

# Bridging Theory and Practice in Semantic Data Science

Dmitrii Orlov<sup>1,\*</sup>, Lise Stork<sup>1</sup>, Nikolay Yakovets<sup>2</sup> and Daphne Miedema<sup>1</sup>

<sup>1</sup> University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

<sup>2</sup> Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, The Netherlands

## Abstract

Representations of data science pipelines proposed in research papers often emphasize technical stages such as data preprocessing and model training. However, to what extent these technical pipelines reflect real-world practice remains unclear. This work presents a two-part exploratory study: (1) development of a codebook of data science activities from the literature; and (2) annotation of Kaggle notebooks, used as a case study of real-world data science practice, implementing knowledge graph workflows through this codebook. Our findings reveal two gaps between theory and practice. While literature focuses on technical steps and complex knowledge processing (e.g., ontology construction, knowledge integration, reasoning, refinement), simple data preprocessing activities prevail in real-world notebooks. Vice versa, the notebooks show a focus on documentation and dissemination that the literature largely lacks. These gaps highlight the potential of empirical studies to reveal aspects often overlooked by conceptual models in the literature. Understanding the material realities of data and knowledge work is crucial for improving tools and methodologies that support knowledge engineering activities. All resources for this work are available at: <https://doi.org/10.6084/m9.figshare.30305962>.

## Keywords

Data Science Practice, Knowledge Graphs, Data Science Pipelines, Kaggle Notebooks, Dissemination

## 1. Introduction

Knowledge graph (KG) workflows form an important part of the broader data science landscape and have become essential for representing and analyzing complex, interconnected data [1]. Although many papers conceptually represent data science pipelines, a recent survey reports major discrepancies between theory and practice [2]. Similar to the broader data science landscape, existing models of KG workflows often focus on technical pipeline stages, overlooking the collaborative and communicative aspects central to real-world practice. Understanding how practitioners build and work with KGs is crucial for designing better tools and methodologies. Therefore, in this poster paper we explore the following research question: **To what extent do knowledge graph workflows in practice align with conceptual data science pipelines described in the literature?**

We address our research question with a two-part study. First, we developed a structured codebook grounded in data science activities in the literature. Second, we annotated KG notebooks sourced from Kaggle with the derived codes. Through this qualitative exploratory case study, we compare pipelines in theory and KG practice on Kaggle. In this paper, we describe our study approach and core findings.

## 2. Context

Several studies have investigated data science and knowledge graph development pipelines from different conceptual perspectives. Biswas et al. [2] examined theoretical representations and identified “cyber and physical components, [with] only a few human processes”. They then use the technology-focused theoretical model to classify activities in the data science projects’ code. For KGs, Tamašauskaitė

*Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium*

\*Corresponding author.

✉ [d.s.orlov@uva.nl](mailto:d.s.orlov@uva.nl) (D. Orlov); [l.stork@uva.nl](mailto:l.stork@uva.nl) (L. Stork); [n.yakovets@tue.nl](mailto:n.yakovets@tue.nl) (N. Yakovets); [d.e.miedema@uva.nl](mailto:d.e.miedema@uva.nl) (D. Miedema)

🌐 <https://lisestork.github.io/> (L. Stork); <https://yakovets.ca/> (N. Yakovets); <https://daphnemiedema.nl/> (D. Miedema)

🆔 0009-0003-3831-5534 (D. Orlov); 0000-0002-2146-4803 (L. Stork); 0000-0002-1488-1414 (N. Yakovets); 0000-0003-3507-677X (D. Miedema)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and Groth [3] conducted a systematic review yielding a unified six-step pipeline. Similarly, their representation emphasizes technical stages but gives limited attention to human processes.

Prior research demonstrates that data work is inherently human and collaborative, requiring continuous interpretation, decision-making, and trust negotiation [4, 5, 6]. Within the semantic web community, studies on crowdsourcing and collaborative validation emphasize that KG engineering needs human involvement rather than fully automated processing [7, 8]. These studies position data and knowledge graph work as a sociotechnical process, needing both human and technical analyses.

To study data science in practice, Kaggle has become a prominent site for empirical research on computational notebooks [2, 9]. However, within the knowledge engineering field, they are still often seen merely as “portals” to external systems [1]. In contrast, surveys show that notebooks act not only as execution environments, but as information-rich “stories” that interweave code with explanations [10, 9]. Their narrative structure aids collaboration and reuse [10, 11, 12], encapsulating human interaction.

