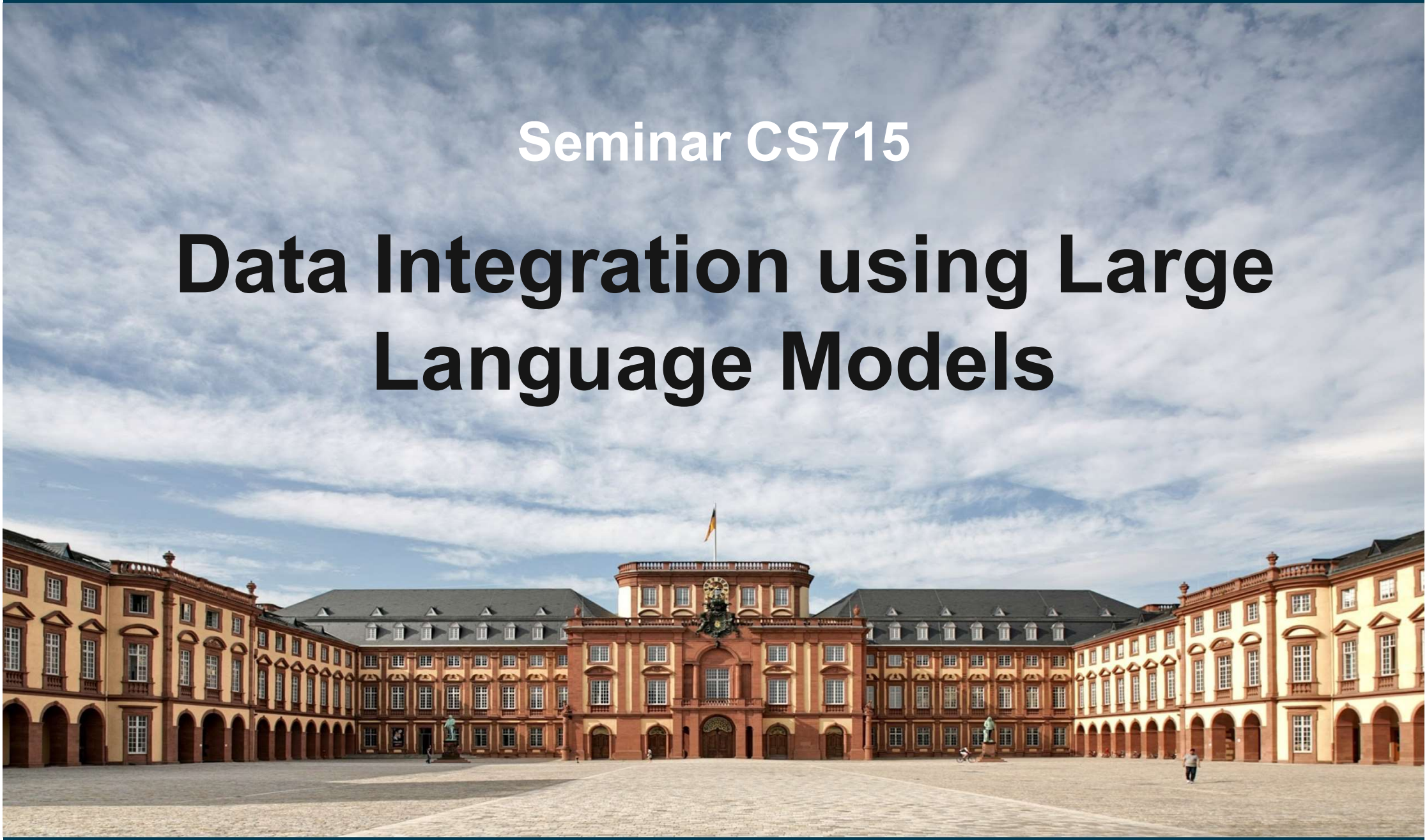
UNIVERSITY OF MANNHEIM

Seminar CS715

Data Integration using Large Language Models



Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web Data Integration
  - Data and Web Mining
  - Deployment of 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

Hallo

- **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

Hallo

- **M. Sc. Wi-Inf. Keti 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 integration projects?
- Participants
  1. Baumann, Nick
  2. Chen, Chun-Yi
  3. Vogli, Aleksandro
  4. Golchha, Pujit
  5. Joseph, Abhay
  6. Chauhan, Vishal
  7. Steiner, Aaron
  8. Shyamsundar, Sharan
  9. Joo, Eun
  10. Lichwa, Mateusz

