Smart Data Research Review: Conceptual Analysis, Value Orientation, Key Technologies, and Application Framework (Postprint)

Authors: Zhang Yunzhong, Liu Jialin

Date: 2023-04-01T16:02:49+00:00

Abstract

[Purpose/Significance] Smart data is a new concept in the field of data science under the background of "Smart Earth," and currently both its theoretical exploration and practical applications are developing rapidly. Reviewing the evolution of academic understanding, building consensus, and distinguishing differences are of great significance for clarifying the theoretical system of smart data and promoting the development of smart data applications. [Methods/Process] Based on extensive and in-depth reading of relevant domestic and foreign literature in the field, the research trajectory of smart data is organized into four facets: conceptual connotation, value orientation, key technologies, and application framework. Through comparison and analysis, we summarize three conceptual perspectives, five characteristic features, five value orientations, three clusters of key technologies, and five major application areas of smart data. [Results/Conclusions] The study finds that the essence of smart data lies in achieving data standardization, precision, and value addition through data evolution or structural design. Its value orientation exhibits diversified and composite characteristics. Its technical system aims to provide hierarchical evolutionary support for "computable-understandable-conversable." The core of its application framework lies in precisely achieving intelligent interaction between "data" and "users." The future theoretical system of smart data needs to be continuously improved under the broad "data science" perspective, focusing on theoretical system construction, data rights governance, balanced technology development, service level deepening, and integration of theory and practice.

Full Text

2. Conceptual Discrimination of Smart Data

2.1 The Concept of Smart Data

To date, no unified concept or conceptual system for Smart Data has been established. Current scholarly definitions of Smart Data can be broadly categorized into three perspectives: value perspective, structure perspective, and process perspective. We have compiled relevant definitions from these different viewpoints in Table 1 . From our perspective, the value perspective treats Smart Data as a blueprint that emphasizes the hierarchical structure of data value but fails to provide concrete implementation schemes. The structure perspective focuses on achieving the goal in one step, stressing top-level refinement of data frameworks and models. The process perspective emphasizes stepwise evolution, highlighting bottom-up progressive extraction to create new value. These three perspectives each have their own focus but converge on data standardization, precision, and value addition. Notably, the three-stage evolutionary view of "digitalization—datafication—intelligentization" under the process perspective has been widely accepted at present: digitalization creates computable structures for utilization and storage; datafication highlights understandability, advancing through semantic enrichment and visualization to expand intelligent applications; and intelligentization presents a conversational form, expressing concrete value and focusing on the manifestation of dynamic intelligence [13].

2.2 Characteristics of Smart Data

Current academic discussions on the characteristics of Smart Data remain unsystematic, with only a few scholars explicitly proposing three key characteristics that Smart Data must possess in the business domain: accuracy, operability, and agility [14]. Building upon this foundation, we have examined Smart Data definitions across different fields and dimensions, identifying five key characteristics: integrability, precision, practicality, value-added nature, and decision-orientation.

Integrability refers to the fusion of data sources, structures, and adaptability. In terms of sources, Smart Data must integrate five major domains: spatial data collection, social data, logistics data, office data, and personal terminal data [19]. Structurally, it transforms chaotic, invisible, unassociated, and low-utilization data into well-organized, manageable, discoverable, interrelated, and reusable forms [5]. Regarding adaptability, Smart Data must align with the specific characteristics of particular domains, themes, and industries [16].

Precision encompasses both accurate content description and deterministic forms of data self-protection. The ideal state of Smart Data involves possessing high-quality, reliable, and accurate data [17] with sufficient precision to drive value. Simultaneously, it requires adequate privacy protection and data rights demarcation, such as converting from plain digital formats to non-plaintext

digital strings to avoid potential legal and technical issues, enabling lawful and reasonable data access and use [18].

Practicality represents the combination of operability and agility. Operability requires data to support scalable operations to maximize decision-making objectives across platforms [14]. Agility demands that data be real-time available, ready for immediate use, and flexibly adaptable to changing environmental requirements [14]. Smart Data meeting these criteria is more practical than other data forms, offering the superior characteristic of being "available on demand and ready for immediate use."

