Data Integration and Machine Learning: A Natural Synergy

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

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

ABSTRACT

There is now more data to analyze than ever before. 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.

KEYWORDS

Data integration, Machine learning

ACM Reference Format:
Xin Luna Dong and Theodoros Rekatsinas. 2018. Data Integration and Machine Learning: A Natural Synergy. In *Proceedings of 2018 International Conference on Management of Data (SIGMOD'18).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3183713.3197387

1 INTRODUCTION

The ties between data integration (DI) and machine learning (ML) have always been apparent [6]. However, the sheer volume and variety of data consumed by modern analytical pipelines have greatly strengthened the connections between data integration and machine learning. Data integration systems are increasingly looking to use machine learning to automate parts of different integration tasks. Examples include data cataloging and inferring the schema of raw data [19], data alignment [17], and transformation recommendations for data normalization [13]. At the same time, machine learning algorithms are only as good as the data used for training [20], which means that one must utilize data from the greatest possible variety of sources. Data integration and machine learning making each other more effective is a true example of a powerful synergy.

Goal. The goal of this tutorial is to delineate the interplay between modern data integration techniques and modern machine learning. Specifically, we review (1) how recent advancements in machine learning (such as highly-scalable inference engines and deep learning) are revolutionizing data integration, and (2) how incorporating data integration tasks in machine learning pipelines leads to more accurate and usable systems for analytics. This tutorial will highlight the strong connections between data integration and machine learning, review related technical challenges and recent solutions, and outline open problems that remain to be solved.

Scope. This tutorial highlights how recent advancements in machine learning are shaping the area of data integration and highlights the role of data integration methods in modern machine learning pipelines. In contrast to previous tutorials that either focus on specific DI problems such as entity resolution [17] and data fusion [11] or discuss large-scale DI [12], we review machine learning-based solutions (including deep learning solutions) that are revolutionizing solutions across the entire data integration stack. In addition, our tutorial touches aspects of data cleaning—a problem closely related to data integration. Specifically, we focus on recent data cleaning solutions that adopt statistical semantics and can used to address data preparation challenges in machine learning pipelines. This is in contrast to tutorials on data cleaning that put emphasis on rule-based approaches [3, 47]. Finally, our tutorial is complementary to tutorials that review broader data management challenges pertinent to machine learning such as data exploration, feature engineering, and scalable model serving [28, 39].

Target Audience. This tutorial targets all researchers and practitioners interested in data quality and data management 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.

Outline. This 1.5-hour tutorial is split into three parts:

*The secretary disavows any knowledge of this author's actions.

SIGMOD'18, June 10–15, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-4703-7/18/06…$15.00
https://doi.org/10.1145/3183713.3197387

(1) **A DI and ML primer**: In this introductory part of the tutorial, we review the problems that constitute a typical data integration stack [6]: (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 (Section 2).
(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 (Section 3).
(4) **Future opportunities**: Finally, we outline several open research problems as potential directions for new research in this area (Section 4).

2 ML SOLUTIONS FOR AUTOMATED DI

We start describing the synergy from the perspective of data integration. The data integration field applied ML techniques from its beginning for understanding semantics of data, and aligning schema and entities. Recent progress in ML significantly improved the results, and is making revolutionary changes to this field.

We describe how ML has been reshaping different tasks of data integration. We choose not to follow the ordering of the different tasks in the data integration stack, but follow the importance of the tasks in seamlessly integrating data from different sources. As we describe the techniques, we emphasize the difference of the techniques applied 10 years ago and applied now and describe how the technique shifting has influenced data integration results.

2.1 ML for Entity Resolution

Entity resolution identifies records that refer to the same real-world entity. It was almost 50 years since entity resolution was first proposed [14]. It is an unavoidable and arguably the most important problem in integrating data from different sources. Whereas schema alignment is important too, it can often be solved manually because the size of a schema is typically small; in contrast, we often need to match at least thousands of entities from different sources, making manual solutions seldom an option.

Entity resolution proceeds in three steps: (1) blocking records that are likely to refer to the same real world entity; (2) comparing pairs of records to decide if it is a match; and (3) clustering records according to pairwise matching results, such that each cluster corresponds to a real-world entity. For a long while entity resolution is solved using rule-based methods or unsupervised learning: blocking is rule based (e.g., blocking person records by name and phone numbers); pairwise matching is rule-based (often through a linear combination of attribute similarities) [14, 15]; and clustering is rule-based (e.g., transitive closure, MERGE-CENTER) or by optimizing a particular objective function (e.g., Markov clustering and correlation clustering) [21], without any need of training data.

