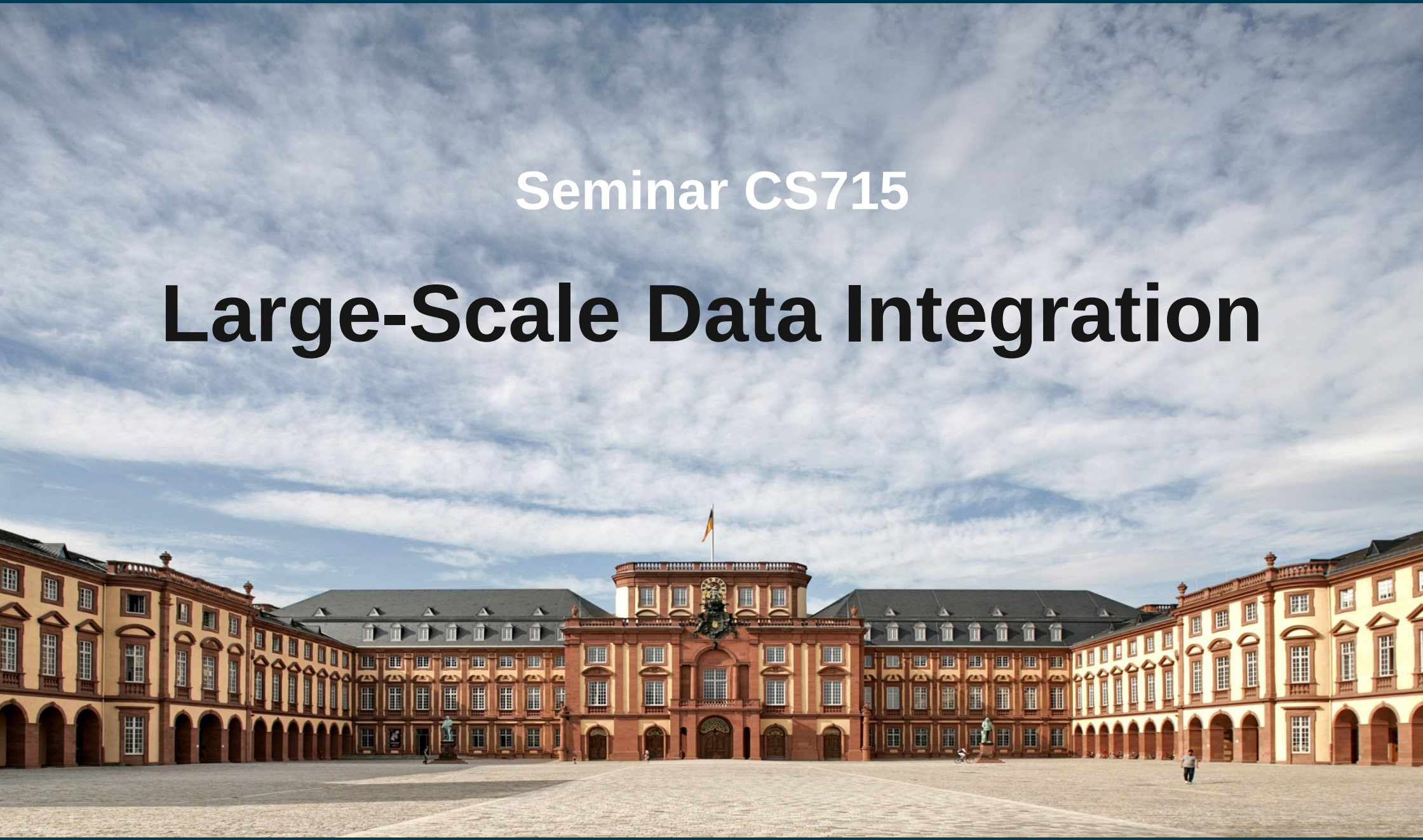
UNIVERSITY OF MANNHEIM

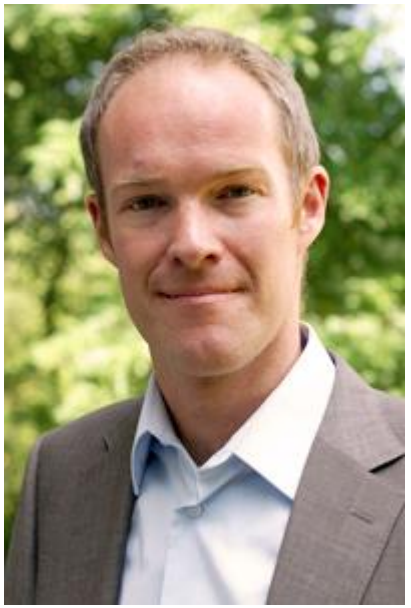
Seminar CS715

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

- **Yaser Oulabi**
- Graduate Research Associate
- Research Interests:
  - Web Data Integration
  - Knowledge Base Completion
  - Data Trustworthiness and Fusion
- Room: B6, 26, C 1.03
- eMail: yaser@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, 20.02.2019	Send list of preferred topics via eMail
Monday, 25.02.2019, 10:00	Kick-off meeting and topic assignment (today)
	Read papers about your topic and search for additional literature
Friday, 08.03.2019	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 29.03.2019	Meet with your mentor to discuss your presentation
	Prepare draft of your presentation
Until Sun. 28.04.2019	Send draft presentation to your mentor
	Finalize your presentation
Friday, 10.05.2019 **Monday, 13.05.2019**	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 23.06.2019	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.

