

Data Understanding for Data-Centric AI – Framework Development and Review of Current Methods

Joshua Holstein, Philipp Spitzer, Samuel Gensch, Marieke Höll, Michael Vössing, Niklas Kühl

Business & Information Systems Engineering (2026)

Appendix (available online via <http://link.springer.com>)

Appendix

Dimension	Count	2nd Order Theme	Count	Methods
Foundations	14	Infrastructure	7	Dimitriadou et al. (2016); Peng et al. (2011); Po and Sorrentino (2011); Aydin and Anderson (2017); Belhajjame et al. (2013); Scott et al. (2014); Zhang et al. (2014)
		Provenance	3	Lee-Post and Pakath (2019); Huynh et al. (2018); Luo et al. (2024)
		Characterization and Familiarization	4	Kang et al. (2020); Sellam and Kersten (2016); Bikakis et al. (2021); Wongsuphasawat et al. (2016)
Collection and Selection	12	Data Collection	3	Aydin and Anderson (2017); Liu and Li (2018); Luo et al. (2024)
		Selecting Relevant Data	10	Dimitriadou et al. (2016); Owczarek et al. (2023); Vellido et al. (2013); Puolamaki et al. (2018); de Almeida et al. (2020); Lehrmann et al. (2013); Liu and Li (2018); Liu and Yoon (2024); Scott et al. (2014); Zhang et al. (2014)
		Supplemental Data	1	Liu and Yoon (2024)
Contextualization and Integration	12	Integration	7	Peng et al. (2011); Po and Sorrentino (2011); Liu and Yoon (2024); Lin et al. (2020); Belhajjame et al. (2013); Scott et al. (2014); Su et al. (2010)
		Domain Knowledge	4	Dimitriadou et al. (2016); Caballero et al. (2022); Woodall et al. (2013); Cabrera et al. (2019)
		Linking Data to Real-World	1	Mahfoud et al. (2018)

Continued on next page

Dimension	Count	2nd Order Theme	Count	Methods
Exploration and Discovery	40	Exploration	25	Kang et al. (2020); Dimitriadou et al. (2016); Song et al. (2020); Abello et al. (2006); Chung et al. (2020); Vellido et al. (2013); Sellam and Kersten (2016); Ben Lahmar and Herschel (2021); Dibia and Demiralp (2019); Yu et al. (2023); Zhang et al. (2019); Ehsan et al. (2018); Paganelli et al. (2020); Cabrera et al. (2019); Nestorov et al. (2019); Bikakis et al. (2021); Joglekar et al. (2016); Puolamaki et al. (2018); Lafon et al. (2013); Eirinaki et al. (2014); Saghafi et al. (2022); El et al. (2020); Wongsuphasawat et al. (2016); Aydin and Anderson (2017); Mahfoud et al. (2018)
		Cluster, Patterns and Relationships	30	Pournarakis et al. (2017); Kang et al. (2020); Hajian and Domingo-Ferrer (2013); Dimitriadou et al. (2016); Chung et al. (2020); Vellido et al. (2013); Zhang et al. (2023); Sellam and Kersten (2016); Ben Lahmar and Herschel (2021); Ruggieri et al. (2010a); Yu et al. (2023); Ruggieri et al. (2010b); Ehsan et al. (2018); Sluban et al. (2014); Paganelli et al. (2020); Cabrera et al. (2019); Nestorov et al. (2019); Pastor et al. (2021); Joglekar et al. (2016); Sim et al. (2010); Schäfer and Leser (2022); de Almeida et al. (2020); Lafon et al. (2013); Eirinaki et al. (2014); Lehrmann et al. (2013); Mahfoud et al. (2018)
		Visualizations	24	Kang et al. (2020); Owczarek et al. (2023); Abello et al. (2006); Chung et al. (2020); Vellido et al. (2013); Zhang et al. (2023); Sellam and Kersten (2016); Ben Lahmar and Herschel (2021); Dibia and Demiralp (2019); Ehsan et al. (2018); Sluban et al. (2014); Cabrera et al. (2019); Nestorov et al. (2019); Pastor et al. (2021); Bikakis et al. (2021); Puolamaki et al. (2018); Lafon et al. (2013); El et al. (2020); Lehrmann et al. (2013); Migut et al. (2015); Wongsuphasawat et al. (2016); Peng et al. (2011); Aydin and Anderson (2017); Mahfoud et al. (2018)

Continued on next page

Dimension	Count	2nd Order Theme	Count	Methods
Insights	25	Deliverables	0	–
		Data Quality	20	Heinrich et al. (2018); Chung et al. (2020); Caballero et al. (2022); Vellido et al. (2013); Woodall et al. (2013); Yu et al. (2023); Zhang et al. (2019); Sluban et al. (2014); Paganelli et al. (2020); Pastor et al. (2021); Puolamaki et al. (2018); Timmerman and Bronselaer (2019); Heinrich and Klier (2015); de Almeida et al. (2020); Lee-Post and Pakath (2019); Huynh et al. (2018); Fürber and Hepp (2011); Westin and Sein (2015); Luo et al. (2024); Belhajjame et al. (2013)
		Decision-Making	7	Pournarakis et al. (2017); Heinrich et al. (2018); de Almeida et al. (2020); Migut et al. (2015); Peng et al. (2011); Liu and Yoon (2024); Mahfoud et al. (2018)

Table 1: Distribution of approaches based on research focus

Target Group	Count	Methods
Method Experts	41	Kang et al. (2020); Hajian and Domingo-Ferrer (2013); Dimitriadou et al. (2016); Owczarek et al. (2023); Song et al. (2020); Abello et al. (2006); Heinrich et al. (2018); Chung et al. (2020); Caballero et al. (2022); Vellido et al. (2013); Zhang et al. (2023); Ben Lahmar and Herschel (2021); Ruggieri et al. (2010a); Dibia and Demiralp (2019); Zhang et al. (2019); Ehsan et al. (2018); Sluban et al. (2014); Paganelli et al. (2020); Nestorov et al. (2019); Pastor et al. (2021); Joglekar et al. (2016); Puolamaki et al. (2018); Sim et al. (2010); Timmerman and Bronselaer (2019); Heinrich and Klier (2015); Schäfer and Leser (2022); de Almeida et al. (2020); Huynh et al. (2018); Fürber and Hepp (2011); Westin and Sein (2015); Lehrmann et al. (2013); Wongsuphasawat et al. (2016); Po and Sorrentino (2011); Aydin and Anderson (2017); Liu and Li (2018); Luo et al. (2024); Lin et al. (2020); Mahfoud et al. (2018); Belhajjame et al. (2013); Zhang et al. (2014); Su et al. (2010)
Domain Experts	11	Dimitriadou et al. (2016); Vellido et al. (2013); Zhang et al. (2023); Ruggieri et al. (2010b); Cabrera et al. (2019); Lee-Post and Pakath (2019); Lafon et al. (2013); Saghafi et al. (2022); Migut et al. (2015); Scott et al. (2014); Liu and Yoon (2024)
Decision Makers	3	Pournarakis et al. (2017); Peng et al. (2011); Liu and Yoon (2024)
General Users	10	Sellam and Kersten (2016); Ben Lahmar and Herschel (2021); Dibia and Demiralp (2019); Yu et al. (2023); Ruggieri et al. (2010b); Bikakis et al. (2021); Eirinaki et al. (2014); Saghafi et al. (2022); El et al. (2020); Woodall et al. (2013)

Table 2: Distribution of approaches based on target group