Automated Relational Data Explanation using External Semantic Knowledge

Sainyam Galhotra
The University of Chicago
sainyam@uchicago.edu

Udayan Khurana
IBM Research AI
ukhurana@us.ibm.com

ABSTRACT

In data science problems, understanding the data is a crucial first step. However, it can be challenging and time intensive for a data scientist who is not an expert in that domain. Several downstream tasks such as feature engineering and data curation depend on the understanding of data semantics. In this demonstration, we present, *ADE (Automated Data Explanation)*, a novel system that uses *maximum likelihood estimation approach* through ensembles for automatically labeling and explaining relational data by taking advantage of openly available semantic knowledge bases, webtables and Wikipedia. It helps a user to understand concepts of various columns and their relationships, an abstract summary about the overall data, and additional context not present in the data. It reduces the need for cumbersome search queries or expert consultation and can also receive inputs or corrections from a user, making it a mixed-initiative automation system.

PVLDB Reference Format:
Sainyam Galhotra and Udayan Khurana. Automated Relational Data Explanation using External Semantic Knowledge. PVLDB, 15(12): 3562 - 3565, 2022.
doi:10.14778/3554821.3554844

1 INTRODUCTION

Understanding the nature of a given problem in data science or data management is essential from the perspective of many tasks such as data pre-processing, feature engineering or deciding appropriate metrics for evaluation. For a data scientist[1] unfamiliar with the domain of the problem or the specific data accompanying the problem, it is a time-consuming process to achieve an appropriate level of understanding before proceeding with the various components of the data science pipeline. Analysts typically iterate over the dataset records manually to understand the set of entities and relate them across different input tables. In this work, we ask the questions: '*Can we leverage web datasets and knowledge graphs to automatically generate semantically coherent explanations for relational data?*'.

There are many crucial aspects to help data scientists understand a dataset. First, we need to identify the semantic meaning of values in various columns, and identify present entities and their attributes. Secondly, we need to abstract the individual row-level concepts to a larger real-world relationship. Finally, we need to summarize and explain the given dataset and machine learning problem in a way that is understandable by a variety of users. The impact of such mapping and abstraction is to aid the data scientist's decision-making in the following areas: (a) Summarize the input dataset collection; (b) Provide a wider context with respect to external world knowledge; (c) Identify additional features when added to the given data that could potentially improve understanding and in training a machine learning model; (d) Perform data preparation related tasks such as missing value imputation, string value encoding; (e) Find sensitive attributes useful for de-biasing the dataset and perform fairness aware learning; (f) Identify an appropriate metric of evaluation for the problem.

To this end, we propose *ADE (Automated Data Explanation)*[2], a novel system that helps summarize an input collection of relational tables to achieve above-mentioned requirements. ADE is based upon effective use of openly available web resources in the following forms: (a) Knowledge bases (KB) such as DBPedia, Wikidata, custom ones that represent facts in form of triplets, or ontologies; (b) Openly available web tables, such as those on wikipedia and government websites; (c) Expert (or community) edited abstracts (or reviews) on topics such as those found on Wikipedia or other domain-specific websites such as Investopedia for financial investments. The externally available information is robustly combined using a *maximum likelihood estimation approach* that relies on different dataset signals for understanding the semantics of the dataset and generating an explanation.

To the best of our knowledge, we are not aware of a system that provides a user with the capability to quickly understand a given data. We demonstrate scenarios where ADE provides useful summary helping a data scientist with actionable insights into the data. This system does not interrupt the flow of an existing manual or AutoML pipeline, but can serve as a complimentary add-on to help the scientist get more insights into the data with minimum effort. In this paper, we discuss the essential aspects of ADE, provide some examples along with a plan to demonstrate it to audience members from diverse technical backgrounds at the conference. ADE's explanation generation and identification of new features from external web sources is complimentary to prior work in the domain of automated machine learning [4, 11] and analysis [3].

The problem of identifying semantic concepts of columns of tabular data is relevant to the objective of data understanding. In order to identify semantic meaning of a column with non-numerical data, most approaches annotate by performing an exact match with KB entities and their relationships. Systems such as ColNet [2] and Sherlock [6] exploit and extract contextual semantics from

Proceedings of the VLDB Endowment, Vol. 15, No. 12 ISSN 2150-8097.
doi:10.14778/3554821.3554844

[1] We refer to a data scientist as the conductor of a machine learning/data mining application process as well as the user of the presented system.

