Data and AI Model Markets: Opportunities for Data and Model Sharing, Discovery, and Integration

Jian Pei
Duke University
Durham, NC, USA
j.pei@duke.edu

Raul Castro Fernandez
The University of Chicago
Chicago, IL, USA
raulcf@uchicago.edu

Xiaohui Yu
York University
Toronto, ON, Canada
xhyu@yorku.ca

ABSTRACT

The markets for data and AI models are rapidly emerging and increasingly significant in the realm and the practices of data science and artificial intelligence. These markets are being studied from diverse perspectives, such as e-commerce, economics, machine learning, and data management. In light of these developments, there is a pressing need to present a comprehensive and forward-looking survey on the subject to the database and data management community. In this tutorial, we aim to provide a comprehensive and interdisciplinary introduction to data and AI model markets. Unlike a few recent surveys and tutorials that concentrate only on the economics aspect, we take a novel perspective and examine data and AI model markets as grand opportunities to address the long-standing problem of data and model sharing, discovery, and integration. We motivate the importance of data and model markets using practical examples, present the current industry landscape of such markets, and explore the modules and options of such markets from multiple dimensions, including assets in the markets (e.g., data versus models), platforms, and participants. Furthermore, we summarize the latest advancements and examine the future directions of data and AI model markets as mechanisms for enabling and facilitating sharing, discovery, and integration.

PVLDB Reference Format:
Jian Pei, Raul Castro Fernandez, and Xiaohui Yu. Data and AI Model Markets: Opportunities for Data and Model Sharing, Discovery, and Integration. PVLDB, 16(12): 3872 - 3873, 2023.
doi:10.14778/3611540.3611573

1 OVERVIEW

The markets for data and AI models are rapidly emerging and increasingly significant in the realm and the practices of data science and artificial intelligence. These markets are being studied from diverse perspectives, such as e-commerce, economics, machine learning, and data management. Recently, data and AI model markets have become a prominent topic of discussion in leading academic venues for data management, such as VLDB, SIGMOD, and ICDE. The premier conferences published more and more papers on the subject, and several workshops dedicated to data markets are scheduled for SIGMOD 2023 and VLDB 2023 (after the success of the VLDB 2022 workshop on data science for data marketplaces). ICDE 2023 had a special session on "marketplaces, economics, and insights". In light of these developments, there is a pressing need to present a comprehensive and forward-looking survey on the subject to the database and data management community. Therefore, we develop this tutorial for VLDB 2023.

In this tutorial, we aim to provide a comprehensive and interdisciplinary introduction to data and AI model markets. Unlike a few recent surveys and tutorials that concentrate only on the economics aspect, we take a novel perspective and examine data and AI model markets as grand opportunities to address the long-standing problem of data and model sharing, discovery, and integration. This is a core theme in data management and data science.

We define **data and AI model markets** as *places and mechanisms that enable multiple parties to share, discover, and integrate data and AI resources and generate added value.* In this tutorial, we motivate the importance of data and model markets using practical examples, present the current industry landscape of such markets, and explore the modules and options of such markets from multiple dimensions, including assets in the markets (e.g., data versus models), platforms, and participants. Furthermore, we summarize the latest advancements and examine the future directions of data and AI model markets as mechanisms for enabling and facilitating sharing, discovery, and integration.

We are organizing a mini symposium on data markets as a half-day workshop at VLDB 2023. This mini symposium aims to explore various topics related to data markets. This tutorial is an excellent complement to the workshop and adds highly visible value to VLDB 2023. The tutorial provides attendees with an opportunity to gain an in-depth and systematic understanding of the background, industry landscape, and research frontiers of the subject matter. We envision the tutorial as a focused session that delves into specific areas related to data and AI model markets, providing attendees with practical experience and express learning opportunities. By combining theory and practice, the tutorial provides a more comprehensive and engaging learning experience for participants.

Our target audience comprises both academic researchers and industry practitioners in the fields of data management, data science, AI and machine learning, and their respective applications. For academic researchers, our tutorial showcases a novel research frontier and offers a range of well-motivated research problems. For industry practitioners, our tutorial introduces the latest techniques and tools that can be useful for business applications. As we structure the materials around the theme of data and AI model sharing, discovery, and integration, attendees of VLDB are expected to enjoy a strong connection to this emerging area from the well-established roots of data management.

Proceedings of the VLDB Endowment, Vol. 16, No. 12 ISSN 2150-8097.
doi:10.14778/3611540.3611573

Our tutorial is self-contained and does not require any specific background from the audience. Instead, we establish a connection between the audience's background in data management and applications to this new area, and use a wide range of examples in data management, databases, and data mining to motivate the need for data and model markets, explain the technical challenges and the existing solutions, and discuss the limitations and future directions. By leveraging these familiar examples, we make the topic accessible to attendees from various backgrounds and foster a better understanding of the subject matter.

