Data Integration and Machine Learning: A Natural Synergy

Xin Luna Dong
lunadong@amazon.com
Amazon
Seattle, WA, USA

Theodoros Rekatsinas
thodrek@cs.wisc.edu
University of Wisconsin-Madison
Madison, WI, USA

ABSTRACT

As data volume and variety have increased, so have the ties between machine learning and data integration become stronger. For machine learning to be effective, one must utilize data from the greatest possible variety of sources; and this is why data integration plays a key role. At the same time machine learning is driving automation in data integration, resulting in overall reduction of integration costs and improved accuracy. This tutorial focuses on three aspects of the synergistic relationship between data integration and machine learning: (1) we survey how state-of-the-art data integration solutions rely on machine learning-based approaches for accurate results and effective human-in-the-loop pipelines, (2) we review how end-to-end machine learning applications rely on data integration to identify accurate, clean, and relevant data for their analytics exercises, and (3) we discuss open research challenges and opportunities that span across data integration and machine learning.

ACM Reference Format:
Xin Luna Dong and Theodoros Rekatsinas. 2019. Data Integration and Machine Learning: A Natural Synergy. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332296

1 TARGET AUDIENCE

This tutorial targets all researchers and practitioners interested in data quality challenges in end-to-end data science pipelines. The goal is to inform the audience about the class of problems that exist in the intersection of data integration and machine learning as well as recent breakthroughs that are results of the synergistic effect between the two. We also aim to motivate further research in the area of ML-based data integration solutions. We assume general familiarity with common ML terms but do not require prior knowledge of specific algorithms or system internals.

2 BIOGRAPHICAL SKETCHES OF TUTORS

In-person presenters: Xin Luna Dong, Theodoros (Theo) Rekatsinas
Corresponding tutor: Theo (thodrek@cs.wise.edu)

Xin Luna Dong is a Principal Scientist at Amazon, leading the efforts of constructing Amazon Product Knowledge Graph. She was one of the major contributors to the Google Knowledge Vault project, and has led the Knowledge-based Trust project, which is called the "Google Truth Machine" by Washington's Post. She has co-authored book "Big Data Integration", was awarded ACM Distinguished Member, VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion", and Best Demo award in SIGMOD 2005. She serves in VLDB endowment and PVLDB advisory committee, and is a PC co-chair for VLDB 2021, ICDE Industry 2019, VLDB Tutorial 2019, SIGMOD 2018 and WAIM 2015. She has given more than 10 tutorials on data integration, knowledge collection, and graph mining in top-tier conferences.

Theodoros (Theo) Rekatsinas is an Assistant Professor in the Department of Computer Sciences at the University of Wisconsin-Madison. He is a member of the Database Group. He earned his Ph.D. in Computer Science from the University of Maryland and was a Moore Data Postdoctoral Fellow at Stanford University. His research interests are in data management, with a focus on data integration, data cleaning, and uncertain data. Theo's work has been recognized with an Amazon Research Award in 2018, a Best Paper Award at SDM 2015, and the Larry S. Davis Doctoral Dissertation award in 2015.

3 OUTLINE

This 3-hour tutorial is split into three parts:

(1) **A DI and ML primer**: In this introductory part of the tutorial, we review the problems that constitute a typical data integration stack [5]: (1) data extraction, (2) schema alignment, (3) entity resolution, and (4) data fusion. We also discuss ML-related concepts, including supervised, semi-supervised, and unsupervised learning setups, pertinent to ML-based solution for data integration. We also review components of typical end-to-end ML-based analytics to introduce parts for which data integration solutions are key.
(2) **ML solutions for automated DI**: In the first technical part of the tutorial, we focus on classical problems along the data integration stack. We motivate a *ML-based view* for these problems and review algorithmic frameworks and systems that build upon machine learning methods to introduce automated solutions for each of these problems.
(3) **DI for effective ML pipelines**: In the second technical part of the tutorial, we review how data integration tasks form critical parts of modern machine learning and play a crucial role in obtaining highly accurate results. We focus on two tasks that form the major bottlenecks in any machine learning pipeline: creation of large-scale training datasets, and cleaning of the data used for training or inference.
(4) **Future opportunities**: Finally, we outline several open research problems as potential directions for new research in this area.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3332296