Supervised learning approaches to DI started about 20 years ago, and used Decision trees, Logistic regression, and SVM until early 2010's. ML-based models typically compute attribute-wise value similarity and use that as features. The survey by Kopcke et al. [26] shows that early supervised approaches such as SVM and Decision tree with 500 training labels obtain similar results with rule-based methods: $\sim$ 90% F-measure for easy data sets (e.g., Bibliography) and $\sim$ 70% F-measure for harder ones (e.g., E-commerce). Recent ML models, such as Random Forest, significantly improved pairwise matching. Das et al. [5] show that training Random Forest on around 1,000 labels can obtain $\sim$ 95% F-measure for easy data sets, and $\sim$ 80% F-measure for harder data sets. Deep learning allows comparing long text values by their embedding representations (e.g., Word2Vec), and starts to show promise when matching texts and dirty data. Finally, logic-based learning methods (e.g., probabilistic soft logic) enable linking entities of multiple types at the same time, called *collective linkage* [40].

The high precision and recall achieved recently for entity resolution make it ready for industry production. However, the good performance comes with a cost: the cost of generating training labels. A recent study shows that obtaining a precision of 99% and recall of 99% (required for production) on linking a pair of fairly clean data sets requires 1.5M training labels [7]. This is only for one type of entities from two data sources; no need to mention multiple types and multiple sources. This challenge motivates research on active learning to collect training labels [5, 48].

2.2 ML for Data Fusion

Data fusion resolves conflicts from different data sources. Li et al. [29] show that even in quality-sensitive domains (e.g., stock, flight), authoritative sources can provide conflicting and erroneous values. Access to highly accurate data is critical for industry applications, such as knowledge graph search, so data fusion is often an important step in data integration. Recently, it evolves to *knowledge fusion* [8–10] to clean both data and extraction errors, playing an important role in automatic knowledge graph construction.

Data fusion also started with rule-based methods [11], such as averaging and voting, and data mining methods, such as HITS [25, 37]. The large body of work on data fusion resorts to Graphical model [10, 16, 29] to model the relationship between data correctness, source accuracy, and source correlation (e.g., copy relationship) and uses EM to obtain the solution. It is mainly unsupervised learning, but can also leverage ground truths in parameter initialization so allows semi-supervised learning. Recently, SLiMFast[45] is proposed as a *discriminative* model that also enables considering other features of data sources (e.g., update date, number of citations, etc.) for fusion; in presence of sufficient labeled data SLiMFast uses empirical risk minimization (ERM).

Table 1: Summary of ML techniques used for data integration.

DI tasks	Hyperplanes (e.g., Log Reg)	Kernal (e.g., SVM)	Tree-based (e.g., Random forest)	Graphical models (e.g., CRF)	Logic programs (e.g, soft logic)	Neural networks (e.g., RNN)
Entity resolution	X	X	X		X	X
Data fusion	X			X		
DOM extraction	X					
Text extraction	X			X		X
Schema alignment	X			X		X

2.3 ML for Data Extraction

Data extraction allows obtaining structured data from unstructured data such as texts, and semi-structured data such as Web DOM trees. Entity linkage and data fusion techniques make it possible to align data extracted from different sources. Up to now extraction results are still error-prone, so not fully ready for industry use; however, latest ML progress has brought revolutionary changes to extraction techniques.

We first discuss extraction for semi-structured data, which is shown to contribute to 80% knowledge extracted by Knowledge Vault [8] from the web. A decade ago extraction from semi-structured data is mainly conducted by *wrapper induction*; that is, based on annotations on a few webpages from a website, inducing the XPaths that can extract values of given attributes from the whole website [1]. This method requires limited annotations for a website, but each website requires its own annotations, making it infeasible for extraction from the whole website.

Recently, distant supervision is applied to extraction from semi-structured data. Distant supervision was originally applied to texts: instead of manually creating labels, distant supervision leverages existing seed data to automatically create (oftentimes noisy) labels, and enables learning on such labels [23, 32]. Applying the same techniques on semi-structured data is able to extract 1.3M (entity, attribute, value) knowledge triples from the web, with an accuracy of $\sim$ 60% [8], and this accuracy is improved to over 90% [7].

