

Seminar

Data Stream Management and Analysis

Kickoff Meeting 2024-03-26

PROF. DR. SANDRA GEISLER, PROF. DR. CHRISTOPH QUIX,
ANASTASIIA BELOVA, SOO-YON KIM, LIAM TIRPITZ

The Seminar Team



Prof. Dr. Sandra Geisler
Junior Professorship DSMA
geisler@dsma.rwth-aachen.de



Prof. Dr. Christoph Quix
Christoph.quix@fit.fraunhofer.de



Anastasiia Belova, M. Sc.
belova@dbis.rwth-aachen.de



Soo-Yon Kim, M. Sc.
kim@dbis.rwth-aachen.de



Liam Tirpitz, M. Sc.
tirpitz@dbis.rwth-aachen.de

Agenda



Dive into the research field



Seminar Topics



Match making



Organization of the seminar



What you will learn

Data Stream Management and Analysis

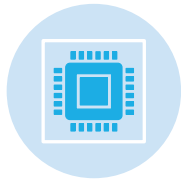
Data-driven Environments



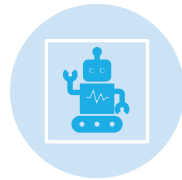
Sensors, mobile devices, and IoT devices leverage smart & connected environments



Devices produce high volumes of data as streams rapidly



Industry and consumers demand exploiting the data to realize low latency and high-level applications



Big data technologies and AI can empower data-driven processes & organizations



Source: <https://www.pcwelt.de>

But: challenges get bigger due to increasing data volumes and complexity

Social Networks



Users share more than **95 million photos and videos** on Instagram each day.

A total of **4.2 billion likes** are registered on Instagram each day.

There are about **350,000 tweets/posts** sent every minute on X

<https://www.meltwater.com/en/blog/instagram-statistics>
<https://famewall.io/statistics/twitter-stats/>

What are Data Streams?

Synonyms: Data in motion, Online data, (near) real-time data, streaming data

Some first informal definitions:

“[..] **time-varying, volatile, unpredicted** and possibly **unbounded** information [..]”
[Patroumpas & Sellis, 2006]

“[..] a data set that is produced **incrementally** over time, rather than being available in full before its processing begins.”
[Golab & Özsu, 2003]

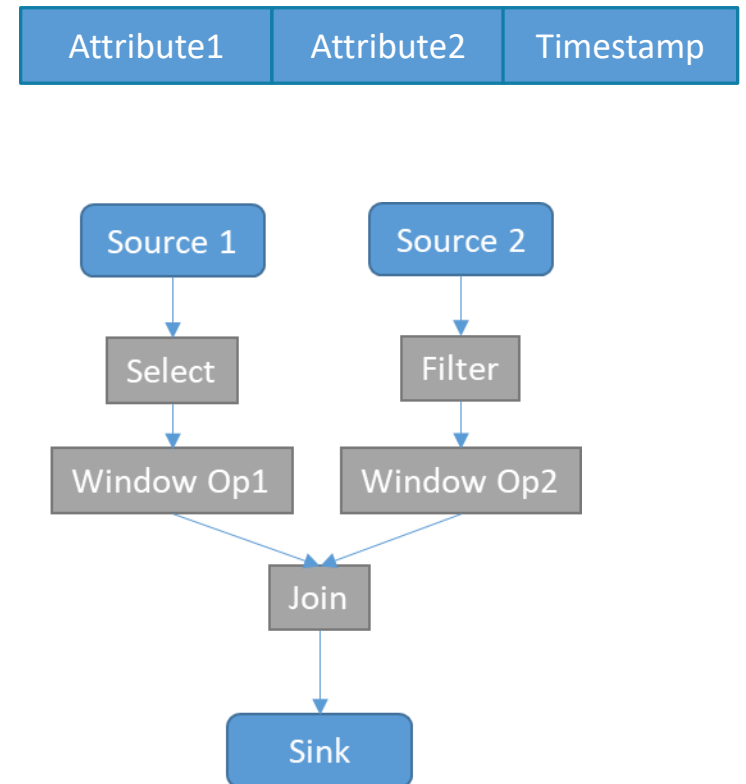
“[..] **continuous** data flows with unknown size and end [..]”
[Geisler, 2010]

Data Stream – Basic Concepts

- Data Stream S : Unbounded multiset of data stream elements (s, τ)
[Geisler, 2016]

$$S(\tau_i) = \{ \langle (s_0, \tau_0), m_0 \rangle, \dots, \langle (s_i, \tau_i), m_i \rangle \}$$

- Continuous queries, one-pass processing
[Golab and Öszu, 2010; Babu and Widom, 2001]
- DBMS-active human-passive [Stonebraker et al., 2005]
- Query operators:
Projection, Filtering, Union, Join, Group, Apply, **Windows**
- Various data models, query languages, algorithms
[Babcock et al. 2002; Aggarwal, 2013; Geisler, 2013]