We apply thematic analysis to build an empirical codebook from visual pipeline representations in the literature [13]. To evaluate real-world workflows, we mine the public Meta Kaggle dataset [14] and deductively apply the developed codebook to annotate the notebooks. This systematic qualitative approach enables a well-grounded comparison between the activities present in theory and practice.

### 3. Approach

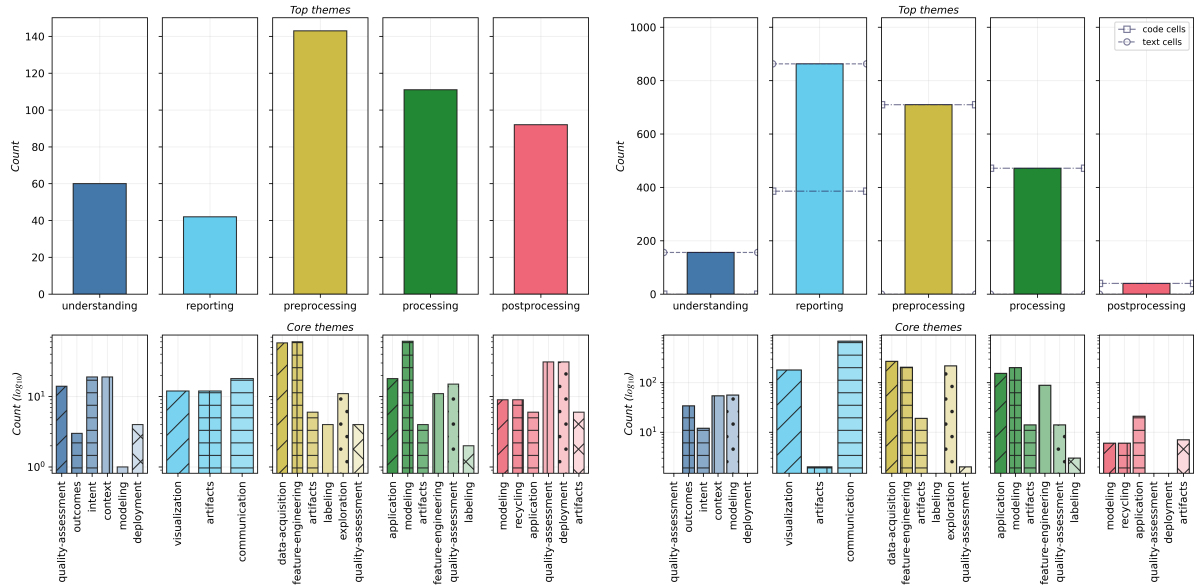
Our approach consists of two systematic procedures. First, we construct an activity-based codebook using thematic analysis of visual pipeline representations in academic literature. We queried the ACM Digital Library for articles published between 2018 and 2024, yielding an initial sample of 278 papers. Applying strict selection criteria narrowed this set to 35 articles, from which we manually extracted 51 visual pipeline representations, including 5 focused specifically on knowledge graphs. Removing the pipelines used in more than one paper (i.e. duplicate occurrences) resulted in a final sample of 42 unique representations. We extracted all pipeline activities into a tabular format and, iteratively resolving semantic misalignments, constructed a codebook with 5 top-level and 15 core themes (Figure 1a).

Second, we used the codebook to deductively annotate computational notebooks mined from an October 2023 Meta Kaggle snapshot. To target workflows utilizing KGs, we queried author-provided notebook titles with a domain-specific regular expression (e.g., matching terms like *knowledge*, *graph*, or *semantic*). Restricting the initial 2,485 candidate notebook versions to their latest stages yielded 163 distinct notebooks. We manually filtered the sample to ensure the presence of at least one KG activity from the model of Tamašauskaitė and Groth [3], resulting in a final sample of 39 notebooks. We annotated both textual markdown and Python code cells using a custom XML structure: `<DC> (<DCA t="..." s="...">...</DCA>)+ </DC>`, with attributes for top-level (t) and core (s) themes.

We specifically use ACM DL and Kaggle to engage with high-quality, open, and publicly available information in our research. We use local compute and storage to download all data and manually annotate both academic articles and computational notebooks, thus incurring no external service and/or resource use. Subsequent analysis is carried out locally using Python data science libraries (e.g. Pandas, NumPy, Matplotlib). Our data and scripts are available at: <https://doi.org/10.6084/m9.figshare.30305962>.