Value-added nature refers to enhanced data value density, requiring both high efficiency and high utility. High efficiency demands the fusion of data structure and data adaptability. High utility requires superior semantic expression, using more refined semantic interpretations to assist in achieving decision-making objectives [20]. Data value addition is essentially a stepwise analytical process oriented toward semantic and contextual Smart Data, with its stage-wise knowledge outputs forming the cornerstone of the knowledge pyramid [21].

Decision-orientation involves the coordination of foresight, systematicness, and evaluability. Foresight means transcending data appearances and existing knowledge levels to anticipate and identify signs of potential impacts [22]. Systematicness refers to taking optimal actions by combining resource allocation with system information acquisition in complex and uncertain environments [23]. Evaluability includes analyzing decision conditions and judging post-decision value, aiming to make accurate assessments of highly dynamic and time-sensitive data [24].

2.3 Clarification of Related Concepts

Smart Data vs. Data Wisdom: The core debate centers on whether the wisdom is human or data-driven. Data itself cannot acquire consciousness, and attempting to make data recognize human consciousness and wisdom would be a strategic error [25]. True wisdom results from human-computer interaction, with human expertise and innovative thinking being the internal drivers of wisdom generation [26].

Smart Data vs. Big Data: The primary differences lie in data volume and form. Big Data, after preprocessing to enable analysis, processing, interpretation, and structured access [27], can be transformed into Smart Big Data. In the future, Big Data will inevitably develop toward intelligence, and Smart Data will inevitably develop toward large-scale datafication.

Smart Data vs. Domain Data and Scientific Data: Smart Data governs data across all domains, building bridges between cross-domain datasets [28]. Scientific data represents a specific category of domain data [16], with some research data possessing the rudimentary form of Smart Data, including basic standardization and quality control.

3. Value Orientation of Smart Data

Smart Data is value-centric, processing raw data to obtain valuable information, particularly insights relevant to decision-making [29]. The value orientation of Smart Data aims to identify its fundamental value positions and attitudes, which we categorize into management value, economic value, innovation value, cultural-educational value, and legal value.

The management value of Smart Data aims to elevate Smart Data services from "doing things right" to "doing the right things." On one hand, M. S. Javan and M. K. Akbari argue from a data resource allocation perspective that Smart Data provides a data fusion abstraction framework for data collection and conversion operations [30], enabling cross-platform observation and manipulation of multiple valuable datasets to provide more comprehensive decision-making visions. On the other hand, J. Chen and colleagues, from a data resource identification perspective, contend that Smart Data integrates real-time interactive functions such as targeted evaluation, reliability analysis, and automatic optimal strategy construction [31], providing decision support for intelligent services that fully perceive the world, dynamically understand events, and enable adequate human-computer interaction [32].

The economic value of Smart Data refers to excavating economic value buried in massive datasets [33], identifying and eliminating non-value-added segments in data processing to create profits. In product manufacturing, S. Belkadi notes that Smart Data revitalizes lean production concepts [34], optimizing production decisions with refined critical information to ultimately achieve cost-effective revenue generation. In after-sales service, UK government research on Smart Data's role in consumer markets found that Smart Data technologies enable remote collection, analysis, and feedback of consumer interest metrics, identifying competitive advantages based on customer needs to promote healthy product market development [35].

The innovation value of Smart Data lies in its ability to maximally activate data value and knowledge potential, drive information technology innovation, and enhance people's insight and decision-making power for innovation and creation, moving from "known-unknown" to "unknown-unknown" [27]. R. Kitchin believes that advanced resource models can meet the extremely high demands of scientific innovation for data validity, integrity, and dynamic change [36]. Zeng Lei points out that Smart Data enhances data expression capabilities through linked data, context awareness, knowledge discovery, and scenario reconstruction [27]. Liu Wei, in organizing the digital humanities technology system, notes that intelligent services will accelerate the design and discovery of potential research projects [37].