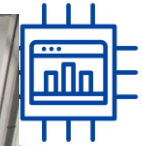
Smart Equipment
Maintenance

Self-
regulating



Product Quality
Control

Self-
executing

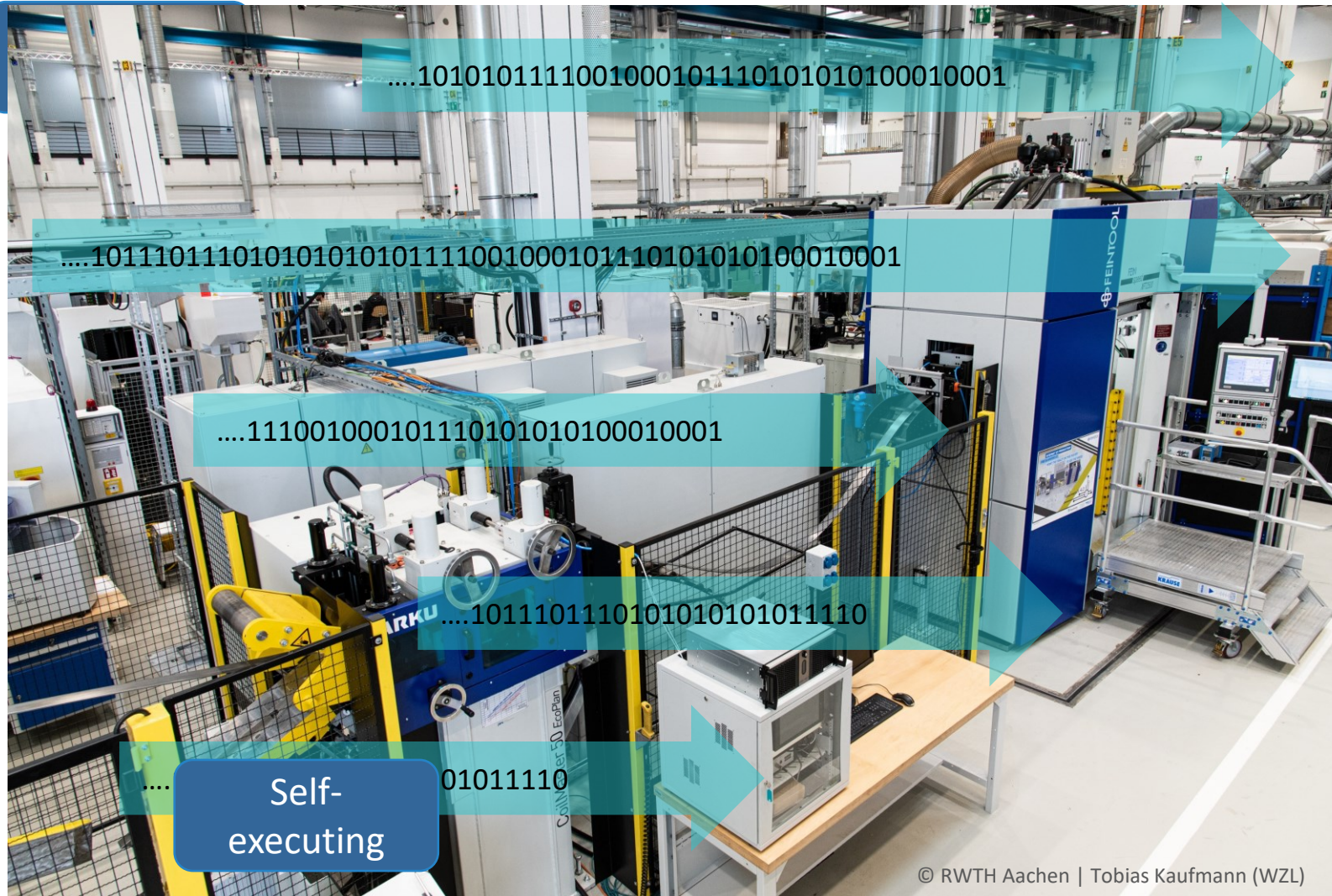


Process
Monitoring

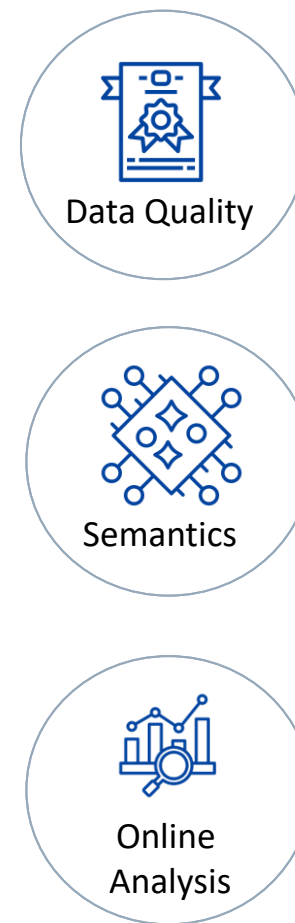
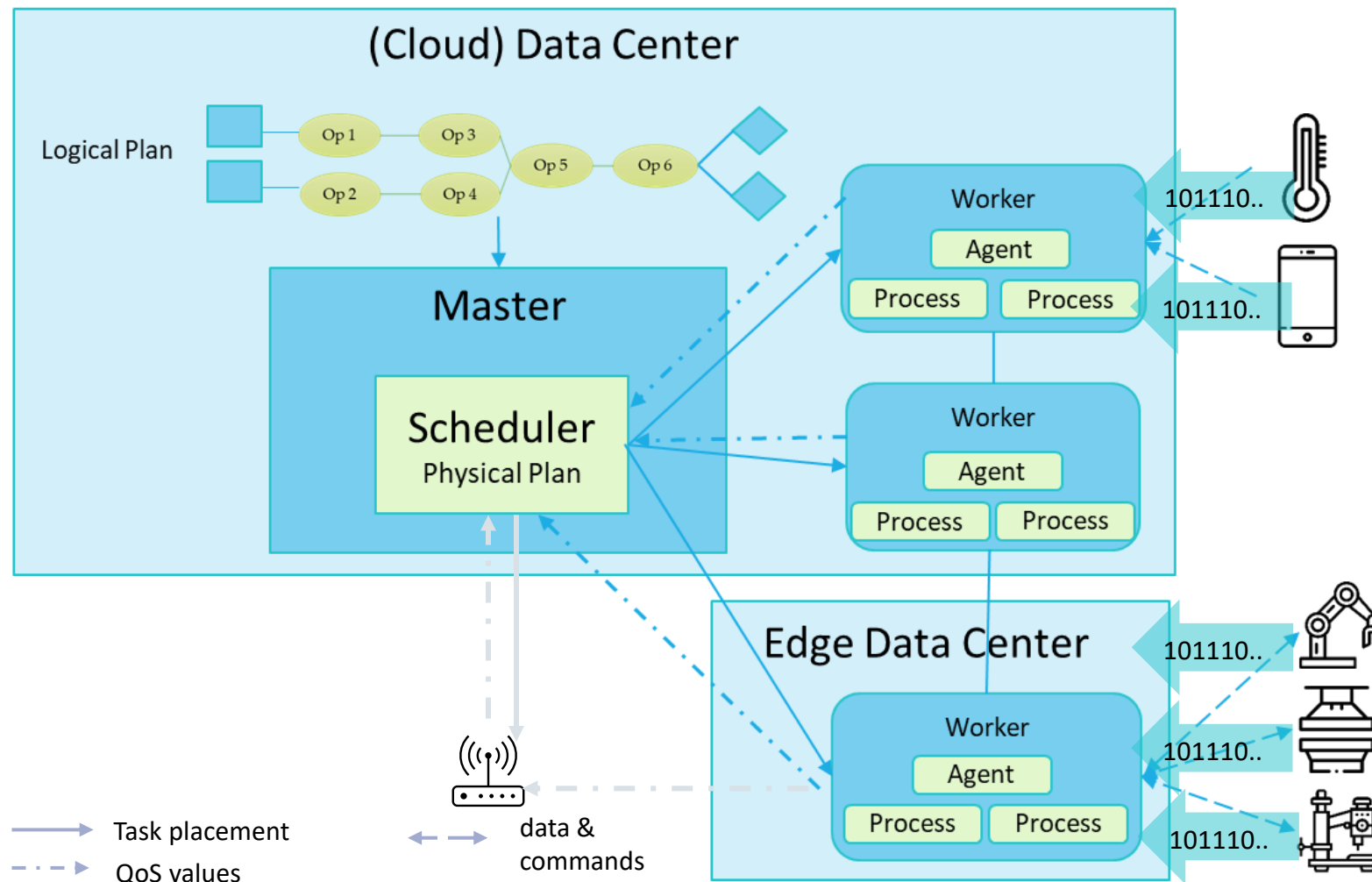
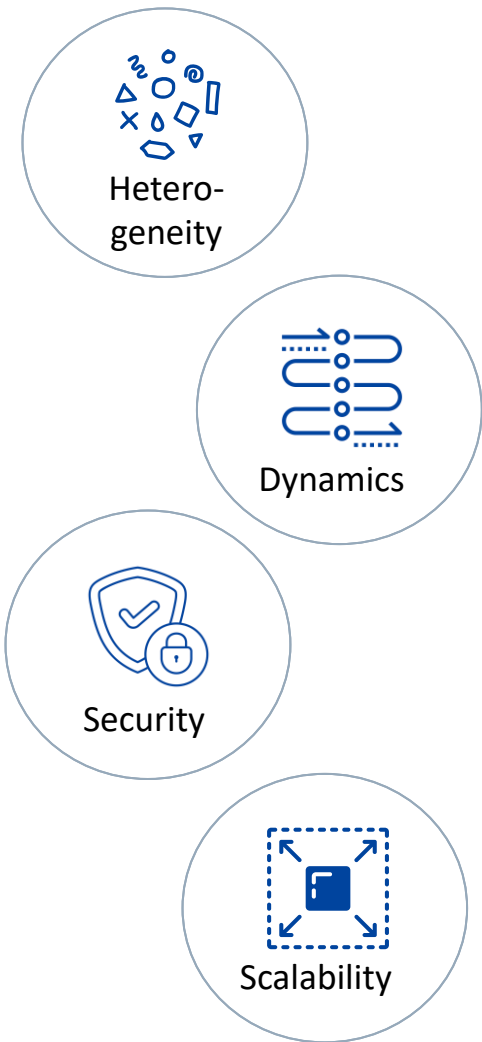
Self-
organizing



Material
Distribution &
Tracking



© RWTH Aachen | Tobias Kaufmann (WZL)



Seminar Topics



Topics

- (1) Cross-platform Data Processing (Tirpitz)
- (2) Data Stream Processing on Programmable Network Devices (Tirpitz)
- (3) Data Mesh and its Application in Manufacturing (Belova)
- (4) Self-adaptive Data Stream Processing (Geisler)
- (5) FAIRification of Data Streams (Geisler)
- (6) Neural Networks for Data Streams (Kim)
- (7) Large Language Models and Data Streams (Kim)
- (8) Automating Data Science (Quix)

Match Making

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Preferences														
2		1	2	3	4	5	6	7	8	9	10	11	12		
3	1													0	
4	2			1				2			3			6	
5	3						2					3	1	6	
6	4				1			2					3	0	
7	5													0	
8	6													0	
9	7													0	
10	8	1							2				3	0	
11	9													0	
12	10													0	
13	11			3	1								2	6	
14	12													0	
15		0	0	4	1	0	2	2	0	0	3	3	3		
16															

Seminar Organization

Schedule

What	Submission	When
Kickoff		26.03.2024
Resignation Deadline		16.04.2024
Abstract & outline	4 pages	23.04.2024
Draft version, teaser	12-15 pages	21.05.2024
Complete draft	17-20 pages	18.06.2024
Slides	PDF	25.06.2024
Final version	17-20 pages	02.07.2024
Review	0.5 - 1 page	09.07.2024
Final presentations	30 min. + discussion	16.07.2024

Important: Page limits and dates are binding!

