From Raw Materials to Assets: Challenges and Reflections on Data Assetization (Postprint)

Authors: Wu Chao

Date: 2023-03-19T00:00:00+00:00

Abstract

Embedded and wearable devices are proliferating among the general public, and various sensors are now capable of collecting users'sensitive data. The ubiquitous Internet, together with widespread cloud computing and storage facilities, has also made the transmission and management of this data increasingly convenient. Models such as deep learning have begun to fully exploit the value of this data. However, from initially serving as raw material to ultimately becoming a product delivered to users, data must undergo a series of processing and value-added processes, during which economic factors will become the greatest driving force. This article discusses the issue of data capitalization. In this process, to promote the value chain from data to data products, many key economic issues need to be considered, among which core issues include the pricing of data as an asset, as well as privacy protection.

Full Text

From Raw Materials to Assets: Challenges and Reflections on Data Capitalization

Abstract: Embedded and wearable devices are becoming pervasive, with various sensors capable of collecting users' sensitive data. Ubiquitous Internet and widespread cloud computing and storage infrastructure have made transferring and managing this data increasingly convenient, while deep learning models can now fully mine the value within this data. Nevertheless, data must undergo a series of processing and value-added steps from its initial state as raw material to its final delivery as a product to users, with economic factors serving as the primary driving force throughout this process. This article discusses issues in data capitalization, examining the value chain from data to data products. Many key economic considerations must be addressed, with core challenges including data pricing as an asset and privacy protection.

Keywords: data capitalization, data pricing, privacy protection

Author Information: WU Chao, Senior Engineer, Computer Network Information Center, Chinese Academy of Sciences; Research Fellow, Imperial College London. Research interests: data analysis and modeling for smart cities and healthcare. E-mail: chao.wu@imperial.ac.uk

Computing technology and capabilities have become fully pervasive. The observation, integration, analysis, and interpretation of data continuously create new knowledge, driving scientific and technological progress and social development. Embedded and wearable devices are now widespread among the public, with various portable sensors capable of collecting users' sensitive data—for instance, smartphones contain GPS, accelerometers, proximity and light sensors, cameras, gyroscopes, fingerprint sensors, and even heart rate monitors as data collection and perception devices. Ubiquitous Internet and widespread cloud computing and storage infrastructure have also made transferring and managing this collected data increasingly easy. This collected data can be utilized in two ways: (1) establishing statistical models to help public and private sectors understand various aspects of societal operations, such as early detection of epidemics; and (2) providing personalized services at the micro level, such as product and service recommendations for individual users.

Before the emergence of deep networks, machine learning models did not require large amounts of training data; even with more data, models could not be trained to perform better (they entered a saturation state) [1]. In contrast, deep networks are data-hungry—because they are sufficiently deep and have enough parameters to train, the more data available, the deeper the network that can be constructed, and the better its performance. This is the role of big data. Today, this artificial intelligence technology, represented by the combination of big data and deep neural networks, is profoundly influencing all aspects of social life. As a raw material, data can generate new value through processing and mining via data analysis and modeling, becoming a new source of productivity and an asset.

Numerous cases have demonstrated the application value of data [2,3]. However, for a technology to profoundly advance social development, it must evolve from having mere application value to possessing dual value—both application and economic. Viewing big data from an economic perspective, we can see that so-called "data" occupies the starting position in the entire value chain. From its initial state as raw material to its final delivery as a product to users, data undergoes a series of processing and value-added steps, including cleaning [4], semanticization [5], fusion [6], analysis [7], modeling [8], knowledge extraction [9], application [10], and distribution [11]. Like an industrial product moving from raw material to final form to market, this constitutes a complex value chain requiring sophisticated coordination. In most current big data research, the focus remains solely on the technical foundations of these specific processes. We believe that as the entire ecosystem becomes established, economic factors behind each step will become the most significant driving force.

