8-2021

Data pricing and data asset governance in the AI Era

Jian PEI

Feida ZHU
Singapore Management University, fdzhu@smu.edu.sg

Zicun CONG

Luo XUAN

Liu HUIWEN

See next page for additional authors

Data Pricing and Data Asset Governance in the AI Era*

Jian Pei
jpei@cs.sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Feida Zhu
fdzhu@smu.edu.sg
Singapore Management University
Singapore

Zicun Cong
zicun_cong@cs.sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Xuan Luo
xuan_luo@cs.sfu.ca
Simon Fraser University
Burnaby, BC, Canada

Huiwen Liu
hwliu.2018@phdcs.smu.edu.sg
Singapore Management University
Singapore

Xin Mu
mux@pcl.ac.cn
Peng Cheng Laboratory
Shenzhen, Guangdong, China

ABSTRACT

Data is one of the most critical resources in the AI Era. While substantial research has been dedicated to training machine learning models using various types of data, much less efforts have been invested in the exploration of assessing and governing data assets in end-to-end processes of machine learning and data science, that is, the pipeline where data is collected and processed, and then machine learning models are produced, requested, deployed, shared and evolved. To provide a state-of-the-art overall picture of this important and novel area and advocate the related research and development, we present a tutorial addressing two essential problems. First, in the pipeline of machine learning, how can data and machine learning models be priced properly so that contributions from various parties can be assessed and recognized in a fair manner? Second, in the collaboration among many parties in building, distributing and sharing machine learning models, how can data as assets be managed? Accordingly, the first part of our proposal surveys data and model pricing in the pipeline of machine learning, while the second part discusses data asset governance for collaborative artificial intelligence. Each part is self-contained. At the same time, the two parts echo each other and connect a series of interesting and important problems into a dynamic big picture.

CCS CONCEPTS

• **General and reference** → **Surveys and overviews**; • **Information systems** → **Data mining**; **Electronic commerce**; **Electronic data interchange**; • **Social and professional topics** → **Digital rights management**; **Database protection laws**; **Soft intellectual property**; **Privacy policies**.

KEYWORDS

data assets, data pricing, data products, machine learning, AI, data governance

*This research is supported in part by the NSERC Discovery Grant program. All opinions, findings, conclusions and recommendations in this paper are those of the author and do not necessarily reflect the views of the funding agencies.

KDD '21, August 14–18, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8332-5/21/08.
https://doi.org/10.1145/3447548.3470818

ACM Reference Format:
Jian Pei, Feida Zhu, Zicun Cong, Xuan Luo, Huiwen Liu, and Xin Mu. 2021. Data Pricing and Data Asset Governance in the AI Era. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3447548.3470818

1 INTRODUCTION

Data is one of the most critical resources in the AI Era. While substantial research has been dedicated to training machine learning models using various types of data, much less efforts have been invested in the exploration of assessing and governing data assets in end-to-end processes of machine learning and data science, that is, the pipeline where data is collected and processed, and then machine learning models are produced, requested, deployed, shared and evolved.

There are two essential issues about data assets in the end-to-end processes of AI and data science. First, building powerful machine learning models, particularly deep learning models, requires large amounts of data. Much data may be acquired from external sources. Moreover, many parties share their machine learning models as a service (MLaas) so that they can monetize their data assets and intellectual property in a timely manner. As data is used in every step of building and delivering machine learning models and services, how can data and machine learning models be priced properly so that contributions from various parties can be assessed and recognized in a fair manner? Second, the nature of big data today entails an increasingly decentralized setting where data from various sources would be contributed to achieve data intelligence in a collaborative manner. How can we govern data assets peculiar to decentralized data collaboration under the two principles of trust and incentive, including consensus, privacy, data auditing, data accounting and incentive design?

To provide a state-of-the-art overall picture of this novel and important area and advocate the related research and development, we present a tutorial addressing the above two essential problems. Accordingly, our tutorial consists of two parts. The first part surveys data and model pricing in the pipeline of machine learning. The second part discusses data asset governance for collaborative artificial intelligence. Each part is elf-contained and takes 3 hours. At the same time, the two parts echo each other and connect a series of interesting and important problems into a dynamic big picture.

2 DATA AND MODEL PRICING IN THE PIPELINE OF MACHINE LEARNING

In the first part of the tutorial, we systematically review the state-of-the-art research and development in the end-to-end process of machine learning, and discuss the principles, opportunities, and challenges. We start with a quick introduction to end-to-end supply chain of data and machine learning, and review the essential principles in pricing data and machine learning models. Then, we focus on the practice of pricing in the three important components of machine learning, namely pricing in data collection, pricing in collaborative training of machine learning models, and pricing in machine learning model deployment.

