Dataset Discovery and Exploration: State-of-the-art, Challenges and Opportunities

Norman W. Paton
norman.paton@manchester.ac.uk
University of Manchester
Manchester, UK

Zhenyu Wu
zhenyu.wu@manchester.ac.uk
Univerity of Manchester
Manchester, UK

ABSTRACT

Dataset discovery and exploration involve identifying and understanding the available data, thereby informing users as to what data analyses may be possible. Discovering and exploring the relationships between datasets benefits from tool support, and in this tutorial, we specifically consider techniques that underpin *dataset search*, *data navigation*, *dataset annotation* and *schema inference*. Although there are are significant results in each of these areas, in practice they are far from independent of each other, and can share both objectives and underlying techniques. As a result, this tutorial not only seeks to provide insights into the challenges and opportunities of these areas in isolation, but also points out how they can complement and inform each other. The tutorial is associated with a Python Notebook to illustrate the concepts and techniques discussed in practice.

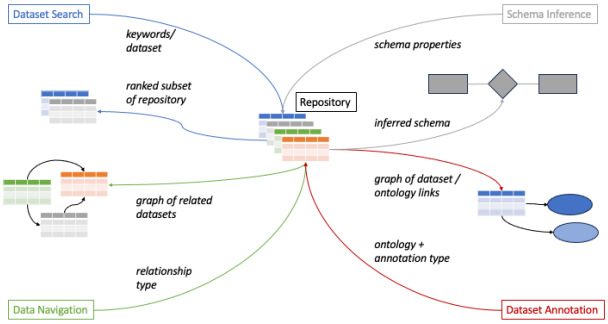

Figure 1: Dataset discovery and exploration functionalities reviewed in the tutorial.

1 INTRODUCTION

Data scientists are employed to obtain value from data, but survey evidence indicates that they spend in the region of 80% of their time preparing the data for analysis [1]. Drilling into this figure, 19% of the time is being spent on *collecting data sets* and 60% on *cleaning and organising data*. The subject of this tutorial is Dataset Discovery and Exploration, and in particular techniques and tools that aim to enhance the productivity of data scientists and engineers when faced with obtaining insights from large, potentially valuable but minimally curated data repositories. In relation to how data scientists spend their time, this likely includes all of the time spent on *collecting data sets* and at least some of the time spent on obtaining the knowledge necessary for *organising* the data. Relevant repositories could include data lakes [2], open data [3] or web tables [4].

Data management researchers have long recognised that there are challenges and opportunities obtaining an understanding of the available datasets and relationships between them, and there are results of substance in several different areas. In this tutorial, we will specifically cover:

- *Dataset search*: given a request, for example in the form of keywords or an available dataset, return a ranked list of datasets from a repository that meet the request (e.g., [5–9]).
- *Data navigation*: given a collection of datasets, identify a set of relationships between those datasets that represent features such as shared domains or joinability (e.g., [10–14]).
- *Dataset annotation*: given a collection of datasets and an ontology, associate datasets or their properties with concepts from the ontology that represent their semantics (e.g., [15–19]).
- *Schema inference*: given a collection of datasets, return a schema that summarises the structural features of the datasets in the repository (e.g., [20–24]).

All of these areas have been associated with significant technical results, and indeed there are surveys on several of these areas, typically from area-specific communities [7, 21, 23]. However, this emphasis on individual areas means that it can be difficult to see the wood for the trees: these areas address different aspects of the wider problem of Dataset Discovery and Exploration, and this tutorial seeks to make explicit recurring challenges, shared technical approaches and potential for synergies across these areas.

There have been several proposals that bring together different aspects of dataset discovery and exploration. For example, Aurum supports both search and join discovery [11], and Blend [25] proposes a collection of operators for column search and comparison along with a framework for composing these operators. There have also been usability evaluations of systems that support several different aspects of dataset search and navigation [26, 27]. However, more integrated investigations are in the minority, and we hope to encourage further integrated studies through the tutorial.

