References Part I: Introduction

Bengio, Y., Goodfellow, I.J. & Courville, A., 2015. Deep learning. *Nature*, 521(7553), pp.436–444.

Bishop, C.M., 2016. *Pattern Recognition and Machine Learning*, Springer New York.

Doan, A., Halevy, A.Y. & Ives, Z.G., 2012. *Principles of Data Integration*, Morgan Kaufmann.

Domingos, P., 2012. A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, 55(10), pp.78–87.

Dong, X. et al., 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, NY, USA: ACM, pp. 601–610.

Dong, X.L. & Srivastava, D., 2015. Big data integration. *Synthesis Lectures on Data Management*, 7(1), pp.1–198.

Dong, X.L. & Srivastava, D., 2013. Big Data Integration. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 6(11), pp.1188–1189.

Getoor, L. & Machanavajjhala, A., 2012. Entity resolution: theory, practice & open challenges. *PVLDB*, 5(12), pp.2018–2019.

Goodfellow, I. et al., 2016. *Deep learning*, MIT press Cambridge.

Halevy, A., Norvig, P. & Pereira, F., 2009. The Unreasonable Effectiveness of Data. *IEEE intelligent systems*, 24(2), pp.8–12.

Konda, P. et al., 2016. Magellan: Toward Building Entity Matching Management Systems. *PVLDB*, 9(12), pp.1197–1208.

References Part I: Introduction

Kumar, A., Boehm, M. & Yang, J., 2017. Data Management in Machine Learning: Challenges, Techniques, and Systems. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 1717–1722.

Lockard, C. et al., 2018. CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web. *arXiv [cs.AI]*. Available at: http://arxiv.org/abs/1804.04635.

Mohri, M., Rostamizadeh, A. & Talwalkar, A., 2012. *Foundations of Machine Learning*, MIT Press.

Polyzotis, N. et al., 2017. Data Management Challenges in Production Machine Learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 1723–1726.

Ratner, A. et al., 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *PVLDB*, 11(3), pp.269–282.

Rekatsinas, T. et al., 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *PVLDB*, 10(11), pp.1190–1201.

Wu, S. et al., 2018. Fonduer: Knowledge Base Construction from Richly Formatted Data. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 1301–1316.

Zheng, G. et al., 2018. OpenTag: Open Attribute Value Extraction from Product Profiles. In *KDD*. Available at: https://people.mpi-inf.mpg.de/~smukherjee/research/OpenTag-KDD18.pdf.

References Part II: Entity Linkage

Bhattacharya, I. & Getoor, L., 2006. A latent dirichlet model for unsupervised entity resolution. In *SDM*. SIAM, pp. 47–58.

Das, S. et al., 2017. Falcon: Scaling Up Hands-Off Crowdsourced Entity Matching to Build Cloud Services. In *Sigmod*. pp. 1431–1446.

Doan, A. et al., 2017. Human-in-the-Loop Challenges for Entity Matching: A Midterm Report. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2017, Chicago, IL, USA, May 14, 2017*. pp. 12:1–12:6.

Fellegi, I.P. & Sunter, A.B., 1969. A Theory for Record Linkage. *Journal of the Americal Statistical Association*, 64(328), pp.1183–1210.

Getoor, L. & Machanavajjhala, A., 2012. Entity resolution: theory, practice & open challenges. *PVLDB*, 5(12), pp.2018–2019.

Gokhale, C. et al., 2014. Corleone: Hands-off Crowdsourcing for Entity Matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’14. New York, NY, USA: ACM, pp. 601–612.

Hassanzadeh, O. et al., 2009. Framework for Evaluating Clustering Algorithms in Duplicate Detection. *PVLDB*, 2(1), pp.1282–1293.

Ji, H., 2014. Entity Linking and Wikification Reading List. Available at: http://nlp.cs.rpi.edu/kbp/2014/elreading.html.

Konda, P. et al., 2016. Magellan: Toward Building Entity Matching Management Systems. *PVLDB*, 9(12), pp.1197–1208.

Kopcke, H., Thor, A. & Rahm, E., 2010. Evaluation of entity resolution approaches on real-world match problems. *PVLDB*, 3(1), pp.484–493.

References Part II: Entity Linkage

Mudgal, S. et al., 2018. Deep Learning for Entity Matching: A Design Space Exploration. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 19–34.

Pujara, J. & Getoor, L., 2016. Generic Statistical Relational Entity Resolution in Knowledge Graphs. In *AAAI*.

Rakshit Trivedi, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos, Jun Ma and Hongyuan Zha., LinkNBed: Multi-Graph Representation Learning with Entity Linkage. In *56th Annual Meeting of the Association for Computational Linguistics*. ACL.

