Master Thesis

Karlsruher Institut für Technologie

Institut für Programmstrukturen und Datenorganisation (IPD)
Lehrstuhl für Systeme der Informationsverwaltung, Prof. Böhm

**Master Thesis**

# Quantifying the Effect of UCI Data Publication on Citation Count

Transparency and reproducibility are indispensable for good research practice. To fulfill these requirements, research data should be Findable, Accessible, Interoperable, and Reusable (FAIR[a]). In recent years, data sharing has become more widespread [1]. Next to benefits for the scientific community as a whole, sharing data yields citations and promotes visibility [2, 3, 4].

Research in machine learning often requires many datasets, for instance, to provide strong experimental evidence of the benefits of a novel algorithm. This demand creates an opportunity for any data-driven research paper to increase the citation count by publishing data. In particular, one of the most popular dataset sources used by the machine learning community, the UCI repository, allows requesting certain citations in exchange for data. There is no limit — one can ask to cite many papers at once[b]. Arguably, this ‘obligation to cite’ creates an unjust situation in science. This is because paper citation count starts to depend on data availability and loses connection to research quality.

This thesis aims to estimate **the influence of data sharing via the UCI repository on citation count and the effect of an explicit citation request**. It will be interesting to compare the results to findings from similar research that however does not account for data usage. An exciting challenge is quantifying the effects of two factors that might exist and act simultaneously:

- Articles that make their data available tend to be better.
- Requesting citations for data increases citation count even for mediocre papers.

We expect existing research on causality [5] to provide the necessary tools for this kind of analysis. In case of delayed data publication for instance, one may study the dynamics of a citation count before and after publishing the data. Work on the assignment consists of the following tasks:

- Literature review with a focus on causality.
- Proposing a model to estimate the effects of data sharing on UCI.
- Collecting necessary data to estimate the model parameters (most likely via web scrapping).
- Estimating model parameters, verifying the validity of assumptions, analyzing results.

To help you with this task, we offer:

- Thorough mentoring and recurrent meetings with your advisor.
- Established collaboration with an expert in causal inference.

During this work, the student will acquire practical experience in Machine Learning, causal inference, web scrapping, and potentially contribute to justice in research by quantifying the effect of trading data for citations and raising scientists’ awareness of it.

[1] L. Tedersoo et al. “Data sharing practices and data availability upon request differ across scientific disciplines”. In: *Scientific data* 8.1 (2021).
[2] H. A. Piwowar et al. “Sharing detailed research data is associated with increased citation rate”. In: *PloS one* 2.3 (2007).
[3] G. Christensen et al. “A study of the impact of data sharing on article citations using journal policies as a natural experiment”. In: *PloS one* 14.12 (2019).
[4] G. Colavizza et al. “The citation advantage of linking publications to research data”. In: *PloS one* 15.4 (2020).
[5] J. Pearl. “The seven tools of causal inference, with reflections on machine learning”. In: *Commun. ACM* 62.3 (2019).

[a] Principles of research data publishing
[b] And some use this opportunity, see, e.g., this and that

**Ansprechpartner**

Vadim Arzamasov, Ph.D.
Am Fasanengarten 5

vadim.arzamasov@kit.edu
76131 Karlsruhe

Raum: 340
Gebäude: 50.34