Text extraction is hard because of the whims in composing texts in expressing a meaning. Early techniques rely on lexical and syntactic features extracted from texts. These features are used to train logistic regression first [32], later CRF to model correlation between attributes [23], and then Markov logic network to allow rule specification [52, 53]. RNNs and word embeddings have enabled deep understanding of texts without much, if any, feature engineering [31]. Recently, Bi-LSTM and attention are combined with other models to significantly improve text extraction [22, 30].

2.4 ML for Schema Alignment

Schema alignment matches types and attributes. It is one of the first problems studied for data integration and adopted ML techniques from the beginning, such as Naive Bayes and stacking [46]. Although automatic schema mapping seems an overkill when we align data between two data sources with typical sizes of schemas, it is important when we consider millions of sources from the web. For example, Pimplikar et al. [38] studies how to apply graphical model to align webtables with knowledge bases, by aligning entities and schemas at the same time.

Universal schema [46] has revolutionized schema alignment. It is motivated by OpenIE knowledge extraction: unlike traditional information extraction that extracts knowledge according to a predefined ontology (i.e., schema), OpenIE extracts (subject, predicate, object) triples, where the predicate can be any word or phrase from texts. Reasoning over the predicates and mapping them to existing ontology predicates is important to broaden applications for OpenIE results. Such relationships can be asymmetric; for example, "employed_by" can be inferred from "teach at", but not vice versa. Universal schema is proposed for this purpose: instead of outputting mappings between predicates, it adds inferred triples. Original solutions for universal schema leverages matrix factorization [46]; recently, it is improved using RNN and can infer a relationship be composing two or more relationships (e.g., "Melinda-*lives-in*-Seattle" can be inferred from "Melinda-*spouse*-Bill-*Chairman*-Microsoft-*HQ-in*-Seattle") [4, 35].

Summary: Finally, we summarize the types of ML models that have been explored for various DI tasks in the literature using Table 1. We also highlight that despite the progress, one critical challenge for DI is to create training labels for many different cases from many different data sources, and point out research directions to address this challenge at the end of our tutorial.

3 DI FOR EFFECTIVE ML PIPELINES

We now review how DI methods help address two major bottlenecks in ML pipelines: training data generation and data cleaning. We begin this part by focusing on the creation of large volumes of training data. Specifically, we discuss how DI solutions pertinent to entity resolution and data fusion play a key role in recent state-of-the-art methods for creating large training datasets fast. We then focus on the second biggest bottleneck in modern ML pipelines: cleaning dirty data to ensure high-quality predictions. Here, we review state-of-the-art data cleaning approaches that only require lightweight human input to perform data cleaning.

3.1 Creation of Large-Scale Training Data

Machine learning pipelines rely heavily on labeled training data to achieve high quality. The collection of labeling data via manual annotation can be a particularly tedious and non-scalable process. This has motivated a series of recent works around the paradigm of *weak supervision* with the goal to use higher-level and noisier input from experts and heterogeneous data sources to train ML systems [33].

Various forms of weak supervision have been studied in the literature but the most effective techniques focus on *weak labels*,

i.e., a set of noisy labeled examples. These may come from crowd workers [43], be the output of heuristic rules [42], or the result of distant supervision [32], where one or multiple external knowledge bases are heuristically used to guide the training of an ML system. *DI techniques play an instrumental role in all aforementioned weak supervision methods.*

Distant supervision relies on *entity linking* [24], a task similar to that of entity resolution (see Section 2.1) to match facts from a knowledge base to corresponding *mentions* in the input data. Specifically, for the problem of relation extraction from textual data, distant supervision relies on the same text-similarity metrics as entity resolution to compute the similarity between text excerpts from the input corpus to fact references in a knowledge base. Distance supervision requires that a DI task is solved accurately so that high-quality training data is obtained.

