

How to read a research paper?

Information Systems Group

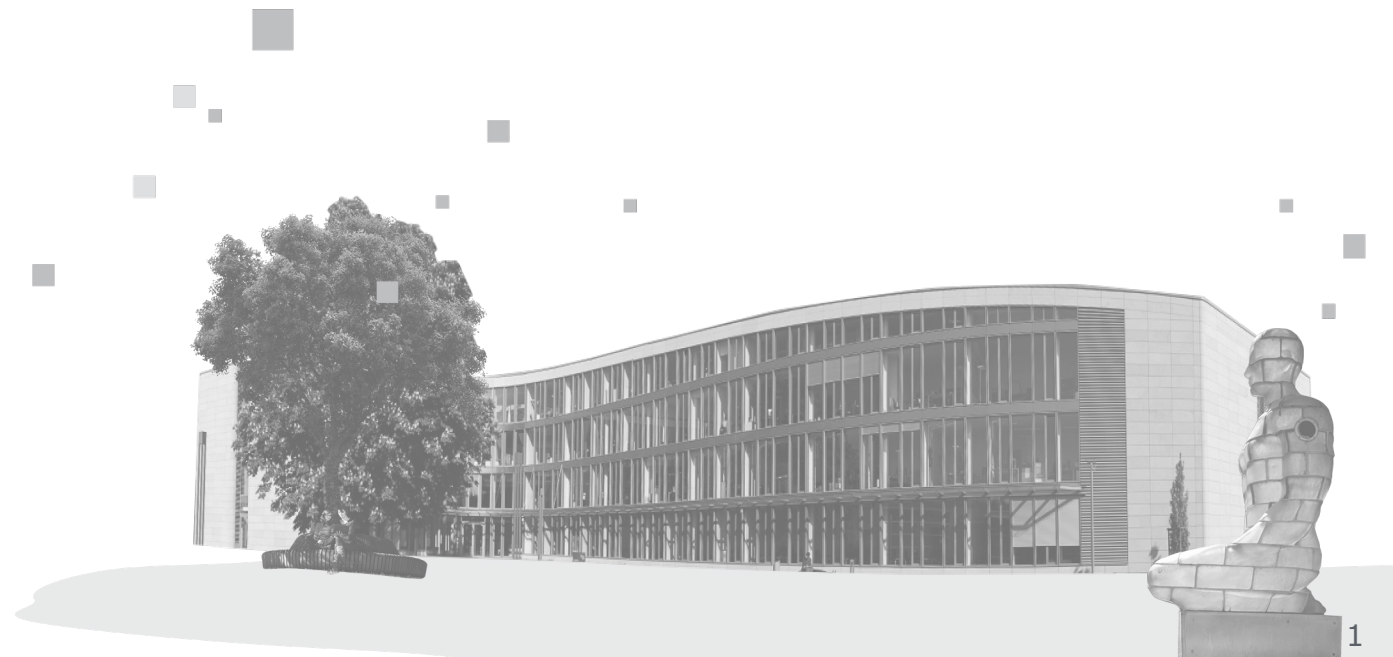
Common session of the master seminars 2024:

- DQ4AI: Data Quality Assessment
Lisa Ehrlinger, Sedir Mohammed
- Table Representation Learning
**Christoph Hönes, Lukas Laskowski,
Francesco Pugnaroni**

