Tokenized Data Pricing for the Data Economy

Candidate XCDF6

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Engineering

of

University College London.

Department of Computer Science

University College London

26th April 2023

Declaration

I, Candidate XCDF6, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

This thesis presents a first-principles approach to pricing datasets and implements a data pricing engine. The growth and development of machine learning and the wider Data Economy has necessitated the need for data markets and ”data liquidity”- the ease with which data can be bought and sold. Pricing has the potential to unlock these data markets [1], [2] and transform data into a hyper-liquid, tradeable and valuable asset. This hypothesis has been put forward by previous work [3] on data supply chains which postulates that the absence of pricing transparency is the cause for 1) decreased liquidity in data markets and 2) asymmetric risk distribution between buyers and sellers of data.

Our study creates an ensemble of two data valuation and pricing methodologies based on **apparent** (i.e., objective measures such as data quality) and **latent** (i.e., subjective measures; consumer-dependent) data value sources. To test these methodologies, we first unify disparate theoretical measures of data value and develop an automatic pipeline that extracts value-adding attributes. We then create a novel value-inference model on a repository of about 23,000 datasets sourced from Kaggle. Finally, this work materializes in the form of an end-to-end valuation engine that provides decision-aiding tools (i.e., uncertainty reduction mechanisms) for data buyers while ensuring data security for sellers. To increase the scalability of our pricing engine to machine and deep learning contexts we implement advanced value-estimation techniques that reduce buyer costs and enable the engine’s deployment to high volume, many-to-many markets.

We have structured our study around the following series of three experiments:

1. **Apparent Value**. The first study conducts an in-depth analysis of current literature on the objective and intrinsic quality of datasets. We unify disparate measures of quality, which we refer to as *data attributes*, into a comprehensive list of data value-drivers. To scale the valuation process, we automate attribute extraction to create transparent data summaries relevant to a dataset’s worth. We then perform a large-scale investigation on public datasets and build a model to predict data value on a relative scale. Finally, we provide a theoretical framework for applications to data search and selection.

2. **Latent Value**. This study investigates complementary approaches to apparent valuation. The

objective is to integrate extrinsic measures of value such as data utility to provide a truthful and transparent price. We consider market scenarios where buyers have an analytical model (i.e., statistical or machine learning) which they aim to improve by acquiring new datasets. In this setting, we develop a novel uncertainty reduction framework that provides buyers with decision-aiding information about datasets. Further, we demonstrate the superiority of this approach using Monte Carlo experiments.

3. **End-to-end Pricing Engine**. This final study describes the implementation of our end-to-end pricing engine which combines the two valuation/pricing methodologies explored in the first experiments. We implement state-of-the-art value-estimation techniques to address the issue of scalability and computational complexity. Finally, we evaluate the feasibility of our proposed engine as a function of value-prediction accuracy and time benchmarks.

We highlight the following original scientific contributions of the thesis:

1. Analyzed the apparent **value of public datasets** (i.e., from the Kaggle platform). To our best knowledge, there is only a single similar study [3]. However, it restricts its investigation to the relationship of price and data topics with no consideration for the intrinsic characteristics of the datasets.

2. Developed a **scalable value model** using intrinsic data quality attributes. The novel methodology ranks data assets on a relative scale by leveraging previous buyer preferences.

3. **Support buyer decision-making** in data markets using a novel uncertainty reduction framework. We further empirically demonstrate the benefits of this approach in model-based tasks, where a buyer's objective is to purchase data that improves the performance of its analytical model.

4. Created an **end-to-end pricing engine** by combining the proposed apparent and latent valuation methods into a single model which can be easily deployed to various data market settings (i.e., one-to-one, one-to-many, many-to-many).

5. A position **paper presenting the framework for the responsible use, valuation and monetisation of data**. The paper additionally explored the mechanics of tokenizing data, including the applications, opportunities and implications for an equitable Data Economy. Available at: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4419590`.

Candidate XCDF6, *Tokenized Data Pricing for the Data Economy*[1]

Supervisors: Prof. Philip Treleaven and Dr. Hirsh Pithadia

[1] Source code available at: `https://github.com/rechim25/pricing-for-data-economy`.