[2] We present a demonstration of ADE's functionalities at https://semanticannotation.github.io/

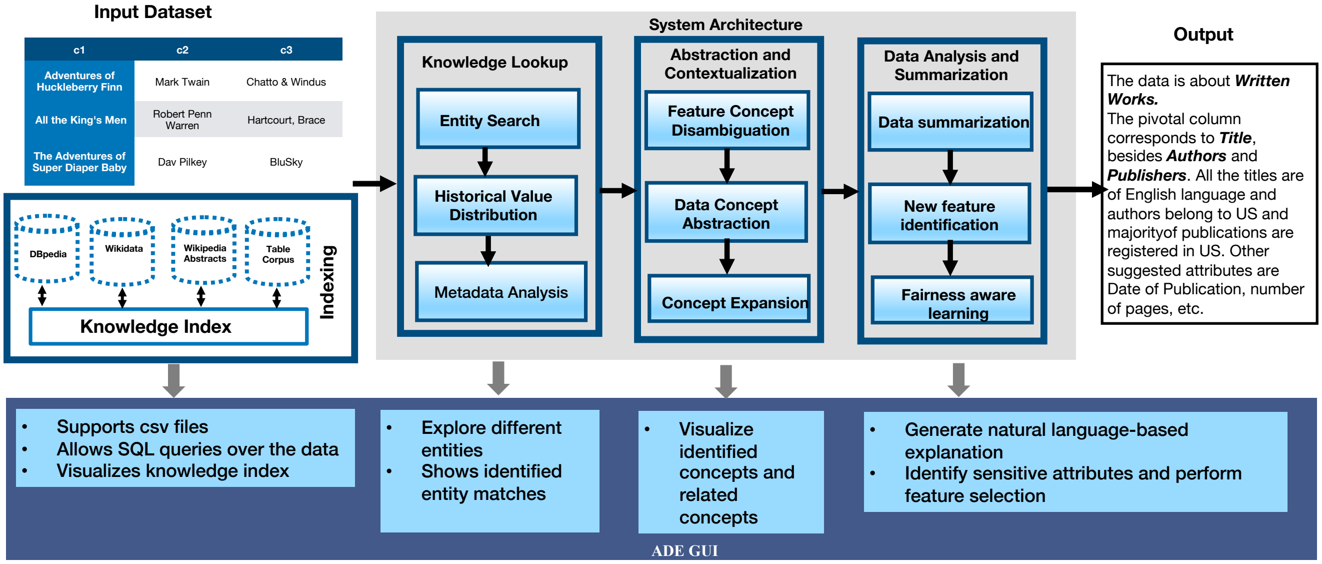

Input Dataset
c1
c2
c3
Adventures of Huckleberry Finn
Mark Twain
Chatto & Windus
All the King's Men
Robert Penn Warren
Hartcourt, Brace
The Adventures of Super Diaper Baby
Dav Pilkey
BluSky
DBpedia
Wikidata
Wikipedia Abstracts
Table Corpus
Indexing
Knowledge Index
System Architecture
Knowledge Lookup
Entity Search
Historical Value Distribution
Metadata Analysis
Abstraction and Contextualization
Feature Concept Disambiguation
Data Concept Abstraction
Concept Expansion
Data Analysis and Summarization
Data summarization
New feature identification
Fairness aware learning
Output
The data is about Written Works.
The pivotal column corresponds to Title, besides Authors and Publishers. All the titles are of English language and authors belong to US and majorityof publications are registered in US. Other suggested attributes are Date of Publication, number of pages, etc.
Supports csv files
Allows SQL queries over the data
Visualizes knowledge index
Explore different entities
Shows identified entity matches
Visualize identified concepts and related concepts
Generate natural language-based explanation
Identify sensitive attributes and perform feature selection
ADE GUI

Figure 1: ADE System Architecture and key highlights of ADE user interface.

the tabular data using neural networks. Compared to textual data, numerical data is much harder to annotate due to the inherent ambiguity. Most previous works either identify a set of features for high-dimensional data [1] or develop a background KB from sources like Wikidata [7] and then apply approaches such as transformation [10] and k-nearest neighbors search [9] to collect candidate semantic labels. These approaches are complimentary to the ideas described in this paper, and our work presented here extends upon the ideas of column concept determination.