2 FUNCTIONALITIES

There is likely no fully accepted definition as to what constitutes data discovery and exploration, but here are some possible definitions:

- *Dataset discovery* is the process of identifying datasets that may meet an information need. This may, e.g., be

done directly through a search, by navigating from related datasets, or by browsing the datasets with a specific annotation.
- *Dataset exploration* is the process of understanding the properties of datasets and the relationships between them. This may, for example, be carried out by exploring the relationships of a given dataset, by viewing shared annotations at dataset or attribute level, or by exploring relationships that are shared by several datasets in an inferred schema.

Both dataset discovery and dataset exploration tend to build on notions of similarity, which can be *syntactic* or *semantic*:

- *Syntactic similarity* captures relationships between tokens or collections of tokens. Some techniques that use syntactic similarity may be exact; if the aim is to identify joinable columns, then a join may only succeed when an equality predicate is applied to completely consistent representations. However, syntactic similarity may also involve partial values, for example by applying Jaccard similarity to sets of n-grams. Thus syntactic similarity can accommodate some level of representational inconsistencies, but is unlikely to cope with *USA* in relation to *United States of America* or *record* in relation to *album*.
- *Semantic similarity* captures relationships between the meaning of table, column, row or cell values, which may or may not be syntactically similar. Semantic similarity often builds on word embeddings [28] or language models [29] to construct multi-dimensional representations for data model constructs.

It would be possible to classify functionalities to support dataset discovery and exploration in different ways, but here we consider the areas illustrated in Figure 1.

2.1 Dataset Search

Given a repository of data, a search aims to retrieve datasets that, individually or together, satisfy an information need. A request may take different forms, for example, as Keywords, a Dataset or a Query.

- *Keyword search* allows users to target data with minimal knowledge of dataset structure and relationships. Keyword search can apply to a dataset's header and/or body and/or annotations [7].
- *Dataset-driven search*, given a dataset perhaps from other search or navigation tasks, looks for other datasets with similar Header and/or Body values to identify *more data like this* [6]. A version of this type of search is the Table Union Search problem that prioritises unionable tables rather than considering other notions of similarity [8].
- *Query driven search* seeks to support more precise requests. For example, Aurum [11] provides a language for querying a model in which nodes are columns in a repository, and edges represent relationships.

Search results tend to be provided as a ranked list, and as a result techniques are evaluated for effectiveness using metrics such as *mean reciprocal rank* or *precision @ k*.

2.2 Dataset Navigation

Having identified a dataset, for example through Dataset Search, it is often useful to identify related datasets that may provide additional information or context. In practice, different types of relationship have been investigated, through:

- *Join Path Discovery,* which identifies relationships between datasets that can underpin joins. Joins depend on extensional similarity, so searching for candidate join paths involves identifying joinable values at the instance level (e.g., [30, 31]).
- *Semantic similarity,* which identifies similarity relationships between datasets that may not support a join, for example because of disjoint or inconsistent representations. Thus semantic similarity highlights where related concepts are represented in a repository (e.g., [10, 13]).
- *Domain Discovery,* which identifies dataset attributes that share a domain, where a domain is a collection of values that instantiate an application concept (e.g. [14, 32]). For example, *REM* and *U2* are both in the domain of *Rock Bands*.

The effectiveness of navigation techniques typically involves relating results to a manually produced ground truth, though note that some results may be subjective. For example, what is a suitable granularity for domains? Should *Rock Band* be a domain, or is the more general *Music Group* more suitable?

2.3 Dataset Annotation

In a large repository, there are likely to be a variety of naming conventions, for example because the data was originally produced by different publishers. Data annotation is the process of associating intensional or extensional data items with terms from a vocabulary or concepts from an ontology. Datasets may be annotated at different levels of detail:

- *Entity Annotation* acts at the instance level, to pin down the interpretation of a token or string (e.g., [17]), for example to make explicit if the string *Texas* is referring to the city (http://dbpedia.org/resource/Texas) or the Scottish alternative rock group (http://dbpedia.org/resource/Texas_(band)).
- *Semantic Type Annotation* associates a complete dataset with an annotation (e.g., [19, 33]). The semantic type of a table could be inferred from the entity annotations of its subject attribute, from column headers, etc.
- *Property Annotation* provides annotations to individual attributes, potentially with literal types (e.g., phone numbers) or semantic types (e.g., work phone numbers) (e.g., [19, 34]).