Grades



Paper (70%)

- Introduction
- Technically sound
- Complete
- Readability and structure
- References
- Independence of work



Presentation (30%)

- Slides
- Talk
- Structure
- Questions




Registration

- Formal, binding registration by **April 16th, 2024**
 - If you do not resign from the seminar until this date, any later resignation will be considered as a **failure** (i.e., grade = 5.0)
 - If you cannot continue with the seminar for whatever reason, then **please tell us!**

Data Stream Management and Analysis (SE) [24.12.00031]

[Kurs](#) [Einstellungen](#) [Teilnehmer*innen](#) [Bewertungen](#) [Berichte](#) [Mehr ▾](#)

▾ Allgemeines

 [Ankündigungen](#)

Als erledigt

Welcome to the Seminar "Data Stream Management and Analysis"

In this course you will learn how to

- search for and read relevant literature for a certain topic,
- structure your research,
- write a scientific paper about your topic,
- present and defend your topic,
- read other researchers' works and write a review about it,
- moderate a paper session

Finally, of course you also learn something about the field of data stream management and analysis

We will organize the seminar along the following agenda and deadlines:

Course in RWTHmoodle

New Submission for DSMA'24

Follow the instructions, step by step, and then use the "Submit" button at the bottom of the form. The required fields are marked by *.

Author Information

For each author please fill out the form below. Some items on the form are explained here:

- **Email address** will only be used for communication with the authors. It will not appear in public Web pages of this conference. The email address can be omitted for not corresponding authors. These authors will also have no access to the submission page.
- **Web page** can be used on the conference Web pages, for example, for making the program. It should be a Web page of the author, not the Web page of her or his organization.
- Each author marked as a **corresponding author** will receive email messages from the system about this submission. There must be at least one corresponding author.

Author 1 ([click here to add yourself](#)) ([click here to add an associate](#))

First name*:

Last name*:

Email*:

Country/region*:

Affiliation*:

Web page:

☒ corresponding author

Reviewing in EasyChair



Scientific Writing

Literature - Where to search?

- Literature meta search engines: Google Scholar, Web of Science, Scopus...
- Domain-specific search engines: PubMed, DBLP, ...
- Publishers: ACM, Springer, IEEE, ...
- Conference websites
- Libraries: University, Computer Science Library, ...
- Company websites
- Blogs of reknown researchers and other experts (e.g. CEOs)

→ Depends on what you need: book, journal/conference, white papers, ...



How to search?

- Questions your advisor gave you
- Search terms & questions extracted from the starting papers
- Note down new questions while reading
- Be aware of commonly used terms and synonyms!
- Use refined web searches: “data anonymization“, -health, ...



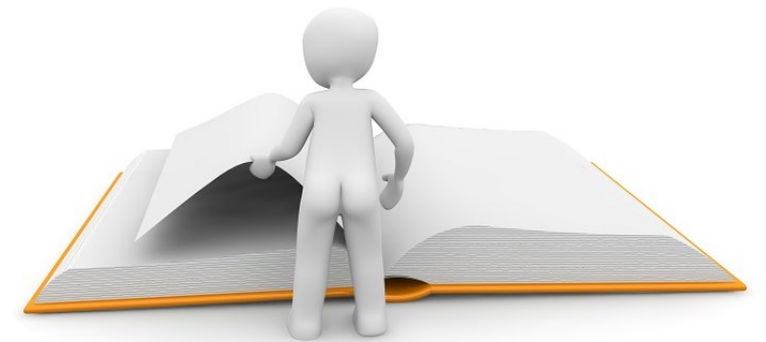
What are hints for a "good" hit?

- Year of the publication
- Authors (Do you know them already from some other publication?)
- Publication type (journal vs. workshop paper)
- Affiliation of the authors
- Publisher of the journal or the conference proceedings
(ACM, IEEE, Elsevier, Springer, ...)
- Number of citations



How to read a paper?

- Type & aim of the paper: Journal, conference, survey paper, workshop paper, vision paper, white paper
 - Title & abstract: Will the paper answer my questions?
 - Conclusion: Are the results convincing? What are open questions?
 - References: What is cited and how?
 - Figures & measurements: Credible setup & results?
- Always be critical and objective (resist the advertisement)!



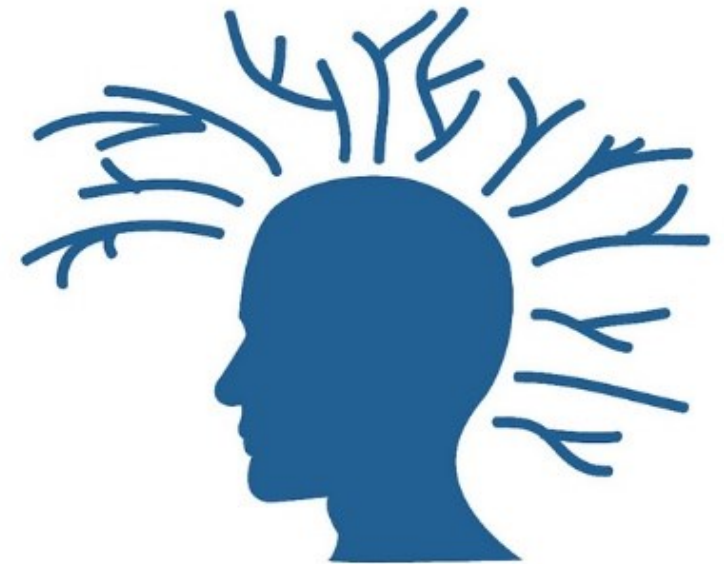
How to find more papers?