The cultural-educational value of Smart Data combines humanistic value in smart culture with educational value in smart education. For cultural resources, Zeng Lei observes that Smart Data methods have long been recognized and applied in the transformation of Smart Data resources in digital humanities,

aiming to meet the cultural service demands of "Internet + cultural scenarios" [27]. For education and teaching, Luo Lin proposes a "data-person-knowledge" knowledge flow framework, wherein Smart Data serves as knowledge yeast that, through effective knowledge conversion and utilization, can intelligently respond to human knowledge needs and become a valuable knowledge integration entity for human use [38].

The legal value of Smart Data is human-centric, controlling data privacy boundaries as a security safeguard. A. Cavoukian proposes a user-centered Smart Data privacy framework that regulates data access and use operations through principles of finality, legality, limitation, transparency, security, and accountability [8]. Meanwhile, D. Roman notes that Smart Data applies blockchain, artificial intelligence, and other technologies to protect sensitive data, proactively incorporating privacy and security into data self-protection [39], achieving both transparent data use and traceability, which is key to the lawful and reasonable use of Smart Data.

4. Key Technologies of Smart Data

4.1 Surveying Smart Data Key Technologies Through Conference Research Topics

What key technologies does Smart Data encompass? To date, neither theoretical circles nor industry have provided authoritative interpretations. We contend that Smart Data technology represents a collection of several key technology clusters, including digitalization, datafication, and intelligentization, whose essence is an evolving technical system. This study attempts to identify core Smart Data technologies based on topics from six IEEE Smart Data Conferences held between 2015 and 2020. By comparing technologies across the three stages of Smart Data evolution—digitalization (computable), datafication (understandable), and intelligentization (conversational)—we explore similarities and differences in key supporting technologies at different stages, as shown in Table 3 .

4.2 Analysis of Key Technology Clusters in Smart Data Evolution Paths

Digitalization aims to transform data into computable formats, completing the mapping from the real world to the digital world and forming the data foundation of Smart Data. The challenge of digitalization lies in integrating dispersed data and fusing multi-source heterogeneous data under the premise of data security [40], completing integrated data management and consolidating Smart Data infrastructure construction. Key technologies at this stage include: (1) Data management technologies: data collection and transmission using IoT, sampling, scanning, and web crawlers; data conversion and integration using ETL tools (Kettle, Talend, Apatar) or NoETL tools (Athena); data storage using distributed file/database systems and NewSQL/NoSQL databases (e.g.,

Neo4j, JSON-LD); and data query and indexing using SQL and SPARQL. (2) Data security technologies: identity verification, iris recognition, and authentication protocols; decentralized technologies like blockchain and smart contracts. (3) Data infrastructure: open-source Smart Big Data systems (Hadoop, Spark, Flink, Storm) and new computing modes such as edge computing, cloud computing, parallel computing, and stream computing. Recent advances in digitalization key technologies focus primarily on data management, with development shifting from "data storage" to "data security." While NoSQL and NewSQL systems have replaced relational databases for efficient, scalable analysis of unstructured data, their evolution from key-value to graph structures has insufficient security considerations, prompting in-depth research on security threats to distributed storage and processing [42-43].

Datafication aims to transform data into semantic formats, enhancing the understandability of digital content and revealing its rich connotations. The challenge of datafication lies in avoiding difficulties in data analysis, understanding, and visualization caused by data heterogeneity, semantic absence, and hidden knowledge. Key technologies include: (1) Knowledge discovery technologies: using data mining, machine learning, and deep learning for automatic analysis to quickly uncover hidden relationships in multi-granularity data. (2) Semantic technologies: the Semantic Web [44] and ontologies are core to current semantic technologies, showing specialized development trends and spawning emerging semantic technologies like linked data and knowledge graphs, achieving a shift in knowledge representation from "string" to "thing" and laying the foundation for Smart Data "understandability." (3) Visualization technologies: including specialized digital profiling techniques that integrate static and dynamic feature data to visualize user or institutional characteristics, and general visualization methods using VR/AR for multi-dimensional perception and virtual interaction [45] or GPS/GIS for 3D data visualization. Recent advances in datafication key technologies concentrate on human-computer interactive visualization, with semantic annotation corpora expanding from specific domains to social network science, focusing on image understanding, annotation, and retrieval [46-47]. Visualization technologies are evolving from graphics fundamentals to interactive visual design, shifting from data presentation to data exploration [48,49].