2 ADE OVERVIEW

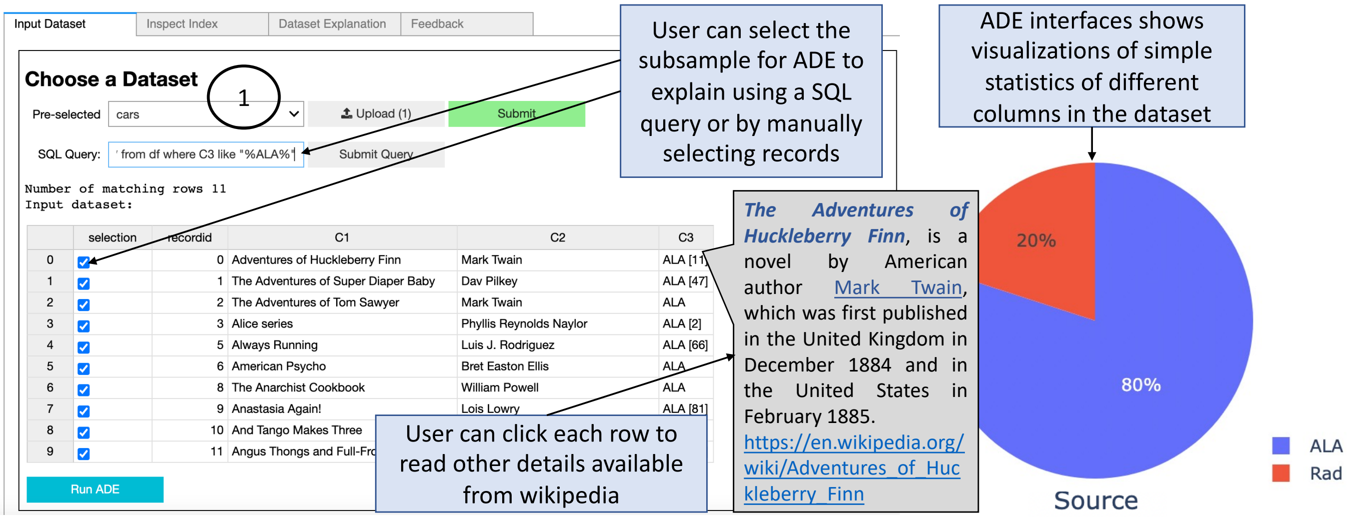
Figure 1 illustrates the system architecture along with the corresponding visualization of ADE's user interface. The input to the system is a relational data (in form of CSV files or SQL data) and the output is (a) an annotation of the columns, (b) the overall data title, (c) description of the features in terms of their attributes and, (d) additional possible attributes that provide a larger context to this data. The user can input any csv file and perform SQL queries to identify a smaller subset for subsequent processing and explanations.

In addition to the input dataset, ADE leverages a knowledge index constructed from information available in (a) knowledge graphs such as DBPedia and Wikidata, (b) textual descriptions of entities such as Wikipedia abstracts, and (c) more than 32 million webtables from wikipedia, blogs and data repositories such as Viznet [5]. The knowledge index comprises efficient data structures based on our previous work on semantic annotation [8]. Specifically, it iterates over the millions of tables obtained from openly available web sources and produces inverted entity-concept index for textual values, interval tree index for numerical values, pattern tree index for other attributes, and column co-occurence index over all columns. Index construction is a one-time process that helps to efficient retrieval of relevant information subsequently. As a part of the visualization, ADE allows users to search and inspect values over the knowledge index to understand different entities and their related concepts mined from web sources.

The ADE pipelines works in three phases, i) Knowledge Lookup ii) Abstraction and Contextualization, iii) Data analysis and summarization. We now provide a high-level overview of the three components and corresponding user interfaces. For algorithmic details of the building blocks, we refer the reader to C^2 [8].

Knowledge Lookup component aims to link each of the textual values in the input dataset to entities in knowledge graphs or values in external webtable corpus. For the numerical values, it considers meta-information of the columns such as distribution of values, range overlaps and compare these with columns observed in millions of historical instances for numerical columns effectively using modified interval tree indexes. Additionally, it analyzes meta-data such as column names and/or any available column description to match with property values from knowledge bases or column names of historical data. This mechanism leverages indexing techniques discussed in [8] for efficient lookup.