Agenda of Today‘s Kickoff Meeting

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

1. Seminar Organization

Learning Goals

- Writing a seminar thesis as an exercise for your master thesis
- Understanding and presenting state-of-the-art scientific work
- Designing experiments and present experimental results
- Searching and citing scientific papers / journal articles
- How to structure your thesis and presentation
- How to write a scientific paper using LaTeX

Schedule

Date	Session
Thursday, 02.03.2023 (10:00-11:30)	Kick-off meeting and topic/mentor assignment
	Read papers about your topic and search for additional literature Design experimental setup (if applicable) Prepare outline and argumentation line for the presentation
Until 24.03.2023	Meet with your mentor to discuss outline and argumentation
	Prepare draft of your presentation
Until 21.04.2023	Send draft presentation to your mentor
	Finalize your presentation
Friday, 12.05.2023 (10:00-12:30)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 09.07.2023	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
  - use latex template of Springer Computer Science Proceedings
    - http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines
- Final grade
  - 70% written report
  - 30% presentation

Which template to use?

my title

author1 name[1] and author2 name[1,2]

[1] University 1
email1@gmail.com
[2] University 2
email2@gmail.com

Abstract. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

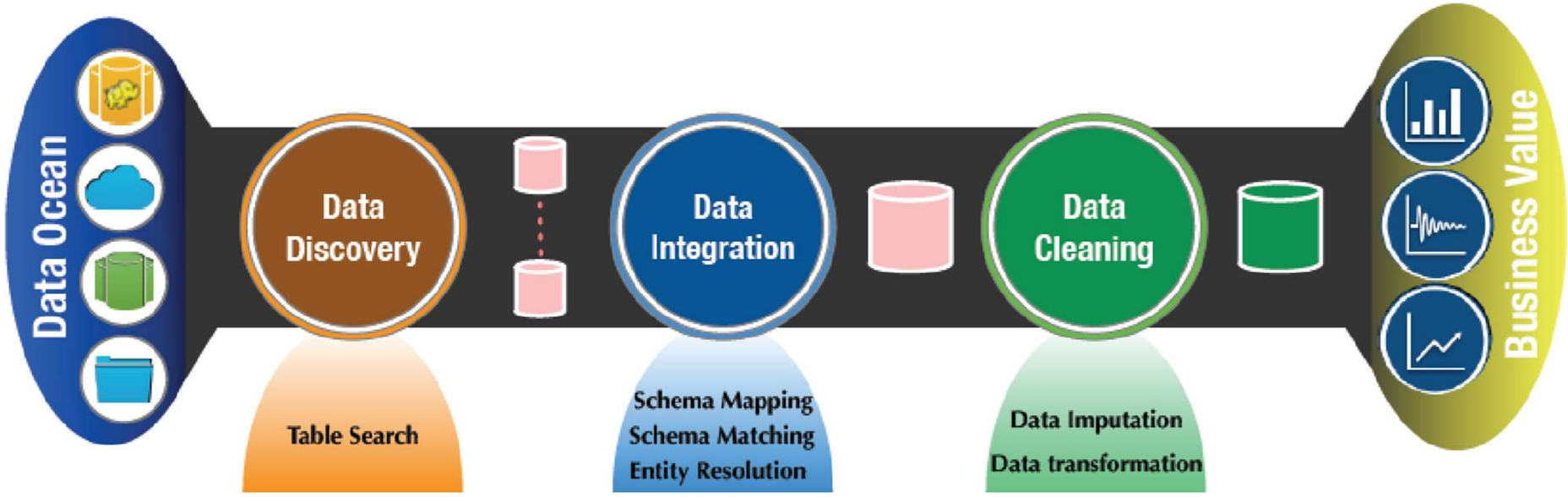
1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

http://www.springer.com/de/it-informatik/lncs/conference-proceedings-guidelines

2. Seminar Topics and Topic Assignment

Data Integration using Large Language Models (LLM)

