Table 1: Labeled data science pipelines from the subject studies. ACQ: Data acquisition, PRP: Data preparation, STR: Data storage, FTR: Feature engineering, MDL: Modeling, TRN: Training, EVL: Evaluation, PRD: Prediction, INT: Interpretation, CMN: Communication, DPL: Deployment.

Overall goal: 🟩 Describe/propose pipeline, 🟪 Survey/compare/review, 🟥 DS optimization, 🟧 Introduce new method/application

Type	References	Preprocessing			Modeling					Post-processing			Involves		
		ACQ	PRP	STR	FTR	MDL	TRN	EVL	PRD	INT	CMN	DPL	Cyber	Physical	Human
Machine learning process	Olson et al., 2016 [40]	🟥	🟥	-	🟥	🟥	🟥	🟥	🟥	-	-	-	🟥	-	-
	Miao et al., 2017b [36]	🟩	-	-	-	🟩	🟩	🟩	🟩	-	-	🟩	🟩	-	-
	Garcia et al., 2018 [18]	-	🟩	-	-	🟩	🟩	🟩	🟩	-	🟩	-	🟩	-	-
	Hong and Hunter, 2017 [24]	🟩	🟩	-	🟩	🟩	-	-	-	-	-	-	🟩	-	-
	Microsoft Blog, 2019 [37]	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	-	🟩	🟩	-	-
	Zhou, 2019 [71]	-	🟩	-	🟩	🟩	🟩	-	🟩	-	-	-	🟩	-	-
	Shibuya, 2017 [54]	-	🟩	-	-	🟩	🟩	-	-	-	-	-	🟩	-	-
	Polyzotis et al., 2018 [42]	-	🟪	-	🟪	🟪	🟪	🟪	-	-	-	🟪	🟪	-	-
	Roh et al., 2019 [46]	🟪	🟪	-	🟪	🟪	-	-	-	-	🟪	-	🟪	🟪	-
	Miao et al., 2017a [35]	🟥	-	🟥	-	🟥	-	-	-	-	-	-	🟥	-	-
	Sparks et al., 2017 [58]	-	🟥	-	🟥	🟥	🟥	-	-	-	-	-	🟥	-	-
	Guo, 2017 [22]	🟩	🟩	-	🟩	🟩	🟩	🟩	🟩	-	-	-	🟩	-	-
	Baylor et al., 2017 [7]	-	🟧	-	🟧	🟧	🟧	🟧	-	-	🟧	🟧	🟧	-	-
	Abadi et al., 2016 [1]	-	🟧	🟧	-	🟧	🟧	-	-	-	-	-	🟧	-	-
	Chilimbi et al., 2014 [11]	-	-	-	🟧	🟧	🟧	-	🟧	-	-	-	🟧	-	-
	Kraska et al., 2013 [31]	-	-	🟧	🟧	🟧	🟧	🟧	-	🟧	-	-	🟧	-	-
	Sculley et al., 2015 [50]	🟪	-	-	🟪	🟪	-	-	-	-	-	🟪	🟪	-	-
	Chang, 2017 [9]	🟩	🟩	-	🟩	🟩	🟩	🟩	🟩	-	-	🟩	🟩	-	-