The effectiveness of techniques for inferring annotations tends to be measured using metrics such as precision and recall against benchmark datasets (e.g., [35]). While using external ontologies brings new evidence to dataset discovery and exploration tasks, effective annotation depends on ontology coverage, and there is a risk that the available ontology may not reflect the aspects of the domain that are relevant to the task at hand.

2.4 Schema Inference

In a large repository, there are often multiple datasets describing a single concept. Thus, if the schema of a repository is the union of the schemas of the datasets in the repository, there may be many more datasets in the schema of the repository than there are concepts in the represented repository. Thus schema inference over a collection of datasets infers a schema that represents the data in the repository, but is typically much smaller than the schema of the repository. Schema inference may support:

- *Querying*, in which case the aim is to infer a schema that precisely describes the underlying data, so that it can be

queried. This approach does not resolve representational inconsistencies, and generally assumes that the underlying documents are somewhat similar. This approach has primarily been explored for XML and JSON [36, 37].

- *Documenting*, in which case the aim is to identify recurring features in the underlying repository, such as many datasets representing *company* data, which can be grouped together as a single type in an inferred schema. It is common to infer a schema that broadly retains the same level of detail in sources, but abstracts over (some) representational inconsistencies (e.g., [38–40]).
- *Summarizing*, in which case the aim is to identify the most important features in the underlying repository and to present them as representative of the repository as a whole (e.g., [24, 41]).

For evaluation, Schema Inference for Querying tends to produce a schema that is correct in the sense that the inferred schema describes every dataset, so goals tend to relate to the size or strictness of the result. In contrast, for Documenting and Summarizing, the correct answer is likely to be subjective and thus human judgements on suitability may come into play.

3 SYNERGIES, SIMILARITIES, CHALLENGES AND OPPORTUNITIES

Although the functionalities in Section 2 are superficially distinct, we can identify overlaps and synergies between them. For example, *Dataset Search* can make use of relationships between datasets, so that a search can identify sets of tables that when joined may satisfy a search request [6]. Furthermore, *Schema Inference* may act over the results of a *Dataset Search*, rather than generating the potentially voluminous schema of a complete repository. In addition, the results of *Dataset Annotation* could be used to inform searches, or to identify which datasets are part of the same type in *Schema Inference*.

There has been a lot of good work on diverse topics / approaches. However, individual research results tend to be narrow and deep, and thus the extent to which the results are in a position to significantly advance the productivity of data scientists using data catalogs is not yet clear. Perhaps there is now a significant opportunity for deployment of results – the transfer of techniques from the research community to products, open source systems and data platforms. Such an activity may well also flush out research gaps and challenges. An orthogonal opportunity likely also exists to make fuller use of recent developments in AI, such as deep clustering [42], foundation models [29] and retrieval augmented generation [43].

4 SUPPLEMENTARY MATERIAL

The tutorial is complemented by:

- *Demonstration Software:* we make available a Python Notebook[1] that allows experimentation with (original or reimplemented versions of) techniques that support all of:
  - *dataset search*, specifically D^3L [6].
  - *data navigation*, specifically *Aurum* [11].
  - *dataset annotation*, specifically *TableMiner+* [19].
  - *schema inference*, specifically clustering over similarities derived using *Starmie* [44].
- *A survey:* we have a paper in ACM Computing Surveys on *Dataset Discovery and Exploration* [45] that provides a systematic comparison of proposals in each of *dataset search*, *data navigation*, *dataset annotation* and *schema inference* at a level of detail well beyond that provided here.

[1] https://github.com/PierreWoL/EDBTDemo

5 BIOGRAPHY

Norman Paton is a Professor of Computer Science at the University of Manchester, where he has held a variety of roles including as Head of School. Prior to working at Manchester, he was a lecturer at Heriot-Watt University (Edinburgh) and obtained a PhD from Aberdeen University. His current research interests relate to data discovery, exploration and integration, with an emphasis on the use of AI techniques to inform automation. He is on the Editorial Boards of Philosophical Transactions of the Royal Society A and Distributed & Parallel Databases, and has had Track/PC Chair roles at CIKM, EDBT, ICDE and VLDB. He is a member of Academia Europaea.

Zhenyu Wu is a Ph.D. student in the Department of Computer Science at the University of Manchester, prior to which she was a Research Assistant at the Singapore University of Technology and Design. Her research interests lie in data exploration in data lakes. She holds a Master's degree in High Performance Computing with Data Science from the University of Edinburgh.