Sarawagi, S. & Bhamidipaty, A., 2002. Interactive deduplication using active learning. In *SIGKDD*.

Singla, P. & Domingos, P., 2006. Entity Resolution with Markov Logic. In *ICDM*. Washington, DC, USA: IEEE Computer Society, pp. 572–582.

Stonebraker, M. et al., 2013. Data Curation at Scale: The Data Tamer System. In *CIDR*.

Verroios, V., Garcia-Molina, H. & Papakonstantinou, Y., 2017. 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*. pp. 1133–1148.

References Part II: Data Extraction

Das, R. et al., 2017. Chains of reasoning over entities, relations, and text using recurrent neural networks. In *EACL*.

Dong, X. et al., 2014. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’14. New York, NY, USA: ACM, pp. 601–610.

Dong, X.L., 2017. Challenges and Innovations in Building a Product Knowledge Graph. In *AKBC*.

Gulhane, P. et al., 2011. Web-scale information extraction with vertex. In *2011 IEEE 27th International Conference on Data Engineering*. pp. 1209–1220.

He, R. et al., 2017. An Unsupervised Neural Attention Model for Aspect Extraction. In *ACL*.

Hoffmann, R. et al., 2011. Knowledge-based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 541–550.

Limaye, G., Sarawagi, S. & Chakrabarti, S., 2010. Annotating and Searching Web Tables Using Entities, Types and Relationships. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 3(1-2), pp.1338–1347.

Lockard, C. et al., 2018. CERES: Distantly Supervised Relation Extraction from the Semi-Structured Web. *arXiv [cs.AI]*. Available at: http://arxiv.org/abs/1804.04635.

References Part II: Data Extraction

Mintz, M. et al., 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Mitchell, T. et al., 2018. Never-ending Learning. *Communications of the ACM*, 61(5), pp.103–115.

Neelakantan, A., Roth, B. & McCallum, A., 2015. Compositional vector space models for knowledge base completion. In *ACL*.

Riedel, S. et al., 2013. Relation Extraction with Matrix Factorization and Universal Schemas. In *HLT-NAACL*.

Shin, J. et al., 2015. Incremental Knowledge Base Construction Using DeepDive. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 8(11), pp.1310–1321.

Wu, S. et al., 2018. Fonduer: Knowledge Base Construction from Richly Formatted Data. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, pp. 1301–1316.

Zhang, C. et al., 2017. DeepDive: Declarative Knowledge Base Construction. *CACM*, 60(5), pp.93–102.

References Part II: Data Fusion

Dawid, A.P. & Skene, A.M., 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 28(1), pp.20–28.

Dong, X.L. et al., 2014. From Data Fusion to Knowledge Fusion. *PVLDB*.

Dong, X.L. et al., 2015. Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 8(9), pp.938–949.

Dong, X.L. & Naumann, F., 2009. Data Fusion: Resolving Data Conflicts for Integration. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 2(2), pp.1654–1655.

Gao, J. et al., 2016. Mining Reliable Information from Passively and Actively Crowdsourced Data. In *KDD*. pp. 2121–2122.

Jaffe, A., Nadler, B. & Kluger, Y., 2015. Estimating the accuracies of multiple classifiers without labeled data. In *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. pp. 407–415.

Li, H., Yu, B. & Zhou, D., 2013. Error rate analysis of labeling by crowdsourcing. In *ICML Workshop: Machine Learning Meets Crowdsourcing. Atlanta, Georgia, USA*.

Li, Q. et al., 2014. A Confidence-aware Approach for Truth Discovery on Long-tail Data. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 8(4), pp.425–436.

References Part II: Data Fusion

Li, X. et al., 2013. Truth Finding on the Deep Web: Is the Problem Solved? *PVLDB*, 6(2).

Li, Y. et al., 2016. A Survey on Truth Discovery. *SIGKDD Explor. Newsl.*, 17(2), pp.1–16.

Nickel, M. et al., 2016. A Review of Relational Machine Learning for Knowledge Graphs. *Proceedings of the IEEE*, 104(1), pp.11–33.

Pasternack, J. & Roth, D., 2010. Knowing what to believe (when you already know something). In *COLING*. pp. 877–885.

Platanios, E. A., Dubey, A., & Mitchell, T. (2016, June). Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning*(pp. 1416-1425).

Rekatsinas, T. et al., 2017. SLiMFast: Guaranteed Results for Data Fusion and Source Reliability. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. New York, NY, USA: ACM, pp. 1399–1414.

