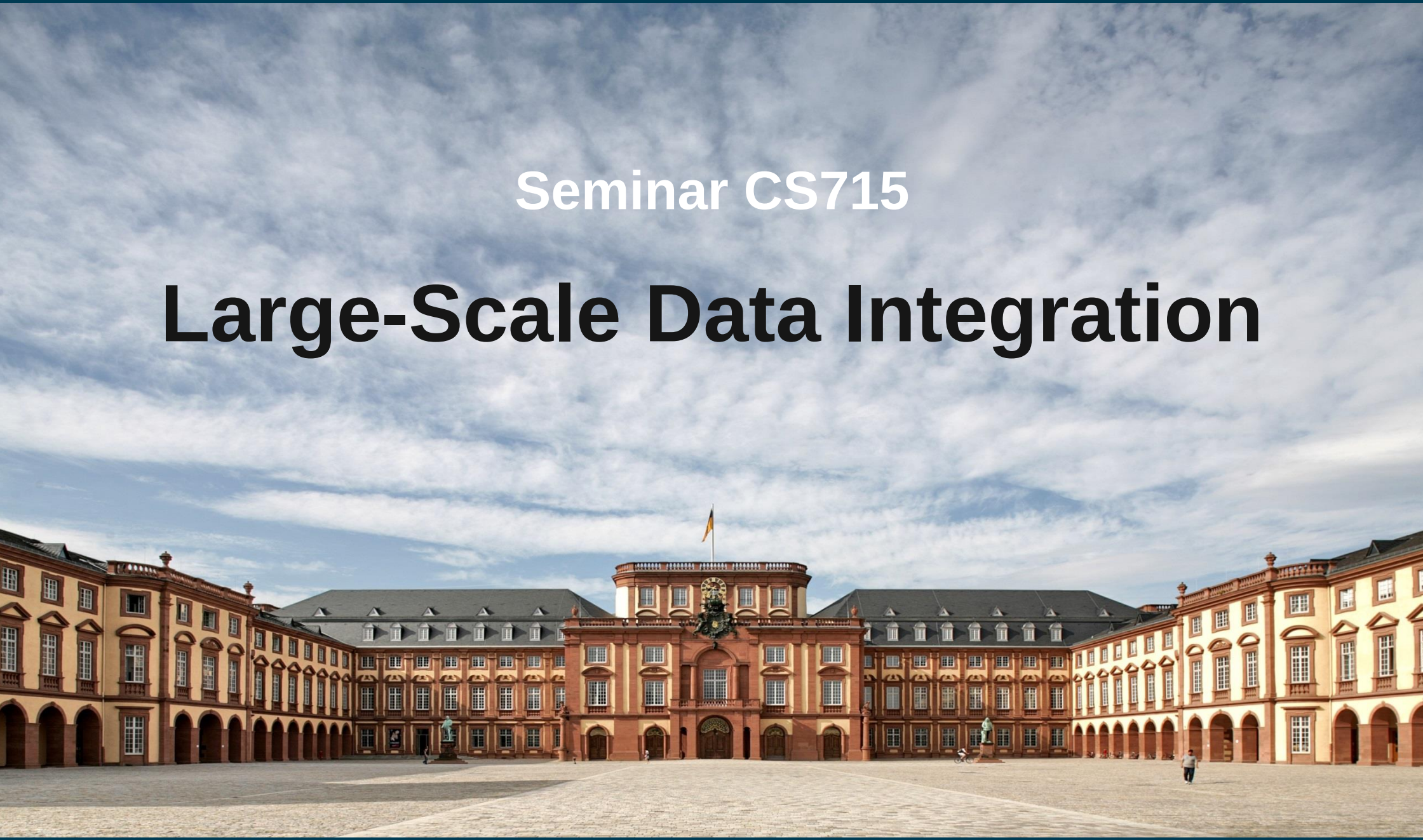
UNIVERSITY OF MANNHEIM

Seminar CS715

Large-Scale Data Integration



Hallo

- **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: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30