ACKNOWLEDGEMENTS

Zhenyu Wu is supported by a Dean's Scholarship from The University of Manchester. We are grateful for an ongoing collaboration in this space with Jiaoyan Chen.

REFERENCES

[1] G. Press, "Cleaning big data: Most time-consuming, least enjoyable data science task, survey says," *Forbes*, 2016.
[2] R. Hai, C. Koutras, C. Quix, and M. Jarke, "Data lakes: A survey of functions and systems," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12 571–12 590, 2023. [Online]. Available: https://doi.org/10.1109/TKDE.2023.3270101
[3] J. Attard, F. Orlandi, S. Scerri, and S. Auer, "A systematic review of open government data initiatives," *Gov. Inf. Q.*, vol. 32, no. 4, pp. 399–418, 2015. [Online]. Available: https://doi.org/10.1016/j.giq.2015.07.006
[4] O. Lehmberg, D. Ritze, R. Meusel, and C. Bizer, "A large public corpus of web tables containing time and context metadata," in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, 2016, pp. 75–76. [Online]. Available: https://doi.org/10.1145/2872518.2889386
[5] D. Brickley, M. Burgess, and N. Noy, "Google dataset search: Building a search engine for datasets in an open web ecosystem," in *WWW*. ACM, 2019, p. 1365–1375. [Online]. Available: https://doi.org/10.1145/3308558.3313685
[6] A. Bogatu, A. A. A. Fernandes, N. W. Paton, and N. Konstantinou, "Dataset discovery in data lakes," in *36th IEEE ICDE*. IEEE, 2020, pp. 709–720.
[7] A. Chapman, E. Simperl, L. Koesten, G. Konstantinidis, L. Ibáñez, E. Kacprzak, and P. Groth, "Dataset search: a survey," *VLDB J.*, vol. 29, no. 1, pp. 251–272, 2020. [Online]. Available: https://doi.org/10.1007/s00778-019-00564-x
[8] F. Nargesian, E. Zhu, K. Q. Pu, and R. J. Miller, "Table union search on open data," *Proc. VLDB Endow.*, vol. 11, no. 7, pp. 813–825, 2018. [Online]. Available: http://www.vldb.org/pvldb/vol11/p813-nargesian.pdf
[9] A. Khatiwada, G. Fan, R. Shraga, Z. Chen, W. Gatterbauer, R. J. Miller, and M. Riedewald, "SANTOS: relationship-based semantic table union search," *Proc. ACM Manag. Data*, vol. 1, no. 1, pp. 9:1–9:25, 2023. [Online]. Available: https://doi.org/10.1145/3588689
[10] R. C. Fernandez, E. Mansour, A. A. Qahtan, A. K. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, M. Stonebraker, and N. Tang, "Seeping semantics: Linking datasets using word embeddings for data discovery," in *34th IEEE ICDE*, 2018, pp. 989–1000. [Online]. Available: https://doi.org/10.1109/ICDE.2018.00093
[11] R. C. Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, and M. Stonebraker, "Aurum: A data discovery system," in *34th IEEE ICDE*, 2018, pp. 1001–1012. [Online]. Available: https://doi.org/10.1109/ICDE.2018.00094
[12] K. Li, Y. He, and K. Ganjam, "Discovering enterprise concepts using spreadsheet tables," in *Proc. 23rd ACM SIGKDD*, 2017, pp. 1873–1882. [Online]. Available: https://doi.org/10.1145/3097983.3098102
[13] F. Nargesian, K. Q. Pu, E. Zhu, B. G. Bashardoost, and R. J. Miller, "Organizing data lakes for navigation," in *Proc. ACM SIGMOD*, 2020, pp. 1939–1950.