- References in the current paper
- Which papers cited the current paper → delivers more current ones
- Some search engines propose related articles, such as IEEE Xplore or ACM
- Identify leading authors and sift through their work
- Use different search engines

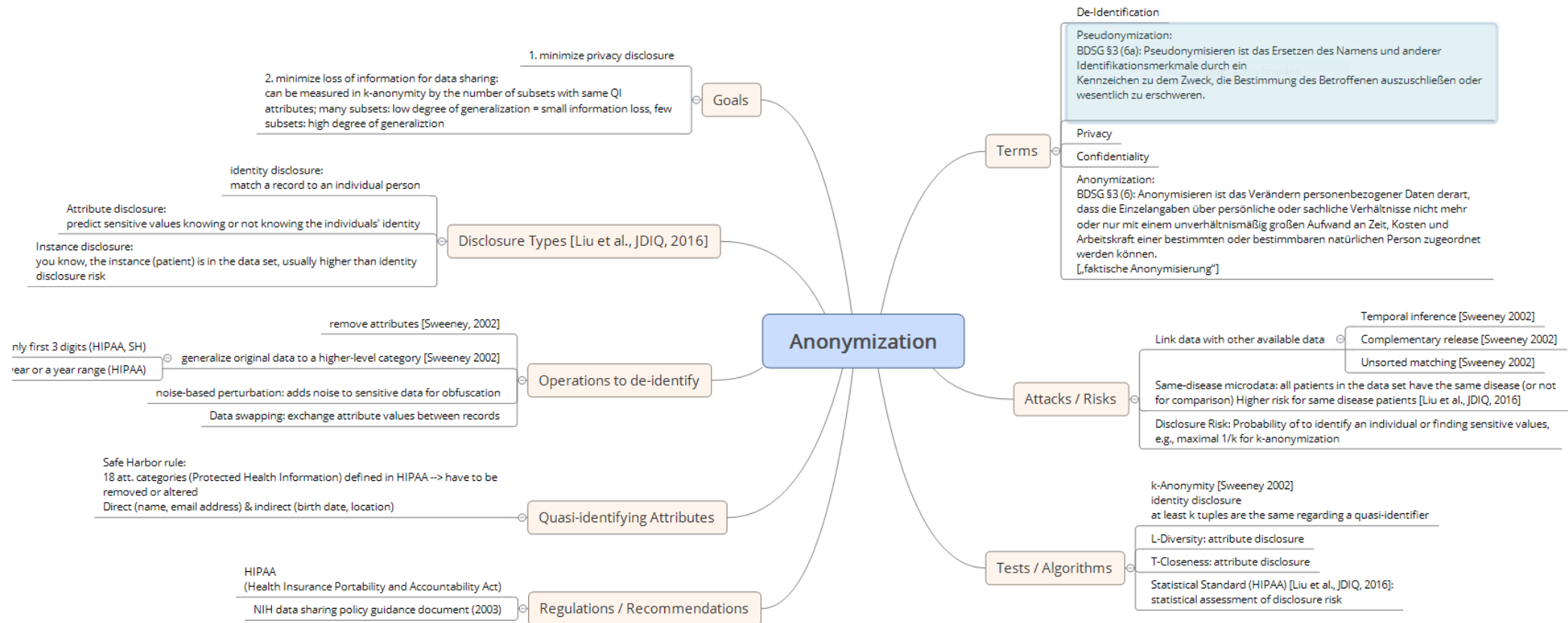


Structuring Knowledge for Your Topic

- Mind mapping, concept mapping, argument mapping:
 - Graphs with objects and relationships among them
 - You can create a map for one text or for one topic
- Common Goals [Davies, 2011]:
 - Understand relationships, remember them, and be able to analyze their component parts → deeper understanding
 - Easier to understand and follow
 - More active way of learning
 - Easier to memorize a diagram than a text