	Google Cloud Blog, 2019 [21]	🟩	🟩	-	-	🟩	🟩	🟩	🟩	-	-	-	🟩	-	-
	Amershi et al., 2019 [4]	🟪	🟪	-	🟪	🟪	🟪	🟪	🟪	-	-	🟪	🟪	-	-
	Van Der Weide et al., 2017 [63]	🟥	🟥	-	🟥	-	-	-	🟥	-	-	-	🟥	-	🟥
	Hill et al., 2016 [23]	🟩	-	-	🟩	🟩	-	🟩	-	-	-	-	🟩	-	-
	Shang et al., 2019 [52]	-	🟥	-	🟥	🟥	-	-	-	-	-	-	🟥	-	-
	Zhang et al., 2016 [68]	-	🟥	-	🟥	🟥	-	-	-	-	-	-	🟥	-	-
	Gil et al., 2018 [19]	-	🟩	-	🟩	🟩	-	🟩	🟩	-	-	-	🟩	-	-
	Sadiq et al., 2018 [48]	🟩	🟩	🟩	🟩	🟩	-	🟩	-	🟩	-	-	🟩	-	-
	Zhou et al., 2020 [70]	🟩	🟩	-	-	🟩	-	-	🟩	🟩	-	🟩	🟩	🟩	-
	Aggarwal et al., 2019 [3]	-	🟩	-	🟩	🟩	🟩	-	🟩	-	-	-	🟩	-	-
	Toreini et al., 2020 [61]	🟩	🟩	-	🟩	-	🟩	🟩	🟩	-	-	-	🟩	-	-
	Ashmore et al., 2021 [5]	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪
	Shashanka, 2019 [53]	🟩	🟩	-	🟩	🟩	🟩	🟩	🟩	🟩	–	–	🟩	–	–
	MLOps, 2020 [38]	🟩	🟩	🟩	-	🟩	🟩	🟩	-	🟩	–	–	🟩	–	–
	Daumé III, 2016 [12]	🟧	🟧	-	-	🟧	🟧	🟧	🟧	-	-	-	🟧	-	-
Big data management	Todd and Dietrich, 2017 [60]	🟧	🟧	🟧	🟧	🟧	-	-	🟧	-	🟧	-	🟧	-	🟧
	Zhang et al., 2017 [69]	🟩	🟩	🟩	-	🟩	-	-	🟩	-	🟩	-	🟩	🟩	-
	Sapp, 2017 [49]	🟪	🟪	🟪	🟪	🟪	🟪	🟪	🟪	-	🟪	🟪	🟪	🟪	🟪
	Landset et al., 2015 [32]	-	-	🟥	-	🟥	🟥	🟥	🟥	-	-	🟥	🟥	-	-
	Polyzotis et al., 2017 [41]	-	-	-	-	🟪	🟪	-	🟪	-	-	🟪	🟪	-	-
	Hu et al., 2014 [25]	🟪	🟪	🟪	-	-	-	-	-	-	-	-	🟪	🟪	🟪
	Demchenko et al., 2012 [14]	🟩	🟩	🟩	-	-	-	-	-	-	🟩	-	🟩	-	-
	Khan et al., 2017 [29]	-	🟪	-	🟪	🟪	-	🟪	🟪	🟪	🟪	-	🟪	-	-
	El Arass and Souissi, 2018 [15]	🟩	🟩	🟩	🟩	-	-	🟩	-	–	–	🟩	🟩	-	-
	Hummer et al., 2019 [26]	-	🟧	-	-	🟧	🟧	🟧	🟧	-	🟧	🟧	🟧	-	🟧
	Yildiz, 2020 [67]	🟩	🟩	🟩	🟩	🟩	-	🟩	-	🟩	🟩	🟩	🟩	-	-
	Glen, 2019 [20]	🟩	🟩	-	-	🟩	🟩	🟩	-	-	🟩	-	🟩	-	-
	Jones, 2018 [28]	-	🟩	-	🟩	🟩	🟩	🟩	🟩	-	–	🟩	🟩	-	-