### 4. Workflows in the Literature

Our thematic analysis resulted in a codebook with 5 top-level themes and 15 core themes (Figure 1a). These themes represent distinct operational phases without implying a specific standard or ordering. The top themes are defined as follows: *Understanding* captures evaluative processes: e.g. reflection and assessment of workflows, applications, data, impacts. *Reporting* corresponds to narrative-based dissemination or representation of data, processes, and outcomes. *Preprocessing* encompasses tasks that treat original data without generating novel insights. *Processing* features technical executions leading directly to insights using statistics, machine learning, graph analyses, and other techniques. *Postprocessing* covers operations dedicated to downstream application, maintenance, and deployment.



(a) Theme frequency in academic representations. (b) Theme frequency in notebooks. Horizontal lines indicate occurrences in code and text cells.

(c) Theme distribution across the notebook contents, by their relative position: 0% is the start of the notebook and 100% is the end of the notebook.

**Figure 1:** From top left, clockwise: theme frequency over all 42 annotated academic representations; theme frequency over all 39 annotated notebooks; theme distribution by position in a notebook.

Theme frequencies show that literature prioritizes technical over human aspects. Specifically, *preprocessing* is the most frequent, followed by *processing* and *postprocessing*. Conversely, human aspects, i.e. *understanding* and *reporting*, are less represented. Within the top themes, local maxima represent characteristic core themes. For instance, *feature-engineering* and *data-acquisition* distinguish theoretical *preprocessing*, but *modeling* is most prominent for *processing*. *Postprocessing* contains the most *quality-assessment* and *deployment*, whereas *understanding* prioritizes *intent* and *context*. *Reporting* theme is more evenly spread but *communication* is the most frequent.

Manually studying pipelines w.r.t. their articles’ domains highlights how research aims affect which activities are emphasized or obscured. Technically focused fields, e.g. in engineering and application domains, view data science mainly as a technical progression from raw data to practical assets. This is contested by other paradigms, asserting that “[other] contributing elements [...] should be properly recognized” [15, p.110]. To this end, empirical studies on data work emphasize *postprocessing* and *understanding* and focus on responsibilities, work distribution, collaboration patterns, and deployment impacts. The sole representation from the ethics study field omits activities other than *preprocessing* and *processing* to focus on select technical steps prone to ethical issues and algorithmic bias.

## 5. Workflows in Practice

Figure 1b shows how the theme frequency in the 39 Kaggle notebooks contrasts the academic models. *Reporting* is the most often observed activity, present both in markdown and code cells. Within it, textual *communication* is the most common core activity, used for narrative shaping, e.g. organizing sections and clarifying technical goals, followed by *visualization* with data charts and KG renderings. Technical themes *preprocessing* and *processing* are the two next most frequent. However, *postprocessing*

activities are the least represented. Crucially, formal knowledge structuring, ontology development, and semantic-enabled analyses are rarely represented in the annotated notebooks. Instead, machine learning algorithms for named entity and relation extraction during KG construction are more common, with local *application* briefly demonstrating the results’ potential. The higher number of *understanding* over *postprocessing* activities also contrasts with our literature insights, suggesting that notebook authors prioritize reflection over the operationalization of the KGs.

Analyzing theme positions in notebooks (in Figure 1c) reveals a coexistence of a relative global ordering and iterative cycles across theme transitions. On the top-theme level, notebooks maintain a general progression from introductory *preprocessing* to intermediate *processing* and final *postprocessing* blocks. However, this sequence is continuously surrounded by consistent occurrences of *reporting* and *understanding* activities, spanning the entire notebook content. Globally, we also observe two three-step iterations: *preprocessing*, *reporting*, *processing*, and back. Zooming into the core themes shows frequent transitions from *data-acquisition* to *feature-engineering*, and from *modeling* to immediate graph *application*. Additionally, *communication* and *visualization* act as cross-theme bridges. For example, textual narratives transition into technical *data-acquisition*, while graph visualizations follow graph *modeling* steps to render their resulting topologies. This structural co-existence indicates that KG practice in notebooks is locally iterative but forms a linear narrative process on the high level.

