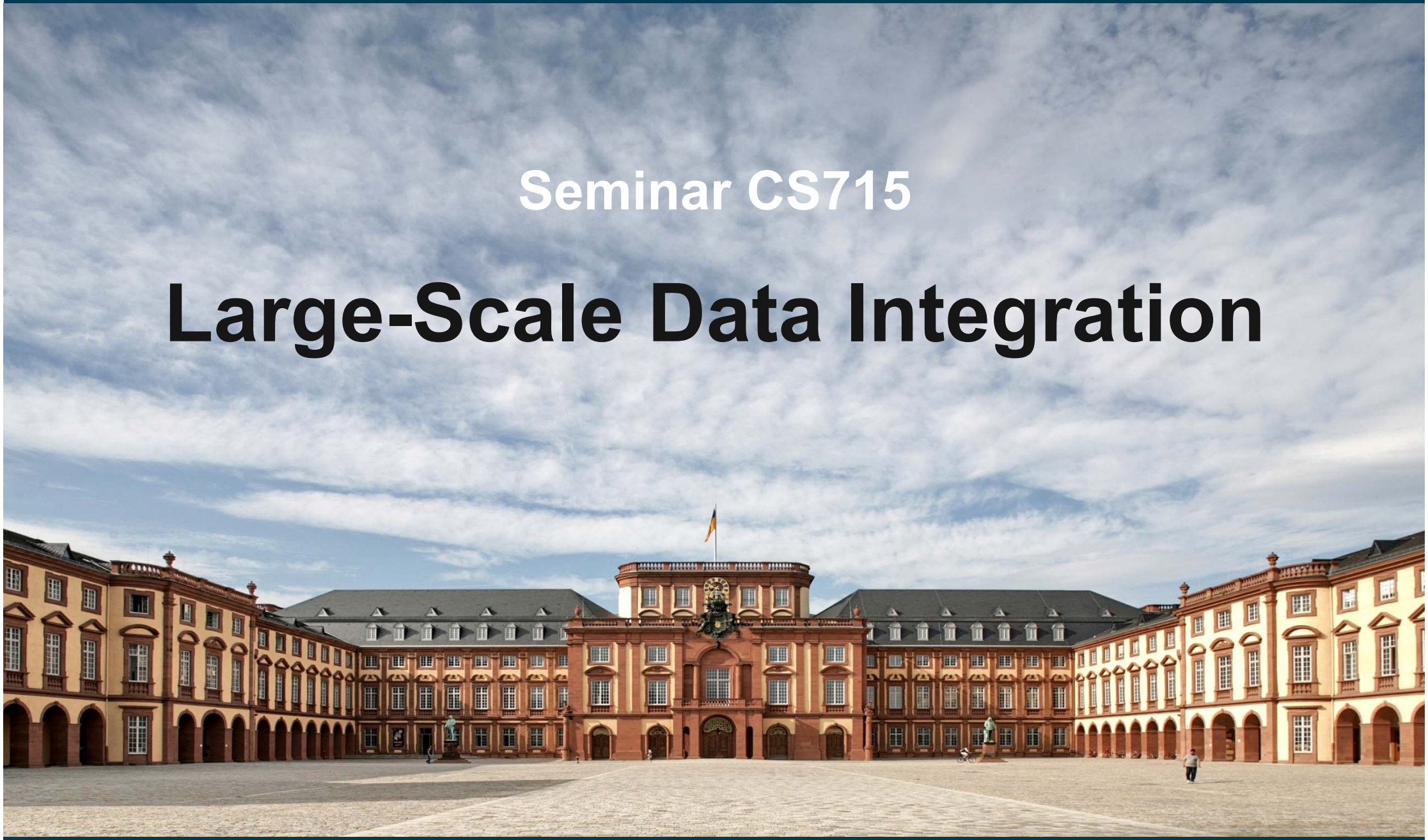
UNIVERSITY OF MANNHEIM

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

Hallo

- **M. Sc. Wi-Inf. Alexander Brinkmann**
- Graduate Research Associate
- Research Interests:
  - Data Search using Deep Learning
  - Product Data Categorization
- Room: B6, 26, C 1.03
- eMail: alex.brinkmann@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
Sunday, 7.03.2021	Send list of preferred topics via eMail
Wednesday, 17.03.2021, 11:00	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 Friday 23.04.2021	Meet with your mentor to discuss your presentation
	Prepare draft of your presentation
Until Sun. 17.05.2021	Send draft presentation to your mentor
	Finalize your presentation
Friday, 28.05.2021 (10:00-12:30)	Presentation and discussion of your topic (30 % of your final grade)
	Write seminar thesis
Sunday, 28.06.2021	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.