In the second phase – **Abstraction and Contextualization**, (a) all the results from the first phase are aggregated, homogenized to a common context and the most likely explanations for column concepts are identified (based on a *maximum likelihood estimation approach* described in [8]). This also involves disambiguation, such as "book" is considered more relevant concept than "movie" when put in context with an "author" column. This disambiguation is performed based on the column co-occurence index. (b) The inferred concepts for the columns are used to identify the overall concept of the table(s). This phase also identifies key patterns of the columns, like all authors are from US origin. Note that identified patterns are not only based on the input attributes of the dataset but also other attributes available from millions of web sources. Finally, (c) the list of column concepts is expanded by leveraging information available from knowledge graphs, e.g., if a "title" and "author" was found in the data, the system also suggests "publication date", "page count",



SQL Query: ' from df where C3 like "%ALA%"

```
Number of matching rows 11
Input dataset:
```

	selection	recordid	C1	C2	C3
0	☑	0	Adventures of Huckleberry Finn	Mark Twain	ALA [11
1	☑	1	The Adventures of Super Diaper Baby	Dav Pilkey	ALA [47]
2	☑	2	The Adventures of Tom Sawyer	Mark Twain	ALA
3	☑	3	Alice series	Phyllis Reynolds Naylor	ALA [2]
4	☑	5	Always Running	Luis J. Rodriguez	ALA [66]
5	☑	6	American Psycho	Bret Easton Ellis	ALA
6	☑	8	The Anarchist Cookbook	William Powell	ALA
7	☑	9	Anastasia Again!	Lois Lowry	ALA [81]
8	☑	10	And Tango Makes Three		
9	☑	11	Angus Thongs and Full-Fro		

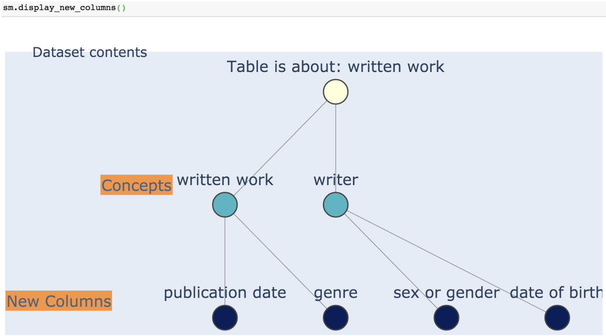
User can select the subsample for ADE to explain using a SQL query or by manually selecting records

ADE interfaces shows visualizations of simple statistics of different columns in the dataset

The Adventures of Huckleberry Finn, is a novel by American author Mark Twain, which was first published in the United Kingdom in December 1884 and in the United States in February 1885. https://en.wikipedia.org/wiki/Adventures_of_Huckleberry_Finn

User can click each row to read other details available from wikipedia

Figure 2: ADE interface with the input screen and other tabs.



```
sm.display_new_columns()
```

Figure 3: Concepts of the dataset along with new columns

etc., as the other attribute columns for this type of data. One of the central ideas that guarantee robustness of concept identification and contextualization is the ability of our approach to ensemble different sources of information based on their quality [8].

The third phase, **Data analysis and summarization** takes the identified concepts to generate an easily understandable natural language summary of the dataset. The explanation is complemented with high-level statistics of newly identified features such as correlation with other attributes, and representative patterns. These statistics (often referred to as data profiles) are helpful in explaining the importance of attributes available from other data sources which can be used by the end user to choose useful features for the respective analytics tasks. In addition to the new features, ADE provides a fairness score of each feature (whenever relevant) that captures its correlation with the sensitive attributes. These scores allow the end-user to perform fairness aware analysis and train trustworthy classifiers.

Implementation ADE backend is implemented in Python 3.6, which is supported by interactive visualizations and an API that helps users utilize it in Jupyter notebooks, as demonstrated in this paper. In addition to the features presented in Figure 1, it also provides the ability for a user to correct specific predictions from the ADE system at any step and use the overridden choices instead. All the changes made by the user are remembered for any future reference to any of these columns. We highlight each of the steps in the corresponding figures.