[14] M. Ota, H. Mueller, J. Freire, and D. Srivastava, "Data-driven domain discovery for structured datasets," *Proc. VLDB Endow.*, vol. 13, no. 7, pp. 953–965, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol13/p953-ota.pdf

[15] C. S. Bhagavatula, T. Noraset, and D. Downey, "TabEL: Entity Linking in Web Tables," in *Proc. 14th ISWC, Part I*, ser. LNCS, vol. 9366. Springer, 2015, pp. 425–441. [Online]. Available: https://doi.org/10.1007/978-3-319-25007-6_25

[16] J. Chen, E. Jiménez-Ruiz, I. Horrocks, and C. Sutton, "Learning semantic annotations for tabular data," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 2088–2094. [Online]. Available: https://doi.org/10.24963/ijcai.2019/289

[17] X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, "TURL: table understanding through representation learning," *Proc. VLDB Endow.*, vol. 14, no. 3, pp. 307–319, 2020. [Online]. Available: http://www.vldb.org/pvldb/vol14/p307-deng.pdf

[18] V. Efthymiou, O. Hassanzadeh, M. Rodriguez-Muro, and V. Christophides, "Matching web tables with knowledge base entities: From entity lookups to entity embeddings," in *The Semantic Web - ISWC Part I*, ser. LNCS, vol. 10587. Springer, 2017, pp. 260–277. [Online]. Available: https://doi.org/10.1007/978-3-319-68288-4_16

[19] Z. Zhang, "Effective and efficient semantic table interpretation using tableminer$^+$," *Semantic Web*, vol. 8, no. 6, pp. 921–957, 2017. [Online]. Available: https://doi.org/10.3233/SW-160242

[20] M. A. Baazizi, D. Colazzo, G. Ghelli, and C. Sartiani, "Parametric schema inference for massive JSON datasets," *VLDB J.*, vol. 28, no. 4, pp. 497–521, 2019. [Online]. Available: https://doi.org/10.1007/s00778-018-0532-7

[21] S. Cebiric, F. Goasdoué, H. Kondylakis, D. Kotzinos, I. Manolescu, G. Troullinou, and M. Zneika, "Summarizing semantic graphs: a survey," *VLDB J.*, vol. 28, no. 3, pp. 295–327, 2019. [Online]. Available: https://doi.org/10.1007/s00778-018-0528-3

[22] N. Kardoulakis, K. Kellou-Menouer, G. Troullinou, Z. Kedad, D. Plexousakis, and H. Kondylakis, "Hint: Hybrid and incremental type discovery for large RDF data sources," in *SSDBM 2021: 33rd International Conference on Scientific and Statistical Database Management*, 2021, pp. 97–108. [Online]. Available: https://doi.org/10.1145/3468791.3468808

[23] K. Kellou-Menouer, N. Kardoulakis, G. Troullinou, Z. Kedad, D. Plexousakis, and H. Kondylakis, "A survey on semantic schema discovery," *The VLDB Journal*, vol. 31, p. 675–710, 2022.

[24] C. Yu and H. V. Jagadish, "Schema summarization," in *Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006*, 2006, pp. 319–330. [Online]. Available: http://dl.acm.org/citation.cfm?id=1164156

[25] M. Esmailoghli, C. Schnell, R. J. Miller, and Z. Abedjan, "Blend: A unified data discovery system," 2023.

[26] A. Bogatu, N. W. Paton, M. Douthwaite, and A. Freitas, "Voyager: Data discovery and integration for onboarding in data science," in *Proc. 25th EDBT*, 2022, pp. 2:537–2:548. [Online]. Available: https://doi.org/10.48786/edbt.2022.47

[27] C. Liu, A. Usta, J. Zhao, and S. Salihoglu, "Governor: Turning open government data portals into interactive databases," in *Proc CHI*, 2023, pp. 415:1–415:16. [Online]. Available: https://doi.org/10.1145/3544548.3580868

[28] F. Torregrossa, R. Allesiardo, V. Claveau, N. Kooli, and G. Gravier, "A survey on training and evaluation of word embeddings," *Int. J. Data Sci. Anal.*, vol. 11, no. 2, pp. 85–103, 2021. [Online]. Available: https://doi.org/10.1007/s41060-021-00242-8

[29] A. Narayan, I. Chami, L. J. Orr, and C. Ré, "Can foundation models wrangle your data?" *Proc. VLDB Endow.*, vol. 16, no. 4, pp. 738–746, 2022. [Online]. Available: https://www.vldb.org/pvldb/vol16/p738-narayan.pdf

[30] E. Zhu, D. Deng, F. Nargesian, and R. J. Miller, "Josie: Overlap set similarity search for finding joinable tables in data lakes," in *Proc. ACM SIGMOD*. ACM, 2019, p. 847–864. [Online]. Available: https://doi.org/10.1145/3299869.3300065

[31] E. Zhu, Y. He, and S. Chaudhuri, "Auto-join: Joining tables by leveraging transformations," *Proc. VLDB Endow.*, vol. 10, no. 10, pp. 1034–1045, 2017.