This part contains the following sections.

(1) Introduction: machine learning pipeline, data and machine learning models as economic goods
(2) Essentials of pricing data and machine learning models: data markets, data and model pricing desiderata, pricing strategies
(3) Pricing in data collection – pricing raw data: pricing general data sets, pricing crowdsourcing tasks and data, pricing data queries, compensating privacy loss
(4) Pricing in data collection – pricing data labels: gold task-based pricing models, peer prediction-based pricing models
(5) Pricing in collaborative training of machine learning models: revenue allocation by Shapley value, other revenue allocation methods
(6) Pricing in machine learning model deployment: pricing machine learning models, pricing raw data versus machine learning models
(7) Summary and future directions

3 DATA ASSET FOR COLLABORATIVE INTELLIGENCE

The second part of the tutorial is devoted to the task of establishing data as a new class of assets for collaborative intelligence setting. We start with a systematic introduction to the notion of data asset, examining the three key components of value, right and control. As the sustainable success of any data economy relies on sound data asset governance, we then focus on the two critical governing principles of "trust" and "incentive". Various components of each principle are explored based on a wide range of technologies including consensus protocols, distributed ledgers, federated learning, data auditing and tokenomics design.

This part is presented with the following sections.

(1) Background and motivation of data asset
(2) Three core components of data asset: value, right and control
(3) Data asset governance for decentralized collaborative intelligence – principles, dimensions and mechanisms
(4) "Trust" for data asset governance – attacking models, agreement, accounting, auditing and privacy
(5) "Incentive" for data asset governance – focusing on tokenomics design
(6) Data economy ecosystem case studies for both personal data and B-to-B application settings
(7) Summary and future directions

SPEAKER BIBLIOGRAPHIES

Jian Pei is a Professor in the School of Computing Science and an associate member of the Department of Statistics and Actuarial Science at Simon Fraser University, Canada. His research areas include data science, big data, data mining, and database systems. He is a productive and influential author in data mining, database systems, and information retrieval. His publications have been cited extensively. His research has generated remarkable impact substantially beyond academia. His algorithms have been adopted by industry in production and popular open source software suites. He is responsible for several commercial systems of unprecedentedly large scale. Jian Pei received many prestigious awards. He is recognized as a Fellow of the Royal Society of Canada (RSC), the Canadian Academy of Engineering (CAE), ACM and IEEE.

Feida Zhu is a tenured Associate Professor in the School of Computing and Information Systems at Singapore Management University. His research interests include data mining and machine learning, blockchain and data asset. He was the founding director of the Pinnacle Lab for Analytics with China Ping An Insurance Group and DBS-SMU Life Analytics Lab. Feida Zhu has been the Founder and Chief Scientist of SYMPHONY, a blockchain-based protocol to empower democratized and personalized intelligence with privacy by design. Feida has over 100 peer-reviewed research publications at top international venues and has won several Best Paper Awards.

Zicun Cong is currently a Ph.D. student at the School of Computing Science, Simon Fraser University, Canada. His research interests lie in machine learning and data mining, with an emphasis on explainable artifactual intelligence and data pricing. He has worked extensively on interpreting the internal mechanisms of complex machine learning and statistical models. Currently, he is focusing on designing efficient, scalable, and interpretable algorithms for data and model pricing.

Xuan Luo is currently a Ph.D. student at the School of Computing Science, Simon Fraser University, Canada. Her research interests lie in blockchain and data mining with an emphasis on data pricing. She has worked for years in designing a scalable payment solution for hybrid token exchange and applying machine learning algorithms to analyze blockchain transaction data. Currently, her focus is on designing efficient algorithms for data pricing.

Huiwen Liu is currently a Ph.D. student at the School of Computing and Information Systems at Singapore Management University. Her research interests lie in data mining, machine learning, data asset and governance in distributed settings. She has worked extensively on evaluating consensus protocols in both classic settings and more recent applications of societal scale. Currently, she is focusing on designing novel consensus protocols for decentralized learning environment with collaborative data governance.

Xin Mu is currently working as a postdoc at Peng Cheng Laboratory, Shenzhen, China. His research interests lie primarily in machine learning and data mining, including ensemble methods, stream mining, collective intelligent and blockchain. He is currently focusing on data auditing for decentralized data collaboration. He received his Ph.D. degree in the LAMDA Group at Nanjing University, advised by Professor Zhi-Hua Zhou.