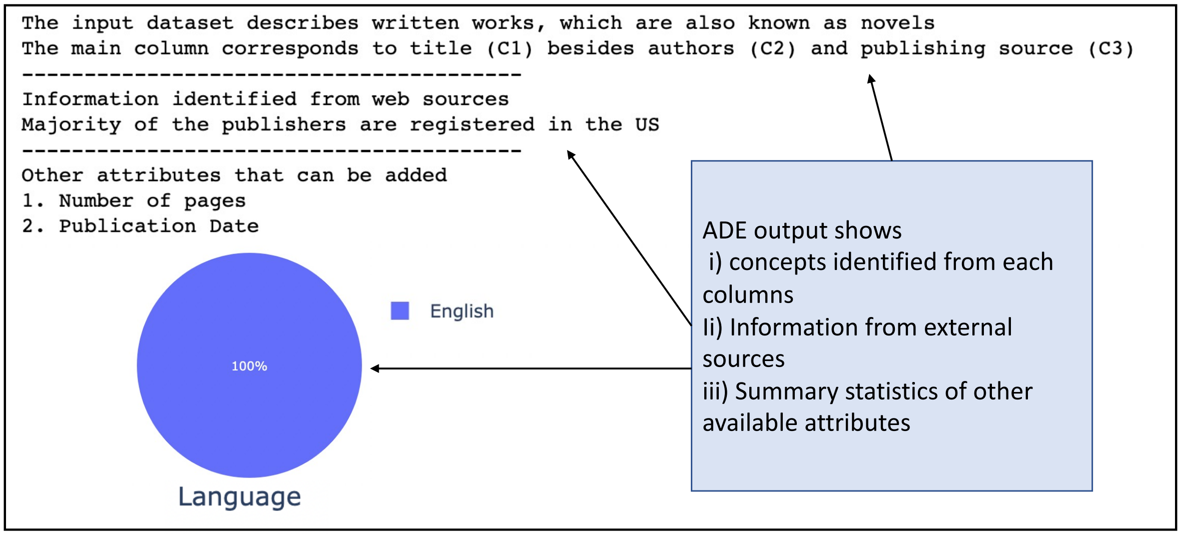
3 DEMONSTRATION PLAN

We plan to demonstrate ADE through a `Jupyter notebook` with interactive IPyWidget components, that will engage the audience to highlight various functionalities. We allow the user to choose from five preloaded datasets (books, cars, animals, covid, and IMDB) or upload a new dataset in csv files. The user can also use all functionalities as an API call to effectively plug-in ADE's output into their machine learning pipelines. During the demonstration, we will guide the participants through the following steps. Figure 2 shows a snapshot of the user interface for an example dataset.

Step (1) **(Select input dataset):** First, the user selects a dataset or a data repository to input for ADE to generate explanation. ADE shows a snapshot of the input dataset and provides a SQL query option to identify a smaller sub-sample for subsequent steps.

Step (2) **(Run ADE):** After uploading the dataset, the explanations of the dataset are generated as soon as the user presses 'Run ADE' (Figure 2).

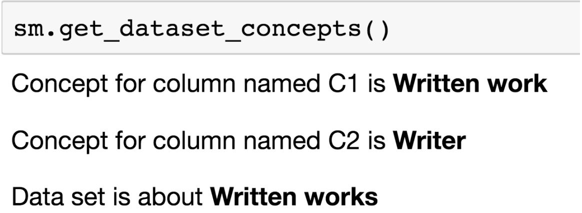
Step (3) **(Inspect the index):** The 'inspect index' tab allows the user to peek through a small subset of the knowledge index constructed over millions of datasets. This index demonstrates the set of available resources and the user can inspect specific entities by typing corresponding keywords. Figure 3 shows a snapshot of the knowledge index for the columns in the input dataset and its connection to datasets available externally. The user can further inspect the confidence of concept predictions for input columns to understand the uncertainty in the output.



```
The input dataset describes written works, which are also known as novels
The main column corresponds to title (C1) besides authors (C2) and publishing source (C3)
----------------------------------------
Information identified from web sources
Majority of the publishers are registered in the US
----------------------------------------
Other attributes that can be added
1. Number of pages
2. Publication Date
```

ADE output shows
i) concepts identified from each columns
Ii) Information from external sources
iii) Summary statistics of other available attributes

Figure 4: ADE's output for the input example.



```
sm.get_dataset_concepts()
```

Concept for column named C1 is **Written work**

Concept for column named C2 is **Writer**

Data set is about **Written works**

Figure 5: Python based API call that allows users to plugin ADE's output in their analytics pipelines.