Privacy Protection in Data Capitalization

Privacy protection has become a critical issue in the process of data capitalization. Data ownership and privacy rights have long been core concerns in the information industry [12]. Privacy can be viewed as users' control over the extent and manner of information flow. Traditional privacy protection research has focused primarily on access control and the removal of personal information before data publication, preventing the recovery of removed information after fusion of multiple data sources. However, with the development of big data, mobile collection devices, and machine learning technologies, privacy protection during the data collection phase presents a new challenge. As data becomes increasingly important for building effective models, privacy protection in data collection should exist in a state of trade-off. Solving privacy protection issues cannot be viewed in isolation; rather, it should be placed within a larger framework that balances users' privacy rights with the services and resources obtained from their data, achieving optimality in the current context. Therefore, a privacy protection mechanism that supports multi-party win-win outcomes needs to be established: on one hand, ensuring user privacy is controllable to facilitate data transaction and circulation; on the other hand, promoting the healthy development of data-driven business models and ecosystems.

Data collection, as a key link in developing innovative and personalized, context-aware applications, exists in a "legal gray area" from a privacy perspective. Currently, most applications only indicate their market price without explicit agreements on the scope and granularity of data collection. For example, a navigation software application can continuously collect large amounts of user data in the background without the user' s knowledge. Taking mobile applications as an example, 91% of iOS apps and 83% of Android apps exhibit at least one type of risky behavior that leaks user privacy [13]. Companies such as Facebook, Apple, Twitter, Yelp, and Path have all become the focus of litigation for allegedly releasing privacy-violating mobile applications [14]. Applications (particularly mobile apps) often ambiguously describe data collection information (such as types and quantities). Although data collection is usually mentioned in end-user agreements (as in the Apple App Store), users typically do not read these lengthy documents and simply choose to accept the terms. Moreover, licensing statements in end-user agreements are often vague and misleading, while in practice, large amounts of sensitive user data are collected. Furthermore, privacy protection in data collection is not a binary problem [15], but rather a matter of degree. Although some application stores (such as Google Play Store) provide certain control mechanisms for application access to user data, they still lack support for granularity of data access. While Google Play Store indicates the types of data an app needs to access, it does not specify the quantity and frequency of data collection, which are often critical [16].

A compromise and balance between privacy protection and data utility is required [17-19], and an ecological environment must be constructed at the technical solution level. In this context, governments worldwide have introduced a

series of policies and regulations. For example, Europe' s General Data Protection Regulation (GDPR) was implemented in May 2018. Determann [20] discussed the differences between GDPR and privacy protection norms in other countries. Post [21] analyzed Google' s privacy violation investigation in the EU (Spain), its profound impact, and the subsequent changes in the EU' s legal environment. China' s Cybersecurity Law, officially implemented on June 1, 2017, emphasizes the protection responsibilities and penalty measures that network operators within China must undertake for the personal information they collect.

Data Pricing and Trading in Data Capitalization

To promote the value chain from data to data products, many key economic issues must be considered, with one core problem being the pricing of data as an asset. Data differs from other raw materials in four significant aspects: (1) Data usage does not result in data consumption; data development is not exclusive but can even be altruistic; (2) Aggregated data is more valuable than individual data and should command higher prices; (3) For the same type of data, different sources yield different values, which is particularly prominent in medical data; (4) The same data holds different value for different users. Under these special conditions, pricing data assets is a difficult problem. We believe that adopting a market-negotiated price may be more realistic and feasible.

Pricing alone is insufficient; trading is also necessary. For data assets to generate value, they must be circulated. Early research on data circulation focused on data accessibility and distributed system reliability [31]. However, "information asymmetry"persists throughout data collection and trading: users currently lack awareness of data collection and thus remain in a disadvantaged position. Although some studies propose systemic solutions based on law and transaction frameworks, they lack practical technical implementations. Public surveys we conducted at the Imperial Festival and the UK Digital Economy Conference revealed that most users are unaware of how much data is being collected by applications. Illegal data trading impacts the security of high-value information such as personal data [22], and both buyers and accomplices in illegal data trading should be penalized. Particularly for pricing, traditional utility-based or cost-based pricing models are not applicable [23]. While financial asset pricing theories offer valuable insights, data provided by suppliers is difficult to precisely match with the application directions of demanders, making it impossible to resolve the problem of supply-demand mismatch. Additionally, demanders find it difficult to offer high prices when uncertain whether a data resource can truly generate revenue for their organization. Liu et al. [24] argue that in big data trading, the lack of sufficient historical references makes it difficult to determine transaction prices for data resources, thus proposing a Rubinstein model based on a bidding mechanism for bargaining between trading parties to reach an equilibrium price. Li and Miklau [25] proposed three principles for data market pricing and the basic structure of pricing functions. Valz [26] dynamically

