# Towards NLP-Enhanced Data Profiling Tools

Immanuel Trummer
Cornell University
Ithaca (NY)
itrummer@cornell.edu

**Abstract.** Data profiling tools obtain valuable information via natural language processing (NLP) of column and table names. If profiling time is limited, NLP helps to prioritize data analysis methods and targets.

**Data Profiling.** The goal of data profiling [3] is to derive useful meta-data about a data set. Common profiling goals include, for instance, data correlation analysis, identifying candidate keys, or finding foreign key relationships. The resulting meta-data can be used for a variety of tasks, ranging from refinements of the database schema (e.g., by introducing foreign key constraints) to query optimization (e.g., taking into account data correlations leads to more accurate execution cost estimates for query plans). However, data profiling is expensive, motivating methods that help narrowing down profiling scope. This paper advocates for the use of NLP methods to prioritize expensive profiling operations.

**NLP Background.** The area of NLP has seen transformational advances over the past few years. These advances have been enabled by new neural network architectures, in particular the Transformer architecture, as well as by pre-trained language models [5]. Such models implement the idea of transfer learning: they are pre-trained on tasks for which large amounts of training data are readily available (e.g., predicting obfuscated words in text). For sufficiently large models, pre-training can be sufficient to achieve impressive performance even on novel NLP tasks that relate only marginally to the original pre-training objective [1]. Furthermore, pre-trained models can be specialized efficiently to new tasks via fine-tuning, requiring comparatively moderate amounts of computation time and training samples [2].

**NLP for Data Profiling.** Human database administrators can often form an educated guess about data properties, based on column and table names alone. The names of schema elements are chosen with the goal to convey an intuition for their semantics. For instance, considering columns “maker” and “model" in a table called “cars”, most users would assume a correlation between the two columns. Advanced NLP methods may enable automated data profiling tools to come to similar conclusions. Of course, such conclusions are not reliable (already since the names of schema elements may be chosen poorly, misleading human users and NLP-enhanced tools alike). However, text related to the database (which includes the names of schema elements as well as supporting documents such as the data dictionary) may provide useful hints to guide expensive profiling efforts. For instance, searching for correlated columns under profiling time constraints, a tuning tool could prioritize analysis of column pairs that are likely correlated, based on their names.

**First Prototype.** The author implemented a simple prototype of an NLP-enhanced data profiling tool (available online at https://github.com/itrummer/DataCorrelationPredictionWithNLP). This prototype finds correlated column pairs under a profiling cost budget. It supports multiple types of correlation, including, for instance, correlation according to the Pearson coefficient or according to Theil’s U. Instead of analyzing correlation for column pairs in random order, the prototype prioritizes columns that are likely correlated. To assess the likelihood of correlation for two specific columns, the prototype uses a pre-trained language model. This model is fine-tuned for predicting correlation based on column names.

**Proof of Concept.** The author performed first experiments to show that NLP on schema element names can speed up data profiling. The profiling goal was to find column pairs with an uncertainty coefficient (also known as Theil’s U) of at least 0.95. A collection of 3,952 data sets, obtained from the Kaggle Web site and featuring 119,384 column pairs, was used as test cases (the generation of these test cases as well as the training process is described in more detail in a technical report [4]). For instance, if allowed to analyze 25% of column pairs for each data set, the prototype finds twice as many correlated column pairs when prioritizing columns that are likely correlated, according to NLP-based predictions. This trend holds for other percentages and other measures of correlation, including Pearson and Spearman correlation, as well.

**Conclusion.** Natural language analysis of text associated with database schema elements may unlock efficiency gains in data profiling. First experimental results support this hypothesis.



## REFERENCES

[1] T. Brown, B. Mann, N. Ryder et al. 2020. Language models are few-shot learners. *https://arxiv.org/pdf/2005.14165.pdf*. 1-75.
[2] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. 2018. *ACL*. 328-339.
[3] F. Naumann, 2013. Data profiling revisited. *SIGMOD Record.* 42, 4 (2013), 40-49.
[4] I. Trummer. 2021. Can deep neural networks predict data correlations from column names? *https://arxiv.org/pdf/2107.04553.pdf* . 1-12.
[5] T. Wolf, D. Lysandre, V. Sanh et al. 2020. Transformers: state-of-the-art natural language processing. *EMNLP*. 38-45.