[32] F. Piai, P. Atzeni, P. Merialdo, and D. Srivastava, "Fine-grained semantic type discovery for heterogeneous sources using clustering," *VLDB J.*, vol. 32, no. 2, pp. 305–324, 2023. [Online]. Available: https://doi.org/10.1007/s00778-022-00743-3

[33] J. Chen, E. Jiménez-Ruiz, I. Horrocks, and C. Sutton, "Colnet: Embedding the semantics of web tables for column type prediction," in *The 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, The 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*. AAAI Press, 2019, pp. 29–36. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.330129

[34] S. Neumaier, J. Umbrich, J. X. Parreira, and A. Polleres, "Multi-level semantic labelling of numerical values," in *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*, 2016, pp. 428–445. [Online]. Available: https://doi.org/10.1007/978-3-319-46523-4_26

[35] D. Ritze, O. Lehmberg, and C. Bizer, "Matching HTML tables to dbpedia," in *Proc. 5th Intl Conf on Web Intelligence, Mining and Semantics, WIMS*. ACM, 2015, pp. 10:1–10:6. [Online]. Available: https://doi.org/10.1145/2797115.2797118

[36] M. A. Baazizi, H. Ben Lahmar, D. Colazzo, G. Ghelli, and C. Sartiani, "Schema inference for massive JSON datasets," in *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017*. OpenProceedings.org, 2017, pp. 222–233. [Online]. Available: https://doi.org/10.5441/002/edbt.2017.21

[37] G. J. Bex, F. Neven, and S. Vansummeren, "Inferring XML schema definitions from XML data," in *Proc. 33rd VLDB*, 2007, pp. 998–1009. [Online]. Available: http://www.vldb.org/conf/2007/papers/research/p998-bex.pdf

[38] N. Barret, I. Manolescu, and P. Upadhyay, "Computing generic abstractions from application datasets," in *Proceedings 27th International Conference on Extending Database Technology, EDBT 2024, Paestum, Italy, March 25 - March 28*, 2024, pp. 94–107. [Online]. Available: https://doi.org/10.48786/edbt.2024.09

[39] A. Bonifati, S. Dumbrava, and N. Mir, "Hierarchical clustering for property graph schema discovery," in *Proc. 25th EDBT*, 2022, pp. 2:449–2:453. [Online]. Available: https://doi.org/10.48786/edbt.2022.39

[40] K. Christodoulou, N. W. Paton, and A. A. A. Fernandes, "Structure inference for linked data sources using clustering," *Trans. Large Scale Data Knowl. Centered Syst.*, vol. 19, pp. 1–25, 2015. [Online]. Available: https://doi.org/10.1007/978-3-662-46562-2_1

[41] X. Yang, C. M. Procopiuc, and D. Srivastava, "Summarizing relational databases," *Proc. VLDB Endow.*, vol. 2, no. 1, pp. 634–645, 2009. [Online]. Available: http://www.vldb.org/pvldb/vol2/vldb09-784.pdf

[42] H. T. Rauf, A. Freitas, and N. W. Paton, "Deep clustering for data cleaning and integration," in *Proceedings 27th International Conference on Extending Database Technology, EDBT*, 2024.

[43] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, Q. Guo, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024.

[44] G. Fan, J. Wang, Y. Li, D. Zhang, and R. J. Miller, "Semantics-aware dataset discovery from data lakes with contextualized column-based representation learning," *Proc. VLDB Endow.*, vol. 16, no. 7, pp. 1726–1739, 2023. [Online]. Available: https://www.vldb.org/pvldb/vol16/p1726-fan.pdf

[45] N. W. Paton, J. Chen, and Z. Wu, "Dataset discovery and exploration: A survey," *ACM Comput. Surv.*, vol. 56, no. 4, 2024. [Online]. Available: https://doi.org/10.1145/3626521