

Seminar CS715

Large-Scale Data Integration

Data Integration using Deep Learning



- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
 - Web Data Integration
 - Data and Web Mining
 - Data Web Technologies
- Room: B6 - B1.15
- eMail: christian.bizer@uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30



Hallo

- **M. Sc. Wi-Inf. Ralph Peeters**
- Graduate Research Associate
- Research Interests:
 - Entity Matching using Deep Learning
 - Product Data Integration
- Room: B6, 26, C 1.04
- eMail: ralph.peeters@uni-mannheim.de



- **M. Sc. Wi-Inf. Alexander Brinkmann**
- Graduate Research Associate
- Research Interests:
 - Data Search using Deep Learning
 - Product Data Categorization
- Room: B6, 26, C 1.04
- eMail: alexander.brinkmann@uni-mannheim.de



- **M. Sc. Wi-Inf. Ketı Korini**
- Graduate Research Associate
- Research Interests:
 - Table Annotation using Deep Learning
 - Schema Matching
- Room: B6, 26, C 1.03
- eMail: kkorini@uni-mannheim.de



You and Your Experience

- A Short Round of Introductions
 - What are you studying?
 - Which DWS courses did you attend?
 - What kind of experience do you have with Data Science/Data Engineering projects?
- Participants
 - Vu, Tuananh
 - Jansen, Tim
 - Der, Reng
 - Elzamarany, Ahmed
 - Pradeep Kumar, Sreehari
 - Zong, Xu

Agenda of Today's Kickoff Meeting

1. Seminar organization
2. Seminar topics
3. How to structure your seminar paper / presentation?
4. Your Questions

1. Organization

Learning Targets

- Writing a seminar thesis as an exercise for your master thesis
- Understanding and presenting state-of-the-art scientific work
- Searching and citing scientific papers / journal articles
- How to structure your thesis and presentation
- How to argue, how to explain, how to write!
- How to write a nicely formatted paper using LaTeX

Schedule

Date	Session
Wednesday, 14.09.2022 (10:00-11:30)	Kick-off meeting and topic assignment
	Read papers about your topic and search for additional literature Prepare outline and argumentation line for the presentation
Until 7.10.2022	Meet with your mentor to discuss your presentation
	Prepare draft of your presentation
Until 28.10.2022	Send draft presentation to your mentor
	Finalize your presentation
Friday, 18.11.2022 (10:00-12:30)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 22.01.2023	Submission of your seminar thesis (70 % of your final grade)

Formal Requirements

- Presentation
 - 12 minutes + 8 minutes discussion
 - should be 100% understandable for all participants
- Written report (paper)
 - 12-15 pages single column
 - including abstract and appendixes
 - not including bibliography
 - every additional page reduces your grade by 0.3
 - written in English language
 - use latex template of Springer Computer Science Proceedings
 - <http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>
- Final grade
 - 70% written report
 - 30% presentation

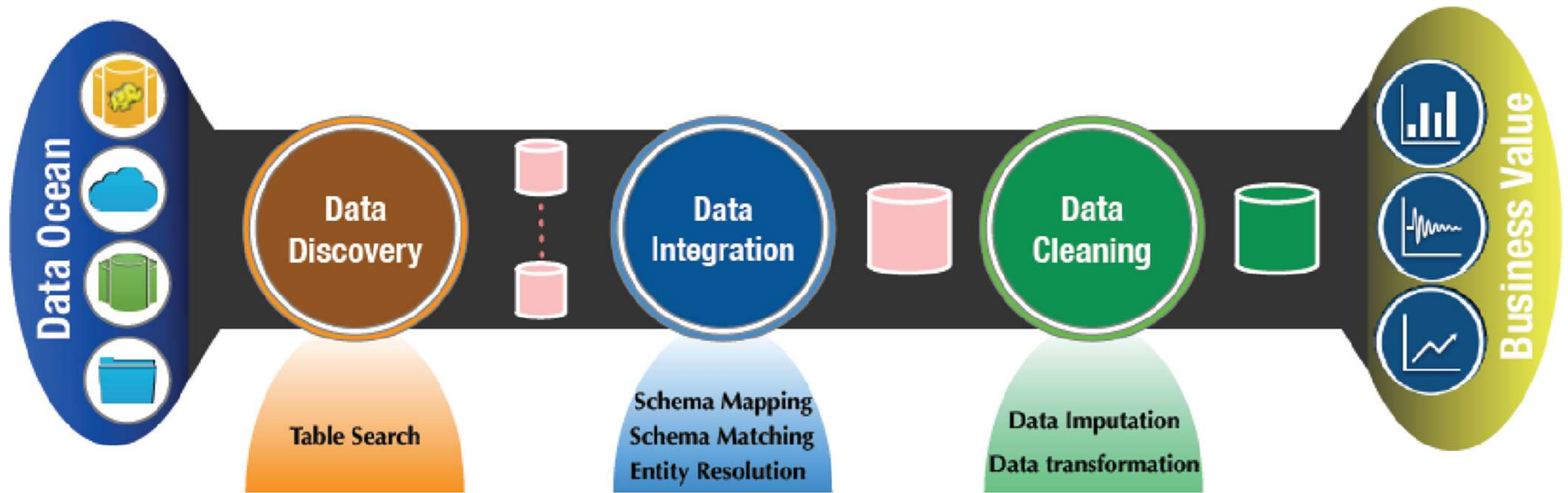
Which template to use?