Hallo

- **Anna Primpeli**
- Graduate Research Associate
- Research Interests:
  - Data Extraction
  - Web Data Integration
  - Active Learning
  - Structured Data on the Web
- Room: B6, 26, C 1.04
- eMail: anna@informatik.uni-mannheim.de

Hallo

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

Agenda of Today‘s Kickoff Meeting

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

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, 19.02.2020	Send list of preferred topics via eMail
Monday, 02.03.2020, 10:00	Kick-off meeting and topic assignment (today)
	Read papers about your topic and search for additional literature
Friday, 13.03.2020	Drop-out deadline (A drop-out after this deadline will be graded with 5.0)
	Prepare outline and argumentation line for the presentation
Until Friday 03.04.2020	Meet with your mentor to discuss your presentation
	Prepare draft of your presentation
Until Sun. 03.05.2020	Send draft presentation to your mentor
	Finalize your presentation
Friday, 15.05.2020 (9:00-12:30)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 28.06.2020	Submission of your seminar thesis (70 % of your final grade)

Formal Requirements

- Presentation
  - 15 minutes + 10 minutes discussion
  - should be 100% understandable for all participants
- Written report (paper)
  - 10-12 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/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

Motivation of the Seminar

1. The number of data sources on the Web as well as in enterprise contexts steadily increases.
   - Linked Data, Web Tables, Schema.org data, Excel files on the intranet
   - Data Lakes
2. Traditional data integration techniques
   1. do not scale to these new requirements
   2. do not exploit the resulting new opportunities
3. Thus, this seminar covers the question how to adjust integration techniques so that they scale to new settings and properly exploit the new opportunities.

General Literature

- Dong/Srivastava: Big Data Integration. Morgan & Claypool, 2015.
- Doan/Halevy: Principles of Data Integration. Morgan Kaufmann, 2012.
- Christophides/Efthymiou: Entity Resolution in the Web of Data. Morgan & Claypool, 2015.

Topics

1. Dataset Search (Egi, Chris)

- Chapman, A., Simperl, E., Koesten, L. *et al.* Dataset search: a survey. *The VLDB Journal* (2019) doi:10.1007/s00778-019-00564-x
- Embedding based appraoch: http://www.cse.lehigh.edu/~brian/pubs/2019/BigData/Improved_Table_Retrieval.pdf
- Trabelsi, et al.: Improved Table Retrieval Using Multiple Context Embeddings for Attributes. Big Data 2019.
- WWW2018 Tutorial: Deep Learning for Matching in Search and Recommendation

2. Profiling Semantic Annotations for Dataset Search (Marius, Chris)

- Natascha Noy: Discovering millions of datasets on the web https://blog.google/products/search/discovering-millions-datasets-web/
- Anna Primpeli: WebDataCommon Schema.org Data from November 2019 Common Crawl

Supervisor e-mail: chris@informatik-uni-mannheim.de

Topics

3. Profiling of Relational Data for Data Integration (Ruyue, Anna)

- Naumann, Felix. "Data profiling revisited." *ACM SIGMOD Record* 42.4 (2014): 40-49.
- Dong, Xin Luna, and Divesh Srivastava. "Big data integration." *Synthesis Lectures on Data Management* 7.1 (2015): 1-198. (Chapter 5.3)
- Abedjan, Ziawasch, Lukasz Golab, and Felix Naumann. "Profiling relational data: a survey." *The VLDB Journal—The International Journal on Very Large Data Bases* 24.4 (2015): 557-581.
- Data Profiling Tool: https://pypi.org/project/pandas-profiling/
- Fan, Wenfei, et al. "Discovering conditional functional dependencies." *IEEE Transactions on Knowledge and Data Engineering* 23.5 (2010): 683-698.
- Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, and Felix Naumann. 2015. Data profiling with metanome. Proc. VLDB Endow. 8, 12 (August 2015), 1860-1863.
- Metanome: https://github.com/HPI-Information-Systems/metanome-algorithms