Mind Map - Example



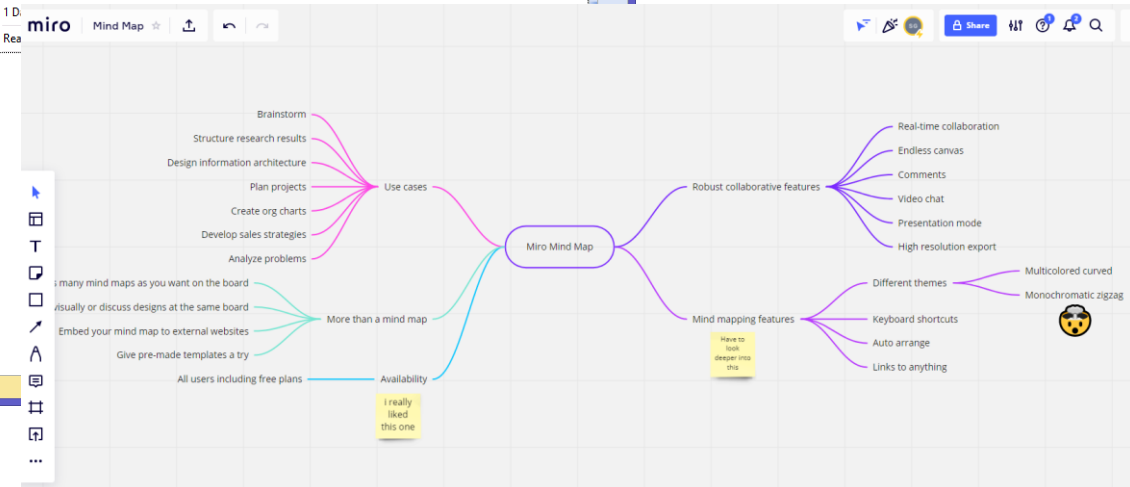
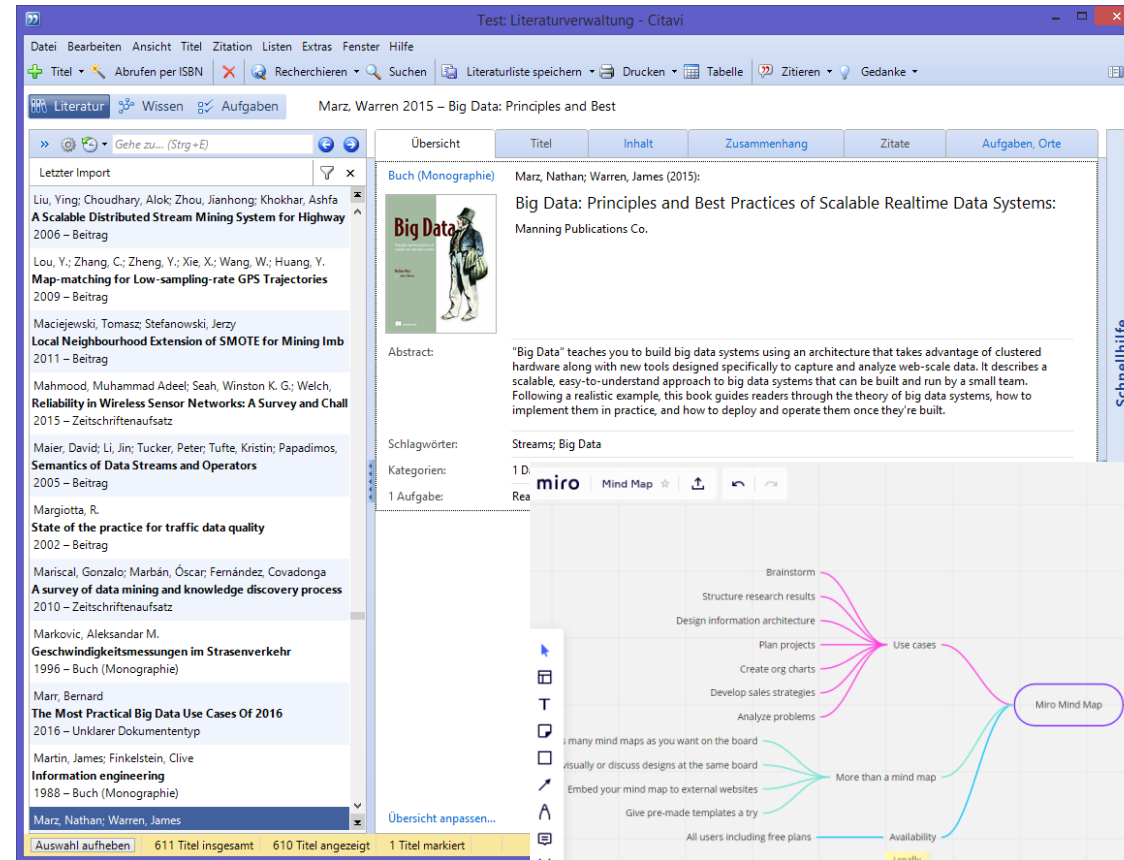
Tools

■ Literature management

- JabRef
- Citavi

■ Mind mapping

- Miro (education license)
- Xmind
- FreeMind
- MindManager
- DocEar
- PowerPoint



Plagiarism

What is plagiarism?

- Drawing any idea or language from another author without crediting them adequately,
- whether it be intentionally or unintentionally.¹

Why you should not do it: Plagiarism is

- cheating
- stealing the fruit of other researchers' hard work by misleading the readers to believe that you created the idea

1: HARVARD COLLEGE WRITING PROGRAM (N.D.). HARVARD GUIDE TO USING SOURCES. [HTTPS://USINGSOURCES.FAS.HARVARD.EDU/WHAT-CONSTITUTES-PLAGIARISM](https://usingsources.fas.harvard.edu/what-constitutes-plagiarism)

In this seminar, we consider as plagiarism

If you copy **word-by-word** from any source. It does not matter whether there is a reference to that source or not.

Word-by-word copies are allowed **only for short quotations** using quotes ("), e.g., 1-2 key statements from a text can be quoted.

Also, **formal definitions** can be copied as-is from the original source.

If you **rephrase the original text only marginally** and keep the same line of argumentation and the same structure as the original text.

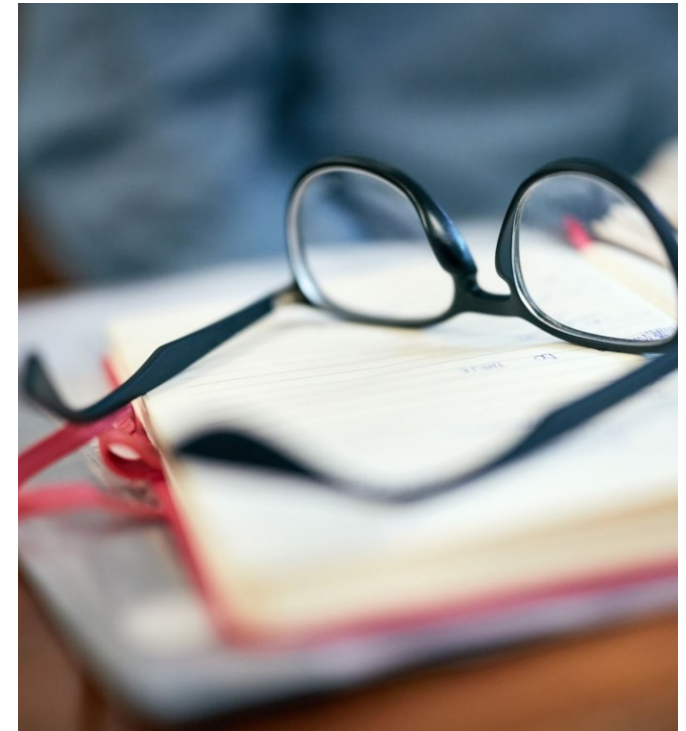
We call this sentence-by-sentence copies or "**structural plagiarism**".