**Design IT.
Create Knowledge.**

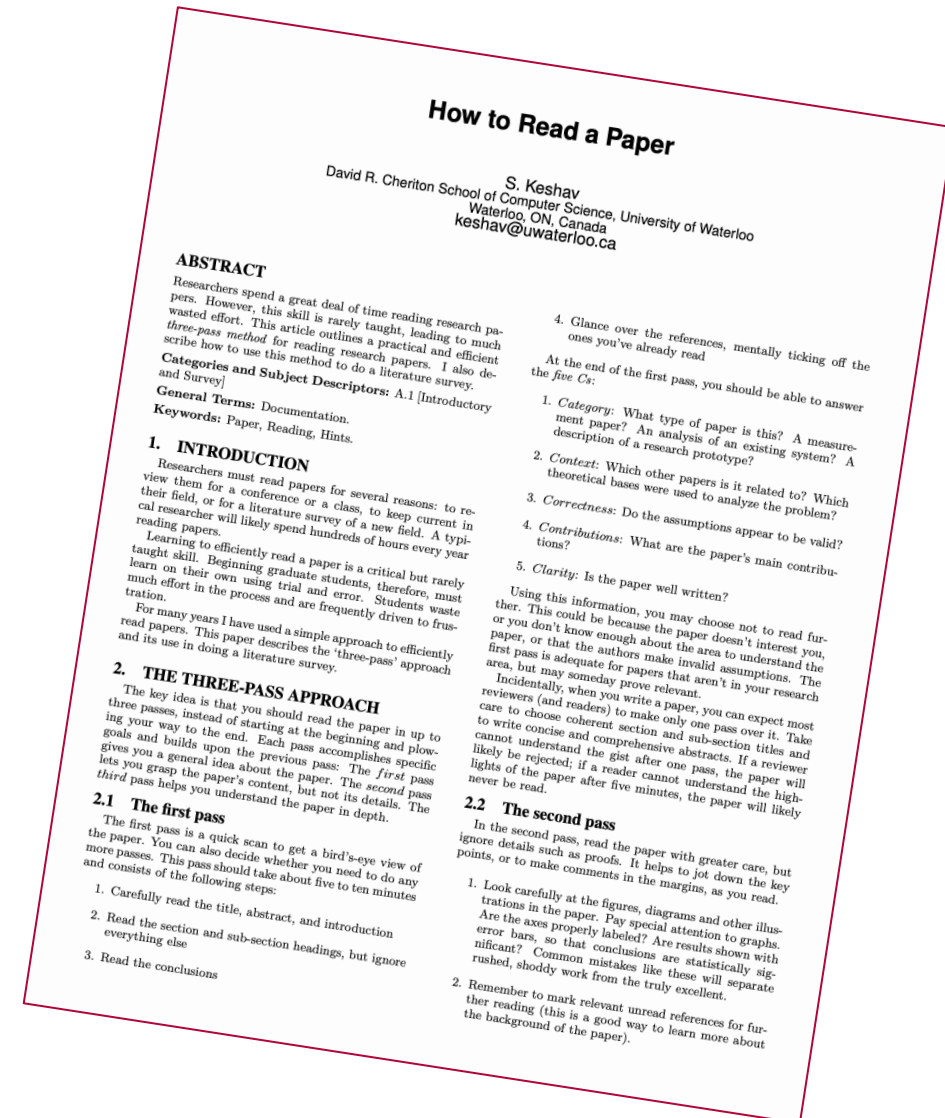
www.hpi.de

How to read a research paper?



Agenda

- Code of conduct
- Publication process and paper types
- The three pass approach on how to read a paper



Code of Conduct - Overview

At DEF/HPI, we are committed to providing a high-quality learning as well as research environment and building a community where students and staff can thrive scientifically and personally. Everyone should expect a safe, supportive, and inclusive environment in all our spaces.

Our Code of Conduct helps us meet this goal. Words or actions that are disrespectful, racist, discriminatory, hostile, or harassing are not acceptable.

Examples of these include:

- Offensive comments about others' ethnicity, accent, religion, nationality, gender, sexual orientation, or other personal traits
- Refusing to work with someone based on these personal traits
- Physical or verbal threats and assaults
- Using sexualized or vulgar language or actions
- Disrupting another person's work experience

Code of Conduct - Help and Support

Violations of this code are taken seriously. If you witness or experience any inappropriate behavior, report it to a lecturer or any DEF/HPI contact point. All reports will be handled confidentially and with care.