The other two forms of weak supervision that involve collecting labels from weak sources (i.e., crowd workers or heuristic rules), are closely related to *data fusion* (see Section 2.2). Here, one needs to deal with weak sources that are noisy, can provide conflicting labels, and might be highly correlated. To deal with the uncertain labels obtained by weak sources, state-of-the-art frameworks built around weak supervision, such as Snorkel [41] and NELL [34] entail three tasks: They (1) learn the accuracy of each weak supervision source by leveraging the agreement and disagreement across different labeling, (2) they model the correlations of weak supervision sources by employing structure learning techniques, and (3) they model the expertise of different sources of weak supervision for specific data inputs. Notice, that all three above tasks are integral to data fusion [11] and methods developed by the database community are directly applicable.

3.2 Data Cleaning

State-of-the-art ML pipelines rely heavily on the high-effort task of data cleaning to obtain high-quality results [39]. Data cleaning is typically split into three tasks: (1) error detection, where data inconsistencies such as duplicate data, violations of logical constraints that assert the consistency of the data, and incorrect data values are identified, (2) data repairing, which involves updating the available data to remove any detected errors, and (3) data imputation, which derives and fills in missing data from existing data. In this part of the tutorial we focus on end-to-end frameworks for data cleaning that only rely on lightweight high-level human supervision to detect and repair data errors.

Specifically, we focus on a new breed of error detection and data repairing frameworks that has recently emerged in the database community and rely on statistical approaches to perform data cleaning. More precisely, we will describe systems such as Data X-ray [51] and MacroBase [2] that rely on quantitative statistics to identify unusual trends (i.e., outliers) in data, frameworks such as HoloClean [44] that employ statistical learning and probabilistic inference to repair errors in data, and approaches such as ActiveClean [27] that leverage sampling to perform on-demand data cleaning while targeting downstream machine learning models explicitly.

4 FUTURE OPPORTUNITIES

We finally discuss several open challenges and opportunities in the intersection of data integration and machine learning. The open problems we cover aim to attract the attention of the database community to how data integration and machine learning can help each other be more effective. By no means is this an exhaustive list.

Multi-modal DI. Traditionally data integration has focused on textual data. However, there is an abundance of image, sensory, and audio data that is rarely integrated with textual data into a common queryable knowledge repository. This is to a certain extent due to the inherently different methods required to process each aforementioned data mode. Nonetheless, state-of-the-art deep learning methods can potentially provide the necessary tools and formalisms required for multi-modal data integration. Recent results in multi-modal information extraction [52] and multi-modal deep learning [36] certainly provide positive evidence.

Fast and Cheap Training Data for DI. Machine learning models for data integration can require large amounts of training data when applied over domains with reach domain-specific semantics. Obtaining large number of training examples can be resource-intensive in many practical scenarios. Recent approaches in the database literature have focused on active learning methods to solicit human supervision more effectively [18, 49, 50]. A promising direction is to understand how these methods relate to weak supervision methods recently introduced in the machine learning community [42]. Overall, there is an immediate need for new algorithmic frameworks and systems for collecting large amounts of training data for DI more effectively.

Human-in-the-Loop DI. Machine learning models, or any other automatical approaches, can hardly obtain a 100% accuracy on DI, which is a very complex task. It is thus important to involve human in the loop, conducting labelling, verifications, and auditing. A future direction is for a system to automatically identify when, where, and how to get human involved, by applying active learning, transitive learning, and reinforcement learning.

Efficient Model Serving for DI. Model serving for DI entails several resource intensive operations such as data normalization and blocking before entity resolution or data fusion is performed. Existing methods execute each step in isolation without taking into account the computation performed in subsequent steps along the DI pipeline. Open questions here include abstractions that will enable RDBMS-style plan generation and optimization to serve DI models efficiently by avoiding redundant computation or by reusing computation across different steps.

Declarative Interfaces for DI. Historically, DI solutions have focused on isolated problems, such as entity resolution or schema mapping, rather than end-to-end solutions. However, the recent results in DI suggest that machine learning can provide a common formal footing for all different problems (see Section 2) along the data integration stack. A systematic study is required to identify the common abstractions across different ML-based solutions in the data integration stack. These abstractions can in turn lead to a declarative framework for data integration.

Effective Data Augmentation for ML-pipelines. Data augmentation refers to a class of data enrichment techniques that are useful for controlling the generalization error of machine learning models. Recent works in machine learning have focused on devising formal methods for data augmentation. However, most approaches rely on transformation of the data points already present in a seed dataset. A major opportunity in this context would be to explore how the efforts of the database community in data cataloging [19] and source selection [13] can be applied in the context of data augmentation to improve the quality of the learned ML models.