The aim of writing a seminar paper is that you build your own text structure.

If you use an existing **figure without putting a reference** to its source.

Tips to prevent Plagiarism [Booth et al., 2003]

1. Cite every source properly, including every quotation, paraphrase, summary.
2. Even if you cited the source in previous texts, cite it again if the current sentence is a different idea from where you cited for the first time.
3. Use the correct notation for direct quotation: If the quotation is fewer than four typed lines of prose or three lines of verse, it is counted as a short quotation (use quotation marks ""); otherwise, it is a long quotation (use an indented block). [Gibaldi, 2009]
4. As long as the idea is not created in your own head, cite it. Do not take any chances to lead your readers to think that you created the ideas/methods.
5. Rephrase the ideas in the references with your own words.



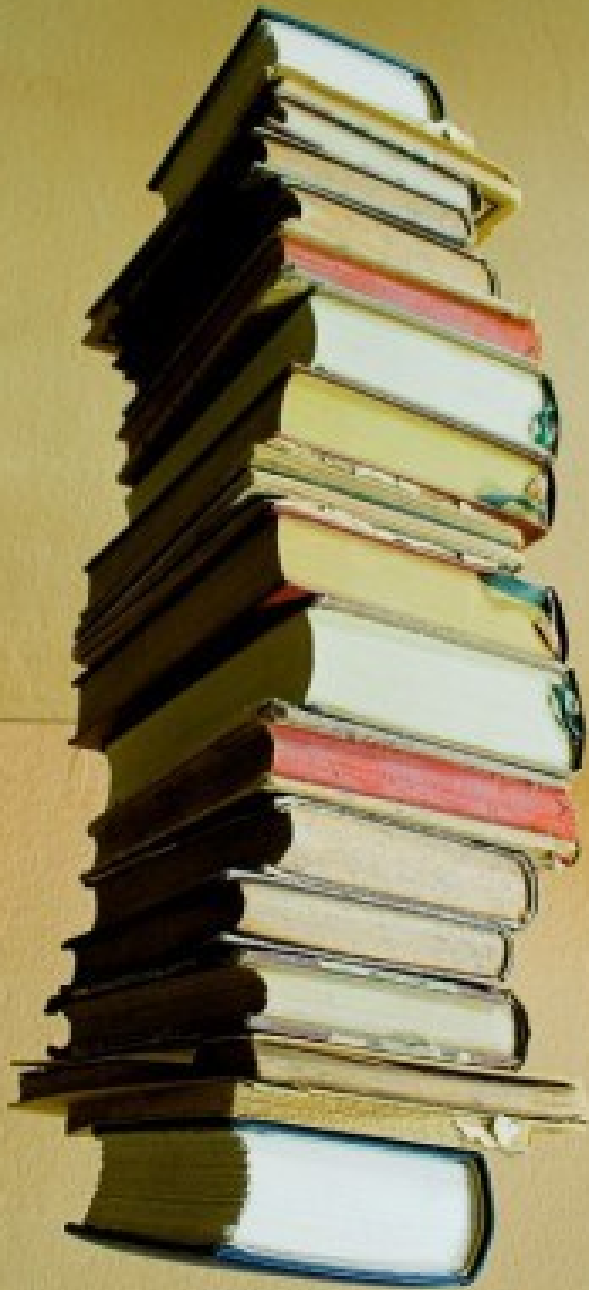


Further Tips [Booth et al., 2003]

How not to paraphrase too closely?

1. Read carefully the texts which you need to summarize.
2. Move your eyes from the texts, organize the idea in your head and write down the idea via your own words.
3. Compare the sentences you wrote in Step 2 with the original texts in Step 1. If both contain similar ideas, and they share the same ordering of synonyms, then go back to Step 1 and restart the process.

Repeat Step 1 – Step 3 until you cannot find the parallel synonyms between your sentences and the original texts, though both may express the same idea.

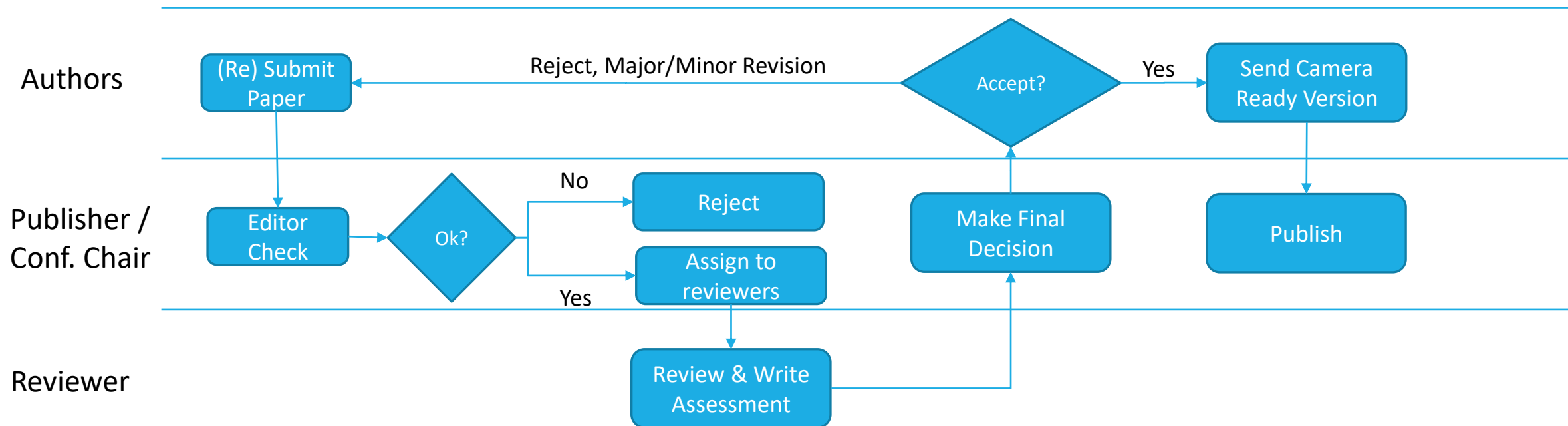


How to cite?