Please be aware of **further contact points and support** structures at DEF/HPI, including:

- Equal Opportunities Officers (Charlotte Weiss, Florence Böttger, Oliwia Gust)
- Diversity Manager (Oliwia Gust)
- Ombudsman for good scientific practice (Prof. Tilmann Rabl)
- Student Trusted Advisors (Hanna, Joscha and Zero)
- Psychological counseling hotline (0800 7777015)
- Incident Response System (www.safecampus.hpi.de)
- as well as the respective offers of the University of Potsdam (Mental Health Counseling Service, Psychosocial Counseling of Studentenwerk, Nightline)
- www.uni-potsdam.de/en/discrimination-free-university/consulting-and-support/overview-of-counseling-and-advising-services

Publications in the database community

- Publications (papers) should clearly describe scientific findings
- In computer science (specifically the database community) the following holds:
 - **Conferences** are primarily important
 - **Journals** of secondary importance
- Conference papers can come in different shapes (see following slides)
 - About 2 month of reviewing
 - About 15%-25% acceptance rate for good conferences
 - In case of an accept: travel to conference and presentation of the paper
- Workshop papers are similar to conference papers with a shorter reviewing period
- Journal papers have an iterative reviewing period
 - Multiple feedback rounds
 - Up to 3 years until paper is published

Types of papers

- Journal article (10–50 pages)
 - Often at the end of a project or thesis
- Article in proceedings (2–10 pages)
 - Often about non-finished work or project overviews
- Conference paper (6–12 pages)
 - Research results
- Demo (at conference) (2–4 pages)
 - Description of a prototype demo
- Poster (at conference) (3–5 pages)
 - „small paper“
- Workshop paper (6–12 pages)
 - Preliminary research results
- Technical report (10–30 pages)
 - Long version of a conference paper
 - Usually published by an institution
- Preprints (no limit)
 - No peer-reviewing process
 - Early dissemination of research results
 - Typically published on <https://arxiv.org>

Important database conferences

- Very Large Databases (VLDB)
 - <https://vldb.org>
- ACM Special Interest Group – Management of Data (SIGMOD)
 - <https://sigmod.org>
 - GI – Gesellschaft für Informatik
- IEEE International Conference on Database Engineering (ICDE)
- International Conference on Extending Database Technology (EDBT)
- Conference on Database Systems for Business, Technology and Web (BTW)
- And many more smaller conferences and specialized workshops exist ...

Most common publishers

- Publishing communities oversee the publication process of research papers
- Publisher websites are a good source to find research papers

- ACM (Association for Computing Machinery) Digital Library: <https://dl.acm.org>
- SpringerLink: <https://link.springer.com>
- IEEE (Institute of Electrical and Electronics Engineers): <https://ieeexplore.ieee.org/>
- Elsevier: <https://www.elsevier.com>

Typical publication workflow

- Idea(s)
- Implementation and experiments (or proofs)
- Writing the article
- Submit in time (typically through an online submission system like CMT)
- Reviewing process
 - Typically done by 2-4 (anonymous) experts in the field
 - Sometimes double-blind
 - Sometimes with rebuttal
- Final decision about accept/reject made by the program chair
 - Notification to authors typically per email
- In case of an accept: prepare camera-ready copy of paper
- Register for conference and organize conference travel
- Present research paper at the conference

Conference papers

- Typical length of 12 pages
- English language
- Research papers on conferences can come in different shapes
 - Regular research
 - Vision
 - Industry
 - Experiment, Analysis & Benchmark
→ incl. Reproducibility

Four Paper Categories

There are four equally important categories of papers in the research track:

Regular Research Papers

Scalable Data Science Papers (SDS)

Experiment, Analysis & Benchmark Papers (EA&B)

Vision Papers

VLDB Call for Papers

Typical paper structure

- Abstract
- Introduction
- Related Work
- Notation, Definitions, Architecture
- Main Idea(s)

- (Extensions, pot. after Exp.)
- Experiments
- (Related Work, if not after Intro)
- Conclusion and Outlook
- (Acknowledgements)
- References

Example:

Papenbrock and Naumann 2016: A Hybrid Approach to Functional Dependency Discovery

Abstract

1. Functional Dependencies
 2. Related Work
 3. Hybrid FD Discovery
 4. The HyFD Algorithm
 5. Preprocessing
 6. Sampling
 7. Induction
 8. Validation
 9. Memory Guardian
 10. Evaluation
 11. Conclusion & Future Work
- Acknowledgements
1. References