Intelligentization aims to form conversational human-computer interaction patterns. The challenge lies in training machines to simulate human thinking and cognition to complete friendly human-computer interaction and intelligent decision-making, expanding application scenarios. Key technologies include: (1) Cognitive technologies: traditional approaches use genetic algorithms, artificial neural networks, and expert systems to represent complex data distributions and features; emerging technologies 主要指 cognitive computing and computational neuroscience [50], studying cognitive systems and human brain neurons to explain human cognitive behavior and provide "interfaces" for optimizing human-computer interaction. (2) Conversational technologies: natural language processing generates text to express intentions and uses syntactic and semantic analysis to understand true meanings, moving toward cognitive intelligence [51];

intelligent Q&A technologies enable deep reasoning and multi-round interactive Q&A through recommendation systems that recognize social needs and emotional cues [52]. Recent advances in intelligentization key technologies focus on conversational technologies, with human-computer dialogue evolving from symbolic rule-based systems and statistical machine learning to data-driven deep learning systems [53]. Task-oriented dialogue systems are moving toward large-scale commonsense integration, while open-domain non-task-oriented systems are developing toward personalized conversations [54].

5. Smart Data Application Framework

5.1 Smart Data Application Mechanism

Smart Data applications constitute a systematic service framework with Smart Data, Smart technologies, Smart products, and Smart functions as object boundaries. Currently expanded into multiple dimensions including Smart Business, Smart Living and Mobility, Smart Healthcare, Smart Culture and Education, and Smart Science and Innovation, typical applications across these domains show clear field-specific differences in four key questions: "what data resources to rely on, what key technologies to adopt, what data forms to create, and what application scenarios to serve," as illustrated in Figure 1 [Figure 1: see original paper].

5.2 Smart Business

Smart Business represents a Smart Data-driven business model encompassing Smart Finance, Smart Logistics, Smart Commerce, and other economic application scenarios. Smart Business primarily involves information flow data, capital flow data, and logistics data, which are both dispersed across networks and concentrated in data centers and clouds [55]. Application scenarios include: (1) Using blockchain and security authentication technologies to promote digital signatures and smart contracts, enhancing mobile payment security, robustness, and privacy [56] while simplifying financial asset transaction processes. (2) Applying machine learning to analyze financial information, effectively detecting fraud characteristics and predicting financial and business trends to support investment decisions [57]. (3) Employing GPS/GIS visualization to build Smart Logistics supply chain frameworks for rapid identification, location, sorting, and distribution, making supply chains perceivable, visible, and controllable [58]. (4) Utilizing big data prediction to analyze potential service items and implicit marketing thinking in business data markets, increasing profit margins [59].

5.3 Smart Living and Mobility

Smart Living and Mobility focuses on multiple aspects of Smart City construction related to daily life, including Smart Communities, Smart Transportation, and Smart Energy. It requires semantic fine-grained fusion of physical world data, information space data, and human society data across ternary spaces.

Application scenarios include: (1) Using sensor technology to integrate various infrastructures into a supply network [60], promoting interconnectivity among government, enterprises, social organizations, and citizens in community governance [61]. (2) Applying IoT technology to build Smart Decision Systems using ternary space data and reverse monitoring and planning for Smart Living and Mobility applications [62]. (3) Using profiling technology to extract and label community features for community data portraits [63], providing data dashboards for user needs, environmental monitoring, and energy planning. (4) Employing AI technology to deeply integrate automation and mechanization for Smart Technology Centers [64], managing security, privacy, and risks for various information flows in Smart Living and Mobility [65].