<http://www.springer.com/de/it-informatik/Incs/conference-proceedings-guidelines>

2. Seminar Topics

Motivation and Overview



Symbolic Representations → **Distributed Representations**

Traditional Machine Learning → **Neural Nets**

Thirumuruganathan, Tang, Ouzzani, Doan: Data Curation with Deep Learning. EDBT 2020.

1. Entity Matching using Deep Learning

- Student: Elzamarany, Ahmed
- Mentor: Ralph Peeters
- N. Barlaug and J. A. Gulla, “Neural Networks for Entity Matching: A Survey,” *ACM Transactions on Knowledge Discovery from Data*, vol. 15, no. 3, p. 52:1-52:37, Apr. 2021.
- S. Mudgal *et al.*, “Deep Learning for Entity Matching: A Design Space Exploration,” in *Proceedings of the 2018 International Conference on Management of Data*, New York, NY, USA, 2018, pp. 19–34.
- Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, “Deep entity matching with pre-trained language models,” *Proceedings of the VLDB Endowment*, vol. 14, no. 1, pp. 50–60, Sep. 2020.
- More references and benchmarks: [Papers with Code: Entity Resolution](#)

2. Entity Matching using Contrastive Learning

- Student: Pradeep Kumar, Sreehari
- Mentor: Ralph Peeters
- R. Peeters and C. Bizer, “Supervised Contrastive Learning for Product Matching.” in *Companion Proceedings of the Web Conference 2022*, Lyon, France, April 2022.
- R. Wang, Y. Li, and J. Wang, “Sudowoodo: Contrastive Self-supervised Learning for Multi-purpose Data Integration and Preparation.” arXiv:2207.04122 [cs], Jul. 08, 2022.
- M. Almagro, D. Jiménez, D. Ortego, E. Almazán, and E. Martínez, “Block-SCL: Blocking Matters for Supervised Contrastive Learning in Product Matching.” arXiv:2207.02008 [cs], Jul. 05, 2022.

3. Active Learning for Entity Matching

- Student: Der, Reng
- Mentor: Ralph Peeters
- J. Huang, W. Hu, Z. Bao, Q. Chen, and Y. Qu, “Deep entity matching with adversarial active learning,” *The VLDB Journal*, Apr. 2022.
- A. Jain, S. Sarawagi, and P. Sen, “Deep indexed active learning for matching heterogeneous entity representations,” *Proceedings of the VLDB Endowment*, vol. 15, no. 1, pp. 31–45, Sep. 2021.
- A. Bogatu, N. W. Paton, M. Douthwaite, S. Davie, and A. Freitas, “Cost-effective Variational Active Entity Resolution,” in *2021 IEEE 37th International Conference on Data Engineering*, Apr. 2021, pp. 1272–1283.
- More references and benchmark results: [Papers with Code: MusicBrainz20K](#)

4. Representation Learning for Tabular Data

- Student: Vu, Tuananh
- Mentor: Ketii Korini
- D. Wang, P. Shiralkar, C. Lockard, B. Huang, X. L. Dong, et al., “TCN: Table Convolutional Network for Web Table Interpretation,” in Proceedings of the Web Conference 2021, New York, NY, USA, Apr. 2021, pp. 4020–4032
- H. Iida, D. Thai, V. Manjunatha, and M. Iyyer, “TABBIE: Pretrained Representations of Tabular Data,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, Jun. 2021, pp. 3446–3456
- Z. Wang et al., “TUTA: Tree-based Transformers for Generally Structured Table Pre-training,” in Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, New York, NY, USA, Aug. 2021, pp. 1780–1790
- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: table understanding through representation learning,” Proceedings of the VLDB Endowment, vol. 14, no. 3, Nov. 2020, pp. 307–319