Plagiarism: considerations when writing a paper

- ACM definition: “**misrepresentation of another's writings or other creative work** (including unpublished and published documents, data, research proposals, computer code, or other forms of creative expression, including electronic versions) **as one's own**”
- See details on ACM website: <https://www.acm.org/publications/policies/plagiarism-overview>
- Different types of plagiarism (not always considered plagiarism):
 - verbatim copying, near-verbatim copying, or intentionally paraphrasing portions of another's work;
 - *using automated tools that rephrase existing work as one's text without proper attribution;*
 - copying elements of another's work, such as equations, tables, charts, illustrations, presentation, or photographs that are not common knowledge, or copying or intentionally paraphrasing sentences without proper or complete source citation;
 - verbatim copying of portions of another's work with incorrect source citation
 - Self-plagiarism: verbatim or near-verbatim reuse of significant portions of one's own published work without citing the original source.

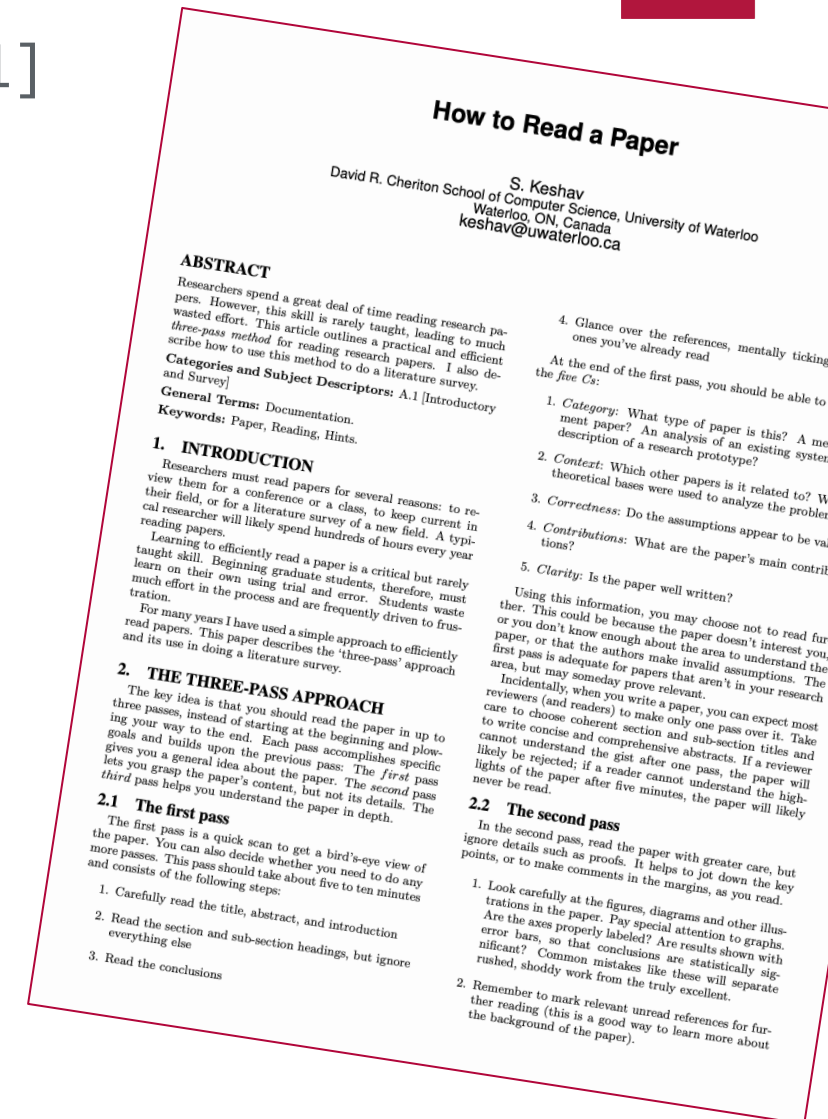
The three-pass approach on how to read a paper [1]

The key idea is reading a paper in three passes instead of once from the beginning to the end, each pass accomplish a specific goal:

1. First pass: gives a general idea about the paper
2. Second pass: gives an idea about the content but overlooks the details
3. Third pass: gives a deep understanding of the paper

Why?