REFERENCES

[1] P. Bailis, E. Gan, S. Madden, D. Narayanan, K. Rong, and S. Suri. Macrobase: Prioritizing attention in fast data. In *SIGMOD*, pages 541–556, 2017.
[2] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. Data cleaning: Overview and emerging challenges. In *SIGMOD*, pages 2201–2206, 2016.
[3] R. Das, A. Neelakantan, D. Belanger, and A. McCallum. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *EACL*, 2017.
[4] S. Das, P. S. G. C., A. Doan, J. F. Naughton, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, and Y. Park. Falcon: Scaling up hands-off crowdsourced entity matching to build cloud services. In *SIGMOD*, pages 1431–1446, 2017.
[5] A. Doan, A. Y. Halevy, and Z. G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.
[6] X. L. Dong. Challenges and innovations in building a product knowledge graph. In *SigKDD*, 2018.
[7] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *SIGKDD*, 2014.
[8] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. *PVLDB*, 2014.
[9] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. In *VLDB*, 2015.
[10] X. L. Dong and F. Naumann. Data fusion–resolving data conflicts for integration. *PVLDB*, 2009.
[11] X. L. Dong and D. Srivastava. Big data integration. *Proc. VLDB Endow.*, 6(11):1188–1189, Aug. 2013.
[12] X. L. Dong and D. Srivastava. Big data integration. *Synthesis Lectures on Data Management*, 7(1):1–198, 2015.
[13] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the Americal Statistical Association*, 64(328):1183–1210, 1969.
[14] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. pages 371–380, 2001.
[15] J. Gao, Q. Li, B. Zhao, W. Fan, and J. Han. Mining reliable information from passively and actively crowdsourced data. In *KDD*, pages 2121–2122, 2016.
[16] L. Getoor and A. Machanavajjhala. Entity resolution: theory, practice & open challenges. *PVLDB*, 5(12):2018–2019, 2012.
[17] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu. Corleone: Hands-off crowdsourcing for entity matching. In *SIGMOD*, pages 601–612, 2014.
[18] P. Gulhane, A. Madaan, R. Mehta, J. Ramamirtham, R. Rastogi, S. Satpal, srinivasan H. Sengamedu, A. Tengli, and C. Tiwari. Web-scale information extraction with vertex. In *ICDE*, 2011.
[19] A. Halevy, F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, and S. E. Whang. Goods: Organizing google's datasets. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 795–806, New York, NY, USA, 2016. ACM.
[20] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, Mar. 2009.
[21] O. Hassanzadeh, F. Chiang, R. J. Miller, and H. C. Lee. Framework for evaluating clustering algorithms in duplicate detection. *PVLDB*, 2(1):1282–1293, 2009.
[22] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier. An unsupervised neural attention model for aspect extraction. In *ACL*, 2017.
[23] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, 2011.
[24] H. Ji. Entity linking and wikification reading list. http://nlp.cs.rpi.edu/kbp/2014/elreading.html, 2014.
[25] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, 1998.
[26] H. Kopcke, A. Thor, and E. Rahm. Evaluation of entity resolution approaches on real-world match problems. *PVLDB*, 3(1):484–493, 2010.
[27] S. Krishnan, J. Wang, E. Wu, M. J. Franklin, and K. Goldberg. Activeclean: Interactive data cleaning for statistical modeling. *Proc. VLDB Endow.*, 9(12):948–959, Aug. 2016.
[28] A. Kumar, M. Boehm, and J. Yang. Data management in machine learning: Challenges, techniques, and systems. SIGMOD '17, pages 1717–1722, 2017.
[29] X. Li, X. L. Dong, K. B. Lyons, W. Meng, and D. Srivastava. Truth finding on the Deep Web: Is the problem solved? *PVLDB*, 6(2), 2013.
[30] C. Lockard, X. L. Dong, A. Einolghozati, and P. Shiralkar. Ceres: Distantly supervised relation extraction from the semi-structured web. In *PVLDB*, 2018.
[31] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *ACL*, 2016.