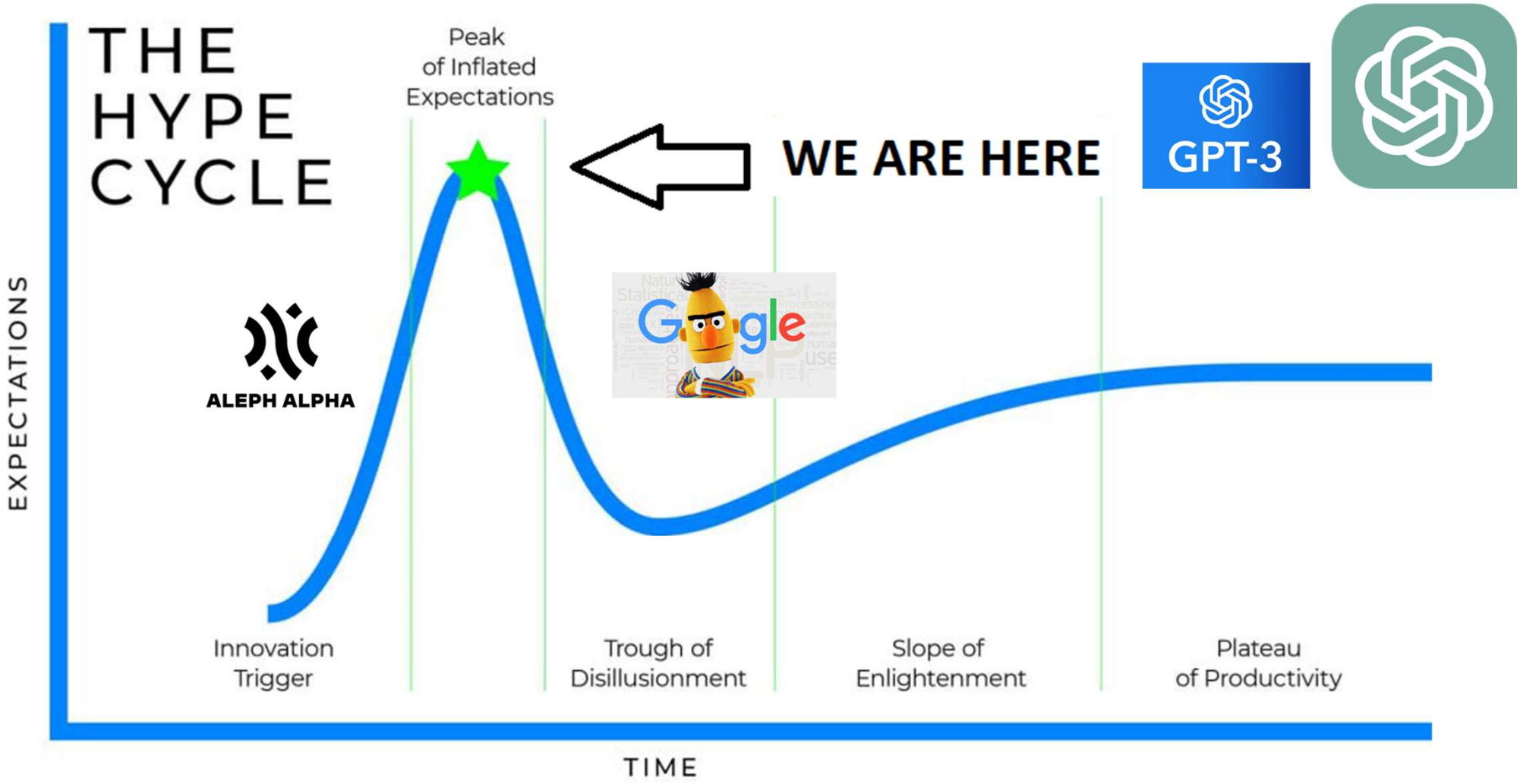


Symbolic Representations → **Distributed Representations**

Traditional Machine Learning → **Deep Neural Networks**

Thirumuruganathan, Tang, Ouzzani, Doan: Data Curation with Deep Learning. EDBT 2020.
Avanika Narayan et al.: Can Foundation Models Wrangle Your Data? PVLDB Vol.16 4, 2022.

Large Language Models



Topics

1. Deep Learning for Table Search

- Student: Vogli, Aleksandro
- Mentor: Ralph Peeters

- G. Fan, J. Wang, Y. Li, D. Zhang, and R. Miller, “Semantics-aware Dataset Discovery from Data Lakes with Contextualized Column-based Representation Learning.” arXiv:2210.01922 [cs], January 2023.
- A. Bogatu, A. A. A. Fernandes, N. W. Paton, and N. Konstantinou, “Dataset Discovery in Data Lakes,” in 2020 IEEE 36th International Conference on Data Engineering (ICDE), Apr. 2020, pp. 709–720.
- A. Das Sarma et al., “Finding related tables,” in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, May 2012, pp. 817–828.