Type	References	Preprocessing			Modeling					Post-processing			Involves		
		ACQ	PRP	STR	FTR	MDL	TRN	EVL	PRD	INT	CMN	DPL	Cyber	Physical	Human
Team process	Pouchard, 2016 [43]	🟩	🟩	🟩	-	-	-	-	-	-	🟩	-	🟩	-	🟩
	Severtson, 2017 [51]	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	🟩	🟩	🟩	-	🟩
	Berman et al., 2018 [8]	🟪	🟪	🟪	-	🟪	-	-	-	-	🟪	🟪	🟪	🟪	🟪
	Agarwal, 2018 [2]	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	🟩	-	🟩	-	🟩
	Nguyen et al., 2019 [39]	🟪	🟪	-	🟪	🟪	🟪	🟪	🟪	-	-	-	🟪	-	-
	Rüegg et al., 2014 [47]	🟩	🟩	-	-	-	-	-	-	🟩	🟩	-	🟩	-	-
	Gandomi and Haider, 2015 [17]	🟪	🟪	🟪	🟪	🟪	-	🟪	-	🟪	-	-	🟪	-	-
	Ball, 2012 [6]	🟪	-	🟪	-	-	-	🟪	-	-	🟪	-	🟪	-	-
	Wing, 2019 [65]	🟩	🟩	🟩	-	-	-	🟩	-	🟩	-	-	🟩	🟩	🟩
	Rehman et al., 2016 [44]	🟥	🟥	-	-	🟥	-	🟥	🟥	-	-	-	🟥	-	-
	Chen and Zhang, 2014 [10]	🟪	🟪	-	-	-	-	🟪	-	🟪	-	-	🟪	-	-
	Jagadish, 2015 [27]	🟪	🟪	-	🟪	🟪	-	🟪	-	🟪	-	-	🟪	-	🟪
	Larson and Chang, 2016 [33]	🟪	🟪	-	-	🟪	-	🟪	🟪	-	🟪	-	🟪	-	🟪
	Rizvi et al., 2017 [45]	🟥	🟥	🟥	🟥	🟥	-	-	🟥	-	-	-	🟥	-	-
	Demchenko et al., 2016 [13]	🟥	-	🟥	🟥	🟥	-	🟥	🟥	-	🟥	-	🟥	-	🟥
	Wolf et al., 2016 [66]	🟪	🟪	🟪	🟪	-	-	🟪	-	-	-	-	🟪	-	-
	Sinaeepourfard et al., 2016 [55]	🟩	🟩	🟩	🟩	-	-	-	-	-	🟩	-	🟩	🟩	-
	Kim et al., 2016 [30]	🟪	🟪	-	🟪	🟪	-	🟪	🟪	🟪	-	🟪	🟪	-	🟪
	Fisher, 2017 [16]	🟪	🟪	-	-	🟪	-	🟪	-	🟪	🟪	🟪	🟪	🟪	-
	Turkay et al., 2018 [62]	🟪	🟪	-	🟪	🟪	-	🟪	-	🟪	-	-	🟪	-	🟪
	Smith et al., 2017 [57]	-	🟧	🟧	-	-	-	-	-	-	🟧	🟧	🟧	-	🟧
	Wang et al., 2019 [64]	🟪	🟪	-	🟪	🟪	-	🟪	🟪	🟪	-	-	🟪	-	-
	Lo et al., 2020 [34]	🟪	🟪	-	🟪	🟪	🟪	🟪	🟪	-	🟪	-	🟪	-	-
	Siva, 2020 [56]	🟩	🟩	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	-
	Stodden, 2020 [59]	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	🟩	🟩	🟩	🟩	🟩	🟩