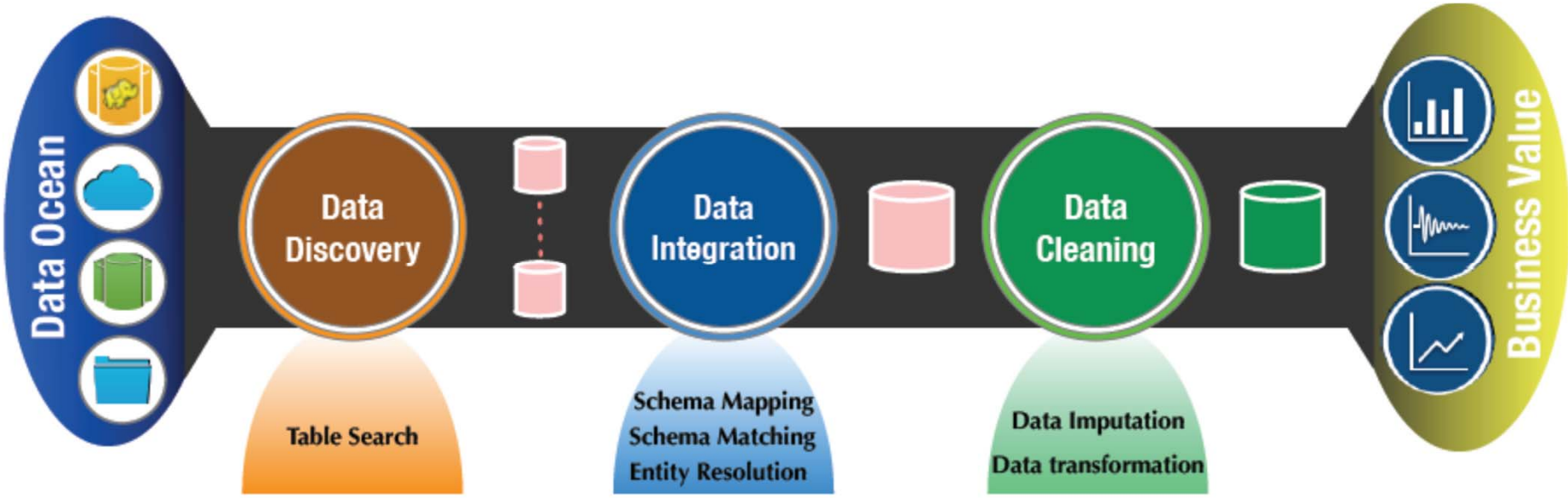
1 Introduction

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 and Overview



V|p erdf#Uhsuhvhqwdwlrqv# → G lvwulexwhg#Uhsuhvhqwdwlrqv

Wudglwlrqdo#P dfklqh#Ohduqlqj# → Q hxudo#Q hwv

Wkluxp xuxjdqdwkdq/#Wdqj/#Rx}}dql/#Grdq=##Gdwd#Fxudwlrq z lwk Ghhs#Ohduqlqj1#HGEW#53531

Topics

1. Table Search using Deep Learning (Giang, mentor: Alex)

- S. Zhang and K. Balog, “Web Table Extraction, Retrieval and Augmentation: A Survey.“ in Proceedings of the 11th ACM Transactions on Intelligent Systems and Technology, pages 1-35. ACM, 2020
- M. Trabelsi, Z. Chen, B. D. Davison and J. Heflin, “A Hybrid Deep Model for Learning to Rank Data Tables“ in Proceedings of the IEEE Internationl Conference on Big Data, 2020
- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: Table Understanding through Representation Learning,” arXiv:2006.14806 [cs], Jun. 2020.

Topics

2. Entity Matching using Deep Learning (Keti, mentor: Ralph)

- Y. Li, J. Li, Y. Suhara, A. Doan, and W.-C. Tan, “Deep entity matching with pre-trained language models,” in Proceedings of the VLDB Endowment, vol. 14, Sep. 2020.
- J. Shao, Q. Wang, A. Wijesinghe, and E. Rahm, “ErGAN: Generative Adversarial Networks for Entity Resolution,” arXiv:2012.10004 [cs], Dec. 2020.
- Z. Wang, B. Sisman, H. Wei, X. L. Dong, and S. Ji, “CorDEL: A Contrastive Deep Learning Approach for Entity Linkage,” arXiv:2009.07203 [cs], Sep. 2020.

Topics

3. Schema Matching using Deep Learning (Xuehui, mentor: Ralph)

- J. Chen, E. Jimenez-Ruiz, I. Horrocks, and C. Sutton, “ColNet: Embedding the Semantics of Web Tables for Column Type Prediction,” arXiv:1811.01304 [cs], Nov. 2018.
- M. Hulsebos et al., “Sherlock: A Deep Learning Approach to Semantic Data Type Detection,” arXiv:1905.10688 [cs, stat], May 2019.
- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: Table Understanding through Representation Learning,” arXiv:2006.14806 [cs], Jun. 2020.

Topics

4. Data Imputation using Deep Learning (Yifan, mentor: Ralph)

- R. Wu, A. Zhang, I. F. Ilyas, and T. Rekatsinas, “Attention-based Learning for Missing Data Imputation in HoloClean,” in Proceedings of the 3rd MLSys Conference, 2020.
- J. Yoon, J. Jordon, and M. van der Schaar, “GAIN: Missing Data Imputation using Generative Adversarial Nets,” arXiv:1806.02920 [cs, stat], Jun. 2018.
- X. Deng, H. Sun, A. Lees, Y. Wu, and C. Yu, “TURL: Table Understanding through Representation Learning,” arXiv:2006.14806 [cs], Jun. 2020.

Topics

5. Deep Active Learning (Andreas, mentor: Anna)

- J. Kasai, K. Qian, S. Gurajada, Y. Li, and L. Popa, “Low-resource Deep Entity Resolution with Transfer and Active Learning,” arXiv:1906.08042, 2019.
- Y. Nafa et al., “Active Deep Learning on Entity Resolution by Risk Sampling,” arXiv:2012.12960, 2020.
- Siméoni, Oriane, et al. „Rethinking deep active learning: Using unlabeled data at model training.“ arXiv preprint arXiv:1911.08177 (2019).

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
   - 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/
  - 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?