Topics

2. Experimental Topic: Evaluating Large Language Models on the Task of Entity Matching

- Student: Steiner, Aaron
- Mentor: Ralph Peeters

- P. Wang et al.: PromptEM: Prompt-tuning for low-resource generalized entity matching. Proceedings of the VLDB Endowment. Volume 16, Issue 2, pp 369–378. November 2022.
- A. Narayan, I. Chami, L. Orr, S. Arora, and C. Ré. 2022. Can Foundation Models Wrangle Your Data? PVLDB Vol.16 Issue 4, 2022.
- A. Srivastava et al., “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. arXiv:2206.04615 [cs], June 2022.
- Q. Dong et al., “A Survey for In-context Learning”. arXiv:2301.00234 [cs], Dec. 2022.
- A. Venkatesh et al., “On Evaluating and Comparing Open Domain Dialog Systems.” arXiv:1801.03625 [cs], Dec. 2018.

Topics

3. Representation Learning for Missing Value Imputation

– Student: Joseph, Abhay

– Mentor: Alexander Brinkmann

- Richard Wu, Aoqian Zhang, Ihab Ilyas, and Theodoros Rekatsinas. 2020. Attention-based Learning for Missing Data Imputation in HoloClean. MLSys 2020, 307–325.
- Avanika Narayan et al.: Can Foundation Models Wrangle Your Data? PVLDB Vol.16 4, 2022.
- Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. TURL: Table Understanding through Representation Learning. SIGMOD 2022, 33–40.
- J. Yoon, J. Jordon, and M. Schaar. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. ICML 2018, 5689–5698.
- Ihab F. Ilyas and Theodoros Rekatsinas. 2022. Machine Learning and Data Cleaning: Which Serves the Other? JDIQ, 14, 3 (September 2022), 1–11.

Topics

4. Experimental Topic: Evaluating Large Language Models on the Task of Missing Value Imputation for Knowledge Graph Completion

- Student: Chen, Chun-Yi
- Mentor: Alexander Brinkmann

- A. Venkatesh et al., “On Evaluating and Comparing Open Domain Dialog Systems.” arXiv:1801.03625 [cs], 2018.
- A. Srivastava et al., “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. arXiv:2206.04615 [cs], June 2022.
- Q. Dong et al., “A Survey for In-context Learning”. arXiv:2301.00234 [cs], Dec. 2022.
- Richard Wu, Aoqian Zhang, Ihab Ilyas, and Theodoros Rekatsinas. 2020. Attention-based Learning for Missing Data Imputation in HoloClean. MLSys 2020, 307–325.
- Avanika Narayan et al.: Can Foundation Models Wrangle Your Data? PVLDB Vol.16 4, 2022.
- https://paperswithcode.com/task/knowledge-graph-completion

Topics

5. Schema Matching using Deep Learning

- Student: Golchha, Pujit
- Mentor: Keti Korini

- Zhang, Jing, et al. “SMAT: An attention-based deep learning solution to the automation of schema matching.” European Conference on Advances in Databases and Information Systems. Springer, Cham, 2021.
- Shraga, Roee, Avigdor Gal, and Haggai Roitman. “Adnev: Cross-domain schema matching using deep similarity matrix adjustment and evaluation.” Proceedings of the VLDB Endowment 13.9 (2020): 1401–1415.
- Koutras, Christos, et al. “REMA: Graph Embeddings-based Relational Schema Matching.” EDBT/ICDT Workshops. 2020.
- Rahm, E., Bernstein, P. A survey of approaches to automatic schema matching. The VLDB Journal 10 (2001), 334–350.

Topics

6. Cell Entity Annotation in Tabular Data

- Student: Lichwa, Mateusz
- Mentor: Keti Korini

- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: table understanding through representation learning,” Proc. VLDB Endow., vol. 14, no. 3, Nov. 2020, pp. 307–319
- Huynh, V.P., Liu, J., Chabot, Y., Labbé, T., Monnin, P. and Troncy, R., DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. In SemTab@ ISWC, Nov. 2020, (pp. 27–39).
- Chen, S., Karaoglu, A., Negreanu, C., Ma, T., Yao, J.G., Williams, J., Jiang, F., Gordon, A. and Lin, C.Y. LinkingPark: An automatic semantic table interpretation system. Journal of Web Semantics, 74, 2022, p.100733.
- More references and benchmarks: Papers with Code: Cell Entity Annotation