## 6. Discussion

Our results strongly resonate with the work of Biswas et al. [2], highlighting that practical workflows differ significantly from theoretical pipelines on paper. However, the strongest divergence in our study is the prevalence and even spread of *reporting* activities throughout Kaggle notebooks, contrasting sharply with their scarcity in academic models. Communication, documentation, and the display of results represent both a fundamental strength and an ongoing challenge in data science practice [10, 16]. When academic representations claim to be typical or generic yet only capture a technical snapshot of the practice, they introduce critical blind spots [17]. Uninformed reuse of these technology-centric models risks oversimplifying KG practitioners’ efforts and obscuring the collaborative and narrative dimensions of their artifacts. We claim that a structural recognition of communication activities in theoretical frameworks is needed to align paper models with notebook users’ realities [17].

Our findings also demonstrate that practical KG engineering is highly compatible with, and can be effectively discussed through, the lens of shared data science and knowledge engineering taxonomies. By evaluating notebooks qualitatively, we pivot from a traditional distributed systems lens on KG systems [1], stepping into the domain of empirical studies on these artifacts [9]. Our annotations reveal a significant role for technical operations alongside continuous visual and textual narrative-building. However, they also expose a lack of core KG formalisms: e.g. automated semantic reasoning, explicit ontology use, data integration, dynamic entity and relationship inference, etc. [3]. Compared to idealized views, practical notebooks simultaneously skip essential technical stages while introducing complex, human-centric narrative elements omitted by theory [3].

Hence, our findings suggest that KGs in Kaggle notebooks are primarily used for organizing and representing data on demand in graph form, rather than steps toward engineering extensive KG systems. Nevertheless, these environments serve a powerful pedagogical role, providing an accessible medium to demonstrate and share specialized KG construction methods with a broader data science community [11]. We thus encourage future work to consider notebooks as narrative environments for KG work, valuable for cross-disciplinary knowledge transfer, communication, and documentation [10].

## 7. Conclusion

We conducted a two-part empirical study examining how data science and knowledge engineering pipeline activities are reflected in practical Knowledge Graph (KG) workflows. Through a thematic analysis of academic pipeline representations, we developed an activity-based codebook and manually

annotated 39 KG notebooks from Kaggle. Our findings reveal a significant gap between theory and practice. While academic models emphasize technical stages, real-world KG notebooks often lack common discipline formalisms and advanced processing steps. Instead, large parts are devoted to reporting and communication activities, showing that computational environments act as narrative artifacts rather than mere interfaces for external KGs. For tool developers, this implies that to support KG work in notebooks, systems should prioritize narrative building, intermediate result visualization, and knowledge sharing, in addition to their focus on performance.

Future work should investigate practitioners' reasons behind the prevalence of reporting over elaborate application and KG formalisms at Kaggle. Expanding this analysis to other platforms and computational environments will clarify whether the specific theory-practice gap we observe reflects broader patterns. Ultimately, recognizing these human practices is essential for developing both better theoretical models and tools that effectively support practitioners working with knowledge graphs.

## Acknowledgments

This research was partially supported by Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, <https://www.hybridintelligence-centre.nl/>, under Grant No. (024.004.022)

## Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 3.5 Flash and Grammarly in order to perform grammar and spelling checks. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

- [1] S. Sakr, A. Bonifati, H. Voigt, A. Iosup, K. Ammar, R. Angles, W. Aref, M. Arenas, M. Besta, P. A. Boncz, K. Daudjee, E. D. Valle, S. Dumbrava, O. Hartig, B. Haslhofer, T. Hegeman, J. Hidders, K. Hose, A. Iamnitchi, V. Kalavri, H. Kapp, W. Martens, M. T. Özsu, E. Peukert, S. Plantikow, M. Ragab, M. R. Ripeanu, S. Salihoglu, C. Schulz, P. Selmer, J. F. Sequeda, J. Shinavier, G. Szárnyas, R. Tommasini, A. Tumeo, A. Uta, A. L. Varbanescu, H.-Y. Wu, N. Yakovets, D. Yan, E. Yoneki, The future is big graphs: a community view on graph processing systems, *Communications of the ACM* 64 (2021) 62–71. URL: <https://dl.acm.org/doi/10.1145/3434642>. doi:10.1145/3434642.
- [2] S. Biswas, M. Wardat, H. Rajan, The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large, in: *Proceedings of the 44th International Conference on Software Engineering*, ACM, Pittsburgh Pennsylvania, 2022, p. 2091–2103. URL: <https://dl.acm.org/doi/10.1145/3510003.3510057>. doi:10.1145/3510003.3510057.
- [3] G. Tamašauskaitė, P. Groth, Defining a Knowledge Graph Development Process Through a Systematic Review, *ACM Transactions on Software Engineering and Methodology* 32 (2023) 1–40. URL: <https://dl.acm.org/doi/10.1145/3522586>. doi:10.1145/3522586.
- [4] I. Cha, J. Oh, C. Y. Park, J. Han, H. Lee, Unlocking the tacit knowledge of data work in machine learning, in: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ACM, Hamburg Germany, 2023, p. 1–7. URL: <https://dl.acm.org/doi/10.1145/3544549.3585616>. doi:10.1145/3544549.3585616.
- [5] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. Vera Liao, C. Dugan, T. Erickson, How data science workers work with data: Discovery, capture, curation, design, creation, *Conference on Human Factors in Computing Systems - Proceedings* (2019) 1–14.
- [6] S. Passi, S. J. Jackson, Trust in data science: Collaboration, translation, and accountability in corporate data science projects, *Proceedings of the ACM on human-computer interaction* 2 (2018) 1–28. doi:10.1145/3274405.