adjusts pricing based on data content. Zhai et al. [27] evaluated the value of big data resources from the perspective of option value of assets, noting that data is constantly changing and updating, and that the non-exclusive nature of data may lead to value depreciation, ultimately constructing an evaluation model to calculate data asset value by synthesizing these factors. Markets help enable reasonable data pricing [28], and Iyilade and Vassileva [29] proposed a privacy-preserving data trading algorithm whose basic idea is to optimize data sharing among applications through market mechanisms. However, these pricing methods share a common problem: significant concerns about security issues and privacy leakage in data trading have left large amounts of data sources unactivated [30]. Although data exhibits clear commodity characteristics, it possesses strong non-traditional commodity attributes, such as near-zero replication costs, non-exclusivity, and time sensitivity. Consequently, despite the establishment of some data exchanges (such as the closed Microsoft Azure DataMarket in 2017), data trading has struggled to achieve scale, and data remains difficult to circulate and realize its value.

Currently, most applications are transitioning from advertising-based business models to models centered on personal data collection. However, under the current data collection paradigm, users cannot receive rewards for their contributed data. While this model may appear to benefit application services on the surface, it actually hinders the sustainable development of their business models when potential legal consequences are considered. The unclear ownership of user data leads to difficulties in effective data circulation.

The PBD Model

We propose a new mobile privacy protection model—the PBD model [32] (Pay-by-Data). PBD explicitly treats data as a means of payment for application effectiveness, establishing an agreement between users and data collectors regarding collection and feedback, thereby achieving reasonable data pricing through privacy protection.

(1) Introduce a Data Pricing Agreement (DPA) between data consumers and data providers. The DPA uses data (privacy) as a pricing instrument, defining a new type of payment method for application services that allows users to trade their data (privacy) to obtain services or other incentives. The DPA details the types of data accessed by applications, the frequency of data collection, and the rewards users receive; it also establishes different pricing mechanisms for different data qualities. Consequently, the collection of micro-level user data is explicitly regulated by the data pricing agreement, reducing arbitrary violations of user privacy.

(2) Improve the communication between applications and underlying mobile services, as well as the methods for requesting user data, through customized platforms such as Android. User data access is controlled by a data pricing authentication service, providing finer-grained support. The

data pricing agreement is implemented on blockchain-based smart contracts, thereby ensuring fair execution and traceability. Additionally, new data access development APIs are provided for application development.

(3) Investigate market mechanisms to find a balance between privacy protection and data collection. Transparent and trustworthy data collection explicitly defines the compensation (i.e., resources and services) corresponding to users' data collection, generating incentives; and thus constructs a data pricing and trading method where data is used as a currency to purchase services and resources provided by applications (including real currency), enabling these applications and users to reach pricing equilibrium through effective market mechanisms.

References

[1] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. Journal of Machine Learning Research, 2010, (9): 249-256.

[2] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity. America: McKinsey Global Institute, 2011.

[3] McAfee A, Brynjolfsson E. Big data: The management revolution. Harvard Business Review, 2012, 90(10): 60-66.

[4] Bifet A, Holmes G, Kirkby R, et al. Data mining: Data stream mining. Encyclopedia of Database Systems, 2009: 855-855.

[5] Auer S, Bizer C, Kobilarov G, et al. DBpedia: A Nucleus for a Web of Open Data. In: The Semantic Web. Heidelberg: Springer Berlin Heidelberg, 2007: 11-15.

[6] Hall D L, Llinas J. An introduction to multisensor data fusion. Proceedings of the IEEE, 2002, 85(1): 6-23.