Topics

1. **Product Data Matching using Embeddings and Deep Neural Networks (**Vasili Bocicariov, **Chris)**
   - Mudgal, S. et al.: Deep Learning for Entity Matching: A Design Space Exploration. In: Proceedings of the 2018 International Conference on Management of Data (2018)
   - Shah, K., Kopru, S., Ruvini, J.D.: Neural Network based Extreme Classication and Similarity Models for Product Matching. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. (2018)
2. **Product Taxonomy Matching and Integration (**Anne Katrin Michael, **Chris)**
   - Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag, Berlin Heidelberg (2013).
   - Park, S., Kim, W.: Ontology Mapping Between Heterogeneous Product Taxonomies in an Electronic Commerce Environment. Int. J. Electronic Commerce. 12, 69–87 (2007).
3. **Push Interest Prediction using Machine Learning (**Janik Gassner, **Chris)**
   - Industry Topic together with STOCARD

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

Topics

4. **A comparison of two families of active learning query strategies: committee based and structure based (density weighted) methods (**Marius Bock, **Anna)**
   - Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. 6, 1, 1–114 (2012 (Chapters 3&5).
   - Settles, Burr, and Mark Craven. „An analysis of active learning strategies for sequence labeling tasks.“ Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2008.

5. **Hierarchical Categorization of Products using Embeddings and Deep Neural Networks (**Nils Richter, **Anna)**
   - Silla, Carlos N., and Alex A. Freitas. „A survey of hierarchical classification across different application domains.“ Data Mining and Knowledge Discovery 22.1-2 (2011): 31-72.
   - Xiong, Tengke, and Putra Manggala. „Hierarchical Classification with Hierarchical Attention Networks.“ (2018).

6. **Profiling schema.org Event Data on the Web (**Yin-Feng Li, **Anna)**
   - Robert Meusel, Petar Petrovski and Christian Bizer: The WebDataCommons Microdata, RDFa and Microformat Dataset Series. 13th International Semantic Web Conference (ISWC), 2014.
   - Anna Primpeli: WebDataCommon Schema.org Data Extracted from November 2018 Common Crawl.

Topics

7. Data Streams Clustering (Basil Sattler, **Anna)**

- Guha, S. et al.: Clustering data streams: theory and practice. IEEE Transactions on Knowledge and Data Engineering. 15, 3, 515–528 (2003).
- Ahn, K. et al.: Correlation clustering in data streams. In: International Conference on Machine Learning. pp. 2237–2246 (2015).
- Ntoutsi, I. et al.: Density-based projected clustering over high dimensional data streams. In: Proceedings of the 2012 SIAM international conference on data mining. pp. 987–998 SIAM (2012).

8. Transfer Learning of Matching Knowledge (Sebastian Ziegler, **Anna)**

- S. N. Negahban, Scaling multiple-source entity resolution using statistically efficient transfer learning. In 21st international conference on Information and knowledge management (2012)
- Thirumuruganathan, Saravanan, Shameem A. Puthiya Parambath, Mourad Ouzzani, Nan Tang, and Shafiq Joty: Reuse and Adaptation for Entity Resolution through Transfer Learning (2018)

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

Topics

9. Blocking for Large-scale N:M Entity Matching (Jonas Kändler, **Yaser)**

- Christophides, Vassilis et al: „Entity resolution in the web of data.“ Synthesis Lectures on the Semantic Web 5.3 (2015): 55-72.
- G. Papadakis et al: Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data. In Proc. of the 5th Int. Conf. on Web search and data mining, ACM, 2012..

10. Truth Discovery on the Web (Ayyasamy Shangeetha, **Yaser)**

- Zhao, Bo, et al. „A bayesian approach to discovering truth from conflicting sources for data integration.“ Proceedings of the VLDB Endowment 5.6 (2012): 550-561.
- Li, Qi, et al. „Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation.“ Proceedings of the 2014 ACM SIGMOD international conerence on Management of data. ACM, 2014..

11. Best Practices for Building Training and Evaluation Sets for Entity Matching (Sarah Fathallah, **Yaser)**

- Köpcke, H., Rahm, E.: Training selection for tuning entity matching. In: QDB/MUD. pp. 3–12 (2008)
- Bianco, G.D. et al.: A Practical and Effective Sampling Selection Strategy for Large Scale Deduplication. IEEE Transactions on Knowledge and Data Engineering. 27, 9, 2305–2319 (2015)

Supervisor e-mail: yaser@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 who 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?