5 BIOGRAPHICAL SKETCHES

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 Post. She has got the VLDB Early Career Research Contribution Award for "advancing the state of the art of knowledge fusion". She co-authored book "Big Data Integration", is the PC co-chair for SIGMOD 2018 and WAIM 2015, and is serving in the Board of Trustees of the VLDB Endowment. She has given several tutorials on data integration and knowledge collection 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.

REFERENCES

[1] Web-scale information extraction with vertex. In *ICDE*, 2011.
[2] P. Bailis, E. Gan, S. Madden, D. Narayanan, K. Rong, and S. Suri. Macrobase: Prioritizing attention in fast data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 541–556, New York, NY, USA, 2017. ACM.
[3] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang. Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 2201–2206, New York, NY, USA, 2016. ACM.
[4] R. Das, A. Neelakantan, D. Belanger, and A. McCallum. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *EACL*, 2017.
[5] 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.
[6] A. Doan, A. Y. Halevy, and Z. G. Ives. *Principles of Data Integration*. Morgan Kaufmann, 2012.
[7] X. L. Dong. Challenges and innovations in building a product knowledge graph. In *AKBC*, 2017.
[8] 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.
[9] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. *PVLDB*, 2014.
[10] 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.
[11] X. L. Dong and F. Naumann. Data fusion–resolving data conflicts for integration. *PVLDB*, 2009.
[12] X. L. Dong and D. Srivastava. Big data integration. *Proc. VLDB Endow.*, 6(11):1188–1189, Aug. 2013.
[13] X. L. Dong and D. Srivastava. Big data integration. *Synthesis Lectures on Data Management*, 7(1):1–198, 2015.
[14] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the Americal Statistical Association*, 64(328):1183–1210, 1969.
[15] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. pages 371–380, 2001.
[16] 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.
[17] L. Getoor and A. Machanavajjhala. Entity resolution: theory, practice & open challenges. *PVLDB*, 5(12):2018–2019, 2012.
[18] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu. Corleone: Hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 601–612, New York, NY, USA, 2014. ACM.
[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. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 1717–1722, New York, NY, USA, 2017. ACM.
[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] X. Ma and E. Hovy. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *ACL*, 2016.
[31] C. Manning. Representations for language: From word embeddings to sentence meanings. https://simons.berkeley.edu/talks/christopher-manning-2017-3-27, 2017.
[32] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL*, 2009.
[33] T. Mitchell. Learning from limited labeled data (but a lot of unlabeled data). https://lld-workshop.github.io/slides/tom_mitchell_lld.pdf, 2017.
[34] 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.
[35] A. Neelakantan, B. Roth, and A. McCallum. Compositional vector space models for knowledge base completion. In *ACL*, 2015.
[36] 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.
[37] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*, pages 877–885, 2010.
[38] R. Pimplikar and S. Sarawagi. Answering table queries on the web using column keywords. *PVLDB*, 5(10):908–919, 2012.
[39] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, pages 1723–1726, New York, NY, USA, 2017. ACM.
[40] J. Pujara and L. Getoor. Generic statistical relational entity resolution in knowledge graphs. In *AAAI*, 2016.
[41] 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.
[42] 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.

[43] 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.
[44] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré. Holoclean: Holistic data repairs with probabilistic inference. *PVLDB*, 10(11):1190–1201, 2017.
[45] 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.
[46] S. Riedel, L. Yao, B. M. Marlin, and A. McCallum. Relation extraction with matrix factorization and universal schemas. In *HLT-NAACL*, 2013.
[47] B. Saha and D. Srivastava. Data quality: The other face of big data. In *ICDE*, pages 1294–1297, 2014.
[48] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *SIGKDD*, 2002.
[49] 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.
[50] 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.
[51] X. Wang, X. L. Dong, and A. Meliou. Data x-ray: A diagnostic tool for data errors. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 1231–1245, New York, NY, USA, 2015. ACM.
[52] S. Wu, L. Hsiao, X. Cheng, B. Hancock, T. Rekatsinas, P. Levis, and C. Ré. Fonduer: Knowledge base construction from richly formatted data. In *Proceedings of the 2018 ACM International Conference on Management of Data*, SIGMOD '18, 2018.
[53] C. Zhang, C. Ré, 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.