- Depending on the reason why you are reading a paper, you need different levels of understanding
- This approach is flexible and can be used for all the categories of papers



[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

First pass (about 5-10min)

Steps:

- Carefully read title, abstract, and introduction
- Read headings of sections, sub-sections, and paragraphs, ignoring the actual content
- Read the conclusions
- Glance the references to look for known work

Expected learnings (the five Cs):

- **Category** (Which is the category of the paper?)
- **Context** (Which are the related works? Which are the theoretical bases?)
- **Correctness** (Are the assumptions valid?)
- **Contributions** (Which are the main contributions?)
- **Clarity** (Is the paper well written?)

[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

First pass – What to do after?

Do you want to continue with the second pass?

- The paper may be not interesting
- The paper may be out of your knowledge area, but may still be useful for future reading
- The assumptions made by the authors may be invalid

As a side note: when you write a paper expect that most readers and reviewers may only make one pass over it, hence

- Take care of choosing coherent titles for sections and sub-sections
- Write concise and comprehensive abstract and conclusion

[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

Second pass (up to an hour)

Read the whole paper ignoring details, e.g., proofs

- Add comments and mark unread relevant references for future reading
- Focus on figures, diagrams, and illustration:
- Are axes labeled properly?
- Are results clear?

Expected learnings:

- You should be able to summarize the main thrust of the paper, providing evidence

[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

Second pass – What to do after?

Do you want to continue with the third phase?

- You may be interested in the paper, but it is not in your research area and decide to stop
- The paper may be poorly written → set the paper aside
- The paper may require knowledge that you do not have, in that case evaluate if studying the necessary literature is worth the effort

[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

Third pass (4/5 hours)

Read the whole paper in detail

Virtually re-implement the paper:

- Make the same assumptions as the authors
- Re-create (virtually) the work yourself by thinking about how you would have done it
- Compare your ideas with the actual paper to identify:
 - Paper's innovations
 - Hidden failings
 - Hidden assumptions
 - Ideas for future work

[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

Third pass – Expected learnings

You are expected to be able to reconstruct the structure of the paper by memory

You are expected to be able to identify strong and weak points:

- Implicit assumptions
- Missing citations to relevant work
- Potential issues in the evaluation

[1] Srinivasan Keshav. 2007. How to read a paper. *Comput. Commun. Rev.* 37, 3 (2007), 83–84.

Bonus – Doing a literature survey

Definition: Paper that summarizes and organizes recent research results in a novel way that integrates and adds understanding to work in the field”

Challenge: you need to read tens of papers in an unfamiliar field


Approach:

- Identify a set of *keywords* that describe what you are looking for
- Use academic search engine, e.g., Google Scholar, to find 3-5 recent papers in the area
- For each paper:
 - Do one pass
 - Read related work to find relevant papers in the field
- If you found a survey, read it and decide if you want to go further
- If you did not, identify sources and author names that appear in the bibliographies
- Put the relevant papers aside for now
- Visit the websites of the authors and check the conferences where they published recently to identify the top conferences in the field
- Go through the recent proceedings of the conferences to identify recent high-quality work
- Make two passes on these papers and the ones that you previously put aside
- If they all cite a paper that you did not find earlier, read it, otherwise you have the first version of your survey

Search for literature / related work

- Backward search
 - Search for referenced papers
 - Search for longer version of the same work
 - Search for earlier version of the same work
- Forward search
 - Search for articles which cite the current one (starting paper)
 - Search for articles of the same author (e.g., journal version of the work)
 - Search for articles about the same topic but from different authors
 - Search for surveys (overview articles) about a topic
- Popular search engines
 - DBLP: <https://dblp.org/db/conf/dbpl/index.html>
 - Google Scholar: <http://scholar.google.com>

Google Scholar

Questions?

**Design IT.
Create Knowledge.**

www.hpi.de