Topics

7. Deep Tabular Learning for Domain-Specific Prediction Tasks

– Student: Shyamsundar, Sharan

– Mentor: Keti Korini

- Yoon, Jinsung, et al. “Vime: Extending the success of self-and semi-supervised learning to tabular domain.” Advances in Neural Information Processing Systems 33 (2020).
- Somepalli, Gowthami, et al. “Saint: Improved neural networks for tabular data via row attention and contrastive pre-training.” arXiv preprint arXiv:2106.01342 (2021).
- Gharibshah, Zhabiz, and Xingquan Zhu. “Local Contrastive Feature Learning for Tabular Data.” Proceedings of the 31st ACM International Conference on Information & Knowledge Management (2022).
- Stefan Hegselmann, et al. “TabLLM: Few-shot Classification of Tabular Data with Large Language Models” arXiv:2210.10723 [cs.CL] (2022).
- Borisov, Vadim, Tobias Leemann, et al. “Deep neural networks and tabular data: A survey.” IEEE Transactions on Neural Networks and Learning Systems (2022).

Topics

8. Information Extraction for E-Commerce Product Data

– Student: Baumann, Nick

– 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. WWW 2022, 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. ACL 2019, 5214–5223.
- Qifan Wang, et al. 2020: Learning to Extract Attribute Value from Product via Question Answering: A Multi-task Approach. SIGKDD 2020.

Topics

9. Experimental Topic: Evaluating Large Language Models on the Task of Product Information Extraction

- Student: Chauhan, Vishal
- Mentor: Alexander Brinkmann

- A. Venkatesh et al. 2018, “On Evaluating and Comparing Open Domain Dialog Systems.” arXiv:1801.03625 [cs], Dec. 2018.
- A. Srivastava et al., “Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models”. arXiv:2206.04615 [cs], June 2022.
- Q. Dong et al., “A Survey for In-context Learning”. arXiv:2301.00234 [cs], Dec. 2022.
- Li Yang et al. 2022: MAVE: A Product Dataset for Multi-source Attribute Value Extraction. WSDM 2022.
- P. Petrovski et al. 2016: The wdc gold standards for product feature extraction and product matching. EC-Web 2016.

Topics

10. Experimental Topic: Combining WebAPIs and Large Language Models for Question Answering via In-Context Learning

- Student: Joo, Eun
- Mentor: Ralph Peeters

- Omar Khattab, et al.: Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. arXiv:2212.14024 [cs.CL], Dec. 2022.
- Q. Dong et al., “A Survey for In-context Learning”. arXiv:2301.00234 [cs], Dec. 2022.
- Christopher Potts: Stanford online seminar – GPT-3 & Beyond. Starting from minute 28:13, Jan 2023.
- Example Task: Ask ChatGPT or GPT3 questions about restaurants or hotels in Mannheim using TripAdvisor data and in-context learning.

3. How to Structure Your Paper / Presentation

Goals of Literature and Experimental Papers

- Goals of Literature Papers
  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

- Goals of Experimental Papers
  1. describe the problem / task
  2. design an experimental setup for evaluating LLM on the task
  3. present and discuss the results of your experiments

How to Structure Your Literature 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 (likely two)
   - 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 using a set of comparison criteria
4. Conclusion
   - What did the comparison of the methods and evaluation results show?
   - Can something be concluded for future work?
5. Bibliography

How to Structure Your Experimental Paper?

1. Introduction and Problem Statement
   - Which problem/task is addressed? Why is the problem important?
   - Overview of existing methods and benchmarks used for evaluation
   - Structure of your paper
2. Description of Experimental Design
   - How to you select examples for which challenges from which benchmark?
   - Which prompt designs do you test?
   - Which Large Language Models do you benchmark?
3. Presentation of Experimental Results
   - Present the results of your experiments (tables containing values and deltas).
   - Present the results of your error analysis (types of errors, frequency of these types)
4. Conclusion
   - What did the experiments and the error analysis show?
   - What can be concluded for future work?
5. Bibliography

Learn from Examples

- Read survey articles and previous experimental papers and identify the structure from the previous slides
  - 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., documents like W3C Specifications
  - Do not cite Wikipedia, ever!
  - Use footnotes to refer to project pages, download pages, or technical documentation
- 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?