Step 4 (Analyze individual columns): analyze the content and meta-data for individual columns and identify the concepts they represent along with a list of top-5 candidates, and abstract the idea of what the overall dataset is about.

Step 5 (Explain the dataset): provide a larger context to the given dataset in the form of related attributes and allows user to add new columns based on their properties like highly correlated or sensitive attributes, which can then be used for downstream tasks. Additionally, the natural language summary of the dataset or the involved machine learning problem is helpful for the user (Figure 4). ADE presents externally available information in the form of statistics like patterns and vizualizations as shown in Figure 4.

Step 6 (Correct ADE output): demonstrate the ability of a user to contribute in addition to the system's predictions. In this component, a user can edit the ADE's natural language output, which is used by the feedback pipeline to correct the backend mapping for future datasets and analysis.

API call. ADE's graphical user interface is accompanied by an API which allows easy plugin of the library in python. Figure 5 shows an example API call which helps user to quickly identify concepts and plugin the output in their codebase. The API call provides one-line commands to run all functionalities supported in GUI.

Demonstration Engagement. After the guided demonstration, users will be able to use ADE to generate explanations for other preloaded datasets like IMDB, cars, Animals, etc or upload new datasets. We will provide the choice of five preloaded openly available datasets from different domains for use. Through this demonstration, we will show how ADE can help users tap into external data sources to explain datasets, identify patterns and useful attributes for a fairness-aware and effective data discovery and analytics. The key takeaway is that semantic annotation of entities is helpful to improve user understanding and external sources can play a key role in this regime.

REFERENCES

[1] Firas Abuzaid, Peter Kraft, Sahaana Suri, Edward Gan, Eric Xu, Atul Shenoy, Asvin Ananthanarayan, John Sheu, Erik Meijer, Xi Wu, et al. 2018. Diff: a relational interface for large-scale data explanation. *PVLDB* 12, 4 (2018), 419–432.

[2] J Chen, E Jiménez-Ruiz, I Horrocks, and C Sutton. 2019. Colnet: Embedding the semantics of web tables for column type prediction. In *AAAI*.

[3] Noëlie Cherrier, Jean-Philippe Poli, Maxime Defurne, and Franck Sabatié. 2019. Consistent feature construction with constrained genetic programming for experimental physics. In *2019 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 1650–1658.

[4] Sainyam Galhotra, Udayan Khurana, Oktie Hassanzadeh, Kavitha Srinivas, Horst Samulowitz, and Miao Qi. 2019. Automated Feature Enhancement for Predictive Modeling using External Knowledge. In *2019 International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1094–1097.

[5] Kevin Hu, Neil Gaikwad, Michiel Bakker, Madelon Hulsebos, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. 2019. VizNet: Towards a large-scale visualization learning and benchmarking repository. In *CHI*. ACM.

[6] M Hulsebos, K Hu, M Bakker, E Zgraggen, A Satyanarayan, T Kraska, Ç Demiralp, and C Hidalgo. 2019. Sherlock: A Deep Learning Approach to Semantic Data Type Detection. *KDD* (2019).

[7] Ernesto Jiménez-Ruiz, Vasilis Efthymiou, Jiaoyan Chen, Vincenzo Cutrona, Oktie Hassanzadeh, Juan Sequeda, Kavitha Srinivas, Nora Abdelmageed, Madelon Hulsebos, Daniela Oliveira, and Catia Pesquita (Eds.). 2022. *Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching*. CEUR Workshop Proceedings, Vol. 3103. CEUR-WS.org.

[8] Udayan Khurana and Sainyam Galhotra. 2021. Semantic Annotation for Tabular Data. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management* (2021).

[9] S Neumaier, J Umbrich, JX Parreira, and A Polleres. 2016. Multi-level semantic labelling of numerical values. In *ICWS*.

[10] P Nguyen and H Takeda. 2018. Semantic labeling for quantitative data using Wikidata. (2018).

[11] Kavitha Srinivas, Takaaki Tateishi, Daniel Karl I. Weidele, Udayan Khurana, Horst Samulowitz, Toshihiro Takahashi, Dakuo Wang, and Lisa Amini. 2022. Semantic Feature Discovery with Code Mining and Semantic Type Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.