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 265–283.

[2] Sudeep Agarwal. 2018. Understanding the Data Science Lifecycle. http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle.

[3] Charu Aggarwal, Djallel Bouneffouf, Horst Samulowitz, Beat Buesser, Thanh Hoang, Udayan Khurana, Sijia Liu, Tejaswini Pedapati, Parikshit Ram, Ambrish Rawat, et al. 2019. How can ai automate end-to-end data science? *arXiv preprint arXiv:1910.14436* (2019).

[4] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software Engineering for Machine Learning: A Case Study. In *Proceedings of the 41st International Conference on Software Engineering*. ACM.

[5] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges. *ACM Comput. Surv.* 54, 5, Article 111 (may 2021). https://doi.org/10.1145/3453444

[6] Alex Ball. 2012. *Review of data management lifecycle models*. University of Bath, IDMRC.

[7] Denis Baylor, Eric Breck, Heng-Tze Cheng, Noah Fiedel, Chuan Yu Foo, Zakaria Haque, Salem Haykal, Mustafa Ispir, Vihan Jain, Levent Koc, et al. 2017. TFX: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1387–1395.

[8] Francine Berman, Rob Rutenbar, Brent Hailpern, Henrik Christensen, Susan Davidson, Deborah Estrin, Michael Franklin, Margaret Martonosi, Padma Raghavan, Victoria Stodden, et al. 2018. Realizing the potential of data science. *Commun. ACM* 61, 4 (2018), 67–72.

[9] Maurice Chang. 2017. 4 Stages of the Machine Learning (ML) Modeling Cycle. https://www.linkedin.com/pulse/4-stages-machine-learning-ml-modeling-cycle-maurice-chang.

[10] CL Philip Chen and Chun-Yang Zhang. 2014. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information sciences* 275 (2014), 314–347.

[11] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. 2014. Project Adam: Building an efficient and scalable deep learning training system. In *11th {USENIX} Symposium on Operating Systems Design and Implementation (OSDI 14)*. 571–582.

[12] Hal Daumé III. 2016. What Is a Machine Learning Pipeline? https://nlpers.blogspot.com/2016/08/debugging-machine-learning.html.

[13] Yuri Demchenko, Fatih Turkmen, Cees de Laat, Christophe Blanchet, and Charles Loomis. 2016. Cloud based big data infrastructure: Architectural components and automated provisioning. In *2016 International Conference on High Performance Computing & Simulation (HPCS)*. IEEE, 628–636.

[14] Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, and Cees De Laat. 2012. Addressing big data challenges for scientific data infrastructure. In *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. IEEE, 614–617.

[15] Mohammed El Arass and Nissrine Souissi. 2018. Data lifecycle: from big data to SmartData. In *2018 IEEE 5th international congress on information science and technology (CiSt)*. IEEE, 80–87.

[16] Douglas Fisher. 2017. A selected summary of AI for computational sustainability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[17] Amir Gandomi and Murtaza Haider. 2015. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management* 35, 2 (2015), 137–144.

[18] Rolando Garcia, Vikram Sreekanti, Neeraja Yadwadkar, Daniel Crankshaw, Joseph E Gonzalez, and Joseph M Hellerstein. 2018. Context: The missing piece in the machine learning lifecycle. In *KDD CMI Workshop*, Vol. 114.

[19] Yolanda Gil, Ke-Thia Yao, Varun Ratnakar, Daniel Garijo, Greg Ver Steeg, Pedro Szekely, Rob Brekelmans, Mayank Kejriwal, Fanghao Luo, and I-Hui Huang. 2018. P4ML: A phased performance-based pipeline planner for automated machine learning. In *AutoML Workshop at ICML*.

[20] Stephanie Glen. 2019. The Lifecycle of Data. https://www.datasciencecentral.com/profiles/blogs/the-lifecycle-of-data.

[21] Google Cloud Blog. 2019. Machine Learning Workflow. https://cloud.google.com/ml-engine/docs/tensorflow/ml-solutions-overview.

[22] Yufeng Guo. 2017. The 7 Steps of Machine Learning. https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

[23] Charles Hill, Rachel Bellamy, Thomas Erickson, and Margaret Burnett. 2016. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 162–170.

[24] Sue Ann Hong and Tim Hunter. 2017. Build, Scale, and Deploy Deep Learning Pipelines with Ease. https://databricks.com/blog/2017/09/06/build-scale-deploy-deep-learning-pipelines-ease.html.

[25] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. 2014. Toward scalable systems for big data analytics: A technology tutorial. *IEEE access* 2 (2014), 652–687.

[26] Waldemar Hummer, Vinod Muthusamy, Thomas Rausch, Parijat Dube, Kaoutar El Maghraoui, Anupama Murthi, and Punleuk Oum. 2019. Modelops: Cloud-based lifecycle management for reliable and trusted ai. In *2019 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 113–120.

[27] HV Jagadish. 2015. Big data and science: Myths and reality. *Big Data Research* 2, 2 (2015), 49–52.