5.4 Smart Healthcare

Smart Healthcare is patient-centered, providing personalized specialized medical management. It requires organizing complex data from medical equipment, treatment plans, and medical records. Application scenarios include: (1) Using IoT to form a medical IoT [66], integrating equipment and team data across hospitals to build shared medical resources. (2) Applying medical data mining through deep learning of hidden relationships in medical literature and clinical treatment data to effectively screen precision medical solutions. (3) Developing cloud-based medical applications using cloud computing, with electronic health records as the core to establish hospital information platforms, breaking spatiotemporal limitations for cloud-based treatment and saving time and economic costs for both patients and providers. (4) Using AI to integrate human-computer interaction service models into various Smart Healthcare scenarios, improving electronic health records and expanding medical knowledge Q&A communities.

5.5 Smart Culture and Education

Smart Culture and Education aims to broaden the semantic expression of cultural/educational institution collections and enhance all-around experiences in cultural/educational spaces, creating favorable environments for users to perceive and learn cultural/educational knowledge [67]. It requires integrating collection data and service data from cultural heritage and library/archive/museum institutions. Application scenarios include: (1) Using IoT to enable Smart Library/Archive/Museum digital collection provision, shortening information distance between collections and users in digital space [68]. (2) Applying semantic technologies to associate explanatory descriptions and multimedia content with collection objects based on collection features and user needs, expanding collection data connotations [69]. (3) Using VR/AR to fuse virtual and real space boundaries, providing interactive experiences of heterogeneous spatiotemporal "dialogue" with cultural heritage and developing virtual culture [70]. (4) Applying profiling technology to create thematically fused, spatiotemporally attributed user dynamic portraits for specific cultural services [71], and in educational contexts, to assist in evaluating teacher-student interaction, analyzing

course effectiveness, and assessing teaching quality to allocate educational resources and build Smart Education environments [72].

5.6 Smart Science and Innovation

Smart Science and Innovation is a cross-learning process of knowledge flow across organizations and domains that enhances Smart Data innovation value through collaborative innovation. It requires broad integration of literature data, research achievement data, and institutional service data. Application scenarios include: (1) Using knowledge graphs to visually display associations and structures among scientific knowledge, discovering patterns and revealing development overviews [73]. (2) Applying natural language processing for machine translation to enable rapid reading and automatic generation of high-quality data reports to reduce manual workload [74]. (3) Using knowledge discovery to mine, cluster, and analyze user behavior and social relationship data to uncover deep information insights and improve knowledge management [75]. (4) Applying intelligent Q&A to upgrade knowledge products like databases and knowledge platforms for intelligence analysis, providing information retrieval and deep Q&A based on user queries to meet personalized retrieval and analysis needs [76].

6. Conclusion and Discussion

This study systematically reviews relevant research achievements in the Smart Data field, examining four fundamental dimensions—conceptual discrimination, value orientation, key technologies, and application framework—while deeply discussing development models and application mechanisms. Six key conclusions are summarized:

(1) The concept of Smart Data can be traced to data hierarchy models, with its core essence lying in achieving data standardization, precision, and value addition through data evolution or structural design. Smart Data possesses integrability, precision, practicality, value-added nature, and decision-orientation. Its "intelligence" results from human-computer interaction, with human expertise and innovative thinking as the internal drivers of wisdom generation.

(2) Smart Data is value-centric, presenting diversified composite value orientations manifested in five dimensions: management value, economic value, innovation value, cultural-educational value, and legal value. The value orientation of Smart Data serves as the prerequisite for its application value and utility.

(3) The Smart Data key technology system is a collection of technology clusters including digitalization, datafication, and intelligentization, essentially providing technical support for the evolutionary path of "computable-understandable-conversational" Smart Data. The technol-

ogy categories and representative solutions are not static but dynamically updated.

(4) Current typical Smart Data applications include but are not limited to Smart Healthcare, Smart Living and Mobility, Smart Business, Smart Culture and Education, and Smart Science and Innovation—all subdivisions under the Smart City umbrella. The Smart Data application framework aims to solve how to precisely achieve intelligent interaction between "data" and "users," addressing core questions of which data sources to collect, how data evolves, what forms Smart Data should take, and what functions and application scenarios to support.