[7] Trnka A. Big data analysis. European Journal of Science and Theology, 2014, 10(1): 143-148.

[8] Chen H, Chiang R H L, Storey V C. Business intelligence and analytics: from big data to big impact. MIS Quarterly, 2012, 36(4): 1165-1188.

[9] Fan W, Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, 2013, 14(2): 1-5.

[10] Murdoch T B, Detsky A S. The inevitable application of big data to health care. JAMA, 2013, 309(13): 1351-1352.

[11] Viktor M S, Cukier K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. London: John Murray, 2013.

[12] Petrie C. The Proper Use of the Internet: Digital Private Property. IEEE Internet Computing, 2016, 20(2): 92-94.

[13] O' Brien K J. Data-gathering via apps presents a gray legal area. New York Times, [2012-10-29]. https://www.nytimes.com/2012/10/29/technology/mobile-apps-have-a-ravenous-ability-to-collect-personal-data.html.

[14] van Grove J. Your address book is mine: Many iPhone apps take your data. VB Mobile, [2012-02-14]. https://venturebeat.com/2012/02/14/iphone-address-book/.

[15] Xu L, Jiang C, Wang J. Information security in big data: privacy and data mining. IEEE Access, 2014, 2: 1149-1176.

[16] de Montjoye Y A, Hidalgo C A, Verleysen M, et al. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, 2013, 3(6): 1376.

[17] Xu J, Wang W, Pei J, et al. Utility-based anonymization for privacy preservation with less information loss. ACM SIGKDD Explorations Newsletter, 2006, 8(2): 21-30.

[18] Gionis A, Tassa T. k-Anonymization with Minimal Loss of Information. IEEE Transactions on Knowledge & Data Engineering, 2008, 21(2): 206-219.

[19] Ghosh A, Roth A. Selling privacy at auction. In: Proceedings of the 12th ACM conference on Electronic commerce. ACM, 2011: 199-208.

[20] Determann L. Adequacy of Data Protection in the EU - General Data Protection Regulation as Global Benchmark for Privacy Laws? [2017-06-23]. https://papers.ssrn.com/sol3/papers.cfm?abstract_{id}=2902228.

[21] Post R. Data Privacy: The Google Spain Decision and the Right to Be Forgotten. In: Privacy and Identity Management. FIP Advances in ICT. Springer, Cham, 2017: 3-17.

[22] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and l-diversity. IEEE Transactions on Knowledge & Data Engineering, 2013, 26(1): 97-109.

[23] Liu Zhaoyang. Big data pricing problem analysis. Library and Information Knowledge, 2016, (1): 57-64.

[24] Liu Hongyu, Zhang Xiaoyu, Hou Xilin. Research on big data transaction pricing based on bargaining game model. China Metallurgical Education, 2015, (6): 86-91.

[25] Li C, Miklau G. Pricing aggregate queries in a data marketplace. WebDB, 2012.

[26] Valz E. Pricing data: A look at past, present and future. KDNuggets, [2015-08-18]. http://www.kdnuggets.com/2015/08/pricing-data-past-present-future.html.

[27] Zhai Lili, Wang Jiani, He Xiaoyan. Research on data asset evaluation methods for mobile cloud computing alliance enterprises. Price Theory and Practice, 2016, (2): 153-156.

[28] Fricker S A, Maksimov Y V. Pricing of Data Products in Data Marketplaces. In: Software Business. Springer, 2017: 49-66.

[29] Iyilade J, Vassileva J. A framework for privacy-aware user data trading. In: User Modeling, Adaptation, and Personalization. Springer, 2013: 310-317.

[30] Li H, Dai Y, Lin X. Filling the Gap between Data and Applications: A Survey of Data Pricing Mechanisms. IEEE Access, 2017, 5: 13560-13577.

[31] Bernstein P A, DeWitt D, Heuer A. Building an integrated, ubiquitous, semantic fabric of world-wide data. IEEE Data Engineering Bulletin, 2000, 23(3): 3-13.

[32] Wu C, Guo Y. Enhanced user data privacy with pay-by-data model. In: IEEE International Conference on Big Data. IEEE, 2013: 53-57.

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