[28] M. Tim Jones. 2018. Data, structure, and the data science pipeline. https://developer.ibm.com/technologies/data-science/articles/ba-intro-data-science-1/.

[29] Samiya Khan, Xiufeng Liu, Kashish A Shakil, and Mansaf Alam. 2017. A survey on scholarly data: From big data perspective. *Information Processing & Management* 53, 4 (2017), 923–944.

[30] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 96–107.

[31] Tim Kraska, Ameet Talwalkar, John C Duchi, Rean Griffith, Michael J Franklin, and Michael I Jordan. 2013. MLbase: A Distributed Machine-learning System.. In *Cidr*, Vol. 1. 2–1.

[32] Sara Landset, Taghi M Khoshgoftaar, Aaron N Richter, and Tawfiq Hasanin. 2015. A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data* 2, 1 (2015), 24.

[33] Deanne Larson and Victor Chang. 2016. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management* 36, 5 (2016), 700–710.

[34] Sin Kit Lo, Qinghua Lu, Chen Wang, Helen Paik, and Liming Zhu. 2020. A systematic literature review on federated machine learning: From a software engineering perspective. *arXiv preprint arXiv:2007.11354* (2020).

[35] Hui Miao, Amit Chavan, and Amol Deshpande. 2017. Provdb: Lifecycle management of collaborative analysis workflows. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. ACM, 7.

[36] Hui Miao, Ang Li, Larry S Davis, and Amol Deshpande. 2017. Towards unified data and lifecycle management for deep learning. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, 571–582.

[37] Microsoft Blog. 2019. What are ML pipelines in Azure Machine Learning service? https://docs.microsoft.com/en-us/azure/machine-learning/service/concept-ml-pipelines.

[38] Valohai MLOps. 2020. What Is a Machine Learning Pipeline? https://valohai.com/machine-learning-pipeline/.

[39] Giang Nguyen, Stefan Dlugolinsky, Martin Bobák, Viet Tran, Álvaro López García, Ignacio Heredia, Peter Malík, and Ladislav Hluchỳ. 2019. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. *Artificial Intelligence Review* (2019), 1–48.

[40] Randal S Olson, Nathan Bartley, Ryan J Urbanowicz, and Jason H Moore. 2016. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*. ACM, 485–492.

[41] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2017. Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 1723–1726.

[42] Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. 2018. Data Lifecycle Challenges in Production Machine Learning: A Survey. *ACM SIGMOD Record* 47, 2 (2018), 17–28.

[43] Line Pouchard. 2016. Revisiting the data lifecycle with big data curation. *International Journal of Digital Curation* 10, 2 (2016), 176–192.

[44] Muhammad Habib Rehman, Victor Chang, Aisha Batool, and Teh Ying Wah. 2016. Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management* 36, 6 (2016), 917–928.

[45] Syed Ali Asad Rizvi, Elmarie Van Heerden, Arnold Salas, Favour Nyikosa, Stephen J Roberts, Michael A Osborne, and Elmer Rodriguez. 2017. Identifying Sources of Discrimination Risk in the Life Cycle of Machine Intelligence Applications under New European Union Regulations. In *2017 AAAI Spring Symposium Series*.

[46] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A Survey on Data Collection for Machine Learning: a Big Data-AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[47] Janine Rüegg, Corinna Gries, Ben Bond-Lamberty, Gabriel J Bowen, Benjamin S Felzer, Nancy E McIntyre, Patricia A Soranno, Kristin L Vanderbilt, and Kathleen C Weathers. 2014. Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment* 12, 1 (2014), 24–30.

[48] Shazia Sadiq, Tamraparni Dasu, Xin Luna Dong, Juliana Freire, Ihab F Ilyas, Sebastian Link, Miller J Miller, Felix Naumann, Xiaofang Zhou, and Divesh Srivastava. 2018. Data quality: The role of empiricism. *ACM SIGMOD Record* 46, 4 (2018), 35–43.