4. Transfer Learning for Entity Matching (Xi, Anna)

- S. N. Negahban, B. I. Rubinstein, and J. G. Gemmell. Scaling multiple-source entity resolution using statistically efficient transfer learning. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 2224–2228. ACM, 2012.
- Thirumuruganathan, Saravanan, Shameem A. Puthiya Parambath, Mourad Ouzzani, Nan Tang, and Shafiq Joty. "Reuse and Adaptation for Entity Resolution through Transfer Learning." *arXiv preprint arXiv:1809.11084* (2018).
- Tan, Chuanqi, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. "A Survey on Deep Transfer Learning." In *International Conference on Artificial Neural Networks*, pp. 270-279. Springer, Cham, 2018.

Topics

5. Active Learning for Entity Resolution (Yuxin, Anna)

- Settles, B.: Active Learning. Synthesis Lectures on AI and ML. 6, 1, 1–114 (2012).
- Qian, K. et al.: Active Learning for Large-Scale Entity Resolution. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 1379–1388 ACM (2017).
- Ngomo, A.-C.N., Lyko, K.: Eagle: Efficient active learning of link specifications using genetic programming. In: Extended Semantic Web Conference. pp. 149–163 Springer (2012)..
- Primpeli, Anna, and Christian Bizer. "Robust active learning of expressive linkage rules." Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics. 2019
- Isele, R., Bizer, C.: Active learning of expressive linkage rules using genetic programming. Web Semantics: Science, Services and Agents on the World Wide Web. 23, 2–15 (2013).
- Chen, X. et al.: Heterogeneous Committee-Based Active Learning for Entity Resolution (HeALER). In: Welzer, T. et al. (eds.) Advances in Databases and Information Systems. pp. 69–85 Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-28730-6_5..

Supervisor e-mail: anna@informatik-uni-mannheim.de

Topics

6. Explaining entity resolution methods (Daniel, Ralph)

- A. Ebaid, S. Thirumuruganathan, W. G. Aref, A. Elmagarmid, and M. Ouzzani, “EXPLAINER: Entity Resolution Explanations,” in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 2019, pp. 2000–2003, doi: 10.1109/ICDE.2019.00224.
- S. Thirumuruganathan, M. Ouzzani, and N. Tang, “Explaining Entity Resolution Predictions : Where are we and What needs to be done?,” p. 6. *HILDA'19* (2019).
- L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining Explanations: An Overview of Interpretability of Machine Learning,” in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89, doi: 10.1109/DSAA.2018.00018.

7. Matching Numeric Attributes using symbolic and subsymbolic features (Cagdas, Ralph)

- Phuc Nguyen, et al: EmbNum+: Effective, Efficient, and Robust Semantic Labeling for Numerical Values. New Generation Computing Journal, 2019.
- Sebastian Neumaier, et al.: Multi-level Semantic Labelling of Numerical Values. International Semantic Web Conference, 2016.
- Ibrahim, Yusra, et al. "Bridging quantities in tables and text." *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2019.
- Ibrahim, Yusra, and Gerhard Weikum. "Exquisite: explaining quantities in text." *The World Wide Web Conference*. 2019.

Supervisor e-mail: ralph@informatik-uni-mannheim.de

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
- Goal of seminar paper
  - Describe the problem, describing several existing methods for handling the problem, comparing the methods and their evaluation using a systematic comparison schema

How to Structure Your Paper?

1. Introduction and Problem Statement
   - Which problem 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 two selected methods
3. Evaluation
   - Comparison and discussion of the used evaluation tasks, datasets, metrics
   - Comparison of the evaluation results
4. Discussion and Conclusion
   - What does the comparison of the methods and evaluation results show?
   - What can 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/
  - Open consulting hour: every Wednesday 10 am - noon

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
  - 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?