in 2002. His research areas focus on evaluatology, evaluatology-based design automation, and optimization automation. His exceptional expertise is exemplified by his introduction to the discipline of evaluatology, an endeavor that encompasses the science and engineering of evaluation; within this discipline, his proposition of a universal framework for evaluation encompasses essential concepts, terminologies, theories, and methodologies for application across various disciplines. He has made substantial and effective efforts to transfer his academic research into advanced technology to impact general-purpose production systems. Several technical innovations and research results, including 35 patents from his team, have been adopted in benchmarks, operating systems, and cluster and cloud system software with direct contributions to advancing parallel and distributed systems in China or worldwide. Over the past two decades, he has supervised over ninety graduate students, post-doctors, and engineers. Dr. Jianfeng Zhan is the founder and chairman of BenchCouncil. He also holds the role of Co-EIC of BenchCouncil Transactions on Benchmark, Standards, and Evaluations alongside Prof. Tony Hey. He has been honored with several prestigious awards for his exceptional contributions. These include the second-class Chinese National Technology Promotion Prize in 2006, the Distinguished Achievement Award of the Chinese Academy of Sciences in 2005, the IISWC Best Paper Award in 2013, and the Test of Time Paper Award from the Journal of Frontier of Computer Science.