[49] Carlton E Sapp. 2017. Preparing and architecting machine learning. *Gartner Technical Professional Advice* (2017), 1–37.

[50] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*. 2503–2511.

[51] Roald Bradley Severtson. 2017. What is the Team Data Science Process? https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview.

[52] Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, and Tim Kraska. 2019. Democratizing data science through interactive curation of ML pipelines. In *Proceedings of the 2019 International Conference on Management of Data*. ACM, 1171–1188.

[53] M Shashanka. 2019. What is a Pipeline in Machine Learning? How to create one? https://medium.com/analytics-vidhya/what-is-a-pipeline-in-machine-learning-how-to-create-one-bda91d0ceaca.

[54] Naoki Shibuya. 2017. Pipelines, Mind Maps and Convolutional Neural Networks. https://towardsdatascience.com/pipelines-mind-maps-and-convolutional-neural-networks-34bfc94db10c.

[55] Amir Sinaeepourfard, Jordi Garcia, Xavier Masip-Bruin, and Eva Marín-Torder. 2016. Towards a comprehensive data lifecycle model for big data environments. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. ACM, 100–106.

[56] Sivakar Siva. 2020. The "Generic" Data Science Life-Cycle. https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc.

[57] Micah J Smith, Roy Wedge, and Kalyan Veeramachaneni. 2017. FeatureHub: Towards collaborative data science. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 590–600.

[58] Evan R Sparks, Shivaram Venkataraman, Tomer Kaftan, Michael J Franklin, and Benjamin Recht. 2017. Keystoneml: Optimizing pipelines for large-scale advanced analytics. In *2017 IEEE 33rd international conference on data engineering (ICDE)*. IEEE, 535–546.

[59] Victoria Stodden. 2020. The data science life cycle: a disciplined approach to advancing data science as a science. *Commun. ACM* 63, 7 (2020), 58–66.

[60] Stephen Todd and David Dietrich. 2017. Computing resource re-provisioning during data analytic lifecycle. US Patent 9,619,550.

[61] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 272–283.

[62] Cagatay Turkay, Nicola Pezzotti, Carsten Binnig, Hendrik Strobelt, Barbara Hammer, Daniel A Keim, Jean-Daniel Fekete, Themis Palpanas, Yunhai Wang, and Florin Rusu. 2018. Progressive data science: Potential and challenges. *arXiv preprint arXiv:1812.08032* (2018).

[63] Tom Van Der Weide, Dimitris Papadopoulos, Oleg Smirnov, Michal Zielinski, and Tim Van Kasteren. 2017. Versioning for end-to-end machine learning pipelines. In *Proceedings of the 1st Workshop on Data Management for End-to-End Machine Learning*. ACM, 2.

[64] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[65] Jeannette M Wing. 2019. The Data Life Cycle. *Harvard Data Science Review* (2019).

[66] Christof Wolf, Dominique Joye, Tom W Smith, and Yang-chih Fu. 2016. *The SAGE handbook of survey methodology*. Sage.

[67] Mehmet Yildiz. 2020. Big Data Lifecycle Management. https://medium.com/technology-hits/big-data-lifecycle-management-629dfe16b78d.

[68] Yuyu Zhang, Mohammad Taha Bahadori, Hang Su, and Jimeng Sun. 2016. FLASH: fast Bayesian optimization for data analytic pipelines. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2065–2074.

[69] Yingfeng Zhang, Shan Ren, Yang Liu, Tomohiko Sakao, and Donald Huisingh. 2017. A framework for Big Data driven product lifecycle management. *Journal of Cleaner Production* 159 (2017), 229–240.

[70] Baifan Zhou, Yulia Svetashova, Tim Pychynski, Ildar Baimuratov, Ahmet Soylu, and Evgeny Kharlamov. 2020. SemFE: facilitating ML pipeline development with semantics. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3489–3492.

[71] Linda Zhou. 2019. How to Build a Better Machine Learning Pipeline. https://www.datanami.com/2018/09/05/how-to-build-a-better-machine-learning-pipeline.