5. Column Type Annotation in Tabular Data

- Student: Jansen, Tim
- Mentor: Ketil Korini
- Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, Ç. et al. “Annotating columns with pre-trained language models.” In Proceedings of the 2022 International Conference on Management of Data, 2022, pp. 1493–1503
- Zhang, D., Suhara, Y., Li, J., Hulsebos, M., Demiralp, Ç. et al. Sato: Contextual semantic type detection in tables. arXiv preprint arXiv:1911.06311. 2019
- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: table understanding through representation learning,” Proceedings of the VLDB Endowment, vol. 14, no. 3, Nov. 2020, pp. 307–319
- More references and benchmarks: [Papers with Code: Column Type Annotation](#)

6. Information Extraction for E-Commerce Product Data

- Student: Zong, Xu
- Mentor: Alexander Brinkmann
- Xinyang Zhang, Chenwei Zhang, Xian Li, Xin Luna Dong, Jingbo Shang, Christos Faloutsos, and Jiawei Han. 2022. OA-Mine: Open-World Attribute Mining for E-Commerce Products with Weak Supervision. In *Proceedings of the ACM Web Conference 2022*, ACM, Virtual Event, Lyon France, 3153–3161.
- Huimin Xu, Wenting Wang, Xin Mao, Xinyu Jiang, and Man Lan. 2019. Scaling up Open Tagging from Tens to Thousands: Comprehension Empowered Attribute Value Extraction from Product Title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 5214–5223.
- Gilad Fuchs and Yoni Acriche. 2022. Product Titles-to-Attributes As a Text-to-Text Task. In *Proceedings of The Fifth Workshop on e-Commerce and NLP*, Association for Computational Linguistics, Dublin, Ireland, 91–98.

3. How to Structure Your Paper / Presentation

Goal of Seminar Paper

- A seminar paper differs significantly from a master thesis
 - The topic is already defined
 - No need to implement or develop algorithms
 - No need to perform experiments
 - Primarily: reproduction and re-organization of content that is already available
- Goals of seminar paper
 1. Describe the problem / task
 2. describe several existing methods/systems for handling the task,
 3. compare the methods/systems and their evaluation using a systematic set of comparison criteria

How to Structure Your Paper?

1. Introduction and Problem Statement

- Which problem/task is addressed? Why is the problem important?
- Structure of your paper

2. Description of Existing Approaches

- Overview of existing methods and features used by the methods
- Detailed description of **selected methods**
- Comparison of the selected methods using a **set of comparison criteria**

3. Evaluation

- Comparison and **discussion of the evaluation tasks**, metrics
- Comparison of the evaluation results

4. Conclusion

- What did the comparison of the methods and evaluation results show?
- Can something be concluded for future work?

5. Bibliography

Learn from Examples

- Read survey articles and identify the structure from the previous slide
 - Why can this paragraph be found at that position?
 - What is the purpose of some section / subsection?
- Important
 - Read survey articles!
 - Read conference or journal papers.
- Textbook on how to write a thesis
 - Zobel: Writing for Computer Science, 3rd Edition, Springer 2014.
- University Library: Academic Writing Consultancy
 - <https://www.bib.uni-mannheim.de/en/writing-consultancy/>

Citing different Types of Publications

- Journal article
 - Good to cite, current research results
 - Survey articles (very good for an overview)
- Conference and workshop paper
 - Good to cite, current research results
- Books (sometimes cited)
 - Textbooks
 - Collections of articles/papers => Cite specific paper in book
- Websites
 - better not cited, exceptions are, e.g., W3C Specifications
 - Wikipedia is not an exception!!! **Do not cite Wikipedia, ever!**
- Slide sets (especially from our lectures)
 - **Never cite!**

How to Find Relevant Publications?

- Use Standard Search Engines
- **Use Google Scholar**
 - we use it a lot ourselves
- Search Engines of the University's library
 - see slides from the library course
- **Exploit references:** Given a relevant document x
 - Follow references in the past: papers y that x has cited
 - Follow references in the future: papers y that cited x („**cited by**” functionality in Google scholar)

4. Questions?