[32] C. Manning. Representations for language: From word embeddings to sentence meanings. https://simons.berkeley.edu/talks/christopher-manning-2017-3-27, 2017.
[33] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL*, 2009.
[34] T. Mitchell. Learning from limited labeled data (but a lot of unlabeled data). https://lld-workshop.github.io/slides/tom_mitchell_lld.pdf, 2017.
[35] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, and J. Welling. Never-ending learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.
[36] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, and V. Raghavendra. Deep learning for entity matching: A design space exploration. In *Sigmod*, 2018.
[37] A. Neelakantan, B. Roth, and A. McCallum. Compositional vector space models for knowledge base completion. In *ACL*, 2015.
[38] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 689–696, USA, 2011. Omnipress.
[39] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*, pages 877–885, 2010.
[40] R. Pimplikar and S. Sarawagi. Answering table queries on the web using column keywords. *PVLDB*, 5(10):908–919, 2012.
[41] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data management challenges in production machine learning. In *SIGMOD*, pages 1723–1726, 2017.
[42] J. Pujara and L. Getoor. Generic statistical relational entity resolution in knowledge graphs. In *AAAI*, 2016.
[43] A. Ratner, S. H. Bach, H. R. Ehrenberg, J. A. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. *PVLDB*, 11(3):269–282, 2017.
[44] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, and C. Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, pages 3567–3575, 2016.
[45] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *J. Mach. Learn. Res.*, 11:1297–1322, Aug. 2010.
[46] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. Holoclean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(11):1190–1201, 2017.
[47] T. Rekatsinas, M. Joglekar, H. Garcia-Molina, A. Parameswaran, and C. Ré. Slimfast: Guaranteed results for data fusion and source reliability. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 1399–1414, New York, NY, USA, 2017. ACM.
[48] S. Riedel, L. Yao, B. M. Marlin, and A. McCallum. Relation extraction with matrix factorization and universal schemas. In *HLT-NAACL*, 2013.
[49] B. Saha and D. Srivastava. Data quality: The other face of big data. In *ICDE*, pages 1294–1297, 2014.
[50] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *SIGKDD*, 2002.
[51] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. Data curation at scale: The data tamer system. In *CIDR*, 2013.
[52] K. Toutanova, D. Chen, P. Pantel, H. Poon, P. Choudhury, and M. Gamon. Representing text for joint embedding of text and knowledge bases. In *EMNLP*, 2015.
[53] R. Trivedi, B. Sisman, J. Ma, C. Faloustos, H. Zha, and X. L. Dong. Linknbed: Multi-graph representation learning with entity linkage. In *ACL*.
[54] P. Verga, A. Neelakantan, and A. McCallum. Generalizing to unseen entities and entity pairs with row-less universal schema. In *ACL*, 2017.
[55] V. Verroios, H. Garcia-Molina, and Y. Papakonstantinou. Waldo: An adaptive human interface for crowd entity resolution. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1133–1148, 2017.
[56] X. Wang, X. L. Dong, and A. Meliou. Data x-ray: A diagnostic tool for data errors. In *SIGMOD*, pages 1231–1245, 2015.
[57] S. Wu, L. Hsiao, X. Cheng, B. Hancock, T. Rekatsinas, P. Levis, and C. Ré. Fonduer: Knowledge base construction from richly formatted data. In *SIGMOD*, SIGMOD '18, 2018.
[58] C. Zhang, C. RÃČÂĪ, M. Cafarella, C. D. Sa, A. Ratner, J. Shin, F. Wang, and S. Wu. Deepdive: Declarative knowledge base construction. *CACM*, 60(5):93–102, 2017.
[59] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li. Opentag: Open attribute value extraction from product profiles. In *SigKDD*, 2018.