- [7] M. Jakubowski, D. Tomaszuk, K. Hose, Bridging gaps in rdf validation: Insights and innovation opportunities in rdf validation practices, in: *Linking Meaning: Semantic Technologies Shaping the Future of AI: Proceedings of the 21st International Conference on Semantic Systems*, 3-5 September 2025, Vienna, Austria, SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2025, pp. 70–84.
- [8] K. Siorpaes, E. Simperl, Human Intelligence in the Process of Semantic Content Creation, *World Wide Web* 13 (2010) 33–59. doi:10.1007/s11280-009-0078-0.
- [9] L. Quaranta, F. Calefato, F. Lanubile, Eliciting Best Practices for Collaboration with Computational Notebooks, *Proceedings of the ACM on Human-Computer Interaction* 6 (2022) 87:1–87:41. URL: <https://dl.acm.org/doi/10.1145/3512934>. doi:10.1145/3512934.
- [10] M. B. Kery, M. Radensky, M. Arya, B. E. John, B. A. Myers, The Story in the Notebook: Exploratory Data Science using a Literate Programming Tool, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–11. URL: <https://dl.acm.org/doi/10.1145/3173574.3173748>. doi:10.1145/3173574.3173748.
- [11] K. Subramanian, N. Hamdan, J. Borchers, Casual Notebooks and Rigid Scripts: Understanding Data Science Programming, in: *2020 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 2020, pp. 1–5. URL: <https://ieeexplore.ieee.org/document/9127207>. doi:10.1109/VL/HCC50065.2020.9127207.
- [12] W. Epperson, A. Y. Wang, R. DeLine, S. M. Drucker, Strategies for reuse and sharing among data scientists in software teams, in: *Proceedings of the 44th International Conference on Software Engineering: Software Engineering in Practice, ICSE-SEIP '22*, Association for Computing Machinery, New York, NY, USA, 2022, p. 243–252. URL: <https://doi.org/10.1145/3510457.3513042>. doi:10.1145/3510457.3513042.
- [13] V. Braun, V. Clarke, N. Hayfield, G. Terry, Thematic Analysis, in: P. Liamputtong (Ed.), *Handbook of Research Methods in Health Social Sciences*, Springer, Singapore, 2019, pp. 843–860. URL: [https://doi.org/10.1007/978-981-10-5251-4\\_103](https://doi.org/10.1007/978-981-10-5251-4_103). doi:10.1007/978-981-10-5251-4\_103.
- [14] M. Risdal, T. Bozsolik, Meta kaggle, 2022. URL: <https://www.kaggle.com/ds/9>. doi:10.34740/KAGGLE/DS/9.
- [15] M. T. Özsu, Data Science—A Systematic Treatment, *Communications of the ACM* 66 (2023) 106–116. URL: <https://dl.acm.org/doi/10.1145/3582491>. doi:10.1145/3582491.
- [16] A. Willis, P. Charlton, T. Hirst, Developing Students' Written Communication Skills with Jupyter Notebooks, *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (2020) 1089–1095. URL: <https://dl.acm.org/doi/10.1145/3328778.3366927>. doi:10.1145/3328778.3366927.
- [17] S. Rahman, E. Kandogan, Characterizing Practices, Limitations, and Opportunities Related to Text Information Extraction Workflows: A Human-in-the-loop Perspective, in: *CHI Conference on Human Factors in Computing Systems*, ACM, New Orleans LA USA, 2022, pp. 1–15. URL: <https://dl.acm.org/doi/10.1145/3491102.3502068>. doi:10.1145/3491102.3502068.