2 SCOPE AND STRUCTURE

The outline of our tutorial is as follows.

(1) Why and what are data and AI model markets?
   (a) Motivation
   (b) Definition of data and AI model markets
(2) Decomposition and categorization of data and AI model markets
   (a) Categorization based on assets in the markets, data versus models
   (b) Categorization based on platforms (e.g., data brokers, online data marketplaces, private infrastructures, and open repositorie)
   (c) Categorization based on participants (e.g., contributing agents and consuming agents)
(3) Data and AI model sharing through markets
   (a) Incentivization
   (b) Data sharing mechanisms
   (c) AI model sharing mechanisms
   (d) Privacy and security
   (e) Administration and auditing
(4) Data and AI model discovery through markets
   (a) Data discovery
   (b) AI model discovery
   (c) Data acquisition in data and model markets
(5) Data and AI model integration through markets
   (a) Data assemblage in data markets
   (b) Data integration in data markets
   (c) Model integration in model markets
   (d) Data and model matching and integration in markets
   (e) Valuation of data and models in coalitions
(6) Data and model market systems
   (a) Industry practice
   (b) Considerations and challenges
   (c) Several frameworks and prototypes
(7) Summary and future directions

The tutorial is a **3-hour lecture-style** tutorial. Our tutorial begins with a 15-minute motivational session, introducing the state-of-the-art practices and ideas behind data and model markets in the industry. This initial presentation provides a solid practical foundation, piques the audience's interest, and generates excitement for the upcoming content.

The following 30 minutes focus on establishing the core concepts of data and model markets. Even if an attendee only attends this first 45 minutes of the tutorial, they will have a solid understanding of the basic ideas and concepts of this exciting new area.

The next 1 hour and 45 minutes delve into reviewing the representative and important technical methods and results of data and model markets. Finally, we reserve the last 30 minutes of the tutorial to summarize and reinforce the key takeaways and ideas presented, as well as to brainstorm future directions.

In order to guarantee a superior learning experience, we prioritize interactivity and dialogue between the presenters and the audience during the tutorial. This helps to promote engagement and facilitate a deeper understanding of the concepts being presented.

Additionally, we incorporate ample examples throughout the tutorial to provide motivation and clarity on the concepts and methods being introduced. These examples help to illustrate the practical applications of the material and enable attendees to better internalize the content.

A small subset of the highly relevant references that are covered by the tutorial is as follows: [1–6]. Our tutorial will cover a broader range of related work beyond this list.

3 BIOGRAPHIES

Jian Pei is the Arthur S. Pearce Distinguished Professor at Duke University, whose research focuses on data science, data mining, database systems, machine learning, and information retrieval. With his expertise in developing effective and efficient data analysis techniques for novel data-intensive applications and transferring them to products and business practice, he has been recognized as a Fellow of the Royal Society of Canada, the Canadian Academy of Engineering, ACM, and IEEE. He received several prestigious awards, such as the 2017 ACM SIGKDD Innovation Award, the 2015 ACM SIGKDD Service Award, and the 2014 IEEE ICDM Research Contributions Award. He has previously served as the chair of ACM SIGKDD, the Editor-in-Chief of IEEE TKDE, and a PC co-chair of VLDB 2018.

Raul Castro Fernandez is an assistant professor in the Computer Science Department at the University of Chicago. His research interests are in the economics and value of data, data markets, data flow governance, data discovery, and, more broadly data management and data science. Before the University of Chicago, he was a postdoctoral researcher at MIT, and before then, he completed his Ph.D. at Imperial College London.

Xiaohui Yu is an associate professor at York University. His research interests lie in the broad area of data science, with a particular focus on the intersection of data management and machine learning (ML), including data acquisition for ML in data markets, algorithms and systems for large-scale ML, and ML-enabled query processing. He obtained his Ph.D. from the University of Toronto.

REFERENCES

[1] R. Castro Fernandez. Protecting data markets from strategic buyers. SIGMOD: 1755–1769, 2022.
[2] R. C. Fernandez, P. Subramaniam, and M. J. Franklin. Data market platforms: Trading data assets to solve data problems. PVLDB, 13(12):1933–1947, Sep 2020.
[3] Y. Li, X. Yu, N. Koudas: Data acquisition for improving machine learning models. PVLDB, 14(10): 1832-1844, June 2021.
[4] J. Liu, J. Lou, J. Liu, L. Xiong, J. Pei, and J. Sun. Dealer: An end-to-end model marketplace with differential privacy. PVLDB, 14(6):957–969, April 2021.
[5] X. Luo, J. Pei, Z. Cong, and C. Xu. On Shapley value in data assemblage under independent utility. Proc. VLDB Endow., 15(11):2761–2773, Sept. 2022.
[6] J. Pei. A survey on data pricing: From economics to data science. IEEE Transactions on Knowledge and Data Engineering, 34(10):4586–4608, 2022.