- Only central thoughts.
- Directly: quote with reference
 - Doe states in [14]: “Databases are great”.
- Indirectly:
 - Oblique speech
 - Reference usually after statement
 - Doe states in [14], that databases are great.
 - Databases are great [14].
 - If you refer extensively the thoughts of others, announce this in advance
 - The following statements are derived from [14], where Doe discusses the importance of database systems. Databases are great systems, which

Paper Review

Process to assess the quality of a paper and its contents before publication in an organ



Question	
Overall Rating	Accept
Detailed Comments	<p>This paper describes a framework that (1) uses a simulator to generate moving cars; (2) map matches to find the road section associated with the moving cars; (3) uses data stream management system to find the average speed, number of hard braking vehicles etc; (4) uses the parameters in the previous step to determine if a road section contains queue end. The prediction accuracy is evaluated under different setups.</p> <p>I had some questions regarding why queue end should be detected in the way that the paper is proposing. On the second thought, it makes sense. Finding queue end from arbitrary segmented road sections seems very inaccurate and sensitive to thresholding values. It seems tailgating warnings should be done in an individual car basis. For an individual car, one is more interested in knowing the car is approaching something fast. However, the proposed method will find the early signs (road sections) of decelerating speeds.</p> <p>The paper describes an interesting problem and the pieces needed to solve the problem. It would be a nice workshop paper to trigger further discussion.</p> <p>Some comments:</p> <ol style="list-style-type: none"> 1. "Characteristics of data stream mining algorithms are, for example, the handling of a limited memory size and the ability to detect concept drifts in the data to react in an appropriate way. " - not a sentence. 2. "In which traffic situations deliver the methods satisfactory results?" - not a correct sentence.

Example Review

Bibliography & References

- Journals, conference proceedings, thesis, books:
 - Authors, title, name of conference/journal, page numbers, year, and URL if available online
 - Not only URL, if paper was published in journal or conference!
- important information to estimate the quality of the cited article.
- Citation styles: APA, MLA, Chicago/Turabian etc. See also BibTex bibliography styles.

Examples

[14] J. Doe: Databases are great. *Wald&Wiesen Zeitschrift*, Vol. 1, No. 1, S. 3-5, 1972.

[JaKo84] M. Jarke, J. Koch: Query Optimization in Database Systems. *ACM Computing Surveys*, Vol. 16, No. 2, S. 111-152, 1984.
<http://dx.doi.org/10.1145/356924.356928>

- For websites, which are not blog posts or articles, but, e.g., software, initiatives etc.
→ use footnotes!

Bibliography

- [Aggarwal, 2013] Aggarwal, C. C., editor (2013). Managing and Mining Sensor Data. Springer Science & Business Media.
- [Babcock et al., 2002] Babcock, B., Babu, S., Datar, M., Motwani, R., and Widom, J. (2002). Models and Issues in Data Stream Systems. In Popa, L., editor, Proc. 21st ACM Symposium on Principles of Database Systems (PODS), pages 1–16, Madison, Wisconsin.
- [Babu & Widom, 2001] Babu, S. and Widom, J. (2001). Continuous Queries Over Data Streams. SIGMOD Record, 30(3):109–119.
- [Booth et al., 2003] Booth, W. C., Colomb, G. G. & Williams, J. M. (2003). The craft of research (2nd ed.). The University of Chicago Press.
- [Booth et al., 2008] Booth, W. C., Colomb, G. G. & Williams, J. M. (2008). The craft of research (3rd ed.). The University of Chicago Press.
- [Davies, 2011] Davies, M. (2011). Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? Higher Education, 62(3), 279-301.
- [Devlin, 2005] Devlin, M. (2005). Avoiding plagiarism and cheating: A guide for students at Swinburne University of Technology (2nd ed.). Swinburne University of Technology.
https://www.swinburne.edu.au/media/swinburneeduau/current-students/docs/pdf/plagiarism_guide.pdf

Bibliography

- [Geisler, 2013] Geisler, S. (2013). Data Stream Management Systems. In Kolaitis, P. G., Lenzerini, M., and Schweikardt, N., editors, Data Exchange, Integration, and Streams, volume 5 of Dagstuhl Follow-Ups. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany.
- [Geisler et al., 2016] Geisler, S. (2016). A systematic evaluation approach for data stream-based applications (Doctoral dissertation, Universitätsbibliothek der RWTH Aachen).
- [Gibaldi, 2009] Gibaldi, J. (2009). MLA handbook for writers of research papers (7th ed.). Modern Language Association of America.
- [Golab & Özsu, 2010] Golab, L. and Özsu, M. T. (2010). Data Stream Management. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.
- [Stonebraker, 2005] Stonebraker, M., Cetintemel, U., and Zdonik, S. B. (2005). The 8 requirements of real-time stream processing. SIGMOD Record, 34(4):42–47.
- [Tao et al, 2018] Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, 48, 157-169.