(5) Through the dual integration of theoretical research and practical exploration, Smart Data has gradually formed a development model that deeply presents data asset value: in data structure, from focusing on "stepwise evolution" to emphasizing "top-level design"; in data processing, from back-end "data analysis" to front-end "conversational interaction"; in technical support, from data science "general technologies" to Smart Data "specialized technologies"; in application fields, from "classic domains" to "all industries"; and in application levels, from "smart operation" to "smart decision-making."

(6) The Smart Data application mechanism is shown in Figure 2 [Figure 2: see original paper]: (1) Problems/users are the starting point of value creation, proposing diverse needs including service and product requirements; (2) Scenarios/functions are the transfer stations of value addition, concretizing Smart Product prototypes and core functions while determining required data evolution forms and patterns; (3) Resources are the target objects of value processing, with the processing evolving high-quality, standardized Smart Data through Smart Data technologies to match scenario/function operational needs and satisfy initial problem/user requirements.

However, this review also reveals current research limitations worthy of deeper investigation:

(1) Smart Data theoretical system research under the big "data science" view. As a new concept in data science, Smart Data's theoretical system is still forming and imperfectly constructed. Future research should highlight Smart Data's unique theoretical characteristics while emphasizing its high correlation with Big Data and other data science domains, exploring interdisciplinary integration centered on Smart Data.

(2) Expanding from "value" to "rights" discussions. Current research focuses on value exploration but insufficiently addresses the legality of Smart Data use. Future studies should expand into Smart Data rights governance, examining the generation logic and composition of Smart Data rights and exploring technical and institutional means for rights protection such as data security and desensitization across domains.

(3) Balancing "stepwise evolution" and "normative design" in the Smart Data technology system. Although academia has formed two cognitions of Smart Data "evolving from" or "designed from" processes, the former is more widely accepted, which explains the predominance of "digitalization, datafication, intelligentization" tools. Future Smart Data technology systems need to continuously absorb new technologies and tools from the normative design perspective, enabling "one-step" Smart Data generation and forming a comprehensive technology system with balanced development.

(4) Further upgrading Smart Data application fields. As an applied data science, Smart Data has deepened into business, transportation, healthcare, culture, and science and innovation, but primarily at the smart operation level. Future Smart Data applications should provide precise matching, intelligent interaction, and mutual benefit user experiences, achieving a capability upgrade from smart operation to smart decision-making for comprehensive, cross-platform high-quality Smart applications.

Author Contributions: Zhang Yunzhong: conceptualization, research framework design, methodology, paper revision; Liu Jialin: literature comparison and analysis, drafting and revision.

Abstract: [Purpose/significance] Smart Data is a new concept in the field of data science under the development of "Smart Earth," which theoretical exploration and practical application are developing rapidly. Combing the cognitive veins of the academic circles, gathering consensus and analyzing differences is of great significance to clarifying the theoretical system of Smart Data and promoting its application and development. [Method/process] Based on extensive and in-depth reading of relevant literature in domestic and foreign fields, this study divides Smart Data research into four aspects: conceptual connotation, value orientation, key technologies, and application framework. Through comparison and analysis, it summarizes three conceptual perspectives, five characteristic features, five types of value orientations, three clusters of key technologies, and five application areas of Smart Data. [Result/conclusion] The study finds that the essence of Smart Data lies in achieving data standardization, precision, and value addition through data evolution or structural design. Its value orientation presents diversified composite characteristics. Its technical system aims to provide step-by-step evolution support of "computable-understandable-conversational." The core of its application framework lies in precisely achieving intelligent interaction between "data" and "users." In the future, the Smart Data theoretical system needs continuous improvement under the view of big "data science," focusing on theoretical system construction, data rights governance, balanced technology development, service level deepening, and integration of theory and practice.

Keywords: Smart Data; data science; value; key technology; application; re-

view

Note: Figure translations are in progress. See original paper for figures.