Shaham, U. et al., 2016. A Deep Learning Approach to Unsupervised Ensemble Learning. In *International Conference on Machine Learning*. International Conference on Machine Learning. pp. 30–39.

Wang, Q. et al., 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE transactions on knowledge and data engineering*, 29(12), pp.2724–2743.

Yin, X., Han, J. & Yu, P.S., 2007. Truth discovery with multiple conflicting information providers on the web. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1048–1052.

References Part II: Data Fusion

Zhang, Y. et al., 2014. Spectral Methods meet EM: A Provably Optimal Algorithm for Crowdsourcing. In Z. Ghahramani et al., eds. *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., pp. 1260–1268.

Zhao, B. et al., 2012. A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 5(6), pp.550–561.

References Part III: Training Data Creation

Chapelle, O., Scholkopf, B. & Eds., A.Z., 2009. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 20(3), pp.542–542.

Dawid, A.P. & Skene, A.M., 1979. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 28(1), pp.20–28.

Mintz, M. et al., 2009. Distant supervision for relation extraction without labeled data. In *ACL*.

Mitchell, T., 2017. Learning from Limited Labeled Data (But a Lot of Unlabeled Data). Available at: https://lld-workshop.github.io/slides/tom_mitchell_lld.pdf.

Platanios, E.A., Dubey, A. & Mitchell, T., 2016. Estimating Accuracy from Unlabeled Data: A Bayesian Approach. In *International Conference on Machine Learning*. International Conference on Machine Learning. pp. 1416–1425.

Ratner, A. et al., 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. *PVLDB*, 11(3), pp.269–282.

Ratner, A.J. et al., 2016. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*. pp. 3567–3575.

Raykar, V.C. et al., 2010. Learning From Crowds. *Journal of machine learning research: JMLR*, 11, pp.1297–1322.

Recht, B. et al., 2018. Do CIFAR-10 Classifiers Generalize to CIFAR-10? *arXiv [cs.LG]*. Available at: http://arxiv.org/abs/1806.00451.

References Part III: Training Data Creation

Roth, B. & Klakow, D., 2013. Combining generative and discriminative model scores for distant supervision. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. pp. 24–29.

Russell, S. & Stefano, E., 2017. Label-free supervision of neural networks with physics and domain knowledge. *Proceedings of AAAI*.

Salimans, T. et al., 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. pp. 2234–2242.

Schapire, R.E. & Freund, Y., 2012. Boosting: Foundations and Algorithms. Adaptive computation and machine learning.

References Part III: Data Cleaning

Bailis, P. et al., 2017. MacroBase: Prioritizing Attention in Fast Data. In *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD ’17. New York, NY, USA: ACM, pp. 541–556.

Chu, X. et al., 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data*. SIGMOD ’16. New York, NY, USA: ACM, pp. 2201–2206.

Chandola, V., Banerjee, A. & Kumar, V., 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.*, 41(3), pp.15:1–15:58.

Galhardas, H. et al., 2001. Declarative data cleaning: Language, model, and algorithms. In *VLDB*. pp. 371–380.

Hellerstein, J.M., 2008. Quantitative data cleaning for large databases. *Statistical journal of the United Nations Economic Commission for Europe*. Available at: http://db.cs.berkeley.edu/jmh/papers/cleaning-unece.pdf.

Ilyas, I.F., 2016. Effective Data Cleaning with Continuous Evaluation. *IEEE Data Eng. Bull.*, 39, pp.38–46.

Krishnan, S. et al., 2016. ActiveClean: Interactive Data Cleaning for Statistical Modeling. *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases*, 9(12), pp.948–959.

Krishnan, S. et al., 2017. BoostClean: Automated Error Detection and Repair for Machine Learning. *arXiv [cs.DB]*. Available at: http://arxiv.org/abs/1711.01299.

References Part III: Data Cleaning

Mayfield, C., Neville, J. & Prabhakar, S., 2010. ERACER: A Database Approach for Statistical Inference and Data Cleaning. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*. SIGMOD '10. New York, NY, USA: ACM, pp. 75–86.

Rekatsinas, T. et al., 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *PVLDB*, 10(11), pp.1190–1201.

Wang, X., Dong, X.L. & Meliou, A., 2015. Data X-Ray: A Diagnostic Tool for Data Errors. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. New York, NY, USA: ACM, pp. 1231–1245.

Yakout, M., Berti-Équille, L. & Elmagarmid, A.K., 2013. Don'T Be SCAREd: Use SCalable Automatic REpairing with Maximal Likelihood and Bounded Changes. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. SIGMOD '13. New York, NY, USA: ACM, pp. 553–564.